Assurances et gestion des risques
Insurance and Risk Management

Zero-inflated and over-dispersed data models: Empirical evidence from insurance claim frequencies

Imen Karaa et Habib Chabchoub

Volume 84, numéro 3-4, décembre 2017

URI : https://id.erudit.org/iderudit/1043358ar
DOI : https://doi.org/10.7202/1043358ar

Aller au sommaire du numéro

Éditeur(s)

Faculté des sciences de l'administration, Université Laval

ISSN

1705-7299 (imprimé)
2371-4913 (numérique)

Découvrir la revue

Citer cet article

Karaa, I. & Chabchoub, H. (2017). Zero-inflated and over-dispersed data models: Empirical evidence from insurance claim frequencies. *Assurances et gestion des risques / Insurance and Risk Management*, *84*(3-4), 103–128. https://doi.org/10.7202/1043358ar

Résumé de l'article

The main objective of this paper is to model automobile claim frequency by using standard count regression and zero-inflated regression models. The use of the latter model is mainly motivated by its ability to handle the over dispersion and zero-inflation phenomenon. The sample data consist of claims data obtained from one randomly selected automobile insurance company in Tunisia for a single year, 2009, containing beginning drivers and drivers who have had a license for less than three years. Our estimation results show that many exogenous variables can explain the frequency of claims; they are not taken into account in calculating the basic insurance premium. Moreover, the ZI binomial negative regression outperforms the standard count models and the ZI Poisson model in handling zero-inflated and additional over dispersed claim count data.

ZERO-INFLATED AND OVER-DISPERSED DATA MODELS: EMPIRICAL EVIDENCE FROM INSURANCE CLAIM FREQUENCIES

Imen KARAA[1], Habib CHABCHOUB[2]

ABSTRACT

The main objective of this paper is to model automobile claim frequency by using standard count regression and zero-inflated regression models. The use of the latter model is mainly motivated by its ability to handle the over dispersion and zero-inflation phenomenon. The sample data consist of claims data obtained from one randomly selected automobile insurance company in Tunisia for a single year, 2009, containing beginning drivers and drivers who have had a license for less than three years. Our estimation results show that many exogenous variables can explain the frequency of claims; they are not taken into account in calculating the basic insurance premium. Moreover, the ZI binomial negative regression outperforms the standard count models and the ZI Poisson model in handling zero-inflated and additional over dispersed claim count data.

JEL Classification:

Keywords: Claim frequency, over dispersion, zero inflated, Poisson, negative binomial.

1. Introduction

In Tunisia, the rapid increase of Automobile Parks and the number of traffic claims have a significant impact on society. For example, in 2012[3], there were 9,351 traffic claims in Tunisia and 15,767 victims (1,623 fatalities and 14,144 injuries) for only eleven million inhabitants. Therefore, on average 4.5 persons are killed every day, and there are approximately 17 people killed for every 100 reported claims. This latter number is very high compared to deaths in France, where six were killed for 100 reported claims in 2012. The National Traffic Observatory in Tunisia explains that the main causes of road claims are speeding and drunk driving.

There is no quick fix, but it is necessary to act to save lives. All stakeholders (government, monitors, insurers, car manufacturers, parents and concerned individuals) must work for the same purpose. The goal is to continue to reduce the frequency of road traffic claims. In this paper, we are interested in the positioning of the insurers dealing with claims. Insurers can play a role in road safety by demanding higher premiums for high-risk drivers, thus encouraging them to drive more cautiously.

Vehicle insurance is the most prevalent insurance line in Tunisia; it made up 45.7% of the whole market in 2012. In addition, it is the largest sector in non-life insurance, with 54.84% of the market. However, this sector produced a technical deficit of 2.286 billion TND (TND = Tunisian Dinars, 1 TND = 0.49 USD in 2016) in the same year. For this reason, our aim in this study is to analyze the factors that influence the frequency of claims. Two major factors usually play an important role in insurance vehicle claims. The first is related to the policyholders' characteristics, and the second relates to vehicle characteristics.

The usual starting point when modeling the number of claims is to use the Poisson regression model. This is characterized by the equality of mean and variance. However, this equality is rarely confirmed in practice. Therefore, the data are normally over-dispersed, i.e., the variance is much greater than the mean, which means the use of the Poisson regression is inappropriate. This leads to the idea that a distribution with over dispersion is more suitable for describing discrete events such as the negative binomial (NB) regression model, which is a Poisson-gamma mixture. A more detailed description of this model analysis in a cross-sectional framework can be found in Washington et al. (2003). The NB model has been applied in a panel model; see e.g., Boucher and Guillén (2011).

A second important feature of claims data, largely validated in empirical studies, is the large proportion of zeros. In that case, the zero-inflated (ZI) model or hurdle model are more appropriate. Among the different ZI models proposed in the statistical analysis, we found the zero-inflated Poisson model (ZIP) and the negative binomial zero-inflated (ZINB) model. The ZIP model has been discussed with cross-sectional data (see Yip and Yan 2005; Yang *et al.* 2007).

The ZIP model was applied in different fields. As examples, in the finance area, Mouatassim *et al.* (2012) fitted Poisson and ZIP regression models to operational risk frequency. In the manufacturing area, Lambert (1992) analyzed defects in manufacturing equipment using simulations. He concluded that ZIP models are difficult to interpret.

In the insurance area, Mouatassim and Ezzahid (2012) fitted Poisson regression and ZIP regression to Moroccan private health insurance count data. Based on the test of Vuong and the probability integral transforms (PIT), the authors concluded that ZIP regression fits excess of zeros count data better than a standard Poisson regression. In addition, Benlagha *et al.* (2012) fitted the same models as Mouatassim and Ezzahid (2012) to a Tunisian automobile insurance market. The authors observed that the ZIP model provides a better fit than the Poisson model.

If the data are over dispersed and exhibit an excess of zeros, the ZINB model is commonly used as an alternative to the ZIP model (see Yau *et al.* 2003). These models assume that the counts arise from a mixed model where one component is a point mass at zero and the other follows a Poisson or NB distribution. Without confusion, over dispersion can be the result of excess zeros or some other cause.

Using data from a Spanish automobile insurer, Melgar *et al.* (2005, 2006) concluded that the ZINB model is the best model to use to describe the underling claim frequency distribution. Such a model has also been employed by Vasechko *et al.* (2009). Like Melgar *et al.* (2005), the authors concluded that the ZINB model was better than the ZIP and standard count models at analyzing a portfolio of claims from a French insurer.

Yip and Yau (2005) applied various zero-inflated models to motor insurance claim frequency data. Their data were taken from the SAS Enterpriser Miner database in 1998. Their results show that the application of zero-inflated count data models provide a good fit to the data compared to the standard count model.

This model was also applied in public health by Ismail and Zamani (2013). The authors analyzed German healthcare count data. For general overview of count data, we refer the reader to Hilbe (2015).

The ZI models have been extended to panel data by Denuit *et al.* (2007) and in Boucher *et al.* (2009) when they give an analysis of the dichotomy between the number of claims. For general overview of count panel data, we refer the reader to Cameron and Triverdi (2015). These models have also been applied in diverse disciplines such as demography, sociology, psychology, health economics, manufacturing and engineering. Another related study on a copula approach for insurance claim numbers with excess zeros and time-dependence can be found in Zhao and Zhou (2012).

The main contribution of this paper is to examine factors explaining the claim frequency declared by the policyholder to his insurance company. Modeling and determining factors that explain the number of claims is very interesting for actuaries, insurance companies, policymakers, researchers and regulators for several reasons: (1) It helps in understanding the behavior of the policyholders, (2) it can be useful for solvency purposes, (3) it allows to insurers to improve the priori pricing of coverage, and finally, it allows the identification of the claims process.

In actuarial sciences, the insurer seeks to determine the factors that help to explain the claims. By building risk classes, these factors allow the insurer to segment its portfolio and prioritize these classes, with help of the pure premium.

Motivated by modeling the claim frequency, our study differs from existing studies in two points. First, it investigates insurance vehicle claims in the context of a developing country, Tunisia. To the best of our knowledge, this is the first paper that handles over dispersion and the zero-inflation phenomenon in Tunisia, a country characterized by a high number of claims and high percentages of death per 100 claims. Second, it focuses on policyholders with four-wheel cars who are characterized by a lack of experience driving, namely, beginning drivers. Our strategy is simple and consists of two steps. In the first step, we use the classical discrete distribution to model the characteristics of policyholders and the vehicles. In the second step, we employ the non-classical discrete distribution. Comparison between the two approaches is made using the likelihood ratio test, the information criteria and the test of Vuong.

Our main findings can be summarized as follows: first, our estimation results show that there are many exogenous variables that can explain the frequency of claims; they are not taken into account in calculating the basic insurance premium. Second, through the information criteria and the Vuong test, we find that the zero-inflated negative binomial model exhibits the best fit for our data and for handling zero-inflated and additional over dispersed claim count data.

The remainder of this paper is organized as follows. The next section discusses the different models used in this paper and the tests for the comparison of the models. Section 3 describes the sample and the variables employed. Section 4 presents the empirical results. Section 5 states the key findings and the limitations of the study.

2. ECONOMETRIC MODELS

This section is devoted to describing the Poisson, negative binomial, zero inflated Poisson, and zero inflated negative binomial models as well as the different tests used to select between different models.

2.1. The standard count models

The Poisson regression model is defined by the following equation:

$$P(Y = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

where $P(y_i)$ is the probability of claims occurring over one year or one period.

λ_i is the expected insurance claim frequency.

The mean and the variance of Poisson regression model are:

$$\mathrm{E}(y_i) = V(y_i) = \lambda_i = \exp(X_i^{'}\beta)$$

$X_i^{'} = X_{i1}, X_{i2}, \ldots.., X_{ip}$ contains the exogenous variables (characteristics of drivers and their vehicles).

$\beta^{'} = \beta_1, \beta_2, \ldots., \beta_p$ is the vector of estimable coefficients. In the automobile insurance claim context, y_i is the number of at-fault-claims reported by driver i to his insurer.

The log-likelihood (L) of the Poisson regression model is given by:

$$\ln L_{Poisson} = \sum_{i=1}^{n} \ln \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \sum_{i=1}^{n} \left[y_i \ln \lambda_i - \ln(y_i!) - \lambda_i \right]$$

This model was rejected because the mean and the variance of the dependent variables are different, indicating substantial over dispersion in the data. Such over dispersion suggests a negative binomial model that we describe in the next paragraph.

In the negative binomial regression model, the probability of having y_i claims is given by:

$$P(Y = y_i) = \frac{\Gamma\left(y_i + \alpha^{-1}\right)}{y_i!\,\Gamma\left(\alpha^{-1}\right)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}} \left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i}\right)^{y_i}$$

The mean and the variance of NB regression model are:

$$E(y_i) = \lambda_i \text{ and } V(y_i) = \lambda_i + \alpha\lambda_i^2,$$

α denotes the over dispersion parameter, if $\alpha = 0$, the negative parameter collapses to the Poisson model. $\hat{\varepsilon}_i$ is a random error that has a gamma distribution.

It should be noted that the parameter α is the over dispersion parameter in the count models. The rejection of the null hypothesis, $\ln(\alpha) = 0$, can be interpreted as non-suitability of the Poisson model for our data set. To decide which subset of independent variables should be included in the insurance claim estimation model, information criteria was used.

2.2. The Zero Inflated Models

As previously discussed, the Poisson regression model is more suitable for data not characterized by a large proportion of zeros. However, as insurance claims data is generally characterized by a large proportion of zeros, the zero-inflated models of Lambert (1992) are supposed to be more suitable of our data compared to standard count models. The ZI model includes two parts. The first part corresponds to the standard count model (for Y^*, to estimate the number of claims when the insured is in a situation of declaration). The second part relates to the zeros inflation (Logit) that explains the probability of non-declaration.

In a situation of non-declaration ($y_i = 0$), the sample can be composed of two types of policyholders. The first refers to drivers who make the decision to report claims when they happen. The value zero shows that the policyholder did not have a claim during the insurance policy period.

The second type refers to drivers who do not report a claim to the company. It is common knowledge that sometimes drivers do not report small collisions when they want to avoid being penalized with the bonus-malus coefficient or when the cost of the claim is less than the deductible.

The zero-inflated models are generally used to distinguish between different types of policyholders. The dependent stochastic variable Y is the mixture of a binary distribution and a standard count distribution.

$Y = Z \times Y^*$ with Z is modeled using a logistic regression to estimate $P(y_i = 0)$.

$$\text{For policyholder } i, \begin{cases} z_i = 0, \text{ if the driver did not declare} \\ z_i = 1, \text{ in the opposite case.} \end{cases}$$

Y* corresponds to the standard count model (Poisson distribution or negative binomial distribution). It used to predict the value of Y for the drivers who reported a claim ($z_i = 1$). This equation estimates the mean of y_i.

For the zero-inflated Poisson model, we denote q_i, the probability of $z_i = 0$ (no reported claim) and λ_i, the Poisson parameter for the claim's frequencies that depends, as previously, on the explanatory variables. Then the probability distribution Y is as follows:

$$P(Y = y_i) = \begin{cases} q_i + (1 - q_i)e^{-\lambda_i}, & y_i = 0; \\ (1 - q_i)e^{-\lambda_i} \dfrac{\lambda_i^{y_i}}{y_i\,!}, & y_i > 0. \end{cases}$$

where $q_i = \dfrac{e^{X_i^{'}\beta}}{1 + e^{X_i^{'}\beta}}$ with i=0,1,2...

The probability's number of claims conditionally to $z_i = 1$ is equal to the unconditional probability of the unobserved variable y_i^*.

The mean and the variance of ZIP regression model are:

$$E(y_i) = (1 - q_i)\lambda_i,$$

$$V(y_i) = (1 - q_i)(\lambda_i + q_i\lambda_i^2)$$

One can clearly observe that the ZIP regression model reduces to the Poisson model when $q_i = 0$.

The log-likelihood of the ZIP regression model is given by:

$$\ln L_{ZIP} = \sum_{y_i=0} \ln\left(e^{\alpha} + e^{-\lambda_i}\right) + \sum_{y_i>0} \left[y_i \ln(\lambda_i) - \ln(y_i\,!) - \lambda_i\right] - n\ln(1 + e^{\alpha})$$

An alternative to modeling excess zero and over dispersion in count data is to start from a ZIP regression model and add a multiplicative random effect to represent unobserved heterogeneity. This leads to the ZINB regression model.

The probability function of the ZINB distribution can be formulated as follows:

$$P(Y = y_i) = \begin{cases} q_i + (1 - q_i)\left(\dfrac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}}, & y_i = 0; \\ (1 - q_i)\dfrac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\dfrac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}}\left(\dfrac{\lambda_i}{\alpha^{-1} + \lambda_i}\right)^{y_i}, & y_i > 0. \end{cases}$$

The mean and the variance of the ZINB regression model are:

$$E(y_i) = \lambda_i(1 - q_i),$$

$$V(y_i) = \lambda_i(1 - q_i)(1 + q_i\lambda_i + \alpha\lambda_i) = E(y_i)(1 + q_i\lambda_i + \alpha\lambda_i)$$

The log-likelihood can be written as:

$$\ln L_{ZINB} = \sum_{y_i=0} \ln\left(q_i + (1 - q_i)t_i^{\alpha^{-1}}\right) + \sum_{y_i>0}\left[\ln(1 - q_i) + \ln\left(\frac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\right) + \alpha^{-1}\ln(t_i) + y_i\ln(1 - t_i)\right]$$

where $t_i = \dfrac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}$

The zero counts are captured by the first process with probability q_i and the second process with $(1 - q_i)$. The standard count and the ZI models are related to each other. If $\alpha = 0$, the ZINB model will be reduced to the ZIP model. If $\alpha = 0$ and $q_i = 0$, the ZINB model will be reduced to the standard Poisson regression model.

2.3. Tests for the comparison of the models

The commonly used approach to test between different models is to use statistical tests and information criteria. In this paper, to compare between pairs of models, we use the log-likelihood test (LR) and two information criteria, the Akaike information criteria (AIC) and the Bayesian information criteria (BIC), for full and nested models. These information criteria are functions of parameters of the models and the size of the sample. Finally, the Vuong statistics test for non-nested models was applied.

Likelihood ratio test:

The likelihood ratio test can be performed for testing over dispersion in the Poisson versus NB regression models and over dispersion in ZIP versus ZINB regression models, $H_0{:}\alpha = 0$ *vs.* $H_1{:}\alpha > 1$.

Since the null hypothesis is on the boundary of parameter space, the asymptotic distribution for the LRT statistic is a mixture of half of probability mass at zero and half of chi square with one degree of freedom (Lawless, 1987, Stram and Lee 1994, 1995).

Information criteria:

The Akaike information criterion (AIC) is defined as follows:

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

The Bayesian information criterion (BIC) has the calculated value as follows:

$$BIC = -2\log(L) + 2\log(n)k$$

where k represents the number of parameters of the model and n is the size of the portfolio. The rule for selecting the better model using the information criteria is to select the model with the lower AIC and/or BIC value.

Vuong (1989) test:

Vuong (1989) developed some general tests of non-nested models. He developed one to compare two models. Then, Greene (1994) adapted this test to the ZIP cases versus Poisson and ZINB versus NB models.

Vuong's (1989) test fits to the same data by maximum likelihood. The test statistic is simply the average log-likelihood ratio suitably normalized.

The Vuong statistic model is $V = \dfrac{\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} m_i)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2}} = \dfrac{\sqrt{n}\bar{m}}{s_m}$,

where n is the total number of observations, $\bar{m}$ is the mean $\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$

and s_m^2 is the variance of the variable m_i , $s_m^2 = \frac{1}{n-1}\sum_{i=1}^{n}(m_i - \bar{m})^2$.

$$m_i = \log\left[\frac{f_1(y_i)}{f_2(y_i)}\right] = \log\left[\frac{\Pr(y_i - ZIP)}{\Pr(y_i - Poisson)}\right]$$

The hypotheses of the Vuong test, in selecting between the ZIP and Poisson models and ZINB and NB models are:

$$H_0 : E\left[m_i\right] = 0$$
$$H_1 : E\left[m_i\right] \neq 0$$

As described in Vuong (1989), this statistic has a standard normal distribution and is directional. The null hypothesis of the test is that the two non-nested models are equivalent. With large positive values of V, the ZIP/ZINB model is statistically better than the Poisson/NB model. However, with large negative values, the nonzero-inflated version of the Poisson/NB model exhibits the best fit to the data. If absolute values are close to zero, the test favors neither model. The test requires an estimation of both models and the computation of the sample of predicted probabilities.

3. Data description

The data set used in this study consists of the claim frequency reported by a policyholder to his insurer in the policy year; in our case in 2009. Our database is from one of the top five companies (out of 24) operating in the Tunisian insurance market. Indeed, it comes from a subsample of a study by Karaa and Benlagha (2015)[4], where the source of the data was also revealed. Only drivers who had a driver's license for less than three years, i.e., "beginner drivers", have been considered in this database, as well as drivers who have a bonus malus class higher than eight or seven for private use vehicles and rest of use vehicles, respectively[5]. In addition, only policyholders who insure their four-wheeled cars have been considered in our study.

The data contain 15053 observations. The dummy exogenous variables used in our analysis are presented in Table 1. This table also reports basic descriptive statistics in terms of mean and standard deviation.

In our analysis, we consider only two types of coverage as a signal to discriminate between policyholders in terms of unobservable risk, namely, third party coverage versus comprehensive coverage rather than partial insurance versus full insurance.

The drivers who subscribe to comprehensive coverage insurance were approximately 73.4% of the whole sample, while only 26.6% had third party coverage, as shown in Table 2.

The portfolio contains 11,689 men (77.7%) and 3,355 women (22.3%), as indicated in Table 2. Policyholders aged between 31 and 55 years old were approximately 68.8% of the whole portfolio, while only 16.5% of the drivers were younger than 30 and 22% of drivers were older than 75 years old. Approximately 30% live on the coast, 28% live in greater Tunis, 12.9% live in southern Tunisia, and only 5.4% live in the center. Looking at the characteristics of vehicles, 49.9% of the cars are for private use and 50.1% are for the rest of use. In addition, it is seen that 58.8% of the vehicles are made in France, while less than 1% for the vehicles are made in the USA, England or elsewhere.

TABLE 1 *Summary of dummy exogenous variables*

VARIABLE	DESCRIPTION	MEAN	STD DEV
Coverage	= 1 for comprehensive coverage and 0 otherwise.	0.734	0.442
Gender	= 1 for male and 0 for female.	0.777	0.416
Use	= 1 for a rest of use car and 0 for private use.	0.501	0.500
Age 1	= 1 for drivers aged >22 years and 0 otherwise.	0.004	0.253
Age 2	= 1 for drivers aged 22-30 years and 0 otherwise.	0.161	0.368
Age 3	= 1 for drivers aged 31-55 years and 0 otherwise.	0.688	0.463
Age 4	= 1 for drivers aged 56-75 years and 0 otherwise.	0.125	0.331
Age 5	= 1 for drivers aged <75 years and 0 otherwise.	0.022	0.146
Job 1	= 1 for an official and 0 otherwise.	0.214	0.409

[…]

VARIABLE	DESCRIPTION	MEAN	STD DEV
Job 2	= 1 for a senior executive and 0 otherwise.	0.028	0.166
Job 3	= 1 for a middle manager and 0 otherwise.	0.020	0.142
Job 4	= 1 for a crafts man and 0 otherwise.	0114	0.317
Job 5	= 1 for an employee and 0 otherwise.	0.490	0.499
Job 6	= 1 for a retired person and 0 otherwise.	0.053	0.224
Job 7	= 1 for an unemployed person and 0 otherwise.	0.080	0.272
Brand 1	= 1 for a car made in France and 0 otherwise	0.588	0.492
Brand 2	= 1 for a car made in Italy and 0 otherwise.	0.129	0.335
Brand 3	= 1 for a car made in Germany and 0 otherwise.	0.177	0.382
Brand 4	= 1 for a car made in China, Korea or Japan and 0 otherwise.	0.093	0.289
Brand 5	= 1 for a car made in the USA and 0 otherwise.	0.004	0.063
Brand 6	= 1 for a car made in England and 0 otherwise.	0.003	0.052
Brand 7	= 1 for a car made in other countries and0 otherwise	0.006	0.079
Zone 1	= 1 if the policyholder lives in greater Tunis and 0 otherwise.	0.282	0.449
Zone 2	= 1 if the policyholder lives in the North and 0 otherwise.	0.127	0.333
Zone 3	= 1 if the policyholder lives in the Northwest and 0 otherwise.	0.109	0.312
Zone 4	= 1 if the policyholder lives in the Coast and 0 otherwise.	0.299	0.458
Zone 5	= 1 if the policyholder lives in the Center and 0 otherwise.	0.054	0.226
Zone 6	= 1 if the policyholder lives in the South and 0 otherwise.	0.129	0.335

22% of drivers were older than 75 years old. Approximately 30% live on the coast, 28% live in greater Tunis, 12.9% live in southern Tunisia, and only 5.4% live in the center. Looking at the characteristics of vehicles, 49.9% of the cars are for private use and 50.1% are for the rest of use. In addition, it is seen that 58.8% of the vehicles are made in France, while less than 1% for the vehicles are made in the USA, England or elsewhere.

From Table 2, the average premium is 274 TND and the maximum premium is 3,687 TND. The average indemnity paid by the company to his policyholder is 132 TND, and the maximum indemnity is 48,669 TND. The number of claims stretches from no claims up to eight claims, where the mean value is 0.181, indicating a low frequency of claims.

TABLE 2 *Summary of continuous and discrete variables*

	MEAN	STD. DEV.	MIN	MAX
Premium	274.698	161.889	0	3,686.574
Indemnity	132.220	1,211.336	0	48,669.47
Number of claim	0.181	0.593	0	8

Table 3 shows that 29.8% of drivers with at least one claim are women, while 70.2% are men. However, 21% of women had no claim during the insurance contract year, while 79% of men had no claim. In addition, it shows that 50.2% of vehicles declared no at-fault-claim in the period under study for private use and 49.8% declared no claim for the rest of use.

TABLE 3 *Distribution of claim occurrence by gender and the use of vehicle*

	GENDER		TOTAL
	MALE	FEMALE	
Y = 0	10,553 (78.62)	2,870 (21.38)	13,423 (100)
Y = 1	1,145 (70.24)	485 (29.76)	1,630 (100)
Total	11,698 (77.71)	3,355 (22.28)	15,053

	USE		TOTAL
	PRIVATE	REST	
Y = 0	6,742 (50.23)	6,681 (49.77)	13,423 (100)
Y = 1	769 (47.18)	861 (52.82)	1,630 (100)
Total	7,511 (49.89)	7,542 (50.11)	15 053

Note: Percentages are in parentheses.

Table 4 provides a summary of the frequency statistics for our variable of interest: claim frequencies reported for the year. In Table 4, our data exhibit a high number of zero values, which means we have a very large number of drivers who did not declare any at-fault claim (almost 90% of the whole portfolio). We note also that our data includes 822 policyholders who declared one claim during the contract year, 620 insured who declared two claims and 187 who claimed more than two claims. No drivers in our data incurred more than eight claims in 2009.

TABLE 4 *Goodness of fit of the marginal models*

Y_i	OBSERVED		FITTED	
	FREQUENCY	PERCENTAGE	POISSON	NB
0	13,424	89.18	12,561.250	13,406.633
1	822	5.46	2,273.091	1,052.979
2	620	4.12	205.670	339.335
3	114	0.76	12.406	136.922
4	56	0.37	0.561	60.810
5	10	0.07	0.020	28.489
6	5	0.03	0.006	13.809
7	0	0.00	0.000	6.854
8	2	0.01	0.000	3.461
λ	–	–	0.181	0.139
mu	–	–	–	0.181
χ^2	–	–	2,776.944	290.726

15053 observations

In addition to the descriptive analysis previously presented, we test data using the Chi-Squared test, of which of the NB model versus Poisson model can better fit and describe the claims frequencies. The statistic of the Chi-Squared test is given by

$$\chi^2_{k-1} = \sum_{i=1}^{n} \frac{\left(\mathrm{Y_i} - \hat{\mu}_i\right)^2}{Var(\mathrm{Y_i})} = \sum_{i=1}^{n} \frac{(\text{observed-fitted})^2}{\textit{fitted}}$$

where n is the total number of observations and k is the number of categories (in our case, degrees of freedom = 8). Observed is observed frequency and fitted is the fitted claims frequency. The Chi-Squared statistic has degrees of freedom equal to k-1. Note that the critical value for the Chi-Squared statistic at the 5% level is approximately 15.51.

The goodness of fit results for the marginal models are reported in Table 4 below, where we compare the observed and the corresponding fitted claim frequencies. For example, 13,424 (89.18%) persons are observed to have no claims over the year under study. The claim frequencies predicted by the two candidate models, Poisson and NB, are

12,561 and 13,406, respectively. Because of the subject heterogeneity introduced by covariates, the fitted frequency is calculated as the summation of the marginal probability of each policyholder in the sample. The smaller distance between observed and fitted claims suggests a better fit.

From Table 4, it is apparent that the negative binomial distribution fit the data better than the Poisson regression when we compared the fits by observed claim frequencies. This result is consistent with Yip and Yau (2005). However, the two statistics (2,777 and 290.7) are both much larger than the critical value for the χ^2. Therefore, the observed data compared to the negative binomial regression shows an excess probability of zero claims and a significantly lower probability at a count of one.

FIGURE 1 *The insurance frequency claim data fit with Poisson and Negative binomial model*

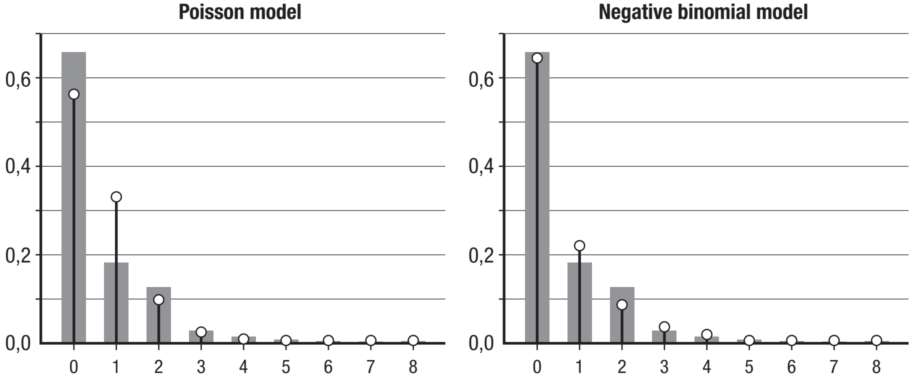

Figure 1 shows the bar chart for the claim frequency, which is fitted by the Poisson model (the graphic on the left). The bar chart on the right presents data fitted by the negative model. As discussed previously within Table 4, it can be seen that the Negative binomial model provides a much better fit to the data.

4. Results and Discussion

The estimation results of the count standard models for the Tunisian database are reported in Table 5. The results of the Poisson model are reported in columns 2-3 and the results of the negative binomial model in columns 4-5.

The results of the Poisson model indicate that the coefficient relative to the variable “insurance coverage” is positive and significant, which means that the drivers who purchased full insurance have a higher probability of claims than those who subscribed to partial coverage. Hence, this result can be interpreted by the existence of asymmetric information. This finding is consistent with that of Karaa and Benlagha (2015) when they used the bivariate probit model to test for residual asymmetric information.

As expected, the estimated coefficient for Gender is negative and statistically significant at the 1 percent level. This result indicates that men have a lower probability of at-fault claims compared to women. Our empirical results also show that the type of vehicle use variable positively determines the at-fault claim at the 5 percent level of significance. This means that the rest of use vehicles have a higher probability of an at-fault claim compared to private use.

We found also evidence on the fact that the origin of the automobile has a significant impact on the probability of at-fault-claim. In particular, we found that the German and China/Korea/Japan brand vehicles increase the probability of an at-fault claim for both Poisson and negative binomial regressions. However, we found that for the Poisson regression the USA brand vehicles decreased the probability of an at-fault claim.

Regarding the variable zone of residence, the results show that the claim frequency is positively related with policyholders who live in greater Tunis, the northwest and the coast. Moreover, the results show that the job of policyholder variables had a positive and significant coefficient, except for employees and retired people, who had a negative coefficient. Moreover, our results show that the parameter α of over dispersion is positive and highly significant (t-stat= 4.470, p-value=0.000).

TABLE 5 *Poisson regression and negative binomial regression results*

INDEPENDENT VARIABLES	POISSON		NB	
	COEFFICIENT	STD. ERR.	COEFFICIENT	STD. ERR.
Coverage	0.397	0.049***	0.433	0.068***
Gender	–0.257	0.042***	–0.251	0.064***
Use	0.122	0.038***	0.129	0.056**
Age 2	–0.117	0.306	–0.087	0.460
Age 3	–0.152	0.303	–0.116	0.457
Age 4	–0.878	0.313***	–0.782	0.466*
Age 5	1.046	0.309***	1.280	0.477***
Job 1	0.237	0.070***	0.392	0.107***
Job 2	0.415	0.100***	0.472	0.164***
Job 3	0.319	0.119***	0.425	0.188**
Job 4	-0.090	0.085	–0.033	0.124
Job 5	-0.492	0.071***	–0.437	0.103***
Job 6	-0.449	0.113***	–0.489	0.162***
Brand 1	-0.039	0.269	0.014	0.371
Brand 2	-0.036	0.274	0.031	0.377
Brand 3	0.469	0.270*	0.619	0.373*
Brand 4	–0.657	0.284**	–0.556	0.386
Brand 5	1.055	0.317***	1.177	0.498**
Brand 6	0.369	0.386	0.602	0.591
Zone 1	0.872	0.079***	0.860	0.102***
Zone 2	0.134	0.099	0.063	0.126
Zone 3	0.498	0.093***	0.448	0.123***
Zone 4	0.325	0.083***	0.279	0.106***
Zone 5	0.074	0.126	0.037	0.161
Constant	–2.146	0.417***	–2.361	0.607***
Over dispersion	–	–	4.470	0.236***
Log Likelihood	-7475.1951		-6660.5844	
LR chi2(24)	1638.40		781.53	
Prob> chi2	0.000		0.0000	

Note: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Turning now to the comparison of these two models, the log-likelihood ratio statistics are reported in the last two rows of Table 5. For both models, LR statistics are much greater than chi-squared statistics (24), which are equal to 36.42. However, we show that the introduction of an over dispersion parameter (alpha term) improves the fit compared with the Poisson regression model. Therefore, it is concluded that the negative binomial model is preferred to the Poisson model.

The estimation results of the marginal effects of the Poisson regression and the negative binomial regression models are provided in Table 6. The significant exogenous variables of the two count models are relatively analogous. Moreover, the marginal effects estimates of the two models are similarly quite comparable. The results, which are provided in columns 1 and 2 for this table, shows that the probability of having a claim increases by 10% when we consider drivers subscribed to comprehensive coverage rather than who chose third-party coverage. In addition, the probability of having a claim was 17% higher for the rest of use vehicles.

Concerning the job variable, the probability of having an at-fault claim increased by 35%, 69% and 51% for official, senior executive and middle manager policyholders, respectively. However, it decreased by 67% and 50% for employees and retirees, respectively.

Table 7 provides the estimation results of the zero-inflated Poisson model and the zero-inflated negative binomial model. The ZI models are two parts of the regression; one corresponds to the standard count model of the claim frequency, and the other relates to the logistic model. The estimation results in Table 7 are approximately similar to those of Table 5 regarding the claim frequency equation.

The first part, which relates to the standard count regression (Poisson or negative binomial), indicates that the coverage choice variable is statistically significant at the one percent level. This variable is positively associated with the claim frequency in both models. Concerning the use of vehicles variable, it is significant at the five percent level in the ZINB model, while it is insignificant in the ZIP model. This variable has a positive value of 0.132 (with a standard error of 0.055). Therefore, the claim frequency rises with drivers who have vehicles for the rest of use.

TABLE 6 *Marginal effects of the standard count models*

INDEPENDENT VARIABLES	POISSON		NB	
	dy/dx	X	dy/dx	X
Coverage	0.010***	0.734	0.053***	0.734
Gender	–0.038***	0.777	–0.036***	0.777
Use	0.017***	0.501	0.017**	0.501
Age 2	–0.015	0.161	–0.011	0.161
Age 3	–0.021	0.688	–0.0146	0.688
Age 4	–0.089***	0.125	–0.079**	0.125
Age 5	0.246**	0.022	0.337	0.022
Job 1	0.035***	0.214	0.059***	0.214
Job 2	0.069***	0.028	0.079**	0.028
Job 3	0.051**	0.020	0.070*	0.020
Job 4	–0.012	0.113	–0.004	0.113
Job 5	-0.067***	0.491	–0.058***	0.491
Job 6	-0.050***	0.052	–0.053***	0.053
Brand 1	–0.005	0.588	0.002	0.588
Brand 2	–0.005	0.129	0.004	0.129
Brand 3	0.075	0.177	0.102	0.177
Brand 4	–0.069***	0.093	–0.059*	0.093
Brand 5	0.254**	0.004	0.298	0.004
Brand 6	0.061	0.003	0.110	0.003
Zone 1	0.049***	0.282	0.143***	0.282
Zone 2	0.148	0.167	0.008	0.127
Zone 3	0.019***	0.109	0.072***	0.109
Zone 4	0.083***	0.299	0.039**	0.299
Zone 5	0.047	0.054	0.005	0.045
Marginal effects	0.13637073		0.13345326	

Note: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

In both models, the frequency of claims increased with male drivers, drivers older than 75 years and with drivers who live in greater Tunis, in the Northwest and in the Coast. In addition, the claim frequency is positively related with officials, senior executives and middle managers, while it is negatively related with craftsmen and employees. Interesting, only the vehicles made in China/Korea/Japan have a significant and negative sign in the ZIP model. However, only the vehicles made in the USA have a significant but positive sign in the ZINB model.

For the second part (logistic model), it is interesting to note that the variable risk exposure is significant at the one percent level in both models. The negative sign indicates that the number of at-fault claims declines with the driver's exposure, which means the value zero shows that the drivers did not have a claim during the insurance policy period.

The over dispersion parameter was found to be significant. Therefore, the ZI negative binomial formulation is better than the ZI Poisson model.

TABLE 7 *ZIP model and ZINB model results*

INDEPENDENT VARIABLES	ZIP		ZINB	
	COEFFICIENT	STD. ERR.	COEFFICIENT	STD. ERR.
Coverage	0.324	0.059***	0.428	0.068***
Gender	0.192	0.066***	0.211	0.100**
Use	0.027	0.045	0.132	0.055**
Age 2	0.251	0.337	-0.013	0.448
Age 3	0.208	0.333	-0.033	0.445
Age 4	-0.385	0.346	-0.704	0.455
Age 5	0.661	0.339*	1.253	0.463***
Job 1	0.116	0.085	0.306	0.106***
Job 2	0.238	0.116**	0.405	0.161**
Job 3	0.169	0.139	0.343	0.184*
Job 4	-0.047	0.104	-0.104	0.123
Job 5	-0.455	0.087***	-0.509	0.102***
Job 6	-0.444	0.134***	-0.561	0.160***
Brand 1	-0.033	0.322	0.017	0.369
Brand 2	-0.031	0.328	0.050	0.375

[…]

INDEPENDENT VARIABLES	ZIP		ZINB	
	COEFFICIENT	STD. ERR.	COEFFICIENT	STD. ERR.
Brand 3	0.286	0.323	0.578	0.371
Brand 4	-0.609	0.339*	-0.559	0.384
Brand 5	0.463	0.370	1.057	0.488**
Brand 6	0.226	0.448	0.555	0.579
Zone 1	0.716	0.093***	0.872	0.102***
Zone 2	0.136	0.116	0.083	0.125
Zone 3	0.496	0.109***	0.451	0.122***
Zone 4	0.324	0.097***	0.293	0.105***
Zone 5	0.169	0.150	0.062	0.161
Constant	-1.168	0.485**	-2.369	0.596***
Inflation model : logistic regression				
Risk Exposure	-0.637	0.095***	-14.275	747.916***
Constant	1.353	0.0496***	-0.518	0.222**
Over dispersion	-	-	2.807	0.297
Log Likelihood	-6692.246		-6645.591	
LR chi2(24)	451.76		545.07	
Prob> chi2	0.000		0.000	

Note: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

5.2. Model comparison and selection

Table 8 reports the parameters log likelihood, AIC and BIC for the four models.

TABLE 8 *Comparison of LL, AIC and BIC for the four models*

MODELS	CRITERION			
	LOG LIKELIHOOD	PARAMETERS	AIC	BIC
Poisson	–7475.195	25	15000.39	15190.87
Negative Binomial	–6660.584	26	13373.17	13571.27
ZIP	–6692.246	27	13438.49	13644.21
ZINB	–6645.591	28	13347.18	13560.52

For testing the over dispersion in Poisson vs. NB regressions, the likelihood ratio is $2[-6660.5844-(-7475.1951)]=1629.2214>\chi^2$, indicating that H_0 is rejected and the data is over dispersed. Therefore, the introduction of an alpha term (over dispersion) improves the fit compared with the Poisson regression model. In addition, based on AIC and BIC, the NB regression model has the lowest value for both criteria, indicating that the NB is more adequate than the Poisson model.

For testing the over dispersion in ZIP vs. ZINB regression models, the likelihood ratio is 75.31. This result indicates that the data is over dispersed, and ZINB is better than the ZIP regression model. In addition, the information criteria favor the ZINB model.

For choosing between ZIP vs. Poisson and ZINB vs. NB regression models, Table 9 reports the results of the Vuong test. Because the z-value is both positive and significant at the 0.1% level, the test of Vuong (1989) shows a strong preference for the ZIP model over an ordinary Poisson regression model (z value equal to 14.42). This result is consistent with Mouatassim and Ezzahid (2012).

TABLE 8 *Test of Vuong*

V STATISTIC	*z*	PR > *z*
Vuong test of ZIP vs. standard Poisson	14.42	0.0000
Vuong test of ZINB vs. standard NB	2.61	0.0046

In addition, Vuong's test shows that the ZINB regression model is better than the standard NB regression model (with a z value equal to 2.61, which is significant at the 1% level).

When choosing between all the models, there is not much difference between the NB, ZIP and ZINB regression models based on the log-likelihood. The ZINB regression model shows superior fit compared to the other models, while the Poisson regression model is inferior to the other models. The ZI models fit better than their corresponding standard count models; this finding suggests the best fitting model needs to account for both over dispersion and zero-inflation in the observed data. This result is consistent with that found by Vasechko *et al.* (2009), where they used the same models to analyze the French automobile insurance market.

5. Conclusion

This study models the number of at-fault claims that occurred over a one-year time period for beginner drivers who had a license for less than three years. This paper shows powerful econometric models suitable for application in estimating the at-fault claim frequency. The methodology applied employs four well-known regression models in modeling the insurance claims: the Poisson and negative binomial, zero-inflated Poisson and zero-inflated negative binomial regression models.

Empirically, we show that the explanatory variables of the claim frequency are appreciably the same as those with the standard count models. The claim frequency increases with the choice coverage, the gender and the use of the vehicle. However, it decreases with the brand of automobile.

An important finding is that there are many exogenous variables that can explain the frequency of claims that are not taken into account in calculating the basic insurance premium such as coverage choice, gender of the policyholder, and the residence or job of the policyholder. Therefore, there is room to improve ratemaking for car insurance. By comparing the models, the log-likelihood, the information criteria, and the Vuong tests indicate that the zero-inflated binomial negative regression is the best model for handling zero-inflated and additional over dispersed claim count data.

While it will be more appealing to examine the number of claims within a panel data set, as for example in Boucher *et al.* (2009), we leave this for future work.

References

Benlagha, N., Charfeddine, L. and Karaa, I., (2012), Modelling claim occurrence in car insurance implementation on Tunisian data, *Asian-African Journal of Economics and Econometrics*, 12(2), 395-406.

Benlagha, N. and Karaa, I., (2017), Evidence of adverse selection in automobile insurance market: A seemingly unrelated probit modelling, *Cogent Economics and Finance*, 5(1), 1330303.

Boucher, J.P., Denuit, M. and Guillén, M., (2009), Number of claims or number of claims? An approach with zero-inflated Poisson models for panel data, *The Journal of Risk and Insurance*, 76,821-846.

Boucher, J.P. and Guillén, M., (2011), A semi-non parametric approach to model panel count data, *Communications in Statistics- Theory and Methods*, 40, 622-634.

Cameron, C. and Trivedi, P.K., (2015), Count panel data, Baltagi B.H.,The oxford handbook of Panel data, *Oxford University Press USA*, 233-256.

Denuit, M.X., Marechal, S. Pitrebois, and Walhin JF., (2007), Actuarial modeling of claim counts: risk classification, credibility and bonus-malus systems, *Wiley*, New York.

FTUSA, Fédération Tunisienne des Sociétés d'Assurance, (2013), *Annual Report,* December, Tunisia.

Greene, W. H. (1994), Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working paper, Stern School of Business, NYU EC-94-10.

Hilbe, J.M. (2014), Modeling count data, *Cambridge University Press.*

Karaa, I., and Benlagha, N. (2015), Testing for asymmetric information in Tunisian automobile insurance market, *Mediterranean Journal of Social Sciences*, 6, 455-464.

Lambert, D. (1992), Zero-inflated Poisson regressions, with an application to defects in manufacturing, *Technometrics*, 34, 1-14.

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

Melgar, M. C., Ordaz Sanz, J.A. and Guerrero, M., (2005), Diverses alternatives pour déterminer les facteurs significatifs de la fréquence d'accidents dans l'assurance automobile, *Insurance and Risk Management*, 73(1), 31-54.

Melgar, M. C., Ordaz Sanz, J.A. and Guerrero, M., (2006), Une étude économétrique du nombre d'accidents dans le secteur de l'assurance automobile, *Brussels Economic Review*, 49(2), 169-183.

Mouatassim, Y., and Ezzahid, E., (2012), Poisson regression and zero-inflated Poisson regression: application to private health insurance data, *European Actuarial Journal*, 2, 187-204.

Mouatassim, Y., Ezzahid E. and Belasri, Y. (2012), Operational Value-at-risk in case of zero-inflated frequency, *International Journal of Economic and Finance*, 4(6), 70-77.

Stram, D.O., and Lee, J.W. (1994), Variance components testing in the longitudinal mixed effects model, *Biometrics*, 50, 1171-1177.

Stram, D.O., and Lee, J.W. (1995), Correction to "Variance components testing in the longitudinal mixed effects model", *Biometrics*, 51, 1196.

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

Vuong, Q.H., (1989), Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica,* 57 (2), 307–333.

Washington, S., karlaftis, M.G., and Mannering, F.L., (2003), Statistical and economic methods for transportation data analysis, Chapman and Hall/ CRC Press.

Yang, Z., Hardin, J. W., Addy, C. L. and Vuong, Q. H. (2007), Testing approaches for over dispersion in Poisson regression versus the generalized Poisson model, *Biometrical Journal*, 49, 565-584.

Yau, K. K., Wang, K. and Lee, A. H. (2003), Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros, *Biometrical Journal*, 45, 437-452.

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

Ismail, N., and Zamani H., (2013), Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models, *Casualty Actuarial Society E-Forum*.

Zhao, X., and Zhou, X., (2012), Copula models for insurance claim numbers with excess zeros and time-dependence, *Insurance: Mathematics and Economics*, 50, 191-199.

NOTES

[1] Economic, Management and Computer Sciences Doctorate School of FSEG Sfax University of Sfax, Road of airport, BP 1088, Sfax, 3018, Tunisia
E-mail: Karaa.imen66@yahoo.com

[2] International School of Business, Sfax, Tunisia
E-mail: Habib.chabchoub@gmail.com

[3] All the numbers and percentages presented in this section are based on the notes of the 2013 FTUSA annual report.

[4] Karaa and Benlagha (2015) examined coverage-risk correlation in the Tunisian automobile insurance market. By using a bivariate probit model, they observed the presence of asymmetric information for beginning drivers. Therefore, asymmetric information seems to be at most a negligible phenomenon in the market for experienced drivers. Therefore, in this paper we choose to model the data only with bad drivers, i.e., beginning drivers. In addition, the data was used by Benlagha and Karaa (2017) to test adverse selection phenomenon.

[5] In Tunisia, there are 11 and 8 risk classes' bonus malus for private use and the rest of use, respectively. The evolution of the bonus malus depends only on claims occurrence. New drivers who have less than three years of driving experience are placed in classes 8 and 7 for private use and the rest of use, respectively. Good drivers can descend one class after two consecutive insurance years without any claim. A lower class creates more incentives for safe driving. Bad drivers will be punished by reclassification of one or two classes for a material/ body claim.