PREDICTION OF TOTAL CLAIM AMOUNT USING LONGITUDINAL DATA

Alicja Wolny-Dominiak[1], Stanisław Wanat[2]

[1] University of Economics in Katowice
Faculty of Economics, Department of Statistical and Mathematical Methods in Economics
1-go Maja 50, 40-283 Katowice, Poland
E-mail: alicja.wolny-dominiak@uekat.pl

[2] Cracow University of Economics
Department of Mathematics,
27 Rakowicka St. 31-510 Cracow, Poland
E-mail: wanats@uek.krakow.pl

Abstract: *The paper focuses on a priori ratemaking in casualty/property insurance. The objective is to predict the value of the random variable being interpreted as the total claim amount for single risk in mass insurance portfolio. The portfolio is fully described by the data in the longitudinal framework. We propose the model, in which two aspects are taken into account: factors influencing the total claim amount and the dependency between variables occurring in time, which is modelled by the copula. The Monte Carlo simulation is taken advantage to estimate the proper expected value. To illustrate the model in practice the portfolio of MOD risks from an insurance company operating on the Polish market is analyzed. Risks are insured every year in the period of 4 years. R packages are carried out in estimation of model parameters.*

Key words: *longitudinal data, copula, Tweedie compound Poisson regression*

JEL codes: *G17, C53, C33, C15*

1. Introduction

Modeling total claim amounts for individual risk in the insurance portfolio is a critical component in the ratemaking process for property and casualty insurance. For pure premium calculations, the common practice is to model total claims amount using Tweedie's compound Poisson regression as in (Jørgensen and Souza, 1994; Wolny-Dominiak, 2014a). The data is usually within a cross-sectional framework only. However, usually risks are observed over a few years and the insurer is able to better assess the policyholder's potential claims amount. As a consequence, there is a possibility to make appropriate premium adjustments according to risk characteristics.

In contrast to the cross-sectional data, in a longitudinal data framework, observations are made repeatedly over time. In our case, since every policy in the portfolio is renewable year after year the panel data is collected. Therefore the dependence between total claims amounts for individual risks over years should be assumed. To accommodate this dependence one can use mixed models instead of Tweedie's compound Poisson regression limited to independent observations (cf. Wolny-Dominiak, 2014b). The other approach is to use copulas as tools in multivariate joint distribution construction. There are some papers taking up this topic (cf. Joe, 1997; Wanat, 2012; Krämer et al., 2013; Shi et al., 2015; Wolny-Dominiak and Wanat, 2016). The object is usually to model the entire predictive distribution of the number of claims, not only point prediction, as in (Sun et al., 2008).

The interest of our application is the prediction of total claims amount of single risk for the year $T+1$ based on the history from years $1,...,T$, that is used in the ratemaking as the pure premium (cf. Frees et al., 2016). We propose copula-based model within a longitudinal framework as in (Leong and Valdez, 2005). In the first part of the paper basics of copulas are described. The second part contains the proposed model as well as the application to real data. In all calculation **R** software is used and the code is available on the webside http://web.ue.katowice.pl/woali/.

2. The Best Predictor

Let us consider the portfolio of n risks insured over the period of T years. Random variables Y_{it}, $i=1,...,n$, $t=1,...,T$ represent total claim amounts, where *i* denotes individual risk and *t* denotes the time period. In this part of the paper we focus on one particular *i*-th risk allowing total claim amounts dependence across time. Our interest is to predict the value $\hat{Y}_{T+1}$ for the year $T+1$. In that case the conditional expectation being the best predictor according to the mean squared prediction error (see (Leong and Valdez, 2005)) may be used:

$$E\left(Y_{T+1} \mid \mathbf{Y}_T = \mathbf{y}_T\right). \tag{1}$$

where $\mathbf{Y}_T = (Y_1,...,Y_T)'$ and $\mathbf{y}_T = (y_1,...,y_T)'$. In the paper (Leong and Valdez, 2005) authors proved that:

$$E\left(Y_{T+1} \mid \mathbf{Y}_T = \mathbf{y}_T\right) = \int_{-\infty}^{\infty} y_{i,T+1} \frac{c_{T+1}(\mathbf{u}_{T+1})}{c_T(\mathbf{u}_T)} dF_{T+1}(y_{T+1}), \tag{2}$$

where $c_{T+1}(\mathbf{u}_{T+1})$, $c_T(\mathbf{u}_T)$ are respectively the density of the copulas associated with the total claim amounts $\mathbf{Y}_{T+1}$ and $\mathbf{Y}_T$ and $F(Y_{T+1})$ is the cumulative distribution function of the total claim amount. If the dependence between variables $Y_1,...,Y_T$ is accommodated by Gaussian copula with correlation matrix $\mathbf{\Sigma}_T$, the search best predictor is given by the following formula:

$$E\left(Y_{T+1} \mid \mathbf{Y}_T = \mathbf{y}_T\right) = E_Z\{F_{T+1}^{-1}\left(\Phi(\mu_{T+1}^Z + \sigma_{T+1}^Z Z)\right)\}. \tag{3}$$

The expectation on the right-hand side is calculated for a standard normal random variable *Z*. Function F_{T+1}^{-1} is inverse to $F_{T+1}(Y_{T+1})$ with parameters:

$$\mu_{T+1}^Z = \boldsymbol{\rho}'_{T+1,T} \mathbf{\Sigma}_T^{-1} \boldsymbol{\varsigma}_T, \tag{4}$$

$$\sigma_{T+1}^Z = 1 - \boldsymbol{\rho}'_{T+1,T} \mathbf{\Sigma}_T^{-1} \boldsymbol{\rho}_{T+1,T}. \tag{5}$$

The mean and the standard deviation depend on the vector:

$$\boldsymbol{\varsigma}'_T = (\Phi^{-1}(u_1),...,\Phi^{-1}(u_T)), \tag{6}$$

which is constructed using values $u_1 = F_1(y_1),...,u_T = F_T(y_T)$ and Φ as a cdf of standardized normal distribution. The vector:

$$\boldsymbol{\rho}'_{T+1,T} = (\rho_{1,T+1},..., \rho_{T,T+1}) \tag{7}$$

corresponds to correlations of Y_{T+1} with each element of $(Y_1,...,Y_T)$.

3. The Estimation

The model has a following assumptions:

1. the dependence is captured by the *T*-dimensional parametric copula $C(\cdot;\Sigma)$,
2. marginal distributions comes from Tweedie's compound Poisson with a regression component $\mu_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_t$, where $\boldsymbol{\beta}_t = (\beta_{0t},...,\beta_{Kt})'$ is a vector in *K* regressor parameters, $\mathbf{x}'_{it}$ is *i*th row in design matrix **X** for *t*th year, constant dispersion parameter ϕ_t and power p_t.

Thus the vector of parameters is of the form:

$$(\beta_{01},...,\beta_{K1},...,\beta_{0T},...,\beta_{KT},\phi_1,...,\phi_T,p_1,...,p_T,\Sigma). \tag{8}$$

To estimate the unknown parameters a two-step procedure called the IFM (Inference Functions for Margins) method based on $i = 1,...,n$ observations $(y_{i1},...,y_{iT})$ is used. In the first step maximum likelihood (ML) estimates of marginal distributions are fitted and a pseudo sample is constructed as:

$$\hat{\mathbf{u}}_i = (\hat{u}_{i1},....\hat{u}_{iT}) = \left(F_{i1}(y_{i1};\hat{\mu}_{i1},\hat{\phi}_1,\hat{p}_1),...,F_{iT}(y_{iT};\hat{\mu}_{iT},\hat{\phi}_T,\hat{p}_T)\right),\ i = 1,...,n, \tag{9}$$

where $F_{i1},...,F_{iT}$ are marginal cumulative distribution functions (cdf). In the second step the matrix of the parameters of the copula is fitted by maximization of the log-likelihood function:

$$l(\Sigma;\hat{\mathbf{u}}_i) = \sum_{i=1}^{n}\log c(\hat{u}_{i1},...,\hat{u}_{iT};\Sigma) \tag{10}$$

Hence, the procedure of the estimation is summed up as

1. estimating the parameter vector of margins $(\mu_{it},\phi_t,p_t)'$, $t = 1,...,T$,
2. transforming $(y_{i1},...,y_{iT})$ to $(u_{i1},...,u_{iT})$ as in (9),
3. maximizing the log-likelihood (10).

Formula (3) is adopted herein to obtain the value of the predicted total claim amount by means of the Monte Carlo simulation. The procedure has the following steps:

1. Calculation of σ^Z_{T+1} according to (5).
2. For each risk:
   a. calculation of μ^Z_{T+1} according to (4),
   b. generation of *s* values of $z_1,...,z_s$ from the standardized normal distribution,
   c. calculation of $\psi_j = F^{-1}_{T+1}\left(\Phi(\mu^Z_{T+1} + \sigma^Z_{T+1}z_j)\right)$ for each generated value.

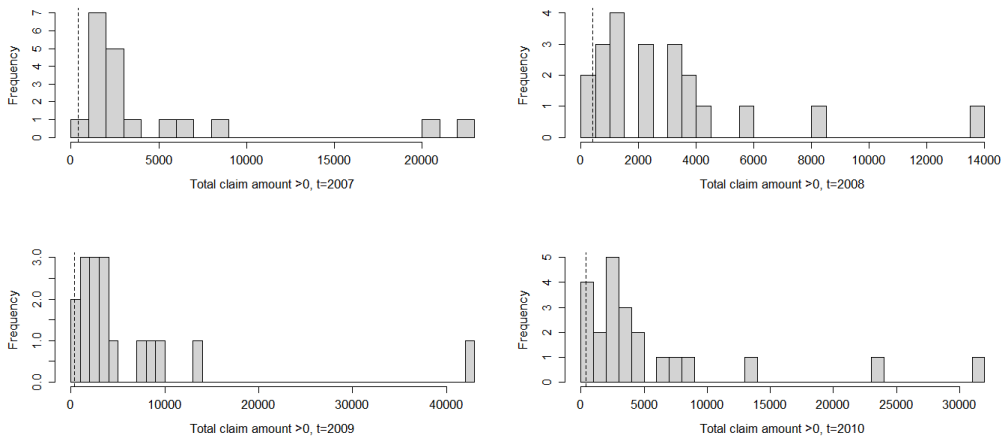
Finally, the predicted total claim amount is calculated according to

$$\hat{Y}_{T+1} = \frac{1}{s}\sum_{j=1}^{s}\psi_j\ . \tag{11}$$

4. The Application

The proposed model will be illustrated using a portfolio provided by a Polish insurance company. The data concern 245 car insurance risks taken out in the years 2007-2010. The distribution of the non-zero total claim amounts $Y_1,...,Y_4$ in subsequent years is shown in figure 1.

Fig. 1 Histograms of the non-zero total claim amount in subsequent years



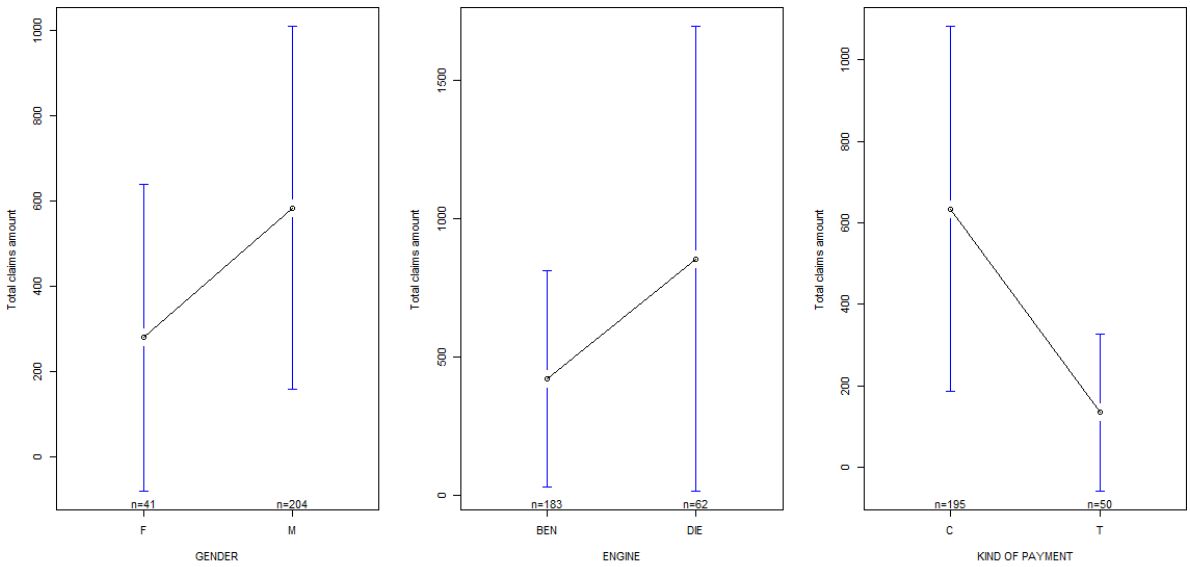
Source: Own calculations.

There are three categorical covariates:

1. GENDER (F – Female, M – Male),
2. ENGINE (BEN – petrol, DIE – diesel),
3. KIND OF PAYMENT (C – cash, T – transfer)

affecting the total claim amount and exposure to risk.
The plots give a general idea of the effect the covariates have on the response. Figure 2 visualizes the means of the total claim amount and the confidence interval in all categories of the rating factors in 2010.

Fig. 2 Means of the total claim amount of rating factors in 2010 by categories



Source: Own calculations.

The plots show that the means of the categories of each covariate under consideration are included in a fairly wide range between the highest and the lowest value. This leads to the presumption that all the factors differentiate the value of the total claim amount. For this reason, the covariates mentioned above are introduced into the GLM.

The aim of this case study is to predict the total claim amount for a single risk in 2011 taking account of correlated observations from the years 2007-2010. The copula-based model is described by the following assumptions:

1. Marginal models – the Tweedie compound Poisson regression, where responses are $Y_{2007},...,Y_{2010}$ with parameter vectors $(\mu_{it} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta}), \phi_t, p_t)'$, $t = 2007,...,2010$,
2. The dependence accommodates a 4-dimentional Gaussian copula for which the correlation structure for $\boldsymbol{\Sigma}$ is autoregressive of order 1 (AR1).

The AR1 structure is selected due to the fact that the analysis concerns relations in the time dimension (between claims for individual risks in the years 2007-2010) and the need to create a possibility of forecasting, in a simple manner, the vector (7), which is indispensable for the practical use of the presented model.

This means that the parameter vector is of the form $(\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_4,\phi_1,...,\phi_4, p_1,..., p_4,\rho)'$, where ρ corresponds to the parameter matrix

$$\Sigma_{AR1} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

The parameter estimators are established by implementing the IFM method. First, the margins are fitted and the copula parameters are determined. Function cpglm{cplm} is used to give the Tweedie compound Poisson regression. The results are presented in table 1.

Tab. 1 Margin parameters

Parameter	Year 2007	Year 2008	Year 2009	Year 2010
$\hat{p}_t$	1,498	1,397	1,503	1,486
$\hat{\phi}_t$	474,79	523,07	531,67	492,45
Intercept	6,54 (0,75)	6,06 (0,63)	5,57 (0,84)	5,61 (0,97)
ENGINE,DIE	-0,06 (0,87)	-0,16 (0,69)	1,63 (0,70)	0,75 (0,71)
GENDER,F	-0,87 (0,83)	-0,70 (0,68)	-0,36 (0,89)	0,70 (1,01)
KIND OF PAYMENT, T	0,61 (1,20)	0,64 (0,80)	0,66 (0,86)	-1,66 (1,08)

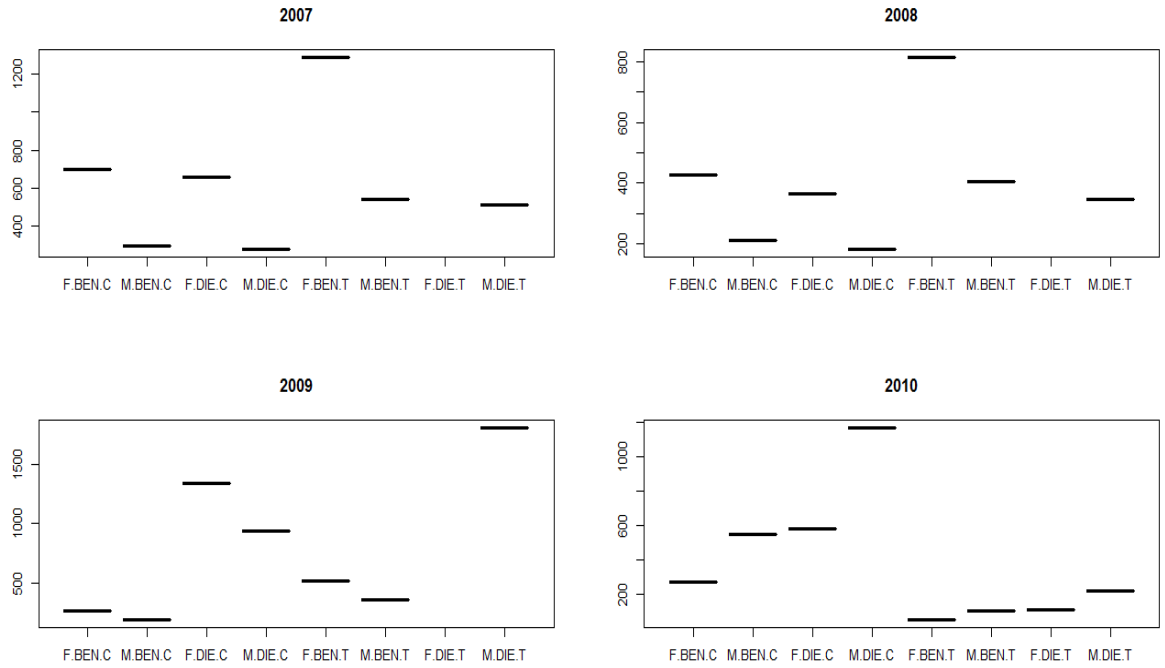
Source: Own calculations.

The estimated parameters enabled determination of mean values of the total claim amounts for homogeneous groups of policies, i.e. risks with the same values of regressors.

1. F.BEN.C – (Female, Petrol, Cash),
2. M.BEN.C - (Male, Petrol, Cash),
3. F.DIE.C - (Female, Diesel, Cash),
4. M.DIE.C - (Male, Diesel, Cash),
5. F.BEN.T - (Female, Petrol, Transfer),
6. M.BEN.T – (Male, Petrol, Transfer),
7. F.DIE.T - (Female, Diesel, Transfer),
8. M.DIE.T - (Male, Diesel, Transfer).

Figure 3 illustrates estimated total claim amounts $\hat{Y}_{2007},...,\hat{Y}_{2010}$ according to the regressors.

Fig. 3 Fitted values of the total claim amounts depending on the risk group



Source: Own calculations.

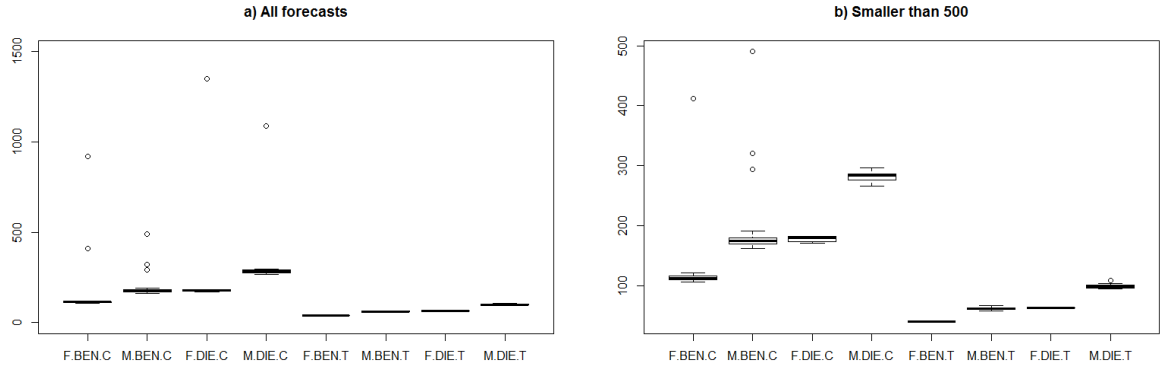
It indicates that in 2007 and 2008 the highest total claim amount values occurred in the F.BEN.T group (i.e. in the group of petrol engine cars owned by women, for which the premium was paid by a bank transfer), whereas the lowest – for the M.DIE.C group (a diesel engine car owned by a man, premium paid in cash). In 2009, the highest and the lowest mean total claim amounts were recorded for the M.DIE.T and the M.BEN.C group, respectively; in 2010, the highest mean total claim amounts were in the M.DIE.C group, whereas the F.BEN.T group had the lowest. It can also be stated that in the risk groups under analysis the mean total claim amounts were at a similar level in the years 2007 and 2008. In 2009 and 2010 the situation was substantially different (compared both to the two previous years and to each other).

Next, the 4-dimentional copula is fitted. In order to estimate the copula parameters, fitCopula{copula} is used assuming the Gaussian family with AR1 correlation structure. The estimated parameter is $\hat{\rho} = 0{,}947$. Finally, the predictor of the total claim amount for a single risk in 2011 is obtained using (3) and assuming that the margins in 2011 and 2010 are the same and the parameter vector is of the form $\boldsymbol{\rho}_{T+1,T} = (\rho^4, \rho^3, \rho^2, \rho)$.

Figures 4-5 present the results obtained depending on the risk groups under consideration. Figure 4 presents the distribution of forecasts for the total claim amounts by means of box plots (4a – all forecasts, 4b – smaller than 500) and figure 5 – density functions[1] (obtained using kernel estimators) of the distribution of forecasts for individual groups (5a – females, 5b – males).

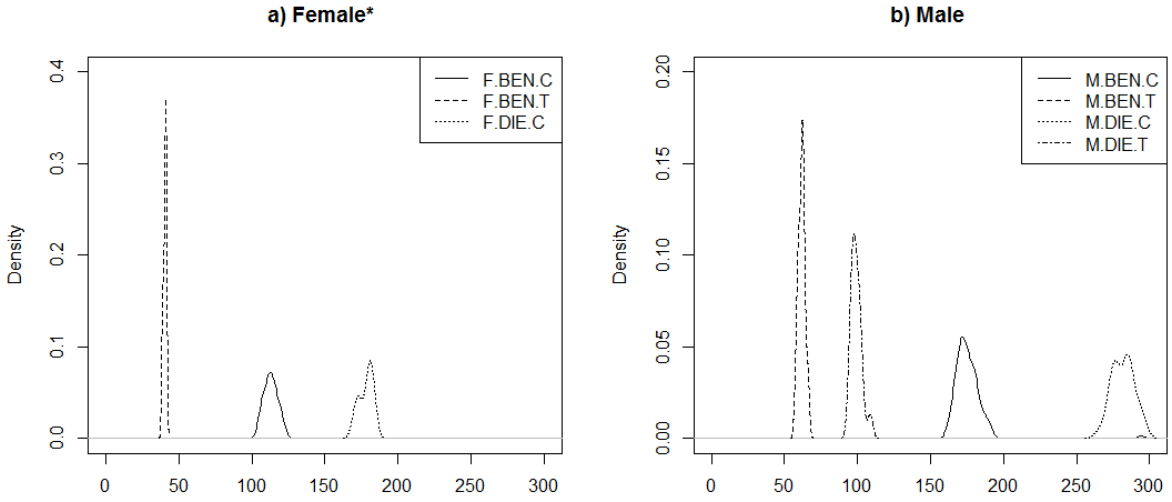
[1] To be precise, the density function values are given for arguments smaller than 300 to ensure better clarity of the figure.

Fig. 4 Predictors for the total claim amounts in individual risk groups



Source: Own calculations.

Fig. 5 Density of the predicted total claim amounts in 2011 for the risk groups under consideration

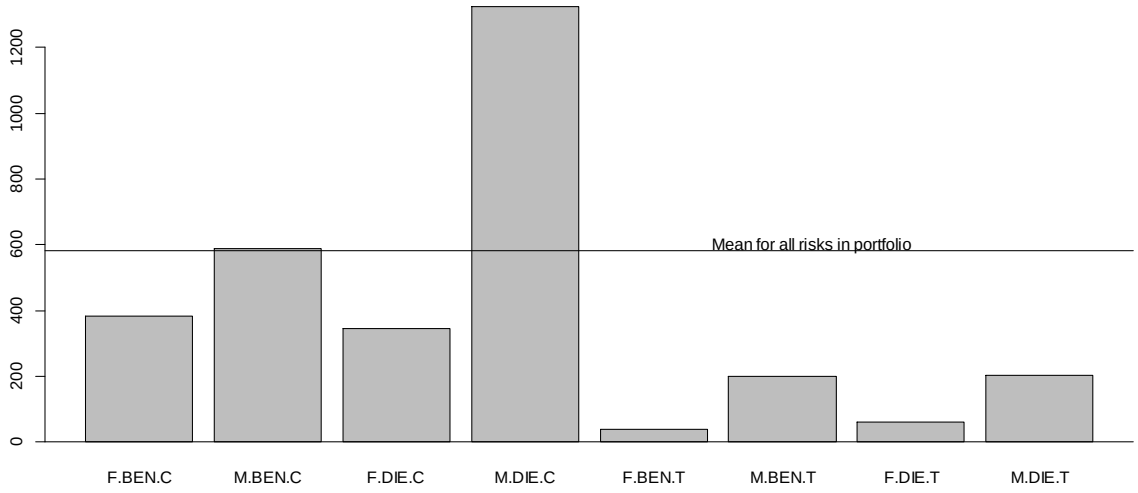


*) F.DIE.T group contains only one risk

Source: Own calculations.

From the practical point of view, the mean values of predictors presented in figure 6 can be treated as pure premiums for risks belonging to a given risk group which are determined taking account of the individual risk of the policies, the risk portfolio, as well as the relations between the risk portfolio in individual years. The distribution density functions (cf. figure 5), on the other hand, can be used to assess the level of uncertainty of the premium calculation by means of this method for policies from a given risk group.

Fig. 6 Pure premiums in individual risk groups



Source: Own calculations.

Generally, high dispersion of the values of future claims can be observed. The figures show that policyholders owning diesel engine cars and paying in cash generate a much higher value of future claims compared to those driving petrol engine cars and using bank transfers. Considering practical purposes of prediction, the predictors of mean total claim values for individual risk groups (figure 6) are of special importance.

5. Conclusions

The article considered the problem of prediction of the total claims amount in the next period based on individual claims history. This is of course a fundamental problem of credibility theory. In our paper, differently from the classical credibility models (where it is assumed conditional independence or independence), the relationship between claims occurring in subsequent periods is allowed. The structure of this relationship is captured with Gaussian copula. We extend the model offered in (Leong and Valdez, 2005) by proposing the use of Tweedie's compound Poisson distribution with a regression component in margins modelling. The proposed approach has been illustrated on real data from the Polish insurance market.

Acknowledgments

Publication was financed from the funds granted to the Faculty of Finance and Law at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

References

Frees E.W., Derrig R.A., Meyers G. (2016): *Predictive modeling applications in actuarial science*. *Volume I: Predictive Modeling Techniques.* Cambridge University Press, New York.

Joe H. (1997): *Multivariate models and multivariate dependence concepts*. CRC Press, Boca Raton.

Jørgensen B., Souza M.C. (1994): *Fitting Tweedie's compound poisson model to insurance claims data*. "Scandinavian Actuarial Journal", Vol. 1994(1), pp. 69-93. doi:10.1080/03461238.1994.10413930

Krämer N., Brechmann E.C., Silvestrini D., Czado C. (2013): *Total loss estimation using copula-based regression models*. "Insurance: Mathematics and Economics", Vol. 53(3), pp. 829-839. doi:10.1016/j.insmatheco.2013.09.003

Leong Y., Valdez E.A. (2005): *Claims Prediction with Dependence using Copula Models*. Retrieved May 4, 2016, from https://www.bing.com/cr

Shi P., Feng X., Ivantsova A. (2015): *Dependent frequency–severity modeling of insurance claims*. "Insurance: Mathematics and Economics", Vol. 64, pp. 417-428. doi:10.1016/j.insmatheco.2015.07.006

Sun J., Frees E.W., Rosenberg, M.A. (2008): *Heavy-tailed longitudinal data modeling using copulas*. "Insurance: Mathematics and Economics", Vol. 42(2), pp. 817-830. doi:10.1016/j.insmatheco.2007.09.009

Wanat S. (2012): *Modele zależności w agregacji ryzyka ubezpieczyciela (Dependence models in the agregating of insurer risks)*. Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Wolny-Dominiak A. (2014a): *Jednomodelowa taryfikacja a priori w krótkoterminowych ubezpieczeniach majątkowych*. „Ekonometria", Vol. 4(46), pp. 34-42. doi:10.15611/ekt.2014.4.03

Wolny-Dominiak A. (2014b): *Taryfikacja w ubezpieczeniach majątkowych z wykorzystaniem modeli mieszanych*. Wydawnictwo UE Katowice, Katowice.

Wolny-Dominiak A., Wanat, S. (2016): *Taryfikacja a priori z wykorzystaniem kopuli (On the use of copula in ratemaking)*. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 415, Wydawnictwo UE Wrocław, Wrocław, pp. 258-265. doi:10.15611/pn.2016.415.24