Une macro SAS d'estimation du modèle de sélection par la méthode d'Heckman*

Xavier d'Haultfœuille

4 septembre 2009

Résumé

Cette note présente la macro `%heckit` qui implémente sous SAS la méthode d'estimation en deux étapes d'Heckman (1979) du modèle de sélection.

1 Survol de la théorie

On considère le modèle suivant :

$$\left\{ \begin{array}{rcl} Y & = & X'\beta + \varepsilon \\ D & = & \mathbb{1}\{Z'\gamma + \eta \geq 0\} \end{array} \right. , \quad (\varepsilon, \eta) \perp\!\!\!\perp (X, Z).$$

où Y n'est observé que lorsque $D = 1$. L'exemple classique, développé par Gronau (1974) et Heckman (1974) est le modèle d'offre de travail (féminin souvent). En effet, les salaires potentiels ne sont observés que si l'individu se porte sur le marché du travail (i.e., $D = 1$, où D est l'indicatrice d'activité). Si l'on ignore ce problème de sélection en estimant β par l'estimateur des MCO de Y sur X sur l'échantillon des actifs, on introduit un biais en général. En effet, se porter sur le marché du travail est un choix qui est a priori corrélé aux salaires que l'on s'attend à recevoir (i.e., η est en général corrélé à ε).

Pour corriger ce biais, Heckman (1979) a proposé une méthode en deux étapes, basée sur l'hypothèse ci-dessous :

H1. $\eta \sim \mathcal{N}(0, 1)$ et $\varepsilon = \eta\delta + \nu$ avec $E(\nu|\eta) = 0$.

Sous l'hypothèse H1, on a :

$$\begin{array}{rcl} E(Y|D = 1, X, Z) & = & X'\beta + E(\eta|\eta \geq -Z'\gamma)\delta + E(\nu|X, Z, \eta \geq -Z'\gamma) \\ & = & X'\beta + \lambda(Z'\gamma)\delta \end{array}$$

où $\lambda(x) = \varphi(x)/\Phi(x)$, φ (resp. Φ) étant la densité (resp. fonction de répartition) d'une loi normale standard. Pour estimer β on procède alors comme suit :

1. Estimation de γ par un probit de D sur Z ;

*Je remercie Romain Aeberhardt pour ses conseils et sa relecture.

2. Estimation de β et δ par MCO de Y sur X et $\lambda(X'\widehat{\gamma})$, sur les individus tels que $D = 1$.

Cette procédure est très simple à implémenter. En revanche, les écarts-types des estimateurs de β et δ sont plus délicats à obtenir, du fait de deux problèmes. Tout d'abord, même si ν est homoscédastique, ce que nous supposerons par la suite[1], le modèle est hétéroscédastique par construction. En effet, on a

$$Y = X'\beta + \lambda(Z'\gamma)\delta + \nu + \zeta,$$

avec $\zeta = (\eta - \lambda(Z'\gamma))\delta$, et l'on peut montrer que

$$V(\zeta|D = 1, X, Z) = \delta^2 \left[1 + \lambda(Z'\gamma)(-Z'\gamma - \lambda(Z'\gamma))\right].$$

Le deuxième problème est que l'on ne régresse pas Y sur $(X, \lambda(Z\gamma))$, mais sur $(X, \lambda(Z\widehat{\gamma}))$. L'erreur commise sur γ en première étape affecte la précision des estimateurs de deuxième étape. Ainsi, même en utilisant la matrice de variance de White, l'estimateur obtenu sera non convergent. Un estimateur convergent de la variance de $(\widehat{\beta}, \widehat{\delta})$ est décrit par Greene (1995), section 22.4.3[2]. C'est cet estimateur qui est calculé dans la macro `%heckit` décrite maintenant.

2 La macro

L'objectif de la macro `%heckit` est de calculer $\widehat{\gamma}, \widehat{\beta}, \widehat{\delta}$ et leur précision. La syntaxe de la macro est la suivante. Les instructions facultatives sont indiquées par $< . >$, et les notations mathématiques utilisées ci-dessous sont celles de la section précédente.

```
heckit(table=, <librairie= >, var_y = , var_selection = , x_eq_y = ,
       x_eq_selection = , <table_sortie_coeff = >, <table_sortie_cov = >);
```

- `table` correspond au nom de la table d'entrée.
- `librairie` est le nom de la librairie dans laquelle est stockée la table. Par défaut, `librairie` = work.
- `var_y` est le nom de la variable Y.
- `var_selection` est le nom de la variable D (comme précédemment, $D = 1$ lorsque Y est renseigné, $D = 0$ sinon).
- `x_eq_y` est la liste (sans virgule) des variables X.
- `x_eq_selection` est la liste (sans virgule) des variables Z.
- `table_sortie_coeff` est le nom de la table stockant les coefficients $(\widehat{\beta}, \widehat{\delta})$, leurs écarts-types, leurs statistiques de test et les p-value associées. Par défaut, aucune table n'est créée.
- `table_sortie_cov_beta` est le nom de la table stockant la matrice de variance covariance de $(\widehat{\beta}, \widehat{\delta})$. Par défaut, aucune table n'est créée.

[1] Par hypothèse, ν, comme fonction de (ε, η), est indépendant de (X, Z) et donc $V(\nu|X, Z)$ ne dépend pas de (X, Z). Mais comme l'on se restreint ici aux individus tels que $D = 1$, on doit considérer $V(\nu|X = x, Z = z, D = 1)$, égal à $V(\nu|\eta \geq -z'\gamma)$ et qui dépend donc en général de z. Pour éviter cette complication on suppose ici que $V(\nu|\eta)$ est indépendant de η.

[2] Une erreur apparaît dans cette section, dans la 5ème édition de l'ouvrage de Greene. Il faut lire $\widehat{\delta}_i = \widehat{\lambda}_i(\widehat{\lambda}_i + \mathbf{w}_i'\widehat{\gamma})$ et non $\widehat{\delta}_i = \widehat{\lambda}_i(\widehat{\lambda}_i - \mathbf{w}_i'\widehat{\gamma})$.

2.1 Les sorties

La macro renvoie dans l'output les résultats du probit de première étape, puis $(\widehat{\beta}, \widehat{\delta})$ accompagné de ses écart-types, de la statistique de test de l'hypothèse nulle $\widehat{\beta}_k = 0$ (ou $\widehat{\delta} = 0$) et de la p-value correspondante. Enfin, elle affiche les estimateurs de $\rho = \text{Corr}(\varepsilon, \eta)$ et $\sigma = \sqrt{V(\varepsilon)}$.

2.2 Remarques

1. Il est possible d'estimer avec la `proc qlim` le modèle de sélection précédent par maximum de vraisemblance. Par rapport à l'estimateur du maximum de vraisemblance, l'intérêt d'utiliser la procédure d'Heckman en deux étapes est double. D'une part, l'hypothèse H1 qui assure la convergence de cet estimateur est moins forte que l'hypothèse de normalité jointe de (ε, η) nécessaire à la convergence du maximum de vraisemblance. La méthode d'Heckman est donc plus robuste que le maximum de vraisemblance. D'autre part, la vraisemblance du modèle est délicate à maximiser, et la `proc qlim` peut être beaucoup plus lente à tourner que la macro `%heckit`.

2. Il existe également un programme SAS, écrit par D. Jaeger, implémentant la méthode d'Heckman en deux étapes[3]. L'inconvénient de ce programme est qu'il est plus lent que la macro `%heckit`. Il peut même s'interrompre pour cause de manque de mémoire si la table SAS d'entrée est trop grande : typiquement, quand (nombre d'observations) $\times$ (nombre de variables explicatives) $\geq 2 \times 10^7$.

3. La macro crée et supprime des tables SAS préfixées par temp_table_. Il est donc préférable d'éviter de créer dans la libraire de travail des tables ayant un tel préfixe. Par ailleurs, la macro crée (dans des tables temporaires) les variables `xgamma`, `lambda`, `delta`, `temp_poids` et des variables préfixées par `temp_var`. Il faut donc éviter que l'une des variables utilisée dans la macro ait ce nom.

Références

Greene, W. H. (1995), *Econometric Analysis*, Prentice Hall.

Gronau, R. (1974), 'Wage comparisons : A selectivity bias', *Journal of Political Economy* **82**, 1119–1149.

Heckman, J. J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica* **42**, 679–693.

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica* **47**, 153–161.

[3]Ce programme, bien qu'officieux, est disponible sur le site internet d'aide SAS. Tel qu'il apparaît dans ce site, le code nécessite deux modifications. D'une part, il est conseillé d'enlever la commande finale endsas qui termine la session SAS. D'autre part, il faut rajouter un "T" à "INTERCEP" à la ligne 284.