DE GRUYTER OPEN

Scientific Annals of the "Alexandru Ioan Cuza" University of Iaşi Economic Sciences
62 (2), 2015, 151-168
10.1515/aicue -2015-0011

MODELING THE FREQUENCY OF AUTO INSURANCE CLAIMS BY MEANS OF POISSON AND NEGATIVE BINOMIAL MODELS

Mihaela DAVID*
Danut-Vasile JEMNA**

Abstract

Within non-life insurance pricing, an accurate evaluation of claim frequency, also known in theory as count data, represents an essential part in determining an insurance premium according to the policyholder's degree of risk. Count regression analysis allows the identification of the risk factors and the prediction of the expected frequency of claims given the characteristics of policyholders. The aim of this paper is to verify several hypothesis related to the methodology of count data models and also to the risk factors used to explain the frequency of claims. In addition to the standard Poisson regression, Negative Binomial models are applied to a French auto insurance portfolio. The best model was chosen by means of the log-likelihood ratio and the information criteria. Based on this model, the profile of the policyholders with the highest degree of risk is determined.

Keywords: claim frequency, count data models, Poisson model, overdispersion, mixed Poisson models, negative binomial models, risk factors

JEL classification: G22

1. INTRODUCTION

We are living in a society that needs to manage risks of various types and with a significant economic impact. In the context of a risk based civilization, the need of protection has become more pronounced, having as a consequence the request of financial security against possible losses. Therefore, the emergence and development of the insurance business are related to the urgent need to protect the individuals and their assets against a possible loss caused by a particular event. The entire process of insurance consists in offering an equitable method of transferring the risk of a contingent or uncertain loss in exchange for payment.

* Faculty of Economics and Business Administration, Alexandru Ioan Cuza University of Iaşi, Romania; e-mail: *mihaela_david88@yahoo.com.*

** Faculty of Economics and Business Administration, Alexandru Ioan Cuza University of Iaşi, Romania; e-mail: *danut.jemna@uaic.ro.*

Non-life insurance business, especially auto insurance branch, holds an increased interest because it is required to manage a large number of situations (both the number of insured vehicles and of accidents) with a wide variety of risks.

A fundamental goal of insurance companies is to calculate an appropriate insurance price or premium corresponding to an insured in order to cover a certain risk. A well-known method to calculate the premium is to multiply the conditional expectation of the claim frequency with the expected cost of claims. Therefore, modelling frequency of claims, also known in theory as count data, represents an essential step of non-life insurance pricing. As sustained in Boucher and Guillen (2009), count regression analysis permits the identification of the risk factors and the prediction of the expected frequency of claims given the risk characteristics.

In the past years there has been considerable interest in count data models, particularly in the actuarial literature. As mentioned in Cameron and Trivedi (1998), an important milestone in the development of models for count data is reached by the emergence of Generalized Linear Models (GLMs). The Poisson regression is a special case of GLMs that was first developed by Nelder and Wedderburn (1972) and detailed later in the papers of Gourieroux *et al.* (1984a, 1984b) and in the work on longitudinal or panel count data models of Hausman *et al.* (1984). Within non-life insurance context, McCullagh and Nelder (1989) demonstrate that the usage of the GLMs techniques, in order to estimate the frequency of claims, has an a priori Poisson structure. Antonio *et al.* (2012) present the Poisson distribution as the modelling archetype of claim frequency.

Although it offers a favourable statistical support, Gourieroux and Jasiak (2001) emphasizes that the Poisson distribution presents significant constraints that limit its use. The Poisson distribution implies equality of variance and mean, a property called equidispersion that, as sustained in Cameron and Trivedi (1999), is a particular form of unobserved heterogeneity. One of the well-known consequences of unobserved heterogeneity in count data analysis is overdispersion which means that the variance exceeds the mean. Other explanation is provided by Jong and Heller (2013) who termed the overdispersion as extra-Poisson variation because this type of data displays far greater variance than that predicted by the Poisson model.

Vasechko *et al.* (2009) state that the problem of overdispersion, inherent to the Poisson model, implies the underestimation of standard errors of the estimated parameters, which leads to the rejection of the null hypothesis, according to which the regression coefficients are not statistically relevant. Consequently, the restrictive nature of Poisson model has sustained the development of numerous techniques proposed for both testing and handling overdispersed data. An exhaustive analysis of these tests is provided in Hausman *et al.* (1984), Cameron and Trivedi (1990), 1998), Gurmu (1991), Jorgensen (1997) or in more recent studies such as Charpentier and Denuit (2005), Jong and Heller (2013), Hilbe (2014).

The alternative distributions used most frequently in order to correct the overdispersion are known as compound or mixed distributions. According to the literature, a particular example of this class is the negative binomial distribution which consists of simple and efficient techniques that oversee the limits of the Poisson distribution and offer results qualitatively similar. In the statistical literature there are presented many ways to construct the negative binomial distribution, however the most used are the NB1 and NB2 forms, introduced by Cameron and Trivedi (1998). Among the recent studies, Denuit *et al.* (2007) give a comprehensive image concerning the mixed Poisson models and they highlight that negative binomial distribution is a satisfactory alternative to Poisson distribution in order to

estimate the claim frequency for an auto insurance portfolio. Working with cross-sectional insurance data, Boucher *et al.* (2007) sustain that the comparison of the log-likelihoods for the two distributions reveals that the extra parameter of the negative binomial distribution improves the fit of data in comparison with the Poisson distribution. For longitudinal or panel data, an excellent account of claim frequency distributions can be consulted in Boucher *et al.* (2008), Boucher and Guillen (2009) and Antonio and Valdez (2010), in which the authors analyse and emphasize the practical use of negative binomial models for auto insurance data.

In the literature of non-life insurance pricing a current research topic is how to identify the variable and the types of variables that allow estimating the frequency of a certain insured risk. A standard classification would include: age and gender-marital status of insured, usage purpose of the insured vehicle, geography (location of garage) and other factors such as whether the vehicle is a sport car or not (Antonio and Valdez, 2010). A more systematic classification is provided by Kouki (2007) who identifies three categories of risk factors: the driver (age, sex, age of driving license, bonus-malus coefficient), the vehicle (power, age) and the insurance contract. In this context, the empirical studies are valuable because they permit the evaluation of a theoretical hypothesis while projecting these factors on an insurance portfolio (Charpentier and Denuit, 2005; Yip and Yau, 2005; Denuit *et al.*, 2007; Allain and Brenac, 2012; Boucher *et al.*, 2013). The results of these studies are considered by the insurance companies while assessing their calculation tools and proposing new solutions according to changes in the behaviour and characteristics of clients.

The present study lines up with the current focus of the auto insurance literature. The aim of this study can be highlighted at two levels. The first is theoretical and methodological and aims to present synthetically the econometric modelling methodology of auto claim frequency. The second objective is related to the empirical part of this research. Working with a French auto insurance portfolio, we estimate an econometric model for claim frequency. On this level, the main contribution of the study is represented by a specific set of explanatory variables that take into account a number of updates concerning the data registered by insurance companies. For example, in this study, we introduce as risk factors the variables occupation of insured, GPS and value of vehicle. Also, in comparison with similar studies, we use a different classification of the insured based on age intervals on the assumption that more homogenous groups will be obtained and the calculation of premiums will better correspond to the reality of studied phenomenon. Although the results cover a portfolio of a French insurance company, the methodology of data count models can be applied to other insurance portfolios of companies from other European countries such as Romania.

The paper is structured as follows. Section 2 deals with a brief presentation of the used data and aspects related to methodology of count data models. Section 3 includes our empirical study. Concluding remarks are summarized in Section 4.

2. DATA AND METHODOLOGY

In this paper, we worked with an auto insurance portfolio of a company operating in France. The analysed phenomenon concerns the third party liability for the damages of the vehicles, for which the insurance is covering the losses within the limits of the insured amount.

2.1. Sample

The sample contains 150021 policies observed during the period 2007-2009. We use 9 exogenous variables for every policy, as well as the total frequency of claims at fault that were reported within the yearly period. Therefore, except the explained variable, the *frequency of claims*, the other ones are considered risk factors that are known a priori by the insurer. In comparison with similar empirical studies, we group the risk factors into three categories that reflect the policyholder characteristics: *age*, *occupation*; the vehicle features: *value*, *type*, *category*, *use*, *GPS*; the insurance policy characteristics: *insurance policy duration*, *bonus-malus coefficient*. Table 1 summarizes the information available about each policyholder.

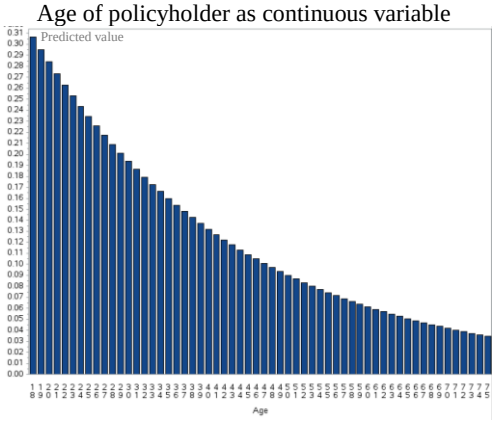
Table no. 1 – List of variables

Variable	Description	Values
Count	Frequency of claims	From 0 to 5 claims declared
Age	Age of policyholder	From 18 to 75 years
Occup	Occupation of policyholder	Employed, Housewife, Retired, Self-employed, Unemployed
Type	Type of vehicle	A, B, C, D, E, F
Categ	Category of vehicle	Large, Medium, Small
Use	Purpose of vehicle usage	Private, Professional, Other
Value	Value of vehicle	From 1000 to 50000 Euros
GPS	GPS device	Yes, No
Bonus	Bonus-malus coefficient	From -50 to 150
Duration	Duration of insurance policy	From 0 to 15 years

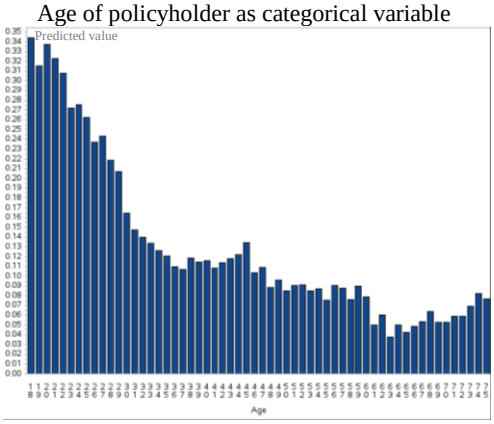
Among these variables, *bonus-malus coefficient* presents a particular interest for auto insurance pricing, having a specific meaning according to the insurance system of each country. Within the French bonus-malus system, this coefficient indicates an increase or a decrease of the insurance premium depending on the number of claims declared by a policyholder. Therefore, if the policyholder does not cause any responsible accident, he receives a bonus, meaning that the insurance premium will be reduced. Contrary, if the policyholder is responsible for the accident, he is penalized by applying a malus, which will have the consequence of a premium increase. These increases and decreases are based on a standard tariff defined by the insurer, depending on which the premium is multiplied by a coefficient. The basic coefficient is 1 and it corresponds to the reference premium of the insurance company. If the *bonus-malus coefficient* is lower than this value, a bonus is applied, and if it is higher, a malus is considered. More specifically, the French bonus-malus system involves a malus of 25% for a claim declared and a bonus of 5% for the non-declaration of any claims in the reference period, usually a year. In this way, the system aims the encouragement of prudent insured drivers and the discouragement of those who, for various reasons, register severe losses. In the studied portfolio, the calculations corresponding to the *bonus-malus coefficient* are already generated, registering negative and positive values, which indicate a decrease or an increase of the insurance premium, respectively.

To assess how best to enter policyholder's age variable into the count model, we examined the difference between the fitted frequency of claims, considering age as a single risk factor, once being introduced as continuous variable and once as categorical variable. Figure 1 illustrates simultaneously the distribution of the expected frequency of claims explained by the policyhoder's age both as continuous and categorical variable (with 58 categories of age).

Age of policyholder as continuous variable



Age of policyholder as categorical variable



Source: Data processed within SAS 9.3

Figure no. 1 – The fitted frequency of claim depending on the age of policyholder

In the first case, a decrease of the claim frequency can be observed along with an increase in the *age of policyholders*. In the case with the age categorical, is also noted the concave shape of the fitted frequency of claims, obtaining high values for the category of young drivers, an obvious decrease over the years, but a slightly increase in elderly drivers category. Taking into account that among certain age groups the estimated frequencies of claims do not differ significantly, this variable could be grouped into fewer categories considering the breakpoints that can be easily observed on the right side of the graph. Therefore, based on this graphic representation, the policyholder's age could be grouped on year intervals as follows:

$$AgeGroup = \begin{cases} Beginner \ (18\text{-}22 \text{ years}); \\ Young \ (23\text{-}29 \text{ years}); \\ Experienced \ (30\text{-}60 \text{ years}); \\ Senior \ (61\text{-}67 \text{ years}); \\ Elderly\ driver \ (68\text{-}75 \text{ years}). \end{cases}$$

Further, the age of the policyholder will be considered in analysis as a risk factor (with the five categories established) in order to obtain homogeneous groups of policyholders, and thereby an accurate assessment of their risk level.

2.2. Econometric models

Within non-life insurance, when actuaries are interested in estimating the frequency of claims, the Poisson model is often considered. Although the literature sustains that it offers a favourable statistical support for count data, the Poisson model implies the equidispersion assumption that is a drawback in practical use when data is overdispersed. The literature presents several reasons why data can be overdispersed and also many models to address the variety of overdispersion found in data. In general, if the cause of overdispersion in Poisson model is not diagnosed, the negative binomial models are commonly recommended. There are a wide number of negative binomial models used, but for insurance data the more intuitive ones are considered the NB1 and NB2 forms of the negative binomial distribution.

The following part of this paper deals with aspects related to insurance count data and applied methodology, presenting at length the properties, the empirical evidence and the comparison of the three applied count data models.

Poisson model

An excellent definition of the law of rare events is given in Cameron and Trivedi (1998). The authors state that the total number of events will follow, approximately, the Poisson distribution if an event may occur in any of a large of trials but the probability of occurrence in any given trial is small. In the context of actuarial literature, comprehensive references on Poisson distribution, used as the main tool in estimating the claim frequency, are Dionne and Vanasse (1989), 1992), Denuit and Lang (2004), Gourieroux and Jasiak (2004), Yip and Yau (2005) and many others.

If the discrete random variable Y_i (claim frequency or observed number of claims), conditioned by the vector of explanatory variables X_i (the insured's characteristics), is assumed to be Poisson distributed, the probability density function of Y_i is:

$$f(y_i|x_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{1}$$

Therefore, the relation (1) represents the probability that the random variable Y_i takes the value y_i ($y_i \in \mathbb{N}$), considering the characteristics of policyholders.

Although the Poisson distribution is considered to be the benchmark model in non-life insurance data, McCullagh and Nelder (1989) sustain that it implies a particular form of heteroskedasticity, leading to the equidispersion hypothesis. This assumption is emphasized by Gourieroux and Jasiak (2001) as a severe drawback that limits the model use because it implies that the conditioned mean and variance of claim frequency are equal. Therefore, the Poisson distribution parameter represents at the same time the mean and the variance of distribution:

$$E(y_i|x_i) = V(y_i|x_i) = \lambda_i \tag{2}$$

Within GLMs framework, the mean of the dependent variable is related to the linear predictor through the so called link function. It is well known fact from the literature that a logarithmic function is the natural link function for the Poisson distribution:

$$ln(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \Rightarrow \lambda_i = e^{x_i^t \beta} \tag{3}$$

Estimations of the parameters are done by maximum likelihood. In order to find the maximum likelihood of (1), the likelihood function is defined as follows:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-e^{x_i^t\beta}}(e^{x_i^t\beta})^{y_i}}{y_i!}$$

Using a logarithm in both sides of the previous equation, the log-likelihood function is obtained:

$$LL(\beta) = \sum_{i=1}^{n} [y_i ln\lambda_i - \lambda_i - lny_i!] = \sum_{i=1}^{n} [y_i x_i^t \beta - e^{x_i^t \beta} - lny_i!] \quad (5)$$

The maximum likelihood estimators $\hat{\beta}_j$ are the solutions of the equations obtained by differentiating the log-likelihood in terms of regression coefficients and solving them to zero. The equations forming the system are not generating explicit solutions and therefore they need to be solved numerically by using an iterative algorithm. Charpentier and Denuit (2005) consider that the most common iterative methods are either Newton-Raphson or Fisher information.

Although Poisson distribution is often used to estimate the frequency of claims, the empirical evidences from literature show that it is usually too restrictive for this type of data. The fundamental problem of the Poisson distribution is that for count data the variance usually exceeds the mean, a feature known as overdispersion.

Overdispersion can result from many reasons. Hilbe (2007) provides an excellent discussion of this issue, differentiating between apparent and real overdispersion. Apparent overdispersion can occur as a result of outliers, the exclusion of relevant risk factors or interaction terms. In this respect, Denuit *et al.* (2007) highlight that overdispersion arises because differences in driving behaviour among individuals cannot be observed by the insurer, such as swiftness of reflexes, aggressiveness behind the wheel, consumption of drugs, etc. When the resolution of these issues does not have a conclusive response, the overdispersion is assumed to be real and it may be due to unobserved heterogeneity related to the equidispersion hypothesis.

The literature presents numerous techniques developed in order to test the assumption of overdispersion. In this regard, Cameron and Trivedi (1990) propose a test for overdispersion by estimating the Poisson model, constructing fitted values $\hat{\lambda}_i = exp(x_i^i \hat{\beta})$, and performing the auxiliary OLS regression without intercept:

$$\frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\hat{\lambda}_i} = \alpha \frac{g(\hat{\lambda}_i)}{\hat{\lambda}_i} + u_i \quad (6)$$

where u_i is the error term and $g(\hat{\lambda}_i)$ is a known function, most commonly $g(\hat{\lambda}_i) = \hat{\lambda}_i^{\,2}$ or $g(\hat{\lambda}_i) = \hat{\lambda}_i^{\,2}$. The first function corresponds to the NB2 form of negative binomial distribution, and the second is related to the NB1 form of negative binomial distribution, both forms being discussed at length in the following part of methodology. The null hypothesis of no overdispersion ($H_0: \alpha = 0$) can be tested against the alternative hypothesis of overdispersion ($H_1: \alpha > 0$) using the t statistic for α.

Another practical and reliable test for overdispersion is introduced by Greene (2002) and is based on the Lagrange Multiplier test (LM). This statistics follow the χ^2 distribution with one degree of freedom and it is given by:

$$LM = \frac{(\sum_{i=1}^{n} \lambda_i^{\,2} - n\bar{y}_i)^2}{2\sum_{i=1}^{n} \lambda_i^{\,2}} \quad (7)$$

If after comparing the statistics calculated value with the theoretical one the test appears to be significant, then the hypothesis of no overdispersion is rejected. Therefore, the approach of the various alternatives of Poisson model is preferred.

Negative Binomial models

The alternative to the Poisson distribution used most frequently in order to handle count data when the variance is appreciably greater than the mean is the negative binomial distribution.

The negative binomial distribution is employed as a functional form that relaxes the equidispersion restriction of the Poisson model. The literature presents many ways to construct the negative binomial distribution, but Boucher *et al.* (2008) argue that the more intuitive one is the introduction of a random heterogeneity term θ of mean 1 and variance α in the mean parameter of the Poisson distribution. This general approach is discussed at length by Gourieroux *et al.* (1984a, 1984b), Cameron and Trivedi (1986, 1990, 1998), Winkelmann (2004) and Greene (2008). Regarding the usage on the insurance data, a classic example arises from the theory of accident proneness which was developed by Greenwood and Yule (1920). This theory sustains that the number of accidents is Poisson distributed, but there is gamma-distributed unobserved individual heterogeneity reflecting the fact that the true mean is not perfectly observed. Within the actuarial literature, the problem of mixed models is also illustrated and developed in the studies of McCullagh and Nelder (1989), Lawless (1987), Dionne and Vanasse (1989), Denuit and Lang (2004), Boucher *et al.* (2007), Hilbe (2014).

The traditional negative binomial is derived from a Poisson-gamma mixture distribution. Therefore, if the variable θ is considered to be gamma distributed, with the following density distribution:

$$f(\theta) = \frac{(1/\alpha)^{1/\alpha}}{\Gamma(1/\alpha)} \theta^{1/\alpha - 1} \exp(-\theta/\alpha) \tag{8}$$

it is well known that the negative binomial is the resultant overall distribution of claim frequency.

McCullagh and Nelder (1989) sustain that a random variable Y_i is called a negative binomial distributed count with parametersλ_i and α (> 0) if the probability mass function is given by:

$$f(y_i, \lambda_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}} \left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i}\right)^{y_i} \tag{9}$$

In (9) the term α plays the role of a dispersion factor and it is a constant. When α goes to 0, it is obvious that the NB distribution reduces to the standard Poisson distribution with parameter λ_i.

Cameron and Trivedi (1986) consider a more general class of negative binomial distribution (NBp) having the same mean λ_i, but a variance of the form $\lambda_i + \alpha\lambda_i^p$. Cameron and Trivedi (1998) there are presented the two most commonly known and utilized variants of the negative binomial distribution. When p is set to 1, it leads to the NB1 distribution and a model with $p = 2$, is called the NB2 distribution, also referred to as a quadratic negative binomial distribution.

The probability mass function of the NB1 model is:

$$f(y_i, \lambda_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1}\lambda_i)}{\Gamma(y_i + 1)\Gamma(\alpha^{-1}\lambda_i)}(1+\alpha)^{-\lambda_i/\alpha}(1+\alpha^{-1})^{-y_i}) \quad (10)$$

The first two moments of the NB1 are the following:

$$E[y_i] = \lambda_i \quad (11)$$

$$V[y_i] = \lambda_i + \alpha\lambda_i = \phi\lambda_i \quad (12)$$

For the NB1 model, the log-likelihood function (LL) is presented below:

$$LL(\alpha, \beta) = \sum_{i=1}^{n}\left\{\left(\sum_{k=0}^{y-1}(k + \alpha^{-1}\lambda_i)\right) - lny_i! - (y_i + \alpha^{-1}\lambda_i)\, ln(1+\alpha) + y_i ln\alpha\right\} \quad (13)$$

Estimation based on the first two moments of the NB1 density, as suggested in Cameron and Trivedi (1998), yields the Poisson GLMs estimator, which is also called the NB1 GLMs estimator.

For the NB2, the probability mass function coincides with the general negative binomial density function as in (9). The two first moments of the NB2 distribution are:

$$E[y_i] = \lambda_i = e^{x_i^t \beta} \quad (14)$$

$$V[y_i] = \lambda_i(1 + \alpha\lambda_i) \quad (15)$$

The log-likelihood function corresponding to the NB2 model is given as follows:

$$LL(\alpha, \beta) = \sum_{k=1}^{n}\left\{-log(y_i) + \sum_{k=1}^{y_i} log(\alpha y_i - k\alpha + 1) - (y_i + \alpha^{-1})log(1 + \alpha\lambda_i) + y_i log(y_i)\right\} \quad (16)$$

Cameron and Trivedi (1998) argue that the differences of estimating α have received little attention in the literature, mostly because the interest lies in estimation of β_j, with α as a nuisance parameter. The same authors highlight that even then they are important, as the standard error estimates of $\hat{\beta}_j$ depend considerably on $\hat{\alpha}$.

Boucher and Guillen (2009) state that the process of parameters' estimation is approximately the same for all three models. However it is highlighted that this situation is expected when a satisfactory number of conditions for consistency are satisfied. For the NB models, Vasechko *et al.* (2009) and Allain and Brenac (2012) state that the estimated parameters $\hat{\beta}_j$ and estimated values $\hat{\lambda}_i$ differ slightly while compared with the results achieved after applying Poisson model, but the standard errors values of the estimators $\hat{\beta}_j$ increase significantly after applying the model's alternatives. Boucher and Guillen (2009), analysing the auto claim frequency through the NB models, observe that although the regression coefficients do not change significantly, they allow a better assessment of the standard errors of the estimates that may be underestimated by Poisson model. A similar

point of view belongs to Hilbe (2014) who sustains that the NB distribution presents similar properties as Poisson distribution, knowing that the mean is understood in the same manner as the Poisson mean, but the variance has a much wider scope than is allowed by the Poisson distribution.

Criteria for assessing goodness of fit

The literature presents many statistics that can be used to select and to assess the performance of count data models.

As discussed in Denuit and Lang (2004), the standard measure of goodness of fit that can be used to assess the adequacy of various models is the likelihood ratio (LR) that follows a $\chi^2_{\alpha,p}$ distribution for a level of significance α of 0.05 and with p degrees of freedom, where p represents the number of explicative variables included in the regression model. This statistics test is obtained from the difference between the deviance of the regression model without covariates (D_0) and the deviance of the model including the independent variables (D_p):

$$LR = D_0 - D_p \tag{17}$$

Charpentier and Denuit (2005) define the deviance as twice the difference between the maximum log-likelihood achievable ($y_i = \lambda_i$) and the log-likelihood of the fitted model:

$$D = 2(LL(y_i|y_i) - LL(\lambda_i|y_i) \tag{18}$$

A value of the likelihood ratio higher than the statistics theoretical value ($LR > \chi^2_{\alpha,p}$) suggests that the regression model explains well the analysed data.

In order to compare the models, there are used some tests based on the log-likelihood function. In this regard, a standard method of comparison between the Poisson and NB models is to use the likelihood ratio, given by the expression: $LR = -2(LL_P - LL_{NB})$, where LL_P and LL_{NB} are the values of the log-likelihood under the Poisson and negative binomial models, respectively. This statistics follows the χ^2 distribution with one degree of freedom. A calculated value of the test higher than the theoretical value ($LR > \chi^2_{2\alpha;1}$) underlines that the NB models are chosen to the detriment of Poisson regression.

A convenient method used to discriminate between the two NB models is the comparison of the log-likelihood function values. Another standard method to distinguish the discussed models refers to information criteria, which is also based on the fitted log-likelihood function. Boucher *et al.* (2007) sustain that the criteria used to compare the models must penalize the one with a large number of parameters, considering the fact that the likelihood increases with the addition of parameters. The standard criteria refers to Akaike Information Criteria ($AIC = -2LL + 2p$) and the Bayesian Information Criteria ($BIC = -2LL + pln(n)$), where p represents the number of parameters introduced in the regression model, n indicates the sample volum, and LL is the model log-likelihood function. The literature proposes many others information criteria that employ a penalty term associated with the number of parameters (p) and the sample size (n), but AIC and BIC criteria are the most often used in practice. An overview of penalized criteria is presented at duration by Kuha (2004).

According to the literature, mixed results were obtained concerning the models employed to estimate the claim frequency. In the application of Cameron and Trivedi

(1998), the NB2 model is preferred to NB1 as it has higher log-likelihood, with the same number of parameters. In contrary, the results of a more recent study of Boucher *et al.* (2007) confirm that the NB1 was one of the best models to fit the auto insurance data with which they worked. Analysing different type of insurance data, including auto portfolios, Hilbe (2014) argues that NB1 and NB2 are valuable models both to diagnose and to adjust overdispersion in data. Hence, there is no empirical evidence in the literature to support a certain count model, given the fact that the overdispersion appears for several reasons as discussed previous in this section.

3. EMPIRICAL RESULTS

3.1. Descriptive statistics

Analysing the insurance portfolio structure, it can be noticed that the maximum frequency of declared accidents or claims by a policyholderis 5. More specifically, throughout the analysed period, 131838 (87.88%) of policyholders did not declare any accident, 15395 (10.26%) of policyholders declared one claim, 2202 (1.47%) of them informed the insurance company of the occurrence of two accidents, 442 (0.29%) of policyholders had three claims declared, 100 (0.07%) policyholders with four claims and only 44 (0.03%) of them declared to the company five claims. The distribution of the claim frequency suggests that the portfolio is heterogeneous, an aspect that can be easily deduced from the results shown in Table 2.

Table no. 2 – Variables Analysis

Variable	Mean	Median	Std Dev	Min	Max
Count	0.1449	0	0.4299	0	5
Bonus	-6.8603	-30	48.7486	-50	150
Duration	5.4975	4	4.6031	0	15

Source: Data processed within SAS 9.3

Thus, with a mean of 0.1449 and a variance of 0.1848, the variance of claim frequency exceeds its mean. In addition, the distribution of the independent variable, *bonus-malus coefficient*, shows that more than 50% of policyholders benefit from a bonus because they did not have any accidents declared. On one hand, it indicates a lack of homogeneity in the data, but on the other hand, it shows that the policyholders present a low risk for the insurance company. These results admit the assumption of heterogeneity and justify the a priori differentiation of policyholders.

3.2. Econometric models

Within SAS, the GENMOD procedure is used to fit the Poisson and NB2 regression models in the framework of GLMs. The Type 3 analysis, generated by using this procedure, permits a test of the relevance of one variable taking all the others into account. For the fit of NB1 model, the used procedure in SAS is COUNTREG.

Poisson model

The results obtained from the Poisson regression for the insurance portfolio presented in Section 2 are shown in Table 3 (LR statistics for analysed variables) and Table 5 (regression coefficients' estimations). In Table 3, in column *Chi-square* is calculated, for each variable, two times the difference between the log-likelihood of the model which includes all the independent variables and the log-likelihood of the model obtained by deleting one of a specified variable. This test follows the asymptotic $\chi^2_{\alpha,p}$ distribution for a level of significance α of 0.05 and with p degrees of freedom that represent the number of parameters associated to the analyzed variable.

Table no. 3 – LR Statistics for Type 3 Analysis

Source	Poisson Regression(*)		Poisson Regression(**)	
	Chi-Square	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**
AgeGroup	3822.01	<.0001	3977.74	<.0001
Occup	1126.40	<.0001	1129.56	<.0001
Type	375.13	<.0001	375.27	<.0001
Categ	1.25	0.5358	-	-
Use	1690.01	<.0001	1794.57	<.0001
GPS	608.82	<.0001	608.92	<.0001
Value	3.79	0.0516	-	-
Bonus	7450.25	<.0001	7454.48	<.0001
Duration	325.67	<.0001	325.92	<.0001

(*) Poisson regression including all the explanatory variables
(**) Poisson regression including only the significant explanatory variables

Source: Data processed within SAS 9.3

It can be observed that the variable denoting the *category of vehicle* is not statistically significant as it yields a *p-value* of 0.5358 greater than the level of significance α of 0.05. In consequence, this variable is excluded from the model and the analysis will continue in the same manner until it is obtained the optimal combination of factors (*p-values* < 0.05) which can explain the variation of claim frequency. After excluding from the model the non-significant factors (*category* and *value of vehicle*), it is noticed that all the other predictors appear to significantly contribute to the process of understanding and predicting the frequency of claims made on vehicle insurance policies. Nevertheless, if the equidispersion assumption of Poisson distribution is not fulfilled, we are dealing with overdispersed data, and thereby the *p-values* tell us nothing about the relationship of the predictor and response. Therefore, it is imperative to test the equidispersion assumption, meaning the equality between the conditioned mean and variance of claim frequency, when constructing and interpreting a Poisson model. In this context, the method of Cameron and Trivedi and the test of Greene are used. Table 4 shows the results obtained after estimating and testing the parameter α for both forms of the known function $g(\hat{\lambda}_i)$ developed by Cameron and Trivedi.

Table no. 4 – Parameter Estimates

Form of function $g(\hat{\lambda}_i)$	Parameter (α)	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
$g(\hat{\lambda}_i) = \hat{\lambda}_i$	α_0	1	0.01412	0.00553	2.55	<.0107
$g(\hat{\lambda}_i) = \hat{\lambda}_i^{\,2}$	α_1	1	0.22806	0.04904	4.65	<.0001

Source: Data processed within SAS 9.3

At a level of significance of 0.05, the values of t statistic obtained for both α_0 and α_1 parameters leads to the rejection of the null hypothesis of equidispersion, indicating that there is overdispersion in the data and that the use of both NB1 and NB2 forms of negative binomial distribution is justified. With a *p-value* under 0.05, the Lagrange Multiplier test (LM = 1896.19) appears to be significant, and thereby the hypothesis of no overdispersion is again rejected. Both test statistics indicate very strong evidence against the fit of the Poisson model to the data and thus to correct the overdispersion, the alternative mixed models presented in Section 2 are used.

Negative binomial models

Both NB1 and NB2 regressions are based on the same explanatory variables of the claim frequency, leading to similar results as the ones from the Poisson regression. For all parameters, the *p-value* is under 0.05. Analyzing the results from Table 5, it can be observed that the parameters and the estimated values are very close to those obtained in the previous model. The standard errors of parameter estimates are slightly higher than those obtained for the Poisson model, but this does not impact the statistical significance of the regression coefficients. The two adjusted models do not provide further details in comparison with the Poisson regression in terms of risk factors that explain the variation of claim frequency, but managing these enhanced models could make a difference in terms of adjusting the Poisson overdispersion in the data.

Table no. 5 – Analysis of Parameter Estimates

Parameter	Poisson Regression		NB1 Regression		NB2 Regression	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Intercept	-0.1644	0.0339	-0.1775	0.0351	-0.1714	0.0362
AgeGroup (Elderly)	-0.8917	0.0504	-0.8796	0.0517	-0.8849	0.0521
AgeGroup(Experienced)	-1.1281	0.0184	-1.1167	0.0191	-1.1230	0.0195
AgeGroup (Senior)	-1.1332	0.0541	-1.1139	0.0551	-1.1255	0.0555
AgeGroup (Young)	-0.4328	0.0203	-0.4285	0.0211	-0.4296	0.0217
Occup (Employed)	-0.2417	0.0183	-0.2404	0.0189	-0.2403	0.0194
Occup (Retired)	-1.1899	0.0445	-1.1870	0.0455	-1.1907	0.0454
Occup (Self-employed)	-0.3772	0.0216	-0.3761	0.0223	-0.3762	0.0227
Occup (Unemployed)	-0.0749	0.0207	-0.0738	0.0214	-0.0763	0.0220
Type (A)	-0.4009	0.0292	-0.3992	0.0301	-0.4025	0.0310
Type (B)	-0.3072	0.0296	-0.3040	0.0306	-0.3103	0.0315
Type (C)	-0.2443	0.0314	-0.2405	0.0325	-0.2449	0.0334
Type (D)	-0.1308	0.0295	-0.1318	0.0305	-0.1317	0.0314
Type (E)	-0.0693	0.0317	-0.0734	0.0328	-0.0732	0.0338
Use (Other)	-0.9163	0.0280	-0.9048	0.0288	-0.9156	0.0290
Use (Private)	-0.4624	0.0141	-0.4582	0.0145	-0.4619	0.0149
GPS (Yes)	-0.3568	0.0148	-0.3531	0.0152	-0.3480	0.0155
Bonus	0.0106	0.0001	0.0105	0.0001	0.0106	0.0001
Duration	-0.0272	0.0015	-0.0267	0.0016	-0.0271	0.0016
Scale	1.0000	0.0000	0.2824	0.0148	0.3952	0.0239

Source*: Data processed within SAS 9.3*

Reviewing the coefficient signs from Table 5, a decrease of the claim frequency can be observed along with an increase in the *duration of the insurance contracts*. When the *bonus-malus coefficient* increases, the frequency of claims increases as well. The interpretation of Poisson and negative binomial models is the same. Based on the regressions coefficients, the profile of policyholders with the higher risk for the company can be established. This profile corresponds to policyholders from the beginner's age group, housewife, having insured a vehicle of type F, using it in professional purposes, not having a GPS device, with a malus of 150 applied to their premiums and being the client of the insurance company for one year.

The estimated frequency of claims represents one of the components of the insurance premium for those new clients of the insurance company who present the same characteristics that correspond to one of the policyholders' groups. In order to obtain the estimated value of claim frequency for these groups, we have to take into consideration that the link function for Poisson or negative binomial distribution is the logarithm function as presented in the methodology section of this paper. Considering the regression coefficients for NB2 model, the estimated value of claim frequency for the most risky policyholders' group is obtained by the following calculation:

$$\lambda_{riskiest\ class} = e^{-0.1644+0.0106*150-0.0272*1} = 4.0487$$

which represents the expected value of claim frequency for the clients who present the same characteristics as those with the riskiest profile for the insurance company.

In this paper, the used factors that differ from other similar studies, has a significant impact on the frequency of claims, with the exception of the vehicle's value. Taking into account the occupation of the insured, it can be noticed that there are significant differences between all 5 categories of occupation and the policyholders corresponding to the *housewife* group present the highest risk for the insurer. As consequence, the insurance company could exclude from the insurance portfolio the new clients that fall into this category. Another important factor is represented by the age of insured grouped into 5 categories as presented in Section 2.1. The results obtained show that the drivers from the *beginner* group present the highest level of risk for the company, as has been shown in other studies. Nevertheless, working with more years' intervals, in comparison with other empirical results, allows a more accurate differentiation of policyholders and more homogeneous groups of clients, so that the introduction of this variable in pricing analysis will not be considered a discouraging factor while choosing insurance services. In addition, the introduction of GPS as risk factor is significant for the determination and implementation of some protection measures that could be taken by the policyholders in order to prevent the accidents or could be included in the pricing policy of the insurance company. Concerning the value of vehicle, although it does not have a significant impact on the frequency of claims, the insurance company could take it into consideration while assessing the second component of insurance premium, the cost of claims.

Based on the 95% confidence intervals for the dispersion parameters of NB1 regression ($\alpha\epsilon(0.2534; 0.3114)$) and NB2 regression ($\alpha\epsilon(0.3484; 0.4420)$, it can be sustained that dispersion is significantly different from 0 and the application of the negative-binomial models is justified. Moreover, the NB2 model indicates a higher level of dispersion in comparison with the NB1 model, meaning that the first one could be considered more effective in correcting the overdispersion.

Models' goodness-of-fit

An essential step in the econometrical analysis represents the validation of models by comparing the calculated values with the observed ones. Examine the relationship between the expected and observed values, respectively (Table 6), the negative binomial models appear to be a substantial improvement over the Poisson model and this confirms the conclusion of the last paragraph that NB2 model provides the best fit to our insurance data.

Table no. 6 – Observed Claim Frequency versus Predicted

Claim Frequency	Observed	Model		
		Poisson	NB1	NB2
0	131838	134147.67	134225.58	134717.33
1	15395	14134.97	14010.41	13235.19
2	2202	1506.78	1532.11	1678.07
3	442	195.76	210.09	296.68
4	100	29.74	34.63	67.17
5	44	5	6.49	18.08

Source: Data processed within SAS 9.3

To conclude the comparisons between the analyzed count data models, Table 7 summarizes the results obtained for the goodness-of-fit tests.

Table no. 7 – Criteria for Assessing Goodness of Fit

Criterion	Model		
	Poisson	NB1	NB2
Log Likelihood	-55144.7684	-55050.4282	-54942.9562
AIC (smaller is better)	116020.9824	110138.8563	115619.3581
BIC (smaller is better)	116209.4344	110327.3084	115817.7287

Source: Data processed within SAS 9.3

The obtained values of the likelihood ratio test ($LL_{NB2-P} = 403.62$ and $LL_{NB1-P} = 188.68$) are greater than the theoretical one ($\chi^2_{2\alpha;1} = 2.706$) for both NB1 and NB2 models in comparison with Poisson regression.The resultsunderline that NB1 and NB2 models give a better fit of the data as opposed to Poisson regression. The remaining comparison between the negative binomial models indicates that NB2 model is preferred here to NB1 as it has higher log-likelihood ($LL_{NB2} = -54942.9562 > LL_{NB1} = -55050.4282$). The validity of these statements can also be confirmed by the information criteria. The lowest values of AIC and BIC comparative-fit tests are obtained for the NB2 model which underlinesthat this one is chosen to the detriment of both NB1 and Poisson models.

Eventually, to determine whether the data is better modeled using NB2, we considered the likelihood ratio test discussed in Section 2. The log likelihood for the full model is $LL_{NB2(7)} = -54942.9562$ and for the null model is $LL_{NB2(0)} = -62236.1707$. The likelihood ratio value obtained is $LR = 2(-54942.9562 + 62236.1707) = 14586.429$ and since the full model includes seven predictor variables, the statistics theoretical value is $\chi^2_{0.05,7} = 14.067$. This yields a *p-value* < 0.0001, highlighting once more that the NB2 is the best model to adjust the basic Poisson algorithm in order to estimate our insurance data.

4. CONCLUSIONS

An accurate insurance pricing system allows insurance companies to cover expected losses, expenses and make adequate the provision for contingencies. The first step in auto insurance pricing is the modeling of claim frequency, which represents an essential part for obtaining a reasonable and equitable insurance premium.

In this paper, it was considered an analysis of the classical and mixed count data models employed to estimate the frequency of claims made on vehicle insurance policies, focusing on the factors used to explain the insured risk. After a distinct analysis of insured's age variable, we obtained five categories of age depending on different years intervals in comparison with similar studies. This classification is used in the econometric modeling of insurance premiums.

After testing the equidispersion assumptions of Poisson distribution, both statistics presented in this paper reach the same conclusion, meaning the existence of overdispersion within the studied insurance portfolio. Results of these tests showed that NB models correct the overdispersion, providing a better fit to the data in comparison to the Poisson model. Furthermore, the comparison of NB1 and NB2 models indicated that the last one is preferred. By using the likelihood ratio in order to test the fit of the NB2 model, the results suggest that this model is the most appropriate to deal with the problem of overdispersion and to predict the claim frequency for the analyzed auto insurance portfolio.

While using Poisson and negative binomial models in the framework of GLMs, the risk factors that appeared to explain significantly the frequency of claims was the age-group and occupation of policyholders, the type, use and GPS device of vehicle, the bonus-malus coefficient and duration of the insurance policy. Based on the obtained results, we observed a decrease of claim frequency along with an increase of the insurance contracts duration, and also an increase of the frequency of claims along with the increase of bonus-malus coefficient. For these variables, there were obtained results which are similar with other actuarial studies and also consistent with the reality of the studied phenomenon.

The results obtained for the three variables introduced as risk factors indicates that the insured's occupation and GPS device appears to be significant, while the value of vehicle does not explain the frequency of claims. The modeling results could be considered as interesting sugestions for the insurance companies while implementing their pricing policy. Thus, the company could work with more age groups in order to evaluate the risk level of each insured and implicitly to calculate the insurance premium. The insured's occupation represents another valid factor that could be considered by the company in order to group the insurance portfolio in homogenous classes. Based on the GPS variable, the company could implement some precautionary measures, suggesting the new insured to use a GPS device. All this aspects aim at obtaining reasonable premium that corresponds to the risk level of each insured, and therby respecting the principle of equity in insurance.

Our empirical study could be useful to the policy-makers by allowing a better control on the insured risks and an accurate assessment of the insurance company liabilities leading to solvency and profitability.

References

Allain, E., and Brenac, T., 2012. Modèles linéaires généralisés appliqués à l'étude des nombres d'accidents sur des sites routiers: le modèle de Poisson et ses extensions. *Recherche Transports Sécurité, 72*, 3-18.

Antonio, K., Frees, E. W., and Valdez, E. A., 2012. A multilevel analysis of intercompany claim counts. *ASTIN Bulletin, 40*(1), 150-177.

Antonio, K., and Valdez, E. A., 2010. 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., and Guillen, M., 2007. Risk classification for claims counts - A comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal, 11*(4), 110-131.

Boucher, J. P., Denuit, M., and Guillen, M., 2008. Models of insurance claim counts with time dependence based on generalization of Poisson and Negative Binomial Distributions. *Advancing the Science of Risk Variance, 2*(1), 135-162.

Boucher, J. P., and Guillen, M., 2009. A survey on models for panel count data with applications to insurance. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales, 103*(2), 277-295.

Boucher, J. P., Perez-Marin, A. M., and Santolino, M., 2013. Pay-as-you-drive insurance: the effect of the kilometers on the risk of accident. *Anales del Instituto de Actuarios Españoles, 19*(3), 135-154.

Cameron, A. C., and Trivedi, P. K., 1986. Econometric models based on count data. Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics, 1*(1), 29-53.

Cameron, A. C., and Trivedi, P. K., 1990. Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics, 46*(3), 347-364.

Cameron, A. C., and Trivedi, P. K., 1998. *Regression Analysis of Count Data*. New York: Cambridge University Press.

Cameron, A. C., and Trivedi, P. K., 1999. Essentials of Count Data Regression (Chapter 15). In B. B.H. (Ed.), *A Companion to Theoretical Econometrics*. Malden, MA: Blackwell Publishing Ltd. .

Charpentier, A., and Denuit, M., 2005. *Tome II: Tarification et provisionnement*. Paris: Economica.

Denuit, M., and Lang, S., 2004. Nonlife ratemaking with bayesian GAM's. *Insurance: Mathematics and Economics, 35*(3), 627-647.

Denuit, M., Maréchal, X., Pitrebois, S., and Walhin, J. F., 2007. *Modeling of claim counts. Risk Classification, Credibility and Bonus-Malus Systems*. Chichester: Wiley.

Dionne, G., and Vanasse, C., 1989. A generalization of auto insurance rating models: the negative binomial distribution with a regression component. *ASTIN Bulletin, 19*(2), 199-212.

Dionne, G., and Vanasse, C., 1992. Auto insurance ratemaking in the presence of asymmetrical information. *Journal of Applied Econometrics, 7*(2), 149-165.

Gourieroux, C., and Jasiak, J., 2001. Dynamic Factor Models. *Econometric Reviews, Taylor & Francis Journals, 20*(4), 385-424.

Gourieroux, C., and Jasiak, J., 2004. Heterogeneous INAR(1) model with application to car insurance. *Insurance: Mathematics and Economics, 34*(2), 177-192.

Gourieroux, C., Monfort, A., and Trognon, A., 1984a. Pseudo Maximum Likelihood Methods: Theory. *Econometrica, 52*(3), 681-700.

Gourieroux, C., Monfort, A., and Trognon, A., 1984b. Pseudo Maximum Likelihood Methods: Applications to Poisson Models. *Econometrica, 52*(3), 701-720.

Greene, W. H., 2002. *Econometric Analysis*. New Jersey: Prentice Hall.

Greene, W. H., 2008. Functional forms for the negative binomial model for count data. *Economics Letters, 99*(3), 585-590.

Greenwood, M., and Yule, G. U., 1920. An inquiry in to the nature of frequency distributions of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents. *Journal of the Royal Statistical Society A, 83*, 255-279.

Gurmu, S., 1991. Tests for detecting overdispersion in the positive Poisson regression model. *Journal of Business and Economic Statistics, 9*(2), 215-222.

Hausman, J., Hall, B., and Griliches, Z., 1984. Economic models for count data with an application to the patents - R&D relationship. *Econometrica, 52*(4), 909-938.

Hilbe, J. M., 2007. *Negative Binomial Regression*. New York: Cambridge University Press.

Hilbe, J. M., 2014. *Modeling Count Data*. New York: Cambridge University Press.

Jong, P., and Heller, G. Z., 2013. *Generalized Linear Models for Insurance Data* (5th ed.). New York: Cambridge University Press.

Jorgensen, B., 1997. *The Theory of Dispersion Models*. London: Chapman and Hall.

Kouki, M., 2007. *Conducteurs novices et conducteurs expérimentés: Approche économétrique sur la sinistralité et la couverture d'assurance*. Working Paper.

Kuha, J., 2004. AIC and BIC comparisons of assumptions and performance. *Sociological Methods and Research, 33*, 188-229.

Lawless, J. F., 1987. Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics, 15*(3), 209-225.

McCullagh, P., and Nelder, J. A., 1989. *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Nelder, J. A., and Wedderburn, R. W. M., 1972. Generalized linear interactive models. *Journal of the Royal Statistical Society A, 135*(3), 370-384.

Vasechko, O. A., Grun-Réhomme, M., and Benlagha, N., 2009. Modélisation de la fréquence des sinistres en assurance auto. *Bulletin Français d'Actuariat, 9*(18), 41-63.

Winkelmann, R., 2004. Co-payments for prescription drugs and the demand for doctor visits - Evidence from a natural experiment. *Health Economics, 13*(11), 1081-1089.

Yip, K., and Yau, K., 2005. On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics, 36*(2), 153-163.