©*The Journal of Risk and Insurance*, 2002, Vol. 69, No. 3, 325-340

Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims

Manuel Artís
Mercedes Ayuso
Montserrat Guillén

Abstract

The insurance industry is concerned with the detection of fraudulent behavior. The number of automobile claims involving some kind of suspicious circumstance is high and has become a subject of major interest for companies. This article demonstrates the performance of binary choice models for fraud detection and implements models for misclassification in the response variable. A database from the Spanish insurance market that contains honest and fraudulent claims is used. The estimation of the probability of omission provides an estimate of the percentage of fraudulent claims that are not detected by the logistic regression model.

Introduction

An insurance contract involves an agreement between the insurer and the insured for the company to cover a risk. The behavior of the insured, who will receive compensation in the event of an accident, is not always honest. Automobile insurance fraud has become a subject of major concern for companies and consumers.

Some studies on the theory of insurance fraud concern the use of auditing patterns to control fraud (Bond and Crocker, 1997; Picard, 1996). This approach, known as costly state verification,[1] assumes that the insurer can obtain some information about

Manuel Artís, Mercedes Ayuso, and Montserrat Guillén (corresponding author) work with the Grup de recerca: Risc en Finances i Assegurances, Departament d'Econometria, Estadística i Economia Espanyola, Universitat de Barcelona. The authors are grateful for the comments received from two anonymous referees; R.A. Derrig, R. Gagné, and S. Tennyson; and valuable suggestions from the participants in the 26th Seminar of the European Group of Risk and Insurance Economists held in Madrid, September 20–22, 1999, and in the 4th International Congress on Insurance: Mathematics and Economics held in Barcelona, July 24–26, 2000. This work has received support from the Spanish grant CICYT SEC1999-0693. We thank the Linguistic Advice Centre for correcting the manuscript.

[1] Another important discussion is one about costly state falsification (Crocker and Morgan, 1998; Crocker and Tennyson, 1999). These models are specified on the hypothesis that the insurer cannot audit the claims—that is, the insurer cannot tell whether the claim is true or false. The optimal contract design considers that the claimant can exaggerate the loss amount but at a cost greater than zero.

the claim with an auditing cost. It implicitly assumes that the audit process (normally using auditing technology) always discerns whether a claim is fraudulent or honest. The main discussion concerns the design of a contract that minimizes the insurer costs, including the total claim payment plus the cost of the audit. Normally, models suggested in this framework use the amount of the claim to make a decision on applying monitoring techniques (Bond and Crocker, 1997; Crocker and Tennyson, 1999). In other works, deterministic audits are compared with random audits (see Picard, 2000), or it is assumed that the insurer will commit to an audit strategy (Boyer, 1999).

However, the auditing technology applied to the detection of fraud may not be perfect, and claims may finally be misclassified. The two questions that we consider in this article are: (1) Are all claims classified as honest claims really honest claims? and (2) Can all of these claims really be honest claims? Our purpose is to improve the methodology to account for the presence of errors in the existing information on claims.

Some authors find a modeling approach that sheds some light on the empirical investigation of fraud. Weisberg and Derrig (1993) specify a multiple linear regression model to select indicators of different types of suspicion of fraud. Derrig and Weisberg's (1998) results are presented considering unsettled claims.[2] Derrig and Ostaszewski (1995) apply fuzzy set techniques to classify claims. Brockett et al. (1998) propose a self-organizing neural network to transform claim characteristics into claim types. Belhadji and Dionne (1997) and Artís et al. (1999) suggest discrete choice models to estimate the probability of the presence of fraud in a claim, based on previous knowledge of the insured's behavior in a portfolio.

The aim of the existing modeling techniques in this context is to provide new tools to recognize the presence of suspicious circumstances and to help claim supervisors identify fraud. Nevertheless, one of the main criticisms of the methods used to date is that information on previous claims is imperfect. In some data sets, previous information on claims is limited, because only indicators of suspicious characteristics of claims are available. In addition, sometimes more than one opinion has been used to classify claims in fraud groups. In the data set used by Artís et al. (1999), claims classified as "fraudulent claims" only concerned claims for which, although no legal prosecution had taken place, the insured finally admitted fraud. And while "honest claims" involved the claims with no fraud, they could also involve some claims with undetected fraud—namely, fraudulent claims that the existing methodology was unable to identify. The same data set is used here. Our definition of fraud is limited to observed fraud only. Observation of fraud means that the insurer has no doubt about the existence of fraud. In this context, a claim can be either fraudulent or honest.

Caron and Dionne (1999) estimated the total fraud level in the automobile insurance industry in Quebec assuming that fraud cannot always be observed (and

[2] Information on the investigation process is taken into account at different stages of the claim process.

assuming that all suspected fraudulent claims can be fraudulent). Furthermore, they concluded that the conditional probability of claim adjustment staff detecting fraud, assuming that a claim is fraudulent, is 1/3. Therefore, beside further considerations on the commitment of insurers, it seems that the existing audit strategy is not perfect.

The purpose of this article is to develop an audit technology that takes into account that previous information on fraud is available and that uncertain knowledge on the nature of all the other claims (either honest or undetected fraud) can also be used. The key feature is to allow for misclassification error in the existing method that separates fraudulent and honest claims.

We suspect that previous claims may have been misclassified, but only in one direction. Only honest claims may contain a portion of fraudulent claims that cannot be identified. Following the statistical literature (see Bollinger and David, 1997; Poterba and Summers, 1995), this is usually referred to as omission error. We claim that the classification of statistical information is based upon previous technologies that may have failed to identify fraud.

As a result, we obtain an estimate of the percentage of fraudulent claims that were not detected by the adjustment staff, but should have been detected with the available information. We claim that further improvements for investigating fraud should be sought, but that the existing information is not used efficiently.

The rest of the article is organized as follows. In the "Data" section, we describe the data and the variables included in the fixed part of the model. In the "Model and the Estimation Method" section, we present standard binary choice models and the extended models when omission error is considered. Maximum likelihood estimation results are discussed in "Results." We conclude that the estimated proportion of honest claims that were not identified as fraudulent is significant, although moderate. The "Final Remarks" section emphasizes the conclusions on the use of misclassification assumptions in logistic regression models.

The Data

The data correspond to a sample of claims for car damages from accidents that occurred from 1993 to 1996 in Spain (see Artís et al., 1999). The insurer classified claims into two categories: honest or fraudulent. The variable of interest was, therefore, dichotomous and was coded using zeros for honest claims and ones for fraudulent claims.

The insurer has no doubt that fraud exists if the insured admits fraud. In Spain, legal prosecution of insurance fraud is extremely rare. When the insurer has some suspicion on a claim, a negotiation with the policy holder usually ensues. The insurer denies complete payment or announces that the contract will be canceled. As a consequence, fraudulent policy holders usually admit fraud, and no further action is undertaken.

Table 1 defines explanatory variables that are used later in the model. Half of the claims were legitimate, and the other half are claims that were identified as

TABLE 1
Variables Used in the Model

Y	Observed type of claim (fraudulent equals 1 and legitimate equals 0)
Characteristics of the Insured/Claimant/Policy:	
AGE	Age of insured driver when the accident occurred
LICENSE	Number of years since the insured obtained first driver's license
RECORDS	Number of previous claims of the insured
COVERAGE	Third-party liability equals 1; extended coverage equals 0
DEDUCTIBLE	Existence of a deductible equals 1; otherwise equals 0
ACCESSORI	Coverage for accessories equals 1; otherwise equals 0
Characteristics of the Vehicle:	
VEHUSE	Vehicle for private use equals 1; other uses equal 0
VEHAGE	Age of the vehicle
Characteristics of the Accident:	
FAULT	Insured accepts the blame for the accident equals 1; otherwise equals 0
NONURBAN	Accident occurred in a nonurban area equals 1; otherwise equals 0
NIGHT	Accident occurred at night equals 1; otherwise equals 0
WEEKEND	Accident occurred during a weekend equals 1; otherwise equals 0
WITNESS	Existence of witnesses equals 1; otherwise equals 0
POLICE	Existence of police report equals 1; otherwise equals 0
ZONE1	Zone with high level of accidents equals 1; otherwise equals 0
ZONE3	Zone with low level of accidents equals 1; otherwise equals 0
REPORT	Existence of a suspicious textual report equals 1; otherwise equals 0. This variable indicates that the claimant reported unusual circumstances for the accident.
NAMES	Same family name for insured and the other vehicle driver equals 1; otherwise equals 0.
PROXIM	Accident occurred between the policy issue date and the policy effective starting date equals 1; otherwise equals 0.
DELAY	Claim not reported to the company within the established period equals 1; otherwise equals 0

fraudulent. The sample is not strictly random, because there was an over-sampling of fraudulent claims in order to obtain a good representation for this group.[3]

The data set has 1,995 claims: 998 were legitimate and 997 were fraudulent. The data contain information on the accident (place, report of the police officer, and so on), the insured driver (history of previous claims), and the vehicle (date of manufacture). The information contained in the samples was obtained from the claim statement or the policy.

Table 2 shows some descriptive measures for the overall sample and for the two subsamples of fraudulent and legitimate claims. The variables used in our study correspond to information obtained from the claim reports, including prior claim information. We agree that, especially in accidents involving injured individuals, fraud may be incurred during medical treatment, and we argue that other indicators should be used in that case to identify fraudulent claims. The article by Derrig and Weisberg (1998) discusses the kind of indicators that are taken into account when looking for fraud in medical treatment.

THE MODEL AND THE ESTIMATION METHOD

One uses qualitative response models to analyze a categorical dependent variable that, in the dichotomous case, may indicate the presence or absence (occurrence or nonoccurrence) of a particular event. Nevertheless, as we justified in the Introduction, this dichotomous variable may contain some source of error. Due to the qualitative nature of the dependent variable, this is not called a measurement error but a misclassification error. *Omission errors* refer to the fact that the event occurred but was not recorded, so that the observed category is "non-occurrence." In other words, when the observation of the dependent variable indicates "absence," but the true category is "presence," omission error occurs. A commission error occurs when the observed category indicates "presence," but the true category is "absence." In the situation herein, we will only have omission errors because we can only have undetected fraudulent claims; however, we know for sure that all fraudulent claims are correctly classified, so that commission error does not exist.

Misclassification error in the response variable has been studied in the context of discrete choice models. Bollinger and David (1997) use this approach to evaluate the impact of ignoring misclassification in an application devoted to consumption data. Hausman et al. (1998) also apply this method to analyze the causes of employment change. We shall apply the same method.

Logistic models are usually defined in a latent variable context. Let Y_i^* be unobservable. This latent variable indicates the utility of the ith individual to embrace one of the two

[3] The estimation procedure should include weights in order to correct for oversampling if the interest of the model is predictive. In that case, we would assume that the population percentage of legitimate claims is approximately 78 percent and that, correspondingly, the population proportion of fraud is 22 percent. A correction for choice-based sampling would then be implemented, and only the intercept parameter estimates would change. Even if no correction were introduced in the estimation procedure, the rest of the parameter estimates of the linear predictor would be consistently estimated (Cosslett, 1993) using the maximum likelihood method.

TABLE 2
Summary Statistics

	Total Sample		Observed Fraudulent Claims		Observed Honest Claims	
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
AGE	38.02	12.32	36.81	11.51	39.23	12.98
LICENSE	14.23	9.09	13.42	8.64	15.04	9.45
RECORDS	1.42	1.80	1.62	1.92	1.22	1.65
COVERAGE	0.91	0.29	0.94	0.24	0.87	0.34
DEDUCTIBLE	0.03	0.16	0.01	0.11	0.04	0.20
ACCESORI	0.07	0.25	0.06	0.24	0.08	0.27
VEHUSE	0.88	0.32	0.84	0.36	0.92	0.27
VEHAGE	6.17	4.48	6.26	4.55	6.09	4.42
FAULT	0.32	0.47	0.47	0.50	0.17	0.38
NONURBAN	0.07	0.26	0.09	0.28	0.05	0.23
NIGHT	0.13	0.34	0.22	0.41	0.05	0.22
WEEKEND	0.27	0.44	0.31	0.46	0.24	0.42
WITNESS	0.01	0.08	0.01	0.08	0.01	0.08
POLICE	0.11	0.31	0.03	0.18	0.19	0.39
ZONE1	0.14	0.34	0.10	0.30	0.17	0.38
ZONE3	0.49	0.50	0.59	0.49	0.39	0.49
REPORT	0.59	0.49	0.64	0.48	0.55	0.50
NAMES	0.06	0.24	0.10	0.30	0.03	0.17
PROXIM	0.02	0.13	0.03	0.16	0.01	0.08
DELAY	0.24	0.43	0.37	0.48	0.12	0.32

possible outcomes. For $i = 1, \ldots, n$, where n is the number of individuals, assume a regression model for Y_i^* such that:

$$Y_i^* = \beta' X_i + e_i, \tag{1}$$

where X_i is a column vector of the observed explanatory variables, β is the vector of unknown parameter, and e_i is a disturbance term in the model.

Since Y_i^* cannot be observed, Equation (1) leads to a qualitative choice model, where only the outcome is observed. Let $\tilde{Y}_i$ be a dichotomous variable indicating presence of fraud such that:

$$\begin{aligned} \tilde{Y}_i &= 1, \quad \text{if } Y_{\mathrm{i}}^* > 0 \\ \tilde{Y}_i &= 0, \quad \text{otherwise.} \end{aligned} \tag{2}$$

If there is no measurement error in the response, $\tilde{Y}_i$ indicates the true outcome with the following probability:

$$\mathrm{Prob}(\tilde{Y}_i = 1|X_i) = \mathrm{Prob}(Y_i^* > 0|X_i) = \mathrm{Prob}(e_i > -\beta' X_i) = F(\beta' X_i), \tag{3}$$

where $F(\cdot)$ is the (symmetric) cumulative probability function assumed for e_i. The intuition of the model is that the claims with explanatory variables similar to those designated as fraud should also be designated as fraud.

Within the response misclassification framework, assume that the dependent variable can report an incorrect outcome. Call the observed binary variable Y_i. Assume that the probability of a response error only depends on the true value $\tilde{Y}_i$ and that the misclassification error does not depend on any other explanatory variables. Assume that the probability of misclassification is as follows:

$$\begin{aligned} \gamma_0 &= \mathrm{Prob}(Y_i = 1|\tilde{Y}_i = 0), \\ \gamma_1 &= \mathrm{Prob}(Y_i = 0|\tilde{Y}_i = 1), \end{aligned} \tag{4}$$

where γ_0 is the probability that the true outcome is 0, but the observed outcome is 1 (commission error). On the other hand, γ_1 is the probability that the true outcome is 1, but the observed outcome is 0 (omission error).[4]

With no response misclassification, Y_i is always equal to $\tilde{Y}_i$, and its conditional expectation is given by Equation (3). When response misclassification exists, the conditional expectation of the observed dependent variable is given by:

$$\begin{aligned} E(Y_i|X_i) &= \mathrm{Prob}(Y_i = 1|X_i) \\ &= \mathrm{Prob}(Y_i = 1|\tilde{Y}_i = 1)\mathrm{Prob}(\tilde{Y}_i = 1|X_i) + \mathrm{Prob}(Y_i = 1|\tilde{Y}_i = 0)\mathrm{Prob}(\tilde{Y}_i = 0|X_i) \\ &= (1 - \gamma_1)F(\beta' X_i) + \gamma_0(1 - F(\beta' X_i)) = \gamma_0 + (1 - \gamma_0 - \gamma_1)F(\beta' X_i), \end{aligned} \tag{5}$$

which is equal to Equation (3) if $\gamma_0 = \gamma_1 = 0$.

In the absence of response misclassification, we can obtain the estimates of the parameter vector using the maximum likelihood method (see Hausman et al., 1998). The log-likelihood function (L) is:

$$\ln L = \frac{1}{n}\left[\sum_{i=1}^{n} Y_i \ln F(\beta' X_i) + (1 - Y_i)\ln(1 - F(\beta' X_i))\right]. \tag{6}$$

When we consider response misclassification, the likelihood function depends two further parameters. That is to say, the corrected log-likelihood function (L_{co}) is:

$$\begin{aligned} \ln L_{co} = \frac{1}{n}\Bigg[\sum_{i=1}^{n} & Y_i \ln(\gamma_0 + (1 - \gamma_0 - \gamma_1)F(\beta' X_i)) \\ & + (1 - Y_i)\ln(1 - \gamma_0 - (1 - \gamma_0 - \gamma_1)F(\beta' X_i))\Bigg], \end{aligned} \tag{7}$$

If we maximize with respect to parameters, γ_0, γ_1 and β, we obtain the following first-order conditions:

[4] Note that Equation (4) is a simple assumption because one could argue that omission error may depend on exogenous observable factors.

$$\frac{\partial \ln L_{co}}{\partial \gamma_0} = \frac{1}{n}\left[\sum_{i=1}^{n} Y_i \frac{1-F(\beta' X_i)}{\gamma_0 + (1-\gamma_0-\gamma_1)F(\beta' X_i)} + (1-Y_i)\frac{F(\beta' X_i)-1}{(1-\gamma_0-(1-\gamma_0-\gamma_1)F(\beta' X_i))}\right] = 0,$$

$$\frac{\partial \ln L_{co}}{\partial \gamma_1} = \frac{1}{n}\left[\sum_{i=1}^{n} Y_i \frac{-F(\beta' X_i)}{\gamma_0 + (1-\gamma_0-\gamma_1)F(\beta' X_i)} + (1-Y_i)\frac{F(\beta' X_i)}{(1-\gamma_0-(1-\gamma_0-\gamma_1)F(\beta' X_i))}\right] = 0,$$

$$\frac{\partial \ln L_{co}}{\partial \beta} = \frac{1}{n}\left[\sum_{i=1}^{n}\left[Y_i \frac{(1-\gamma_0-\gamma_1)f(\beta' X_i)}{\gamma_0 + (1-\gamma_0-\gamma_1)F(\beta' X_i)} + (1-Y_i)\frac{-(1-\gamma_0-\gamma_1)f(\beta' X_i)}{(1-\gamma_0-(1-\gamma_0-\gamma_1)F(\beta' X_i))}\right]X_i\right] = 0.$$

We use a scoring method to obtain the maximum likelihood estimates as in Hausman et al. (1998). The information matrix is then:

$$I = -E\begin{bmatrix} \frac{\partial^2 \ln L_{co}}{\partial \gamma_0^2} & \frac{\partial^2 \ln L_{co}}{\partial \gamma_0 \partial \gamma_1} & \frac{\partial^2 \ln L_{co}}{\partial \gamma_0 \partial \beta'} \\ \frac{\partial^2 \ln L_{co}}{\partial \gamma_1 \partial \gamma_0} & \frac{\partial^2 \ln L_{co}}{\partial \gamma_1^2} & \frac{\partial^2 \ln L_{co}}{\partial \gamma_1 \partial \beta'} \\ \frac{\partial^2 \ln L_{co}}{\partial \beta \partial \gamma_0} & \frac{\partial^2 \ln L_{co}}{\partial \beta \partial \gamma_1} & \frac{\partial^2 \ln L_{co}}{\partial \beta \partial \beta'} \end{bmatrix}$$

$$= E\begin{bmatrix} \frac{(1-F)^2}{P(1-P)} & -\frac{F(1-F)}{P(1-P)} & \frac{(1-\gamma_0-\gamma_1)f(1-F)}{P(1-P)}X' \\ -\frac{F(1-F)}{P(1-P)} & \frac{F^2}{P(1-P)} & -\frac{(1-\gamma_0-\gamma_1)fF}{P(1-P)}X' \\ \frac{(1-\gamma_0-\gamma_1)f(1-F)}{P(1-P)}X & -\frac{(1-\gamma_0-\gamma_1)fF}{P(1-P)}X & \frac{(1-\gamma_0-\gamma_1)^2 f^2}{P(1-P)}XX' \end{bmatrix},$$

In the previous expressions, we denote $f \equiv f(\beta' X_i)$, $F \equiv F(\beta' X_i)$ and $P \equiv \gamma_0 + (1-\gamma_0-\gamma_1)F(\beta' X_i)$. We have also eliminated subscript i for brevity.

Using the logit model specification, which follows from the assumption of a logistic distribution for the disturbance term in Equation (3),

$$F(\beta' X_i) = \frac{e^{\beta' X_i}}{1+e^{\beta' X_i}} \tag{8}$$

$$f(\beta' X_i) = \frac{F(\beta' X_i)}{1+e^{\beta' X_i}}. \tag{9}$$

In this application, we do not admit commission errors. Instead, we focus on the estimation of the omission error (γ_1). Assume that the commission error (γ_0) is fixed and known. So γ_0 is equal to zero because we are sure that no misclassification is present in the observed fraudulent claims group. The estimation algorithm has been implemented in SAS/IML, and the routines are available from the authors upon request.

Results

Maximum likelihood estimates for the logit model with omission error appear below. The dependent variable is the type of claim observed. If fraud is observed, the dependent variable equals one; otherwise, it is zero.

Table 3 presents the parameter signs and expected parameter signs. Except for the variable regarding the presence of witnesses, all parameter signs are in accordance to what is expected and to what was obtained in previous studies (see Artís et al., 1999). The results of the logit model with omission error are not very different from the results that were obtained for the standard binary logit model (without omission error).

The likelihood ratio test is 722.994 with 21 degrees of freedom, which indicates that a significant improvement occurs in the model when one includes the explanatory variables and the omission error parameter, if one compares it with the restricted model with only the constant term and no omission error. When one compares the estimation results of the model with all the explanatory variables and the omission error parameter and the model with no omission error, the likelihood ratio test is equal to 2.094 with 1 degree of freedom, which is only a slight improvement.

Table 3
Comparison for the Obtained and the Expected Parameter Signs

	Obtained Signs	Expected Signs
AGE	–	–
LICENSE	NS	–
RECORDS	+	+
COVERAGE	+	+
DEDUCTIBLE	NS	–
ACCESORI	–	–
VEHUSE	–	–
VEHAGE	NS	+
FAULT	+	+
NONURBAN	+	+
NIGHT	+	+
WEEKEND	+	+
WITNESS	+	–
POLICE	–	–
ZONE1	+	+
ZONE3	+	+
REPORT	+	+
NAMES	+	+
PROXIM	+	+
DELAY	+	+

Note: NS means nonsignificant.

When one compares the results of the model with omission error and the model without omission error, the same significant parameter estimates show up except for the regressor, indicating the existence of an extended coverage including accessories in the car. Otherwise, none of the estimates in the two models differ from one another. The omission error parameter estimate is significant at the 5 percent level.

One expected result is the estimated positive sign of the RECORDS coefficient. The insured who has accidents becomes aware of the claim. The influence of the insured's age is also significant, showing that younger insureds have a higher probability of fraudulent claims. Another variable, which might be related to the latter, is the number of years since the driver obtained a driver's license. No significant effect is found for this variable in either model.

From the estimation results, one concludes that when the insured has only a third-party-liability contract, *ceteris paribus*, the probability of the insured committing fraud is higher than that of another driver with extended coverage. However, private-use vehicles have a lower probability of being involved in dishonest claims than non-private-use vehicles, which include trucks and motorbikes. These last two conclusions can be drawn from the statistical significance of the parameter estimates accompanying the variables COVERAGE and VEHUSE.

The coefficient corresponding to the dichotomous DEDUCTIBLE variable is not statistically significant, so the existence of a deductible (apparently) does not affect the probability of fraud. The effect of this characteristic may be justified by the existence of some correlation with other variables, so that its effect is already captured by some other regressors, such as age, type of vehicle, or type of coverage.

The influence of an extension of the policy contract to cover accessories is significant at the 10 percent significance level in the omission error model and is not significant in the logit model. Indeed, when the contract already covers these parts, the insured is not going to be tempted to inflate the costs to receive compensation for them (see Table 4).

We also used a dummy variable to illustrate a situation when the insured driver and a third party involved in the accident (usually another driver) have the same family name. In fact, the effect of this variable indicates that whenever the names are the same, the probability of fraud rises. This may indicate a family relationship between the drivers involved in the accident, who may have planned it to obtain a benefit from the insurance coverage.

Another interesting result is that insureds who accept the blame in the accident are more likely to be cheating the company. They may have reached an agreement with the third party. They may think that if they accept the blame, they are less likely to be audited because they will already receive a penalty due to the bonus-malus system.

Vehicle age does not have a great influence. The sign of the parameter estimate is positive, as expected, because drivers with older cars may be tempted to obtain the cash value of the car from the insurance company before buying a new car.

The variable DELAY indicates whether the insured waited too many days to report the claim. Spanish regulation requires a maximum delay of one week. We found evidence that a claim is more likely to be fraudulent if there is a long delay in contacting the

TABLE 4
Estimation Results for the Logit Model With and Without Omission Error

	Logit Model With Omission Error*		Logit Model Without Omission Error	
	Coefficients	*P* Value	Coefficients	*P* Value
CONSTANT	−1.457[a]	0.000	−1.440[a]	0.000
AGE	−0.023[a]	0.006	−0.021[a]	0.006
LICENSE	0.005	0.684	0.003	0.762
RECORDS	0.200[a]	0.000	0.177[a]	0.000
COVERAGE	0.876[a]	0.001	0.795[a]	0.001
DEDUCTIBLE	−0.335	0.468	−0.303	0.488
ACCESORI	−0.420[b]	0.084	−0.350	0.114
VEHUSE	−0.562[a]	0.008	−0.507[a]	0.006
VEHAGE	0.010	0.487	0.012	0.354
FAULT	1.565[a]	0.000	1.388[a]	0.000
NONURBAN	0.594[a]	0.013	0.559[a]	0.008
NIGHT	1.787[a]	0.000	1.488[a]	0.000
WEEKEND	0.317[a]	0.021	0.274[a]	0.026
WITNESS	1.466[a]	0.043	1.140[b]	0.081
POLICE	−1.943[a]	0.000	−1.805[a]	0.000
ZONE1	0.345[b]	0.086	0.320[b]	0.084
ZONE3	0.712[a]	0.000	0.642[a]	0.000
REPORT	0.624[a]	0.000	0.562[a]	0.000
NAMES	1.284[a]	0.000	1.172[a]	0.000
PROXIM	1.989[a]	0.004	1.716[a]	0.001
DELAY	1.315[a]	0.000	1.212[a]	0.000
γ_1	0.047[a]	0.043	—	—
γ_0	0.000	—	—	—
Log-likelihood:	−1021.331		−1022.378	

Sample size = 1995; Restricted log-likelihood = −1382.828
[a] Indicates significance at the 5 percent level.
[b] indicates significance at the 10 percent level.
* LR test = 722.994; Pseudo-R^2= 26.15%; Performance index = 0.32.

company. This finding is consistent with the general idea that time is required to build up an accident with fraud.

The time and place of the accident also have a significant influence on the probability of the existence of fraud. Those accidents that occur at night (between 11 p.m. and 5 a.m.), those that occur on the weekend, or those that occur in a nonurban area are more likely to involve some fraud.

The influence of the variable PROXIM demonstrates that planned fraud may be as evident as contracting the policy after having the accident. In some unusual circumstances, a contract may have an effective date that precedes the issuing date. An accident between those two dates is reported and therefore covered by the insurance contract. Therefore, one practical recommendation for companies is to further control the way agents issue new policies.

The existence of witnesses at the scene of the accident is also an indicator of a higher probability of fraud. This fact could not be derived from the previous univariate analysis. But it does make clear the need to audit the nature of these witnesses and their relationship to the insured.

All variables related to the driving zone are based on the definition that is given by UNESPA (the Spanish association of insurance companies). They refer to three geographical regions in the country, with high, medium, and low level of accidents, respectively. The highest zone corresponds to the north of the country, which has the toughest weather conditions; the medium zone includes Madrid and Catalonia, where the number of vehicles is higher; while the low zone indicates the rest. The significance of these coefficients indicates an important difference in claimant behavior across the country.

Another interesting analysis included the textual study of the written reports. Claiming requires the insured to report a short explanation of the accident circumstances. We found that some language was more related to fraud than other language. This is a very preliminary result, but it provides evidence that fraud is more likely if the accident occurs under certain circumstances (parking, driving backward, overtaking).

As expected, the presence of the police at the site of the accident is a significant deterrent to fraud.

Finally, the parameter estimating the probability of omission error is significantly different from zero. The estimation results show that approximately 5 percent of honest claims are misclassified. While these results were indeed found within the context of our model, with these explanatory variables, they confirm that an improvement in fraud control is necessary to identify these claims. We do not propose that this figure must be directly extrapolated to the whole portfolio, but that it is an estimation of the percentage of undetected fraud within honest claims using current auditing technology. Nevertheless, since in practice the number of claims that investigators classify under the honest category is comparatively larger than the amount of detected fraud, our intuition is that the overall percentage of undiscovered fraud is moderate.

The novelty of this result is that our approach demonstrated the existence of a significant number of undetected fraudulent claims, conditional on the information in the data set. The performance of the two models in terms of explaining the determinants of a fraudulent claim is similar. The omission error parameter inclusion shifts the probability of the presence of fraud upward because it accounts for the shortcomings of the system that was used to generate the classification in the available data set. The probability increase due to the presence of the omission error parameter in the model is not systematic, because of the nonlinear nature of the model specification.

In addition, for the logistic regression model with omission error, parameters are not directly related to the measurement of the odds ratio due to the presence of the

TABLE 5
Classification Table for the Logit Model With Omission Error

	Predicted Type		
	Legitimate	Fraudulent	Total
Observed Type			
Legitimate	698	300	998
Fraudulent	220	777	997
Total	918	1077	1995

When the estimated probability of fraud exceeded 0.45, the predicted type was fraud.

TABLE 6
Classification Table for the Logit Model Without Omission Error

	Predicted Type		
	Legitimate	Fraudulent	Total
Observed Type			
Legitimate	708	290	998
Fraudulent	229	768	997
Total	937	1058	1995

When the estimated probability of fraud exceeded 0.45, the predicted type was fraud.

parameter that estimates the probability of omission error. This makes direct interpretation of the parameter estimates less straightforward, and we can only see the influence of the explanatory variables in the sign of the parameter estimate. Therefore, parameter estimates shown on the left side of Table 3 are only interpretable as log-odds, as they would come within the context of a logit model, if we think of the outcome as the true category.

We also report here some measurements of model adequacy. Tables 5 and 6 show the classification table using the same sample that was used for estimation purposes, with and without omission error. We set the threshold equal to 0.45 to produce optimal classification results (the highest percentage of correct dichotomous predictions). This means that whenever a claim had a predicted (adjusted) probability of fraud higher than 0.45, it was classified under the fraudulent group. Otherwise, the model predicted that it belonged to the legitimate group. We obtained the threshold level using a grid search method and choosing a compromise between the best overall classification and the best percentage for fraudulent claims.[5]

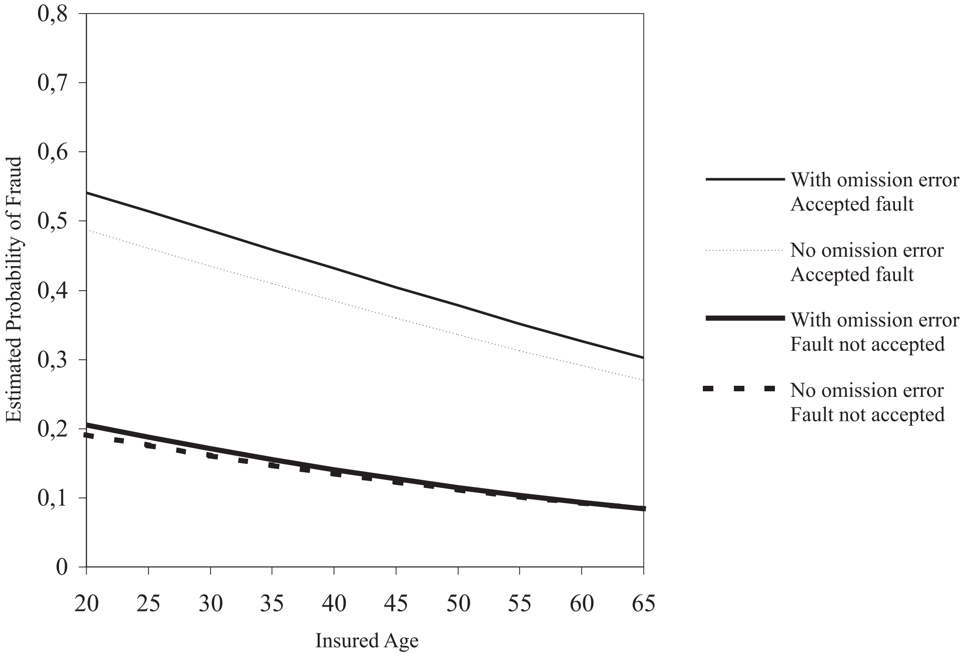
[5] Note that this threshold is also intuitive if we consider that the percentage of true fraudulent claims in the sample is 55 percent, since we estimated approximately 5 percent of undetected fraudulent claims in the honest group.

The total percentage of correct classification was 74 percent, which is acceptable. The conditional percentage of fraudulent claims that were correctly classified was 78 percent. However, the conditional percentage of legitimate claims that were correctly classified was 70 percent.

We also present some examples to show prediction results. Two predictions are compared: (1) the estimated probability of fraud when the omission error is assumed to be zero, and (2) the estimated probability of (true) fraud when the omission error is assumed to be larger than zero.

Figure 1 shows how the estimated probability of fraud for one type of claim decreases when the insured's age rises. The insured age ranged between about 20 and 65 years old in this research. We chose a claim with the following characteristics: The driver is a man who has 12 years of driving experience. The policy is for third-party coverage. There is no deductible or accessories coverage. The car is for private use and is 3 years old. The accident takes place at night, in a nonurban area with a low level of accidents, on a weeknight, without witnesses. There is a police report. The written report is not suspicious. The delay in communicating the accident to the company is normal. The insured person and the other party do not have coincident family names. Finally, the accident does not take place between the policy issue date and the effective starting date. In the figure, we see the two groups of curves. The upper curves represent a situation in which the insured driver accepts the blame; the lower curves represent a situation in which the insured does not. In each case, two curves are presented: The

FIGURE 1
Estimated Probability of Fraud for a Particular Type of Claim When Considering Omission Error (solid line) and When No Omission Error Is Assumed (dotted line)



solid line uses the parameter estimates for the logit model with omission error and the dotted line plots the estimated curve that would result if no omission error were assumed.

Figure 1 shows that the model with omission error results in a larger estimated probability of fraud. The probability of fraud decreases as the insured age increases. The effect of not considering omission error is not the same in all cases. A big difference exists between the same claim when considering whether or not the insured driver accepted the blame.

Final Remarks

Asymmetric information between agents in insurance contracts causes distortion of the insurance process. Unexpected inflation of costs distorts the rating mechanism.

Different authors have shown that the existence of moral hazard and adverse selection is important in insurance policy underwriting and claims auditing. In econometric studies, attention is devoted to finding models that provide a way to estimate the probability of the presence of fraud in a claim.

We have shown how to correct the model to take into account misclassification of the type of claim. Results are discussed for a sample of claims, and we show that a small, significant proportion of the claims that are observed to be honest are likely to contain omission error. In other words, a significant part of the fraud is not detected. The usefulness of this method for insurers is twofold. First, it estimates the proportion of fraudulent claims that is not detected by the auditing technology. And second, when omission error is significant, this method indicates that the model may not be correctly specified or that the number of exogenous variables is not sufficient.

A cost-benefit analysis is required to determine the audit strategy in the portfolio. The logit model that we have proposed does not imply that the insurer should audit all claims that are predicted as fraud. Costs have to be taken into account in any further step. Our model can be useful to assess the performance of an existing classification method, as it can test for the presence of omission error and can also evaluate its magnitude, conditional on the exogenous information on claims. In other words, it can be used to assess the “efficiency” of warning indicators.

References

Artís, M., M. Ayuso, and M. Guillén, 1999, Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market, *Insurance: Mathematics and Economics*, 24: 67-81.

M. Ayuso, Modelos Econométricos para la Detección del Fraude en el Seguro del Automóvil, 1998, Ph.D. thesis, Universidad de Barcelona.

Belhadji, E. B., and G. Dionne, 1997, Development of an Expert System for the Automatic Detection of Automobile Insurance Fraud, Working paper 97-06, École des Hautes Études Commerciales, Université de Montréal.

Bollinger, C. R., and M. H. David, 1997, Modelling Discrete Choice With Response Error: Food Stamp Participation, *Journal of the American Statistical Association*, 92: 827-835.

Bond, E. W., and K. J. Crocker, 1997, Hardball and the Soft Touch: The Economics of Optimal Insurance Contracts with Costly State Verification and Endogenous Monitoring Costs, *Journal of Public Economics*, 63: 239-264.

Borch, K., 1990, *Economics of Insurance*, Advanced Textbooks in Economics, Vol. 29 (Amsterdam: North-Holland).

Boyer, M., 1999, When Is the Proportion of Criminal Elements Irrelevant? A Study of Insurance Fraud When Insurers Cannot Commit, in: G. Dionne and C. Laberge-Nadeau, eds., *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation* (Boston, Mass.: Kluwer).

Brockett, P. L., R. A. Derrig, and X. Xia, 1998, Using Kohonen's Self Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud, *Journal of Risk and Insurance*, 65: 245-274.

Caron, L., and G. Dionne, 1999, Insurance Fraud Estimation: More Evidence From the Quebec Automobile Insurance Industry, in: G. Dionne and C. Laberge-Nadeau, eds., *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation* (Boston, Mass.: Kluwer).

Cosslett, S. R., 1993, Estimation from Endogenously Stratified Samples, in: G. S. Maddala, C. R. Rao, and H. D. Vinod, eds., *Handbook of Statistics*, Vol. 11, Econometrics (Amsterdam: North-Holland).

Crocker, K. J., and J. Morgan, 1998, Is Honesty the Best Policy? Curtailing Insurance Fraud Through to Optimal Incentive Contracts, *Journal of Political Economy*, 106: 355-375.

Crocker, K. J., and S. Tennyson, 1999, Costly State Falsification or Verification? Theory and Evidence from Bodily Injury Liability Claims, in: G. Dionne and C. Laberge-Nadeau, eds., *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation* (Boston, Mass.: Kluwer).

Derrig, R. A., and K. M. Ostaszewski, 1995, Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification, *Journal of Risk and Insurance*, 62: 447-482.

Derrig, R. A., and H. I. Weisberg, 1998, AIB PIP Claim Screening Experiment Final Report. Understanding and Improving the Claim Investigation Process, *AIB Filing on Fraudulent Claims Payment*, DOI Docket R98-41 (Boston, Mass.: Department of Insurance).

Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton, 1998, Misclassification of the Dependent Variable in a Discrete-Response Setting, *Journal of Econometrics*, 87: 239-269.

Picard, P., 1996, Auditing Claims in the Insurance Market With Fraud: The Credibility Issue, *Journal of Public Economics*, 63: 27-56.

Picard, P., 2000, Economic Analysis of Insurance Fraud, in: G. Dionne, ed., *Handbook of Insurance* (Boston, Mass.: Kluwer).

Poterba, J. M., and L. H. Summers, 1995, Unemployment Benefits and Labour Market Transitions: A Multinomial Logit Model with Errors in Classification, *The Review of Economics and Statistics*, 77: 207-216.

Weisberg, H. I., and R. A. Derrig, 1993, *Quantitative Methods for Detecting Fraudulent Automobile Bodily Injury Claims* (Boston, Mass.: Automobile Insurers Bureau of Massachusetts).