Gómez-Déniz, Emilio; Calderín-Ojeda, Enrique

Article

A priori ratemaking selection using multivariate regression models allowing different coverages in auto insurance

Risks

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Gómez-Déniz, Emilio; Calderín-Ojeda, Enrique (2021) : A priori ratemaking selection using multivariate regression models allowing different coverages in auto insurance, Risks, ISSN 2227-9091, MDPI, Basel, Vol. 9, Iss. 7, pp. 1-18, https://doi.org/10.3390/risks9070137

risks

MDPI

Article

A Priori Ratemaking Selection Using Multivariate Regression Models Allowing Different Coverages in Auto Insurance

Emilio Gómez-Déniz [1,*] and Enrique Calderín-Ojeda [2]

[1] Department of Quantitative Methods, Faculty of Economics, University of Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain
[2] Department of Economics, University of Melbourne, Melbourne 3031, Australia; enrique.calderin@unimelb.edu.au
* Correspondence: emilio.gomez-deniz@ulpgc.es

Abstract: A comprehensive auto insurance policy usually provides the broadest protection for the most common events for which the policyholder would file a claim. On the other hand, some insurers offer extended third-party car insurance to adapt to the personal needs of every policyholder. The extra coverage includes cover against fire, natural hazards, theft, windscreen repair, and legal expenses, among some other coverages that apply to specific events that may cause damage to the insured's vehicle. In this paper, a multivariate distribution, based on a conditional specification, is proposed to account for different numbers of claims for different coverages. Then, the premium is computed for each type of coverage separately rather than for the total claims number. Closed-form expressions are given for moments and cross-moments, parameter estimates, and for a priori premiums when different premiums principles are considered. In addition, the severity of claims can be incorporated into this multivariate model to derive multivariate claims' severity distributions. The model is extended by developing a zero-inflated version. Regression models for both multivariate families are derived. These models are used to fit a real auto insurance portfolio that includes five types of coverage. Our findings show that some specific covariates are statistically significant in some coverages, yet they are not so for others.

Keywords: automobile insurance; conditional distribution; coverage; insurance pricing; multivariate zero-inflated models; regression

Citation: Gómez-Déniz, Emilio, and Enrique Calderín-Ojeda. 2021. A Priori Ratemaking Selection Using Multivariate Regression Models Allowing Different Coverages in Auto Insurance. *Risks* 9: 137. https://doi.org/10.3390/risks9070137

Academic Editor: Mogens Steffensen

Received: 9 June 2021
Accepted: 14 July 2021
Published: 20 July 2021

1. Introduction

In the automobile insurance sector, it is natural to calculate the a priori premium taking into account the number of claims and individual characteristics of each insured, such as gender, age, years of validity of the policy, etc. This procedure to compute the a priori premium is usually completed via parametric models rather than using the ordinary regression model, which can predict values of the number of claims even if negative. For this purpose, parametric models based on the use of the Poisson, negative binomial, and Poisson-inverse Gaussian distributions, among others, are the standard models considered in the univariate case. As of today, most insurance companies distinguish, apart from the total number of claims, individualized claims for different coverages, such as windscreen claims, thefts and fire claims, etc. So far, most actuarial models aim to differentiate only between two types of coverage when computing an appropriate premium based on different coverage. Perhaps one of the reasons for that is due to the lack of models capable of describing more than two coverages. The most often considered approach to tackle this problem is the one based on the bivariate Poisson distribution (see Bermúdez 2009; Bermúdez and Karlis 2017, among others). See also Gómez-Déniz (2016); Gómez-Déniz and Calderín-Ojeda (2018); Gómez-Déniz and Calderín-Ojeda (2020); Denuit et al. (2009), and Frees (2010) for more details related to this topic. Alternative references for a review of count regression are Cameron and Trivedi (1986); Cameron and Trivedi (1998); Winkelmann (2003), and Boucher et al. (2007).

A copula-based correlated random effects model that accommodates dependence between claim frequency and severity was examined in Oh et al. (2020).

Traditionally, the business associated with insurance consists of selling risk coverage to buyers. In particular, in automobile insurance, the insurer provides financial protection against physical damage or bodily injury resulting from an incident (see Frees et al. 2016). However, it is common today, mainly due to the existing competition, that the insurance companies offer coverage of different claims within the same product not only to gain in competitiveness, but also to benefit from risk diversification and volatility. In this paper, we consider a motor vehicle insurance portfolio with policies observed during some time period that contain, apart from other known factors (gender, age, years of validity of the driver's license, etc.), information about the claims number concerning different coverages that are considered as response variables. This includes windscreen, parking, theft and fire, etc.

Therefore, it is assumed that the insurance company collects information on the claims for these coverages and the total number of claims given by the sum of the claims in all the coverages. Thus, every policyholder generates a sequence of claims numbers for each coverage; one of them is the total claims number, which includes the sum of the coverages' claims. Then, based on a conditional specification, a multivariate model that allows a simple way to describe the use of a finite but sufficiently large number of coverages is proposed. The resulting multivariate discrete distribution obtained enables us to study the dependence structure of a limited number of coverages in automobile insurance and include covariates such as gender, age, etc. We start by using a Poisson model for the random variable total claims number, and then by conditioning, we introduce the remaining variables in a branch architecture structure. Finally, closed-form expressions are given for parameter estimates, and a priori premiums are provided when different premium principles are used.

The purpose of this paper is to introduce a novel methodology based on a multivariate distribution via a conditional specification, proposed to account for different numbers of claims in different coverages and also for the total claims frequency. This approach enable us to examine the dependence structure of a finite number of coverages in motor vehicle insurance and also incorporate heterogeneity in the model through explanatory variables. Then, we use this procedure to calculate premiums based only on the claims frequency. Next, we show that the amount of claims can be incorporated into this multivariate model to derive multivariate claims' severity distributions. For this, we assume that the claims size in the joint coverages follows a multivariate Erlang distribution. As multivariate probability distributions are complex, it is argued that analytical solutions are highly unlikely as compared to those derived under univariate and bivariate cases (see Cummins and Wiltbank 1983, 1984); nevertheless, in this work, we derive a multivariate model where the total number of claims that affect the portfolio is the result of the interaction of multivariate processes. The main advantage of the modelization presented in this work is that it avoids working with copulas (see, for instance, Balakrishnan and Lai 2009, chp. 1, p. 59). Although the copula approach for modeling multivariate models has been proven to be very useful, it has also been criticized due to the difficulty of choosing an appropriate copula structure and the complication of estimating the parameters that control the dependency. In addition, a multivariate zero-inflated model to account for the excess of common zeros in the empirical distribution is developed. Finally, these two multivariate distributions can be reparameterized to incorporate covariates to determine which factors and explanatory variables have an influence on the mean of the corresponding coverage. As an illustration, in this work, we use the French Motor Personal Line datasets available in the package "CASdatasets" in **R**, which include five response variables.

Although the modeling proposed here was developed ad hoc for the auto insurance market, it is unquestionable that other insurance lines in general insurance might benefit from it. For example, in home insurance, the whole premium could be split into different

coverages such as moisture damage, theft, pipe repairs, locksmiths, and even protection against tenant rent default.

The rest of the paper is structured as follows. Section 2 describes the primary model and some of its properties. Then, premium calculations based on this basic model are discussed. Finally, a multivariate zero-inflated model and multivariate regression procedures are shown. Some methods of estimation are provided in Section 3. Next, a numerical application pertaining to a private motor French insurer is developed in Section 4. Finally, conclusions are drawn in Section 5.

2. The Branch Architecture Model

Let us consider a portfolio with N observed policies during T periods of time and also assume that the insurance company gathers information on the number of claims related to several types of coverages. Example of these coverages may include windscreens, fire and theft, etc. Therefore, the insurer collects information about these coverages, as well as the total number of claims for each policyholder given by the sum of the claims in all different coverages. For the ith policyholder, we consider the multivariate random variable expressed as the following sequence, $N_{ji} = (N_{ji1}, N_{ji2}, \ldots, N_{jiT})'$ of claims numbers for coverage j, with $j = 1, 2, \ldots, \mathcal{J}$, assuming that one of them, i.e., the first one, is the total number of claims, which includes the sum of the claims for all types of coverages purchased by this policyholder.

Furthermore, we assume that N_{1i}, the total number of claims recorded in the auto insurance portfolio, follows a Poisson distribution with mean $\Theta_{1i} > 0$ for $i = 1, 2, \ldots, N$, where N is the total number of policyholders. Now, let us suppose that the policyholders have purchased some of the types of coverages, such as windscreen protection, fire and theft, parking, etc. That is, once the policyholder has made a claim, this can be of any of these types. Let us denote by $Z_{1i\kappa}$, $\kappa = 1, \ldots, N_{1i}$, a random variable associated with the number of claims corresponding to the first type of coverage and policyholder i, resulting from the κth claim of the total claims reported by the ith policyholder assumed to be independent and identically distributed, following also a Poisson distribution with mean $\Theta_{2i} > 0$. Then, the conditional distribution of N_{2i} given $N_{1i} = n_{1i}$, $N_{2i} = \sum_{\kappa=1}^{N_{1i}} Z_{1i\kappa}$, the total number of claims of this first coverage, among the N_{1i} total claims is a Poisson distribution with parameter $n_{1i}\Theta_{2i}$ and the joint distribution of (N_{1i}, N_{2i}) has a probability function given by,

$$\begin{aligned} f(n_{1i}, n_{2i}|\Theta_{1i}, \Theta_{2i}) &= \Pr(n_{2i}|n_{1i}, \Theta_{2i}) \Pr(n_{1i}|\Theta_{1i}) \\ &= \frac{1}{n_{1i}! n_{2i}!} \Theta_{1i}^{n_{1i}} (\Theta_{2i} n_{1i})^{n_{2i}} \exp[-(\Theta_{1i} + \Theta_{2i} n_{1i})], \end{aligned}$$

for n_{1i}, $n_{2i} = 0, 1, \ldots,$ and $i = 1, 2, \ldots, N$ and with the convention that $0^j = 0$ for $j = 0$ and 1 otherwise. This bivariate distribution appears in Leiter and Hamdan (1973) (see also Cacaoullos and Papageorgiou 1980; Johnson et al. 1996, chp. 37, p. 136 in the context of accident analysis)

Let $Z_{2i\kappa}$, $\kappa = 1, \ldots, N_{1i}$ now be a random variable associated with the total number of claims corresponding to the second type of coverage and policyholder i, resulting from the κth claim of the total claims reported by the ith policyholder assumed to be independent and identically distributed Poisson distribution with mean $\Theta_{3i} > 0$ and conditionally independent of N_{2i}. Then, the conditional distribution of N_{3i} given $N_{1i} = n_{1i}$, $N_{3i} = \sum_{\kappa=1}^{N_{1i}} Z_{2i\kappa}$, the total number of claims of this second coverage, among the N_{1i} total claims is a Poisson distribution with parameter $n_{1i}\Theta_{3i}$, and now, the joint distribution of (N_{1i}, N_{2i}, N_{3i}) has a probability function given by,

$$\begin{aligned} f(n_{1i}, n_{2i}, n_{3i}|\Theta_{1i}, \Theta_{2i}, \Theta_{3i}) &= \Pr(n_{1i}, n_{2i}) \Pr(n_{3i}|(n_{1i}, n_{2i})) = \Pr(n_{1i}, n_{2i}) \Pr(n_{3i}|n_{1i}) \\ &= \frac{\Theta_{1i}^{n_{1i}}}{n_{1i}!} \prod_{j=2}^{3} \frac{(\Theta_{ji} n_{1i})^{n_{ji}}}{n_{ji}!} \exp\left[-\left(\Theta_{1i} + n_{1i} \sum_{j=2}^{3} \Theta_{ji}\right)\right], \end{aligned}$$

where the hypotheses of conditional independence between the two types of coverages were assumed.

Following the same argument, it is easy to see that if we have $\mathcal{J}$ types of coverages, then the joint probability function of $(N_{1i}, N_{2i}, \ldots, N_{\mathcal{J}i}) \in \mathbb{N}^{\mathcal{J}}$ is given by:

$$f(n_{1i}, \ldots, n_{\mathcal{J}i}|\Theta_i) = \frac{\Theta_{1i}^{n_{1i}}}{n_{1i}!} \prod_{j=2}^{\mathcal{J}} \frac{(\Theta_{ji} n_{1i})^{n_{ji}}}{n_{ji}!} \exp\left[-\left(\Theta_{1i} + n_{1i} \sum_{j=2}^{\mathcal{J}} \Theta_{ji}\right)\right], \tag{1}$$

where $\Theta_i = (\Theta_{1i}, \ldots, \Theta_{\mathcal{J}i})$. For this multivariate distribution, it is allowed that n_{1i} takes larger or smaller values than n_{ji}; however, in the proposed model, it is verified that n_{1i} is larger than or equal to $n_{\mathcal{J}i}$ for all $\mathcal{J} > 1$. In this case, it is obvious that Θ_1 must be larger than Θ_j, $j = 2, \ldots, \mathcal{J}$. The latter statement is confirmed in the numerical application section.

This distribution is a multivariate extension of the bivariate one proposed in Leiter and Hamdan (1973) (see also Cacaoullos and Papageorgiou 1980).

The ordinary probability-generating function of $(N_1, \ldots, N_{\mathcal{J}})$ with the probability mass function (pmf) given in (1) is given by:

$$G_{N_1,\ldots,N_{\mathcal{J}}}(z_1, \ldots, z_{\mathcal{J}}) = \exp\left\{\Theta_1 \left[z_1 \exp\left(\sum_{j=2}^{\mathcal{J}} \Theta_j (z_j - 1)\right) - 1\right]\right\},$$

for $|z_j| \leq 1$, $j = 1, \ldots, \mathcal{J}$.

From here, it is easy to see that the marginal distribution of N_{1i} is Poisson with parameter Θ_{1i}, while N_{ji}, $j = 1, \ldots, \mathcal{J}$ have a Neyman Type A distribution with parameters Θ_{1i} and Θ_{ji}. Recall that the probability function of the Neyman Type A distribution (see Neyman 1939; Douglas 1955; Kemp 1967; Johnson et al. 2005, chp. 8, among others) is given by:

$$f(n_{ji}) = \frac{\Theta_{ji}^{n_{ji}} \exp(-\Theta_{1i})}{n_{ji}!} \sum_{n_{1i}=0}^{\infty} \left[\Theta_{1i} \exp(-\Theta_{ji})\right]^{n_{1i}} \frac{n_{1i}^{n_{ji}}}{n_{1i}!}, \quad n_{ji} = 0, 1, \ldots, \tag{2}$$

for $j = 1, \ldots, \mathcal{J}$.

Some computations provide the marginal and cross-moments, which are given by:

$$\begin{aligned}
E(N_{1i}|\Theta_{1i}) &= \Theta_{1i}, & (3)\\
E(N_{ji}|\Theta_{1i}, \Theta_{ji}) &= \Theta_{1i}\Theta_{ji}, \quad j = 2, \ldots, \mathcal{J}, & (4)\\
E(N_{1i}N_{ji}|\Theta_{1i}, \Theta_{ji}) &= \Theta_{1i}(1 + \Theta_{1i})\Theta_{ji}, \quad j = 2, \ldots, \mathcal{J},\\
E(N_{ji}N_{li}|\Theta_{1i}, \Theta_{ji}) &= \Theta_{1i}(1 + \Theta_{1i})\Theta_{ji}\Theta_{li}, \quad j, l = 2, \ldots, \mathcal{J},\ j \neq l,
\end{aligned}$$

from which it is simple to see that:

$$\begin{aligned}
cov(N_{1i}, N_{ji}) &= \Theta_{1i}\Theta_{ji}, \quad j = 2, \ldots, \mathcal{J}\\
cov(N_{ji}, N_{li}) &= \Theta_{1i}\Theta_{ji}\Theta_{li}, \quad j, l = 2, \ldots, \mathcal{J},\ j \neq l,
\end{aligned}$$

and therefore, the model admits only a positive correlation between pairs of random variables. The marginal variances are given by:

$$\begin{aligned}
var(N_{1i}|\Theta_{1i}) &= \Theta_{1i}, & (5)\\
var(N_{ji}|\Theta_{1i}, \Theta_{ji}) &= \Theta_{1i}\Theta_{ji}(1 + \Theta_{ji}), \quad j = 2, \ldots, \mathcal{J}. & (6)
\end{aligned}$$

Observe that, using (5) and (6) together with (3) and (4), the model is equidispersed (variance equal to the mean) for N_{1i} and overdispersed (variance larger than the mean) for the rest of the coverages.

Finally, the correlation can easily be computed as:

$$\rho(N_{1i}, N_{ji}) = \sqrt{\frac{\Theta_{ji}}{1+\Theta_{ji}}}, \quad j = 2, \ldots, \mathcal{J}, \tag{7}$$

$$\rho(N_{ji}, N_{li}) = \sqrt{\frac{\Theta_{ji}\Theta_{li}}{(1+\Theta_{ji})(1+\Theta_{li})}}, \quad j, l = 2, \ldots, \mathcal{J},\ j \neq l. \tag{8}$$

One can be interested also in the distribution of N_{1i} given $N_{ji} = n_{ji}$. The probability-generating function of this conditional distribution is given by:

$$G_{N_{1i}|N_{ji}=n_{ji}}(z) = \exp\left[\Lambda_j(z-1)\right] \frac{B_{n_{ji}}(\Lambda_j z)}{B_{n_{ji}}(z)}, \quad |z| \leq 1,$$

where $\Lambda_j = \Theta_{1i} \exp\left[-\Theta_{ji}\right]$ and $B_n(\tau)$ are the Bell numbers given by:

$$B_n(\tau) = \sum_{k=0}^{n} S(n,k)\tau^k,$$

with:

$$S(n,k) = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^n$$

being the Stirling number of the second kind.[1]

Now, the conditional mean of N_{1i} given by $N_{ji} = n_{ji}$ can be written as:

$$E\left(N_{1i}|N_{ji} = n_{ji}\right) = \frac{B_{n_{ji}+1}(\Lambda)}{B_{n_{ji}}(\Lambda)}.$$

2.1. Some Results in Risk Theory

Observe that due to the model construction, we have that $\sum_{j=2}^{\mathcal{J}} \Theta_{ji} = 1$, i.e., every claim in coverage j is a proportion of the total claims N_{1i}. Then, if the actuary decides to use the net premium principle, i.e., $P(X) = E(X)$, to compute the premium, then for the ith policyholder and coverage s with $s \in \{2, \ldots, \mathcal{J}\}$, the premium results $P_{si} = \Theta_{1i}\Theta_{si} = P_{1i}\Theta_{si}$, where $P_{1i} = \Theta_{1i}$ is the net premium for the total coverage, that is the sum of the premiums in each of the coverages purchased. A similar result is obtained by using the expected value principle. A catalog of premium principles can be found in Young (2006).

Let us now consider that the actuary decides to use the variance premium principle, i.e., $P(X) = E(X) + var(X)/E(X)$ with $E(X) > 0$, to calculate the premium. Then, in this case, we obtain that $P_{1i} = 1 + \Theta_{1i}$ and $P_{ji} = 1 + \Theta_{ji}P_{1i}$. However, in this case, we have that $\sum_{j=2}^{\mathcal{J}} P_{ji} = \mathcal{J} - 1 + P_{1i}$, which is different from P_{1i}, except for the case in which $\mathcal{J} = 1$ and no coverages exist.

However, a model solely based on the number of claims is not realistic. In risk theory, it is common to incorporate the amount associated with each of the claims to build the compound model. That is, the property and/or casualty ratemaking are generally based on a claim frequency distribution and a loss distribution. Due to the complex derivation of this multivariate compound model, the subscript i is removed from the text in the remainder of this section. For this purpose, let us now assume that $Y_j = \sum_{i=1}^{N_j} E_i$, $j = 1, \ldots, \mathcal{J}$, where E_i is the random variable denoting the size or amount of the ith claim, following an exponential distribution with probability density function (pdf) $h(y_j) = \sigma^{-1} \exp(-y_j/\sigma)$. Furthermore, we assume that $E_1, E_2, \ldots$, are independent and identically distributed random variables and also independent of the number of claims N_j. It is well known (see for example Rolski et al. 1999) that Y_j follows a piecewise distribution with pdf given by

$g(y_j) = \sum_{n_j=1}^{\infty} f_{N_j}(n_j) h^{*n_j}(y_j)$, $y_j > 0$, and $g(0) = f_{N_j}(0)$. Then, by following the methodology given in Lee and Lin (2012), we have that $\boldsymbol{Y} = (Y_1, \ldots, Y_{\mathcal{J}})$ follows a multivariate Erlang distribution with scale parameter $\sigma > 0$ and shape parameter $n_j > 0$, $j = 1, 2, \ldots, \mathcal{J}$. Their marginal distributions are a univariate Erlang mixture.

Then, simple computations provide,

$$g(y_1) = \begin{cases} \sqrt{\frac{\Theta_1}{y_1 \sigma}} \exp\left[-\left(\Theta_1 + \frac{y_1}{\sigma}\right)\right] I_1\left(2\sqrt{\frac{\Theta_1 y_1}{\sigma}}\right), & y_1 > 0, \\ \exp(-\Theta_1), & y_1 = 0. \end{cases}$$

Here, $I_1(\cdot)$ represents the modified Bessel function of the first kind, which admits the following series representation,

$$I_\nu(z) = \sum_{k=0}^{\infty} \frac{1}{\Gamma(k+\nu+1)k!} \left(\frac{z}{2}\right)^{2k+\nu}.$$

The distribution for the coverages can be computed by using (2) in the following way,

$$g(y_j) = \sum_{n_j=1}^{\infty} \frac{\Theta_j^{n_j} \exp(-\Theta_1)}{n_j!} \sum_{n_1=0}^{\infty} \left[\Theta_1 \exp(-\Theta_j)\right]^{n_1} \frac{n_1^{n_j}}{n_1!} \frac{y_j^{n_j - 1} \exp(-y_j/\sigma)}{\sigma^{n_j} \Gamma(n_j)}.$$

Now, taking into account that:

$$\sum_{n_j=1}^{\infty} \frac{(\Theta_j n_1 y_j / \sigma)^{n_j}}{n_j! \Gamma(n_j)} = I_1\left(2\sqrt{n_1 y_j \Theta_j / \sigma}\right)$$

we finally obtain the aggregate claim size pdf for the different coverages given by,

$$g(y_j) = \begin{cases} \sqrt{\frac{\Theta_j}{y_j \sigma}} \exp\left[-\left(\Theta_1 + \frac{y_j}{\sigma}\right)\right] \sum_{n_1=0}^{\infty} \frac{\sqrt{n_1} \Lambda_j^{n_1}}{n_1!} I_1\left(2\sqrt{\frac{\Theta_j n_1 y_j}{\sigma}}\right), & y_j > 0, \\ \exp\left(-\Theta_1 + \Lambda_j\right), & y_j = 0, \end{cases}$$

for $j = 2, \ldots, \mathcal{J}$. Thus, they are also given as a piecewise distribution. For practical purposes, the infinite sum that appears in this expression can be replaced by a finite sum from one to k, where k can take values around one-hundred. From the assumption of the independence between the number of claims and the claims size, we have that:

$$\begin{aligned} E(Y_1) &= \sigma\Theta_1, \\ E(Y_j) &= \sigma\Theta_1\Theta_j, \quad j = 2, \ldots, \mathcal{J}, \end{aligned}$$

which can be considered as the net premium when both the number and size are considered at the same time.

Finally, we have that:

$$\begin{aligned} var(Y_1) &= 2\sigma^2\Theta_1, \\ var(Y_j) &= \sigma^2\Theta_1\Theta_j(2 + \Theta_j), \quad j = 2, \ldots, \mathcal{J}, \end{aligned}$$

while the covariance (see Lee and Lin 2012) is given by:

$$cov(Y_j, Y_l) = \sigma^2 cov(N_j, N_l), \quad j \neq l.$$

2.2. Multivariate Zero-Inflated Model

In many automobile insurance portfolios, the claims are rarely observed as compared to the no-claims situation. Univariate and bivariate zero-inflated models have been introduced in the statistical literature in many fields. In the setting of auto insurance, we refer to Boucher et al. (2007) and Frees et al. (2016) for the univariate case and Bermúdez (2009)

and Bermúdez and Karlis (2017) for the bivariate case. Multivariate ones are scarce in the general statistical literature. References in the statistical literature are Li et al. (1999) and Liu and Tian (2015). In the actuarial literature, there are no references of models of this nature that go beyond the two variables. However, multivariate zero-truncated models were considered in Zhang et al. (2020).

A multivariate zero-inflated model can be constructed as a mixture of the multivariate distribution given in (1) and a point mass at $(0,\ldots,0) \in \mathbb{R}^{\mathcal{J}}$ in the following way,

$$g(n_{1i},\ldots,n_{\mathcal{J}i}|\Theta_i) = \begin{cases} 1-\Phi+\Phi f(0,\ldots,0|\Theta_i), & (n_{1i},\ldots,n_{\mathcal{J}i}) = (0,\ldots,0), \\ \Phi f(n_{1i}\ldots,n_{\mathcal{J}i}|\Theta_i), & (n_{1i},\ldots,n_{\mathcal{J}i}) \neq (0,\ldots,0), \end{cases} \tag{9}$$

where $0 \leq \Phi \leq 1$ is an inflation parameter. Obviously, this model reduces to (1) for $\Phi = 1$. Under this model, the marginal means and cross-moments are given by:

$$\begin{aligned} E(N_{1i}|\Theta_{1i}) &= \Phi\,\Theta_{1i}, \\ E(N_{ji}|\Theta_{1i},\Theta_{ji}) &= \Phi\,\Theta_{1i}\Theta_{ji}, \quad j=2,\ldots,\mathcal{J}, \\ E(N_{1i}N_{ji}|\Theta_{1i},\Theta_{ji}) &= \Phi\,\Theta_{1i}(1+\Theta_{1i})\Theta_{ji}, \quad j=2,\ldots,\mathcal{J}, \\ E(N_{ji}N_{li}|\Theta_{1i},\Theta_{ji}) &= \Phi\,\Theta_{1i}(1+\Theta_{1i})\Theta_{ji}\Theta_{li}, \quad j,l=2,\ldots,\mathcal{J},\ j\neq l, \end{aligned}$$

from which the covariance between pairs of marginal random variables can be obtained. They are given by:

$$\begin{aligned} cov(N_{1i},N_{ji}) &= \Phi\,\Theta_{1i}\Theta_{ji}(1+\bar{\Phi}\,\Theta_{1i}), \quad j=2,\ldots,\mathcal{J} \\ cov(N_{ji},N_{li}) &= \Phi\,\Theta_{1i}\Theta_{ji}\Theta_{li}(1+\bar{\Phi}\,\Theta_{1i}), \quad j,l=2,\ldots,\mathcal{J},\ j\neq l, \end{aligned}$$

where $\bar{\Phi} = 1-\Phi$.

Again, if the actuary computes the premium by using the net premium principle, then for each coverage, the premiums are not affected by the inflation parameter Φ. A complete model would allow inflating each coverage with inflation parameters Φ_j; however, they are not included in this work due to the computational cost of estimating a large number of parameters.

The marginal variances are given by:

$$\begin{aligned} var(N_{1i}|\Theta_{1i}) &= \Phi\,\Theta_{1i}(1+\bar{\Phi}\,\Theta_{1i}), \\ var(N_{ji}|\Theta_{1i},\Theta_{ji}) &= \Phi\,\Theta_{1i}\Theta_{ji}\left[1+\Theta_{ji}(1+\bar{\Phi}\,\Theta_{1i})\right], \quad j=2,\ldots,\mathcal{J}. \end{aligned}$$

Finally, the correlations are:

$$\rho(N_{1i},N_{ji}) = \sqrt{\frac{\Theta_{ji}(1+\bar{\Phi}\,\Theta_{1i})}{1+\Theta_{ji}(1+\bar{\Phi}\,\Theta_{1i})}}, \quad j=2,\ldots,\mathcal{J}, \tag{10}$$

$$\rho(N_{ji},N_{li}) = \sqrt{\frac{\Theta_{ji}\Theta_{li}(1+\bar{\Phi}\,\Theta_{1i})^2}{\left[1+\Theta_{ji}(1+\bar{\Phi}\,\Theta_{1i})\right]\left[1+\Theta_{li}(1+\bar{\Phi}\,\Theta_{1i})\right]}}, \tag{11}$$

for $j,l=2,\ldots,\mathcal{J}$, $j\neq l$.

2.3. A regression Model

For the sake of convenience, the model (1) can be rewritten in a different way to facilitate the implementation of covariates to determine which factors and explanatory variables have an influence on the mean of the corresponding coverage. Then, by equating Θ_{1i} to μ_{1i} and Θ_{ji} to μ_{ji}/μ_{1i}, $j=2,\ldots,\mathcal{J}$, $\mu_{1i}\neq 0$, we obtain the normalized joint distribution, which can be expressed as,

$$f(n_{1i},\ldots,n_{\mathcal{J}i}) = \mu_{1i}^{n_{1i}}\prod_{j=2}^{\mathcal{J}}\left[\frac{1}{n_{ji}!}\left(\frac{n_{1i}\mu_{ji}}{\mu_{1i}}\right)^{n_{ji}}\right]\exp\left[-\left(\mu_{1i}+\frac{n_{1i}}{\mu_{1i}}\sum_{j=2}^{\mathcal{J}}\mu_{ji}\right)\right], \tag{12}$$

for $n_{1i} = 0, 1, \ldots, n_{ji} = 0, 1, \ldots, n_{1i}$, $j = 2, \ldots, \mathcal{J}$.

The probability function (12) satisfies the condition that the marginal means are given by $E(N_{ji}) = \mu_{ji}$, $j = 1, 2, \ldots, \mathcal{J}$, assuming that $\mu_{ji} \neq 0$, for all j. Thus, it is suitable for including covariates. Then, to carry out this regression model, we suppose that the observed counts $(N_{1i}, \ldots, N_{\mathcal{J}i})$ have independent distributions given by (12) with $E(N_{ji}) = \mu_{ji}$, $j = 1, 2, \ldots, \mathcal{J}$. Now, it is assumed that a set of observable covariates useful to subdivide the portfolio into classes of risks with homogeneous characteristics are included in the linear predictor, η_{ji}. To guarantee a positive expected value of the response variables, it is reasonable and common to use a logarithmic link for this function and therefore express the mean as:

$$\eta_{ji} = \log \mu_{ji} \quad = \quad \underline{x}_{ji} \underline{\gamma}_j^\top, \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, \mathcal{J},$$

where $\underline{x}_{ji} = (x_{ji1}, \ldots, x_{jim})$ is a vector of m covariates for the ith observation μ_{ji} and $\underline{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jm})$ denotes the corresponding vector of regression coefficients to be estimated, which usually includes a constant term. Without loss of generality, it is assumed that for each $j = 1, \ldots, \mathcal{J}$, N_{ji} is related to the same set of covariates. In addition, one of the covariates may be identified as an exposure term to calibrate the size of a potential outcome variable by assuming that the mean varies proportionally with the exposure e_{ji} (see Frees 2010; Frees et al. 2016),

$$\mu_{ji} \quad = \quad e_{ji} \exp\{\underline{x}_{ji} \underline{\gamma}_j^\top\}, \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, \mathcal{J}.$$

Similarly, the covariates can be implemented in the multivariate zero-inflated regression model by simply regressing the mean value of the different coverages. It should be pointed out, although it is not considered here, that it could also be assumed that the inflated parameter Φ could depend on certain regressors. This issue seems not to be possible here, and thus, it could be a subject that merits further investigation in future research.

3. Estimation of the Parameters

In this section, we firstly describe the methodology for the maximum likelihood estimation and derivation of the entries of Fisher's information matrix for the basic model. Next, the same development is illustrated for the associated regression model. Finally, the expression of the log-likelihood function, score equations, and the second derivative of the log-likelihood function with respect to the parameters for the zero-inflated model are exhibited.

In general, the statistical inference for multivariate models is not trivial, and the computational procedure is often expensive (see, for instance, Selch and Scherer 2010). Nevertheless, the estimation procedure for the model proposed here is straightforward. To see this, we first consider the case without covariates. Let us assume that a sample $\tilde{n} := \{(\tilde{n}_{11}, \ldots, \tilde{n}_{\mathcal{J}1}), \ldots, (\tilde{n}_{1n}, \ldots, \tilde{n}_{\mathcal{J}n})\}$ that includes n independent observations in each one of the $\mathcal{J}$ types of coverage is collected. The log-likelihood function is proportional to:

$$\ell(\Theta_1, \ldots, \Theta_{\mathcal{J}}; \tilde{n}) \propto \sum_{i=1}^{n} \tilde{n}_{1i} \log \Theta_1 + \sum_{i=1}^{n} \sum_{j=2}^{\mathcal{J}} \tilde{n}_{ji} \log \Theta_j - \sum_{i=1}^{n} \Theta_1 - \tilde{\Theta} \sum_{i=1}^{n} \tilde{n}_{1i},$$

where $\tilde{\Theta} = \sum_{j=2}^{\mathcal{J}} \Theta_j$. After differentiating the latter expression, it is possible to obtain in closed-form the maximum likelihood estimators of the parameters. They are given by:

$$\begin{aligned} \widehat{\Theta}_1 &= \bar{n}_1, \\ \widehat{\Theta}_j &= \frac{\bar{n}_j}{\bar{n}_1}, \quad j = 2, \ldots, \mathcal{J}, \end{aligned}$$

where $\bar{n}_k$, $k = 1, \ldots, \mathcal{J}$, represents the sample mean, i.e., $\sum_{i=1}^{n} \tilde{n}_{1k}/n$.

The elements that provide the entries of Fisher's information matrix are as follows:

$$\begin{aligned} E\left(-\frac{\partial^2 \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}};\tilde{n})}{\partial \Theta_1^2}\right) &= \frac{n}{\Theta_1}, \\ E\left(-\frac{\partial^2 \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}};\tilde{n})}{\partial \Theta_j^2}\right) &= \frac{n\Theta_1}{\Theta_j}, \quad j=2,\ldots,\mathcal{J}, \\ E\left(-\frac{\partial^2 \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}};\tilde{n})}{\partial \Theta_l \partial \Theta_k}\right) &= 0, \quad l,k=1,\ldots,\mathcal{J},\ l\neq k. \end{aligned}$$

For the regression model, the log-likelihood function contains $\mathcal{J}\times m$ parameters, and it is proportional to:

$$\begin{aligned} \ell(\underline{\gamma}_1,\ldots,\underline{\gamma}_{\mathcal{J}};\tilde{n}) &\propto \sum_{i=1}^{n}\tilde{n}_{1i}\log\mu_{1i}+\sum_{i=1}^{n}\sum_{j=2}^{\mathcal{J}}\tilde{n}_{ji}\log\mu_{ji}-\sum_{i=1}^{n}\sum_{j=2}^{\mathcal{J}}\tilde{n}_{ji}\log\mu_{1i} \\ &- \sum_{i=1}^{n}\mu_{1i}-\sum_{i=1}^{n}\frac{\tilde{n}_{1i}}{\mu_{1i}}\sum_{j=2}^{\mathcal{J}}\mu_{ji}, \end{aligned}$$

where $\mu_{ji}=\exp\{\underline{x}_{ji}\underline{\gamma}_j^{\top}\}$, $i=1,2,\ldots,n$ and $j=1,\ldots,\mathcal{J}$.

The score equations are given by:

$$\begin{aligned} \frac{\partial \ell(\underline{\gamma}_1,\ldots,\underline{\gamma}_{\mathcal{J}};\tilde{n})}{\partial \gamma_{1k}} &= \sum_{i=1}^{n}\left(\frac{\tilde{n}_{1i}}{\mu_{1i}}-1\right)x_{ik}+\sum_{i=1}^{n}\sum_{j=2}^{\mathcal{J}}\left(\frac{\tilde{n}_{1i}\mu_{ji}}{\mu_{1i}^2}-\frac{\tilde{n}_{ji}}{\mu_{1i}}\right)x_{ik} \\ \frac{\partial \ell(\underline{\gamma}_1,\ldots,\underline{\gamma}_{\mathcal{J}};\tilde{n})}{\partial \gamma_{jk}} &= \sum_{i=1}^{n}\left(\frac{\tilde{n}_{ji}}{\mu_{ji}}-\frac{\tilde{n}_{1i}}{\mu_{1i}}\right)x_{ik}, \end{aligned}$$

with $j=2,\ldots,\mathcal{J}$ and $k=1,\ldots,m$.

Fisher's information matrix is made up of four blocks, as can be seen below:

$$E\left(\begin{array}{c|ccc} -\frac{\partial^2\ell}{\partial\gamma_{1k}\gamma_{1l}} & -\frac{\partial^2\ell}{\partial\gamma_{1k}\gamma_{2l}} & \cdots & -\frac{\partial^2\ell}{\partial\gamma_{1k}\gamma_{\mathcal{J}l}} \\ \hline -\frac{\partial^2\ell}{\partial\gamma_{2l}\gamma_{1k}} & -\frac{\partial^2\ell}{\partial\gamma_{2k}\gamma_{2l}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial^2\ell}{\partial\gamma_{\mathcal{J}l}\gamma_{1k}} & \mathbf{0} & \cdots & -\frac{\partial^2\ell}{\partial\gamma_{\mathcal{J}k}\gamma_{\mathcal{J}l}} \end{array}\right)$$

where:

$$\begin{aligned} E\left(-\frac{\partial^2\ell}{\partial\gamma_{1k}\gamma_{1l}}\right) &= \sum_{i=1}^{n}\left(\frac{1}{\mu_{1i}}+\sum_{j=2}^{\mathcal{J}}\frac{\mu_{ji}}{\mu_{1i}^2}\right)x_{ik}x_{il}, \\ E\left(-\frac{\partial^2\ell}{\partial\gamma_{1k}\gamma_{jl}}\right) &= -\sum_{i=1}^{n}\frac{1}{\mu_{1i}}x_{ik}x_{il}, \\ E\left(-\frac{\partial^2\ell}{\partial\gamma_{jk}\gamma_{jl}}\right) &= \sum_{i=1}^{n}\frac{1}{\mu_{ji}}x_{ik}x_{il},\ E\left(-\frac{\partial^2\ell}{\partial\gamma_{jk}\gamma_{hl}}\right)=\mathbf{0},\ \text{with } j\neq h, \end{aligned}$$

where $j,h=2,\ldots,\mathcal{J}$; $l,k=1,\ldots,m$, and $\mathbf{0}$ is the zero matrix with dimension $m\times m$.

For the zero-inflated model, the log-likelihood is proportional to:

$$\begin{aligned} \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}},\Phi;\tilde{n}) &\propto n^*\log(1-\Phi+\Phi\exp(-\Theta_1))+(n-n^*)[\log\Phi-\Theta_1] \\ &+\log\Theta_1\sum_{i=n^*+1}^{n}n_{1i}-\tilde{\Theta}\sum_{i=n^*+1}^{n}n_{1i}+\sum_{i=n^*+1}^{n}\sum_{j=2}^{\mathcal{J}}n_{ji}\log(n_{1i}\Theta_j), \end{aligned}$$

where n^* is the number of zeroes of the random variable N_{1i}. The normal equations that provide the maximum likelihood estimates are given by,

$$\frac{\partial \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}},\Phi;\tilde{n})}{\partial \Phi} = \frac{n^*(\exp(-\Theta_1)-1)}{1-\Phi+\Phi\exp(-\Theta_1)}+\frac{n-n^*}{\Phi}=0, \quad (13)$$

$$\frac{\partial \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}},\Phi;\tilde{n})}{\partial \Theta_1} = -\frac{n^*\Phi\exp(-\Theta_1)}{1-\Phi+\Phi\exp(-\Theta_1)}-(n-n^*)+\frac{1}{\Theta_1}\sum_{i=n^*+1}^{n} n_{1i}=0, \quad (14)$$

$$\frac{\partial \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}},\Phi;\tilde{n})}{\partial \Theta_j} = -\sum_{i=n^*+1}^{n} n_{1i}+\frac{1}{\Theta_j}\sum_{i=n^*+1}^{n} n_{ji}=0, \quad j=2,\ldots,\mathcal{J}. \quad (15)$$

From (13), we obtain that $\Phi = (n^*-n)/(n(\exp(-\Theta_1)-1))$, which can be carried out to (15) to obtain the estimator of Θ_1, say $\widehat{\Theta}_1$. This value is carried out now to (13) to obtain the estimator of the inflated parameter, Φ. Finally, from (15), the estimator of Θ_j, $j=2,\ldots,\mathcal{J}$ is obtained in the closed-form expression given by $\widehat{\Theta}_j = \sum_{i=n^*+1}^{n} n_{ji} / \sum_{i=n^*+1}^{n} n_{1i}$.

The second partial derivatives are as follows,

$$\begin{aligned}
\frac{\partial^2 \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}},\Phi;\tilde{n})}{\partial \Phi^2} &= -\frac{n^*(\exp(-\Theta_1)-1)^2}{(1-\Phi+\Phi\exp(-\Theta_1))^2}-\frac{n-n^*}{\Phi^2},\\
\frac{\partial^2 \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}},\Phi;\tilde{n})}{\partial \Phi\partial \Theta_1} &= -\frac{n^*\exp(-\Theta_1)}{(1-\Phi+\Phi\exp(-\Theta_1))^2},\\
\frac{\partial^2 \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}},\Phi;\tilde{n})}{\partial \Theta_1^2} &= \frac{n^*\Phi(1-\Phi)\exp(-\Theta_1)}{(1-\Phi+\Phi\exp(-\Theta_1))^2}-\frac{1}{\Theta_1^2}\sum_{i=n^*+1}^{n} n_{1i},\\
\frac{\partial^2 \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}},\Phi;\tilde{n})}{\partial \Theta_j^2} &= -\frac{1}{\Theta_j^2}\sum_{i=n^*+1}^{n} n_{ji}, \quad j=2,\ldots,\mathcal{J},\\
\frac{\partial^2 \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}},\Phi;\tilde{n})}{\partial \Phi\partial \Theta_j} &= 0, \quad j=2,\ldots,\mathcal{J},\\
\frac{\partial^2 \ell(\Theta_1,\ldots,\Theta_{\mathcal{J}},\Phi;\tilde{n})}{\partial \Theta_1\partial \Theta_j} &= 0, \quad j=2,\ldots,\mathcal{J},
\end{aligned}$$

Observe that the analytic expressions of $\sum_{i=n^*+1}^{n} n_{1i}$ and $\sum_{i=n^*+1}^{n} n_{ji}$ are not feasible. For computational reasons, for large values of n, this is evaluated by ignoring the expectation operator and replacing it by $\sum_{i=n^*+1}^{n} n_{1i}$ and $\sum_{i=n^*+1}^{n} n_{ji}$. The asymptotic variance–covariance matrix is approximated by inverting the observed information matrix.

When covariates are introduced under the inflated model, we proceed first by replacing in (9) the pmf f by its corresponding (12), where again, $\mu_{ji} = \exp\{\underline{x}_{ji}\underline{\gamma}_j^{\top}\}$, $i=1,2,\ldots,n$, and $j=1,\ldots,\mathcal{J}$. In practice, as shown in the numerical applications below, the parameter estimation and computation of standard errors were carried out by the method of maximum likelihood using Mathematica® v.12.0. We directly maximized the log-likelihood function by using different maximum search methods available in the FindMaximum built-in function in the Mathematicasoftware package. This software package also provides at least two methods of obtaining the elements of the Hessian matrix. The first one consists of retrieving them from the Cholesky factors (this package is available on the web upon request). The second one, which is faster, derives them by finite differentiation. Results were also confirmed with WinRATS v.7.0.

4. Numerical Application

For our empirical analysis, we used the French Motor Personal Line datasets available in the package "CASdatasets" in **R**. This is a collection of ten datasets that comes from a private motor French insurer. Each dataset includes risk features such as claim amount, risk area, gender of the policyholder, number of claims for different coverages, etc. In particular, we chose the freMPL10 dataset, which includes 32,100 policies for the year 2004. In our study, we considered six response variables, which are shown in Table 1.

Note that the dependent variable Claims for each policyholder comprises the sum of the individual claims in all other variables. The details of the joint claims frequency for all types of coverage and the total number of claims are illustrated in Appendix A (Table A1). Note that the maximum number of claims reported by an insured is six. The number of policyholders who did not report a claim is 12,257 (38.18%), and the number of customers that only declared a claim in any of the coverages is 10,803 (33.65%).

Table 1. Description of the response variables considered.

Claims	Total number of claims made by the policyholder.
Nonresponsible	Number of nonresponsible claims in the four preceding years.
Responsible	Number of responsible claims in the four preceding years.
Parking	Number of parking claims in the four preceding years.
Windscreen	Number of windscreen claims in the four preceding years.
Fire and theft	Number of fire and theft claims in the four preceding years.

Together with all the responses, this dataset includes a set of explanatory variables. Table 2 below describes the factors and explanatory variables used in the investigation. We also considered an offset variable when modeling the claims frequency, exposure, the time exposed to risk during the investigation period.

In Table 3, the parameter estimates and their corresponding p-values are provided for the basic and zero-inflated models without covariates. Some measures of model selection are also provided in the bottom part of this table. For comparisons purposes, we used the multivariate negative binomial distribution (MNB) provided in (Johnson et al. 1996, chp. 36, p. 94) with the pmf given by:

$$\Pr(n_1,\ldots,n_k) = \frac{\Gamma(\alpha+\sum_{i=1}^{k} n_i)}{\Gamma(\alpha)\prod_{i=1}^{k} n_i!}\Theta_1^{\alpha}\prod_{i=2}^{k}\Theta_i^{n_i},$$

where $\alpha > 0$, $0 < \Theta_i < 1$, $i = 1,2,\ldots,k$, and $\sum_{i=1}^{k}\Theta_i = 1$. As can be seen in Table 3, the multivariate Poisson distribution studied in this paper has a better performance than the MNB for this dataset. Furthermore, it is observable that the the zero-inflated model improves the basic one due to the high frequency of zeros. On the other hand, we also tried to fit the two multivariate Poisson distributions provided in Bermúdez and Karlis (2011); however, we were unable to derive the maximum likelihood estimates of this model for this dataset.

Table 2. Description of factors and explanatory variables considered.

Variable	**Description**
lic age	the driving license age, in months;
veh age	takes the value 1, if the age of the vehicle is less than or equal to five years, 0 otherwise;
gender	takes the value 1 if male, 0 if female;
status	takes the value 0 if alone, 1 if other;
private 1	takes the value 1 if the usage of the vehicle is private, 0 otherwise;
private 2	takes the value 1 if the usage of the vehicle is private + trip to office, 0 otherwise;
professional	takes the value 1 if the usage of the vehicle is professional, 0 otherwise;
driver age	the driver age, in years;
has km limit	takes the value 1 if there is a km limit, 0 otherwise;
risk area	Unknown risk area between 1 and 13, possibly ordered;
bonus	takes the value 1 if the numerical value for bonus/malus is less than 100, 0 otherwise;
malus	takes the value 1 if the numerical value for bonus/malus is larger than 100, 0 otherwise.

Table 3. Parameter estimates and p-values for the basic model (second and third columns) and zero-inflated model (fourth and and fifth columns). Some measures of model selection are also given for comparison purposes.

Parameter	MNB		Basic Model		Zero-Inflated Model	
	Estimate	p-Value	Estimate	p-Value	Estimate	p-Value
$\widehat{\alpha}$	1.043	<0.001				
$\widehat{\Theta}_1$	0.329	<0.001	1.060	<0.001	1.197	<0.001
$\widehat{\Theta}_2$	0.335	<0.001	0.254	<0.001	0.254	<0.001
$\widehat{\Theta}_3$	0.092	<0.001	0.274	<0.001	0.274	<0.001
$\widehat{\Theta}_4$	0.085	<0.001	0.057	<0.001	0.057	<0.001
$\widehat{\Theta}_5$	0.019	<0.001	0.367	<0.001	0.367	<0.001
$\widehat{\Theta}_6$	0.123	<0.001	0.047	<0.001	0.047	<0.001
Inflation parameter, $\widehat{\Phi}$					0.886	<0.001
$\ell_{\max}$	−118,828.250		−106,895.00		−106,692.00	
AIC	237,671.00		213,803.00		213,398.00	
BIC	237,729.00		213,853.00		213,457.00	
CAIC	237,736.00		213,859.00		213,464.00	

Table 4 below exhibits the empirical Pearson's correlation between the different frequencies associated with each response variable for the total number of claims, and each one of the different coverages (first row), the correlation derived computed via the basic model (second row) and zero-inflated model (third row), and that computed by using (7) and (10), respectively, are also shown. It is observable that there exists a weak positive correlation between Claims and the rest of the dependent variables for each one of the coverages, and the empirical values are near the theoretical values. These figures were calculated before incorporating the effect of the explanatory variables for the different coverages. We also calculated the correlation coefficient between the rest of the response variables. Again, there is a weak positive correlation, ranging from 0.0480 between Responsible and Nonresponsible and 0.0035 between Parking and Windscreen.

Table 4. Correlation between empirical frequencies associated with the total claims number and each response variables (first row) and the correlation derived by means of the basic model (second row) and zero-inflated model (third row).

	Nonresponsible	Responsible	Parking	Windscreen	Fire and Theft
Claims	0.5150	0.5606	0.2606	0.6287	0.2378
	0.4499	0.4637	0.2334	0.5183	0.2123
	0.4733	0.4874	0.2479	0.5428	0.2258

Empirical marginal distributions and fitted marginals under the basic model (Fit 1) and zero-inflated model (Fit 2) are illustrated below in Table 5 using the estimates computed in the previous section. Note that the total number of observations equals 32,100.

We fit the multivariate regression model in (12) and the zero-inflated regression model described in the second section. Parameter estimates and their corresponding p-values are displayed in Tables 6 and 7, respectively. It is observable that for the former regression model, the explanatory variables private 1 and risk area are statistically significant at the 5% significance level. This is also verified by the intercept term of the model, i.e., constant. Furthermore, some other covariates (private 2, profession and has km limit) are significant at the same level for all the responses except for Fire and Theft; similarly, driver age is not significant for the dependent variable Responsible. On the other hand, with respect to the zero-inflated regression model, the explanatory variables risk area and constant are statistically significant at the same significance level for all responses; moreover, the covariate private 1 is not significant for the response variable Parking and

has km limit for Fire and Theft. In terms of the four measures of model selection considered, the zero-inflated regression model is preferable over the model (12).

Table 5. Empirical and fitted homogeneous marginal distributions using the basic model (Fit 1) and zero-inflated model (Fit 2).

Count	Claims			Non Responsible			Responsible		
	Obs.	Fit 1	Fit 2	Obs.	Fit 1	Fit 2	Obs.	Fit 1	Fit 2
0	12,257	11,121.20	12,257.00	24,694	25,310.10	25,407.1	24,343	24,898.80	25,008.50
1	10,803	11,788.50	10,279.50	6311	5283.81	5125.42	6426	5498.24	5322.11
2	5571	6247.91	6153.96	970	1222.22	1254.89	1124	1360.30	1392.85
3	2296	2207.59	2456.09	110	235.15	255.965	183	279.81	303.63
4	794	585.01	735.183	15	40.95	47.1004	19	52.13	59.75
5	274	124.02	176.05	0	6.63	8.02949	3	9.03	10.90
6	87	21.91	35.13	0	1.01	1.28651	2	1.47	1.87
≥7	18	3.81	7.03	0	0.16	0.22	0	0.26	0.35
Count	**Parking**			**Windscreen**			**Fire and Theft**		
	Obs.	Fit 1	Fit 2	Obs.	Fit 1	Fit 2	Obs.	Fit 1	Fit 2
0	30,287	30,251.40	30,257.80	22,306	23,173.20	23,345.9	30,595	30,567.70	30,572.00
1	1686	1743.15	1730.42	7607	6249.06	5992.86	1407	1460.07	1451.36
2	110	100.41	106.135	1772	1990.30	2013.38	93	69.36	73.43
3	15	4.82	5.42	327	525.77	562.70	5	2.74	3.093
4	1	0.21	0.25	72	126.06	142.33	0	0.09	0.11
5	1	0.01	0.01	13	28.19	33.46	0	0.00	0.00
6	0	0.00	0.00	3	5.95	7.41	0	0.00	0.00
≥7	0	0.00	0.00	0	1.47	1.94	0	0.00	0.00

Now, we are interested in comparing the six mixed random variables' aggregate claims amount for Claims (Y_1) and the different coverages, i.e., Nonresponsible, Responsible, Parking, Windscreen, and Fire and Theft, $(Y_2, \ldots, Y_6)$. In order to estimate the scale parameter σ of the exponential distribution, we considered the variable ClaimAmount available in the dataset freMPL10. The estimate of this parameter is $\widehat{\sigma} = 1.96265$. The pdf/pmf associated with the mixed random variables is displayed in Figure 1. As expected, the density of the random variable Claims fades away to zero slower than the random variables of the different coverages. Among the different coverages, the Responsible variable is the one that approaches zero faster compared to the other coverages.

Table 6. Parameter estimates and p-values for the regression model with the density function (12). Some measures of the model selection are also exhibited.

Parameter	Claims		Non Responsible		Responsible		Parking		Windscreen		Fire and Theft	
	Estimate	p-Value	Estimate	p-Value	Estimate	p-Value	Estimate	p-Value	Estimate	p-Value	Estimate	p-Value
lic age	−0.206	0.000	−0.238	0.000	−0.211	0.000	−0.127	0.279	0.008	0.874	−0.565	0.000
veh age	0.021	0.084	0.047	0.093	−0.003	0.907	0.279	0.000	0.004	0.843	−0.034	0.558
gender	0.073	0.000	0.006	0.796	0.015	0.538	−0.095	0.050	0.196	0.000	0.024	0.643
status	−0.002	0.857	−0.074	0.035	0.022	0.511	0.000	0.995	0.079	0.010	−0.299	0.000
private 1	−0.592	0.000	−0.615	0.000	−0.579	0.000	−0.493	0.001	−0.584	0.000	−0.422	0.036
private 2	−0.430	0.000	−0.500	0.000	−0.342	0.000	−0.517	0.000	−0.431	0.000	−0.142	0.461
professional	−0.380	0.000	−0.397	0.000	−0.394	0.000	−0.556	0.000	−0.322	0.000	−0.204	0.301
driver age	0.346	0.000	1.103	0.000	0.128	0.263	0.806	0.000	−0.417	0.000	0.635	0.006
has km limit	−0.378	0.000	−0.379	0.000	−0.289	0.000	−0.317	0.000	−0.522	0.000	−0.110	0.260
risk area	0.018	0.000	0.015	0.001	0.058	0.000	0.083	0.000	−0.036	0.000	0.125	0.000
bonus	−0.761	0.000	−1.827	0.000	−0.204	0.038	−0.150	0.535	−0.263	0.002	−0.245	0.209
malus	0.056	0.230	0.102	0.190	0.067	0.571	0.078	0.788	−0.183	0.096	−0.445	0.086
constant	1.752	0.000	−1.205	0.000	0.523	0.017	−4.317	0.000	2.125	0.000	−1.477	0.001
$\ell_{\max}$	−117,736.00											
AIC	23,5628.00											
BIC	23,6281.00											
CAIC	23,6359.00											
Observations	32,100											

Table 7. Parameter estimates and *p*-values for the zero-inflated regression model. Some measures of the model selection are also exhibited.

Parameter	Claims		Non Responsible		Responsible		Parking		Windscreen		Fire and Theft	
	Estimate	***p*-Value**	**Estimate**	***p*-Value**	**Estimate**	***p*-Value**	**Estimate**	***p*-Value**	**Estimate**	***p*-Value**	**Estimate**	***p*-Value**
lic age	0.050	0.091	0.316	0.000	−0.036	0.527	−0.049	0.663	0.332	0.000	−0.516	0.000
veh age	−0.011	0.422	−0.045	0.109	0.007	0.786	0.168	0.003	−0.029	0.233	−0.105	0.073
gender	0.082	0.000	−0.014	0.583	0.039	0.119	−0.071	0.142	0.202	0.000	0.024	0.647
status	−0.059	0.000	−0.177	0.000	0.035	0.307	−0.006	0.929	0.019	0.534	−0.368	0.000
private 1	−0.175	0.000	−0.299	0.000	−0.180	0.031	0.048	0.762	−0.170	0.024	−0.434	0.012
private 2	−0.099	0.021	−0.329	0.000	−0.086	0.279	−0.054	0.730	−0.106	0.135	−0.256	0.116
professional	−0.048	0.275	−0.218	0.006	−0.117	0.152	−0.061	0.701	−0.012	0.861	−0.265	0.115
driver age	−0.211	0.000	−0.035	0.771	−0.439	0.000	0.580	0.009	−1.111	0.000	0.374	0.099
has km limit	−0.303	0.000	−0.291	0.000	−0.159	0.001	−0.291	0.001	−0.426	0.000	−0.081	0.422
risk area	0.018	0.000	0.014	0.003	0.056	0.000	0.077	0.000	−0.035	0.000	0.126	0.000
bonus	−0.436	0.000	−1.510	0.000	0.013	0.882	−0.502	0.004	0.135	0.134	0.740	0.004
malus	0.157	0.001	0.311	0.000	−0.036	0.757	−0.743	0.003	−0.021	0.849	0.056	0.859
constant	2.067	0.000	−0.099	0.650	1.403	0.000	−3.671	0.000	2.513	0.000	−1.327	0.006
Inflation parameter, Φ	0.841	0.000										
$\ell_{\max}$	−116,856.00											
AIC	233,870.00											
BIC	234,532.00											
CAIC	234,611.00											
Observations	32,100											

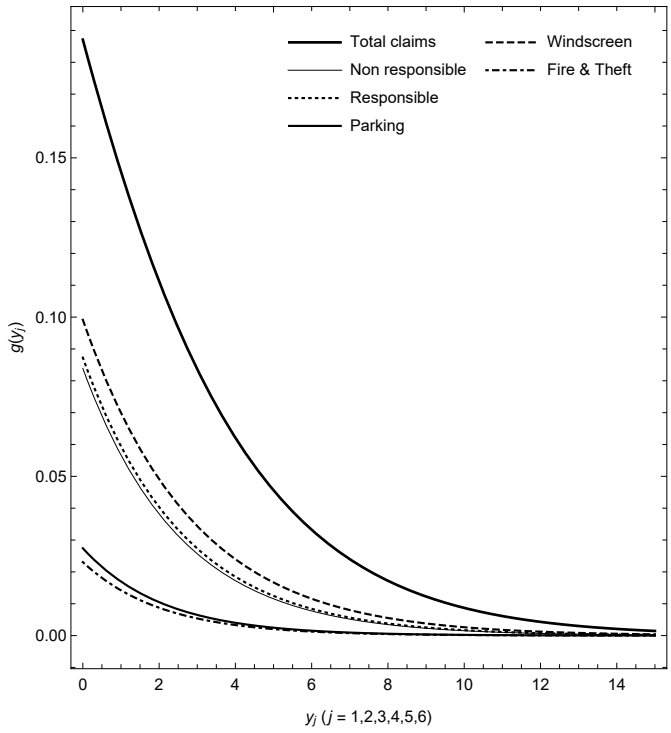

Figure 1. pdf/pmf associated with the mixed random variables of the aggregate claims amount for Claims (Y_1) and the different coverages $(Y_2, \ldots, Y_6)$ obtained from the estimated value of the mean of the claims size, $\widehat{\sigma} = 1.96265$.

5. Final Comments and Future Research

It is common today, mainly due to the existing competition, that insurers offer coverage of different claims within the same product not only to gain in competitiveness, but also to benefit from risk diversification and volatility. Up to date, most insurance companies differentiate, apart from the total number of claims, individualized claims for different coverages, such as windscreen claims and thefts and fire claims, among others. Therefore, it seems reasonable to assume that every policyholder generates a sequence of claims numbers for each coverage; one of them is the total claims number, which includes the sum of the claims in all the coverages. In this work, we introduced a new methodology based

on a multivariate discrete distribution via conditional specification to explain the claims frequency in different coverages and the total claims number. This procedure allows us to analyze the dependence structure of a finite number of coverages in motor vehicle insurance and also to include heterogeneity via explanatory variables. Closed-form expressions were given for model parameter estimates, and a priori premiums were provided when different premiums principles were used. Numerical applications revealed that specific covariates are statistically significant in some coverages, yet they are not so for others. In this way, it allows us to discern how the different explanatory variables affect each coverage when calculating the corresponding premiums.

The approach introduced in this work avoids the use of copula-based modeling. The latter methodology has been very useful, but at the same time, very criticized in the statistical literature when modeling multivariate data. Although there exists a wide catalog of copulas, it has been mentioned that a weakness of the copula approach is in choosing an appropriate copula structure for the model at hand (Balakrishnan and Lai 2009, chp. 1, p. 59). Furthermore, any copula includes a parameter that controls the dependence structure, and this parameter is sometimes difficult to estimate since it must fall into the admissible support. As explored in the second section of this work, the model depends extremely on the parameter Θ_1, and for that reason, a more flexible dependence structure based on multivariate subordination is an issue that deserves to be studied. In this regard, using this approach would be interesting to compare this family of distributions with the multivariate regression model based on the multivariate Sarmanov distribution, similar to the models derived in Bolancé and Vernic (2019). This model could be used to explain situations where the policyholder wishes to extend the third-party motor vehicle insurance to account for different coverages that adapt to their personal needs. Alternatively, it could be feasible to implement a multivariate version with elliptical copula-based models to accommodate a wide range of dependence. It is essential to mention that the properties of the copula are not the same as for continuous random variables since the probability of ties in the data is positive. Thus, the estimation cannot be directly carried out, and a continuous extension of integer-valued random variables is needed by using the approach proposed by Denuit and Lamber (2005).

The purpose of the work is not to compare other models, as models of this nature are not known to our knowledge in the actuarial literature. However, the cases with two coverages were discussed via the bivariate Poisson case (see Bermúdez and Karlis 2017) and the case with all the coverages using the multivariate negative binomial distribution in (Johnson et al. 1996, chp. 36, p. 94). Obviously, the fit obtained with the proposed modeling does not seem entirely reasonable (as judged by the chi-squared test statistics, which was not shown in the paper). Then, the model could be improved by using a similar model, but assuming that the total number of claims and all the coverages follow a negative binomial distribution instead. It would be also possible to zero-inflate all the different coverages. This issue could be the subject of future research.

Author Contributions: Conceptualization, E.G.-D. and E.C.-O.; methodology, E.G.-D. and E.C.-O.; software, E.G.-D. and E.C.-O.; validation, E.G.-D. and E.C.-O.; formal analysis, E.G.-D. and E.C.-O.; investigation, E.G.-D. and E.C.-O.; resources, E.G.-D. and E.C.-O.; data curation, E.G.-D. and E.C.-O.; writing—original draft preparation, E.G.-D. and E.C.-O.; writing—review and editing, E.G.-D. and E.C.-O.; visualization, E.G.-D. and E.C.-O.; supervision, E.G.-D. and E.C.-O.; project administration, E.G.-D. and E.C.-O. All authors have read and agreed to the published version of the manuscript.

Funding: E.G.-D.'s work was partially funded by Grant ECO2017–85577–P (Ministerio de Economía, Industria y Competitividad. Agencia Estatal de Investigación).

Acknowledgments: The authors are grateful to the four anonymous referees for their constructive comments and suggestions, which greatly helped us improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1 displays the empirical joint claims frequency for all types of coverage and the total number of claims.

Table A1. Empirical joint claims frequency corresponding to the response variables Claims (N_1), Nonresponsible (N_2), Responsible (N_3), Parking (N_4), Windscreen (N_5), and Fire and Theft (N_6).

N_1	N_2	N_3	N_4	N_5	N_6	Count	N_1	N_2	N_3	N_4	N_5	N_6	Count	N_1	N_2	N_3	N_4	N_5	N_6	Count	N_1	N_2	N_3	N_4	N_5	N_6	Count	N_1	N_2	N_3	N_4	N_5	N_6	Count
0	0	0	0	0	0	12,257	3	2	0	0	1	0	147	4	2	1	1	0	0	13	5	2	2	1	0	0	4	6	3	2	0	1	0	2
1	0	0	0	0	1	447	3	2	0	1	0	0	24	4	2	2	0	0	0	19	5	2	3	0	0	0	7	6	4	1	0	1	0	2
1	0	0	0	1	0	3947	3	2	1	0	0	0	137	4	3	0	0	1	0	11	5	3	0	0	2	0	3	7	0	5	0	2	0	1
1	0	0	1	0	0	590	3	3	0	0	0	0	47	4	3	0	1	0	0	2	5	3	0	1	1	0	2	7	0	6	0	1	0	2
1	0	1	0	0	0	2924	4	0	0	0	2	2	4	4	3	1	0	0	0	12	5	3	1	0	0	1	4	7	1	0	1	5	0	1
1	1	0	0	0	0	2895	4	0	0	0	3	1	4	4	4	0	0	0	0	7	5	3	1	0	1	0	4	7	1	1	0	5	0	2
2	0	0	0	0	2	28	4	0	0	0	4	0	36	5	0	0	0	4	1	4	5	3	1	1	0	0	4	7	1	4	0	2	0	1
2	0	0	0	1	1	193	4	0	0	1	3	0	7	5	0	0	0	5	0	7	5	3	2	0	0	0	3	7	2	0	2	3	0	1
2	0	0	0	2	0	817	4	0	0	3	1	0	2	5	0	0	1	2	2	2	5	4	0	0	1	0	2	7	3	2	0	2	0	2
2	0	0	1	0	1	34	4	0	1	0	1	2	3	5	0	0	1	3	1	1	5	4	1	0	0	0	4	7	3	2	2	0	0	1
2	0	0	1	1	0	238	4	0	1	0	2	1	15	5	0	0	2	1	2	2	6	0	0	0	6	0	1	8	0	2	5	1	0	1
2	0	0	2	0	0	33	4	0	1	0	3	0	60	5	0	0	2	2	1	1	6	0	1	0	4	1	2	8	1	4	0	3	0	2
2	0	1	0	0	1	150	4	0	1	1	0	2	1	5	0	0	3	2	0	4	6	0	1	0	5	0	1	8	3	3	0	2	0	2
2	0	1	0	1	0	1038	4	0	1	1	1	1	3	5	0	1	0	3	1	1	6	0	1	1	4	0	2	10	1	3	0	6	0	2
2	0	1	1	0	0	180	4	0	1	1	2	0	20	5	0	1	0	4	0	6	6	0	2	0	3	1	1	2	0	2	0	0	0	436
4	0	1	2	0	1	3	5	0	1	1	3	0	4	6	0	2	0	4	0	2	2	1	0	0	0	1	138	4	0	1	2	1	0	9
5	0	1	2	0	2	2	6	0	2	1	2	1	2	2	1	0	0	1	0	954	4	0	2	0	0	2	2	5	0	1	4	0	0	1
6	0	2	1	3	0	1	2	1	0	1	0	0	151	4	0	2	0	1	1	13	5	0	2	0	0	3	1	6	0	3	0	3	0	2
2	1	1	0	0	0	783	4	0	2	0	2	0	48	5	0	2	0	1	2	1	6	0	3	1	2	0	3	2	2	0	0	0	0	398
4	0	2	1	0	1	1	5	0	2	0	2	1	4	6	0	4	0	0	2	1	3	0	0	0	0	3	2	4	0	2	1	1	0	10
5	0	2	0	3	0	8	6	0	4	0	2	0	2	3	0	0	0	1	2	19	4	0	2	2	0	0	4	5	0	2	1	2	0	1
6	0	4	2	0	0	1	3	0	0	0	2	1	55	4	0	3	0	0	1	3	5	0	2	3	0	0	3	6	0	5	0	1	0	2
3	0	0	0	3	0	136	4	0	3	0	1	0	17	5	0	3	0	1	1	12	6	1	0	0	5	0	2	3	0	0	1	0	2	2
4	0	3	1	0	0	2	5	0	3	0	2	0	5	6	1	1	0	3	1	2	3	0	0	1	1	1	11	4	0	4	0	0	0	4
5	0	3	1	1	0	10	6	1	1	0	4	0	6	3	0	0	1	2	0	45	4	1	0	0	0	3	2	5	0	4	0	1	0	2
6	1	1	1	3	0	1	3	0	0	2	0	1	1	4	1	0	0	1	2	1	5	1	0	0	3	1	4	6	1	2	0	1	2	2
3	0	0	2	1	0	7	4	1	0	0	2	1	20	5	1	0	0	4	0	13	6	1	2	0	3	0	8	3	0	0	3	0	0	4
4	1	0	0	3	0	50	5	1	0	1	3	0	2	6	1	2	1	2	0	1	3	0	1	0	0	2	8	4	1	0	1	1	1	2
5	1	1	0	1	2	1	6	1	2	2	1	0	1	3	0	1	0	1	1	45	4	1	0	1	2	0	16	5	1	1	0	2	1	7
6	1	2	3	0	0	2	3	0	1	0	2	0	268	4	1	0	2	0	1	1	5	1	1	0	3	0	23	6	1	3	0	2	0	4
3	0	1	1	0	1	10	4	1	0	2	1	0	7	5	1	1	1	1	1	1	6	1	4	1	0	0	2	3	0	1	1	1	0	49
4	1	1	0	1	1	22	5	1	1	1	2	0	3	6	2	0	0	3	1	1	3	0	1	2	0	0	12	4	1	1	0	2	0	78
5	1	2	0	0	2	2	6	2	0	0	4	0	1	3	0	2	0	0	1	23	4	1	1	1	0	1	5	5	1	2	0	1	1	2
6	2	0	2	1	1	1	3	0	2	0	1	0	211	4	1	1	1	1	0	12	5	1	2	0	2	0	20	6	2	1	0	3	0	1
3	0	2	1	0	0	20	4	1	1	2	0	0	10	5	1	2	1	1	0	2	6	2	1	1	2	0	1	3	0	3	0	0	0	74
4	1	2	0	0	1	7	5	1	3	0	0	1	2	6	2	2	0	1	1	1	3	1	0	0	0	2	7	4	1	2	0	1	0	62
5	1	3	0	1	0	7	6	2	2	0	2	0	4	3	1	0	0	1	1	52	4	1	2	1	0	0	16	5	1	3	1	0	0	7
6	2	2	1	1	0	2	3	1	0	0	2	0	257	4	1	3	0	0	0	20	5	1	4	0	0	0	2	6	2	3	0	1	0	3
3	1	0	1	0	1	9	4	2	0	0	0	2	5	5	2	0	0	2	1	1	6	2	3	1	0	0	1	3	1	0	1	1	0	59
4	2	0	0	1	1	5	5	2	0	0	3	0	6	6	2	4	0	0	0	2	3	1	0	2	0	0	9	4	2	0	0	2	0	35
5	2	1	0	1	1	9	6	3	0	0	2	1	3	3	1	1	0	0	1	40	4	2	0	1	0	1	1	5	2	1	0	2	0	13
6	3	0	0	3	0	1	3	1	1	0	1	0	284	4	2	0	1	1	0	25	5	2	1	1	1	0	3	6	3	0	1	2	0	2
3	1	1	1	0	0	61	4	2	0	2	0	0	2	5	2	1	2	0	0	2	6	3	1	0	2	0	2	3	1	2	0	0	0	144
4	2	1	0	0	1	8	5	2	2	0	0	1	2	6	3	1	1	1	0	2	3	2	0	0	0	1	17	4	2	1	0	1	0	47
5	2	2	0	1	0	21	6	3	2	0	0	1	1																					

Note

[1] The Stirling number of the second kind can be computed in Mathematica with the command `StirlingS2`$[n, \tau]$ (see, for instance, Ruskeepaa 2009; Olver et al. 2010).

References

Balakrishnan, Narayanaswamy, and Chin-Diew Lai. 2009. *Continuous Bivariate Distributions*, 2nd ed. New York and London: Springer.

Bermúdez, Lluis. 2009. A priori ratemaking using bivariate Poisson regression models. *Insurance: Mathematics and Economics* 44: 135–41.

Bermúdez, Lluis, and Dimitris Karlis. 2011. Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics* 48: 226–36. [CrossRef]

Bermúdez, Lluis, and Dimitris Karlis. 2017. A priori ratemaking using bivariate Poisson models. *Scandinavian Actuarial Journal* 2: 148–58. [CrossRef]

Bolancé, Catalina, and Raluca Vernic. 2019. Multivariate count data generalized linear models: Three approaches based on the Sarmanov distribution. *Insurance: Mathematics and Economics* 85: 89–103. [CrossRef]

Boucher, Jean, Michel Denuit, and Montserrat Guillén. 2007. Risk classification for claim counts: A comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal* 11: 110–31. [CrossRef]

Cacaoullos, Teophilos, and Haralambos Papageorgiou. 1980. On some bivariate probability models applicable to traffic accidents and fatalities. *International Statistical Review* 48: 345–56. [CrossRef]

Cameron, Colin, and Pravin Trivedi. 1986. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* 1: 29–54. [CrossRef]

Cameron, Colin, and Pravin Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Cummins, David, and Laurel Wiltbank. 1983. Estimating the total claims distribution using multivariate frequency and severity. *The Journal of Risk and Insurance* 50: 377–403. [CrossRef]

Cummins, David, and Laurel Wiltbank. 1984. A multivariate model of the total claims process. *ASTIN Bulletin* 14: 45–52. [CrossRef]

Denuit, Michel, and Philippe Lamber. 2005. Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis* 93: 40–57. [CrossRef]

Denuit, Michel, Xavier Marèchal, Sandra Pitrebois, and Jean-Francois Walhin. 2009. *Actuarial Modelling of Claim Counts Risk Classification, Credibility and Bonus-Malus Systems*. Hoboken: John Wiley & Sons.

Douglas, Jim. 1955. Fitting the Neyman Type A (two parameter) contagious distribution. *Biometrics* 11: 149–173. [CrossRef]

Frees, Edward. 2010. *Regression Modeling with Actuarial and Financial Applications*. Cambridge: Cambridge University Press.

Frees, Edward, Gee Lee, and Lu Yang. 2016. Multivariate frequency-severity regression models in insurance. *Risks* 4: 4. [CrossRef]

Gómez-Déniz, Emilio. 2016. Bivariate credibility bonus-malus premiums distinguishing between two types of claims. *Insurance: Mathematics and Economics* 70: 117–24.

Gómez-Déniz, Emilio, and Enrique Calderín-Ojeda. 2018. Multivariate credibility in bonus–malus systems distinguishing between different types of claims. *Risks* 6: 34. [CrossRef]

Gómez-Déniz, Emilio, and Enrique Calderín-Ojeda. 2020. A survey of the individual claim size and other risk factors using credibility bonus–malus premiums. *Risks* 8: 20.

Johnson, Norman, Adrienne Kemp, and Samuel Kotz. 2005. *Univariate Discrete Distributions*. Hoboken: John Wiley, Inc.

Johnson, Norman, Samuel Kotz, and Narayanaswamy Balakrishnan. 1996. *Discrete Multivariate Distributions*. Hoboken: John Wiley, Inc.

Kemp, Charles. 1967. On a contagious distribution suggested for accident data. *Biometrics* 23: 241–55. [CrossRef]

Lee, Simon, and Sheldon Lin. 2012. Modeling dependent risks with multivariate Erlang mixtures. *ASTIN Bulletin* 42: 153–80.

Leiter, Robert E., and M. A. Hamdan. 1973. Some bivariate probability models applicable to traffic accidents and fatalities. *International Statistical Review* 41: 87–100. [CrossRef]

Li, Chin-Shang, Jye-Chyi Lu, Jinho Park, Kyungmoo Kim, Paul A. Brinkley, and John P. Peterson. 1999. Multivariate zero-inflated Poisson models and their applications. *Technometrics* 41: 29–38. [CrossRef]

Liu, Yin, and Guo-Liang Tian. 2015. Type I multivariate zero-inflated Poisson distribution with applications. *Computational Statistics & Data Analysis* 83: 200–22.

Neyman, Jerzy. 1939. On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics* 10: 35–57. [CrossRef]

Oh, Rosy, Peng Shi, and Jae Youn Ahn. 2020. Bonus-Malus premiums under the dependent frequency-severity modeling. *Scandinavian Actuarial Journal* 3: 172–95. [CrossRef]

Olver, Frank, Daniel Lozier, Ronald Boisvert, and Charles Clark. 2010. *NIST Handbook of Mathematical Functions*. Cambridge: Cambridge University, New York, NY.

Rolski, Tomasz, Hanspeter Schmidli, Volker Schmidt, and Jozef Teugel. 1999. *Stochastic Processes for Insurance and Finance*. Hoboken: John Wiley & Sons.

Ruskeepaa, Heikki. 2009. *Mathematica Navigator. Mathematics, Statistics, and Graphics*, 3rd ed. Cambridge: Academic Press.
Selch, Daniela, and Matthias Scherer. 2010. *A Multivariate Claim Count Model for Applications in Insurance*. Berlin: Springer International Publishing.
Winkelmann, Rainer. 2003. *Econometric Analysis of Count Data*. Berlin: Springer Science & Business Media.
Young, Virginia. 2006. Premium principles. In *Encyclopedia of Actuarial Science*. New York: John Wiley & Sons, pp. 1–14.
Zhang, Pengcheng, Enrique Calderín-Ojeda, Shuanming Li, and Xueyuan Wu. 2020. On the Type I multivariate zero-truncated hurdle model with applications in health insurance. *Insurance: Mathematics and Economics* 90: 35–45. [CrossRef]