Econométrie des données de panel: Rappels

Guillaume Horny*

*Banque de France

Master 2 MASERATI

Rappels

Plan

Plan

Les sommes

Soit $\{x_i : i = 1, \ldots, n\}$ une suite de n nombres. Leur somme s'écrit :

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \ldots + x_n,$$

La somme a les propriété suivantes :

- Pour toute constante c, on a :

$$\sum_{i=1}^{n} c = nc,$$

$$\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i$$

- Pour a et b constants et $\{(x_i, y_i) : i = 1, \ldots, n\}$

$$\sum_{i=1}^{n} (ax_i + by_i) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i$$

La moyenne

Leur moyenne s'écrit :

$$x_{.} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

Propriétés :

- la somme des écarts à la moyenne est nulle :

$$\sum_i (x_i - x_{.}) = \sum_i x_i - \sum_i x_{.} = \sum_i x_i - nx_{.} = nx_{.} - nx_{.} = 0.$$

- on peut aussi montrer :

$$\sum_i (x_i - x_{.})^2 = \sum_i x_i^2 - nx_{.}^2$$

- ou encore :

$$\sum_i (x_i - x_{.})(y_i - y_{.}) = \sum_i x_i y_i - nx_{.}y_{.}$$

Plan

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 !)

Plan

Espérance

Soit X une variable discrète prenant les valeurs $\{x_1, \ldots, x_k\}$. Sa densité de probabilité est $f_X()$. Son espérance est :

$$\mathrm{E}(X) = x_1 f_X(x_1) + x_2 f_X(x_2) + \ldots + x_k f_X(x_k) \equiv \sum_{j=1}^{K} x_j f_X(x_j).$$

Lorsque X est continue :

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x f_X(x) dx.$$

Pour toute fonction $g(.)$ de X, on a :

$$\mathrm{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx.$$

Densité conditionnelle

On s'intéresse généralement aux relations entre une variable aléatoire Y et une ou plusieurs autres. Supposons qu'il n'y ait qu'une seule autre variable aléatoire appelée X. L'influence de X sur Y est visible dans la distribution conditionnelle de Y sachant X.

La densité de probabilité de Y conditionnellement à X est :

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \forall x \text{ tel que } f_X(x) > 0.$$

Lorsque X et Y sont discrètes :

$$f_{Y|X}(y|x) = \mathrm{P}(Y = y|X = x).$$

Espérance conditionnelle

Supposons que l'on connaisse $f_{Y/X}(.)$ et la valeur x prise par X. On peut alors calculer $\mathrm{E}(Y|X=x)$, parfois notée de façon plus courte $\mathrm{E}(Y|X)$. Généralement, lorsque x change, $\mathrm{E}(Y|X)$ change également.

Pour Y discrète, on a :

$$\mathrm{E}(Y|X) = \sum_{j=1}^{K} y_j f_{Y|X}(y_j|x).$$

Il s'agit toujours d'une moyenne pondérée des y_j, mais les poids dépendent ici de la valeur prise par X.

Pour Y continue :

$$\mathrm{E}(Y|X) = \int_{-\infty}^{+\infty} y f_{Y|X}(y|x)dy.$$

Espérance conditionnelle : Exemple

Supposons que (X, Y) représente l'ensemble des salariés, avec X une mesure des années d'étude et Y le niveau de salaire. Le salaire moyen des personnes ayant étudié 12 ans (atteint le bac) est alors mesuré par $\mathrm{E}(Y|X = 12)$. On peut faire ce calcul pour différentes valeurs de X et voir ainsi comment le salaire est relié au niveau d'études.

On peut faire le calcul toutes les durées d'éducation et reporter les résultats dans un tableau. Par simplicité, on préfère typiquement supposer que l'espérance conditionnelle prend la forme d'une fonction standard. Si on suppose qu'elle est linéaire :

$$\mathrm{E}(\mathit{salaire}|\mathit{etudes}) = \beta_0 + \beta_1 \mathit{etudes}.$$

Si cette hypothèse est raisonnable, alors chaque année supplémentaire d'étude rapporte en moyenne β_1. Le salaire moyen après 17 ans d'études est $\beta_0 + 17\beta_1$.

Propriétés des espérances conditionnelles

- $\mathrm{E}(Y|X) = \mathrm{E}(Y)$ ssi X et Y sont indépendants
- **Loi des espérances itérées :** $\mathrm{E}_X[\mathrm{E}_{Y|X}(Y|X)] = \mathrm{E}_Y(Y)$.

Preuve de la loi des espérances itérées : écrire les espérance sous forme d'intégrales ou de sommes.

Illustration de la loi des espérances itérées

Soit $Y = salaire$ et $X = etudes$. Supposons que :

$$\mathrm{E}(salaire|etudes) = 1130 + 50etudes.$$

La loi des espérances itérées implique :

$$\begin{aligned}\mathrm{E}(salaire) &= \mathrm{E}_{etudes}(1130 + 50etudes) \\ &= 1130 + 50\mathrm{E}_{etudes}(etudes).\end{aligned}$$

On a de plus $\mathrm{E}(etudes) = 12$. Alors :

$$\mathrm{E}(salaire) = 1130 + 50 * 12 = 1730.$$

Note : $\mathrm{E}(salaire|etudes) = g(etudes)$. Elle dépend de *etude* et pas de *salaire* !

Propriétés des espérances conditionnelles (suite)

Théorème des projection successives :

$$\mathrm{E}(Y|X) = \mathrm{E}_{Z|X}[\mathrm{E}_{Y|X,Z}(Y|X,Z)|X].$$

On peut obtenir $\mathrm{E}(Y|X)$ en deux étapes :

1. Trouver l'expression de $\mathrm{E}_{Y|X,Z}(Y|X,Z)$, où Z est n'importe quelle variable aléatoire
2. Prendre son espérance conditionnelle à X.

Ce qui est important ici est que $X \subset (X,Z)$. Intuitivement, la première espérance dépend à la fois de X et de Z, elle est plus "détaillée" que la seconde qui ne retient que l'influence de X. En prenant la seconde espérance, on prend la moyenne par rapport à Z, ce qui permet de l'évacuer du calcul.

Plan

Modèle

Le modèle linéaire pour l'individu $i (i = 1, \ldots, N)$ s'écrit :

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK} + \epsilon_{it}.$$

Son écriture matricielle est :

$$Y = X\beta + \epsilon,$$

où Y est un vecteur $(Nx1)$, X une matrice $Nx(K+1)$, β un vecteur de dimension $((K+1)x1)$ et ϵ un vecteur $(Nx1)$.

Estimateur OLS

La condition de premier ordre de la minimisation par rapport à β de $\sum_i (y_i - x_i\beta)^2$ s'écrit sous forme matricielle :

$$X'(Y - X\widehat{\beta}) = 0.$$

Elle a pour solution :

$$\widehat{\beta} = (X'X)^{-1}X'Y.$$

Remarques

- C'est uniquement lorsque X est une matrice carrée qu'on peut écrire :

$$\widehat{\beta} = (X'X)^{-1}X'Y = X^{-1}(X')^{-1}X'Y = X^{-1}Y.$$

- A l'optimum, les conditions de premier ordre s'écrivent sous forme matricielle :

$$\begin{aligned} X'(Y - X\widehat{\beta}) &= 0. \\ \Leftrightarrow X'\widehat{\epsilon} &= 0. \end{aligned}$$

La covariance entre n'importe quelle variable explicative et les résidus OLS est toujours nulle.

Propriété : Absence de biais

$$\begin{aligned}\widehat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + \epsilon) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon \\ &= \beta + (X'X)^{-1}X'\epsilon.\end{aligned}$$

Si on suppose $\mathsf{E}(\epsilon|X) = 0$, alors :

$$\begin{aligned}\mathsf{E}(\widehat{\beta}|X) &= \beta + (X'X)^{-1}X'\mathsf{E}(\epsilon|X) \\ &= \beta.\end{aligned}$$

Expression de la variance

Sous l'hypothèse d'homoscédasticité ($\text{var}(\epsilon) = \sigma^2 Id$) :

$$\begin{aligned}
\text{var}(\widehat{\beta}|X) &= \text{var}[(X'X)^{-1}X'Y] \\
&= \text{var}[\beta + (X'X)^{-1}X'\epsilon] \\
&= \text{var}[(X'X)^{-1}X'\epsilon] \\
&= (X'X)^{-1}X'\text{var}(\epsilon)X(X'X)^{-1} \\
&= (X'X)^{-1}X'(\sigma^2 Id)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}.
\end{aligned}$$

Un estimateur sans biais de la variance est :

$$\widehat{\sigma}^2 = \frac{\widehat{\epsilon}'\widehat{\epsilon}}{N - K - 1}.$$

Théorème de Gauss-Markov

Sous les hypothèses $\mathrm{E}(\epsilon|X) = 0$ et d'homoscédasticité, on a :

Théorème : L'estimateur OLS est l'estimateur linéaire sans biais de plus petite variance.

Propriété : Normalité asymptotique

Sous l'hypothèse :

$$\epsilon \sim N(0, \sigma^2).$$

On a :

$$\widehat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1}).$$

Sous l'hypothèse de modèle linéaire à erreurs gaussiennes homoscédastiques, on peut montrer que l'estimateur OLS est un estimateur du maximum de vraisemblance.