Une nouvelle méthode pour résoudre le problème de sélection endogène

Xavier d'Haultfœuille

INSEE-D3E

24 mars 2009

Le problème

Supposons que l'on observe Z et $D \in \{0, 1\}$ mais Y seulement lorsque $D = 1$. Comment faire de l'inférence sur (D, Y, Z) ?

- Problème très courant : non-réponse, modèle de sélection, variables contrefactuelles inobservées, troncature...
- Sans restriction supplémentaire, la distribution des données n'est pas identifiée.
- L'approche la plus usuelle consiste à supposer $Y \perp\!\!\!\perp D \mid Z$ (sélection "ignorable").
- Elle peut être restrictive car elle suppose qu'il n'y a pas de sélection sur inobservables (due par exemple à de l'autosélection basée sur de l'information privée).

Le problème

- L'approache instrumentale usuelle consiste à chercher un instrument Z affectant D et tel que :

$$Y \perp\!\!\!\perp Z \tag{1}$$

- Approche très classique pour les modèles d'offre de travail (cf. Heckman, 1976) ou la littérature sur les effets de traitement (cf. Angrist et Imbens, 1994, Heckman et Vytlacil, 2005...).
- Manski (1994, 2003) montre qu'elle permet d'identifier partiellement la distribution de Y.

L'approche considérée ici

- L'approche instrumentale standard peut être difficile à maintenir lorsque D dépend principalement de Y : si par exemple $D = 1\{Y > \eta\}$, où η est inobservée et $\eta \perp\!\!\!\perp (Y, Z)$.
- Je suppose ici la condition d'exclusion suivante :

$$D \perp\!\!\!\perp Z | Y \tag{2}$$

- Une condition de rang entre Y and Z est également requise.
- L'hypothèse (2) a été également considérée par Chen (2001), Tang et al. (2003), Hemvanich (2004) et Ramalho et Smith (2007). C'est également l'hypothèse qui sous-tend le calage généralisé (cf. Deville, 2002). Les premiers papiers se concentrent sur l'estimation paramétrique de modèles sur (Y, Z), celui de Deville sur l'estimation paramétrique de $P(D = 1|Y)$. Ici je me focalise principalement sur l'identification non-paramétrique de la loi de (D, Y, Z) sous (2).

Exemples

- Non-réponse non-ignorable. Y est observée uniquement lorsque les individus répondent à la question correspondante ($D = 1$). La non-réponse dépend directement de Y.

  Par exemple, dans une enquête sur la drogue, supposons que l'on pose la question "Avez-vous consommé au moins une fois de la drogue au cours du dernier mois ?" La méthode peut être appliquée s'il existe un instrument affectant Y mais pas directement D : le prix de la drogue au niveau local par exemple (si celui-ci est disponible, un proxy de ce dernier sinon).

- Modèle de Roy avec secteur inobservé. Soit par exemple Y le salaire potentiel d'un individu, η son salaire de réserve et D l'indicatrice de participation. Alors $D = 1\{Y > \eta\}$ et on observe Y lorsque $D = 1$. Pour identifier un tel modèle, on utilise d'habitude des variables affectant η mais pas Y. Ici c'est l'inverse. Exemple possible : le taux d'emploi local.

Plan de la présentation

Plan de la présentation

Résultat principal

- La condition de rang entre Y et Z s'écrit en terme de condition de "complétude" : pour toute fonction h bornée inférieurement,

$$E(h(Y)|Z) = 0 \text{ p.s.} \Longrightarrow h(Y) = 0 \text{ p.s.} \tag{3}$$

- Quand Support$(Y) = \{y_1, ..., y_s\}$ et Support$(Z) = \{z_1, ..., z_t\}$, cela revient à supposer

$$\text{rang}(M) = s,$$

où M est la matrice d'élément $P(Y = y_i|Z = z_j)$. Quand (Y, Z) est continue, Newey and Powell (2003) montrent qu'une condition suffisante est que la densité conditionnelle de Y sachant Z soit de type exponentiel. Je montre qu'elle est également satisfaite si

$$Y = \mu\left(\nu(Z) + \varepsilon\right)$$

avec $\nu(.)$ à large support et ε qui satisfait des conditions techniques (principalement que sa fonction caractéristique ne s'annule pas).

Résultat principal

Théorème

Sous les conditions 2 et 3, et si $P(Y) \equiv P(D = 1|Y) > 0$ *alors la distribution de* (D, Y, Z) *est identifiée.*

Idée de la preuve : on peut identifier les solutions $Q(.)$ de

$$E\left(\frac{D}{Q(Y)}\middle|Z\right) = 1 \tag{4}$$

Or $$E\left(\frac{D}{Q(Y)}\middle|Z\right) = E\left(\frac{E(D|Y,Z)}{Q(Y)}\middle|Z\right) = E\left(\frac{P(Y)}{Q(Y)}\middle|Z\right).$$

Donc P est une des solutions de (4). Sous l'hypothèse (3) cette solution est en fait unique, donc P est identifiée. On montre ensuite que cela permet d'identifier la distribution de (D, Y, Z).

Bornes sous des conditions de monotonicité

- Si l'on n'est pas prêt à faire l'hypothèse d'indépendance précédente, on peut la remplacer par la condition plus faible suivante que pour tout (y, z) :

$$\begin{array}{rcl} z' & \mapsto & P(D=1|Y=y, Z'=z') \\ y' & \mapsto & P(D=1|Y=y', Z=z) \end{array} \quad \text{sont croissantes.}$$

  où a priori, $Z' \neq Z$.

- Soit $Q(.)$ une solution de (4), on a pour toute fonction h appartenant à un certain ensemble :

$$E\left(\frac{Dh(Y)}{Q(Y)}\right) \leq E\left(h(Y)\right) \leq E\left[E(h(Y)|Z', D=1)\right].$$

  De plus les deux bornes sont minimales (i. e., peuvent être atteintes) en général.

Bornes sous des conditions de monotonicité

- Par conséquent, on peut obtenir des bornes finies même si $h(Y)$ est non bornée. Résultat similaire à celui de Manski et Pepper (2000), dans un cadre différent.

- La borne inférieure (resp. supérieure) vaut $E(h(Y))$ si $D \perp\!\!\!\perp Z|Y$ (resp. $D \perp\!\!\!\perp Y|Z'$). On peut donc les rapprocher en choisissant Z et Z' de façon appropriée.

Identification paramétrique

- Question : peut-on identifier la distribution de (D, Y, Z) en utilisant seulement $E(Z)$ et non sa distribution complète ? Utile pour des raisons d'implémentation ou de disponibilité de Z.
- On retient une approche correspondant au calage généralisé (Deville, 2002).

Théorème

Supposons la condition 2 vérifiée et $P(D = 1|Y) = F(Y'\beta)$, *où* F *est connue. Alors :*

- β *est identifié localement si*

$$rang(E(DZY'F'(Y'\beta_0)/F^2(Y'\beta_0))) = dim(Y);$$

- β *est identifié globalement si* $E(Z|Y, D = 1) = \Gamma Y$ *où* Γ *est de plein rang.*

Plan de la présentation

Identification

Estimation

Estimation paramétrique

- Soit $Y^* = DY$. On considère ici l'estimation de $P(Y)$ à partir d'un échantillon $((D_1, Y_1^*, Z_1), ..., (D_n, Y_n^*, Z_n))$ de copies indépendantes de (D, Y^*, Z).
- Si $P(D = 1|Y) = F(Y'\beta)$, on peut estimer β par GMM en utilisant :

$$E\left(\frac{DZ}{F(Y^{*\prime}\beta)} - 1\right) = 0$$

- Si (Y, Z) a un support fini, on peut de même s'appuyer sur

$$E\left(\frac{D\mathbb{1}\{Z = z_k\}}{\sum_j P(y_j)\mathbb{1}\{Y^* = y_j\}} - 1\right) = 0 \quad k = 1, ..., t$$

Estimation non-paramétrique

- Si Y et Z sont continues, il s'agit de faire des GMM avec un paramètre fonctionnel (la fonction P). Supposons $(Y, Z) \in [0,1]^2$, soit $f = 1/P$ et notons

$$T : \phi \mapsto E\left(D\phi(Y^*)|Z\right).$$

  Alors il s'agit de résoudre en f l'équation $T(f) = 1$. C'est un "problème inverse mal posé".
- On considère un estimateur basé sur une régularisation de Tikhonov, comme Darolles et al. (2002), Hall et Horowitz (2005) ou Horowitz et Lee (2007). Soit

$$\widehat{T}(\phi)(z) = \frac{\sum_{i=1}^n D_i \phi(Y_i^*) K_{h_n}(z - Z_i)}{\sum_{i=1}^n K_{h_n}(z - Z_i)}$$

  Soit $D_M = \{\phi / M \geq \phi \geq 1 \ p.s.\}$ pour une constante $M < \infty$ et définissons, pour toute fonction g de carré intégrable,

$$||g||^2 = \int g(u)^2 du$$

Estimation non-paramétrique

- L'estimateur de f satisfait

$$\widehat{f} \in \arg\min_{\phi \in D_M} \left\| \widehat{T}(\phi) - 1 \right\|^2 + \alpha_n \|\phi\|^2.$$

où α_n est le paramètre de régularisation.

- Soit $\delta_n = h_n + 1/nh_n$. Supposons que $h_n \to 0$, $\delta_n \to 0$ et $\delta_n/\alpha_n \to 0$, alors

$$\left\| \widehat{f} - f \right\| \xrightarrow{L^2} 0$$

- On peut alors estimer $\theta = E(g(Y, Z))$ par

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

On peut montrer que :

$$\widehat{\theta} \xrightarrow{L^1} \theta.$$

Conclusion

- Le papier développe une stratégie instrumentale inspirée du calage généralisé pour traiter de la sélection endogène. Cette stratégie est particulièrement adaptée lorsque Y affecte directement la sélection.
- L'identification non-paramétrique ponctuelle ou partielle est obtenue sous des conditions d'indépendance ou de monotonicité.
- Le papier présente également les propriétés à distance finie des estimateurs précédents et une application à l'évaluation de l'efficacité du redoublement à l'école primaire.