European Research Studies,
Volume XIII, Issue (4), 2010

Use and Extension of Count Data Models in the Determination of Relevant Factors for Claims in the Automobile Insurance Sector

José Antonio Ordaz[1] María del Carmen Melgar[2] M. Kazim Khan[3]

Abstract:

Using real, Spanish data, different specifications of zero-inflated models are provided in this paper to estimate the number of accidents declared by policyholders. These count data models seem to be the most appropriate solutions to study this question. Our work is completed with the estimations of the number of clients that do not declare their actual accidents and the number of these accidents. The analysis of all these factors could become useful for insurers to improve their efficiency. We conclude with a final theoretical discussion on the possible advantages given by other alternative models, like the so-called thinned models.

JEL Classification: C52, G22

Keywords: Accident, zero-inflated model, 'extra-zeros' estimation.

[1] *José Antonio Ordaz – Universidad Pablo de Olavide – Carretera de Utrera Km. 1, 41013 Sevilla, Spain – Tel.: +34 954348548 – Fax: +34 954349339 – e-mail: jaordsan@upo.es*

[2] *María del Carmen Melgar – Universidad Pablo de Olavide – Carretera de Utrera Km. 1, 41013 Sevilla, Spain – Tel.: +34 954348549 – Fax: +34 954349339 – e-mail: mcmelhir@upo.es*

[3] *M. Kazim Khan – Kent State University – P.O. Box 5190, Kent, Ohio 44242, USA – e-mail: kazim@math.kent.edu*

* *This work has received support from the Spanish Ministry of Science and Innovation and FEDER grant ECO2008-01223/ECON.*

1. Introduction

From the earlier 70's, we can find a significant number of theoretical analysis about the contractual relationships in the insurance industry where there exists asymmetric information between the participants. Nevertheless, empirical estimates of assurance models referring to adverse selection and moral hazard phenomena appear more recently. In this sense, the works by Dionne (1998) and Abbring, Chiappori, Heckman and Pinquet (2003) are good examples of reference. One of the main objectives of the empirical approach focused on the asymmetry of the information has been to prove that higher coverage is positively related to higher number of accidents.

Although we can find this type of analysis in different sectors, automobile insurance market has developed as one of the most appropriate fields to carry out these studies (Chiappori, 1999). In this sense we can point out the works by Dahlby (1983) and Boyer and Dionne (1989). Both of these works do not deny the existence of asymmetric information. However, data employed by Dahlby (1983) are of aggregated kind and it is not clear enough that the consideration of individual data leads to similar results. Some years later we can find another work by Dalhby (1992) where aggregated data (from Canada) are used again.

Puelz and Snow (1994) are the first authors that employ individual data; in this case, this data came from an insurance company of Georgia, USA. This work is considered in the literature as the seminal one in this research area. They provided a two-equation model. The first one was referred to the insurance companies' policies price and the second described the fixation process of a franchise by the companies, considering the premium, personal characteristics of the insured people and the occurrence of an accident. This study concludes that higher accident rates are related with people that choose the lowest franchises and so on, the higher coverage levels. Major criticisms to this work are referred to the use of linear specifications in the model estimating process and to the no consideration of variables related to the risk.

Chiappori and Salanié (1997) proposed a very general approach potentially useful to any situation with asymmetric information. The main idea consists in the simultaneous estimation of two non-linear equations. The first one refers to the chosen franchise, only depending on the particular characteristics of the insured people. The second one establishes a dummy variable, indicating if there is, or not, an accident. The simultaneous estimating process determines the existing relationship between higher accident rates and higher coverage levels.

Richaudeau (1999) follows this research guideline and offers and important advance. His work does not only indicate the occurrence of an accident. He goes beyond this point and provides a count data model in order to estimate the number of accidents. In this task, he uses a negative binomial model.

Studies by Chiappori and Salanié (1997, 2000) and Richaudeau (1999), where French automobile insurance data are used, do not show any relationship between accident rates and levels of insurance coverage. Nevertheless, it is important to point out that the research by Chiappori and Salanié (1997) is based on individuals with only a few years of driving experience. The non-existence of significant correlation between coverage and accident rates could be due to the driving inexperience of this group. This fact does not necessarily imply a lack of correlation in the group of drivers with more experience.

Cohen (2005) uses a thorough database from Israel in relation with insured people with different intervals of driving experience years. His study confirms the concluding remarks noted by Chiappori and Salanié about the absence of correlation between coverage levels and accident rates in inexperienced drivers. However, he finds this type of correlation in drivers that have more than two years of driving experience. Cohen's work provides a Poisson model, although this model may not be the most appropriate one because of the high percentage of insured people that did not have (or did not declare) any accident.

The present study refers to the automobile insurance as well. Our main purpose consists in estimating the number of accidents that are declared by policyholders. By this process, we do not only analyze the existing potential correlation between this number and the levels of coverage, but we also explain their most significant factors.

Additionally, we can deduce from our methodology that a large part of the policyholders do not declare they have accidents to their company. We can estimate this number of policyholders and the number of accidents they do not declare. This extension of our research is not really usual and it could come in useful for insurance companies to consider the benefits of theirs 'bonus-malus' policies.

After this first introductory section, we describe in Section 2 the database we have employed in this work. Then, in Section 3 we present the main features of the count data models that are more commonly used in these types of works (Cameron and Trivedi, 1998). At this point, we justify why zero-inflated models can be the most appropriate solution for the situation we study. Section 4 provides the results we have obtained from our econometric modeling process. In this section we indicate the most important variables that explain the number of accidents in the automobile insurance. We also estimate the number of policyholders that do not declare any accident although they have had someone. This is a very interesting aspect of our work.

To conclude, we discuss in Section 5 some problems related to the interpretation of the zero-inflated models, which can be considered from alternative points of view and need to be used with some care. These aspects could be the beginning for a future research work. The paper finishes with the appendices and the full detailed references we have indicated along this paper.

2. Descriptive analysis of the database

The database we use in this study has been kindly provided by a Spanish private insurance company that works in the automobile sector. We initially have information from 63900 clients of this company. After a first debugging process, we have selected 60000 policies from 16th June 2002 to 15th June 2003 and, finally, we have taken a random sample of 15000 registers due to computational reasons. This size is large enough to be representative.

Available information has been classified in four different categories: variables related to the insured vehicle; about the personal characteristics of policyholders; features of the insurance policy; and characteristics of the declared accidents. The exact description of all these variables is explained in the Appendix 1 at the end of the work.[4]

2.1 Characteristics of the insured vehicle

Types and uses of vehicles are the variables we have considered in this point. The vehicle's type offers five different possibilities. "Car or Van" represents 80.50% of the whole of vehicles. After that, we can find the categories referred to "Special Vehicle" and "Motorcycle"; they represent 10.37% and 7.69%, respectively. The other categories ("Truck" and "Coach") are jointly only 1.44%.

With respect to the uses of vehicles, original data have been grouped in three categories of use. The "Private" use is the most relevant one, representing almost 80% of the whole of vehicles, in particular 79.76%. After this category, we find "Professional" with 19.63%. Finally, "Other uses" are only 0.61%.

2.2 Characteristics of policyholders

The most relevant characteristics of policyholders for the insurance companies are, basically, age, gender, years of driving license experience and the usual area of traffic. The date of the reference period in our research is the 15th December 2002. The average age of the drivers of our database is quite high: 48 years old. Only 3.08% of them are less than 26 years old.

Related to the gender, 84.83% of policyholders of our database are males. Driving experience of policyholders is another relevant aspect that is taken into account by insurance companies to fix the premiums; they give a different treatment to insured drivers with a license experience less than 2 years. These drivers only represent 0.71% of the whole of registers.

Traffic area is the last characteristic that insurance companies take into consideration in this section. Insurers usually take the address of policyholders as a proxy of their usual area of traffic. Initially, our data refers to the 52 Spanish provinces. We have grouped all of them in 8 areas corresponding to the Spanish NUTS-1 or geographic groups of regions that Eurostat considers for statistical purposes in the case of Spain. According to this classification, we find that the most

[4] *Dionne, Gouriéroux and Vanasse (1999) and Cohen (2005) use a similar classification. Nevertheless, we must note information on drivers and vehicles required by insurance companies is less extent in Spain than in other countries.*

represented region is the “Southern” with 46.33% of all the insured people. In relation to the rest of regions, we could point out the “Central”, that represents 16.84%, “Northwestern” with 15.43%, and “Eastern” that is 12.07%. The other 4 regions only get jointly 9.33%.

2.3 Characteristics of the policies

The two main elements that define an insurance policy are the premium and the level of coverage. The premium is referred to the annual amount that insured people should pay to their companies. We have grouped this variable in 4 categories. More than a half of policyholders pay less than 400 € (58.93%). The highest premium (corresponding to more than 600 €) is the category with the least frequency; it represents 17.84% of the whole of policies.

With respect to the levels of insurance coverage, warranties that can be contracted are: compulsory responsibility, supplementary responsibility, defense and claim of damages, own damages, fire, crash damages, total damages, stealing, breakage of windows, death, disability, travel assistance, deprivation of driving license and total loss. We have defined three levels of coverage depending on the contracted warranties in policies and on the types of vehicles: low, medium and high (Appendix 2). Globally, over half of all the policyholders have the lowest level of coverage (54.28%). When this level increases, the portion of drivers who contract it decreases: 37.77% have the medium level and, finally, 7.95% subscribe the highest one.

2.4 Characteristics of the accidents

Our database contains the exact date of accidents, their cost, description and guilt. In this work, we have only focused on the occurrence and number of accidents associated to each policy. We want to study if these accidents (and their number) are declared or not to the company by the policyholders.

Table 1 shows the distribution of the number of declared accidents. One of the most relevant aspects we can point out is that 77.05% of insured people have not declared any accident along the studied period. The total number of policyholders with registered accidents is 3442. Most of them have only declared 1 accident; they represent 68.71% of the cases. People with 2 accidents are also a relevant group, representing 21.59% and only 9.70% of the whole have declared between 3 and 7 accidents. The average number of accidents is 1.46 per policy.

Table 1: Distribution of the number of registered accidents

No of accidents	Frequencies	Including 0		Not including 0	
		Percentages	Accumulative percentages	Percentages	Accumulative percentages
0	11558	77.05	77.05		
1	2365	15.77	92.82	68.71	68.71
2	743	4.95	97.77	21.59	90.30
3	223	1.49	99.26	6.48	96.78
4	78	0.52	99.78	2.26	99.04
5	19	0.13	99.91	0.55	99.59
6	10	0.06	99.97	0.29	99.88
7	4	0.03	100.00	0.12	100.00
TOTAL	15000	100.00		100.00	

Source: *Own study from the database.*

Table 2 shows the results of the relationship between accidents and characteristics of insured vehicles. If we consider the different types of vehicles, we can see on the one hand that the categories referred to “Special Vehicle” and “Motorcycle” have few accidents: only 6.75% and 7.03% of vehicles of each respective type have had and declared some accident along the analyzed period. On the other hand, the category with the highest accident rate is “Coach” (52.17%). The types of “Car or van” (the most represented in the sample) and “Truck” present similar accident rates: 26.46% and 25.26%, respectively.

In relation to the uses of insured vehicles, from Table 2 we can note the majority group, i.e. “Private”, is which have the highest accident rate: 24.66% of these policies have declared some accident. The figure for the “Professional” use is quite lower: 16.33%. Finally, the category that includes the rest of possible uses, “Other use”, shows the lowest figure: only 11.96%.

Table 2: Accident rates by characteristics of insured vehicles

Variables	Accidents %		
	No	Yes	Total
Types of vehicles			
Car or van	73.54	26.46	100.00
Truck	74.74	25.26	100.00
Coach	47.83	52.17	100.00
Motorcycle	92.97	7.03	100.00
Special Vehicle	93.25	6.75	100.00
Uses of vehicles			
Private	75.34	24.66	100.00
Professional	83.67	16.33	100.00
Other	88.04	11.96	100.00
TOTAL	77.05	22.95	100.00

Source: *Own study from the database.*

If we observe accident rates related to the age of policyholders (Table 3), we do not find large differences between the three groups we have considered from 14 to 70 years old. However, the eldest group (with more than 70 years old) shows a significantly lower accident rate: 15.93%. The reason of this behavior can rely on the fact that people from this group do not already use their vehicles very much because of their age, so there is less probability they register an accident. Anyway, this group is not very important in terms of its number.

The results of the introduction of the gender of policyholders in the analysis of accidents are shown in Table 3 as well. Females have an accident in the 26.46% of cases. This figure is higher than is presented by males, who suffer someone in the 22.32% of cases. However, we must note the average number of accidents is almost the same for both genders: 1.45 for females and 1.46 for males.

With respect to the years of experience of the driving license, we can note most inexpert drivers present an accident rate that is close to 13 points higher than those who have 2 years or more of experience: 35.51% and 22.86%, respectively (Table 3). Nevertheless, the average number of accidents for inexperience drivers is lower than for experience drivers.

If we jointly consider the usual area of traffic and the accidents happening, Table 3 also shows there are maybe four regions whose behavior is noticeably different from the rest of them. On the one hand, we can observe that “Madrid” has an accident rate quite higher than the others: 28.71%. On the other hand, the “Northwestern”, “Canarias” and “Central” regions have accident rates lower than the average. In any case, it should be taken into account that the actual significance of the data of these areas depends on their weight in the sample.

Table 3: Accident rates by characteristics of policyholders

	Accidents %		
Variables	No	Yes	Total
Groups of age			
[14-25] years old	76.62	23.38	100.00
[26-45] years old	75.84	24.16	100.00
[46-70] years old	77.28	22.72	100.00
More than 70 years old	84.07	15.93	100.00
Gender			
Male	77.68	22.32	100.00
Female	73.54	26.46	100.00
Driving experience			
Less than 2 years	64.49	35.51	100.00
2 years or more	77.14	22.86	100.00
Usual area of traffic			
Canarias	78.69	21.31	100.00
Central	81.04	18.96	100.00
Ceuta-Melilla	75.00	25.00	100.00
Eastern	75.59	24.41	100.00
Madrid	71.29	28.71	100.00
Northeastern	75.63	24.37	100.00
Northwestern	77.28	22.72	100.00
Southern	76.04	23.96	100.00
TOTAL	77.05	22.95	100.00

Source: *Own study from the database.*

The study of the relationship of accidents with the policy premiums shows a positive correlation between these two variables (Table 4). As it is well known, when an accident happens, the premiums increase. Our data confirms this point. In addition, the average number of accidents corresponding to each level of premiums increases as well.

Finally, in Table 4 we also can see the relationship between accidents and policyholders' levels of coverage. It is very interesting to observe that higher levels of coverage seem to be directly associated to higher accident rates. While only 16.13% of policyholders with the lowest level of coverage had some accident, this percentage grows up to 39.43% for policyholders with the highest level of coverage. It is also remarkable that the average number of accidents increases with the levels of coverage: 1.36, 1.48 and 1.64 are the figures associated to the low, medium and high levels, respectively. So, our data seem to show a strong positive correlation between accident rates and levels of coverage of insured drivers, as it is expected in frameworks where there exists asymmetric information between insured and insurer sides.

Table 4: Accident rates by characteristics of policies

	Accidents %		
Variables	No	Yes	Total
Groups of annual premiums (€)			
(0,300]	88.22	11.78	100.00
(300,400]	77.44	22.56	100.00
(400-600]	71.91	28.09	100.00
> 600	63.08	36.92	100.00
Levels of coverage			
Low	83.87	16.13	100.00
Medium	70.72	29.28	100.00
High	60.57	39.43	100.00
Total	77.05	22.95	100.00

Source: *Own study from the database.*

3. Methodology

The most proper models to be employed in the estimation procedures of discrete variables with nonnegative integer values are the count data ones. In this sense the traditional models are the Poisson and the negative binomial regression models. When it is known up front that the zero counts are inflated, there is a qualitative difference between the positive values versus the zero values. In other words, the zero values may have multiple sources. In such situations, the zero-inflated negative binomial (ZINB) and the zero-inflated Poisson (ZIP) models have found a wide variety of applications (Greene, 1997; Cameron and Trivedi, 1986, 1998; Jones, 2001; Winkelmann, 2003; Yau, Wang and Lee, 2003; Melgar, Ordaz and Guerrero, 2004; Melgar and Ordaz, 2005).

More generally, if *Z* is any random variable taking nonnegative integer values, the zero-inflated version of *Z*, denoted by *Y*, has the density:

$$\begin{aligned} P(Y=0) &= q+(1-q)P(Z=0) \\ P(Y=k) &= (1-q)P(Z=k), \qquad k=1,2,... \end{aligned} \tag{1}$$

The random variable *Y* may be viewed as a discrete mixture of the density of *Z* with the density of a degenerate random variable at zero (Cameron and Trivedi, 1998). In the context of insurance data, *Z* could represent the actual number of accidents that a specified *i* client will have during the year and $1-q_i$ is his/her probability of reporting them to the insurance company (the so-called probability of *participation*). The significant proportion of zero values in our dependent variable can have two different meanings: on one side, perhaps the policyholder has not actually suffered an accident; and on the other side, the policyholder can have suffered an accident but he has decided not to declare it to the insurance company in order not to be punished in their premiums.

Our model for the number of accidents declared by the *i*-th client may be expressed as $Y_i = Z_i I_i$, where I_i is an independent Bernoulli random variable with $P(I_i = 1) = 1 - q_i$, and:

$$q_i = F\big(\tau(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_n X_{in})\big). \tag{2}$$

In this expression, *F* is a cumulative distribution function distribution, typically chosen to be either logistic or standard normal (leading to the logit or probit models respectively), $X_{i1},\ldots,X_{in}$ are the explanatory variables, and $\tau,\beta_0,\beta_1,\ldots,\beta_n$ are the unknown parameters to be estimated.

Let *N* represent the total number of clients in the population. According to any zero-inflated model, the number of clients who did not declare accidents, N_0, and the resulting number of undeclared accidents, A_u, are equal to:

$$N_0 = \sum_{i=1}^{N} \chi_{\{z_i>0, I_i=0\}} \tag{3}$$

and

$$A_u = \sum_{i=1}^{N} Z_i \, \chi_{\{z_i>0, I_i=0\}} , \tag{4}$$

where the notation $\chi_{\{A\}}$ represents the indicator function of the event *A*. Hence the expected number of clients who did not report and the resulting number of undeclared accidents are:

$$E(N_0) = \sum_{i=1}^{N} q_i \big(1 - P(Z_i = 0)\big), \qquad E(A_u) = \sum_{i=1}^{N} q_i \, E(Z_i) . \tag{5}$$

The respective variances, $Var(N_0)$ and $Var(A_u)$, are given by:

$$\sum_{i=1}^{N} q_i \left(1 - P(Z_i = 0)\right)\left\{1 - q_i \left(1 - P(Z_i = 0)\right\},\right.$$

$$\sum_{i=1}^{N} q_i \left\{Var(Z_i) + E(Z_i)^2 (1 - q_i)\right\}. \tag{6}$$

Depending upon the choice of the model for Z_i, we may estimate their parameters and then obtain the estimates of the above expressions.

For instance, in the case when Z_i is assumed to be a negative binomial random variable, the zero-inflated negative binomial model becomes:

$$P(Y_i = k) = q_i\left(1 - \min\{k,1\}\right) + \left(1 - q_i\right)\frac{\Gamma(k+\nu)}{\Gamma(k+1)\cdot\Gamma(\nu)}\left(\frac{\nu}{\nu+\lambda_i}\right)^{\nu}\left(\frac{\lambda_i}{\nu+\lambda_i}\right)^{k}, \quad k = 0,1,2,\ldots \tag{7}$$

where $\nu > 0$ and:

$$\lambda_i = \exp\left\{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_n X_{in}\right\},\ q_i = F(\tau \ln \lambda_i),\ \tau \in R\,. \tag{8}$$

The distribution *F* is taken to be the probit or logit model (Greene, 1997). If we introduce the so-called ‘precision parameter’ α (Jones, 2001), where $\alpha = \frac{1}{\nu}$, it can be proved that $E(Y|_{X_i}) = \lambda_i$ and $Var(Y|_{X_i}) = \lambda_i + \alpha\lambda_i^2$. As α goes to zero the mean and the variance become equal, that is a feature of the Poisson regression model. For this reason, α could be considered as a measure of the data’s over-dispersion level and if we could contrast its statistical significance, we could determine the validity of the Poisson model against the negative binomial model.

After estimating the parameters of this model, we can obtain the most important factors that determine the number of declared accidents.

In addition, with the help of this model, the expected number of clients who did not report and the resulting number of undeclared accidents are:

$$E(N_0) = \sum_{i=1}^{N} q_i\left(1 - \left(\frac{\nu}{\nu+\lambda_i}\right)^{\nu}\right), \qquad E(A_u) = \sum_{i=1}^{N} q_i\,\lambda_i\,. \tag{9}$$

The respective variances of N_0 and A_u are

$$Var(N_0) = \sum_{i=1}^{N} q_i\left(1 - \left(\frac{\nu}{\nu+\lambda_i}\right)^{\nu}\right)\left\{1 - q_i\left(1 - \left(\frac{\nu}{\nu+\lambda_i}\right)^{\nu}\right)\right\},$$

$$Var(A_u) = \sum_{i=1}^{N} q_i\,\lambda_i\left(2 + \frac{\lambda_i}{\nu} - q_i\right). \tag{10}$$

Using $\hat{q}_i$, $\hat{\lambda}_i$ and $\hat{\nu}$ as the maximum likelihood (ML) estimates of q_i, λ_i and ν respectively, we may obtain the estimates of the number of clients who did not report their accidents and the total number of undeclared accidents.

For the sake of completeness, in the following we also provide the expressions for the case when Z_i is assumed to be a Poisson random variable. In this case, the zero-inflated Poisson model becomes:

$$P(Y_i = k) = q_i(1-\min\{k,1\}) + (1-q_i)e^{-\lambda_i}\frac{\lambda_i^k}{k!}, \quad k = 0,1,2,\ldots, \tag{11}$$

and λ_i and q_i are as in (8). In this case the expected counts are:

$$E(N_0) = \sum_{i=1}^{N} q_i\left(1-e^{-\lambda_i}\right), \qquad E(A_u) = \sum_{i=1}^{N} q_i\,\lambda_i\,. \tag{12}$$

The respective variances of N_0 and A_u are:

$$Var(N_0) = \sum_{i=1}^{N} q_i\left(1-e^{-\lambda_i}\right)\left\{1-q_i\left(1-e^{-\lambda_i}\right)\right\},$$

$$Var(A_u) = \sum_{i=1}^{N} q_i\,\lambda_i\left(1+\lambda_i\,(1-q_i)\right). \tag{13}$$

Once again one may use the ML method to estimate the parameters.

In a similar way to the ZINB model, we can also estimate the expected number of clients who did not report their accidents and the total number of undeclared accidents.

To conclude this section, we can note that the choice between the zero-inflated specifications of models against their usual forms can be done by using the Vuong statistic (Vuong, 1989):

$$V = \frac{\sqrt{N}\left[\frac{1}{N}\sum_{i=1}^{N} m_i\right]}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(m_i - \overline{m}\right)^2}}, \tag{14}$$

where $m_i = \log\left(\frac{P_1(Y = y_i)}{P_2(Y = y_i)}\right)$, $P_1(Y = y_i)$ and $P_2(Y = y_i)$ are the functions of distribution corresponding to the zero-inflated and the 'traditional' specifications of the models, respectively, and $\overline{m}$ is the mean of m_i, $i = 1,\ldots,N$.

Vuong proves that this statistic follows a reduced normal distribution. When its value is higher than 1.96, the zero-inflated model is then the best estimation procedure. On the other hand, when the value of this statistic is lower than -1.96, the 'traditional' specifications of the models are most desirable. In the range between these two values, the decision remains unclear.

4. Main results

In this section we show the main results provided by our econometric study. We have determined the most significant explanatory variables in the estimation process of the number of accidents that have been declared by the drivers to their insurance company, pointing up the existence of correlation between the highest levels of coverage and the highest accident rates.

As our endogenous variable only have nonnegative discrete values, count data models become the best analytic solution to be used. In addition, we must take into consideration that this variable has a large number of zero values; in particular, 77.05% of all the policyholders of our database declared they have not had any accident. As mentioned earlier, under these circumstances zero-inflated models have revealed as the most appropriate ones, being preferred to their 'traditional' forms. They can explain this situation in the best way because it can be supposed that a large number of these zeros are not actual zeros. There are some people who do not declare their accidents to the insurance company in order not to be punished in their premiums.

We have used *Limdep v. 7.0* as econometric software to carry out the estimation of the parameters of our model. After having compared a large number of different possible regression models, we have finally chosen the specifications as shown in Table 5. They correspond to the ZINB and the ZIP models, and their results are very similar. As it can be seen from the significance of Alpha and Tau coefficients, these models are preferred to their respective simple or 'traditional' forms (Greene, 1995). Additionally, the value and significance of the Vuong statistic justifies our choice.

We can observe from Table 5 that there are 10 significant explanatory variables (including the constant), considering a level of confidence of 95%.

If we analyze the *types of vehicles*, we find that the categories referred to "Coach", "Motorcycle" and "Special vehicle" show a significantly different behavior in relation with all the rest, i.e. cars, vans, and trucks. On the one hand, coachs have a higher positive relationship with the number of declared accidents; on the other hand, the relationship of motorcycles and special vehicles is lower compared to the other vehicles. Initially, the negative sign of the parameter associated to motorcycles could be considered surprising, but the reason could rely on the hard conditions that our company imposes on the motorcycle drivers that maybe lead them to prefer not to declare their accidents.

Related to the *uses of vehicles*, "Other uses" appears negatively correlated with the number of declared accidents in comparison with the "Private" and "Professional" uses.

Table 5: Final output of the zero-inflated (ZINB and ZIP) models

Dependent variable: NUMACC
Logistic distribution used for splitting model
Total number of included observations: 15000
Actual zeros:11558

	ZINB model			ZIP model		
Variable	Coeff.	*z*-stat.	*P*-value	Coeff.	*z*-stat.	*P*-value
CONSTANT	-0.36406	-10.399	0.0000	-0.33009	-11.061	0.0000
COACH	0.81861	4.335	0.0000	0.77351	4.780	0.0000
MOTORCYC	-0.93411	-7.896	0.0000	-0.82654	-7.960	0.0000
SP_VEH	-0.84024	-8.647	0.0000	-0.74643	-8.684	0.0000
OTH_USE	-0.62848	-3.014	0.0026	-0.56710	-3.072	0.0021
EXP<2Y	0.58427	4.070	0.0000	0.52877	4.323	0.0000
CENTRAL	-0.19388	-4.743	0.0000	-0.17363	-4.796	0.0000
NORTWEST	-0.11234	-3.134	0.0054	-0.10281	-3.246	0.0012
COV_MED	0.31675	8.743	0.0000	0.28886	9.054	0.0000
COV_HIGH	0.59024	10.310	0.0000	0.54430	10.872	0.0000
Over-dispersion parameter: Alpha	0.72680	3.133	0.0017			
Zero-inflation model parameter: Tau	-0.74384	-4.090	0.0000	-0.99467	-5.261	0.0000
Log. Likelihood	-10761.0			-10765.8		
Predicted zeros	11946.3			11909.3		
Vuong statistic	9.6978			39.3839		

Source: *Own study.*

The *driver's experience*, observed throughout the years of his/her driving license, is another relevant correlated variable with the number of declared accidents. In particular, insured drivers with less than 2 years of driving experience (denoted by EXP<2Y) have higher probability of having (or declaring) accidents.[5]

With respect to the *region of policyholders' residence*, only two of them appear as significant against the rest of the country: the "Central" and the "Northwestern" regions. The negative sign of their respective associated parameters indicates that policyholders that come from these regions have a lower probability of having (or declaring) accidents than those of the rest of Spain.

The last significant variable in our analysis is the *level of insurance coverage*. Associated coefficients to each one of the different levels show an increasingly positive relationship with claims as well. So, the higher levels of insurance coverage the higher accident rates. This result suggests the existence of problems related to adverse selection and moral hazard and confirms the theoretical

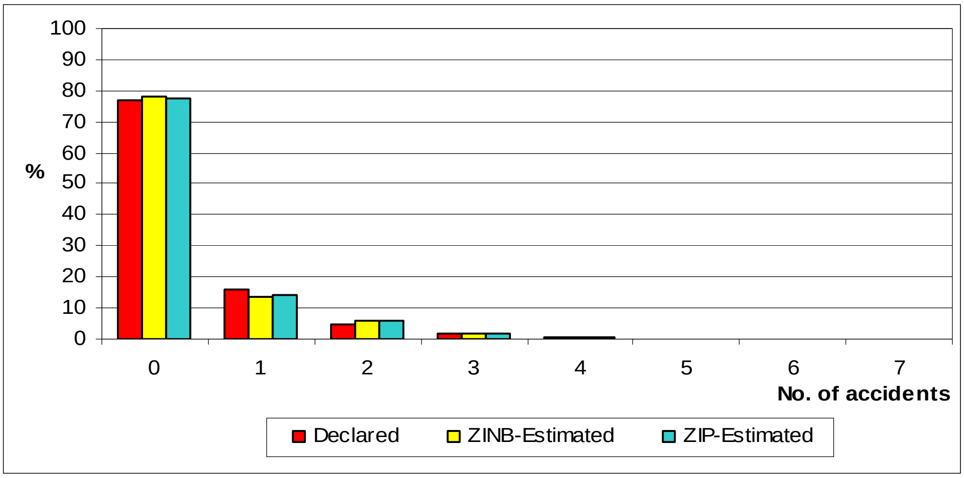
[5] *At this point, we must note that the age of policyholders has not been chosen in the estimating process of the model to avoid possible problems of collinearity with the driving experience. In some way, its effect must be present through this experience.*

aspects pointed up by other empirical studies in the literature as, for instance, Dionne, Gouriéroux and Vanasse (1999), Richaudeau (1999) and Cohen (2005).[6]

Finally, we must note that the variable referred to *gender of policyholders* have not been significant enough in any case in our study.

The goodness of fit of our models in terms of number of declared accidents and estimated probability for each one of these numbers can be observed graphically in Figure 1.

Figure 1: Declared and estimated probability of number of accidents by ZINB and ZIP models



Source: *Own study.*

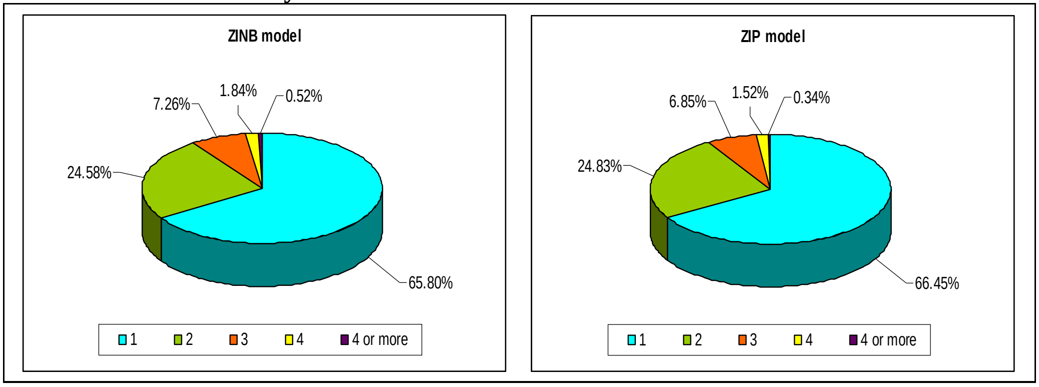
After having estimated the zero-inflated specifications of our models, ZINB and ZIP, we have passed to a second step in our work to determine the number of ‘extra-zeros” suggested by the model, that is, the number of policyholders that do not declare their accidents although they have had someone. Additionally, we have estimated the number of accidents non-reported by this way. This aspect was explained in the methodology section and it was implemented via a *Matlab* code we have specifically designed for this purpose. This extension is not common in this type of research.

Table 6 shows that the number of estimated ‘extra-zeros’ is 4048 in the case of the ZINB model and 4304 if we consider the ZIP model. This leads to the estimated number of non-reported accidents of 5945 and 6221 associated to each

[6] *As to the premiums, we have not included them in the final econometric analysis because of their high collinearity with the level of coverage, although the amount of premiums is not necessarily related to this level (for instance, this amount depends on the type and the use of vehicle as well).*

model, respectively. Figure 2 illustrates, for both models, the distribution of policyholders that do not declare their accidents depending on this number.

Figure 2: Distribution of policyholders that do not declare their accidents depending on this number by ZINB and ZIP models



***Source:* Own study.**

If we take into consideration, for instance, the ZINB model from Table 6, the analysis of their figures suggests there exists a 35.02% of policyholders that declared they had no accidents, yet they actually had at least one, and the total of estimated number of non-declared accidents would represent 55.37% of the theoretical whole of happened accidents. These policyholders likely opted for this decision in order not to be punished by the company.

This information can be very useful for insurers because they could then evaluate the success of the implementation of their ‘bonus-malus’ policies and improve their efficiency.

Table 6: Number of estimated ‘extra-zeros’ and undeclared accidents

	ZINB-Estimated (95% confidence interval)	ZIP-Estimated (95% confidence interval)
Number of ‘extra-zeros’ / Policyholders who did not declare their accidents	4048 (3943-4155)	4304 (4196-4412)
Number of undeclared accidents corresponding to the ‘extra-zeros’	5945 (5760-6131)	6221 (6040-6402)

Source:* *Own study.

5. Final discussion

To conclude, we should point out that zero-inflated models can have another interpretation, which indicates that such models need to be used with some care. As pointed out in the methodology section of this paper, all zero-inflated models can be viewed as $Y_i = Z_i I_i$, where Z_i is a random variable representing the actual number of accidents, I_i is an independent Bernoulli random variable, and Y_i is the number of reported accidents. The density of Y_i is precisely the zero-inflated models. However, since $Y_i = Z_i I_i$, one may argue that the independent coin toss experiment (denoted by I_i) takes place at the end of the year (or at the beginning of the year), resulting in classifying the individual as the one who reports all or non of his/her accidents.

This feature becomes further evident when one considers the question of the expected number of undeclared accidents given that the person reported some accidents, i.e., $E(Z_i - Y_i |_{Y_i>0})$. This expression is always zero when Y_i is taken to have any zero-inflated model. These features indicate that there is a need to update the zero-inflated models which are both tractable as well as represent the more realistic scenarios where the client may report some of his/her accidents but not necessarily all of them. For instance, one may propose a zero-inflated model of the following type: $Y_i = I_{i0} + I_{i1} + I_{i2} + \cdots + I_{iZ_i}$, where Z_i is the total number of accidents that the *i*-th client has over the year and I_{ik} indicates whether the *k*-th accident will be reported or not, and we take $I_{i0} = 0$ with probability one.

The tractability of this model depends on the assumptions one makes about $I_{i1}, I_{i2}, \ldots, I_{iZ_i}$. The standard zero-inflated models are all based on the assumption that $I_{i1} = I_{i2} = \cdots = I_{iZ_i} = I_i$ which is independent of Z_i.

Arguably this assumption may be unrealistic in various zero-inflated count data situations. Another possibility is to assume that: $I_{i1}, I_{i2}, I_{i3}, \ldots$ are independent and identically distributed as $Bernoulli(1-q_i)$, which leads to the case of Y_i is distributed as $Poisson(\lambda_i(1-q_i))$, and Y_i is distributed as $Negative\,Binomial(v,(v/(v+\lambda_i(1-q_i))))$. Such models may be called the thinned models. A bit more generally, if one assumes that $I_{i1}, I_{i2}, I_{i3}, \ldots$ are exchangeable *Bernoulli* random variables, to allow a dependence structure on $I_{i1}, I_{i2}, I_{i3}, \ldots$, the resulting models then become less tractable.

These are some alternatives that we would like to study in depth in the near future.

Appendix 1: Definition of the employed variables in the econometric analysis

Dependent variable	
NUMACC	Number of declared accidents.
Explanatory variables	
VEH_CAT	Types of vehicles. - Dummy variables: TRUCK (truck), COACH (coach), MOTORCYC (motorcycle), SP_VEH (special vehicle: it includes overall industrial and agricultural vehicles). - Excluded category: car or van.
VEH_USE	Uses of vehicles. - Dummy variables: PROF_USE (professional use) and OTH_USE (other uses). - Excluded category: private use.
AGE	Age of policyholders (years old). - Dummy variables: AG26_45 (between 26 and 45 years old), AG46_70 (between 46 and 70 years old) and AG71_ (more than 70 years old). - Excluded category: between 14 and 25 years old.
FEMALE	Gender of policyholders: 1 for female; 0 otherwise.
NUTS-1	Large regions or areas (NUTS-1) of usual traffic. - Dummy variables: CANARIAS (Islas Canarias), CENTRAL (Central region: Castilla-La Mancha, Castilla y León and Extremadura), CEU_MEL (Autonomous Cities of Ceuta and Melilla), EASTERN (Eastern region: Cataluña, C. Valenciana and Islas Baleares), MADRID (Madrid), NORTEAST (Northeastern region: Aragón, Euskadi, La Rioja and Navarra), NORTWEST (Northwestern region: Asturias, Cantabria and Galicia). - Excluded category: Southern region (Andalucía and Region of Murcia).
EXP<2Y	Driving experience: 1 for less than two years' driving experience; 0 otherwise.
PREMIUM	Annual premiums (€). - Dummy variables: P301_400 (between 301 and 400 €), P401_600 (between 401 and 600 €) and P601_ (more than 600 €). - Excluded category: less than 301 €.
LEV_COV	Levels of insurance coverage. - Dummy variables: COV_MED (level of medium coverage) and COV_HIGH (level of high coverage). - Excluded category: level of low coverage.

Appendix 2: Definition of the levels of insurance coverage

Levels of coverage	Warranties
Low	Compulsory responsibility, supplementary responsibility, defense and claims of damages, death and/or disability and travel assistance.
Medium	Low level + fire and/or breakage of windows and/or stealing and/or deprivation of driving license.
High	Low level + own damages, general damages or total loss.

References

1. Abbring, J.H., P.A. Chiappori, J.J. Heckman and J. Pinquet, 2003, "Adverse Selection and Moral Hazard in Insurance: Can Dynamic Data Help to Distinguish?", Journal of the European Economic Association 1 (Papers and Proceedings), 512-521.
2. Boyer, M. and G. Dionne, 1989, "An Empirical Analysis of Moral Hazard and Experience Rating", Review of Economics and Statistics 71, 128-134.
3. Cameron, A.C. and P.K. Trivedi, 1986, "Econometric Models Based on Count Data: Comparison and Applications of Some Estimators and Tests", Journal of Applied Econometrics 1, 29-54.
4. Cameron, A.C. and P.K. Trivedi, 1998, "Regression Analysis of Count Data" (Cambridge University Press, Cambridge).
5. Chiappori, P.A., 1999, "Asymmetric Information in Automobile Insurance: An Overview", in: Dionne, G and C. Laberge-Nadeau, C. (eds.), "Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation", 1-11 (Kluwer Academic Publishers, Boston, MA).
6. Chiappori, P.A. and B. Salanié, 1997, "Empirical Contract Theory: The Case of Insurance Data", European Economic Review 41, 943-950.
7. Chiappori, P.A. and B. Salanié, 2000, "Testing for Asymmetric Information in Insurance Markets", Journal of Political Economy 108 (1), 56-78.
8. Cohen, A., 2005, "Asymmetric Information and Learning: Evidence from the Automobile Insurance Market", Review of Economics and Statistics 87 (2), 197-207 .
9. Dahlby, B., 1983, "Adverse Selection and Statistical Discrimination: An Analysis of Canadian Automobile Insurance Market", Journal of Public Economics 20, 121-131.
10. Dahlby, B., 1992, "Testing for Asymmetric Information in Canadian Automobile Insurance", in: Dionne, G (ed.), "Contributions to Insurance Economics", 423-443 (Kluwer Academic Publishers, Boston, MA).
11. Dionne, G., 1998, "La Mesure Empirique des Problèmes d'Information", Cahier de recherche 9833 (U.F.R. de Sciences Économiques, Gestion, Mathématiques et Informatique, Paris).
12. Dionne, G., C. Gouriéroux and C. Vanasse, 1999, "Evidence of Adverse Selection in Automobile Insurance Markets", in: Dionne, G and C. Laberge-

Nadeau (eds.), “Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation”, 13-46 (Kluwer Academic Publishers, Boston, MA.).

13. Greene, W. H., 1995, “Limdep Version 7.0: User's Manual” (Bellport, NY: Econometric Software).
14. Greene, W.H., 1997, “Econometric Analysis” (3rd ed) (MacMillan, New York.)
15. Jones, A. M., 2001, “Applied Econometrics for Health Economists - A Practical Guide” (Office of Health Economics, London).
16. Melgar, M.C., J.A. Ordaz and F.M. Guerrero, F.M., 2004, “The Main Determiants of the Number of Accidents in the Automobile Insurance: An Empirical Analysis”, Études et Dossiers – Working Paper Series of the Geneva Association 286, 45-54.
17. Melgar, M.C. and J.A. Ordaz, “La siniestralidad en el seguro de automóviles en Andalucía frente al resto de España. Un análisis a través de un modelo count data”, Economic Analysis Working Paper 4 (15).
18. Puelz, R. and A. Snow, A., 1994, “Evidence on Adverse Selection: Equilibrium Signaling and Cross-Subsidization in the Insurance Market”, Journal of Political Economy 102 (2), 236-257.
19. Richaudeau, D., 1999, “Automobile Insurance Contracts and Risk of Accident: An Empirical Test Using French Individual Data”, Geneva Papers on Risk and Insurance Theory 24, 97-114.
20. Vuong, G. H., 1989, “Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses”, Econometrica 57, 307-333.
21. Winkelmann, R., 2003, “Econometric Analysis of Count Data” (Springer-Verlag, Berlin).
22. Yau, K.K.V., K. Wang and A.H. Lee, 2003, “Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros”, Biometrical Journal 45, 437-452.