Econométrie des données de panel

Guillaume Horny*

*Banque de France

Master 2 MASERATI

Introduction

Plan

Plan

Un peu de vocabulaire...

Que sont des données de panel ?

- ce sont des données relatives à des unités statistiques observées à plusieurs reprises dans le temps
- un panel est ainsi une répétition de coupes

On suit généralement des individus, des entreprises, des pays, etc. C'est pourquoi on parle parfois de données **longitudinales**.

On parle de panel **cylindré** (*balanced*) lorsque toutes les unités sont suivies à chaque date (pas de trou).

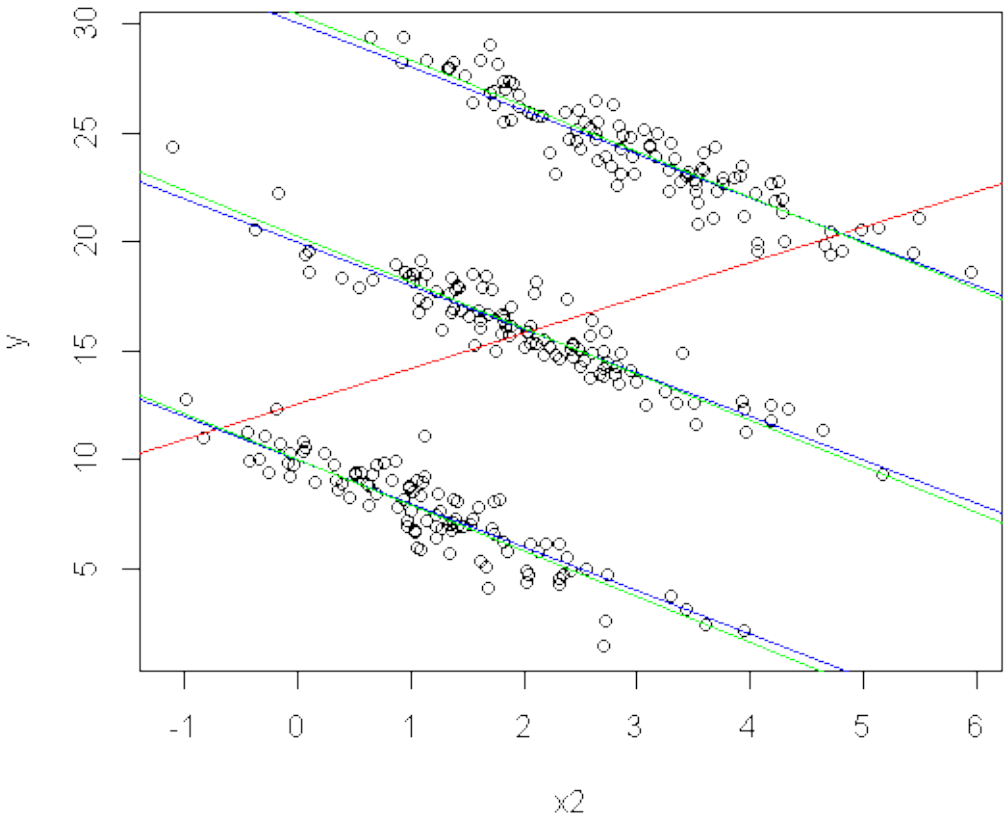
Un exemple de ce à quoi peuvent resembler des données de panel

```
R> library(plm)
R> data("Grunfeld", package="plm")
R> head(Grunfeld, 30)
firm  year    inv  value capital
   1 1935 317.6 3078.5     2.8
   1 1936 391.8 4661.7    52.6
   1 1937 410.6 5387.1   156.9
 .....
<SNIP>
 .....
   1 1954 1486.7 5593.6  2226.3
   2 1936  355.3 1807.1    50.5
   2 1937  469.9 2676.3   118.1
   2 1938  262.3 1801.9   260.2
```

Pourquoi un cours dédié ?

- les données de panel sont de fait très répandues à l'heure actuelle. Les administrations procèdent au suivi des assurés/contribuables, les entreprises au suivi de leurs clients, de leurs filiales, de leurs commerciaux...
- les techniques mises au point pour des données en coupe, lorsqu'elles sont appliquées à un panel, produisent des résultats erronés

Exemple (1/3)



Les droites issues des vraies valeurs sont en bleu, la droite de régression (OLS) en rouge, les régressions ajustées pour la dimension panel en vert

Exemple (2/3) : simulation de données

```
R> rm(list = ls())
R> id <- rep(1:3, each = 100)
R> date <- rep(1:100, 3)
R> x1 <- id * 10
R> x2 <- id + rnorm(300)
R> y <- x1 - 2 * x2 + rnorm(300)
R> donnees <- cbind(id, date, y, x1, x2)
R> donnees[1:10, ]
```

Exemple (3/3) : création de graphiques

```
R> plot(x2, y)
R> abline(lm(y ~ x2), col = "red")
R> abline(coef = c(10, -2), col = "blue")
R> abline(coef = c(20, -2), col = "blue")
R> abline(coef = c(30, -2), col = "blue")
R> z <- lm(y ~ 0 + x2 + as.factor(id))
R> abline(coef = c(coef(z)[2],coef(z)[1]),col="green")
R> abline(coef = c(coef(z)[3],coef(z)[1]),col="green")
R> abline(coef = c(coef(z)[4],coef(z)[1]),col="green")
```

Introduction

La question de l'hétérogénéité

> *"the most important discovery is the evidence on the pervasiveness of heterogeneity and diversity in economic life"*

Heckman, discours de remise du prix Nobel en 2001.

Ce constat paraît si naturel aujourd'hui qu'il semble étrange de le considérer comme une découverte aussi importante.

La question de l'hétérogénéité

La même idée se retrouve toutefois dans le domaine très différent de la médecine (Aalen, 1998) :

"It is a basic observation of medical statistics that individuals are dissimilar. Still, there is a tendency to regard this variation as a nuisance, and not as something to be considered seriously in its own right. Statisticians are often accused of being more interested in averages, and there is some truth to this."

Hétérogénéité observée et inobservée

- **l'hétérogénéité** est la différence entre les facteurs pertinents lors de la prise de décision et connus des agents (Cunha, 2005). Ainsi, nous sommes en présence d'hétérogénéité dès lors que les goûts, anticipations, capacités ou contraintes ne sont pas les même d'un agent à l'autre.
- Nous sommes présence **d'hétérogénéité non observée** lorsque diffèrent les facteurs pertinents et connus des agents, qui sont de plus inconnus de l'économètre (Browning, 2005). Il s'agit donc d'un type particulier d'hétérogénéité, caractérisé par le manque d'information sur les individus.

L'hétérogénéité observée renvoie aux différences entre les observations mesurées par les variables explicatives, et l'hétérogénéité inobservée aux autres différences.

L'hétérogénéité inobservée

Les variables explicatives sont rarement toutes observées : certaines peuvent ne pas être mesurables, codifiables ou encore être absentes des données. L'analyste peut avoir conscience du problème, mais se trouver dans l'impossibilité de les prendre en compte dans son modèle. En prenant en compte l'hétérogénéité inobservée, on accepte l'idée qu'il existe des déterminants inobservés par l'économètre. On ne les identifie pas forcément, mais on en contrôle les effets.

La question de l'hétérogénéité inobservée n'est importante que dans les **applications** : savoir si une variable n'est pas ou mal observable n'est pas une question centrale dans l'élaboration de modèles théoriques. Il revient à l'économètre d'en tenir compte pour éviter que ses résultats ne soient erronés du fait d'une mauvaise spécification d'une forme réduite.

Introduction

Notations

- On indice par i les individus ($i = 1, \ldots, N$) et t les périodes ($t = 1, \ldots, T$). On a donc NxT observations
- On note y_{it} la variable dépendante (ici un scalaire).
- On note x_{it} un vecteur de $Kx1$ variables, notées x_{it}^k.

Notations

- La moyenne inter-temporelle de la $k$ème variable pour l'individu i est :

$$x_{i.}^{k} = \frac{1}{T} \sum_{t=1}^{T} x_{it}^{k}$$

- La moyenne inter-individuelle de la $k$ème variable à la date t est :

$$x_{.t}^{k} = \frac{1}{N} \sum_{i=1}^{N} x_{it}^{k}$$

- La moyenne inter-individuelle et inter-temporelle de la $k$ème variable est :

$$x_{..}^{k} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}^{k}$$

Décompositions de la variance

Savoir qu'elle est l'origine de la variance est souvent informatif. La dispersion des salaires est-elle due à des différences permanentes (qualification, grilles de salaires) ou temporaires ?

La variabilité des salaires est :

$$\text{varb } y = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - y_{..})^2$$

On peut écrire :

$$\text{varb } y = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - y_{i.} + y_{i.} - y_{..})^2$$

Décompositions de la variance

$$\text{varb}\, y = \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-y_{i.})^2+\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i.}-y_{..})^2+2\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-y_{i.})(y_{i.}-y_{..})$$

Le dernier terme est nul car $\sum_t(y_{it}-y_{i.})=0$ (les esthètes préféreront faire référence au théorème des projections successives). D'où :

$$\text{varb}\, y = \text{varb}_{\text{intra-individuelle}}\, y + \text{varb}_{\text{inter-individuelle}}\, y$$

Dans l'exemple des salaires, cette formule nous permet de décomposer la variabilité totale en :

- variabilité temporelle propre à l'individu (part variable...)
- variabilité permanente entre individus (formation initiale, talent...)

Décompositions de la variance

D'autres décompositions sont possibles

- On a :

$$\text{varb } y = \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - y_{.t})^2 + \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{.t} - y_{..})^2$$

D'où :

$$\text{varb } y = \text{varb}_{\text{intra-temporelle}}\ y + \text{varb}_{\text{inter-temporelle}}\ y$$

- Ou encore des décompositions en variabilité intra-individuelle-temporelle, inter-individuelle et inter-temporelle (exercice !)

Introduction

Avantages des données de panel (1/4)

- Elles proviennent de l'**accumulation** de données dans le temps. Les échantillons panels sont donc plus grands que n'importe quelle coupe issue de la même source, d'où une meilleur convergence ainsi qu'une précision accrue des estimateurs
- Ceci explique pourquoi on parlera notamment des estimateurs MCG, biaisés à distance finie mais néanmoins convergents
- Une nuance : on a généralement avec un panel micro $N \to \infty$ mais pas $T \to \infty$ (et l'inverse avec les panels macro). D'où une convergence dans la dimension individuelle de meilleure qualité que dans la dimension temporelle avec les panels micro.

Avantages des données de panel (2/4)

- Ce sont avant tout des **données individuelles**, on peut donc connaître le rôle des différences entre individus dans le comportement de la variable dépendante. Surtout, le **suivi** des individus nous renseigne sur la manière dont leurs situations évoluent dans le temps.
- Exemple : deux coupes peuvent nous permettre de mesurer un taux de chômage de 10% à deux dates, mais il nous faut un panel pour savoir s'il s'agit des mêmes personnes.
- En d'autres termes, la double dimension individuelle et temporelle permet de séparer les effets des caractéristiques individuelles des évolutions temporelles

Avantages des données de panel (3/4)

- On verra qu'on est également capables de mesurer l'impact des caractéristiques individuelles inobservables permanentes dans le temps
- Par exemple, à partir de données individuelles de salaires, on peut distinguer la part due aux caractéristiques observables (expérience professionnelle...), de celle due aux inobservables (motivation, implication, talent...), de celle due à la conjoncture.
- Autre exemple, un panel de pays nous permet de rendre compte de l'influence de différences structurelles (systèmes juridiques, politques...) sur leur croissance économique

Avantages des données de panel (4/4)

Intuition : Soit v_i une variable aléatoire inobservée permanente dans le temps. Un modèle linéaire serait :

$$y_{it} = \beta_0 + x_{it}^{'}\beta + v_i + \epsilon_{it}.$$

Si v_i est sans corrélation avec les x_{it}, on peut estimer le modèle par OLS (attention toutefois à l'hétéroscédasticité). Sinon, l'estimateur OLS n'est pas convergent. On peut toutefois écrire, grâce au suivi des individus :

$$\Delta y_{it} = \Delta x_{it}^{'}\beta + \Delta \epsilon_{it}.$$

L'estimateur OLS de ce modèle est convergent, si $\Delta x_{it}^{'}$ est inversible et si $\Delta x_{it}^{'}$ est sans corrélation avec $\Delta \epsilon_{it}$.

Inconvénients des données de panel (1/3)

- Ce sont avant tout des données individuelles, l'information est potentiellement riche, mais sa **fiabilité** est parfois douteuse
- Par exemple, de nombreuses données d'entreprises sont déclarées par des membres de l'entreprise qui ne sont pas statisticiens, qui ont souvent beaucoup d'autres choses à faire, et qui ont parfois intérêt à manipuler les informations qu'ils déclarent

Inconvénients des données de panel (2/3)

- Les influences des **observations aberrantes** ne se compensent généralement pas dans ce contexte (à l'inverse des données en coupe qui sont souvent plus clémentes). Les estimations peuvent être sensibles à un nombre, même faible, de points aberrants
- Les **observations manquantes ou incomplètes** tendent à être fréquentes en pratique. Celà s'explique souvent par les difficultés à effectuer le suivi longitudinal des individus, mais aussi parfois par des comportements stratégiques demandant à être explicitement intégrés dans le modèle, sous peine de biais sévères.

$\Rightarrow$ le repérage et la correction (ou l'élimination) des observations aberrantes et manquantes est encore plus important dans le cas des panels

Inconvénients des données de panel (3/3)

- Les comportements sont généralement stables dans le temps, d'où des modèles où les erreurs sont souvent autocorrélées. Les écarts-types des coefficients doivent être évalués avec attention, au risque d'avoir des t de Student fortement surévalués.

Introduction

Format des données (1/2)

Les données se présentent généralement dans le format de l'exemple plus haut, appelé format **long** :

$$\begin{array}{cccc} y_{11} & x_{11}^1 & \cdots & x_{11}^K \\ \vdots & \vdots & \cdots & \vdots \\ y_{1T} & x_{1T}^1 & \cdots & x_{1T}^K \\ \vdots & \vdots & \cdots & \vdots \\ y_{N1} & x_{N1}^1 & \cdots & x_{N1}^K \\ \vdots & \vdots & \cdots & \vdots \\ y_{NT} & x_{NT}^1 & \cdots & x_{NT}^K \end{array}$$

Les données peuvent ici être stockées dans une matrice $(NTx(K+1))$

Format des données (2/2)

Les données sont aussi parfois au format **large** :

$$\begin{array}{cccccccccc} y_{11} & \cdots & y_{1T} & x_{11}^1 & \cdots & x_{1T}^1 & x_{11}^K & \cdots & x_{1T}^K \\ y_{N1} & \cdots & y_{NT} & x_{N1}^1 & \cdots & x_{NT}^1 & x_{N1}^K & \cdots & x_{NT}^K \end{array}$$

Les données peuvent ici être stockées dans une matrice $(Nx(KT + T))$

La plupart des logiciels s'attendent à ce que les données soient au format long lorsqu'on appelle les fonctions propres aux données de panel. Si elles sont au format large : `reshape` (R et Stata).

Logiciels

- Les logiciels usuels d'économétrie (SAS, Stata, R...) permettent de traiter des données de panel et d'estimer assez facilement les modèles que nous verrons dans ce cours
- Sans vouloir trop déflorer le suspens, les estimateurs que nous verront reposent sur de l'algèbre linéaire et parfois des transformations simples de données. N'importe quel logiciel de calcul matriciel peut donc faire l'affaire.
- Pour les modèles plus avancés, je préfère personnellement Stata. À garder en tête si vous envisagez d'investir à plus long terme dans ce domaine.

Bibliographie

- Patrick Sevestre (2002) : *Économétrie des données de panel*, Dunod.
- Jeffrey Wooldridge (2008) : *Econometric Analysis of Cross Section and Panel Data*, MIT Press.
- Colin Cameron et Pravin Trivedi (2005) : *Microeconometrics - Methods and Applications*, Cambridge University Press.