Le traitement du biais de sélection endogène par modèle de Heckman

Laura Castell
Patrick Sillard

Insee Mesurer pour comprendre

01 LA SÉLECTION ENDOGÈNE DANS LES ENQUÊTES AUPRÈS DES MÉNAGES

MÉTHODE DE CORRECTION SUR OBSERVABLES

- **Estimation de la probabilité de participer à partir des variables disponibles pour l'ensemble de l'échantillon**
  - Variables explicatives de la participation et des variables d'intérêt
  - Estimateur d'Horvitz-Thompson sans biais

- **Hypothèse d'indépendance conditionnelle : le mécanisme de non-réponse est ignorable (MAR)**
  - La participation est indépendante des variables d'intérêt, conditionnellement aux observables

SITUATION EN CAS DE SÉLECTION ENDOGÈNE

- **Une variable, explicative de la participation et de la variable d'intérêt, conditionnellement aux observables, est omise dans le modèle**

  → Le mécanisme de non-réponse est non ignorable (MNAR)

  → Les méthodes usuelles de correction sur observables conduisent à des estimations biaisées

  Le biais est d'autant plus important que :

  - **le taux de réponse est faible,**
  - **la variable omise est corrélée à la variable d'intérêt**
  - **la variance de la variable d'intérêt est élevée**

LA SÉLECTION ENDOGÈNE EN PRATIQUE

- **Exemple classique de variable inobservée : l'intérêt pour la thématique de l'enquête**
  - Pas un problème si cet intérêt est fortement corrélé à des caractéristiques socio-démographiques classiques disponibles (âge, sexe, revenu…)
- **Un risque accru avec l'utilisation croissante d'Internet**
  - Une démarche de participation plus pro-active
  - Des taux de réponse relativement faibles

02 CORRECTION DE LA SÉLECTION ENDOGÈNE PAR MODÈLE DE HECKMAN

UNE MODÉLISATION SIMULTANÉE DE LA PARTICIPATION ET DE LA VARIABLE D'INTÉRÊT

- **Le modèle :**

$$\begin{cases} \text{(i)} & y_i = c^1 + \mathbf{z}_i\chi + \epsilon_i^1 \\ \text{(ii)} & r_i^* = c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \epsilon_i^0 \\ \text{(iii)} & r_i = \mathbb{1}(r_i^* \geqslant 0) \end{cases}$$

$$\begin{cases} E\left(\begin{pmatrix}\epsilon_i^0 \\ \epsilon_i^1\end{pmatrix} \middle| \mathbf{z}_i, \mathbf{w}_i\right) = 0 \\ \begin{pmatrix}\epsilon_i^0 \\ \epsilon_i^1\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \Sigma\right) \\ \Sigma = \begin{pmatrix}1 & \varrho\sigma \\ \varrho\sigma & \sigma^2\end{pmatrix} \end{cases}$$

- **Identification de la sélection endogène par la corrélation des aléas ϱ**
- **Modèle applicable aux variables binaires, en remplaçant l'équation (i) par la variable latente**

MÉTHODES D'ESTIMATION

- **Par imputation des y pour les non-répondants**
  - $E(y_i|z_i,w_i,r_i=0)$
  - Méthode classique d'utilisation du modèle de Heckman
- **Par repondération**
  - $E(r_i|z_i,w_i,y_i)$
  - On utilise cette estimation comme modélisation de la participation, puis on applique les mêmes étapes de correction qu'usuellement si besoin (GRH, calage sur marges)
  - Méthode originale, davantage adaptée aux pratiques de diffusion dans la statistique publique

HYPOTHÈSES ET LIMITES

- **Un poids afférent à une variable d'intérêt**
  - Pour faciliter l'utilisation, on peut chercher une pondération qui corrige en grande partie le biais lié à plusieurs variables d'intérêt

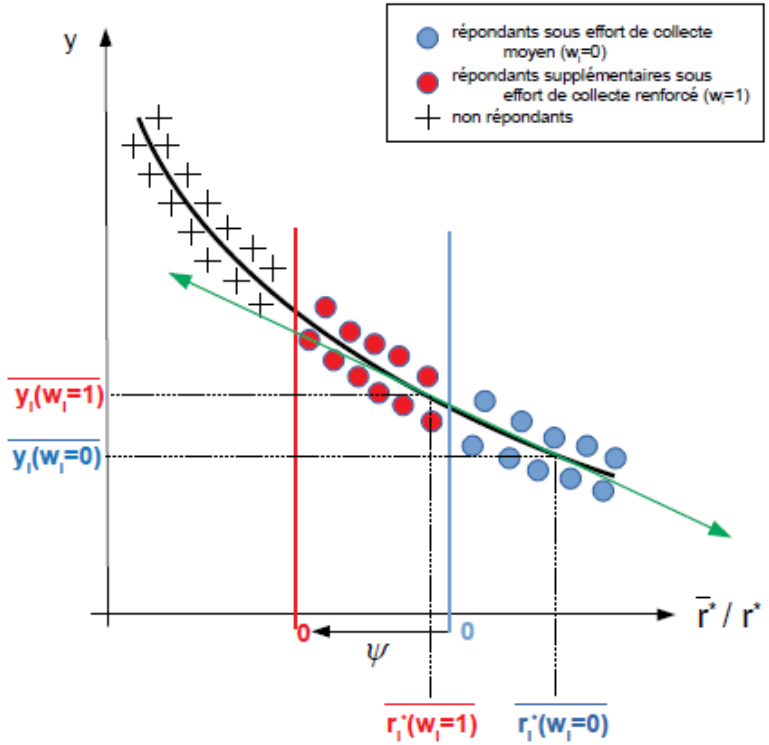
    Exemple : poids associé à l'indicatrice d'au moins un symptôme, dans l'enquête EpiCov pour l'estimation des différents symptômes
- **Existence d'un instrument w, explicatif de la participation (ii) mais pas de la variable d'intérêt (i)**
  - Nécessaire à la convergence du modèle
  - Principe de monotonie : $r_i^*(\mathrm{w}_i = 1) - r_i^*(\mathrm{w}_i = 0) = \psi$

    principe valable pour tous les individus et pas seulement en moyenne

– Linéarité des aléas

- La relation identifiée entre r_i et y_i est supposée linéaire
- Le modèle identifie cette relation au voisinage des taux de participation observée, puis extrapole cette relation sur l'ensemble de l'échantillon

  → plus les taux de participation observés sont éloignés, moins l'extrapolation sera biaisée.
- Des méthodes utilisant d'autres distributions ou moins paramétriques existent



03 QUELLES CONFIGURATIONS D'ENQUÊTE POUR UTILISER CETTE MÉTHODE DE CORRECTION ?

UN PROTOCOLE ADAPTÉ

- **Des sous-échantillons tirés aléatoirement conduisant à des taux de participation différents, et qui respectent le principe de monotonie**
  - Explique la participation mais pas les variables d'intérêt
  - Les participants au protocole le moins incitatif participeraient forcément au protocole renforcé si on leur proposait
- **Exemples :**
  - Un lot avec des incitations financières
  - Un lot avec un protocole de relance renforcé
  - Un lot multimode avec le même mode que dans le lot monomode (protocole « emboîté »)

UNE HYPOTHÈSE FORTE : L'ABSENCE D'EFFET DE MESURE

- **Cas avec un effet de mesure**

$$\begin{cases} \text{(i)} & y_i = c^1 + \mathbf{z}_i\chi + \alpha m_i^A + \epsilon_i^1 \\ \text{(ii)} & r_i^* = c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \epsilon_i^0 \\ \text{(iii)} & r_i = \mathbb{1}(r_i^* \geqslant 0) \end{cases}$$

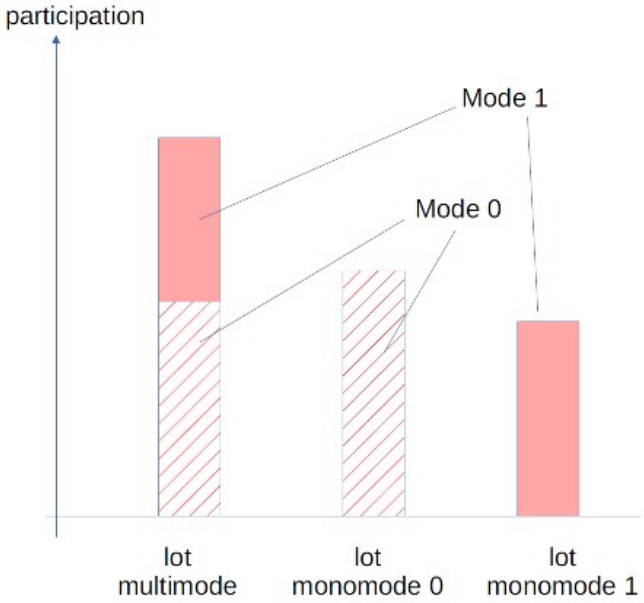
- Si le mode est omis dans l'équation d'outcome : les conditions d'exclusion du modèle ne sont plus respectées

$$E\left(\begin{pmatrix}\epsilon_i^0 \\ 1 \\ \epsilon_i^1\end{pmatrix}\middle|\, \mathbf{z}_i, \mathbf{w}_i\right) = 0$$

- Si le mode est inclus dans l'équation d'outcome : le modèle n'est pas identifiable car le mode m^A est colinéaire avec l'instrument w

IDENTIFIER EFFET DE SÉLECTION ET EFFET DE MESURE

- Un protocole avec trois sous-échantillons tirés aléatoirement
- Disposer de deux instruments pour :
  - estimer l'effet de mesure puis le corriger
  - estimer la sélection endogène en faisant l'hypothèse d'absence d'effet de mesure.



QUELLES PERSPECTIVES ?

- Des possibilités d'identifier et de corriger une sélection endogène et un effet de mesure
- Autres méthodes moins paramétriques envisageables
- Nécessité de prévoir des protocoles adaptés
- Cas d'usage :
  - Réaliser des sous-échantillons pur Internet pour des extensions territoriales, des analyses plus fines par sous-catégories...
  - Réaliser des tests de nouveaux protocoles multimodes

Retrouvez-nous sur

insee.fr

Laura Castell
Patrick Sillard

Insee Mesurer pour comprendre