Flexible (panel) regression models for bivariate count-continuous data with an insurance application

Yang Lu

Flexible (Panel) Regression Model for Bivariate Count-Continuous Data with Insurance Application

Yang Lu *

April 6, 2019

Abstract

We propose a flexible regression model that is suitable for mixed count-continuous panel data. The model is based on a compound Poisson representation of the continuous variable, with bivariate random effect following a polynomial expansion based joint density. Besides the distributional flexibility it offers, the model allows for closed form forecast updating formulas. This property is especially important for Insurance applications, in which the future individual insurance premium should be regularly updated according to one's own past claim history. An application to vehicle insurance claims is provided.

Key words: mixed data, polynomial expansion, random effect, sequential forecasting/pricing.

1 Introduction

This paper introduces a flexible regression model for mixed count-continuous panel data, with flexible correlated bivariate random effects (or unobserved heterogeneity). Such a model is useful for a wide range of applications, such as:

- number/cost of visits to physicians in health economics [see e.g. Duan et al. (1983)],
- number/amount of shopping in marketing [see e.g. Paull (1978)],
- number/cost of trips in tourism forecasting [see e.g. Englin and Shonkwiler (1995)].

Our paper will be focused on an area which necessitates sequential forecasting, that is the analysis of longitudinal insurance claims. Indeed, in many other applications, only cross-sectional data are used and the scientific interest is often limited to explaining the count and continuous variables by observable individual characteristics (or covariates). In Insurance, regulations require the yearly

*University of Paris-13, Department of Economics (CEPN). Correspondence to: luyang000278@gmail.com

insurance premium to be updated regularly, in order to take into account each policyholder's own past history.[1] In this context, the bivariate count-continuous panel data arises since data are typically aggregated annually and are available in terms of annual claim count and the corresponding total claim cost. While the traditional pricing method has focused on the counts only [see e.g. Dionne and Vanasse (1989)], there are several benefits to consider both total cost and count, as well as their interdependence. First, it increases the forecast precision of the total cost, which is the ultimate variable of interest for an insurance company. Second, when only the count variable is taken into account, the pricing system penalizes unfairly customers with small previous claims compared to those who file large claims. This induces an efficiency loss for the insurance market [see Einav et al. (2010)], and gives individuals the incentive to (strategically) not hide minor accidents, if the expected reimbursement is less than the future premium increase[2].

Therefore, of vital importance is a model that *i*) takes into account both observable covariates, as well as random effects; *ii*) captures, in a flexible way, the dependence between the count and continuous (size) variable; *iii*) allows for tractable Bayesian forecasting formulas. Our paper proposes a model unifying these properties.

More precisely, we introduce a bivariate regression model with correlated random effects. They induce not only contemporaneous dependence between the two components, but also serial correlation, which serves as the basis for the forecasting and pricing of future claims. Our joint distribution of the heterogeneities is semi-parametric, on the contrary to the standard gamma heterogeneity assumption. This leads to flexible (marginal and joint) distributions for the observable bivariate count-continuous variable.

The proposed joint density function of the random effects takes the form of polynomial expansion with respect to a reference density and can approximate any distribution function arbitrarily well. It has a similar spirit as the model of Gurmu et al. (1999); Gurmu and Elder (2012)] for cross-sectional count data. Our contribution to this latter literature is twofold. First, we propose a simple, and formal justification of this method. This also leads us to a comparison with another stream of literature, which employs a different polynomial expansion method to approximate the conditional transition density of the asset price dynamics. In particular, it is shown that our approach requires a much weaker assumption, and ensures positivity of the density function. Second, we propose an equivalent parameterization of the polynomial expansion-based density function. This allows us to derive significantly simpler expressions for the likelihood function and, more importantly, the forecasting formulas. This should further increase the appeal of this semi-parametric class of models in empirical studies.

[1] Such pricing method is called bonus-malus.
[2] This is called hunger for bonus phenomenon.

The rest of the paper is organized as follows. Section 2 presents the model. Section 3 provides an empirical illustration using a database of vehicle insurance claims. Section 4 concludes. Proofs are gathered in Appendix.

2 The model

2.1 The general setting

For each individual i, we denote by $N_{i,t}, t = 1, ..., T$ a sequence of count variables and $Y_{i,t}$ taking values in $[0, \infty[$, as well as a set of covariates X_i. For expository purpose we assume the latter to be time-invariant and the extension to the time-varying case is straightforward. In insurance applications, the count variable corresponds to the number of claims during period t, and the continuous variable corresponds to the total cost of these $N_{i,t}$ claims, respectively. As a convention, $Y_{i,t}$ is 0 if $N_{i,t}$ equals zero. We assume that conditional on (static) unobserved heterogeneities $U_{i,1}, U_{i,2}$ and observable covariates X_i, the count variable $N_{i,t}$ is Poisson distributed with parameter $\lambda_i U_{1,i}$, and $Y_{i,t}$ is Gamma distributed, with shape parameter $\delta N_{i,t}$, and rate parameter $\beta_i U_{i,2}$, where:

$$\lambda_i = \exp(d_1' X_i), \qquad \beta_i = \exp(d_2' X_i). \tag{1}$$

From now on, for expository purpose, we will assume that $(X_i, (N_{i,t})_t, (Y_{i,t})_t)$ are i.i.d. across individuals and will thus omit the index i.

Thus we have: $\mathbb{E}[Y_t | U_1, U_2, N_t] = \frac{\delta N_t}{\beta U_2}$, which is proportional to N_t. Moreover, variable Y_t has the following (conditional) compound representation:

$$Y_t = \sum_{j=1}^{N_t} Z_{j,t}, \tag{2}$$

where $Z_{j,t}, j = 1, ..., N_t$ are the (unobservable) costs of individual claims reported during period t. Conditionally on U_1, U_2, these individual claim cost follow the gamma distribution[3] with shape parameter δ and rate parameter βU_2.

In this model, for a given individual, the values of U_1, U_2 are unknown, but do not change of time. This induces serial correlation between (N_t, Y_t) at different dates t, and allows the insurance company to better predict the future insurance claim cost according to past ones.

Thus the dependence between N_t and Y_t is the combination of two effects:

1. conditionally on random effects U_1, U_2, the responses variables N_t and Y_t are dependent

[3] In particular, if U_2 follows gamma marginal distribution, then the distribution of each individual cost is called generalized beta of the second kind (GB2) [see e.g. McDonald (1984); Cummins et al. (1990); Gouriéroux (1999); Frangos and Vrontos (2001)].

due to the representation (2);

2. random effects U_1 and U_2 are dependent.

Moreover, the dependence between U_1 and $1/U_2$ can be either positive or negative. In the first case, individuals with a higher claim frequency tend to have more severe claims. This can be explained by the fact that both the count and total cost reflect, to some extent, the skill of the driver. Alternatively, this association can also be negative. One potential explanation is the so-called "hunger for bonus" phenomenon. Indeed, some policyholders strategically choose not to report claims whose cost is smaller than the cumulated future premium increase [see e.g. Lemaire (1995)]. This would lead to a higher (resp. lower) claim frequency, but a lower (resp. higher) cost per claim for honest individuals (resp. arbitrageurs who do not report small claims). The evidence of such negative dependence has recently been found by Garrido et al. (2016).

As a comparison, the existing literature on mixed discrete-continuous regression usually considers only either of these two types of dependence. For instance, Fitzmaurice and Laird (1995); Yang et al. (2007); de Leon and Wu (2011) specify the joint density between a discrete and a continuous variable via the marginal distribution of the discrete variable and the conditional distribution of the continuous variable given the discrete one. This corresponds to the first type of dependence mentioned above. Due to the lack of random effects, these models are not suitable when extended to a panel context, as serial correlation cannot be introduced between observations of different periods.

On the other hand, Catalano and Ryan (1992); Sammel et al. (1997); Gueorguieva and Agresti (2001) introduce dependence via random effects only. However, these papers either concern binary discrete variable and not count variable, or involve rather non tractable, simulation-based algorithm for estimation and forecasting [see e.g. Sammel et al. (1997); Dunson (2000)].

Let us now propose a specification for the joint density of (U_1, U_2), before studying its implications for the estimation and forecasting.

2.2 The model for the random effects

We assume that the joint density function of (U_1, U_2) is:

$$g(u_1, u_2) = \frac{1}{M} e^{-c_1 u_1 - c_2 u_2} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} \Big[\sum_{j=0}^{J} \sum_{k=0}^{J} b_{j,k} u_1^j u_2^k \Big]^2, \tag{3}$$

where:

- coefficients $(b_{j,k})$ are real numbers, and $b_{0,0} = 1$ for identification purpose.

- the normalization constant is equal to:

$$\begin{aligned} M &= \sum_{j_1,k_1,j_2,k_2=0}^{J} b_{j_1,k_1} b_{j_2,k_2} \frac{\Gamma(j_1+j_2+\alpha_1)\Gamma(k_1+k_2+\alpha_2)}{c_1^{j_1+j_2+\alpha_1} c_2^{k_1+k_2+\alpha_2}} \\ &= \sum_{j=0}^{2J}\sum_{k=0}^{2J} \frac{\Gamma(j+\alpha_1)\Gamma(k+\alpha_2)}{c_1^{j+\alpha_1} c_2^{k+\alpha_2}} \Pi_{j,k} \\ &= W_1(0)'\Pi W_2(0), \end{aligned} \tag{4}$$

where matrix $(\Pi_{j,k})_{0\le j,k\le 2J}$ is defined by:

$$\Pi_{j,k} = \sum_{\substack{j_1+j_2=j \\ 0\le j_1,j_2\le J}} \sum_{\substack{k_1+k_2=k \\ 0\le k_1,k_2\le J}} b_{j_1,k_1} b_{j_2,k_2}, \tag{5}$$

and vector functions $W_1 = (W_{1,j})_{0\le j\le 2J}$, $W_2 = (W_{2,k})_{0\le k\le 2J}$ are given by:

$$W_{1,j}(s_1) = \frac{\Gamma(j+\alpha_1)}{(c_1+s_1)^{j+\alpha_1}}, \qquad W_{2,k}(s_2) = \frac{\Gamma(k+\alpha_2)}{(c_2+s_2)^{k+\alpha_2}}, \qquad \forall 0 \le j,k \le 2J, s_1, s_2 \ge 0. \tag{6}$$

Thus $g(u_1,u_2)$ is a linear combination[4] of gamma product densities:

$$f_J(u_1,u_2) = \frac{1}{M}\sum_{j=0}^{2J}\sum_{k=0}^{2J} e^{-c_1u_1}u_1^{\alpha_1+j-1} e^{-c_2u_2}u_2^{\alpha_2+k-1}\Pi_{j,k} = e^{-c_1u_1}u_1^{\alpha_1-1} e^{-c_2u_2}u_2^{\alpha_2-1}\frac{X(u_1)'\Pi X(u_2)}{W_1(0)'\Pi W_2(0)},$$

where vector functions $X(u) := (1, u, u^2, ..., u^{2J})'$ for each u.

The resulting model is semi-parametric in the sense that the distribution of the unobserved heterogeneity is flexible and can approximate any distributions, whereas the regressors λ and β involve a finite dimensional parameter d. Similar polynomial expansions have been previously proposed by Gallant and Nychka (1987); Gurmu et al. (1999); Bierens (2008). Its background is the orthonormal projection of $\sqrt{g_0}$ in an appropriate $\mathcal{L}^2$ space, using a polynomial basis, whereas the square ensures the positivity of the density [see e.g. Gallant and Nychka (1987)]. However, unlike the conventional approach which explicitly introduces these orthogonal polynomials, in this paper we use, without loss of generality, the canonical polynomials. This is due to the fact that both the orthogonal polynomials and the canonical polynomials are basis of the set of all polynomials, but the latter is much easier to deal with when used to the square (since the product of two canonical polynomials is still a canonical polynomials). This choice has the advantage of greatly simplify the computation of the likelihood function in the full model.

From the representation of combination of gamma densities, we derive the marginal distribu-

[4]But with possibly negative weights, since the entries of matrix B need not to be necessarily nonnegative.

tions of U_1 and U_2:

$$g_1(u_1) = e^{-c_1 u_1} u_1^{\alpha_1 - 1} \frac{X(u_1)' \Pi W_2(0)}{W_1(0)' \Pi W_2(0)}, \tag{7}$$

$$g_2(u_2) = e^{-c_2 u_2} u_2^{\alpha_2 - 1} \frac{W_1(0)' \Pi X(u_2)}{W_1(0)' \Pi W_2(0)}, \tag{8}$$

which are also linear combinations of gamma densities. Thus we have the following property:

Proposition 1. *The two components U_1 and U_2 are independent if and only if all the coefficients b_{jk} are separable, that is, if for all j and k, we have: $b_{jk} = b_{j,0} b_{0,k}$.*

In other words, our model can accommodate for the special case where the two random effects are independent.

Proof. See Appendix. □

Similarly, the joint Laplace transform has also a closed form expression:

$$\mathcal{L}(s_1, s_2) = \iint_0^\infty e^{-u_1 s_1 - u_2 s_2} g(u_1, u_2) \mathrm{d}u_1 \mathrm{d}u_2 = \frac{W_1(s_1)' \Pi W_2(s_2)}{W_1(0)' \Pi W_2(0)}, \qquad \forall s_1, s_2 \geq 0.$$

The density of (N_t, Y_t). Conditional on U_1, U_2, the joint distribution of (N_t, Y_t) has two components:

- The first is the point mass at $(0, 0)$, with probability:

$$\mathbb{P}[N_t = Y_t = 0 \mid U_1, U_2] = e^{-\lambda U_1}.$$

- One continuous component with respect to the measure $\mu_1 \otimes \mu_2$, where μ_1 is the Lebesgue measure on positive integers, and μ_2 is the Lebesgue measure on positive real numbers. This component has a density:

$$f(n, y \mid U_1, U_2) = \frac{\lambda^n U_1^n e^{-\lambda U_1}}{n!} \frac{\beta^{\delta n} U_2^{\delta n} y^{\delta n - 1} e^{-\beta U_2 y}}{\Gamma(\delta n)}, \qquad \forall n \geq 1, y > 0. \tag{9}$$

By integrating out the distribution of (U_1, U_2), we get the unconditional distribution of (N_1, Y_1). The elementary probability of the degenerate component is:

$$\mathbb{P}[N_t = Y_t = 0] = \mathbb{E}[e^{-\lambda U_1}] = \frac{1}{M} \sum_{j=0}^{2J} \sum_{k=0}^{2J} \frac{\Gamma(j + \alpha_1)\Gamma(k + \alpha_2)}{(c_1 + \lambda_t)^{j + \alpha_1} c_2^{k + \alpha_2}} \Pi_{j,k} = \frac{V_1'(0, \lambda) \Pi V_2(0, 0)}{W_1(0)' \Pi W_2(0)}, \tag{10}$$

where vector functions V_1, V_2 are defined by:

$$V_{1,j}(n, s_1) = \frac{\Gamma(j + \alpha_1 + n)}{(c_1 + s_1)^{j+\alpha_1+n}}, \qquad V_{2,k}(n, s_2) = \frac{\Gamma(k + \alpha_2 + \delta n)}{(c_2 + s_2)^{k+\alpha_2+\delta n}}, \qquad \forall n, s_1, s_2, j, k.$$

Similarly, the density of the continuous component is:

$$\begin{aligned} f(n, y) &= \frac{\lambda^n \beta^{\delta n} y^{\delta n - 1}}{n!\Gamma(\delta n)} \mathbb{E}[e^{-\lambda U_1 - \beta U_2 y} U_1^n U_2^{\delta n}] \\ &= \frac{\lambda^n \beta^{\delta n} y^{\delta n - 1}}{n!\Gamma(\delta n)} \frac{1}{M} \sum_{j=0}^{2J} \sum_{k=0}^{2J} \frac{\Gamma(j + \alpha_1 + n)}{(c_1 + \lambda_1)^{j+\alpha_1+n}} \frac{\Gamma(k + \alpha_2 + \delta n)}{(c_2 + \beta_1 y)^{k+\alpha_2+\delta n}} \Pi_{j,k}, \\ &= \frac{\lambda^n \beta^{\delta n} y^{\delta n - 1}}{n!\Gamma(\delta n)} \frac{V_1(n, \lambda)' \Pi V_2(n, \beta y)}{W_1(0)' \Pi W_2(0)} \qquad \forall y > 0, n \in \mathbb{N} \setminus \{0\}. \end{aligned} \tag{11}$$

Accounting for the heavy-tail of the loss distribution. One of the desirable properties of loss distributions in Insurance is their heavy-tail. Let us first prove that our model allows for heavy-tailed loss distribution:

Proposition 2. *The distribution of the total cost Y_t has a heavy tail. In particular, it has a finite mean if and only if:*

$$\alpha_2 > 1. \tag{12}$$

Proof. See Appendix. □

From now on we assume this condition to be satisfied, since otherwise the risk is not insurable.

2.3 Estimation and forecasting algorithm

The log-likelihood function. Let us assume that for each individual i, we observe $(N_{i,t})_{1 \le t \le T}$, $(Y_{i,t})_{1 \le t \le T}$ and covariates X_i. These pairs $(N_{i,t}, Y_{i,t})$ are independent conditional on (U_1, U_2). Then the log-likelihood function of the model is:

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{I} \log \mathbb{P}\Big[N_{i,t} = n_{i,1}, Y_{i,t} = y_{i,t}, t = 1...T \mid X_i\Big] \\ &\propto \sum_{i=1}^{I} \left[\log \mathbb{E}\Big[e^{-\Lambda_{1iT} U_1} U_1^{\overline{N}_{iT}} U_2^{\delta \overline{N}_{iT}} e^{-\Lambda_{2iT} U_2}\Big] + \sum_{t, N_{i,t} \neq 0} \delta N_{i,t} \log y_{i,t} \beta_{i,t} - \log \Gamma(\delta N_{i,t}) \right] \\ &\propto \sum_{i=1}^{I} \left[\log \frac{V_1(\overline{N}_{iT}, \Lambda_{1iT})' \Pi V_2(\overline{N}_{iT}, \Lambda_{2iT})}{W_1(0)' \Pi W_2(0)} + \sum_{t, N_{i,t} \neq 0} \log \frac{\lambda_{i,t}^{N_{i,t}} \beta^{\delta N_{i,t}} y^{\delta N_{i,t}}}{\Gamma(\delta N_{i,t})} \right], \end{aligned}$$

where $\overline{N}_{iT}$ is the sum of counts, $\Lambda_{1T} := \sum_{t=1}^{T} \lambda = T\lambda$ is the cumulative intensity, and $\Lambda_{2T} := \sum_{t=1}^{T} \beta y_t$ is the normalized cumulative cost, and the set of parameters is equal to

$$\theta = (\alpha_1, \alpha_2, (b_{j,k})_{j,k}, \delta, c_1, c_2, d_1, d_2).$$

Forecasting formulas. Conditional on past claim history $(N_t), (Y_t), t = 1, ..., T$, the distribution of (U_1, U_2) is:

$$f_T(u_1, u_2 | (N_t), (Y_t), t = 1, ..., T) \propto f(u_1, u_2) \prod_{t=1}^{T} u_1^{N_t} e^{-\lambda u_1} \prod_{t=1}^{T} u_2^{\delta N_t} e^{-\beta u_2 y_t}$$

Thus the expected total cost at time $T+1$, given claim count and cost history is:

$$\begin{aligned}
&\mathbb{E}[Y_{T+1} \mid (N_t, Y_t), t = 1, ..., T] = \frac{\delta}{\beta_t} \lambda_t \mathbb{E}[U_1/U_2 \mid (N_t, Y_t), t = 1, ..., T] \\
&= \frac{\delta\lambda}{\beta} \frac{\mathbb{E}[U_1^{\overline{N}_T+1} U_2^{\delta\overline{N}_T-2} e^{-\Lambda_{1T}U_1 - \Lambda_{2T}U_2}]}{\mathbb{E}[U_1^{\overline{N}_T} U_2^{\delta\overline{N}_T-1} e^{-\Lambda_{1T}U_1 - \Lambda_{2T}U_2}]} \\
&= \frac{\delta\lambda}{\beta} \frac{\sum_{j,k=0}^{2J} \Pi_{j,k} \frac{\Gamma(j+\alpha_1+\overline{N}_T+1)}{(c_1+\Lambda_{1T})^{j+\alpha_1+\overline{N}_T+1}} \frac{\Gamma(k+\alpha_2+\delta\overline{N}_T-1)}{(c_2+\Lambda_{2T})^{k+\alpha_2+\delta\overline{N}_T-1}}}{\sum_{j,k=0}^{2J} \frac{\Gamma(j+\alpha_1+\overline{N}_{1T})}{(c_1+\Lambda_{1T})^{j+\alpha_1+\overline{N}_{1T}}} \frac{\Gamma(k+\alpha_2+\delta\overline{N}_T)}{(c_2+\Lambda_{2T})^{k+\alpha_2+\delta\overline{N}_T}} \Pi_{j,k}} \\
&= \frac{\delta\lambda}{\beta} \frac{V_1(\overline{N}_T + 1, \Lambda_{1T})' \Pi V_2(\overline{N}_T - 1/\delta, \Lambda_{2T})}{V_1(\overline{N}_T, \Lambda_{1T})' \Pi V_2(\overline{N}_T, \Lambda_{2T})}.
\end{aligned} \tag{13}$$

We can also rewrite this expectation as a weighted average, indeed from (13) we have:

$$\mathbb{E}[Y_{T+1} \mid (N_t, Y_t), t = 1, ..., T] = \frac{\delta\lambda}{\beta} \sum_{j,k=0}^{2J} \frac{(j + \alpha_1 + \overline{N_T})(c_2 + \Lambda_{2T})}{(c_1 + \Lambda_{1T})(k + \alpha_2 + \delta\overline{N}_T - 1)} \omega_{j,k} \tag{14}$$

with:

$$\omega_{j,k} = \frac{\Pi_{j,k} \frac{\Gamma(j+\alpha_1+\overline{N}_T)}{(c_1+\Lambda_{1T})^{j+\alpha_1+\overline{N}_T}} \frac{\Gamma(k+\alpha_2+\delta\overline{N}_T)}{(c_2+\Lambda_{2T})^{k+\alpha_2+\delta\overline{N}_T}}}{\sum_{j',k'=0}^{2J} \frac{\Gamma(j'+\alpha_1+\overline{N}_T)}{(c_1+\Lambda_{1T})^{j'+\alpha_1+\overline{N}_T}} \frac{\Gamma(k'+\alpha_2+\delta\overline{N}_T)}{(c_2+\Lambda_{2T})^{k'+\alpha_2+\delta\overline{N}_T}} \Pi_{j',k'}}.$$

In terms of insurance pricing, the term $\frac{(j+\alpha_1+\overline{N_T})(c_2+\Lambda_{2T})}{(c_1+\Lambda_{1T})(k+\alpha_2+\delta\overline{N}_T-1)}$ in (14) is the product of two terms:

- The first one $\frac{j+\alpha_1+\overline{N_T}}{c_1+\Lambda_{1T}}$ is the posterior expectation of U_1 conditionally on previous counts only, if U_1 was gamma distributed;
- The numerator of the second term $\frac{c_2+\Lambda_{2T}}{k+\alpha_2+\delta\overline{N}_T-1}$ is, up to an additive constant, linear combination of previous claim costs whereas the denominator is, up to an additive constant, the cumulative claim counts. Thus this second term measures, to some extent, the average cost per claim during the first T periods.

Note also that due to the dependence between U_1 and U_2, the posterior expected cost is generically not the product of expected future claim count and expected future average cost per claim[5], that is:

$$\mathbb{E}[N_{T+1} \mid (N_t, Y_t), t = 1, ..., T]\mathbb{E}[Y_{T+1}/N_{T+1} \mid (N_t, Y_t), t = 1, ..., T] \neq \mathbb{E}[Y_{T+1} \mid (N_t, Y_t), t = 1, ..., T].$$

3 Numerical application

In this section we fit our model to an insurance database, and analyze the potential bias in terms of pricing, when an independence assumption between U_1 and U_2 is imposed.

3.1 The data

Our data concerns an Australian vehicle insurance portfolio. It contains only one period of observation, hence in the following, the time index t will be omitted. It has initially been analysed by De Jong and Heller (2008) using generalized linear models and is available free of charge at the following website:

`www.businessandeconomics.mq.edu.au/our_departments/Applied_Finance_and_Actuarial_Studies/research/books/GLMsforInsuranceData/data_sets`

The database consists of 65324 (vehicle damage) policies, of which 4423 (6.8 percent) had at least one claim. For each policyholder, we observe the total claim amount (continuous valued, in 10000 \$) and the total claim count (integer valued), as well as the following covariates[6]:

- type of the vehicle (categorical, with 6 different values[7])
- vehicle's age (categorical, with 4 different values)
- gender (male or female)
- driver's area of residence (categorical, with 6 different values)
- driver's age (categorical, with 4 different values)

The following table summarizes the distribution of the count variable.

[5]Except when U_1, U_2 are independent. In this case, coefficients $b_{j,k} = b_j b_k$ are separable and as a consequence coefficients $\Pi_{j,k}, \omega_{j,k}$ are also separable.

[6]Although the value of the vehicle is also available, it has not been taken into account. This is in line with the existing literature [see e.g. Pinquet et al. (2001); Czado et al. (2012); De Jong and Heller (2008)] and is motivated by several reasons: *i*), the value of the vehicle is partially correlated with other covariates, in particular the type of the vehicle; *ii*) for some policies, the recorded vehicle value is recorded as zero, which suggests a lack of credibility of this variable. One simple alternative would be to divide the amount of the claim by the value of the vehicle to get the "normalized claim cost". However this alternative has not been chosen due to certain policies with a zero vehicle value.

[7]There are initially 13 different types of vehicles. Seven of which (bus, convertible, motorized caravan, minibus, coupe, panel van, and roadster) account for, in total, only 2200 policies, or 3.5 % of the sample. These policies are dropped for illustration purpose and we are left with five major types of vehicles (utility cars, sedan, trucks, station wagons, hardtop, as well as hatchbacks).

Number of claims	Number of policies with the given number of claims
0	60901
1	4149
2	255
3	17
4	2

Table 1: Number of policies according to the number of claims

The average number of claims per individual is approximately 0.073, whereas the variance of the number of claims is equal to 0.077. Thus the count variable is slightly over-dispersed.

Figure 1 plots the histogram of $Y|N > 0$, that is the distribution of total claim cost, for those who reported claims. We can remark that this distribution has both a heavy left tail, corresponding to a large proportion of minor-sized claims, as well as a heavy right tail, corresponding to extreme value claims.

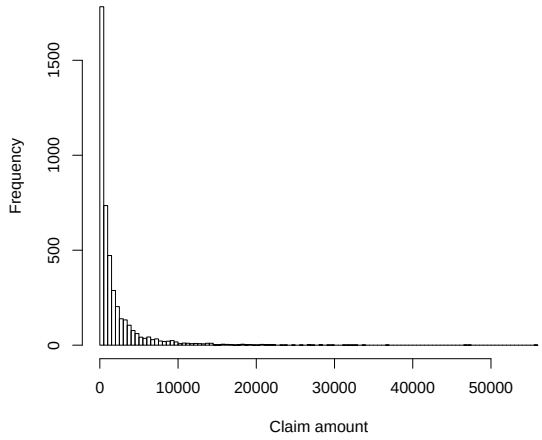

Figure 1: Histogram of total claims.

In average, the expectation of total cost of claims, given that there is at least one claim, is approximately 2000 dollars. In order to better illustrate the heavy right tail of the claim cost, let us provide the values of some quantiles of the distribution of the observed claim cost, among policyholders who actually reported claims.

Percentage	90%	95 %	97%	98 %	99%	100%
Quantile	0.49	0.80	1.06	1.28	1.70	5.59

Table 2: Some quantiles of the cost variable, in 10000 dollars.

We can see, for instance, that the 99% quantile of the cost variable is approximately two (resp.

four) times larger than the 95 % (resp. 90 %) quantile, whereas the biggest claim is another three times larger. The heavy tail of the distribution of cost implies that standard models based on gamma-distributed claim costs [see e.g. Czado et al. (2012)] may not be appropriate for this data.

The following table summarizes the available covariates. All covariates are discrete and transformed into binary variables[8]. For identification issue, the constant is *not* included as a covariate, and the means $\mathbb{E}[U_1]$, $\mathbb{E}[U_2]$ are not constrained to be, say, 1.

Variable	Description
X_1	1 if the policyholder is male
X_2	1 if the area of residence is "A"
X_3	1 if the area of residence is "B"
X_4	1 if the area of residence is "C"
X_5	1 if the area of residence is "D"
X_6	1 if the area of residence is "E"
X_7	1 if the driver's age is "1" (youngest)
X_8	1 if the driver's age is "2"
X_9	1 if the driver's age is "3"
X_{10}	1 if the driver's age is "4"
X_{11}	1 if the driver's age is "5"
X_{12}	1 if the vehicle's age is "1" (newest)
X_{13}	1 if the vehicle's age is "2"
X_{14}	1 if the vehicle's age is "3"
X_{15}	1 if the vehicle is a hatchback
X_{16}	1 if the vehicle is a sedan
X_{17}	1 if the vehicle is a hardtop
X_{18}	1 if the vehicle is a truck
X_{19}	1 if the vehicle is a station wagon

Table 3: Summary of the binary variables included in the regressors.

Thus in the log-likelihood function, both d_1 and d_2 are of dimension 19. In other words, constants are not included as regression parameter. This is because our specification of the unobserved heterogeneities does not restrict the expectations $\mathbb{E}[U_1], \mathbb{E}[U_2]$ to be constant.

3.2 Estimation

In this subsection we estimate the model for three different values of J:

- Model M0, with $J = 0$, that is $f(u_1, u_2) \propto e^{-c_1 u_1 - c_2 u_2} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1}$, which corresponds the density of independent gamma variables;

[8]For instance, if one categorical variable can take 4 different values, then it is transformed into three linearly independent binary variables.

- Model M1, with $J = 1$, that is

$$f(u_1, u_2) = \frac{1}{M} e^{-c_1 u_1 - c_2 u_2} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} (1 + b_{01} u_1 + b_{10} u_2 + b_{11} u_1 u_2)^2,$$

- Model M2, with $J = 2$, that is:

$$f(u_1, u_2) = \frac{1}{M} e^{-c_1 u_1 - c_2 u_2} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} (1 + b_{01} u_1 + b_{02} u^2 + b_{10} u_2 + b_{11} u_1 u_2 + b_{12} u_1 u_2^2 + b_{20} u_1^2 + b_{21} u_1^2 u_2 + b_{22} u_1^2 u_2^2)^2.$$

- In order to compare with the unconstrained Model M2, and quantify the induced potential actuarial pricing error[9], we also estimate Model M2bis, that is the constrained version of M2, under the independence condition $b_{j,k} = b_{j,0} b_{0,k}$ [see Proposition 1]. More precisely we have:

$$f(u_1, u_2) = \frac{1}{M} e^{-c_1 u_1 - c_2 u_2} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} (1 + b_{01} u_1 + b_{20} u_1^2)^2 (1 + b_{10} u_2 + b_{02} u_2^2)^2.$$

Due to the large number of parameters, the numerical optimization of the log-likelihood function is highly complicated. Thus it is essential to find an initial value for the optimization that is close to the real value of the parameter. Such an initial value can be found via two seemingly unrelated[10] non-linear least square estimation, also called pseudo maximum likelihood [see e.g. Gourieroux et al. (1984)]. Indeed, we can remark that:

$$\mathbb{E}[N_i \mid X_i] = \exp(d_1' X_i) \mathbb{E}[U_1] \tag{15}$$

$$\mathbb{E}[Y_i \mid X_i] = \mathbb{E}\Big[\mathbb{E}[Y_i \mid X_i, N_i] \mid X_i\Big] = \exp\left[(d_1' - d_2') X_i\right] \mathbb{E}[U_1/U_2]. \tag{16}$$

Thus d_1 and $d_1 - d_2$, as well as $\mathbb{E}[U_1], \mathbb{E}[U_1/U_2]$ can be estimated by minimizing:

$$\sum_{i=1}^{I} (N_i - \exp(d_1' X_i) \mathbb{E}[U_1])^2,$$

and

$$\sum_{i=1}^{I} \Big[Y_i - \exp\big((d_1' - d_2') X_i\big) \mathbb{E}[U_1/U_2] \Big]^2.$$

Once the initial values of the regression parameters are obtained, we estimate the different models by maximum likelihood. The following table reports first the estimates of the regression coefficients.

[9] The current actuarial literature usually assumes independence between U_1 and U_2, see e.g. Tzougas et al. (2014).

[10] That is to say, by neglecting the dependence between N and Y. See e.g. Zellner (1962) for a discussion.

Variable	Regression coefficient for the count variable				Regression coefficient for the cost variable			
	M0	M1	M2	M2bis	M0	M1	M2	M2bis
X_1	-0.0201	-0.0235	-0.0226	-0.0217	0.166	0.178	0.152	0.176
X_2	-0.158	-0.146	-0.142	-0.153	0.436	0.468	0.421	0.456
X_3	-0.114	-0.132	-0.125	-0.0120	0.426	0.453	0.409	0.431
X_4	-0.131	-0.146	-0.126	-0.128	0.340	0.369	0.324	0.352
X_5	-0.287	-0.264	-0.259	-0.271	0.416	0.436	0.462	0.427
X_6	-0.185	-0.169	-0.199	-0.175	0.260	0.245	0.236	0.254
X_7	0.418	0.403	0.426	0.423	-0.361	-0.335	-0.374	-0.365
X_8	0.254	0.229	0.260	0.249	-0.120	-0.135	-0.114	-0.126
X_9	0.228	0.196	0.264	0.221	-0.052	-0.046	-0.038	-0.044
X_{10}	0.198	0.185	0.201	0.193	-0.031	-0.029	-0.036	-0.033
X_{11}	0.0156	0.0192	0.0144	0.0164	0.107	0.095	0.114	0.112
X_{12}	0.0872	0.0893	0.0869	0.086	0.107	0.113	0.093	0.112
X_{13}	0.225	0.218	0.197	0.230	0.0896	0.0878	0.0852	0.905
X_{14}	0.0838	0.0866	0.0842	0.0893	0.606	0.572	0.632	0.664
X_{15}	0.146	0.154	0.142	0.153	0.106	0.098	0.120	0.114
X_{16}	0.172	0.169	0.182	0.184	0.254	0.268	0.224	0.263
X_{17}	0.358	0.326	0.335	0.349	-0.026	-0.018	-0.014	-0.029
X_{18}	0.214	0.228	0.241	0.223	-0.127	-0.132	-0.142	-0.128
X_{19}	0.243	0.226	0.238	0.239	0.122	0.136	0.114	0.126

Table 4: Estimates of the regression coefficients.

As expected, we can check that the estimators of the regression coefficients are quite similar across different models. Somehow surprisingly, the sign of the regression coefficients before the same covariate might be different for the count and the cost variable. In other words, the covariates might have opposite effects on the two response variables; this confirms also that it is important to accommodate for flexible dependence between the random effects U_1, U_2.

The following table reports the estimates of the remaining parameters, that are δ, as well as parameters characterizing the joint distribution of the unobserved heterogeneities.

	M0	M1	M2	M2bis
α_1	3.19	3.36	2.90	3.09
c_1	49.4	50.1	52.7	49.7
α_1	2.64	2.28	2.25	2.52
c_2	0.26	0.12	0.356	0.24
δ	1.68	1.91	1.76	1.76
b_{01}	—	1.98	-0.557	1.44
b_{10}	—	-0.24	-1.89	-0.12
b_{11}	—	0.06	-1.59	$= b_{01}b_{10} = -0.17$
b_{02}	—	—	0.335	0.87
b_{20}	—	—	-1.95	1.13
b_{12}	—	—	0.221	$= b_{10}b_{02} = -0.10$
b_{21}	—	—	-0.839	$= b_{20}b_{01} = 1.63$
b_{22}	—	—	-0.112	$= b_{20}b_{02} = 1.63$

Table 5: Estimates of the remaining parameters for the three models. The symbol "—" indicates that a parameter is not involved in a particular model.

Finally, let us provide a comparison of the four models in terms of information criteria.

	M0	M1	M2	M2bis
log-likelihood	-11209	-11154	-11087	-11198
BIC	22462	22385	22306	22430

Table 6: Information criteria of the four models.

We can observe that Model 2 is the best model in terms of BIC, although the difference between M2 and M1 is rather small. On the other hand, we can see that the benchmark model M0, as well as the constrained model M2bis have a BIC that is much larger than their more flexible competitors. To further illustrate this, let us analyze the heterogeneity distribution for the estimated model M2. The following figures plot the p.d.f. of the marginal distributions of U_1 and that of U_2.

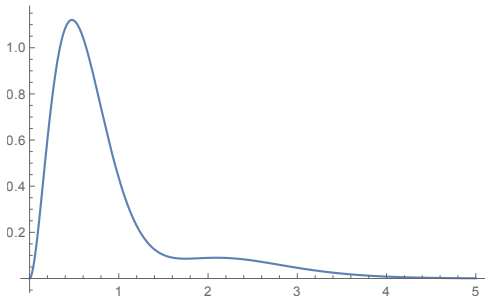

Figure 2: Marginal p.d.f of U_1 obtained from Model M2.

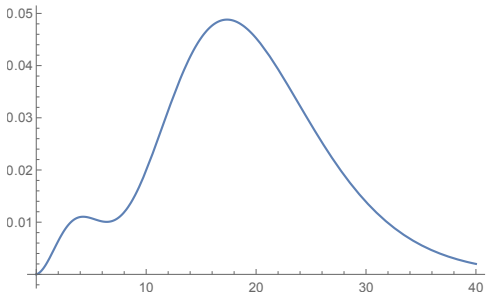

Figure 3: Marginal p.d.f of U_2 obtained from Model M2.

We can see that Model M2 indicates a significant difference of the marginal distributions of both U_1 and U_2 from the gamma distribution assumption. Let us now plot the iso-densities of the joint distribution of (U_1, U_2).

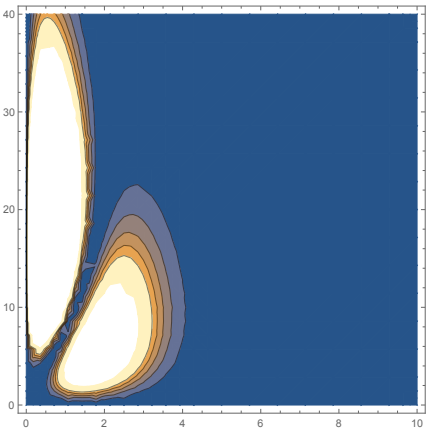

Figure 4: Iso-densities of the joint distribution of (U_1, U_2) obtained from Model 2. The lighter a region is, the larger the joint p.d.f. is in that region.

We can observe that the joint p.d.f. of (U_1, U_2) have two distinct local maxima. This is to be compared with the joint p.d.f. in Model M0, in which U_1, U_2 are independent gamma (in this case, there is only one global maximum). This implies that the joint p.d.f. of $(U_1, U_1/U_2)$ has also two modes, one mode corresponds to individuals with both larger values of U_1 and $1/U_2$, or in other words, the riskier group; the other one corresponds to individuals with smaller values of U_1 and $1/U_2$, that are the safer group. This positive association between U_1 and $1/U_2$ is also reflected by the correlation coefficient between these two terms, which is equal to 0.33.

3.3 Impact of wrong dependence assumptions on actuarial pricing

Let us now study the impact on the actuarial premium of the incorrect assumption of independence between U_1 and U_2. Given the claim counts $N_1, N_2, ..., N_T$, as well as total costs $Y_1, Y_2, ..., Y_T$, we compute next year's actuarial fair premium, that is the conditional expected future cost Y_{T+1} of the claim, under Model M2, as well as the nested model[11] M2b, which assumes independence between U_1 and U_2.

For illustration purpose, we consider a policyholder for which the values of the covariates are all equal to 0, that is, $X_1 = X_2 = ... = X_{17} = 0$. In this case, we have $\lambda_t = \beta_t = 1$ for all t, and $\Lambda_{1T} = T$ is equal to the number of periods, whereas $\Lambda_{2T} = \overline{Y}_T$ is equal to the cumulative claim cost. As a consequence, in this special case, the premium is a function of cumulative counts $\overline{N}_T = N_1 + \cdots + N_T$, and cumulative costs $\overline{Y}_T = Y_1 + \cdots + Y_T$, and T only. The following table provides some examples of the values of the premium for different values of $\overline{N}_T$, $\overline{Y}_T$, and T.

	Model M2b	Model 2
$\overline{N}_T = 0, \overline{Y}_T = 0, T = 0$	147 \$	147 \$
$\overline{N}_T = 0, \overline{Y}_T = 0, T = 1$	144 \$	142 \$
$\overline{N}_T = 0, \overline{Y}_T = 0, T = 3$	139 \$	134 \$
$\overline{N}_T = 1, \overline{Y}_T = 300\,\$, T = 1$	159 \$	163 \$
$\overline{N}_T = 1, \overline{Y}_T = 300\,\$, T = 3$	153 \$	154 \$
$\overline{N}_T = 1, \overline{Y}_T = 2000\,\$, T = 1$	263 \$	419 \$
$\overline{N}_T = 1, \overline{Y}_T = 2000\,\$, T = 3$	253 \$	396 \$
$\overline{N}_T = 1, \overline{Y}_T = 5000\,\$, T = 1$	454 \$	904 \$
$\overline{N}_T = 1, \overline{Y}_T = 5000\,\$, T = 3$	436 \$	853 \$

Table 7: Examples of posterior premium at period $T + 1$, given claim counts and costs of the first T periods.

We observe that the two models provide the same a priori premium (first row of the previous table), that is when no previous claim history is available. This is expected, since the a priori premium is equal to $\mathbb{E}[U_1/U_2]$, which can be estimated consistently by the same quasi likelihood method. Once past claim history begins to accumulate, the premium given by the two models start to differ quite significantly. From the second and third row, we can see that Model 2 provides more premium reduction than Model 2, when 0 past claim has been reported during the past observation period. This is expected, since in Model 2, the unobserved heterogeneities U_1 and $1/U_2$ are positively correlated. Thus the absence of claims indicates that the individual has, in average, a smaller value of U_1, and, as a consequence, a smaller value of $1/U_2$. From the

[11] The choice of the alternative model M2bis is motivated by the fact that M2bis is flexible when it comes to modeling the marginal distributions of U_1 and U_2. Thus Models M2 and M2bis differ only in their ability of capturing the dependence. One could also compare, say, Models M2 and M0, but in this case the pricing errors would be more difficult to analyze, since both the marginal distributions and the dependence structure contribute differ between the two models.

fourth and fifth row, when one accident of value 300 $ is reported (which is significantly smaller than the average claim cost, for those who report claims), Model 2 charges a higher premium increase than Model M2bis. This can be explained by the following. The presence of a claim indicates that the individual has a larger value of U_1, but the small size of the claim indicates a smaller value of $1/U_2$. In Model M2bis, where dependence is not taken into account, these two effects are well compensated. This is less the case in Model 2, where due to the positive dependence between U_1 and $1/U_2$, the two effects are only partially compensated, hence the higher premium in Model 2. For the same reason, in the sixth and seventh row, where a claim of average value (2000 $) is reported, or in the eighth and ninth row, where a claim of large value (5000 $) is reported, Model M2 charges a higher premium than Model M2bis. Moreover, the larger the claim size, the more important the difference between the premium given by the two models. To summarize, compared to Model M2, Model M2bis overcharges low-risk customers, but undercharges high-risk ones. This constitutes a huge disadvantage on a competitive market. Indeed, an insurer using Model M2bis, say A, is likely to lose low-risk customers (who prefer its competitor, say B, operating under Model M2 and hence provide lower premium for these customers). On the other hand, Insurer A is more likely to take in high-risk customers (who leave insurer B to seek lower premium), who are undercharged by insurer A *compared to their real risk*. As a consequence, insurer A will face solvency issues in the long-run, since the premium income is likely to be insufficient to cover the realized claims. Therefore, Model M2bis provides not only a less satisfactory fit[12], but also significant pricing bias.

4 Conclusion

We have proposed a general, flexible random effect panel data model for bivariate count-continuous variables, that is especially adapted for insurance forecast and pricing. The model allows for convenient, matrix-based expressions for the likelihood function, as well as the forecast formulas. Finally, we have demonstrated the benefit such a model offers for the forecasting of future insurance claims, compared to existing parametric specifications.

[12] In terms of BIC, for instance.

Appendix 1: proofs

Proof of Proposition 1. We have independence if and only if $g(x,y) = g_1(x)g_2(y)$, which is equivalent to:

$$\frac{1}{M}\Big(\sum_{j=0}^{J}\sum_{k=0}^{J} b_{j,k}u_1^j u_2^k\Big)^2 = \frac{g(u_1,u_2)}{e^{-c_1u_1-c_2u_2}u_1^{\alpha_1-1}u_2^{\alpha_2-1}} = \frac{g_1(u_1)g_2(u_2)}{e^{-c_1u_1-c_2u_2}u_1^{\alpha_1-1}u_2^{\alpha_2-1}} = P_1(u_1)P_2(u_2),$$

where $P_1 = \frac{E_1(u_1)'\Pi W_2(0)}{W_1(0)'\Pi W_2(0)}, P_2 = \frac{W_1(0)'\Pi E(u_2)}{W_1(0)'\Pi W_2(0)}$ are polynomials. By taking $u_2 = 0$, we have, $P_1(u_1) = \frac{1}{MP_2(0)}(\sum_{j=0}^{J} b_{j,0}u_1^j)^2$; by taking $u_1 = 0$, we have $P_2(u_2) = \frac{1}{MP_1(0)}(\sum_{k=0}^{J} b_{0,k}u_2^k)^2$; by taking $u_1, u_2 = 0$, we have $P_1(0)P_2(0) = \frac{1}{M}b_{00} = \frac{1}{M}$. Thus we have finally:

$$\Big(\sum_{j=0}^{J}\sum_{k=0}^{J} b_{j,k}u_1^j u_2^k\Big)^2 = \Big(\sum_{k=0}^{J} b_{0,k}u_2^k\Big)^2\Big(\sum_{j=0}^{J} b_{j,0}u_1^j\Big)^2,$$

which yields $b_{jk} = b_{j,0}b_{0,k}$ for integers j,k between 0 and J.

Proof of Proposition 2. The marginal density of y has two components, one mass at zero, and one continuous component on $\mathbb{R}_{>0}$. The density of this component is:

$$f(y) = \sum_{n=1}^{\infty} f(n,y).$$

By equation (11), the dominant term of this function is proportional to $y^{-1-\alpha_2}$, when y goes to infinity. Thus $\mathbb{E}[Y_t]$ exists if and only if $\alpha_2 > 1$.

Appendix 2: A simulation study

In order to get a sense of the finite sample behavior of the estimation method as well as its impact on pricing, in this section we propose a Monte-Carlo simulation exercise. For a fixed vector of parameters θ, we simulate 50 databases[13], which one consisting of 10000 i.i.d. samples of (N_1, Y_1, X_1). Then we conduct maximum likelihood estimation on these 100 simulated databases.

[13]This relatively small number of replications is explained by the fact that simulating each database, as well as computing the maximum likelihood estimator, is quite computationally intensive.

The simulated data and summary statistics of the estimates

The Data Generating Process (DGP) we propose is the following: in the joint density formula of (U_1, U_2), we take:

$$\alpha_1 = \alpha_2 = 2, \quad c_1 = c_2 = 1, \quad J = 1, \quad b_{00} = 1, b_{01} = b_{10} = 0, \quad b_{11} = 0.5,$$

where b_{00} is set to one by convention. Note also that for illustration purpose, we have assigned nonnegative values to all the parameters $b_{j,k}$. The advantage of this set of parameters is that the resulting joint density is a mixture distribution and is therefore easy to simulate.[14]

Then we set the two regressors in (1) as:

$$\lambda = \exp(\beta_1 X_1), \qquad \beta = \exp(\beta_2 X_2),$$

where $\beta_1 = 0.5, \beta_2 = 0.6$, and X_1, X_2 are i.i.d. $\mathcal{N}(0,1)$ distributed. Finally, parameter δ is set to 1. The following table reports the theoretical mean and variances of the two variables N, Y.

Variable	Mean	Variance	Variance/Mean
N	7.6	30.6	4.04
Y	3.2	29.3	9.2

Table 8: Summary statistics of the couple (N, Y).

We can see that for this DGP, the count variable is quite heavily overdispersed. As pointed out by a referee, this situation is commonly encountered in practice for low count data.

The following table reports the sample mean and standard deviation of the estimates, obtained respectively from the 50 simulated databases.

Variable	Theoretical value	Empirical Mean	Standard Deviation
β_1	0.500	0.501	0.12
β_2	0.600	0.598	0.11
α_1	2.00	2.02	0.13
α_2	2.00	2.10	0.14
c_1	1	1.06	0.07
c_2	1	0.97	0.08
b_{01}	0	0.01	0.06
b_{10}	0	0.01	0.05
b_{11}	0.5	0.052	0.2
δ	1	0.97	0.2

Table 9: Table of the estimates along with their standard deviation.

[14] Nevertheless, even when some of the coefficients are negative, it is still possible to use a acceptance-rejection algorithm to simulate samples of (U_1, U_2). See Gallant and Tauchen (1993) for details.

All of the parameter estimates are significant, except those of b_{01} and b_{10}. This is expected since their true values are exactly zero.

Impact on pricing

Next, let us examine how the sampling error might affect the forecasting performance. To this end, for each of the 50 simulated databases, we use the estimated model to compute the conditional expectation $\mathbb{E}[Y_2|Y_1, N_1, X]$ using (13). We then compute the sample mean and variance of $\mathbb{E}[Y_2|Y_1, N_1, X]$. The following figure plots the histogram of these estimated forecasts, for $X_1 = X_2 = 0$, $N = 1$ and $Y = 1.6$.

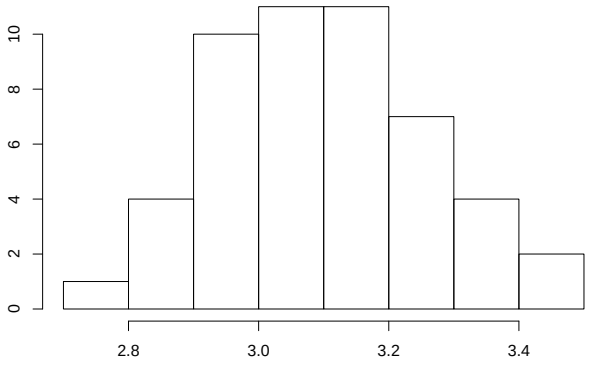

Figure 5: Histogram of the estimated forecasts.

Of these 50 samples of $\mathbb{E}[Y_2|Y_1, N_1, X]$, the largest error is around 0.3, which is about 10 % of its theoretical value.

It is useful to compare this simulation exercise with the real data application of Section 3. There are several differences between the previous simulation and the insurance data. Firstly, the insurance data contains more than 60000 observations, that is more than 6 times larger than the simulated datasets, which should contribute to lower the sampling error of the maximum likelihood estimator. Secondly, as is often the case in practice, the real data we analyzed contains much more covariates than the simulation experiment.[15] One can reasonably argue that this spells a larger number of parameters and thus can adversely affect the precision of the maximum likelihood estimator. Nevertheless, we argue that in practice this impact can be mitigated by the

[15]In our simulation experiment, we have deliberately chosen only two covariates, in order not to mix the error induced by adding too many regression coefficients, as well as those induced by the novel specification of the random effects (U_1, U_2). Such a choice is rather standard in this literature and has been previously made in Gurmu et al. (1999); Gurmu and Elder (2012) for the analysis of count data.

fact that these regression parameters can be quite easily estimated separately by nonlinear least square [see equations (15) and (16)]. Also, Lasso-like methods can also be used to diminish the number of regression parameters, although it is clearly out of the scope of the present paper.

References

Bierens, H. J. (2008). Semi-Nonparametric Interval-Censored Mixed Proportional Hazard Models: Identification and Consistency Results. *Econometric Theory*, 24(03):749–794.

Catalano, P. J. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87(419):651–658.

Cummins, J. D., Dionne, G., McDonald, J. B., and Pritchett, B. M. (1990). Applications of the gb2 family of distributions in modeling insurance loss processes. *Insurance: Mathematics and Economics*, 9(4):257–272.

Czado, C., Kastenmeier, R., Brechmann, E. C., and Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 2012(4):278–305.

De Jong, P. and Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press.

de Leon, A. R. and Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, 30(2):175–185.

Dionne, G. and Vanasse, C. (1989). A generalization of automobile insurance rating models: the negative binomial distribution with a regression component. *Astin Bulletin*, 19(2):199–212.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics*, 1(2):115–126.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):355–366.

Einav, L., Finkelstein, A., and Schrimpf, P. (2010). Optimal mandates and the welfare cost of asymmetric information: Evidence from the uk annuity market. *Econometrica*, 78(3):1031–1092.

Englin, J. and Shonkwiler, J. S. (1995). Estimating social welfare using count data models: an application to long-run recreation demand under conditions of endogenous stratification and truncation. *Review of Economics and Statistics*, pages 104–112.

Fitzmaurice, G. M. and Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American statistical Association*, 90(431):845–852.

Frangos, N. E. and Vrontos, S. D. (2001). Design of optimal bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance. *Astin Bulletin*, 31(01):1–22.

Gallant, A. R. and Nychka, D. W. (1987). Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica*, 55(2):363–390.

Gallant, A. R. and Tauchen, G. (1993). A nonparametric approach to nonlinear time series analysis: estimation and simulation. In *New directions in time series analysis*, pages 71–92. Springer.

Garrido, J., Genest, C., and Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70:205–215.

Gouriéroux, C. (1999). The econometrics of risk classification in insurance. *Geneva Papers on Risk and Insurance Theory*, 24(2):119–137.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to poisson models. *Econometrica*, 52(3):701–720.

Gueorguieva, R. V. and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96(455):1102–1112.

Gurmu, S. and Elder, J. (2012). Flexible bivariate count data regression models. *Journal of Business & Economic Statistics*, 30(2):265–274.

Gurmu, S., Rilstone, P., and Stern, S. (1999). Semiparametric estimation of count regression models. *Journal of Econometrics*, 88(1):123–150.

Lemaire, J. (1995). Bonus-malus systems in automobile insurance. *Insurance: Mathematics and Economics*, 3(16):277.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, 52(3):647–663.

Paull, A. E. (1978). A generalized compound poisson model for consumer purchase panel data analysis. *Journal of the American Statistical Association*, 73(73):706–713.

Pinquet, J., Guillén, M., and Bolancé, C. (2001). Allowance for the age of claims in bonus-malus systems. *Astin Bulletin*, 31(02):337–348.

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):667–678.

Tzougas, G., Vrontos, S., and Frangos, N. (2014). Optimal bonus-malus systems using finite mixture models. *Astin Bulletin*, 44(02):417–444.

Yang, Y., Kang, J., Mao, K., and Zhang, J. (2007). Regression models for mixed poisson and continuous longitudinal data. *Statistics in Medicine*, 26(20):3782–3800.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368.