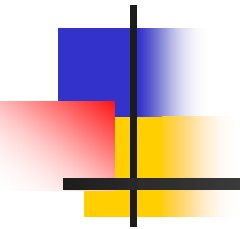


# Selection bias and auditing policies on insurance claims



Zürich,

September 5, 2005

Jean Pinquet,

Montserrat Guillén & Mercedes Ayuso

# Selection bias



---

- Occurs when the range of derivation of a statistical model is smaller than the range of application of this model.
- This happens if a selection mechanism is forced on the working sample (on which the model is estimated), but not on the application range.
- Example for auditing policies: fraud risk is estimated on audited claims, but the statistical model is applied on incoming claims which are assessed before any audit decision.



# Expected impact of selection bias on the statistical analysis of fraud

---

- Given observable characteristics, fraud risk is expected to be less important for a claim exempted from audit by the expert (claim adjuster) 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.

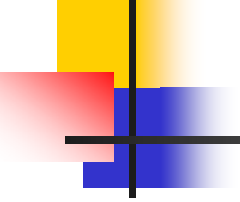


# Random auditing: a controlled experiment to counteract selection bias

---

- Random auditing strategy: select claims at random, then audit all of them.
- A fraud risk model estimated from these claims is not subject to selection bias.
- Insurance companies probably fear the cost of such a strategy. Random auditing is not often carried out in the real world.

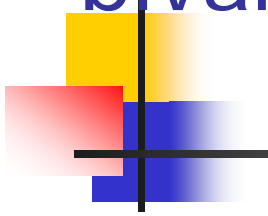
# Counteracting selection bias with a bivariate probit model on fraud and audit equations (I)



---

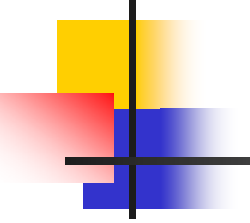
- The model is estimated on claims subject to a usual audit strategy.
- We consider a bivariate model (audit and fraud equations)
- The audit and fraud variables are sign indicators of latent variables. The key parameter is the correlation coefficient between these two latent variables.

# Counteracting selection bias with a bivariate probit model on fraud and audit equations (II)



---

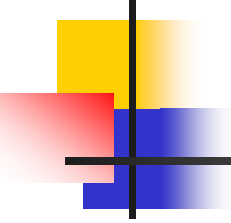
- The estimated correlation coefficient is expected to be positive since it reflects the ability of claims adjusters to assess hidden characteristics in fraud distributions through the audit decision.
- The estimation is straightforward (uses bivariate Gaussian copulas).
- The model allows to derive fraud probabilities depending on claim characteristics, but also on the audit variable.



## Issue addressed by the paper: can a selection model replace a random auditing policy?

---

- Our claims data base is split into two populations. Claims selected at random (one out of five) are recommended for audit, whereas there is no specific recommendation for the other ones.
- The paper uses the latter claims as a working sample, and those selected at random as a hold out sample.
- The goal is to verify with the claims selected at random that selection bias is corrected properly by the bivariate model.



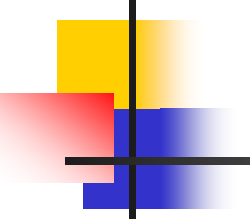
# Three levels for incoming claims: suspicious, audited and fraudulent

---

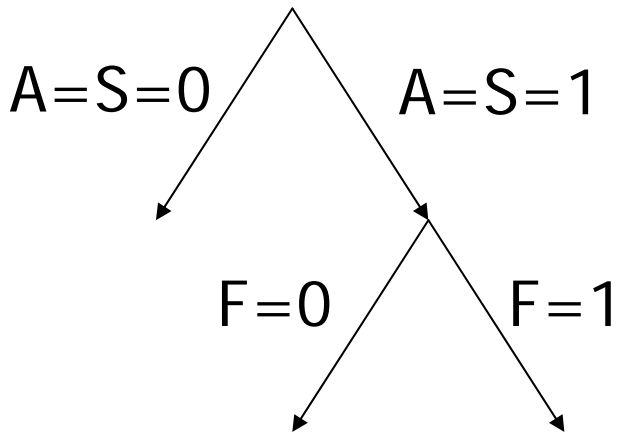
- All the incoming claims are actually not likely to be audited. Why?
- They are all in the first place checked by the adjusters.
- A claim the adjuster does not find suspicious will be exempted from audit and settled routinely whatever the audit strategy. We denote  $S$ ,  $A$  and  $F$  the variables related to fraud suspicion, audit and fraud.



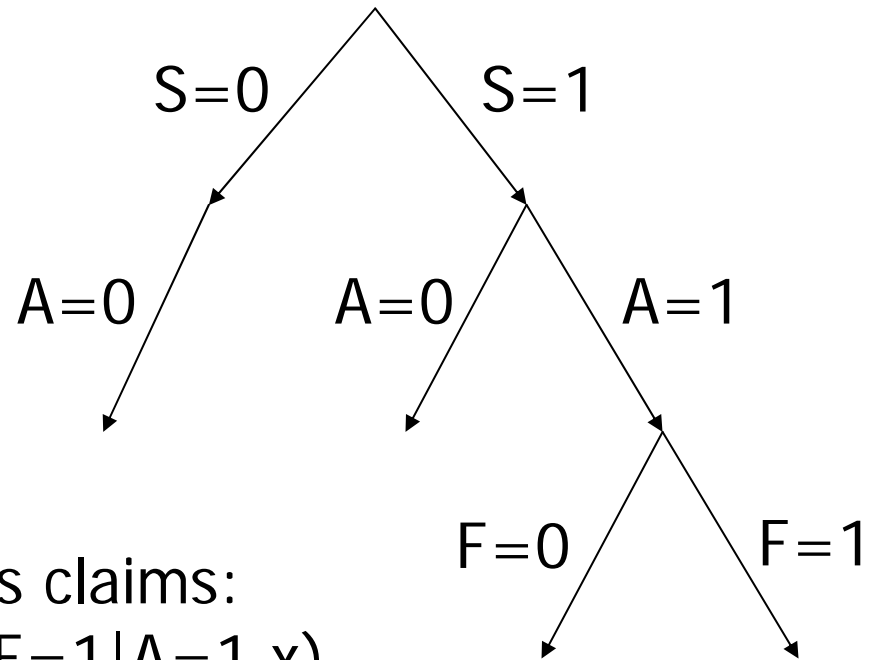
# The binary variables S, A and F and audit strategies



Random auditing  
Strategy ( $A=S$ )



Usual auditing strategy  
( $S=0$  implies  $A=0$ )

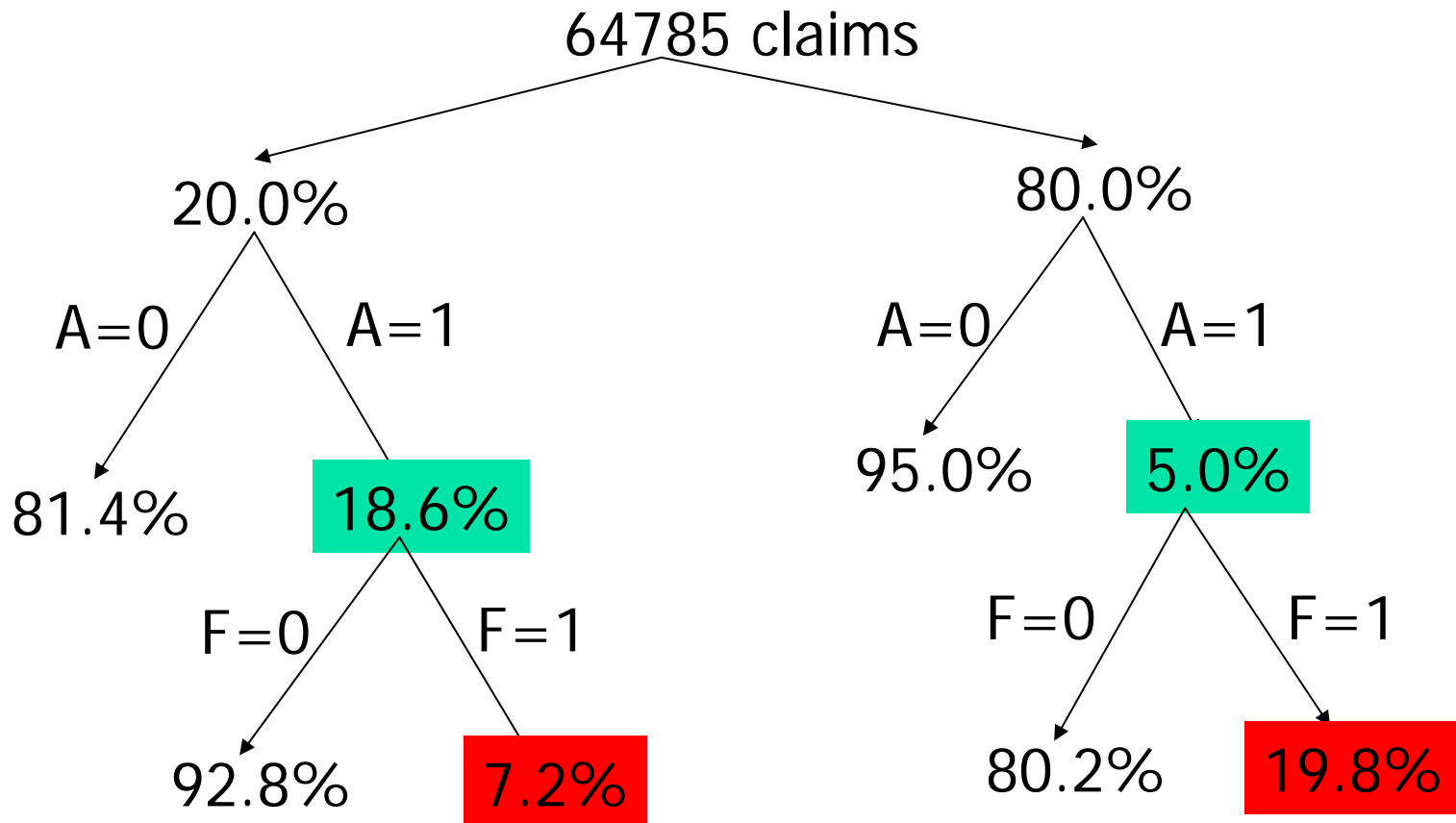


Selection bias on suspicious claims:  
 $E(F=1|x)$  confused with  $E(F=1|A=1,x)$   
( $x$  observable claim characteristics)

# The claims data base (percentages w.r.t. parent node)

Hold out sample  
(random auditing)

Working sample (usual  
auditing strategy)





# Selection bias: causes and consequences

---

- Cause of selection bias: the audit rate on incoming claims is 5% if claims adjusters are left free to audit or not the suspicious claims, and close to 20% if they are not.
- Consequence: the fraud rate on audited claims is 19.8% in the first case, and 7.2% without selection bias. This difference reflects the ability of claims adjusters.
- Question: is the bivariate probit model able to predict results of random auditing (which tell the truth) from the usual auditing strategy?



# Estimation and applications (I)

---

- The estimated correlation coefficient depends on the set of regression components. It decreases if more covariates are retained. It varies between 0.36 and 0.64 in our different trials.
- 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 fraud probability if such a claim is audited.

# Estimation and applications (II)

- For an average suspicious claim in the working sample:
- $P[A=1]=0.266$ ;  $P[F=1 | A=1]=0.198$ .
- We compute  $P[F=1]$  as a function of the correlation coefficient  $\rho$ . The true value for  $P[F=1]$  (from random auditing) is close to 0.072.
- $\rho=0.36$  (large set of covariates):  $P[F=1]=0.106$ .
- $\rho=0.51$  (medium set) :  $P[F=1]=0.082$ .
- $\rho=0.64$  (minimum set):  $P[F=1]=0.068$ .



# Comments

---

- Positive result: The preceding estimation encompass the actual fraud rate, which can only be reached through random auditing.
- A weakness of censored selection models is that estimation results (and the importance of selection bias) highly depend on the regression components set.
- Constraint on the model: you need to know which incoming claims are suspicious w.r.t. fraud.

# Applications to auditing policy design



---

- If selection bias is neglected, fraud probabilities on incoming suspicious claims are overestimated.
- An audit policy derived from such a statistical model will propose audit too often (since the expected gain from audit increases with fraud probability).
- Results are given in the paper.



# Conclusions

---

- The bivariate probit model applied on our data is able to counteract selection bias. However
  1. Results depend on the set of regression components.
  2. You need to know which claims are suspicious.
  3. Understanding the sign of the estimated correlation coefficient from the data is not easy.

Paper available upon request (as a reply to any mail sent to [pinquet@u-paris10.fr](mailto:pinquet@u-paris10.fr) or [jpinquet@free.fr](mailto:jpinquet@free.fr) )