DOI: 10.5604/01.3001.0014.1211

PRZEGLĄD STATYSTYCZNY
R. LXIII – ZESZYT 3 – 2016

ALICJA WOLNY-DOMINIAK[1]

THE COPULA-BASED TOTAL CLAIM AMOUNT REGRESSION MODEL WITH AN UNOBSERVED RISK FACTOR

1. INTRODUCTION

The basic characteristic of an insurance portfolio is its heterogeneity, which means that individual risks generate different claim amounts. In view of this, assigning a single premium to each risk is unfair. Therefore, a common practice of any insurance company is ratemaking, which is defined as the process of classification of the risk portfolio into risk groups where the same premium corresponds to each risk. The grouping is done based on what is referred to as risk factors, which cause the portfolio homogeneity. The risk factors may be divided into:

– observed factors (observed at the conclusion of an insurance contract) – these are the factors that describe an insured person and an insurance subject, as well as a spatial variable (in the sense of the geographical region),
– unobserved factors – such as a driver's skills, the safety of a district where a property is located, a factor specific to each risk treated as a random variable with a certain distribution.

The current practice of insurance companies is to carry out ratemaking in two stages determined by the risk factors that are taken into consideration (cf. Dionne, 1989). The first stage is *a priori* ratemaking, which means dividing the risk portfolio into groups of risks that are homogeneous in terms of the observed factors. Then *a posteriori* ratemaking is carried out, when the unobserved risk factor is taken into account individually for each risk.

The ratemaking problem comes down to determining a premium for a homogeneous risk group, where a premium is understood as the expected total claim amount for a single risk. In the estimation, two separate models – the average value of claims (called a claim severity model) and the number of claims (called a claim frequency model) – are applied to a single risk. Due to the character of risk portfolios and insurance data, a common practice applied by insurance companies is to use generalized linearized models (GLM's – cf. De Jong, Heller, 2008; Frees, 2009; Ohlsson, Johansson, 2010; Antonio, Valdez, 2012; Wolny-Dominiak, Trzpiot, 2013; Wolny-

[1] University of Economics in Katowice, Faculty of Economics, Department of Statistical and Mathematical Methods in Economics, 50 1th May St., 40-287 Katowice, Poland, e-mail: woali@ue.katowice.pl.

-Dominiak, 2014). Owing to the progress in numerical algorithms for finding maximum values of the log-likelihood function and their numerical implementation in commercial and non-commercial software, GLM's have become a common practice in the Polish insurance market as well.

The above approach to ratemaking requires the independence between an average value of claims and the number of claims. The reason for this is that the expected total claim amount is understood as the product of the expected claim frequency and the expected claims severity. However, in the literature this assumption is called into question, as in Krämer et al. (2013) or Shi et al. (2015). The dependence between two random variables is accommodated by the copula and the authors propose a copula-based regression model in order to estimate the total claim amount. The interest of this paper is to extend this model taking into account an unobservable risk factor in the claim frequency model. This factor, called also unobserved heterogeneity, is treated as a random variable influencing the number of claims. Typically, in such a situation a mixed Poisson distribution is assumed, but for our purposes we propose to apply the zero-truncated distribution. The goal is then to estimate the expected value of the product of two random variables: the average value of claims and the number of claims for a single risk assuming the dependence between the average value of claims and the number of claims for a single risk and the dependence between the number of claims for a single risk and the unobservable risk factor. In the model, we construct the bivariate distribution, which gives us the opportunity to estimate this expected value using the Monte Carlo (MC) simulation.

In the paper we give the details of the theoretical aspects of the model as well as the empirical example. To acquaint the reader with the model operation, every step of the process of the expected value estimation is described and the **R** code is available for download (see http://web.ue.katowice.pl/woali/ and R code Team, 2014).

2. TOTAL CLAIM AMOUNT MODEL UNDER INDEPENDENCE

A starting point for *a priori* ratemaking is the total claim amount model, in which the random variables – the average value of claims and the number of claims for a single risk – are independent. Consider a portfolio of n property risks where the risk is understood as a random variable with a certain distribution, hereinafter denoted as S_i, $i = 1,\ldots,n$, representing the total claim amount for the i-th risk. If the number of claims for the i-th risk in the portfolio is marked as N_i and if i denotes the value of a single claim, the variable S_i may be expressed in the following form:

$$S_i = Y_{i1} + \ldots + Y_{iN_i},\ S_i = 0 \text{ if } N_i = 0. \tag{1}$$

The considerations presented below take into account only the risks for which at least one claim has occurred. Assuming that variables $Y_{i1},\ldots,Y_{iN_i}$, are independent and have

identical distributions, and that they are independent of N_i, the expected value and the variance of variable S_i may be expressed as follows:

$$\begin{aligned} E[S_i] &= E[Y_i N_i] = E[Y_i]E[N_i], \\ Var[S_i] &= E^2[Y_i]Var[N_i] + E[N_i]Var[Y_i]. \end{aligned} \tag{2}$$

The expected value $E[S_i]$ corresponds to the so-called *pure premium* for a single risk. This is the premium covering the risk, without any additional costs of insurance. If an insurance company has a mass portfolio of risks, which is the case for example in motor third part liability (MTPL) and motor own damage (MOD) insurance or in immovable property insurance, the claim frequency model and the claims severity model are used to estimate the pure premium $E[S_i]$. The parameters of the models are estimated using data included in insurance policies. This practice is described in detail in works authored by, for example, De Jong, Heller (2008), Frees (2009), Ohlsson, Johansson (2010), Cizek et al. (2011).

Modelling the total claim amount (not the pure premium) for a single risk, the following assumptions are commonly made in this approach:

– In the claim frequency model the number of claims for a single risk has the Poisson distribution $N_i \sim Pois(\lambda_i)$,
– In the claim severity model variables Y_{ik} have identical distributions coming from the exponential dispersion family of distributions with the same dispersion parameter $Y_i \sim EDM(\mu_i, \varphi_Y)$.

The heterogeneity of an insurance portfolio is described by regression coefficients introduced to the mean of both models:

$$\mu_i = \exp(\mathbf{x}_i^Y \boldsymbol{\beta}^Y), \ \lambda_i = E_i \exp(\mathbf{x}_i^N \boldsymbol{\beta}^N), \tag{3}$$

where $\boldsymbol{\beta}^Y = (\beta_0^Y, \beta_1^Y, ..., \beta_k^Y)^T$, $\boldsymbol{\beta}^N = (\beta_0^N, \beta_1^N, ..., \beta_k^N)^T$ are fixed-effect vectors corresponding with observed risk factors; $\mathbf{x}_i^Y$, $\mathbf{x}_i^N$ are i-th rows of the matrix of models $\mathbf{X}^Y$ and $\mathbf{X}^N$, respectively. E_i denotes the risk exposure (typically – the time of the policy duration). Then the total claim amount for a single risk is simply:

$$E[S_i] = \mu_i \lambda_i = E_i \exp(\mathbf{x}_i^Y \boldsymbol{\beta}^Y) \exp(\mathbf{x}_i^N \boldsymbol{\beta}^N). \tag{4}$$

It should be noticed that if no claim has occurred for the i-th risk, the number of claims $N_i = 0$, which means, naturally, that the value of variable Y_i should also be zero. However, only the average claim non-zero value is assumed in the claims severity model. Therefore, the zero-truncated distribution of the number of claims is assumed in the case under analysis. Assuming the Poisson distribution for the number of claims, the probability mass function with deleted zero values has the following form:

$$\Pr^{ZTPois}(N_i = k_i \mid k_i > 0, \lambda_i) = \frac{\Pr^{Pois}(N_i = k_i \mid \lambda_i)}{1 - \Pr^{Pois}(N_i = 0 \mid \lambda_i)} = \frac{\lambda_i^{k_i}}{[\exp(\lambda_i) - 1]k_i!} \tag{5}$$

where $\Pr^{Pois}(N_i = 0 \mid \lambda_i) = \exp(-\lambda_i)$. The expected value and the variance are $E(N_i) = \frac{\lambda_i \exp(\lambda_i)}{\exp(\lambda_i) - 1}$ and $Var(N_i) = \frac{\lambda_i \exp(\lambda_i)}{\exp(\lambda_i) - 1}[1 - \frac{\lambda_i}{\exp(\lambda_i) - 1}]$ respectively (cf. Cruyff, van der Heijden, 2008).

The parameters of frequency and severity models are usually estimated separately, using the maximum likelihood method. Finally, the estimated value of the expected total claim amount is obtained in the point estimation by plugging in coefficient estimators into formula (4). The same strategy can be used with respect to the variance value taking the formula (2).

Example 1 – total claim amount model under independence

In order to demonstrate the current practice, the insurance portfolio taken from (Wolny-Dominiak, Trzesiok, 2014) is investigated herein. The data comes from the former Swedish insurance company Wasa and concerns partial *casco* insurance for motorcycles in the period of 1994–1998. The frequency and severity models are assumed as $Y_i \sim Gamma(\mu,\varphi_Y)$ and $N_i \sim ZTPois(\lambda)$ without regressors. We use the maximum likelihood method (MLE) in the estimation. The fitted parameters are presented in table 1 below.

Table 1.

Estimates of parameters in claim severity-frequency model

Model	Parameters	Mean	Variance
Severity	$\hat{\mu} = 25437$ $\hat{\varphi}_Y = 2.03$	$\hat{E}[Y_i] = 25437$	$\sqrt{\hat{Var}[Y_i]} = 38651$
Frequency (without exposure)	$\hat{\lambda} = 0.04$	$\hat{E}[Y_i] = 1.024$	$\sqrt{\hat{Var}[Y_i]} = 0.16$

Source: own calculations.

Plugging values from table 1 into formula (4), estimated characteristics of the total claim amount are obtained. The quantiles of $\hat{E}[S_i]$ are presented in figure 1, taking into account the exposure to each risk.

The left-hand figure displays quantiles of the order from 0 to 0.95, while the right-hand one – quantiles of the order from 0.95 to 1.

Insurance companies use the above-described practice only if an assumption is made that the claim amount value Y_i is independent of the claim number N_i for the risk. If this assumption is rejected, a dependence between variables has to be accommodated. And this could be done using a copula.

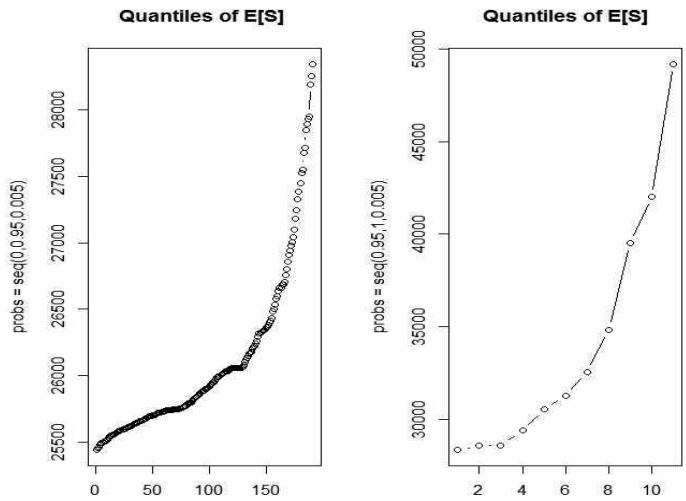

Figure 1. Quantiles of total claim amount

Source: own calculations.

3. DEPENDENCE WITH BIVARIATE COPULAS

The theory of copulas is frequently referred to in literature as in Joe (1997), Nielsen (1999), Wanat (2012). Here we give a short introduction for those who are not familiar with the subject. A bivariate copula $C(\cdot)$ is a two-dimensional cumulative distribution function (cdf) C : [0, 1] x [0, 1] $\rightarrow$ [0, 1] whose univariate margins are uniform on [0, 1]. For continuous random variables (X_1, X_2) with marginal cdf's $F_1(\cdot)$, $F_2(\cdot)$ and densities $f_1(\cdot)$, $f_2(\cdot)$, random variables of the form $U_1 = F_1(X_1)$, $U_2 = F_2(X_2)$ are also uniform on [0, 1]. According to the Sklar theorem (1959):

$$F_{X_1,X_2}(x_1, x_2) = C(F_{X_1}(x_1), F_{X_2}(x_2)). \tag{6}$$

Hence, the joint distribution $F(\cdot)$ is decomposed into marginal distributions and the copula $C(u_1, u_2)$, which captures the structure of the relation between X_1 and X_2. The corresponding joint density $f_{X_1,X_2}(\cdot)$ is then as follows:

$$f_{X_1,X_2}(x_1,x_2) = c(F_{X_1}(x_1), F_{X_2}(x_2)) f_{X_1}(x_1) f_{X_2}(x_2), \tag{7}$$

where $c(\cdot)$ denotes the copula density.

Generally, if a bivariate cdf of (X_1, X_2) exists, also a bivariate copula $C(\cdot)$ exists, and in the case of continuous random variables the copula is unique. However, the model proposed herein assumes mixed continuous and discrete variables.

Let us assume N is the count variable with a density function $f_N(\cdot)$ and consider a continuous-discrete random variable (Y, N). Let us focus on the parametric bivari-

ate copula with one parameter θ, such as the Gauss, Clayton or Frank copulas, which separates the dependence structure from margins. Denoting the partial derivative of the copula with respect to variable Y as $D(u_1,u_2)=\frac{\partial}{\partial u_1}C(u_1,u_2)$, $u_1, u_2 \in (0, 1)$, according to the formula (7), as is shown in Krämer et al. (2013) in the case of mixed outcomes, the joint density function $f_{Y,N}(\cdot)$ may be expressed as follows:

$$f_{Y,N}(y,k)=f_Y(y)\big(D(F_Y(y),F_N(k))-D(F_Y(y),F_N(k-1))\big). \tag{8}$$

In order to construct the above density function, the parameter vector of marginal distributions has to be estimated as well as the copula parameter θ. The inference functions for margins (IFM) method is used in this paper. It consists in estimating univariate parameters from separately maximized univariate likelihoods, and then estimating the copula parameter θ. Like in the above-described formula (8), only the margin of the first variable appears as the proper log-likelihood function giving the estimated value of θ in the following form:

$$l(\theta)=\sum_{i=1}^{n}\log\big(D(F_Y(y_i),F_N(k_i))-D(F_Y(y_i),F_N(k_i-1))\big). \tag{9}$$

Hence, the IFM method consists of three main steps (A1):

1. obtaining estimates of the vector parameters of margins,
2. transforming (y_i, k_i) to (u_{1i}, u_{2i}) as
   $u_{1i}=F_Y(y_i\,|\,\varphi_Y)$, $u_{2i}=F_N(k_i\,|\,\varphi_N)$, $u_{3i}=F_N(k_i-1\,|\,\varphi_Y)$,
3. optimizing $l(\theta)=\sum_{i=1}^{n}\log\big(D(u_{1i},u_{2i})-D(u_{1i},u_{2i})\big)$.

The example below illustrates the construction of density function $f_{Y,N}(y, k)$ for different types of one-parameter copulas $C(\cdot|\theta)$.

Example 2 – copula-based bivariate density construction

This example makes use of simulated data. The margins are taken as: $Y \sim Gamma(\mu, \phi_Y)$ with a mean μ and a dispersion ϕ_Y and $N \sim Poisson(\lambda)$ with a mean λ. Data (y_i, k_i), $i = 1,\ldots,100$ are drawn from $Gamma(\mu = 300, \phi = 1.5)$ and $Poisson(\lambda = 1)$. Assuming the parameter vector of margins as (300, 1.5, 1), observations (y_i, k_i) are transformed into (u_{1i}, u_{2i}) in the following way:

$$u_{1i}=F_Y(y_i\,|\,300,1.5),\ u_{2i}=F_N(k_i\,|\,1),\ u_{3i}=F_N(k_i-1\,|\,1) \tag{10}$$

assuming that $k_i = 0$ for $k_i < 0$. Finally, the copula parameter θ is estimated using the Gauss and Frank copulas and the copula-based density function $f_{Y,N}(\cdot)$ is constructed.

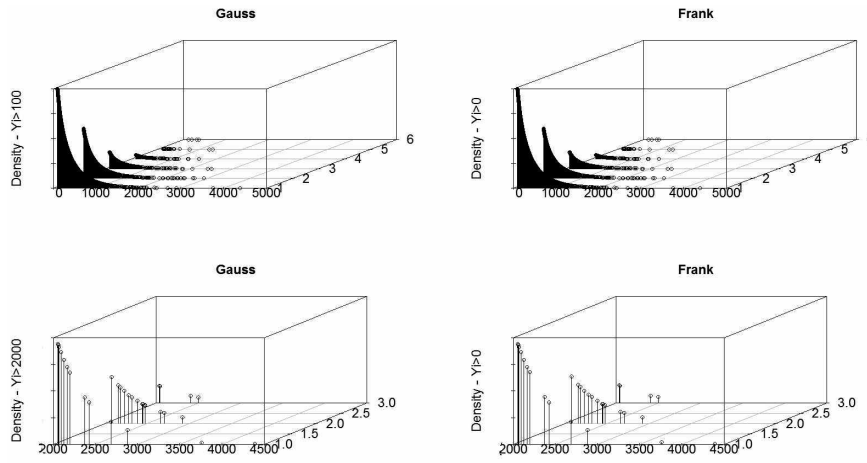

Fig. 2. Bivariate density of the random variable (Y_i, N_i)
Source: own calculations.

The IFM method is useful for models with the closure property of parameters being expressed in lower-dimensional margins. In addition, due to the fact that each inference function derives from a log-likelihood of a marginal distribution, the inference does not have to be obtained explicitly and numerical optimizations can be carried out for the log-likelihoods of margins. For this purpose, the BFGS algorithm implemented in **R** is used in this paper (see optim function).

4. COPULA-BASED TOTAL CLAIM AMOUNT MODEL

If it is assumed that the average claim value Y_i and the number of claims N_i are dependent random variables, the total claim amount S_i is defined as the following product:

$$S_i = Y_i N_i, \quad i = 1,\ldots,n. \tag{11}$$

The variable obtained in this way is a continuous variable with positive values. Due to the occurrence of interrelations between random variables Y_i, N_i, the expected value of variable S_i has the following form:

$$\mathrm{E}[S_i] = E[Y_i N_i], \tag{12}$$

which means that the frequency-severity model does not apply here. Using therefore the basic formula for the expected value, the following is obtained:

$$E[S_i] = E[Y_i N_i] = \int_0^{+\infty} s_i f_S(s_i \mid \varphi_Y, \varphi_N, \theta) ds_i, \tag{13}$$

where $s_i = y_i$, k_i, $y_i > 0$, $k_i = 1, 2, 3,\ldots$, and $f_S(\cdot)$ is the density function of the variable S_i. If it is assumed that the relation between variables Y_i, N_i is described by the copula $C(\cdot|\theta)$, then according to theorem 6 in Krämer et al. (2013) the distribution of the total claim amount is given by the following density function:

$$f_S(s_i) = \\ = \sum_{k_i=1}^{\infty}[D(F_Y(\frac{s_i}{k_i}\,|\,\varphi_Y), F_N(k_i\,|\,\varphi_N)) - D(F_Y(\frac{s_i}{k_i}\,|\,\varphi_Y), F_N(k_i - 1\,|\,\varphi_N))]\frac{s_i}{k_i} f_Y(\frac{s_i}{k_i}\,|\,\varphi_Y) \tag{14}$$

for $s_i > 0$. It can be seen that the function has a complex form and the expected value $E[S_i]$ cannot be determined analytically and a numerical procedure has to be used. This paper puts forward the following algorithm (A2):

1. obtaining the vector parameters of margins and the copula parameter $C(\cdot|\theta)$ using the IFM method $(\varphi_Y, \varphi_N, \theta)'$ under the assumption of the family of copulas,
2. obtaining the value of $f_S(s_i\,|\,\hat{\varphi}_Y, \hat{\varphi}_N, \hat{\theta})$ according to (14).

It gives the opportunity to obtain the value of expectation $E[S_i]$ and the value of variance $Var(S_i) = \int_0^{+\infty} s_i^2 f_S(s_i) ds_i - \hat{E}^2[S_i]$ through numerical integration.

The advantage of the proposed procedure is its flexibility. Any model can be used to determine the initial values needed to estimate the copula parameters in point 1. In the case of insurance applications, it is convenient to adopt the frequency and severity model with the independence assumption (cf. Section 2 above). Unfortunately, the downside of the algorithm is its relatively slow operation, which is the effect of the need to sum up in step 2 and perform numerical integration in steps 3 and 4.

Example 3 – estimation of the total claim amount expectation using the copula-based model without unobserved heterogeneity

The model is illustrated using the same portfolio as in Example 1, but the structure of the relation between Y_i and N_i changes. It is accommodated by the two-dimensional copula C with the parameter θ. Assuming margins $Y_i \sim Gamma(\mu_i, \varphi_Y)$ and $N_i \sim ZTPois(\lambda)$, the algorithm (A2) is run in the case of four families of parametric bivariate copulas: the Gauss, Clayton, Gumbel and Frank copulas. As the IFM method is applied, the parameters of margins are the same as in Example 1. Using these values, the copula parameters and the corresponding Kendall coefficient τ are obtained. The results are listed in table 2.

Table 2.

Estimation of Kendall's tau and copula parameter

Copula	Gauss	Clayton	Gumbel	Frank
$\hat{\theta}$	0.48	2.8)	1.21	4.71
$\hat{\tau}$	0.32	0.58	0.17	0.44

Source: own calculations.

Based on the estimators presented above and using formula (14), the copula-based density of the total claim amount is constructed. Next, the expected values $E[S_i]$, $i = 1,\ldots,666$ are estimated through numerical integration. Figure 3 displays histograms of $\hat{E}[S_i]$ for different copulas.

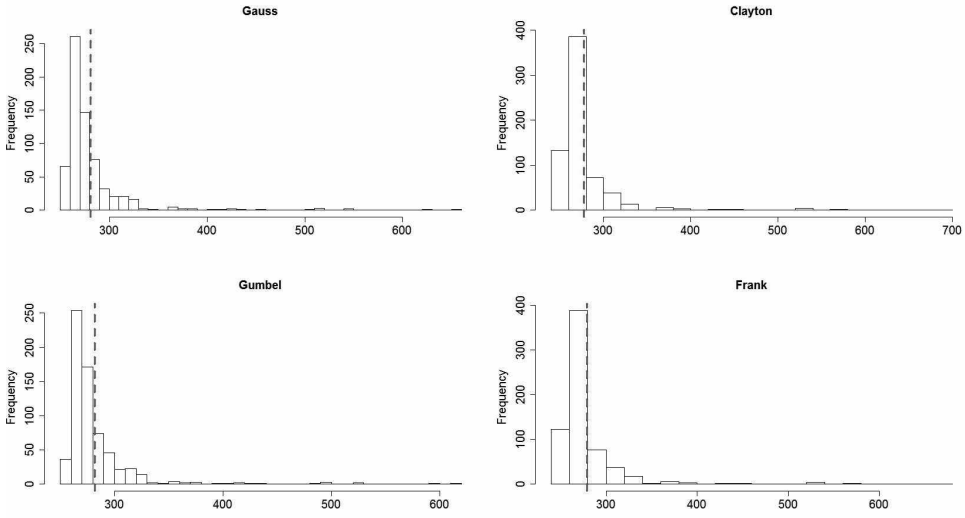

Figure 3. Histogram of expected total claim amount

Source: own calculations.

5. THE COPULA-BASED TOTAL CLAIM AMOUNT MODEL WITH AN INDIVIDUAL UNOBSERVABLE RISK FACTOR

Another starting point for *a posteriori* ratemaking are total claim amount models where the individual unobserved factor for the *i*-th risk, referred to as the risk profile (cf. Bühlmann, Gisler, 2005), is taken into account. This risk profile is usually taken into consideration in the claim frequency model using cross-sectional data (cf. Dimakos, Rattalma, 2002; Denuit et al., 2007; Boucher et al., 2007; Wolny-Dominiak, 2014) or longitudinal data (cf. Boucher et al., 2009; Wolny-Dominiak, 2014). It is well-known that this quantity is also affected by individual unobserved factors. One example is

motor insurance, where the unobserved factor is equated with a driver's (an insured person's) individual features that have an impact on a given risk loss burden. A driver with a strong aversion to driving fast, with little children etc., will display a weaker tendency towards causing claims to arise than a daring driver. Most frequently, the unobserved risk factor is treated as a realization of a certain random variable with a pre-set probability distribution.

5.1. MARGINAL CLAIM FREQUENCY

Let us assume that the unobserved risk factor corresponding to unobserved heterogeneity defines the continuous random variable V with the density function $f_V(\cdot)$ with the parameter vector φ_V. In the copula-based total claim amount model, a proposal is made to introduce the factor into the marginal frequency model as a random effect V. Consequently, as in the mixed Poisson model (cf. Denuit et al., 2007), the parameter λ_i of the model $ZTPois(\lambda_i)$ is randomized by $\lambda_i V$, which gives a conditional distribution of the number of claims $N_i \mid V \sim ZTPois(\lambda_i V)$ with the mass probability function defined by the following formula:

$$P[N_i = k_i \mid V] = \frac{(\lambda_i V)^{k_i}}{[\exp(\lambda_i V) - 1]k_i!}, \quad k_i > 0. \tag{15}$$

The claim number distribution requires a transition from the conditional distribution to the marginal one. One possibility is the direct use of the conditional distribution and a formula for the infinite mixture of distributions of the number of claims and the unobserved factor:

$$P[N_i = k_i \mid k_i > 0] = \int_o^{+\infty} P^{ZTPois}[N_i = k_i \mid u] f_V(v \mid \varphi_V) du. \tag{16}$$

As it can be seen, for any density function $f_V(\cdot)$ the estimation of the distribution parameters is a complex task due to the occurrence of the random effect $\lambda_i V$. The direct use of formula (16) then requires numerical integration, which involves considerable lengthening of the computation time. Another possibility is to use the Expectation-Maximization (EM) method, which is also rather time-consuming (cf. Karlis, 2001; Trzęsiok, Wolny-Dominiak, 2015). On the other hand, the probability function (16) can sometimes be determined analytically. One example is the popular negative-binomial (NB) distribution, which is a *Poisson-Gamma* distribution mixture. Assuming that $V \sim Gamma(\alpha)$ and $N_i \mid V \sim ZTPois(\lambda_i V)$, the marginal distribution of the number of claims is a first-order NB distribution.

The zero-truncated distribution with an unobserved factor can be obtained easily in the same way as in the case of the ZTPois distribution.

$$\Pr^{ZT}(N_i = k_i \mid k_i > 0) = \frac{\Pr(N_i = k_i)}{1 - \Pr(N_i = 0)}. \tag{17}$$

For example, the zero-truncated NB (ZTNB) distribution has the following probability mass function:

$$\Pr^{ZTNB}(N_i = k_i \mid k_i > 0, \lambda_i, \alpha) = \frac{\Pr^{NB}(N_i = k_i \mid \lambda_i, \alpha)}{1 - \Pr^{NB}(N_i = 0 \mid \lambda_i, \alpha)}, \tag{18}$$

where $\alpha > 0$ is a dispersion parameter. The probability of the occurrence of zero is then: $\Pr^{NB}(N_i = 0 \mid \lambda_i, \alpha) = (1 + \alpha\lambda_i)^{-\alpha^{-1}}$ and the expectation $E_{ZTNB}[N_i] = \dfrac{\lambda_i}{1 - (1 + \alpha\lambda_i)^{-\alpha^{-1}}}$.

In order to estimate ZTNB parameters one can use the MLE method. The log-likelihood function is defined as follows:

$$\begin{aligned} l(\boldsymbol{\beta}^N, \alpha) = \sum_{i=1}^{n} [&\log\Gamma(k_i + \frac{1}{\alpha}) - \log\Gamma(\frac{1}{\alpha}) - \log k_i! - (k_i + \frac{1}{\alpha})\log(1 + \alpha\lambda_i) + \\ &+ k_i \log \alpha\lambda_i - \log(1 - (1 + \alpha\lambda_i)^{-\alpha^{-1}})], \end{aligned} \tag{19}$$

where regression coefficients are introduced into the model through the parameter $\lambda_i = E_i \exp(\mathbf{x}_i^N \boldsymbol{\beta}^N)$.

Example 4 – parameter estimation and construction of a ZTNB distribution

To illustrate the ZTP model with unobserved heterogeneity Gamma distributed, which gives a ZTNB distribution, we simulate the sample $n = 500$ of the numbers of claims distributed as $N_i \sim NB(\lambda = 2, \alpha = 0.67)$. Then, we truncate the sample receiving zero-truncated data. Maximizing the log-likelihood (19) with the BFGS method, the estimated parameters are $\hat{\lambda} = 1.91$, $\hat{\alpha} = 0.64$. Figure 4 provides the probability function and the cdf of the constructed ZTNB (equivalent to ZTP-Gamma) based on the NB with parameters $(\hat{\lambda}, \hat{\alpha})$.

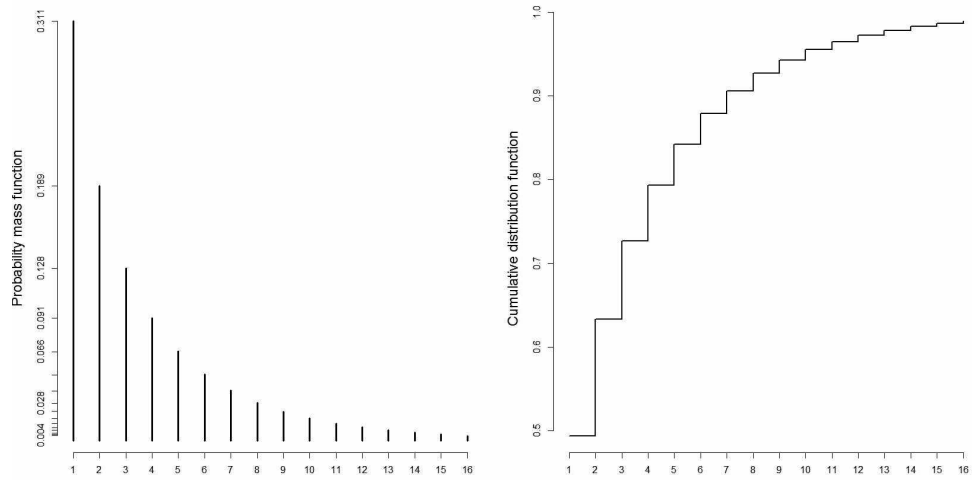

Figure 4. The probability function and the cumulative distribution function of ZTNB
Source: own calculations.

5.2. TOTAL CLAIM AMOUNT

Proceeding to the copula-based total claim amount model with the unobserved factor, three random variables are considered: the average claim value Y_i, the number of claims N_i and the unobserved factor V. The total claim amount is defined according to formula (2), except that the distribution of variable N_i is the marginal distribution of the two-dimensional variable (N_i, V). A proposal is made in this paper to determine the expected value of the total claim amount using the ZTNB distribution. It means that the unobserved factor is taken into account in the margin of the number of claims. The new procedure (A3) has the following steps:

1. obtaining the vector parameters of the number of claims (λ_i, α) assuming $N_i \sim ZTNB(\lambda_i, \alpha)$ and the regression component $\lambda_i = E_i \exp(\mathbf{x}_i^N \boldsymbol{\beta}^N)$,
2. obtaining the vector parameters of the average value of claims φ_Y assuming $Y_i \sim EDM(\mu_i, \varphi_Y)$ and the regression component $\mu_i = \exp(\mathbf{x}_i^Y \boldsymbol{\beta}^Y)$,
3. obtaining the copula parameter $C(\cdot|\theta)$ using the IFM method under the assumption of the copula type,
4. obtaining the value of $f_S(s_i \mid \hat{\mu}_i.\hat{\varphi}_Y, \hat{\lambda}_i, \hat{\alpha}, \hat{\theta})$ according to (14).

The constructed density of the total claim amount for a single risk gives the opportunity to estimate a pure premium. In the example below the proposed model is illustrated using real data from a Polish insurance company. As data is confidential, one can use another database in the **R** code.

Example 5 – total claim amount model with unobserved heterogeneity

We consider the portfolio that consists of 1,276 MOD (Motor Own Damage) policies insured in 2010 with the observed average value of claims Y_i and the number of claims N_i for every policy. The exposure E_i is taken as the duration of the policy. The histograms of random variables are shown in figure 5. The right-hand side is generated by the product Y_iN_i. The red lines represent the means.

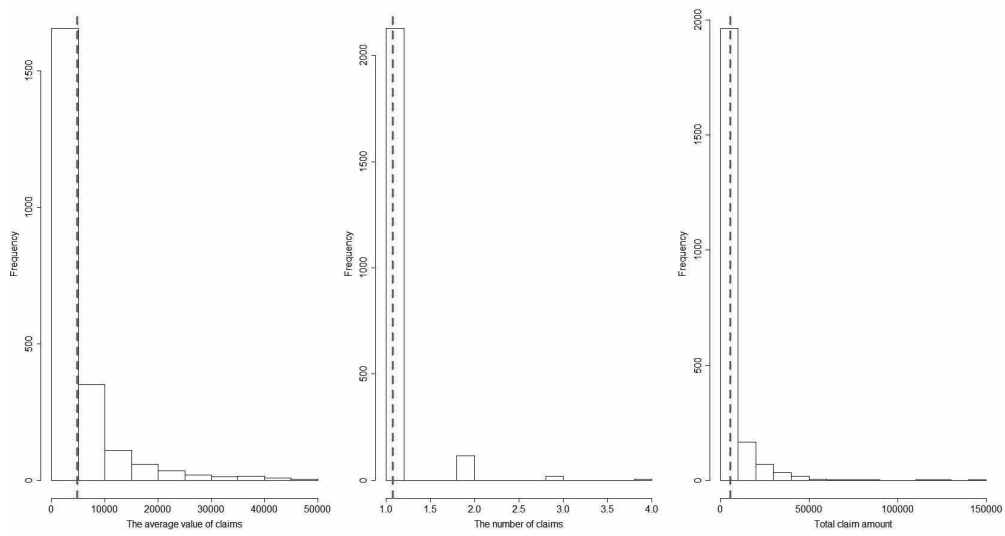

Figure 5. Histograms of the average value of claims, the number of claims and total claim amount for single risk

Source: own calculations.

The portfolio consists of three categorical covariates. Details on the factors are given in table 3.

Table 3.

Details on rating factors

Rating Factors	Categories/Number of observations		
POWER RANGE	0–66 269	67–124 803	125+ 187
GENDER	0 (Female) 416	1 (Male) 843	
PREMIUM_SPLIT	0 (No split) 754	1 (Split) 505	

Source: own calculations.

First, we analyze marginal models. As we see, the skew histogram of the average value of claims in the figure 4, the gamma distribution $Y_i \sim Gamma(\mu_i,\varphi_Y)$ is assumed. Figure 6 provides the boxplot divided according to the factor GENDER.

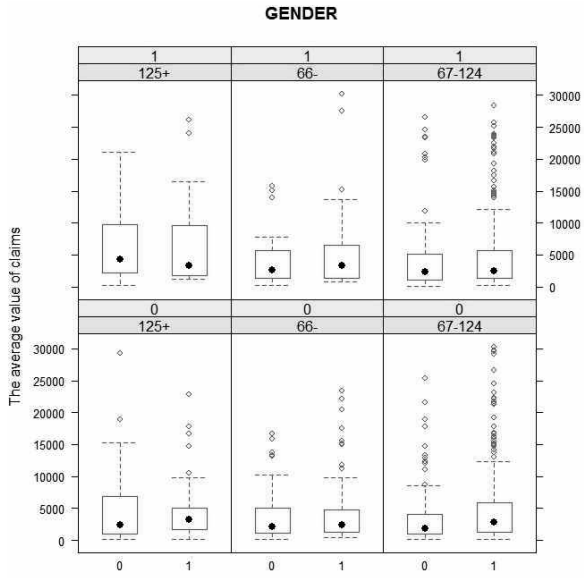

Figure 6. Boxplots of the average value according to the factor GENDER
Source: own calculations.

The Gamma assumption gives the opportunity to estimate the model parameters using the IWSL algorithm as in the standard practice in the GLM. As no claims policies are observed, the number of claims is modelled using the $N_i \sim ZTNB(\lambda_i,\alpha)$ distribution. It allows us to take into account unobserved heterogeneity in the total claim amount estimation. In order to estimate the model parameters and fit the claim frequency, we use the numerical optimization in the MLE method taking the log-likelihood function as in the formula (19).

All three coefficients are statistically significant according to the Wald test, but only in the GLM Gamma. For the number of claims no factors have significant coefficients on a level of 0.05. Therefore, we estimate λ parameter to be the same for every policy. The regression coefficient estimators in GLM Gamma are presented in table 4.

Table 5 shows fitted values of the average value of claims for all combinations of regression coefficients.

We observe a relatively high variability in the fitted claims amount. The lowest value is given by the cars with low power and a female driver, who pays the premium without splitting the payment, while the highest value is generated by high-power cars and a male driver paying in instalments.

Table 4.

GLM Gamma parameters

Rating Factors	$\hat{\beta}_j$	Standard error
Intercept	8.65	0.13
POWER RANGE (0–66)	-0.52	0.15
POWER RANGE (67–124)	-0.40	0.13
GENDER (1)	0.27	0.10
PREMIUM_SPLIT (1)	0.27	0.10
Dispersion parameter	$\hat{\phi} = 1.37$	-

Source: own calculations.

Table 5.

The fitted average value of claims $\hat{Y}_i$ in groups

POWER RANGE.GENDER.PREMIUM SPLIT	Fitted value
125+.0.0	4582.11
66-.0.0	**3650.85**
67-124.0.0	3957.34
125+.1.0	5278.62
66-.1.0	4205.80
67-124.1.0	4558.88
125+.0.1	5132.54
66-.0.1	4089.42
67-124.0.1	4432.72
125+.1.1	**5912.72**
66-.1.1	4711.03
67-124.1.1	5106.52

Source: own calculations.

Afterwards we analyze the number of claims for a single risk. In order to take into account the unobserved factor, the distribution is assumed as $N_i \sim ZTNB(\lambda_i,\alpha)$. No factors have significant coefficients on a level of 0.1. Therefore, we estimate λ parameter, the same for every policy, receiving $\hat{\lambda} = 0.0003$, $\hat{\alpha} = 461.95$ with standard errors equal to 120.1 and 0.23 respectively. Thus, plugging this values into the $E_{ZTNB}[N_i]$ and multiplying by the exposure E_i the expected number of claims for a single risk is obtained. In the portfolio only 35 risks are not covered in the whole period ($E_i < 1$). Hence, most risks have $\hat{E}_{ZTNB}[N_i] = 1.08$ with $E_i = 1$.

Using the received estimated values of parameters we consider four type of copulas: Gaussian, Clayton, Gumbel and Frank. Maximizing the log-likelihood (9) we

choose the Gumbel copula with fitted $\hat{\theta} = 1.19$, which is equivalent to Kendall's $\tau = 0.16$. This type of the copula gives the smallest AIC value.

Finally, we construct the copula-based density of the total claim amount $f_S(\cdot)$ according to the formula (14) and using estimated parameters $\hat{\mu}_i$, $\hat{E}_{ZTNB}[N_i]$, $\hat{\theta}$. It gives us full information about this random variable and the possibility of estimating the expected value of the total claim amount. Figure 7 on the left-hand side provides the plots of values of the density for risks from the analyzed portfolio. For comparison, we also present the density plot based on the kernel estimation (cf. Sheather, Jones, 1991).

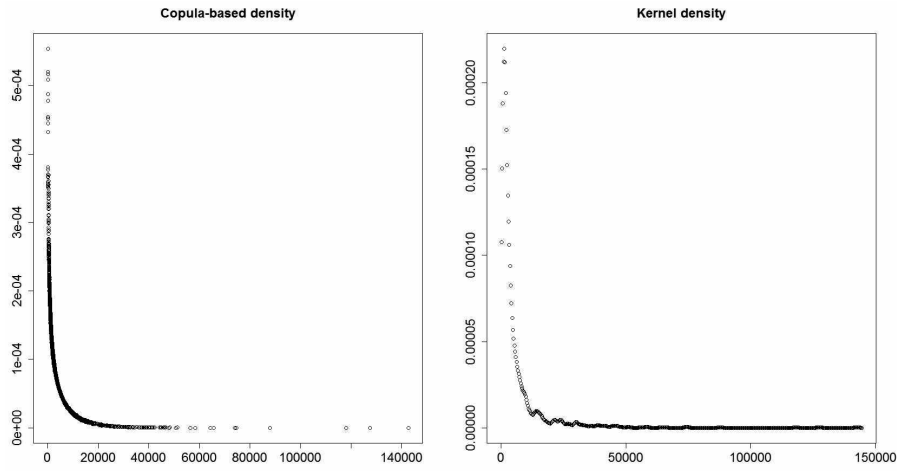

Figure 7. The density of total claim amount

Source: own calculations.

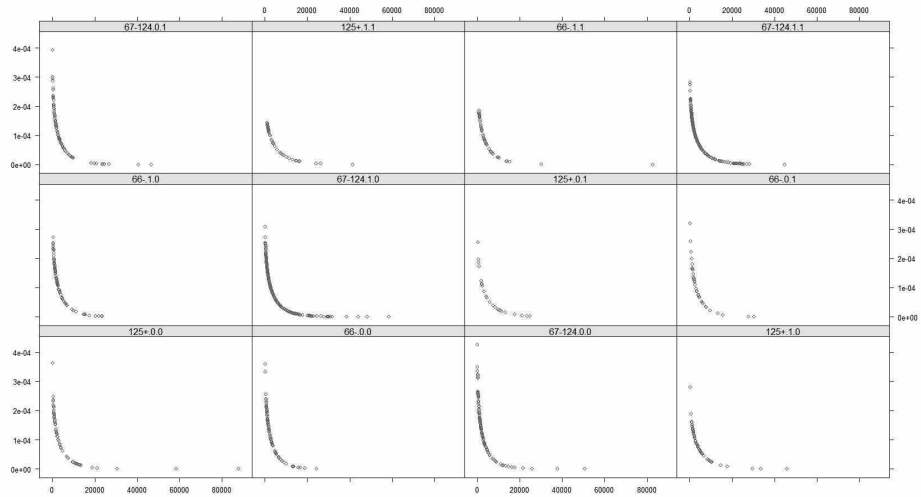

Figure 8. Copula-based density of total claim amount in groups

Source: own calculations.

In figure 8, we notice that the distributions in all groups are generally left-skewed. This is natural, as the margins of the average value of claims are Gamma distributed.

After that the copula-based expected total claim amount is determined using the MC simulation. This simulation provides values $\hat{E}[S_i]$ received via numerical integration. Figure 9 provides the summary.

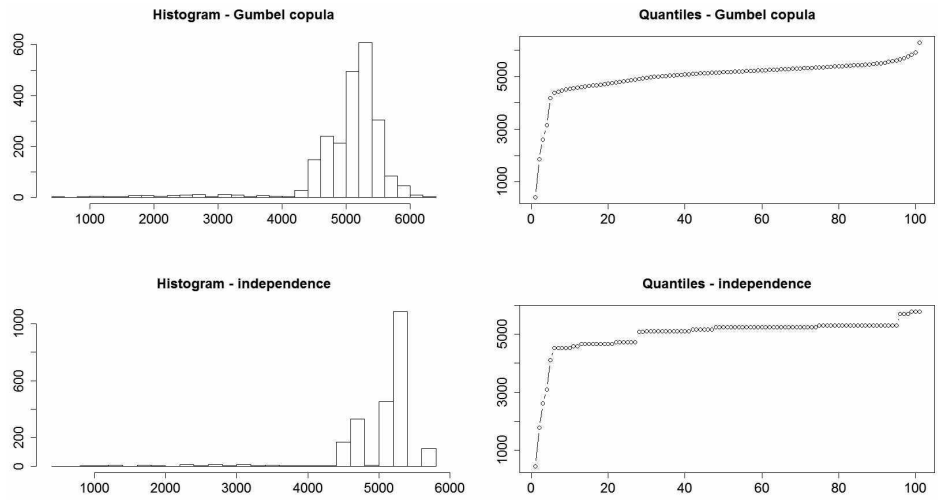

Figure 9. Expected total claim amount – MC integration

Source: own calculations.

The results show that values received in the copula-based model are slightly higher than the values in the model under the independence assumption. This fact is observed in the histograms as well as in the quantile plots. It can suggest that models commonly applied by insurance companies underestimate total claim amounts and hence pure premiums for a single risk. To visualize the variability of the expected total claim amount in groups according to the combinations of regressors taken in the Gamma GLM, the boxplot is displayed in figure 10.

It shows low variability in all groups appearing rather for risks with the low value of the claim amount. Except that the means (the black dots) are decisively higher for males with power 125+ than for females with any power, which is the intuitive result.

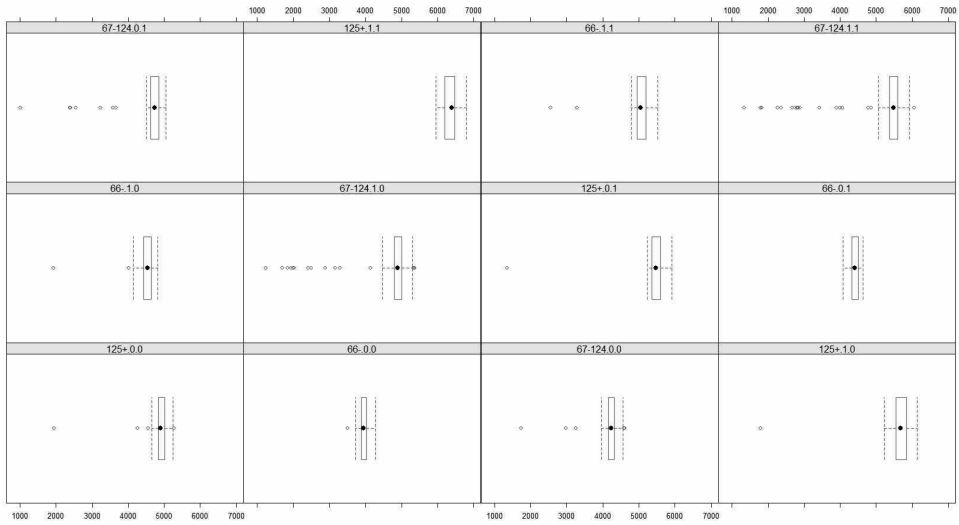

Figure 10. Copula-based expected total claim amount in groups
Source: own calculations.

6. CONCLUSIONS

In this paper, we model an average value of claims and the number of claims in the case of dependence between both random variables. The proposed model provides exact distributions of individual total claim amounts, which tend to be left-skewed. Moreover, we also show how to numerically construct the density of the bivariate random variable. This gives the possibility of estimating the expected total claim amounts in the portfolio using e.g. MC integration in pricing. As we use the ZTNB distribution, heterogeneity is taken into account. It corresponds to credibility representing the unobservable factor influencing the number of claims for a single risk. However, there are no obstacles to use another mixed Poisson model (cf. Karlis, 2001; Wolny-Dominiak, Trzęsiok, 2015). Nowadays the statistical modelling cannot do without computation, so the numerical examples discussed in this paper required strong programming work. Therefore, the full **R** code with a complete description is available for download.

REFERENCES

Antonio K., Valdez E. A., (2012), Statistical Concepts of A Priori and A Posteriori Risk Classification in Insurance, *Advances in Statistical Analysis*, 96 (2), 187–224.

Boucher J. P., Denuit M., Guillén M., (2007), Risk Classification for Claim Counts: A Comparative Analysis of Various Zeroinflated Mixed Poisson and Hurdle Models *North American Actuarial Journal*, 11 (4), 110–131.

Boucher J. P., Denuit M., Guillén M., (2009), Number of Accidents or Number of Claims? An Approach with Zero-Inflated Poisson Models for Panel Data, *Journal of Risk and Insurance*, 76 (4), 821–846.
Bühlmann H., Gisler A., (2005), *A Course in Credibility Theory and its Applications*, Springer.
Cizek P., Härdle W. K., Weron R., (2011), *Statistical Tools for Finance and Insurance*, Springer Science & Business Media.
De Jong P., Heller G. Z., (2008), *Generalized Linear Models for Insurance Data*, Cambridge University Press.
Denuit M., Maréchal X., Pitrebois S., Walhin J. F., (2007), *Actuarial Modelling of Claim Vounts: Risk Classification, Credibility and Bonus–malus Systems*, John Wiley & Sons, Chichester.
Dimakos X. K., Di Rattalma A. F., (2002), Bayesian Premium Rating with Latent Structure, *Scandinavian Actuarial Journal*, 2002 (3), 162–184.
Frees E. W., (2009), *Regression Modeling with Actuarial and Financial Applications*, Cambridge University Press.
Joe H., (1997), *Multivariate Models and Multivariate DSependence Concepts*, CRC Press.
Karlis D., (2001), A General EM Approach for Maximum Likelihood Estimation in Mixed Poisson Regression Models, *Statistical Modelling*, 1 (4), 305–318.
Krämer N., Brechmann E. C., Silvestrini D., Czado C., (2013), Total Loss Estimation using Copula-based Regression Models, *Insurance: Mathematics and Economics*, 53, 829–839.
Nelsen R. B., (1999), *An Introduction to Copulas*, Springer Science & Business Media.
Ohlsson E., Johansson B., (2010), *Non-life Insurance Pricing with Generalized Linear Models*, Springer.
R Core Team (2014), R: A *Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R–project.org/.
Sheather S. J., Jones M. C., (1991), A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation, *Journal of the Royal Statistical Society*, Series B (Methodological), 683–690.
Shi P., Feng X., Ivantsova A., (2015), Dependent Frequency–severity Modeling of Insurance Claims, *Insurance: Mathematics and Economics*, 64, 417–428.
Sklar M., (1959), *Fonctions de Répartition à n Dimensions et Leurs Marges*, Université de Paris 8.
Sun J., Frees E. W., Rosenberg M. A., (2008), Heavy-tailed Longitudinal Data Modeling Using Copulas, *Insurance: Mathematics and Economics*, 42 (2), 817–830.
Trzęsiok M., Wolny-Dominiak A., (2015), Złożony Mieszany Rozkład Poissona – Zastosowania Ubezpieczeniowe, *Studia Ekonomiczne*, 227, 85–95.
Wanat S., (2012), *Modele Zależności w Agregacji Ryzyka Ubezpieczyciela*. Zeszyty Naukowe, Uniwersytet Ekonomiczny w Krakowie, Seria Specjalna, Monografie.
Wolny-Dominiak A., (2014), *Taryfikacja w Ubezpieczeniach Majątkowych z Wykorzystaniem Modeli Mieszanych*, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.
Wolny-Dominiak A., Trzęsiok M., (2014), *insuranceData: A Collection of Insurance Datasets Useful in Risk Classification in Non-life Insurance*, R package version 1.0.
Wolny-Dominiak A., Trzpiot G., (2013), GLM and Quantile Regression Models in A Priori Ratemaking, *Insurance Review*, 4, 49–57.

REGRESYJNY MODEL ŁĄCZNEJ WARTOŚCI SZKÓD Z UWZGLĘDNIENIEM NIEOBSERWOWALNEGO CZYNNIKA RYZYKA

Streszczenie

W masowych portfelach ryzyk zakłady ubezpieczeń przeprowadzają tzw. taryfikację, której celem jest wyznaczenie składki czystej dla pojedynczego ryzyka. Modele statystyczne stosowane obecnie w praktyce należą najczęściej do klasy uogólnionych modeli liniowych (GLM), w których szacuje się w osobnych modelach wartości oczekiwane dwóch zmiennych losowych: średniej wartości szkody oraz

liczby szkód dla ryzyka. Składka czysta definiowana jest wtedy jako iloczyn uzyskanych wartości. Takie podejście wymaga założenia niezależności pomiędzy rozpatrywanymi dwoma zmiennymi losowymi. Jednak w literaturze to założenie jest podważane. Celem tego artykułu jest zaproponowanie modelu z kopulą uwzględniającego nieobserwowalny czynnik ryzyka w modelowaniu liczby szkód. Model ten służy do oszacować oczekiwanej wartości iloczynu dwóch zmiennych losowych: średniej wartości szkody oraz liczby szkód dla pojedynczego ryzyka przy założeniu zależności oraz występowaniu czynnika nieobserwowalnego. W pracy szczegółowo opisano aspekty teoretyczne związane z budową modelu oraz szacowaniem wartości oczekiwanej. Ponadto w licznych przykładach przedstawiono numeryczne rozwiązania obliczeniowe w programie **R**. Dodatkowo udostępniono kody programu **R** na stronie internetowej http://web.ue.katowice.pl/woali/.

Słowa kluczowe: taryfikacja, GLM, nieobserwowalny czynnik ryzyka, kopula

THE COPULA-BASED TOTAL CLAIM AMOUNT REGRESSION MODEL WITH AN UNOBSERVED RISK FACTOR

Abstract

Nowadays a common practice of any insurance company is ratemaking, which is defined as the process of classification of the mass risk portfolio into risk groups where the same premium corresponds to each risk. As generalised linear models are usually applied, the process requires the independence between the average value of claims and the number of claims. However, in literature this assumption is called into question. The interest of this paper is to propose the copula-based total claim amount model taking into account an unobservable risk factor in the claim frequency model. This factor, called also as unobserved heterogeneity, is treated as a random variable influencing the number of claims. The goal is to estimate the expected value of the product of two random variables: the average value of claims and the number of claims for a single risk assuming the dependence between the average value of claims and the number of claims for a single risk and the dependence between the number of claims for a single risk and the unobservable risk factor. We give details of the theoretical aspects of the model as well as the empirical example. To acquaint the reader with the model operation, every step of the process of the expected value estimation in described and the **R** code is available for download, see http://web.ue.katowice.pl/woali/.

Keywords: ratemaking, GLM, unobserved factor, copula