ELSEVIER

Insurance: Mathematics and Economics 24 (1999) 67–81

INSURANCE MATHEMATICS & ECONOMICS

Modelling different types of automobile insurance fraud behaviour in the Spanish market

Manuel Artís, Mercedes Ayuso, Montserrat Guillén *

Dept. d'Econometria, Estadística i Economia Espanyola, Universitat de Barcelona, Diagonal 690, 08034 Barcelona, Spain

Received 1 September 1997; received in revised form 1 June 1998

Abstract

From a microeconomic point of view, the control of insurance fraud requires a detailed knowledge of the insureds' behaviour. In this paper, we present discrete-choice models for fraud behaviour and we estimate the influence of the insured and claim characteristics on the probability of committing fraud. Data correspond to a Spanish sample. Correction for choice-based sampling is introduced in the estimation due to the oversampling of fraud claims. The structure of the Spanish automobile insurance market is also discussed. Our results differ according to the type of fraud behaviour that is under consideration.

Keywords: Automobile insurance; Fraud; Discrete-choice models; Microeconometrics

1. Introduction

During the past decades, models of fraud control have received increasing attention among insurance theorists. Fraud is considered a type of moral hazard (see Cummins and Tennyson, 1996 and Dionne et al., 1993) because the insured has more information on the accident than the insurer.

Although the theoretical aspects of automobile insurance fraud can be found in the literature, applied work remains limited. In this paper, we present some empirical results based on an econometric model of fraud behaviour.

The automobile insurance provides an excellent opportunity for studying fraud and the results are of great interest to the industry. One main problem is that the automobile insurance products and legislations are still fairly different from country to country. Moreover, the insured behaviour is also very diverse. Some aspects related to fraud in medical expenses (Marter and Weisberg, 1991) which are important in the US market, seem less relevant in the European context.

Previous studies of automobile insurance fraud fall into two broad categories: those that use aggregate industry data and study the general effect of fraud on the economic system and those that identify general trends of suspicious or fraudulent claims.

* Corresponding author. Tel.: +34-93-4021409; fax: +34-93-4021821; e-mail: guillen@eco.ub.es

0167-6687/99/$ – see front matter
PII: S0167-6687(98)00038-9

Picard (1996) characterizes the equilibrium of an insurance market where opportunist policyholders may file fraudulent claims and shows that the insurance market benefits from transferring the monitoring costs to a budget-balanced common agency.

Derrig and Ostaszewski (1995) apply fuzzy set theories to classify and identify characteristics of fraudulent claims. In a previous work, Weisberg and Derrig (1993) has already established a linear regression model in order to analyse the statistical significance of some fraud indicators (insured characteristics, accident circumstances, medical treatment,. . .). Using those fraud indicators, Brockett et al. (1995) apply neural networks to classify claims. In fact, these works establish flags aimed to detect fraud in the underwriting and claims processes.

Our approach begins by studying the behaviour of individual fraud from a theoretical perspective and finally an econometric model is estimated.

Section 2 establishes the general background of utility functions. Section 3 refers to the econometric methodology used in the paper. A general choice tree is presented and there is a discussion on how the framework of discrete choice models for consumers may be implemented to deal with automobile insurance fraud.

Section 4 gives an overview of the evolution of the Spanish automobile insurance market. We also discuss the main types of automobile insurance fraud in the Spanish context. The data set and simple classification of claims are presented. We distinguish between legitimate claims and several kinds of fraud. In fact, our classification is analogous to the one suggested by Weisberg and Derrig (1991, 1993). Additionally, we also consider a sequential decision-making process. In other contexts (see, for example, Dionne et al. (1996), where classification takes place in several steps), it is usually the case that choices occur with different probability depending on previous actions. Moreover, it is generally accepted that an insured would first decide not to plan a fraud, but could possibly choose to take advantage of an accident in order to fabricate an opportunistic fraud.

In Section 5, the estimation results of a multinomial and a nested discrete-choice model are presented. Their properties are discussed in the context of fraud assessment.

Exploratory studies such as the work by Weisberg and Derrig (1993) are devoted to a quantitative analysis. Their objective is finding quantitative indicators of fraud in order to derive implications for improved claim handling. Therefore, they study the potential value of fraud indicators. Belhadji and Dionne (1997) also present a binary choice model to detect fraud. In our analysis, we focus on a causal model of fraud that shows how different individual characteristics may affect the probability of fraud. We also compare how different types of fraud may have different causes (or may have the same causes but with a different intensity).

Finally, several relevant empirical results concerning the Spanish market are also presented.

2. The model for fraudulent behaviour

Utility is essentially a measure of overall benefit to the individual. The consumer behaviour can be modelled using the theory of utility maximization. Individuals are assumed to act according to the choice that is most favourable to them. Hoyt (1990) states that the expected utility of fraudulent behaviour in insurance is equal to the sum of two components

$$E[U(F)] = (1 - p)U(f) + pU(P), \tag{1}$$

where, $E[U(F)]$ is the expected utility from fraud; f is the amount of fraudulent payout; P is the amount of punishment; $U(f)$ is the utility of fraudulent payout f; $U(P)$ is the disutility of punishment P; $(1 - p)$ is the probability of fraudulent payout and p is the probability of punishment.

Thus, the first component is associated to the amount of fraudulent payout. The second term refers to the disutility of punishment. This approach is based on the monetary benefit of committing fraud and the penalization when fraud is detected. Besides, the probability of fraudulent payout and the probability of penalization are also taken into account.

Another important study of fraud from the expected utility point of view can be found in Picard (1996).

Our aim is to provide a model that is useful to study causes of fraud at an individual level. It should account for different types of fraud and explain individual behaviour. We propose to distinguish the type of fraud chosen, because we assume that different kinds of fraud may produce quite different benefits to the individual. The causes affecting each behaviour may also differ.

Several approaches can be used when dealing with each fraud–legitimate classification. For example, the data set used by Weisberg and Derrig (1993) takes into account the following types of claims: legitimate, planned fraud, build-up fraud and opportunistic fraud. This kind of information is not available for the Spanish sample that is used in our application. Our approach is adapted to the classification in our data set. Nevertheless, the theoretical and methodological presentation is also applicable to other divisions of the fraud classes.

Essentially, our data set considers two possibilities for fraud: (1) the insured wants to obtain a payout for himself or (2) the insured is trying to benefit a third person (usually another driver).

A discrete-choice model provides a suitable framework to estimate the influence of exogenous variables in the observed behaviour (Agresti, 1990). Hence, we propose the application of a multinomial logit model or a nested logit model.

In spite of the particular structure of the possible set of choices, there exists a latent economic criterion in the choice: the insured wants the largest utility. In order to consider several types of fraud, we extend the expected utility model (Hoyt, 1990), by introducing two separate equations. We consider the expected utility of fraud for personal profit and the expected utility of fraud to benefit a third person. This approach is appropriate in the Spanish automobile insurance framework, where premium coverages are divided into two main classes: only liability insurance or full insurance. We argue that, under these circumstances, drivers who are involved in a particular accident may reach an agreement. Thus, fraud takes places not only when the insured wants to obtain a payout, but also when a third party does.

Our model for expected utility of fraud is as follows. $E(U(F_1))$ indicates the expected utility of fraud for self-benefit and $E(U(F_2))$ is the expected utility of fraud to benefit a third party. In general, let us assume that

$$E(U(F_r)) = (1 - p_r)U_r + p_r U(P), \quad r = 1, 2, \tag{2}$$

where p_r is the probability of detecting each type of fraud, U_r the utility of the fraudulent payout and $U(P)$ is the disutility associated with an amount of punishment. The expected utility of a legitimate claim is also considered.

In fact, as suggested by Hoyt (1990), fraud control should focus on reducing the expected utility of insurance fraud. We study the probability of making fraud. The probability of detecting fraud may also differ for each type of fraud. For example, it is generally more difficult to find out if the insured tries to benefit another driver. On the other hand, the amount received in this case is lower if compared to the fraudulent payout when there is a personal benefit.

The econometric model shows which variables influence the utility of fraud. The model provides a decision tool to identify suspicious claims, thus increasing the probability of detection. We assume that the punishment is quite stable from one individual to another. Insurance companies tend to apply a balanced rule in order to protect the rights of legitimate claimants. It is very common that punishment implies returning the accident costs and/or putting the insured in a lower bonus-malus class.

The theory of random utility is the theoretical foundation of the econometric model that is used to discuss the choice among the set of possible alternatives. Like in the binary logit model (when there are only two possible choices), the utility refers to an unobserved variable indicating the propensity of each individual to choose one of the possible alternatives of the choice set. Therefore, the utility of choosing a legitimate claim is also considered, because it is another possible behaviour. The theory of utility maximization has a stochastic character. It has a deterministic part associated to a set of explanatory variables. Additionally, it contains a random component associated to a certain error term, accounting for the associated effect of variables not included in the deterministic part of the model that may also influence the choice (Maddala, 1983, McFadden, 1983 and Greene, 1997).

Let us consider a random utility function such as

$$U_{ir} = \beta_r' x_i + e_{ir}, \tag{3}$$

Table 1
Expected cost of the claim

Type of claim	Classified as	
	Legitimate	Fraud
Legitimate	M	$c_1 + M$
Fraud	M	c_2

M is the cost of the accident that must be paid according to the policy contract, c_1 is the cost of investigating a suspicious claim which is not fraudulent (legitimate) and c_2 is the amount spent in identifying a fraud.

where i indicates the ith individual in the sample ($i = 1, \ldots, n$), r the alternative to which the utility refers ($r = 0, \ldots, R-1$), x_i the vector of explanatory variables of the ith individual and β_r are the vectors of unknown parameters to be estimated. Finally, e_{ir} is the error term associated to the rth utility of the ith individual. For our purpose $R = 3$, because the set of choices includes three alternatives.

We consider a situation where the insurance company must control fraud. In fact, once an accident has taken place, the company must pay according to the policy agreement but it must follow a suitable strategy in order to identify fraudulent claims.

The classification of a claim and its associated costs are detailed in Table 1. The arrangement is very simple, since the two types of fraud have been collapsed into only one cell.

Given the set of individual characteristics, let us assume that the probability of committing fraud is equal to $(1-q)$, the probability of classifying a legitimate claim as fraud is equal to p_1. On the other hand, a fraud claim is not identified with probability p_2.

Then, the expected cost of the claim, $E(c)$, can be written as

$$E(c) = (1-q)[p_2 M + (1-p_2)c_2] + q[(1-p_1)M + p_1(c_1 + M)]. \tag{4}$$

This expression is equivalent to

$$E(c) = M + p_1 c_1 + (1-q)[(1-p_2)(c_2 - M) - p_1 c_1]. \tag{5}$$

Our model provides a method to estimate the probability of (each type of) fraud given the claim characteristics. Since the cost parameters (M, c_1 and c_2) are known and p_1, p_2, are the conditional error probabilities that can also be inferred from the results, it follows that the expected claim costs can also be estimated. This framework can be generalized to accommodate distinct fraud divisions.

The econometric model will provide rules in the tarification process and establish guidelines for the issuing of policies, once the most influential variables are identified.

3. Discrete-choice models

In the framework of discrete-choice models, our fraud model classifies each claim into one of several different classes: legitimate, fraud for personal profit and fraud for a third party benefit.

We have envisioned two different approaches: firstly we will consider a multinomial logit model (MNL) and, secondly, a nested multinomial logit model (NMNL) will be used.

3.1. Multinomial logit model

In the multinomial logit model context, we assume that for each individual i ($i = 1, \ldots, n$) a vector of dependent variables is observed (Y_{i0}, Y_{i1}, Y_{i2}). Y_{i0} is equal to one if the ith claim is legitimate and zero otherwise. Similarly, Y_{i1} equals one when the claim is fraud for self-benefit and Y_{i2} equals one for fraudulent claims that benefit a third party.

In this case, the insured has to make a single choice among three alternatives. As the unordered-choice models can be motivated by a random utility model (McFadden (1983)), we can use the utility function presented in (3) in order to calculate the estimated individual probabilities.

If, for example, the insured chooses to commit fraud for self-benefit ($Y_{i1} = 1$), we assume that U_{i1} is the maximum among the three utilities. Hence, the observed variable, Y_{i1}, is defined as

$$Y_{i1} = f(U_{i1}) = 1 \quad \text{if} \quad U_{i1} = \text{Max}(U_{i0}, U_{i1}, U_{i2}), \qquad Y_{i1} = f(U_{i1}) = 0 \quad \text{otherwise.} \tag{6}$$

Likewise, we obtain accordingly definitions for Y_{i0} and Y_{i2}.

The statistical model is based on the probability that choice k is made, which is

$$P_{ik} = P(Y_{ik} = 1|x_i) = P(U_{ik} > U_{im}, \ \text{for all} \ m \neq k|x_i) \quad k = 0, 1, 2. \tag{7}$$

Therefore, [1]

$$P(\beta_k' x_i + e_{ik} > \beta_m' x_i + e_{im}) = P(e_{im} < e_{ik} + \beta_k' x_i - \beta_m' x_i). \tag{8}$$

As shown by Maddala (1983), if the residuals e_{ir} are independent and identically distributed with the type I extreme-value distribution, [2] we can write

$$\begin{aligned} P(Y_{ik} = 1|x_i) &= \int_{-\infty}^{\infty} \left[\prod_{m \neq k} F(e_{ik} + \beta_k' x_i - \beta_m' x_i) \right] f(e_{ik}) \mathrm{d}e_{ik} \\ &= \frac{\mathrm{e}^{\beta_k' x_i}}{\sum_{r=0}^{R} \mathrm{e}^{\beta_r' x_i}} = \frac{\mathrm{e}^{\beta_k' x_i}}{\sum_{r=0}^{2} \mathrm{e}^{\beta_r' x_i}}, \quad k = 0, 1, 2. \end{aligned} \tag{9}$$

This model is called the multinomial logit model. The estimated equations provide a set of estimated conditional probabilities for the three choices given the insured characteristics x_i. To solve an indeterminacy in the model, a convenient normalization is to assume that $\beta_0 = 0$. Then the probabilities become

$$P(Y_{ik} = 1|x_i) = \frac{\mathrm{e}^{\beta_k' x_i}}{1 + \sum_{r=1}^{2} \mathrm{e}^{\beta_r' x_i}}, \quad k = 1, 2, \tag{10}$$

$$P(Y_{i0} = 1|x_i) = \frac{1}{1 + \sum_{r=1}^{2} \mathrm{e}^{\beta_r' x_i}}. \tag{11}$$

As we can see, the model contains a set of parameters (β_r) associated to each alternative (except for the first choice, due to the identification condition).

Parameter estimates are obtained by maximum likelihood, together with their standard error estimates. In this paper, an iterative method is used as implemented in LIMDEP (Greene, 1995).

The coefficients in the multinomial model are difficult to interpret and, therefore, some authors use the marginal effects of the regressors on the probabilities, based on the parameter estimates. Log-odds probability ratios may also be computed using the insured characteristics and the parameter estimates (Greene, 1997).

3.2. Nested multinomial logit model

Let us consider that the final insured behaviour is decided step by step. In the first level, the individual considers the possibility of committing fraud or not. If the first alternative is chosen, the insured decides between two possible fraud types. Fig. 1 shows the nested choice decision tree.

[1] Note that conditioning on x_i has been eliminated because we assume that the explanatory variables are fixed.

[2] The cumulative type I extreme-value distribution function is $F(\varepsilon_i < \varepsilon) = \exp(-\mathrm{e}^{-\varepsilon})$ and the probability density function is $f(\varepsilon_i) = \exp(-\varepsilon_i - \mathrm{e}^{-\varepsilon_i})$ (see Maddala, 1983).

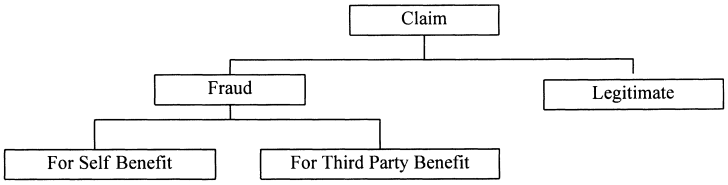

Fig. 1. Tree structure specified for the nested multinomial logit model.

The insured is assumed to make a choice among three final alternatives (legitimate, self-benefit fraud and third party benefit fraud) in the choice set. These are the elementary alternatives. The model contains two branches: fraud or legitimate claim. Since there is only one limb (claim), the model has two levels.

Individuals choose one of the alternatives at the lowest level of the tree. They maximize their utility step by step rather than globally. Thus, they choose one and only one branch.

In Greene (1995), we can find the mathematical specification of the model. We denote by $k|j$ the choice of alternative k in branch j (legitimate, fraud). The number of alternatives in branch one is $N_{k|1} = 2$ and the number of alternatives in branch two is $N_{k|2} = 1$. This happens because once an honest behaviour has been chosen, there is no other alternative. On the other hand, if the insured is dishonest, two alternatives are possible.

The conditional probability of alternative k in branch j, $k|j$ is

$$P(k|j, x_i) = \frac{e^{\beta' x_{k|j}}}{\sum_{r|j} e^{\beta' x_{r|j}}} = \frac{e^{\beta' x_{k|j}}}{e^{J_j}}, \tag{12}$$

where $x_{k|j}$ includes variables related to the individual as well as to the choices (as well as variables that interact with choice dummy variables), β the first parameter vector of the model, and J_j is the inclusive value for branch j $(J_j = \log\sum_{r|j} e^{\beta' x_{r|j}})$. It summarizes the attributes of the alternatives below a branch (expected aggregate utility of the subset of choices).

At the next level up the tree, we define the conditional probability of choosing a particular branch (j = legitimate, fraud),

$$P(j|x_i) = \frac{e^{\alpha' z_j + \tau J_j}}{\sum_m e^{\alpha' z_m + \tau J_m}}, \tag{13}$$

where z_j may include variables that interact with branch-specific dummy variables, α the second group of parameters to be estimated and τ is the inclusive value coefficient.

The probability of the final choice made by an individual is $P(k, j|x_i) = P(k|j, x_i)P(j|x_i)$.

Note that when using this model, we will distinguish which variables influence the choice at the first level and which do influence the choice between types of fraud. When the coefficient of the inclusive value is set equal to 1.0, the nested nature of the choice tree disappears.

The nested logit model, like the multinomial model, can be derived from the theory of stochastic utility maximization (McFadden (1978)), but the choosing mechanism differs in both models.

The nested choice model can be estimated by a sequential method [3] (two step estimation). In this case, at the first step, the parameters of the conditional log-likelihood are estimated (lower level). The simple discrete-choice model provides estimates of β. The vector of parameter estimates and the sample observations are then used to compute each individual's inclusive values, J_j. At the second step, the upper level is considered, where the inclusive value is one of the exogenous variables and there is only a single coefficient of the inclusive value regardless of the number of branches. This procedure results in consistent (though not efficient) parameter estimates.

[3] An interesting application can be found in (Koujianou, 1995). In this case, the author applied a sequential method to estimate a nested model with five levels.

Table 2
Evolution of the insurance industry in Spain

Year	GNP (1000 billion pesetas)	Earned premiums (1000 billion pesetas)	Earned premiums/GNP (%)
1990	39.02	1.23	3.1
1991	39.89	1.38	3.5
1992	40.17	1.45	3.6
1993	39.73	1.47	3.7
1994	40.51	1.76	4.3

Source: INE, Estadistíca de Seguros Privados, 1985–1994 (Servicio Actuarial de Unespa, 1995) and the authors.

Table 3
The automobile insurance industry in Spain

Year	Total earned premiums in the automobile sector (1000 billion pesetas)	Ratio of automobile premiums/total premiums (%)	Mean number of claims (Claims/Policies) (%)	Average claim cost in constant pesetas
1990	0.59	34.7	34.7	65 106
1991	0.70	33.3	34.8	68 109
1992	0.77	31.6	35.1	60 651
1993	0.85	31.8	35.6	60 598
1994	0.89	26.5	36.1	58 147

Source: Estadística de Seguros Privados, 1985–1994 (Servicio Actuarial de Unespa, 1995) and the authors.

We also used the full information maximum likelihood (FIML) estimation technique as implemented in LIMDEP (Greene, 1995).

4. Application to the Spanish automobile insurance market

This section summarizes the results of the fraud choice model for our Spanish data set. Firstly, we present a descriptive analysis of the automobile insurance industry in Spain. Afterwards, the estimation results are presented.

4.1. The Spanish automobile insurance market: A perspective

Automobile insurance as a component of the insurance industry has experimented an enormous expansion in the Spanish market. Companies are increasingly interested in establishing efficient methods to control the evolution of automobile accidents and their associated costs (see Ayuso and Guillen, 1999).

In the recent years, there has been a significant increase of earned premiums, which has come together with an increase in the number of claims. Companies believe that the technical disequilibrium observed in the Spanish automobile insurance market is due to the rise in the total number of claims (UNESPA, 1995a,1995b).

A comparison of the evolution of the national economy and the insurance industry using a ratio shows that the sector is becoming more relevant. Table 2 presents the GNP and the earned premiums in the recent years. Earned premiums are considered to be a reliable indicator of the evolution of the insurance industry in Spain. The last column summarizes the evolution of the insurance industry compared to the GNP, in constant prices, in order to eliminate the effects of inflation. For the GNP a general deflating index was used. In the case of earned premiums, the deflating index of the services sector (in which insurance is included) was chosen. The conclusion drawn from Table 2 is that the insurance industry is improving its relative position in the Spanish economy.

We have focussed on the automobile insurance, which is a compulsory insurance for all drivers. According to the Spanish regulation, every driver must sign an insurance contract covering at least liability (*Responsabilidad Civil Obligatoria*).

The importance of the automobile insurance industry is described in the first two columns of Table 3. In the past decade, there has been an increase in the number of personal auto policies. Nevertheless, the importance of

automobile insurance within the insurance sector has decreased, probably due to the presence of new insurance products, namely life and some forms of health insurance. Roughly speaking, the automobile insurance industry represented one-third of all the insurance written in Spain at the beginning of this decade, but it has dropped to one-quarter.

The last two columns of Table 3 show that the mean number of claims has slightly increased every year, but the average cost of claims has started decreasing in the most recent years. The industry has a general feeling that the decrease is much less than expected and so, companies suspect that fraud may have entered the Spanish automobile insurance scene in a significant way.

4.2. Fraud in the Spanish automobile insurance market

Fraud is part of the insurance consumer behaviour. Companies believe that there has been a relatively recent effort to reduce fraud in the automobile insurance industry.

In Spain the attitude towards fraud has been characterized by a passive position of companies. The industry has been offering new attractive products. Firms have entered a battle of lowering premiums and have devoted a lot of effort towards increasing their market share. There has been a complete lack of coordination among companies with little control on adverse selection. There is no official institution controlling fraud or devoted to fraud detection, nor have the companies implemented systems to control fraud other than the inspection of claims.

It seems obvious that the behaviour of the Spanish market has allowed for the presence of different kinds of fraud. Products such as the *CIDE/ASCIDE* procedure (an agreement between companies or No-Fault) speed-up compensations and deter screening and control of claims. On the other hand, from an outer perspective, insurance companies are still perceived by consumers as organizations which make large profits.

It is estimated that the presence of fraud in the Spanish automobile industry ranges from 15% to 60% (CES (1992)). The wide range of this interval is caused by the difference between large and smaller firms and their estimation criteria. Larger firms accept that about 15–20% of the claims contain some form of fraud, while smaller companies are much more heterogeneous, so that they may reach 60%. One of the largest companies (see Cobo, 1993) states that 22% of the claims contain some suspicious circumstance. These figures are similar to those reported by Hoyt (1990) for the US (15% of fraud in the insurance industry) and by Clarke (1990) for Germany (11% in the automobile industry). A similar overview is given by Picard (1996).

The main elements of fraud in the Spanish context may be summarized in the following groups:

(i) *Presence of false data in the contract*. Age or driving experience may incorrectly be recorded as well as the quality of the car driven (model and vehicle age).
(ii) *Multiple contracts*. An agreement with several companies may lead to several compensations for the same accident.
(iii) *False claim*. Typically pre-existing damages are included in the claim. Those damages were not reported previously in order not to be moved to a worse bonus-malus category. The insured may also claim a false theft. Another common practice is to build-up an accident in order to benefit another driver (usually a relative or a friend).

Since the first and second situations are easier to detect, we have focussed on false claims. Our data set includes claims classified by the company. Information about the kind of fraud that was found is given. We assume that claims classified as legitimate correspond to honest insureds. Otherwise, we should take into account that the dependent variable is measured with error. Fraud claims correspond to those cases when the insured finally admitted that he had incurred fraud.

4.3. The data

The database has been obtained from a sample of claims of a Spanish company. Data were collected from 1993 to 1996. Half of the claims are legitimate, the other half are claims that had been identified as fraudulent. Each fraud

Table 4
Variables in the data set

CLAIM	Type of fraud (legitimate, fraud for self-benefit, fraud for third party benefit)
NFILES	Number of files associated to the claim. Each file indicates a kind of loss claimed (property, bodily-injury,. . .)
NFAULT	Insured did not accept fault equals 1, otherwise 0.
POLICE	Police officer reported about the accident equals 1, otherwise 0.
WITNESS	Presence of witnesses equals 1, otherwise 0.
NONURBAN	Accident took place in a nonurban area equals 1, otherwise 0.
RECORDS	Number of previous claims
VEHAGE	Number of years since vehicle fabrication

Table 5
Descriptive statistics

Variable	Total sample [a]		Fraud		Non-fraud	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
NFILES	1.53	0.57	1.00	0.06	1.86	0.50
NFAULT	0.70	0.46	0.48	0.50	0.84	0.37
POLICE	0.14	0.35	0.04	0.20	0.20	0.40
WITNESS	0.01	0.08	0.006	0.08	0.008	0.09
NONURBAN	0.06	0.24	0.08	0.27	0.05	0.22
RECORDS	1.31	1.65	1.61	1.79	1.13	1.53
VEHAGE	6.58	4.58	6.55	4.78	6.60	4.46

Unweighted sample statistics.

has been classified as being for personal benefit or for a third party benefit. All other types of fraud, such as repair store implication or driving under the influence of alcohol, have been excluded from the analysis. The sample is not strictly random. There has been an oversampling of fraud claims in order to obtain a good representation for this group. The estimation procedure includes a weighting of the observations in order to correct for this phenomenon. We have assumed that in the population, the percentage of legitimate claims is approximately 78%, and we have also assumed that different types of fraud are equally frequent. So, using these population proportion estimates and the corresponding sample rates, a correction for choice-based sampling has been implemented.

The final data set has 1357 claims: 846 are legitimate claims, 247 are fraudulent claims aimed to obtain a benefit for the insured and 264 are fraudulent claims that benefit third parties. Data contain information on the accident characteristics (place, report of the police officer,. . .), the insured driver characteristics (history of previous claims,. . .) and vehicle characteristics (date of fabrication). The information contained in the sample has been obtained from the claim statement or the policy. Table 4 gives a list of variables used in the econometric model.

Table 5 shows some descriptive measures for the overall sample and for the two subsamples of fraud and legitimate claims.

5. Estimation results

5.1. Estimation results: Multinomial discrete-choice model

All the explanatory variables included in the model refer to the insured or to the accident. They do not change according to the alternative chosen. The dependent variable is the type of claim (legitimate, fraud for own benefit, fraud to benefit a third person).

Table 6
Results for fraud MNL model

Variable	Coefficient	*t*-test	*P*-value
Fraud for own benefit			
Constant	15.57	59.61	0.00
NFILES	−17.11	−98.24	0.00
NFAULT	1.65	6.62	0.00
POLICE	−1.66	−3.38	0.00
WITNESS	−11.56	−14.10	0.00
NONURBAN	1.15	2.51	0.01
RECORDS	0.22	3.01	0.00
Fraud for third party			
Constant	5.63	4.83	0.00
NFILES	−5.97	−5.05	0.00
NFAULT	−0.85	−3.24	0.00
POLICE	−1.42	−2.95	0.00
WITNESS	3.94	4.53	0.00
NONURBAN	0.10	0.21	0.83
RECORDS	0.23	3.11	0.00

Number of observations: 1357; chi-squared: 859.34; log-likelihood function: −486.03; degree of freedom: 12; restricted log-likelihood: −915.70; significance level: 0.00.

The estimation was obtained using maximum likelihood with the correction for choice-based sampling in order to take into account the effect of the over-representation of fraud claims. Therefore, weights were included in the estimation procedure. The parameter estimates are shown in Table 6.

The first set of seven parameter estimates corresponds to the parameter vector in the probability of committing fraud for self-benefit. The second group of estimates refers to the probability of choosing fraud for a third party. Both specifications use the same variables. Table 6 also shows the asymptotic *t*-values.

The likelihood ratio test unambiguously rejects the hypothesis that the coefficients in the two estimating equations are equal to zero; the log-likelihood falls from −486 to −915, while 12 degrees of freedom are gained, when the equality constraints are imposed.

All coefficients are significant at the 5% level except one. The location of the accident (road or urban area) seems to have no influence in the choice of fraud for a third party if compared to legitimate claims.

The results show several interesting facts about the estimated probability of fraud. All estimates have the expected signs. The number of files associated to a claim is related to a higher probability of a legitimate claim. A higher number of files indicates a type of accident with a lot of damages and therefore causes deeper investigations for the company to detect fraud.

Another characteristic associated with legitimate claims is the report of the police, which clearly leaves little room for fraud.

Some parameter estimates suggest that the influence of the associated variables is opposite in the two equations. For example, the presence of witnesses is linked to a smaller probability of fraud for self-benefit, while it is positively related to the probability of choosing the third alternative.

When the fault is accepted by a third party, the probability of fraud in benefit of the insured increases. Accordingly, when the insured denies his fault the probability of committing fraud for a third party benefit decreases, *ceteris paribus*.

Finally, as expected, the number of previous claims by the insured is associated to a higher probability of defrauding. In this case, the impact is similar in the two possible types of fraud. The interpretation of the significance of this coefficient may be confusing unless we take into account that we do not include fraud against the insured.

Table 7
Classification results MNL model

Actual	Predicted			
	Legitimate	Own	Third	TOTAL
Legitimate	766	56	24	846
Own	50	179	18	247
Third	123	51	90	264
TOTAL	939	286	132	1357

One possible criticism about this approach is the possible violation of the independence of irrelevant alternatives (IIAs) property. To test this model against a restricted situation, in which one type of fraud is eliminated from the choice set, the Hausman–McFadden (1984) statistic was calculated. [4]

The results of the Hausman test are: 0.2658 and 0.0008, respectively, when each type of fraud is eliminated. Therefore, the test cannot reject the IIA property for the type of fraud and the odds ratios such as P_{i0}/P_{i1} remain constant when the third alternative is not taken into account.

The adjusted probability of each choice can be calculated as shown in Appendix A.

Table 7 presents the classification results in the sample used for the model estimation.

As shown in Table 7, 1035 claims were classified in the correct group. This means that the model achieved 76.3% of total correct classification. 90.5% of legitimate claims are predicted by the model in their own group. 72.5% of fraud for self-benefit claims were also detected, while the worst group corresponds to the third alternative with only 34% of correct classification. We can see that the rate of correct classification is very poor in the fraud for third party benefit. This may be due to the difficulty to predict this kind of behaviour. The threshold for classification could also be changed in order to optimize classification rates in practice. It should also be emphasized that a cross-classification method (or a bootstrap approach) would be more appropriate to validate the model results, which, as presented here, might be too optimistic.

The marginal effects of the regressors on the probabilities were also calculated but they are not shown here for brevity. [5] Besides, since several dichotomous variables are used, interpretation becomes cumbersome.

The multinomial model provides an interesting tool to simulate several scenarios. Fig. 2 shows the behaviour of the predicted probability of fraud in the most frequent type of claimant, when the number of previous claims increases. It is shown that the estimated probability of fraud for a third party increases when the claimant has more previous claims. The solid line is the estimated probability of fraud, it is the sum of the estimated probabililty of fraud for self-benefit and the estimated probability of fraud for a third party. We see that the overall estimated probability of fraud has a larger increase for a history of a few previous claims. The type of claimant that has been chosen for showing the evolution of the estimated probabilities when the number of previous claims increases, is the most frequent type of individual in the sample. The vehicle age is six years, the accident took place in an urban area, there is no police report, there are no witnesses, the claimant denied his fault and the number of files equals one. As we can see, the estimated probability of fraud exceeds 50%.

For every individual, the model gives point estimates of $(1-q)$ in (4), which equals the probability of choosing the second alternative plus the probability of choosing the third.

Since p_1 and p_2 can be inferred from the general classification results shown in Table 7, the probability of classifying a legitimate claim as fraud is $p_1 = (56+24)/846 = 9.5\%$, and the probability of classifying a fraud claim in any of the fraud types is $(179+18+51+90)/(247+264) = 66.1\%$. Then, $p_2 = 33.9\%$.

[4] The statistic is equal to $T = (\hat{\beta}_r - \hat{\beta}_u)'(\hat{V}_r - \hat{V}_u)^{-1}(\hat{\beta}_r - \hat{\beta}_u)$, where $\hat{\beta}_u$ and $\hat{\beta}_r$ refer to the parameter estimates obtained by applying the maximum likelihood method on the unrestricted and restricted choice sets, respectively, and $\hat{V}_u$ and $\hat{V}_r$ are the estimate covariance matrices corresponding to the two choice sets. The test statistic under the null hypothesis follows a Chi-squared distribution with seven degrees of freedom, in this case.

[5] When evaluating at the mean vector of the explanatory variables the following expression is used: $\partial P_k/\partial x_i = P_k[\beta_k - \sum_{r=1}^{2} P_r \beta_r]$. In this case, the estimated values are very low.

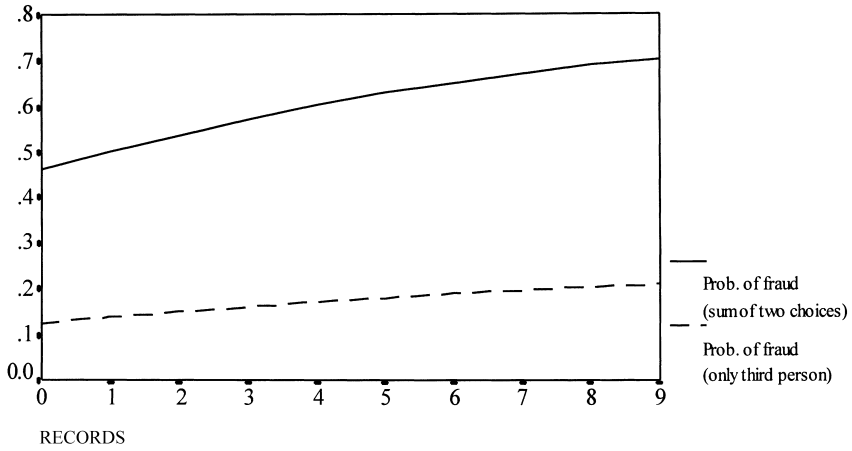

Fig. 2. Estimated probability of fraud for the most frequent claimant.

Finally, given M (the accident cost), c_1 and c_2 as in (4), our model gives an estimate of the expected cost of the claim. For example, if we assume that $c_1 = c_2 = 0.1\,M$, it follows that the estimated expected cost of a claim equals $M[-0.6044(1-q)+1.0095]$, where $(1-q)$ can be estimated from the model. If the aggregate estimated probability of fraud is 0.41, then the expected cost of the claim is 23.83% smaller than the accident cost.

5.2. Estimation results: Nested multinomial logit model

In order to fit a nested logit model we have used the full information maximum likelihood method and the two step estimation approach. The results are very similar and we only report the first procedure.

Following Section 3.2, we assume that the model to be estimated is

$$P(\text{OWN}_i|\text{FRAUD}_i) = \frac{\exp(\beta_1 + \beta_2 \text{NFAULT}_i + \beta_3 \text{POLICE}_i)}{e^{J_i}} \tag{14}$$

$$P(\text{THIRD}_i/\text{FRAUD}_i) = 1 - P(\text{OWN}_i|\text{FRAUD}_i) \tag{15}$$

$$P(\text{FRAUD}_i) = \frac{e^{\alpha_1 \text{RECORDS}_i + \tau J_i}}{1 + e^{\alpha_1 \text{RECORDS}_i + \tau J_i}}, \tag{16}$$

$$P(\text{NOFRAUD}_i) = 1 - P(\text{FRAUD}_i). \tag{17}$$

Table 8 shows the estimation results. We note that there has been a correction for choice-based sampling due to the oversampling of fraud claims. The log-likelihood function for the branch model is higher that the log-likelihood obtained when only constants were included in the model.

All coefficients are significant at the 5% level and they have the expected signs. As we can see, the existence of a police report is associated to the choice of a particular type of fraud (self-benefit), once the insured decided to commit fraud. On the other hand, when the claimant does not accept his fault when defrauding, this is positively associated to the choice of fraud for his own benefit.

The value of the coefficient associated to the number of previous claims is also positive and significant, thus indicating that the insureds with many previous claims tend to commit fraud. One possible explanation for this fact is that they probably are in the lowest bonus-malus class. Since the company punishment is generally a change of the bonus-malus class, they think that their situation cannot become worse.

From the interpretation of the model results, we conclude that the previous history of the claimant influences his probability to commit fraud. This means that personal circumstances may have a great impact on the choice of the

Table 8
Results for fraud nested logit model

Variable	Coefficient	*t*-test	*P*-value
Lower level (FRAUD)			
Constant	−1.13	6.67	0.00
NFAULT	2.23	9.54	0.00
POLICE	1.11	5.48	0.00
Upper level			
RECORDS	0.16	4.16	0.00
Inclusive values AIV			
NOFRAUD	1.00	−−	–
FRAUD	−1.34	−8.63	0.00

Number of observations: 4071; chi-squared: 1003.84; log-likelihood function: −792.88; degree of freedom: 5; restricted log-likelihood: −1294.80; significance level: 0.00. (each individual is replicated once for every choice)

Table 9
Classification results nested logit model

Actual	Predicted			
	Legitimate	Own	Third	TOTAL
Legitimate	688	91	67	846
Own	188	30	28	247
Third	159	31	75	264
TOTAL	1035	152	171	1357

claimant in the first level of the decision tree. If we had financial variables available in the data set, we could test whether insureds incurred fraud as a way to make profits. The vehicle age seemed to have no impact at this step. At the lower level of the model, we have concluded that the significant variables affecting the type of fraud chosen have to do with the particular circumstances of the accident. Variables related to the police report or the fault are those found to be influencing the choice of the type of fraud.

Appendix shows how to calculate the estimated fraud probabilities for this model.

The classification results for the nested multinomial logit model are shown in Table 9. The percentage of correct classification is 58.4%. This indicates that the model prediction performance within the sample is only acceptable. We think that these results suggest that the nested logit model does not provide a good prediction technique, but we should emphasize that its usefulness lies on the fact that it provides an explanation of the decision process. Using the nested multinomial logit model, we have found which variables influence each choice and we have found that different variables influence each step of the decision process. We are also aware that there might be omitted variables in the model, so we think that these results only show the applicability of this methodology in this context. Again, validation of this model predictive performance should consider either more variables and a cross-validation approach.

6. Concluding remarks

Our orientation in this paper has been methodological rather than theoretical. We have shown how discrete-choice models may be useful to study fraudulent behaviour. The application of this technique to a Spanish sample reveals a number of interesting findings. Some of the claimant characteristics as well as accident circumstances that influence the fraudulent behaviour are identified. Utility functions are introduced and it is shown that the causes of fraud behaviour may be different when several types of fraud are considered. We have seen that the prediction performance of the multinomial logit model is much better.

When assuming that the decision is made step by step, the suitable modelization is a nested multinomial logit model, but the estimation procedure becomes more complicated. The results show the impact of claim characteristics in either the twigs or the branches of the decision tree.

The main conclusion of this paper is that flags indicating the presence of fraud are quite different depending on the kind of fraudulent behaviour. We have linked the utility model to its corresponding econometric model. The question of selecting the multinomial logit model or the nested multinomial logit model remains open, because it is a consequence of the assumption about the mechanism underlying the decision process.

Our study is restricted to analyse legitimate claims versus the two types of fraud most prevalent in the Spanish market. We have seen that the presence of a police report discourages fraudulent practices, whereas the historical behaviour of the insured has a great impact on his attitude in the claiming process. Companies operating in Spain are concerned about the presence of moral hazard. Here, the analysis of individual choice is based on the data available in the claim statement. Our approach aims to provide a predictive tool, but it is also an explanatory model.

Our results are comparable to those proposed by other authors, but we have accounted for a unified framework with several kinds of fraud. We suggest that including a measurement error in the dependent variable should be the next step, specially because some legitimated claims might be fraud claims that could not be detected. The generalization of these models to include measurement error in the dependent variable remains for future work.

Acknowledgements

The authors thank the comments received from R.A. Derrig and suggestions from the participants of the International Conference on Insurance: Mathematics and Economics held in Amsterdam from 25th to 27th August 1997. We also thank the anonymous referee for the general and specific comments.

Appendix A

A.1. Estimated probabilities for the MNL model

The calculation of the estimated probabilities under the multinomial logit model can be computed using the following expressions:

$$P(Y_{i0}=1)=\frac{1}{1+B_{i1}+B_{i2}},\qquad P(Y_{i1}=1)=\frac{B_{i1}}{1+B_{i1}+B_{i2}},\qquad P(Y_{i2}=1)=\frac{B_{i2}}{1+B_{i1}+B_{i2}},$$

where

$$\begin{aligned}B_{i1}=&\exp(15.57-17.11\,\mathrm{NFILES}_i+1.65\,\mathrm{NFAULT}_i-1.66\,\mathrm{POLICE}_i-11.56\,\mathrm{WITNESS}_i\\&+1.15\,\mathrm{NONURBAN}_i+0.22\,\mathrm{RECORDS}_i),\\B_{i2}=&\exp(5.63-5.97\,\mathrm{NFILES}_i-0.85\,\mathrm{NFAULT}_i-1.42\,\mathrm{POLICE}_i+3.94\,\mathrm{WITNESS}_i\\&+0.10\,\mathrm{NONURBAN}_i+0.23\,\mathrm{RECORDS}_i).\end{aligned}$$

A.2. Estimated probabilities for the nested logit model

At the lower level of the tree the estimated conditional probabilities, $P(k|j,i)$ can be calculated as

$$P(1|1,1)=P(\mathrm{OWN}_i|\mathrm{FRAUD}_i)=\frac{\exp(-1.1297+2.2323\,\mathrm{NFAULT}_i+1.1067\,\mathrm{POLICE}_i)}{1+\exp(-1.1297+2.2323\,\mathrm{NFAULT}_i+1.1067\,\mathrm{POLICE}_i)},$$

$$P(2|1,1)=P(\mathrm{THIRD}_i|\mathrm{FRAUD}_i)=1-P(1|1,1),$$

$$P(1|2,1)=P(\mathrm{NOFRAUD}_i|\mathrm{NOFRAUD}_i)=1.$$

At the upper level, the adjusted probabilities $P(j|i)$ follow from the expression below:

$$P(1|1) = P(\text{FRAUD}_i) = \frac{\exp(0.15980\,\text{RECORDS}_i - 1.3424\,\text{BRANCHIV}_i)}{e^{\text{LIMBIV}_i}},$$

$$P(2|1) = P(\text{NOFRAUD}_i) = \frac{1}{e^{\text{LIMBIV}_i}}$$

where

$$\text{BRANCHIV}_i = \log[\exp(-1.1297 + 2.2323\,\text{NFAULT}_i + 1.1067\,\text{POLICE}_i) + 1]$$

$$\text{LIMBIV}_i = \log[\exp(0.15980\,\text{RECORDS}_i - 1.3424\,\text{BRANCHIV}_i) + 1],$$

and finally, $P(1)$ is 1.

The estimated unconditional probability (given the values of the explanatory variables) of the choice made by an individual is: $P(k, j, i) = P(k|j, i)P(j|i)P(i)$.

References

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.

Ayuso, M., Guillen, M., 1999. Modelos de Detección de Fraude en el Seguro de Automóvil. Cuadernos Actuariales, in press.

Belhadji, E.B., Dionne, G., 1997. Development of an Expert System for the Automatic Detection of Automobile Insurance Fraud. Working Paper 9706. École des Hautes Études Commerciales, Université de Montréal.

Brockett, P.L., Xiaohua, X., Derrig, R.A., 1995. Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud. Special Actuarial Seminar, Automobile Insurers Bureau, Boston, 25 January.

CES, 1992. El Fraude en el Seguro de Automóviles. Centro de Estudios del Seguro, Madrid.

Clarke, M., 1990. The control of insurance fraud. A comparative view. The British Journal of Criminology 30 (1), 1–23.

Cobo, P., 1993. Manual de Investigaciòn de Siniestros y Lucha contra el Fraude en el Seguro de Automóviles. Mapfre, Madrid.

Cummins, J.D., Tennyson, S., 1992. Controlling automobile insurance costs. Journal of Economic Perspectives 6 (2), 95–115.

Cummins, J.D., Tennyson, S., 1996. Moral hazard in insurance claiming: evidence from automobile insurance. Journal of Risk and Uncertainty 12 (1), 29–50.

Derrig, R.A., Ostaszewski, K.M., 1995. Fuzzy techniques of pattern recognition in risk and claim classification. The Journal of Risk and Insurance 62 (3), 447–482.

Dionne, G., Artis, M., Guillen, M., 1996. Count data models for a credit scoring system. Journal of Empirical Finance 3 (3), 303–325.

Dionne, G., Gibbens, A., St-Michel, P., 1993. An economic analysis of insurance fraud. In: Fortin, J.L., Girand, J.D. (Eds.), Insurance Fraud, University of Montreal Press.

Greene, W.H., 1995. LIMDEP, Version 7.0, User's Manual. Econometric Software, Inc., New York.

Greene, W.H., 1997. Econometric Analysis. 3rd ed., Prentice-Hall, New York.

Hausman, J., McFadden, D., 1984. Specification tests for the multinomial logit model. Econometrica 52 (5), 1219–1240.

Hoyt, R.E., 1990. The effect of insurance fraud on the economic system. Journal of Insurance Regulation 8 (3), 304–315.

Koujianou, P., 1995. Product differentiation and oligopoly in international markets: the case of the US automobile industry. Econometrica 63 (4), 891–951.

Maddala, G.S., 1983. Limited-dependent and Qualitative Variables in Econometrics. Econometric Society Monographs. Cambridge University Press, Cambridge.

Marter, S., Weisberg, H.I., 1991. Medical costs and automobile insurance: a report on bodily injury liability claims in Massachusetts. Journal of Insurance Regulation 9 (3), 381–422.

McFadden, D., 1978. Modelling the choice of residential location. In: Karlquist, A., et al. (Eds), Spatial Interaction Theory and Planning Models, North-Holland, Amsterdam, pp. 75–96.

McFadden, D.L., 1983. Qualitative response models. In: Griliches, Z., Intrilligator, M. (Eds), Handbook of Econometrics, vol. II, Chapter 24. Noth-Holland, Amsterdam, pp. 1395–1457.

Picard, P., 1996. Auditing claims in the insurance market with fraud: the credibility issue. Journal of Public Economics 63 (1), 27–56.

UNESPA, 1995a. Estadística de Seguros. Cifras de Avance e Informe Económico 1994. Aseguradora, UNESPA, Madrid.

UNESPA, 1995b. Estadística de Seguros Privados, 1985–1994. Aseguradora, UNESPA, Madrid.

Weisberg, H.I., Derrig, R.A., 1991. Fraud and automobile insurance: A report on the baseline study of bodily injury claims in Massachusetts. Journal of Insurance Regulation 9 (4), 497–541.

Weisberg, H.I., Derrig, R.A., 1993. Quantitative methods for detecting fraudulent automobile bodily injury claims. AIB Cost Containment/Fraud Filing, pp. 49–82.