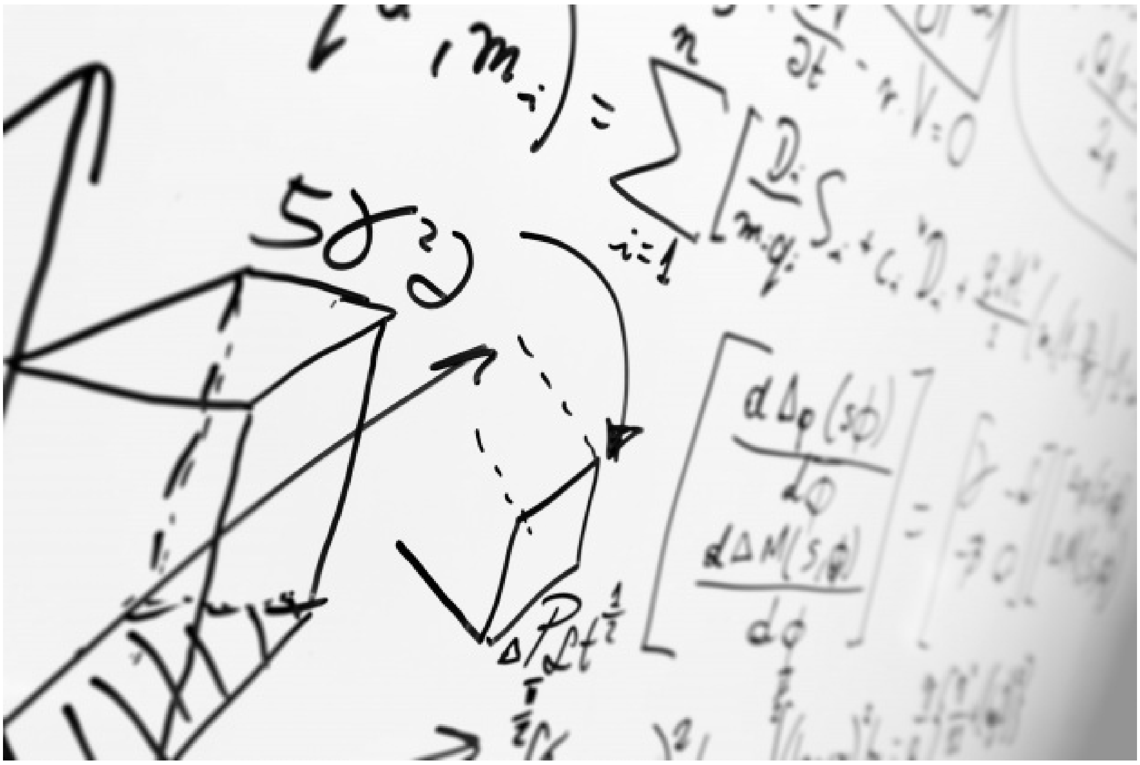
Leçon 5 : Introduction aux données de Panel

M. KEHO

04 /05/2017
1.0

Table des matières

Introduction



Il existe trois types d'échantillons de données. On distingue en premier lieu les données temporelles ou séries chronologiques où les variables représentent des phénomènes observés à intervalles réguliers. C'est ce type de données qu'on utilise dans la plupart des applications en macroéconomie lorsqu'on travaille sur un pays donné. On a en second lieu les données en coupe instantanée où les variables représentent des phénomènes observés au même instant sur plusieurs individus. Il s'agit généralement des données d'enquête ponctuelle auprès d'individus, de ménages ou d'entreprises. En troisième lieu, on a les données de panel dans lesquelles les variables sont observées sur plusieurs individus et sur plusieurs périodes. Les panels combinent donc les dimensions temporelle et individuelle des données. L'utilisation des panels permet de contourner la difficulté liée au manque de données longues dans la dimension temporelle. Elle permet de rendre plus puissants les tests lorsqu'on augmente la dimension individuelle. Cependant, l'analyse des données de panel requiert des procédures d'estimation très précises et fait apparaître des difficultés quant au traitement de l'hétérogénéité individuelle. Elle constitue aujourd'hui une spécialité dans l'économétrie (économétrie des données de panels) qui a donné lieu à de nombreux développements. Ici, nous en faisons une brève présentation.

Spécification d'un modèle en données de panel

I

Considérons la fonction de consommation du chapitre 11. La spécification en panel s'écrit :

$C_{it} = a_0 + a_1 R_{it} + a_2 P_{it} + a_3 G_{it} + e_{it}$ (15.1) $C_{it} = a_0 + a_1 R_{it} + a_2 P_{it} + a_3 G_{it} + e_{it}$ (15.1) $C_{it} = a_0 + a_1 R_{it} + a_2 P_{it} + a_3 G_{it} + e_{it}$ (15.1)

[redacted]

où l'indice *i* est pour l'individu i (par exemple le pays) et l'indice t pour la période t (l'année).

On suppose qu'il existe *n* individus et *T* périodes.

On suppose que l'échantillon est cylindré ou complet (balanced) : Chacun des n individus est observé sur T périodes de sorte qu'on dispose de T observations pour chacun des n individus, ce qui fournit *N* x T observations. Lorsque le panel n'est pas cylindré, il peut exister des problèmes d'hétéroscédasticité et/ou d'autocorrélation des erreurs aléatoires. Aujourd'hui, la plupart des logiciels économétrique sont capables de gérer des panels non cylindrés.

Pour tirer profit de la double dimension, individuelle et temporelle des données, différentes spécifications ont été proposées.

- **Modèles à effets fixes**

Ces modèles supposent une uniformité des coefficients d'un individu à l'autre sauf le terme constant : $C_{it} = \alpha_i + \phi + a_1 R_{it} + a_2 P_{it} + a_3 G_{it} + e_{it}$ (15.2) $C_{it} = \alpha_i + \phi + a_1 R_{it} + a_2 P_{it} + a_3 G_{it} + e_{it}$ (15.2) $C_{it} = \alpha_i + \phi + a_1 R_{it} + a_2 P_{it} + a_3 G_{it} + e_{it}$ (15.2) [redacted]

α_i est l'effet individuel, appelé « effet spécifique » de l'individu *i*. Il permet de capter l'hétérogénéité individuelle. Il est possible d'inclure un effet temporel non aléatoire.

- **Modèles à effets aléatoires**

$C_{it} = \phi + a_1 R_{it} + a_2 P_{it} + a_3 G_{it} + e_{it}$ (15.3) $C_{it} = \phi + a_1 R_{it} + a_2 P_{it} + a_3 G_{it} + e_{it}$ (15.3) $C_{it} = \phi + a_1 R_{it} + a_2 P_{it} + a_3 G_{it} + e_{it}$ (15.3)

[redacted]

avec $e_{it} = \alpha_i +$ $\mu_{it}\mu_{it}\mu_{it}$ [redacted], où α_i et $\mu_{it}\mu_{it}\mu_{it}$ [redacted] sont des perturbations aléatoires non corrélées. En fait, l'erreur du modèle est composée de deux termes :

α_i: effet individuel

$\mu_{it}\mu_{it}\mu_{it}$ [redacted]: effet résiduel

D'où le nom de modèle à « erreurs composées ». Dans cette spécification, on considère comme aléatoire, c'est-à-dire une perturbation propre à chaque individu.

- **Modèles à coefficients variables**

On suppose que les coefficients varient d'un individu à l'autre et d'une période à l'autre.

Le modèle est spécifié comme suit :

$C_{it} = \phi + a_{1it} R_{it} + a_{2it} P_{it} + a_{3it} G_{it} + e_{it}$ (15.4) $C_{it} = \phi + a_{1it} R_{it} + a_{2it} P_{it} + a_{3it} G_{it} + e_{it}$ (15.4) $C_{it} = \phi + a_{1it} R_{it} + a_{2it} P_{it} + a_{3it} G_{it} + e_{it}$ (15.4)

[redacted]

$$a_{kit} - b_k + a_{ki} + a_{kt} a_{kit} = b_k \mid a_{ki} \mid a_{kt} a_{kit} - b_k + a_{ki} + a_{kt}$$

avec [redacted]

Ici, deux individus ayant les mêmes caractéristiques observables et faisant face au même environnement n'auront pas nécessairement, en espérance, la même consommation.

Tests de stationnarité

II

Comme dans le cas des séries temporelles, il existe des tests de stationarité pour les modèles de données de panel. Les principaux tests de stationnarité (racine unitaire) sur données de panel sont :

- Test de Levin, Lin et Chu (2002)
- Test de Im, Pesaran et Shin (2003)
- Test de Breitung (2000)
- Test de Maddala et Wu (1999)
- Test de Choi (2001)
- Test de Hadri (2000)
- Tests ADF et PP

Ces tests reposent sur la spécification autorégressive suivante :

$y_{it} = \rho_i y_{it-1} + X_{it}\delta_i + e_{it}$ (15.4) $y_{it} = \rho_i y_{it-1} + X_{it}\delta_i + e_{it}$ (15.4) $y_{it} = \rho_i y_{it-1} + X_{it}\delta_i + e_{it}$ (15.4)

[redacted]

$|\rho_i| = 1 |\rho_i| \quad 1 |\rho_i| = 1$

La série admet une racine unitaire si [redacted].

Les tests précédents diffèrent suivant l'hypothèse sur la constance ou non du coefficient autorégressif p_i . Les tests de Levin, Lin et Chu (LLC), Breitung et Hadri supposent que $p_i = p$.

Les tests IPS, ADF et PP supposent que le coefficient p_i varie selon les individus.

Les tests de LLC et Breitung considèrent la spécification de l'équation ADF suivante :

$\Delta y_{it} = \alpha y_{it-1} + \sum_{j=1}^{p_i} \beta_{ij} \Delta y_{it-j} + X_{it}\delta + e_{it}$ (15.5) $\Delta y_{it} = \alpha y_{it-1} + \sum_{j=1}^{p_i} \beta_{ij} \Delta y_{it-j} + X_{it}\delta + e_{it}$ (15.5) $\Delta y_{it} = \alpha y_{it-1} + \sum_{j=1}^{p_i} \beta_{ij} \Delta y_{it-j} + X_{it}\delta + e_{it}$ (15.5)

[redacted]

On teste alors :

$H_0 : \alpha = 0$ contre $H_a : \alpha < 0$ $H_0 : \alpha = 0$ contre $H_a : \alpha < 0$ $H_0 : \alpha = 0$ contre $H_a : \alpha < 0$

[redacted]

Les tests de Levin, Lin et Chu (LLC) et de Breitung considèrent comme hypothèse nulle l'existence d'une racine unitaire tandis que le test de Hadri considère l'absence de racine unitaire comme hypothèse nulle. Il est donc similaire au test KPSS.

L'application de ces tests dans Eviews ne pose aucune difficulté particulière.

Estimation et interprétation

III

Il existe plusieurs procédures pour estimer un modèle en données de panel. Pour le modèle à effet fixe, on peut utiliser les variables muettes et estimer le modèle par les moindres carrés ordinaires. On peut également utiliser le théorème de Frisch-Waugh en appliquant les MCO sur les variables transformées en écarts à la moyenne individuelle. Dans le cas du modèle à effets aléatoire, la bonne méthode d'estimation est celle des moindres carrés généralisés (MCG).

Un modèle en panel peut s'estimer aisément sur les logiciels Eviews et STATA. Après l'estimation, il faut procéder aux différents tests. En particulier, il faut choisir entre un modèle à effets fixes et un modèle à effets aléatoires. Pour cela, on recourt au test d'Hausman. Dans Eviews, cela se fait en cliquant dans l'onglet **View/Fixed/Random Effects Testing/Correlated Random Effects - Hausman Test...**

Les coefficients d'un modèle en données de panel s'interprètent de la même façon comme dans les modèles en données strictement temporelles. Dans le cas d'un modèle à correction d'erreur estimé par la méthode en deux étapes, il est possible de calculer les coefficients de court terme et ceux de long terme.