De l'observation du revenu déclaré en tranches à la prédiction du vrai revenu fiscal..

… ou comment estimer un modèle économétrique sans disposer d'observations

Marc CHRISTINE

(ex-Insee)

PLAN

1. Position et origine du problème
2. Première approche : non paramétrique
3. Deuxième approche : paramétrique
4. Troisième approche : économétrique
5. Perspectives

1. POSITION ET ORIGINE DU PROBLÈME

- Le revenu : variable essentielle du comportement des ménages

- Difficulté : disposer d'une observation correcte de cette variable

- Si échantillon tiré dans les fichiers fiscaux :

revenu « vrai » = revenu fiscal déclaré à l'IRPP

Sinon : observation directe dans les enquêtes

GROUPE « MARGES » de l'Insee

Comment intégrer dans le processus de calage des enquêtes Ménages les données déclarées sur les revenus

- **.. pour les caler sur un vrai total fiscal**
- **.. en respectant certains quantiles de distribution de ces revenus dans la population ?**

Problèmes posés par l'utilisation des revenus déclarés par rapport à une référence fiscale :

- adéquation conceptuelle entre revenus fiscaux et revenus « ressentis », quel champ couvert ?
- sous-estimations systématiques par rapport au revenu « vrai »
- décalages temporels entre donnée d'enquête et source de calage
- erreurs de mesure (confusion *avant* / *après* impôt)
- non-réponse
- déclaratif tronqué : ***réponses en tranches***

EXEMPLES :

- **Tronc commun des ménages (TCM)** : nombreuses questions

détaillant l'existence des différents types de ressources et :
« *quel est actuellement le montant mensuel des ressources*
de l'ensemble de votre ménage ? »

(réponse en valeur ou « *à combien les estimez-vous pour un mois ordinaire ?* » en tranches)

- **Suivi de la demande touristique (SDT)** : « *somme des ressources mensuelles nettes du foyer avant prélèvement à la source* » :
- **Treize tranches :**

Moins de 300 euros / 300-600 / 601-900 / 901-1200 / 1201-1500 /
1501-1900 /1901-2300 / 2301-2700 / 2701-3000 / 3001-3800 / 3801-5300 /
5301-6900 / 6901 ou plus

- **Donc il n’est pas possible de caler directement les revenus déclarés sur des totaux fiscaux.**

- Principe de la méthode proposée :

« Transformer » les revenus déclarés en revenus vrais.

- On suppose que l’on se place :

- en aval d’un processus de correction de la non-réponse
- en aval d’un processus d’analyse, de correction ou d’élimination des points aberrants
- ***on ne sait pas observer dans les enquêtes le vrai revenu fiscal.***

- Cependant, le problème peut avoir une solution grâce aux procédures d’***appariements***, dont la qualité augmente et qui se généralisent, ou de ***tirage direct*** dans les bases fiscales.

2. PREMIÈRE APPROCHE : NON PARAMÉTRIQUE

- **X : revenu fiscal vrai, fonction de répartition F_0**
- **Revenus observés Y_i , variables $i.i.d$, de fonction de répartition F.**

- On cherche à imputer des revenus vrais X_i^* : $X_i^* = T(Y_i)$, $T\ croissante$,

.. de telle sorte que ce revenu ait une fonction de répartition donnée F_0.

- On doit donc avoir :

$$P\{T(Y) < x\} = F_0(x) \Leftrightarrow P\{Y < T^{-1}(x)\} = F_0(x)$$

- **Solution :** $T = F_0^{-1} \circ F.$

Mise en œuvre sur un échantillon

- La fonction F_0 (***définie sur la population à partir des données fiscales,*** taille N) est exprimée soit sous forme analytique, soit sous forme de répartition discrète : $F_0(x) = \frac{1}{N}\sum_{i=1}^{N} 1_{X_i<x}$.
- On dispose d'un échantillon s, probabilité d'inclusion π_i.
- On remplace la fonction de répartition F de Y par l'estimation de la fonction de répartition empirique dans l'échantillon :

$$\hat{F}_Y(y) = \frac{\sum_{i\in s}\frac{1_{Y_i<y}}{\pi_i}}{\sum_{i\in s}\frac{1}{\pi_i}}$$

D’où la transformation approchée :

$$\hat{T} = F_0^{-1} \circ \hat{F}_Y$$

$$\text{ou } \hat{T} = F_0^{-} \circ \hat{F}_Y \text{ avec :}$$

$$F_0^{-}(y) = Inf\ \{z\,; F_0(z) > y\}$$

- Et la *prédiction du revenu vrai* :

$$\hat{X}_j = \hat{T}(Y_j) = F_0^{-}\left[\hat{F}_Y(Y_j)\right] = F_0^{-}\left[\frac{\sum_{i\in s}\frac{1_{Y_i<Y_j}}{\pi_i}}{\sum_{i\in s}\frac{1}{\pi_i}}\right].$$

- **Si l'on ordonne les revenus déclarés :** $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$,
- On obtient une expression de la fonction de répartition empirique :

$$\hat{F}_Y\left(Y_{(j)}\right) = \frac{\sum_{i \in s} \frac{1_{Y_i < Y_{(j)}}}{\pi_i}}{\sum_{i \in s} \frac{1}{\pi_i}} = \frac{\sum_{i=1}^{j-1} \frac{1}{\pi_{(i)}}}{\sum_{i \in s} \frac{1}{\pi_i}}$$

- Et de la prédiction du revenu vrai pour l'unité (j) :

$$\hat{X}_{(j)} = F_0^{-}\left[\hat{F}_Y\left(Y_{(j)}\right)\right] = F_0^{-}\left[\frac{\sum_{i=1}^{j-1} \frac{1}{\pi_{(i)}}}{\sum_{i \in s} \frac{1}{\pi_i}}\right]$$

- **Cette solution ne prend pas en compte les valeurs des revenus observés mais seulement leurs *rangs* dans l'échantillon.**

3. DEUXIÈME APPROCHE : PARAMÉTRIQUE

▶ **On suppose que les deux fonctions de répartition ne diffèrent que par des paramètres de position et d'échelle :**

$$F(y) = F_0\ (\alpha + \beta y),\ \beta > 0.$$

- D'où l'expression des revenus vrais prédits :

$$X_i = F_0^{-1} \circ F(Y_i) = \alpha + \beta Y_i.$$

$\Longrightarrow$ **Deux paramètres à estimer : α, β, à partir de l'échantillon**

- En imposant deux types de conditions :
  - Calage sur le vrai revenu total des ménages R_T.
  - Calage sur des quantiles q_τ de la distribution de ces vrais revenus.

D’où un système d’équations :

$$\sum_{i \in s} \frac{X_i}{\pi_i} = \alpha \sum_{i \in s} \frac{1}{\pi_i} + \beta \sum_{i \in s} \frac{Y_i}{\pi_i} = R_T \ \textbf{(C1)}$$

$$\sum_{i \in s} \frac{1_{X_i < q_\tau}}{\pi_i} = \sum_{i \in s} \frac{1_{\frac{Y_i - \alpha}{\beta} < q_\tau}}{\pi_i} = \tau \sum_{i \in s} \frac{1}{\pi_i} \ \textbf{(C2)}$$

- $\hat{q}_\tau(Y)$ = **quantile empirique d’ordre τ de la distribution des Y_i dans l’échantillon :**

$$\hat{q}_\tau(Y) = Inf \left\{ y \, ; \, \sum_{i \in s} \frac{1_{Y_i < y}}{\pi_i} > \tau \sum_{i \in s} \frac{1}{\pi_i} \right\} :$$

$$\textbf{(C2)} \ \textbf{\textit{devient}} \ : \alpha + \beta \, \hat{q}_\tau(Y) = q_\tau.$$

Avec différentes valeurs des quantiles τ : médiane, quartiles, déciles…

▶ Résolution exacte

- Calage sur un seul quantile : la ***médiane***
- Système de deux équations à deux inconnues

▶ Résolution approchée.

- On cale approximativement sur les quantiles
- On cherche alors :

  - $Min\ \sum_{\tau}[\alpha + \beta \hat{q}_{\tau}(Y) - q_{\tau}]^2$

  - sous la condition :

$$\alpha \sum_{i \in s} \frac{1}{\pi_i} + \beta \sum_{i \in s} \frac{Y_i}{\pi_i} = R_T$$

Solution :

$$\hat{\beta} = \frac{\sum_{\tau} \left[\hat{q}_{\tau}(Y) - \hat{R}_m(Y)\right] \left[q_{\tau} - \hat{R}_m\right]}{\sum_{\tau} \left[\hat{q}_{\tau}(Y) - \hat{R}_m(Y)\right]^2}$$

$$\hat{\alpha} = \frac{R_T - \hat{\beta} \sum_{i \in s} \frac{Y_i}{\pi_i}}{\sum_{i \in s} \frac{1}{\pi_i}}$$

- $\widehat{\boldsymbol{R}}_{\boldsymbol{m}}(\boldsymbol{Y}) = \frac{\sum_{i \in s} \frac{Y_i}{\pi_i}}{\sum_{i \in s} \frac{1}{\pi_i}}$ = **estimateur du revenu moyen déclaré.**

- $\widehat{\boldsymbol{R}}_{\boldsymbol{m}} = \frac{R_T}{\sum_{i \in s} \frac{1}{\pi_i}}$ = **estimateur du revenu moyen vrai.**

Application au calage ultérieur

- on a des revenus vrais *prédits* : $\hat{X}_i = \hat{\alpha} + \hat{\beta} Y_i$.

- on les injecte dans les équations de calage sur le revenu vrai total et les quantiles de revenus vrais

- **Calage sur le revenu total (nouveaux poids $\boldsymbol{\omega_i}$) :**

$$\sum_{i \in s} \omega_i \hat{X}_i = \hat{\alpha} \sum_{i \in s} \omega_i + \hat{\beta} \sum_{i \in s} \omega_i Y_i = R_T \ \textbf{(C1)}$$

- **Calage sur les quantiles :**

$$\sum_{i \in s} \omega_i 1_{\hat{X}_i < q_\tau} = \sum_{i \in s} \omega_i 1_{\frac{Y_i - \hat{\alpha}}{\hat{\beta}} < q_\tau} = \tau \sum_{i \in s} \omega_i \ \textbf{(C2)}.$$

- compte tenu de la construction des revenus vrais prédits, *ces équations sont déjà satisfaites avec les poids initiaux* $\frac{\mathbf{1}}{\boldsymbol{\pi_i}}$

- on peut rajouter d'autres variables socio-démographiques dans les conditions de calage

- on fait tourner la procédure avec l'ensemble des variables utilisées.

Améliorations possibles de la solution :

- avoir différents modèles selon des catégories de logements ou de ménages (âge, taille, habitat)...

- y compris définies à partir des revenus vrais (classes de quantiles)

- en calant sur les revenus totaux des classes de quantiles considérées.

4. TROISIÈME APPROCHE : ÉCONOMÉTRIQUE

- ***Idée de base : postuler un modèle économétrique aléatoire reliant le revenu déclaré Y au revenu vrai X :***

$$\boldsymbol{Y = a + b\,X + U, b > 0}$$

où on suppose que les variables relatives à la population résultent d'un ***modèle aléatoire de génération des données.***

4.1 Préliminaires probabilistes

X **et** U **indépendantes**

U **: aléa normal :** $U \rightsquigarrow \mathcal{N}(0, \sigma^2)$

X **: variable aléatoire de loi donnée <u>connue</u> : Densité** f_0,

Fonction de répartition : F_0

▶ **La loi de** $\boldsymbol{Y = a + b\, X + U}$ **a une densité :**

$$g(y) = \int_{-\infty}^{+\infty} \frac{1}{b} f_0\left(\frac{y-u}{b}\right) \frac{1}{\sigma} h\left(\frac{u-a}{\sigma}\right) du = \frac{1}{\sigma} \int_{-\infty}^{+\infty} f_0(z)\, h\left(\frac{y-a-bz}{\sigma}\right) dz\,.$$

h = densité de la loi $\mathcal{N}(0, 1)$.

- **Pour n observations des Y_i indépendantes :**

- **Vraisemblance :**

$$L(y_1, \dots, y_n) = \prod_{i=1}^{n} g(y_i)$$
$$= \frac{1}{\sigma^n} \prod_{i=1}^{n} \left[\int_{-\infty}^{+\infty} f_0(z) h\left(\frac{y_i - a - bz}{\sigma}\right) dz \right].$$

- **Log-vraisemblance :**

$$ln\, L(y_1, \dots, y_n) = -n\, ln\, \sigma + \sum_{i=1}^{n} ln \left[\int_{-\infty}^{+\infty} f_0(z) h\left(\frac{y_i - a - bz}{\sigma}\right) dz \right]$$

- **Estimation des paramètres par maximum de vraisemblance :**

- **3 équations :**

$$\sum_{i=1}^{n} \frac{1}{g(y_i)} \left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i - a - bz}{\sigma}\right) dz \right] = 0$$

$$\sum_{i=1}^{n} \frac{1}{g(y_i)} \left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i - a - bz}{\sigma}\right) z\, dz \right] = 0$$

$$\sigma = -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{g(y_i)} \left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i - a - bz}{\sigma}\right) (y_i - a - bz)\, dz \right].$$

- Compte tenu des deux premières, la dernière s'écrit :

$$\sigma = -\frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{g(y_i)} \left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i - a - bz}{\sigma}\right) dz \right].$$

▲ Attention : problème d'identifiabilité

- Si la variable X est normale, les effets conjoints de X et de U sont « noyés » dans une même loi normale.

- Si $X \leadsto \mathcal{N}(m_0, \mu_0^2)$, alors :

$$Y = a + bX + U \leadsto \mathcal{N}\left(a + bm_0, b^2\mu_0^2 + \sigma^2\right).$$

- La loi normale dépend de deux paramètres mais trois sont inconnus.

- **On peut mieux le voir aussi si l'on utilise la méthode des moments :**

- ***Moments de Y:***

$$EY = a + b\,EX$$

$$VY = b^2 VX + \sigma^2$$

$$E(Y - EY)^3 = b^3 E(X - EX)^3 = b^3 \mu_3$$

Équations aux moments :

$$\hat{b}^3 = \frac{1}{\mu_3}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^3$$

$$\hat{a} = \bar{Y} - \hat{b}\,EX$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \hat{b}^2 VX$$

- Il faut que : $\mu_3 \neq 0$.
- **L'identifiabilité sera assurée avec une distribution de X *asymétrique*.**

- **Prédicteur optimal du revenu vrai**

- On cherche la meilleure approximation de X par une variable aléatoire dépendant de Y, de la forme

$$Z = u(Y),$$

en minimisant l'écart quadratique moyen $E(X - Z)^2$

▶ **Solution : espérance conditionnelle : $Z^* = u^*(Y) = E(X/Y)$,**

$$E(X/Y = y) = \int x \frac{f_{(X,Y)}(x,y)}{f_Y(y)} dx = \frac{\int_{-\infty}^{+\infty} x\, f_0(x) h\left(\frac{y-a-bx}{\sigma}\right) dx}{\int_{-\infty}^{+\infty} f_0(x) h\left(\frac{y-a-bx}{\sigma}\right) dx},$$

- Pour chaque observation y_j de Y, **prédicteur optimal correspondant de X :**

$$X^*(y_j) = \frac{\int_{-\infty}^{+\infty} x f_0(x) h\left(\frac{y_j - \hat{a}_s - \hat{b}_s x}{\hat{\sigma}_s}\right) dx}{\int_{-\infty}^{+\infty} f_0(x) h\left(\frac{y_j - \hat{a}_s - \hat{b}_s x}{\hat{\sigma}_s}\right) dx}$$

On peut aussi écrire :

$$\int_{-\infty}^{+\infty} f_0(x) h\left(\frac{y_j - \hat{a}_s - \hat{b}_s x}{\hat{\sigma}_s}\right) dx = \frac{\hat{b}_s}{\hat{\sigma}_s} \int_{-\infty}^{+\infty} F_0(x)\, h'\left(\frac{y_j - \hat{a}_s - \hat{b}_s x}{\hat{\sigma}_s}\right) dx$$

$$\int_{-\infty}^{+\infty} x\, f_0(x) h\left(\frac{y_j - \hat{a}_s - \hat{b}_s x}{\hat{\sigma}_s}\right) dx = - \int_{-\infty}^{+\infty} F_0(x) \left[h\left(\frac{y_j - \hat{a}_s - \hat{b}_s x}{\hat{\sigma}_s}\right) \right.$$

Autres méthodes d'estimation

Fonction caractéristique de X :

$$\varphi_Y(t) = Ee^{itY} = Ee^{it(a+bX+U)} = e^{ita}\left(Ee^{itbX}\right)\left(Ee^{itU}\right)$$

$$= e^{ita}\,\varphi_X(bt)\,e^{-\frac{t^2\sigma^2}{2}}$$

estimée par : $\frac{1}{n}\sum_{j=1}^{n} e^{itY_j}$,

D'où différentes équations pour diverses valeurs de t

… qu'on peut résoudre par moindres carrés.

Remarques

- Il peut être plus pertinent de travailler en logarithmes :

$$\boldsymbol{ln\, Y = a + b\; ln\, X + U,\; b > 0}$$

… à conditions que la vraie loi de X ne soit pas lognormale.

- La méthode reste inchangée. Le prédicteur optimal de X devient :

$$X^*(y_0) = \frac{\int_0^{+\infty} f_0(ln\, x)\, h\left(\frac{ln\, y_0 - a - bx}{\sigma}\right) dx}{\int_0^{+\infty} \frac{f_0(ln\, x)}{x}\, h\left(\frac{ln\, y_0 - a - bx}{\sigma}\right) dx}.$$

4.2 Mise en œuvre dans le cadre d'un échantillon

- population finie de taille N
- génération aléatoire des données : $(X_i, U_i), i = 1, \dots, N$

- X_i et U_i i.i.d.
- Observations des réalisations y_i des $Y_i = a + bX_i + U_i$ (Revenus *déclarés*)
- **On n'observe ni les X_i ni les U_i mais on connaît la distribution des X_i *(revenus vrais).***

- **- échantillon s de taille n tiré avec un plan de sondage $p(s)$**
  - indépendant du processus de génération des données
  - connu par l'intermédiaire des probabilités d'inclusion d'ordre 1 : π_i

On estime sans biais, *conditionnellement aux observations* y_i, la log-vraisemblance à partir de l'échantillon :

$$(ln\ \hat{L})_s = -N\ ln\sigma + \sum_{i\in s}\frac{1}{\pi_i} ln\left[\int_{-\infty}^{+\infty} f_0(z)h\left(\frac{y_i - a - bz}{\sigma}\right)dz\right]$$

- **On estime les paramètres par maximum de vraisemblance :**

$$\hat{a}_s, \hat{b}_s, \hat{\sigma}_s.$$

- **Prédicteur optimal du revenu vrai, pour une observation y_j de Y_j :**

$$X^*(y_j) = \frac{\int_{-\infty}^{+\infty} x f_0(x)h\left(\frac{y_j - \hat{a}_s - \hat{b}_s x}{\hat{\sigma}_s}\right)dx}{\int_{-\infty}^{+\infty} f_0(x)h\left(\frac{y_j - \hat{a}_s - \hat{b}_s x}{\hat{\sigma}_s}\right)dx}$$

4.3 Cas où le revenu déclaré est en tranches

On n'observe que l'appartenance du revenu déclaré Y_i à une tranche T_k (k = 1, …, $K \geq 2$).

- **On observe le nombre total de variables tombant dans chaque tranche : N_k**
  - **Les N_k suivent une loi multinomiale :**

$$\boldsymbol{\mathcal{M}(N, p_1, \ldots, p_K),\ p_k = P\{Y \in T_k\}}$$

  - **Vraisemblance :** $L(n_1, \ldots, n_K) = P\{N_k = n_k, \forall k\} = \frac{N!}{n_1! \ldots n_K!} \prod_{k=1}^{K} p_k^{n_k}$.

- **Les T_k sont des intervalles** $\Longrightarrow$ les p_k s'expriment au moyen de la fonction de répartition G de Y (dépendant de la fonction connue F_0) :

$$G(y) = P\{Y < y\} = P\left\{X + \frac{U}{b} < \frac{y-a}{b}\right\} = \int_{-\infty}^{+\infty} F_0\left(\frac{y-a-\sigma v}{b}\right) h(v)\,dv\,.$$

Log-vraisemblance :

$$ln\, L(n_1,\dots,n_K) = ln\left(\frac{N!}{n_1!\dots n_K!}\right) + \sum_{k=1}^{K} n_k\, ln\, p_k$$

- **Maximisation** $\Longrightarrow$ **équations du 1^er^ ordre :**

$$\frac{\partial lnL}{\partial \Phi} = \sum_{k=1}^{K+1} \frac{n_k}{p_k}\frac{\partial p_k}{\partial \Phi} = 0, \Phi = (a, b, \sigma).$$

- **Les** $\frac{\partial p_k}{\partial \Phi}$ s'expriment au moyen des $\frac{\partial G}{\partial \Phi}$,

$$p_k = P\{s_{k-1} \leq Y < s_k\} = G(s_k) - G(s_{k-1})$$

$$\frac{\partial G}{\partial a}(y) = -\frac{1}{b}\int_{-\infty}^{+\infty} f_0\left(\frac{y-a-\sigma v}{b}\right) h(v)dv$$

$$\frac{\partial G}{\partial \sigma}(y) = -\frac{1}{b}\int_{-\infty}^{+\infty} f_0\left(\frac{y-a-\sigma v}{b}\right) v\, h(v)dv$$

$$\frac{\partial G}{\partial b}(y) = -\frac{1}{b^2}\int_{-\infty}^{+\infty} f_0\left(\frac{y-a-\sigma v}{b}\right) (y-a-\sigma v)\, h(v)dv$$

$$\Longrightarrow \textit{estimateurs des paramètres } a, b, \sigma$$

- **Prédicteur optimal de X**

$$X_j^* = \sum_{k=1}^{K} u_k^* 1_{Y_j \in T_k}, \text{ réalisant } Min\ E(X_j^* - X_j)^2$$

$$u_k^* = \frac{E(X 1_{Y \in T_k})}{E(1_{Y \in T_k})} = \frac{E(X 1_{Y \in T_k})}{P\{Y \in T_k\}} = \frac{E(X 1_{Y \in T_k})}{p_k}.$$

- Calculable à partir de la seule connaissance de la distribution des X_i
- En remplaçant les paramètres par leurs estimations.

Mise en œuvre sur un échantillon :

- On remplace les n_k par leurs estimations sans biais, *conditionnellement aux* Y_i :

$$(\hat{n}_k)_s = \sum_{i \in s} \frac{1_{y_i \in T_k}}{\pi_i}$$

  - et on maximise la log-vraisemblance estimée :

$$(ln\,\hat{L})_s = ln\left(\frac{n!}{(\hat{n}_1)_s! \ldots (\hat{n}_K)_s!}\right) + \sum_{k=1}^{K} (\hat{n}_k)_s ln\, p_k.$$

- **La suite de la procédure comme ce qui précède.**

4.4 Résumé de la démarche

• Observer les tranches de revenus déclarées.

• Estimer les paramètres du modèle linéaire reliant le revenu déclaré au vrai revenu, par maximum de vraisemblance, ***en supposant connue la distribution théorique des vrais revenus.***

• Imputer le meilleur prédicteur du vrai revenu à chaque ménage à partir de la connaissance de sa tranche de revenu déclarée.

Inconvénient **: un seul revenu prédit par tranche observée !**

► Il est possible d'estimer les paramètres du modèle reliant les revenus observés aux revenus vrais en ne connaissant que les observations tronquées des revenus déclarées en tranches, et sans observer les valeurs des revenus vrais mais seulement leur distribution théorique.

• Appliquer les procédures de calage habituelles en utilisant ces valeurs de revenu imputées (plus celles d'autres variables éventuelles) pour caler sur un vrai revenu total et d'autres totaux connus.

5. PERSPECTIVES ET TRAVAUX ULTÉRIEURS

- tester ces méthodes sur données réelles avec des algorithmes de résolution numérique des équations du maximum de vraisemblance ou des méthodes de simulation (Monte-Carlo)

- tester les différentes procédures proposées de reconstitution des revenus vrais à partir des données collectées dans le TCM ou les enquêtes.

- regarder la stabilité des modèles et des paramètres quand on travaille sur des sous-catégories de population

- comparer les prédicteurs des vrais revenus obtenus par ces procédures avec les revenus fiscaux réels, quand on peut en disposer individuellement grâce aux appariements

- enrichir les modèles économétriques :
- analyser les distributions empiriques des revenus déclarés et des revenus vrais
- peut-on introduire d'autres variables explicatives ?
- tester des aléas non normaux, *asymétriques* ?
- modifier la modélisation : *modèle probit* ?

Merci de votre attention !

Marc Christine

mchristine7577@gmail.com

BIBLIOGRAPHIE

- BECK S., Réflexions sur l'allègement du module Revenus de l'enquête Patrimoine 2017-18, note n°622/DG75-F350/SB/ML du 9 avril 2015
- CHRISTINE M., rapport du Groupe Marges, version révisée novembre 2013, note interne Unité Méthodes Statistiques (UMS), Insee, *unpublished paper*
- CHRISTINE M., communication au 12è Colloque francophone sur les Sondages, 22-24 mars 2023, Campus Cordorcet - Paris
- PIRANDELLO L, « Trois petits papiers en quête d'auteur », Actes des 14ès Journées de méthodologie statistique de l'Insee (2022), sur http://jms-insee.fr
- Tronc commun des ménages (TCM), version 2022, document Insee du 29 octobre 2021