Associate Professor Mihaela COVRIG, PhD
E-mail: mihaela.covrig@csie.ase.ro
The Bucharest University of Economic Studies
Professor Dumitru BADEA, PhD
The Bucharest University of Economic Studies

SOME GENERALIZED LINEAR MODELS FOR THE ESTIMATION OF THE MEAN FREQUENCY OF CLAIMS IN MOTOR INSURANCE

Abstract. *One of the most important problems that a non-life actuary faces is constructing a fair pricing. In particular, claim counts modeling is one of the components of motor insurance ratemaking. This paper aims to describe the econometric modeling of the mean frequency of claims in a motor insurance portfolio using generalized linear models. The main frequency distributions of count data are presented together with the generalized linear models. Numerical illustration presents and compares the different proposed regression models, using annual CASCO insurance data from a Romanian insurance company. The main findings are that the Negative Binomial regression model performs better than the Poisson model and quantifies overdispersion. The figures, the estimations and the tests are done in the open source soft* R.

Keywords: *claim counts modeling, count data, motor insurance, GLM,*R.

JEL Classification: C510, C520, C890, G220.

1. Introduction

One of the most important problems that a non-life actuary faces is constructing a fair pricing. In particular, claim counts modeling is one of the components of motor insurance ratemaking.

Both in non-life insurance, and in life insurance, one comes across random variables that model situations such as the number of damage claims per motor insurance policy, or the number of insured people in a life insurance policy portfolio, and the possible applications are not limited to these situations. Estimating the mean frequency of claims in a motor insurance portfolio is one of the crucial components of its pricing, along with estimating the average damage severity.

This paper aims to describe the econometric modeling of the mean frequency of claims in a motor insurance portfolio using generalized linear models. The first to be presented are the main frequency distributions of count data, followed by the generalized linear models. We assumed that the explained variable is Poisson distributed, followed by mixed Poisson distributions, which reformulates the restrictive assumption that the mean is equal to the dispersion, allowing the modeling of the overdispersion phenomenon.

Numerical illustration presented and compared the different proposed regression models using annual CASCO insurance data from a Romanian insurance company. The main findings are that the Negative Binomial regression model performs better than the Poisson model and quantifies overdispersion. The paper ends with conclusions that point out the possibility of new directions of research or approach in the estimation of annual number of claims reported to the insurer.

The graphic representations, the estimates of the various regression models, as well as the tests of the different statistic hypotheses were done in `R`, using, among others, `glm` routines, and packages such as `COUNT`, `msme`, `MASS`.

2. Distributions of counting random variables and generalized linear models

The number of occurrences of a particular event on a specified time horizon or space can be modeled by a random variable, and let us denote it *Y*. The first choice for the distribution of *Y* is the Poisson distribution, *Y~Poisson*(λ), with the probability mass function:

$$P(Y=k;\lambda)= f(k;\lambda)=\frac{\lambda^k}{k!}\cdot e^{-\lambda},\ k\in\mathbf{N}, \lambda>0. \qquad (1)$$

This distribution is characterized by the equidispersion property: the expected value and variance are equal, that is $E[Y]=Var[Y]$, and $E[Y]=\lambda$. The maximum likelihood estimator of the parameter λ is the sample mean. It is an unbiased, consistent and efficient estimator.

In practice, particularly in the field of non-life insurance, and in the case of those variables which indicate the number of damage claims, one can notice that the variance is bigger than the mean, a phenomenon called overdispersion, leading to the identification of other distributions capable to model or capture this aspect. A possible solution consists in resorting to mixed Poisson distributions.

Thus, in line with the theories of De Jong and Heller (2008) and of Denuit, Marechal, Pitrebois and Walhin (2007), parameter λ of the Poisson distribution is considered to be a realization of a positive continuous random variable Λ, *Y~Poisson*(Λ), so that the distribution of the variable *Y*, given Λ=*λ*, is $P(Y=k|\Lambda=\lambda)=e^{-\lambda}\cdot\frac{\lambda^k}{k!},\ k\in\mathbf{N}$ and $Var[Y_{\text{mixed Poisson}}]>Var[Y_{\text{Poisson}}]=E[Y_{\text{Poisson}}]$.

If the random variable Λ follows a *Gamma* distribution, with the parameterization $\Lambda\sim Gamma\left(\frac{1}{\delta},\delta\mu\right)$, and the probability density function

$$g_\Lambda(\lambda)=\frac{1}{\Gamma\left(\frac{1}{\delta}\right)\cdot(\delta\mu)^{\frac{1}{\delta}}}\cdot\lambda^{\frac{1}{\delta}-1}\cdot e^{-\frac{1}{\delta\mu}\cdot\lambda},\ \lambda>0,\delta>0,\mu>0, \qquad (2)$$

with $E[\Lambda]=\mu$ and $Var[\Lambda]=\delta\mu^2$, then the probability mass function of the mixed distribution of *Y* is given, following De Jong and Heller (2008), by

$$\begin{aligned}P(Y=k)&=\int_0^\infty P(Y=k|\Lambda=\lambda)\cdot g_\Lambda(\lambda)d\lambda\\ &=\int_0^\infty e^{-\lambda}\cdot\frac{\lambda^k}{k!}\cdot\frac{1}{\Gamma\left(\frac{1}{\delta}\right)\cdot(\delta\mu)^{\frac{1}{\delta}}}\cdot\lambda^{\frac{1}{\delta}-1}\cdot e^{-\frac{1}{\delta\mu}\cdot\lambda}d\lambda\\ &=\frac{\Gamma\left(k+\frac{1}{\delta}\right)}{k!\cdot\Gamma\left(\frac{1}{\delta}\right)}\cdot\left(\frac{1}{1+\delta\mu}\right)^{\frac{1}{\delta}}\cdot\left(1-\frac{1}{1+\delta\mu}\right)^k,\ k\in\mathbf{N}\\ &=\frac{\Gamma(k+r)}{k!\cdot\Gamma(r)}\cdot p^r\cdot(1-p)^k,\ k\in\mathbf{N},\end{aligned}$$

that is $Y\sim Negativ\ Binomial(r=\frac{1}{\delta},p=\frac{1}{1+\delta\mu})$, $r>0,p\in(0,1)$ and $E[Y]=\mu$, $Var[Y]=\mu(1+\delta\mu)=\mu+\delta\mu^2>\mu$.

Next we shall use the parameterization *Y ~ Negativ Binomial* (*μ*, *δ*).

One could take into account other mixed Poisson distributions: if one presupposes that Λ follows an *Inverse Gaussian* distribution, the result will be a *Poisson-Inverse Gaussian* distribution, or, if one presupposes that Λ follows a *Log normal* distribution, the result will be a *Poisson-Log normal* distribution. However, these have complex expressions of the probability mass function, or they are defined by recurrence relationships (Denuit, Marechal, Pitrebois and Walhin (2007)). Recent developments of statistical software make possible the estimation of the parameters of such mixed distributions.

Generalized linear models (GLMs) expand the classical regression model due to the fact that the theoretical distribution of the dependent variable is not necessarily normal, but it belongs to a special class of distributions, while the relationship between the mean of the dependent variable and a linear function of explanatory variables is given by a link function. GLMs were introduced in the seminal paper of Nelder and Wedderburn (1972).

A generalized linear model (GLM) consists of three elements, as explained by McCullagh and Nelder (1989), Denuit and Charpentier (2005), and Frees, Derrig and Meyers (2014).

The first element is represented by the random component of the model, that is, by the independent random variables $Y_1, Y_2, ..., Y_n$, their distribution is in the exponential family, θ_i are the canonical parameters and the common parameter ϕ is the scale or dispersion parameter:

$$f(y_i;\theta_i,\phi)=e^{\frac{y_i\theta_i-b(\theta_i)}{\phi}-c(y_i,\phi)}, y_i \in S, \qquad (3)$$

where $b(\cdot)$ and $c(\cdot,\cdot)$ are known functions, the support set $S \subseteq \mathbf{N}$ or **R**, while $\theta_i \in \Theta$, an open set in **R**. The expected value μ_i and the variance of Y_i are $\mu_i = M[Y_i] = b'(\theta_i)$, $Var[Y_i] = \phi \cdot \mu'_i(\theta_i)$, so the dispersion varies with the mean.

The second element is represented by the systematic component of the model, built from *p*+1 parameters $\beta = (\beta_0, \beta_1, ..., \beta_p)^t$ and with *p* explanatory variables. The linear predictor is

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, i = 1,2,...,n. \qquad (4)$$

The design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)^t$, of size $n \times (p+1)$, has full rank $p+1 < n$, so that the square matrix $\boldsymbol{X^t} \cdot \boldsymbol{X}$ is non-singular.

The third element is represented by a link function *g* between the random and the systematic components, that is monotone and differentiable, so that

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \qquad (5)$$

The estimation of the parameters $\beta_0, \beta_1, ..., \beta_p$ is done by the maximum likelihood method. That leads to solving a non-linear system of equations that can only be solved through numerical methods, such as Newton-Raphson.

The Poisson distribution and the Negative Binomial distribution belong to the exponential family of distributions. The Poisson regression model is a GLM with the response variable $Y \sim Poisson(\mu)$, $\mu > 0$, $f(y;\mu) = P(Y = y;\mu) = e^{-\mu} \cdot \frac{\mu^y}{y!}$, $y \in \mathbf{N}$, in the case of which the relationship between the canonic parameter θ and the parameter μ of the Poisson distribution is given by $\theta = \ln \mu$, the dispersion parameter $\phi = 1$ and $E[Y] = \mu = Var[Y]$. If the selection variables $Y_i \sim Poisson(\mu_i)$, *i*=1,2,…,*n*, then the linear predictor is given by the relationship (4), while the link function is one of log type, $\eta_i = g(\mu_i) = \ln \mu_i$, and thus the mean frequency will be estimated or adjusted by the model

$$\mu_i = e^{\beta_0} \cdot e^{\beta_1 x_{i1}} \cdot e^{\beta_2 x_{i2}} \cdot \ldots \cdot e^{\beta_p x_{ip}}, \qquad (6)$$

also referred to as the multiplicative model.

The Negative Binomial regression model is a GLM with the response variable *Y~ Negativ Binomial* (μ, δ), μ>0, δ>0,

$f(y;\mu,\delta) = P(Y = y;\mu,\delta) = \dfrac{\Gamma\left(y+\dfrac{1}{\delta}\right)}{y!\cdot\Gamma\left(\dfrac{1}{\delta}\right)} \cdot \left(\dfrac{1}{1+\delta\mu}\right)^{\frac{1}{\delta}} \cdot \left(1-\dfrac{1}{1+\delta\mu}\right)^{y}, y \in \mathbf{N}$, for which $\theta = \ln\dfrac{\delta\mu}{1+\delta\mu}$, $\phi = 1$, $E[Y] = \mu < \mu + \delta\mu^2 = Var[Y]$. If the selection variables $Y_i \sim Binomial\ Negativ\left(\mu_i, \delta\right)$, i=1,2,…,n, then the linear predictor is given by the relationship (4) in the definition of the GLMs, while the link function is conveniently chosen as one of the log type, $\eta_i = g(\mu_i) = \ln \mu_i$, and thus the mean frequency will be fitted also by a multiplicative model such as (6), that allows for an easy interpretation of the coefficients.

Although the two regression models have the same form, the estimators of the parameters do not coincide, as they represent solutions of different systems of equations that are built with the aid of the likelihood functions. Moreover, for the Negative Binomial regression model, an additional parameter, δ, is estimated. The introduction of this parameter ensures a bigger flexibility in the adjustment of data by the Negative Binomial regression as compared to the Poisson model, helping in the modeling of the overdispersion phenomenon.

The dispersion parameter ϕ of a generalized linear model is estimated by

$$\hat{\phi} = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_{i,\ Pearson}^{2}}{n-p-1}, \qquad (7)$$

where $\hat{\varepsilon}_{i,\ Pearson} = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$ and $\hat{\varepsilon}_{i,\ Pearson} = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + \hat{\delta}\cdot\hat{\mu}_i^2}}$ are the Pearson residuals of a Poisson regression model and a Negative Binomial regression model, respectively.

Practically speaking, in order to estimate the Poisson or Negative Binomial regression models, De Jong and Heller (2008) recommend that the explanatory variables included in a GLM should be dummy covariates. These are obtained either from non-numeric variables, or from numeric variables. Thus, each response category of a non-numeric variable or each class of variation for a numeric variable is assigned a dichotomous variable. Moreover, for each variable, we retain the category with the biggest exposure in the sample, and the corresponding dummy variables are not included in the GLM as explanatory variables, but they define the baseline profile of an elementary unit in the sample. The estimate of the mean frequency for the baseline profile is achieved with the aid of the intercept parameter.

Next we present a review of the main results in the specialized literature of the applications in the actuarial area of the quantitative methods, of the generalized linear models and, particularly, of the regression models used to estimate the mean frequency of the occurrence of a specific event.

The characteristic of the count data coming from the motor insurance area is the excess of zeros. This aspect may be explained by the reserve of policy owners to report a claim that they consider minor, or insignificant, for fear their damage claim history might influence their premiums in the future. Despite this, such a situation leads to the fact that data do not exhibit equidispersion, specific to Poisson distributions, but overdispersion. Mullahy (1986) introduced the zero-inflated models and since then a lot of studies have focused on this issue and have highlighted the importance of zero inflation in actuarial studies, for instance: Yip and Yau (2005), Boucher, Denuit and Guillen (2007), Perumean-Chaney, Morgan et al. (2013), Wolny-Dominiak (2013), Sarul and Sahin (2015), and also, in other fields, such as psychology, we mention Coxe, West and Aiken (2009).

In Covrig, Mircea et al. (2015), detailed explanations on estimating or fitting GLMs using R are provided. Quantitative methods applied in actuarial science are developed in Tănăsescu and Mircea (2014), where the authors analyze the difficult topic of ruin probability for an insurance company from a quantitative method perspective.

Generalizations of the Negative Binomial regression model are given in Vangala, Lord and Geedipally (2015) or Shirazi, Lordet et al. (2016). More advanced and recent research has focused on longitudinal and panel data, such as Boucher and Inoussa (2014), or on some complex mixed models, such as Poisson Inverse Gaussian GLMs or Bayesian models that allow for the modeling of highly overdispersed data: Zha, Lord and Zou (2016), Klein, Denuit et al. (2014), or Gómez-Déniz, Ghitany and Gupta (2016).

3. Using GLMs on count data of a Romanian motor insurance portfolio

We illustrate the regression models presented in the previous section through the estimation of the mean frequency of claims per motor insurance policy using a sample of CASCO policies provided by a Romanian insurance company. The motor insurance policies included in the sample cannot produce other events as their validity period is closed and they are prior to 2014.

The sample represents a CASCO motor insurance policy selection, valid for 1 year, in the portfolio of a Romanian insurance company. In addition to the variable which indicates the number of claims, we took into account other variables, some of which characterize the policy owner - such as age, in years, or the number of renewals - while others characterize the insured automobile, such as: the brand;

cylindrical capacity, expressed in cm^3; the engine power, expressed in kw; the vehicle age, in years; the insured amount, in EUR, or the county where the insurance policy was issued, most commonly corresponding to the area where the vehicle is used.

Table 1 includes the initial variables, except the automobile brand, their transformation in categorical variables, with the corresponding coding, as well as the reference categories, marked in bold letters. We name reference category the category with the highest exposure in the sample. Romanian development regions, numbered from 1 to 8, are defined at http://www.mdrap.ro/dezvoltare-regionala/-2257/programul-operational-regional-2007-2013/-2975. The most frequently CASCO insured automobile brand in the sample is Volkswagen.

The baseline profile of a CASCO insurance policy in this sample was the following: a legal entity (LE) purchased the insured vehicle and the policy was issued for the first time, the vehicle had a cylinder capacity under 1400 cm^3 and an age between 3 and 6, the insured amount is in the interval 5001 and 10000 EUR, the brand is Volkswagen, and the automobile is driven in the region Bucharest-Ilfov.

Table 1.Characteristics of a CASCO insurance policy in the sample

Initial variables	Categorical variables	
Owner’s age, in years	Cat_age	
	1	18-30
	2	31-40
	3	41-55
	4	over 56
	5	**LE**
Cylinder capacity, in cm^3	Cat_cil_cap	
	1	**under 1400 cm^3**
	2	1401-1800 cm^3
	3	1801-2200 cm^3
	4	2201-3000 cm^3
	5	over 3001 cm^3
Automobile age, in years	Cat_auto_age	
	1	0-2
	2	**3-6**
	3	7-10
	4	over 11
Insured amount, in Eur	Cat_insured_amount	
	1	under 5000 Eur
	2	**5001-10000 Eur**
	3	10001-30000 Eur
	4	over 30001 Eur

Romanian development region	Region	
	1	North-East
	2	South-East
	3	South Muntenia
	4	South-West Oltenia
	5	West
	6	North-West
	7	Center
	8	**Bucharest-Ilfov**
Number of policy renewals	Cat_nb_renew	
	0	**0**
	1	1
	2	2
	3	3
	4	4, 5, 6, 7

The sample relative frequency distribution of the variable that indicates the number of claims per policy is presented in Table 2.

Table 2. The sample relative frequency distribution of the number of claims per policy

Number of claims per policy	Proportion
0	0.7525
1	0.1443
2	0.0604
3	0.0368
4	0.0047
5	0.0010
6	0.0004
Total	1

The sample mean frequency is calculated as a weighted mean of the number of claims per policy, $\hat{\mu} = \sum_{k=0}^{6} k \times p_k$, where p_k is the proportion of policies with k claims in the sample. The obtained result is $\hat{\mu} = 0.4015$, that is an annual observed mean frequency of 40.15% per policy.

We considered it suggestive to represent the dependence of the observed mean frequency of claims on each possible explanatory variable for which there is available information, with a purpose of introducing them in the regression models, as can be seen in Figure 1. Thus Figure 1 (a) shows that the data in the sample confirms the fact that the mean frequency of claims is inversely proportional with the owner's age, decreasing from 57% in the case of the young drivers, to approximately 30% in the case of the elder. A possible explanation is the fact that aging is associated with more experience, as well as with a decreasing usage of the automobile. The cylindrical capacity influences the mean frequency of claims directly, the evolution being a spectacular one, as we can see in Figure 1 (b). As regards the automobile age, in the case of new vehicles, the observed mean frequency of claims is very close to the sample mean, and it is followed by a descending trend, from a value above average, of 47%, to approximately 29%, for the category of the oldest cars, Figure 1 (c). Figure1 (f) emphasizes the decreasing tendency of the mean number of claims with the number of renewals, so that, if, in the case of a policy with under 3 renewals the mean frequency varies a little around the average value of 40.15%, starting with 4 renewals the average frequency drops under the average value, to a minimum of 5% in the case of 7 renewals.

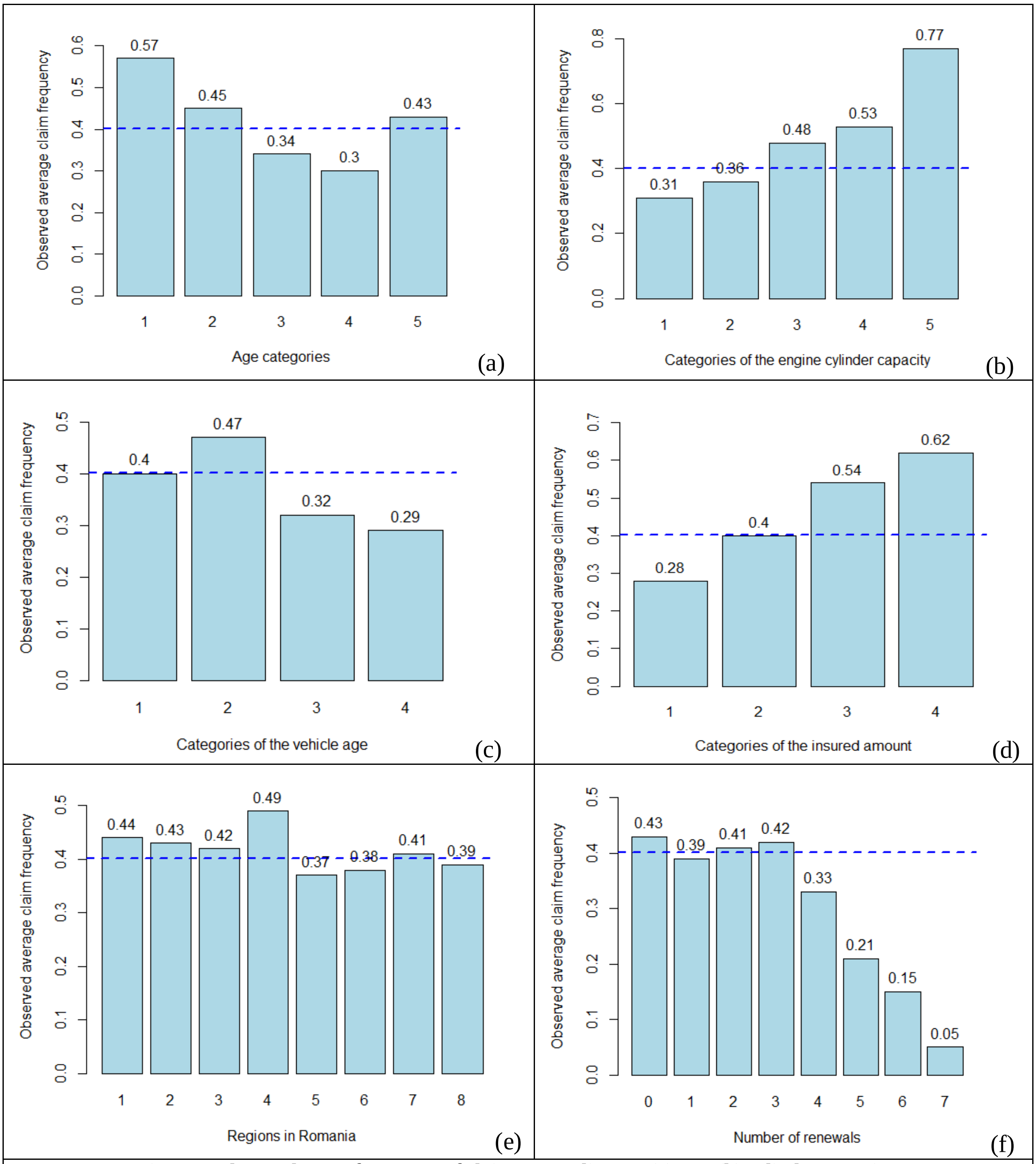

Figure 1.Observed mean frequency of claims per policy on: (a) age, (b) cylinder capacity, (c) automobile age, (d) insured amount, (e) region, and (f) number of renewals

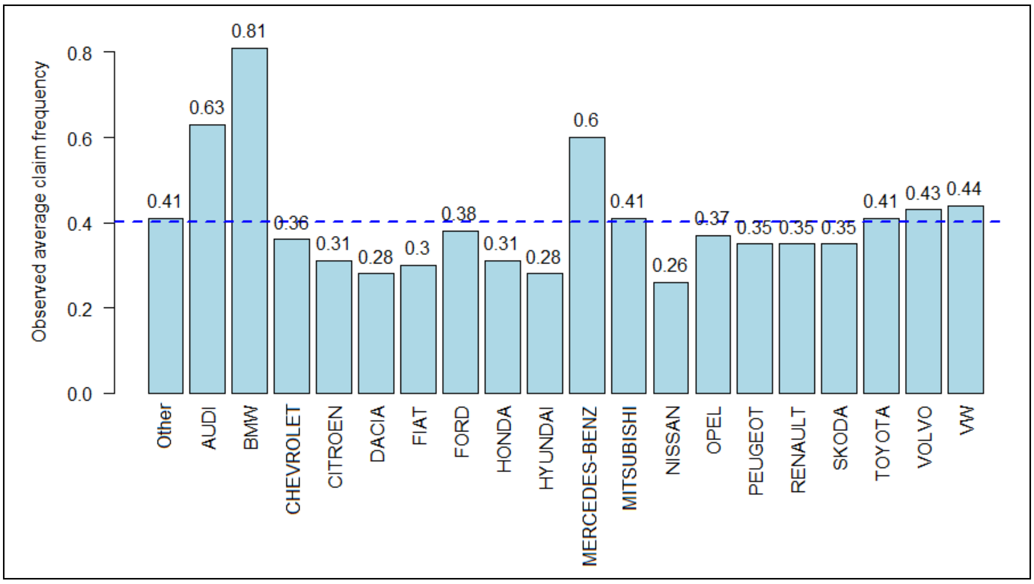

Figure 2. Observed average claim frequency, varying with the brand of the insured automobile.

We notice that the observed mean claim frequency of German automobile brands, such as Audi, BMW, Mercedes-Benz or Volkswagen, is above the sample average value of 40.15%.

A first approach in the estimation of the mean number of claims per policy would be to presuppose that *Y*, the variable describing the number of damage claims, follows a Poisson distribution. The estimation of the corresponding Poisson regression model was done in `R` with the `glm` routine, specifying the response variable, the reference levels of the categorical covariates, the type of distribution and the link function. Among the seven considered explanatory variables, only the insured amount was not statistically significant.

Table 3 illustrates the results of using the `drop1` function, corresponding to the application of some likelihood ratio tests, which help us compare, in turn, the all variable model to each of the models in which one variable was eliminated at a time. If the calculated significance level (*p-value*) is under 0.05, then that particular variable may be kept in the model.

Table 3.Statistical significance tests of the explanatory variables in the Poisson regression model

	Df	Deviance	AIC	LRT	Pr(>Chi)
Full model		14471	21749		
(Cat_age,"5")	4	14556	21826	84.908	$<2.2\cdot10^{-16}$***
(Cat_brand,"VW")	19	14612	21852	141.377	$<2.2\cdot10^{-16}$***

(Cat_cil_cap,"1")	4	14544	21814	73.540	$4.056 \cdot 10^{-15}$***
(Cat_auto_age, "2")	3	14586	21858	115.316	$<2.2 \cdot 10^{-16}$***
(Region,"8")	7	14513	21777	42.296	$4.561 \cdot 10^{-7}$***
(Cat_nb_renew,"0")	4	14528	21798	57.686	$8.880 \cdot 10^{-12}$***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

An improved version of the estimation of a Poisson regression model would be to consider the robust standard errors with probabilistic properties superior to those standard errors in a regular Poisson regression model. Consequently, Table 4 below presents the output of such a situation.

Table 4. The estimated Poisson regression model with robust standard errors

Variables	$\hat{\beta}_j$	Robust SE	Z	Pr(>|z|)
(Intercept)	-0.8487	0.0780	-10.8808	0.0000
(Cat_age,"5")1	0.2639	0.0769	3.4317	0.0006
(Cat_age,"5")2	0.0698	0.0486	1.4362	0.1509
(Cat_age,"5")3	-0.1739	0.0498	-3.4920	0.0005
(Cat_age,"5")4	-0.2380	0.0638	-3.7304	0.0002
(Cat_brand,"VW") Other	-0.0841	0.0762	-1.1037	0.2697
(Cat_brand,"VW")AUDI	0.2335	0.0746	3.1300	0.0018
(Cat_brand,"VW")BMW	0.4880	0.1158	4.2142	0.0000
(Cat_brand,"VW")CHEVROLET	-0.0416	0.1525	-0.2728	0.7851
(Cat_brand,"VW")CITROEN	-0.2197	0.1495	-1.4696	0.1418
(Cat_brand,"VW")DACIA	-0.2939	0.0880	-3.3398	0.0008
(Cat_brand,"VW")FIAT	-0.1896	0.1500	-1.2640	0.2063
(Cat_brand,"VW")FORD	-0.0621	0.0851	-0.7297	0.4657
(Cat_brand,"VW")HONDA	-0.4301	0.1815	-2.3697	0.0178
(Cat_brand,"VW")HYUNDAI	-0.3468	0.1340	-2.5881	0.0096
(Cat_brand,"VW")MERCEDES-BENZ	0.1837	0.0927	1.9817	0.0476
(Cat_brand,"VW")MITSUBISHI	-0.1360	0.1510	-0.9007	0.3680
(Cat_brand,"VW")NISSAN	-0.4860	0.1628	-2.9853	0.0028
(Cat_brand,"VW")OPEL	-0.0753	0.0906	-0.8311	0.4059
(Cat_brand,"VW")PEUGEOT	-0.0635	0.1087	-0.5842	0.5592
(Cat_brand,"VW")RENAULT	-0.0823	0.0871	-0.9449	0.3449
(Cat_brand,"VW")SKODA	-0.2171	0.0871	-2.4925	0.0127
(Cat_brand,"VW")TOYOTA	0.0111	0.0960	0.1156	0.9078
(Cat_brand,"VW")VOLVO	-0.0690	0.1323	-0.5215	0.6017
(Cat_cil_cap,"1")2	0.1529	0.0513	2.9805	0.0029
(Cat_cil_cap,"1")3	0.2952	0.0556	5.3094	0.0000
(Cat_cil_cap,"1")4	0.2746	0.0718	3.8245	0.0001
(Cat_cil_cap,"1")5	0.7063	0.1252	5.6414	0.0000
(Cat_auto_age,"2")1	-0.1217	0.0613	-1.9853	0.0471
(Cat_auto_age,"2")3	-0.3311	0.0408	-8.1152	0.0000
(Cat_auto_age,"2")4	-0.4350	0.2007	-2.1674	0.0302
(Region,"8")1	0.2055	0.0543	3.7845	0.0002
(Region,"8")2	0.0982	0.0712	1.3792	0.1678
(Region,"8")3	0.1187	0.0598	1.9849	0.0474
(Region,"8")4	0.1860	0.1277	1.4565	0.1453

(Region,"8")5	-0.1631	0.0902	-1.8082	0.0705
(Region,"8")6	-0.0547	0.0682	-0.8021	0.4225
(Region,"8")7	0.0529	0.0816	0.6483	0.5168
(Cat_nb_renew,"0")1	-0.0998	0.0482	-2.0705	0.0384
(Cat_nb_renew,"0")2	-0.0165	0.0588	-0.2806	0.7792
(Cat_nb_renew,"0")3	0.0298	0.0714	0.4174	0.6760
(Cat_nb_renew,"0")4	-0.3932	0.0753	-5.2218	0.0000

As the sample variance of variable Y is $s_y^2 = 0.6699 > 0.4015 = \hat{\mu} = \bar{y}$, we intend to investigate if the data exhibits overdispersion. First we calculate the sum of the squared Poisson residuals, namely the Pearson statistic χ^2=20042.0631, and then we estimate the dispersion parameter, $\hat{\phi} = 1.6120$, for which we obtain a value greater than 1. A Poisson regression model is correctly specified if the estimate $\hat{\phi}$ is very close to 1.

Identifying the presence of overdispersion presupposes to test the null hypothesis that the data does not exhibit overdispersion. This can be done with two tests suggested in Hilbe (2014) and in Cameron and Trivedi (2013). Applying these tests leads to a calculated level of significance of $2\cdot10^{-16}$, therefore the data in the sample support the alternative hypothesis.

Further we assume that the distribution of Y is Negative Binomial and we estimate the corresponding regression model, using the `glm.nb` function in the `MASS`package, or the `nbinomial` function in the `COUNT` package. The results presented in the table below show that the same explanatory variables are statistically significant, just as in the Poisson model.

Table 5.Statistical significance tests of the explanatory variables in the Negative Binomial model

	Df	Deviance	AIC	LRT	Pr(>Chi)
Full model		8462.3	20371		
(Cat_age,"5")	4	8512.0	20412	49.676	$4.219\cdot10^{-10}$***
(Cat_brand,"VW")	19	8541.9	20412	79.571	$2.204\cdot10^{-9}$***
(Cat_cil_cap,"1")	4	8502.1	20402	39.853	$4.642\cdot10^{-8}$***
(Cat_auto_age, "2")	3	8529.7	20432	67.445	$1.504\cdot10^{-14}$***
(Region,"8")	7	8485.7	20380	23.433	0.001433**
(Cat_nb_renew,"0")	4	8495.9	20396	33.643	$8.821\cdot10^{-7}$***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The comparison of the Poisson and Negative Binomial regression models can be performed applying a likelihood ratio test with the null hypothesis that the Poisson model better adjusts data with respect to the Negative Binomial model. The calculated level of significance is $1.209\cdot10^{-301}$, so we prefer the Negative Binomial model. Table 6 presents the output of the estimated Negative Binomial regression model.

Table 6. The estimated Negative Binomial regression model

Variables	$\hat{\beta}_j$	SE	Z	p-value
(Intercept)	-0.8315	0.0819	-10.1509	3.28E-24
(Cat_age,"5")1	0.2672	0.0875	3.0519	0.0023
(Cat_age,"5")2	0.0649	0.0516	1.2575	0.209
(Cat_age,"5")3	-0.1704	0.0522	-3.2633	0.0011
(Cat_age,"5")4	-0.2547	0.0620	-4.1111	3.94E-05
(Cat_brand,"VW") Other	-0.0891	0.0819	-1.0879	0.277
(Cat_brand,"VW") AUDI	0.2518	0.0839	3.0012	0.0027
(Cat_brand,"VW") BMW	0.5223	0.1371	3.8112	0.0001
(Cat_brand,"VW") CHEVROLET	-0.0580	0.1489	-0.3897	0.697
(Cat_brand,"VW") CITROEN	-0.2243	0.1528	-1.4673	0.142
(Cat_brand,"VW") DACIA	-0.3071	0.0935	-3.2824	0.0010
(Cat_brand,"VW") FIAT	-0.2074	0.1439	-1.4413	0.149
(Cat_brand,"VW") FORD	-0.0668	0.0925	-0.7227	0.47
(Cat_brand,"VW") HONDA	-0.4497	0.1643	-2.7365	0.0062
(Cat_brand,"VW") HYUNDAI	-0.3397	0.1334	-2.546	0.0109
(Cat_brand,"VW") MERCEDES-BENZ	0.1768	0.1057	1.6731	0.0943
(Cat_brand,"VW") MITSUBISHI	-0.1503	0.1516	-0.9909	0.322
(Cat_brand,"VW") NISSAN	-0.4649	0.1672	-2.7813	0.0054
(Cat_brand,"VW") OPEL	-0.0974	0.0951	-1.0241	0.306
(Cat_brand,"VW") PEUGEOT	-0.0766	0.1107	-0.6916	0.489
(Cat_brand,"VW") RENAULT	-0.0995	0.0932	-1.0672	0.286
(Cat_brand,"VW") SKODA	-0.2176	0.0878	-2.478	0.0132
(Cat_brand,"VW") TOYOTA	0.0018	0.0992	0.0184	0.985
(Cat_brand,"VW") VOLVO	-0.0681	0.1370	-0.4974	0.619
(Cat_cil_cap,"1")2	0.1535	0.0534	2.8774	0.00401
(Cat_cil_cap,"1")3	0.2928	0.0577	5.0749	3.88E-07
(Cat_cil_cap,"1")4	0.2527	0.0787	3.2106	0.0013
(Cat_cil_cap,"1")5	0.7141	0.1443	4.9477	7.51E-07
(Cat_auto_age,"2")1	-0.1109	0.0669	-1.6578	0.0973
(Cat_auto_age,"2")3	-0.3378	0.0417	-8.1054	5.26E-16
(Cat_auto_age,"2")4	-0.3927	0.2128	-1.8452	0.065
(Region,"8")1	0.1971	0.0587	3.3567	0.0008
(Region,"8")2	0.0969	0.0754	1.2841	0.199
(Region,"8")3	0.0942	0.0627	1.5013	0.133
(Region,"8")4	0.1605	0.1403	1.1444	0.252
(Region,"8")5	-0.1797	0.0942	-1.9079	0.0564
(Region,"8")6	-0.0785	0.0684	-1.1471	0.251
(Region,"8")7	0.0462	0.0846	0.5458	0.585
(Cat_nb_renew,"0")1	-0.0927	0.0505	-1.8352	0.0665
(Cat_nb_renew,"0")2	-0.0141	0.0621	-0.2275	0.82
(Cat_nb_renew,"0")3	0.0213	0.0767	0.2777	0.781
(Cat_nb_renew,"0")4	-0.3918	0.0716	-5.4721	4.45E-08

For this regression model, the Pearson statistics is χ^2=11927.85, and the estimate of the dispersion parameter is $\hat{\phi} = 0.9594$, a value very close to 1. The R functions that provide the parameters estimates $\hat{\beta}_j$ for the Negative Binomial regression model also return information about the parameter δ of the distribution, giving the estimate of θ, where $\theta = \frac{1}{\delta}$. In our case, $\hat{\theta} = 0.5532$, and then $\hat{\delta} = \frac{1}{\hat{\theta}} = 1.8075$.

For the observed data on the number of claims per policy y_i, respectively for the fitted values $\hat{\mu}_i$ in the regression model, we calculated the mean and the variance. Thus, $\text{mean}(y_i)$=0.4015, $\text{mean}(\hat{\mu}_i)$=0.4019, $\text{variance}(y_i)$=0.6699, $\text{variance}(\hat{\mu}_i) = \text{mean}(\hat{\mu}_i) + \hat{\delta} \cdot (\text{mean}(\hat{\mu}_i))^2$=0.6940, and we could notice that the variances are very close, so the Negative Binomial regression model successfully quantified overdispersion.

By applying the goodness of fit tests, namely Deviance Goodness of fit test and Pearson Goodness of fit test, in which the null hypothesis is that the response variable *Y* follows a Negative Binomial distribution, we obtained the calculated levels of significance (*p-value*) 1, and 0.9994 respectively, so that we cannot reject the null hypothesis.

4. Conclusions

The main results of this research paper are the identification and correction of overdispersion, starting from the data in the CASCO motor insurance policy sample, provided by a Romanian insurance company.

The detection of overdispersion was achieved through the successful application of the tests suggested by Hilbe (2014) and by Cameron and Trivedi (2013). The correction was done through the identification of a Negative Binomial regression model, the goodness of fit tests indicating that this model better adjusted the data. Thus, this model offered better predictions of the average number of claims per CASCO insurance policy than the Poisson regression model. The variables with significant influence on the mean frequency were: the age of the insurance policy owner, the automobile brand, its cylinder capacity, the vehicle age, the region where the policy was issued and its number of renewals. The insured amount of the vehicle was not statistically significant in any of the estimated regression models.

As far as we know, this paper is the first research study conducted on claim counts modeling done on the basis of data from a Romanian insurance company, so at the same time it represents a useful instrument for insurance providers.

Another important aspect that needs to be mentioned is the exclusive use of the open source R software in our processing of the data in order to obtain the graphical displays, as well as the estimates, and to perform statistical hypotheses tests referring to goodness of fit tests or likelihood ratio tests.

Regarding the limits of our research, we need to mention that these were independent from us, and perhaps a bigger sample is needed for more generalizable results, one to include more profiles of insurance policy owners, as well as more detailed information about the owners and about the insured vehicles.

Regarding the possible directions in which we could continue our research, we aim to carry on our estimations with other types of regression models, such as Poisson-Gaussian Inverse regression model or Zero-Inflated regression model, as well as to compare these models to the models fitted in this paper.

REFERENCES

[1] **Boucher, J.-P., Denuit, M., Guillen, M. (2007**), ***Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models***; *North American Actuarial Journal*, 11, 110-131;

[2] **Boucher, J.-P., Inoussa, R. (2014),** ***A Posteriori Ratemaking with Panel Data***; *ASTIN Bulletin*, 44, 587-612;

[3] **Cameron, A. C., Trivedi, P. K. (2013),** ***Regression Analysis of Count Data***; *Cambridge University Press*, Cambridge;

[4] **Covrig, M., Mircea, I., Zbăganu, G., Coşer, A., Tindeche, A. (2015),** ***Using* R *in Generalized Linear Models***; *Romanian Statistical Review*, 3/2015, 33-45;

[5] **Coxe, S., West, S. G., Aiken, L. S. (2009),*The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives***; *Journal of Personality Assessment*, 91, 121-136;

[6] **De Jong, P., Heller, G. Z. (2008),** ***Generalized Linear Models for Insurance Data***; *Cambridge University Press*, Cambridge;

[7] **Denuit, M, Charpentier, A. (2005),** ***Mathematiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement***; *Collection Economie et Statistique Avancees*, Economica, Paris;

[8] **Denuit, M., Marechal, X., Pitrebois, S., Walhin, J.-F. (2007),** ***Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems***; *Wiley*, New York;

[9] **Frees, E. W., Derrig, R. A., Meyers, G. (eds.) (2014)**, ***Predictive Modeling Applications in Actuarial Science;*** Volume 1, *Predictive Modeling, Cambridge University Press*, New York;

[10] **Gómez-Déniz, E., Ghitany, M. E., Gupta, R. C. (2016),** ***Poisson-mixed Inverse Gaussian Regression Model and Its Application***; *Communications in Statistics - Simulation and Computation*, 45, 2767-2781;

[11] **Hilbe, J. M. (2014),** ***Modeling Count Data***; *Cambridge University Press,* New York;

[12] **Klein, N., Denuit, M., Lang, S., Kneib, T. (2014),** ***Nonlife Ratemaking and Risk Management with Bayesian Generalized Additive Models for Location, Scale and Shape***; *Insurance: Mathematics and Economics*, 55, 225-249;

[13] **McCullagh, P., Nelder, J. A. (1989)** ***Generalized Linear Models***; *Chapman&Hall*, London;

[14] **Mullahy, J. (1986),** ***Specification and Testing of some Modified Count Data Models***; *Journal of Econometrics,* 33, 341–365;

[15] **Nelder, J. A., Wedderburn, R. W. M. (1972),** ***Generalized Linear Models***; *Journal of the Royal Statistical Society*, *Series A*, 135, 370-384;

[16] **Perumean-Chaney, S. E., Morgan, C., McDowall, D., Aban, I. (2013),** ***Zero-inflated and Overdispersed: What's one to do?***, *Journal of Statistical Computation and Simulation,* 83, 1671-1683;

[17] **Sarul, L. S., Sahin, S. (2015),** ***An Application of Claim Frequency Data Using Zero Inflated and Hurdle Models in General Insurance***; *Journal of Business, Economics&Finance,* 4, 732-743;

[18] **Shirazi, M., Lord, D., Dhavala, S. S., Geedipally, S. R. (2016),** ***A Semiparametric Negative Binomial Generalized Linear Model for Modeling Over-Dispersed Count Data with a Heavy Tail: Characteristics and Applications to Crash Data***; *Accident Analysis&Prevention*, 91, 10–18;

[19] **Tănăsescu, P., Mircea, I. (2014),** ***Assessment of the Ruin Probabilities***; *Economic Computation and Economic Cybernetics Studies and Research*, 48, 79-98, *ASE Publishing*;

[20] **Vangala, P., Lord, D., Geedipally, S. R. (2015),** ***Exploring the Application of the Negative Binomial–Generalized Exponential Model for Analyzing Traffic Crash Data with Excess Zeros***; *Analytic Methods in Accident Research*, 7, 29–36;

[21] **Wolny-Dominiak, A. (2013),** ***Zero-inflated Claim Count Modeling and Testing – A Case Study***; *Ekonometria,* 39, 144-151;

[22] **Yip, K. C. H., Yau, K. K. W. (2005),** ***On Modeling Claim Frequency Data in General Insurance with Extra Zeros***; *Insurance: Mathematics and Economics*, 36, 153-163;

[23**] Zha, L., Lord, D., Zou, Y. (2016)**,***The Poisson Inverse Gaussian (PIG) Generalized Linear Regression Model for Analyzing Motor Vehicle Crash Data***; *Journal of Transportation Safety&Security*, 8, 18-35;

[24] http://cran.r-project.org/

[25] http://www.mdrap.ro/dezvoltare-regionala/-2257/programul-operational-regional-2007-2013/-2975.