Les données de panel en 50mn

Thierry Magnac

IDEI & Université des Sciences Sociales, Toulouse

Enquêtes longitudinales

Autres que françaises …

- Emploi: NLSY, GSOEP, Panel européen
- Consommation: CEX
- Revenus: PSID
- Patrimoine, épargne: HRS, ELSA
- Panels de pays: Penn Tables

Avantages des enquêtes longitudinales/ coupes transversales

Premier argument: **Identification**

Exemple: En coupe transversale, on ne peut pas distinguer les effets d'âge et de génération, ce que l'on peut faire avec des données de panel.

On montre que les salaires ont leur maximum à un âge de 50-55 ans.

Avantages (suite)

Deuxième argument: **Contrôle de l'hétérogénéité inobservable**

Exemple: Des variables omises, constantes au cours du temps, expliquent à la fois la variable dépendante et les variables indépendantes.

Dans des fonctions de salaire, les effets mesurés des variables sectorielles sont très différents en coupe transversale et en panel.

Avantages (fin)

Troisième argument: *Dynamique*

Certains phénomènes ne peuvent s'expliquer qu'en fonction de leur passé.

La demande de travail des entreprises à chaque période dépend de l'effectif à la période précédente (*i.e.* coûts d'ajustement).

Coûts: Collecte & Traitement

- **Attrition**: Perte par mobilité et non réponse des primo-répondants.

Coûteux d'augmenter le taux de rétention (à 10 ans il est néanmoins de 90% dans NSLY79, voir site Web)

Exemple: Enquête Santé anglaise, le coût peut être multiplié par 10 pour un individu mobile.

- **Traitement** des données incomplètes

Plan

- *Equations statiques:* à chaque période, une relation entre une dépendante et des indépendantes.
- *Equations dynamiques:* relation entre une dépendante et son passé.
- *Données incomplètes et coupes transversales répétées.*

Cadre: Analyse causale (Voir Endogénéité, Jean-Marc Robin)

- **Paramètres d'intérêt**: Impact de variations exogènes des variables explicatives
- Principe de l'isolement d'une cause en contrôlant toutes les autres causes possibles (y compris celles qui sont inobservables comme des effets individuels constants au cours du temps).

Equation statique: Les salaires en fonction de l'éducation et de l'expérience

$$y_{it} = x_{it}\beta + \varepsilon_{it}$$

$$= x_{it}\beta + \alpha_i + u_{it}$$

N grand, T petit: des effets temporels sont toujours inclus parmi les x_{it}.

Paramètres d'intérêt: β

Hypothèses: Effets individuels fixes

$$x_i = (x_{i1}, .., x_{iT})$$

$$H_1 : Ex_i' u_{it} = 0, Eu_{it} = 0$$

$$H_2^a : u_i \text{ et } u_j \text{ sont indépendants}$$

$$H_2^b : E(u_i u_i' \mid x_i) = \sigma_u^2 . I_T$$

Identification

$$x_{it} = (x_{it}^{(1)}, x_i^{(2)})$$

$$y_{it} = x_{it}^{(1)}\beta_1 + x_i^{(2)}\beta_2 + \alpha_i + u_{it}$$

$$= x_{it}^{(1)}\beta + \alpha_i^{(2)} + u_{it}$$

et $(0, \alpha_i^{(2)})$ n'est pas distinguable de (β, α_i)

L'hétérogénéité qu'elle soit observable ou non est traitée de la même façon.

Dimensions inter et intra-individuelles

$$y_{i.} = \frac{1}{T} \sum_{t=1}^{T} y_{it} \text{ (dimension "between")}$$

$$y_{it} - y_{i.} \text{ (dimension "within")}$$

Pour le vecteur empilé, on définit des opérateurs
B = between, W = within

Modèle empilé

$$Y = X\beta + \sum_{i=1}^{N} \alpha_i E_i + U$$

$$\Leftrightarrow \begin{cases} WY = WX\beta + WU \\ BY = BX\beta + \sum_{i=1}^{N} \alpha_i E_i + BU \end{cases}$$

Orthogonalité des 2 dimensions (si cylindré!),

$$WB = 0$$

Estimation

Trois arguments:

- La dimension interindividuelle n'apporte aucune information sur le paramètre d'intérêt.
- Les deux dimensions sont orthogonales
- Il n' y a pas de contrainte entre les paramètres des deux dimensions

Alors le meilleur estimateur (BLUE) est celui qui utilise la dimension « **within** ».

Effets individuels aléatoires

- X_1 variables au cours du temps

donc WX_1 est de plein rang

- X_2 constantes dans le temps

$$BX_2 = X_2$$

$$Y = X_1\beta_1 + X_2\beta_2 + \sum_{i=1}^{N} \alpha_i E_i + U$$

On suppose aussi:

$$H_3 : E\alpha_i = 0, E\alpha_i u_{it} = 0$$

Corrélation entre effets individuels/variables explicatives

- *Aucune:* Modéle à erreurs composées
- *Non restreinte:* Effets fixes ou effets individuels « corrélés ».
- *Partiellement restreinte:* Hausman & Taylor (81), Breusch, Mizon & Schmidt (89)
- *Cadre général:* Arellano & Bover (95)

Modèle de Mundlak

Projection linéaire

$$\alpha_i = x_{i.}^{(1)}\theta_1 + x_i^{(2)}\theta_2 + v_i = z_i\theta + v_i$$

de telle façon que $E(z_i' v_i) = 0$.

$$Y = X_1\beta_1 + X_2\beta_2 + \sum_{i=1}^{N} \alpha_i E_i + U$$

$$= WX_1\beta_1 + BX_1(\beta_1 + \theta_1)$$

$$+ BX_2(\beta_2 + \theta_2) + \sum_{i=1}^{N} v_i E_i + U$$

Estimation

$$Y = WX_1\beta_1 + BX_1\gamma_1 + BX_2\gamma_2 + \tilde{\varepsilon}$$

et donc:

$$\begin{cases} WY = WX_1\beta_1 + W\tilde{\varepsilon} \\ BY = BX_1\gamma_1 + BX_2\gamma_2 + B\tilde{\varepsilon} \end{cases}$$

Estimateur Within est **BLUE**

Modèle à erreurs composées

Pas de corrélation entre α_i et $x_{it}^{(1)}$ et $x_i^{(2)}$

$\Rightarrow \theta_1, \theta_2$ sont nuls $\Rightarrow \gamma_1 = \beta_1$

Une contrainte entre paramètres des 2 dimensions

a deux conséquences:

i. Test de cette contrainte

ii. Gain en précision si elle est satisfaite (estimateur MCG ou MCQG)

Recherche de spécification

1. Distinguer les variables, variables au cours du temps et constantes au cours du temps
2. Formuler les « conditions d'orthogonalité »: certaines corrélations entre erreurs et variables sont nulles.
3. Etudier l'identification des paramètres d'intérêt.
4. Estimer et tester les hypothèses de départ si suridentification. Si rejet, revenir au point 2.

Autres points

- Hétéroscédasticité?
- Autocorrélation?

Si absence de structure précise, corriger en se servant de méthodes à la White.

Dynamique

Deux formes canoniques :

$$1.\ y_{it} = \alpha y_{it-1} + x_{it}\beta + v_i + u_{it}$$

$$u_{it} \text{ est } MA(q)$$

$$2.\ y_{it} = x_{it}\beta + v_i + u_{it}$$

$$u_{it} \text{ est } ARMA(p,q)$$

$$\Rightarrow \theta(L)y_{it} = \theta(L)x_{it} + \theta(1)v_i + \theta(L)u_{it}$$

$$\text{où } \theta(L)u_{it} \text{ est } MA(q)$$

Différences Statique/Dynamique

Biais: Les estimateurs utilisés dans un cadre statique sont **biaisés** dans un cadre dynamique si le nombre de périodes T est fixe (en particulier, l'estimateur de la covariance).

Argument: L'estimateur Within ne corrige pas de la corrélation entre une variable explicative (la variable dépendante retardée) et le terme d'erreur qui vient de la présence d'un effet individuel.

Principe d'estimation convergente

Traiter les deux problèmes de présence d'effets individuels et d'endogénéité

1. On construit les premières différences:

$$\Delta y_{it} = y_{it} - y_{it-1}$$

2. On instrumente l'équation résultante par la variable en niveau décalée deux fois

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta u_{it}$$

$$E(y_{it-2} \Delta u_{it}) = 0$$

(car u_{it} est supposé bruit blanc)

Conditions initiales

Argument : Il y a corrélation entre l'effet individuel et la variable dépendante à la première période.

Deux approches :

1. $E(y_{i0}v_i) = \omega_0$ (Paramètre de contrôle)

2. $y_{i0} = \frac{v_i}{1-\alpha} + \sum_{\tau=0}^{-\infty} \alpha^\tau u_{i\tau} = \frac{v_i}{1-\alpha} + \varepsilon_{i0}$

Le processus est "initialisé en $-\infty$.

$$y_{it} = x_{it}\beta + v_i + u_{it}$$

Extensions

- Ecrire toutes les conditions d'orthogonalité et utiliser la méthode généralisée des moments (GMM).
- Utiliser une approche par vraisemblance pour utiliser les conditions d'orthogonalité les plus informatives.
- Utiliser des approximations « T est grand » au lieu de « N est grand » et/ou des conditions sur N/T pour les approximations asymptotiques

Données incomplètes

Plusieurs hypothèses:

- *Données manquantes au hasard*: Toutes les méthodes précédentes s'appliquent.
- *Sélection sur les exogènes*: Renforcer les « conditions d'orthogonalité » en « indépendance en moyenne » au lieu de non corrélation.
- *Sélection générale*: voir modèles non linéaires

Pseudo-panels: Utiliser des coupes transversales répétées

Idée: Construire dans chaque coupe des individus fictifs par agrégation que l'on suit au cours du temps

Cohortes d'âge, d'éducation, sexe et toute autre variable constante.

Puis, utiliser la linéarité du modèle pour agréger les modèles individuels.

Un nouveau voyage?

- Hétérogénéité des coefficients & non stationarité dans les panels
- Les modèles non linéaires de panel (et en particulier ceux qui servent à corriger l'attrition).