Equilibrium
Volume 6 Issue 3, 2011
ISSN 1689-765X

José Antonio Ordaz
Universidad Pablo de Olavide, Spain
María del Carmen Melgar
Universidad Pablo de Olavide, Spain
M. Kazim Khan
Kent State University in Ohio, United States

An Analysis of Spanish Accidents in Automobile Insurance: the Use of the Probit Model and the Theoretical Potential of Other Econometric Tools*

JEL Classification Codes: *C1*

Keywords: *automobile insurance, claims, probit model, zero-inflated models, thinned models*

Abstract: Automobile insurance is one of the main pillars of the entire insurance industry in the developed economies. Knowing as much as possible about the factors related to the accidents is an essential issue for the insurance companies so that they may improve their levels of efficiency. Therefore, in this paper we focus on studying the most relevant variables that help explain the registration of claims in the automobile insurance sector. For this purpose, we fit a probit model specification using a database from a Spanish insurance company. Our research points out the significance of certain variables, such as the policyholders' driving experience, their region of residence as well as their levels of insurance coverage, towards the likelihood of registering an insurance claim. However, probit analysis represents only one of the multiple perspectives which we can consider to examine the question of accidents and their reporting. Very briefly, we also discuss the utility of zero-inflated count data models to study the number of accidents declared by policyholders. Finally, we point out the possibilities that thinned models could offer for this type of research.

Introduction

Automobile insurance is one of the most important branches of the whole insurance industry in all modern countries. In the case of Spain, in 2009 the overall

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

amount of the premiums of this sector represented 35.97% of total revenue from non-life insurance, and 18.97 % of all insurance business (DGSFP 2010). Therefore, any researching work referred to this economic sector could be significant.

The main objective of the present analysis is to identify which variables are the most relevant in the determination of the probability process that the policyholder makes or not claims. Characteristics of the insured vehicle, such as its category and use, others relating to the driver, such as age, gender, driving experience and area of residence, and those relating to the policies, as its annual premium and the chosen level of insurance coverage, are some of the variables that are ordinarily taken into consideration by insurance companies. To know the factors that may affect the claims, it is a matter of great interest for insurers (Cohen 2005; Ordaz and Melgar 2010a). The availability of a good risk model would allow firms to establish more precisely the premium that must be paid by their policyholders, which would give greater efficiency to this important issue.

To achieve this objective, in this study we take the information on the variables above outlined from a Spanish insurer included in an important Swiss multinational group, to which we apply a probit discrete binary choice model to explain where the variable is defined so that it reflects the report of claims, by assigning the values 1 and 0 respectively.

One of the most important results we find in our econometric analysis refers to the evidence of positive relationship between the claims and the level of insurance coverage contracted by the policyholders, shown by other authors as well. It is the case, for instance, of the work by Puelz and Snow (1994), that offers a similar conclusion with other methodology, based in a two-equation model and data of Georgia, USA. This point may reflect moral hazard and/or adverse selection behaviours, very common situations in the insurance world.

But this is only one of the different possibilities of studying the accidents in the sector. Zero-inflated models or even thinned models are other econometric tools one may use for dealing with this topic in depth from other perspectives. We make a brief review of the most important theoretical features of these techniques as a secondary objective of this work.

The paper consists of 6 sections. After the introduction, Section 2 contains a description of the main features of the dataset. In Section 3, we point out the main characteristics of the well known probit model we have used for our analysis. The results are then presented in Section 4. Section 5 briefly discusses a few other potential extensions that can be considered for an in depth study of this topic of research. Section 6 provides the main conclusions of this work. Acknowledgements and References are provided at the end of the paper.

Descriptive analysis of the database

The findings of this research paper are based on a database, containing insurance information of 130,000 policies, which was provided to us by a private Spanish insurer. This company belongs to one of the most solid multinational insurer group in Europe, which has its head office in Switzerland.

The time interval for this dataset covers the period from June 16, 2002 to 15 June 2003. For computational reasons, a random sample of 15,000 policies has been used[1], of which we know certain characteristics related to the type and use of the vehicle; age, gender, years of driving experience and area of residence of the policyholder; and also the annual premium he/she pays and the level of insurance coverage of the policy. These variables, or a categorical version of them, are taken as explanatory variables. On the other hand, we have considered a binary variable (that we have labelled CLAIM) whose 1 and 0 values reflect if the insured has made or has not made some kind of claim, respectively.

Since our primary objective is to analyze which factors are the most significant in determining the report or non-report of any type of claim, we especially focus on the differences that arise in each of the available variables regarding this issue.

First of all, we must emphasize the large number of zeros that appear for the dependent variable: 11,558 policyholders have not declared any loss, representing a rate of 77.1% from the total number of insured drivers of our database.

The vehicles have been classified into five groups according to the type they belong: “tourism or van”, “truck”, “coach”, “motorcycle” and “special vehicle”. The category that includes tourisms and vans is the most common one, accounting for 80.5% of the total. After them, special vehicles represent 10.3% and motorcycles appear with 7.7%. Trucks and coaches jointly give the remaining 1.5%[2].

As to the claims in each category, Table 1 shows 26.5% of cars or vans have registered some claim in the reference period, and the figure for trucks is very similar: 25.3%. In contrast, the behaviour exhibited on the one hand, by coaches, and on the other hand, by motorcycles and special vehicles, is very different: 52.2% of coaches have reported some claim, but only 7% of motorcycles and 6.8% of special vehicles registered claims.

[1] The software we use in our subsequent econometric analysis shows some problems with the total size of the database, so we have selected a random sample that is enough large to be representative of all the data (more than 10%). Additionally we have previously made the descriptive analysis we offer in this Section with the entire population: the results in percentage terms have been very similar in all the analyzed distributions.

[2] The distribution of the vehicles of our database is quite similar to the official figures of the total number of the Spanish vehicles (DGT 2004), thus showing how representative can be our analysis as a proxy of the Spanish market. Cars and vans represent 83.3% and motorcycles account for 6.0%; the other vehicles appear with some little differences due to the several classification criteria can be used, basically in the case of trucks and special industrial vehicles.

Table 1. Claim rates by types of insured vehicles in %

	Claims		
Types of vehicles	No	Yes	Total
Car or van	73.5	26.5	100.0
Truck	74.7	25.3	100.0
Coach	47.8	52.2	100.0
Motorcycle	93.0	7.0	100.0
Special vehicle	93.2	6.8	100.0
All categories	77.1	22.9	100.0

Source: own study from the database.

The descriptive analysis of the figures for the main use of the insured vehicle indicates that 79.8% of them are for "personal" use. With respect to "professional" use (which includes public service, industrial uses, freight transport, school transport, passenger transport and general farming), it accounts for 19.6% and finally, the category of "other" (which was rental concerns, driving school, sale and withdrawal of driving licenses) is only 0.6% of the total.

Table 2 presents the details of claims for each one of the uses we have indicated. One can see the professional and, indeed, any other uses show lower claim rates, 16.3% and 12.0% respectively, than the ones which are registered in the case of personal use: 24.7%.

Table 2. Claim rates by uses of insured vehicles in %

	Claims		
Uses of vehicles	No	Yes	Total
Personal	75.3	24.7	100.0
Professional	83.7	16.3	100.0
Other	88.0	12.0	100.0
All categories	77.1	22.9	100.0

Source: own study from the database.

Among the characteristics of the insured people, age is the first variable we analyze. Four intervals were considered: "18–25 years old", "26–45 years old", "46–70 years old" and "more than 70 years old". The majority of considered drivers belong to the middle intervals. In particular, policyholders between 26 and 45 years old represent 39.8% of the total and those 46 to 70 years old, 51.8%. The remaining 8.4% is distributed so that the younger group of 14 to 25 years old is accounting for 3.1% and that of the older ones, for 5.3%.

In regard to the claims, Table 3 shows that the percentages of policyholders who have someone are for the first three age groups around 22-24%. The category of policyholders with more than 70 years old, meanwhile, shows a remarkable lower figure of claim rate: only 15.9%.

Table 3. Claim rates by age of policyholders in %

	Claims		
Groups of age	No	Yes	Total
[14–25] years old	76.6	23.4	100.0
[26–45] years old	75.8	24.2	100.0
[46–70] years old	77.3	22.7	100.0
More than 70 years old	84.1	15.9	100.0
All categories	77.1	22.9	100.0

Source: own study from the database.

We have considered the gender of the policyholders as well. The descriptive analysis of this question indicates that 85.3% are men, and 22.3% of them made some claim. Regarding women, they show a slightly higher figure, which is 26.5%.

The driving experience is another aspect to be taken into consideration. This was done through the variable referring to the time of possession of a driving license. Considering all the insured drivers, only 0.8% have less than 2 years of experience. However, its claim rate accounts for 35.5%. This is a much higher percentage than the one of the experienced drivers, namely 22.9%.

The area of residence is also a highly relevant variable. This variable is normally taken as a proxy for the policyholders' usual driving area. We have worked with the division of the Spanish territory at the level of NUTS-1 Regions, according to the criterion of Eurostat[3]. The "Southern" region is the most represented one, bringing together 46.3% of the total insured. We should then mention the following regions: "Central", which accounts for 16.8%; "North-western" with 15.4%; and "Eastern", which includes 12.1% of the whole of policyholders. The other four regions share the remaining 9.4%

As to the claims found in each of the regions, Table 4 shows that residents in the first region ("Southern") presented claims in 24.0% of cases. Of the rest, it must be pointed out the significantly higher figure of "Madrid", where the percentage of policyholders with claims reaches 28.7%. At the other extreme, we see the "Canarias" and, especially, the "Central" region, where the figures of claims are 21.3% and 19.0% respectively.

[3] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction.

Table 4. Claim rates by areas of residence of policyholders in%

	Claims		
Areas of residence	No	Yes	Total
Canarias	78.7	21.3	100.0
Central	81.0	19.0	100.0
Ceuta-Melilla	75.0	25.0	100.0
Eastern	75.6	24.4	100.0
Madrid	71.3	28.7	100.0
North-eastern	75.6	24.4	100.0
North-western	77.3	22.7	100.0
Southern	76.0	24.0	100.0
All categories	77.1	22.9	100.0

Source: own study from the database.

The last block of analyzed variables refers to features directly related to the policies. In particular, it has been taken into consideration the annual amount paid as premium and the level of insurance coverage.

With respect to the amount of the premium, it has been divided into four intervals (in € = euros): "(0–300]", "(300–400]", "(400–600]", and "more than 600". The majority of policyholders belong to the interval of cheapest premiums, representing 32.2% of the insured drivers of our database. The two middle intervals provide similar figures, representing 26.8% and 23.2% respectively. Finally, the premiums above 600 € are only 17.8% of the total.

As to the claim rates, it is very noticeable the positive and growing relationship between the amount of the premium and the report of claims shown in Table 5. While the percentage of claims of the policies of less than 300 € is 11.8%, this number is gradually rising from finally reaching the 36.9% in the case of policies with premiums in excess of 600 €.

Table 5. Claim rates by amount of policies' annual premiums in%

	Claims		
Groups of annual premiums (in €)	No	Yes	Total
(0–300)	88.2	11.8	100.0
(300–400)	77.4	22.6	100.0
(400–600)	71.9	28.1	100.0
More than 600	63.1	36.9	100.0
All categories	77.1	22.9	100.0

Source: own study from the database.

Regarding the coverage of the policy, it has been divided into three levels based on the guarantees of the insurance contract. The "low" level includes only the compulsory guarantees under the law; policies with this level of coverage account for 54.3% of the total. Those who want any additional optional guarantee, such as that concerning the glass breakage, fire and/or theft, are integrated in the level of coverage that we have labelled as "medium". This is the type chosen by 37.8% of insured drivers of our database. Finally, the "high" level also covers the own damage of the vehicles; here is the 7.9% of the total insured.

The analysis of claims for each one of the levels of insurance coverage can be seen in Table 6. In this, one can observe how the percentages will grow as does the level of insurance coverage. Thus, for the lowest level, the percentage of cases with claims that is collected is 16.1%, for the intermediate is 29.3%, and for the highest one is 39.4%.

Table 6. Claim rates by policies' insurance coverage levels in%

	Claims		
Levels of insurance coverage	No	Yes	Total
Low	83.9	16.1	100.0
Medium	70.7	29.3	100.0
High	60.6	39.4	100.0
All categories	77.1	22.9	100.0

Source: own study from the database.

This result is very interesting. Even though this should not necessarily imply that policyholders with different levels of insurance coverage differ in risk, it is true that from the perspective of insurers they really find theses differences in the claim rates. On the one hand, the relationship of this variable with claims could indicate a situation of moral hazard arising from behaviour by those excessively careless drivers who enjoy a wide coverage. Additionally, on the other hand, it may also reflect the existence of adverse selection behaviour as a driver aware of his/her proneness to claims would generally contract a higher coverage for reassurance. Both issues are among the main problems that are seen in the insurance market.

We will see afterwards if this last empirical result is statistically supported by our subsequent econometric analysis.

Basic description of the econometric model

As we noted in the Introduction of this work, we use an econometric model with the described dataset to make our analysis. In particular, a probit model is provided in this research. The binary discrete choice models[4] such as probit, are characterized by the endogenous variable *Y* only takes two values, 1 and 0, corresponding to each of the two possible scenarios that are considered.

In this study, the endogenous variable Y_i takes the issue of whether the *i*-th policyholder made or not some type of claim to the insurer such that:

$$Y_i = \begin{cases} 1 & \text{if the } i\text{ - th policyholder made some claim} \\ 0 & \text{otherwise} \end{cases}$$

If we assume that the variable *Y* depends on a set of explanatory variables *X*, following the general econometric specification:

$$Y_i = F(X_i\beta) + \varepsilon_i \qquad [1]$$

where ε_i represents the usual random disturbance error, this model estimates the probability that the policy of the *i*-th individual records any claim:

$$\hat{Y}_i = \hat{P}_i = F(X_i\hat{\beta}) \qquad [2]$$

From this general approach, common to any binary discrete choice model[5], the probit model is characterized by using the distribution function for a standard normal: Φ. So, we will have:

$$F(X_i\beta) = \Phi(X_i\beta) = \Phi(Z_i) = \int_{-\infty}^{Z_i} \phi(s)\,ds \qquad [3]$$

where:

$$\phi(s) = \frac{1}{(2\pi)^{1/2}} e^{-\frac{s^2}{2}} \qquad [4]$$

is the density function of normal distribution and *s* is a 'latent' integration variable with mean 0 and variance 1.

Regarding the interpretation of the model, the estimated parameters do not directly determine the marginal effect of changes in exogenous variables X_j on the estimated probability (as in the case of a linear model). Its sign and mag-

[4] These models are very common in the econometric field. For this reason we only describe their most basic characteristics. Gujarati (2003) can be indicated as a good reference for further details.

[5] Probit and logit models are quite similar in its most essential aspects. Our choice between both of them has been based on the better goodness-of-fit shown by the probit model in our subsequent empirical analysis.

nitude, however, are indicative of the direction of change and the relevance of these variations. The marginal effect is then computed as a result of the product of the density function of standard normal distribution at a determined point (the policy of the *i*-th individual) and the corresponding parameter:

$$\frac{\partial P_i}{\partial X_j} = \frac{\partial \Phi(X_i \beta)}{\partial X_j} = \phi(X_i \beta)\beta_j \qquad [5]$$

The magnitude of the variation of probability is based on the values of each and every one of the explanatory variables and their respective coefficients in the particular observation we want to consider. Therefore, in order to obtain a representative value of these marginal effects, they are usually evaluated for the mean values of the regressors.

If X_j is a dummy variable, which is the case with most of the explanatory variables in our model, the analysis of their average effect is done through the difference of the values provided by:

$$E[Y_i \mid X_k = 1] \text{ and } E[Y_i \mid X_k = 0] \qquad [6]$$

With respect to the estimation of the model, it will be done through the maximum likelihood method that provides consistent and asymptotically efficient estimators.

To test the individual significance of each parameter (and consequently of the corresponding explanatory variable) is used the Wald test, whose *z*-statistic, follows a standard normal distribution. To evaluate the goodness-of-fit of this type of models, there exist different alternatives, such as the McFadden R^2, the LR-statistic or likelihood ratio and pseudo-R^2 of prediction-evaluation. Finally, as for the detection of the existence of possible problems of endogeneity in the model, one can use the so-called Hausman test (Hausman 1978, pp. 1251–1272).

Estimation of the model and structural analysis of results

Table 7 shows the probit model specification of register of claims which has finally been selected from among the various tests that have been carried out[6].

This choice is based on the significance of the explanatory variables (at $\alpha < 0{,}6$ and is also made to ensure goodness-of-fit and its global significance. In this regard it is noted that the value of the pseudo-R^2 of the prediction-evaluation of the

[6] As mentioned above, we have also used a logit model in our tests. We have finally chosen a probit specification due to goodness-of-fit criteria.

chosen specification is 76.99%. This value is quite significant.[7] Regarding the endogeneity between the variables of the model, the Hausman test confirmed the presence of this question. This limitation is usual in this type of research and is generally assumed.

Table 7. Model estimation output

Dependent variable: CLAIM					
Model: binary probit Method: Maximum likelihood					
Included observations: 15,000					
Variable	Coefficients	Marginal effects	Standard error	z-Statistic	P-value
CONSTANT	-0.791757	-0.258	0.020301	-39.00008	0.0000
COACH	0.756319	0.288	0.264005	2.864792	0.0042
MOTORCYC	-0.727330	-0.182	0.060677	-11.98681	0.0000
SP_VEH	-0.696053	-0.177	0.052810	-13.18024	0.0000
OTH_USE	-0.531261	-0.141	0.168153	-3.159385	0.0016
EXP<2Y	0.628696	0.234	0.128287	4.900711	0.0000
CENTRAL	-0.141053	-0.045	0.033210	-4.247302	0.0000
MADRID	0.220726	0.078	0.098379	2.243618	0.0249
NORTWEST	-0.100458	-0.032	0.033295	-3.017240	0.0026
COV_MED	0.295716	0.092	0.025649	11.52917	0.0000
COV_HIGH	0.554107	0.186	0.041820	13.24989	0.0000
Mean dependent variable		0.229467	LR statistic		904.1646
St. deviation dependent variable		0.420504	Degrees of freedom		10
Log likelihood		-7,627.384	Probability of LR statistic		0.000000
Restricted log likelihood		-8,079.467	McFadden R^2		0.055954
Expectation-Prediction evaluation for binary specification (success cut-off: C = 0.5)					
Correct predictions for dependent variable = 0		11,530	Correct predictions for dependent variable = 1		18
Pseudo-R^2 (%) 76.99					

Source: own study.

All variables introduced in the model are qualitative, so their entry is done through dummy variables[8]. It should also be highlighted that some of the initially selected variables have not been significant enough in some specifications

[7] All the econometric results shown in Table 7 have been carried out with *EViews* v.6, except that references to “marginal effects” of the explanatory variables which has externally been calculated according to equations [6].

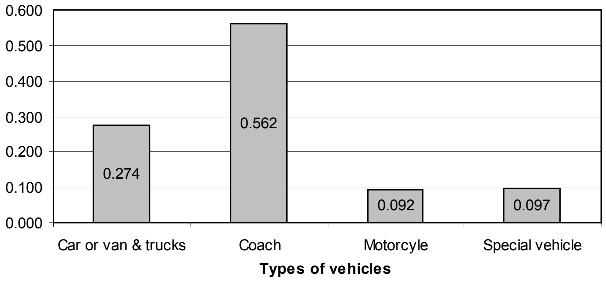
[8] The introduction of dummy variables is performed additively, thus avoiding problems that could arise when including interaction terms (Ronis and Harrison 1988, pp. 361–372).

we have made, or have shown evidence of multicollinearity; for that reason, they have not been considered in our final adjustment[9].

The results of this estimate, together with the structural analysis has been performed subsequently from them, allow the following conclusions:

The first group of variables is devoted to the different types of vehicles. We have taken as the base category the cars and vans. In comparison, all categories have proved statistically significant except for the trucks. The incidence of these categories in the register of claims is unequal both quantitatively and in the sign. So, while the drivers of coaches show a greater propensity for claims that the set of categories that do not appear explicitly, motorcycles and special vehicles have less chance of claims. Figure 1 shows the results of structural analysis performed on this variable. It can be seen how the estimated average probability of claims[10] for coaches is 0.562. Meanwhile, for cars or vans, jointly with trucks, it is 0.274. Motorcycles and special vehicles, however, offer substantially lower figures, specifically, 0.092 and 0.097, respectively[11].

Figure 1. Estimated average probability of claims by types of insured vehicles



Source: own study.

Another relevant variable is the use of vehicles. In particular, it has been significant through the category related to "other" uses (OTH_USE), which includes all other uses different from personal and professional ones, as defined in the descriptive analysis of the data given in Section 2. Compared to these two uses, the "other" category shows a negative relationship to the claims; in particular, the probability of making a claim in this case is 14.1% lower.

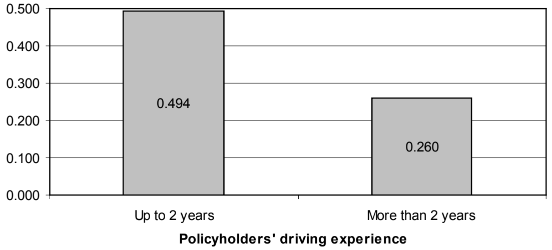
[9] That is the case of the age and the gender of the insured drivers, and the annual premium of the policy, as we will discuss later.

[10] These values are obtained by always taking the mean values of the other explanatory variables.

[11] It is noted that the result of motorcycles, in principle, could be striking. However, this may be due to the hard requirements the insurance company may be imposing to the policyholders of such vehicles.

The experience of drivers revealed as one of the most important variables in explaining the claims in the sector. As expected, the lack of experience is a decisive factor in the occurrence of accidents. Structural analysis of results leads us to verify that policyholders with their licences less than 2 years old have an average probability of suffering a loss equal to 0.494, while this probability for those who possess a driving licence for more than 2 years is 0.260 (figure 2).

Figure 2. Estimated average probability of claims by policyholders' driving experience

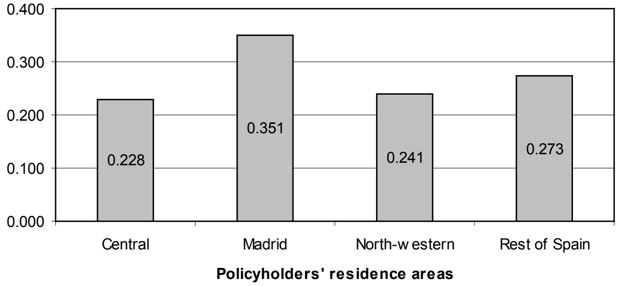


Source: own study.

The area of residence of the insured driver and therefore their usual traffic area is another significant variable to explain claims in automobile insurance. Of the eight great regions in which it divides the Spanish territory, three have behaved significantly differently from the rest: the "Central", "Madrid" and "North-western". The first one refers to the Autonomous Communities of Castilla y León, Castilla-La Mancha and Extremadura, the second one corresponds to the Autonomous Community of Madrid and the third one concerns Cantabria, Galicia and Asturias. While the influence of the "Central" and "North-western" regions is negative in the claims, the "Madrid" region shows a positive relationship and also greater in quantitative terms than others. Figure 3 gives the numbers of the structural analysis from the modelling results on this variable. It can be seen how the estimated average probability of claims is considerably greater in Madrid (0.351) than in the rest of the Spanish State (0.273). However, the other two regions that we have highlighted appear with lower numbers.

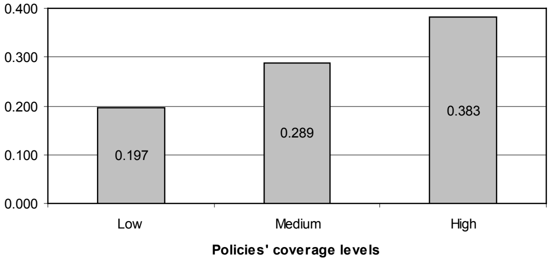
The last variable that has shown its relevance is the extent of policies' insurance coverage. Starting from the lowest level as base category, the other two categories we have considered, i.e. the intermediate (COV_MED) and the highest levels (COV_HIGH), play an important role in the model. The influence of this variable on claim rates is clearly positive and increasing. As can be seen in Figure 4, the estimated average probability of claims for each one of the possible levels of insurance coverage, from lowest to highest, is 0.197, 0.289 and 0.383. Our econometric analysis confirms the results we already saw in our previous descriptive analysis about this question. As we have already said, this can involve some inherent behaviours of the insurance market such as moral hazard and/or adverse selection. We feel that this is one of our most important results, obtaining for the Spanish case the same conclusion shown by other authors referred to other geographical areas, as we indicated in the beginning of this paper.

Figure 3. Estimated average probability of claims by policyholders' residence areas



Source: own study.

Figure 4. Estimated average probability of claims by policies' coverage levels



Source: own study.

Finally, it is necessary to note that the variables related to age and gender of the insured derivers and the policy premiums, initially considered in the descriptive analysis, have not been retained in the econometric estimation of the model. In the case of age, their categories have not been significant enough; its effect, perhaps, is most likely felt indirectly through the variable experience of the driver. Regarding gender, it has not been significant either. And as the premiums are concerned, because of problems of endogeneity in the extent of policy coverage, we decided against its entry into the final specification of the model.

Other possibilities in the analysis of accidents

There exist other important ways to deal in depth the topic of accidents in automobile insurance industry. To show the most significant theoretical characteristics of some of them constitutes an additional objective of our research.

The specific features of this sector make it suitable for deploying econometric models to test the validity of certain theoretical conclusions regarding markets with asymmetric information. The works by Boyer and Dionne (1989), Puelz and Snow (1994), Dionne et al (1999), and Chiappori and Salanié (2000) are some seminal references on this topic.

Other studies have focused on analyzing the number of casualties suffered by drivers, as well as identifying the factors that have significance in this process. In this sense, we can find the works by Shankar et al. (1997), and Lee et al. (2002). Melgar et al. (2005, 2006), and Ordaz and Melgar (2010b) have also analyzed this issue. Their papers used zero-inflated count data models to determine the most important variables when estimating the number of claims that policyholders make to their companies.

If *Z* is a random variable taking nonnegative integer values, the zero-inflated version of *Z*, denoted by *Y*, has the density:

$$\begin{aligned} &P(Y_i = 0) = q_i + (1 - q_i)P(Z_i = 0) \\ &P(Y_i = k) = (1 - q_i)P(Z_i = k) \qquad k = 1, 2, ... \end{aligned} \qquad [7]$$

or alternatively:

$$P(Y_i = k) = q_i\left(1 - \min\{k, 1\}\right) + (1 - q_i)P(Z_i = k) \quad k = 0, 1, 2, ... \qquad [8]$$

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 a specified client will have during the year and $1 - q$ is his/her probability of reporting them to the insurance company.

The number of declared accidents by the *i*-th client may be expressed as $Y_i = Z_i \cdot I_i$, where I_i is an independent Bernoulli random variable defined as:

$$I_i = \begin{cases} 1, \text{ if the policyholder decide to declare their accidents} \\ 0, \text{ otherwise} \end{cases}$$

with $P(I_i = 1) = 1 - q_i$ (the so-called probability of *participation*, i.e. the probability of declaring accidents, associated to each *i*-policyholder), and the probability of *no participation* is given by:

$$q_i = F\left(\tau(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_n X_{in})\right) \qquad [9]$$

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

Depending upon the choice of the probability distribution function for Z_i, Poisson or negative binomial, we will obtain the zero-inflated specification for the Poisson or for the negative binomial model, respectively:

$$P(Y_i = k) = q_i \left(1 - \min\{k,1\}\right) + \left(1 - q_i\right) e^{-\lambda_i} \frac{\lambda_i^k}{k!}, \quad k = 0,1,2,\ldots, \qquad [10]$$

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

$$k = 0,1,2,\ldots \qquad [11]$$

In comparison with the ‘simple’ count data models, the great advantage of zero-inflated count data models relies on their capacity to explain the difference between the policyholders that do not really have accidents against the policyholders that declare they have no accidents, yet they actually have at least one in order not to be punished by their insurance company.

Nevertheless, at this point we should point out that all zero-inflated models have another interpretation, which indicates that such models need to be used with some care. Since $Y_i = Z_i \cdot I_i$, 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., $E(Z_i - Y_i \mid Y_i > 0)$. 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: $Y_i = I_{i0} + I_{i1} + I_{i2} + \cdots + I_{Z_i}$, 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 $I_{i1}, I_{i2}, \ldots, I_{Z_i}$. The standard zero-inflated models are all based on the assumption that $I_{i1} = I_{i2} = \cdots = I_{Z_i} = I_i$ 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 $Negative\,Binomial(\nu, (\nu /(\nu + \lambda_i(1-q_i))))$, depending on the probability distribution function we take for Z_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 leave open for further research in the future.

Conclusions

The weight that automobile insurance industry nowadays represents in the whole insurance business in the developed economies, and the importance for companies of knowing anything related to its activity, are the fundamental reasons that have motivated this work. Thus, the main focus of the analysis we have carried out has been the determination of the most significant variables in the register of claims.

To this end, we have worked with data relating to 15,000 policies provided by a Spanish private insurance company belonging to a relevant Swiss insurer group to which we have applied a probit binary model, since we consider an endogenous variable taking only 1 and 0 values, depending on whether the policy has or has not recorded some claim.

After developing an initial exploratory descriptive analysis, econometric estimation was performed using the probit model. The estimated model provides the most important variables of the database in relation to the accident claims reports. It also points out the influence of each one of the variables with respect to the claims and allows us to estimate their marginal effects. Furthermore, the model helps us to conduct a structural analysis of the results and estimate the average odds of claims for each considered category.

Highlights of this structural analysis are the importance in claims of the type of vehicle (for example, coach), as well as of the policyholders' driving experience. Thus, a coach can be up to 28.8% more likely to claim than most vehicles. Regarding driving license, people whose experience is less than 2 years can increase their probability of claims up to a 23.4% against those who are more expert.

Also notable results have been obtained from the use of vehicle and the area of residence of the insured driver. While other uses than personal and professional ones have exhibited less proneness to register a claim (particularly up to

14.1%), to live in regions such as the Autonomous Community of Madrid make the probability of a claim to be 7.8% higher than in most of the Spanish territory.

Finally, what deserves a special mention is the positive relationship that has been observed between the claims and the variable measuring the levels of insurance coverage. It was found that there is an increased register of claims with increasing level of insurance coverage of the policy. The risk of claims is 18.6% higher in cases in which the insured enjoys the greatest level of coverage against the lowest level, the minimum allowed legally. This may provide evidence of moral hazard and/or adverse selection situations. Both aspects are closely linked to insurance markets with asymmetric information and our analysis appears to indicate them.

To conclude, we have indicated other possible ways we can take into account to analyse the accidents in automobile insurance industry. Models such as zero-inflated count data specifications can be an appropriate option to study the number of claims declared by policyholders to their companies. In this sense, the authors of the present paper have done some work on this topic. As an extension of this field of research, thinned models appear as a good new way to explore in the future.

Literature

Boyer M., Dionne G. (1989), *An Empirical Analysis of Moral Hazard and Experience Rating*, „Review of Economics and Statistics", vol. 71, No. 1.

Cameron A. C., Trivedi P. K. (1998), *Regression Analysis of Count Data*, Cambridge University Press, Cambrigde.

Chiappori P. A., Salanié B. (2000), *Testing for Asymmetric Information in Insurance Markets*, „Journal of Political Economy", vol. 108, No. 1.

Cohen A. (2005), *Asymmetric Information and Learning: Evidence from the Automobile Insurance Market*, „Review of Economics and Statistics", vol. 87, No. 2.

DGSFP – Dirección General de Seguros y Fondos de Pensiones (2010), *Seguros y Fondos de Pensiones. Informe 2009*, Ministerio de Economía y Hacienda, Madrid.

DGT – Dirección General de Tráfico (2004), *Anuario Estadístico General 2003*, Ministerio del Interior, Madrid.

Dionne G., Gouriéroux C., Vanasse C. (1999), *Evidence of Adverse Selection in Automobile Insurance Markets*, [in:] Dionne G., Laberge-Nadeau C. (eds.), *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*, Kluwer Academic Publishers.

Gujarati, D. N. (2003), *Basic Econometrics*, 4th ed., McGraw-Hill.

Hausman J. A. (1978), *Specification Tests in Econometrics*, „Econometrica", vol. 46, No. 6.

Lee A. H., Stevenson M. R., Wang K., Yau K. K. W. (2002), *Modeling Young Driver Motor Vehicle Crashes: Data with Extra Zeros*, „Accident Analysis and Prevention", vol. 34, No. 4.

Melgar M. C., Ordaz J. A., Guerrero F. M. (2005), *Diverses Alternatives pour Déterminer les Facteurs Significatifs de la Fréquence d'Accidents dans l'Assurance Automobile*, „Assurances et Gestion des Risques-Insurance and Risk Management", vol. 73, No. 1.

Melgar M. C., Ordaz J. A., Guerrero F. M. (2006), *Une étude économétrique du nombre d'accidents dans le secteur de l'assurance automobile*, „Brussels Economic Review – Cahiers Economiques de Bruxelles", vol. 49, No. 2.

Ordaz J.A., Melgar M. C. (2010a), *Covariate-Based Pricing of Automobile Insurance*, „Insurance Markets and Companies: Analyses and Actuarial Computations", vol. 1, No. 2.

Ordaz J.A., Melgar M. C. (2010b), *The Utility of Zero-Inflated Models in the Estimation of the Number of Accidents in the Automobile Insurance Industry*, „Equilibrium", vol. 2, No. 5.

Puelz R., Snow, A. (1994), *Evidence on Adverse Selection: Equilibrium Signaling and Cross-Subsidization in the Insurance Market*, „Journal of Political Economy", vol. 102, No. 2.

Ronis D. L., Harrison K. A. (1988), *Statistical Interactions in Studies of Physician Utilization*, „Medical Care", vol. 26, No. 4.

Shankar V., Milton J., Mannering F. (1997), *Modeling Accident Frequencies as Zero-Altered Probability Processes: an Empirical Inquiry*, „Accident Analysis and Prevention", vol. 29, No. 6.