Régression avec double censure emboîtée et dépendance : Une application au cas de la fraude aux cotisations sociales

Anne ALIGON[a] Denisa BANULESCU-RADU[b]
Sylvain BENOIT[c] Raphaël BROSSEAU[a]
Christophe HURLIN[b] Cédric VALLÉE[a]

[a] Caisse centrale MSA, Direction des Statistiques, des Etudes et des Fonds University
[b] Université d'Orléans, Laboratoire d'Economie d'Orléans
[c] Université Paris Dauphine - PSL

JMS, 30 mars 2022

Plan

Enjeu de politiques publiques

Fraude aux cotisations sociales, évasion sociale, économie non observée, manque à gagner... ?

- Difficulté à évaluer un phénomène par nature occulte
- Manque d'évaluation fiabilisée de l'ampleur du phénomène
- Une pluralité de méthodes d'estimation :
  1. indirectes
  2. directes
- Attentes des pouvoirs publics :
  1. cohésion sociale
  2. impact sur les ressources de la sécurité sociale

Définition

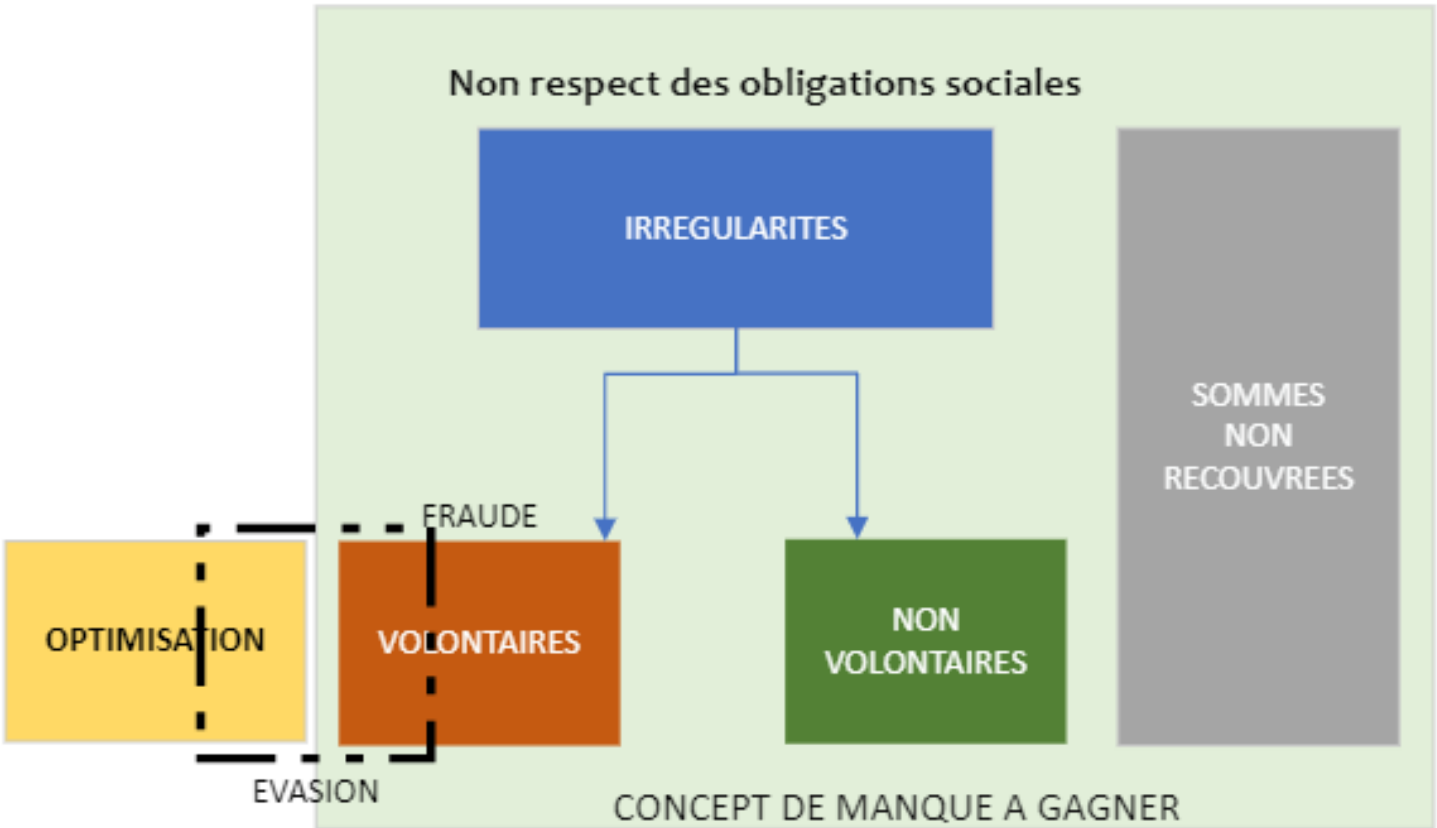


FIGURE – Définition du manque à gagner [1]

1. « La fraude aux prélèvements obligatoires », Cour des comptes, nov. 2019

Challenge économétrique

Une estimation fondée sur une **approche dite « directe » ou micro-économétrique** ($\neq$ des méthodes dites de post-stratification ou des tirages aléatoires)

- Précision de l'estimation
- Disponibilité des résultats des contrôles MSA
- Outil d'aide au ciblage
- Ressources limitées :
  1. MSA = Guichet unique (cotisations+maladie+ATMP+famille+retraite)
  2. Coûts induits par les contrôles aléatoires (perte d'efficacité)
- Mais travail de recherche académique à mener !

Évaluation empirique

Deux exercices pour communication au HCFIPS [2]

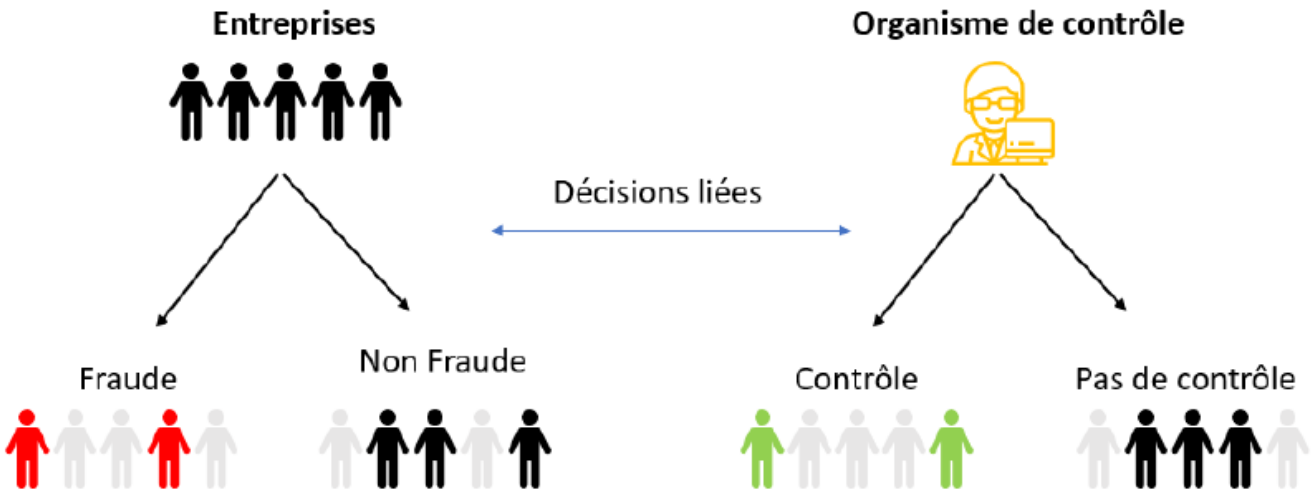
Estimation en milliard d'euros	**2014**	**2015**	**2016**
Travaux 2019	0,17	n.d	n.d
Actualisation 2021	0,18	0,17	0,17
Manque à gagner CCA / recettes	1,5%	1,5%	1,4%

- Périmètre du contrôle comptable d'assiette uniquement
- Résultats en ligne avec ceux diffusés par les autres régimes

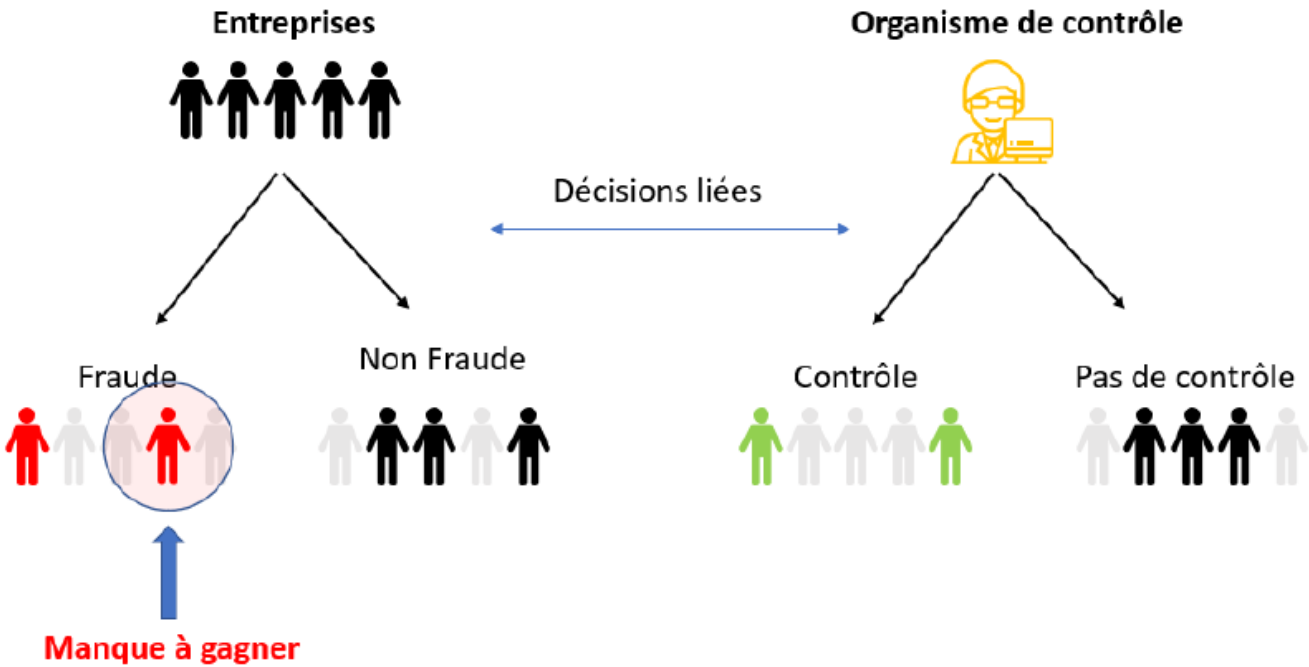
2. Haut Conseil pour le Financement de la Protection Sociale.

Plan

Qu'est ce que le manque à gagner ?



Qu'est ce que le manque à gagner ?



Comment le définir ?

Définition statistique du manque à gagner

La variable MAG_i est une **variable aléatoire**, tout comme la variable agrégée MAG.

Notre objectif est de caractériser les deux premiers **moments théoriques** de sa distribution, i.e. l'espérance $\mathbb{E}(MAG_i)$ et la variance $\mathbb{V}(MAG_i)$.

Comment l'estimer ?

Puisque le MAG_i n'est pas observable, il faut postuler une **distribution paramétrique** :

1. Nous postulons un **processus générateur de données (DGP ou modèle)** sur la **distribution conditionnelle** de la variable MAG_i.
2. Nous déduisons de ce modèle les **formules théoriques** pour les **deux premiers moments**

$$\mathbb{E}\left(MAG_i | \mathbf{X}_i = \mathbf{x}_i\right) = f\left(\mathbf{x}_i; \beta\right) \qquad \mathbb{V}\left(MAG_i | \mathbf{X}_i = \mathbf{x}_i\right) = g\left(\mathbf{x}_i; \beta\right)$$

   où $\mathbf{X}_i$ désigne un ensemble de variables explicatives.
3. On **estime** les paramètres β du modèle pour obtenir une estimation convergente des moments, sous la forme $f(\mathbf{x}_i; \widehat{\beta})$ et $g(\mathbf{x}_i; \widehat{\beta})$.

Sources d'erreur dans l'évaluation du MAG

1. Le **processus générateur des données (DGP) est mal specifié** (choix des variables, hypothèses non valides, effets non linéaires...), dès lors la formule théorique du MAG est mal spécifiée.
2. La **définition théorique du MAG** associée au processus générateur des données (DGP) n'est pas correcte.
   Example : on considère MAG $= f(X;\theta)$*, mais la formule* $f(.)$ *est fausse.*
3. Les **paramètres du modèle sont mal estimés**, i.e. θ n'est pas estimé de façon convergente, $\widehat{\theta}$ est biaisé...

Description du DGP

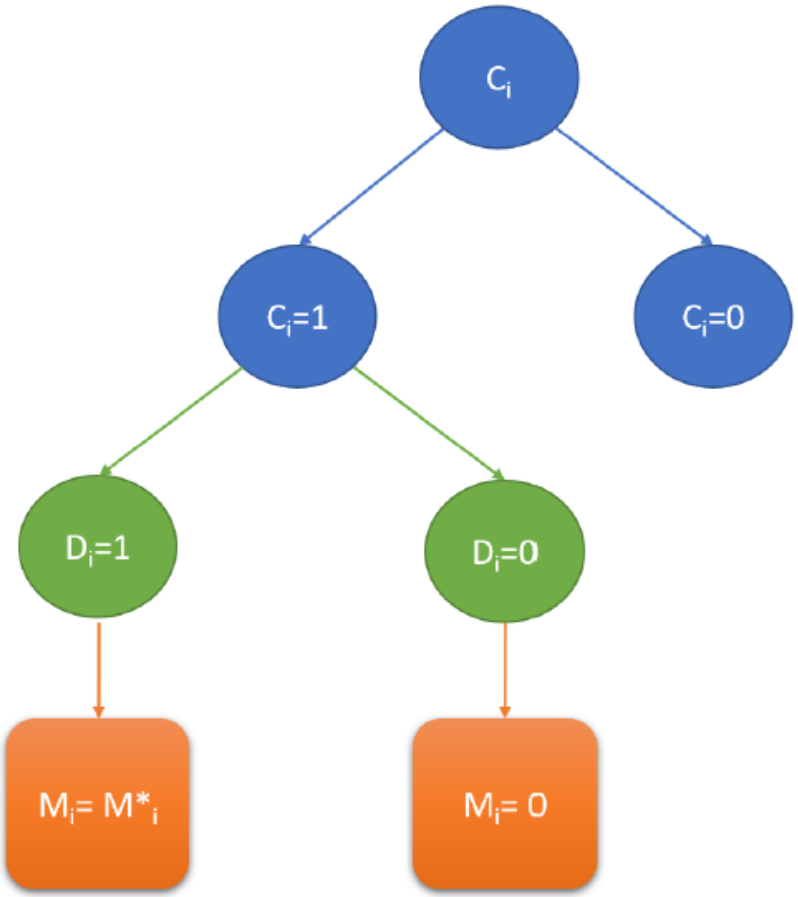


FIGURE – Description schématique du DGP

Equation de contôle du DGP

Le contrôle

Soit C_i la variable indiquant si l'entreprise i a été contrôlée ($C_i = 1$) ou non ($C_i = 0$), telle que :

$$C_i = \begin{cases} 1 & \text{si } C_i^* = \mathbf{X}_{c,i}\beta_c + \varepsilon_{c,i} > 0 \\ 0 & \text{sinon} \end{cases} \qquad \forall i = 1, \ldots, n$$

où C_i^* est une variable latente, $\mathbf{X}_{c,i}$ un ensemble de k_c facteurs, β_c un vecteur de paramètres et $\varepsilon_{c,i}$ i.i.d. avec $\mathbb{E}(\varepsilon_{c,i}) = 0$ et $\mathbb{V}(\varepsilon_{c,i}) = \sigma_c^2$.

Equation de détection du DGP

La fraude

Soit $\widetilde{D}_i$ la variable dichotomique indiquant si l'entreprise i fraude ($\widetilde{D}_i = 1$) ou ne fraude pas ($\widetilde{D}_i = 0$), tel que :

$$\widetilde{D}_i = \begin{cases} 1 & \text{si } D_i^* = \mathbf{X}_{d,i}\beta_d + \varepsilon_{d,i} > 0 \\ 0 & \text{sinon} \end{cases} \qquad \forall i = 1, \ldots, n$$

où D_i^* est une variable latente, $\mathbf{X}_{d,i}$ un ensemble de k_d facteurs, β_d un vecteur de paramètres et $\varepsilon_{d,i}$ i.i.d avec $\mathbb{E}(\varepsilon_{d,i}) = 0$ et $\mathbb{V}(\varepsilon_{d,i}) = \sigma_d^2$.

Equation de redressement du DGP

Le montant potentiel de fraude

Soit M_i^* la variable latente désignant le montant potentiel en euros du redressement adressé à l'entreprise agricole i telle que :

$$M_i^* = \begin{cases} \mathbf{X}_{m,i}\beta_m + \varepsilon_{m,i} & \text{si } \widetilde{D}_i = 1 \\ 0 & \text{sinon} \end{cases} \qquad \forall i = 1, \ldots, n,$$

avec $\mathbf{X}_{m,i}$ un ensemble de k_m facteurs, β_m un vecteur paramètres et où le terme d'erreur $\varepsilon_{m,i}$ vérifie $\mathbb{E}(\varepsilon_{m,i}) = 0$ et $\mathbb{V}(\varepsilon_{m,i}) = \sigma_m^2$.

Equation de redressement du DGP

Le montant potentiel n'est observable que pour les individus ayant été (i) effectivement contrôlés par les caisses de la MSA, et (ii) redressés suite à la constatation d'une fraude.

Le montant du redressement

On note M_i le montant du redressement effectivement observé tel que :

$$M_i = \begin{cases} \mathbf{X}_{m,i}\beta_m + \varepsilon_{m,i} & \text{si } C_i = 1 \text{ et } \widetilde{D}_i = 1 \\ 0 & \text{sinon} \end{cases} \qquad \forall i : C_i = 1,$$

Définition théorique du manque à gagner

MAG agrégé

Le MAG agrégé est défini par :

$$
\begin{aligned}
MAG &= \sum_{i:(C_i=0)\cap(\widetilde{D}_i=1)} M_i^* \\
&= \sum_{i:C_i=0} \underbrace{M_i^*}_{\text{v.a.}} \times \underbrace{\mathbf{1}_{(\widetilde{D}_i=1)}}_{\text{v.a.}} = \sum_{i=1}^{n} \underbrace{M_i^*}_{\text{v.a.}} \times \mathbf{1}_{(C_i=0)} \times \underbrace{\mathbf{1}_{(\widetilde{D}_i=1)}}_{\text{v.a.}}
\end{aligned}
$$

où $\mathbf{1}_{(.)}$ est la fonction indicatrice valant 1 si la condition est vraie et 0 sinon.

Définition théorique du manque à gagner

Moments conditionels du MAG

Sous les hypothèses H1-H5 et le processus générateur de données, les moments conditionnels du MAG agrégé vérifient

$$\mathbb{E}_X\left(MAG\right) = \sum_{i:C_i=0} \mathbb{E}_X\left(M_i^* | \left(C_i = 0\right) \cap \left(\widetilde{D}_i = 1\right)\right) \times \Pr(\widetilde{D}_i = 1 | C_i = 0)$$

$$\mathbb{V}_X\left(MAG\right) = \sum_{i:C_i=0} \mathbb{V}_X\left(M_i^* | \left(C_i = 0\right) \cap \left(\widetilde{D}_i = 1\right)\right) \times \Pr(\widetilde{D}_i = 1 | C_i = 0)$$

où $\mathbf{X} = (\mathbf{X}_c : \mathbf{X}_d : \mathbf{X}_m)$ désigne l'ensemble des variables explicatives du modèle, et où $\mathbb{E}_X\left(.\right)$ et $\mathbb{V}_X\left(.\right)$ désignent respectivement l'espérance et la variance conditionnelle par rapport à $\mathbf{X}$.

Probabilité conditionnelle de fraude

Sous les hypothèses H1-H5, il vient :

$$\begin{aligned}
\Pr(\widetilde{D}_i = 1 | C_i = 0) &= P_{\widetilde{D}_i=1} \\
&= \Pr\left(\varepsilon_{d,i} > -\mathbf{X}_{d,i}\beta_d \middle| \varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \\
&= 1 - \Pr\left(\varepsilon_{d,i} < -\mathbf{X}_{d,i}\beta_d \middle| \varepsilon_{c,i} < -\mathbf{X}_{c,i}\beta_c\right) \\
&= 1 - \frac{\Phi_2\left(-\mathbf{X}_{c,i}\beta_c, -\mathbf{X}_{d,i}\beta_d; \Sigma_{cd}\right)}{\Phi\left(-\mathbf{X}_{c,i}\beta_c/\sigma_c\right)}
\end{aligned}$$

où Σ_{cd} désigne la matrice de variance covariance du vecteur $(\varepsilon_{c,i}, \varepsilon_{d,i})'$, $\Phi_2(.,.;\Sigma_{cd})$ la cdf de la loi normale bivariée d'espérance nulle et de matrice de variance covariance Σ_{cd}, et $\Phi(.)$ désigne la cdf de la loi normale standard univariée.

Espérance du manque à gagner

Sous les hypothèses H1-H5 et le processus générateur de données, **l'espérance conditionnelle du MAG agrégé** est définie par :

$$\begin{aligned}\mathbb{E}_X\left(MAG\right) \quad = \quad & \sum_{i:C_i=0} \mathbf{X}_{m,i}\ \beta_m\ P_{\widetilde{D}_i=1} \\ & + \delta_c \sum_{i:C_i=0} \mathbb{E}_X\left(\varepsilon_{c,i} \middle| \left(\varepsilon_{c,i} < b_{c,i}\right) \cap \left(\varepsilon_{d,i} > a_{d,i}\right)\right) \times P_{\widetilde{D}_i=1} \\ & + \delta_d \sum_{i:C_i=0} \mathbb{E}_X\left(\varepsilon_{d,i} \middle| \left(\varepsilon_{c,i} < b_{c,i}\right) \cap \left(\varepsilon_{d,i} > a_{d,i}\right)\right) \times P_{\widetilde{D}_i=1}\end{aligned}$$

où les seuils de troncature sont définis par $b_{c,i} = -\mathbf{X}_{c,i}\beta_c$ et $a_{d,i} = -\mathbf{X}_{d,i}\beta_d$, et où δ_c et δ_d désignent les coefficients de corrélation partielle de $\varepsilon_{m,i}$ sur $\varepsilon_{c,i}$ et $\varepsilon_{d,i}$ respectivement.

Variance du manque à gagner

Sous les hypothèses H1-H5 et le processus générateur de données, **la variance conditionnelle du MAG agrégé** est définie par :

$$
\begin{aligned}
\mathbb{V}_X\left(MAG\right) &= \delta_c^2 \sum_{i:C_i=0} \mathbb{V}_X\left(\varepsilon_{c,i} \middle| \left(\varepsilon_{c,i} < b_{c,i}\right) \cap \left(\varepsilon_{d,i} > a_{d,i}\right)\right) \times P_{\widetilde{D}_i=1} \\
&+ \delta_d^2 \sum_{i:C_i=0} \mathbb{V}_X\left(\varepsilon_{d,i} \middle| \left(\varepsilon_{c,i} < b_{c,i}\right) \cap \left(\varepsilon_{d,i} > a_{d,i}\right)\right) \times P_{\widetilde{D}_i=1} \\
&+ 2\delta_c\delta_d \sum_{i:C_i=0} \mathbb{C}ov_X\left(\varepsilon_{c,i}, \varepsilon_{d,i} \middle| \left(\varepsilon_{c,i} < b_{c,i}\right) \cap \left(\varepsilon_{d,i} > a_{d,i}\right)\right) \times P_{\widetilde{D}_i=1}
\end{aligned}
$$

où les seuils de troncature sont définis par $b_{c,i} = -\mathbf{X}_{c,i}\beta_c$ et $a_{d,i} = -\mathbf{X}_{d,i}\beta_d$, et où δ_c et δ_d désignent les coefficients de corrélation partielle de $\varepsilon_{m,i}$ sur $\varepsilon_{c,i}$ et $\varepsilon_{d,i}$ respectivement.

Prévision du MAG

La prévision du MAG agrégé, notée $\widehat{MAG}$, et l'intervalle de confiance à $1-\alpha\%$ associé à cette prévision sont définis par :

$$\widehat{MAG} = \mathbb{E}_X\left(MAG\right)$$

$$IC_{1-\alpha} = \left[\mathbb{E}_X\left(MAG\right) \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\mathbb{V}_X\left(MAG\right)}\right]$$

Résultats pour un tirage

Dans le cadre de cette simulation, nous pouvons vérifier la validité de nos formules de prévision et d'intervalle de confiance sur le MAG agrégé.

$$\widehat{MAG} = \mathbb{E}_X(MAG) = 33\ 420$$

$$\mathbb{V}_X(MAG) = 1\ 590$$

$$IC_{95\%} = [33\ 342, 33\ 498]$$

réalisation du MAG : $mag = 33\ 796$

Résultats pour 1000 tirages

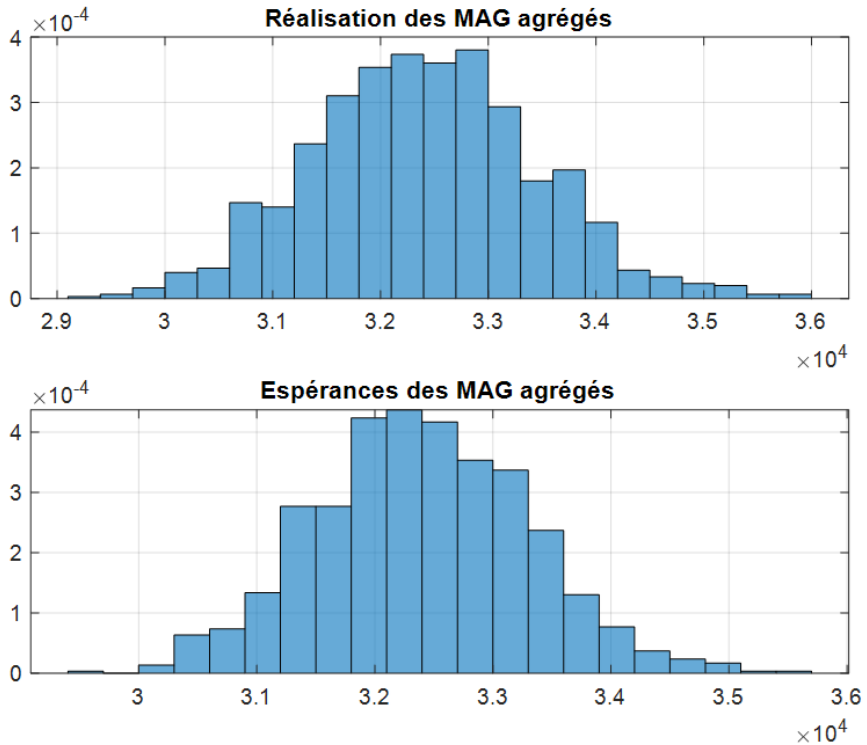


FIGURE – Réalisations et prévisions des MAG agrégés

Résultats pour 1000 tirages

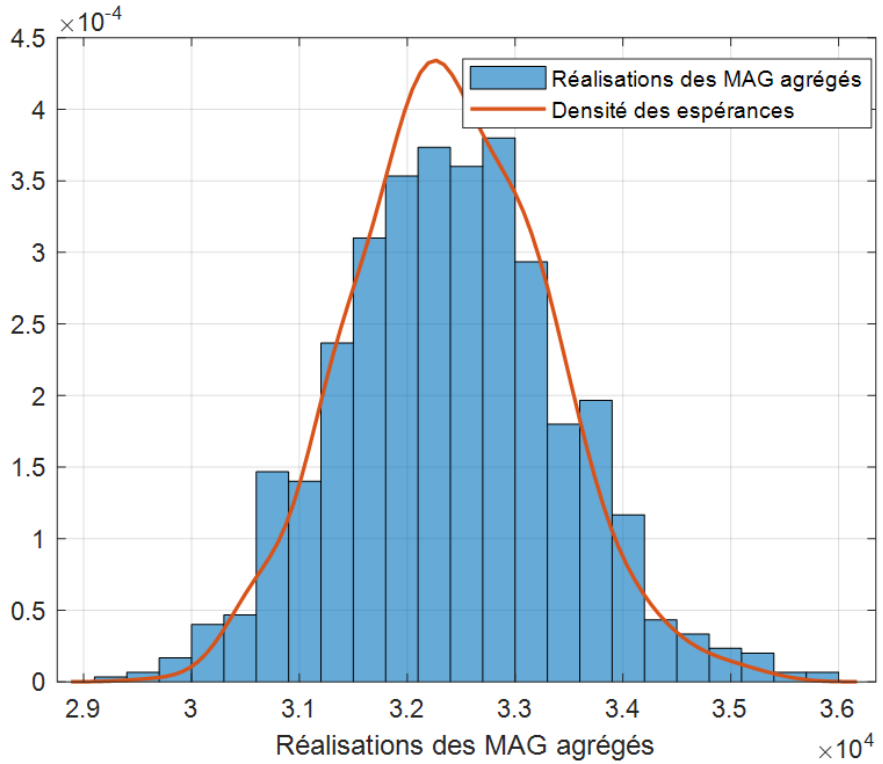


FIGURE – Histogrammes des réalisations du MAG agrégé et densité des prévisions

Plan

Estimation du modèle

- On s'intéresse à présent à l'**estimation** des $k_c + k_d + k_m + 4$ paramètres $\theta = (\beta_c' \ \beta_d' \ \beta_m' \ \rho_{cd} \ \rho_{cm} \ \rho_{dm} \ \sigma_m)'$.
- Cette estimation est réalisée à partir d'un **échantillon d'observations** de variables $\{C_i, D_i, M_i\}$.
- Nous introduisons ici une variable *observable* D_i égale à la variable $\widetilde{D}_i$ représentant la *détection* de la fraude, *uniquement* pour les entreprises contrôlées.
- Cette variable dichotomique D_i indique si l'entreprise contrôlée i a été redressée ($D_i = \widetilde{D}_i = 1$) ou non ($D_i = \widetilde{D}_i = 0$) à la suite du contrôle.

$$D_i = \widetilde{D}_i \quad \forall i : C_i = 1$$

Décisions de contrôle et de détection

Definition

Forme du modèle emboîté Les décisions de contrôle et de détection peuvent se représenter sous la forme d'une structure de type **Probit emboité (nested probit)** avec dépendance.

- Les décisions de fraude $\widetilde{D}_i$ et de contrôle C_i sont supposées non-emboîtées (même si elles sont liées), mais les évenements de détection D_i et de contrôle C_i sont des événements emboités.
- Ce modèle est peu commun, car d'une part les modèles emboités sont plus souvent de type logit et d'autre part les modèles emboités sont rarement avec dépendances.
- Ce modèle correspond aussi à un modèle de **type bi-Probit avec censure** des observations de D_i des entreprises pour lesquelles $C_i = 0$.

Décisions de contrôle et de détection

Log-vraisemblance

La log-vraisemblance du modèle bi-probit avec censure associé aux décisions de contrôle et de détection s'écrit :

$$\begin{aligned}\ell_n(\theta; C, D) = & \sum_{i:C_i=0} \ln\left(\Phi\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c\right)\right) \\ & + \sum_{i:D_i=0} \ln\left(\Phi_2\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c, q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d; -\rho_{cd}\right)\right) \\ & + \sum_{i:D_i=1} \ln\left(\Phi_2\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c, q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d; \rho_{cd}\right)\right)\end{aligned}$$

où $\Phi(u)$ désigne la cdf de la loi normale standard et $\Phi_2(u, v; \rho)$ la cdf de la normale bivariée $\mathcal{N}(\mathbf{0}, \Sigma_2)$ avec $\Sigma_2 = [1\ \rho\ ;\ \rho\ 1]$.

Modèle complet avec redressement

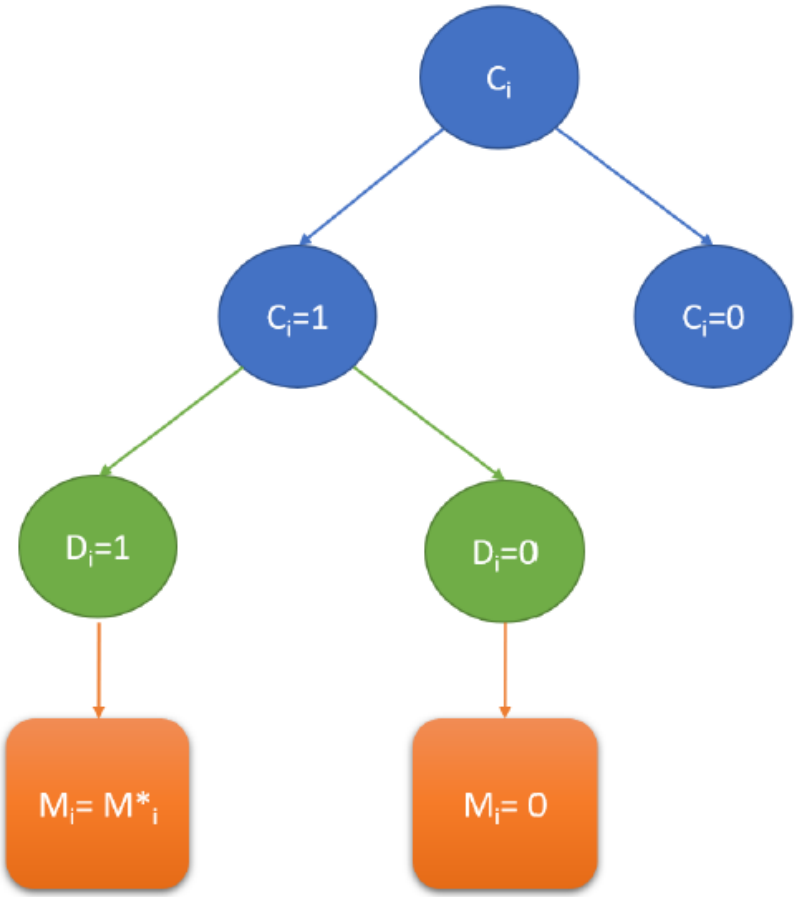


FIGURE – Modèle tobit de type II avec censure de type probit emboité

Modèle complet avec redressement

Forme du modèle complet

La structure du modèle complet s'apparente à un modèle de censure sur le redressement M_i de **type Tobit II** (Amemiya (1984)) avec un mécanisme de censure représenté par un modèle bi-probit lui-même censuré, ou de façon équivalente un modèle probit emboîté avec dépendances.

Modèle complet avec redressement

La log-vraisemblance associée aux événements de détection, de contrôle et de redressement est définie par :

$$\begin{aligned}\ell_n(\theta; C, D, M) &= \sum_{i:C_i=0} \ln\left(\Pr\left(C_i = 0\right)\right) + \sum_{i:M_i=0} \ln\left(\Pr\left(\left(D_i = 0\right) \cap \left(C_i = 1\right)\right)\right) \\ &\quad + \sum_{i:M_i\neq 0} \ln\left(\begin{array}{c} f_{M|C,D}\left(M_i^* \mid D_i^* > 0, C_i^* > 0\right) \\ \times \Pr\left(\left(D_i = 1\right) \cap \left(C_i = 1\right)\right) \end{array}\right)\end{aligned}$$

- Les deux premiers termes de la vraisemblance sont identiques à ceux présentés précédemment.
- Reste à caractériser **la densité conditionnelle du redressement** M_i^* sachant que l'entreprise a été contrôlée et détectée comme fraudeuse, i.e., $D_i^* > 0$ et $C_i^* > 0$, noté $f_{M|C,D}\left(u, v\right)$.

Modèle complet avec redressement

Log-vraisemblance

La log-vraisemblance du modèle complet associé aux décisions de contrôle, de détection et de redressement s'écrit :

$$\begin{aligned}\ell_n(\theta; C, D, M) = & \sum_{i:C_i=0} \ln\left(\Phi\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c\right)\right) + \sum_{i:M_i=0} \ln\left(\Phi_2\left(q_{c,i}\mathbf{X}_{c,i}\widetilde{\beta}_c, q_{d,i}\mathbf{X}_{d,i}\widetilde{\beta}_d; -\rho_{cd}\right)\right) \\ & + \sum_{i:M_i\neq 0} \ln\left(\frac{1}{\sigma_m}\phi\left(\frac{M_i^* - \mathbf{X}_{m,i}\beta_m}{\sigma_m}\right)\Phi_2\left(\mu_{c,i}, \mu_{d,i}; \Sigma_{CD|M}\right)\right)\end{aligned}$$

où $\Phi(u)$ désigne la cdf de la loi normale standard, $\Phi_2(u, v; \rho)$ la cdf de la normale bivariée $\mathcal{N}(\mathbf{0},\Sigma_2)$ avec $\Sigma_2 = [1\ \rho\ ;\ \rho\ 1]$, et $\Phi_2\left(u, v; \widetilde{\Sigma}_2\right)$ la cdf de la normale bivariée $\mathcal{N}\left(\mathbf{0}, \widetilde{\Sigma}_2\right)$.

Modèle complet avec redressement

Maximum Likelihood: Modèle complet

Estimated	True (ratios)	
-6.5172	-6.5549	paramètres équation de contrôle
0.7049	0.7071	
-1.4059	-1.4142	
2.1029	2.1213	
-2.8077	-2.8284	
-7.1104	-7.1630	paramètres équation de détection
3.5034	3.5355	
-4.2350	-4.2426	
33.6295	33.7500	paramètres équation de redressement
7.0433	7.0000	
0.7738	0.8000	corrélations
0.3004	0.3000	
0.5626	0.5000	
2.2651	2.2361	écart-type terme d'erreur sur le redressement

FIGURE – Estimation par MV du modèle complet

Plan

Conclusion

1. Nous avons proposé un **modèle sur la distribution conditionnelle** du MAG (DGP).
2. De ce DGP, nous avons dérivé des formules explicites pour les deux **premiers moments** du MAG agrégé, validée par simulations.
3. Nous avons proposé une **méthode d'estimation par maximum de vraisemblance** qui permet d'estimer de façon convergente l'ensemble des paramètres du modèle, y compris les corrélations.
4. Nous avons montré que l'approche à la Heckman (ratio de Mills) n'est pas valide dans ce contexte, car non seulement elle ne permet pas d'estimer les corrélations, mais qu'elle n'est pas adaptée dans le cas d'un modèle non-linéaire.
5. **Estimation du MAG sur données réelles**.

Plan

Processus générateur des données

Hypothèse H1 (normalité) : *On suppose que les termes d'erreurs* $\varepsilon_{c,i}$ *et* $\varepsilon_{d,i}$ *suivent une distribution normale admettant* ρ_{cd} *pour corrélation.*

Hypothèse H2 (identification) : *Il existe au moins un facteur explicatif de la décision de contrôle* $X_{c,i,u} \in X_{c,i}$ *et un facteur explicatif de la décision de fraude* $X_{d,i,v} \in X_{d,i}$ *tels que* $\mathbb{C}ov(X_{c,i,u}, X_{d,i,v}) = 0$, $\mathbb{C}ov(X_{c,i,u}, \varepsilon_{d,i}) = 0$, *et* $\mathbb{C}ov(X_{d,i,v}, \varepsilon_{c,i}) = 0$.

Processus générateur des données

Nous posons deux hypothèses techniques supplémentaires :

Hypothèse H3 (normalité) : *On suppose que le terme d'erreur* $\varepsilon_{m,i}$ *admet une distribution normale, et qu'il peut être lié aux termes d'erreur* $\varepsilon_{c,i}$ *et* $\varepsilon_{d,i}$*. On note respectivement* ρ_{cm} et ρ_{dm} *les corrélations correspondantes.*

Hypothèse H4 (identification) : *Il existe au moins un facteur explicatif du montant du redressement* $X_{m,i,u} \in \mathbf{X}_{m,i}$ et *un facteur explicatif de la décision de contrôle* $X_{c,i,v} \in \mathbf{X}_{c,i}$ *tels que* $\mathbb{C}ov(X_{m,i,u}, X_{c,i,v}) = 0$, $\mathbb{C}ov(X_{m,i,u}, \varepsilon_{c,i}) = 0$, *et* $\mathbb{C}ov(X_{c,i,v}, \varepsilon_{m,i}) = 0$.

Processus générateur des données

Hypothèse H5 (structure informationnelle) : *On suppose que l'on évalue les moments de la distribution du MAG agrégé alors que la décision de contrôle a été prise par la MSA et est observée, mais que la décision de fraude des entreprises n'est pas observable.*

Plan

Ratio de Mills

Dans le cas où les décisions de contrôle et de fraude ne sont pas liées, c'est-à-dire si $\rho_{cd} = 0$, on fait apparaître les inverses de ratio de Mills

$$
\begin{aligned}
\mathbb{E}_X(MAG) &= \sum_{i:C_i=0} \mathbf{X}_{m,i}\beta_m P_{\widetilde{D}_i=1} + \delta_c \sum_{i:C_i=0} \mathbb{E}_X\left(\varepsilon_{c,i} \middle| \varepsilon_{c,i} < b_{c,i}\right) \times P_{\widetilde{D}_i=1} \\
&\quad + \delta_d \sum_{i:C_i=0} \mathbb{E}_X\left(\varepsilon_{d,i} \middle| \varepsilon_{d,i} > a_{d,i}\right) \times P_{\widetilde{D}_i=1} \\
&= \sum_{i:C_i=0} \mathbf{X}_{m,i}\beta_m P_{\widetilde{D}_i=1} - \delta_c \sum_{i:C_i=0} \sigma_c \frac{\phi\left(b_{c,i}/\sigma_c\right)}{\Phi\left(b_{c,i}/\sigma_c\right)} \times P_{\widetilde{D}_i=1} \\
&\quad + \delta_d \sum_{i:C_i=0} \sigma_d \frac{\phi\left(a_{d,i}/\sigma_d\right)}{1-\Phi\left(a_{d,i}/\sigma_d\right)} \times P_{\widetilde{D}_i=1}
\end{aligned}
$$

avec $P_{\widetilde{D}_i=1} = \Pr(\widetilde{D}_i = 1 | C_i = 0) = 1 - \Phi\left(a_{d,i}/\sigma_d\right).$

Ratio de Mills

- Dans le cas général $\rho_{cd} \neq 0$, l'expression des moments conditionels du redressement M_i^* fait intervenir les moments d'une **loi normale bi-variée avec double troncature**.
- Manjunath et Wilhelm (2010) étendent ces résultats dans le cas d'une distribution multivariée avec double troncature arbitraire et fournissent un package R package tmvtnorm.
- Kan et Robotti (2018) proposent une approche alternative basée sur des relations de récurrence entre les intégrales qui impliquent la densité de la normale multivariée, et proposent différents codes Matlab permettant de calculer les moments pour des lois multivariées avec des double troncatures arbitraires.
- Le point important est que ces moments à double troncature peuvent être très différents des **inverses des ratios de Mills** habituellement utilisés pour traiter des problèmes de sélection.