“Covariate-based pricing of automobile insurance”

AUTHORS	José Antonio Ordaz María del Carmen Melgar

ARTICLE INFO	José Antonio Ordaz and María del Carmen Melgar (2010). Covariate-based pricing of automobile insurance. *Insurance Markets and Companies*, *1*(2)
RELEASED ON	Tuesday, 07 September 2010
JOURNAL	"Insurance Markets and Companies"
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES
0

NUMBER OF FIGURES
0

NUMBER OF TABLES
0

© The author(s) 2026. This publication is an open access article.

José Antonio Ordaz (Spain), María del Carmen Melgar (Spain)

Covariate-based pricing of automobile insurance

Abstract

This paper analyzes the most significant variables that explain the existence of claims in the automobile insurance sector. This question is a key issue for insurers. Knowing all these factors, they could eventually fix more precisely their premiums and reach higher levels of efficiency.

To achieve our target, a probit model specification is provided using a database from a Spanish private insurance company. The results of our work point out the significance of some variables such as the policyholders' driving experience or their region of residence. Additionally, our research shows evidence of the existence of relationships between the claims and the increasing policies' levels of insurance coverage, thus suggesting the presence of some usual problems in the insurance markets such as the moral hazard and the adverse selection.

Keywords: automobile insurance, claims, probit model.

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 2008 the overall amount of the premiums of this sector represented 37.24% of total revenue from non-life insurance, and 20.42 % of all insurance business (DGSFP, 2009).

These figures justify why automobile insurance is the focus of much research. Its characteristics make it also conducive to the implementation of econometric models that seek to test the validity of certain theoretical results given in the markets in the presence of asymmetric information. The works by Boyer and Dionne (1989), Puelz and Snow (1994), Dionne, Gouriéroux and Vanasse (1999), and Chiappori and Salanie (2000) are some essential references on this topic. Other studies have focused on analyzing the number of casualties suffered by drivers, as well as identifing the factors influencing it. In this sense, Melgar, Ordaz and Guerrero (2004, 2005, 2006) have used count data models to determine the most important variables when estimating the number of claims that policyholders make to their companies. Shankar, Milton and Mannering (1997), Richaudeau (1999), and Lee, Stevenson, Wang and Yau (2002) are other papers which deal with this issue.

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 to insurers. Finally, the availability of a good risk model would allow insurance 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 the objective, in this study we take the information on the variables above outlined from a Spanish insurer, 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.

The paper is structured into 5 sections. After the introduction, section 1 contains a description of the main features of the data with which we have worked. In section 2, we explain the econometric model to be used. The results are then presented in section 3. The final section offers some brief conclusions. The paper finishes with the References and an Appendix containing the list of variables used in the present work.

1. Analysis of the sample

In order to carry out the research that we intend to perform, we have a database with information on a total of 130,000 policies, which has been provided by a Spanish private insurer. The time interval for this dataset covers the period from June 16, 2002 to June 15, 2003. For computational reasons, a random sample of 15,000 policies has been used, 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) which 1 and 0 values reflect if the insured has made or not some kind of claim, respectively[1].

© José Antonio Ordaz, María del Carmen Melgar, 2010.

[1] As mentioned above, all variables we used are defined in the Appendix at the end of this paper.

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 exhibits 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 only give, jointly, the remaining 1.5%. 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 behavior 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.

Table 1. Claim rates by types of insured vehicles

	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%

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

	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%

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 some 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

	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%

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% has 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%.

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 area of residence is also a highly relevant variable. This variable is normally taken as a proxy for the policyholder 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. The exact definition of each one of them can be seen in the Appendix of the present research. 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.

Table 4. Claim rates by areas of residence of policyholders

	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%

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, the positive and growing relationship is very noticeable 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

	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%

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

	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%

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 these 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 behavior by those excessive careless drivers who enjoy a wide coverage. Additionally, on the other hand, it may also reflect the existence of adverse selection behavior 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.

2. Model specification

A probit econometric model is provided in this research. The binary discrete choice models such as probit, are characterized by the endogenous variable Y which only takes two values, 1 and 0, corresponding to each of the two possible scenarios that are considered[1].

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

[1] There exist other binary discrete choice models, such as the linear probability model (LPM) and the logit model. As it is well known in the literature, the first one has some theoretical limitations such as the assumption of the constant marginal effects of the explanatory variables on the studied probability. With respect to the logit model, their results are usually quite similar to those of the probit model. 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.

explanatory variables X, following the general econometric specification:

$$Y_i = F(X_i\beta) + \varepsilon_i, \quad (1)$$

where ε represents the usual random disturbance error, then taking into account the assumption that $E[\varepsilon_i \mid X_i] = 0$, we have:

$$E[Y_i \mid X_i] = E[F(X_i\beta) \mid X_i] + E[\varepsilon_i \mid X_i] = F(X_i\beta). \quad (2)$$

Moreover, if we calculate the conditional expectation value of Y in terms of probabilities, then we will obtain:

$$E[Y_i \mid X_i] = \sum_i Y_i \cdot P(Y_i \mid X_i) =$$

$$= 1 \cdot P(Y_i = 1 \mid X_i) + 0 \cdot P(Y_i = 0 \mid X_i) =$$

$$= P(Y_i = 1 \mid X_i). \quad (3)$$

From this it follows that:

$$E[Y_i \mid X_i] = F(X_i\beta) = P(Y_i = 1 \mid X_i). \quad (4)$$

Considering that the variable Y_i can only take the values 1 and 0, meaning the model implies, therefore, it allocates a certain conditional probability that $Y_i = 1$, denoted by P_i, i.e.:

$$P(Y_i = 1 \mid X_i) = P_i = F(X_i\beta), \quad (5)$$

and consequently:

$$P(Y_i = 0 \mid X_i) = 1 - P_i = 1 - F(X_i\beta). \quad (6)$$

Finally, the model estimates the probability that the policy of the i-th individual records any claim:

$$\hat{Y}_i = \hat{P}_i = F(X_i\hat{\beta}). \quad (7)$$

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

$$F(X_i\beta) = \Phi(X_i\beta) = \Phi(Z_i) = \int_{-\infty}^{Z_i} \phi(s)\, ds, \quad (8)$$

where:

$$\phi(s) = \frac{1}{(2\pi)^{1/2}} e^{-\frac{s^2}{2}} \quad (9)$$

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 magnitude, 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:

$$\frac{\partial P_i}{\partial X_{ji}} = \frac{\partial \Phi(X_i\beta)}{\partial X_{ji}} = \phi(X_i\beta)\beta_j. \quad (10)$$

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:

$$E[Y_i \mid X_{ki} = 1] \text{ and } E[Y_i \mid X_{ki} = 0]. \quad (11)$$

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) the Wald test is used, whose z-statistic, follows a standard normal distribution. In such models, where the endogenous variable takes only values 1 or 0, the usual coefficient of determination R^2 is not valid as a measure of goodness-of-fit. Instead, other alternatives have been developed, such as the McFadden R^2, ranging between 0 and 1 (although its interpretation is not directly comparable to the linear 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 (1978).

3. Estimated 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[1].

[1] As mentioned in section 2, we have also used a logit model in our tests. We have finally chosen a probit specification due to goodness-of-fit criteria.

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					

This choice is made based on the significance of the explanatory variables (at $\rho < 0.05$) and also 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 chosen specification is 76.99%. This value, although not very high, is quite significant[1]. 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.

All variables introduced in the model are qualitative, so their entry is done through dummy variables[2].

It should also be highlighted that some of the initially selected variables have not been significant enough in some specifications we have made, or have shown evidence of multicollinearity; for that reason, they have not been considered in our final adjustment[3].

The results of this estimate, together with the structural analysis that 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 as 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 that the estimated average probability of claims[4] 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[5].

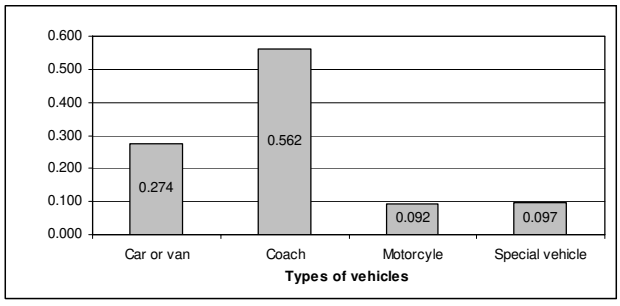

Fig. 1. Estimated average probability of claims by types of insured vehicles

[1] 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 have externally been calculated according to equations (11).

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

[3] 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.

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

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

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 1. 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.

The experience of drivers is 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).

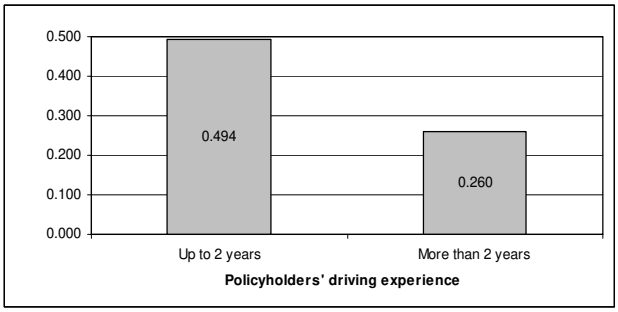

Fig. 2. Estimated average probability of claims by policyholders' driving experience

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 to which the Spanish territory is divided, three have behaved significantly different 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 which is 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 that 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.

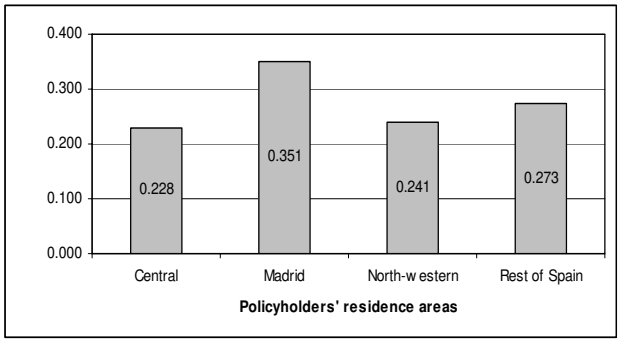

Fig. 3. Estimated average probability of claims by policyholders' residence areas

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 saw in our previous descriptive analysis. This can involve inherent behaviors of the insurance market such as moral hazard and/or adverse selection. We find that this is one of our most important results.

Finally, it is necessary to note that the variables related to age and gender of the insured drivers 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.

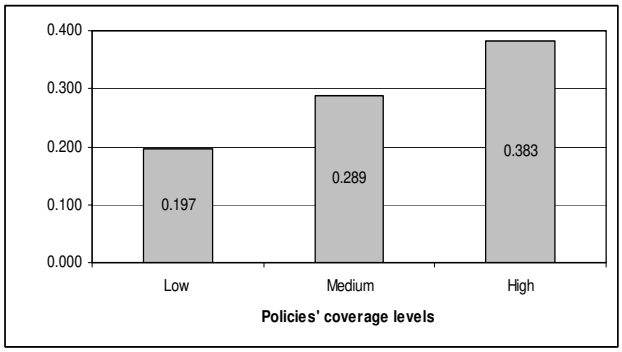

Fig. 4. Estimated average probability of claims by policies' coverage levels

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 focus of 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 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, the most important variables contained in the database in relation to claims report have been carried out to perform the econometric estimation of a probit model. Our final selected specification has allowed us to point out what is the influence of each of these variables with respect to the claims and to estimate their marginal effects, conducting a structural analysis of the results and estimating the average odds of claims for each category considered.

Highlights of this structural analysis are of 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 makes 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 must notice that this last result leads us to think that it could be probably preferable to establish separate models for each level of insurance coverage. In the same way, it could be appropriate to define different models for each type of vehicles, their uses or, even, the driving experience or the area of residence of the policyholders. In this paper, we have preferred to consider jointly all the variables and their respective categories in order to clarify their particular significances in the analysis as a whole. Our conclusions can be now a good starting point to further researches in the future.

Acknowledgements

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

References

1. Boyer M., G. Dionne. An Empirical Analysis of Moral Hazard and Experience Rating // *Review of Economics and Statistics*, 1989, No 71, pp. 128-134.
2. Chiappori P.A., B. Salanié. Testing for Asymmetric Information in Insurance Markets // *Journal of Political Economy*, 2000, No 108 (1), pp. 56-78.
3. DGFSP. Seguros y Fondos de Pensiones. Informe 2008, Madrid, Dirección General de Seguros y Fondos de Pensiones (DGSFP), Ministerio de Economía y Hacienda, 2009, 300 pp.
4. Dionne G., C. Gouriéroux, C. Vanasse. Evidence of Adverse Selection in Automobile Insurance Markets // *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*, Dionne G., C. Laberge-Nadeau (eds.), 1999, pp. 13-46.
5. Hausman J.A. Specification Tests in Econometrics // *Econometrica*, 1978, No 46, pp. 1251-1272.
6. Lee A.H., M.R. Stevenson, K. Wang, K.K.W. Yau. Modeling Young Driver Motor Vehicle Crashes: Data with Extra Zeros // *Accident Analysis and Prevention*, 2002, No 34, pp. 515-521.
7. Melgar M.C., J.A. Ordaz, F.M. Guerrero. The Main Determinants of the Number of Accidents in the Automobile Insurance: An Empirical Analysis // *Études et Dossiers*, 2004. – Nº 286. – pp. 45-56.
8. Melgar M.C., J.A. Ordaz, F.M. Guerrero. 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*, 2005, No 73 (1), pp. 31-54.
9. Melgar M.C., J.A. Ordaz, F.M. Guerrero. Une étude économétrique du nombre d'accidents dans le secteur de l'assurance automobile // *Brussels Economic Review – Cahiers Economiques de Bruxelles*, 2006, No 49 (2), pp. 169-183.

10. Puelz R., A. Snow. Evidence on Adverse Selection: Equilibrium Signalling and Cross-Subsidization in the Insurance Market // *Journal of Political Economy*, 1994, No 102 (2), pp. 236-257.
11. Richaudeau D. Automobile Insurance Contracts and Risk of Accident: An Empirical Test Using French Individual Data // *Geneva Papers on Risk and Insurance Theory*, 1999, No 24, pp. 97-114.
12. Ronis D.L., K.A. Harrison. Statistical Interactions in Studies of Physician Utilization // *Medical Care*, 1988, No 26 (4), pp. 361-372.
13. Shankar V., J. Milton, F. Mannering. Modeling Accident Frequencies as Zero-Altered Probability Processes: an Empirical Inquiry // *Accident Analysis and Prevention*, 1997, No 29 (6), pp. 829-837.

Appendix 1. Definition of the used variables in the econometric analysis

Dependent variable	
CLAIM	Binary variable: 1 for any made claim; 0 otherwise
Explanatory variables	
VEH_TYPE	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_USES	Uses of vehicles: ♦ Dummy variables: PROF_USE (professional use) and OTH_USE (other uses). ♦ Excluded category: personal 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.
EXP<2Y	Driving experience: 1 for less than two years' driving experience; 0 otherwise.
NUTS-1	Large regions or areas (NUTS-1) of policyholders' residence: ♦ Dummy variables: CANARIAS (Islas Canarias), CENTRAL (Central region: Castilla-La Mancha, Castilla y León and Extremadura), CEU_MEL (Ceuta and Melilla), EASTERN (Eastern region: Cataluña, Comunidad Valenciana and Islas Baleares), MADRID (Madrid), NORTEAST (North-eastern region: Aragón, Euskadi, La Rioja and Navarra), NORTWEST (North-western region: Asturias, Cantabria and Galicia). ♦ Excluded category: Southern region (Andalucía and Murcia).
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 (medium coverage level) and COV_HIGH (high coverage level). ♦ Excluded category: low coverage level.