La régression sur données de panel

I. Définition

Les données utilisées en économétrie sont le plus souvent des séries chronologiques, tel le nombre de naissances enregistrées par an dans le département de la Mayenne de 1953 à 2007. On peut à l'inverse disposer de données concernant une période donnée, dites *en coupe instantanée*, tel le nombre de naissances enregistrées en 1971 dans chacun des départements français.

Les *données de panel*, ou *données croisées*, possèdent les deux dimensions précédentes et rapportent les valeurs des variables considérées relevées pour un ensemble, ou *panel*, d'individus sur une suite de périodes.

Ainsi la série de panel mesurant la natalité annuelle par département de 1953 à 1997.

Remarque : si en marketing ou en statistique, le mot panel désigne généralement un échantillon fixe de consommateurs interrogés à différentes périodes, en économétrie, le terme de données de panel est simplement synonyme de données croisées ayant généralement une dimension temporelle.

On utilise une notation naturelle à deux indices : x_{it} note l'observation de la variable x pour l'individu i à la période t.

Si on fixe l'individu observé, on obtient la série chronologique, ou *coupe longitudinale*, le concernant, tandis que si l'on fixe la période examinée, on obtient une *coupe transversale*, ou *instantanée*, pour l'ensemble des individus.

Il est possible d'envisager des données croisées de plus de deux dimensions, on y reviendra dans l'avant-dernier paragraphe.

De nombreux modèles économétriques, notamment dans le domaine des études internationales, peuvent être confrontés à des données croisées, le caractère particulier de celles-ci invite à considérer des spécifications et des méthodes d'estimation adaptées.

II. Modèles et méthodes

Différents modèles ont été proposés, certains apparaissent également dans d'autres chapitres de l'économétrie et la terminologie n'est pas totalement

standardisée. Il convient par suite, lorsqu'on lit un article et plus encore si l'on souhaite mettre en œuvre la méthode indiquée, de s'assurer que l'on a bien identifié celle-ci.

On considère, pour fixer les idées, l'équation économétrique

(1) $y = a + b.x + c.z + \varepsilon$

De variable endogène y, d'explicatives x et z (et la constante), de coefficients a, b et c, et d'aléa ε, supposée s'appliquer aux données de panel étudiées.

La relation s'écrit encore, pour l'observation it :

(2) $y_{it} = a + b.x_{it} + c.z_{it} + \varepsilon_{it}$

La modélisation particulière porte uniquement sur la spécification des aléas ε_{it}. La forme de base s'écrit simplement :

(3) $\varepsilon_{it} = u_i + v_t + w_{it}$

où u_i désigne un terme, constant au cours du temps, ne dépendant que de l'individu i, v_t un terme ne dépendant que de la période t, et w_{it} un terme aléatoire croisé.

La suite dépend des hypothèses retenues quant aux composantes u_i, v_t et w_{it} et à leurs relations…

Méthode "naïve"

Une première méthode, naïve, consiste à appliquer simplement les mco sur l'ensemble des données mises bout-à-bout sans se préoccuper de leur nature particulière ni de celle de l'aléa ε.

Modèle à effets fixes

Ce modèle, également appelé *modèle de la covariance*, suppose que u_i et v_t sont des effets constants, non aléatoires, qui viennent donc simplement modifier la valeur de la constante a de l'équation (1) selon les valeurs de i et de t.

L'estimation s'opère par les mco, après ajout aux explicatives des indicatrices, ou dummy variables, associées aux individus i et aux périodes t (moins un individu et une période pour ne pas créer de colinéarité avec la constante).

Si on suppose que les perturbations aléatoires croisées w_{it} satisfont aux hypothèses classiques des mco (ie elles sont centrées, homoscédastiques, indépendantes et normales), les estimations sont optimales et permettent notamment les tests de Fisher pour éprouver la nécessité des termes u_i ou v_t.

Modèle à effets aléatoire

Ce modèle, encore appelé *modèle à erreur composée*, suppose les u_i et v_t véritablement aléatoires. La spécification de base suppose :

- les u_i, v_t et w_{it} centrés (ie d'espérance nulle)
- les u_i, v_t et w_{it} homoscédastiques et d'écarts type respectifs σ_u, σ_v et σ_w
- les u_i, v_t et w_{it} non corrélés et indépendants les uns des autres.

L'idée de cette modélisation est que les trois effets ne s'exercent plus sur la constante du modèle (1), mais véritablement sur la perturbation aléatoire ε. La méthode vise ensuite à préciser ces effets pour en tenir compte pour affiner l'estimation.

Sous les hypothèses indiquées, la variance de l'alea ε est :

(4) $\quad \text{var}(\varepsilon) = \sigma_u^2 + \sigma_v^2 + \sigma_w^2$

L'estimation du modèle, tels les doubles moindres carrés ou la méthode des variables instrumentales, procède en deux étapes : la première estime les composantes de la variance apparaissant dans la relation (4), ces estimations sont ensuite utilisées pour estimer l'équation (1) par les moindres carrés généralisés, la structure de variance-covariance des aléas étant approximativement connue.

Bien que les modèles à effets fixes et à effets aléatoires paraissent de nature différentes, le second est généralement recommandé. Des tests qu'on ne détaillera pas ici permettent d'éprouver les deux hypothèses. Si enfin l'objectif principal est l'estimation des coefficients des variables autres que la constante et s'ils diffèrent peu, la question du choix perd de son acuité.

En effets fixes comme en effets aléatoires, les économètres commencent généralement par estimer et tester un modèle avec le seul *effet individu*, l'*effet temps* étant souvent inexistant ou très mineur.

Modèle autorégressif

On a vu dans les leçons précédentes la fréquence des problèmes d'autocorrélation lorsqu'on traite des séries chronologiques.

La spécification (3) est remplacée par la relation :

(5) $\varepsilon_{it} = \rho_i.\varepsilon_{i,t-1} + w_{it}$

où ρ_i est le coefficient d'autocorrélation temporelle associé à l'individu i et w_{it} la part d'innovation dans l'aléa ε_{it}, avec

(6) $var(\varepsilon_{it}) = \sigma_i^2$ (homoscédasticité – à i fixé)
(7) $cov(\varepsilon_{it}, \varepsilon_{jt}) = \sigma_{ij}^2$ (corrélation contemporaine stable – à i et j fixés)

Ce qui peut être déduit d'hypothèses convenables concernant l'innovation w_{it}, notamment :

- les w_{it} sont centrés
- les w_{it} sont indépendants pour i fixé et t variant
- les w_{it} et w_{js} sont indépendants pour s différent de t
- les covariances $cov(w_{it}, w_{jt})$ sont non nulles mais non dépendantes de t

On ne détaillera pas davantage cet aspect mathématique des choses.

La procédure d'estimation opère encore en deux étapes : la première estime les coefficients ρ_i et $\sigma_{ij,}^2$ que la seconde utilise pour estimer l'équation (1) par les moindres carrés généralisés.

Autres modèles

D'autres modèles plus raffinés ont été proposés, dérivés tels le précédent des méthodes élaborées de traitement des séries chronologiques; ils ne sont pas décrits ici.

III. Modèles de gravité

Inspirés librement du modèle physique de la gravitation universelle (qui voit celle-ci décroître comme le carré de la distance, selon la *loi de Newton*) les *modèles de gravité*, notamment utilisés pour expliquer les échanges internationaux, introduisent un ou plusieurs termes de nature géographique, ou politique, destinés à prendre en compte la proximité ou l'éloignement entre les acteurs considérés.

En formulation linéaire, les grandeurs étant le plus souvent mesurées par leur logarithme, la forme générale est :

(8) $Y_{ij} = a + b.X_i + c.Z_j + f.D_{ij} + \varepsilon_{ij}$

Où Y_{ij} est le flux du pays i vers le pays j pour le ou les biens étudiés, les variables explicatives X_i et Z_j, des variables économiques telles que le PNB, la population, et éventuellement d'autres variables plus spécifiques liées au domaine considéré, et enfin D_{ij}, la ou les variables de distance retenues. Celles-ci peuvent être la distance des capitales ou encore la distance moyenne entre les deux pays, la longueur de frontière commune, une dummy variable notant l'appartenance à une même organisation économique, etc.

De par leur nature même, les modèles de gravité relèvent du domaine des données de panel. Si toutes les observations sont contemporaines, le choix d'une formalisation (non temporelle) inspirée des modèles précédents permet l'estimation, et en particulier la mesure de l'*effet-distance* associé aux variables D_{ij}.

Si les observations sont en outre temporelles (Y_{ijt}, X_{it}, Z_{jt}, D_{ijt}), les données sont à trois dimensions et il faut fixer i, j ou t pour utiliser les logiciels usuels qui n'en admettent que deux, ou recourir à des procédures avancées ou à la programmation, après formalisation précise du modèle.

IV. Mise en œuvre en SAS

La procédure SAS de régression sur données de panel est la procédure TSCSREG (pour Time Series Cross Section REGression).

Il est tout d'abord nécessaire que parmi les données figurent les séries *ident* des indicateurs i des individus et *time* dénotant la date t des observations, et que le fichier soit trié selon ces deux variables par une commande SORT préalable :

```
PROC SORT;
    BY ident time;
```

La suite ressemble à l'appel de la procédure REG classique :

```
PROC TSCSREG   [options];
    ID ident time;
    MODEL y = x z … / options;
    [label: TEST …;]
```

Parmi les options de l'instruction MODEL :

- FIXTWO indique le modèle à effets fixés.

- FIXONE le même sans effet temps.
- RANTWO indique le modèle à effets aléatoires, c'est le choix par défaut.
- RANONE le même sans effet temps.
- PARKS indique le modèle autorégressif.

Exemple

```
/* Exemple repris et modifié de la doc SAS */
/* Analyzing Demand for Liquid Assets      */

DATA ts;
      INPUT state $ year d y rd rt rs;
      LABEL       d =  'Per Capita Demand Deposits'
                  y =  'Permanent Per Capita Personal Income'
                  rd = 'Service Charge on Demand Deposits'
                  rt = 'Interest on Time Deposits'
                  rs = 'Interest on S & L Association Shares';

DATALINES;
Ca   1949   6.2785  7.2056  -1.0700  0.1080  1.0664
Ca   1950   6.4019  7.2889  -1.0106  0.1501  1.0767
Ca   1951   6.5058  7.3827  -1.0024  0.4008  1.1291
…                   (7 états fois 11 ans)
;

PROC REG   DATA = ts;
      MODEL   d = y rd rt rs;

PROC SORT   DATA = ts;
      BY state year;

PROC TSCSREG   DATA = ts;
      MODEL   d = y rd rt rs   / rantwo parks;
      ID state year;

RUN;
```

Résultats résumés

On se limite à donner en un tableau les coefficients estimés par les trois méthodes demandées; on a figuré les t de Student sous les coefficients estimés.

Méthode	constante	y	rd	rt	rs
mco	-3.49658 (-6.15)	1.31094 (15.86)	-0.48430 (-13.57)	0.00137 (0.02)	-0.13384 (-0.82)
Rantwo	-1.23606 (-1.70)	1.064058 (10.23)	-0.29094 (-5.53)	0.039388 (1.42)	-0.32662 (-2.86)
Parks	-2.66565 (-8.49)	1.222569 (28.87)	-0.43591 (-21.71)	0.041237 (1.97)	-0.26683 (-4.08)

----===ooOoo===----

(30.05.2008)