

Économétrie des données de panel

Thomas Chuffart

thomas.chuffart@univ-fcomte.fr

Économétrie des données de panel

Informations générales

- 18h de cours - Généralement le mercredi matin
- Moodle - Clef : ie_econometrics - Coming soon
- Pré-requis : MCO, GMM, notions de convergence
- Pré-requis 2 : R à jour avec Rstudio, package plm
- https://github.com/archuff/cours_econometrics_R
- Examen écrit
- Projet : groupes de 2-3

Économétrie des données de panel

Le projet

Vous avez deux choix :

- Vous répliquez un papier existant (attention à la disponibilité des données)
 - 5 points sur la forme
 - 5 points réplique des résultats
 - 5 points code R
 - 5 critiques et extensions
 - 3 points si extensions intéressantes
- Construction de votre base de données :
 - 5 points sur la forme
 - 5 points originalité de la question économique
 - 5 points résultats
 - 5 points code R

Économétrie des données de panel

Le projet

- Quelle sont les questions de recherche à laquelle vous souhaitez répondre ?
- Pourquoi cette question vous semble-t-elle intéressante ?
- Quels travaux académiques ou issus d'instituts statistiques avez-vous consultés pour définir votre question de recherche ?
- Quelles données allez-vous utiliser ?

Introduction

Definition (Panel dataset)

Un jeu de données longitudinal, ou panel, suit un échantillon d'individus sur plusieurs périodes. Un individu a plusieurs observations pour la même variable.

Introduction

Exemple :

Id	Time	X_1	...	X_k
1	2012	10	...	-0.5
1	2013	12	...	-0.3
1	2014	10	...	1.2
2	2012	10	...	-0.5
2	2013	12	...	-0.3
2	2014	10	...	1.2
...
N	2012	10	...	-0.5
N	2013	12	...	-0.3
N	2014	10	...	1.2

Introduction

Notations :

- Individu : pays, région, état, personne etc. $i = 1, \dots, N$
- Série temporelle : séquence prise à des moments successifs équidistants. $t = 1, \dots, T$

Introduction

Definition (Micro-panel data set)

La dimension temporelle T est beaucoup plus petite que la dimension individuelle N . Exemple : The University of Michigan's Panel Study of Income Dynamics, PSID (15000 individus depuis les années 70)

Definition (Macro-panel data set)

La dimension temporelle est similaire à la dimension individuelle. Exemple : Les états américains sur les 20 dernières années.

Introduction

Definition (Balanced panel)

Un panel est dit *balanced* (équilibré) si il y a le même nombre de périodes pour chaque individus. Pour un panel *unbalanced*, la dimension temporelle T est spécifique à chaque individu.

Remarque :

- Les methodes économétriques d'un panel *unbalanced* et *balanced* sont similaires
- Il faut par contre se demander pourquoi le panel est *unbalanced*

Introduction

Quels sont les avantages des données de panel, à quoi servent-elles ?

- Hsiao, C., (2003), Analysis of Panel Data, second edition, Cambridge University Press.
- Mátyás and Sevestre, (2008), The Econometrics of Panel Data, Handbook, third edition, Springer

Introduction

Les avantages des modèles et des données de panel :

- 1 Beaucoup de données ($N \times T$), donc plus de degrés de liberté
 \Rightarrow Moins de collinéarité entre les variables explicatives
- 2 ATTENTION : Plus de données \Rightarrow Plus d'hétérogénéité
- 3 Ouvre de nouvelles perspectives d'études économiques.
- 4 La nature des données de panel permet d'isoler des effets de traitements, de politiques publiques.
- 5 Les données de panel peuvent résoudre des problèmes économétriques classiques. Ex : présence de variables non observées qui sont corrélées avec des variables explicatives.

Introduction

Exemple (Ben-Porath (1973))

Si dans des données en coupe transversale de femmes mariées nous trouvons qu'elles participent au marché du travail à 50% :

- 1ère interprétation : Chaque femme dans une population homogène a 50% de chance d'être sur le marché du travail quelque que soit l'année
- 2ème interprétation : Cela peut aussi impliquer que 50% des femmes dans une population hétérogène travaillent toujours et que 50% ne travaillent jamais.

Pour trouver quelle interprétation est bonne, les données individuelles sur plusieurs années sont utiles.

Introduction

Example

Soit le modèle de régression suivant :

$$y_{it} = \alpha + \beta' x_{it} + \gamma' z_{it} + \varepsilon_{it} \quad i = 1, \dots, N \quad t = 1, \dots, T \quad (1)$$

$$y_{it} = \alpha + \beta' x_{it} + u_{it} \quad (2)$$

- x_{it} ($K_1 \times 1$) et z_{it} ($K_2 \times 1$) sont des vecteurs de variables exogènes, z_{it} non observées
- α est une constante, β ($K_1 \times 1$) et γ ($K_2 \times 1$) sont des vecteurs de paramètres
- $\varepsilon_{it} \sim IID(0, \sigma_\varepsilon^2)$, $u_{it} = \gamma' z_{it} + \varepsilon_{it}$
- $Cov(x_{it}, z_{it}) \neq 0$

Introduction

Pour résoudre ce problème :

- Hypothèse : $z_{it} = z_i$, i.e. les valeurs de z varient entre les individus mais restent constantes dans le temps
(Caractéristiques individuelles par exemple)

$$y_{it} = \alpha + \beta' x_{it} + \gamma' z_i + \varepsilon_{it} \quad i = 1, \dots, N \quad t = 1, \dots, T$$

$$y_{it} - y_{i,t-1} = \beta'(x_{it} - x_{i,t-1}) + \varepsilon_{it} - \varepsilon_{i,t-1}$$

L'estimation par MCO est désormais sans biais et consistante pour β .

Introduction

Pour résoudre ce problème :

- Hypothèse : $z_{it} = z_t$, i.e. facteurs communs

$$y_{it} = \alpha + \beta' x_{it} + \gamma' z_i + \varepsilon_{it} \quad i = 1, \dots, N \quad t = 1, \dots, T$$

$$y_{it} - \bar{y}_t = \beta'(x_{it} - \bar{x}_t) + \varepsilon_{it} - \bar{\varepsilon}_t$$

Avec \bar{y}_t la moyenne empirique des individus sur une période.
L'estimation par MCO est désormais sans biais et consistante.

Introduction

Exemple Pooled OLS :

$$\ln \text{CDS}_{it} = \alpha + \beta_1 \text{SM}_{it} + \beta_2 \text{FX}_{it} + \varepsilon_{it}$$

Data : Sovereign CDS spreads and their determinants

Estimation sans prendre en compte l'hétérogénéité

```
library(plm)
```

```
## Loading required package: Formula
```

```
load('panel_psid_exemple1.RData')
```

```
summary(plm(log(wage) ~ educ + experience, data = df, index = c("pid", "year"),  
          model = "pooling"))
```

```
## Pooling Model
```

```
##
```

```
## Call:
```

```
## plm(formula = log(wage) ~ educ + experience, data = df, model = "pooling",  
##     index = c("pid", "year"))
```

```
##
```

```
## Balanced Panel: n = 2084, T = 4, N = 8336
```

```
##
```

```
## Residuals:
```

	Min.	1st Qu.	Median	3rd Qu.	Max.
##	-10.842813	-0.316410	0.077228	0.454271	4.745900

```
##
```

```
## Coefficients:
```

##		Estimate	Std. Error	t-value	Pr(> t)
##	(Intercept)	8.75821402	0.06090874	143.792	< 2.2e-16 ***
##	educ	0.12262350	0.00424355	28.896	< 2.2e-16 ***
##	experience	0.02374722	0.00099522	23.861	< 2.2e-16 ***
##	---				

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## Total Sum of Squares:      6765.9
```

```
## Residual Sum of Squares:  5790.3
```

```
## R-Squared:                0.14419
```

```
## Adj. R-Squared: 0.14398
```

```
## F-statistic: 701.979 on 2 and 8333 DF, p-value: < 2.22e-16
```

Estimation en prenant en compte l'hétérogénéité

```
summary(plm(log(wage) ~ educ + experience, data = df, index = c("pid", "year"),  
          model = "within"))
```

```
## Oneway (individual) effect Within Model  
##  
## Call:  
## plm(formula = log(wage) ~ educ + experience, data = df, model = "within",  
##      index = c("pid", "year"))  
##  
## Balanced Panel: n = 2084, T = 4, N = 8336  
##  
## Residuals:  
##      Min.      1st Qu.      Median      3rd Qu.      Max.  
## -6.792246 -0.117163  0.011723  0.149364  2.722799  
##  
## Coefficients:  
##              Estimate Std. Error t-value Pr(>|t|)  
## educ           0.0397879  0.0092493  4.3017 1.721e-05 ***  
## experience 0.0187873  0.0012910 14.5525 < 2.2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Total Sum of Squares:    1502.8  
## Residual Sum of Squares: 1448.3  
## R-Squared:              0.036252  
## Adj. R-Squared:        -0.28525  
## F-statistic: 117.55 on 2 and 6250 DF, p-value: < 2.22e-16
```

Introduction

Plan du cours :

- Chapitre 1 : Modèles linéaires et hétérogénéité
- Chapitre 2 : Modèles de panel dynamique
- Chapitre 3 : Sélection de modèle

Introduction

$$y_i = \beta' x_i + \varepsilon_i$$

- Certains facteurs pour un individu particulier peuvent manquer.
- La variable économique y est supposée distribuée selon la loi de probabilité $P(Y|\theta)$
- θ est identique pour tous les individus ce qui rend l'hypothèse sur le DGP pas très réaliste.

Introduction

Definition (Biais d'hétérogénéité)

Si les effets individuels ou temporelles spécifiques à chaque individu sont ignorés comme dans les données en coupe transversale et série temporelle, l'estimation des paramètres peut-être biaisée et pose un problème de spécification.

⇒ Il faut donc contrôler par ces effets individuels non-observables.

Introduction

Example

Soit une fonction de production Cobb-Douglas avec deux facteurs observée pour N entreprises sur T périodes :

- y_{it} est le logarithme de la production pour l'entreprise i à la date t
- n_{it} est le logarithme du nombre d'employés pour l'entreprise i à la date t
- k_{it} est le logarithme du stock de capital pour l'entreprise i à la date t

$$y_{it} = \alpha_i + \beta_i k_{it} + \gamma_i n_{it} + \varepsilon_{it} \quad \varepsilon_{it} \sim IID(0, \sigma_\varepsilon^2)$$

Introduction

Example

Dans cette spécification, α_i et β_i sont spécifiques à chaque entreprise. Mais d'autres spécifications peuvent être considérées :

$$y_{it} = \alpha + \beta k_{it} + \gamma n_{it} + \varepsilon_{it} \quad \varepsilon_{it} \sim IID(0, \sigma_\varepsilon^2)$$

$$y_{it} = \alpha_i + \beta k_{it} + \gamma n_{it} + \varepsilon_{it} \quad \varepsilon_{it} \sim IID(0, \sigma_\varepsilon^2)$$

Modèles linéaires et hétérogénéité

Chapitre 1 - Outline

- 1 **Tests de spécification**
- 2 Effets linéaires non-observés
- 3 Effets fixes
- 4 Effets aléatoires
- 5 Tests de spécification
- 6 Panel hétérogènes : coefficients aléatoires

Modèles linéaires et hétérogénéité

1. Tests de spécification

Le modèle linéaire usuel pour étudier des effets quantitatifs ou qualitatifs est défini par :

$$y_{it} = \alpha_{it} + \beta'_{it}x_{it} + \varepsilon_{it} \quad i = 1, \dots, N \quad t = 1, \dots, T \quad (3)$$

- α_{it} (1×1) et $\beta_{it} = (\beta_{1it}, \beta_{2it}, \dots, \beta_{Kit})$ ($1 \times K$) sont des vecteurs de paramètres qui varient en fonction de i et de t ,
- $x_{it} = (x_{1it}, x_{2it} \dots, x_{Kit})'$ ($K \times 1$) est un vecteur de variables exogènes,
- ε_{it} (1×1) est le terme d'erreur.

Modèles linéaires et hétérogénéité

1. Tests de spécification

On peut appliquer des restrictions sur ce modèle :

- Homogénéité de la pente
- Homogénéité de la constante
- Stabilité temporelle des coefficient : nous posons comme hypothèse que le paramètre sont constants sur l'horizon temporel mais peuvent varier entre les individus.

Modèles linéaires et hétérogénéité

Tests de spécification

Definition (Panel homogène/hétérogène)

- Un modèle de panel homogène est un modèle où tous les paramètres sont identiques pour chaque individu
- Un modèle de panel hétérogène est un modèle où tous les paramètres varient pour chaque individu

Modèles linéaires et hétérogénéité

1. Tests de spécification

On étudie donc le modèle :

$$y_{it} = \alpha_i + \beta'_i x_{it} + \varepsilon_{it} \quad i = 1, \dots, N \quad t = 1, \dots, T$$

- Modèle à effets individuels :

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it} \quad (4)$$

- Modèle de panel homogène (pooled)

$$y_{it} = \alpha + \beta' x_{it} + \varepsilon_{it} \quad (5)$$

Modèles linéaires et hétérogénéité

1. Tests de spécification

1er test :

$$H_0 : y_{it} = \alpha + \beta' x_{it} + \varepsilon_{it} \quad (6)$$

$$H_0 : \alpha_i = \alpha \quad \text{et} \quad \beta_i = \beta \quad \forall i \in [1, N]$$

$$H_1 : \exists (i, j) \in [1, N], \quad \alpha_i \neq \alpha_j \quad \text{ou} \quad \beta_i \neq \beta_j$$

Lemma

Sous l'hypothèse que les $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$, les tests de Fisher peuvent être utilisés.

Modèles linéaires et hétérogénéité

1. Tests de spécification

- Sous H_1 , il y a NK β à estimer et N constantes
- La somme des carrés résidus du modèle sous l'alternative suit donc une χ^2 à $NT - NK - N$ DDL.

Definition

Sous la null, $H_0 : \beta = \beta_i \quad \alpha = \alpha_i \quad \forall i \in [1, N]$ la statistique de Fisher F s'écrit :

$$F = \frac{(SCR_c - SCR)/(NK + N - K - 1)}{SCR/(NT - NK - N)} \quad (7)$$

Modèles linéaires et hétérogénéité

1. Tests de spécification

- Sous H_1 , la SCR est égale à la somme des N résidus provenant des N régressions individuelles :

$$SCR = \sum_{i=1}^N SCR_i = \sum_{i=1}^N \hat{\varepsilon}_{it}$$

- Sous H_0 , On retrouve une SCR classique :

$$SCR = \sum_{i=1}^N \sum_{t=1}^T \hat{\varepsilon}_{it}$$

- Si on ne rejette pas l'hypothèse nulle, le modèle à utiliser est un modèle de panel homogène,

Modèles linéaires et hétérogénéité

1. Tests de spécification

Deuxième test :

$$H_0 : y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it}$$

$$H_0 : \beta_i = \beta \quad \forall i \in [1, N]$$

$$H_1 : \exists (i, j) \in [1, N], \quad \beta_i \neq \beta_j$$

Modèles linéaires et hétérogénéité

1. Tests de spécification

- Sous H_1 , il y a NK β à estimer et N α
- La somme des carrés résidus du modèle sous l'alternative suit donc une χ^2 à $NT - NK - N$ DDL.
- Si on rejette l'hypothèse nulle, le modèle à utiliser est un modèle de panel hétérogène,

Definition

Sous la null, $H_0 : \beta = \beta_i \quad \forall i \in [1, N]$ la statistique de Fisher F s'écrit :

$$F = \frac{(SCR_c - SCR)/(NK - K)}{SCR/(NT - NK - N)} \quad (8)$$

Modèles linéaires et hétérogénéité

1. Tests de spécification

Troisième test :

$$H_0 : y_{it} = \alpha + \beta' x_{it} + \varepsilon_{it}$$

$$H_0 : \alpha_i = \alpha \quad \forall i \in [1, N]$$

$$H_1 : \exists (i, j) \in [1, N], \quad \alpha_i \neq \alpha_j$$

Modèles linéaires et hétérogénéité

1. Tests de spécification

- Sous H_1 , il y a N α à estimer et K β .
- La somme des carrés résidus du modèle sous l'alternative suit donc une χ^2 à $NT - NK$ DDL.

Definition

Sous la null, $H_0 : \alpha = \alpha_i \quad \forall i \in [1, N]$ la statistique de Fisher F s'écrit :

$$F = \frac{(SCR_c - SCR)/(N - 1)}{SCR/(NT - N - K)} \quad (9)$$

Modèles linéaires et hétérogénéité

Outline

- 1 Tests de spécification
- 2 **Effets linéaires non-observés**
- 3 Effets individuels ou aléatoires ?
- 4 Effets fixes
- 5 Effets aléatoires
- 6 Tests de spécification
- 7 Panel hétérogènes : coefficients aléatoires

Modèles linéaires et hétérogénéité

2. Effets linéaires non-observés

Definition

Un modèle de panel linéaire à effets individuels non observés est défini par :

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it} \quad (10)$$

- α_i est un scalaire ;
- β est un vecteur $K \times 1$;
- x_{it} est un vecteur $K \times 1$;
- $\varepsilon_{it} \sim IID$ avec $E(\varepsilon_{it}) = 0$ et $E(\varepsilon_{it}^2) = \sigma_\varepsilon^2$.

Modèles linéaires et hétérogénéité

2. Effets linéaires non-observés

Les α_i vont être notre "problème", selon le type de spécification adoptée ils peuvent être :

- des effets fixes individuels ou effets non-observés
- des variables latentes (modèle à effets aléatoires)

Modèles linéaires et hétérogénéité

2. Effets linéaires non-observés

On peut récrire le modèle en écriture vectorielle et/ou matricielle.
Notation :

$$y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{iT} \end{pmatrix} \quad x_i = \begin{pmatrix} x_{1,i1} & x_{2,i1} & \dots & x_{K,i1} \\ x_{1,i2} & x_{2,i2} & \dots & x_{K,i2} \\ \dots & \dots & \dots & \dots \\ x_{1,iT} & x_{2,iT} & \dots & x_{K,iT} \end{pmatrix}$$

$$e = \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix} \quad \varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \dots \\ \varepsilon_{iT} \end{pmatrix}$$

Modèles linéaires et hétérogénéité

2. Effets linéaires non-observés

Definition

Un modèle de panel linéaire à effets individuels non observés est défini par :

$$y_i = e\alpha_i + x_i\beta + \varepsilon_i \quad (11)$$

- α_i est un scalaire ;
- e est un vecteur $T \times 1$;
- β est un vecteur $K \times 1$;
- x_i est une matrice $T \times K$;
- y_i est un vecteur $T \times 1$;
- $E(\varepsilon_i) = 0$ et $E(\varepsilon_i\varepsilon_i') = \sigma_\varepsilon^2 I_T$.

Modèles linéaires et hétérogénéité

2. Effets linéaires non-observés

On peut aussi stacker tous ces vecteurs pour obtenir la version matricielle. Notation :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} \quad X = \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_N \end{pmatrix}$$

$$\tilde{e} = \begin{pmatrix} e_T & 0_T & \dots & 0_T \\ 0 & e_T & \dots & 0_T \\ \dots & \dots & \dots & 0_T \\ 0_T & 0_T & \dots & e_T \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{pmatrix}$$

Modèles linéaires et hétérogénéité

2. Effets linéaires non-observés

Definition

Un modèle de panel linéaire à effets individuels non observés est défini par :

$$Y = \tilde{e}\tilde{\alpha} + X\beta + \varepsilon \quad (12)$$

- $\tilde{\alpha}$ est un vecteur $N \times 1$;
- \tilde{e} est une matrice $NT \times N$;
- β est un vecteur $K \times 1$;
- X est une matrice $NT \times K$;
- Y est un vecteur $NT \times 1$;

Modèles linéaires et hétérogénéité

2. Effets linéaires non-observés

Comment modéliser les α_i ?

- Vieille école : Effets fixes (paramètre à estimer) vs effets aléatoires (variable aléatoire)
- Wooldridge (2010) : Mauvaise approche. Si on a beaucoup d'individus, cela fait sens de traiter les α_i comme des tirages aléatoires.
- Mundlack (1978) : Est-ce que les α_i sont corrélés avec les variables explicatives ?

$$\text{cov}(x_{it}, \alpha_i) = 0, \quad \forall t, \quad \forall i$$

Modèles linéaires et hétérogénéité

2. Effets linéaires non-observés

Definition

Le terme effet fixe ne signifie pas que les α_i sont traités comme déterministe. On admet qu'il puisse exister de la corrélation entre les α_i et les x_{it} .

- Wooldridge (2010) ne se réfère jamais aux termes *effet fixe* ou *effet aléatoire*
- Nous l'utiliserons tout de même pour présenter deux méthodes d'estimation connues sous ces dénominations.

Modèles linéaires et hétérogénéité

2. Estimation des effets fixes

On considère le modèle :

$$y_i = e\alpha_i + X_i\beta + \varepsilon_i, \quad \forall i = 1, \dots, N \quad (13)$$

- α_i est modélisé comme un terme constant satisfaisant $E(\alpha_i | x_{i1}, \dots, x_{iT}) = 0$
- Le terme d'erreur ε_{it} est iid $\forall i, t$ avec :
 - $E(\varepsilon_{it}) = 0$
 - $E(\varepsilon_{it}\varepsilon_{is}) = \sigma_\varepsilon^2$ si $t = s$, 0 sinon (écriture vectorielle ?)
 - $E(\varepsilon_{it}\varepsilon_{js}) = 0$, $\forall (t, s)$ (écriture vectorielle ?)

- 1 Tests de spécification
- 2 Effets linéaires non-observés
- 3 **Effets fixes**
- 4 Effets aléatoires
- 5 Tests de spécification
- 6 Panel hétérogènes : coefficients aléatoires

Modèles linéaires et hétérogénéité

3. Modèle à effets fixes

Theorem

Sous les hypothèses précédentes, l'estimateur par MCO est BLUE

Definition

Dans ce contexte, l'estimateur des MCO de $\hat{\beta}$ est appelé LSDV.

Modèles linéaires et hétérogénéité

3. Modèle à effets fixes

L'estimateur par MCO des α_i et β sont obtenus via le programme d'optimisation :

$$\left\{ \hat{\alpha}_i, \hat{\beta}_{LSDV} \right\} = \underset{}{\operatorname{argmin}} \sum_{i=1}^N \varepsilon_i' \varepsilon_i \quad (14)$$

- Quels sont les conditions de première ordre ?
- A partir des CPO, trouver $\hat{\beta}_{LSDV}$.

Modèles linéaires et hétérogénéité

3. Modèle à effets fixes

Definition

Sous les hypothèses posées plus haut, l'estimateur à effets fixes de β est défini par :

$$\hat{\beta}_{LSDV} = \left(\sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) \right) \quad (15)$$

Modèles linéaires et hétérogénéité

3. Modèle à effets fixes

- $\hat{\beta}$ ne nécessite pas l'estimation des α_j .
- Il faut simplement transformer les variables explicatives et la variable expliquée en les centrant sur les moyennes de chaque coupe transversale
- Cette transformation revient à appliquer l'opérateur que l'on appellera *Within*

Modèles linéaires et hétérogénéité

3. Modèle à effets fixes

On note W_N cet opérateur qui est une matrice idempotent :

$$W = I_T - B \quad (16)$$

avec B l'opérateur *Between* :

$$B = e_T(e_T' e_T)^{-1} e_T' = \frac{e_T e_T'}{T} \quad (17)$$

Modèles linéaires et hétérogénéité

3. Modèle à effets fixes

- Cet opérateur permet de transformer une variable en lui retirant sa moyenne individuelle,
- Les observations individuelles sont mesurées via les déviations à la moyenne individuelle.

$$Wy_i = (I_T - \frac{1}{T}e_T e_T')y_i = y_i - \frac{1}{T} \sum_{i=1}^N y_i e = y_i - \bar{y}_i e \quad (18)$$

Ou de façon matricielle :

$$W_N Y = (I_{NT} - B_N)Y \quad (19)$$

avec $B_N = I_N \otimes B$.

Modèles linéaires et hétérogénéité

3. Modèle à effets fixes

Definition

Sous les hypothèses usuelles, l'estimateur LSDV ou Within est défini par :

$$\hat{\beta}_{LSDV} = \left(\sum_{i=1}^N x_i W x_i' \right)^{-1} \sum_{i=1}^N x_i W y_i \quad (20)$$

ou encore

$$\hat{\beta}_{LSDV} = (X W_N X')^{-1} X W_N Y \quad (21)$$

Modèles linéaires et hétérogénéité

3. Modèle à effets fixes

Theorem

L'estimateur LSDV β est sans biais et convergent :

$$\hat{\beta}_{LSDV} \xrightarrow[NT \rightarrow \infty]{p} \beta \quad (22)$$

Preuve ?

Theorem

L'estimateur des α_i est sans biais et convergent :

$$\hat{\alpha}_i \xrightarrow[T \rightarrow \infty]{p} \alpha_i \quad (23)$$

Modèles linéaires et hétérogénéité

3. Modèle à effets fixes

Theorem

La variance asymptotique de $\hat{\beta}_{LSDV}$ est donnée par :

$$V(\hat{\beta}_{LSDV}) = \sigma_{\varepsilon}^2 \left(\sum_{i=1}^N x_i' W x_i \right)^{-1} \quad (24)$$

Modèles linéaires et hétérogénéité

Chapitre 1 - Outline

- 1 Tests de spécification
- 2 Effets linéaires non-observés
- 3 Effets fixes
- 4 **Effets aléatoires**
- 5 Tests de spécification
- 6 Panel hétérogènes : coefficients aléatoires

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

Definition

La modélisation des effets non-observés par des effets aléatoires (random effects) est un cas particulier d'un modèle à erreur composée :

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

$$u_{it} = \varepsilon_{it} + \lambda_t + \alpha_i$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

- Hypothèse d'exogénéité stricte :

$$E[\varepsilon_{it} | x_{i1}, \dots, x_{iT}] = 0, \quad \forall \quad i = 1, \dots, N, \quad t = 1, \dots, T$$

$$E[\alpha_i | x_{i1}, \dots, x_{iT}] = 0$$

- Non-corrélation et homoskedasticité :

$$E[\alpha_i \alpha_j] = \sigma_\alpha^2 \quad \text{si } i = j, \quad 0 \quad \text{sinon}$$

$$E[\varepsilon_{it} \varepsilon_{js}] = \sigma_\varepsilon^2 \quad \text{si } i = j \quad \text{et } t = s, \quad 0 \quad \text{sinon}$$

$$E[\alpha_j \varepsilon_{it}] = 0 \quad \forall \quad j, i, t$$

$$E[\alpha_i x'_{it}] = E[\varepsilon_{it} x'_{it}] = 0$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

- Le modèle à effets fixe revient à faire l'hypothèse que les α_i est un paramètre aléatoire avec

$$\text{cov}(\alpha_i, x'_{it}) \neq 0$$

- Le modèle aléatoire postule que :

$$\text{cov}(\alpha_i, x'_{it}) = 0$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

Remarques :

- Si les effets individuels ont une moyenne non nulle, on peut poser $\alpha_i^* = \mu + \alpha_i$ avec $E[\alpha_i] = \mu$.
- On a :

$$\text{Var}(y_{it}|x_{i1}, \dots, x_{it}) = \sigma_{y|x}^2 = \sigma_{\alpha}^2 + \sigma_{\varepsilon}^2$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

Ce modèle peut aussi sous forme vectorielle :

$$y_i = x_i\beta + u_i, \quad u_i = e\alpha_i + \varepsilon_i \quad (25)$$

ou

$$y_i = \tilde{x}_i\gamma + u_i, \quad u_i = e\alpha_i + \varepsilon_i \quad (26)$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

Definition

Sous les hypothèse postulées, la matrice variance-covariance de u_i est définie par :

$$\Omega = E[u_i u_i'] = \sigma_\alpha^2 e e' + \sigma_\varepsilon^2 I_T \quad (27)$$

et

$$\begin{aligned}\Omega^{-1} &= \frac{1}{\sigma_\varepsilon^2} \left(I_T - \left(\frac{\sigma_\alpha^2}{\sigma_\varepsilon^2 + T \sigma_\alpha^2} \right) e e' \right) \\ \Omega^{-1} &= \frac{1}{\sigma_\varepsilon^2} \left(W - \phi \frac{1}{T} e e' \right)\end{aligned}$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

$$\Omega = \begin{pmatrix} \sigma_{\alpha}^2 + \sigma_{\varepsilon}^2 & \sigma_{\alpha}^2 & \dots & \sigma_{\alpha}^2 \\ \sigma_{\alpha}^2 & \sigma_{\alpha}^2 + \sigma_{\varepsilon}^2 & \dots & \sigma_{\alpha}^2 \\ & & \dots & \sigma_{\alpha}^2 \\ & & & \sigma_{\alpha}^2 + \sigma_{\varepsilon}^2 \end{pmatrix}$$

Modèles linéaires et hétérogénéité

Modèle à effets aléatoires

$$\begin{aligned} \text{Var}(U|X) &= \begin{pmatrix} \text{Var}(u_1|X) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \text{Var}(u_N|X) \end{pmatrix} = \begin{pmatrix} \Omega & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Omega \end{pmatrix} \\ &= I_N \otimes \Omega \end{aligned}$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

Theorem

L'estimateur LSDV dans le cas où les effets non observés sont modélisés comme des erreurs aléatoires n'est pas BLUE.

Néanmoins, il est sans biais et consistant sous les hypothèses précédentes.

Hint : Matrice Ω pas standard.

Theorem

Dans le cadre du modèle à effets aléatoires, l'estimateur GLS (MCG, moindres carrés généralisés) est BLUE

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

On considère le modèle suivant :

$$y_i = \tilde{x}_i' \gamma + u_i, \quad u_i = e \alpha_i + \varepsilon_i \quad (28)$$

Theorem

Sous l'hypothèse Ω connue, l'estimateur par MCG est donné par :

$$\hat{\gamma}_{GLS} = \left(\sum_{i=1}^N x_i' V^{-1} x_i \right)^{-1} \left(\sum_{i=1}^N x_i V^{-1} y_i \right) \quad (29)$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

On peut réécrire $\hat{\gamma}_{MCG}$:

$$\hat{\gamma}_{GLS} = \left[\sum_{i=1}^N x_i' \left(\frac{1}{\sigma_\varepsilon^2} \left(W - \phi \frac{1}{T} ee' \right) \right) x_i \right]^{-1} \left[\sum_{i=1}^N x_i \left(\frac{1}{\sigma_\varepsilon^2} \left(W - \phi \frac{1}{T} ee' \right) \right) y_i \right]$$

$$\hat{\gamma}_{GLS} = \left[\sum_{i=1}^N x_i' W x_i - x_i' \phi \frac{1}{T} ee' x_i \right]^{-1} \left[\sum_{i=1}^N x_i W y_i - x_i \phi \frac{1}{T} ee' y_i \right]$$

$$\hat{\gamma}_{GLS} = \left[\sum_{i=1}^N (x_i' W x_i) - \frac{\phi}{T} \sum_{i=1}^N (x_i' ee' x_i) \right]^{-1} \left[\sum_{i=1}^N (x_i W y_i) - \frac{\phi}{T} \sum_{i=1}^N (x_i ee' y_i) \right]$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

A l'aide de propriétés sur les matrices, on peut montrer que :

Definition

Si la matrice Ω est connue, on a :

$$\hat{\beta}_{MCG} = \left[\frac{1}{T} \sum_{i=1}^N x_i' W x_i - \phi \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right]^{-1} \left[\frac{1}{T} \sum_{i=1}^N x_i' W y_i - \frac{\phi}{T} \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})' \right]$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

L'estimateur des MCG est une moyenne pondérée des estimateurs *within* et *between* : $\hat{\beta}_{MCG} = \theta \hat{\beta}_{be} + (I_N - \theta) \hat{\beta}_{LSDV}$

Definition

L'estimateur *between* de β correspond à l'estimation OLS du modèle :

$$\bar{y}_i = \bar{x}_i \beta + u_i$$

$$\hat{\beta}_{be} = \left[\sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right]^{-1} \left[\sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})' \right]$$

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

Rappel : $\phi = \frac{\sigma_{\alpha}^2}{T\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2}$

- Ce paramètre mesure le poids donné à la variation entre les individus.
- Si $\phi = 0$, cette variation est ignorée et on se retrouve avec le modèle à effets fixes.
- Si $\phi = 1$, on ajoute les variations entre les individus et les périodes.

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

Theorem

Quand T tend vers l'infini, l'estimateur MCG converge vers l'estimateur LSDV :

$$\hat{\beta}_{MCG} \xrightarrow{T \rightarrow \infty} \hat{\beta}_{LSDV}$$

- Chaque individu à un nombre d'observation infini
- α_i peut-être alors considérée comme une variable aléatoire.

Modèles linéaires et hétérogénéité

4. Modèle à effets aléatoires

Definition

Si la matrice de variance-covariance Ω n'est pas connue, on utilise les MCG *faissable* :

- 1 step : estimations des composants de la variance
- 2 step : on utilise ces estimations dans l'estimation des MCG.

$$\hat{\sigma}_{\alpha}^2 = \frac{\sum_{i=1}^N \left(\bar{y}_i - \hat{\beta}'_{LSDV} \bar{x}_i \right)^2}{N - K - 1} - \sigma_{\varepsilon}^2$$
$$\hat{\sigma}_{\varepsilon}^2 = \frac{\sum_{i=1}^N \sum_{t=1}^T \left((y_{it} - \bar{y}_i) - \hat{\beta}_{LSDV} (x_{it} - \bar{x}_i) \right)^2}{N(T - 1) - K}$$

Chapitre 1 - Outline

- 1 Tests de spécification
- 2 Effets linéaires non-observés
- 3 Effets fixes
- 4 Effets aléatoires
- 5 **Tests de spécification**
- 6 Panel hétérogènes : coefficients aléatoires

Modèles linéaires et hétérogénéité

5. Tests de spécification

Comment discriminer entre les deux types d'effets ?

- Quand T est grand, peu importe!!!
- Sinon, pas facile d'y répondre... il peut y avoir des grosses différences entre les différents estimateurs !

Modèles linéaires et hétérogénéité

5. Tests de spécification

Hypothèses fondamentales du modèle à effets aléatoires :

- α_i sont des tirages aléatoires
- Tous les composants du terme d'erreur sont orthogonaux aux variables explicatives. Ces dernières sont donc strictement exogène.

Que se passe-t'il si on néglige la corrélation entre les α_i et les variables explicatives ?

Modèles linéaires et hétérogénéité

5. Tests de spécification

Definition

Soit :

$$\alpha_i = \bar{x}_i' a + \alpha_i^*$$

$$y_{it} = \mu + \beta' x_{it} + \bar{x}_i' a + u_{it}, \quad u_{it} = \alpha_i^* + \varepsilon_{it}$$

Modèles linéaires et hétérogénéité

5. Tests de spécification

Definition

Sous les hypothèses précédentes, l'estimation des paramètres de la spécification de Mundlak est :

$$\hat{\mu}_{MCG}^* = \bar{y} - \bar{x}' \hat{\beta}_{be}$$

$$\hat{\beta}_{MCG}^* = \theta \hat{\beta}_{be} + (I_N - \theta) \hat{\beta}_{LSDV}$$

$$\hat{a}_{MCG}^* = \hat{\beta}_{be} - \hat{\beta}_{LSDV}$$

Modèles linéaires et hétérogénéité

5. Tests de spécification

L'estimateur *between* est obtenu en estimant le modèle :

$$\bar{y}_i = c + (\beta + a)' \bar{x}_i + \varepsilon_i \quad (30)$$

Que se passe-t'il si on néglige la corrélation entre α_i et x_{it} ?

Modèles linéaires et hétérogénéité

5. Tests de spécification

$$\hat{\beta}_{be} \xrightarrow[N \rightarrow \infty]{p} \beta + a$$
$$\hat{\beta}_{LSDV} \xrightarrow[N \rightarrow \infty]{p} \beta$$

Donc

$$\hat{\beta}_{MCG} \xrightarrow[N \rightarrow \infty]{p} \beta + a\bar{\theta}$$

Modèles linéaires et hétérogénéité

5. Tests de spécification

Hausman (1978), test des effets fixes vs effets aléatoires (Econometrica).

- Soit le modèle $y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it}$
- Soit deux estimateurs convergents $\bar{\beta}$ et $\tilde{\beta}$
- Sous H_0 , $\bar{\beta}$ qui a les bonnes propriétés asymptotiques
- Sous H_1 , $\tilde{\beta}$ est biaisé et ne converge pas vers sa vraie valeur

Donc si $\tilde{\beta}$ et $\bar{\beta}$ sont proches, on ne rejette pas H_0 , sinon on rejette.

Modèles linéaires et hétérogénéité

5. Tests de spécification

Definition

La statistique du test d'Hausman est définie par :

$$H = \left(\hat{\beta}_{MCG} - \hat{\beta}_{LSDV} \right)' \left(\text{Var} \left(\hat{\beta}_{MCG} - \hat{\beta}_{LSDV} \right) \right)^{-1} \left(\hat{\beta}_{MCG} - \hat{\beta}_{LSDV} \right) \quad (31)$$

Sous l'hypothèse nulle, cette statistiques de test suit une distribution du $\chi^2(K)$.

Modèles linéaires et hétérogénéité

Chapitre 1 - Outline

- 1 Tests de spécification
- 2 Effets linéaires non-observés
- 3 Effets fixes
- 4 Effets aléatoires
- 5 Tests de spécification
- 6 **Panel hétérogènes : coefficients aléatoires**

Modèles linéaires et hétérogénéité

Panel hétérogènes : coefficients aléatoires

Que se passe-t-il si

- il y a un changement structurel ?
- un changement socioéconomique ?
- un changement climatique/naturel ?

β peut varier !

Modèles linéaires et hétérogénéité

6. Panel hétérogènes : coefficients aléatoires

Definition

Le modèle de panel hétérogène ou à coefficients aléatoires s'écrit

$$y_{it} = \sum_{k=1}^K \beta_{itk} x_{kit} + \varepsilon_{it}, \quad i = 1, \dots, N \quad t = 1, \dots, T \quad (32)$$

La constante est ici incluse directement dans les variables explicative : $x_{1it} = 1$. Nous ferons l'hypothèse que les coefficients ne varient pas au cours du temps. β peut varier en fonction des individus.

Modèles linéaires et hétérogénéité

6. Panel hétérogènes : coefficients aléatoires

Remarques :

- Ce modèle revient à estimer une régression pour chaque cross-section
- mais les erreurs ε_{it} peuvent être corrélées entre les individus.
- β_i peuvent être définis comme des variables aléatoires avec une fonction de répartition. identique

Modèles linéaires et hétérogénéité

6. Panel hétérogènes : coefficients aléatoires

Definition

$$y_{it} = \sum_{k=1}^K (\beta_k + \psi_{ki}) x_{kit} + \varepsilon_{it} \quad (33)$$

avec β le vecteur de paramètres communs à chaque cross-section et ψ_i le vecteur de la déviation individuelle.

- Estimation des β ,
- Prédiction de chaque composante individuelle β_i ,
- Estimation de la dispersion entre les β_i .

Modèles linéaires et hétérogénéité

6. Panel hétérogènes : coefficients aléatoires

Selon la spécification des ψ_i , on a deux modèles possibles :

- Si les observations individuelles sont hétérogènes, ψ_i peut-être considérés comme des (modèle SURE de Zellner (1962))
- Si les observations individuelles sont des tirages aléatoires d'une population homogène, on émet l'hypothèse que les ψ_i sont des variables aléatoires avec une moyenne nulle $E(\psi_i|x_{kit}) = 0$ et une matrice variance-covariance constantes $Var(\psi_i) = \Delta$.

Modèles linéaires et hétérogénéité

6. Panel hétérogènes : coefficients aléatoires

Definition (Modèle à coefficients aléatoires)

Chaque coeff est une v.a avec une fonction de densité commune :

$$y_{it} = \sum_{k=1}^K (\beta_k + \psi_{ki}) x_{kit} + \varepsilon_{it} \quad (34)$$

$$\beta_i \sim IID(\beta, \Delta) \quad (35)$$

- $E(\psi_i) = 0, \quad E(\varepsilon_i) = 0$
- $E(\psi_i \psi_j') = \Delta \quad i = j$
- $E(x_{it} \psi_j') = E(\psi_i \varepsilon_j') = 0$
- $E(\varepsilon_i \varepsilon_j') = \sigma_i^2 I_T, i = j$

Modèles linéaires et hétérogénéité

6. Panel hétérogènes : coefficients aléatoires

La matrice de variance-covariance Δ s'écrit :

$$\Delta = E [(\beta_i - \beta) (\beta_i - \beta)'] = \begin{pmatrix} \sigma_{\beta_1}^2 & \sigma_{\beta_1, \beta_2} & \cdots & \sigma_{\beta_1, \beta_K} \\ \sigma_{\beta_2, \beta_1} & \sigma_{\beta_2}^2 & \cdots & \sigma_{\beta_2, \beta_K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\beta_K, \beta_1} & \sigma_{\beta_K, \beta_2} & \cdots & \sigma_{\beta_K}^2 \end{pmatrix}$$

Modèles linéaires et hétérogénéité

6. Panel hétérogènes : coefficients aléatoires

Definition

On peut aussi écrire le modèle à coefficients aléatoires :

$$y_i = x_i\beta + u_i$$

$$u_i = x_i\psi_i + \varepsilon_i = x_i(\beta_i - \beta) + \varepsilon_i$$

Modèles linéaires et hétérogénéité

6. Panel hétérogènes : coefficients aléatoires

La matrice de variance-covariance du terme d'erreur est donc donnée par :

$$\begin{aligned}\Omega_i &= E(u_i u_i') \\ &= E[(x_i \psi_i + \varepsilon_i)(x_i \psi_i + \varepsilon_i)'] \\ &= x_i \Delta x_i' + \sigma^2 I_T\end{aligned}$$

- L'estimateur MCO est unbiased mais,
- pas efficace à cause de la matrice variance-covariance.

Modèles linéaires et hétérogénéité

6. Panel hétérogènes : coefficients aléatoires

Definition

L'estimateur par MCG est blue :

$$\hat{\beta}_{MCG} = \left(\sum_{i=1}^N x_i' \Omega_i^{-1} x_i \right)^{-1} \left(\sum_{i=1}^N x_i' \Omega_i^{-1} y_i \right) \quad (36)$$

Modèles linéaires et hétérogénéité

Panel hétérogènes : coefficients aléatoires

Estimation d'un modèle à coefficients aléatoires :

- Estimer les N régressions individuelles $y_{it} = \beta_i' x_{it} + \varepsilon_{it}$.

- Calculer

- $\hat{\sigma}^2 = \frac{1}{T-K} \sum_{t=1}^T \hat{\varepsilon}_{it}$

- $\hat{\Delta} = \frac{1}{n-1} \sum_{i=1}^N \left(\left(\hat{\beta}_i - \frac{1}{N} \sum_{i=1}^N \hat{\beta}_i \right) \left(\hat{\beta}_i - \frac{1}{N} \sum_{i=1}^N \hat{\beta}_i \right)' \right)$

- Construire $\hat{\beta}_{MCG} = \left(\sum_{i=1}^N x_i' \hat{\Omega}_i^{-1} x_i \right)^{-1} \left(\sum_{i=1}^N x_i' \hat{\Omega}_i^{-1} y_i \right)$

Modèles linéaires et hétérogénéité

6. Panel hétérogènes : coefficients aléatoires

Bonhomme2015

- Hétérogénéité time-varying par groupe dans les modèles linéaires de panel.
- L'appartenance à un groupe est définie par les données.