
LISBOA
SCHOOL OF
ECONOMICS &
MANAGEMENT

# MASTER

## ACTUARIAL SCIENCE

## MASTER´S FINAL WORK

DISSERTATION

# MODELLING DEPENDENCE BETWEEN FREQUENCY AND SEVERITY OF INSURANCE CLAIMS

INÊS DUARTE CARREIRA

OCTOBER 2017


LISBOA
SCHOOL OF
ECONOMICS &
MANAGEMENT

# MASTER

## ACTUARIAL SCIENCE

# MASTER´S FINAL WORK

DISSERTATION

## MODELLING DEPENDENCE BETWEEN FREQUENCY AND SEVERITY OF INSURANCE CLAIMS

INÊS DUARTE CARREIRA

**SUPERVISION:**

JOÃO ANDRADE E SILVA

OCTOBER 2017

# Abstract

The estimation of the individual loss is an important task to price insurance policies. The standard approach assumes independence between claim frequency and severity, which may not be a realistic assumption. In this text, the dependence between claim counts and claim sizes is explored, in a Generalized Linear Model framework. A Conditional severity model and a Copula model are presented as alternatives to model this dependence and later applied to a data set provided by a Portuguese insurance company. At the end, the comparison with the independence scenario is carried out.



**Keywords:** Frequency; Severity; Policy Loss; Dependence; Generalized Linear Model; Conditional Model; Copula Model; Hurdle model

# Resumo

A estimação da perda individual é uma importante tarefa para calcular os preços das apólices de seguro. A abordagem padrão assume independência entre a frequência e a severidade dos sinistros, o que pode não ser uma suposição realística. Neste texto, a dependência entre números e montantes de sinistros é explorada, num contexto de Modelos Lineares Generalizados. Um modelo de severidade condicional e um modelo de Cópula são apresentados como alternativas para modelar esta dependência e posteriormente aplicados a um conjunto de dados fornecido por uma seguradora portuguesa. No final, a comparação com o cenário de independência é realizada.

**Palavras-chave:** Frequência; Severidade; Perda por apólice; Dependência; Modelos Lineares Generalizados; Modelo Condicional; Modelo de Cópula; Modelo de Barreira

# Acknowledgments

I would like to thank my supervisor, Professor João Andrade e Silva, for his support, help, and availability to answer all my questions during the elaboration of this thesis.

I am also grateful to my parents and to my sister, for their support and encouragement through this master and throughout my life, as well as to my friends, for their motivation and patience.

Finally, I express my gratitude to the insurance company that provided the data needed for the application of the models presented in this text.

# Table of contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

In our daily lives, we face many risks, such as car accidents, work injuries and house damages. And to protect ourselves from possible losses in the future, we agree to pay a premium to an insurer who undertakes the risk.

On the other side, the goal of an insurer is to charge an accurate premium to the policyholder, to avoid losing policies to a competitor. To accomplish this, the insurer should perform an adequate estimation of the individual's expected loss. Therefore, in the past years, actuaries have been investigating and developing techniques to continue to improve the existing methods and to obtain an estimate that best reflects the reality.

Since the individual loss is the total amount paid due to the occurrence of claims, then it is given by the sum of the amounts of each claim. As a result, two components are usually investigated when it comes to estimate it: claim frequency and claim severity. The first one refers to the number of claims, and the second one to the cost associated with each claim. Moreover, these two quantities vary from policy to policy, due to the characteristics of each policyholder, the characteristics of the product insured or to other factors. Therefore, the modeling of insurance claims is widely done using regression models, namely in the GLM framework.

A common approach is to assume that claim counts and claim amounts are independent, which simplifies the computation of the expected loss by just being the product of their expected values [Klugman et al. (2008)]. As a consequence, we can model the severity and frequency components separately. Another approach, also under the independence assumption, is to use a Tweedie model.

But is the independence assumption realistic? On the one hand, it makes the computation easier; on the other hand, not considering the dependence between claim counts and claim amounts can lead to under or over-estimation of the total loss, which can lead to improper pricing of insurance policies and future losses to the insurer. In fact, it is very likely that these two components are correlated. For instance, a negative

association is often found in automobile insurance, where there may be policyholders that have frequent claims with small amounts, if we think of a policyholder living in urban areas.

Thus, a relaxation of the independence assumption seems to be needed. To account for dependence between claim sizes and claim numbers, some recent approaches have been proposed in the literature and have shown that, in fact, we can have cases where this dependence is significant and should not be ignored.

The two main approaches, which are the focus of this paper, are the Conditional approach and the Copula regression approach. The first one uses a conditional severity model and allows the number of claims to enter the model as a covariate. It was considered by Gschlößl and Czado (2005), when developing a study about spatial modelling of frequency and severity, and by Garrido et al. (2016), in a ratemaking perspective. It was also investigated by Frees et al. (2011), to model health care expenditures. In all the mentioned studies, the results were improved compared to the independent model. The second approach uses a copula to link the marginal frequency and severity GLMs and to model the dependence. Czado et al. (2012) and Krämer et al. (2013) followed a mixed copula approach proposed by Song (2007) to estimate the total loss and made an application to a German car-insurance dataset, which presented a moderate positive dependence between the two components. The former used a Gaussian copula and the latter made an extension to other copula families (Clayton, Gumbel, and Frank), which showed that better results are obtained when the appropriate copula is selected. Both approaches will be presented in this text to estimate the policy loss, without assuming independency between frequency and severity.

In order to illustrate both methods, they will be applied to a car insurance data set. The final purpose is to compare both independent and dependent models, as to check if relaxing the independence assumption improves the model. Moreover, in the literature mentioned above, the application of the copula model was done to a truncated data set. In this text, a complete data set is used. To allow the inclusion of zeros, and to facilitate the comparison with the independent scenario, a Hurdle model will be applied to claim frequency, instead of the standard Poisson regression. This model also has the advantage of dealing with the excess of zeros, which is a common feature in non-life insurance data.

The outline of the text is as follows: Chapter 2 provides an overview of Generalized Linear Models. Chapter 3 presents the independent case between claim frequency and claim severity, as well as the two standard models used under this assumption. Chapter 4 addresses the dependence problem and presents an overview of two models that take it into account. First, a conditional model is discussed, and then a copula based model is presented. Finally, Chapter 5 studies the application of the dependent models to real car insurance data and compares the results with the independent model.

# Chapter 2

# Generalized Linear Models

To estimate the individual's expected loss, the insurers make use of explanatory variables or covariates, such as the characteristics of the policyholder and of the insured products. This information allows the insurer to charge a fair premium to each policyholder, that is, to charge the amount that best reflects the expected loss transferred to the insurance company.

A widely used modelling process is the Generalized Linear Models (GLM) approach, which allows us to model a mean's transformation of the dependent variable as a linear function of the covariates. The ingredients to these models, following Ohlsson and Johansson (2010), are:

1. A distribution for the dependent variable. It is assumed that each component of Y has a distribution from the exponential dispersion family, that is,

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

where a(.), b(.) and c(.) are some specific functions; $\theta$ is the natural parameter; and $\phi$ is the dispersion parameter. Therefore, its mean and variance are given by

$$E[Y] = \mu = b'(\theta)$$

and

$$Var(Y) = b''(\theta)a(\emptyset) = V(\mu)a(\emptyset)$$

where $V(\mu) = b''(\theta)$ is called the variance function.

It can easily be shown that distributions like Gamma and Poisson (frequently used to model claim sizes and claim counts, respectively), as well as the Normal (classical linear models), belong to this family.

2. A linear predictor. It is a function of a set of covariates $x_j$,

$$\eta = \sum_{j=1}^{p} x_j\beta_j,$$

where $\beta_j$ corresponds to the parameters that should be estimated.

3. A link function. It is a function g(.) which connects the linear predictor η to the mean response

$$g(\mu) = \eta\,.$$

Each distribution has a canonical link that can be used, but other link functions can be considered, taking into account the possible values of $\mu$.

To find the maximum likelihood estimates (MLE) of the regression parameters, in a GLM framework, let's consider *m* independent observations. As a result, the log-likelihood function will be

$$\ell(\theta, \phi|\boldsymbol{y}) = \sum_{i=1}^{m} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} \right\} + \sum_{i=1}^{m} c(y_i, \phi).$$

After some computations, we arrive at the following system of $p$ equations, from which we can obtain the desired estimates,

$$\sum_{i=1}^{m} \frac{y_i - \mu_i}{a(\phi_i)V(\mu_i)g'(\mu_i)} x_{ij} = 0, \;\; j = 1, \dots, p.$$

Additionally, under general conditions, MLE are asymptotically normally distributed. Therefore, for large samples, we have that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{I}_{\boldsymbol{\beta}}^{-\mathbf{1}}\right),$$

where $\boldsymbol{I}_{\boldsymbol{\beta}} = -E[\frac{\partial^2 \ell}{\partial^2 \beta}]$ is the Fisher's information matrix.

# Chapter 3

# The independent case

The aggregate loss, over a fixed period, for policyholder *i* (or risk cell *i*) can be defined as

$$S_i = Y_{i0} + Y_{i1} + \cdots + Y_{iN_i},$$

where $N_i$ and $Y_{ij}$ $(j = 0, 1, \dots, N_i)$, with $Y_{i0} \equiv 0$, are random variables that represent the number of claims and the claim amounts, respectively, for the *i*th policyholder. Now, the problem is how to estimate this loss.

When estimating the total loss, the standard approach is to assume independence between frequency and severity, that is, to assume that there is no association between those two random variables. As presented by Klugman et al. (2008), and dropping the index *i*, the independence assumptions are given by:

- Conditionally on $N = n$, the individual claim sizes $Y_1, \dots, Y_n$ are mutually independent and identically distributed (i.i.d.);
- Conditionally on $N = n$, the common distribution of $Y_1, \dots, Y_n$ doesn't depend on claim numbers $N$;
- The distribution of the claim numbers $N$ does not depend on the values of $Y_1, Y_2, \dots$

Usually, under these independence assumptions, one of the following approaches is chosen by the actuaries: to model the total loss directly, using Tweedie models; or to fit separate models to frequency and severity, and then use them to obtain the distribution of the total loss, as well as its moments.

One of the advantages of the first approach is that only one model needs to be fitted, which means that it uses fewer parameters and that it is less time consuming, when compared to the second approach. On the other hand, the latter approach has the advantage of using a different set of covariates to explain claim sizes and claim numbers, or to allow for different effects, even with different directions, of the same covariate in

both components. For instance, the number of kilometers driven by the policyholder can have an effect on the claim numbers, but no influence on the claim sizes.

## 3.1. Tweedie Models

The Tweedie family of distributions includes distributions such as the Gamma, the Poisson or the Compound Poisson. They are characterized by a variance function given by

$$V(\mu) = \mu^p, p \in \mathbb{R}.$$

More details can be found in Jørgensen (1997).

When modelling the individual aggregate loss, $S$, as the response variable, the interest falls in the class of Tweedie models with $1<p<2$, which is defined as a Poisson sum of Gamma random variables, also called the compound Poisson-Gamma distribution. It is a mixed distribution with a positive probability at zero and a continuous distribution for positive real numbers.

Furthermore, the Tweedie model belongs to the exponential dispersion models, and consequently it can be modelled in the context of GLMs. Therefore,

$$E[S_i] = g^{-1}(x_i{}'\alpha)$$

where $x_i$ is the set of covariates and $\alpha$ is the vector of regression parameters, for the *i*th policyholder.

Jørgensen and Souza (1994) used the Tweedie models to estimate the pure premium, considering a Poisson process for the arrival of claims and a Gamma distribution for individual costs. They applied this method to a Brazilian private motor insurance portfolio where they found a value for $p$ of 1.37.

## 3.2. Frequency-Severity model

Although the Tweedie model gives a good approximation to the expected loss in many cases, the standard approach is to separate frequency and severity. By following this approach, we can obtain more accurate information about how the rating factors affect the individual loss, as previously mentioned.

In fact, the expected value of the individual aggregate loss, under the independence assumption, can be obtained by computing the product of the expected claim counts and the expected claim amounts, i.e.,

$$E[S_i] = E\left[E\left(\sum_{j=1}^{N_i} Y_{ij} \middle| N_i\right)\right] = E\left[N_i E[Y_i]\right] = E[N_i]E[Y_i] = \mu_{N_i} . \mu_{Y_i}$$

An equivalent expression can be found if we model the average claim size, $\overline{Y}_i = \frac{S_i}{N_i}$ for $N_i > 0$, instead of the individual claim size:

$$S_i = N_i \overline{Y}_i \Rightarrow E[S_i] = E[E(N_i \overline{Y}_i | N_i)] = E[N_i]E[\overline{Y}_i] = \mu_{N_i} . \mu_{Y_i}, \quad (3.1)$$

where $E[\overline{Y}_i] = E\left[E[\overline{Y}_i | N_i]\right] = E[Y_i] = \mu_{Y_i}$. When $N_i = 0$, $\overline{Y}_i = 0$ and $S_i = 0$.

Therefore, a regression model can be fitted to each component separately and, in the end, they are put together to obtain the expected loss. As regards the independent model, the result (3.1) will be used in this text.

### 3.2.1. Frequency regression model

The marginal distributions are fitted to the data considering the characteristics of the random variables. Furthermore, they can depend on a set of covariates. Thus a GLM is considered.

For the frequency component, we can think in the Poisson GLM, which is appropriate to model claim counts. However, in non-life insurance, it is common to find an excessive number of policies that did not report any claims (excess of zeros), that is, the number of observed zeros can be far more than what would be expected for this distribution. Therefore, the use of a Hurdle model [Mullahy J (1986); Zeileis et al. (2008)], to accommodate this feature, appears to be appropriate. This choice is also motivated by the fact that it makes easier to compare the independent model with the dependent Copula model presented in section 4.2., which is one of the purposes of this text.

The Hurdle model has two components: one for the zeros (hurdle component) and another for the positive counts (truncated component). As defined in Zeileis et al. (2008), the probability mass function is given by

$$f_N\,(n_i;\, x_i^N,\, z_i,\, \beta^N, \gamma) = \begin{cases} f_{zero}(0; z_i, \gamma) & , n_i = 0 \\ \big(1 - f_{zero}(0; z_i, \gamma)\big) \dfrac{f_{count}\big(n_i;\, x_i^N, \beta^N\big)}{1 - f_{count}\big(0; x_i^N, \beta^N\big)} & , n_i > 0 \end{cases}$$

where $z_i = (1, z_{i1}, \ldots., z_{iq})'$ and $x_i^N = (1, x_{i1}^N, \ldots., x_{ik}^N)'$ are the sets of explanatory variables; $\gamma = (\gamma_0, \ldots, \gamma_q)$ and $\beta^N = (\beta_0^N, \ldots., \beta_k^N)'$ are the unknown regression parameters; and $f_{zero}$ and $f_{count}$ are the probability functions for the zero component and for the claim numbers (before truncation), respectively.

Thus, to model the claim frequency, a Binomial GLM can be implemented for the hurdle component and a truncated Poisson GLM can be chosen for the positive claim counts. A logit link and a log-link will be considered in the first and second case, respectively. Additionally, and because not all the policies are in force the whole year, let *e* be the exposure time, in years, for each policy. Then, for policy *i*,

$$\mu_{count} = e \times \exp\big(x^{N\prime}\beta^N\big) \quad \text{and} \quad 1 - f_{zero}(0; z, \gamma) = \frac{\exp(z'\gamma)}{1+\exp(z'\gamma)}\,.$$

Putting it all together, we obtain the following frequency mean for policy *i*,

$$\mu_N = e \times exp\left(x^{N\prime}\beta^N\right) \times \frac{1-f_{zero}(0;z,\gamma)}{1-f_{count}(0;x^N,\beta)}. \tag{3.2}$$

### 3.2.2. Severity regression model

When the response variable is the claim size (or the average claim size), which is a positive continuous random variable, models like a Gamma GLM can be chosen. In this text, a log-link will be assumed. This is a common practice in the insurance industry, as it yields a multiplicative rating structure. See Ohlsson and Johansson (2010) for more details.

The expected (average) claim size is given by

$$\mu_{Y_i} = \exp(x_i^{Y\prime}\beta^Y) \tag{3.3}$$

where $x_i^Y = (1, x_{i1}^Y, \ldots., x_{ip}^Y)'$ is the set of explanatory variables for the claim severity and $\beta^Y = (\beta_0^Y, \ldots., \beta_p^Y)'$ are the unknown regression parameters.

Note that, if the conditional individual claim sizes follow a Gamma distribution with mean $\mu_{Y_i}$ and scale parameter $\phi$, then, by the convolution property, the conditional

average claim size follows a Gamma distribution with mean $\mu_{Y_i}$ and scale parameter $\frac{\phi}{N_i}$. Therefore, if the number of claims is included as weight in the model for the average claim size, it will be equivalent to the model for the individual claim sizes.

### 3.2.3. Estimation

As the model for frequency is fitted separately from the model for severity, the estimation of its parameters is also done separately. Additionally, for the frequency part, it is important to notice that the parameters from the hurdle component and the ones from the count component can, also, be estimated separately. The estimates can be found by maximizing each likelihood function.

For the claim frequency, the log-likelihood is given by

$$\ell\left(\gamma, \beta^N \middle| \boldsymbol{n}\right) = \sum_{i=1}^{m} \log(f_N(n_i|\gamma, \beta^N)) = \ell(\gamma|\boldsymbol{n}) + \ell\left(\beta^N \middle| \boldsymbol{n}\right),$$

where

$$\ell(\gamma|\boldsymbol{n}) = \sum_{i:\, n_i=0} \log(f_{zero}(0|\gamma)) + \sum_{i:n_i>0} \log(1 - f_{zero}(0|\gamma)),$$

and

$$\ell\left(\beta^N \middle| \boldsymbol{n}\right) = \sum_{i:n_i>0} [\log(f_{count}(n_i|\beta^N)) - \log\left(1 - f_{count}\left(0\middle|\beta^N\right)\right].$$

As a result,

$$\hat{\gamma} = \arg\max_{\hat{\gamma}} \ell(\gamma|\boldsymbol{n}) \text{ and } \widehat{\beta^N} = \arg\max_{\widehat{\beta^N}} \ell(\beta^N|\boldsymbol{n}).$$

On the other hand, for the claim severity,

$$\widehat{\beta^Y} = \arg\max_{\beta^Y} \ell(\beta^Y|\boldsymbol{y}),$$

where $\ell(\beta^Y|\boldsymbol{y}) = \sum_{i=1}^{m} \log(f_Y(y_i|\beta^Y))$.

The computation of these maximum likelihood estimates, in a GLM framework, is done as presented in Chapter 2.

### 3.2.4. Policy loss

Putting equations (3.2) and (3.3) together, as mentioned in (3.1), we obtain the following individual expected loss

$$E[S_i] = \mu_{N_i} \times \ \mu_{Y_i} = \frac{1-f_{zero}(0;z_i,\gamma)}{1-f_{count}(0;x_i^N,\beta)} \times \ e_i \times \exp(x_i^{N\prime}\beta^N \ + x_i^{Y\prime}\beta^Y). \qquad (3.4)$$

Furthermore, in the frequency-severity independent model, the variance of a policy aggregate loss can also be easily derived, using the iterated expectations:

$$Var(S_i) = E[Var(N_i\bar{Y}_i|N_i)] + Var(E[N_i\bar{Y}_i|N_i]) = E\left[{N_i}^2 Var(\bar{Y}_i|N_i)\right] + \left(\mu_{Y_i}\right)^2 Var(N_i)$$
$$= \mu_{N_i} Var(Y_i) + \left(\mu_{Y_i}\right)^2 Var(N_i) \qquad (3.5)$$

If we consider *m* independent policyholders, with independent policy losses $S_1, \dots, S_m$, and define the total loss for the insurer as

$$S = \sum_{i=1}^{m} S_i \ ,$$

then the expected total loss and its variance are, respectively,

$$\mu_S = \sum_{i=1}^{m} E[S_i]$$

and

$$\sigma_S^2 = \sum_{i=1}^{m} Var(S_i).$$

Therefore, by applying the central limit theorem, the asymptotic distribution of the total loss S is normal, i.e.,

$$\frac{\sqrt{m}}{\sqrt{\sigma_S^2}}(S - \mu_S) \xrightarrow{D} \mathcal{N}(0,1).$$

**Hurdle-Poisson model** – If the claim frequency follows a Hurdle-Poisson model, that is, if the number of claims follows a Poisson($\lambda$) distribution, with $\lambda = \mu_{count}$; and the hurdle component follows a Bernoulli (1-p) distribution, with $p = f_{zero}(0)$, then:

$$\mu_{N_i} = \lambda_i \frac{1-p_i}{1-e^{-\lambda_i}} \text{ and } Var(N_i) = \frac{1-p_i}{1-e^{-\lambda_i}}\lambda_i(\lambda_i+1) - \left(\frac{1-p_i}{1-e^{-\lambda_i}}\lambda_i\right)^2 = \mu_{N_i}(\lambda_i+1) - \left(\mu_{N_i}\right)^2$$

On the other hand, if the average claim amount follows a Gamma$\left(\mu_{Y_i}, \frac{\emptyset_Y}{N_i}\right)$ distribution, then

$$Var(\bar{Y}_i) = \frac{\emptyset_Y}{N_i}\left(\mu_{Y_i}\right)^2.$$

Therefore, the policy expected loss and variance, for this model, can be obtained by replacing these expressions in (3.4) and (3.5), respectively.

For the variance, we get:

$$Var(S_i) = E\left[{N_i}^2 \frac{\emptyset_Y}{N_i}\left(\mu_{Y_i}\right)^2\right] + \left(\mu_{Y_i}\right)^2\left[\mu_{N_i}(\lambda_i + 1) - \left(\mu_{N_i}\right)^2\right]$$

$$= \left(\mu_{Y_i}\right)^2 \mu_{N_i}\left[\emptyset_Y + \lambda_i + 1 - \mu_{N_i}\right].$$

# Chapter 4

# The dependent case

Although the independence assumption seems very helpful to simplify the model, it can lead to inaccurate results when the frequency and the severity are associated. In fact, there are cases where this assumption has been proved to be unrealistic and, consequently, models to account for dependence have been developed.

## 4.1. Conditional model

A good starting point to introduce this part is to enter in the conditional probability framework, as proposed by Gschlößl and Czado (2007) and Garrido et al. (2016), i.e, to consider models for the claim sizes that are conditional on the claim counts.

### 4.1.1. Policy loss

Without assuming independence, the expected individual aggregate loss becomes

$$E[S_i] = E[N_i \bar{Y}_i] = E\big[N_i E[\bar{Y}_i|N_i]\big], \tag{4.1}$$

Therefore, if we do not assume that claim sizes are independent from claim numbers, we can no longer use the product of their expected values. The problem that arises is, then, how to estimate this expected value.

The solution developed by the aforementioned authors starts by fitting a GLM, with a log-link, to the conditional severity, given the frequency. In this case, a modification is made in the severity component, by allowing the claim numbers to enter the model as a covariate. As a result, the conditional mean severity will be given by

$$E[\bar{Y}_i|N_i] = e^{\tilde{\boldsymbol{x}}_i{}'\tilde{\beta}^Y + \delta N_i} = e^{\tilde{\boldsymbol{x}}_i{}'\tilde{\beta}^Y} e^{\delta N_i} = \tilde{\mu}_{Y_i} e^{\delta N_i}, \tag{4.2}$$

where $\tilde{\beta}^Y$ and $\delta$ are the regression parameters, $\tilde{\boldsymbol{x}}_i$ and $N_i$ are the respective covariates, and $\tilde{\mu}_{Y_i} = e^{\tilde{\boldsymbol{x}}_i{}'\widetilde{\beta^Y}}$ is a modified marginal mean severity.

It can be easily derived that when $\delta \neq 0$, the vector of the regression parameters $\tilde{\beta}$ will be different from $\beta$ (independent case). This happens due to the existence of one more covariate (the claim numbers), which will affect the model. On the other hand, if $\delta = 0$, then the expected value will be reduced to the one of the independent case:

$$E[\bar{Y}_i|N_i] = \mu_{Y_i} \implies E[S_i] = \mu_{Y_i}\mu_{N_i}$$

Therefore, as in Garrido et al.(2016), we conclude that the independent model is nested in the dependent one.

By replacing (4.2) in (4.1), we get the formula for the expected individual loss,

$$E[S_i] = E\left[N_i\tilde{\mu}_{Y_i}e^{\delta N_i}\right] = \tilde{\mu}_{Y_i}M'_{N_i}(\delta) \tag{4.3}$$

where $M'_{N_i}(\delta)$is the derivative of the moment generating function (m.g.f.) of $N_i$, $M_{N_i}(s)$, defined at point $s = \delta$.

Then, to estimate the policy loss, a three-step approach can be followed. First, fit a regression model to the claim counts, $N_i$, like a Hurdle model, which allows to obtain $\hat{\mu}_{N_i}$ (equivalent to the frequency model under independence). Secondly, conditional on $N_i > 0$, fit a GLM regression model, such as a Gamma GLM, to the average claim size with the claim numbers as a covariate, and obtain $\hat{\tilde{\mu}}_{Y_i}$ and $\hat{\delta}$ . Lastly, assuming that the dispersion parameter for the number of claims is known, the individual expected loss can be estimated by replacing these estimates in (4.3).

The variance of the policy loss is more complex. Using some computation, it can be proved that

$$\begin{aligned}
Var(S_i) = Var(N_i\bar{Y}_i) &= Var(E[N_i\bar{Y}_i|N_i]) + E\left[Var[N_i\bar{Y}_i|N_i]\right] \\
&= Var(N_iE[\bar{Y}_i|\,N_i]) + E\left[{N_i}^2 Var[\bar{Y}_i|N_i]\right] \\
&= Var\left(N_i\tilde{\mu}_{Y_i}e^{\delta N_i}\right) + E\left[{N_i}^2\frac{\phi^D}{N_i}V_Y(\tilde{\mu}_{Y_i}e^{\delta N_i})\right] \\
&= {\tilde{\mu}_{Y_i}}^2\left[E\left[{N_i}^2e^{2\delta N_i}\right] - \left(E\left[N_ie^{\delta N_i}\right]\right)^{\mathbf{2}}\right] + \phi^D E\left[N_iV_Y(\tilde{\mu}_{Y_i}e^{\delta N_i})\right] \\
&= {\tilde{\mu}_{Y_i}}^2\left[\frac{1}{4}M''_{N_i}(2\delta) - \left(M'_{N_i}(\delta)\right)^{\mathbf{2}}\right] + \phi^D E\left[N_iV_Y(\tilde{\mu}_{Y_i}e^{\delta N_i})\right]
\end{aligned}$$

where $\phi^D$is the severity dispersion parameter in the dependent model and $M''_{N_i}(s)$ is the second derivative of the m.g.f. of $N_i$, defined at point $s = 2\delta$.

**Hurdle-Poisson model –** When the Poisson($\lambda$)-Hurdle model is considered, the m.g.f. is given by

$$M_{N_i}(\delta) = p + (1-p)\frac{e^{\lambda e^{\delta}}-1}{e^{\lambda}-1},$$

and its derivatives by

$$M'_{N_i}(\delta) = (1-p)\lambda e^{\delta}\frac{e^{\lambda e^{\delta}}}{e^{\lambda}-1} = \mu_{N_i}\exp\{\lambda\left(e^{\delta}-1\right)+\delta\}$$

and

$$M''_{N_i}(\delta) = \mu_{N_i}\exp\{\lambda\left(e^{\delta}-1\right)+\delta\}\left(1+\lambda e^{\delta}\right).$$

Furthermore, if a gamma distribution is considered, $V_Y\left(\tilde{\mu}_{Y_i}e^{\delta N_i}\right) = \left(\tilde{\mu}_{Y_i}e^{\delta N_i}\right)^2$ and

$$E\left[N_i V_Y(\tilde{\mu}_{Y_i}e^{\delta N_i})\right] = \frac{1}{2}\left(\tilde{\mu}_{Y_i}\right)^2 M'_{N_i}(2\delta).$$

Therefore, considering the conditional approach,

$$Var(S_i) = \tilde{\mu}_{Y_i}{}^2\mu_{N_i}\left[\lambda\exp\{\lambda\left(e^{2\delta}-1\right)+4\delta\}+(\phi^D+1)\exp\{\lambda\left(e^{2\delta}-1\right)+2\delta\}-\right.$$
$$\left.\mu_{N_i}\exp\{2\lambda\left(e^{\delta}-1\right)+2\delta\}\right].$$

As expected (because the models are nested), if we set $\delta = 0$ in this variance, we get the same result as the one of the independent case (section 3.2.4).

### 4.1.2. Estimation

To estimate the regression parameters, a maximum-likelihood approach is usually followed. To perform this task, the joint distribution of frequency and severity will be needed. It can be decomposed by

$$f_{\bar{Y},N}(y_i,n_i) = f_{\bar{Y}|N}(y_i)f_N(n_i)$$

Therefore, the joint likelihood and log-likelihood, considering *m* policyholders, will be, respectively,

$$L\left(\gamma,\beta^N,\tilde{\beta}^Y,\delta;\boldsymbol{y},\boldsymbol{n}\right) = \prod_{i=1}^{m} f_{\bar{Y},N}\left(y_i,n_i|\gamma,\beta^N,\tilde{\beta}^Y,\delta\right) = \prod_{i=1}^{m} f_{\bar{Y}|N}\left(y_i|\tilde{\beta}^Y,\delta\right)f_N(n_i|\gamma,\beta^N)$$

and

$$\ell\left(\gamma,\beta^N,\tilde{\beta}^Y,\delta;\boldsymbol{y},\boldsymbol{n}\right) = lnL\left(\gamma,\beta^N,\tilde{\beta}^Y,\delta;\boldsymbol{y},\boldsymbol{n}\right) = \ell_{\bar{Y}|N}\left(\tilde{\beta}^Y,\delta;\boldsymbol{y}|\boldsymbol{n}\right) + \ell_N(\gamma,\beta^N;\boldsymbol{n}).$$

To obtain the estimates of the regression parameters, a maximization of the log-likelihood function is done. From this formalization, it follows that the estimation of $\gamma$ and $\beta^N$ will only depend on the marginal log-likelihood $\ell_N(\gamma, \beta^N; \boldsymbol{n})$, as in the independent model. For $\tilde{\beta}^Y$ and $\delta$, we only need the conditional log-likelihood $\ell_{\bar{Y}|N}(\tilde{\beta}^Y, \delta; \boldsymbol{y}|\boldsymbol{n})$. Therefore, the estimation can be performed separately. Properties of conditional maximum likelihood estimators are discussed in Andersen (1970).

## 4.2. Copula regression model

Another way to allow for dependence is using a copula to construct a joint model by linking the marginal distributions of claim sizes and claim counts, as done by Czado et al. (2012) and Krämer et al. (2013). Furthermore, it allows to model also nonlinear correlations between them, in contrast with the conditional approach.

There are two main steps that should be followed in this approach: first marginal models should be fitted to each variable; and second, a parametric copula should be selected. The first step is identical to the one described in sections 3.2.1. and 3.2.2. (independent case), where a Hurdle model can be fitted to the frequency and a Gamma GLM to the severity. After this step, the marginal regressions are combined using a bivariate copula.

### 4.2.1. The bivariate Copula

A Copula is a multivariate distribution function whose univariate marginal distributions are uniformly distributed. It is used to model the dependence structure between random variables. In this text, the interest relies on bivariate copulas.

**Definition 4.1.** *Bivariate Copula*

A Bivariate (2-dimensional) Copula is a function C: $[0,1]^2 \rightarrow [0,1]$ with the following properties:

(1) $\forall u, v \in [0,1]$:
$C(u,0) = C(0,v) = 0$ and $C(u,1) = u$ and $C(1,v) = v$;

(2) $\forall u_1, u_2, v_1, v_2, \in [0,1]$ with $u_1 \leq u_2$ and $v_1 \leq v_2$:
$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$ .

Nelsen (2006) is a good introduction to copulas and their properties and, Frees and Valdez (1998) explore their application in actuarial science.

One of the most important results in the theory of copulas was established by Sklar (1959).

**Theorem 4.1. (Sklar's Theorem).** Let $F$ be a $n$ -dimensional distribution function with univariate marginals $F_1, \dots, F_n$. Then there exists a copula $C$ with uniform marginals such that

$$F(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)). \tag{4.4}$$

Conversely, if C is a copula and $F_1, \dots, F_n$ are distribution functions, then the function F defined by (4.4) is a joint distribution function with marginals $F_1, \dots, F_n$. ◂

Additionally, Sklar showed that, if the variables are continuous, then there is a unique copula representation.

A very convenient implication of this theorem is that we can model the marginals and the dependence separately.

## 4.2.2. The joint density function

In this section we will consider, again, the claim numbers, $N$, and the average claim size, $\bar{Y}$. Its joint density/probability mass function can be defined by

$$f_{N,\bar{Y}}(n, y|\theta) = \begin{cases} f_{zero}(0) & , n = 0 \; and \; y = 0 \\ \left(1 - f_{zero}(0)\right) \times f_{N,\bar{Y}|N>0}(n, y|\theta) & , n > 0 \; and \; y > 0 \end{cases} \tag{4.5}$$

where $f_{N,\bar{Y}|N>0}(n, y|\theta)$ can be expressed using a copula, as done by Czado et al. (2012) and Krämer et al. (2013).

Thus, considering only positive counts and amounts, with $F_{N,\bar{Y}|N>0}$ the joint distribution function, and $F_{N|N>0}$ and $F_{\bar{Y}}$ the univariate marginal distributions, according to Sklar's Theorem, there exists a bivariate copula C:[0,1]x[0,1]→[0,1] such that

$$F_{N,\bar{Y}|N>0}(n, y|\theta) = C(F_{N|N>0}(n), F_{\bar{Y}}(y)|\theta).$$

The parameter $\theta$ is the copula parameter and allows us to model de dependence between the variables. However, since we are dealing with discrete and continuous random variables, the copula C is not unique. Nonetheless, it continues to be appropriate to describe the dependence between them [Genest and Nešlehová (2007)].

Keeping in mind that one of the random variables is discrete ($F_{N,\bar{Y}|N>0}(n,y)$ is not differentiable with respect to $n$), $f_{N,\bar{Y}|N>0}\,(n,y|\theta)$ can be obtained by doing

$$f_{N,\bar{Y}|N>0}(n,y) = \frac{\partial}{\partial y} P(\bar{Y} \le y, N = n | N > 0)$$

$$= \frac{\partial}{\partial y}\left[P(\bar{Y} \le y, N \le n | N > 0) - P(\bar{Y} \le y, N \le n-1 | N > 0)\right]. \quad (4.6)$$

Using the Copula formulation and letting the copula' partial derivative, with respect to the first variable, be

$$D_1(u,v|\theta) \coloneqq \frac{\partial}{\partial u} C(u,v|\theta)\,,$$

the joint density (4.6) is given by

$$f_{N,\bar{Y}|N>0}(n,y|\theta)$$

$$= \frac{\partial}{\partial y}\left[C\left(F_{\bar{Y}}(y), F_{N|N>0}(n)\middle|\theta\right) - C\left(F_{\bar{Y}}(y), F_{N|N>0}(n-1)\middle|\theta\right)\right]$$

$$= f_{\bar{Y}}(y)\left[\,D_1\left(F_{\bar{Y}}(y), F_{N|N>0}(n)\middle|\theta\right) - \,D_1\left(F_{\bar{Y}}(y), F_{N|N>0}(n-1)\middle|\theta\right)\right]$$

A wide range of bivariate copulas and their properties can be found in Nelson (2006), such as the Elliptical copulas (Gaussian, Student-t) or the Archimedean copulas (Frank, Gumbel, Clayton). Some derivatives can also be found in Schepsmeiner and Stöber (2014). Table 4.1 contains information about some widely used copulas.

| **Family** | **Copula $C(u,v|\theta)$** | **Range of $\theta$** |
|---|---|---|
| Gauss | $\Phi_2(\Phi^{-1}(u), \Phi^{-1}(v)\|\theta)$ | $\theta \in ]-1,1[$ |
| Frank | $-\frac{1}{\theta}\log\left(1 + \frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}\right)$ | $\theta \in \mathbb{R}\backslash\{0\}$ |
| Gumbel | $\exp\left\{-\left((-log\,u)^{\theta} + (-log\,v)^{\theta}\right)^{\frac{1}{\theta}}\right\}$ | $\theta \in [1,\infty[$ |
| Clayton | $(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$ | $\theta \in ]0,\infty[$ |

**Table 4.1 – Characteristics of some (one-parameter) copula families**

To identify the most appropriate copula family, we should look for their proprieties, such as the tail behavior, as well as for the dependence structure and choose the one that corresponds to the data. Once a copula's family has been selected, the copula parameter must be estimated. Two copula families can also be compared, after the estimation, using the log-likelihood ratio test developed by Vuong (1989), which is appropriated to compare two non-nested models (further details in section 4.3).

### 4.2.3. Estimation

The estimation of the unknown parameters can be done using maximum-likelihood techniques. If $\boldsymbol{\varphi} = (\gamma, \beta_c^N, \beta_c^Y, \phi_c, \theta)$ is the vector of unknown parameters, the log-likelihood function will be

$$\ell(\boldsymbol{\varphi}|\boldsymbol{n}, \boldsymbol{y}) = \ell(\gamma|\boldsymbol{n}) + \ell(\beta_c^N, \beta_c^Y, \phi_c, \theta|\boldsymbol{n}, \boldsymbol{y}),$$

where $\ell(\gamma|\boldsymbol{n})$ is as defined in section 3.2.3 and

$$\begin{aligned}\ell(\beta_c^N, \beta_c^Y, \phi_c, \theta|\boldsymbol{n}, \boldsymbol{y}) &= \sum_{i:n_i>0} \log\left(f_{N,\bar{Y}|N>0}(n_i, y_i|\beta_c^N, \beta_c^Y, \phi_c, \theta)\right) \\ &= \sum_{i:n_i>0} \log\left(f_{\bar{Y}}(y_i|\beta_c^Y, \phi_c)\right) \\ &+ \sum_{i:n_i>0} \log\Big[ D_1\left(F_{\bar{Y}}(y_i|\beta_c^Y, \phi_c), F_{N|N>0}(n_i|\beta_c^N)\middle|\theta\right) \\ &- D_1\left(F_{\bar{Y}}(y_i|\beta_c^Y, \phi_c), F_{N|N>0}(n_i - 1|\beta_c^N)\middle|\theta\right)\Big].\end{aligned}$$

The parameters estimates will be given by

$$\hat{\boldsymbol{\varphi}} = \arg\max_{\varphi} \ell(\boldsymbol{\varphi}|\boldsymbol{n}, \boldsymbol{y}).$$

For $\gamma$, the estimation can be done separately and the estimates will be the same as for the independent case. For the second term of the log-likelihood, $\ell(\beta_c^N, \beta_c^Y, \phi_c, \theta|\boldsymbol{n}, \boldsymbol{y})$, the maximization should be done numerically and a variety of methods to do it can be found in the literature. Czado et al. (2012) used an algorithm based on maximization by parts (MBP) published in Song et al. (2005), to estimate the model parameters. Krämer et al. (2013) applied the BFGS optimization algorithm, which is a quasi-Newton method, to maximize the log-likelihood. Additionally, based on this last work, the package

**CopulaRegression**[1] in R was developed to describe the joint distribution of positive discrete and continuous random variables using various bivariate copulas (Gauss, Clayton, Gumbel and Frank). In this text, we proceed with the BFGS algorithm.

**BFGS algorithm**

The BFGS algorithm was published simultaneously by Broyden, Fletcher, Goldfarb, and Shanno in 1970. It is a quasi-Newton method used to solve nonlinear optimization problems without constraints. The algorithm uses the values of the objective function and its first and second derivatives. Furthermore, approximations of the hessian matrix (here denoted by H) are considered instead of the exact one.

Let $g(x) = -\ell(x)$, where $x$ is used instead of the vector $(\beta_c^N, \beta_c^Y, \phi, \theta)$ (to follow the usual notation in optimization). The problem that we seek to solve can be expressed as

$$\min_{x \in \mathbb{R}^n} g(x).$$

**Algorithm 4.1.** (BFGS algorithm)

Step 0: Let $k = 0$. Set an initial value $x^{(0)}$ and $H_0$ (usually the identity matrix);

Step 1: If stopping criteria are met, stop. Otherwise, continue.

Step 2: Compute the search direction, $p_k$, that satisfies $H_k p_k = -\nabla g\left(x^{(k)}\right)$;

Step 3: Compute step length $\alpha_k > 0$ that minimizes $g\left(x^{(k)} + \alpha_k p_k\right)$, and set $x^{(k+1)} = x^{(k)} + \alpha_k p_k$

Step 4: Compute $H_{k+1} = H_k - \frac{H_k s^{(k)} \left(s^{(k)}\right)^T H_k}{\left(s^{(k)}\right)^T H_k s^{(k)}} + \frac{y^{(k)} \left(y^{(k)}\right)^T}{\left(y^{(k)}\right)^T s^{(k)}}$

where $s^{(k)} = x^{(k+1)} - x^{(k)}$ and $y^{(k)} = \nabla g\left(x^{(k+1)}\right) - \nabla g(x^{(k)})$

Step 5: Set $k = k + 1$ and return to step 1. ▪

Further detailed information can be consulted in Nocedal and Wright (2006).

[1] https://cran.r-project.org/web/packages/CopulaRegression/CopulaRegression.pdf

The estimates $\widehat{\beta^N}, \widehat{\beta^Y}$ and $\hat{\phi}$, obtained by fitting the marginal models, will be used as initial values for the BFGS algorithm. For $\theta$, following the strategy proposed by Krämer et al (2013), the initial value will be the one that maximizes

$$\ell(\theta|\boldsymbol{u},\boldsymbol{v}) = \sum_{i=1}^{m} \log\left[ D_1(u_i, v_i|\theta) - D_1(u_i, w_i|\theta)\right],$$

where $u_i := F_{\bar{Y}}\left(y_i|\widehat{\beta^Y}, \hat{\phi}\right), v_i := F_{N|N>0}\left(n_i|\widehat{\beta^N}\right)$ and $w_i := F_{N|N>0}\left(n_i - 1|\widehat{\beta^N}\right)$ .

Additionally, given that the copula parameter $\theta$ is, in general, restricted (see Table 4.1), a transformation $h$: $\Theta \to \mathbb{R}$ is performed.

After running the algorithm, the estimates for the model parameters are obtained, as well as an approximation to the hessian matrix, which will be used to estimate the standard errors.

### 4.2.4. Dependence

To analyze the degree of dependence between the two random variables, a measure of association can be used. Since copulas are invariant under monotone transformations, then a scale-invariant measure is more appropriate, such as Kendall's tau and Spearman's rho, instead of the usual correlation coefficient [Ohlsson and Johansson (2010)].

**Definition 4.2.** Kendall's tau (population version)

Let $(X_1, Y_1)$ and $(X_2, Y_2)$ be i.i.d. random vectors. The population version of Kendall's tau is defined as

$$\tau_{X,Y} = \mathrm{P}\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - \mathrm{P}\{(X_1 - X_2)(Y_1 - Y_2) < 0\},$$

i.e., as the probability of concordance minus the probability of discordance. ▪

The relationship between the copula parameter $\theta$ and Kendall's tau, for the copula families mentioned in Table 4.1, can be found in Table 4.2.

| **Copula** | Gauss | Frank | Gumbel | Clayton |
|---|---|---|---|---|
| $\boldsymbol{\tau}$ | $\tau = \frac{2}{\pi} arcsin\,(\theta)$ | $1 - \frac{4}{\theta}\left[1 - \frac{1}{\theta}\int_0^{\theta} \frac{t}{e^t - 1} dt\right.$ | $\tau = \frac{\theta - 1}{\theta}$ | $\tau = \frac{\theta}{\theta + 2}$ |
| **Range of $\boldsymbol{\tau}$** | $\tau \in \mathbb{R}$ | $\tau \in \mathbb{R}\backslash\{0\}$ | $\tau \in [0, \infty[$ | $\tau \in ]0, \infty[$ |

**Table 4.2 – Relationship between the copula parameter, $\theta$, and Kendall's Tau, $\tau$.**

### 4.2.5. Policy loss

In the Copula model presented in this section, there is no direct formula for the expected policy loss, as in the Independent model or in the Conditional model. To obtain its estimate, Czado et al. (2012) proceeded with the use of Monte-Carlo Estimators and Krämer et al (2013) proceeded with the derivation of the policy loss' distribution (for positive losses). In the last case, they proved that, for policy $i$,

$$f_{S|S>0}(s|\theta) = \sum_{n=1}^{\infty}\left[D_1\left(F_{\bar{Y}}\left(\frac{s}{n}\right), F_{N|N>0}(n)\middle|\theta\right) - \right.$$
$$\left. - D_1\left(F_{\bar{Y}}\left(\frac{s}{n}\right), F_{N|N>0}(n-1)\middle|\theta\right)\right] \times \frac{1}{n} f_{\bar{Y}}\left(\frac{s}{n}\right). \quad (4.7)$$

Given that the density function of $S_i$ can be written as

$$f_S(s|\theta) = \begin{cases} f_{S|S>0}(s|\theta) \times (1 - f_{zero}(0)) & , s > 0 \ (n, y > 0) \\ f_{zero}(0) & , s = 0 \ (n = y = 0) \end{cases},$$

and using (4.7), the required expected value can be obtained by doing

$$E[S_i] = \int_0^{\infty} s \, f_{S|S>0}(s_i|\theta) ds \times (1 - f_{zero}(0)) .$$

Similarly, the variance will be

$$Var[S_i] = E\left[{S_i}^2\right] - (E[S_i])^2$$

where $E\left[{S_i}^2\right] = \int_0^{\infty} s^2 \, f_{S|S>0}(s_i|\theta) ds \times \left(1 - f_{zero}(0)\right)$.

## 4.3. Dependent vs Independent model

After fitting both models, overall goodness-of-fit measures can be used to compare them. For instance, the models can be ranked according to the Akaike Information Criterion (AIC), where the one with the lowest value is considered the best. The AIC is defined as

$$AIC = 2k - 2\log(L),$$

where $k$ denotes the number of parameters and $L$ denotes the value of the maximum likelihood.

Furthermore, in the conditional approach, as explained in section 4.1.1., the independent model can be obtained by imposing the restriction $\delta = 0$ in the dependent one, that is, the latter is nested in the former. Therefore, to investigate if the dependent model is significant, we should test if $\delta = 0$. This can be achieved with a two-tailed

hypothesis testing, where the null hypothesis is $H_o: \delta = 0$ and the test statistic is given by:

$$\mathrm{T}=\frac{\hat{\delta}}{\sqrt{\mathrm{Var}(\hat{\delta})}} \overset{a}{\sim} N(0,1)$$

Alternatively, since the severity models are nested, we can compare their deviances using the test statistic

$$\frac{D(y,\widehat{\mu_{Y_i}})-D(y,\widehat{\mu_{Y_i}^{d}})}{\hat{\phi}} = 2[\ell^d(\gamma,\beta^N,\widetilde{\beta^Y},\delta;\boldsymbol{y},\boldsymbol{n}) - \ell^I(\gamma,\beta^N,\beta^Y;\boldsymbol{y},\boldsymbol{n})] \sim \chi^2_{p_d-p_I} \qquad (4.8)$$

where $\ell^d(\cdot)$ and $\ell^I(\cdot)$ are the log-likelihood of the dependent and independent model, respectively; and $p_d - p_I$ is the excess of parameters of the dependent model over the independent model (which is 1 in this case).

Since in the copula approach the models are not nested, then the Vuong's test can be used. The test statistic is defined by

$$V = \frac{LR_m}{\sqrt{m}\widehat{\omega_m}} \overset{a}{\sim} N(0,1)$$

where $LR_m = \sum_{i=1}^{m} \ell_i^{(1)} - \sum_{i=1}^{m} \ell_i^{(2)}$, with $\ell_i^{(j)}$ the pointwise log-likelihood of model $j$ (j=1,2); and $\widehat{\omega_m}^2 = \widehat{Var}(\ell_i^{(1)} - \ell_i^{(2)})$. Hence, at 5%-significance level, model 1 is preferred for an observed test statistic higher than 1.96, while model 2 is preferred for an observed value smaller than -1.96. For other values, the test is inconclusive. This test can also be used to select the most appropriate copula's family to include in the Copula model.

# Chapter 5

# Data analysis

The models presented in the previous chapters will be now applied to insurance data. First, the description and treatment of the data set will be performed. Then the independent and dependent scenarios will be analyzed and compared, to investigate the effect of dependence in the estimation of both the policy and total losses. For this purpose, the software R was used.

## 5.1. Data description

The data set that will be analyzed in this chapter was provided by a Portuguese insurance company. It contains data on a portfolio of motor own damage insurance from a period of the beginning of this decade. Each policy is characterized by the policy number and the unit of risk (each vehicle), resulting in a total of 127 571 observations and in a total exposure of 103 478.8. For each observation, there is information about the claim numbers, the total claim amount and the exposure time, as well as a set of explanatory variables. Besides that, a new variable, called average claim amount, was created. For policies with at least one claim, it is given by the total claim amount divided by the number of claims. For policies with no claims, the average claim amount is zero.

The analysis of the given data revealed that most of the policies (around 94%) did not make any claim, and a maximum of 5 claims was registered. Furthermore, a total of 8 072 claims was found, i.e., the average claim frequency was 0.078 per policy/year. Information about the absolute and relative frequency of the claim counts can be found in Table 5.1.

| **Number of claims** | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **Frequency (number of observations)** | 120 101 (94.144%) | 6 943 (5.443%) | 467 (0.366%) | 48 (0.038%) | 9 (0.007%) | 3 (0.002%) |
| **Frequency (total exposure)** | 96 906.680 (93.649%) | 6 094.222 (5.889%) | 423.860 (0.410%) | 42.808 (0.041%) | 8.315 (0.008%) | 2.915 (0.003%) |

**Table 5.1 – Claim Count distribution**

Given that at least one claim had occurred, the mean of the average claim amount was 2 684.96 m.u. (monetary units) and a total loss of 21 673 003 m.u. was registered. Furthermore, Table 5.2 shows the mean of the average claim amount for each claim count. From this table, it can be observed that, in general, as the number of claims increases, the average severity decreases. This is also supported by Figure 5.1, where the plot of the severity against frequency reveals a possible negative association between these two variables, reflected by the negative slope of the regression line.

| **Number of claims** | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Average severity per claim** | 2 824.324 | 1 909.317 | 1 402.712 | 2 042.938 | 325.581 |

**Table 5.2 – Average Claim Severity for each claim count**

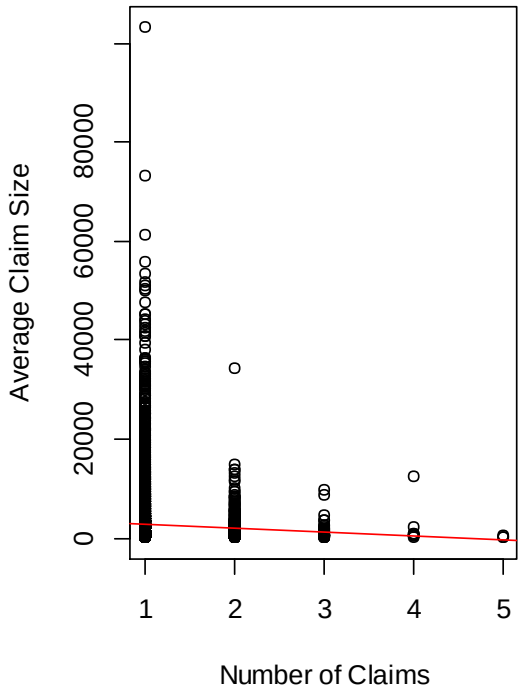


**Figure 5.1 – Plot of the Number of Claims (positive) against the Average Claim Size, with the corresponding regression line**

In addition, information on a total of 13 explanatory variables was available. Some of these variables are related to policyholder's characteristics, namely gender, age, driving license's age and his/her geographical area. However, due to a very high percentage of missing values for the policyholder's age, this variable was not considered. Furthermore, the policyholder's gender was also not considered due to moral reasons. Vehicle characteristics, such as fuel, type of vehicle, age, weight, number of seats and engine displacement were also given, as well as the capital insured, the deductible, and the bonus/malus class of the policyholder. All these covariates will enter the model as regressors and their significance will be evaluated. Since most of the variables were continuous, then a division into classes was performed for each one. Their labels, description, and categorization are presented in Table B.1 of appendix B.

## 5.2. Independent Model

To find the marginal models that best fit the data, an estimation of several generalized linear models was performed for the claim frequency and for the claim severity. For each component, first a model with all explanatory variables as main effects was considered and, then, the significance of the coefficients was evaluated. The final model was obtained by eliminating the non-significant covariates, with the help of significance and deviance tests. Interaction between covariates was not considered to keep the model simple.

### 5.2.1. Poisson-Hurdle model for frequency

A Binomial regression, with a logit link, was applied to "claims" vs "no claims" and a truncated Poisson regression, with a logarithmic link, was performed to the positive part (with at least one claim). This was done in R, using the function hurdle from the **pscl**[2] package. The final model can be found in Table B.2 of the appendix B, with all selected variables being significant, at least, at the 10% significance level. The remaining variables were excluded from the model and the exposure was included in the model as an offset.

From the estimated model, we observe that the bonus/malus classes, the insured capital and the deductible are statistically important for both the zero and the count part.

[2] https://cran.r-project.org/web/packages/pscl/pscl.pdf

The geographical area, fuel, driver's license's age and type and engine displacement of the vehicle are only statistically significant for the first part, while the weight of the vehicle is significant for the second part of the frequency model. Due to problems of significance, and to improve the model, some levels of some factors were merged. This happened for the capital insured (classes 2 to 4 and classes 5 to 7), the vehicle type (classes "MT", "MV" and "TT") and the driver's license's age (classes 3 and 4 and classes 7 to 11).

Furthermore, some conclusions about the claim experience can be taken. First, it is worth noticing that the intercept parameter in each model represents the value of the linear predictor for the reference group and that each estimated parameter represents the differential effect (positive or negative, depending on the sign) in the linear predictor with respect to that group. For the Hurdle model (as a whole), the reference group is composed of new drivers from the north of Portugal, in the highest level of bonus and in the lowest of capital insured, with no deductible and with a passenger diesel car with low weight and low engine displacement. Compared to this group, the claim experience is aggravated when the bonus class decreases or the malus class increases, or when the capital insured increases. On the other hand, a decrease in the reported claims is found when the deductible, weight or engine displacement of the vehicle or the license's age increases, as well as when another geographical area, type of vehicle or fuel is considered. Some of these results are intuitive and expected.

Note that the different sets of explanatory variables used for each model, in addition to support the choice of the Hurdle model, show that most of the factors are more important to explain the occurrence or not of a claim, than to explain the number of reported claims.

### 5.2.2. Gamma GLM for severity

Given the occurrence of claims, the average claim amount is a continuous, positive and right-skewed random variable, as supported by the histogram represented in Figure 5.2. Therefore, the gamma family is a justified option. Again, to obtain a multiplicative mean structure, a logarithmic link function was used.

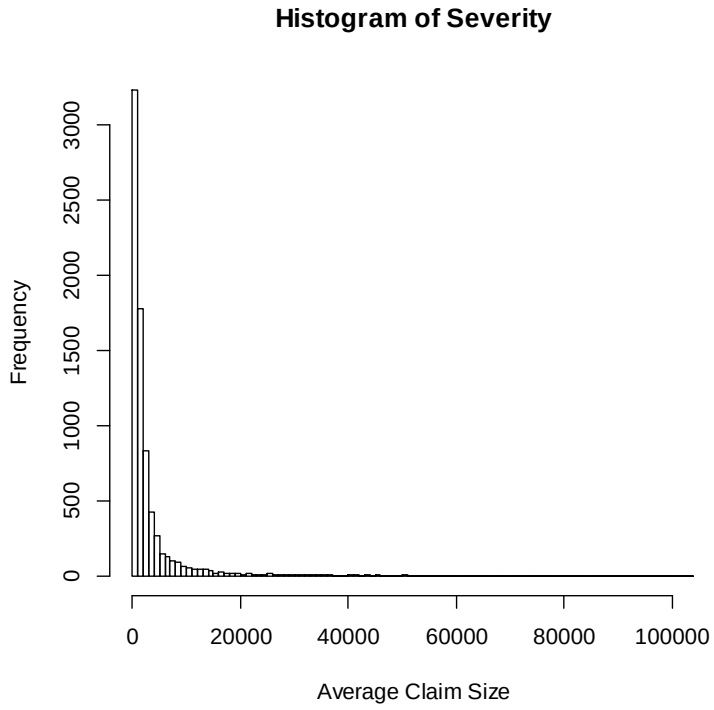


**Figure 5.2 – Histogram of Average Claim Size**

Since we are dealing with average amounts, then the claim numbers were included in the model as weights. Additionally, only the dataset containing positive claims was used to model the severity component, which comprises a total of 7 470 observations.

Using the same strategy as the one used for frequencies, the final model was obtained and the parameter estimates can be found in Table B.3 of appendix B. In this case, the classes of capital insured, deductible, vehicle age and license's age turned out to be significant at the 5% significance level. After merging the classes "MV" and "TT" of the variable vehicle type, as well as the classes "north" and "center" of the variable zone, both variables were also considered significant and were included in the model.

Here, the reference group is formed by new drivers from the north of Portugal, with capital insured up to 5000, with no deductible, and with a passenger car with 1 year or less. With respect to these policyholders, the average claim size increases with the increase of the capital insured, the deductible or the vehicle age, as well as when

considering a commercial car. A decrease is found for older drivers and for other zones and types of vehicles.

It is important to notice that, both in this final model and in the final model for frequency, no problems of multicollinearity were found among the covariates considered.

### 5.2.3. The independent model

After fitting the marginal models to claim frequency and claim severity, and assuming independence between the two, the individual expected losses can be estimated, as well as the total loss, as discussed in section 3.2.4.

Therefore, the estimate of the individual expected loss is obtained by only computing the product, for each policyholder, of the fitted value for the frequency and the fitted value for the severity. It ranged from 22.04 to 3 018 m.u.. Summing up all policies, an estimated total loss of 20 774 215 m.u. was found, with a standard deviation of 455 605.7 m.u.. Furthermore, the 98% confidence interval, for this model, is [19 714 476, 21 833 954].

The log-likelihood of the whole model is given by the sum of the marginal log-likelihoods, resulting in a total of – 100 508.1. Because the number of estimated parameter is 73, the AIC value is equal to $2 \times 73 - 2 \times (-100\,508.1) = 201\,162.2$.

## 5.3. Dependent Model and Comparisons

In this section, the independence assumption will be relaxed. First, the Conditional approach will be discussed. Then, the Copula model will be analyzed. In each case, a comparison to the independent model will be performed. Additionally, a final comparison between both models that account for dependence will be done, in order to see which one best captures this feature.

### 5.3.1. Conditional Approach

Using the Conditional model, only the marginal regression for the severity component needs to be investigated. As presented before, the marginal model for frequency remains

the same and, to account for dependence, the Gamma GLM for severity is modified by including now the number of claims as a covariate. As done by Garrido et al (2016), the number of claims was included as a discrete variable and not as a factor like the remaining variables of the model.

The estimated model is presented in Table B.3 of the appendix B. It was found that all the variables included in the independent case remained significant, with small changes in the estimated coefficients. The claim counts also turned out to be statistically significant, with an estimated parameter, $\hat{\delta}$, equal to -0.22716, which means that an increase in the number of claims decreases the average claim size. Thus, for a unit increase in the number of claims, the average claim size is expected to suffer a decrease of 20% ($1 - e^{-0.22716} = 1 - 0.7968 = 0{,}2032$).

Moreover, the statistical significance of the coefficient associated with the number of claims shows that there is statistical evidence that the dependence parameter is different from zero, which means that the independent model can be improved by considering dependence between claim numbers and amounts.

Besides the individual significance test, the other tests presented in section 4.3 can also be performed. The severity model under independence has an AIC value of 141 449, while under dependence it has a value of 141 373, which shows that the dependent model is slightly better. Moreover, because the models are nested, a comparison of the change in their deviances with the $\chi^2_{(1)}$ distribution can be done. Since the critical value at 5% of the $\chi^2_{(1)}$ distribution is 3.841, which is much smaller than 82.401 [the observed value of the test statistic (4.8)], the change in the deviance is statistically significant, meaning that the dependent model is indeed an improvement over the independent one.

When it comes to the total loss, an estimate of 20 857 168 m.u. was obtained, with a standard deviation of 455 684.9 m.u. and a 98% confidence interval equal to [19 797 245, 21 917 091]. The individual estimates ranged between 22.3 and 3 084 m.u.. These values are slightly higher than the ones estimated in the independent model. A possible explanation for this result might be the negative dependence found in the data, which causes an increase in the average claim size when the number of claims decreases. Thus, the fact that this data set is mostly constituted by a small number of claims (given the occurrence of claims, around 93% of the policies reported only one claim), leads to higher

estimates being obtained. Furthermore, when comparing the individual estimated losses, an average increase of 0.46% is found when dependence is considered.

### 5.3.2. Copula Approach

When it comes to the Copula model, the estimation is more complex. First, a copula must be selected. From the data analysis, and supported by the Conditional model, the dependence between the average claim size and the number of claims appears to be negative. Therefore, options like the Clayton or the Gumbel (standard) copulas are not appropriate, as they are not defined for negative values of Kendall's tau. The Gaussian copula, however, delivers good results when applied to the data, namely the strong significance of the dependence parameter.

Secondly, the truncated part of the Hurdle model and the Gamma GLM (both presented in section 5.2) allow us to select the covariates that will enter in the Copula model, as well as to obtain the parameter estimates, which are used as initial values in the BFGS algorithm. Note that the zero-Hurdle part remains unchanged, as its estimation is done separately and is already at the optimum. Therefore, only the data with at least one claim is needed to obtain the remaining final estimates ($\widehat{\beta_c^N}, \widehat{\beta_c^Y}, \widehat{\phi_c}$ and $\hat{\theta}$). This was done with the help of the **CopulaRegression** package in R, with changes in the function's code to accommodate the features of our model. The function used can be found in the appendix C.1, which is an adaptation of the function `copreg`**.** The main change was to include the estimated coefficients of the independent model as input in the copula regression. If this change had not been done, then these initial coefficients would be estimated only considering the truncated data set, instead of the complete one. Further details about the sub-functions used can be found in **CopulaRegression** and **VineCopula**[3] packages.

The final estimated model, using a gaussian copula, is presented in Table B.2 and Table B.3 of the appendix B. The additional parameter, $\theta$, which is the parameter that reflects the dependence, was estimated in - 0.233, with a standard error of 0.0231. This implies that the dependence is statistically significant. Additionally, it is equivalent to an

[3] https://cran.r-project.org/web/packages/VineCopula/VineCopula.pdf

estimated Kendall's tau of -0.1497, thereby implying, one more time, the existence of some negative dependence between the main variables. Note that, in the case of the gaussian copula, the dependence parameter $\theta$ is the same as the correlation coefficient.

When compared to the marginal regressions in the independent model, the parameter estimates kept their signs and suffered small changes. The standard errors decreased and all the variables were still significant.

The estimated total loss was 20 799 332 m.u., with a standard deviation of 386 914.2, and the 98% confidence interval was found to be [19 899 369, 21 699 294]. For each policyholder, the minimum estimated loss was 21.94 m.u. while the maximum was 3 033 m.u.. These values where obtained by first integrating the joint density function (4.7), after replacing its parameters by its estimates. The functions dpolicy_loss and epolicy_loss from **CopulaRegression** were used for this matter. The values obtained were then multiplied by $1 - \widehat{f_{zero}}(0)$, as described in section 4.2.5. Additionally, under this approach, an average increase of 0.16% in the individual losses is observed, compared to the independent case.

Finally, to the log-likelihood of the joint part of the copula model (- 67 594.32), we must join the log-likelihood referent to the zero-hurdle part (- 27 740.52), which makes a total log-likelihood of - 95 334.84. Since the total number of parameters estimated in the copula model is 74, then the resulting AIC value is 190 817.7. This value is smaller than the AIC of the independent model, which provides support to the Copula approach.

Besides the comparison of the AIC values, the Vuong test was performed, as the two models are not nested. The observed test statistic was found to be around 15.94. This value is much larger than the positive critical value at 5% of the standard Normal distribution (1.96), which means that the Copula model is an improvement over the independent frequency-severity model. This test was performed on R by applying the function testV, described in section C.2 of the appendix C.

Additionally, the Vuong test can also be used to compare the two dependent models presented in this text. In this case, the test statistic had an observed value of 21.74, meaning that the Copula approach outperforms the Conditional approach. This is not

surprising, as with the conditional approach only the linear association is captured. Nevertheless, both methods provided significant results.

Note that the observed total loss falls inside of all the confidence intervals. Furthermore, by comparing all the estimated total losses with the one observed, we can detect an underestimation in every model. Nonetheless, as only one observation is available, no conclusions can be drawn regarding under/overestimation.

Furthermore, in Figure 5.3 can be found the conditional density functions of the average claim size, given the number of claims (using the copula approach). If these quantities were independent, then the conditional distribution would be the same for all possible values of claim counts (and the same as the unconditional distribution). However, this is not the case. In Figure 5.3, changes in the conditional densities can be detected as the number of claims increases, illustrating the negative association found on the data. More details about the conditional density function can be consulted in sections A.3 and C.3 of the appendices A and C, respectively.

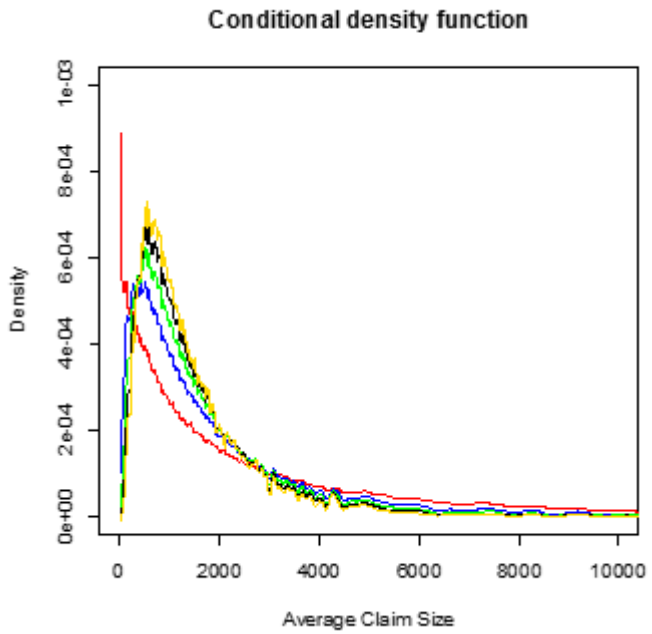


**Figure 5.3 – Conditional density functions of the Average Claim Size, given the number of claims (red: N=1; blue: N=2; green: N=3; black: N=4; gold: N=5)**

### 5.3.3. Estimated Premiums: case study

To conclude, Table 5.3. contains information on the estimated premiums, under each model, for four groups of policyholders. Furthermore, the absolute and relative variations of the dependent premiums, with respect to the independent ones, are also presented. The groups where chosen from a total of 44 742 risk cells, considering values around the independent premiums' quintiles (20%, 40%, 60% and 80%) with a minimum total exposure (38.38, 39.32, 36.85 and 32.28, respectively). Their characteristics are presented in the appendix B.1.

| | | **Conditional model** | | | **Copula model** | | |
|---|---|---|---|---|---|---|---|
| | **Independent model** | **Premium** | Δ | Δ% | **Premium** | Δ | Δ% |
| **1** | 101.91 | 99.57 | -2.34 | -2.29% | 100.42 | -1.49 | - 1.46% |
| **2** | 133.80 | 138.99 | +5.18 | +3.87% | 135.46 | +1.66 | +1.24% |
| **3** | 170.79 | 172.30 | +1.51 | +0.88% | 170.72 | -0.06 | - 0.04% |
| **4** | 271.57 | 284.76 | +13.19 | +4.86% | 275.18 | +3.60 | +1.33% |

**Table 5.3 – Estimated premiums under each model and the absolute (Δ) and relative (Δ%) variation with respect to the independent premium**

Regarding these risk cells, a change in the premiums was verified when dependence was considered. However, the difference between the models is not very large. Besides the policyholders' group 4, where the relative difference of the dependent premium, under the conditional model, is almost 5%, in the remaining groups no "significant" difference was found. When it comes to the whole set of risk cells, the scenario is identical and there are only few cases where the difference is indeed significant (increase/decrease larger than 5%). For instance, for policyholders belonging to risk cell number 5 (appendix B.1) the copula's premium was found to be 12.17% lower than the independent one (100.59 against 114.53, respectively). This group, however, has a small total exposure.

Nevertheless, the variation was more accentuated in the conditional dependent model than in the copula model, which was, in general, a common event in all the other risk cells of this data set.

# Chapter 6

# Conclusions

In this research, the assumption of independence between claim sizes and claim counts was relaxed to get more accurate premiums. In a Generalized Linear Model framework, two models that account for dependence were presented as alternatives to the widely used frequency-severity model, where the premiums are found by just computing the product between both components.

The first model considered was a conditional severity model, where the severity GLM was extended by including the number of claims as a covariate. This model has the interesting particularity of having the independent scenario as special case (the independent model is nested in the dependent one). Furthermore, it also has a closed form formula for the individual expected loss, which simplifies the computations. The improvement introduced by this model, in contrast with the independent scenario, was seen in the application made to the motor own damage insurance portfolio provided by a portuguese insurance company. The number of claims turned out to be highly significant in the severity regression and showed a negative relationship between both quantities: when the claim numbers increases one unit, a decrease of 20% in the average claim size is found.

The second model described was a copula regression model, which linked the marginal distributions and provided a joint model for frequency and severity. Information about the degree of dependence was given by the additional estimated parameter - the copula parameter. In the application, a gaussian copula was chosen and a negative dependence was found, with an estimated correlation coefficient equal to -0.233. With the help of Vuong tests, the copula model revealed to be the preferred one not only when compared to the independent case, but also when compared to the dependent conditional model. This last finding can be justified by the limitation of the conditional approach, which only models a linear dependence.

Both the conditional and the copula models resulted in slightly higher estimated total losses when compared to the independent scenario, as well as in an average increase

of the individual estimated losses (0.46% and 0.16%, respectively). The combination of a portfolio of small claim counts and the existence of negative dependence is one possible explanation for this finding. When it comes to the dispersion, the conclusions were very similar. From the analysis of the premiums, the differences between the models turned out to be small, with the copula model providing more conservative estimates than the conditional model. Although having considered dependence has improved the model, ignoring it does not lead to much different premiums in this portfolio.

The copula model, however, has the disadvantage of being computationally demanding and more time consuming, as the estimation is done using numerical methods. Oppositely, the conditional approach has a much simpler application.

Nevertheless, the three models have a common first step of fitting a marginal GLM to both the frequency and severity components. Although better fittings to each component could be found, the Gamma GLM for severity and the Hurdle-Poisson model for frequency were chosen. These choices allowed the use of a complete data set (including the zeros), instead of a truncated one as used by the authors mentioned in this text, and facilitated the comparison between the final models, which was one of our goals.

To sum up, insurance companies should question if assuming independence between claim counts and claim amounts is the best option when constructing tariffs, or if more accurate premiums could be achieved by considering the existence of dependence. As showed in this text, there are already models that provide good results in this field. This can help the insurer avoiding bigger losses in the future and becoming more competitive.

Future research could focus on analyzing more than two variables, for instance by including other type of coverages and considering the possibility of dependence between them. Additionally, other types of copulas (including copulas with more than one parameter) could be analyzed. The comparison between each copula would give a better understanding about the type of dependence of the data. Unfortunately, in the application made in this text only the gaussian copula provided good results, mostly due to the fact that the other considered copulas are only defined for positive dependence. A possible solution would be to make a copula rotation/transformation.

Nevertheless, a lot of interesting topics can be pursued and extensions can be made to other areas, namely when it comes to copulas.

# Appendix A

# Background

## A.1. Distributions

### A.1.1. Gamma distribution

If $Y \sim Gamma(\mu, \phi)$, with μ the mean and ø the dispersion parameter, then

$$f_Y(y|\mu, \emptyset) = \frac{1}{y\Gamma\left(\frac{1}{\phi}\right)}\left(\frac{y}{\mu\emptyset}\right)^{\frac{1}{\phi}} \exp\left(-\frac{y}{\mu\emptyset}\right), y > 0$$

and

$$E[Y] = \mu \text{ and } Var(Y) = \phi\mu^2.$$

### A.1.2. Poisson distribution and Zero-truncated Poisson distribution

If $N \sim Poisson(\lambda)$, then

$$f_N(n|\lambda) = \frac{\lambda^n}{n!}e^{-\lambda}, n = 0,1,2,\ldots$$

and

$$E[N] = Var(N) = \lambda.$$

However, if only positive values are assumed, then the zero-truncated Poisson variable N' is considered, with

$$f_{N'}(n|\lambda) = \frac{f_N(n|\lambda)}{1 - f_N(0|\lambda)} = \frac{\lambda^n}{n!\,(1 - e^{-\lambda})}e^{-\lambda}, n = 1,2, \ldots$$

and $[N'] = \frac{\lambda}{1-e^{-\lambda}}$.

### A.1.3. Bernoulli distribution

If $X \sim Bernoulli(p)$, then

$$f_X(x|p) = p^x(1-p)^{1-x}, \;\; x = 0,1$$

and

$$E[X] = p \text{ and } Var(X) = p(1-p)$$

## A.2. Bivariate Gaussian Copula

The partial derivative, with respect to the first variable, of the gaussian copula, as defined in Table 4.1., is given by

$$D_1(u, v|\theta) = \Phi\left(\frac{\Phi^{-1}(v) - \theta\Phi^{-1}(u)}{\sqrt{1-\theta^2}}\right)$$

The derivation of this result can be consulted in Czado et al. (2012).

## A.3. Conditional density function

The conditional density function of the average claim size, given the number of claims is defined in the copula approach as

$$f_{\bar{Y}|N}(y|n>0) = \frac{f_{N,\bar{Y}|N>0}(n,y)}{f_{N|N>0}(n)}$$

where $f_{N,\bar{Y}|N>0}(n,y)$ is as established in section 4.2.2. and $f_{N|N>0}(n)$ is the truncated count density function.

# Appendix B

# Categorization and Estimated models

## B.1. Categorization and description of the covariates

| Label | Description |
|---|---|
| **Classcapital** | Class of capital insured (m.u.)<br>Factor with 9 levels:<br>1 – up to 5000 4 – ]11 000, 15 000] 7 – ]21 000, 25 000]<br>2 – ]5 000, 8 000] 5 – ]15 000, 18 000] 8 – ]25 000, 35 000]<br>3 – ]8 000, 11 000] 6 – ]18 000, 21 000] 9 – above 35 000 |
| **Classdeduc** | Class of deductible (m.u.)<br>Factor with 4 levels:<br>1 – no deductible; 3 – deductible of 500;<br>2 – deductible of 125 or 250; 4 – deductible of 1 000 or more |
| **Zone** | Policyholder's geographical area (Portugal)<br>Factor with 3 levels:<br>1 – North (Viana do Castelo, Braga, Vila Real, Bragança, Porto, Aveiro, Viseu and Guarda)<br>2 – Center (Coimbra, Castelo Branco, Leiria, Santarém, Portalegre and Lisboa)<br>3 – South and Islands (Évora, Setubal, Beja, Faro, Açores and Madeira) |
| **classveh_age** | Class of the vehicle age (years)<br>Factor with 5 levels:<br>1 – [0,1] 3 – ]3,5] 5 – above 9<br>2 – ]1,3] 4 – ]5,9] |
| **Fuel** | Type of fuel<br>Factor with 2 levels:<br>1 – Diesel 2 – Gasoline |
| **Classdisplac** | Class of the vehicle's engine displacement<br>Factor with 4 levels:<br>1 – up to 1 600 2 – above 1600 |
| **Numseats** | Class of the vehicle's number of seats<br>Factor with 2 levels:<br>1 – up to 5 seats 2 – above 5 seats |
| **Classweight** | Class of the vehicle's weight (Kg)<br>Factor with 2 levels:<br>1 – up to 1 500 2 – above 1500 |
| **veh_cat** | Type of vehicle<br>Factor with 4 levels:<br>LP - Passenger car MV - Minivan<br>MT - Commercial car TT – Jeep |
| **Bm** | Bonus/Malus class (discount or penalization with respect to the base premium)<br>Factor with 9 levels:<br>1 – discount of 50% 6 – penalization of 10%<br>2 – discount above 30% and below 50% 7 – penalization of 20%<br>3 – discount of 25% or 30% 8 – penalization of 40%<br>4 – discount of 10% or 20% 9 – penalization above 40%<br>5 – no discount or penalization |

| Label | Description | | |
|---|---|---|---|
| **classlic_age** | Policyhoder's license age (years) class<br>Factor with 11 levels: | | |
| | 1 – [0, 4] | 5 – ]16, 20] | 9 – ]32, 46] |
| | 2 – ]4, 8] | 6 – [20, 24] | 10 – ]36,40] |
| | 3 – ]8, 12] | 7 – ]24, 28] | 11 – above 40 |
| | 4 – ]12, 16] | 8 – ]28, 32] | |

**Table B.1 – Categorization and description of the covariates**

The characteristics of the policyholders belonging to the groups presented in section 5.3.3. , are:

1 – class of capital 2, class of deductible 1, center of Portugal, class of vehicle age 4, gasoline, passenger car, class of vehicle displacement 1, class of vehicle weight 1, discount above 30% and below 50%, and class of license age 5;

2 – class of capital 8, class of deductible 2, center of Portugal, class of vehicle age 1, diesel, passenger car, class of vehicle displacement 1, class of vehicle weight 1, 50% discount, and class of license age 8;

3 – class of capital 8, class of deductible 2, north of Portugal, class of vehicle age 1, diesel, passenger car, class of vehicle displacement 1, class of vehicle weight 1, 50% discount, class of license age 5;

4 – class of capital 9, class of deductible 1, north of Portugal, class of vehicle age 1, diesel, passenger car, class of vehicle displacement 2, class of vehicle weight 2, 50% discount, and class of license age 11;

5 – class of capital 4, class of deductible 2, center of Portugal, class of vehicle age 3, diesel, minivan or jeep, class of vehicle displacement 1, class of vehicle weight 2, 50% discount, and class of license age 5.

# B.2. Estimated models

## B.2.1. Estimated Frequency models

| | Zero component | | Truncated count component | | | |
|---|---|---|---|---|---|---|
| | | | Independent Model | | Dependent Copula Model | |
| Variable | Estimates, $\hat{\gamma}$ | Standard error | Estimates, $\widehat{\beta^N}$ | Standard error | Estimates, $\widehat{\beta_c^N}$ | Standard error |
| (intercept) | -2.30983 | 0.10085 | -2.09725 | 0.20691 | -2.02268 | 0.20316 |
| bm2 | -0.06423 | 0.03069 | 0.20315 | 0.10650 | 0.21416 | 0.10537 |
| bm3 | 0.13012 | 0.04456 | 0.44430 | 0.14216 | 0.44061 | 0.14105 |
| bm4 | 0.36063 | 0.04262 | 0.25043 | 0.14378 | 0.26091 | 0.14279 |
| bm5 | 0.46945 | 0.04918 | 0.56556 | 0.15033 | 0.65370 | 0.14677 |
| bm6 | 0.97461 | 0.10708 | 0.71429 | 0.28117 | 0.72977 | 0.27570 |
| bm7 | 1.60320 | 0.10125 | 1.27708 | 0.18833 | 1.31180 | 0.18475 |
| bm8 | 1.43350 | 0.20342 | 1.32535 | 0.34812 | 1.42715 | 0.33590 |
| bm9 | 1.57295 | 0.17945 | 1.75772 | 0.24815 | 1.73481 | 0.24765 |
| classcapital2to4 | 0.09283 | 0.05229 | 0.49483 | 0.20413 | 0.42182 | 0.19980 |
| classcapital5to7 | 0.13153 | 0.05492 | 0.45605 | 0.20887 | 0.35555 | 0.20491 |
| classcapital8 | 0.14941 | 0.06149 | 0.71449 | 0.22349 | 0.59987 | 0.21977 |
| classcapital9 | 0.29270 | 0.06872 | 0.65766 | 0.25062 | 0.46856 | 0.24856 |
| classdeduc2 | -0.43900 | 0.02621 | -0.73629 | 0.09344 | -0.73787 | 0.09231 |
| classdeduc3 | -0.88171 | 0.03789 | -1.17470 | 0.18084 | -1.20707 | 0.18013 |
| classdeduc4 | -1.41116 | 0.11044 | -0.92571 | 0.49909 | -1.04952 | 0.50719 |
| classweight2 | | | -0.17059 | 0.09992 | -0.17555 | 0.09879 |
| zone2 | -0.16149 | 0.02653 | | | | |
| zone3 | -0.14090 | 0.03631 | | | | |
| Fuelgasoline | -0.10388 | 0.03220 | | | | |
| classdisplac2 | -0.05402 | 0.02809 | | | | |
| veh_catMTorMVorTT | -0.06123 | 0.03478 | | | | |
| classlic_age2 | -0.27954 | 0.09620 | | | | |
| classlic_age3&4 | -0.14755 | 0.08334 | | | | |
| classlic_age5 | -0.19772 | 0.08563 | | | | |
| classlic_age6 | -0.22219 | 0.08692 | | | | |
| classlic_age7to11 | -0.14137 | 0.08170 | | | | |
| | | | | | | |
| Log-likelihood | -27 740.52 | | -2 073.329 | | | |
| AIC | 59 713.7 | | | | | |

**Table B.2 – Estimated models for Frequency (Hurdle-Poisson model)**

## B.2.2. Estimated Severity models

| | **Independent model** | | **Dependent Conditional model** | | **Dependent Copula model** | |
|---|---|---|---|---|---|---|
| Variable | Estimates, $\widehat{\beta^Y}$ | Standard error | Estimates, $\hat{\delta}$and $\widehat{\widetilde{\beta^Y}}$ | Standard error | Estimates, $\hat{\delta}$and $\widehat{\beta_c^Y}$ | Standard error |
| (intercept) | 6.8206 | 0.1286 | 7.0902 | 0.1330 | 6.8439 | 0.0989 |
| classcapital2 | 0.3643 | 0.0834 | 0.3748 | 0.0820 | 0.3661 | 0.0640 |
| classcapital3 | 0.6025 | 0.0831 | 0.6192 | 0.0816 | 0.6063 | 0.0637 |
| classcapital4 | 0.7328 | 0.0834 | 0.7467 | 0.0820 | 0.7452 | 0.0638 |
| classcapital5 | 0.9063 | 0.0887 | 0.9118 | 0.0872 | 0.9125 | 0.0673 |
| classcapital6 | 0.9451 | 0.0922 | 0.9602 | 0.0906 | 0.9579 | 0.0702 |
| classcapital7 | 1.1851 | 0.0924 | 1.1895 | 0.0908 | 1.1862 | 0.0699 |
| classcapital8 | 1.2016 | 0.0912 | 1.2204 | 0.0897 | 1.2140 | 0.0687 |
| classcapital9 | 1.7444 | 0.0983 | 1.7713 | 0.0966 | 1.7539 | 0.0741 |
| classdeduc2 | 0.3958 | 0.0344 | 0.3742 | 0.0341 | 0.3911 | 0.0267 |
| classdeduc3 | 0.6036 | 0.0512 | 0.5711 | 0.0507 | 0.5950 | 0.0397 |
| classdeduc4 | 0.8595 | 0.1505 | 0.8220 | 0.1480 | 0.8516 | 0.1180 |
| classveh_age2 | 0.3029 | 0.0449 | 0.2949 | 0.0442 | 0.3043 | 0.0344 |
| classveh_age3 | 0.5900 | 0.0504 | 0.5801 | 0.0495 | 0.5909 | 0.0379 |
| classveh_age4 | 0.7554 | 0.0532 | 0.7426 | 0.0522 | 0.7544 | 0.0393 |
| classveh_age5 | 0.8003 | 0.0774 | 0.7871 | 0.0761 | 0.7969 | 0.0575 |
| veh_catMT | 0.1973 | 0.0820 | 0.2005 | 0.0806 | 0.2043 | 0.0632 |
| veh_catMVorTT | -0.1705 | 0.0523 | -0.1696 | 0.0514 | -0.1718 | 0.0406 |
| classlic_age2 | -0.3087 | 0.1223 | -0.2983 | 0.1201 | -0.3146 | 0.0947 |
| classlic_age3 | -0.3646 | 0.1113 | -0.3709 | 0.1093 | -0.3859 | 0.0862 |
| classlic_age4 | -0.4026 | 0.1071 | -0.4044 | 0.1052 | -0.4149 | 0.0829 |
| classlic_age5 | -0.4807 | 0.1052 | -0.4872 | 0.1034 | -0.4975 | 0.0814 |
| classlic_age6 | -0.5417 | 0.1066 | -0.5480 | 0.1048 | -0.5583 | 0.0824 |
| classlic_age7 | -0.5537 | 0.1091 | -0.5397 | 0.1072 | -0.5571 | 0.0845 |
| classlic_age8 | -0.5595 | 0.1090 | -0.5254 | 0.1070 | -0.5580 | 0.0843 |
| classlic_age9 | -0.7333 | 0.1117 | -0.7401 | 0.1098 | -0.7512 | 0.0863 |
| classlic_age10 | -0.6548 | 0.1105 | -0.6395 | 0.1086 | -0.6599 | 0.0855 |
| classlic_age11 | -0.7426 | 0.1064 | -0.7241 | 0.1045 | -0.7344 | 0.0825 |
| zone2to3 | -0.0659 | 0.0328 | -0.0773 | 0.0322 | -0.0714 | 0.0253 |
| numclaims,$\hat{\delta}$ | - | - | -0.2272 | 0.0330 | - | - |
| | | | | | | |
| Dispersion parameter | 1.989733 | | 1.920036 | | 1.184554 | |
| Theta | - | | - | | - 0.23299 | 0.02311 |
| | | | | | | |
| Log-likelihood | - 70 694.26 | | -70 655.74 | | | |
| AIC | 141 449 | | 141 373 | | | |

**Table B.3 – Estimated models for Severity (Gamma model)**

# Appendix C

# R Functions

## C.1. Copula Regression Function

To run this function, the **CopulaRegression** and **VineCopula** packages are needed.

**Input:**

betaY: estimated coefficients for the severity independent model, $\widehat{\beta^Y}$

betaN: estimated coefficients of the positive count component for the frequency independent model, $\widehat{\beta^N}$

delta: estimated dispersion parameter for the severity independent model, $\hat{\phi}$

x: $n$ observations of the positive Gamma variable

y: $n$ observation of the zero-truncated Poisson variable

R: $n \times p$ matrix of covariates, for the Gamma model

S: $n \times k$ matrix of covariates, for the zero-truncated Poisson model

family: bivariate copula family (1=Gauss, 3=Clayton, 4=Gumbel, 5=Frank)

exposure: exposure time for the zero-tuncated Poisson model

**Output:**

betaY_cop: estimated coefficients for the severity component of the copula model, $\widehat{\beta_c^Y}$

betaN_cop: estimated coefficients for the frequency positive component of the copula model, $\widehat{\beta_c^N}$

delta_cop: estimated dispersion parameter for the severity component of the copula model, $\widehat{\phi_c}$

theta_cop: estimated copula (dependence) parameter, $\hat{\theta}$

tau: estimated Kendall's tau, $\hat{\tau}$

sd.betaY_cop: estimated standard deviation for the severity component of the copula model

sd.betaN_cop: estimated standard deviation for the frequency positive component of the copula model

sd.g.theta_cop: estimated standard deviation of the copula parameter function

loglik: total log-likelihood

npar: number of estimated parameters

ll: log-likelihood evaluated at each observation

**Main-Function:**

```
CR<-function (betaY,betaN,delta,x, y, R, S = R, family = 1, exposure = rep(1, length(x))) {
  mu <- as.vector(exp(R%*% betaY))
  lambda <- as.vector(exp(S %*% betaN)) * exposure
  theta_initial <- BiCopEst(rank(x- mu)/(length(x) + 1), rank(y - lambda)/(length(y) + 1), family=family)$par
  tau_initial = BiCopPar2Tau(par = theta_initial, family = family)
  u <- pgam(x, mu, delta/y)
  v <- pztp(y, lambda)
  vv <- pztp(y - 1, lambda)
  foo <- function(para) {
    theta0 <- z2theta(para, family)
    out <- (-sum(log(D_u(u, v, theta0, family) - D_u(u, vv, theta0, family))))
    return(out)
  }
  para_initial <- theta2z(theta_initial, family)
  para.ifm <- optim(para_initial, foo, method = "BFGS")$par
  theta.ifm <- z2theta(para.ifm, family)
  tau.ifm <- BiCopPar2Tau(par = theta.ifm, family = family)
  joint <- mle_joint(betaY, betaN, theta.ifm, delta, x, y, R, S, family, exposure, TRUE, TRUE)
  betaY_cop <- joint$alpha;
  betaN_cop<- joint$beta;
  delta_cop <- joint$delta;
  theta_cop <- joint$theta;
  tau <- joint$tau
  sd.betaY_cop <- joint$sd.alpha;
  sd.betaN_cop <- joint$sd.beta;
  sd.g.theta_cop<-joint$sd.g.theta
  family <- joint$family
```

```
  ll <- joint$ll
  loglik <- sum(ll)
  npar <- length(betaY_cop) + length(betaN_cop) + 1
  outlist <- list(betaY_cop= betaY_cop, betaN_cop = betaN_cop, delta_cop = delta_cop, theta_cop =
theta_cop, tau = tau, sd.betaY_cop= sd.betaY_cop, sd.betaN_cop = sd.betaN_cop,  sd.g.theta_cop=
sd.g.theta_cop, loglik = loglik, npar = npar, ll = ll)
  class(outlist) = "copreg"
  return(outlist)
}
```

**Sub-functions:**

From the package **CopulaRegression**:

pgam: distribution function of a Gamma variable;

pztp: distribution function of a zero-truncated Poisson variable;

theta2z and z2theta: transformation of the copula parameter and its inverse, respectively.

D_u: copula partial derivative;

mle_joint: returns the estimated coefficients and the estimated copula parameter;

From the package ´VineCopula':

BiCopEst: returns the initial copula parameter;

BiCopPar2Tau: returns the initial kendall's tau.

Note: a change was made in the arguments of the functions pgam and dgam. Whenever these functions appeared, the third argument was changed from delta to delta/y. This was done because $\bar{Y}_i|N_i \sim gamma\left(\mu_{Y_i}, \frac{\phi_Y}{N_i}\right)$.

## C.2. Vuong test

To run this function, the package **nonnest2**[4] is needed.

[4] https://cran.rstudio.com/web/packages/nonnest2/nonnest2.pdf

**Input:**

y: $n$ observations of the number of claims

p: probability of making at least one claim ($n \times 1$ vector)

m_cop: model object returned from CR

mH: model object returned from hurdle

msev: model obect returned from glm (gamma independent or dependent model)

**Main function:**

```
testeV<-function(y,p,m_cop,mH,msev)
{
    ll_copula<-y
    ll_copula[y>0]=log(p[y>0])+ m_cop$ll
    ll_copula[y==0]=log(1-p[y==0])
    ll_hurdle<-llcont(mH)
    ll_gamma<-llcont(msev)
    ll_independent=NSin
    ll_independent[NSin>0]=ll_hurdle[NSin>0]+ll_gamma
    ll_independent[NSin==0]=ll_hurdle[NSin==0]
    kcopula=length(m_cop$betaN)+length(m_cop$betaY)+length(coef(mH,model="zero"))+2
    kindep=length(coef(mH))+ length(coef(msev))+1
    m=ll_copula-ll_independent
    aux=sqrt(length(m)) * mean(m)/sd(m)
    tstat=aux-((kcopula-kindep)*log(length(m))/2)
    return(tstat)
}
```

**Output:**

tstat: value of the test statistic

**Sub-functions:**

llcont: returns the log-likelihood evaluated at each observation. From package **nonnest2**

## C.3. Conditional density function

To run this function, the **CopulaRegression** package is needed.

**Input:**

y: conditioning value ($y = 1,2,3, \ldots$)

x: $n$ observations of the positive Gamma variable

mu: estimated expected value of the positive gamma variable

delta: estimated dispersion parameter of the gamma variable

lambda: estimated parameter of the zero-truncated Poisson

theta: estimated copula parameter

family: bivariate copula family (1=Gauss, 3=Clayton, 4=Gumbel, 5=Frank)

**Output:**

out: vector of the conditional density function

**Main function:**

```
dcond<-function (y, x, mu, delta, lambda, theta, family)
{
  y <- rep(y, length(x))
  out<-density_joint(x, y, mu, delta, lambda, theta, family, TRUE)/dztp(y,lambda)
  out[out < 0] = 0
  out[out > 1] = 1
  return(out)
}
```

**Sub-function:**

density_joint: joint density function from **CopulaRegression** package

# Bibliography

Andersen, E.B., (1970). Asymptotic properties of conditional maximum likelihood estimators, *Journal of the Royal Statistical Society,* 32, 283–301.

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

Frees, E. and Valdez, E.(1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2, 1–25.

Frees, E.W., Gao, J. and Rosenberg, M.A. (2011). Predicting the Frequency and Amount of Health Care Expenditures, *North American Actuarial Journal*, 15:3, 377-392.

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

Genest, C. and Nešlehová, J. (2007). A Primer On Copulas For Count Data, A*stin Bulletin,* 37(2), 475-515.

Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal,* 3, 202-225.

Jørgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall, London.

Jørgensen, B. and De Souza, M. (1994). Fitting Tweedie's compound Poisson model to insurance claim data, *Scandinavian Actuarial Journal*, 69-93.

Klugman, S.A.; Panjer, H.H. and Willmot, G.E. (2008). *Loss Models: From Data to Decisions*, 3rd ed., John Wiley & Sons, Hoboken NJ.

Krämer, N., Brechmann, E.C., Silvestrini, D. and Czado, C. (2013). Total loss estimation using copula-based regression models, *Insurance: Mathematics and Economics*, 53, 829-839.

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

Nelsen, R.B. (2006). *An Introduction to Copulas*, 2nd edition, Springer, Berlin.

Nocedal, J., and Wright, S. (2006). *Numerical Optimization*, 2nd ed., Berlin, New York: Springer-Verlag.

Ohlsson, E. and Johansson, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models, EEA Series*, Springer-Verlag, Berlin, Germany.

Schepsmeiner, U. and Stöber, J. (2014). Derivatives and Fisher information of bivariate copulas. *Statistical Papers,* 55, 525–542.

Sklar, A. (1959). Fonctions de Répartition à n Dimensions et Leurs Marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8, 229-231.

Song, P. (2007). *Correlated data analysis: modeling, analytics, and applications*, 1st ed, Springer, New York.

Song, P., Fan, Y. and Kalbfleisch, J. (2005). Maximization by parts in likelihood inference. *Journal of the American Statistical Association, 100*, 1145–1166.

Vuong, Q. (1989). Ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57, 307-333.

Zeileis, A., Kleiber, C. and Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software,* 27(8).