Endogénéité dans un système d'équations normal bivarié avec variables qualitatives

Journées de Méthodologie Statistique
2002

Pourquoi traiter l'endogénéité ?

- Problème analogue à l'estimation d'une équation séparée dans un système d'équations simultanées
  - aboutit en général à des estimateurs biaisés
  - car les équations ne sont pas indépendantes
- Variables expliquée et explicatives observées et continues : méthode des variables instrumentales et des régressions augmentées (Robin, 2000).
- Présence de variables qualitatives :
  - situation plus complexe, en général EMV
  - méthodes conditionnelles dans certains cas

Rappel variables continues

- Modèle bivarié

$$\begin{cases} y_{1i} = X_{1i}\beta_1 + \gamma\, y_{21} + u_{1i} \\ y_{2i} = X_{2i}\beta_2 + u_{2i} \end{cases}$$

- Avec termes d'erreurs normaux

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \rightarrow N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right]$$

- De sorte que :

$$u_{1i} = \rho\frac{\sigma_1}{\sigma_2}u_{2i} + v_{1i} \qquad \text{et } v_{1i} \text{ normal}$$

Rappel variables continues

- On peut alors réécrire :

$$y_{1i} = X_{1i}\beta_1 + \gamma\, y_{2i} + \rho\frac{\sigma_1}{\sigma_2}(y_{2i} - X_{2i}\beta_2) + v_{1i}$$

- Cause du biais

$$E(y_{1i} / X_{1i}, X_{2i}, y_{2i}) = X_{1i}\beta_1 + \gamma\, y_{2i} + \rho\frac{\sigma_1}{\sigma_2}(y_{2i} - X_{2i}\beta_2)$$

- Estimation de la seconde équation par les mco :

$$y_{1i} = X_{1i}\beta_1 + \gamma y_{2i} + \rho\frac{\sigma_1}{\sigma_2}\hat{u}_{2i} + v_{1i} + \rho\frac{\sigma_1}{\sigma_2}(u_{2i} - \hat{u}_{2i})$$

Rappel variables continues

- Sur grand échantillon, on peut négliger la variance due à l'imputation, soit le terme d'espérance nulle :

$$\rho \frac{\sigma_1}{\sigma_2}(u_{2i} - \hat{u}_{2i})$$

- Et donc aboutir à l'équivalent de la régression augmentée de la méthode des variables instrumentales, avec des tests à distance finie (contrepartie des hypothèses)

Variable expliquée qualitative

- Le même modèle, mais latent :

$$\begin{cases} y_{1i}^{*} &= X_{1i}\beta_1 + \gamma\, y_{21} + u_{1i} \\ y_{2i} &= X_{2i}\beta_2 + u_{2i} \end{cases}$$

- Conduit à la même équation latente :

$$y_{1i}^{*} = X_{1i}\beta_1 + \gamma\, y_{2i} + \rho\frac{\sigma_1}{\sigma_2}\hat{u}_{2i} + v_{1i} + \rho\frac{\sigma_1}{\sigma_2}(u_{2i} - \hat{u}_{2i})$$

- Sur grand échantillon, on peut négliger le terme du à l'imputation, et estimer le modèle ci-dessus comme un modèle « probit augmenté » avec les mêmes tests, notamment l'endogénéité.

Variable expliquée qualitative

- Autre possibilité, maximiser la log-vraisemblance :

$$L_i = Z_{1i}\left[\frac{1}{\sqrt{1-\rho^2}}\left(X_{i1}\beta_1 + \gamma y_{2i} + \frac{\rho}{\sigma_2}(y_{2i} - X_{2i}\beta_2\right)\right]\frac{1}{\sigma_2}\varphi\left(\frac{y_{2i} - X_{2i}\beta_2}{\sigma_2}\right)$$

- Avec

$Z_{1i}(x) = \Phi(x)$ si 1, et $Z_{1i}(x) = 1 - \Phi(x)$ si 0

- Mais cette méthode requiert une programmation spécifique.

Variable explicative qualitative

- On a cette fois le modèle :

$$\begin{cases} y_{1i} & = & X_{1i}\beta_1 + \alpha d_i + u_{1i} \\ y_{2i}^{*} & = & X_{2i}\beta_2 + u_{2i} \end{cases}$$

- Qui peut être réécrit de la façon suivante :

$$y_{1i} = X_{1i}\beta_1 + \alpha d_i + \rho\sigma_1\mu_i + v_{1i}$$

- Où le terme correctif est une sorte de ratio de Mill, et v_{1i} est une variable aléatoire d'espérance nulle conditionnellement à d_i, mais malheureusement non normale
- Plus précisément :

Variable explicative qualitative

$$\mu_i = \lambda_i d_i + \tilde{\lambda}_i (1 - d_i) \qquad \lambda_i = \frac{\varphi(X_{2i}\beta_2)}{\Phi(X_{2i}\beta_2)} \qquad \tilde{\lambda}_i = \frac{\varphi(X_{2i}\beta_2)}{1 - \Phi(X_{2i}\beta_2)}$$

- Ces valeurs peuvent être estimées grace à la deuxième équation. Si on néglige une nouvelle fois la variance due à l'imputation, on aboutit à l'équation :

$$y_{1i} = X_{1i}\beta_1 + \alpha d_i + \rho\sigma_1 \hat{\mu}_i + v_{1i}$$

- L'estimateur des mco est sans biais, mais le terme d'erreur non normal est hétéroscédastique.

Variable explicative qualitative

- La méthode de White permet d'estimer la matrice de variance-covariance de cet estimateur :
- Calculer les résidus de l'estimation par les mco
- Utiliser les carrés de ces résidus comme des estimateurs des termes diagonaux de la matrice de variance-covariance de Ω
- En déduire un estimateur de la matrice de variance-covariance de l'estimateur des mco :

$$(X'_1 X_1)^{-1}(X'_1 \tilde{\Omega} X_1)(X'_1 X_1)^{-1}$$

- Conduire des test asympotiques

Variable explicative qualitative

- Extension aisée à des variables explicatives polytomiques ordonnées (remplacer l'indicatrice unique par des indicatrices des différentes tranches)
- Autre possibilité : utiliser l'estimateur du maximum de vraisemblance

$$L_i = Z_{2i}\left[\frac{1}{\sqrt{1-\rho^2}}\left(X_{21}\beta_2 + \frac{\rho}{\sigma_2}(y_{1i} - X_{1i}\beta_1 - \alpha d_i)\right)\right]\frac{1}{\sigma_1}\varphi\left(\frac{y_{1i} - X_{1i}\beta_1 - \alpha d_i}{\sigma_1}\right)$$

Deux variables qualitatives

- Pas de méthode conditionnelle facile à mettre en œuvre (à ma connaissance)
- Nécessité de réaliser un minimum de programmation ; vraisemblance du modèle bivarié :

$$L_i = \iint_{Y_{1i}^* \in D(Y_{i1});Y_{2i}^* \in D(Y_{2i})} \exp\left(-\frac{U_{1i}^2 + U_{2i}^2 - 2\rho U_{1i}U_{2i}}{2(1-\rho^2)}\right) dY_{1i}^* dY_{2i}^*$$

$$U_{1i} = (Y_{1i}^* - X_{1i}\beta_1 - \alpha d_i)/\sigma_1 \quad U_{2i} = (Y_{2i}^* - X_{2i}\beta_2)/\sigma_2$$

- L'expression fait intervenir des intégrales doubles

Deux variables qualitatives

- Astuce pour ne faire intervenir qu'une intégrale simple :

$$u_1 = \rho_1 v + \varepsilon_1 \qquad u_2 = \rho_2 v + \varepsilon_2$$

- La vraisemblance se réécrit :

$$l_i = \log\left[\int Z_{1i}\left(\frac{X_{1i}\beta_1 + \alpha d_i + \rho_1 v}{s_1}\right) Z_{2i}\left(\frac{X_2\beta_{21} + \rho_2 v}{s_2}\right) dv\right]$$

- Avec des restrictions $s_1 = 1$, $s_2 = \rho_2 = \frac{1}{\sqrt{2}}$
- L'intégrale se calcule au moyen de techniques de quadrature gaussienne