Biais de sélectivité portant sur des variables quantitatives sur données longitudinales

S Lollivier (*)
(*) Insee, DSDS

Introduction

On sait depuis les années 70 que l'estimation d'une équation sur un sous-échantillon obtenu de façon sélective dans la population peut conduire à des biais (Heckman, 1978). L'exemple typique est celui d'une équation de salaire estimée sur les seules femmes actives, alors même que le comportement d'activité relève d'un arbitrage dans lequel le salaire que la personne peut obtenir sur le marché intervient. On est alors confronté à un système à équations simultanées, dans lequel la première ne peut être estimée que sur un sous-échantillon dépendant d'un régime déterminé par la seconde. Intuitivement, il est facile de concevoir pourquoi l'estimation de l'équation de salaire sans tenir compte de l'équation d'activité conduit en général à des résultats biaisés : dans le sous-échantillon pour lequel la première variable est observée, l'espérance du terme d'erreur conditionnelle à la variable explicative décrite dans la seconde équation n'est plus nulle.

En coupe instantanée, certains auteurs font intervenir une caractérisation totalement paramétrique du système, en supposant la normalité jointe des termes d'erreurs des deux équations, afin d'estimer le système par la méthode du maximum de vraisemblance. Outre le fait que cette technique suppose en général dans la pratique d'avoir recours à une programmation spécifique pour le calcul de la vraisemblance, elle conduit à effectuer un maximum d'hypothèses paramétriques pour rendre compte des comportements. C'est la raison pour laquelle on peut lui préférer une procédure en deux étapes, la première consistant en l'estimation d'un modèle probit portant sur la variable qualitative déterminant la sélection, la seconde en une régression linéaire augmentée d'un terme issu de la première étape. L'estimateur des moindres carrés ordinaires portant sur cette régression est alors convergent et asymptotiquement normal (Heckman, 1978). Il est en outre convergent sous des hypothèses moins restrictives que l'estimateur du maximum de vraisemblance, puisque la normalité jointe des deux perturbations n'est pas nécessairement requise. Aisée à mettre en œuvre, cette méthode ne souffre que de la seule complication pratique consistant à réaliser de façon rigoureuse le test permettant de conclure ou non à l'endogénéité de la sélection, la matrice de variance-covariance des estimateurs n'étant pas celle qui ressort de l'estimation des moindres carrés ordinaires.

Moyennant quelques hypothèses supplémentaires, la méthode en deux étapes peut aisément s'étendre aux modèles de sélection sur données longitudinales. En particulier, les travaux de Wooldridge (1995) montrent qu'il est possible d'estimer des modèles analogues aux modèles à effets individuels sur données de panel, tout en corrigeant d'un processus de sélection éventuellement endogène. Même si le recours à des méthodes de maximisation de la vraisemblance reste possible, tous les travaux récents préfèrent limiter au maximum les hypothèses paramétriques portant sur les lois des termes d'erreur.

Rappel : modèles en coupe avec variable de sélection dichotomique

On utilise le modèle de sélection :

$$d_i^* = z_i g + u_i,$$
$$d_t = \mathbb{1}[d_i^* \geq 0],$$
$$w_i = x_i b + e_i, \text{ observé si } d_i = 1,$$

où la première équation décrit une règle de sélection binaire. Lorsque $d_i = 1$, on observe la variable quantitative w_i qui dépend elle aussi d'un terme linéaire.

Les paramètres à déterminer sont b et g, tandis que x_t et z_t sont des vecteurs de variables explicatives dépendant du temps, et composés de variables éventuellement communes. Enfin, u_i et e_i sont des termes d'erreur inobservés, indépendants en *i*. Ils sont en outre supposés exogènes, c'est à dire que la loi de ces perturbations ne dépend pas de x_t et z_t

L'estimation de ce modèle au moyen d'une procédure en deux étapes suppose seulement la normalité de la perturbation e_i et le fait que conditionnellement aux variables explicatives, la régression de e_i sur u_i est linéaire, ce qui peut s'écrire :

$$e_i = r s_1 u_i + v_i ,$$

avec $E(v_i | x_i, z_i, d_i) = 0$, et $V(v_i | x_1, z_i, d_i) = s_1^2 (1 - r^2)$ comme précédemment. On notera que cette hypothèse est vérifiée en cas de normalité jointe de u_i et e_i, comme conséquence des propriétés des lois normales, v_i étant alors lui-aussi normal. Pour le calcul de l'estimateur en deux étapes, cette hypothèse de normalité de v_i n'est pas nécessaire, ce qui permet d'alléger les contraintes paramétriques nécessaires à l'identification du modèle.

Sous ces hypothèses, on peut montrer que :

$$E(w_i | x_i, z_i, d_i = 1) = x_i b + r s_1 l_i \text{, avec } l_i = \frac{j(z_i g)}{\Phi(z_i g)},$$

soit :

$$w_{i1} = x_i b + r s_1 l_i + v_i ,$$

v_i étant une variable aléatoire d'espérance nulle et de variance $V(v_i) = s_1^2 - s_1^2 r^2 a_i$, avec $a_i = (z_i g) l_i + l_i^2$. Dans la littérature, l_i porte le nom d'inverse du ratio de Mill. Les termes d'erreur v_i sont en outre indépendants entre les observations (Lee, Maddala et Trost, 1980).

Il est alors possible d'estimer l'équation de sélection au moyen d'un modèle probit ; on obtient alors un estimateur convergent de g, qui peut servir à calculer $\hat{l}_i$:

$$\hat{l}_i = \frac{j(z_i \hat{g})}{\Phi(z_i \hat{g})}.$$

D'où la nouvelle équation :

$$w_1 = x_1 b + r s_1 \hat{l} + z_i ,$$

avec $z_i = v_i + r s_1 (l_i - \hat{l}_i)$, z_i étant un terme d'erreur dont l'espérance tend vers zéro lorsque le nombre d'observation tend vers l'infini.

Le fait de remplacer dans la régression un terme inobservé par un terme imputé introduit cependant une complication entraînant une corrélation entre les termes d'erreur des différentes observations. Ceci ne gêne en rien le calcul de l'estimateur des moindres carrés ordinaires, qui est convergent et asymptotiquement normal. Ceci complexifie en revanche le calcul de sa matrice de variance-covariance asymptotique (Lee, Maddala et Trost, 1980) .

Cette procédure en deux étapes (probit puis mco) permet par conséquent d'obtenir des estimateurs convergents des paramètres d'intérêt et de leur matrice de variance-covariance asymptotique**.** L'estimateur des mco étant asymptotiquement normal, il est alors possible d'effectuer des tests

asymptotiques, notamment le test du biais de sélectivité. Une possibilité est de calculer un estimateur de cette matrice de variance-covariance en utilisant les formules de Lee, Maddala et Trost (1980). Un autre moyen, plus économe en programmation, est de recourir à des techniques de bootstrap. La méthode la plus simple à mettre en œuvre ici est sans doute celle du bootstrap par paires. Elle consiste à considérer l'ensemble des variables expliquées et explicatives pour un individu donné $((w_1, d_i), (x_i, z_i))$ comme un tirage aléatoire dans une distribution multidimensionnelle (Efron, Tibshirani, 1986). L'algorithme consiste alors à réaliser B tirages aléatoires indépendants **avec remise** de N individus dans l'échantillon d'origine, lui même de taille N. La mise en œuvre de la méthode décrite précédemment fournit un estimateur de tous les paramètres pour chaque tirage ($b = 1,..., B$). Quand B tend vers l'infini, la matrice de variance-covariance empirique des estimateurs bootstrap est un estimateur sans biais de la matrice de variance-covariance des paramètres d'intérêt[1].

Processus de sélection sur données longitudinales

L'objectif est ici de proposer une généralisation des modèles de sélection sur données longitudinales. De la même façon qu'en coupe transversale, l'observation d'un mécanisme peut être perturbée par un processus de sélection corrélé avec celui-ci. Ceci se traduit par des biais si l'on procède à une estimation séparée du mécanisme sans tenir compte de la sélection. Un premier moyen de prendre en compte ces biais serait d'estimer un modèle bivarié par la méthode du maximum de vraisemblance. Mais avec une variable d'intérêt continue, un certain nombre de travaux récents permettent de construire des estimateurs corrigeant les biais, avec des hypothèses moins contraignantes sur les lois des termes d'erreur, surtout concernant la partie quantitative du modèle. Les méthodes d'estimation sont en outre beaucoup moins coûteuses en termes de programmation et de temps de calcul que la maximisation d'une vraisemblance faisant intervenir plusieurs termes d'erreur corrélés entre eux.

Le modèle en niveau de Wooldridge (1995) propose un jeu d'hypothèses permettant de généraliser aux données de panel le modèle d'Heckman. Ceci conduit au modèle suivant pour $i = 1,..., N$, et $t = 1,..., T$:

$$d_{it}^* = z_{it}\gamma + h_i + u_{it},$$
$$d_{it} = 1[d_{it}^* \geq 0],$$
$$w_{it} = x_{it}\beta + a_i + e_{it}, \text{ observé si } d_{it} = 1.$$

La première équation décrit à nouveau une règle de sélection binaire, qui dépend d'un terme linéaire, et d'un effet individuel inobservé et constant dans le temps, mais qui peut cette fois être corrélé avec les variables explicatives. Lorsque $d_{it} = 1$, on observe la variable quantitative w_{it} qui dépend elle aussi d'un terme linéaire et d'un terme individuel inobservé le cas échéant corrélé avec les régresseurs.

Les paramètres à déterminer sont β et γ, tandis que x_{it} et z_{it} sont des vecteurs de variables explicatives dépendant du temps, et composés de variables éventuellement communes. a_i et h_i sont des effets individuels invariants au cours du temps, qui peuvent être corrélés avec les variables explicatives. Enfin, u_{it} et e_{it} sont des termes d'erreur inobservés, indépendants en *i* et en *t*. Ils sont eux-mêmes strictement exogènes et indépendants des a_i et des h_i. Dans l'exposé, les variables x_{it} et z_{it} sont supposées évoluer au cours du temps. Le formalisme s'étend aisément au cas de variables explicatives invariantes dans le temps.

Wooldridge effectue quatre hypothèses (W1 à W4) afin de procéder à la détermination d'estimateurs convergents de β.

[1] A noter que la méthode du bootstrap peut également être utilisée pour construire la loi des paramètres, notamment à distance finie. On ne détaillera pas cette possibilité, l'ouvrage ayant pour parti pris de ne s'intéresser qu'aux propriétés asymptotiques.

W1 : La régression de h_i sur z_i est linéaire.

Wooldridge reprend ici la spécification de Chamberlain (1984) concernant la dépendance entre l'effet individuel et les variables $z_i = (z_{i1},...,z_{iT})$:

$$h_i = z_i d + c_i \text{ , soit, } h_i = \sum_{t=1}^{T} z_{it} d_t + c_i ,$$

où c_i est un terme aléatoire. Avec cette hypothèse, le terme d'erreur c_i est strictement exogène. Pour simplifier l'estimation, et éviter l'introduction de tous les régresseurs présents, passés et futurs, on pourrait également adopter une spécification « à la Mundlak (1978) », selon laquelle la régression ne fait intervenir que les moyennes temporelles $z_{i.}$ des z_{it}, et non la totalité des z_i.

W2 : Le terme d'erreur de l'équation de sélection $v_{it} = c_i + u_{it}$ est normal et son espérance conditionnelle à (x_i, z_i) est nulle.

Cette hypothèse permet d'estimer l'équation de sélection.

W3 : La régression de a_i sur $x_i = (x_{i1},...,x_{iT})$ et v_{it} est linéaire.

On s'intéresse cette fois à l'équation portant sur la variable quantitative. L'hypothèse effectuée permet d'écrire :

$$E(a_i \mid x_i, v_{it}) = x_i y + f_t v_{it} = \sum_{t=1}^{T} x_{it} y_t + f_t v_{it} .$$

A nouveau, plutôt que de prendre compte les valeurs passées, présentes et futures des x_{it}, on pourrait se restreindre à leur moyenne temporelle $x_{i.}$, en adoptant la formulation de Mundlak (1978), reprise par Nijman et Verbeck (1992), et Zabel (1992).

W4 : L'espérance du terme d'erreur e_{it} conditionnelle à (x_i, z_i) et à v_{it} ne dépend pas de (x_i, z_i) et est linéaire en v_{it}.

En d'autres termes,

$$E(e_{it} \mid x_i, z_i, v_{it}) = E(e_{it} \mid v_{it}) = r_t v_{it} .$$

Cette formulation est moins restrictive, tout en aboutissant à une forme analogue à celle qui consiste à supposer la normalité jointe entre e_{it} et v_{it}.

Au total, en combinant W3 et W4,

$$\begin{aligned} E(a_i + e_{it} \mid x_i, z_i, d_{it} = 1) &= E(a_i \mid x_i, z_i, d_{it} = 1) + E(e_{it} \mid x_i, z_i, y_i, d_{it} = 1) \\ &= w_i y + (f_t + r_t) E(v_{it} \mid x_i, z_i, y_i, d_{it} = 1) \\ &= x_i y + (f_t + r_t) l_{it}. \end{aligned}$$

où l_{it} est l'inverse d'un ratio de Mill obtenu à partir de l'équation de sélection :

$$l_{it} = \frac{j(\frac{z_{it}g + z_i d}{s_t})}{\Phi(\frac{z_{it}g + z_i d}{s_t})},$$

où s_t est l'écart-type de v_{it}. Si on note $w_t = f_t + r_t$, il en résulte pour le sous-échantillon sur lequel la variable w_{it} est observée :

$$w_{it} = x_{it}b + x_i y + w_t l_{it} + e_{it},$$

avec

$$E(e_{it} | x_i, z_i, d_{it} = 1) = 0.$$

La procédure d'estimation proposée par Wooldridge est alors la suivante :

(i) Estimer, période par période, et par la méthode du maximum de vraisemblance une équation de sélection sous forme réduite :

$$d_{it}^* = z_i d_t + v_{it}, \text{ ou } d_{it}^* = \sum_{t=1}^{T} z_{it} d_{tt} + v_{it}$$

et en déduire un estimateur convergent $\hat{l}_{it} = \frac{j(z_i \hat{d}_t)}{\Phi(z_i \hat{d}_t)}$ de chacun des l_{it}. Ce modèle est en fait un sur-modèle de celui décrit par l'hypothèse W1. Les paramètres sont en effet moins contraints. Le fait d'estimer cette équation période par période ne permet donc pas de déterminer des estimateurs de g et d sans calcul complémentaire.

(ii) Estimer l'équation :

$$w_{it} = x_{it}b + x_i y + w_t l_{it} + e_{it}$$

par les moindres carrés ordinaires sur l'échantillon empilé, après avoir substitué $\hat{l}_{it}$ à l_{it}. L'estimateur obtenu pour (b, y, w) est convergent et asymptotiquement normal pour *N* infini. Wooldridge fournit également une forme analytique de la matrice de variance-covariance des estimateurs.

On notera toutefois qu'il suppose implicitement dans sa procédure d'estimation les termes d'erreur v_{it} indépendants en *i* et en *t*, puisqu'il procède à l'ajustement d'un modèle probit séparé à chacune des dates $t = 1, ..., T$. Cette démarche se justifie par le fait que cette régression n'est qu'instrumentale, et qu'il cherche seulement à obtenir des estimateurs convergents des paramètres afin de corriger du biais de sélectivité dans l'équation quantitative.

Une autre façon de procéder consiste à estimer le modèle de sélection en tant que tel. Ce modèle a besoin d'une contrainte identifiante, $s_t = 1$ pour tout t. On peut alors montrer que l'estimateur probit simple sur données empilées est un estimateur convergent de b et d, quelle que soit la corrélation temporelle entre les termes d'erreur $c_i + v_{it}$[2]. On peut ensuite recourir à une méthode de bootstrap

[2] On peut montrer qu'il s'agit d'un estimateur du pseudo-maxmum de vraisemblance du modèle de sélection sur données longitudinales.

par paires analogue à celles décrite dans la section précédente pour obtenir une estimation de la matrice de variance-covariance des paramètres (b, y, w)[3].

Références

Efron B., et R. Tibshirani (1986): « Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy », *Statistical Science*, 1, n°1, 54-77.
Heckman J. (1978): « Dummy endogenous variables in a simultaneous equation system », *Econometrica*, 46, 931-959.
Lee L-F, Maddala G., et Trost R. (1980) : « Asymptotic covariance matrices of two-stage probit and two-stage tobit methods for simultaneous equations models with selectivity », *Econometrica*, 48, 491-503.
Mundlak Y. (1978) : « On the pooling of time series and cross section data », *Econometrica*, 46, n°1, 69-85.
Nijman T. et Verbeek M. (1992) : « Nonresponse in panel data: the impact on estimates of a life cycle consumption function », *Journal of Applied Econometrics*, 7, 243- 257
Wooldridge J. M. (1995) : « Selection corrections for panel data models under conditional mean independence assumptions », *Journal of Econometrics*, 68, 115-132.
Zabel J.E.(1992) : « Estimating fixed effects and random effects with selectivity », *Economics Letters*, 40, 269-272.

[3] Voire même utiliser cette technique de bootstrap pour simuler la loi des paramètres à distance finie.