Econométrie des données de panel: Introduction

Guillaume Horny*

*Banque de France

Master 2 MASERATI

Plan

Plan

Un peu de vocabulaire...

Que sont des données de panel ?

- des données relatives à des unités statistiques observées à plusieurs reprises dans le temps
- un panel est ainsi une répétition de coupes, dans lesquelles on suit 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).

Exemple de données de panel

Les données comptables collectées par Grunfeld, portant sur des entreprises de 1935 à 1954

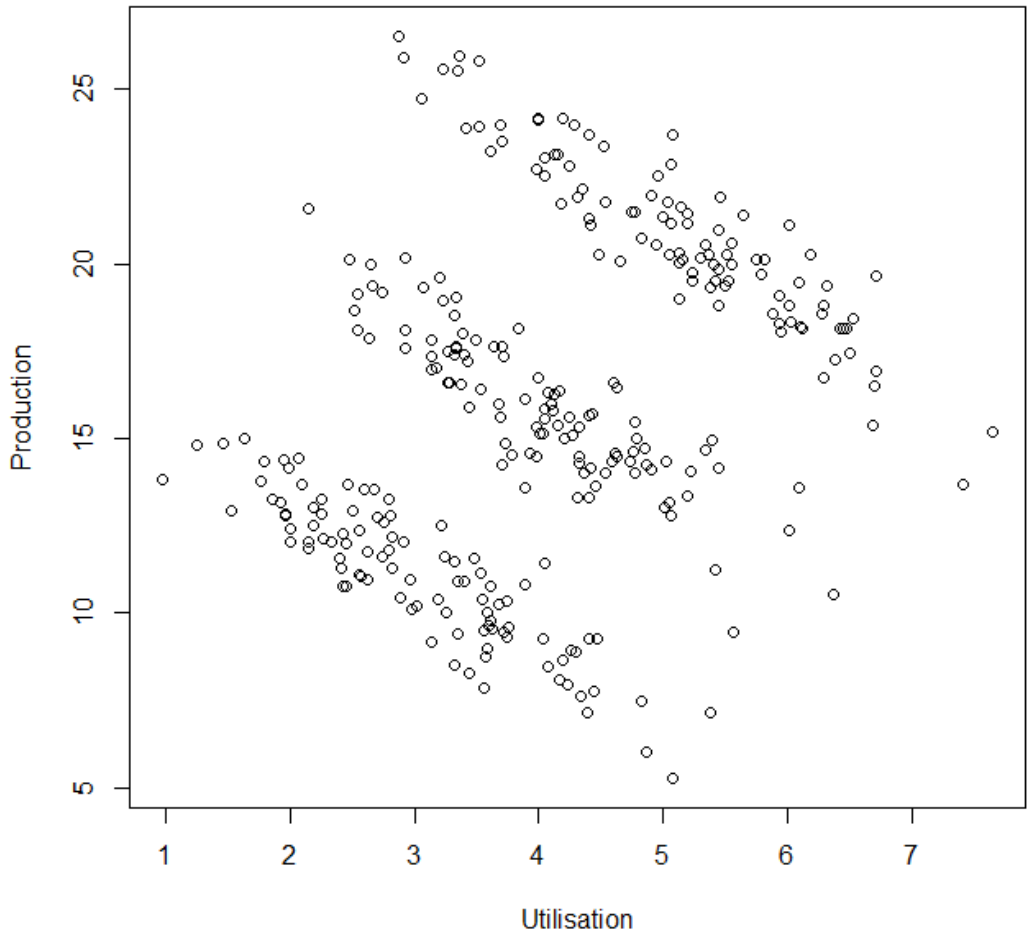
firm	year	inv	value	capital
1	1935	317.6	3078.5	2.8
1	1936	391.8	4661.7	52.6
1	1937	410.6	5387.1	156.9
.....				
1	1954	1486.7	5593.6	2226.3
2	1936	355.3	1807.1	50.5
2	1937	469.9	2676.3	118.1
2	1938	262.3	1801.9	260.2

Pourquoi un cours dédié ?

- les données de panel sont de fait très répandues à l'heure actuelle :
  - les assurances procèdent au suivi des clients/assurés et de leurs risques,
  - les services marketing suivent les comportements clients,
  - l'administration fiscale effectue un suivi des contribuables
  - le superviseur des banques suit les bilans des banques
  - ...
- les techniques mises au point pour des données en coupe, lorsqu'elles sont appliquées à un panel, produisent des résultats erronés

Exemple (1/4)

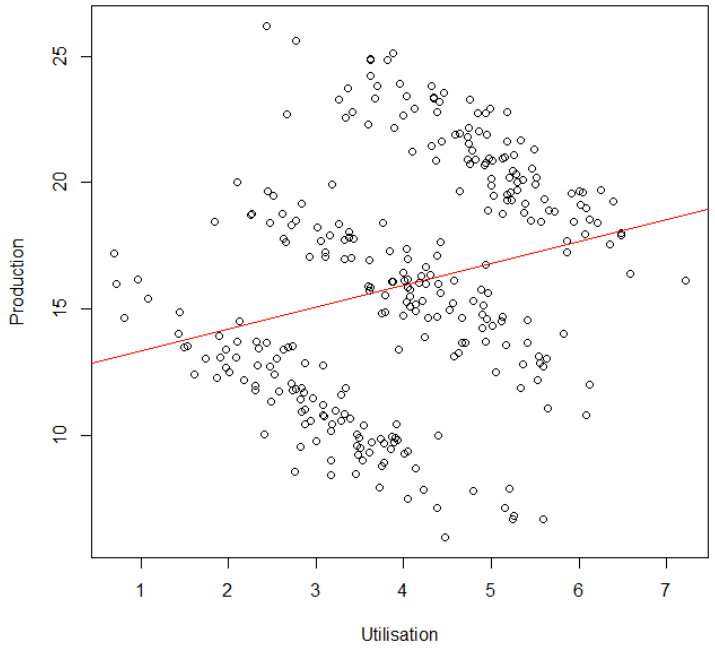
On a des observations sur la production (ordonnées) de machines en fonction du nombre d'heures d'utilisation (abscisses).



Exemple (2/4)

L'estimateur OLS :

- va passer par le point moyen
- s'ajuste à la forme du nuage en minimisant le carré des écarts entre chaque point et la droite de régression

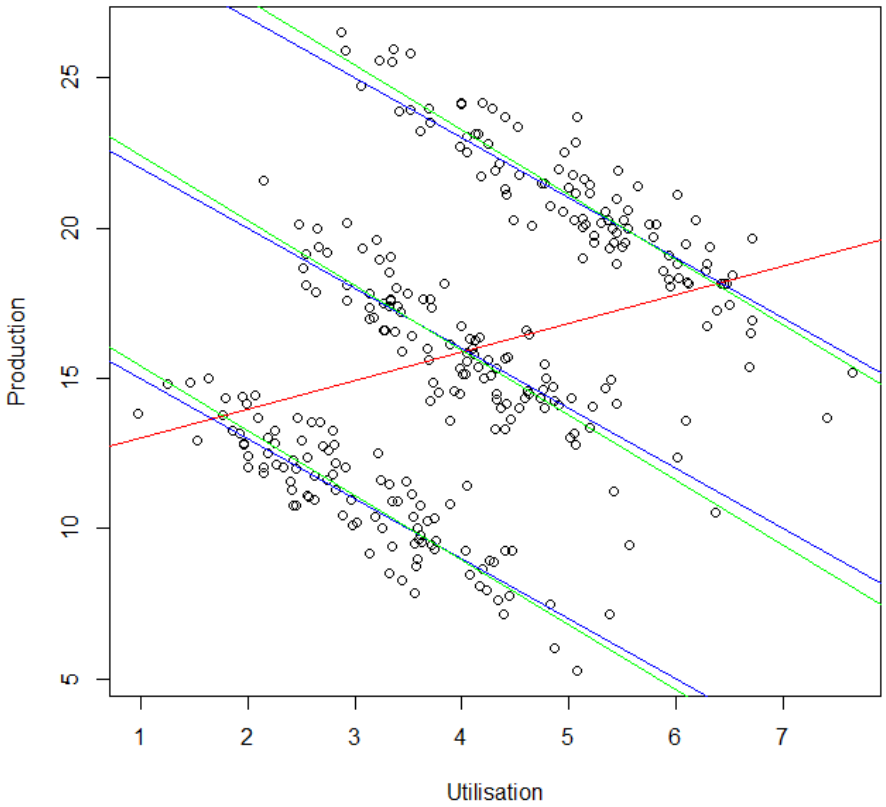


Exemple (3/4)

Que conclure si on apprend que :

- chaque groupe de points correspond à la production d'une machine, chacune appartenant à une génération différente d'équipement
- Le nombre d'heure d'utilisation est le cumul d'heure travaillées par la machine depuis qu'elle a été achetée : les points les plus à gauche sont les plus anciens, les points les plus à droites les plus récents.

Exemple (4/4)



Les droites :

- en rouge sont celles estimées par OLS
- en bleu correspondent à celles ayant servi à simuler les données
- en vert sont des régressions ajustées pour la dimension panel

Introduction

Avantages (1/4)

Ce sont avant tout des **données individuelles** :

- le **suivi** des individus nous renseigne sur la manière dont leurs situations évoluent dans le temps.
- à chaque date, on connait les différences entre individus

$\Rightarrow$ La double dimension individuelle et temporelle permet de séparer les effets des caractéristiques individuelles des évolutions temporelles communes à tous les individus.

Pourquoi ? La question de l'hétérogénéité

Statisticien et économètres tendent à raisonner en moyennes :

"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."

(Aalen, 1998)

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

(Heckman, 2001, discours de remise du prix Nobel)

Hétérogénéité observée et inobservée

- **l'hétérogénéité** renvoie aux facteurs connus des agents et pertinents lors de la prise de décision, et à leurs différences. 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 inobservée** lorsque ces différences sont inconnues de l'économètre. 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, l'hétérogénéité inobservée aux différences qui ne sont pas mesurées par des variables.

L'hétérogénéité inobservée

Les variables explicatives sont rarement toutes observées : certaines peuvent ne pas être mesurables, codifiables ou encore être absentes des données.
Exemple : l'implication d'un chômeur dans sa recherche d'emploi, la créativité d'une équipe de R&D, l'ambiance de travail...

L'analyste peut avoir conscience du problème, mais se trouver dans l'impossibilité de les prendre en compte dans son modèle. Prendre en compte l'hétérogénéité inobservée ne veut pas dire qu'on identifie les éléments inobservés, mais plutôt qu'on cherche à limiter les problèmes de variables omises.

Avantages (2/4)

- Ce sont les données idéales pour les comparaisons de groupes d'individus avant/après une réforme, un choc. Elles sont à la base des évaluations des politiques publiques, des effets de traitements...
- 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,
  $\Rightarrow$ meilleure convergence et plus de précision
- Ceci explique pourquoi on parlera notamment des estimateurs MCG, biaisés à distance finie mais néanmoins convergents

Avantages (3/4)

On verra qu'on est également capables de mesurer l'impact des caractéristiques individuelles inobservables permanentes dans le temps

- **Exemple 1 :** 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.
- **Exemple 2 :** Avec des 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.
- **Exemple 3 :** un panel de pays nous permet de rendre compte de l'influence de différences structurelles (systèmes juridiques, institutions politiques...) sur la croissance économique

Introduction

Inconvénients (1/4)

- Ce sont avant tout des données individuelles, l'information est potentiellement riche, mais sa **fiabilité** est parfois douteuse
- de nombreuses données d'entreprises sont déclarées par des membres de l'entreprise qui :
  - ont souvent beaucoup d'autres choses à faire,
  - ont parfois intérêt (ou pensent avoir intérêt) à manipuler les informations qu'ils déclarent
    Exemple : ne pas communiquer le bilan les mauvaises années ou à l'inverse être systématiquement pessimiste sur l'évolution de l'activité...

Inconvénients (2/4)

- Les influences des **observations aberrantes** ne se compensent généralement pas dans ce contexte, à l'inverse des données en coupe qui tendent à pardonner plus facilement ces erreurs de traitement des données. Les estimations sont 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 pouvant biaiser les estimations.

$\Rightarrow$ Repérer et corriger des observations aberrantes et manquantes, en éliminant ou en imputant des valeurs, est encore plus important dans le cas des panels

Inconvénients (3/4)

- 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.

Inconvénients (4/4)

Autre complication dans le calcul des écarts-types : la notion d'unité statistique n'est pas toujours évidente

Exemple :
L'expérience STAR au Tenessee a conduit à répartir 11 600 élèves d'école primaire et leurs enseignants en classes de taille "normale", petite classe et classes de taille normale avec un enseignant auxiliaire. L'expérience a commencé avec la vague de 1985 et a durée 4 ans. Après celà, tous les élèves ont été remis dans des classes "normales".

On dispose des notes des élèves et on connaît les classes auxquels ils ont été allouées. On a plusieurs explicative définies au niveau de la classe, mais aucune au niveau des élèves. Quelle est l'unité vraiment suivie dans le temps : des élèves ou des classes ?

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
- 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.

Annexes

Affichage de données sous R

```
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
```

Simulation de données de l'exemple

```
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, ]
```

Création du graphique

```
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")
```

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.