Présentation de

« A new instrumental méthod for dealing with endogenous selection »

de Xavier d'Haultfoeuille (Paris I – Crest)

par Benjamin Vignolles (SOeS – Crest)

Le cadre

On s'intéresse aux caractéristiques d'une variable aléatoire Y (moments, distribution, effet d'un traitement, etc.) et on observe (Y*, D) tels que

$$\begin{cases} Y^* = Y \ \ si \ \ D = 1 \\ Y^* = 0 \ \ si \ \ D = 0 \end{cases}$$

avec D = ψ (Y, υ) (la sélection est dépendante de la variable d'intérêt)

Des exemples de cas types

(1) Dans une enquête sur la consommation de drogues, il est probable que la non réponse dépende de la consommation des enquêtés

(2) Sur un marché, on n'observe que le prix des biens ou services ayant effectivement fait 'objet d'une transaction.

(3) Pour évaluer l'impact d'une politique publique, on a besoin d'estimer un contrefactuel (ex: les notes des redoublants s'ils n'avaient pas redoublé).

Remarques sur la notion de non réponse

Seul le cas (1) correspond à une situation de sondage: on n'observa pas la variable à cause du care d'enquête.

Dans les cas (2) et (3), les individus pour lesquels Y n'est pas observée sont tels qu'il n'y a pas de réalisation de Y (ex: absence de salaire pour un chômeur).

Les deux formes de non réponse sont traitées de manière indistincte.

Une solution possible: le modèle d'Heckman

On paramétrise le processus de sélection

$$D = 1(z'\gamma + \varepsilon > 0)$$

avec $(\nu,\ \varepsilon) \mapsto N(0,\ \Sigma)$

et on estime un modèle à deux équations. $\neq$

Problèmes: 1) z doit influencer D mais pas Y et 2) la structure est paramétrique.

Le papier propose une approche différente, par variable instrumentales

Les variables instrumentales: généralités

Classiquement utilisées dans le type de modèle suivant

$y = x'\beta + u, \quad E(u \mid x) \neq 0$

Un instrument z permet d'estimer β sans biais s'il satisfait les conditions

(1) $E(x'z) \neq 0$ (z doit expliquer une partie de x)

(2) $E(u'z) = 0$ (z doit être exogène)

L'utilisation d'un instrument pour corriger la non-réponse

Ici, la non observation de Y n'est pas aléatoire et l'ignorer aboutirait à des estimateurs biaisés.

On cherche un instrument z tel que

(1) $Y=\phi(z, \varepsilon)$ (z doit expliquer en partie la variable d'intérêt)

(2) $D \perp\!\!\!\perp z \mid (Y, x)$ (x des variables de contrôle éventuelles)

Des exemples sur le choix d'un instrument

(1) Le prix de la marijuana sur le marché influe sur la consommation des individus mais probablement pas sur la non réponse.

(2) Des caractéristiques exogènes des acteus sur un marché peuvent influer sur leurs décisions de prendre part à la transaction mais pas sur le prix.

(3) Des résultats scolaires passés peuvent être corrélés aux résultats présents sans directement expliquer le redoublement.

Les hypothèses additionnelles

- ✓ La loi de z est identifiable
- ✓ P(D = 1 |Y) > 0 presque sûrement
- ✓ La distribution de Y est B-complete pour tout z:

$$\Big(E(g(Y)|Z)=0 \quad a.s.\Big) \Longrightarrow \Big(g(Y)=0 \quad a.s.\Big).$$

avec B l'ensemble des fonctions g de Y bornées par valeurs inférieures telles que E(|g(Y|^{q})< $+\infty$.

Estimation (1)

- ✓ Sous ces hypothèses, on a

$$E\left(\frac{D}{Q(Y)}\middle|Z\right)=1$$

- ✓ Trois méthodes en pratique

  - avec support fini : $E\left(\frac{D}{\sum_{k=1}^{s}P(y_k)1\{Y=y_k\}}-1\middle|Z\right)=0$

  - paramétrique $E\left[\left(\frac{D}{F(V'\beta_0)}-1\right)W\right]=0.$

  - non paramétrique

$$T\phi(z)=E(D\phi(Y^*)|Z=z) \qquad Tf=1. \qquad \widehat{T}\phi(z)=\frac{\sum_{i=1}^{n}D_i\phi(Y_i^*)K_{h_n}(z-Z_i)}{\sum_{i=1}^{n}K_{h_n}(z-Z_i)}.$$

Estimation (2)

- ✓ On (re)pondère les individus pour lesquels Y est observés par l'inverse de P(D = 1 | Y)

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} D_i \widehat{f}^{-i}(Y_i^*) g(Y_i^*, Z_i)$$

- ✓ On a

$$\lim_{n\to\infty} E\left(|\widehat{\theta} - \theta|\right) = 0.$$

Simulations

D'après les simulations

- Le convergence et la précision sont sensibles aux choix de fenêtre et de noyau dans le cas non-paramétrique

- L'estimateur paramétrique semble un « bon » compromis (il converge et il est le plus précis).

- Le meilleur estimateur pour P n'est pas forcément le meillleur estimateur pour g(Y).

Une procédure de test pour l'exogénéité de z

- ✓ Elle est fondée sur:
  - le fait que P(Y) est une probabilité, i.e. 0 <= P(Y)<= 1
  - la B-complétude de Y pour z (suridentification du système d'équations estimantes quand Y et Z ont des supports finis).
- ✓ L'idée: z est endogène si l'estimateur n'est pas dans l'intervalle [0; 1] et il suffit de tester cette contrainte en le comparant à un estimateur contraint..

Un relâchement de l'hypothèse d'exogénéité de z

- ✓ On peut remplacer l'hypothèse d'indépendance conditionnelle de D et z par celle de monotonicité: D(-) est une fonction strictement croissante ou décroissante de é.
- ✓ On obtiens alors un encadrement de g(Y) qui permet d'identifier un ensemble auquel elle appartient (« *set identification* »).

Merci de votre attention