

María del Carmen Melgar, José Antonio Ordaz

## **The Utility of Zero-Inflated Models in the Estimation of the Number of Accidents in the Automobile Insurance Industry\***

**Key words:** *Automobile Insurance Industry, Accidents, Applied Econometrics, Zero-Inflated Models*

**Abstract:** The main purpose of the present paper is to provide an econometric model which estimates the number of automobile accidents that policyholders declare to their insurance companies, pointing out those variables that are significant in this process. Our empirical analysis is based on the data supplied by a private insurance company that operates in Spain, and on the zero-inflated count data models as methodology. We find a positive association between the levels of coverage and the accident rates, suggesting the existence of problems related to adverse selection and moral hazard. This result is one of the most important conclusions of our work and confirms the theoretical aspects pointed up by other empirical studies in the literature. Additionally, estimating the number of policyholders that suffered any accident but not declared, and how many these non-declared accidents are, could be very useful information for insurers to evaluate their risk planning. Our model attempts to reach this target as well.

### **INTRODUCTION**

The automobile insurance is nowadays one of the most important sectors of the insurance industry in the developed economies. In the case of Spain, the premiums of this branch and the cost of their accidents represent 22.44% and 20.24%, respectively, of the whole of the national insurance activity (IEA 2009).

The relevance of these figures justifies the large interest this sector offers to researchers from multiple perspectives. Their own particularities lead to the implementation of different econometric models that intend to contrast some theoretic results of markets with asymmetric information. In this sense, the works by Boyer and Dionne (1989, pp. 128–134), Puelz and Snow (1994, pp. 236–257),

---

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

Dionne *et al.* (1999, pp. 13–46), Richaudeau (1999, pp. 97–114) and Chiappori and Salanié (1997, pp. 943–950; 2000, pp. 56–78) are some essential references we can point out.

The main objective of the present paper 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. In Melgar *et al.* (2005, pp. 31–54), there can be found a first reference in relation with this analysis. But additionally, we can deduce from our methodology that a large part of the policyholders do not declare they have had 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 could be a useful starting point for insurance companies to estimate the benefits of their ‘bonus-malus’ policies.

## THE DATABASE

The database we use in this study has been provided by a private insurance firm that operates in the Spanish automobile sector.

Initially, we have information from 60,000 clients of this company referred to one year’s period: from 16<sup>th</sup> June 2002 to 15<sup>th</sup> June 2003, but we have finally considered a random sample of 15,000 observations due to computational reasons.

Available information has been initially classified in three different categories: variables related to the insured vehicle (type and use); about the personal characteristics of policyholders (age, gender, driving experience and area of residence); and features of the policy (premium and level of coverage). Additionally, a fourth type of information is considered: the database contains the number of declared accidents corresponding to each policy; this question is essential for the final purpose of this work.

All these variables are usually chosen (and recorded) by the insurance companies to fix their premiums and estimate their risk pools. Therefore, they constitute the starting point for our study.

Table 1 shows, for every variable, the percentages that each one of their categories represents over the whole and, within them, the percentages that registered some accident.

The vehicle’s type offers five different possibilities. ‘Car or Van’ represents 80.5% of the whole of vehicles. After that, we can find the categories referred to ‘Special vehicle’ and ‘Motorcycle’; they represent 10.3% and 7.7%, respectively. The other categories (‘Truck’ and ‘Coach’), jointly, are only 1.5%. 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. After this category, we find ‘Professional’ with 19.6%. Finally, ‘Other uses’ are only 0.6%.

Table 1. Distribution of categories and accident rates by variables

Variable Category	Percentages over the whole of sample data	Percentages with accidents, over the whole of data of the corresponding category
Types of the insured vehicles		
Car or Van	80.5%	26.5%
Truck	1.3%	25.3%
Coach	0.2%	52.2%
Motorcycle	7.7%	7.0%
Special vehicle	10.3%	6.8%
Uses of the insured vehicles		
Private	79.8%	24.7%
Professional	19.6%	16.3%
Other uses	0.6%	12.0%
Age of policyholders		
[18–25] years old	3.1%	23.4%
(25–45] years old	39.8%	24.2%
(45–70] years old	51.8%	22.7%
More than 70 years old	5.3%	15.9%
Gender of policyholders		
Male	85.3%	22.3%
Female	14.7%	26.5%
Driving experience of policyholders		
Less than 2 years	0.8%	35.5%
2 years or more	99.2%	22.9%
Address area of policyholders		
Southern	46.3%	24.0%
Canary Islands	4.4%	21.3%
Central	16.8%	19.0%
Ceuta & Melilla	0.4%	25.0%
Eastern	12.1%	24.4%
Madrid	1.4%	28.7%
North-eastern	3.2%	24.4%
North-western	15.4%	22.7%

Table 1 – continued

Riable Category	Percentages over the whole of sample data	Percentages with accidents, over the whole of data of the corresponding category
Annual premium		
(0–300] €	32.2%	11.8%
(300–400] €	26.8%	22.6%
(400–600] €	23.2%	28.1%
> 600 €	17.8%	36.9%
Level of coverage		
Low	54.3%	16.1%
Medium	37.8%	29.3%
High	7.9%	39.4%
Total	100.0%	22.9%

Source: own calculations based on the study database.

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 average age of the drivers in our database is quite high: 48 years old and only 3.1% of them are less than 26 years old. With respect to the gender, 85.3% of policyholders of the database are males. Driving experience of policyholders is another relevant feature that is taken in account by insurance companies to fix the premiums. A different treatment is carried out when insured drivers have a driving license experience less than 2 years. They only represent 0.8% of the whole of our registers. Traffic area of policyholders (or their address area taken as a proxy) is the last aspect that insurance companies take into consideration in this section. Related to this variable, we have grouped the original available data in the 8 NUTS-1 or geographic groups of regions that are considered by *Eurostat* for statistical purposes in the case of Spain. According to this classification, we find that the most represented region is the ‘Southern’ with 46.3% of all our insured people.

The two main elements that define an insurance policy are the premium and the level of coverage. The premium is the annual payment of insured people to the company; we have grouped this variable in four categories. More than a half of policyholders pay 400 € as a maximum (59.0%). The highest premium (corresponding to more than 600 €) is the category with the least frequency: it represents 17.8% of policyholders. With respect to the levels of insurance coverage, we have defined three levels as a function of the contracted warranties in policies and the types of vehicles: low, medium and high. Globally, over half of all policyholders have the lowest level of coverage (54.3%). When this level increases,

the portion of drivers who contract it decreases: 37.8% have the medium level of coverage and, finally, 7.9% subscribe the highest level of coverage.

As mentioned earlier, one of the main objectives of this paper is to determine the most significant explanatory variables in the estimation process of the number of accidents that are declared by the policyholders to their companies. In this sense, we have this information in our database, because it contains the exact date of the accidents, their description, associated guilt and cost.

In the present paper, we have only focused on the occurrence and number of accidents associated to each policy. We want to study if the accidents (and their number) are declared or not to the company by the policyholders.

The last column of Table 1 offers information related to the accident rates of each one of the categories of the variables we have just described above. Within all these numbers, the figures related to the different levels of coverage are of special relevance: 16.1%, 29.3% and 39.4% are the percentages corresponding to the lowest, medium and highest levels, respectively. Therefore, we can observe an increasingly positive relationship between accidents and levels of coverage, thus suggesting the existence of some type of strategic behaviour by the insured population. At last, we will try to verify this crucial aspect with the support of the econometric analysis.

On the other hand, in Table 2 we offer the distribution of the number of accidents declared by the insured people that compose the sample. The distribution of policyholders that declare some accident is considered as well. From these figures, we can point out that only 3,442 policyholders declared some type of accident, representing 22.9% (or, more exactly, 22.95%) of the whole sample. Another remarkable aspect is that 90.30% of them only declared a maximum of 2 accidents.

Table 2. Distribution of the number of accidents

Number of accidents	Frequencies	Including 0		Not including 0	
		Percentages	Cumulative percentages	Percentages	Cumulative percentages
0	11,558	77.05	77.05		
1	2,365	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	15,000	100.00		100.00	

Source: own calculations based on the study database.

**METHODOLOGY**

Our endogenous variable, the declared number of accidents, have only nonnegative discrete values. 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 regressions.

In our descriptive analysis, we found that 77.05% of all the policyholders declared that they have not had any accident. But this 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.

So, it can be supposed that a large number of these zeros are not actual zeros. In such situations, zero-inflated count data models (either Poisson or negative binomial specifications) appear as the most appropriate solutions because they can distinguish these qualitative differences (Cameron and Trivedi 1998; Winkelmann 2003).

In these models we start defining the following random variables:  $Z_i$  stands for the actual number of suffered accidents, and  $I_i$  refers to the number of declared accidents, where  $I_i$  is a Bernoulli random variable independent of  $Z_i$  defined as:

$$I_i = \begin{cases} 1, & \text{if the policyholder decideto declaretheir accidents} \\ 0, & \text{otherwise} \end{cases}$$

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

$$q_i = F(\tau(\beta_0 + \beta_1 X_{i1} + \dots + \beta_n X_{in})) \tag{1}$$

where  $X_{in}$  are the explanatory variables and  $F$  is a cumulative distribution function, typically chosen to be either logistic or standard normal, leading to the logit or probit models respectively, where  $\tau$  are the unknown parameters of the model to be estimated.

According to these assumptions, the zero-inflated probability distribution will be given by the expression:

$$\begin{aligned} P(Y_i = 0) &= q_i + (1 - q_i)P(Z_i = 0) \\ P(Y_i = k) &= (1 - q_i)P(Z_i = k) \quad k = 1, 2, \dots \end{aligned} \tag{2}$$

or, alternatively, by:

$$P(Y_i = k) = q_i(1 - \min\{k, 1\}) + (1 - q_i)P(Z_i = k) \quad k = 0, 1, 2, \dots \tag{3}$$

Depending on the particular probability distribution that follows, Poisson or negative binomial, we will then have the zero-inflated Poisson (ZIP) model or the zero-inflated negative binomial (ZINB) model, respectively:

$$P(Y_i = k) = q_i(1 - \min\{k, 1\}) + (1 - q_i)e^{-\lambda_i} \frac{\lambda_i^k}{k!}, \quad k = 0, 1, 2, \dots, \tag{4}$$

$$P(Y_i = k) = q_i(1 - \min\{k, 1\}) + (1 - q_i) \frac{\Gamma(k + \nu)}{\Gamma(k + 1) \cdot \Gamma(\nu)} \left(\frac{\nu}{\nu + \lambda_i}\right)^\nu \left(\frac{\lambda_i}{\nu + \lambda_i}\right)^k, \quad k = 0, 1, 2, \dots \tag{5}$$

where  $\nu$ , in both cases, and  $\lambda_i$  is the characteristic parameter associated to the gamma distribution included in a negative binomial model.

We can obtain additional results in our research. 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:

$$N_0 = \sum_{i=1}^N \chi_{\{z_i > 0, I_i = 0\}} \tag{6}$$

and

$$A_u = \sum_{i=1}^N Z_i \chi_{\{z_i > 0, I_i = 0\}} \tag{7}$$

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:

$$E(N_0) = \sum_{i=1}^N q_i(1 - P(Z_i = 0)) \quad E(A_u) = \sum_{i=1}^N q_i E(Z_i) \tag{8}$$

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

$$\sum_{i=1}^N q_i(1 - P(Z_i = 0))\{1 - q_i(1 - P(Z_i = 0))\} \tag{9}$$

$$\sum_{i=1}^N q_i \{Var(Z_i) + E(Z_i)^2(1 - q_i)\}$$

After choosing the particular probability distribution for  $\lambda$ , we may estimate its parameters using the maximum likelihood method. We can then analyze the significance of the explanatory variables of the model, and finally obtain the estimates for (8) and (9), thus specifying confidence intervals both for the 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, pp. 307–333):

$$V = \frac{\sqrt{N} \left[ \frac{1}{N} \sum_{i=1}^N m_i \right]}{\sqrt{\frac{1}{N} \sum_{i=1}^N (m_i - \bar{m})^2}} \quad (10)$$

where  $f_0$  and  $f_1$  are the functions of distribution corresponding to the zero-inflated and the ‘traditional’ specifications of the models, respectively, and  $\bar{m}$  is the mean of  $m$ .

This statistic follows a standard 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.

## MAIN RESULTS

We have worked with both ZIP and ZINB models and have compared their respective results. Even though they are very similar, we think the ZIP model could theoretically be the best choice because it is logical to suppose that the behaviour of the actual number of accidents taking place during a time interval follows a typical Poisson law of probability distribution. The usual assumed limitation of the Poisson law related to the equality between their mean and their variance should not be seen as a limitation in this case, because when the number of observations is very large (such as in our case, of  $N = 15000$ ) and the probability of the having an accident,  $p$ , is very small (for instance, we could think that  $p = 1/365$  for the case when there is only at most one accident per year), the equality between the mean and the variance is just the result we should find.<sup>1</sup>

<sup>1</sup> There exist statistical methods of choosing between ZIP and ZINB models. The most usual of them consists of studying the significance of the so-called *precision parameter*, defined

We have used *Limdep v. 7.0* as econometric software to carry out the estimation process. After having compared a large number of different possibilities of the ZIP model, we have finally chosen the specification as shown in Table 3, where all the parameters are significant at a level of confidence of 95% (even 99%).

Table 3. Final output of the ZIP model

Dependent variable: NUMACC Logistic distribution used for splitting model Total number of included observations: 15,000 Actual zeros: 11,558			
Zero-Inflated Poisson model			
Variable	Coefficient	z-Statistic	P-value
CONSTANT	-0.33009	-11.061	0.0000
COACH	0.77351	4.780	0.0000
MOTORCYC	-0.82654	-7.960	0.0000
SP_VEH	-0.74643	-8.684	0.0000
OTH_USE	-0.56710	-3.072	0.0021
EXP<2Y	0.52877	4.323	0.0000
CENTRAL	-0.17363	-4.796	0.0000
NORTWEST	-0.10281	-3.246	0.0012
COV_MED	0.28886	9.054	0.0000
COV_HIGH	0.54430	10.872	0.0000
ZI model parameter:	-0.99467	-5.261	0.0000
Log. Likelihood	-10765.8		
Predicted zeros	11909		
Vuong statistic	39.3839		

Source: own study.

As can be seen from the value of the Vuong statistic, and from the significance of the zero-inflated parameter, this model is preferred to its 'traditional' form.

The goodness-of-fit of the model can be studied through the comparison between the recorded distribution of the number of declared accidents and the estimated probability for each one of these numbers (Dionne and Vanasse, 1992, pp. 149–165). Such estimations,, can be calculated by taking into account equation (2), obtaining:

---

as, where  $\lambda$  is the parameter referred in (5). Other methods can be seen in Ridout *et al.*, (2001, pp. 219–223).

$$\hat{P}_k = \frac{1}{N} \sum_{i=1}^N \left( \hat{q}_i + (1 - \hat{q}_i) e^{-\hat{\lambda}_i} \right) \quad k = 0$$

$$\hat{P}_k = \frac{1}{N} \sum_{i=1}^N (1 - \hat{q}_i) e^{-\hat{\lambda}_i} \frac{\hat{\lambda}_i^k}{k!}, \quad k = 1, 2, \dots$$
(11)

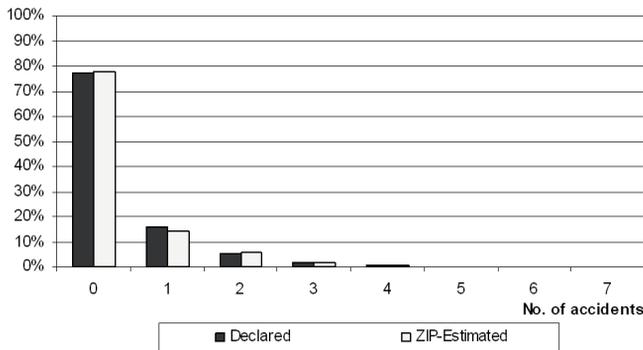
where  $\hat{\lambda}_i = \exp \left\{ \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_n X_{in} \right\}$

and

$$\hat{q}_i = \frac{\exp \left\{ \hat{\tau} \left( \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_n X_{in} \right) \right\}}{1 + \exp \left\{ \hat{\tau} \left( \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_n X_{in} \right) \right\}}$$

Figure 1 shows the graphical representation of the indicated analysis<sup>2</sup>.

Figure 1. Distribution of the number of declared accidents and estimated probability of suffering accidents by ZIP model



Source: own study.

The estimating process shows those variables which are significant to explain the number of declared accidents (Table 3)<sup>3</sup>. If we analyze the *types of vehicles*, we find that the categories referred to ‘Coach’, ‘Motorcycle’ and ‘Special vehicle’ have a significantly different behaviour in relation with all the rest, i.e., cars-vans, and trucks. On the one hand, coaches 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. In

<sup>2</sup> The Wilcoxon signed-rank test (Wilcoxon, 1945, pp. 80–83) could be used to analytically corroborate the statistical similarity between both the observed and the estimated series. The result we obtained confirms this aspect for a confidence level close to 100%.

<sup>3</sup> Even though the marginal effects of each variable can be computed, we have not done it in this study because our main objective was only to analyze their significance in the process.

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

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 have higher probability of having accidents. With respect to the region of policyholders' residence, we also observe that some of them ('Central' and 'North-western') appear as significant against the rest of them.

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. We find that the higher the levels of insurance coverage the higher the accident rates. This result suggests the existence of problems related to adverse selection and moral hazard, and confirms the theoretical aspects pointed up by other empirical studies, such as by Dionne *et al.* (1999, pp. 13–46), Richaudeau (1999, pp. 97–114) and Cohen (2005, pp. 197–207).

Finally, we must note that the gender of policyholders has not been significant enough. The  $p$ -value associated to this variable has not reached, in any case, the value of 0.01.

After having estimated the zero-inflated specification of our Poisson model, we have passed to the second part in our study in order to determine the number of 'extra-zeros' indicated by the model, that is, the number of policyholders that declare no accidents although they have had some. Additionally, we have estimated the number of accidents non-reported by this way. This extension is not usual in the literature.

Applying our ZIP model to the corresponding expressions (8) and (9), the number of estimated 'extra-zeros' is 4,304; this leads to the estimated number of non-reported accidents of 6,221 (Table 4). These figures indicate that 28.69% of the whole of policyholders did not declare their accidents or, alternatively, 37.24% of policyholders that declared they had no accidents, yet they actually had at least one. They likely opted for this decision in order not to be punished by their company. With respect to the estimated number of non-declared accidents, they would represent 55.37% of the theoretical whole of happened accidents.

Table 4. Number of estimated 'extra-zeros' and undeclared accidents

	ZIP-Estimated (95% confidence interval)
Number of 'extra-zeros' / Policyholders who did not declare their accidents	4304 (4196-4412)
Number of undeclared accidents (corresponding to the 'extra-zeros')	6221 (6040-6402)

Source: own study.

## CONCLUSIONS

The more in-depth understanding of the factors and variables that could have some influence on the mechanisms for determining the occurrence and reporting of accidents may be very useful to the insurers, for the sake of achieving improvements in efficiency, in risk management. Based on information from an insurance company database with about 15,000 Spanish automobile policies, we provide an econometric model to estimate outstanding claims by policyholders.

In our preliminary descriptive analysis, we found that 77.05% of all the policyholders declared that they have not had any accident ever. Under the pressure of a 'bonus-malus' insurance system, there could be a number of accidents that happen but that they are not declared, *i.e.* there could be a significant of zero records that are not real zeros. In these cases, zero-inflated count data models (either Poisson or negative binomial) appear as the most appropriate solutions. In particular, we have finally chosen a zero-inflated Poisson specification.

Our modelling process has shown the significance of the variables usually taken into account by the insurance companies to explain the number of claims, such as uses and types of vehicles, driving experience of the policyholder, and their region of residence. Within all these factors, we should point out the relevance of another one: the level of insurance coverage contracted by customers. We confirmed that the higher the levels of insurance coverage the higher the accident rates. This result seems to be a characteristic of markets with asymmetric information like insurance markets are.

The most innovative aspect of this work is the subsequent estimation we have made of the number of policyholders who do not report their accidents and how many they represent. These figures hide an interesting behaviour.

As mentioned above, knowing all this information, the insurance companies could evaluate the success of the implementation of their particular risk policies.

## REFERENCES

- Boyer M., Dionne G., (1989), *An Empirical Analysis of Moral Hazard and Experience Rating*, "Review of Economics and Statistics", 71.
- Cameron A. C., Trivedi P. K., (1998), *Regression Analysis of Count Data*, Cambridge University Press, Cambridge.
- Chiappori P. A., Salanié B., (1997), *Empirical Contract Theory: The Case of Insurance Data*, "European Economic Review", 41.
- Chiappori P. A., Salanié B., (2000), *Testing for Asymmetric Information in Insurance Markets*, "Journal of Political Economy", 108, 1.
- Cohen A., (2005), *Asymmetric Information and Learning: Evidence from the Automobile Insurance Market*, "Review of Economics and Statistics", 87, 2.
- 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.
- Dionne G., Vanasse C., (1992), *Automobile Insurance Ratemaking in the Presence of Asymmetrical Information*, "Journal of Applied Econometrics", 7, 2.
- IEA - Instituto de Estadística de Andalucía, (2009), *Estadísticas del Sistema Financiero en Andalucía. Entidades de seguros. Año 2007*, Junta de Andalucía, Sevilla.
- 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", 73, 1.
- Puelz R., Snow, A., (1994), *Evidence on Adverse Selection: Equilibrium Signaling and Cross-Subsidization in the Insurance Market*, "Journal of Political Economy", 102, 2.
- 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.
- Ridout M., Hinde J., Demétrio C. G. B., (2001), *A Score Test for Testing a Zero-Inflated Poisson Regression Model against Zero-Inflated Negative Binomial Alternatives*, "Biometrics", 57.
- Vuong G. H., (1989), *Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses*, "Econometrica", 57.
- Wilcoxon, F., (1945), *Individual Comparisons by Ranking Methods*, "Biometrics", 1.
- Winkelmann R., (2003), *Econometric Analysis of Count Data*, Springer-Verlag, Berlin.

**ANALIZA LICZBY WYPADKÓW ODNOTOWANYCH  
PRZEZ PRZEMYSŁ UBEZPIECZEŃ SAMOCHODOWYCH:  
UŻYTECZNOŚĆ MODELI ZERO-ZAWYŻONYCH**

**Słowa kluczowe:** *przemysł ubezpieczeń samochodowych, wypadki, ekonometria stosowana, modele zero-zawyżone*

**Abstrakt:** Głównym celem artykułu jest przedstawienie modelu ekonometrycznego, który szacuje liczbę wypadków, zadeklarowaną przez ubezpieczonych w swoich firmach ubezpieczeniowych, zwracając uwagę na takie zmienne, które są istotne w tym procesie. Analiza empiryczna opiera się na danych dostarczonych przez hiszpańską, prywatną firmę ubezpieczeniową i na zero-zawyżonym modelu danych. Istnieje dodatni związek między poziomem okoliczności objętych polisą ubezpieczeniową i liczbą wypadków, co sugeruje istnienie problemów związanych z negatywnym wyborem i hazardem moralnym. Wynik ten jest jednym z najważniejszych wniosków z pracy i potwierdza teoretyczne aspekty wskazane w innych badaniach empirycznych w literaturze. Ponadto szacując liczbę ubezpieczonych, którzy ucierpieli w zrządzeniu losowym, ale nie zadeklarowali tego oraz szacując liczbę niezgłoszonych wypadków, firmom ubezpieczeniowym łatwiej byłoby zarządzać ryzykiem. Przedstawiony model stara się również osiągnąć ów cel.