UNIVERSITE PARIS XII VAL DE MARNE

FACULTE DE SCIENCES ECONOMIQUES ET DE GESTION

Séminaire d'économétrie des panels

Lundi 10 janvier 2000

L'économétrie des données de panel avec SAS : une introduction

P. Blanchard

ERUDITE

Document de travail n°2000-01

Erudite
Université Paris XII Val de Marne
Faculté de Sciences Economiques et de Gestion
58 Av. Didier
94214 Cedex La Varenne Saint-Hilaire
France
Tel : (33) 01 49 76 80 50
Fax: (33) 01 48 85 29 93 et 01 49 76 81 55
Email: blanchard@univ-paris12.fr
Page Web de l'ERUDITE: http://www.univ-paris12.fr/www/labos/erudite/
La dernière version (11/1/2000) de ce document peut être téléchargée par :
http://www.univ-paris12.fr/www/labos/erudite/membres/pb/pb0100.pdf

Sommaire

Introduction

Face au développement croissant des méthodes économétriques appliquées aux données de panel et compte tenu de la double dimension individuelle et temporelle de ces données, les logiciels économétriques ont très rapidement intégré des fonctionnalités leur permettant :
a) de gérer ce type de données,
b) d'appliquer des méthodes économétriques adéquates.

L'offre en matière de logiciels économétriques[1] sur données de panel est conséquente, même s'il n'y a pas de logiciel économétrique spécialement conçu à cette fin :

- *LIMDEP* (7.0) : c'est probablement celui qui offre le plus de méthodes économétriques "en standard" pour les données de panel, mais il offre peu d'instructions de gestion de fichiers et son langage de programmation est puissant mais lourd à manipuler.

- *STATA* (6.0) : moins complet que *LIMDEP*, ce logiciel offre depuis peu (version 6) de nombreux estimateurs pour données de panel et de nombreuses procédures-utilisateur. Il présente les mêmes défauts (moins marqués quand même) que *LIMDEP* en ce qui concerne la gestion des données et le langage de programmation.

- *GAUSS* (pour *Windows 95*) (et ses nombreux "équivalents", *Ox*, *S-Plus*, *AD Model Builder*, par exemple) : GAUSS domine assez nettement le marché des logiciels économétriques sur panel, en ce qui concerne particulièrement les simulations, l'estimations des modèles non linéaires et des modèles dynamiques données de panel. *GAUSS* doit ce succès à sa très grande rapidité d'exécution et à la disponibilité des modules *DPD* (Dynamic Panel Data de Arellano et Bond (1991)), *MAXLIK* (estimation par le maximum de vraisemblance contraint et non contraint, R. Schoenberg) et *Constrained Optimization* (Moindres carrés non linéaires avec ou sans contraintes, R. Schoenberg).

- *RATS* (4.3) et *TSP* (4.4) : ils sont surtout utilisés pour les estimations sur séries temporelles mais ils offrent quelques instructions pour l'estimation des modèles usuels sur panel. A noter que *TSP* (écrit en partie par B. H. Hall) propose l'estimateur Π de Chamberlain (1982). Avec la dernière version de *TSP* (4.4), on trouvera de nombreux programmes sources *TSP* estimant des modèles dynamiques sur données de panel par les méthodes proposées par Arellano et Bond (1991) et Blundell et Bond (1995) (cf. http://elsa.berkeley.edu/tsp_progs.html et http://infoshako.sk.tsukuba.ac.jp/~ykitazaw/homepages/homepage1.html).

- *SAS* (6.12) : Disponible sur quasiment tous les types de machine et de système d'exploitation, *SAS* est probablement le logiciel statistique le plus utilisé dans le monde. Il offre un très grand nombre de méthodes statistiques joint à une très bonne interface avec multi-fenêtrage. Il est notamment le logiciel de référence dans de nombreux instituts d'études économiques français (*INSEE*, *DP*, Banque de France, *CREDOC*…).

- Et bien d'autres *EVIEWS*, *SHAZAM*, …

[1] Le site de Cribari-Neto (http://www.de.ufpe.br/~cribari/) contient de nombreuses références aux logiciels économétriques et aux langages de programmation. Un autre site très complet est The Econometrics Journal on line (http://www.eur.nl/few/ei/links/software.html).

Face à cette offre quasi pléthorique, quel(s) choix peut-on conseiller ? Si l'on cherche à définir les fonctionnalités essentielles que devrait posséder un logiciel économétrique "idéal" sur données de panel, il nous semble que cinq d'entre elles revêtent une importance particulière :

1°) Ce logiciel doit être disponible sur des matériels très différents : gros système, mini-ordinateur (station de travail), micro-ordinateurs (*PC* et/ou *MAC*). En effet, il est de plus en plus fréquent, que ce soit pour des raisons de disponibilité des données, d'hétérogénéité du parc informatique ou d'organisation du travail que les chercheurs utilisent des matériels différents qui les obligent bien souvent à travailler sur différents logiciels.

2°) Il doit posséder de puissantes fonctionnalités de gestion de base de données de panel (appariement de fichier, extraction, tris croisés...). Dans le même ordre d'idée, l'importation et l'exportation des données dans et à partir de formats variés de fichiers (*ASCII*, binaire, *WKS*, *XLS*[2]...) doivent être les plus simples possibles en limitant au maximum la programmation.

3°) Les méthodes standards en matière d'économétrie des données de panel (estimateurs *within* et *mcqg*…) doivent être accessibles sans nécessiter, pour les plus usuelles, une lourde programmation.

4°) Compte tenu du grand volume de données à traiter, le logiciel doit posséder une très grande rapidité en matière de calcul et de lecture des fichiers de données sans imposer de contraintes fortes sur la taille des fichiers à lire ou à écrire.

5°) Enfin, compte tenu du développement continu des techniques économétriques, il doit disposer d'un puissant langage de programmation permettant de mettre en œuvre les méthodes les plus complexes (panels incomplets, modèles dynamiques, modèles non linéaires…).

Tous les logiciels précédemment cités disposent de plusieurs de ces fonctionnalités même si aucun ne les offre toutes. De ces cinq critères de choix, on considère comme essentiel la capacité à gérer les données de panel (contrôle, fusion, appariement...) et la disponibilité de techniques économétriques considérées désormais comme standards (estimateurs *within* et *mcqg*) et d'un langage de programmation, le tout pour des volumes significatifs de données.

Le lecteur un peu averti en matière de logiciels économétriques sera peut-être surpris de l'utilisation de *SAS* en économétrie des données de panel, et donc de l'intérêt que lui porte le présent document, compte tenu des critiques formulées ici et là à son encontre. *SAS* est cher, gourmand en espace disque (un minimum de 190 *Mo* est requis pour installer les modules nécessaires aux économètres (*base, stat, ets, iml, pc-file-formats, graph*…). *SAS* est souvent aussi critiqué pour :

- le nombre limité des méthodes économétriques (en particulier pour les tests) qu'il propose (cf. Korosi et *alii* [1992])

- la lenteur de ses opérations d'entrées-sorties (cf. J. K. MacKie-Mason (1992)) due à sa médiocre gestion de la mémoire virtuelle, ce qui pourrait être un réel handicap pour des études sur des gros fichiers.

[2] Il existe des programmes utilitaires réalisant automatiquement des conversions de fichier non *SAS* vers/à partir de fichier *SAS* : Stat-Transfer et DBMS/Copy. Il est aussi possible d'écrire un programme SAS convertissant un fichier *dBASE*, *WKS* ou *XLS* (version 4.0 et 5.0/95) au format *SAS*, et réciproquement (cf. les programmes cités à la fin de cette introduction). Cette conversion peut être aussi effectuée par menu.

Ces critiques semblent bien définitives. N'y aurait-il donc aucun intérêt à l'utilisation de *SAS* ? Ceci est certainement exagéré car *SAS* présente des avantages considérables :

- *SAS* est certainement le meilleur logiciel en matière de gestion de fichiers. Avec *SAS*, les opérations de fusion, d'appariement, de cylindrage, de nettoyage se font simplement. Il ne faut pas ignorer que ce travail représente une part non négligeable des travaux d'économie appliquée.

- Dans la mesure ou de nombreuses méthodes d'économétrie des panels reviennent à appliquer les *mco* sur données transformées, *SAS* peut estimer assez simplement les modèles de base de l'économétrie des panels.

- Pour les modèles ou les méthodologie de test plus complexes (*VI*, *GMM*, le test d'Hausman,…), le langage *IML* de *SAS* permet des développements comparables à eux que permet *GAUSS*. Par ailleurs, en y prêtant un peu d'attention, il est possible d'écrire des programmes *IML* qui, sans être aussi rapide que ceux écrits en *GAUSS*, s'exécutent néanmoins très rapidement. Par ailleurs la différence de vitesse d'exécution est assez facilement gommée par les facilités de gestion des fichiers de données qu'offre *SAS* en comparaison de celles offertes par *GAUSS*.

- *SAS* fonctionne sur quasiment tous les types d'ordinateurs (sauf les micro-ordinateurs *APPLE*) et sous quasiment tous les systèmes d'exploitation avec des différences minimes (l'interface principalement). En comparaison, *GAUSS* ne fonctionne que *PC* (*MS-DOS* et *Windows)* et sur des stations de travail *UNIX*.

- *SAS* travaillant observation par observation, la taille des fichiers qu'il peut traiter est virtuellement illimitée et n'est donc en fait limitée que par la taille du disque dur.

- Les organismes producteurs de statistiques utilisent presque tous *SAS;* l'obtention des données auprès d'eux ne nécessitera donc la plupart du temps aucune opération de conversion, opération qui peut être complexe, risquée et longue.

- Enfin, *SAS* offre des outils performants de développement (macro-instructions, interfaçage avec le *C*, langage de développement d'application…).

Au total, *SAS* est donc une alternative sérieuse offerte aux économètres/économistes appliquées travaillant sur données de panel. Enfin, bien que ce soit plutôt un argument d'autorité[3], *SAS* semble dominer, avec *GAUSS*, le marché des logiciels économétriques pour les données de panel.

Nous terminons cette introduction en citant quelques ressources Internet pour *SAS* et par une bibliographie commentée.

[3] ou du moins le simple reflet d'un comportement du type "le meilleur logiciel du monde est celui que je connais".

Ressources Internet pour *SAS*

a) L'Université de Berkeley propose notamment des guides *SAS* et des fichiers de données au format *SAS* (en particulier ceux de B. H. Hall et de D. Card) : http://elsa.berkeley.edu/

b) L'Université de Carnegie-Mellon propose des fichiers de données et des programmes dont certains sont écrits en *SAS* : http://lib.stat.cmu.edu/

c) Les sites Web de *SAS*-France (http://www.sas.com/offices/europe/france/index.html)et *SAS*-USA (http://www.sas.com/). Malheureusement, comme c'est souvent le cas des sites commerciaux, les informations sont plutôt de nature publicitaire que technique. A notre connaissance, il n'y a rien de spécifique sur l'utilisation de *SAS* pour les données de panel.

d) Une foule de liens très utiles concernant *SAS* avec http://www.basas.com/pages/hotspot.htm et avec http://www.telepath.com/khobson/sas/hotlist.html .

e) Le groupe de discussion *SAS*. Ce groupe (indépendant de *SAS-Institute*) est très actif, des dizaines de questions/réponses tous les jours. Les répondants sont pour la plupart très compétents et serviables. Pour en faire partie (gratuitement) envoyer un email à Listserv@uga.cc.uga.edu avec dans le corps du texte le message Subscribe SAS-L suivi de votre prénom et votre nom. Les questions (et les solutions) les plus intéressantes sont archivées gopher://jse.stat.ncsu.edu/11/software/sas/ . Malheureusement, il y a peu de choses sur les panels.

f) Y. Guillotin propose des exemples *SAS* à http://pcgains3.univ-lemans.fr/YG5/Sas5.htm.

g) Le site de l'Erudite. Sur ce site, sur la page personnelle de l'auteur de ce document, vous trouverez :

- ce document pb0100.pdf (241Ko) qui correspond à la présentation du 10 Janvier 2000 : « L'économétrie des panels avec SAS. Une introduction » (P. Blanchard), au format PDF),

- des fichiers de données de panel stockés dans **pbdata.zip** (62Ko) au format ZIP, qui contient (pour une description plus détaillée cf. le fichier Word **pbreadsemi.doc** conservé aussi dans **pbdata.zip**) qui contient 4 fichiers (dont 3 de données de panel) :
  1. Les données de Grunfeld (Format XLS, version 4): 6 variables, 10 firmes, 20 années, panel cylindré (Boot et deWitt, International Economic Review (1960)), fichier **grunfeld.xls** (26 Ko).
  2. Les données de Grunfeld (Format XLS, version 95/2000): 6 variables, 10 firmes, 20 années, panel cylindré (Boot et deWitt, International Economic Review (1960)), fichier **grunf99.xls** (51 Ko).
  3. Les données de Baltagi et Griffin (1983) (Format SAS): 6 variables, 18 pays, 19 années, panel cylindré (Baltagi, Econometric Analysis of Panel Data, 1995), **gasoline.sd2 (25Ko)**.
  4. Un fichier d'explication sur les données (**pbreadsemi.doc**) (format Word, 22Ko).

- des programmes *SAS* stockés dans **pbprgsas.zip** (10Ko) au format ZIP qui contient les programmes utilisés dans ce document :

    1) **panel1.sas** : le programme de la section 2
    2) **panel2.sas** : le programme de la section 3-A
    3) **panel3.sas** : le programme de la section 3-B
    4) **panel4.sas** : le programme de la section 3-C
    5) **panel5.sas** : le programme de la section 4
    6) **panel6.sas** : le programme de la section 5

- deux programmes SAS stockés dans **util.zip** (2Ko, format ZIP) :
    1. **convert.sas** qui procède à la conversion de tableaux Excel au format SAS. Deux cas sont présentés : la conversion d'un fichier Excel 4.0 (non classeur) et un fichier au format Excel 5.0/95 (classeur),
    2. **creesimul.sas** générant aléatoirement un panel cylindré de 1000 individus et 8 périodes pour 4 variables (X1 X2 X3 et Y) avec corrélation entre les effets individuels aléatoires et une variable explicative (X2).

Des macro-commandes écrites et diffusées par Emmanuel Duguet pour l'économétrie des panels et des variables qualitatives **macroduguet.zip** (50 Ko, format ZIP), disponibles aussi sur son site http://panoramix.univ-paris1.fr/EUREQUA/annuaire/duguet/duguet.htm. Le fichier contenant les macros pour les panels s'appelle **panel2.cpu**. Pour une description précise, on se reportera à E. DUGUET, Macro-commandes pour l'Econométrie des panels et des variables qualitatives, document de travail INSEE, n°G9914, sept. 1999, que l'on peut aussi télécharger sur ma page personnelle comme un fichier PDF. Par ftp (ftp://panoramix.univ-paris1.fr/pub/CEME/duguet - demander le répertoire macro_commandes), E. Duguet diffuse aussi des polycopiés sur SAS et notamment sur SAS-IML, avec des exercices et leurs corrigés rédigés pour les etudiants de Licence et de Magistère à Paris I.

En cours

- des notes de cours sur *SAS* **pbsascours.pdf** (**xx**Ko) ,

- Des programmes *SAS* **pbpanelsas.zip** (**xx**Ko) :

    1. **classic.sas** : un programme estimant un modèle sur données de panel par les estimateurs *mco*, *intra*, *inter*, *mcqg* et *Swamy* (modèle à coefficients aléatoires) y compris de nombreux test de spécification (LM, Hausman) sur un panel cylindré ou non cylindré.
    2. **gmmld.sas** : estimation d'un modèle dynamique sur données de panel par les *GMM* selon la méthode d'Arellano et Bond (1991), (transposition de *DPD* à *SAS*).
    3. **sysgmm.sas** : estimation d'un modèle dynamique sur données de panel par la méthode "system estimator" de Blundell et Bond (1995), (transposition de *DPD98* à *SAS*).

Bibliographie commentée

Baltagi B. (1995), *The Econometrics of Panel Data*, Wiley.
Ouvrage de référence sur le sujet. Est disponible aussi un "instructor's manual" qui donne les solutions des différents exercices proposés dans le livre. Il n'y a aucun programme associé mais par contre de nombreux exemples sont présentés.

Baltagi B. (1998), *Econometrics*, Wiley.
Un chapitre du manuel est consacré à une introduction à l'économétrie des données de panel. Est disponible aussi un "solutions manual" qui donne les solutions des différents exercices proposés dans le livre. Dans le chapitre 12 du "solutions manual", vous trouverez un programme *SAS* estimant sur un jeu de données ("gasoline data", fourni aussi dans le livre) un modèle par les estimateurs *mco*, *between*, *within*, et *mcqg*. Le programme *SAS* utilise *IML* (le calcul matriciel) et les produits de *Kronecker*.

Dormont B. (1989), *Introduction à l'Econométrie des Données de Panel*, CNRS. Un autre classique, quoiqu'un peu ancien, sur le sujet, en français, mais sans aucun programme.

Duguet E. (1999), Macro-commandes pour l'Econométrie des panels et des variables qualitatives, document de travail INSEE, n°G9914. Ce document de travail contient les macros-instructions pour estimer des modèles sur données de panel par les *mco/between/within/mcqg* (ainsi que des tests) et aussi de nombreuses explications théoriques et pratiques.

Gallant A. R. (1987), *Non Linear Statistical Models*, John Wiley and Sons.
Ce livre propose des données et des programmes *SAS* d'estimation de modèles non linéaires (*GMM*…). Bien que les méthodes présentées ne soient pas spécifiques aux panels, c'est une référence utile.

Hsiao C. (1986), *Analysis of Panel Data*, Cambridge, University University. Un autre classique sur le sujet, mais sans aucun programme et peu d'exemples.

Matyas L. et P. Sevestre (1996), *The Econometrics of Panel Data*, Kluwer Academic Publishers.
L'autre ouvrage de référence sur le sujet.
Une section du chapitre 33 (P. Blanchard) est consacrée à *SAS*. Vous y trouverez un programme *SAS* estimant le modèle de G*runfeld* par les estimateurs *mco*, *between*, *within*, *mcqg* et réalisant un test de Breusch-Pagan. Le programme *SAS*, très proche du programme proposée dans la section 2 de ce document, utilise les instructions de procédures usuelles de *SAS*. Néanmoins, un test d'Hausman est proposé à la fin du programme en utilisant *IML* (très proche du programme de la section 3-B).

Enfin, des articles ou des ouvrages généraux sur *SAS* et l'économétrie, souvent assez anciens.

Lanjouw P. (1992), *The SAS System Version 6*, The Economic Journal, 102, sept., p. 1302-1313.

Brillet J. L. (1989), *Econometric Modelling on Microcomputers: A review of Major Software Packages*, Journal of Applied Econometrics, vol. 4, p 73-92.

Goldstein R., J. Anderson, A. Ash, B. Craig, D. Harrington and M. Pagano, (1989) *Survival Analysis Software on MS/PC-DOS Computers*, Journal of Applied Econometrics.

MacKie-Mason J.K. (1992), *Econometric Software : A User's View*, Journal of Economic Perpectives, vol. 6, n°4, Fall, p. 165-187.

Korosi G., Matyas L. et Szekeley I. (1991), *Practical Econometrics*, Gower

Section 1) Le modèle et les données utilisées

A) Le modèle

Le modèle (cf. Baltagi et Griffin (1983), European Economic Review) s'écrit :

$$Ln\left(\frac{gas}{car}\right)_{it} = a + b_1 \ln\left(\frac{pg}{pp}\right)_{it} + b_2 \ln\left(\frac{y}{n}\right)_{it} + b_3 \ln\left(\frac{car}{n}\right)_{it} + n_{it}$$

où *i* est l'indice du pays et *t* l'indice de l'année,

et avec

gas : la consommation (en volume) d'essence,
car : le nombre de véhicules automobiles en circulation,
pq : l'indice du prix de l'essence,
pg : l'indice des prix à la consommation,
y : le revenu réel disponible,
n : la population totale.

ou encore de façon plus compacte $Lngc_{it} = a + b_1 \ln pgpp_{it} + b_2 \ln yn_{it} + b_3 \ln carn_{it} + n_{it}$ (1)

soit, avec les notations économétriques usuelles :

$$y_{it} = a + \mathbf{x}'_{it}\, b + n_{it} \qquad i = 1,...,N \quad t = 1,...,T \tag{2}$$

où b et $\mathbf{x}_{it}$ sont de dimension $K \times 1$ (avec *k*=3).

Ce modèle peut être modifié notamment dans deux directions, très utiles en pratique :

a) le modèle à effets fixes (individuels) (fixed effects model) tel que

$$y_{it} = a + m_i + \mathbf{x}_{it} b + e_{it} \tag{4}$$

où $e_{it} \sim i.i.d.(0, s_e^2)$ et non auto corrélés.

b) le modèle à erreurs composées (one-way error component model) tel que:

$$y_{it} = a + \mathbf{x}_{it} b + m_i + e_{it} \tag{3}$$

où $m_i \sim i.i.d.(0, s_m^2)$ et $e_{it} \sim i.i.d.(0, s_e^2)$ sont mutuellement indépendants, non autocorrélés et indépendants des **x**.

B) Les données utilisées

Les données sont issues de Baltagi (1995 et 1998)[4] et sont relatives à 18 pays de l'*OCDE* observés sur 19 années (1960-1978). Le tableau *SAS* (**gasoline.sd2**, 25Ko) contient 342 observations (18x19) pour 6 variables utiles pour notre analyse :

country : le nom du pays (chaînes de caractères, 8 car. maxi., en majuscules),
ident : l'identifiant du pays (un numéro allant de 1 à 18, plus simple à utiliser que le nom du pays),
year : l'année d'observation,
lngc : le log. népérien de la consommation d'essence,
lnpgpp : le log. népérien du prix relatif de l'essence par rapport au niveau général des prix,
lncarn : le log. népérien du nombre de voitures automobiles par tête.

Ces données ont été utilisées par Baltagi et Griffin dans leur article de 1983. Ce panel est un panel cylindré. Les programmes qui suivent sont donc écrits dans cette optique. De plus, on remarquera que N et T sont respectivement égaux à 18 et 19, ce qui assez faible pour N et assez élevé pour T par rapport aux panels micro-économiques que l'on a l'habitude de rencontrer.

Les noms des pays sont, dans l'ordre du fichier :

ident	pays	Ident	Pays
1	AUSTRIA	10	JAPAN
2	BELGIUM	11	NETHERLA
3	CANADA	12	NORWAY
4	DENMARK	13	SPAIN
5	FRANCE	14	SWEDEN
6	GERMANY	15	SWITZERL
7	GREECE	16	TURKEY
8	IRELAND	17	U.K.
9	ITALY	18	U.S.A.

Dans tout ce qui suit, nous ne considérons que des modèles avec des effets individuels fixes ou aléatoires, et non avec des effets temporels (joints ou non à des effets individuels).

De même, les méthodes d'estimations présentées ne sont pas robustes à l'hétéroscédasticité. Elles ne portent que sur des panels cylindrés dans lesquels il n'y a pas de régresseur constant dans la dimension temporelle ni de régresseur constant dans la dimension individuelle.

Néanmoins, les programmes s'adaptent pour la plupart d'entre eux assez facilement à ces extensions, notamment la dernière citée. En particulier, *TSCSREG* peut gérer facilement des modèles avec effets individuels **et** temporels aléatoires. Les macro-commandes de E. Duguet gèrent parfaitement la présence de régresseurs constant dans la dimension temporelle ou individuelle et permettent aussi d'obtenir des estimations robustes à l'hétéroscédasticité. Seule, la présence de panel incomplet pose des problèmes de programmation et économétriques qui nécessitent une adaptation non immédiate des programmes *panel3.sas*, *panel4.sas*, *panel5.sas* et *panel3.sas*.

[4] Fournies sur disquette avec le livre de Baltagi (1998) et en annexe dans celui de 1995 Elles peuvent aussi être téléchargées sur son site web http://econ.tamu.edu/bbaltagi/index.htm (ou sur le mien cité à la fin de l'introduction).

Par ailleurs, les programmes *SAS* présentés dans ce document ne doivent pas être considérés comme des solutions clés en main (qui seraient exemptes de tout défaut) mais comme de simples illustrations de différentes façons de programmer certains estimateurs sur données de panel. Il y a toujours des arbitrages à effectuer entre le degré de complexité du programme, le degré de contrôle sur les méthodes utilisées et la rapidité d'exécution.

Enfin, ces programmes ont été testés sur micro-ordinateur; leurs performances respectives, en termes absolus, bien sur, mais aussi en termes relatifs, seront très vraisemblablement assez différentes sur un mini-ordinateur ou un gros système.

Section 2) Estimation *mco*/*within*/*mcqg* avec les procédures SAS

Cette méthode a l'avantage de la simplicité car elle ne requiert que des connaissances de base sur l'étape DATA et sur les PROC, tout en laissant l'utilisateur libre de faire certains choix de programmation. Les étapes à suivre sont simples : après une phase d'initialisation, on procède aux estimations *within* et *between* (les estimations *mco* sont facultatives) afin de pourvoir calculer $\hat{q}$ pour estimer le modèle à erreurs composées par les *mcqg*.

Avant de commencer les estimations, il faut faire un certain nombre d'initialisations. On suppose que le fichier de données **gasoline.sd2** est stocké dans le répertoire **seminair** du disque dur **C:**. Les tableaux temporaires (par exemple **tab1**), crées par les programmes sont censés être dans la librairie *work*, soit **C:\sas\saswork** et détruits à la fin de la session *SAS*.

```
/* panel1.sas */

/* initialisations */

LIBNAME in 'c:\seminair\' ;

PROC SORT DATA = in.gasoline ; BY country year ;

DATA tab1 ; SET in.gasoline ;
n = 18 ; t = 19 ; nt = n*t ; kb = 4 ; kw = 3 ;
```

Un programme SAS est constitué d'une série d'étapes DATA de d'appel à des procédures (PROC) effectuant une certaine action et, éventuellement, créant en sortie, un tableau contenant certains résultats de la procédure. On se souviendra que SAS travaille observation par observation lors d'une étape DATA ou d'une PROC sur un tableau.

On trie le tableau **gasoline** par *country* et *year* (inutile si c'est déjà fait). Pour tout ce qui suit, on supposera que ce tri a été effectué. On met[5] dans le tableau de travail **tab1** les valeurs de *N,T, NT*, *kb* (nombre de régresseurs du *between*) et *kw* (nombre de régresseurs du *within*, *kw=kb-1*) afin de pouvoir réaliser certains calculs facilement en paramétrant au maximum le programme. En effet, *SAS* ne permet pas dans une étape DATA ou PROC de mémoriser certaines variables (scalaire, vecteur ou matrice) autrement que dans le tableau lui-même. On laisse de côté les (indispensables) analyses de statistiques descriptives qu'il faudrait faire.

1) On estime tout d'abord le modèle par les mco, $\hat{b}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ obtenu par empilement des individus.

[5] L'utilisation de l'instruction RETAIN conduirait à une plus grande vitesse d'exécution.

```
/* estimation mco*/

PROC REG DATA = tab1 ;
     MODEL lngc = lnyn lnpgpp lncarn ;
    OUTPUT OUT = fmcores R = umco ;
TITLE estimation par les mco ;
```

Cette étape n'est pas réellement obligatoire mais elle permet de stocker dans un tableau (**fmcores**) les résidus des *mco* pour des analyses ultérieures (test de Breusch et Pagan, notamment). A noter que le tableau **fmcores** contient, outre les résidus des *mco*, les variables présentes dans **tab1**.

2) Il faut ensuite calculer les moyennes individuelles (pour pouvoir obtenir les estimateurs *within* et *between*). Ces moyennes sont stockées dans le tableau **mtab1** qui comporte *N* observations et qui sera fusionné ultérieurement avec **tab1** (*NT* observations). Attention à effectuer le *between* avant la fusion, c'est à dire sur **mtab1**, car le tableau fusionné contiendrait alors *NT* moyennes individuelles (cf. supra) ce qui fausserait l'estimation de la variance des perturbations du *between*.

```
/* calcul des moyennes individuelles */

PROC MEANS DATA = tab1 MEAN NOPRINT ; BY country ;
     VAR lngc lnyn lnpgpp lncarn ;
     OUTPUT OUT = mtab1 MEAN = mlngc mlnyn mlnpgpp mlncarn;
```

3) On estime le modèle par l'estimateur *between* $\hat{b}_{betw} = (\mathbf{X}_b^{'}\mathbf{X}_b)^{-1}\mathbf{X}_b^{'}\mathbf{y}_b$ obtenu en appliquant les *mco* sur les moyennes individuelles des variables telle que, par exemple, $\bar{x}_{i.} = \frac{1}{T}\sum_{t=1}^{T} x_{it}$.

```
/* estimation between */

PROC REG DATA = mtab1 OUTEST = vbetw ;
     MODEL mlngc = mlnyn mlnpgpp mlncarn ;
TITLE estimation between ;

DATA vbetw ( KEEP = sig2b dum ) ; SET  vbetw ;
    sig2b = (_RMSE_)**2 ; dum = 1 ;
```

On utilise directement le tableau contenant les moyennes individuelles. En fait, il n'est pas absolument nécessaire d'afficher les estimations *between* (on peut donc utiliser l'option NOPRINT après le PROC REG) mais nous avons besoin de l'estimation de la variance des perturbations obtenue à l'aide de l'estimateur *between* ($\hat{s}_b^2 = \frac{\mathbf{e}_b^{'}\mathbf{e}_b}{N-K}$ où $\mathbf{e}_b$ représente le vecteur des résidus de

l'estimation *between*) pour pouvoir appliquer les *mcqg* (cf. plus loin). Ceci est fait en récupérant la variable-système _RMSE_qui est stockée dans le tableau **vbetw** compte tenu de l'option OUTEST). On crée la variable *dum* qui vaudra 1 pour toutes les observations du tableau afin d'effectuer ultérieurement la fusion de **vbetw** avec d'autres tableaux. Juste après le PROC REG, il contient :

OBS	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	INTERCEP	MLNYN	MLNPGPP	MLNCARN	MLNGC
1	MODEL1	PARMS	MLNGC	0.19669	2.54163	0.96758	-0.96355	0.79530	-1

Après l'étape DATA, qui suit le PROC REG, il ne contient plus, compte tenu du KEEP, que :

OBS	_RMSE_	DUM
1	0.03868	1

4) L'étape suivante consiste à appliquer l'estimateur *within* (ou *intra*, dit aussi *LSDV* : "least squares dummy variables") $\hat{b}_{with} = (\mathbf{X}_w^{'}\mathbf{X}_w)^{-1}\mathbf{X}_w^{'}\mathbf{y}_w$ obtenu en appliquant les *mco* sur les écarts aux moyennes individuelles des variables tels que, par exemple, $\tilde{x}_{it} = (x_{it} - \bar{x}_{i.})$

On calcule d'abord les écarts aux moyennes individuelles après avoir fusionné **tab1** (qui contient les x_{it} en nombre *NT*) et **mtab1** (qui contient les $x_{i.}$ en nombre *N*).

```
/* estimation within */

DATA within ; MERGE tab1 mtab1 ; BY country ;
    dum = 1 ;
    wlngc   =  lngc   - mlngc   ;
    wlnyn   =  lnyn   - mlnyn   ;
    wlnpgpp =  lnpgpp - mlnpgpp ;
    wlncarn =  lncarn - mlncarn ;

PROC REG DATA = within OUTEST = vwith ;
    MODEL wlngc = wlnyn wlnpgpp wlncarn /NOINT ;
TITLE estimation within ;

DATA vwith (KEEP = _RMSE_ dum ) ; SET vwith ;
    dum = 1 ;
```

Ensuite on estime le modèle par l'estimateur *within* en appliquant les *mco* aux données exprimées en écarts aux moyennes individuelles. Notez l'option NOINT pour exclure la constante du modèle. Comme pour le *between*, on a besoin, pour le calcul de l'estimateur des *mcqg*, de l'estimation de la variance des perturbations du *within* (calculée à partir des résidus du *within*) soit $\hat{s}_e^2 = \frac{\mathbf{e}_w^{'}\mathbf{e}_w}{NT - Kw}$.

Attention *SAS* le calcule en utilisant *NT-kw* comme nombre de degrés de libertés au lieu de (*NT-N-kw*). Ce point sera corrigé un peu plus loin dans le programme. Cette estimation est stockée dans **vwith** (notez encore que, dans les tableaux **within** et **vwith,** on crée *dum*=1 pour des opérations

ultérieures de fusion). Voici les résultats de l'estimation *within* (on ne reproduit pas les résultats des *mco* et du *between*).

```
   estimation within

Model: MODEL1

NOTE: No intercept in model. R-square is redefined.

Dependent Variable: WLNGC

                            Analysis of Variance

                             Sum of          Mean
       Source          DF    Squares        Square      F Value       Prob>F

        Model           3    14.32419       4.77473      591.500       0.0001
        Error         339     2.73649       0.00807
        U Total       342    17.06068

       Root MSE       0.08985     R-square       0.8396
       Dep Mean      -0.00000     Adj R-sq       0.8382
       C.V.          -5.765962E16

                                Parameter Estimates

                         Parameter      Standard    T for H0:
         Variable  DF     Estimate         Error   Parameter=0    Prob > |T|

         WLNYN      1     0.662250    0.07141117         9.274        0.0001
         WLNPGPP    1    -0.321702    0.04291251        -7.497        0.0001
         WLNCARN    1    -0.640483    0.02888017       -22.177        0.0001
```

Il faut noter plusieurs points :

a) Le nombre de degrés de libertés (339) est **faux** car il faut tenir compte des *N* moyennes individuelles calculées. **Par conséquent, l'estimation de la variance des perturbations est biaisée, les écarts-type estimés, les T de Student et les niveaux de significativité marginale aussi.** Il ne semble pas possible de corriger ce problème par une option quelconque de la procédure (on pourra en utilisant une autre procédure, cf. la section 3). Néanmoins, comme nous l'avons dit précédemment, lors de l'estimation par les *mcqg*, on pourra corriger $\hat{s}_e^2$.

b) Il faut aussi remarquer que l'estimateur *within* ne permet pas d'estimer l'effet de variables invariantes dans le temps (comme le sexe, la race, la religion...). Il y a alors stricte colinéarité avec les effets fixes; alors *SAS* élimine automatiquement certaines variables (avec un message) du type de celui que nous aurons au e).

c) Il ne faut pas prêter attention au message «*R-Square is redefined*… ». En effet SAS constatant l'absence de constante dans le modèle calcule le R^2 comme $R_{nc}^2 = 1 - \dfrac{Scr}{\sum_i \sum_t \tilde{y}_{it}^2}$ au lieu de

$R_c^2 = 1 - \dfrac{Scr}{\sum_{i,t}\left(\tilde{y}_{it} - \bar{\tilde{y}}\right)^2}$, ce qui ne pose pas de problème dans notre cas, puisque les $\tilde{y}_{it}$ sont égaux aux $y_{it} - y_{i.}$ dont la moyenne $\bar{\tilde{y}}$ est nulle.

d) Les résultats (et ceux qui suivent) sont un peu différents de ceux de Baltagi (1995, 1998) Baltagi et Griffin (1983) du fait que les données utilisées étaient déjà en logarithme et donc arrondies.

e) Si on considère les estimations sur les écarts aux moyennes individuelles, l'opération de transformation des variables de départ a fait disparaître les effets fixes. Ils sont alors calculés par $\hat{a}_i = \bar{y}_{i.} - \bar{\mathbf{x}}_{i.}\hat{b}_{with}$. Les écarts-type estimés des effets fixes estimés (et par suite, leurs T de Student) peuvent être calculés comme $\hat{V}(\hat{a}_i) = \dfrac{\hat{s}_e^{\,2}}{T} + \bar{\mathbf{x}}'_{i.}\hat{V}(\hat{b}_{wit})\bar{\mathbf{x}}_{i.}$ avec $\hat{s}_e^{\,2} = \dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \hat{a}_i - \mathbf{x}'_{i.}\hat{b}_{with})^2}{NT - N - K_w}$. Néanmoins, les termes $\hat{a}_i$ incorporent la constante. Mais, comme la constante peut être calculée comme $\hat{a} = \bar{y} - \hat{b}_{with}\bar{x}$, il est facile de récupérer les estimations des effets fixes hors effet de la constante, en calculant $\hat{a}_i^* = \hat{a}_i - \hat{a}$ compte tenu de la contrainte d'identification. Avec notre programme, ce calcul serait assez difficile à faire (mais ce sera beaucoup plus facile avec *IML* - cf. **panel5.sas** - et encore plus simple avec PROC TSCSREG et PROC GLM). Ecrire le modèle en écarts aux moyennes individuelles permet de ne pas introduire directement dans **X** des indicatrices individuelles (en nombre *N*) et ainsi de réduire la taille de la matrice à inverser sans que cela modifie les résultats. Par contre, ceci rend difficile la mise en évidence des estimations des effets fixes et de la constante, ainsi que de leurs écarts-type estimés (cf. le point b, ci-dessus). Néanmoins, cette transformation est indispensable si $N \to \infty$. Pour pouvoir disposer, simplement, des estimations des effets fixes, on aurait pu faire :

```
DATA temp ; SET in.gasoline ;
IF ident = 1  THEN dump1  = 1 ; ELSE dump1  = 0 ;                    /* on utilise ident car */
IF ident = 2  THEN dump2  = 1 ; ELSE dump2  = 0 ;                 /* plus simple que, par exemple  */
IF ident = 3  THEN dump3  = 1 ; ELSE dump3  = 0 ;                  /* IF country = "FRANCE"= THEN … */
IF ident = 4  THEN dump4  = 1 ; ELSE dump4  = 0 ;
IF ident = 5  THEN dump5  = 1 ; ELSE dump5  = 0 ;
IF ident = 6  THEN dump6  = 1 ; ELSE dump6  = 0 ;
IF ident = 7  THEN dump7  = 1 ; ELSE dump7  = 0 ;
IF ident = 8  THEN dump8  = 1 ; ELSE dump8  = 0 ;
IF ident = 9  THEN dump9  = 1 ; ELSE dump9  = 0 ;
IF ident = 10 THEN dump10 = 1 ; ELSE dump10 = 0 ;
IF ident = 11 THEN dump11 = 1 ; ELSE dump11 = 0 ;
IF ident = 12 THEN dump12 = 1 ; ELSE dump12 = 0 ;
IF ident = 13 THEN dump13 = 1 ; ELSE dump13 = 0 ;
IF ident = 14 THEN dump14 = 1 ; ELSE dump14 = 0 ;
IF ident = 15 THEN dump15 = 1 ; ELSE dump15 = 0 ;
IF ident = 16 THEN dump16 = 1 ; ELSE dump16 = 0 ;
IF ident = 17 THEN dump17 = 1 ; ELSE dump17 = 0 ;
IF ident = 18 THEN dump18 = 1 ; ELSE dump18 = 0 ;
```

```
PROC REG DATA = temp ;                          /* estimation within sol 1 : cte  + 18 dummies */
    MODEL lngc = lnyn lnpgpp lncarn dump1 dump2  dump3  dump4  dump5  dump6  dump7  dump8
                                    dump9 dump10 dump11 dump12 dump13 dump14 dump15 dump16
                                    dump17 dump18 ;
TITLE estimation within sol 1 : cte  + 18 dummies ;

PROC REG DATA = temp ;                          /* estimation within sol 2 : cte + 17 dummies */
    MODEL lngc = lnyn lnpgpp lncarn dump1 dump2  dump3  dump4  dump5  dump6  dump7  dump8
                                    dump9 dump10 dump11 dump12 dump13 dump14 dump15 dump16
                                    dump17 ;
TITLE estimation within sol 2 : cte + 17 dummies ;

PROC REG DATA = temp ;                       /* estimation within sol3 : sans cte + 18 dummies */
    MODEL lngc = lnyn lnpgpp lncarn dump1 dump2  dump3  dump4  dump5  dump6  dump7  dump8
                                    dump9 dump10 dump11 dump12 dump13 dump14 dump15 dump16
                                    dump17 dump18 /NOINT ;
TITLE estimation within sol3 : sans cte + 18 dummies ;

PROC REG DATA = temp ;                                          /* estimation sous  contrainte */
    MODEL lngc = lnyn lnpgpp lncarn dump1 dump2  dump3  dump4  dump5  dump6  dump7  dump8
                                    dump9 dump10 dump11 dump12 dump13 dump14 dump15 dump16
                                    dump17 dump18 ;
    RESTRICT dump1+dump2+dump3+dump4+dump5+dump6+dump7+dump8+
             dump9+dump10+dump11+dump12+dump13+dump14+dump15+dump16+dump17+dump18 ;
TITLE estimation within sol 4 : cte + 18 dummies + RESTRICT ;
```

Voici les résultats des diverses solutions :

Tableau n°1 (on ne donne que les coefficients estimés et leurs écarts-type estimés)

Noms des Variables	Solution 1			Solution 2 (écarts-type identiques)	Solution 3		Solution 4	
INTERCEP	B	3.055251	0.21959957	3.055251	-	-	2.402670	0.22530938
DUMP1	B	-0.769395	0.04457642	-0.769395	2.285856	0.22832349	-0.116814	0.02115663
DUMP2	B	-0.889700	0.04667508	-0.889700	2.165551	0.21289849	-0.237118	0.02560616
DUMP3	B	-0.013411	0.03059601	-0.013411	3.041840	0.21863504	0.639171	0.03168455
DUMP4	B	-0.665795	0.04576975	-0.665795	2.389456	0.20808633	-0.013214	0.02797022
DUMP5	B	-0.850479	0.05012521	-0.850479	2.204771	0.21646984	-0.197898	0.02617240
DUMP6	B	-0.905382	0.04133605	-0.905382	2.149868	0.21788442	-0.252801	0.02232497
DUMP7	B	-0.718141	0.05547462	-0.718141	2.337110	0.21488472	-0.065560	0.03343003
DUMP8	B	-0.462926	0.05692697	-0.462926	2.592325	0.24368637	0.189656	0.03085633
DUMP9	B	-0.822703	0.06159562	-0.822703	2.232548	0.23954072	-0.170122	0.03355306
DUMP10	B	-0.679323	0.04736544	-0.679323	2.375927	0.21183563	-0.026742	0.02929747
DUMP11	B	-0.820459	0.04495023	-0.820459	2.234791	0.21417238	-0.167878	0.02442462
DUMP12	B	-0.838550	0.04875143	-0.838550	2.216701	0.20304201	-0.185969	0.03225287
DUMP13	B	-1.373474	0.09948542	-1.373474	1.681777	0.16246363	-0.720893	0.08689666
DUMP14	B	-0.028908	0.18246610	-0.028908	3.026343	0.39451137	0.623673	0.18072974
DUMP15	B	-0.652748	0.03471857	-0.652748	2.402503	0.22909278	-0.000167	0.02549931
DUMP16	B	-0.545263	0.05352651	-0.545263	2.509988	0.23565649	0.107318	0.03936958
DUMP17	B	-0.709803	0.04588219	-0.709803	2.345448	0.22728405	-0.057222	0.02209700
DUMP18	0	0	.	-	3.055251	0.21959957	0.652581	0.03618028
							RESTRICT	1.1137E-13

Il est équivalent d'estimer le modèle avec 18 indicatrices sans la constante ou de l'estimer avec 17 indicatrices et la constante ou imposer des contraintes. Néanmoins, les coefficients (et les écarts-type) des indicatrices ne seront pas identiques dans les différents cas (sauf 1 et 2), mais les autres résultats (ceux relatifs aux autres régresseurs, non reproduits ici) seront inchangés. Choisir une de ces solutions est indispensable car les coefficients des 18 indicatrices et de la constante ne sont identifiables (on peut seulement obtenir $a + m_i$). Pour résoudre cette difficulté, on impose donc la contrainte $\sum_i m_i = 0$. En outre, on peut noter que :

- Les solutions 1 et 2 donnent le même résultat à la différence qu'un message de colinéarité[6] stricte est envoyé par *SAS* avec la solution 1 (notez le B pour « biaised »). La dernière indicatrice est automatiquement éliminée par *SAS* dans la solution 1. Dans la solution 2, elle est éliminée volontairement de la liste des régresseurs. Les Etats-Unis, en 18ème position dans le fichier, sont alors la référence (+ la constante générale). Pour retrouver les estimations des effets fixes hors la constante, il faut tout d'abord calculer la moyenne des indicatrices 1 à 18[7] (soit –0.65258). Par conséquent, pour retrouver la contrainte d'identification, il faut que USA = +0.65258. Pour retrouver, la constante, il faut enlever 0.65258 à 3.055251. Pour obtenir les indicatrices hors constante et hors USA, il faut ajouter la valeur des USA à chaque indicatrice 1 à 17 (pas pour la dernière évidemment). On retrouve alors le résultat de la colonne 4.

- En ce qui concerne la solution 3, la moyenne des estimations des effets fixes est égale à 2,40267. Pour retrouver les indicatrices hors la constante, il suffit de retrancher la constante à chaque effet fixe estimé.

- La quatrième solution donne le même résultat que 1,2 ou 3 (après identification) mais directement, sans nécessiter une phase d'identification. Néanmoins, elle peut être fastidieuse à programmer (et exigeante en temps calcul) si *N* est grand (pour des données à un niveau sectoriel très fin).

Attention aux écarts-type estimés, aux T de Student et aux nsm (niveau de significativité marginal). Voici les nsm obtenus pour certains pays pour les 4 solutions.

Tableau n°2

Nom	Solution 1 et 2	Solution 3	Solution 4
Constante	0.0001	-	0.0001
Dump1	0.0001	0.0001	0.0001
Dump3	0.6615	0.0001	0.0001
Dump4	0.0001	0.0001	0.6369
Dump10	0.0001	0.0001	0.3620
Dump14	0.8742	0.0001	0.0006
Dump15	0.0001	0.0001	0.9948
Dump16	0.0001	0.0001	0.0068
Dump17	0.0001	0.0001	0.0100
Dump18	.	0.0001	0.0001

[6] NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased. The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

```
DUMP18  = +1.0000 * INTERCEP -1.0000 * DUMP1    -1.0000 * DUMP2
          -1.0000 * DUMP3    -1.0000 * DUMP4    -1.0000 * DUMP5    -1.0000 * DUMP6
          -1.0000 * DUMP7    -1.0000 * DUMP8    -1.0000 * DUMP9    -1.0000 * DUMP10
          -1.0000 * DUMP11   -1.0000 * DUMP12   -1.0000 * DUMP13   -1.0000 * DUMP14
          -1.0000 * DUMP15   -1.0000 * DUMP16   -1.0000 * DUMP17
```

[7] et non de 1 à 17.

Si l'on est intéressé par les estimations des effets fixes (pour classer les pays par exemple, et en ne jugeant que les différences significatives), la troisième solution semble préférable car elle évite d'avoir à recalculer les écarts-type. Cependant, les nsm 1-2 et 3 sont différents car ils dépendent de la situation de référence. En effet, les hypothèses nulles dans ces différents cas sont :

Solution 1 et 2 : $H0: a + f_i = a + f_{usa}$

Solution 3 : $H0: f_i = a$

Solution 4 : $H0: f_i = 0$

On peut donc tester $f_i = 0$, pas de différence par rapport à la moyenne). De plus, un test de significativité globale des effets fixes, à partir de (3) ou (4) est envisageable.

5) Enfin, on calcule l'estimateur des *mcqg* $\hat{b}_{mcqg} = (\mathbf{X}_g^{'}\mathbf{X}_g)^{-1}\mathbf{X}_g^{'}\mathbf{y}_g$ obtenu en appliquant les *mco* sur les quasi écarts aux moyennes individuelles ("quasi demeaned") des variables telles que, par exemple, $\tilde{x}_{it} = (x_{it} - \hat{q}\bar{x}_{i.})$. De nombreuses solutions ont été proposées pour le calcul de $\hat{q}$. Une méthode, fréquemment retenue, consiste à le définir comme $\hat{q} = 1 - \sqrt{\frac{\hat{s}_e^2}{T\hat{s}_u^2 + \hat{s}_e^2}}$. En effet, si l'on note $v_{it} = m_i + e_{it}$ et si les estimateurs *between* et *within* peuvent être calculés, alors on peut appliquer les étapes suivantes :

- On évalue $\hat{s}_e^2$ à l'aide des estimations *within* tel que $\hat{s}_e^2 = \frac{e_w^{'}e_w}{NT - N - k_w} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\hat{b}_{with})^2}{NT - N - k_w}$ En effet, les perturbations du modèle *within* s'écrivent $V(v_{it} - v_{i.}) = V(e_{it} - e_{i.})$, les effets individuels ont donc disparu. La variance des résidus du *within* peut donc être utilisée pour estimer la variance des perturbations e_{it}.

- Les perturbations du modèle *between* s'écrivent $v_{i.} = m_i + e_{i.}$ donc $V(v_{i.}) = s_m^2 + \frac{s_e^2}{T}$. Donc à l'aide des résidus du *between*, $\hat{V}(v_{i.}) = \hat{s}_b^2 = \frac{\mathbf{e}_b^{'}\mathbf{e}_b}{N - k_b}$ (où $\mathbf{e}_b$ représente le vecteur des résidus de l'estimation *between*), on peut calculer $\hat{s}_u^2 = \hat{s}_b^2 - \frac{\hat{s}_e^2}{T}$ pour obtenir $\hat{q} = 1 - \sqrt{\frac{\hat{s}_e^2}{T\hat{s}_u^2 + \hat{s}_e^2}}$ soit $1 - \sqrt{\frac{\hat{s}_e^2}{T\hat{s}_b^2}}$. Cette solution a été proposée par Swamy et Arora (1972) et c'est celle que nous avons retenue dans les programmes **panel1.sas**, **panel3.sas**, **panel4.sas**, et **panel5.sas**.

```
/* estimation mcqg*/

/* utile pour les tests et pour la fusion */
DATA within ; SET within ; SET fmcores (KEEP = umco) ;

/* fusion */
DATA mcqg ; MERGE  within vbetw vwith ; BY dum ;
    sig2u = sig2b - sig2eps/t ;
    sig2eps = ((_RMSE_)**2)*(nt-kw)/(nt-n-kw) ;  /* on corrige le nb de ddl */
    theta = 1 - SQRT((sig2eps)/(t*sig2b)) ;
    glngc   =  lngc   - theta *mlngc   ;
    glnyn   =  lnyn   - theta *mlnyn   ;
    glnpgpp =  lnpgpp - theta *mlnpgpp ;
    glncarn =  lncarn - theta *mlncarn ;
    gconst  =  1      - theta          ;

DATA presult ; SET mcqg ; IF _N_ = 1 ;
PROC PRINT DATA = presult ; VAR sig2b sig2eps sig2u theta nt n t kb kw ;
TITLE edition de presult ;

PROC REG DATA = mcqg ;
    MODEL glngc = glnyn glnpgpp glncarn gconst / NOINT ;
TITLE estimation mcqg ;
```

Dans le tableau **mcqg**, on met les résidus des *mco*[8] (pour les tests), les données et les estimations des variances *within* et *between* (cf. MERGE ... ; BY dum ;) Puis, dans le même tableau, on fait les transformations sur les variables y compris la constante, après avoir calculé $\hat{q}$. On a bien sûr calculé correctement le nombre de degrés de liberté de l'estimation de la variance des perturbations du *within*.

Voici le contenu de **presult** qui contient la première observation de **mcqg** (ce qui est suffisant pour éditer les valeurs des constantes qui nous intéressent) :

```
                  edition de presult

OBS    SIG2B       SIG2EPS     SIG2U      THETA      NT    N     T    KB    KW
  1    0.038686   .0085249    0.038238   0.89231    342   18    19     4     3
```

Voici les estimations des *mcqg* :

[8] En faisant attention de ne prendre dans **fmcores** que ce qui nous intéresse c'est à dire *umco*, et non pas l'ensemble des variables, déjà présentes dans within.

```
estimation mcqg

Model: MODEL1

NOTE: No intercept in model. R-square is redefined.

Dependent Variable: GLNGC

                            Analysis of Variance

                                 Sum of          Mean
        Source           DF      Squares        Square      F Value       Prob>F

        Model             4     88.18445     22.04611     2418.006       0.0001
        Error           338      3.08171      0.00912
        U Total         342     91.26615

            Root MSE       0.09549     R-square       0.9662
            Dep Mean       0.46268     Adj R-sq       0.9658
            C.V.          20.63764

                                  Parameter Estimates

                     Parameter      Standard    T for H0:
      Variable  DF     Estimate         Error   Parameter=0    Prob > |T|

       GLNYN      1     0.554986    0.05912818         9.386        0.0001
       GLNPGPP    1    -0.420389    0.03997814       -10.515        0.0001
       GLNCARN    1    -0.606840    0.02551504       -23.784        0.0001
       GCONST     1     1.996698    0.18432598        10.832        0.0001
```

Le lecteur intéressé vérifiera que l'on obtient quasiment les mêmes résultats que Baltagi (1998, p. 318 et 1995, p. 21, les différences provenant sans doute du fait que les données fournies par Baltagi étaient déjà en logarithme, d'ou des problèmes d'arrondis).

On peut faire plusieurs remarques sur ces calculs :

- on notera le message sur le R² qui, dans ce cas, n'est pas calculé correctement (*SAS* croit que c'est un modèle sans constante). *SAS* calcule $R^2_{nc} = 1 - \frac{Scr}{\sum_i y_{g_i}^2}$ au lieu de $R^2_c = 1 - \frac{Scr}{\sum_{i,t}\left(y_{g_{it}} - \overline{y_g}\right)^2}$ avec $y_{g_{it}} = y_{it} - \hat{q}\, y_{i.}$. Il n'y a aucune raison de penser que $\overline{y_g}$ est nulle.

  Néanmoins, on pourrait laisser une constante à la place de gconst, seule la valeur estimée de la constante serait affectée (mais on pourrait alors l'identifier, car il suffit de diviser par $1-\hat{q}$). Quoiqu'il en soit, le R² intéressant serait celui que l'on obtiendrait en appliquant les coefficients estimés aux données non transformées (idem pour le *within*).

- Il y a d'autres solutions pour le calcul de $\hat{s}_u^2$ et de q :

a) Si l'estimateur *between* ne peut pas être calculé ($T < kb$), ou si $\hat{s}_u^2 < 0$ (il est donc recommandé de calculer explicitement $\hat{s}_u^2$ par $\hat{s}_u^2 = \hat{s}_b^2 - \frac{\hat{s}_e^2}{T}$ et de vérifier sa positivité), on recommence l'étape précédente à l'aide de $s_b^2 = \frac{\sum_{i=1}^{N} (\bar{y}_{i.} - \bar{\mathbf{x}}_{i.} b_{ols})^2}{N - k_b}$, c'est-à-dire en utilisant toujours les moyennes individuelles mais avec l'estimateur des *mco* (c'est une variante proposée par Wallace et Hussain (1969) et Amemiya (1971)).

b) Si $\hat{s}_u^2$ est toujours négatif, on utilise les effets fixes car $\hat{u}_i = \hat{a}_i - \frac{1}{N}\sum_i \hat{a}_i$, ce qui permet d'en déduire $\hat{s}_u^2 = \frac{1}{N}\sum_i \left(\hat{a}_i - \frac{1}{N}\sum_j \hat{a}_j \right)^2$, estimateur qui est toujours positif. Dans ce cas, $\hat{s}_e^2$ est calculé à partir du *within* mais avec *NT* comme nombre de degrés de libertés (Nerlove (1971) cité par Baltagi (1995)). Cette solution est préférable à celle de Maddala et Mount (1973), qui suggèrent de mettre $\hat{s}_u^2$ à zéro, tout simplement parce que ceci revient à appliquer les *mco* dont les estimations sont déjà fournies (en effet $\hat{q}$ est égal à 0 dans ce cas). Néanmoins, ce n'est pas illogique, car si $\hat{s}_u^2$ est égal à 0, cela signifie qu'il n'y a pas d'effet individuel, et donc les *mco* sont appropriés.

c) A noter que si l'estimateur *within* ne peut pas être calculé, l'estimateur *mcqg* ne peut pas l'être non plus (cf. Greene (1997, p. 628, pour une solution à ce problème).

Néanmoins, les études de simulation ont montré que le fait d'utiliser un estimateur plus ou moins efficace pour la première étape n'avait pas de conséquences essentielles sur la deuxième étape (cf. Maddala et Mount (1973)), il suffit d'utiliser un estimateur convergent. Autant en utiliser un simple. **Par ailleurs, si le calcul de $\hat{q}$ est impossible, cela suggère plutôt un problème de spécification.**

- Comme suggéré par Fuller et Battese (1974), suivant en cela une recommandation de Nerlove (1971), l'estimation de la matrice de variance-covariance de l'estimateur *mcqg* est parfois calculée comme $\hat{V}(\hat{b}_{mcqg}) = \hat{s}_e^2 (\mathbf{X}_g' \mathbf{X}_g)^{-1}$ au lieu de $\hat{V}(\hat{b}_{mcqg}) = \hat{s}_g^2 (\mathbf{X}_g' \mathbf{X}_g)^{-1}$, c'est à dire en utilisant l'estimation de la variance des perturbations du *within* et non pas à partir des résidus estimés sur le modèle transformé (*mcqg*). C'est notamment le cas avec *LIMDEP* (cf. Greene 1993, p.493). Ce n'est pas la solution choisie par Baltagi (1995, 1998). Théoriquement, cela devrait avoir peu de conséquences, même si, parfois, les résultats auxquels cela conduit peuvent différer sensiblement (en particulier pour le test d'Hausman). Nous verrons cela dans la section 4-A les différences en matière de résultats. Les programmes **panel1.sas** et **panel2.sas** appliquent la même solution que Baltagi (par la force des choses), alors que **panel3.sas, panel4.sas** et **panel5.sas** retiennent celle de Fuller et Battese.

- Il est aussi possible d'estimer le modèle par les *mcqg* itérés (*imcqg*) comme suggéré par Breusch (1987) dont les estimations sont asymptotiquement équivalentes au Maximum de Vraisemblance (mais plus faciles à obtenir, cf. Hsiao (1986)). Néanmoins, comme l'ont remarqué Blanchard et Matyas (1998) les estimations des *imcqg* sont peu différentes en pratique des *mcqg* mais un peu plus compliquées à obtenir, ce qui en limite l'intérêt. Pour cette

raison, seuls les *mcqg* sont proposés ici. On peut faire la même remarque pour l'estimateur du maximum de vraisemblance, qui est encore plus complexe à obtenir, et peu différent au final avec celui des *mcqg* (cf. Blanchard et Matyas (1998)).

6) La fin du programme donne deux tests de spécification qui testent l'hypothèse nulle H0 : $s_u^2 = 0$ et illustre la plus ou moins grande difficulté qu'il y a de faire certaines opérations avec les étapes PROC et DATA de *SAS*.

a) Il s'agit de tester $H0 : s_u^2 = 0 \ \ vs \ \ HA : s_u^2 \neq 0$ La statistique de test est :

$$F_{bw} = \frac{\dfrac{T \times \mathbf{e}_b' \mathbf{e}_b}{N - k_b}}{\dfrac{\mathbf{e}_w' \mathbf{e}_w}{N(T-1) - k_w}} = \frac{\dfrac{T \times \left(\hat{s}_u^2 + \dfrac{\hat{s}_e^2}{T} \right)}{N - k_b}}{\dfrac{\hat{s}_e^2}{N(T-1) - k_w}} = \left[\frac{\mathbf{e}^t \mathbf{B}_N \mathbf{e} (N - k_b)^{-1}}{\mathbf{e}^t \mathbf{W}_N \mathbf{e} [N(T-1) - k_w]^{-1}} \right] \text{qui suit un } \textit{Fisher} \text{ à}$$

$(N-k_b), N(T-1)-k_w$ degrés de libertés. Nous verrons un peu plus loin la définition de $\mathbf{W}_n$ et $\mathbf{B}_n$.

On notera que l'on a supposé que $\mathbf{e}_b$ a été calculé sur N observations (sinon on omet le facteur T au numérateur). Dans la mesure où il s'agit seulement d'un test de rapport de deux variances (déjà calculées et disponibles dans le tableau **mcqg**), la mise en oeuvre est élémentaire.

```
/* test du F sig²(u)=0 */

DATA test (KEEP = sig2b sig2eps fbw p nt t n kb kw) ; SET mcqg ;
    fbw = t*sig2b/sig2eps ;
    p = 2*(1-PROBF(fbw,(n-kb),(nt-n-kw))) ;
    IF _N_ le 1 ;

PROC PRINT ; TITLE test du F :  H0 : sig²(u) = 0 ;
```

avec PROBF(x,df) qui donne la probabilité qu'une variable de Fisher soit inférieure ou égal à x avec df degrés de libertés. Pour calculer le nsm, il faut faire 2*(1-PROBF(x,df)). Les résultats sont :

```
      test du F :  H0 : sig²(u) = 0

OBS   N    T     NT    KB    KW     SIG2B       SIG2EPS    FBW       P
 1    18   19   342    4     3    0.038686     .0085249    86.2229   0
```

L'hypothèse nulle est clairement rejetée.

b) Par contre, le deuxième test (test de Breusch-Pagan (1980), dit test LM2, qui est un test bilatéral) qui est aussi un test de $H0 : s_u^2 = 0 \ \ vs \ \ HA : s_u^2 \neq 0$ requiert de calculer la statistique :

$$LM_2 = \frac{NT}{2(T-1)}\left[\frac{\sum_i\left[\sum_t e_{it}\right]^2}{\sum_i\sum_t e_{it}^2} - 1\right]^2 = \frac{NT}{2(T-1)}\left[\frac{\mathbf{e}^t(\mathbf{I}_N \otimes \mathbf{J}_T)\mathbf{e}}{\mathbf{e}^t\mathbf{e}} - 1\right]^2$$ avec e_{it} qui représente les résidus des *mco* et $\mathbf{B}_n = \mathbf{I}_N \otimes \mathbf{J}_T$ (cf. la section 4.B. pour une définition plus précise). Cette statistique suit un c^2 à un degré de liberté. Ce test est un test du modèle à erreurs composées. Pratiquement, ce test n'est pas aisé à mettre en œuvre sous *SAS*, en tout cas avec le type de programmation que nous avons adoptée (ce sera nettement plus simple avec *IML*). La raison essentielle à cette difficulté est qu'avec *SAS* dans les étapes DATA ou PROC, il n'est pas possible d'utiliser des vecteurs ou des matrices mais uniquement des tableaux *SAS*. Il va falloir procéder en plusieurs étapes.

```
/* test LM2 de Breusch-Pagan */

PROC MEANS DATA = mcqg NOPRINT ;            /* somme sur i somme sur t eit^2 */
  ID dum ; VAR umco ; OUTPUT OUT = test3 USS = ssqumco ;    /* dénominateur */

PROC MEANS DATA = mcqg NOPRINT ;                              /* somme sur t */
  ID dum ; VAR umco ; BY country ; OUTPUT OUT = test1 SUM = somteit ;

PROC MEANS DATA = test1 NOPRINT ;                    /* somme sur i someit^2 */
  ID dum ;  VAR somteit ; OUTPUT OUT = test2 USS = sseit2 ;    /* numérateur */

DATA result (KEEP = ssqumco sseit2 nt t n kw kb lmtest p) ;
MERGE test test2 mcqg ; BY dum ;
lmtest = (nt/(2*(t-1)))*((sseit2/ssqumco) - 1)**2 ;
IF _N_ = 1 ; P = 2*(1-PROBCHI(lmtest,1)) ;

PROC PRINT ; TITLE test LM2 de Breusch-Pagan : H0 : sig²(u) = 0 ; RUN ;
```

SUM et USS sont des fonctions *SAS* qui calculent respectivement la somme et la somme des carrés d'une variable (Uncorrected Sum of squares) et PROBCHI(x,df) donne la probabilité qu'une variable du c^2 soit inférieure ou égale à x avec df degrés de libertés.

```
test LM2 de Breusch-Pagan : H0 : sig²(u) = 0

      OBS    SSQUMCO     SSQUMCOI     N     T      NT    KB    KW    LMTEST        P
       1     14.9044      200.024     18    19     342     4     3   1465.55        0
```

Dans ce deuxième cas, on rejette aussi l'hypothèse nulle. On ne rejette donc pas la spécification à erreurs composées. Compte tenu de la complexité de cette partie du programme, on voit les limites à l'utilisation de *SAS* pour ce type de programmation sur panel, et la nécessité de recourir à un autre mode opératoire.

Si l'on voulait faire un test d'Hausman, le programme serait encore plus complexe. Il vaut mieux alors dans ce cas utiliser *IML* comme on va le voir dans la section 4-C (**panel4.sas**) ou en utilisant PROC TSCSREG que nous examinons dans la prochaine section.

Section 3) Les procédures PROC TSCSREG et PROC GLM

SAS offre deux procédures qui permettent d'éviter les fastidieuses programmations précédentes. Néanmoins ces procédures ne proposent que peu de tests statistiques et elles ne peuvent pas s'adapter à des cas plus complexes d'estimation (*VI*, *GMM*, panels incomplets). Pour des estimations usuelles sur panel, elles sont cependant très pratiques. A noter que TSCSREG fait partie du module *SAS-ETS*, alors que PROC GLM fait partie de *SAS-STAT*. Ceci a comme conséquence que la documentation *SAS* sur PROC GLM utilise plutôt un vocabulaire issu de la biométrie alors que celle de PROC TSCSREG utilise un vocabulaire plutôt économétrique.

A) La procédure PROC TSCSREG

Cette procédure (écrite à l'origine par A. R. Gallant et D.J. Drummond, 1979) permet d'estimer un modèle par :

- l'estimateur within avec effets individuels ou effets individuels et temporels.
- par l'estimateur des mcqg avec des effets individuels ou des effets individuels et temporels.

Nous nous limitons dans ce qui suit à des modèles avec des effets individuels (fixes ou aléatoires). Pour cela il suffit de spécifier l'option FIXONE ou RANONE et les dimensions individuelles et temporelles (avec CS=… et TS=…). A noter que ces options sont assez récentes et n'existaient pas dans des versions antérieures de cette procédure (elles ne semblent d'ailleurs documentées uniquement dans l'aide en ligne – *SAS-ETS*).

```
/* panel2.sas */

LIBNAME in 'c:\seminair\' ;

/******************************************************* estimation within */

PROC TSCSREG DATA = in.gasoline CS=18 TS=19 ;      /* 17 dummies + constante */
   MODEL  lngc = lnyn lnpgpp lncarn / FIXONE ;
TITLE within;

/* ou */

PROC TSCSREG DATA = in.gasoline ;
   MODEL  lngc = lnyn lnpgpp lncarn / FIXONE ;
   ID country year ;
TITLE within;

/* ou encore pour avoir 18 dummies sans la constante */

PROC TSCSREG DATA = in.gasoline  CS=18 TS=19 ;
   MODEL  lngc = lnyn lnpgpp lncarn / FIXONE NOINT ;
```

Avec cette instruction, il n'y a pas de possibilité de faire un RESTRICT. Par conséquent, on retrouve les problèmes d'interprétation des estimations des effets fixes, vus dans la section 2.

Pour les *mcqg* :

```
/*********************************************************** estimation mcqg */

PROC TSCSREG DATA = in.gasoline CS=18 TS=19 ;
   MODEL  lngc = lnyn lnpgpp lncarn / RANONE FULLER ;
TITLE mcqg ;


/* ou */

PROC TSCSREG DATA = in.gasoline ;
   MODEL  lngc = lnyn lnpgpp lncarn / RANONE FULLER ;
   ID country year ;
TITLE mcqg ;
```

On notera l'utilisation de l'instruction ID pour définir les noms des variables contenant l'identifiant de l'individu (le pays) et de la dimension temporelle (year) à la place des options CS=… et TS=… **Dans tous les cas (avec CS, TS ou ID), les données doivent avoir été triées selon country et year impérativement.** La méthode FULLER est par défaut (les autres sont PARK et DASILVA).

Voici les résultats du *within* :

TSCSREG Procedure

Dependent Variable: LNGC LNGC

Model Description

Estimation Method	FIXONE
Number of Cross Sections	18
Time Series Length	19

Model Variance

SSE	2.736491	DFE	321
MSE	0.008525	Root MSE	0.09233
RSQ	0.9734		

F Test for No Fixed Effects

Numerator	DF: 17	F value:	83.9608
Denominator	DF: 321	Prob.>F:	0.0000

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|	Variable Label
CS 1	1	-0.769395	0.044576	-17.260136	0.0001	Cross Sec
CS 2	1	-0.889700	0.046675	-19.061555	0.0001	Cross Sec
CS 3	1	-0.013411	0.030596	-0.438311	0.6615	Cross Sec
CS 4	1	-0.665795	0.045770	-14.546614	0.0001	Cross Sec
CS 5	1	-0.850479	0.050125	-16.967099	0.0001	Cross Sec
CS 6	1	-0.905382	0.041336	-21.902977	0.0001	Cross Sec
CS 7	1	-0.718141	0.055475	-12.945402	0.0001	Cross Sec
CS 8	1	-0.462926	0.056927	-8.131920	0.0001	Cross Sec
CS 9	1	-0.822703	0.061596	-13.356517	0.0001	Cross Sec
CS 10	1	-0.679323	0.047365	-14.342175	0.0001	Cross Sec
CS 11	1	-0.820459	0.044950	-18.252618	0.0001	Cross Sec
CS 12	1	-0.838550	0.048751	-17.200526	0.0001	Cross Sec
CS 13	1	-1.373474	0.099485	-13.805781	0.0001	Cross Sec
CS 14	1	-0.028908	0.182466	-0.158431	0.8742	Cross Sec
CS 15	1	-0.652748	0.034719	-18.801125	0.0001	Cross Sec
CS 16	1	-0.545263	0.053527	-10.186781	0.0001	Cross Sec
CS 17	1	-0.709803	0.045882	-15.470127	0.0001	Cross Sec
INTERCEP	1	3.055251	0.219600	13.912827	0.0001	Intercept
LNYN	1	0.662250	0.073386	9.024191	0.0001	LNYN
LNPGPP	1	-0.321702	0.044099	-7.294964	0.0001	LNPGPP
LNCARN	1	-0.640483	0.029679	-21.580447	0.0001	LNCARN

On remarquera que :

- les écarts-type des coefficients estimés (et donc les T de Student et les nsm) sont désormais bien calculés,
- on dispose automatiquement des coefficients estimés (et des écarts-type estimés) des indicatrices individuelles (moins une compte tenu de la présence de la constante). On retrouve

les résultats et les problèmes d'interprétations des estimations des effets fixes vus dans la section 2.

- un test d'absence des effets fixes individuels (test de Fisher) est aussi automatiquement reporté c'est à dire un test de $H0 : a_1 = ... = a_N = a$ vs HA: H0 est fausse en calculant $Fc = \frac{SCR_{mco} - SCR_{with}}{SCR_{with}} \times \frac{NT - N - kw}{N - 1}$ ou SCR_{with} et SCR_{mco} sont respectivement la somme des carrés des résidus du *within* et des *mco* (cf. aussi la section .4.C).

Voici les résultats des *mcqg* :

```
TSCSREG Procedure

Dependent Variable: LNGC   LNGC

                                  Model Description

                          Estimation Method          RANONE
                          Number of Cross Sections   18
                          Time Series Length         19

                             Variance Component Estimates

                     SSE          3.043543    DFE             338
                     MSE          0.009005    Root MSE   0.094892
                     RSQ            0.8302

                  Variance Component for Cross Sections        0.044041
                  Variance Component for Error                 0.008525

                            Hausman Test for Random Effects

                              Degrees of Freedom:        3
                     m value:     125.0442    Prob. > m:       0.0000

                                 Parameter Estimates

                       Parameter      Standard    T for H0:               Variable
        Variable  DF      Estimate         Error   Parameter=0  Prob > |T|     Label

        INTERCEP  1       2.020299      0.188218     10.733800      0.0001  Intercept
        LNYN      1       0.559961      0.060075      9.320968      0.0001  LNYN
        LNPGPP    1      -0.411775      0.040192    -10.245298      0.0001  LNPGPP
        LNCARN    1      -0.608107      0.025764    -23.603198      0.0001  LNCARN
```

On notera deux points :

- Les résultats ne sont pas identiques à ceux obtenus précédemment. La différence provient de la façon avec laquelle on a estimé s_u^2 et s_e^2. Ici c'est la méthode de Battese et Fuller (1974) qui est utilisée (dite aussi «fitting constant méthod » ou méthode d'Henderson n°3 (1953), cf. *SAS-ETS* p. 880-881, Fuller et Battese (1974) et Maddala et Mount (1973)).

- Un test d'Hausman est aussi automatiquement fourni (cf. la section 4-C, cf. **panel5.sas**).

- Il est aussi possible de supposer que les perturbations e_{it} de $v_{it} = m_i + e_{it}$ suivent un processus de type autorégressif ou moyenne mobile (cf. *SAS-ETS* p. 884-886, méthode de Park et de Da Silva).

B) La procédure PROC GLM

Cette procédure représente une autre solution pour obtenir les estimations *within*. Les données doivent aussi avoir été triées selon l'identifiant de l'individu (country) et sur l'année (year). Cette instruction permet aussi d'obtenir les estimations *within* avec des effets temporels (fixes) mais aussi avec des effets individuels et temporels (toujours pas de RESTRICT possible).

```
/******************************************************** estimation within */

PROC GLM DATA = in.gasoline ;
   ABSORB country ;
   MODEL  lngc = lnyn lnpgpp lncarn ;
RUN ;
```

Pour des données comportant un grand nombre d'individus, cette instruction est plus rapide et génère un volume de résultat plus limité car elle omet les estimations des constantes individuelles. Si l'on souhaite obtenir les estimations des effets fixes individuels et de leurs écarts-type, on fera :

```
/******************************************************** estimation within */

PROC GLM DATA = in.gasoline ;                          /* inverse généralisée */
   CLASS country ;
   MODEL  lngc = lnyn lnpgpp lncarn country / solution;

/* ou */

PROC GLM DATA = in.gasoline ;                         /* on omet la constante */
   CLASS country ;
   MODEL  lngc = lnyn lnpgpp lncarn country / NOINT solution;
   RUN ;
```

La première des deux instructions PROC GLM qui précèdent donnent les estimations du modèle à effets fixes, en omettant la dernière indicatrice (les USA) mais avec un terme constant, en utilisant une inverse généralisée du fait de la colinéarité (identique à la solution 1 du tableau 1). La seconde fournit les estimations du même modèle sans la constante mais avec toutes les indicatrices (identique à la solution 3 du tableau 1). L'option « solution » demande à faire afficher les coefficients estimés en plus de l'analyse de la variance. Du fait de l'instruction ABSORB, PROC GLM sera très rapide même en présence d'un grand nombre d'individus.

Voici les résultats du *within* avec l'instruction ABSORB :

General Linear Models Procedure

Number of observations in data set = 342

General Linear Models Procedure

Dependent Variable: LNGC LNGC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	100.00647355	5.00032368	586.56	0.0001
Error	321	2.73649080	0.00852489		
Corrected Total	341	102.74296435			

R-Square	C.V.	Root MSE	LNGC Mean
0.973366	2.149096	0.09233035	4.29624203

Source	DF	Type I SS	Mean Square	F Value	Pr > F
COUNTRY	17	85.68228007	5.04013412	591.23	0.0001
LNYN	1	10.06258302	10.06258302	1180.38	0.0001
LNPGPP	1	0.29143370	0.29143370	34.19	0.0001
LNCARN	1	3.97017676	3.97017676	465.72	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
LNYN	0.6622496560	9.02	0.0001	0.07338604
LNPGPP	-.3217024604	-7.29	0.0001	0.04409925
LNCARN	-.6404828807	-21.58	0.0001	0.02967885

On notera que les estimations des écarts-type des coefficients estimés (et les T de Student et les nsm) sont là aussi bien calculés et que compte tenu de l'instruction ABSORB, on ne dispose pas des estimations des effets fixes, ce qui n'est pas nécessairement un problème. Très fréquemment, ces estimations ne sont pas reportées pour des raisons de gain de place et/ou parce qu'elles ne sont pas utiles à l'analyse des résultats.

Section 4) Estimations *mco*/*within*/*mcqg* avec un programme *IML*

Une deuxième possibilité de programmation des estimateurs *between, within et mcqg* réside dans l'utilisation du langage *IML* qui est un langage de programmation assez classique incluant de plus de puissantes instructions de calcul matriciel. Il y a néanmoins plusieurs variantes possibles.

A) Un programme *IML* sans boucle sur les données et sans les produits de *Kronecker*

La solution la plus simple pour programmer les différents estimateurs vus précédemment réside dans l'utilisation d'*IML* sans utiliser les produits de *Kronecker* ni de boucle sur le fichier de données. Ne pas utiliser de produit de *Kronecker* implique alors de disposer de toutes les variables nécessaires y compris des moyennes individuelles et des écarts aux moyennes individuelles. C'est ce que nous faisons dans une première étape avec la création du tableau **datiml**. Attention, dans ce tableau les *N* moyennes individuelles sont dupliquées *T* fois.

```
/* panel3.sas */

/* initialisations */

LIBNAME in 'c:\seminair\' ;

PROC MEANS DATA = in.gasoline MEAN NOPRINT ; BY country ;
     VAR lngc lnyn lnpgpp lncarn ;
     OUTPUT OUT = mtab1 MEAN = mlngc mlnyn mlnpgpp mlncarn;

DATA datiml (KEEP = constant lngc   lnyn  lnpgpp  lncarn
                          mlngc mlnyn mlnpgpp mlncarn
                          wlngc wlnyn wlnpgpp wlncarn) ;
MERGE in.gasoline mtab1 ; BY country ;
  constant = 1 ;
  wlngc   =  lngc   - mlngc   ;
  wlnyn   =  lnyn   - mlnyn   ;
  wlnpgpp =  lnpgpp - mlnpgpp ;
  wlncarn =  lncarn - mlncarn ;
```

On peut alors commencer le programme *IML*. On lit les variables utiles de **datiml** que l'on stocke dans les vecteurs et matrices dont on aura besoin. L'inconvénient de cette méthode est que la place mémoire requise risque d'être conséquente si *N* et/ou *T* sont grands. Si c'est le cas, *SAS* utilisera la mémoire virtuelle (le disque dur) pour stocker les variables, ce qui ralentira considérablement l'exécution du programme. Ajoutons de plus que *SAS* semble gérer fort mal la mémoire RAM, ce qui l'amène à faire de nombreux accès disque. Au total, un programme stockant toutes les données en mémoire peut être lent.

```
/*------------------------------ procédure IML -------------------------- */

PROC IML ;

/************************************************** lecture des données */

USE datiml ;

READ ALL VAR {lngc} INTO y ;
READ ALL VAR {constant lnyn lnpgpp lncarn} INTO x ;

READ ALL VAR {mlngc} INTO yb ;                          /* NT obs. attention */
READ ALL VAR {constant mlnyn mlnpgpp mlncarn} INTO xb ;

READ ALL VAR {wlngc} INTO yw ;
READ ALL VAR {wlnyn wlnpgpp wlncarn} INTO xw ;
```

On affiche alors des informations sur le fichier de données et sur les variables. Notez qu'on ne peut pas utiliser des étapes PROC ou DATA dans un programme *IML* (qui est une procédure, avec ses propres instructions ; notez cependant que les scalaires, vecteurs et matrices sont stockés en mémoire et non pas dans un tableau, à moins qu'ils ne soient sauvegardés comme tels).

```
/****************************************** info. et stat. des. avec IML */

TITLE informations sur le fichier de données ;

SHOW CONTENTS ; PRINT ' ' / ;

SUMMARY STAT{N Nmiss MEAN STD MIN MAX} VAR _NUM_ ; PRINT ' ' / ;
```

L'instruction SUMMARY est très puissante car elle peut calculer des statistiques par groupes (des moyennes par pays, par exemple, avec l'instruction CLASS) et les stocker dans des vecteurs.

On obtient alors comme résultat :

Pour le SHOW CONTENTS (équivalent du PROC CONTENTS) :

```
informations sur le fichier de données

DATASET : WORK.DATIML.DATA
   VAR NAME   TYPE   SIZE
   LNGC       NUM       8
   LNYN       NUM       8
   LNPGPP     NUM       8
   LNCARN     NUM       8
   MLNGC      NUM       8
   MLNYN      NUM       8
   MLNPGPP    NUM       8
   MLNCARN    NUM       8
   CONSTANT   NUM       8
   WLNGC      NUM       8
   WLNYN      NUM       8
   WLNPGPP    NUM       8
   WLNCARN    NUM       8
   Number of Variables:    13
   Number of Observations: 342
```

Pour le SUMMARY (équivalent du PROC MEANS) :

```
informations sur le fichier de données

   Nobs  Variable    N  NMISS        MEAN         STD         MIN         MAX
   ------------------------------------------------------------------------------
    342  LNGC      342      0     4.29624     0.54891     3.38021     6.15664
         LNYN      342      0    -6.13943     0.63459    -8.07252    -5.22123
         LNPGPP    342      0    -0.52310     0.67822    -2.89650     1.12531
         LNCARN    342      0    -9.04180     1.21890   -13.47518    -7.53618
         MLNGC     342      0     4.29624     0.50127     3.72965     5.76636
         MLNYN     342      0    -6.13943     0.59318    -7.81621    -5.44856
         MLNPGPP   342      0    -0.52310     0.66614    -2.70917     0.73938
         MLNCARN   342      0    -9.04180     1.08439   -12.45886    -7.78109
         CONSTANT  342      0     1.00000           0     1.00000     1.00000
         WLNGC     342      0  -1.558E-16     0.22368    -0.75090     1.29564
         WLNYN     342      0  -6.493E-17     0.22549    -0.73757     0.53978
         WLNPGPP   342      0  1.4284E-17     0.12747    -0.53401     0.38593
         WLNCARN   342      0  1.2466E-16     0.55659    -2.28999     1.35105
```

On vérifie bien que la somme des écarts par rapport à la moyenne est égale à 0! (les 4 dernières lignes). Comme l'instruction SUMMARY permet de calculer des statistiques par groupe (par sexe, par exemple, à l'aide d'une option CLASS) ainsi que de sauvegarder les statistiques descriptives dans des scalaires ou des vecteurs, on aurait donc pu utiliser ici cette possibilité pour calculer les moyennes par individu, plutôt que de les stocker dans le tableau **datiml**. On aurait pu les lire aussi à partir d'un tableau séparé (**mtab1)** mais le fait qu'elles soient alors en nomnre *N* cécessite d'adapter légèrement le programme qui suit lors des estimations en between (cf. supra).

On initialise certains paramètres (*n,t,nt, kb, kw*) utilisés plus loin et le nom des variables utilisés pour l'affichage des résultats (*kb*.est le nombre de régresseurs du *between* et des *mco* car il n'y a pas de régresseur constant dans la dimension temporelle).

On procède alors aux estimations en utilisant les formules usuelles de calcul matriciel avec les variables transformées ou non (*mco*).

```
/******************************************************** initialisations */

t = 19 ; n = 18 ; nt = n * t ; kb = 4 ; kw = 3 ;
nomvar = { constante , lnyn , lnpgpp , lncarn } ; /* pour affichage */
nomvarw = { lnyn , lnpgpp , lncarn } ;            /* des résultats */

/******************************************************** estimation mco */

bmco = INV(x`*x)*x`*y ; rmco = y - x*bmco ;
sig2mco = SSQ(rmco)/(nt-kb) ;  /* kb = k = nb de régresseurs between et mco */
vbmco = sig2mco*INV(x`*x) ;
ecmco = SQRT(VECDIAG(vbmco)) ; stmco = bmco/ecmco ;
nsmmco = 2*(1-PROBT(ABS(stmco),(nt-kb))) ;

TITLE résultats mco ; PRINT "estimation mco" ;
PRINT nomvar bmco ecmco stmco nsmmco ; PRINT "sig²mco " sig2mco ;

PRINT "matrice de var-cov des mco" ; PRINT vbmco ; PRINT ' ' / ;
```

On notera les opérateurs *SAS* suivants qui peuvent porter sur des scalaires, des vecteurs ou des matrices :

- SSQ pour calculer la somme des carrés des éléments d'un vecteur,
- VECDIAG pour extraire la diagonale d'une matrice et la stocker dans un vecteur,
- PROBT pour calculer la probabilité qu'une variable de Student soit plus petite ou égale à x (pour un nombre donné de degrés de liberté).

```
/****************************************************** estimation within */

bwith  = INV(xw`*xw)*xw`*yw ; rwith = yw - xw*bwith ;
sig2eps  = SSQ(rwith)/(nt-n-kw) ;
vbwith = sig2eps*INV(xw`*xw) ;
ecw = SQRT(VECDIAG(vbwith)) ; stw = bwith/ecw ;
nsmw = 2*(1-PROBT(ABS(stw),(nt-n-kw))) ;

TITLE résultats within ; PRINT "estimation within" ;
PRINT nomvarw bwith ecw stw nsmw ;
PRINT "sig²eps " sig2eps ;

PRINT "matrice de var-cov du within" ; PRINT vbwith ; PRINT ' ' / ;
```

Voici les résultats (on ne reproduit pas les résultats des *mco*).

```
résultats within

                              estimation within

                NOMVARW     BWITH       ECW       STW      NSMW
                LNYN    0.6622497  0.073386 9.0241906         0
                LNPGPP  -0.321702 0.0440993 -7.294964 2.355E-12
                LNCARN  -0.640483 0.0296789 -21.58045         0

                                       SIG2EPS
                                sig²eps  0.0085249

                           matrice de var-cov du within

                              VBWITH
                            0.0053855 0.0002896 -0.002049
                            0.0002896 0.0019447  0.0000886
                           -0.002049  0.0000886  0.0008808
```

```
/******************************************************* estimation between */

xb1 = xb/SQRT(t) ;
yb1 = yb/SQRT(t) ;         /* pour estimer sur N observations et non sur NT */

bbetw  = INV(xb1`*xb1)*xb1`*yb1 ; rbetw = yb1 - xb1*bbetw ;
sig2b  = SSQ(rbetw)/(n-kb) ;
vbbetw = sig2b*INV(xb1`*xb1) ;
ecb = SQRT(VECDIAG(vbbetw)) ; stb = bbetw/ecb ;
nsmb = 2*(1-PROBT(ABS(stb),(n-kb))) ;

TITLE résultats between ; PRINT " estimation between" ;
PRINT nomvar bbetw ecb stb nsmb ;
PRINT "sig²b " sig2b ;

PRINT "matrice de var-cov du between" ; PRINT vbbetw ; PRINT ' ' / ;
```

Il faut noter que la régression *within* est estimée sur *NT* observations alors que la régression *between* ne doit l'être que sur *N* individus. Or notre opération de fusion a dupliquée *T* fois les moyennes individuelles. En effet, **datiml** est la fusion de **gasoline** (*NT* obs.) avec **mtab1** (*N* obs.). Pour se ramener à une estimation sur *N* moyennes individuelles et non sur *NT*, on multiplie les variables du modèle *between* par $\frac{1}{\sqrt{T}}$ [9](Baltagi, 1995). Il aurait peut-être été plus simple de ne pas fusionner les *N* moyennes avec les *NT* observations et de les lire dans un tableau séparé (**mtab1**).

Voici les résultats du *Between.*

[9] Une autre solution serait de diviser sig2b et $\mathbf{X}'_b\mathbf{X}_b$ par t.

```
résultats between

                              estimation between

              NOMVAR        BBETW       ECB       STB      NSMB
              CONSTANTE 2.5416298 0.5267844 4.8248004 0.0002697
              LNYN      0.9675764 0.1556662 6.2157123 0.0000225
              LNPGPP     -0.96355 0.1329214 -7.249022 4.2303E-6
              LNCARN    -0.795299 0.0824742 -9.643003 1.4626E-7

                                       SIG2B
                                sig²b  0.0386864

                           matrice de var-cov du between

                              VBBETW
                            0.2775018  0.0443199 -0.014951   0.0012248
                            0.0443199  0.024232  -0.016955  -0.010571
                           -0.014951  -0.016955   0.0176681  0.0088368
                            0.0012248 -0.010571   0.0088368  0.006802
```

On estime alors le modèle par les *mcqg*.

```
/*********************************************************** estimation mcqg */

sig2u = sig2b - sig2eps/t ;
theta  = 1 - SQRT((sig2eps)/(t*sig2b)) ;

yg = y - theta*yb ;
xg = x - theta *xb ;

bmcqg  = INV(xg`*xg)*xg`*yg ; rmcqg = yg - xg*bmcqg ;
sig2g = SSQ(rmcqg) ; vbmcqg = sig2eps*INV(xg`*xg) ;
ecg = SQRT(VECDIAG(vbmcqg)) ; stg = bmcqg/ecg ;
nsmg = 2*(1-PROBT(ABS(stg),(nt-n-kw))) ;

TITLE résultats mcqg ; PRINT "estimation mcqg" ;
PRINT nomvar bmcqg ecg stg nsmg ;
PRINT "matrice de var-cov des mcqg" ; PRINT vbmcqg ; PRINT " " ;
PRINT "sig2eps  = " sig2eps " sig2b  = " sig2b " sig2u = " sig2u "
      " sig2g = " sig2g " theta = " theta ; PRINT ' ' / ;

QUIT ;                                                  /* fin du programme IML */
```

dont les résultats sont :

```
résultats mcqg

              NOMVAR        COEFF        SB         T       NSM
              CONSTANTE  1.9966984  0.1782353  11.202598      0
              LNYN       0.5549857  0.0571744   9.7068895     0
              LNPGPP    -0.420389   0.0386571  -10.87482      0
              LNCARN    -0.60684    0.024672   -24.59636      0

                                      S2
                              sig²  0.0085249

                matrice estimée de Var-Cov des coeff. estimés

                        VCV
                   0.0317678 0.0080955 0.0009754 -0.002278
                   0.0080955 0.0032689 -0.000475 -0.001297
                   0.0009754 -0.000475 0.0014944 0.0003442
                   -0.002278 -0.001297 0.0003442 0.0006087

             SIG2EPS              SIG2B               SIG2U              SIG2G
 sig2eps =  0.0085249   sig2b  =  0.0386864   sig2u =  0.0382377   sig2G =  0.0091175

                               THETA
                    theta =  0.8923067

                    matrice de var-cov des mcqg (avec sig2u)

                      VBMCQG
                   0.0339761  0.0086582  0.0010432 -0.002436
                   0.0086582  0.0034961 -0.000508 - 0.001387
                   0.0010432 -0.000508   0.0015983  0.0003681
                  -0.002436  -0.001387   0.0003681  0.000651
```

On note la différence d'estimation de la matrice de variance-covariance des coefficients estimés par les *mcqg* selon que l'on utilise $\hat{s}_u^2$ ou $\hat{s}_e^2$. Si on présente cette différence en termes d'écarts type estimés des coefficients estimés de Student , on obtient :

Tableau n°3

$\hat{s}_{\hat{b}}$ avec sig2g	$\hat{s}_{\hat{b}}$ avec sig2eps	T avec sig2g	T avec sig2eps
0.184326	0.1782353	10.83243	11.20259
0.059128	0.0571744	9.38614	9.70688
0.039978	0.03866571	-10.51548	-10.87482
0.025515	0.024672	-23.78362	-24.59636

ce qui n'est pas négligeable d'autant plus que ces différences seront amplifiées lors de l'utilisation du test d'Hausman (cf. la section C).

Le principal défaut de ce programme est qu'il nécessite de disposer de toutes les variables de base et transformées (moyennes individuelles et écarts aux moyennes individuelles) et que toutes les variables pour toutes les observations doivent être stockées en mémoire. SI N et/ou T sont grands, il est vraisemblable que l'exécution du programme sera ralentie.

B) Un programme *IML* avec les produits de *kronecker*

```
/* panel4.sas */

LIBNAME in 'c:\seminair\' ;

DATA datiml(KEEP = constant lngc  lnyn  lnpgpp  lncarn) ; SET in.gasoline ;
constant = 1 ;
```

Le début du programme est quasiment identique à celui du programme précédent. On remarquera quand même que seules les variables de base sont lues et la constante créée.

```
/*------------------------------ procédure IML --------------------------- */

PROC IML ;

/************************************************** lecture des données */

USE datiml ;

READ ALL VAR {lngc} INTO y ;
READ ALL VAR {constant lnyn lnpgpp lncarn} INTO x ;

/****************************************** info. et stat. des. avec IML */

TITLE informations sur le fichier de données ;

SHOW CONTENTS ; PRINT ' ' / ;

SUMMARY STAT{N Nmiss MEAN STD MIN MAX} VAR _NUM_ ; PRINT ' ' / ;

/****************************************************** initialisations */

t = 19 ; n = 18 ; nt = n * t ; kb = 4 ; kw = 3 ;
nomvar = { constante , lnyn , lnpgpp , lncarn } ;
nomvarw = { lnyn , lnpgpp , lncarn } ;
```

C'est à ce stade qu'apparaît la spécificité du programme : l'utilisation des produits de *Kronecker* pour créer les matrices $\mathbf{W}_n$ et $\mathbf{B}_n$. $\mathbf{I}_{nt}$ représente la matrice identité de dimension (*NT*,*NT*) et $\mathbf{J}_t$ la matrice unité. Par conséquent, $\mathbf{W}_n\mathbf{Y}$ crée un vecteur dont l'élément caractéristique est $y_{it} - y_{i.}$ et $\mathbf{B}_n\mathbf{Y}$ crée un autre vecteur dont l'élément caractéristique est $y_{i.} = \frac{1}{N}\sum_i y_{it}$. Bien sûr, il ne faut pas appliquer l'opérateur $\mathbf{W}_n$ à la constante. C'est pourquoi la matrice transformée n'est crée qu'à partir de la colonne n°2 de $\mathbf{X}$, la constante étant en première position. Dans le cas, contraire, on obtiendrait une colonne de 0, rendant $\mathbf{X}_w^t\mathbf{X}_w$ non inversible.

```
/************************************************** estimation within */

Wn = I(nt) - (1/t)*I(n)@J(t) ;                      /* Int - (1/t)*(In⊗Jt) */

yw = Wn*y ; xw = Wn*x[,2:kb] ;

bwith  = INV(xw`*xw)*xw`*yw ; rwith = yw - xw*bwith ;
sig2eps  = SSQ(rwith)/(nt-n-kw) ;
vbwith = sig2eps*INV(xw`*xw) ;
ecw = SQRT(VECDIAG(vbwith)) ; stw = bwith/ecw ;
nsmw = 2*(1-PROBT(ABS(stw),(nt-n-kw))) ;

TITLE résultats within ; PRINT "estimation within" ;
PRINT nomvarw bwith ecw stw nsmw ;
PRINT "sig²eps " sig2eps ;
PRINT "matrice de var-cov du within" ; PRINT vbwith ; PRINT ' ' / ;
```

```
/*************************************************** estimation between */

Bn = (1/t)*I(n)@J(t) ;                                  /* (1/t)*(In⊗Jt) */

yb = Bn*y ; xb = Bn*x ;

yb1 = yb/SQRT(t) ; xb1 = xb/SQRT(t) ;

bbetw  = INV(xb1`*xb1)*xb1`*yb1 ; rbetw = yb1 - xb1*bbetw ;
sig2b  = SSQ(rbetw)/(n-kb) ;
vbbetw = sig2b*INV(xb1`*xb1) ;
ecb = SQRT(VECDIAG(vbbetw)) ; stb = bbetw/ecb ;
nsmb = 2*(1-PROBT(ABS(stb),(n-kb))) ;

TITLE résultats between ; PRINT "estimation between" ;
PRINT nomvar bbetw ecb stb nsmb ;
PRINT "sig²b " sig2b ;
PRINT "matrice de var-cov du between" ; PRINT vbbetw ; PRINT ' ' / ;
```

```
/*************************************************** estimation mcqg*/

sig2u = sig2b - sig2eps/t ;
theta  = 1 - SQRT((sig2eps)/(t*sig2b)) ;

yg = y - theta*yb ;
xg = x - theta*xb ;

bmcqg  = INV(xg`*xg)*xg`*yg ; rmcqg = yg - xg*bmcqg ;
sig2g = SSQ(rmcqg) ; vbmcqg = sig2eps*INV(xg`*xg) ;
ecg = SQRT(VECDIAG(vbmcqg)) ; stg = bmcqg/ecg ;
nsmg = 2*(1-PROBT(ABS(stg),(nt-n-kw))) ;

TITLE résultats mcqg ; PRINT "estimation mcqg" ;
PRINT nomvar bmcqg ecg stg nsmg ;
PRINT "matrice de var-cov des mcqg" ; PRINT vbmcqg ; PRINT " " ;
PRINT "sig2eps = " sig2eps " sig2b  = " sig2b  " sig2u = " sig2u
      "sig2g = "   sig2g   "theta = " theta ;
PRINT ' ' / ;

QUIT ;                                                  /* fin du programme IML */
```

Deux effets opposés sont en jeu dans la détermination de la vitesse d'exécution du programme :

- Un effet tendant à une plus grande vitesse d'exécution compte tenu que seules les variables de base sont lues à partir du fichier,

- Un effet négatif du à la création des opérateurs $\mathbf{B}_n$ et $\mathbf{W}_n$ qui sont de dimension (372,372). Il est probable que pour des tailles d'échantillon plus conséquent, l'effet négatif l'emporterait encore plus sur l'effet positif. On peut donc critiquer **panel4.sas** de ce point de vue encore plus que **panels3.sas**.

C) Un programme *IML* avec une boucle sur les données et avec les produits de *kronecker*

Le début du programme est inchangé.

```
/* panel5.sas */

LIBNAME in 'c:\seminair\' ;

DATA datiml ; SET in.gasoline ; constant = 1 ;
```

```
/*------------------------------ procédure IML -------------------------- */

PROC IML ;

/****************************************************** lecture des données */

USE datiml ;

/******************************************************** initialisations */

t = 19 ; n = 18 ; nt = n * t ; kb = 4 ; kw = 3 ;
nomvar = { constante , lnyn , lnpgpp , lncarn } ;
nomvarw = { lnyn , lnpgpp , lncarn } ;

/****************************************** info. et stat. des. avec IML */

TITLE informations sur le fichier de données ;

SHOW CONTENTS ; PRINT ' ' / ;

SUMMARY STAT{N Nmiss MEAN STD MIN MAX} VAR _NUM_ ; PRINT ' ' / ;
```

La première nouveauté est dans le module de lecture des données. Plutôt que de lire et d'amener en mémoire les *NT* observations, on va lire les données individu par individu et faire les calculs de même. Par conséquent, les opérateurs $\mathbf{W}_n$ et $\hat{\mathbf{A}}_n$ ne sont plus que de dimension (*T*,*T*), soit (19,19), ce qui implique un stockage en mémoire très réduit et par suite une accélération des calculs (qui serait encore plus marquée si le modèle était estimé par les *VI*). Il est fort probable que le temps d'exécution de **panel5.sas** serait sensiblement plus faible que celui de **panel4.**sas pour des échantillons plus conséquents (*N* ou *T* très grands). Un avantage supplémentaire de ce programme est qu'il s'adapte facilement aux panels incomplets (il suffit de garder dans le fichier de données une variable *Ti* donnant le nombre d'années de présence de la firme i et d'adapter l'instruction READ en conséquence).

```
/***************************** module de lecture des données par individu */

START readata ;

  deb = i ; fin = i + t - 1 ; rec = deb:fin ;

  READ POINT rec VAR {lngc} INTO y ;
  READ POINT rec VAR {constant lnyn lnpgpp lncarn} INTO x ;

  wn = I(t) - (1/t)@J(t) ;       /* utilisation des produits de kronecker */
  bn = (1/t)@J(t) ;

  yb = bn*y ; xb = bn*x ;
  yw = wn*y ; xw = wn*x[,2:kb] ;

  yb1 = yb/SQRT(t) ; xb1 = xb/SQRT(t) ;

  i = fin ;

FINISH readata ;
```

Afin d'éviter les répétitions dans le programme, on va créer un module (sous-programme ou une procédure) d'édition des résultats nommés imp.

```
/************************************** module impression des résultats */

START imp(methode,nomvar,coeff,sb,t,nsm,s2,vcv) ;

PRINT methode ; PRINT nomvar coeff sb t nsm ; PRINT "sig² " s2 ;

PRINT "matrice estimée de Var-Cov des coeff. estimés" ; PRINT vcv ;

FINISH imp ;
```

On utilise ici le passage de variables à la procédure, ce qui permet d'imprimer des résultats de différentes méthodes avec un seul module. Vous noterez que les variables du module sont locales, il n'y aura donc aucun conflit entre *t* (le nombre d'observations dans la dimension temporelle) et *t* le vecteur de T de Student qui est passé à la procédure.

Dans le même esprit, on définit un module de calcul de l'estimation de la variance des perturbations, de la matrice de variance-covariance des coefficients estimés, de leurs écarts-type, de leurs T de student et de leurs nsm qui sera utilisé lors des *mco*, du *between*, du *within* et des *mcqg* car nous avons 4 fois de suite à faire quasiment les mêmes calculs (des mco sur données transformées ou non). Néanmoins, on récupère les résultats dans des variables globales, c'est à dire dont les valeurs pourront être utilisées dans la suite du programme.

```
/****************************************** module calcul des estimateurs */

START estim(ss,nobs,k,dfcor,xx,b) GLOBAL (s2,vb,ec,st,nsm) ;

  s2  = ss/(nobs-dfcor-k) ;
  vb  = s2*INV(xx) ;
  ec  = SQRT(VECDIAG(vb)) ; st = b/ec ;
  nsm = 2*(1-PROBT(ABS(st),(nobs-dfcor-k))) ;

FINISH estim ;
```

On va donc avoir besoin de plusieurs étapes de lecture des données. Tout d'abord, on calcule, pour chaque estimateur, les matrices $\mathbf{X}'\mathbf{X}$ et $\mathbf{X}'\mathbf{y}$ puis les estimations des coefficients pour chaque estimateur. Ceci fait, on fait une seconde boucle sur le fichier de données pour calculer les résidus de chaque estimateur, ce qui permet de calculer les matrices de variance-covariance estimées des coefficients estimés.

```
/************************************** 1ère boucle sur les observations */

xx = 0 ; xwxw = 0 ; xbxb = 0 ; xy = 0 ; xwyw = 0 ; xbyb = 0 ; ssi = 0 ;

DO i = 1 TO nt ;

  RUN readata ;     /* on lit les variables pour les t obs. de l'individu i */
  xx   = xx   + x`*x   ; xy   = xy   + x`*y   ;
  xwxw = xwxw + xw`*xw ; xwyw = xwyw + xw`*yw ;
  xbxb = xbxb + xb`*xb ; xbyb = xbyb + xb`*yb ;

  xixi = x`*x           ; xiyi = x`*y           ;     /* mco par individu */
  bi = INV(xixi)*xiyi   ; ssi = ssi + SSQ(yi - x*bi)     /* pour les tests */

END ;

bmco   = INV(xx)*xy     ;
bwith  = INV(xwxw)*xwyw ;
bbetw  = INV(xbxb)*xbyb ;
```

Ayant les coefficients estimés par les 3 méthodes, on entame une nouvelle phase de lecture du fichiers de données pour calculer, entre autres, les sommes des carrés des résidus.

```
/************************************** 2ème boucle sur les observations */

ss = 0 ; ssw = 0 ; ssb = 0 ; num = 0 ; denum = 0 ;

DO i = 1 TO nt ;

  RUN readata ;     /* on lit les variables pour les t obs. de l'individu i */

  ss   = ss  + T(y)*y   - T(bmco)*T(X)*y    ;
  ssw  = ssw + T(yw)*yw - T(bwith)*T(Xw)*yw ;
  ssb  = ssb + T(yb)*yb - T(bbetw)*T(Xb)*yb ;

  umco = y - x*bmco ;                                /* utile pour les tests */
  denum = denum + SSQ(umco) ; num = num + SUM(umco)**2 ;

  i = fin ;

END ;
```

On aurait pu aussi calculer les sommes des carrés des résidus de chaque estimateur à partir d'un calcul explicite de ces résidus du style :

```
/************************ autre solution : 2ème boucle sur les observations */

ss = 0 ; ssw = 0 ; ssb = 0 ; num = 0 ; denum = 0 ;

DO i = 1 TO nt ;

  RUN readata ;     /* on lit les variables pour les t obs. de l'individu i */

  resmco = y  - x*bmco   ; ss    = ss   + SSQ(resmco) ;
  resw   = yw - xw*bwith ; ssw   = ssw  + SSQ(resw)   ;
  resb   = yb - xb*bbetw ; ssb   = ssb  + SSQ(resb)   ;

  umco = y - x*bmco ;                                /* utile pour les tests */
  denum = denum + SSQ(umco) ; num = num + SUM(umco)**2 ;

  i = fin ;

END ;
```

A ce stade, on peut afficher les résultats des *mco*, du *within* et du *between* (nous ne reproduisons pas ces résultats, ils sont évidemment identiques à ce que nous avions trouvé précédemment).

```
/************************************************** calculs + affichage */

/**** mco ****/

RUN estim(ss,nt,kb,0,xx,bmco) ;                    /* retour s2,vb,ec,st,nsm */
PRINT " " / ; TITLE résultats mco ;
RUN imp("estimation mco",nomvar,bmco,ec,st,nsm,s2,vb) ;

/**** within ****/

RUN estim(ssw,nt,kw,n,xwxw,bwith) ;
sig2eps = s2 ; vbwith = vb ;                       /* retour s2,vb,ec,st,nsm */
PRINT " " / ; TITLE résultats within ;
RUN imp("estimation within",nomvarw,bwith,ec,st,nsm,sig2eps,vbwith) ;

/**** between ****/

RUN estim(ssb,n,kb,0,xbxb,bbetw) ;                 /* retour s2,vb,ec,st,nsm */
sig2b = s2 ;
PRINT " " / ; TITLE résultats between ;
RUN imp("estimation between",nomvar,bbetw,ec,st,nsm,sig2b,vb) ;
```

Au retour de chaque procédure, on mémorise dans un nom spécifique l'élément qui sera utile pour un calcul ultérieur.

Par suite, on peut calculer les éléments qui nous permettrons de calculer l'estimateur des *mcqg*. Ensuite, on débute la troisième boucle de lecture des données, qui nous permet, connaissant alors $\hat{q}$, de transformer les variables de telles sortes à calculer l'estimateur des *mcqg*. On en profite aussi pour calculer les estimations des effets fixes.

```
/************************************* 3ème boucle sur les observations */

ai = SHAPE(0,1,n) ; vai = SHAPE(0,1,n) ; j = 1 ;  /* pour les effets fixes */

sig2u = sig2b - sig2eps/t ;
theta  = 1 - SQRT((sig2eps)/(t*sig2b)) ;

xgxg = 0 ; xgyg = 0 ;

DO i = 1 TO nt ;

  RUN readata ; /* on lit les variables pour les t obs. de l'individu i */

  yg = y - theta*yb ;
  xg = x - theta*xb ;
  xgxg = xgxg + xg`*xg ; xgyg = xgyg + xg`*yg ;

/* calcul des estimations des effets fixes et de leurs écart-type */

  ai[1,j]  = yb[j] - xb[j,2:kb]*bwith  ;                    /* effets fixes */
  vai[1,j] = sig2eps/t + xb[j,2:kb]*(vbwith)*xb[j,2:kb]`; /* s des eff. fix. */
                                                  /* transpose = `= ALT GR 7 */
  i = fin ; j = j + 1 ;

END ;
```

L'instruction SHAPE(x,l,c) permet d'initialiser à x une matrice (ou un vecteur) ayant l lignes et c colonnes.

On peut alors éditer les effets fixes et leurs écarts-type :

```
                    /* calculs des estimations des effets fixes : 2 façons */

PRINT " " / ; TITLE estimation des effets fixes et de leurs écarts-type ;

/* effets fixes + cte */
sai = SQRT(vai) ; PRINT "ai et sai" ; print ai sai ; PRINT " " / ;

/* estimation de la constante et des ai hors la constante */
cte = SUM(ai)/NCOL(ai) ;  ; ai = ai - cte ;

/* édition des effets fixes hors effet constant */
PRINT "cte et les ai hors la constante " ; print cte ai ;
```

Vous noterez la fonction SAS NCOL qui renvoie le nombre de colonnes d'un vecteur. Les résultats sont :

```
estimation des effets fixes et de leurs écarts-type

                                ai et sai

            AI
     2.2858558 2.1655512 3.0418403  2.389456 2.2047714 2.1498684 2.3371097
   : 2.5923253 2.2325479 2.3759275 2.2347914 2.2167006  1.681777 3.0263426
   : 2.4025028 2.5099881 2.3454476 3.0552509

           SAI
     0.2283235 0.2128985  0.218635 0.2080863 0.2164698 0.2178844 0.2148847
   : 0.2436864 0.2395407 0.2118356 0.2141724  0.203042 0.1624636 0.3945114
   : 0.2290928 0.2356565  0.227284 0.2195996

                     cte et les ai hors la constante

                                    CTE
                                2.4026697

            AI
     -0.116814 -0.237118 0.6391706 -0.013214 -0.197898 -0.252801  -0.06556
   : 0.1896556 -0.170122 -0.026742 -0.167878 -0.185969 -0.720893 0.6236729
   : -0.000167 0.1073184 -0.057222 0.6525812
```

On voit que l'on retrouve les résultats des solutions 3 et 4 du tableau n°1 vu précédemment.

Evidemment, une quatrième et dernière boucle sur le fichier de données est nécessaire afin de calculer les résidus des *mcqg* et donc l'estimation de la matrice de variance-covariance des *mcqg*.

```
/************************************ 4ème boucle sur les observations */

bmcqg = INV(xgxg)*xgyg ;

ssg = 0 ;

DO i = 1 TO nt ;

  RUN readata ; /* on lit les variables pour les t obs. de l'individu i */

  yg = y - theta*yb ; xg = x - theta*xb ;

  ssg  = ssg + T(yg)*yg - T(bmcqg)*T(Xg)*yg ;

  i = fin ;

END ;
```

Il ne reste plus qu'à faire les calculs supplémentaires pour les mcqg puis éditer les résultats en appelant le module d'édition des résultats (imp).

```
/*********************************************calculs + affichage mcgq*/

RUN estim(ssw,nt,kw,n,xgxg,bmcqg) ;              /* retour s2,vb,ec,st,nsm */
sig2g = ssg/(nt-kb) ; vbmcqg = vb ;
PRINT " " / ; TITLE résultats mcqg ;
RUN imp("estimation mcqg",nomvar,bmcqg,ec,st,nsm,sig2g,vbmcqg) ;

PRINT "sig2eps = "  sig2eps " sig2b  = " sig2b " sig2u = "   sig2u   "
      " sig2g = "   sig2g   "theta = "   theta ; PRINT ' ' / ;
```

La vitesse d'exécution du programme dépend de la compensation entre le fait d'avoir à faire plusieurs boucles sur le fichier de données et le recours à une moindre taille des matrices $\mathbf{W}_n$, $\mathbf{B}_n$, $\mathbf{X}$ et $\mathbf{Y}$ utilisées. L'intérêt de ce programme est que sa vitesse d'exécution devrait varier faiblement en fonction du nombre d'observation.

Enfin, on termine par un certain nombre de tests[10]. Pour plus de détails sur les formules utilisées, on pourra se reporter à Blanchard et Matyas (1998).

1. Test de significativité des effets fixes (X + Fi) vs (X + NO Fi) (test de significativité jointe des indicatrices individuelles)

[10] Attention, les tests basés sur un test de Chow sont sensibles à une violation de l'hypothèse de sphéricité des perturbations (cf. Baltagi (1981) qui propose une solution, non mise en oeuvre ici pour des raisons de simplicité).

La statistique de Fisher utilisée est très classiquement $F_c = \frac{(Scr_c - SCR_{nc})}{SCR_{nc}} \times \frac{dfnum}{dfdenum}$ dans laquelle *Scrc= RSSmco, Scrnc= RSSwith, dfnum= N-1* et *dfdenum= NT-N-k*$_w$ avec où *TSS* et *RSS* représentent respectivement la somme des carrés totaux de l'endogène et la somme des carrés des résidus calculées sur les données non transformées (*mco*) et transformées (*with*), soit

$Fmcow = \frac{\frac{\mathbf{e}_{mco}'\mathbf{e}_{mco} - \mathbf{e}_w'\mathbf{e}_w}{N-1}}{\frac{\mathbf{e}_w'\mathbf{e}_w}{N(T-1)-K_w}}$. C'est donc un test d'homogénéité des constantes individuelles. Ce test a déjà été rencontré lors de l'examen de la procédure TSCSREG.

2. Test $H0 : s_u^2 = 0 \ \ vs \ \ HA : s_u^2 \neq 0$ La statistique est (déjà vue dans **panel1.sas)** :

$Fbw = \frac{\frac{T \times \mathbf{e}_b'\mathbf{e}_b}{N-K}}{\frac{\mathbf{e}_w'\mathbf{e}_w}{N(T-1)-K_w}}$ et suit un Fisher à (N - K), (N(T -1) - K$_w$) degrés de libertés

3. Test $H0 : b_1 = b_2 = \cdots = b_N = b \ \ vs \ \ HA \ : H0 \ \ est \ \ fausse$. La statistique de test est :

$Fpool = \frac{\frac{\mathbf{e}_{mco}'\mathbf{e}_{mco} - \mathbf{e}_i'\mathbf{e}_i}{K(N-1)}}{\frac{\mathbf{e}_i'\mathbf{e}_i}{N(T-K)}}$ qui suit un Fisher à $K(N-1), N(T-K)$ degrés de liberté. Dans cette expression, le terme $\mathbf{e}_i$ représente les résidus empilés des *mco* du modèle estimé individu par individu. C'est un test de stabilité des coefficients sur les individus, donc un test d'empilement ('poolability test''). Ce test peut aussi se faire en utilisant les estimations *within, mcqg* ou individu par individu.

4. Test $H0 : s_u^2 = 0 \ \ vs \ \ HA : s_u^2 \neq 0$. La statistique de test est (déjà vue dans **panel1.sas)** :

$LM2 = \frac{NT}{2(T-1)} \left[\frac{\sum_i \left[\sum_t e_{it} \right]^2}{\sum_i \sum_t e_{it}^2} - 1 \right]^2$ avec e_{it} qui représente les résidus des *mco*. Cette statistique suit un c^2 avec un degré de liberté. Ce test est aussi connu sous le nom de test de Breusch-Pagan. L'hypothèse nulle peut aussi s'écrire $(Corr(m_i + n_{it}, m_i + n_{is}) = 0$ et donc ce test est un test du modèle à erreurs composées.

5. Test $H0 : s_u^2 = 0 \ \ vs \ \ HA : s_u^2 > 0$. c'est la version unilatérale du test précédent (UMP, cf. Honda (1985)), dont la statistique de test est :

$$LM1 = \sqrt{\frac{NT}{2(T-1)}}\left[\frac{\sum_i\left[\sum_t e_{it}\right]^2}{\sum_i\sum_t e_{it}^2} - 1\right]$$ avec e_{it} qui représente les résidus des *mco* et qui suit une loi normale centrée réduite $N(0,1)$ [11].

```
/******************************************************* tests d'hypothèse */

TITLE résultat des tests ;

fmcow = ( ((ss - ssw)/(n-1)) / (ssw/(nt - n -kw)) ) ;
nsmfmcow = 2*(1-PROBF(fmcow,n-1,nt-n-kw)) ;
PRINT "test fmcow " fmcow nsmfmcow ;

fpool = ( ((ss - ssi)/(kb*(n-1))) / (ssi/(n*(t-kb))) ) ;
nsmfpool = 2*(1-PROBF(fpool,kb*(n-1),nn*(t-kb))) ;
PRINT "test de poolability " fpool nsmfpool ;

fbw = ( (t*ssb/(n-kb)) / (ssw/(nt - n -kw)) ) ;
nsmfbw = 2*(1-PROBF(fbw,n-kb,nt-n-kw)) ;
PRINT "test fbw " fbw nsmfbw ;

lm2 = (nt/(2*(t-1)))*(num/denum - 1)**2 ;
nsmlm2 = 2*(1-PROBCHI(lm2,1)) ;
PRINT "test LM2 " lm2 nsmlm2  ;

lm1 = SQRT((nt/(2*(t-1))))*(num/denum - 1) ;
nsmlm1 = 2*(1-PROBNORM(lm1)) ;
PRINT "test LM1 " lm1 nsmlm1 ;
```

Voici les résultats de ces différents tests :

[11] Nous n'avons pas suivi, pour des raisons de simplicité, la modification des tests LM1 et LM2 proposée par Moulton et Randolph (1989).

```
résultat des tests

                                    FMCOW      NSMMCOW
                         test fmcow  83.960798         0

                                       FPOOL  NSMFPOOL
                    test de poolability  129.31658         0

                                     FBW     NSMFBW
                          test fbw  86.222945         0

                                     LM2    NSMLM2
                         test LM2  1465.5523         0

                                     LM1    NSMLM1
                         test LM1  38.282532         0
```

Toutes les hypothèses nulles sont (largement) rejetées. On rejette au seuil de 5% l'homogénéité des constantes individuelles, l'empilement et la nullité de la variance des effets individuels aléatoires.

6. Le test d'Hausman

Enfin, un test d'Hausman est calculé et reporté. La statistique de test s'écrit :

$$h = c^2(K) = \left[\hat{b}_1 - \hat{b}_2\right]' \left[\hat{V}(\hat{b}_1) - \hat{V}(\hat{b}_2)\right]^{-1} \left[\hat{b}_1 - \hat{b}_2\right]$$

On peut concevoir les test suivants (cf. Baltagi, 1995, p. 69-71) :

between (1) contre *within* (2)[12]
within (1) contre *mcqg* (2)
between (1) contre *mcqg* (2)

Ces tests sont numériquement équivalents (cf. Baltagi 1995). En pratique le test *within*/*mcqg* est le plus souvent reporté. Il teste l'hypothèse fondamentale du modèle à erreurs composées à savoir la non corrélation entre les effets individuels et les variables explicatives (corrélation qui est vraisemblable avec des données individuelles, par exemple entre les capacités non observées d'un individu et sa scolarité). Dans ce cas, sous l'hypothèse alternative, l'estimateur *mcqg* est biaisé et non convergent alors que le *within* est convergent. En présence d'hétéroscédasticité et/ou d'auto correlation des perturbations, l'estimation des variances des perturbations des *mco*, du *between*, *within* et des *mcqg* sont biaisées, le test d'Hausman est alors inadéquat. Le programme ne propose pas pour l'instant la solution d'Arellano (1993).

Attention : Il est fréquent que la différence entre les deux matrices de variance-covariance ne soit pas définie positive (alors qu'en théorie elle devrait l'être, car sous H0, $\hat{b}_1$ et $\hat{b}_2$ sont convergents mais $\hat{b}_1$ n'est pas efficace et seul le premier est convergent sous HA. Dans ce cas, on calcule une inverse généralisée, mais il est de prendre alors le résultat du test avec prudence.

[12] attention dans ce cas $\hat{V}(\hat{b}_b - \hat{b}_w) = \hat{V}(\hat{b}_b) + \hat{V}(\hat{b}_w)$.

```
/********************************************************** Hausman test */

bmcqg1 = bmcqg(|2:kb|) ; vbmcqg1 = vbmcqg(|2:kb,2:kb|) ; /* dim. différente */

hstat = T(bwith - bmcqg1)*INV((vbwith-vbmcqg))*(bwith-bmcqg) ;
nsm = 1 - PROBCHI(hstat,kw) ;

TITLE résultat du test hausman ;
PRINT "test Hausman avec IML" ;
PRINT nt n t kb kw hstat nsm ; PRINT ' ' / ;
```

Voici son résultat :

```
résultat du test Hausman

                          test Hausman avec IML

          NT         N         T        KB        KW     HSTAT       NSM
         342        18        19         4         3 26.495054 7.5118E-6
```

Sachant que l'hypothèse nulle est «pas de corrélation entre les m_i et les $\mathbf{X}_i$ », (ou encore le modèle à erreurs composées est la bonne spécification), on est amené à rejeter l'hypothèse nulle et donc le modèle à erreurs composées. A noter que si l'estimation de la matrice de variance-covariance des mcqg avait été calculée avec $s_g^2 s_e^2$ au lieu de s_e^2, la statistique d'Hausman aurait été de 302.80375, néanmoins la conclusion aurait été la même. Si on avait fait le test d'Hausman avec les estimations *within* et *between* (ou de façon identique les estimations *between* et *mcqg*), on aurait fait :

```
/* between/within */

hstat=. ; nsm = . ; bbetw1 = bbetw(|2:kb|) ; vbbetw1 = vbbetw(|2:kb,2:kb|) ;
hstat = T(bbetw1 - bwith)*INV((vbbetw1+vbwith))*(bbetw1-bwith) ;
nsm = 1 - PROBCHI(hstat,kw) ;

TITLE résultat du test hausman ; PRINT "test Hausman avec IML (between/within)"
;
PRINT nt n t kb kw hstat nsm ; PRINT ' ' / ;
```

Et on aurait obtenu :

test Hausman avec IML (between/within)

NT	N	T	KB	KW	HSTAT	NSM
342	18	19	4	3	26.495054	7.5118E-6

Section 5) Les macro-commandes de E. Duguet

E. Duguet (1999) a écrit des macro-commandes (ou macro-instructions) SAS relatives à l'économétrie des panels et des variables qualitatives. Pour une présentation de ces macro-commandes, on pourra se reporter à Duguet (1999). Notons simplement que 4 macro-commandes sont utilisables pour estimer[13] des modèles du type de ceux vus précédemment : %MCOROB, %BETWEEN, %BWITHIN et %BWMCQG. Néanmoins, son programme donne en plus :

1. pour chaque estimateur, les estimations robustes des écarts-type des coefficients, cf. White (1980),
2. la décomposition de la variance totale de chaque variable (expliquée et explicative) en variance inter-individuelle (*between*) et intra-individuelle (*within*),
3. la possibilité d'avoir des régresseurs constants dans le temps pour chaque individu (pour les *mco*, le *between* et les *mcqg*), des indicatrices sectorielles par exemple,
4. la possibilité d'avoir des régresseurs constants pour un individu à chaque date (pour les *mco*, le *within* et les *mcqg*), des indicatrices temporelles par exemple,

En outre, d'autres macro-commandes sont disponibles pour estimer un modèle (statique et dynamique) sur panel par les *GMM*. Les macro-commandes utilisent à la fois les procédures et le langage *IML*. Nous en donnons une petite illustration en reproduisant partiellement ici après la macro %BETWEEN (qui n'utilise que les procédures). & est le caractère de macro-substitution.

```
%macro between(tab,y,x,i,t,transf=_between,cova=__cbet,res=__rbet,vieux=__pbet,s=);

proc sort data=&tab; by &i &t;

proc means noprint data=&tab; by &i;
var &y &x &s;
output out=&transf mean=;

title 'ESTIMATION DANS LA DIMENSION INTER-INDIVIDUELLE (BETWEEN)';
run;

%mcorob(&transf,&y,&x,cova=&cova,res=&res,vieux=&vieux);
run;

%mend between;
```

qui sera lancée par :

[13] Et donne aussi un test d'existence d'un effet aléatoire individuel (nommé Fbw dans ce qui précède) et un test d'Hausman (*within*/*between*).

```
DM 'CLEAR LOG ; CLEAR OUTPUT ; ' ;

LIBNAME in1 'c:\seminair' ;

FILENAME macro 'c:\seminair\duguet.sas' ;

%INCLUDE macro  ;

%between(in1.gasoline,lngc,lnyn lnpgpp lncarn,ident,year) ;

RUN ;
```

Avec l'autorisation de l'auteur (eduguet@univ-paris1.fr), j'ai mis ces macros sur le site Web de l'Erudite (sur ma page personnelle, rubrique programmes et données) dans un fichier nommé **macroduguet.zip** (50 Ko, format ZIP). Ces macros peuvent aussi être obtenues sur son site web http://panoramix.univ-paris1.fr/EUREQUA/annuaire/duguet/duguet.htm. Ce fichier contient toutes les macro-commandes écrites par E. Duguet. L'utilisation des macros est très simple, leur adaptation à des problèmes spécifiques (panel incomplet…) l'est un peu moins. Un exemple d'utilisation de ces macros est fourni dans *panel6.sas* qui appelle la macro *duguet.sas* (celle concernant les panels, renommée pour ne pas la confondre avec l'original – panel2.cpu).

Voici les principaux résultats obtenus (certaines analyses de statistiques descriptives, les estimations between et within[14] ne sont pas reproduites).

```
ESTIMATIONS BETWEEN ROBUSTES

                                                    T de
                        Coefficient    Ecart-Type  Student       Seuil de
            Régresseur      MCO          Robuste   Robuste   Significativité

             INTERCEP       2.54163       0.44561    5.7037    .000000011726
             LNYN           0.96758       0.17901    5.4050    .000000064796
             LNPGPP        -0.96355       0.14828    6.4982    .000000000081
             LNCARN        -0.79530       0.07234   10.9945    .000000000000
```

```
ESTIMATIONS WITHIN ROBUSTES

                                                    T de
                        Coefficient    Ecart-Type  Student       Seuil de
            Régresseur      MCO          Robuste   Robuste   Significativité

              LNYN          0.66225       0.15328   4.32054       .0000156
              LNPGPP       -0.32170       0.12228   2.63097       .0085142
              LNCARN       -0.64048       0.09665   6.62658       .0000000
```

[14] On notera pour les estimations *within* le même problème pour le calcul des écarts-type estimés des coefficients estimés (*NT-N-kw*) que celui rencontré dans le programme **panel1.sas**.

Decompositions de la Variance Totale

RES	% Between * T	% Within	Variance Totale
LNGC	99.02	0.98	5.0901655
LNYN	99.28	0.72	7.108816
LNPGPP	99.82	0.18	8.917084
LNCARN	98.7	1.3	23.897179

```
                        Analyse de la Variance

                                                N
                      Nombre d'individus =      18

                                                T
                      Nombre de Periodes =      19

                                               NT
                    Nombre d'Observations =     342

                                      MCB                                DLB
  Moyenne des Carres Between / N = 0.0300894 , Degres de Liberte =       14

                                      MCW                                DLW
Moyenne des Carres Within / (N*T) = 0.0080014 , Degres de Liberte =      321

                                                S2B                       SB
Variance Regression Between (sur N obs.) = 0.0386864 , Ecart-type = 0.1966886

                                                S2W                       SW
Variance Regression Within (sur N*T obs.) = 0.0085249 , Ecart-type = 0.0923303

                                      S2E                    SE
       Variance du bruit blanc = 0.0089985 , Ecart-type = 0.0948604

                                       S2A                    SA
     Variance Effet Individuel   = 0.0382128 , Ecart-type = 0.1954809
```

```
      Test d'existence d'un effet aléatoire individuel (Dormont, pp. 70-71) :

                           DLBC         DLW        FISHER
                  F(        252 ,        321 ) = 3.5625811

                                                  PROBA
                   Seuil de Significativite =         0

       Statistique de Hausman (H0 -> Modele a erreurs composees OK) :
              Test de la difference entre within et between
          (sans constante ni indicatrices, ecart types robustes)

                                      K       HAUSMAN
                    Khi-deux (        3 ) = 12.776769

                                                  PROBA
                  Seuil de Significativite = 0.0051451
```

```
                          Analyse de la Variance                          9

                                          THETA2
                          THETA CARRE =  0.0122422

                                           THETA
                            THETA = 0.1106443
```

- Attention, il s'agit bien de $\hat{q} = \sqrt{\frac{\hat{s}_e^2}{T\hat{s}_u^2 + \hat{s}_e^2}}$ et non de $\hat{q} = 1 - \sqrt{\frac{\hat{s}_e^2}{T\hat{s}_u^2 + \hat{s}_e^2}}$ .Du fait de la correction par *T* et *T-1* lors du between, l'estimation de $1-q$ est un peu différente des résultats obtenus précédemment. En effet, E. Duguet utilise, incorrectement semble-t-il, la correctiopn par *T* et *T-1* en calculant $\hat{s}_e^2 = \left(\frac{e_w' e_w}{NT - N - k_w} \right) \times \frac{T}{T-1}$ au lieu de $\hat{s}_e^2 = \left(\frac{e_w' e_w}{NT - N - k_w} \right)$ (cf. son document de travail pour une justification).

- De plus, le résultat du test d'Hausman est différent (mais cela ne change pas la conclusion du test) car les estimations robustes sont utilisées.

- Enfin, il semble y avoir un problème de programmation dans le calcul du nombre de degrés de libertés du numérateur du test d'existence d'un effet aléatoire individuel.

```
                  CORRELATIONS DES QUASI DIFFERENCES (MCQG)

                       ESTIMATION DU MODELE TRANSFORME (MCQG)
                Il faut diviser le coefficient du terme constant par 'theta'

Model: MODEL1
Dependent Variable: LNGC

                                Analysis of Variance

                                   Sum of         Mean
            Source          DF      Squares       Square      F Value       Prob>F
            Model            3     15.01200      5.00400      546.016       0.0001
            Error          338      3.09763      0.00916
            C Total        341     18.10962

                Root MSE       0.09573     R-square       0.8290
                Dep Mean       0.47535     Adj R-sq       0.8274
                C.V.          20.13901

                                Parameter Estimates

                           Parameter      Standard    T for H0:
          Variable  DF      Estimate         Error   Parameter=0    Prob > |T|
          INTERCEP   1      0.219947    0.02022899        10.873        0.0001
          LNYN       1      0.553237    0.05875699         9.416        0.0001
          LNPGPP     1     -0.423865    0.03989520       -10.624        0.0001
          LNCARN     1     -0.606427    0.02541826       -23.858        0.0001
```

ESTIMATION DU MODELE TRANSFORME (MCQG)
Ecarts-types robuste a l'autocorrelation et a l'heteroscedasticite

Régresseur	Coefficient MCO	Ecart-Type Robuste	T de Student Robuste	Seuil de Significativité
INTERCEP	0.21995	0.05604	3.92483	.00008679
LNYN	0.55324	0.11752	4.70778	.00000250
LNPGPP	-0.42387	0.11677	3.62992	.00028351
LNCARN	-0.60643	0.08776	6.90979	.00000000

Conclusion

On a appliqué les programmes précédents à un panel simulé aléatoirement avec N = 1000 et T = 8, pour 3 exogènes. Les résultats des estimations sont identiques (sauf pour TSCSREG qui utilise la méthode n°3 d'Henderson pour l'estimation de q). Seules les performances diffèrent. On reporte aussi les performances observées avec les données de Baltagi.

Tableau n°4

		N=1000, T = 8	N=18, T = 19
Programme		Temps mis (en secondes)	Temps mis (en secondes)
panel1.sas	PROC REG	4,62	1
panel2.sas	PROC GLM et PROC TSCSREG	44,29 [15]	0,27
panel3.sas	IML	1,43	0,27
panel4.sas	IML + Kronecker	3043,91 (Plus de 50 minutes)	0,33
panel5.sas	IML + Kronecker+boucle sur lers données	2,86	0,33
Macros Duguet[16]	PROC REG+IML	6.92	4,62

Machine utilisée : Pentium III 600Mhz, 128Mo de RAM, 13 Go de disque dur, Windows 98.

Note : voici le programme utilisé pour la génération aléatoire des données :

```
LIBNAME in 'c:\pierre99\semipanel\simul' ;            /* A modifier */

DATA temp ;
seed = 26081966 ;
DO ident = 1 TO 1000 ;
   ui = RANNOR(seed) ;
   DO an = 89 TO 96 ;
      x1 = 1 - RANUNI(seed) ;
      x2 = -1 + RANNOR(seed)/3 + ui ;
      x3 = RANUNI(seed) ;
      y = 4 + 0.5*x1 -0.3*x2 + 4*x3 + ui + RANNOR(seed) ;
      OUTPUT ;
  END ;
END ;

DATA in.simul (KEEP = ident an y x1 x2 x3); SET temp ;
```

[15] dont 43,6 s. pour le modèle à erreurs composés (dû probablement à l'utilisation de la méthode d'Henderson). Le modèle à effet fixes est estimé par PROC GLM (0,69 s.).

[16] Le temps d'exécution de ces macro-commandes n'est pas strictement comparable à celui des autres programmes car elles effectuent des opérations supplémentaires (estimations robustes, décomposition de la variance…).

Si vous avez :

- peu de données,
- besoin uniquement des estimateurs *within* et *mcqg* et du test d'Hausman

➔ PROC TSCSREG et PROC GLM (cf. par exemple panel2.sas)

ou TSP, LIMDEP, STATA.

Si vous avez :

- des données très nombreuses (N est grand)
- des panels incomplets
- des tests particuliers à mettre en œuvre
- besoin des *GMM*[17] ou des *VI* (hétéroscédasticité, auto corrélation, endogénéité des régresseurs)

➔ IML (cf. par exemple panel3.sas ou panel5.sas ou les macro-commande de E. Duguet)

Si vous avez :

- des modèles non linéaires sur données de panel (logit, probit…)

➔ SAS (IML, pas les PROC) mais SAS est sérieusement concurrencé par *GAUSS, ADM MODEL BUILDER*…

[17] La procédure SAS estimant des modèles par les *VI* ou par les *GMM* (PROC SYSLIN) n'est pas compatible avec la structure panel des données.

Références bibliographiques

Amemiya T. (1971), The Estimation of the Variance in a Variance-Components Model, *International Economic Review*, February, 12, p. 1-13.

Arellano M. et S. Bond (1991), Some Tests of Specification for Panel Data : Monte Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, 58, p. 277-297.

Baltagi B. (1995), *The Econometrics of Panel Data*, Wiley.

Baltagi B. (1998), *Econometrics*, Wiley.

Baltagi B. et J. M. Griffin (1983), "Gasoline Demand in the OECD : An Application of Pooling and Testing Procedures", *European Economic Review*, 22, p. 117-137.

Blanchard P. et L. Matyas (1996), Robustness of Tests for Error Components Models to Non-Normality, *Economics Letters*, 51, p. 161-167.

Blanchard P. et L. Matyas (1998), Misspecified Heterogeneity in Panel Data Models, *Statistical Papers*, 39, p. 1-27.

Blundell R. W. et S. Bond (1998), Initial Conditions and Moment Restrictions in Dynamic Panel Data Models, *Journal of Econometrics*, 68, p. 29-51.

Breusch T. S. (1987), Maximum Likelihood Estimation of Random Effects Models, *Journal of Econometrics*, 36, p. 383-389.

Breusch T. S. et A. R. Pagan (1980), The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics, *Review of Economic Studies*, 47, jan., p. 239-253.

Brillet J. L. (1989), Econometric Modelling on Microcomputers: A Review of Major Software Packages, *Journal of Applied Econometrics*, Vol. 4, p 73-92.

Chamberlain G., (1982), Multivariate Regression Models for Panel Data, *Journal of Econometrics*, 18, p. 5-46.

Dormont B. (1989), *Introduction à l'Econométrie des Données de Panel*, CNRS.

Duguet E. (1999), Macro-commandes pour l'Econométrie des panels et des variables qualitatives, document de travail INSEE, n°G9914.

Fuller W. A. et G. E. Battese (1974), Estimation of Linear Models with Cross-Error Structure, *Journal of Econometrics*, 2, p. 67-78.

Goldstein R., J. Anderson, A. Ash, B. Craig, D. Harrington and M. Pagano, (1989) Survival Analysis Software on MS/PC-DOS Computers, *Journal of Applied Econometrics*.

Greene W. H.,(1993), *Econometric Analysis*, Macmillan Publishing Company, New-York.

Greene W. H., (1997), *Econometric Analysis*, 3ème éd., Macmillan Publishing Company, New-York.

Hausman J. A. (1978), Specification Tests in Econometrics, *Econometrica*, 46, nov., p. 1251-1271.

Henderson C. R., (1953) Estimation of Variance and Covariance Components, *Biometrics*, 9, june, p. 226-252.

Hsiao C. (1986), *Analysis of Panel Data*, Econometric Society Monograhs, Cambridge University Press.

Matyas L. et P. Sevestre (1996), *The Econometrics of Panel Data*, Kluwer Academic Publishers.

Korosi G., Matyas L. et Szekeley I. (1991), *Practical Econometrics*, Gower

Lanjouw P. (1992), The SAS System Version 6, *The Economic Journal*, 102, sept., p. 1302-1313.

MacKie-Mason J.K. (1992), Econometric Software : A User's View, *Journal of Economic Perpectives*, vol. 6, N°4, Fall, p. 165-187.

Maddala G. S. et T. D. Mount (1973), A Comparative Study of Alternative Estimators for Variance Components Models Used in Econometric Applications, *Journal of the Statistical Association*, june, vol. 68, n°42, p. 324-328.

Nerlove M. (1971), Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections, *Econometrica*, 39, march, p. 359-382.

Swamy P. A. V. B. et S. S. Arora, (1972), The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models, *Econometrica*, 40, p. 261-275.

Wallace T. D. et A. Hussain (1969), The Use of Error Components Models in Combining Cross Section with Time Series Data, *Econometrica*, 37, january, p. 55-72.