UMR 8184

Thema

THÉORIE ÉCONOMIQUE,
MODÉLISATION ET APPLICATIONS

Econométrie des Données de Panel avec Variables Dépendantes Limitées

Aristide E. Houndetoungan

28 Octobre 2021

Variables dépendantes limitées (VDL)

- Dans un modèle économétrique, la variable dépendante est dite limitée si sa distribution ne peut être estimée en dessous ou au-delà d'une certaine valeur.
- Exemples :
  - Variables binaires (individu vacciné ou non),
  - Variables ordonnées (niveau de satisfaction d'un service : pas du tout satisfait, moyennement satisfait, très satisfait),
  - Variables multinomiales (moyens de transport utilisés : marche, vélo, transport en commun, voiture personnelle, autres),
  - Variables censurées (durée passée au chômage frictionnel par les diplômés de l'année passée),
  - Variables de comptage (nombre d'accidents de circulation).
- Dans plusieurs applications, les chercheurs ignorent la nature « limitée » de la variable dépendante et modélisent cette dernière par le modèle linéaire-en-moyennes. Une telle approche est inefficace.

Maximum de vraisemblance

- Permet de prendre en compte la nature limitée de la variable dépendante.
- Principe :
  - Hypothèse sur la distribution de $\mathbf{y}_{i(1:T)} = \{y_{i1}, \ldots, y_{iT}\}$ conditionnellement à $\mathbf{x}_{i,(1:T)} = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}\}$,
  - L'estimateur de maximum de vraisemblance (MV) consiste à trouver les paramètres de la distribution qui maximise la probabilité des réalisations observées, $\{\mathbf{y}_{1(1:T)}, \ldots, \mathbf{y}_{N(1:T)}\}$.

- Exemple :

- Soit $\mathbf{y}_{1:N} = y_1, \ldots, y_N$ des données binaires (prenant la valeur 0 ou 1).
- On postule que $y_i \overset{iid}{\sim} \mathcal{B}$ernoulli$(\theta)$.
- Probabilité de réalisation de y_i,

$$f(y_i;\theta) = \theta^{y_i}(1-\theta)^{1-y_i}, \tag{1}$$

- Probabilité des réalisations observées (encore appelée vraisemblance),

$$L(\mathbf{y}_{1:N};\theta) = \prod_{i=1}^{N} f(y_i;\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}. \tag{2}$$

- Maximisation de la vraisemblance. Maximiser une quantité sous une forme de produits est souvent complexe avec des solutions numériquement instables. On maximise alors le logarithme de la vraisemblance.

- Avec le logarithme, les produits deviennent des sommes,

$$\log L(\mathbf{y}_{1:N};\theta) = \sum_{i=1}^{N} y_i \log(\theta) + (1-y_i)\log(1-\theta), \tag{3}$$

$$\hat{\theta}_{\text{MV}} = \arg\max_{\theta} \left(\sum_{i=1}^{N} y_i \log(\theta) + (1-y_i)\log(1-\theta) \right).$$

Conditions de premier ordre : $\dfrac{\partial \log L(\mathbf{y}_{1:N};\theta)}{\partial\theta} = 0,$

$$\Longrightarrow \sum_{i=1}^{N} \left(\frac{y_i}{\theta} - \frac{1-y_i}{1-\theta} \right) = 0 \Longrightarrow \sum_{i=1}^{N} \frac{y_i(1-\theta) - \theta(1-y_i)}{\theta(1-\theta)} = 0,$$

$$\Longrightarrow \frac{\sum_{i=1}^{N} (y_i(1-\theta) - \theta(1-y_i))}{\theta(1-\theta)} = 0 \Longrightarrow \sum_{i=1}^{N} (y_i(1-\theta) - \theta(1-y_i)) = 0,$$

$$\Longrightarrow \sum_{i=1}^{N} (y_i - y_i\theta - \theta + y_i\theta) = 0 \Longrightarrow \sum_{i=1}^{N} y_i - N\theta,$$

$$\Longrightarrow \hat{\theta}_{\text{MV}} = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

- Dans un modèle économétrique, l'hypothèse $y_i \overset{iid}{\sim} \mathcal{B}\text{ernoulli}(\theta)$ est très restrictive. Cette distribution dépend d'autres variables explicatives, de plus de paramètres.
- En panel, la distribution dépend aussi des effets individuels pour prendre en compte l'hétérogénéité entre les individus.
- Les effets individuels peuvent être fixes ou aléatoires.

Effets fixes - Effets aléatoires

- Supposons que la probabilité de réalisation de y_{it} est $f(y_{it}, \mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta})$. Avec l'hypothèse que les y_{it} sont iid, la vraisemblance des données de l'individu i peut s'écrire comme,

$$\ell(\mathbf{y}_i; \alpha_i, \boldsymbol{\beta}) = \prod_{t=1}^{T} f(y_{it}, \mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}). \tag{4}$$

- Aussi, si les individus sont indépendants, le log de la vraisemblance de l'ensemble des donnée est,

$$\log L(\mathbf{y}; \alpha_{1:N}, \boldsymbol{\beta}) = \log \left(\prod_{i=1}^{N} \ell(\mathbf{y}_i; \alpha_i, \boldsymbol{\beta}) \right) = \sum_{i=1}^{N} \log \left(\ell(\mathbf{y}_i; \alpha_i, \boldsymbol{\beta}) \right). \tag{5}$$

- Comme dans un modèle linéaire, α_i peut être traité comme étant des effets fixes ou des effets aléatoires. Les effets fixes impliquent que $\mathbb{C}\mathbf{ov}(\alpha_i, \mathbf{x}_{it}) \neq 0$ tandis que les effets aléatoires impliquent que $\mathbb{C}\mathbf{ov}(\alpha_i, \mathbf{x}_{it}) = 0$.

- En panel, il est toujours important de se débarrasser des effets individuels pour **éviter le problème des paramètres incidents**.
- Lorsque les effets sont aléatoires, les chercheurs supposent généralement que $\alpha_i | \mathbf{x}_{it} \overset{iid}{\sim} \mathcal{N}\left(\bar{\alpha}, \sigma_\alpha^2\right)$. En effet, bien que cette condition soit une hypothèse supplémentaire, elle implique quand même que $\mathbb{C}\mathbf{ov}(\alpha_i, \mathbf{x}_{it}) = 0$. Elle s'aligne donc avec l'hypothèse des effets aléatoires.
- Sous l'hypothèse des effets aléatoires, on peut écrire une vraisemblance non conditionnelle à α_i. Ainsi, la vraisemblance de l'individu i devient,

$$\ell(\mathbf{y}_i; \boldsymbol{\beta}, \sigma_\alpha^2) = \int_{\mathbb{R}} \prod_{t=1}^{T} f(y_{it}, \mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}) \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left\{ \frac{1}{2\sigma_\alpha^2} \alpha_i^2 \right\} d\alpha_i. \tag{6}$$

Puisqu'on intègre sur l'effet individuel, $\ell(\mathbf{y}_i; \boldsymbol{\beta}, \sigma_\alpha^2)$ ne dépend plus de α_i.

- La vraisemblance de l'ensemble des données ne dépend plus aussi de α_i.

$$\log L(\mathbf{y};\boldsymbol{\beta},\sigma_\alpha^2) = \sum_{i=1}^{N} \log\left(\ell(\mathbf{y}_i;\boldsymbol{\beta},\sigma_\alpha^2)\right),$$

$$\log L(\mathbf{y};\boldsymbol{\beta},\sigma_\alpha^2) = \sum_{i=1}^{N} \log\left(\int_{\mathbb{R}} \prod_{t=1}^{T} f(y_{it},\mathbf{x}_{it},\alpha_i,\boldsymbol{\beta}) \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left\{\frac{1}{2\sigma_\alpha^2}\alpha_i^2\right\} d\alpha_i\right)$$

  Pour la plupart des modèles, le maximum $\log L(\mathbf{y};\boldsymbol{\beta},\sigma_\alpha^2)$ est implémenté dans tous les logiciels économétriques.

- Lorsque les effets sont fixes, il est extrêmement difficile de spécifier la distribution de $\alpha_i|\mathbf{x}_{it}$. On ne pourra donc pas se débarrasser facilement des effets individuels.
- Avec des VDL, les modèles à effets fixes sont généralement non convergents. Le problème des paramètres incidents n'admet pas toujours de solution efficace (**des exceptions s'appliquent**).
- Les chercheurs supposent souvent que les effets sont aléatoires (même si cette hypothèse est elle aussi très forte).

Variables binaires

- Probabilité de $y_{it} = 1$, conditionnellement à $\mathbf{x}_{it}$, $\boldsymbol{\beta}$ et α_i.

$$F\left(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i\right) = \begin{cases} \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) & \text{modèle Probit,} \\ \Lambda(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) & \text{modèle Logit,} \end{cases} \tag{7}$$

avec Φ et Λ les fonctions de répartition respectives des lois normale et logistique. La vraisemblance des données de l'individu i peut s'écrire comme,

$$\ell(\mathbf{y}_i; \alpha_i, \boldsymbol{\beta}) = \prod_{t=1}^{T} F\left(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i\right)^{y_{it}} \left(1 - F\left(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i\right)\right)^{1-y_{it}}. \tag{8}$$

- Lorsque les effets sont aléatoires, on peut écrire la vraisemblance non conditionnellement à α_i (donc ne dépend pas de α_i).

$$\log L(\mathbf{y}; \boldsymbol{\beta}, \sigma_\alpha^2) = \sum_{i=1}^{N} \log\left(\ell(\mathbf{y}_i; \boldsymbol{\beta}, \sigma_\alpha^2)\right), \text{ avec}$$

$$\ell(\mathbf{y}_i; \boldsymbol{\beta}, \sigma_\alpha^2) = \int_{\mathbb{R}} \ell\left(\mathbf{y}_i; \alpha_i, \boldsymbol{\beta}\right) \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left\{\frac{1}{2\sigma_\alpha^2}\alpha_i^2\right\} d\alpha_i.$$

- Lorsque les effets sont fixes, la vraisemblance dépend de α_i :

$$\log L(\mathbf{y}; \alpha_{1:N}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{t=1}^{T} F\left(y_{it} | \mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i\right)^{y_{it}} \left(1 - F\left(y_{it} | \mathbf{x}_{it}, \boldsymbol{\beta}, \alpha_i\right)\right)^{1-y_{it}}.$$

- Problème des paramètres incidents.
- Il n'y a pas de solution pour le modèle probit (il ne converge pas lorsque les effets sont fixes).
- Pour le modèle Logit, Chamberlain montre que la vraisemblance conditionnellement à $\sum_{t=1}^{T} y_{it}$ ne dépend pas de α_i. La vraisemblance de Chamberlain est implémentée dans les logiciels et peut être utilisée pour estimer $\boldsymbol{\beta}$ de manière convergente.

- La solution de Chamberlain peut être également généralisée vers le modèle ordonné et le modèle multinomial. Les modèles logit ordonné et logit multinomial avec effets fixes peuvent être estimés de manière convergente avec la vraisemblance de Chamberlain. La spécification probit à effets fixes est non convergente.

Applications avec R

- Données binaires : script `binaires.R`
- Données ordonnées : script `ordonees.R`
- Données multinomiales : script `mnomiales.R`