A priori ratemaking using bivariate Poisson regression models[*]

Lluís Bermúdez i Morata [†]

Departament de Matemàtica Econòmica, Financera i Actuarial.

Risc en Finances i Assegurances-IREA. Universitat de Barcelona.

September 22, 2008

Abstract

In automobile insurance, it is useful to achieve *a priori* ratemaking by resorting to generalized linear models, and here the Poisson regression model constitutes the most widely accepted basis. However, insurance companies distinguish between claims with or without bodily injuries, or claims with full or partial liability of the insured driver. This paper examines an *a priori* ratemaking procedure when including two different types of claim. When assuming independence between claim types, the premium can be obtained by summing the premiums for each type of guarantee and is dependent on the rating factors chosen. If the independence assumption is relaxed, then it is unclear as to how the tariff system might be affected. In order to answer this question, bivariate Poisson regression models, suitable for paired count data exhibiting correlation, are introduced. It is shown that the usual independence assumption is unrealistic here. These models are applied to an automobile insurance claims database containing 80,994 contracts belonging to a Spanish insurance company. Finally, the consequences for pure and loaded premiums when the independence assumption is relaxed by using a bivariate Poisson regression model are analysed.

JEL classification: C51; *IM classification:* IM11; *IB classification:* IB40.

Keywords: Bivariate Poisson regression models, Zero-inflated models, Automobile insurance, Bootstrap methods, *A priori* ratemaking.

[*] *Acknowledgements.* The author wishes to acknowledge the support of the Spanish Ministry of Education and FEDER grant SEJ 2007-63298. The author is grateful for the valuable suggestions from the participants in the 12th International Congress on Insurance: Mathematics and Economics in Dalian on July 16-18, 2008.

[†] **Corresponding Author.** Departament de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona, Diagonal 690, 08034-Barcelona, Spain. Tel.: +34-93-4034853; fax: +34-93-4034892; e-mail: `lbermudez@ub.edu`

1 Introduction

Designing a tariff structure for insurance is one of the main tasks for actuaries. Such pricing is particularly complex in the branch of automobile insurance because of highly heterogeneous portfolios. A thorough review of ratemaking systems for automobile insurance, including the most recent developments, can be found in Denuit *et al.* (2007).

One way to handle this problem of heterogeneity in a portfolio -referred to as tariff segmentation or *a priori* ratemaking- involves segmenting the portfolio in homogenous classes so that all insured parties belonging to a particular class pay the same premium. This procedure ensures that the exact weight of each risk is fairly distributed within the portfolio. In the case of automobile insurance, in order to group the policies in homogenous classes, a series of classification variables are used (i.e., age, sex and place of residence of driver or horsepower, class and use of the vehicle). These variables are called *a priori* ratemaking variables, since their values can be determined before the insured party begins to drive.

If all the factors influencing a risk could be identified, measured and introduced in the tariff system, then the classes defined would be homogenous. However, this is not that case as there are important risk factors that are not considered in the *a priori* tariff. Some examples are especially difficult to quantify, such as a driver's reflexes, his or her aggressiveness, or knowledge of the Highway Code, among others. As a result, tariff classes can be quite heterogeneous. Hence, the idea has arisen of considering individual differences in policies within the same class by using an *a posteriori* mechanism, i.e., fitting an individual premium based on the experience of claims for each insured party. This concept has received the name of *a posteriori* tariff, experience rating or the bonus-malus system.

Here, only the first step in pricing is studied, the *a priori* ratemaking. In short, the classification or segmentation of risks involves establishing different classes of risk according to their nature and probability of occurrence. For this purpose, factors are determined in order to classify each risk, and it is statistically tested that the probability of a claim depends on these factors, and hence, their influence can be measured. *A priori* classification based on generalized linear models is the most widely accepted method; see e.g. Dionne and Vanasse (1989), Haberman and Renshaw (1996), Pinquet (1999), Bermúdez *et al.* (2001) and Boucher and Denuit (2006) for applications in the actuarial sciences, and Mc Cullagh and Nelder (1989) or Dobson (1990) for a general overview of the statistical theory.

The most commonly used generalized linear model for this tariff system is the Poisson regression model and its generalizations (Denuit *et al.*, 2007). Introduced by Dionne and Vanasse (1989) in the context of automobile insurance, the model can be applied if the number of claims for each individual policy observation is known. Although it is possible to use the total number of claims as the response variable, the nature of automobile insurance policies (covering different risks) is such that the response variable is the number of claims for each type of guarantee. Therefore, a premium is obtained for each class of guarantee as a function of different factors. Then, assuming independence between types of claim, the total premium is obtained from the sum of the expected number of claims of each guarantee.

Here, two different types of guarantee are assumed: third-party liability automobile insurance and the rest of guarantees. Following the usual methodology, assuming independence between types, the premium paid by the policyholder is obtained by summing the premiums for each type of guarantee and this depends on the rating factors. However, the question remains as to whether the independence assumption is realistic. When this assumption is relaxed, it is interesting to see how the tariff system might be affected.

In this study, a bivariate Poisson regression model is introduced. Holgate (1964) provided a practical basis for the bivariate Poisson distribution but its use has been largely ignored, mainly because of computational difficulties. Therefore, only a few applications can be found, for example, Jung and Winkelmann (1993) used a bivariate Poisson regression in a labour mobility study and Karlis and Ntzoufras (2003) modelled sports data. For a comprehensive review of the bivariate Poisson distribution and its applications (especially multivariate regression), the reader should see Kocherlakota and Kocherlakota (1992, 2001) and Johnson, Kotz and Balakrishnan (1997).

One early application of the bivariate Poisson distribution in the actuarial literature is described in Cummins and Wiltbank (1983). In ruin theory, some applications of this distribution are also to be found, for example Partrat (1994), Ambagaspitiya (1999), Walhin and Paris (2000) and Centeno (2005). Cameron and Trivedi (1998) studied the relationship between type of health insurance and various responses that measure the demand for health care by using a bivariate

Poisson regression. In addition, two studies related to fitting purposes should also be quoted, albeit that no factors are considered. First, Vernic (1997) carried out a comparative study with the bivariate Poisson distribution based on data related to natural events insurance and third-party liability automobile insurance. Second, Walhin (2003) compared bivariate Hofmann and bivariate Poisson distributions by fitting a data set for accidents sustained by members of a sample of 122 shunters in two consecutive 2-year periods. However, in a ratemaking context, bivariate Poisson regression models have not been used to model claim counts that depend on the usual rating factors.

In the next section, the model used here is defined. This model is based on the bivariate Poisson regression model, which is appropriate for modelling paired count data that exhibit correlation. In Section 3 the database obtained from a Spanish insurance company is described. In Section 4 the results are summarised. Finally, some concluding remarks are given in Section 5.

2 Bivariate Poisson regression models

Let N_1 and N_2 be the number of claims for third-party liability and for the rest of guarantees respectively and $N = N_1 + N_2$. The usual methodology to obtain the *a priori* premium under the assumption of independence between types of claims can be described as follows. First, the model assumed is $N_1 \sim Poisson(\lambda_1)$ and $N_2 \sim Poisson(\lambda_2)$ independently, and λ_1 and λ_2 depend on a number of rating factors associated with the characteristics of the car, the driver and the use of the car. Second, with λ_1 and λ_2 estimated for each policyholder and following the net premium principle, the total net premium[1] (π) is obtained as $\pi = E[N] = E[N_1] + E[N_2] = \lambda_1 + \lambda_2$.

However, an amount inflates the net premium to ensure that the insurer will not, on average, lose money. Many well-known premium principles can be applied for this purpose. Here the variance premium principle is used. This principle builds on the net premium by including a risk loading that is proportional to the variance of the risk. Under the above assumptions, the variance is equal to the expected value, and the total loaded premium (π^*) is equal to $\pi^* = E[N] + \alpha V[N] = (1+\alpha)(E[N_1] + E[N_2])$.

In bivariate Poisson regression models, the independence assumption is relaxed. The model

[1] Assuming the amount of the expected claim equals one monetary unit.

can be defined as follows. Let us consider independent random variables $X_i\,(i = 1, 2, 3)$ to be distributed as Poisson with parameters λ_i respectively. Then the random variables $N_1 = X_1 + X_3$ and $N_2 = X_2 + X_3$ follow jointly a bivariate Poisson distribution:

$$(N_1,\, N_2) \sim BP(\lambda_1,\, \lambda_2,\, \lambda_3).$$

This is the so-called trivariate reduction method that leads to the bivariate Poisson distribution. Its joint probability function is given by:

$$P(N_1 = n_1,\, N_2 = n_2) \;=\; e^{-(\lambda_1+\lambda_2+\lambda_3)}\, \frac{\lambda_1^{n_1}}{n_1!}\, \frac{\lambda_2^{n_2}}{n_2!} \sum_{i=0}^{\min(n_1,n_2)} \binom{n_1}{i}\binom{n_2}{i}\, i!\, \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^{i}. \qquad (1)$$

The bivariate Poisson distribution defined above presents several interesting and useful properties. First, it allows for positive dependence between the random variables N_1 and N_2 which is what we expect for these types of claims[2]. Moreover $Cov(N_1,\, N_2) = \lambda_3$ and therefore λ_3 is a measure of this dependence. Obviously, if $\lambda_3 = 0$ the two random variables are independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions, referred to as a double Poisson distribution (Kocherlakota and Kocherlakota, 1992). Second, the marginal distributions for N_1 and N_2 are Poisson with $E[N_1] = \lambda_1 + \lambda_3$ and $E[N_2] = \lambda_2 + \lambda_3$.

Hence, the total net premium can be obtained with $\pi = E[N] = E[N_1] + E[N_2] = \lambda_1 + \lambda_2 + 2\lambda_3$. The variance necessary to obtain the loaded premium is now $V[N] = \lambda_1 + \lambda_2 + 4\lambda_3$. Since λ_3 is expected to be positive, the relaxation of the independence assumption leads to a variance greater than the expected value. Overdispersion has often been observed when modelling claim counts in automobile insurance data (Denuit *et al.*, 2007).

Let us assume that N_{1j} and N_{2j} denote the random variables indicating the number of claims of each type of guarantee for the jth policyholder. If covariates are introduced to model λ_1 , λ_2 and λ_3 , a bivariate Poisson regression model can be defined with the following scheme:

$$\begin{aligned}(N_{1j},\, N_{2j}) &\sim BP(\lambda_{1j},\, \lambda_{2j},\, \lambda_{3j}),\\ \log(\lambda_{1j}) &= \boldsymbol{x'_{1j}}\,\boldsymbol{\beta_1},\\ \log(\lambda_{2j}) &= \boldsymbol{x'_{2j}}\,\boldsymbol{\beta_2},\\ \log(\lambda_{3j}) &= \boldsymbol{x'_{3j}}\,\boldsymbol{\beta_3}, \qquad (2)\end{aligned}$$

[2] In case of negatively correlated claims (not considered here) it would be necessary a more general specification.

where $j = 1, \ldots, n$ denotes the observed policies with sample size n, $\boldsymbol{x_{ij}}$ denotes a vector of explanatory variables and $\boldsymbol{\beta_i}$ denotes the corresponding vector of regression coefficients ($i = 1, 2, 3$).

In the case of the explanatory variables, two aspects should be stressed. First, different covariates can be used to model each parameter λ_{ij}. Second, to facilitate the interpretation, covariates can not be introduced to model λ_3. However, they can be included so as to know more about the influence of the covariates on each pair of variables.

A problem arises when examining the joint probability function given in (1), particularly in $\min(n_1, n_2)$ when the proportion of $(0,\, 0)$ is larger than that of other frequencies. Therefore, it seems reasonable to fit a zero-inflated model.

Few studies to date have discussed zero-inflated models in bivariate discrete distributions. Such models have been proposed by Li *et al.* (1999) and Wang *et al.* (2003) who considered inflation only for the $(0,\, 0)$ cell, or Walhin (2001) who discussed zero-inflated bivariate Poisson models. However, here we follow the zero-inflated bivariate Poisson model proposed by Karlis and Ntzoufras (2005). In fact, they propose an extension of the simple zero-inflated model which inflates the probabilities in the diagonal of the probability table. It seems reasonable to believe, for instance, that there also exists a higher proportion of $(1,\, 1)$ because the same accident can lead to one claim of each type being made.

Taking the bivariate Poisson model (BP) defined above as the starting point, the diagonal inflated bivariate Poisson model (DIBP) is specified by the probability function:

$$f_{DIBP}(N_1,\, N_2) = \begin{cases} (1-p)\, f_{BP}(N_1,\, N_2 \,|\, \lambda_1,\, \lambda_2,\, \lambda_3) & N_1 \neq N_2 \\ (1-p)\, f_{BP}(N_1,\, N_2 \,|\, \lambda_1,\, \lambda_2,\, \lambda_3) + p\, f_D(N_1 \,|\, \boldsymbol{\theta}) & N_1 = N_2, \end{cases} \tag{3}$$

where $f_{BP}(N_1,\, N_2 \,|\, \lambda_1,\, \lambda_2,\, \lambda_3)$ is the joint probability function given in (1), $f_D(N_1 \,|\, \boldsymbol{\theta})$ is a probability function for a discrete distribution $D(N_1 \,|\, \boldsymbol{\theta})$ defined on $\{0,\, 1,\, 2,\, \ldots\}$ with parameter vector $\boldsymbol{\theta}$ and p is a parameter defined in $[0,\, 1]$. Note that two special cases can be obtained from this more general case. First, the bivariate Poisson model, taking $p = 0$, and second, the zero-inflated bivariate Poisson (ZIBP), taking $D(N_1 \,|\, \boldsymbol{\theta})$ as a degenerate function at zero.

In contrast to the bivariate Poisson model, the marginal distributions of N_1 and N_2 of a diagonal inflated model are not Poisson distributed and, as such, they can present underdispersion or overdispersion. Let $E_D[N_1]$ and $E_D[N_1^2]$ be the first two moments of $D(N_1 \,|\, \boldsymbol{\theta})$, the marginal mean and variance for N_1 are:

$$\begin{aligned} E_{DIBP}[N_1] &= (1-p)\,(\lambda_1 + \lambda_3) + p\,E_D[N_1] \\ V_{DIBP}[N_1] &= (1-p)\,\left\{(\lambda_1 + \lambda_3)^2 + (\lambda_1 + \lambda_3)\right\} + p\,E_D[N_1^2] \\ &\quad - \{(1-p)\,(\lambda_1 + \lambda_3) + p\,E_D[N_1])\}^2\,. \end{aligned} \tag{4}$$

Finally, in order to calculate the covariance between N_1 and N_2 for this model, $E_{DIBP}[N_1,\, N_2]$ needs to be calculated. From (3), it follows that:

$$E_{DIBP}[N_1\,N_2] = (1 - p)\,E_{BP}[N_1\,N_2] + p\,E_D[N_1^2]. \tag{5}$$

Since $E_D[N_1] = E_D[N_1^2] = 0$ when only cell $(0,\,0)$ is inflated, the marginal distributions in the zero-inflated model are overdispersed and the marginal mean and variance for N_1 are:

$$\begin{aligned} E_{ZIBP}[N_1] &= (1-p)\,(\lambda_1 + \lambda_3) \\ V_{ZIBP}[N_1] &= (1-p)\,\left\{(\lambda_1 + \lambda_3) + p\,(\lambda_1 + \lambda_3)^2\right\}. \end{aligned} \tag{6}$$

For the analysis presented in the following sections, the covariance between N_1 and N_2 for a zero-inflated model needs to be calculated. First, from (5) a similar expression for the zero-inflated model can be obtained:

$$E_{ZIBP}[N_1\,N_2] = (1 - p)\,\{\lambda_3 + (\lambda_1 + \lambda_3)\,(\lambda_2 + \lambda_3)\}\,.$$

Thus, the covariance for a zero-inflated model is given by:

$$\begin{aligned} Cov_{ZIBP}[N_1,\, N_2] &= (1 - p)\,\{\lambda_3 + (\lambda_1 + \lambda_3)\,(\lambda_2 + \lambda_3)\} \\ &\quad - \left\{(1-p)^2\,(\lambda_1 + \lambda_3)\,(\lambda_2 + \lambda_3)\right\}. \end{aligned} \tag{7}$$

Different algorithms have been provided to implement bivariate Poisson regression models (Ho and Singer, 2001; Kocherlakota and Kocherlakota, 2001; or adopting a Bayesian point of view, Tsionas, 2001; Karlis and Meligkotsidou, 2005). Here an EM algorithm provided by Karlis and Ntzoufras (2005) and its implementation using R (*bivpois* package) is used. Standard errors for the parameter estimates are calculated using standard bootstrap methods (*boot* package in R).

3 The database

The original sample comprised a ten percent sample of the automobile portfolio of a major insurance company operating in Spain in 1995. Only cars categorised as being for private use were considered. The data contains information from 80,994 policy holders. The sample is not representative of the actual portfolio as it was drawn from a larger panel of policyholders who had been customers of the company for at least seven years; however, it will be helpful for illustrative purposes.

Twelve exogenous variables were considered plus the yearly number of accidents recorded for both types of claim. For each policy, the initial information at the beginning of the period and the total number of claims from policyholders at fault were reported within this yearly period.

The exogenous variables, described in Table 1, were previously used in Pinquet *et al.* (2001), Brouhns *et al.* (2003), Bolancé *et al.* (2003) and in Boucher *et al.* (2007). Moreover, in Table 2, the cross-tabulation for the number of claims for third-party liability (N_1) and number of claims for the rest of guarantees (N_2) are shown.

For this study, all customers had had a policy with the company for at least three years. Therefore, variable *v7* was rejected and variable *v8* retained its definition and its baseline was now established as a customer who had been with the company for fewer than five years.

The meaning of those variables referring to the policyholders' coverage should also be clarified. The classification here responds to the most common types of automobile insurance policies available on the Spanish market. The simplest policy only includes third-party liability (claimed and counted as N_1 type) and a set of basic guarantees such as emergency roadside assistance, legal assistance or insurance covering medical costs (claimed and counted as N_2 type). This simplest policy does not include comprehensive coverage (damage to one's vehicle caused by any unknown party, for example, damage resulting from theft, flood or fire) nor collision coverage (damage resulting from a collision with another vehicle or object when the policyholder is at fault). This simplest type of policies conforms the baseline group, while variable *v10* denotes policies which apart from the guarantees contained in the simplest policies also include comprehensive coverage (except fire) and variable *v11* denotes policies which also include fire and collision coverage.

4 Results

4.1 Fitting bivariate Poisson models

First, in order to show the convenience of using the bivariate Poisson model, a simple bivariate Poisson model (with constant λ_1, λ_2 and λ_3) was fitted. The estimated values for these parameters were 0.067, 0.088 and 0.014, respectively. AIC equal to 104,573.9 for the bivariate Poisson model was obtained, which was better than the values obtained for the double Poisson model (106,546.1). Even with a small correlation between N_1 and N_2, including λ_3 in the model produced a better fit for the data used.

Once the effectiveness of the bivariate model had been assessed, covariates to model λ_1, λ_2 and λ_3 were included. In fact, first the same variables for λ_1, and λ_2 were included, maintaining λ_3 constant. In Table 3 the results of fitting the bivariate Poisson regression model (with constant λ_3) and the results for the double Poisson regression model (without λ_3 term) are shown.

It can be seen that the intercept for λ_3 was significant (at the 5% level) indicating that the bivariate Poisson model is more appropriate for this data than is the model that assumes independence between N_1 and N_2 (double Poisson). As regards the fit, the AIC values for these models also indicate the improvement achieved with the bivariate model.

Focusing on λ_1 (claims for third-party liability), for the bivariate Poisson model the parameters from *v4* to *v8* and *v10* were significant. For the double Poisson model no important differences were found except for the parameter *v10* which was not significant. This difference may indicate the convenience of including this covariate to model the covariance term λ_3 (see Table 4).

Following the discussion above concerning claims for third-party liability, driving experience (*v5* and *v6*) reduced the expected number of claims, while driving in northern Spain (*v4*) and drivers with fewer than 5 years in the company (*v8*) caused the expected number of claims to increase for this type of claim. As regards the type of coverage, only in the case of the bivariate model, when including comprehensive coverage except fire (*v10*) was the expected number of claims lowered.

Concentrating on λ_2 (the rest of claims, except third-party liability), most of the parameters were significant and no noticeable differences were found between bivariate and double

Poisson models. In particular, the parameters for *v2* to *v5*, *v8* and *v10* to *v12* were statistically significant.

Here, some differences with the third-party liability claims were found. First, parameters related to the type of coverage (*v10* and *v11*) were always significant and their presence increased the expected number o claims markedly. Second, the car's horsepower was also significant here. When if was greater than or equal to 5500cc (*v12*), the probability of having a claim increased. Finally, driving in an urban area (*v2*) became significant and increased the expected number of claims. As regards the driving zone and driving experience, the sign of the coefficient changed for *v4* and *v5* variables with respect to third-party liability claims.

In order to model the covariance term (λ_3), the covariates were introduced in the bivariate Poisson model with the result that only the parameter for *v10* was significant. In Table 4 the results for this bivariate model with covariate on λ_3 are shown. The improvement in AIC with respect to the bivariate model with constant λ_3 can be observed. However, no substantial differences regarding the coefficients were found with the previous bivariate Poisson models from Table 3. When the policy included comprehensive coverage(*v10*), the correlation between N_1 and N_2 is increased since the parameter for *v10* is positive.

Finally, looking at the entries of Table 2, it is clear that the proportion of $(0,\ 0)$ is larger than that of other frequencies. Therefore, as it was mentioned in Section 2, two additional models were fitted using zero-inflated bivariate Poisson models. In Table 5 the results for these models are shown, the model with constant λ_3 on the left-hand side and the model with regressor (*v10*) on λ_3 on the right-hand side.

The parameter p referring to this zero-inflated model was significant and relatively large. Moreover, the AIC values improved substantially with respect to those of the non zero-inflated models. This suggests that the use of a zero-inflated model is a good choice for fitting this database (Boucher and Denuit, 2008). Other models with inflation in diagonal were fitted, but they were rejected because of the non significance of the respective elements of parameter vector $\boldsymbol{\theta}$. Thus, the existence of a higher proportion of $(1,\ 1)$ or $(2,\ 2)$ cannot be considered for this database.

4.2 Comparing *a priori* ratemaking when introducing dependence

An analysis of the impact of using these models in *a priori* ratemaking was conducted at the same time as the differences between the models proposed in Section 3 were analysed through the mean (*a priori* pure premium) and the variance (necessary for *a priori* loaded premium) of the number of claims per year for some profiles of the insured parties.

Five different, yet representative, profiles were selected from the portfolio (Table 6). The first can be classified as the best profile since it presents the lowest mean score. The second was chosen from among the profiles considered as good drivers, with a lower mean value than that of the average for the portfolio (0.1833). A profile with a mean lying very close to this average was chosen for the third profile. Finally, a profile considered as being a bad driver (with a mean above the average) and the worst driver profile were selected.

Table 7 shows the results for the five profiles and the five models considered. From these results, the differences in ratemaking when using a bivariate Poisson model as opposed to two independent Poisson models can be observed. In general, without distinguishing between bivariate models, such models produce higher means for good risks and lower means for bad risks while maintaining almost equal the average risks. As regards variances, the bivariate models increased them in most cases. A further difference that should be emphasized with the double Poisson model is the overdispersion detected in the bivariate models.

In Table 7, it can be observed that the zero-inflated bivariate models did not present any noticeable differences with the non zero-inflated models in terms of the mean scores, but they were present in the case of the variance. The bivariate Poisson models (BP1 and BP2) increased the variances for the good risks more than they did for the bad ones, while the zero-inflated bivariate models (ZIBP1 and ZIBP2) increased the variances much more for the bad risks.

Finally, the differences between the bivariate models with constant λ_3 (BP1 and ZIBP1) and those that included a covariate on λ_3 (BP2 and ZIBP2) were examined. A comparison of non zero-inflated models showed that the model including covariate (BP2) presented a mean and variance lower than those presented by the BP1 model for good risks, yet higher than those presented by the BP1 model for bad risks. However, no differences were detected between zero-inflated models.

5 Conclusions

This paper has tested the independence assumption between claim types given a set of known risk factors and it has shown that independence should be rejected. The bivariate Poisson model is presented as an instrument that can account for the underlying connection between two types of claims arising from the same policy[3]. The interpretation of a number of bivariate Poisson models has been illustrated in the context of automobile insurance claims and the conclusion is that using a bivariate Poisson model leads to an *a priori* ratemaking that presents larger variances and, hence, larger loadings than those obtained under the independence assumption.

For the five models analysed here there seems to be a relationship between the goodness of fit and the level of overdispersion considered in each model. For the double Poisson model, where the expected value and the variance (conditional on the risk factors) are equal for both the marginal (N_1 and N_2) and the joint (N) distributions, the lowest goodness of fit was obtained according to the AIC criterion. An improvement in the fit was achieved by using the bivariate Poisson model, which considers overdispersion only for the joint distribution since the marginal distributions are Poisson distributed. Finally, the highest goodness of fit is observed for the zero-inflated models where overdispersion is allowed both in the marginal and in the joint distributions.

In short, the main finding is that the independence assumption that is implicitely used when pricing automobile insurance by adding the pure premium for each guarantee (which are obtained using count data regression models) is insufficient because correlations (conditional on the covariates) are ignored. A natural extension for this paper would be to identify other multivariate count data models that might consider correlations in pricing several guarantees simultaneously in automobile insurance.

[3] In Frees and Valdez (2008) a hierarchical model allows to capture possible dependencies of claims among the various types through a t-copula specification.

6 References

Ambagaspitiya, R.S., 1999. On the distributions of two classes of correlated aggregate claims. Insurance: Mathematics & Economics 24 (3), 301–308.

Bermúdez, Ll., Denuit, M., Dhaene, J., 2001. Exponential bonus-malus systems integrating a priori risk classification. Journal of Actuarial Practice 9, 67–98.

Bolancé, C., Guillén, M., Pinquet, J., 2003. Time-varying credibility for frequency risk models: Estimation and tests for autoregressive specification on the random effect. Insurance: Mathematics & Economics 33 (2), 273–282.

Boucher, J.-Ph., Denuit, M., 2006. Fixed versus random effects in Poisson regression models for claim counts: a case study with motor insurance. ASTIN Bulletin 36 (1), 285–301.

Boucher, J.-Ph., Denuit, M., Guillén, M., 2007. Risk classification for claims counts: a comparative analysis of various zero-inflated mixed Poisson and Hurdle models. North American Actuarial Journal 11 (4), 110–131.

Boucher, J.-Ph., Denuit, M., 2008. Credibility premiums for the zero-inflated Poisson model and new hunger for bonus interpretation. Insurance: Mathematics & Economics 42 (2), 727–735.

Brouhns, N., Denuit, M., Guillén, M., Pinquet J., 2003. Bonus-malus scales in segmented tariffs with stochastic migration between segments. Journal of Risk and Insurance 70, 577–599.

Cameron, A.C., Trivedi, P.K., 1998. Regression analysis of count data. Econometric Society Monograph No.30, Cambridge University Press.

Centeno, M.L., 2005. Dependent risks and excess of loss reinsurance. Insurance: Mathematics & Economics 37 (2), 229–238.

Cummins, D.J., Wiltbank, L.J., 1983. Estimating the total claims distribution using multivariate frequency and severity distributions. Journal of Risk and Insurance 50, 377–403.

Denuit, M., Maréchal, X., Pitrebois, S., Walhin, J.F., 2007. Actuarial modelling of claim counts. John Wiley & Sons, London.

Dionne, G., Vanasse, C., 1989. A generalization of actuarial automobile insurance rating models: the Negative Binomial distribution with a regression component. ASTIN Bulletin 19 (2), 199–212.

Dobson, A.J., 1990. An introduction to generalized linear models. Chapman & Hall/CRC, London.

Fress, E.W., Valdez, E.A., 2008. Hierarchical insurance claims modeling. Journal of the American Statistical Association. To appear.

Haberman, S., Renshaw, A., 1996. Generalized linear models and actuarial science. The Statistician 45 (4), 407–436.

Ho, L., Singer, J., 2001. Generalized least squares methods for bivariate Poisson regression. Communications in Statistics-Theory and Methods 30, 263–277.

Holgate, P., 1964. Estimation for the bivariate Poisson distribution. Biometrika 51 (1/2), 241–245.

Johnson, N., Kotz, S., Balakrishnan, N., 1997. Discrete multivariate distributions. Wiley - New York.

Jung, R.C., Winkelmann, R., 1993. Two aspects of labor mobility: a bivariate Poisson regression approach. Empirical Economics 18 (3), 543–556.

Karlis, D., Meligkotsidou, L., 2005. Multivariate Poisson regression with full covariance structure. Statistics and Computing 15 (4), 255–265.

Karlis, D., Ntzoufras, I., 2003. Analysis of sports data using bivariate Poisson models. Journal of the Royal Statistical Society (Statistician) 52, 381–393.

Karlis, D., Ntzoufras, I., 2005. Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. Journal of Statistical Software 14 (10), 1–36.

Kocherlakota, S., Kocherlakota, K., 1992. Bivariate discrete distributions. New York: Marcel Dekker.

Kocherlakota, S., Kocherlakota, K., 2001. Regression in the bivariate Poisson distribution. Communications in Statistics-Theory and Methods 30, 815–827.

Li, C., Lu, J., Park, J., Kim, K., Peterson, J., 1999. Multivariate zero-inflated Poisson models and their applications. Technometrics 41, 29–38.

Mc Cullagh, P., Nelder, J.A., 1989. Generalized linear models. Chapman & Hall, New York.

Partrat, C., 1994. Compound model for two dependent kinds of claim. Insurance: Mathematics & Economics 15 (2-3), 219–231.

Pinquet, J., 1999. Experience rating through heterogeneous models. In Handbook of Insurance, edited by G. Dionne. Kluwer Academic Publishers.

Pinquet, J., Guillén, M., Bolancé, C., 2001. Long-range contagion in automobile insurance data: estimation and implications for experience rating. ASTIN Bulletin 31 (2), 337–348.

Tsionas, E., 2001. Bayesian multivariate Poisson regression. Communications in Statistics-Theory and Method 30, 243–255.

Vernic, R., 1997. On the bivariate generalized Poisson distribution. ASTIN Bulletin 27 (1), 23–31.

Walhin, J.F., Paris, J. 2000. Recursive formulae for some bivariate counting distributions obtained by the trivariate reduction method. ASTIN Bulletin 30 (1), 141–155.

Walhin, J.F., 2001. Bivariate ZIP models. Biometrical Journal 43, 147–160.

Walhin, J.F., 2003. Bivariate Hofmann distributions. Journal of Applied Statistics 30 (9), 1033–1046.

Wang, K., Lee, A., Yau, K., Carrivick, P., 2003. A bivariate zero inflated poisson regression model to analyze occupational injuries. Accident Analysis and Prevention 35, 625–629.

Table 1: Explanatory variables used in the model

Variable	Definition
v1	equals 1 for women and 0 for men
v2	equals 1 when driving in urban area, 0 otherwise
v3	equals 1 when zone is medium risk (Madrid and Catalonia)
v4	equals 1 when zone is high risk (Northern Spain)
v5	equals 1 if the driving license is between 4 and 14 years old
v6	equals 1 if the driving license is 15 or more years old
v7	equals 1 if the client is in the company between 3 and 5 years
v8	equals 1 if the client is in the company for more than 5 years
v9	equals 1 of the insured is 30 years old or younger
v10	equals 1 if includes comprehensive coverage (except fire)
v11	equals 1 if includes comprehensive and collision coverages
v12	equals 1 if horsepower is greater than or equal to 5500cc

Table 2: Cross-tabulation of data

	N_2							
N_1	0	1	2	3	4	5	6	7
0	71087	3722	807	219	51	14	4	0
1	3022	686	184	71	26	10	3	1
2	574	138	55	15	8	4	1	1
3	149	42	21	6	6	1	0	1
4	29	15	3	2	1	1	0	0
5	4	1	0	0	0	0	2	0
6	2	1	0	1	0	0	0	0
7	1	0	0	1	0	0	0	0
8	0	0	1	0	0	0	0	0

$N_1:$ number of claims for third-party liability.
$N_2:$ number of claims for the rest of guarantees.

Table 3: Results for bivariate Poisson and double Poisson models

		Bivariate Poisson (BP1)			Double Poisson (DP)		
	Variables	Coeff.	St. Err.	P-value	Coeff.	St. Err.	P-value
λ_1	Intercept	-2.380	0.126	< 0.01	-2.329	0.103	< 0.01
	v1	0.011	0.051	0.822	-0.003	0.042	0.935
	v2	-0.049	0.036	0.179	-0.050	0.035	0.145
	v3	0.016	0.041	0.695	0.002	0.038	0.958
	v4	0.157	0.040	< 0.01	0.180	0.037	< 0.01
	v5	-0.228	0.100	0.022	-0.217	0.092	0.018
	v6	-0.352	0.110	< 0.01	-0.345	0.104	< 0.01
	v8	0.154	0.043	< 0.01	0.151	0.042	< 0.01
	v9	0.139	0.064	0.031	0.115	0.063	0.068
	v10	-0.302	0.054	< 0.01	0.061	0.053	0.252
	v11	-0.061	0.039	0.120	0.062	0.033	0.062
	v12	0.045	0.045	0.323	0.053	0.038	0.157
λ_2	Intercept	-4.822	0.146	< 0.01	-4.436	0.116	< 0.01
	v1	0.060	0.037	0.107	0.044	0.039	0.264
	v2	0.081	0.032	0.010	0.066	0.032	0.040
	v3	0.172	0.034	< 0.01	0.151	0.033	< 0.01
	v4	-0.220	0.042	< 0.01	-0.146	0.038	< 0.01
	v5	0.324	0.127	0.010	0.273	0.097	< 0.01
	v6	0.121	0.131	0.355	0.077	0.104	0.458
	v8	0.248	0.042	< 0.01	0.235	0.038	< 0.01
	v9	0.098	0.052	0.060	0.085	0.051	0.096
	v10	3.114	0.055	< 0.01	2.887	0.055	< 0.01
	v11	2.150	0.054	< 0.01	1.948	0.053	< 0.01
	v12	0.170	0.053	< 0.01	0.165	0.045	< 0.01
$\lambda 3$	Intercept	-4.437	0.035	< 0.01			
Log-likelihood				-48135.98			-48882.95
AIC				96321.96			97813.9

Table 4: Results for bivariate Poisson model with regressor on λ_3

		Bivariate Poisson (BP2)		
	Variables	Coeff.	St. Err.	P-value
λ_1	Intercept	-2.383	0.129	< 0.01
	v1	0.012	0.052	0.811
	v2	-0.050	0.035	0.154
	v3	0.020	0.042	0.640
	v4	0.157	0.045	< 0.01
	v5	-0.226	0.101	0.025
	v6	-0.348	0.114	< 0.01
	v8	0.154	0.049	< 0.01
	v9	0.146	0.070	0.037
	v10	-0.658	0.067	< 0.01
	v11	-0.032	0.038	0.400
	v12	0.045	0.043	0.293
λ_2	Intercept	-4.823	0.137	< 0.01
	v1	0.062	0.040	0.123
	v2	0.084	0.031	< 0.01
	v3	0.179	0.030	< 0.01
	v4	-0.228	0.041	< 0.01
	v5	0.332	0.114	< 0.01
	v6	0.129	0.120	0.280
	v8	0.249	0.038	< 0.01
	v9	0.102	0.047	0.029
	v10	3.043	0.061	< 0.01
	v11	2.158	0.060	< 0.01
	v12	0.172	0.049	< 0.01
$\lambda 3$	Intercept	-4.867	0.051	< 0.01
	v10	1.767	0.075	< 0.01
Log-likelihood				-47873.37
AIC				95798.74

Table 5: Results for zero-inflated bivariate Poisson models

		Z-I Bivariate Poisson			Z-I Bivariate Poisson		
		constant λ_3 (ZIBP1)			covariate on λ_3 (ZIBP2)		
	Variables	Coeff.	St. Err.	P-value	Coeff.	St. Err.	P-value
λ_1	Intercept	-1.041	0.111	< 0.01	-1.055	0.130	< 0.01
	v1	-0.008	0.047	0.874	0.001	0.047	0.981
	v2	-0.064	0.035	0.065	-0.063	0.037	0.088
	v3	-0.033	0.035	0.345	-0.024	0.044	0.582
	v4	0.211	0.046	< 0.01	0.203	0.041	< 0.01
	v5	-0.254	0.091	< 0.01	-0.249	0.114	0.029
	v6	-0.357	0.102	< 0.01	-0.362	0.126	< 0.01
	v8	0.127	0.047	< 0.01	0.135	0.045	< 0.01
	v9	0.099	0.072	0.170	0.105	0.075	0.162
	v10	-0.054	0.055	0.323	-0.255	0.068	< 0.01
	v10	0.044	0.037	0.227	0.046	0.036	0.205
	v12	0.044	0.041	0.284	0.045	0.041	0.275
λ_2	Intercept	-3.253	0.120	< 0.01	-3.269	0.125	< 0.01
	v1	0.023	0.030	0.446	0.031	0.035	0.367
	v2	0.048	0.024	0.047	0.052	0.025	0.037
	v3	0.108	0.024	< 0.01	0.118	0.031	< 0.01
	v4	-0.095	0.037	0.010	-0.114	0.037	< 0.01
	v5	0.216	0.099	0.030	0.227	0.091	0.012
	v6	0.043	0.102	0.676	0.044	0.092	0.630
	v8	0.184	0.032	< 0.01	0.191	0.034	< 0.01
	v9	0.049	0.047	0.302	0.053	0.045	0.239
	v10	2.917	0.059	< 0.01	2.855	0.061	< 0.01
	v11	2.057	0.059	< 0.01	2.050	0.062	< 0.01
	v12	0.178	0.052	< 0.01	0.180	0.043	< 0.01
λ_3	Intercept	-4.741	0.152	< 0.01	-4.879	0.140	< 0.01
	v10				1.962	0.191	< 0.01
p		0.714	0.005	< 0.01	0.710	0.006	< 0.01
Log-likelihood				-45435.06			-45414.80
AIC				90922.11			90883.6

Table 6: Five different policyholders to be compared

Profile	Kind of Profile	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12
1	Best	0	1	0	0	0	1	0	1	0	0	0	0
2	Good	0	0	1	0	0	1	1	0	0	0	0	1
3	Average	0	1	0	0	0	1	0	1	0	0	1	1
4	Bad	0	0	0	1	0	1	0	1	0	1	0	0
5	Worst	1	1	1	0	1	0	1	0	1	1	0	1

Table 7: Comparision of *a priori* ratemaking

	1st Profile		2nd Profile		3rd Profile		4th Profile		5th Profile	
Model	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
BP1	0.0955	0.1191	0.1207	0.1444	0.1849	0.2086	0.2440	0.2677	0.6725	0.6962
BP2	0.0873	0.1027	0.1131	0.1285	0.1804	0.1958	0.2824	0.3726	0.6920	0.7821
DP	0.0793	0.0793	0.1070	0.1070	0.1866	0.1866	0.2860	0.2860	0.6969	0.6969
ZIBP1	0.0834	0.1057	0.1046	0.1369	0.1905	0.2861	0.2816	0.4845	0.5500	1.3103
ZIBP2	0.0826	0.1037	0.1055	0.1371	0.1898	0.2822	0.2771	0.4963	0.5562	1.3440