

## Correlated Random Effects for Hurdle Models applied to Claim Counts

*2011 CAS ANNUAL MEETING, Chicago*

Jean-Philippe Boucher

\*Quantact / Département de mathématiques  
Université du Québec à Montréal, Canada  
<http://www.quantact.uqam.ca/en/jpboucher.html>

## General Topics

- ▶ Count Data;
- ▶ Risk Classification;
- ▶ Panel Data (Longitudinal Data).

## Insurance

- ▶ *A priori* Ratemaking;
- ▶ *A posteriori* Ratemaking;
- ▶ Number of Claims.

- ▶ Old Ratemaking Technique;
- ▶ Introduced by Bailey and Simon(1960) and Bailey(1963);
- ▶ Has been shown the similarities with some statistical distributions.  
See:
  - ▶ Brown(1988);
  - ▶ Mildenhall(1999);
  - ▶ Holler and Sommer(1999) from the 9<sup>th</sup> exam of the Casualty Actuarial Society, that expose the link between the GLM and the minimum bias technique.

## Probability Distribution

$$\Pr[N_{i,t} = n_{i,t}] = \frac{e^{-\lambda_{i,t}} \lambda_{i,t}^{n_{i,t}}}{n_{i,t}!}, \lambda_{i,t} = \exp(x'_{i,t}\beta)$$

- ▶ Regressors are introduced by a score function;
- ▶ *Law of small numbers*;
- ▶ Exponential Family of Distributions : direct application of the GLM theory;
- ▶ Equidispersion property;

## Probability Distributions - NB2 and NB1

$$NB2 : \Pr[N_{i,t} = n_{i,t}] = \frac{\Gamma(n_{i,t} + \alpha^{-1})}{\Gamma(n_{i,t} + 1)\Gamma(\alpha^{-1})} \left(\frac{\lambda_{i,t}}{\alpha^{-1} + \lambda_{i,t}}\right)^{n_{i,t}} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_{i,t}}\right)^{\alpha^{-1}}$$

$$NB1 : \Pr[N_{i,t} = n_{i,t}] = \frac{\Gamma(n_{i,t} + \alpha^{-1}\lambda_{i,t})}{\Gamma(n_{i,t} + 1)\Gamma(\alpha^{-1}\lambda_{i,t})} (1 + \alpha)^{-\lambda_{i,t}/\alpha} (1 + \alpha^{-1})^{-n_{i,t}}$$

- ▶ Obtained by adding an heterogeneity term  $\theta$  to the mean parameter of the Poisson distribution, when  $\theta$  follows a gamma distribution;
- ▶ NB2 :  $Var[N_{i,t}] = \lambda_{i,t} + \alpha\lambda_{i,t}^2 > E[N_{i,t}]$ ;
- ▶ NB1 :  $Var[N_{i,t}] = \lambda_{i,t} + \alpha\lambda_{i,t} > E[N_{i,t}]$ ;
- ▶ Poisson distribution is the limiting case of NB distributions when  $\alpha \rightarrow 0$ .

## Ideas

- ▶ Classic Poisson and Negative Binomial distributions suppose independence between all the contracts of the same insured;
- ▶ There are advantages of using the information on each policyholder along time for modeling the number of claims;
- ▶ Allowing for time dependence between observations are closer to the data generating process that one can find in practice;
- ▶ Future premiums given the past observations can be calculated (credibility theory).

## Panel Data

An insurance policy  $i$  is observed over  $T$  consecutive years, where the vector of random variables  $(N_{i,1}, \dots, N_{i,T})$  is the random counts to be modeled.

In this setting, a dependence between all the contracts of the same insured can be incorporated.

For non-normal distributions, and more precisely for count distributions, panel data modeling admits 3 possibilities:

- ▶ Marginal approach;
- ▶ Conditional approach with random effects;
- ▶ Transition models.

First order condition for Poisson distributions can be expressed as (GLM):

$$\sum_{i=1}^n \frac{(n_i - \mu_i)}{g'(\mu_i) \text{Var}[N_i]} \mathbf{x}_i = \mathbf{0}$$

The GEE approach for correlated data generalizes the first order condition by using all the contracts of the same insured. First order condition can then be expressed as:

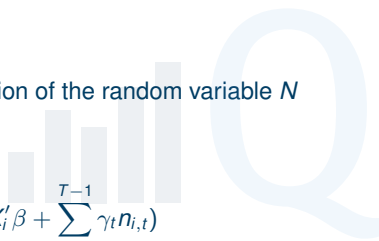
$$\sum_{i=1}^n \left( \frac{\partial}{\partial \beta} E[\mathbf{N}_i] \right)^T \mathbf{V}_i^{-1} (\mathbf{n}_i - E[\mathbf{N}_i]) = \mathbf{0}$$

$$\text{with: } \mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i(\alpha) \mathbf{A}_i^{1/2}$$

where  $\mathbf{A}_i$  is the variance-covariance matrix of the  $N_{i,t}$ s with serial independence, and  $\mathbf{R}_i$  is the "working" correlation matrix.



The idea of the model is to include past realization of the random variable  $N$  into, for example the regressors of the model,



$$E[N_{i,T} | n_{i,1}, \dots, n_{i,T-1}] = g(X_i' \beta + \sum_{t=1}^{T-1} \gamma_t n_{i,t})$$

If we suppose that  $m$  past realizations must be used in the modeling, we must work with :

$$\Pr[N_{i,1}, \dots, N_{i,T}] = \Pr[N_{i,1}, \dots, N_{i,m}] \prod_{j=m+1}^T \Pr[N_{i,j} | n_{i,j-1}, \dots, n_{i,j-m}]$$

$\Rightarrow \Pr[N_{i,1}, \dots, N_{i,m}]$  is evaluated with difficulty...

## Conditional Approach: Random Effects

- ▶ Missing of some important classification variables (swiftness of reflexes, aggressiveness behind the wheel, etc.) in the classification;
- ▶ Hidden features captured by an individual random heterogeneity term  $\theta_i$ ;
- ▶ Given  $\theta_i$ , the annual claim numbers  $N_{i,1}, N_{i,2}, \dots, N_{i,T}$  are independent.
- ▶ The joint probability function of  $N_{i,1}, \dots, N_{i,T}$  is given by

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

- ▶ Models depend on the choices of the conditional distribution of the  $N_{i,t}$  and the distribution of  $\theta_j$ .

## Multivariate Negative Binomial Distribution (MVNB)

When  $N_{i,t}$  is conditionally distributed as a Poisson distribution with random effects following a gamma distribution:

$$\Pr[N_{i,1} = n_{i,1}, \dots, N_{i,T} = n_{i,T}] = \left[ \prod_{t=1}^T \frac{(\lambda_{i,t})^{n_{i,t}}}{n_{i,t}!} \right] \frac{\Gamma(\sum_{t=1}^T n_{i,t} + 1/\alpha)}{\Gamma(1/\alpha)} \left( \frac{1/\alpha}{\sum_{t=1}^T \lambda_{i,t} + 1/\alpha} \right)^{1/\alpha} \times \left( \sum_{t=1}^T \lambda_{i,t} + 1/\alpha \right)^{-\sum_{t=1}^T n_{i,t}}.$$

## Negative Binomial-Beta distribution (NB-Beta)

When  $N_{i,t}$  is conditionally distributed as a NB1 distribution with random effects following a beta distribution:

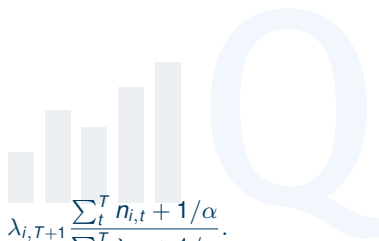
$$\Pr[N_{i,1} = n_{i,1}, \dots, N_{i,T} = n_{i,T}] = \frac{\Gamma(a+b)\Gamma(a + \sum_t \lambda_{i,t})\Gamma(b + \sum_t n_{i,t})}{\Gamma(a)\Gamma(b)\Gamma(a+b + \sum_t \lambda_{i,t} + \sum_t n_{i,t})} \prod_t^T \frac{\Gamma(\lambda_{i,t} + n_{i,t})}{\Gamma(\lambda_{i,t})\Gamma(n_{i,t} + 1)}$$

- ▶ Predictive distributions of panel data with random effects involve Bayesian analysis;
- ▶ At each insured period, the random effects can be updated for past claim experience, revealing some insured-specific informations.

$$\Pr[N_{i,T+1} = n_{i,T+1} | n_{i,1}, \dots, n_{i,T}] = \int \Pr(n_{i,T+1} | \theta_i) g(\theta_i | n_{i,1}, \dots, n_{i,T}) d\theta_i,$$

where  $g(\theta_i | n_{i,1}, \dots, n_{i,T})$  is the *a posteriori* distribution of the random effects  $\theta_i$ .

## Predictive Premiums



$$(MVNB) : E[N_{i,T+1} | N_{i,1}, \dots, N_{i,T}] = \lambda_{i,T+1} \frac{\sum_t^T n_{i,t} + 1/\alpha}{\sum_t^T \lambda_{i,t} + 1/\alpha}.$$

$$(NB - Beta) : E[N_{i,T+1} | N_{i,1}, \dots, N_{i,T}] = \lambda_{i,T+1} \frac{\sum_t^T n_{i,t} + b}{\sum_t^T \lambda_{i,t} + a - 1},$$

## Motivation

Suppose  $N$ , the number of claims as the product of:

- ▶ An indicator variable  $J$  (equal to 1 if at least 1 claim);
- ▶ A counting variable  $K \geq 1$  (giving the number of claims when at least 1 claim has been filed).

## Distribution

$$\Pr[N = n] = \Pr[JK = n] = \begin{cases} \Pr[J = 0] & \text{for } n = 0 \\ \Pr[J = 1] \Pr[K = n] & \text{for } n = 1, 2, \dots \end{cases}$$

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

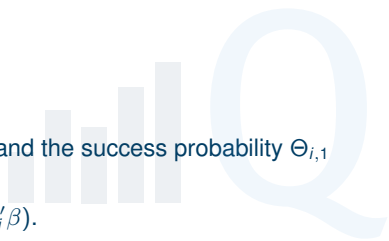
Let  $N_{i,1}, N_{i,2}, \dots, N_{i,T}$  be the number of claims reported by policyholder  $i$  over period 1 to  $T$ .

The joint distribution of  $N_{i,1}, \dots, N_{i,T}$  is given by

$$\begin{aligned} & \Pr[N_{i,1} = n_{i,1}, \dots, N_{i,T} = n_{i,T}] \\ = & \int \int \Pr[N_{i,1} = n_{i,1}, \dots, N_{i,T} = n_{i,T} | \Theta_{i1} = \theta_{i1}, \Theta_{i2} = \theta_{i2}] g(\theta_{i,1}, \theta_{i,2}) d\theta_{i,1} d\theta_{i,2} \\ = & \int \int \prod_{t=1}^T \left( (\Pr[J_{it} = 0 | \Theta_{i1} = \theta_{i1}])^{I(n_{i,t}=0)} \right. \\ & \left. \times (\Pr[J_{it} = 1 | \Theta_{i1} = \theta_{i1}] \Pr[K_{it} = n_{it} | \Theta_{i2} = \theta_{i2}])^{I(n_{i,t}>0)} \right) g(\theta_{i,1}, \theta_{i,2}) d\theta_{i,1} d\theta_{i,2}, \end{aligned}$$

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



- 
- ▶  $J_{it}$  is Bernoulli distributed with mean  $\Theta_{i,1}$ , and the success probability  $\Theta_{i,1}$  is Beta( $a_i, b$ ) distributed.
  - ▶ Covariates enter the model via  $a_i = \exp(x_i' \beta)$ .
  - ▶ 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  $\alpha$ .

Consequently, the joint distribution of all contracts of the same insured is expressed as:

$$\Pr[N_{i,1} = n_{i,1}, \dots, N_{i,T} = n_{i,T}] = \int \int \prod_{t=1}^T \left( \theta_{i,1}^{l(n_{i,t}=0)} (1 - \theta_{i,1})^{l(n_{i,t}>0)} \left( e^{-\gamma_i \theta_{i,2}} \frac{(\gamma_i \theta_{i,2})^{n_{i,t}-1}}{(n_{i,t} - 1)!} \right)^{l(n_{i,t}>0)} \right) g(\theta_{i,1}, \theta_{i,2}) d\theta_{i,1} d\theta_{i,2}$$

The two random effects  $(\Theta_{i1}, \Theta_{i2})$  are likely to be correlated because the same omitted characteristics affect each process.

We use a Gaussian copula to represent the joint distribution of the random effects.


More precisely, we assume that  $g$  can be written as


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

with

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

where  $\Phi$  is the standard Normal distribution function and the marginal density functions  $g_1$  and  $g_2$  are Beta and Gamma, respectively,

- 
- ▶ If the correlation parameter  $\rho$  is equal to 1 then  $\Theta_{i_1}$  and  $\Theta_{i_2}$  are perfectly positively dependent. In this case, the Gaussian copula reduces to the Fréchet-Hoeffding upper bound.
  - ▶ If the correlation parameter  $\rho$  is set to 0 then  $\Theta_{i_1}$  and  $\Theta_{i_2}$  are mutually independent.



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

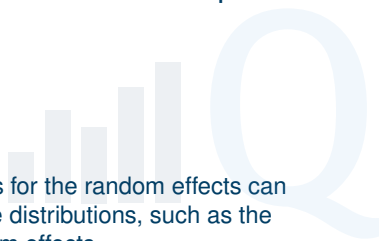
$$\begin{aligned}
 \text{Cov}[N_{i,t}, N_{i,t+j}] &= \text{Var}[\theta_{i,1}] + 2\gamma_i \left( E[\theta_{i,1}^2\theta_{i,2}] - E[\theta_{i,1}\theta_{i,2}]E[\theta_{i,1}] \right) \\
 &\quad + \gamma_i^2 \left( E[\theta_{i,1}^2\theta_{i,2}^2] - E[\theta_{i,1}\theta_{i,2}]^2 \right).
 \end{aligned}$$

No closed form expression is available for the likelihood of the model.  
 Here, we resort to the NL MIXED procedure from the SAS System.

Formally, the predictive distribution is obtained from

$$\begin{aligned}
 & \Pr[N_{i,T+1} = n_{i,T+1} | N_{i,1} = n_{i,1}, \dots, N_{i,T} = n_{i,T}] \\
 = & \frac{\Pr[N_{i,1} = n_{i,1}, \dots, N_{i,T+1} = n_{i,T+1}]}{\Pr[N_{i,1} = n_{i,1}, \dots, N_{i,T} = n_{i,T}]} \\
 = & \int \int \Pr[N_{i,T+1} = n_{i,T+1} | \theta_{i,1}, \theta_{i,2}] \\
 & \times \left( \frac{\Pr[N_{i,1} = n_{i,1}, \dots, N_{i,T} = n_{i,T} | \theta_{i,1}, \theta_{i,2}] g(\theta_{i,1}, \theta_{i,2})}{\int \int \Pr[N_{i,1} = n_{i,1}, \dots, N_{i,T} = n_{i,T} | \theta_{i,1}, \theta_{i,2}] g(\theta_{i,1}, \theta_{i,2}) d\theta_{i,1} d\theta_{i,2}} \right) d\theta_{i,1} d\theta_{i,2} \\
 = & \int \int \Pr[N_{i,T+1} = n_{i,T+1} | \theta_{i,1}, \theta_{i,2}] g(\theta_{i,1}, \theta_{i,2} | n_{i,1}, \dots, n_{i,T}) d\theta_{i,1} d\theta_{i,2},
 \end{aligned}$$

where  $g(\theta_{i,1}, \theta_{i,2} | n_{i,1}, \dots, n_{i,T})$  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.
  - ▶ For other models, such as the correlated random effects hurdle models studied here, these distributions cannot be evaluated analytically.

We use Markov chain Monte Carlo (MCMC) simulations to compute posterior and predictive distributions.

- ▶ 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.
- ▶ 15,179 policyholders who remained with the company for seven complete periods.

Variable	Description
Sex	equals 1 for women and 0 for men
Years with the company (3 – 5) (> 5)	equals 1 if with the company between 3 and 5 years equals 1 if 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 or equal to 5500 cc

Table: Exogenous variables



## Comparison of Models

Models	Number of parameters	Loglikelihood	AIC	BIC
MVNB	7	-26,702.98	53,419.96	53,486.98
Hurdle (ind.)	10	-26,688.70	53,397.40	53,493.14
Hurdle (Gauss.)	11	-26,662.47	53,346.94	53,452.25
Hurdle (F.-H.)	10	-26,663.28	53,346.56	53,442.30

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

Models	Good Profile		Medium Profile		Bad Profile	
	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


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

Models	$T - T_0$	A priori	Sum of claims					
			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: Mean of the Predictive Distribution

Models	$T - T_0$	A priori	Sum of claims					
			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

Table: Variance of the Predictive Distribution

- 
- ▶ *A priori* and predictive premiums are quite close to the MVNB ones;
  - ▶ The corresponding variances greatly differ;
  - ▶ Dependence should not be ignored if one is interested in studying the variance and not only the mean of the the number of claims;
  - ▶ Ignoring dependence would lead to underestimation of the variance and therefore, to inefficient pricing.
  - ▶ Premium calculation is usually loaded by a factor which may depend on the estimated variance.