© *The Journal of Risk and Insurance*, 2005, Vol. 72, No. 4, 539-550

Fraud Detection Using a Multinomial Logit Model With Missing Information

Steven B. Caudill
Mercedes Ayuso
Montserrat Guillén

Abstract

Recently, Artís, Ayuso, and Guillén (2002, *Journal of Risk and Insurance* 69: 325–340; henceforth AAG) estimate a logit model using claims data. Some of the claims are categorized as "honest" and other claims are known to be fraudulent. Using the approach of Hausman, Abrevaya, and Scott-Morton (1998 *Journal of Econometrics* 87: 239-269), AAG estimate a modified logit model allowing for the possibility that some claims classified as honest might actually be fraudulent. Applying this model to data on Spanish automobile insurance claims, AGG find that 5 percent of the fraudulent claims go undetected. The purpose of this article is to estimate the model of AAG using a logit model with missing information. A constrained version of this model is used to reexamine the Spanish insurance claim data. The results indicate how to identify misclassified claims. We also show how misclassified claims can be identified using the AAG approach. We show that both approaches can be used to probabilistically identify misclassified claims.

Introduction

Fraud detection in insurance is a subject of increasing interest because there are now many ways of managing information to help the insurer identify fraudulent claims. Since the early work of Dionne (1984) and the theoretical contributions of Picard (1996, 2000), many studies have followed (see, e.g., Derrig (2002) for an overview of recent developments). There are many alternative approaches to fraud detection and deterrence (see Brockett, Xia, and Derrig, 1998; Dionne and Gagné, 2001; Brockett et al., 2002; Major and Riediger, 2002; Viaene et al., 2002; among others). Currently, identification of fraudulent claims is achieved using a scoring method to implement a claim auditing strategy. In fact, a recent work by Dionne, Giuliano, and Picard (2003)

Steven B. Caudill is at the Department of Economics, Auburn University, AL 36849. The author can be contacted via e-mail: scaudill@business.auburn.edu. Mercedes Ayuso and Montserrat Guillen are at the Departament d'Econometria, Estadística i Economia Espanyola, Universitat de Barcelona, Spain. The authors thank the editors and two anonymous referees for their helpful suggestions. They also acknowledge grants-SEC01-3672, SEC01-2581, and BBVA.

proposes an optimal audit policy, which substantially improves the insurer's savings. Fugate, Marathe, and Scovel (2003) deal with a similar problem and discuss how to deal with choice-based samples.

One method of fraud detection involves the use of multinomial logit models based on misclassified data. When the response is reported or recorded in the wrong category (e.g., a variable is recorded as zero when it should have been recorded as one), the problem is one of misclassification. Insurance claims categorized as "honest" that are actually fraudulent are misclassified. Several approaches to the estimation of logit models based on misclassified data exist in the literature (see, for example, Poterba and Summers, 1995; Bollinger and David, 1997; Hausman, Abrevaya, and Scott-Morton, 1998). Identifying fraudulent claims is very similar to a problem addressed in the medical and epidemiological literature when discussing treatment effects on recovery of an illness. For instance, Magder and Hughes (1997) estimate a logistic regression model based on misclassified data and apply the results to the prediction of smoking cessation.

Recently, Artís, Ayuso, and Guillén (2002) estimate a logit model based on misclassified data using Spanish automobile insurance claims. Some of the claims are categorized as honest and other claims are known to be fraudulent. Using the approach of Hausman, Abrevaya, and Scott-Morton (1998), the authors estimate a modified logit model allowing for the possibility that some claims originally observed as honest might actually be fraudulent. Applying this model to data on Spanish automobile insurance claims, AGG find that 5 percent of the truly fraudulent claims are observed as honest.

The purpose of this article is to estimate the model of AAG using a different approach to fraud detection. The approach presented in this article is based on a logit model with missing information. This model can be estimated by maximum likelihood using the EM algorithm (Dempster, Laird, and Rubin, 1997). A constrained version of this model is used to reexamine the Spanish insurance claim data. The constrained version of the model examined provides a direct estimate of the probability that an honest claim is actually fraudulent. We also show how the probability that an honest claim is actually fraudulent can be calculated in the original AAG model.[1] Both methods indicate that 5 percent of the fraudulent claims go undetected.

This focus on individual misclassification probabilities indicates the weaknesses of the preliminary fraud identification method, which was used to collect the sample of claims. Prediction models developed from this initial sample will only be able to reproduce the sample patterns and might suffer from the same limitations when trying to predict the nature (fraudulent or honest) of new claims.

A LOGIT MODEL WITH MISSING INFORMATION

In this article, the traditional multinomial logit model (MNL) is extended to estimate models in which some choices are not fully observed. To develop our new model, we begin, without loss of generality, with an MNL model in which claims are identified as "honest" or "fraudulent." In this framework, we assume that for each individual i

[1] We thank an anonymous referee for pointing this out to us and providing details.

$(i = 1, \ldots, n)$ a vector of dependent variables is observed $(Y_{i0,}\ Y_{i1})$. If the ith claim is legitimate, then Y_{i0} is equal to one and zero otherwise. Similarly, Y_{i1} equals one for a fraudulent claim. Probabilities in this model are given by

$$P_{ij} = P(Y_{ij} = 1 \mid X_i) = \frac{\exp(\alpha_j + X_i\beta_j)}{\sum_{j=0}^{1} \exp(\alpha_j + X_i\beta_j)}, \quad j = 0, 1. \tag{1}$$

The α's and β's are parameters to be estimated and X_i is a vector of exogenous variables. The log-likelihood function in this MNL model is given by

$$\text{Log}L = \sum_{i=1}^{n} [Y_{i0} \log(P_{i0}) + Y_{i1} \log(P_{i1})], \tag{2}$$

where n is the sample size. Maximum likelihood estimation of the parameters in this model is routine. We wish to extend this model to the case where some of the claims categorized as honest are actually fraudulent. We develop an EM algorithm for maximum likelihood estimation of the model parameters under this assumption.

The EM algorithm has been used previously to estimate similar models. For example, Dempster, Laird, and Rubin (1977) use the EM algorithm to estimate a multinomial probability example with parameter constraints imposed. More recently, Magder and Hughes (1997) use the EM algorithm to estimate a binomial probability model with misclassification.

To begin the development of the EM algorithm, we assume the MNL model has hidden choices. The two observable choices are the classifications, "honest" and "fraudulent." We also assume that the "honest" classification contains two, as yet unnamed, types of claims. The result is a logit model with three alternatives: honest A, honest B, and fraudulent. We know whether claims are categorized as honest or fraudulent, but we do not know which "honest" claims are honest A or honest B. This is the missing information in our logit model.

Our goal is to estimate this model by maximum likelihood. As a first step, we extend the usual MNL model above by denoting the probabilities associated with the alternatives as $P_{i,00}$ (honest A) and $P_{i,01}$ (honest B), and P_{i1} (fraudulent). These probabilities are now given by

$$\begin{aligned} P_{i,0k} &= \frac{\exp(\alpha_{0k} + X_i\beta_{0k})}{\sum_{k=0}^{1} [\exp(\alpha_{0k} + X_i\beta_{0k})] + \exp(\alpha_1 + X_i\beta_1)}, \quad k = 0, 1, \\ P_{i1} &= \frac{\exp(\alpha_1 + X_i\beta_1)}{\sum_{k=0}^{1} [\exp(\alpha_{0k} + X_i\beta_{0k})] + \exp(\alpha_1 + X_i\beta_1)}, \end{aligned} \tag{3}$$

where X_i is a vector of exogenous variables (the same set as used by AAG in the original paper) and the α's and β's represent parameters to be estimated. Probabilities must sum to one, so $P_{i,00} + P_{i,01} + P_{i1} = 1$. The resulting incomplete-data/observed log-likelihood function is given by

$$\mathrm{Log}L = \sum_{i=1}^{n} [Y_{i0} \log(P_{i,00} + P_{i,01}) + Y_{i1} \log(P_{i1})], \tag{4}$$

where Y_{ij} is the usual dummy variable described above indicating "honest" and "fraudulent" claims. Clearly, the parameters of this model cannot be directly estimated by maximum likelihood because whether an "honest" claim is honest A or honest B is unknown. However, rewriting the model as a missing data problem allows one to use the EM algorithm of Dempster, Laird, and Rubin (1977) for maximum likelihood estimation.

If one knew which of the "honest" claims are honest A and which are honest B, maximum likelihood estimation of the model above would simplify to the estimation of the usual MNL model. With that in mind, we denote the set of unobservable indicator variables associated with, respectively, "honest A" and "honest B" by $Y^*_{i,00}$ and $Y^*_{i,01}$. Using these unobservable variables, the complete data log likelihood can be written as

$$\mathrm{Log}L = \sum_{i=1}^{n} \left[Y^*_{i,00} \log(P_{i,00}) + Y^*_{i,01} \log(P_{i,01}) + Y_{i1} \log(P_{i1})\right]. \tag{5}$$

The likelihood function in (5) characterizes the estimation problem as a missing data problem. If the unobserved Y^**'s* were known, estimation would be as simple as estimating the usual MNL model.

The EM algorithm provides a straightforward procedure to obtain parameter estimates for the extended MNL model. In the expectations or "E" step of the EM algorithm, the unobserved Y^**'s* in (5) are replaced with their conditional expectations given the data and values of the unknown parameters. These conditional expectations or probabilities are well known given the nested structure of our logit model and can be calculated as

$$\begin{aligned} E\left(Y^*_{i,00} \mid Y_{i0} = 1\right) &= \frac{\exp(\alpha_{00} + X_i\beta_{00})}{\exp(\alpha_{00} + X_i\beta_{00}) + \exp(\alpha_{01} + X_i\beta_{01})}, \\ E\left(Y^*_{i,01} \mid Y_{i0} = 1\right) &= \frac{\exp(\alpha_{01} + X_i\beta_{01})}{\exp(\alpha_{00} + X_i\beta_{00}) + \exp(\alpha_{01} + X_i\beta_{01})}. \end{aligned} \tag{6}$$

With these conditional expectations inserted into the log-likelihood function in (5) in place of the Y^**'s*, the maximization or "M" step of the EM algorithm maximizes the resulting log-likelihood function. New parameter values are obtained and then the "E" step and the "M" step are repeated. This process continues until the likelihood function in (5) is maximized. Once the maximum has been found, standard errors are obtained using a single iteration of the algorithm of Berndt et al. (1974).

The general model developed above suffers from identification problems and cannot be estimated without the imposition of some constraint on the parameters.[2] The imposition of a constraint solves the identification problem and allows us to characterize the hidden categories in the "honest" claim category. The constraint we impose in the estimation will allow us to describe the two categories hidden in "honest" as truly honest and truly fraudulent. The constraint we need for identification and for the desired characterization of the hidden categories is $\beta_{01} = \beta_1$. This constraint is imposed in the estimation.

This constraint has the effect of pooling observations to allow some of the claims categorized as honest to be fraudulent. The constraint means that the coefficients of the explanatory variables are the same for a fraudulent claim and for an honest claim that is truly fraudulent. To see the impact of this constraint, recall that logit model parameters estimate effects on the logarithms of the odds. Consider the effect of a change in some explanatory variable, X, on the logarithm of the odds of "honest" but truly fraudulent over fraudulent. In general, without our parameter constraint, the marginal effect is given by

$$\frac{\partial \log(P_{i01}/P_{i1})}{\partial X_j} = \beta_{j01} - \beta_{j1}. \tag{7}$$

However, with the constraint imposed, the coefficients are equal and the marginal effect is zero. This means that a change in X cannot change the odds of "honest" but truly fraudulent over fraudulent. This is true because "honest" but truly fraudulent and fraudulent are the same—the constraint pools them.

With the constraint imposed, the probabilities in (6) become, respectively, the conditional probability that a claim is honest given that it is categorized as honest and the conditional probability that a given claim is fraudulent given that the claim is categorized as honest. Thus, the constraint will allow the identification of the model and permit us to identify misclassified claims.

FRAUD DETECTION

In the fraud detection problem, imposition of this constraint allows some of the claims initially classified as honest to be reclassified as fraudulent. The conditional expectations given in (6) then become, respectively, the probability that a claim is actually honest, given that a claim is categorized as "honest," and the probability that a claim is actually fraudulent, given that a claim is categorized as "honest."

Assuming that the parameter restrictions are $\beta_{01} = \beta_1 = 0$, one may write the posterior probability that a claim is actually fraudulent, given that a claim is categorized as "honest" as

[2] Although our focus is on the constrained version of the model, the details of the unconstrained version are presented here because the general EM algorithm may have applicability to other problems in fraud detection. For example, suppose one is interested in separating those committing fraud into fraud type (planned fraud, build-up fraud, and opportunistic fraud). The general EM algorithm presented here might allow one to separate the sample of fraudulent claims provided the identifiablilty issues discussed by Magder and Hughes (1997) can be addressed. They can likely be addressed without resorting to the parameter constraints imposed here.

$$E\left(Y_{i,01}^{*} \mid Y_{i0} = 1\right) = \frac{P_{i01}}{P_{i00} + P_{i01}} = \frac{P_{i01}}{1 - P_{i1}}. \tag{8}$$

Equation (8) shows that the claim-specific probability of misclassification can indeed be computed from model (5). In fact, the claim-specific probability of misclassification can also be calculated in the original AAG framework.

Our model and the original model of AAG are what Magder and Hughes (1997) refer to as constant sensitivity models where sensitivity is defined to be

$$\text{sensitivity} = P(Y^{*} = 1 \mid Y = 1) = \frac{P_1}{P_1 + P_{01}}. \tag{9}$$

In the original AAG article, sensitivity is $1 - \gamma_1$ and in our model sensitivity is $1/(1 + \exp(\alpha_{01}))$. Each of these results can be used to obtain the probability that an "honest" claim is actually a fraud. In our model, the probability is obtained directly from the estimation as indicated in Equation (8) above. In the original AAG framework, this probability is given by

$$P(Y^{*} = 1 \mid Y = 0) = \frac{P_1}{1 - P_1} \frac{\gamma_1}{1 - \gamma_1}. \tag{10}$$

The conditional probability that an honest claim is actually fraudulent can be calculated in the AAG framework, although a little work is needed.

Our new model and the AAG model provide information about probabilities that individual "honest" claims are fraudulent. This is new and important information that cannot be obtained by estimating a naïve logit model with no consideration of the fact that some observations are misclassified. That is, one cannot simply obtain the information about fraudulent claims by estimating a logit model, assuming all claims are correctly classified, and then reassigning those "honest" claims having low probabilities of being honest. The problem is that in the naïve approach, the misclassified observations are used to calculate the probabilities; hence the probabilities must be incorrect. The problem becomes worse as the fraction of observations misclassified increases.[3] Our new approach and the approach of AAG avoid this problem and are able to provide information about which claims are likely misclassified.

RESULTS

The data used for estimating the model correspond to a sample of the year 1995 claims for car damages obtained from a Spanish insurance company. All claims are investigated and they are classified as "fraudulent" or "honest." The variable names and definitions are given in the Appendix. Data were collected from 1993 to 1996. Half of the claims are legitimate, the other half are claims that had been identified as

[3] Hausman, Abrevaya, and Scott-Morton (1998) show that with misclassification of only 2 percent, parameters in a probit model are biased by 15 percent to 25 percent.

fraudulent. Data contain information on the accident characteristics (place, report of the police officer, etc.), the insured driver characteristics (history of previous claims, etc.), and vehicle characteristics (date of fabrication). The information contained in the sample is obtained from the claim statement or the policy. Each fraud has been classified by type of fraud, but this distinction is not taken into account in the present analysis.

The results from estimating our new logit model are given in Table 1. These estimation results are nearly identical to the results in the earlier paper by Artís, Ayuso, and Guillén (2002). The only difference is that the coefficients have all changed signs because the model developed here estimates the probability of an honest claim rather that the probability of a fraudulent claim as in the original work. As the table indicates, the probability that an honest claim is actually fraudulent is estimated from the sums of the probabilities. There are 998 claims categorized as "honest." Our model predicts that 50 of those can be reclassified as "fraudulent," resulting in a fraud probability of 0.05, a value very close to that in the original work.

From the estimation results, we can see that several characteristics are significant and positively related to a higher probability of an honest claim: age of the driver, a car for private use, and the existence of a police report. On the other hand, other indicators are significant and positively related to a higher probability that the claim is fraudulent *ceteris paribus*: number of previous claims, policy covers third party liability, the insured accepts that the accident is his/her fault, the accident occurs in a nonurban area, at night, or during the weekend, the insured person gives names of witnesses, the report shows suspicious claims, there is a family relationship between the drivers involved in the accident, and the claim is filed more than one week after the accident. Finally, other relevant fraud indicators are proximity to the issuing date and the driving zone. Because the results here are nearly identical to the results in AAG, we refer interested readers to the original paper for more details concerning the estimation results and instead focus on the contribution of this approach.

In Table 2, we give some interesting examples of the individual misclassification results obtained from the model estimates. Two different individuals have been selected. The characteristics of these two individuals are listed by rows, showing the value of the corresponding explanatory variable Case 1, shown in column 1, is a claimant selected from the sample data. It corresponds to the claim most likely misclassified as "honest." Conditional estimated probabilities are shown at the bottom. Given the claim is categorized as "honest," the claim has a conditional probability of actually being fraudulent of 72 percent. The estimated probability of fraud is 99 percent. The characteristics of this claim are consistent with the high probability that the claim is fraudulent: the claim is made by a 29-year-old driver who has been driving for 6 years, the driver has filed 13 previous claims, the driver does not have extended coverage, or a deductible, or coverage for accessories, the vehicle is 5 years old and for private use, the insured accepts blame for the accident which occurred in a nonurban area, the accident occurred at night or during the weekend, there are no witnesses and no police report, the accident did not occur in zone 1 or zone 3, there was a suspicious textual report given, there are different family names for the insured and other vehicle drivers, the accident did not occur between the policy issue date and the policy effective starting date, and the claim was not reported to the company within

TABLE 1
Logit Estimation Results

	P_{00}	P_{01}
Constant	1.506	−2.992
	(3.61)	(4.95)
Age	0.023	
	(2.68)	
License	−0.005	
	(0.41)	
Records	−0.200	
	(5.08)	
Coverage	−0.876	
	(3.13)	
Deductible	0.335	
	(0.68)	
Accesori	0.421	
	(1.66)	
Vehuse	0.562	
	(2.64)	
Vehage	−0.010	
	(0.67)	
Fault	−1.566	
	(9.22)	
Nonurban	−0.594	
	(2.41)	
Night	−1.789	
	(6.79)	
Weekend	−0.317	
	(2.25)	
Witness	−1.470	
	(2.62)	
Police	1.944	
	(7.39)	
Zone1	−0.345	
	(1.64)	
Zone3	−0.713	
	(5.08)	
Report	−0.624	
	(4.83)	
Names	−1.284	
	(3.90)	
Proxim	−1.990	
	(2.54)	
Delay	−1.315	
	(7.28)	
	$\sum P_{00} = 947.95$	$\sum P_{01} = 50.05$

Note: Numbers in parentheses are absolute values of *t*-ratios.

TABLE 2
Misclassification Results for Some Selected Cases

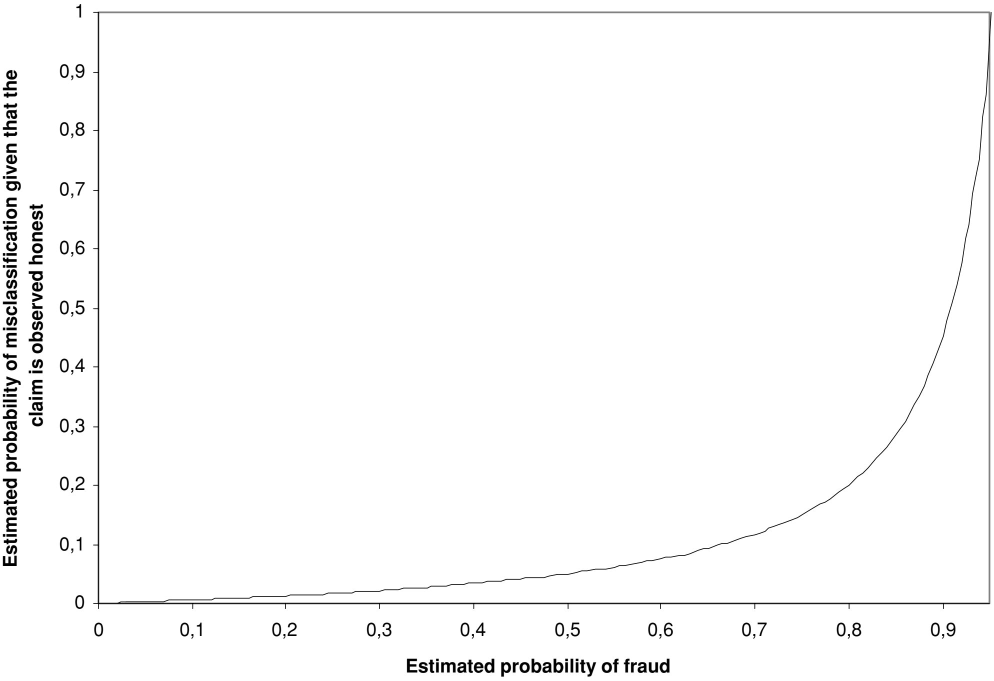
	Case 1	Case 2
Constant	1	1
Age	29	47
License	6	26
Records	13	1
Coverage	0	1
Deductible	0	0
Accesori	0	0
Vehuse	1	1
Vehage	5	8
Fault	1	1
Nonurban	1	1
Night	1	0
Weekend	1	0
Witness	0	0
Police	0	0
Zone1	0	0
Zone3	0	0
Report	1	1
Names	0	0
Proxim	0	0
Delay	1	0
P_{i1}	99%	80%
$E(Y_{i,01}^*\|Y_{i0}=1)$	72%	20%

the established time period. These characteristics do seem to be associated with a suspicious, and possibly fraudulent, claim.

The claim shown in the second column of Table 2 (Case 2) is observed as honest, but the logit model predicts the probability that the claim is fraudulent is 0.80. However, the conditional probability that the claim is fraudulent, given that the claim is recorded as honest, is only 0.20: this claim is filed by a 47-year-old driver, who has been driving for 26 years and has only filed one previous claim, the driver has third-party liability coverage, but no deductible and no coverage for accessories, the vehicle is privately used and is 8 years old, the insured accepted blame for the accident which occurred in a nonurban area, during the daytime, during the week, there were no witnesses and no police report, the accident occurred outside zone 1 and zone 3, a suspicious textual report was filed, there are different family names for the insured and other vehicle drivers, the accident did not occur between the policy issue date and the policy effective starting date, and the claim was reported to the company within the established time period. Collectively, these characteristics point to a claim that is likely honest.

Figure 1 shows a comparison of the relationship between the estimated probability that a claim is fraud and the probability of misclassification, given that it has been observed as honest according to our model estimates. Figure 1 is the graphical depiction of Equation (8). The conditional probability that a claim is actually fraudulent, given

FIGURE 1
Estimated Probability of Fraud Versus Estimated Probability of Misclassification Given That the Claim Is Observed Honest



that the claim is categorized as honest is the probability that the claim is fraudulent and categorized as honest divided by the probability that the claim is honest. This probability is equivalent to the probability that a claim is fraudulent and categorized as honest divided by one minus the probability that the claim is fraudulent. Written this way, one can see that, with the probability that a claim is fraudulent and categorized as honest held constant, increases in the probability that the claim is fraudulent will increase this conditional probability by making the denominator smaller. The plot clearly shows that the relationship is monotonic but nonlinear.

CONCLUSIONS

This article reestimates the model of AAG using a different approach to fraud detection based on a logit model with missing information. This article shows how this model can be estimated using the EM algorithm. A constrained version of this model is used to reexamine the Spanish insurance claim data. This new approach calculates the probability that an honest claim is fraudulent. We also show how this information can be obtained from the original AAG model.

The results obtained here are aimed at revising claims initially classified as honest by the insurer. By implementing this model, one may decide to identify the claims that are more likely to be fraudulent for two different purposes: either to reopen an investigation and more closely examine a claim; or to identify weaknesses in the initial classification system. In this sense, either method can be considered an *ex post*

Appendix

Explanatory Variables Used in the Model

Characteristics of the insured/claimant/policy:	
AGE	Age of insured driver when the accident occurred
LICENSE	Number of years since the insured obtained first driving 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
ACCESORI	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

prediction method because it evaluates the conditional probability that a claim is fraudulent, given that the claim is initially classified as honest.

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.

Artís, M., M. Ayuso, and M. Guillén, 2002, Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims, *Journal of Risk and Insurance*, 69: 325-340.

Berndt, E. R., B. H. Hall, R. E. Hall, and J. A. Hausman, 1974, Estimation and Inference in Nonlinear Structural Models, *Annals of Economic and Social Measurement*, 3: 653-665.

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

Brockett, P. L., R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert, 2002, Fraud Classification Using Principal Component Analysis of RIDITs, *Journal of Risk and Insurance*, 69: 341-371.

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

Cramer, J. S., and G. Ridder, 1991, Pooling States in the Multinomial Logit Model, *Journal of Econometrics*, 47: 267-272.

Derrig, R. A., 2002, Insurance Fraud, *Journal of Risk and Insurance*, 69: 271-288.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977, Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, 39: 1-38.

Dionne, G., 1984, The Effect of Insurance on the Possibilities of Fraud, *Geneva Papers on Risk and Insurance—Issues and Practice*, 9: 304-321.

Dionne, G., and R. Gagné, 2001, Deductible Contracts Against Fraudulent Claims: Evidence From Automobile Insurance, *Review of Economics and Statistics*, 83: 290-301.

Dionne, G., F. Giuliano, and P. Picard, 2003, Optimal Auditing for Insurance Fraud HEC, Chaire de recherche du Canada en gestion des risques. Cahiers de recherche 02–05, Université de Montreal.

Fugate, M., A. Marathe, and C. Scovel, 2003, Logistic Regression with Incomplete Choice-Based Samples, Los Alamos National Laboratory. Available at http://www.c3.lanl.gov/napc/pdf/logistic.pdf.

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

Magder, L. S., and J. P. Hughes, 1997, Logistic Regression When the Outcome Is Measured With Uncertainty, *American Journal of Epidemiology*, 146: 195-203.

Major, J., and D. R. Riedinger, 2002, EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud, *Journal of Risk and Insurance*, 69: 309-324.

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, Ch. 10 in: G. Dionne, ed., *Handbook of Insurance* (Boston: Kluwer Academic Press).

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

Viaene, S., R. A. Derrig, B. Baesens, and G. Dedene, 2002, A Comparison of the State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection, *Journal of Risk and Insurance*, 69: 373-421.