Université de Gafsa

Institut Supérieur d'Administration des Affaires de Gafsa

Économétrie des données de Panel

Niveau : (M1 – MFB) Master Recherche Monnaie, Finance et Banque

Enseignant : Dr. Zayati Montassar

Maître-assistant en méthodes quantitatives : spécialité économétrie

Tél: (00216) 50 074 124

E-mail: zayati_monta@hotmail.fr

Année Universitaire : 2014/2015

Présentation

Ce cours est une initiation, tant sur le plan théorique que sur le plan appliqué, à l'économétrie des données de panel. Effectivement, nous allons présentés les techniques les plus courantes de modélisation des données de panel, et ce par le biais d'un volet théorique et un autre empirique. On débutera par une présentation des problèmes de spécifications de base en économétrie de panel et par les méthodes d'estimation traditionnelles.

L'objectif est de faire en sorte que le lecteur puisse interpréter, de façon exhaustive et relativement approfondie, les résultats de base que donnent les principaux logiciels d'économétrie lorsque l'on envisage des modèles de panel. Nous prendrons ici comme référence les logiciels STATA et Eviews, mais il est bien entendu évident que ces résultats de base sont sensiblement identiques si l'on considère d'autres logiciels comme SAS, Rats ou TSP.

Nous souhaitons, ainsi, présenter les connaissances minimales nécessaires pour pouvoir interpréter un tableau de résultats d'estimation de panel, comme par exemple :

- Les estimateurs Pooled,
- Les estimateurs Between,
- Les estimateurs du modèle à effets individuels fixes (Within),
- L'Error Component Model (modèle à effets individuels aléatoires),
- Les résultats de trois tests de Fischer,
- L'estimateur de la variance des effets individuels,
- Un estimateur de la variance totale,
- La statistique du test d'Hausman.

Sommaire

Introduction

Les données utilisées en économétrie sont le plus souvent des séries chronologiques ou en coupe instantanée concernant une période donnée.

Les données de panel, ou données longitudinales possèdent les deux dimensions précédentes (individuelle et temporelle). En effet, il est souvent intéressant d'identifier l'effet associé à chaque individu (un effet qui ne varie pas dans le temps, mais qui varie d'un individu à un autre). Cet effet peut être fixe ou aléatoire.

Par conséquent, le modèle en données de panel s'écrit comme un modèle à double indice qui prend la forme suivante :

$$Y_{it} = \alpha_i + \sum_k \beta_{ki} x_{kit} + \varepsilon_{it} \qquad \text{avec} \quad \begin{cases} i: 1 \longrightarrow N \\ t: 1 \longrightarrow T_i \end{cases}$$

La double dimension qu'offrent les données de panel est un atout majeur. En effet, si les données en séries temporelles permettent d'étudier l'évolution des relations dans le temps, elles ne permettent pas de contrôler l'hétérogénéité entre les individus. A l'inverse, les données en coupes transversales permettent d'analyser l'hétérogénéité entre les individus mais elles ne peuvent pas tenir compte des comportements dynamiques, puisque la dimension temporelle est exclue du champ d'analyse.

Ainsi, en utilisant des données de panel, on pourra exploiter les deux sources de variation de l'information statistique :

- Temporelle où variabilité intra-individuelle (within)
- et individuelle ou variabilité inter-individuelle (Between).

Remarques 1 :

- L'augmentation du nombre d'observations permet de garantir une meilleure précision des estimateurs, de réduire les risques de multi colinéarité et surtout d'élargir le champ d'investigation.

- Le panel considéré n'est pas nécessairement complet (cylindré) où toutes les unités statistiques sont observés durant la même période considérée. Il peut s'agir d'un panel incomplet, non cylindré.

⟹ Le cylindrage de l'échantillon n'est pas conseillé à cause du risque de biais de sélectivité.

Remarques 2 :

- Théoriquement, les méthodes proposées supposent que la dimension individuelle est infinie (on peut prendre des centaines, ou des milliers d'entreprises) et que la dimension temporelle est finie. D'où l'intérêt de contrôler l'hétérogénéité individuelle qui peut être supposé fixe ou aléatoire.

Chapitre 1 : Les régressions linéaires sur données de panel

La première étape à établir pour un échantillon de données de panel est de vérifier la spécification homogène ou hétérogène du processus générateur de données. La phase de test de spécification revient à déterminer si on a le droit de supposer une fonction de régression identique pour tous les individus (modèle ***Pooled***). Dans ce cas, les élasticités des facteurs exogènes sont identiques ($\beta_i = \beta$) ; et la constante elle aussi identique pour tous les individus ($\alpha_i = \alpha$) selon le modèle suivant :

$$Y_{it} = \alpha + \sum_k \beta_k x_{kit} + \varepsilon_{it}$$

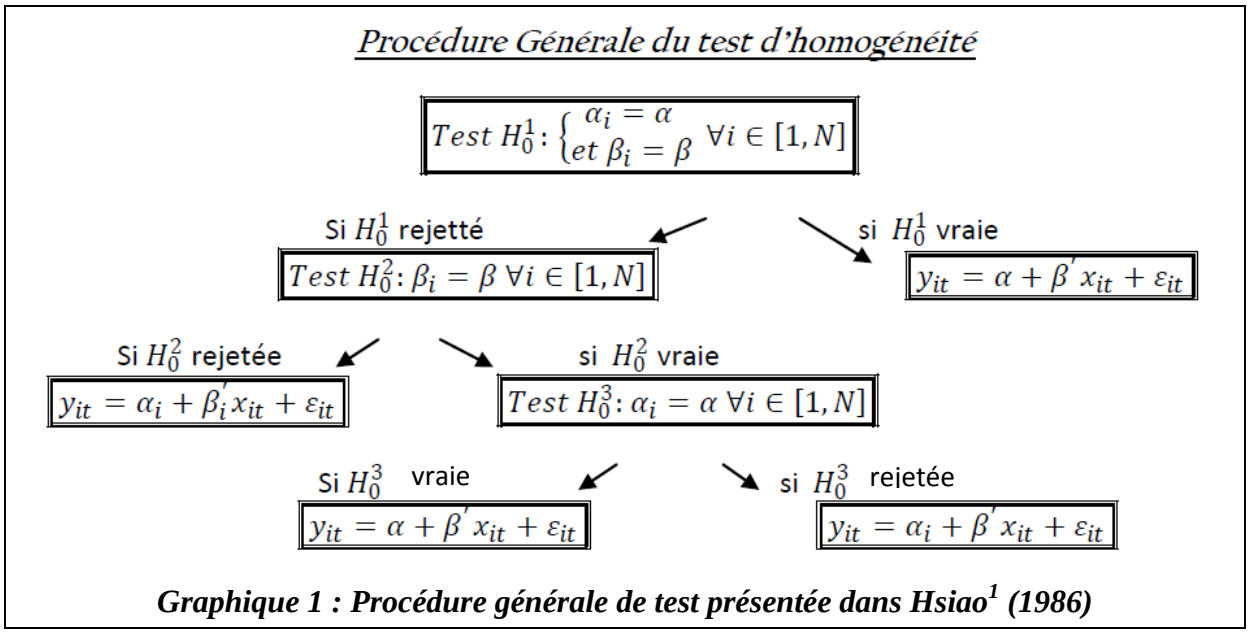
Toutefois, lorsqu'on travaille sur des séries agrégées, il est peu probable que la fonction de régression, soit strictement identique pour tous les individus étudiés. Ainsi, il convient de tester si les élasticités des différents facteurs (β_i) sont identiques. Si ce n'est pas le cas, il n'existe à priori aucune structure d'estimation commune entre les pays (individus), et donc l'utilisation des données de panels ne se justifie pas et peut même conduire à des biais d'estimation. On doit, alors, estimer les fonctions individu par individu.

En revanche, s'il existe bien une relation identique pour toutes les élasticités, alors la source d'hétérogénéité ne peut provenir que des constantes (α_i). Or, rien ne garantit que les pays étudiés possèdent le même niveau moyen de la variable endogène. Au contraire, il se peut parfaitement que des facteurs e-temporels ou structurels (comme la position géographique, le climat, l'éloignement par rapport au grands axes commerciales…) pouvaient conduire à des différences structurelles entre les individus.

Dans ce cas, le niveau moyen des facteurs, déterminer par $E(\alpha_i + \varepsilon_{it}) = \alpha_i$; varie selon les pays même si les élasticités du modèle (β_i) sont les même. On obtient alors un modèle avec ***effets individuels*** qui s'écrit sous la forme :

$$Y_{it} = \alpha_i + \sum_k \beta_k x_{kit} + \varepsilon_{it}$$

Ainsi, la phase de test de spécification revient à déterminer si le processus générateur de données peut être considéré comme homogène, c'est-à-dire unique pour tous les individus, ou si au contraire il apparaît totalement hétérogène, auquel cas l'utilisation des techniques de panel ne peut se justifier. Entre ces deux cas extrêmes se trouve un modèle dit à effets individuels. Il convient, surtout, d'identifier la source d'hétérogénéité pour bien spécifier le modèle.



Graphique 1 : Procédure générale de test présentée dans Hsiao[1] (1986)

1. *Test de Spécification[2] :*

1.1. Hétérogénéité des Comportements :

Plusieurs configurations sont disponibles :

- ✓ Les constantes α_i et les paramètres β_i sont identiques. On qualifie ce panel de panel homogène (**Pooled**).
- ✓ Les N constantes α_i et les N vecteurs de paramètres β_i sont différents selon les individus. On a donc N = 5 (selon le nombre de pays) modèles différents, on rejette la structure de panel.
- ✓ Les N vecteurs de paramètres β_i sont identiques, $\beta_i = \beta$; tandis que les constantes α_i diffèrent selon les individus. On obtient un **modèle hétérogène à effets individuels**.

a) Test d'homogénéité globale (***H_0^1***):

Il s'agit de tester le test suivant :

$$Test\ H_0^1 : \begin{cases}\alpha_i = \alpha \\ \beta_i = \beta\end{cases}\ \forall i \in [1,N]$$

La statistique utilisée est celle de Fisher : $$F_1 = \frac{(SCR_c - SCR)/(N-1)(K+1)}{SCR/[NT-N(K+1)]}$$

Avec :

[1] Hsiao, C., (1986), "Analysis of Panel Data", Econometric society Monographs N°11. Cambridge Universirty Press.

[2] L'explication et l'interprétation des différentes étapes des tests de spécification seront réalisées sur un cas pratique d'une fonction de production de type Cobb Douglass sur un échantillon de 5 pays et une période de 16 ans.

- ✓ N = 5 pays, K= 3 variables exogènes et T = 16 années.
- ✓ SCR c'est la somme des carrés résiduels du modèle (1) : $y_{it} = \alpha_i + \beta_i^{'} X_{it} + \varepsilon_{it}$ $SCR = \sum_{i=1}^{N} SCR_i$ pour chaque individu (pays).
- ✓ SCR_c est celle du modèle contraint (modèle d'homogénéité totale ou *Pooled*) : elle est calculée en estimant le modèle sur l'échantillon complet à NT observations. $y_{it} = \alpha + \beta^{'} X_{it} + \varepsilon_{it}$.

Pour notre modèle les résultats sont les suivantes : $SCR_c = 0.645096$ et $SCR = 0.188292$.

Donc $F_{c1} = \frac{(SCR_c - SCR)/(N-1)(K+1)}{SCR/[NT-N(K+1)]} = \frac{(0.645096 - 0.188292)/(4*4)}{0.188292/[80-5(4)]} = 9{,}097$.

La statistique ainsi calculée est supérieure à F(16 ; 60) ≈ 2.13 du tableau de Fisher ; on rejette alors l'hypothèse H_0^1 d'une parfaite homogénéité du modèle et on passe au deuxième test.

b) Test d'homogénéité des coefficients β_i ($\boldsymbol{H_0^2}$) :

Le test est le suivant : $H_0^2: \beta_i = \beta \ \forall i \in [1, N]$

La statistique du test est la suivante : $F_2 = \frac{(SCR_c^{'} - SCR)/(N-1)K}{SCR/[(NT-N(K+1)]}$

Avec, SCR'_c est la somme des carrés résiduels du modèle contraint à effet individuels :

$$y_{it} = \alpha_i + \beta^{'} X_{it} + \varepsilon_{it}$$

Ici les estimateurs (***Withen***) des paramètres α_i et β sont obtenus en centrant les variables sur les moyennes individuelles respectives. C'est le même que **l'estimateur à effet fixe** donné par Eviews et STATA ou l'estimateur ***Last Squar Damy Variable*** (LSDV) calculé par le logiciel TSP.

Les résultats d'estimation dans notre modèle sont les suivantes : $SCR'_c = 0.261937$

$$F_{c2} = \frac{(SCR_c^{'} - SCR)/(N-1)K}{SCR/[(NT-N(K+1)]} = \frac{(0.261937 - 0.188292)/(4*3)}{0.188292/[(80-5(4)]} = 1{,}9556.$$

Cette valeur est inferieur à celle du tableau de Fisher $F(12\,;\,60) \approx 2.3$. Ainsi, on accepte l'hypothèse H_0^2 et on admet le modèle de panel avec homogénéité des coefficients β_i et on passe au test suivant d'homogénéité des coefficients α_i.

c) Test d'homogénéité des constantes α_i ($\boldsymbol{H_0^3}$) :

$$\text{Test}\, H_0^3: \alpha_i = \alpha \ \forall i \in [1, N]$$

La statistique de ce test est la suivante : $F_3 = \frac{(SCR_c - SCR_c^{'})/(N-1)}{SCR_c^{'}/[N(T-1)-K]}$

$$F_{c3} = \frac{(SCR_c - SCR_c^{'})/(N-1)}{SCR_c^{'}/[N(T-1)-K]} = \frac{(0.645096 - 0.261937)/4}{0.261937/[5(15)-3]} = 26{,}33 > F(4\,;72) \approx 5{,}63$$

On rejette, alors, l'hypothèse $\boldsymbol{H_0^3}$ pour opter à un modèle de panel à effet individuel causé par l'hétérogénéité de la constante $\underline{\alpha_i}$. Il ne reste maintenant que de déterminer la nature de cet effet (fixe ou aléatoire) par le **test d'Hausman.**

d) Test d'Hausman :

C'est un test qui sert à discriminer les effets fixes et aléatoires des effets individuels dans un modèle des données en panel. Il s'agit de tester la présence éventuelle d'une corrélation ou d'un défaut de spécification (corrélation entre des effets individuels et des variables explicatives). Le test est le suivant :

$$\begin{cases} H_0^4\text{: } \mathrm{E}(\alpha_i / X_i)=0 \\ H_0^4\text{: } \mathrm{E}(\alpha_i / X_i) \neq 0 \end{cases}$$

Sous l'hypothèse H_0^4, les effets individuels sont aléatoires, alors la méthode adéquate pour l'estimation est la Méthode des Moindres Carrés Généralisés (MCG). Sinon, sous l'hypothèse H_0^4, l'estimateur utilisé est l'estimateur *Within*.

La statistique du test est la suivante :

$$H = (\beta_{MCG} - \beta_{LSDV})'[Var(\beta_{MCG} - \beta_{LSDV})]^{-1}(\beta_{MCG} - \beta_{LSDV})$$

Les résultats d'estimation étaient les suivantes:

Variables	Withen (LSDV)	MCG
Cte	-	8,523***
		(11,379)
Log(K)	0,9266***	0,9277***
	(45,608)	(32,558)
Log(L)	-1,026***	-1,012***
	(-35,1546)	(-10,968)
IGG	0.0009*	0,0041***
	(1,708)	(3,554)
Test d'Hausman	-	104,378
P-value		(0.0000)***

Les valeurs entre parenthèses sont les t-statistic.
, ** et * sont les significativités respectivement à 10%, 5% et 1%.*
Source : Calcul de l'auteur

La statistique du test d'Hausman est égal à 104,378 et présente une probabilité statistique de 0,0000 donc on va rejeter l'hypothèse H_0, et opter pour un modèle de panel à effet fixe.

2. Modèles à effets individuels

Nous allons à présent nous concentrés sur les modèles de panel hétérogènes, où la seule source d'hétérogénéité provient des constantes individuelles. On suppose ainsi que les coefficients des différentes variables stochastiques explicatives sont identiques pour tous les individus du panel ($\beta_i = \beta$). On suppose en outre que ces coefficients sont des constantes déterministes. Les constantes individuelles α_i; quant à elles, diffèrent selon les individus.

$$Y_{it} = \alpha_i + \sum_k \beta_k x_{kit} + \varepsilon_{it}$$

Les innovations ε_{it} sont supposées être i:i:d: de moyenne nulle, de variance égale à σ_ε^2; $\forall i \in [1; N]$ et sont supposées non corrélées que ce soit dans la dimension individuelle ou dans la dimension temporelle.

Dès lors, dans ce contexte, on doit distinguer deux cas : le cas où les paramètres α_i sont des constantes déterministes (modèle à effets fixes) et le cas où les paramètres α_i sont des réalisations d'un variable aléatoire d'espérance et de variance finie (modèle à effets aléatoires). Nous allons donc successivement envisager ces deux types de modèle.

2.1. Modèle à effets fixes

On fait maintenant l'hypothèse que les effets individuels α_i sont représentés par des constantes (d'où l'appellation modèle à effets fixes). Nous allons déterminer la forme générale des estimateurs des paramètres α_i et β dans ce modèle à effets fixes.

Hypothèses :

- Le modèle à effets fixes individuels présente une structure des résidus qui vérifient les hypothèses standards des MCO. Il s'agit en fait d'un modèle classique avec variables indicatrices individuelles.
- nous allons faire une hypothèse supplémentaire sur la nature du processus des résidus ε_{it}. Cette hypothèse constitue tout simplement la généralisation dans la dimension de panel de la définition d'un *bruit blanc* $\forall i \in [1; N]$ *et* $t \in [1; T]$:
  - $E(\varepsilon_{it}) = 0$
  - $E(\varepsilon_{it}\varepsilon_{is}) = \begin{cases} \sigma_\varepsilon^2 & t = s \\ 0 & \forall t \neq s \end{cases}$
  - $E(\varepsilon_{it}\varepsilon_{js}) = 0 \; \forall j \neq i, \; \forall (t, s)$

a) Estimateur Within ou LSDV (Least Square Dummy Variables)

L'estimateur des Moindres Carrés Ordinaires (MCO) des paramètres α_i et β dans le modèle à effets fixes est appelé estimateur Within; ou estimateur à effets fixes ou estimateur LSDV (Least Square Dummy Variable). Comme nous l'avons vu, le terme Within s'explique par le fait que cet estimateur tient compte de la variance intra groupe de la variable endogène.

La troisième appellation LSDV tient au fait que cet estimateur conduit à introduire des variables dummies.

Les estimateurs de ce modèle par la méthode des MCO sont les meilleurs estimateurs linéaires, sans biais et convergents (*BLUE*[3]). Dans la pratique, l'estimateur des MCO ou LSDV est obtenu à partir d'un modèle transformé où les différentes variables du modèle sont centrées par rapport à leurs moyennes individuelles respectives. On retient, alors, la spécification suivante :

$$\tilde{y}_{it} = \sum_k \beta_k \tilde{x}_{kit} + \tilde{\varepsilon}_{it} \quad \text{Avec} \begin{cases} \tilde{y}_{it} = y_{it} - \bar{y}_{it} \\ \tilde{x}_{it} = x_{it} - \bar{x}_{it} \\ \tilde{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_{it} \end{cases} \text{ et } \bar{y}_{it} = \frac{1}{T_i} \sum_{t=1}^{T_i} y_{it}$$

Les réalisations des estimateurs des constantes α_i sont déduites au point moyen, après estimation des paramètres β_k par MCO sur le modèle transformé précédent.

$$\hat{\alpha}_i = \bar{y}_i - \sum_{k=1}^{p} \hat{\beta}_k \bar{x}_{ki}$$

Remarque importante :

Il est conseillé dans le cas où le panel est non cylindré d'utiliser l'option ***robuste*** (estimateur à effet fixe robuste) de manière à tenir compte de l'hétéroscédasticité des erreurs, puisque la variance des erreurs du modèle transformé n'est pas constante. En effet, on vérifie que : $V(\varepsilon_{it}) = \sigma_\varepsilon^2 \frac{T_i}{T_i - 1}$

Limites :

Outre le fait que la variabilité inter-individuelle n'est pas exploitée pour estimer les paramètres structurels du modèle, une limite inhérente au modèle à effets fixes réside dans le fait que l'impact des facteurs invariants à travers le temps ne peut être identifié. Ceci constitue une limite au niveau de l'analyse économique, puisqu'il revient à restreindre le champ d'analyse économique de l'étude.

2.2. Modèle à effets aléatoires

Dans la pratique standard de l'analyse économétrique, on suppose qu'il existe un grand nombre de facteurs qui peuvent affecter la valeur de la variable expliquée et qui pourtant ne sont pas introduits explicitement sous la forme de variables explicatives. Ces facteurs sont alors approximés par la structure des résidus. Le problème se pose de la façon similaire en économétrie de panel. La seule différence tient au fait que trois types de facteurs omis peuvent être envisagés. Il y a tout d'abord les facteurs qui affectent la variable endogène

[3] Best Linear Unbiased Estimator

différemment suivant la période et l'individu considéré. Il peut en outre exister des facteurs qui affectent de façon identique l'ensemble des individus, mais dont l'influence dépend de la période considérée (effets temporel). Enfin, d'autres facteurs peuvent au contraire refléter des différences entre les individus de type structurelles, c'est à dire indépendantes du temps (effets individuel).

Dès lors le résidu, noté ε_{it}; d'un modèle de panel peut être décomposé en trois principales composantes de la façon suivante (Hsiao 1986) :

$$\forall i \in [1;N] \; et \; t \in [1;T] \;\; ; \quad \varepsilon_{it} = \alpha_i + \lambda_t + \vartheta_{it}$$

Les variables α_i désignent ici les effets individuels qui représentent l'ensemble des spécificités structurelles ou a-temporelles de la variable endogène, qui différent selon les individus. ***On suppose ici que ces effets sont aléatoires***. Les variables aléatoires λ_t représentent quant à elle les effets temporels strictement identiques pour tous les individus. Enfin, le processus stochastique ϑ_{it} désigne la composante du résidu total ε_{it} orthogonale aux effets individuels et aux effets temporels. Généralement, on est conduit à faire un certain nombre d'hypothèses techniques sur cette structure de résidus.

Hypothèses :

On suppose que les résidus $\varepsilon_{it} = \alpha_i + \lambda_t + \vartheta_{it}$ *sont i.i.d. et satisfont les conditions suivantes,* $\forall i \in [1;N] \; et \; t \in [1;T]$:

- $E(\alpha_i) = E(\lambda_t) = E(\vartheta_{it}) = 0$
- $E(\alpha_i \lambda_t) = E(\lambda_t \vartheta_{it}) = E(\vartheta_{it} \alpha_i) = 0$
- $E(\alpha_i \alpha_j) = \begin{cases} \sigma_\alpha^2 & i = j \\ 0 & \forall i \neq j \end{cases}$
- $E(\lambda_t \lambda_s) = \begin{cases} \sigma_\lambda^2 & s = t \\ 0 & \forall s \neq t \end{cases}$
- $E(\vartheta_{it} \vartheta_{js}) = \begin{cases} \sigma_\vartheta^2 & s = t \; ; \; i = j \\ 0 & \forall s \neq t \; ; \; \forall i \neq j \end{cases}$
- $E(\alpha_i x_{it}) = E(\lambda_t x_{it}) = E(\vartheta_{it} x_{it}) = 0$

Sous ces hypothèses, la variance de la variable endogène y_{it} conditionnellement aux variables explicatives x_{it} est alors égale à $\sigma_y^2 = \sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_\vartheta^2$. Les variances $\sigma_\alpha^2, \sigma_\lambda^2$ et σ_ϑ^2 correspondent aux différentes composantes de la variance totale. C'est pourquoi, le modèle à effets aléatoires est aussi appelé modèle à erreurs composés (Error Component Model).

Dans ce cours, en raison de simplification, l'effet temporel est négligé. Nous supposerons qu'il n'existe pas (panel statique).