Econométrie des données de panel: Introduction

Guillaume Horny*

*Banque de France

Master 2 MASERATI

Plan

Plan

Données en coupe, séries temporelles et données de panel

- **Données en coupe** (*cross-sectional data*) :
  - sans dimension temporelle, elles sont relatives à des unités statistiques telles que des ménages, entreprises, salarié, pays...
  - l'ordre des données traduit seulement l'ordre des identifiants d'individus
- **Séries temporelles** (*time series*) :
  - avec une dimension temporelle, elles sont relatives à une variable d'une seule unité statistique observées à plusieurs occasions.
  - l'ordre temporel des données est important, car elles ne sont généralement pas indépendantes dans le temps
- **Données de panel (ou longitudinales)** (*panel data*) :
  - avec une dimension temporelle, elles sont relatives à plusieurs unités statistiques pour lesquelles on observe plusieurs variables à différentes occasions.
  - l'ordre temporel des données est important, car elles ne sont généralement pas indépendantes dans le temps

Que sont des données de panel ?

- **Définition :** "*des données de panel correspondent au suivi d'un échantillon d'unités statistiques dans le temps, et contiennent ainsi plusieurs observations pour chacun des individus de l'échantillon*" (adaptée de Hsiao 2003, première ligne de l'introduction)
- 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

Panels micro, panels macro

- **Panel micro :** un panel pour lequel le nombre d'individus excède largement le nombre de périodes ($T \ll N$)
- **Panel macro :** un panel pour lequel le nombre d'individus est proche du nombre de périodes ($N \simeq T$)

Exemples :

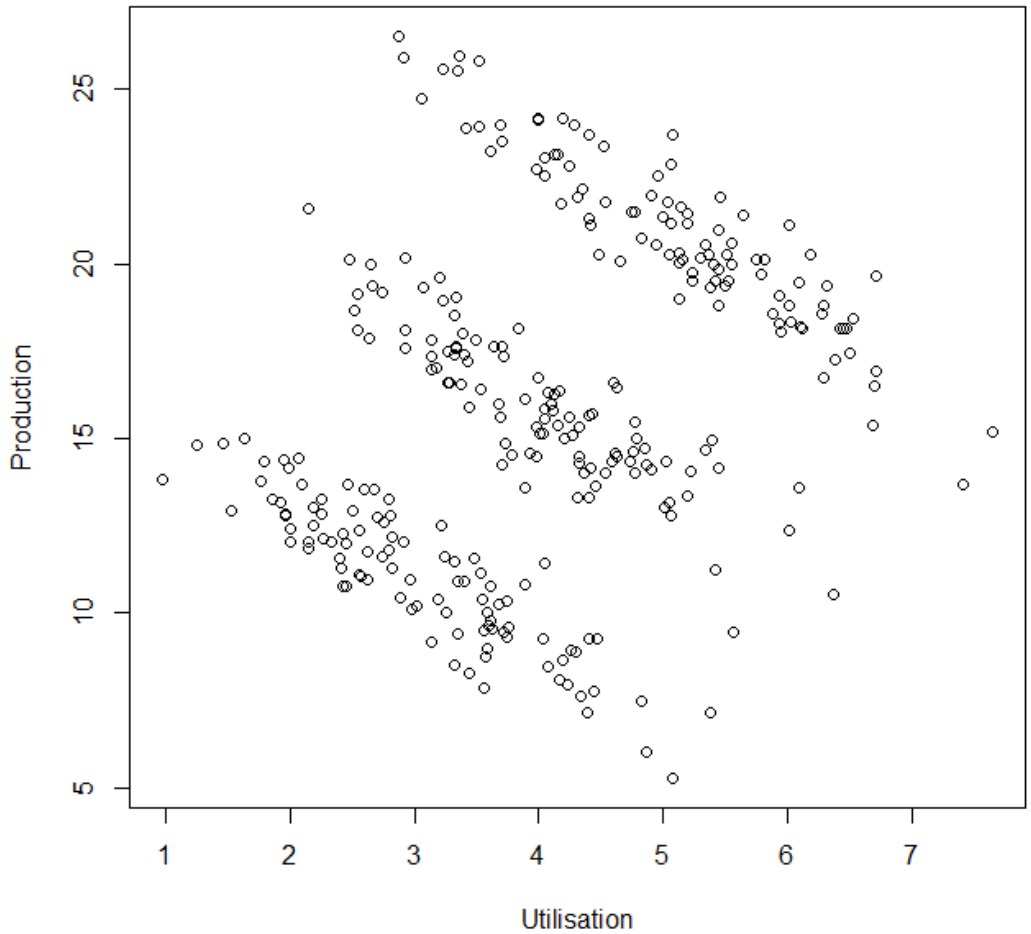
- Le *Bureau of Labor Statistics* finance le *Panel Study of Income Dynamics* (PSID). Le PSID a commencé en 1968 avec 4 800 familles. Le PSID a aujourd'hui collecté des informations sur plus 50 000 individus, certains étant observés sur plusieurs décennies...
- Des panels de pays peuvent facilement être constitués à partir de certaines sources (OCDE, FMI, Banque Mondiale, Eurostat...). Une vingtaine de pays pour lesquels on observe des variables trimestrielles sur 10 ans sera considéré comme un panel macro.

Pourquoi un cours dédié ?

- les données de panel sont de fait très répandues à l'heure actuelle :
  - les banques et assurances suivent les clients et leurs risques,
  - les services marketing suivent les achats des clients,
  - l'administration fiscale suit les revenus 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 (1/5)

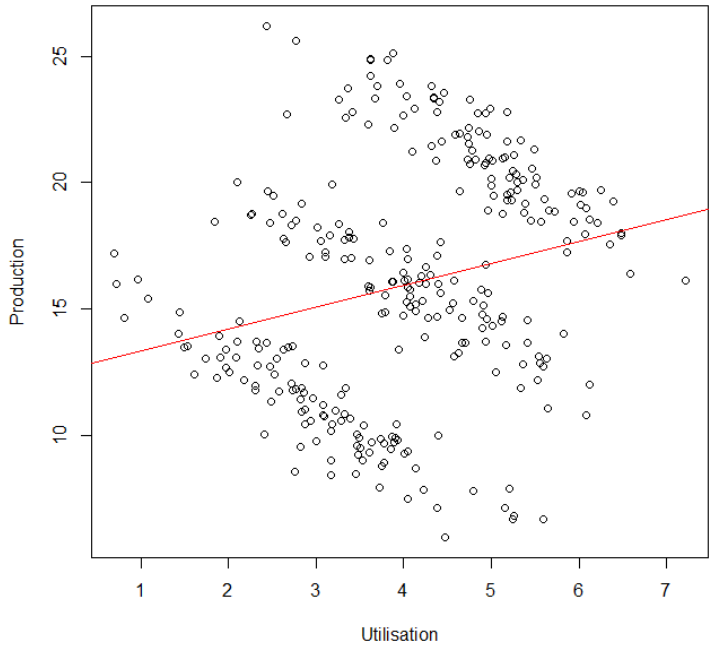
On a des observations sur la production (ordonnées) de machines en fonction de la durée de leur utilisation (abscisses, unité inconnue).



Exemple (2/5)

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

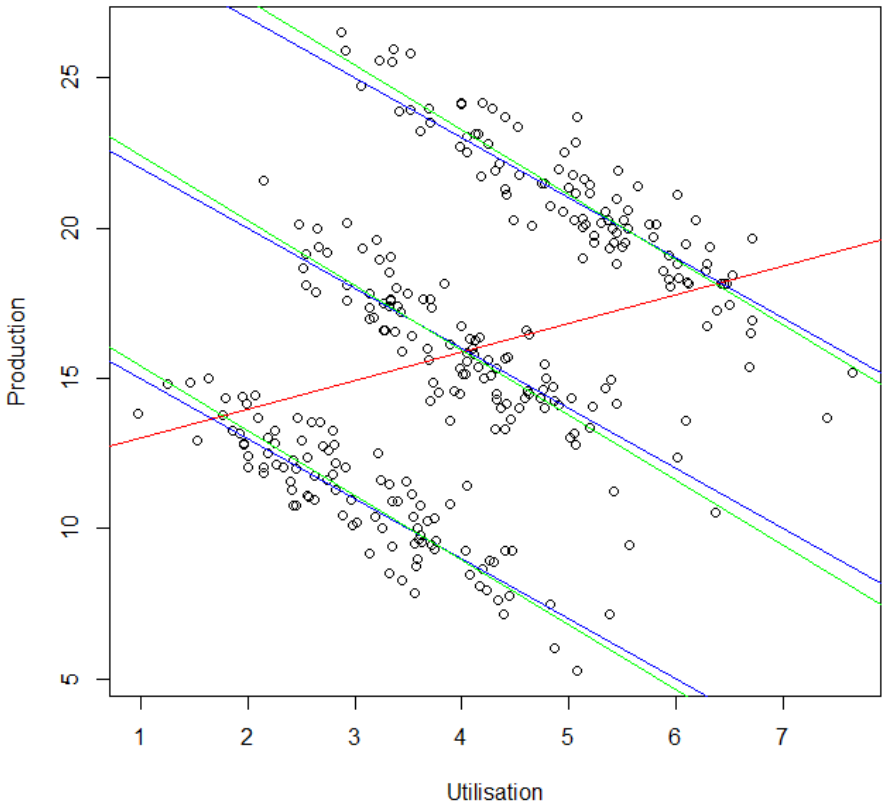


Exemple (3/5)

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 nombre d'heures travaillées par la machine chaque journée

Exemple (4/5)

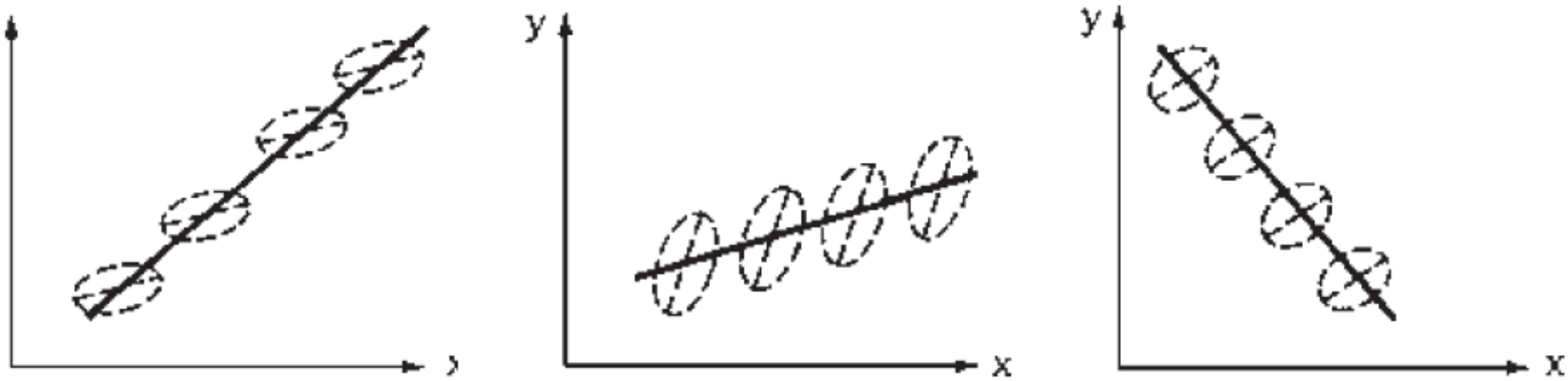


Les droites :

- 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

Exemple (5/5)

Ce type de biais d'hétérogénéité peut aller dans tous les sens et ne peut pas être évalué a priori à partir des données :



Source : Hsiao (2003), cité par Hurlin (2018).

Deaton (1995) sur la productivité agricole

- les petites exploitations sont-elles plus productives que les grandes ?
- des régressions de rendements sur les facteurs de production (surface, heures travaillées, engrais, formation de l'exploitant...) trouvent généralement un coefficient négatif pour la surface de l'exploitation
- Explication économique : difficultés à contrôler l'effort des ouvriers agricoles (aléa moral)
- Explication de Deaton : ces régressions souffrent d'un problème de variable omise, à savoir la qualité des sols, qui est systématiquement corrélée avec la production. Les exploitations sont de grande taille en environnement semi-désertique tandis qu'elles sont petites là où les sols sont fertiles. Au temps pour l'aléa moral...

Introduction

Avantages

1. le **suivi** des individus dans le temps nous permet d'accumuler des observations, et donc d'avoir des échantillons de grande taille relativement à une coupe
2. Identification des effets causaux
3. Hétérogénéité inobservée

Avantage 1 : la taille de l'échantillon

- Avec plus d'observations, on a :
  - moins de problème de colinéarité entre les variables explicatives
  - une meilleure convergence des estimateurs
  - des écarts-types plus petits
- la relativement grande taille des échantillons explique pourquoi on parlera des estimateurs MCG, biaisés à distance finie mais convergents
- Attention toutefois à ne pas mélanger dans l'échantillon des individus avec des comportements trop hétérogènes, ou encore des périodes où les comportements se modifient fortement, au risque de ne plus savoir pour quelle population/période on identifie un comportement (biais de variable omise).

Avantage 2 : identification des effets causaux (1/2)

Les méthodes Diff-in-Diff viennent des panels à deux périodes. En effet, la double dimension individuelle et temporelle nous renseigne :

- sur la manière dont les situations des individus évoluent dans le temps
- sur les caractéristiques des individus et ce qui les distingue les uns des autres à chaque date

$\Rightarrow$ elle nous permet d'identifier les effets causaux chocs, de traitements, de politiques...

Avantage 2 : identification des effets causaux (2/2)

Exemple : un échantillon en coupe nous indique que le taux de chômage est de l'ordre de 15% sur une année de récession suivant une crise financière

- est-ce que ça veut dire que chaque individu a 15% de chances de perdre son emploi l'année qui suit une crise financière ?
- ou bien que les 15% des individus qui ne travaillaient pas avant le choc financier resteront au chômage durant la récession ?

$\Rightarrow$ besoin d'un suivi des individus pour estimer la contribution du choc financier à la probabilité d'entrer au chômage

Avantage 3 : hétérogénéité observée et inobservée (1/2)

- **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, qui manque ainsi 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.

Avantage 3 : hétérogénéité observée et inobservée (2/2)

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...

- les omettre créé un biais de variable omise si elles sont corrélées avec les variables explicatives observées
- Prendre en compte l'hétérogénéité inobservée ne veut pas dire qu'on va chercher à mesurer des éléments inobservés, mais plutôt qu'on cherche à limiter les biais de variables omises.
- les données de panel permettent de prendre en compte l'hétérogénéité inobservée de façon simple, au moyen d'effets fixes ou aléatoires

Introduction

Inconvénients

1. l'importance de la qualité des données
2. hétérogénéité inobservée (à nouveau !)
3. l'évaluation des écarts-types

Inconvénient 1 : la qualité des données (1/2)

- 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 et préfèrent une déclaration rapide à une déclaration précise (erreurs de mesure),
  - 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 (biais de sélection)...

Inconvénient 1 : la qualité des données (2/2)

- 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

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

Inconvénient 2 : hétérogénéité inobservée

- Lorsque des caractéristiques des individus expliquant leurs comportements sont inobservées, l'hypothèse que la variable dépendante est générée par une distribution dont les paramètres sont les mêmes pour tous les individus est souvent erronée.
- le traitement standard du problème avec un panel est de spécifier des effets fixes ou aléatoires
- ce peut être insuffisant, par exemple si $\beta > 0$ pour certains individus et $\beta < 0$ pour d'autres (exemple : liquidités des entreprises et délais de paiement). Dans ce cas, on aura alors un problème de biais du même type que dans l'exemple de la durée d'utilisation des machines

Inconvénient 3 : le calcul des écarts-types (1/2)

- 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 en conséquence, au risque d'avoir des t de Student fortement surévalués.

Inconvénient 3 : le calcul des écarts-types (2/2)

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.