© *The Journal of Risk and Insurance*, 2007, Vol. 74, No. 2, 425-440

Selection Bias and Auditing Policies for Insurance Claims

Jean Pinquet
Mercedes Ayuso
Montserrat Guillén

Abstract

Selection bias results from a discrepancy between the range of estimation of a statistical model and its range of application. This is the case for fraud risk models, which are estimated on audited claims but applied on incoming claims in the design of auditing strategies. Now audited claims are a minority within the parent sample since they are chosen after a severe selection performed by claims adjusters. This article presents a statistical approach that counteracts selection bias without using a random auditing strategy. A two-equation model on audit and fraud (a bivariate probit model with censoring) is estimated on a sample of claims where the experts are left free to take the audit decision. The expected overestimation of fraud risk derived from a single-equation model is corrected. Results are close to those obtained with a random auditing strategy, at the expense of some instability with respect to the regression components set. Then we compare auditing policies derived from the different approaches.

Introduction

Auditing policies derived from statistical analysis and applied to insurance claims (see the section "Fraud Detection . . . Bias Issues" for an overview) face a major selection bias problem. A score which assesses fraud risk is derived from a regression analysis on audited claims, as fraud is checked on the audited claims only. Then the score is

Jean Pinquet is from Université de Paris X, U.F.R. de sciences économiques, 200 avenue de la République, 92001 Nanterre Cedex France; and Ecole Polytechnique, Département d'économie, 91128 Palaiseau Cedex France. Mercedes Ayuso and Montserrat Guillén are from the Departament d'Econometria, Estadistica i Economia Espanyola, Universitat de Barcelona, Diagonal 690, 08034 Barcelona Spain The first author can be contacted via email: pinquet@u-paris10.fr or jean.pinquet@polytechnique.edu. We thank Martin Boyer, Jean Lemaire, the editor, and two referees for their comments. Pinquet acknowledges financial support from the Fédération Française des Sociétés d'Assurance and from the programme national de recherche et d'innovation dans les transports terrestres (PREDIT). Ayuso and Guillén acknowledge financial support from the Spanish Ministry of Education and Science FEDER grant SEJ2004-5052. All the authors acknowledge financial support from the Spanish Ministry of Education and Science FEDER grants SEJ2004-5052 and SEJ2005-0741/ECON.

applied to incoming claims in order to select those which are then recommended for audit. Now audited claims are a minority within the parent sample as they are chosen after a severe selection performed by claims adjusters. This discrepancy between the range of derivation of the risk model (i.e., the audited claims) and the range of application (the incoming claims) creates a selection bias.

Random auditing of claims is the basic strategy which makes it possible to counteract selection bias. A pure random auditing strategy consists in picking claims at random, then in auditing these claims. This controlled experiment eliminates the selection induced by the audit decision. The estimation of a single fraud equation in this sample provides an estimated fraud probability for incoming claims which is not subject to selection bias.

Random auditing was partly carried out on the database we investigated. Twenty percent of the claims were thus selected and recommended for audit. However, only one claim out of five was eventually audited in this population because all the incoming claims were not suspicious with respect to fraud (see the section "Fraud Detection . . . Bias Issues" for more details about auditing processes). Optimal auditing strategies can be designed from the estimation of a fraud equation on the audited claims (see Ayuso et al., 2004, for derivations with the database used in this article).

Most insurance companies are reluctant to carry out a random auditing policy. This is because the long-term influence of an audit decision on the policyholder's value for the company is negative. Indeed, an honest policyholder may take the audit process amiss and his loyalty to the company should decrease as a consequence, as well as his value for the insurer. Hence companies are deterred from performing a systematic audit on part of their claims database. Without a random auditing policy, the effect of selection bias on fraud risk assessment can easily be anticipated. The experts take an audit decision based on the claim characteristics recorded in the company files. However, they are also able to capture idiosyncrasies in fraud distributions which are not summarized in the observable information. Given the observable characteristics, fraud risk is expected to be less significant for a claim exempted from audit by the expert than for another one checked for fraud. A fraud risk model derived without taking this selection problem into account would then overestimate fraud probabilities for the incoming claims.

In the absence of random auditing, a statistical approach can counteract selection bias by using a two equation model. An audit equation is estimated together with another equation for fraud, which is estimated only on the audited claims. From a joint distribution of the random components in each equation, we can condition fraud risk on claim characteristics and on whether or not the claim was retained for audit by the claims adjusters. To our knowledge, selection models have not been applied to date in the literature on claim auditing strategies.

Selection models can be designed with or without censoring for the variable of interest. In our context, data are censored, as fraud is checked on the audited claims only.[1] We will use a bivariate probit model for audit and fraud. The estimated correlation

[1] Heckman (1979) is the seminal paper on selection bias in a censored setting.

coefficient between the two equations is expected to be positive since it reflects the ability of claims adjusters to assess hidden characteristics in fraud distributions.

The article is organized as follows. The section "Fraud Detection . . . Bias Issues" explains how fraud detection strategies for automobile insurance claims are implemented in insurance companies. The section "Presentation of the Database" describes the database. The section "The Bivariate . . . Bias Assessment" presents the bivariate probit model and its application to the correction of selection bias. We consider a natural extension of the single equation probit model for the fraud variable (see Belhadji, Dionne, and Tarkhani, 2000; Artis, Ayuso, and Guillén, 2002 for the inclusion of misclassification risk in the fraud equation). Our claims database is split into two populations. Claims selected at random (one out of five) are recommended for audit, whereas there is no specific recommendation for other claims. We will use the latter population as the working sample, since such claims are subject to selection bias. The claims that are selected at random will be used as a holdout sample, to assess the efficiency of the statistical model. The estimated correlation coefficient in the bivariate probit model was found to be positive for our data. This means that, if we control for observable information on claims, fraud risk is lower for claims exempted from audit by the experts.

Optimal auditing policies which take into account selection bias are presented in the section "Applications of Selection . . . Policies Design" and the section "Conclusions" contains concluding remarks.

Fraud Detection Strategies on Automobile Insurance Claims and Selection Bias Issues

Most property and liability insurers include fraud warning systems in their routine claims handling process. In order to identify fraudulent insurance claims at an early stage, most systems score a new incoming claim using a set of fraud indicators. If the score is high enough, then the claim is audited and fraud (or abuse) may be confirmed.[2] Only claims with a high suspicion level (or score) are selected for investigation.

In most European countries, when an insurer finds a claim which shows evidence of fraud, an agreement with the policyholder is usually reached. Most settlements involve the policy not being renewed. Others lead to a reduction in the claim compensation. This situation is in contrast with the highly litigious U.S. system. However, even if very few fraudulent claims are brought to court, many European insurers are beginning to implement strategies to control fraud.

Let us now formalize the selection bias issue. If all incoming claims can be considered for audit, the selection bias issue can be formalized in the following way: A and F denote the binary variables related to audit and fraud, and x is the vector of variables which describe the claim. A statistical model assessing fraud risk is derived from the audited claims and estimates probabilities of the type $P(F = 1 \mid A = 1, x) = E(F \mid A = 1, x)$. Now an audit policy induced by this model is applied on the incoming claims, and uses the probabilities $P(F = 1 \mid x) = E(F \mid x)$. Selection bias is a consequence of the confusion between the conditional and unconditional probabilities.

[2] See Derrig (2002) for a comprehensive description of insurance fraud types.

In reality, not all the incoming claims are likely to be audited. A description of the audit process will explain why. All the incoming claims are initially checked by the adjusters. They consider the causes of car damages, the circumstances of the accident, and the policy characteristics. A claim that the adjuster does not find suspicious will always be exempted from audit and settled routinely. Audit means that the claim is transferred to a Special Investigation Unit, hereafter referred to as SIU. This unit provides a further assessment of the claim and decides whether to consider it as fraudulent or not.

Hence, a variable related to fraud suspicion is created by the adjuster before the audit decision. We denote it as S. For this binary variable, we have

$$S = 0 \Rightarrow A = 0. \tag{1}$$

The condition given in (1) simply means that the audit decision is nested inside the selection decision induced by S. Since both variables S and A are binary, this condition amounts to $A \leq S$. A claim for which $S = 1$ will be referred to later as suspected of fraud. Conversely, $S = 0$ when the initial screening reveals no fraud suspicion. Another reason (although less frequent) which leads to $S = 0$ is that no loss is expected. This could happen if the insurer of the third party pays for the car damages or if the deductible exceeds the claim cost.

As only the claims suspected of fraud are likely to be audited, selection bias now reflects the confusion between $P(F = 1 \mid A = 1,\ x)$ and $P(F = 1 \mid S = 1,\ x)$, under the condition given in (1). This bias increases with the proportion of suspicious claims that are not audited, and is eliminated if $A = S$ in the sample. Such a sample is created by a random auditing strategy, designed in the following way. First, a population of claims is selected at random, and the treatment (using the vocabulary of experience plans) is to audit all of the suspicious claims. This pair of actions (the random selection of claims, plus treatment) provides a controlled experiment which eliminates selection bias if the fraud equation is estimated from these claims.

The audit decision for suspicious claims is transferred to the adjusters if there is no random auditing strategy. An expert takes this decision on the basis of an implicit trade-off between the estimated audit cost and the gain from fraud detection. Most of the suspicious claims are not audited if the audit decision is left to the experts (see, for instance, the section "Presentation of the Database").

To sum up, Figure 1 describes the links between the three binary variables S, A, and F, depending on the audit strategy.

PRESENTATION OF THE DATABASE

The claims database belongs to an insurance company operating in Spain. The claims are linked to motor insurance contracts and were reported during 2000. The structure of the database is described in Figure 2. The number of claims and the weight with respect to the parent node are given under each node of the tree.

Let us comment on the numbers of claims, according to their type. The proportion of claims selected for random auditing is 20 percent. If random auditing is performed thoroughly, all the claims that are suspected of fraud are audited. Therefore, their weight in the sample is 18.6 percent. From the definition of S given in the section

Figure 1
The Binary Variables *S*, *A*, and *F* and Audit Strategies

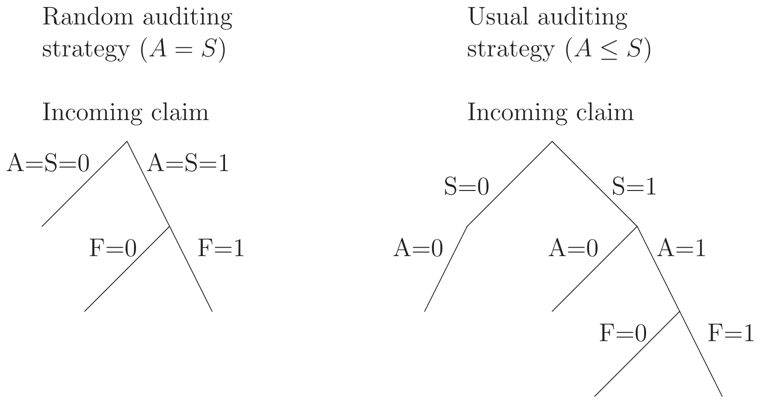
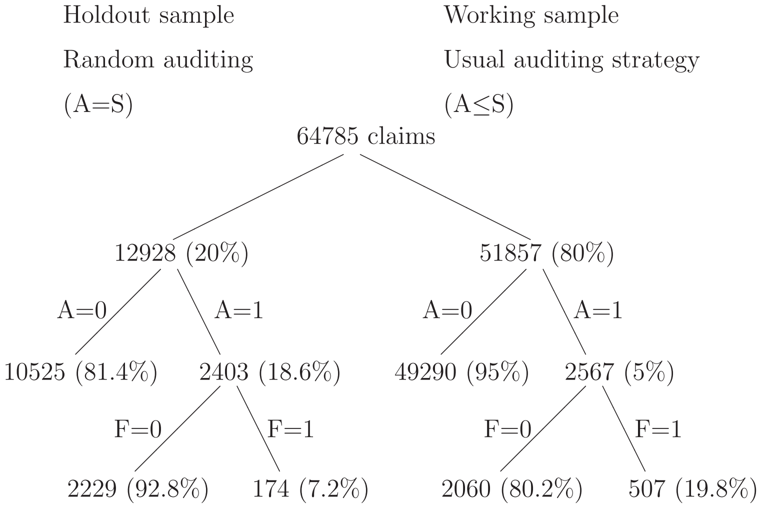

Random auditing strategy ($A = S$)
Usual auditing strategy ($A \leq S$)
Incoming claim
A=S=0
A=S=1
F=0
F=1
Incoming claim
S=0
S=1
A=0
A=0
A=1
F=0
F=1

Figure 2
The Claims Database



"Fraud Detection . . . Bias Issues," this means that the first screening reveals no suspicion of fraud for the other claims or that no loss is expected for the insurance company.

The other sample represents 80 percent of the claims. The audit decision on these claims is left to the adjusters. The audit rate in this sample is close to 5 percent, which is much less than for the first sample of randomly selected claims. The two samples have the same structure as selection in the first sample was performed at random. Therefore, most of the suspicious claims are exempted from audit if the adjusters are left free to take the decision. The selection bias issue arises from the discrepancy between the two audit rates.

As indicated in the Introduction, we will use the claims with no specific audit recommendation as the working sample, as they are subject to selection bias and are processed in a way that is currently used by insurance companies. The claims selected for random auditing give an unbiased fraud probability estimation. They will be used as a holdout sample, to assess the efficiency of the statistical model.

We will first estimate the bivariate probit model on the suspicious claims of the working sample (see the section "Empirical Applications of the Bivariate Probit Model"). Then fraud probabilities which are not conditional on an audit decision will be obtained. After that, we will compare these results with those obtained for the holdout sample. The ability of the bivariate model to replace a costly random auditing strategy will thus be assessed. A similar approach is followed in the section "Applications of Selection . . . Policies Design," where auditing policies are designed from the estimation of fraud probabilities.

The fraud rate on audited claims is 19.8 percent in the working sample. The application to the holdout sample of a fraud equation estimated on the working sample does not modify the average fraud probability, which remains close to 20 percent. On the other hand, the similar fraud rate for the holdout sample is equal to 7.2 percent. Here we have an example of the overestimation induced by selection bias, as mentioned in the Introduction. A single equation model on fraud is unable to even partially fill the gap between the two fraud rates. Determining to what extent a selection model applied to the working sample can fill this gap is the purpose of the statistical analysis which follows.

This difference between the two fraud rates can be explained by the claim screening performed by the experts. It can also be explained by a deterrence effect. The propensity to defraud should decrease with the audit rate, if we assume that policyholders are aware of the audit policy. This argument does not seem convincing in the present situation, as policyholders who were concerned by the random auditing strategy were not informed of their participation in the experiment. Economic models on fraud (see Picard, 2000, for a survey) usually assume that policyholders are aware of their company's auditing policy. Inclusion of deterrence effects in statistical models on fraud is discussed by Tennyson and Salsas (2002) and Dionne, Giuliano, and Picard (2002).

Let us conclude this section with comments on fraud indicators. This database was used in other publications (Artis, Ayuso, and Guillén, 2002; Caudill, Ayuso, and Guillén, 2005), in which possible fraud indicators are listed.[3] We also considered the seniority of the policyholder and the number of previous claims, as used in Belhadji, Dionne, and Tarkhani (2000). We did not have fraud indicators related with bodily injury damages and medical treatment types, like Weisberg and Derrig (1998), so we could not include that kind of regressors.

The Bivariate Probit Model: Theory and Applications to Selection Bias Assessment

The Theoretical Model

Using the notations from the section "Fraud Detection . . . Bias Issues," we define the bivariate probit model on a sample of claims suspected of fraud. Referring to Figure 1,

[3] See for instance Artis, Ayuso, and Guillén (2002, p. 328).

the bivariate model described below is applied to the right part of the tree associated with a usual auditing strategy.

A bivariate model on the audit and fraud equations with a joint distribution of the two random components will be able to assess a fraud probability conditioned by the individual characteristics of the claims and by the audit variable A. Once this estimation has been made, we have a fraud probability for suspicious claims which is unconditional on an audit decision and which can be used in an optimal audit policy.

The bivariate probit model with censoring includes, first of all, an audit equation defined on all the claims that are suspected of fraud. This equation is defined in the following way

$$A_i = 1_{[A_i^* \geq 0]}; \quad A_i^* = (x_A)_i\, \beta_A + (\varepsilon_A)_i; \quad (\varepsilon_A)_i \sim N(0,1). \tag{2}$$

The binary variable A is the sign indicator of a latent variable A^*. The variance of the random variable $(\varepsilon_A)_i$ can be set equal to one without loss of generality, because of the invariance of A with respect to the multiplication of A^* by a positive constant. The regression components in the linear equation can be defined on the policy to which the claim is related, or can be claim-specific. They are represented by a line vector. The parameters are stacked in a column vector which enables a cross product.

The fraud equation is defined on audited claims only. We then write

$$\text{If } A_i = 1 : F_i = 1_{[F_i^* \geq 0]}; \; F_i^* = (x_F)_i\, \beta_F + (\varepsilon_F)_i. \tag{3}$$

The random variable $(\varepsilon_F)_i$ also follows a standard normal distribution.

If we retain a bivariate normal distribution for $((\varepsilon_A)_i, (\varepsilon_F)_i)$, i.e.,

$$\begin{pmatrix} (\varepsilon_A)_i \\ (\varepsilon_F)_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \quad \rho \in [-1,1], \tag{4}$$

we obtain a bivariate probit model. The sign of the correlation coefficient ρ is of paramount importance to the estimation of fraud risk conditional on the audit variable.

In this censored setting, three levels are possible for the dependent variables. Let us compute the corresponding probabilities. We have suppressed the individual index for the sake of simplicity.

1. If the claim is not audited, the fraud variable is not observed. We then have

$$P[A = 0] = P[\varepsilon_A < -x_A\beta_A] = \Phi(-x_A\beta_A) = 1 - \Phi(x_A\beta_A), \tag{5}$$

where Φ is the distribution function of a standard normal variable and

$$\Phi(x_A\beta_A) = p_A = P[A = 1] = E(A).$$

The last equality in (5) results from the symmetry in the distribution of ε_A. The bivariate model is estimated on the suspicious claims only, hence the probability $P[A=0]$ would be described as $P[A=0 \mid S=1]$ if expressed on all the incoming claims as in the section "Fraud Detection . . . Bias Issues."

2. If the claim is audited and if fraud is established, we can write

$$P[A=1,\ F=1] = P[\varepsilon_A \geq -x_A\beta_A,\ \varepsilon_F \geq -x_F\beta_F] = P[\varepsilon_A \leq x_A\beta_A,\ \varepsilon_F \leq x_F\beta_F]$$

$$= P[\Phi(\varepsilon_A) \leq \Phi(x_A\beta_A),\ \Phi(\varepsilon_F) \leq \Phi(x_F\beta_F)].$$

Again, we used symmetry in the distribution of the random components. The variables $\Phi(\varepsilon_A)$ and $\Phi(\varepsilon_F)$ follow a uniform distribution on [0, 1] and the distribution function of $(\Phi(\varepsilon_A), \Phi(\varepsilon_F))$ is a Gaussian copula indexed by ρ. Writing fraud probability as

$$p_F = P[F=1] = P[\Phi(\varepsilon_F) \leq \Phi(x_F\beta_F)] = \Phi(x_F\beta_F),$$

we denote the Gaussian copula in the following way

$$C(\rho, p_A, p_F) = P[\Phi(\varepsilon_A) \leq p_A,\ \Phi(\varepsilon_F) \leq p_F] = P[\varepsilon_A \leq \Phi^{-1}(p_A),\ \varepsilon_F \leq \Phi^{-1}(p_F)].$$

This copula relates a bivariate Gaussian distribution function to its marginal components. Hence the probability of interest is equal to

$$P[A=1,\ F=1] = C(\rho, p_A, p_F). \tag{6}$$

3. The probability of the last level (the claim is audited but not fraudulent) is the complementary value

$$P[A=1,\ F=0] = P[A=1] - P[A=1,\ F=1] = p_A - C(\rho, p_A, p_F).$$

Maximizing the log-likelihood of the sample requires a computation of the Gaussian copula and of its partial derivatives with respect to the parameters. This function does not have a closed form with respect to Φ and Φ^{-1}, unless ρ is equal to the critical values -1, 0, and 1. It can be approximated by Gaussian quadratures (see Dionne, Gagné, and Vanasse, 1998, for an application to panel data with endogenous attrition).

We now present applications of the model to fraud probability prediction.

Under the null hypothesis $\rho = 0$, fraud probability does not depend on the audit variable and we have

$$p_F = P[F=1] = P[F=1 \mid A=1]. \tag{7}$$

Under the alternative assumption $\rho \neq 0$, selection bias arises because in that case

$$P[F=1 \mid A=1] = \frac{C(\rho, p_A, p_F)}{p_A} \neq p_F. \tag{8}$$

These probabilities are confused in the application of a fraud risk model which does not take into account selection bias. Indeed, the conditional probability $P[F = 1 \mid A = 1]$ is estimated on the audited claims whereas the unconditional probability p_F should be applied to the incoming suspicious claims in order to design an audit policy. To gain a better understanding of the influence of selection bias, some basic properties of bivariate Gaussian copulas are recalled and applied to our context.

Three critical values for ρ will be mentioned. First, we have

$$C(0, p_A, p_F) = p_A \times p_F$$

since the random variables ε_A and ε_F are independent if $\rho = 0$. In this case, all the fraud probabilities given in Equations (7) and (8) are equal. Second, ε_A and ε_F are equal a.e. if $\rho = 1$, which implies $C(1, p_A, p_F) = \min(p_A, p_F)$. This value is the upper bound for a bivariate copula, and is known as the upper Fréchet bound. Lastly, we have $\varepsilon_A = -\varepsilon_F$ a.e. if $\rho = -1$, hence $C(-1, p_A, p_F) = \max(p_A + p_F - 1, 0)$. Again, we reach the lower bound for a bivariate copula.

The map $\rho \longrightarrow C(\rho, p_A, p_F)$ is increasing on the interval $[-1, 1]$ for any value of (p_A, p_F).[4] Hence the same result holds for $P[F = 1 \mid A = 1]$ in the bivariate probit model, whereas $P[F = 1 \mid A = 0]$ is a decreasing function of ρ.

Empirical Applications of the Bivariate Probit Model

The selection model should only be applied to the claims which are suspected of fraud, which requires a variable pointing out such claims. Unfortunately, the nature of the claims with respect to fraud suspicion is not available in our database for the claims with no specific recommendation for audit. Among the 49,290 claims which were not audited in the working sample (see Figure 2), some are suspected of fraud ($S = 1$) and exempted from audit due to the adjusters' decision, and some are not suspicious. However, the value of S is not available in the database.

In order to apply the selection model described above, we will create a set of suspicious claims, partly from simulation. This set must contain the 2,567 claims finally transferred to the SIU, as $A = 1$ implies $S = 1$. On the other hand, the proportion of suspicious claims should be the same in the working and the holdout samples since they have the same structure. We estimated an audit equation on the holdout sample, which provided a suspicion probability conditioned on regression components since all the suspicious claims are supposed to be audited in this population. Then we selected claims from those that were not audited in the working sample, using a random sampling scheme and this suspicion equation.[5] Together with the claims transferred to the SIU, these form a set of suspicious claims in the working sample which has the

[4] The distribution function of $(\varepsilon_A, \varepsilon_F)$ and hence the copula are integrals of the bivariate Gaussian density on negative orthants. Hence the increasing link with ρ is not surprising. We did not find a proof of this result in the statistical literature, but it is verified from numerical computations.

[5] Using suspicion probabilities unchanged from the holdout sample would have led to an excess of suspicious claims (22 percent instead of 18.6 percent). The suspicion frequency should be the same in the working and holdout samples, as these samples have the same structure. Hence, we corrected the intercept in order to obtain the 18.6 percent frequency.

same weight as those in the hold out sample (i.e., 18.6 percent). Therefore, we have 18.6 percent × 51,857 $\simeq$ 9,640 suspicious claims in the working sample, 2,567 of which are audited. The other claims were drawn at random from the suspicion equation. This approach is only makeshift but we in any case find it interesting to estimate the selection model. Random auditing is not often carried out in the real world (which creates a selection bias problem), and a suspicion variable should be easy to determine during the first step of claims screening. Since we maintained the proportion of suspicious claims observed with random auditing, the audit rate for suspicious claims in the working sample is $0.0495/0.186 = 0.266$.

Let us give an example to illustrate the importance of selection bias in relation with the bivariate probit model. We will compare the unconditional fraud probability of an incoming suspicious claim with average characteristics, and the fraud probability if such a claim is audited with the audit policy used in the working sample. We expect different results from Equation (8). We will compute $p_F = P[F = 1]$ as a function of ρ under the following constraints

$$p_A = P[A = 1] = 0.266; \quad P[F = 1 \mid A = 1] = 0.198. \tag{9}$$

We just derived the proportion of audited claims among suspicious claims, hence the value given to p_A. The frequency of fraudulent claims among those audited is 0.198, which explains the constraint on $P[F = 1 \mid A = 1]$. The unconditional fraud probability which we denote as $p_F(\rho)$ is a solution of the equation

$$\begin{aligned} C(\rho, p_A, p_F(\rho)) &= P[A = 1,\ F = 1] = P[F = 1 \mid A = 1] \times P[A = 1] \\ &\Leftrightarrow C(\rho,\ 0.266,\ p_F(\rho)) = 0.198 \times 0.266 \simeq 0.0527. \end{aligned} \tag{10}$$

Since the claims adjusters are supposed to contribute expertise in their audit decision, we expect that

$$p_F(\rho) = P[F = 1] < P[F = 1 \mid A = 1] = 0.198.$$

Indeed, the unconditional probability of fraud $p_F(\rho)$ is a weighted average of the two conditional probabilities $P[F = 1 \mid A = 1]$ and $P[F = 1 \mid A = 0]$. The latter probability should be lower than the first one because of the information brought by the audit decision. In this case, ρ should be positive since we have

$$\rho > 0 \Longleftrightarrow p_F(\rho) < 0.198$$

from the equivalence

$$\rho > 0 \Longleftrightarrow 0.198 \times 0.186 = C(\rho,\ 0.186,\ p_F(\rho)) > C(0,\ 0.186,\ p_F(\rho)) = 0.186 \times p_F(\rho).$$

We used the increasing property of the copula with respect to ρ.

Let us give values for $p_F(\rho)$ under the constraints given in (9) and (10). We let ρ increase from 0 to 1 with an increment equal to 0.1. Results are presented in Table 1.

Table 1
Unconditional Fraud Probability for an Incoming Suspicious Claim With Average Characteristics, as a Function of the Correlation Coefficient

ρ	$p_F(\rho)$
0	0.198
0.1	0.166
0.2	0.140
0.3	0.117
0.4	0.099
0.5	0.084
0.6	0.072
0.7	0.063
0.8	0.056
0.9	0.053
1	0.053

Suppose that $\rho = 0.5$. The fraud probability of an incoming suspicious claim with average characteristics is 2.4 times less than if this claim has been audited. This drastically modifies the threshold of a score designed to select claims for audit.

From Table 1, the unconditional fraud probability of an incoming suspicious claim with average characteristics is equal to 7.2 percent (i.e., the fraud rate observed without selection bias in our database) for $\rho = 0.6$. Hence, a positive estimated correlation coefficient is expected on the working sample.

The estimated correlation coefficient strongly depends on the choice of regression components in each equation. On the whole, we noticed that $\hat{\rho}$ decreased with the amount of information used in the regressions. Variations of $\hat{\rho}$ due to the sampling scheme of suspicious claims in the working sample are much less important than those due to the choice of regression components in the bivariate model.

Let us recall the steps of our estimation procedure.

- First, we must create a suspicion variable which is missing for the nonaudited claims in the working sample. To reach this goal, we estimate suspicion probabilities on the holdout sample as it has no missing information. Then we use these probabilities to draw claims at random from the nonaudited claims of the holdout sample. We modified these probabilities to obtain the same frequency (18.6 percent) of suspicious claims in both samples.
- Then the bivariate probit model is estimated on the suspicious claims set of the working sample. This set contains the 2,567 audited claims which are definitely suspicious, and other claims drawn at random in the first step. The set depends then on the sampling scheme, and contains $0.186 \times 51{,}857 \simeq 9{,}640$ claims. As explained in this section, we obtain fraud probabilities which do not depend on the audit decision.

Let us detail for instance estimation results with a medium number of regression components. All of these results are significant at a 1-percent level. With this constraint,

TABLE 2

Regression Results in the Bivariate Probit Model (Working Sample) and in the Equation on Fraud Suspicion (Hold Out Sample)

	Sampling Scheme	Bivariate Probit Model	
Dependent Variables	Fraud Suspicion (S)	Audit (A)	Fraud (F)
Sample	Holdout Sample	Working Sample, $S = 1$	Working Sample, $A = 1$
Size of the sample	12,928	≃9,640	2,567
Parameters	$\hat{\beta}_S$, $(S = A)$	$\hat{\beta}_A$	$\hat{\beta}_F$
Intercept	−1.53	−1.08	−1.90
Seniority of policyholder: less than one year			0.21
Motorbike			0.41
Automobile, private use			0.39
Number of previous claims			0.06
Coverage: third party liability only	0.71	0.74	
Coverage: third party liability + theft, arson and glasses	0.66	0.53	
Third party at fault	0.33		
Use of the no-fault system	0.29		
Age of the policyholder	−0.003	−0.005	

the set of covariates retained in each equation cannot be increased with the variables at our disposal. Maximum likelihood estimations for the equation that generates the sampling scheme and for the bivariate probit model are given in Table 2.

The results depend on the regression components and on random sampling for the bivariate probit model. The estimators are stable with respect to suspicious claims sampling. With this set of regression components, we have

$$\hat{\rho} \simeq 0.51.$$

The estimated unconditional fraud probability for incoming suspicious claims is 8.4 percent on average. The selection model provides a satisfactory result with this set of regression components, as the goal is to reach a 7.2 percent ratio from a starting point—the fraud rate for audited claims in the working sample—which is 19.8 percent.

As mentioned above, the estimated correlation coefficient increases if fewer regression components are retained. The estimated coefficient ranges between 0.36 and 0.64, depending on the set of regression components. The average unconditional fraud probabilities range between 6.9 percent and 10.8 percent.

The positive result obtained in this section is that the selection model is able to get very close to the actual fraud rate, which can only be reached through random auditing. The weakness of selection models applied to censored data is that estimation results are highly dependent on the regression components set.

Applications of Selection Bias Models to Auditing Policies Design

In this section, we will focus on the design of audit decisions based on a short-term analysis. We balance estimated audit costs and gains from fraud detection. The gain is the product of the claim settlement's reserved cost and of the fraud probability.[6] The reserved cost of the claim settlement is determined by the adjuster during the first screening of the claim. Hence it is known by the insurance company before the audit decision, which is not the case with the audit cost. The audit cost recorded in the database corresponds to all the steps of the claim examination, including the first screening.

Let us denote the audit cost related to the SIU examination for claim i as ac_i, and c_i as the reserved claim cost. A transfer of this claim to the SIU generates an expected gain in the short run if

$$\hat{E}[c_i F_i - AC_i] > 0 \Leftrightarrow \hat{E}[F_i] = \hat{P}[F_i = 1] > \frac{\hat{E}[AC_i]}{c_i}. \tag{11}$$

Indeed, $c_i \hat{E}[F_i]$ is the expected gain from audit if claim i is transferred to the SIU. The expected fraud probability is derived from a bivariate probit model in what follows. The audit cost is only known *ex post* (hence its random variable status in this step of the computation).

Information on audit costs is needed to derive the expected values $\hat{E}[AC_i]$. The average audit costs are given in Table 3, according to the status of the claim and the sample.

These results clearly indicate that the more likely a claim is to be fraudulent, the more thoroughly it must be examined by the adjusters. This increases the audit cost. The averages are similar in both samples. The difference between the average audit costs for claims that are transferred to the SIU and those that are not fraudulent can be explained by the stricter selection for SIU in the working sample, which makes these claims more suspicious. Hence selection bias is also seen in audit costs.

The audit costs in Table 3 correspond to all the steps of claims examination, including the mandatory first screening which is not performed by the SIU. The average audit cost charged to the SIU will be set equal to

$$\overline{ac}^{NF} = 67.82 - 37.73 = €30.09; \ \overline{ac}^{F} = 222.84 - 37.73 = 185.11, \tag{12}$$

where $\overline{ac}^{NF}$ and $\overline{ac}^{F}$ relate to nonfraudulent and fraudulent claims. This computation implicitly supposes that the cost of the first screening does not depend on the eventual status of the claim.

The expected audit cost clearly depends on fraud probability since $\overline{ac}^{F} > \overline{ac}^{NF}$. If we condition this expectation on average values for each level of the audited claims, we obtain

[6] In the economic literature on fraud (see Picard, 2000, for a survey) the policyholder holds a private information on the amount of the loss. A fraudulent claim consists of misreporting this amount. The insurer can commit to a compensation scheme, designed in such a way that fair reporting is an optimal strategy for the policyholder. See Boyer (2001) for fraud prevention policies which involve the public authorities through taxes.

TABLE 3
Average Audit Costs (Including the First Screening)

Average Audit Costs of Claims	Holdout Sample $(A = S)$	Working Sample $(A \neq S)$
$A = 0$	€36.97	€37.73
$A = 1, F = 0$	€59.81	€67.82
$A = 1, F = 1$	€231.71	€222.84

$$\begin{aligned}\hat{E}[AC_i] &= \left(\hat{P}[F_i = 1] \times \overline{ac}^F\right) + \left(\hat{P}[F_i = 0] \times \overline{ac}^{NF}\right)\\ &= \overline{ac}^{NF} + \left(\hat{P}[F_i = 1] \times (\overline{ac}^F - \overline{ac}^{NF})\right).\end{aligned}$$

An audit policy can be designed from this assessment of the audit cost. The rule given in (11) suggests that claim i should be transferred to the SIU if

$$c_i > \overline{ac}^F - \overline{ac}^{NF} \ \& \ \hat{P}[F_i = 1] > \frac{\overline{ac}^{NF}}{c_i - (\overline{ac}^F - \overline{ac}^{NF})}. \tag{13}$$

Since the probability threshold must be less than one, the first condition amounts to $c_i > \overline{ac}^F$. We will use the holdout sample to compare the audit policies derived from different fraud models. All suspicious claims are audited if random auditing is performed thoroughly. In this case there is no selection bias and the target audit rate is obtained from a fraud equation estimated on the audited claims in the holdout sample. From the fraud equation specified in the bivariate model of Table 2, with the selection rule and the audit costs given in (13) and (12), we obtain a target audit rate for suspicious claims of 36.7 percent.

Let us assess the influence of selection bias on audit policies. Suppose that the fraud equation specified in Table 2 is estimated on the audited claims of the working sample. The optimal audit rate on suspicious claims of the working sample goes up to 62 percent, which reflects the overestimation of fraud probability if selection bias is neglected.

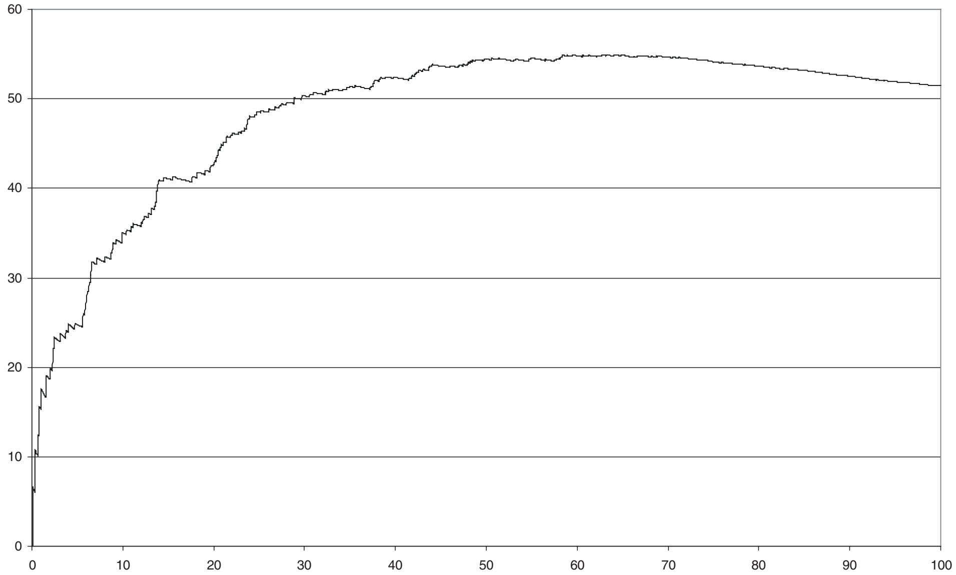
Let us see how the target audit rate can be approximated by a selection model. The proportion of suspicious claims which should be audited in the working sample is 39 percent if we use the selection model estimated in Table 2. This audit rate is very close to the target rate obtained from random auditing (i.e., 36.7 percent), which is a satisfactory result.

If the set of regression components varies in the regression, the optimal audit rate ranges between 33 and 45 percent in our different trials. These rates strongly depend on the set of regression components through the estimated correlation coefficient. This result is obviously negative, as sound risk assessment derived from random auditing cannot be anticipated precisely from the selection model. However, the bivariate probit model points out selection bias and can justify a random auditing strategy which is not often employed by insurance companies.

In what follows, the efficiency of an audit policy is summarized by the average gain per suspicious claim. As for the selection rule given in (13), we compare each audit

FIGURE 3

Average Gain (Euros) per suspicious Claim as a Function of the Percentage of Audited Claims



policy to a zero audit strategy. Hence gains are the reserved costs of the audited claims that are proved to be fraudulent, and losses are the audit costs charged to the SIU. The efficiency of the previously detailed audit policies is almost the same for all of them. This may seem surprising, as the audit rates exhibit great variations. For all these policies, the average gain per suspicious claim is close to €52. This stability with respect to audit rates is explained by the shape of the efficiency curves. Claims are sorted in decreasing order with respect to the expected gain $\hat{E}[c_i F_i - AC_i]$, and the curve is the average gain per suspicious claim expressed as a function of the audit rate. Figure 3 is related to the bivariate model estimated in Table 2.

All the efficiency curves have the same shape. This one reaches its maximum for an audit rate close to 60 percent. The stability of the curve for audit rates greater than 30 percent explains why the different policies have similar efficiencies.

CONCLUSIONS

This article presented a statistical approach which counteracts selection bias without using a random auditing strategy. A two equation model for audit and fraud (a bivariate probit model with censoring) was estimated on a sample of suspicious claims for which the experts were left to take the audit decision. The expected overestimation of fraud risk derived from a single equation model on audited claims was corrected. Results were rather close to those obtained with a random auditing strategy, at the expense of some instability with respect to the regression components set. Then we compared auditing policies derived from a statistical correction of selection bias, and policies obtained from random auditing.

Let us first comment briefly on the instability of selection bias assessment with respect to the set of regression components. This instability is actually consubstantial with the censored character of the data. We presented censored and noncensored selection models in the Introduction. Such models are made up of a selection variable (e.g., audit) and a variable of interest (e.g., fraud). They are censored if the variable of interest is only observed for the selected individuals. In a noncensored model, the influence of the selection variable on the variable of interest is easily derived from a comparison of two samples (selected versus nonselected individuals) with respect to the variable of interest. In a censored context, this influence is only assessed through the variation of estimated selection probabilities on the selected individuals. This estimated probability plays the role of a supplementary covariate in the regression model that is related to the variable of interest. Now this probability is derived from the information already used in the regression. Selection bias reflects a more intricate specification for the distribution related to the variable of interest instead of being based on observed differences in the noncensored setting.[7] If the selection model is unstable for censored data, it is however of greatest interest precisely in this context.

REFERENCES

Artis, M., M. Ayuso, and M. Guillén, 2002, Detection of Automobile Insurance Fraud with Discrete Choice Models and Missclassified Claims, *The Journal of Risk and Insurance*, 69(3): 325-340.

Ayuso, M., M. Guillén, S. Viaene, and D. Van Ghee, 2004, Cost-Sensitive Design of Claim Fraud Screens, *Lecture Notes in Artificial Intelligence*, 3275: 78-87.

Belhadji, E. B., G. Dionne, and F. Tarkhani, 2000, A Model for the Detection of Fraud, *Geneva Papers on Risk and Insurance—Issues and Practice*, 25(5): 517-538.

Boyer, M., 2001, Mitigating Insurance Fraud: Lump-Sum Awards, Premium Subsidies, and Indemnity Taxes, *The Journal of Risk and Insurance*, 68(3): 403-435.

Caudill, S., M. Ayuso, and M. Guillén, 2005, Fraud Detection Using a Multinomial Logit Model with Missing Information, *The Journal of Risk and Insurance*, 72(4): 539-550.

Derrig, R. A., 2002, Insurance Fraud, *The Journal of Risk and Insurance*, 69(3): 271-287.

Dionne, G., R. Gagné, and C. Vanasse, 1998, Inferring Technological Parameters from Incomplete Panel Data, *Journal of Econometrics*, 87: 303-327.

Dionne, G., F. Giuliano, and P. Picard, 2002, Optimal Auditing for Insurance Fraud, Working paper, Available at http://www.hec.ca/gestiondesrisques/cahiers.htm.

Heckman, J. J., 1979, Sample Selection Bias as a Specification Error, *Econometrica*, 47(1): 153-162.

Picard, P., 2000, Economic Analysis of Insurance Fraud, in: Georges Dionne, ed., *Handbook of Insurance* (Kluwer Academic Publishers), 315-362.

Tennyson, S., and P. Salsas-Forn, 2002, Claims Auditing in Automobile Insurance: Fraud Detection and Deterrence Objectives, *The Journal of Risk and Insurance*, 69(3): 289-308.

Weisberg, H. I., and R. A. Derrig, 1998, Detection de la Fraude: Methodes Quantitatives, *Risques*, 35: 75-99 (in English translation).

[7] The seminal paper on selection models with censoring is entitled "Sample Selection Bias as a Specification Error" (Heckman, 1979). Estimated selection probability is included in a linear regression on the variable of interest via the inverse Mills' ratio, which creates the "Heckit."