Econométrie linéaire des panels : une introduction[1]

Thierry MAGNAC

Université des Sciences Sociales, Toulouse

Introduction

Les enquêtes de panel fournissent des observations pour les mêmes individus, ménages, entreprises ou pays, de façon répétée au cours du temps. Elles sont de plus en plus courantes aux Etats-Unis et en Europe mais aussi en France. Certaines enquêtes suivront les achats des mêmes consommateurs au cours d'un certain laps de temps (Secodip, Nielsen). D'autres donneront les revenus et statuts vis à vis de l'emploi des individus et des ménages au cours du temps (Panel Européen, Enquête Emploi, DADS); d'autres encore permettront de reconstituer les investissements, la main d'oeuvre, les salaires des entreprises (ACEMO, BRN). On peut aussi construire des données de panel en utilisant les observations extraites des comptabilités nationales de différents pays au cours des vingt ou trente dernières années (Penn World Tables). On peut enfin utiliser des enquêtes en coupe transversale mais qui donnent des informations rétrospectives sur les histoires éducatives, professionnelles, familiales des individus et des ménages (Formation Qualification Professionnelle, Situations Défavorisées) pour reconstituer des données à double indice, individuel et temporel.

Les gains à utiliser ces données plutôt que des données de séries temporelles sont clairs. On dispose de plus de variabilité, d'une information plus riche et les estimateurs des coefficients d'intérêt sont plus précis. Les variables sont d'ailleurs beaucoup moins souvent colinéaires et les coefficients sont donc beaucoup mieux identifiés. Plus intéressante est la comparaison des données de panel aux données obtenues en une seule coupe transversale. On voit aux données de panel trois avantages principaux.

1) *Elles permettent d'identifier des effets qui ne le sont pas en coupe transversale.* Par exemple, considérons la mesure des rendements monétaires de l'expérience professionnelle ou pour simplifier de l'âge, sur le marché du travail. Pour les mesurer en coupe transversale, on régressera les taux de salaire ou les logarithmes de ces taux, sur l'âge et un terme quadratique en âge pour saisir la décroissance de ces rendements au cours du temps. On utilisera aussi sans doute dans la régression un certain nombre d'autres variables comme l'éducation. De telles estimations linéaires montrent que les rendements de l'expérience sont positifs jusqu'à 45 ou 50 ans puis sont négatifs ensuite. Mais, à y réfléchir à deux fois, une telle mesure des rendements mélange deux effets: un effet de l'âge et un effet de génération.

[1] Cette note doit beaucoup au cours enseigné en troisième année à l'ENSAE, aux commentaires de Stefan Lollivier et Daniel Verger et pour l'application empirique décrite dans la section 6, à la mise en oeuvre critique de Sébastien Roux. Je reste responsable de toutes les erreurs et omissions qui subsistent.

En comparant en coupe transversale deux individus dont les âges sont différents, on compare aussi deux générations. Or celles-ci n'ont pas la même histoire éducative ou professionnelle et leurs situations vis à vis du marché du travail diffèrent. Dès que l'on dispose de données de panel, on peut distinguer les effets de l'âge et les effets de génération puisque pour la même génération, on dispose de plusieurs observations relatives à des âges différents. En estimant alors un tel modèle linéaire, on s'aperçoit d'ailleurs que les rendements de l'âge sont positifs jusqu'à un âge beaucoup plus élevé (55 à 60 ans). L'effet négatif de l'expérience professionnelle aux âges élevés semblent donc cacher des effets de génération et non des effets de l'âge (Lollivier et Payen, 1990).

2) *Elles permettent de contrôler la présence d'hétérogénéité inobservable.* Quand on veut savoir si le marché du travail est segmenté entre différents secteurs par exemple, on pourra estimer en coupe transversale des équations de salaire comme ci-dessus. On incluera à côté des variables précédentes, des variables indicatrices des secteurs. L'interprétation des coefficients de ces variables est pourtant délicate puisqu'ils peuvent refléter une différence dans les salaires offerts par les secteurs mais aussi des différences inobservées, de formation, de qualification, etc, entre les travailleurs de secteurs différents. Les données de panel permettent de contrôler cette hétérogénéité entre travailleurs, inobservable et fixe au cours du temps qui est appelée effet individuel. Comme nous le verrons, les modèles linéaires de panel font reposer l'identification de différences sectorielles de salaires sur les différences de salaires au cours du temps pour des travailleurs qui **changent** de secteur. Ceux-ci ont en effet les mêmes effets individuels inobservables mais les secteurs leur font des offres de salaire différentes, éventuellement[2]. On notera que le contraste entre le point 1 développé plus haut et celui-ci vient de l'observabilité de la variable de génération et de l'inobservabilité des effets individuels.

3) *Elles permettent de formuler des modèles dynamiques.* Les coupes transversales ne sont que des clichés uniques d'un phénomène dynamique alors que les données de panel sont des clichés répétés et permettent donc de saisir un mouvement même si le grand pas de temps entre deux clichés peut donner l'impression de mouvements heurtés. C'est souvent la dynamique de nombreux phénomènes économiques qui est intéressante. Par exemple, les effectifs d'une entreprise à un certain moment dépendent sans doute fortement du nombre d'employés à la période précédente. Ou alors il faudrait s'imaginer que toutes les entreprises ne vivent qu'une période, ou licencient tous leurs travailleurs en fin de période et réembauchent en début de période suivante pour faire table rase du passé. Ces marchés ``spot'' du travail deviennent de plus en plus rares dans nos économies. Il y a en effet des coûts d'ajustement, de licenciement, d'embauche, de création ou de fermeture d'une entreprise. Le modèle linéaire est alors donné par une équation qui fait dépendre le nombre d'employés à la période t du nombre d'employés à la période $t-1$. Ce modèle ne peut pas s'estimer en utilisant des données en coupe transversale et il faut nécessairement disposer de données de panel.

Du point de vue de l'analyste, l'avantage des données de panel semble écrasant. Il ne faut pourtant pas se cacher que les problèmes de collecte de données et donc le coût dans tous les sens du terme, sont plus importants pour les données de panel que pour les données en coupe transversale. Mais au vu des développements de la recherche empirique, les bénéfices l'emportent largement sur les coûts (Heckman et Robb, 1985). Le problème principal de la collecte de données de panel est sans conteste celui de l'attrition des enquêtés, c'est à dire de leur sortie du panel, ou refus de répondre, au fur et à mesure que le temps passe. Ceci n'est gênant que si ces individus se sélectionnent ou sont sélectionnés de manière endogène. Par exemple, l'enquête Emploi est une enquête de logements renouvelés par tiers tous les ans. Si un ménage déménage durant cette période, il est remplacé dans l'enquête par les nouveaux occupants du logement. On concevra donc qu'il est impossible dans ce cas de mener des études sur des phénomènes directement ou

[2]On peut aussi évoquer un second problème qui est celui de l'endogénéité de ces changements de secteur. En effet, ceux-ci sont plus probables pour les travailleurs qui y ont intérêt en terme d'avantages monétaires. Pour traiter de manière générale, l'endogénéité, on se référera à Robin (2000) et dans ce cas particulier, aux modèles dynamiques analysés dans la section 3.

indirectement lié aux migrations des familles car la présence des ménages dans l'échantillon, la sélection, n'est pas indépendante du phénomène que l'on veut étudier. Mais on concevra aussi qu'étudier les migrations dans une enquête en coupe transversale est tout aussi impossible. Il faut donc soit des informations rétrospectives (Formation Qualification Professionnelle par exemple) même si celles-ci posent des problèmes relatifs à la mémoire des individus ou de vraies données de panel (Panel Européen). Il existe cependant des formes intermédiaires de bases de données constituées à partir d'enquêtes en coupe transversale répétée où on peut donc suivre au cours du temps des cohortes d'individus. L'attrition n'existe donc pas dans ces enquêtes dites de pseudo-panels puisqu'on rééchantillonne systématiquement les individus dans les mêmes cohortes. Ces données peuvent être adaptées au cas où on veut surmonter des problèmes d'identification comme ceux qui étaient évoqués dans le point 1 précédent. Elles sont néanmoins beaucoup moins riches en termes d'information et sont donc moins bien faites que les données de panel pour contrôler la présence d'hétérogénéité inobservable (point 2) et très délicates à utiliser dans le cas de modèles dynamiques. Néanmoins, les techniques utilisées dans les modèles linéaires de panel peuvent s'adapter aux pseudo-panels ce qui n'est pas du tout le cas dans des modèles non-linéaires.

Cette note présente donc un état des lieux des méthodes d'estimation de modèles linéaires sur des données à double indice, individuel et temporel. On peut d'abord rester très proche des méthodes économétriques utilisées avec des données en coupe transversale en faisant des hypothèses fortes sur l'exogénéité des variables. Cependant, quand on dispose de plusieurs observations par individu, les perturbations affectant les observations de la variable dépendante et correspondant au même individu, sont sans doute corrélées entre elles. On a donc un problème d'hétéroscédasticité mais les estimateurs des moindres carrés ordinaires (MCO) restent convergents. L'hétéroscédasticité rend incorrecte pourtant l'estimation des écart-types des estimateurs. La correction de ceux-ci peut se faire par des méthodes dérivées du modèle linéaire général et que l'on appelle méthodes à erreurs composées.

On peut aussi profiter de la multiplicité d'observations individuelles pour essayer de contrôler les facteurs individuels non observés et omis dans l'équation à estimer comme dans l'exemple évoqué plus haut, de l'estimation des différences sectorielles de salaires. Si ces facteurs omis sont corrélés avec les variables explicatives, les estimateurs habituels de moindres carrés sont biaisés. L'interprétation des relations économiques comme des relations causales conduit à privilégier ce modèle au précédent. En effet, la mesure de l'impact de la cause sur un effet repose sur le contrôle de toutes les autres causes possibles. C'est le principe de l'analyse ``toutes choses égales par ailleurs''. Contrôler les facteurs individuels non observés et omis, devrait permettre de s'approcher de cette condition idéale d'expérimentation pour mesurer les effets des autres facteurs observables que sont les variables explicatives. Ces méthodes visant à contrôler ces facteurs sont dites à effets fixes mais on préfère maintenant, en comprenant mieux la structure de ces modèles, les appeler modèles à hétérogénéité corrélée.

Ces deux types de méthodes valent pour les équations où on s'intéresse à une relation entre une variable dépendante et des variables explicatives. On appelle ce type de modèles, les modèles statiques. On peut aussi profiter de la disponibilité d'observations au cours du temps pour formuler des modèles dynamiques comme ceux qui ont été évoqués dans le point 3 ci-dessus. La variable dépendante dépend alors des variables dépendantes retardées et des variables explicatives, comme dans le cas prototype d'un modèle autorégressif d'ordre 1, justifié par le modèle économique, décrit plus haut, de détermination des effectifs d'une entreprise. Dans ce modèle, on peut faire les distinctions entre influences de court et de long terme, comme dans une analyse de séries temporelles. Les méthodes d'estimation adaptées sont différentes des méthodes utilisées dans les modèles statiques et on rencontrera des problèmes liés à l'endogénéité des variables qui se traitent par des méthodes à variables instrumentales.

Pour simplifier la présentation de ces méthodes, on aura fait l'hypothèse que les données sont cylindrées. Tous les individus sont observés à toutes les périodes et il n'y a pas d'attrition, c'est-à-dire de sorties de l'échantillon. Dans les enquêtes, cylindrer des données peut être très coûteux en terme de précision puisqu'on omet l'information contenue dans les trajectoires incomplètes et peut être très coûteux en termes de biais si l'attrition, ou sélection, est endogène. On ne traitera pas ici

du problème de sélection qui réclame un traitement préalable des données qualitatives en panel. On ne reviendra sur la question que sous l'hypothèse d'attrition, ou de sélection, exogène. On adapte facilement les méthodes qui ont été développées à des données non cylindrées même si leurs propriétés en termes de précision peuvent être perdues. Finalement, nous avons vu que, dans certains cas, les données de panel ne sont pas nécessaires et on peut utiliser des enquêtes transversales répétées au cours du temps. On appelle ce type de données des données de pseudo-panels puisque sous certaines hypothèses, on peut adapter les méthodes de données de panel à ce cas de figure.

La première section concerne le lecteur qui voudrait comprendre l'apport des données de panel et les méthodes de moindres carrés adaptées à de telles données. On y présente la méthode à effets fixes puis le modèle à effets individuels aléatoires. On étudie ensuite les modèles à erreurs aléatoires non corrélées. La seconde section intéressera le lecteur averti et en particulier celui qui a de bonnes raisons de croire que ses données ne peuvent être comprises que dans un cadre dynamique. On y introduit les méthodes utilisées dans les modèles dynamiques. On traite du cylindrage des données dans la troisième section et des méthodes adaptées aux coupes transversales répétées dans la section suivante. L'application empirique concernant l'estimation de fonctions de salaire pour les hommes en France dans les années 90, est présentée dans la dernière section.

Guide de lecture

Nous conseillons au lecteur de commencer par la section 2 jusqu'au point 2.3 (le modèle à effets aléatoires non corrélés) non compris. Cette partie théorique présente en effet les fondements des modèles de panel et les modèles les plus simples. Elle constitue la base théorique indispensable à une bonne compréhension des méthodes utilisées. Le lecteur pourra alors trouver une application concrète de ces méthodes et de la construction de ces modèles dans la section 6 (jusqu'au point 6.2.1 inclus).

D'autres modèles, construits sur l'absence de corrélation entre les effets individuels et certaines variables explicatives sont exposés à la fin de la section 2 à partir du point 2.3 et leurs applications empiriques sont données aux points 6.2.2 et 6.2.3 de la section 6.

Une fois la section 2 et ses applications bien assimilées, la lecture de la section 4 devient indispensable pour comprendre le rôle du cylindrage des données et les problèmes liés à la sélection des individus. Un exemple de différence des résultats obtenus en cylindrant ou en ne cylindrant pas les données est rapporté à la fin de la section 6.

Les autres sections contiennent des extensions qui peuvent être sautées en première lecture. La section 3, plus difficile, s'adresse à un lecteur averti et la section 5 concerne plus particulièrement les personnes travaillant sur des pseudo-panels.

1. Effets fixes et effets aléatoires

Les données de panel ou données longitudinales sont caractérisées par des données à double indice: un indice individuel et un indice temporel. On observe ainsi une variable dépendante y_{it} et des variables explicatives, x_{it}, pour des unités statistiques (individus, ménages, entreprises ou pays) indicées par $i = 1, \ldots, N$ au cours de périodes indicées par $t = 1, \ldots, T$. On supposera jusqu'à la section 4 **qu'il n'y a aucune valeur manquante.** Pour tout individu et toute période, les observations de y_{it} et de x_{it} sont disponibles. On dit dans ce cas que le panel est **cylindré**. On prendra pour point de départ le modèle linéaire usuel :

$$y_{it} = x_{it}\beta + \varepsilon_{it}$$

où les paramètres d'intérêt que nous cherchons à estimer sont les paramètres β. Le modèle est donc supposé linéaire en β. Il n'est pas nécessairement linéaire en les variables explicatives puisque des termes quadratiques de variables ou des indicateurs d'intervalles de variation pour des variables peuvent librement apparaître dans la liste des variables explicatives, x_{it}. Nous ne relâcherons jamais l'hypothèse de linéarité de l'équation en β.

Les données de panel offrent une information plus riche que les données en coupe puisqu'on dispose de données répétées pour le même individu. En suivant les principes de l'analyse de la variance à deux facteurs, ici individuel et temporel, on peut alors chercher à décomposer plus avant le terme d'erreur en écrivant:

$$y_{it} = x_{it}\beta + \alpha_i + \delta_t + u_{it}$$

où α_i est dit effet individuel et où δ_t est dit effet temporel. Les effets temporels saisissent les chocs agrégés à la date t et sont **communs à tous les individus**. Les effets individuels regroupent toutes les variables individuelles **fixes au cours du temps** et qui ne sont pas observées par l'économètre. C'est ici que se voit le plus clairement un avantage des données de panel par rapport aux données en coupe transversale (point 2 de l'introduction). On peut tenir compte dans l'estimation, de la présence d'hétérogénéité individuelle inobservable constante au cours du temps.

On procède d'abord à une simplification sans conséquence de l'écriture de ce modèle. Les enquêtes les plus usuelles de panel se caractérisent par un grand nombre d'individus N et un petit nombre de périodes T qui est compris entre deux et dix, dans la majorité des cas. Les effets temporels δ_t peuvent alors être considérés comme les paramètres des variables indicatrices de chaque période t et peuvent donc, sans perte de généralité, être inclus dans la liste des paramètres β. Ceci ne fait croître cette liste que de façon modérée. On a ainsi:

$$y_{it} = x_{it}\beta + \alpha_i + u_{it}.$$

où les indicatrices temporelles des périodes font maintenant partie des variables explicatives, x_{it}. Il est d'ailleurs toujours recommandé d'inclure des variables indicatrices temporelles dans ces modèles sous peine de voir tous les coefficients estimés être biaisés par l'oubli de telles variables puisque les chocs agrégés sont en général très significatifs[3].

Ces paramètres ont un intérêt en eux-mêmes d'ailleurs puisqu'ils représentent les effets agrégés et peuvent donc être interprétés économiquement. Par exemple, dans des fonctions de salaire, ils refléteront la croissance annuelle des salaires, toutes choses égales par ailleurs. Pour les effets

[3]Des problèmes d'identification peuvent parfois se poser. Par exemple, l'année courante étant la somme de l'année de naissance et de l'âge en années, il est toujours impossible d'identifier, à la fois, les effets de l'âge, de la période et de la génération (année de naissance). Par exemple, si l'âge et la génération sont des variables explicatives, un des coefficients des indicatrices temporelles ne pourra pas être identifié.

individuels, la situation est différente. Ils sont en effet très nombreux si la dimension individuelle N est grande et ils sont, en général, peu interprétables et peu intéressants. Ce ne sont pas les effets du ménage i ou de l'entreprise i qui nous intéressent en eux-mêmes mais les effets des caractéristiques individuelles sur la variable dépendante et donc le paramètre β. L'utilisation de données de pays pour l'estimation de modèles de croissance peut fournir une exception à la règle puisque les effets pays sont plus intéressants à interpréter.

Deux modélisations pour les effets individuels sont alors possibles. Dans le premier modèle dit à effets fixes, on considèrera que les effets individuels α_i sont des paramètres. On dit que ce sont des paramètres de nuisance par opposition aux paramètres d'intérêt, β, puisque comme nous le verrons, ils nuisent à l'estimation directe des β. Dans le deuxième modèle, on considérera que ces effets individuels sont aléatoires. A première vue, ces deux spécifications semblent différentes mais nous allons voir dans cette section qu'il est facile de les réconcilier dans un tout cohérent.

1.1. Le modèle à effets fixes

On considère le modèle (EF):

$$y_{it} = x_{it}\beta + \alpha_i + u_{it}$$

où β et α_i sont des paramètres. On supposera que les conditions d'exogénéité des variables explicatives sont satisfaites. En notant $x_i = (x_{i1},.,x_{iT})$ le vecteur des variables explicatives à toute période, ces conditions se traduisent d'abord par l'absence de corrélation entre perturbations et variables explicatives à toute période :

$$H_1 \ : \ Ex_i^{'}u_{it} = 0, Eu_{it} = 0$$

et les variables explicatives sont dites fortement exogènes. La deuxième partie de H_1 est la condition habituelle de centrage des perturbations et permet d'identifier α_i. La première partie de cette condition précise que les perturbations individuelles-temporelles et qui affectent la variable dépendante ne sont pas corrélées avec les variables explicatives. Il n'y a donc aucun effet de feedback des variables dépendantes vers les variables explicatives. Cette hypothèse est peu discutable, dans le cas de l'âge par exemple dans une équation de salaire, mais elle est plus discutable dans le cas de l'ancienneté dans l'entreprise. Des chocs très négatifs sur les salaires peuvent donner lieu à un départ de l'entreprise et ceci affecte la variable d'ancienneté. Il y aurait dans ce cas effet de feedback et l'hypothèse H_1 ne serait pas vérifiée.

Ensuite, les perturbations sont supposées indépendantes entre individus. Par simplicité, on suppose aussi qu'elles sont homoscédastiques. Elles ne sont donc pas corrélées au cours du temps pour le même individu et leurs moments du second ordre sont constants :

$$H_2^a \ : \ u_i \text{ et } u_j \text{ sont indépendants}$$
$$H_2^b \ : \ E(u_i u_i^{'} \mid x_i) = \sigma_u^2 . I_T$$

Comme on étudie des données de panel de ménages ou d'entreprises, on ne remettra jamais en cause l'hypothèse H_2^a dans cette note. L'hypothèse d'homoscédasticité H_2^b est faite pour des raisons de simplicité et nous ne discuterons son abandon qu'en fin de section puisque cela ne pose pas de problèmes très difficiles à résoudre.[4]

[4]On notera que l'on a fait des hypothèses d'homoscédasticité conditionnelle alors que les conditions H_1 ne portent que sur des absences de corrélations. On pourrait renforcer H_1 en supposant que les moyennes conditionnelles sont nulles:

Sous les hypothèses H_1 à H_2, le modèle (EF) semble pouvoir s'estimer par les moindres carrés ordinaires puisque c'est un modèle linéaire simple. On montrera que les propriétés asymptotiques usuelles, quand N tend vers l'infini, sont vérifiées pour l'estimateur des moindres carrés ordinaires (MCO) de β. Il y a néanmoins deux difficultés que nous étudions maintenant. L'une est relative à l'identification des paramètres. L'autre est relative à l'estimation directe de β sans estimer les paramètres de nuisance α_i dont les estimateurs MCO ne sont pas convergents quand N tend vers l'infini.

1.2. Identification

Le problème d'identification le plus notable est celui du paramètre d'une variable constante. Si l'une des variables explicatives est la constante, on peut écrire (EF) comme:

$$y_{it} = x_{it}^{(1)}\beta_1 + \beta_0 + \alpha_i + u_{it} = x_{it}^{(1)}\beta_1 + \alpha_i^{(0)} + u_{it}$$

où $x_{it}^{(1)}$ sont les autres variables explicatives, β_0 est le coefficient de la constante et où $\alpha_i^{(0)} = \alpha_i + \beta_0$ est un nouvel effet individuel. Si $\alpha_i^{(0)}$ est connu, il y a une infinité de valeurs possibles pour β_0 et α_i et ces paramètres ne sont donc pas identifiables.

Le problème d'identification est plus profond puisque ce raisonnement s'applique à tout coefficient d'une variable individuelle constante au cours du temps, l'éducation par exemple dans une équation de salaire. Les variables individuelles constantes au cours du temps sont notées $x_i^{(2)}$ et les autres variables sont notées $x_{it}^{(1)}$. Le modèle (EF) s'écrit donc :

$$y_{it} = x_{it}^{(1)}\beta_1 + x_i^{(2)}\beta_2 + \alpha_i + u_{it} = x_{it}^{(1)}\beta + \alpha_i^{(2)} + u_{it}$$

où $\alpha_i^{(2)} = \alpha_i + x_i^{(2)}\beta$ est le nouvel effet individuel. Si $\alpha_i^{(2)}$ est connu, il y a une infinité de valeurs possibles pour β_2 et α_i et ces paramètres ne sont donc pas identifiables. Cela va de soi puisque α_i représente l'hétérogénéité individuelle constante au cours du temps. Que cette hétérogénéité soit observable ou non ne doit pas affecter la modélisation. Il faudra des hypothèses additionnelles, par exemple d'absence de corrélation entre α_i et $x_i^{(2)}$, pour estimer β_2 mais il faudra alors considérer que α_i est un effet individuel aléatoire (voir *infra*).

Pour l'instant, nous ne nous intéresserons donc qu'à l'estimation du modèle (EF) où les variables individuelles x_{it} ne sont pas constantes au cours du temps. Dans ce cas, il est intéressant de décomposer la variabilité de ces variables dans les dimensions inter-individuelle et intra-individuelle, ou ``between'' et ``within'' dans le jargon des données de panel.

$$E(u_{it} \mid x_i) = 0.$$

Celles-ci sont plus fortes que H_1 puisque $E(x_i' u_{it}) = E(x_i' E(u_{it} \mid x_i)) = 0.$

On remarquera aussi que comme:

$$V(u_{it} \mid x_i) = E(u_{it}^2 \mid x_i) - (E(u_{it} \mid x_i))^2 > 0$$

l'hypothèse H_2 entraîne que:

$$\mid E(u_{it} \mid x_i) \mid < \sigma^2$$

ce qui donne une restriction implicite sur l'espérance conditionnelle.

1.3. Dimensions inter-individuelle et intra-individuelle

C'est sous la forme matricielle du modèle linéaire que les calculs que nous allons présenter sont les plus simples. Dans le cas des données de panel, il faut néanmoins faire attention à cette écriture puisque les données sont à double indice. On ``empilera'' les observations dans les vecteurs, en commençant par les données relatives au premier individu, puis au deuxième individu et ceci jusqu'au dernier. Le vecteur de la variable dépendante de dimension NT est, si on l'écrit sous sa forme transposée:

$$Y' = (y_{11} \dots y_{1T} \dots y_{i1} \dots y_{iT} \dots y_{N1} \dots y_{NT})$$

Nous allons décomposer ce vecteur dans ses dimensions inter-individuelle et intra-individuelle. Pour cela, on notera, d'abord, la moyenne individuelle des variables comme:

$$y_{i.} = \frac{1}{T} \sum_{t=1}^{T} y_{it}$$

et on définira deux opérateurs dans l'espace R^{NT} agissant sur le vecteur des observations.

Définition 1 : L'opérateur B dit de moyenne individuelle ou opérateur inter-individuel est l'opérateur qui transforme tout vecteur Y d'éléments y_{it} en le vecteur, BY, de dimension NT, dont les éléments sont les moyennes individuelles, $y_{i.}$.

Définition 2 : L'opérateur W dit opérateur intra-individuel est l'opérateur qui transforme tout vecteur Y d'éléments y_{it} en le vecteur, WY, de dimension NT, dont les éléments sont les écarts aux moyennes individuelles, $y_{it} - y_{i.}$.

On remarque que :

$$W = l - B$$

où I est la matrice identité dans R^{NT}. Par construction, ces opérateurs permettent la décomposition de tout vecteur de R^{NT} en ses projections sur deux sous espaces orthogonaux entre eux. En effet, sans démonstration (voir par exemple, Hsiao, 1985) :

$$l = B + W$$

$$B^2 = B = B' \quad W^2 = W = W'$$

$$BW = 0$$

L'opérateur ``between'' B projette orthogonalement tout vecteur sur le sous-espace des vecteurs n'ayant pas de variabilité temporelle, dit de dimension inter-individuelle. L'opérateur W projette orthogonalement sur le sous-espace des vecteurs dont les moyennes individuelles sont nulles dit de dimension intra-individuelle. Cette décomposition permet d'obtenir directement l'équation de la variance. La variabilité de Y est la somme des variabilités inter-individuelles et intra-individuelles. Notons $\tilde{Y}$ le vecteur d'éléments $y_{it} - y_{..}$ où $y_{..}$ est la moyenne générale de y_{it}. On a alors par le théorème de Pythagore :

$$\left\| \tilde{Y} \right\|^2 = \left\| B\tilde{Y} \right\|^2 + \left\| W\tilde{Y} \right\|^2$$

où $\left\| . \right\|^2 = \sum_{i,t} x_{it}^2$.

On revient pour finir au modèle (EF) en empilant les observations de la même façon que précédemment pour les variables explicatives et les perturbations :

$$Y = X\beta + \sum_{i-1}^{N} \alpha_i E_i + U$$

où les éléments des vecteurs E_i sont tous nuls sauf ceux relatifs à l'individu i, et donc entre les positions $i(T-1)+1$ et iT, qui valent 1. Ainsi :

$$BE_i = E_i \quad WE_i = 0$$

et les vecteurs E_i forment une base de l'espace inter-individuel. De plus, les hypothèses H_1 et H_2 s'écrivent :

$$E(X'U) = E(E_i'U) = 0$$

$$E(UU' \mid X, E_i) = \sigma_u^2 I_{NT}$$

On peut alors décomposer le modèle (EF) sous sa forme matricielle dans ses dimensions inter-individuelle et intra-individuelle:

$$\begin{cases} WY = WX\beta + WU \\ BY = BX\beta + \sum_{i=1}^{N} \alpha_i E_i + BU \end{cases}$$

et par les propriétés des opérateurs, le modèle complet et le « modèle projeté » sont équivalents.

1.4. Estimation

Pour estimer les paramètres β du modèle (EF) par les MCO sans estimer les paramètres α_i, on peut raisonner de deux façons. D'abord, on peut utiliser la propriété de Frish-Waugh rappelée en annexe 3. On projette le modèle sur l'espace orthogonal aux effets individuels et donc aux variables E_i. Or comme celui-ci est l'espace intra-individuel, on utilise le projecteur orthogonal W pour obtenir le modèle projeté :

$$WY = WX\beta + WU$$

L'estimateur des MCO dans ce modèle projeté dans la dimension intraindividuelle est obtenu par la régression des écarts de la variable dépendante à ses moyennes individuelles, $y_{it} - y_{i.}$, sur les écarts individuels des variables explicatives à leurs moyennes individuelles, $x_{it} - x_{i.}$.

L'autre manière de raisonner est d'utiliser le modèle projeté dans les deux dimensions. En utilisant la démarche vue pour l'identification, on notera que la deuxième équation peut se réécrire:

$$BY = \sum_{i=1}^{N} (\alpha_i + x_{i.}\beta) E_i + BU = \sum_{i=1}^{N} \tilde{\alpha}_i E_i + BU$$

Cette partie du modèle ne sert donc à rien pour l'estimation de β et l'équation dans la dimension intraindividuelle résume toute l'information disponible sur β. Ainsi, **la dimension inter-individuelle n'apporte pas d'information sur le paramètre** β sous les hypothèses H_1 à H_3. Ce résultat doit être souligné puisque la seule information sur β, avec des données de coupe transversale, est apportée par la dimension inter-individuelle.

L'estimateur MCO du modèle intraindividuel est appelé estimateur de la covariance ou estimateur within. La condition d'identification du paramètre β est la condition habituelle d'absence de multicolinéarité

$$H_4 \ : \ rg(X'WX) = K.$$

qui est nécessaire et suffisante. Toutes les variables x_{it} doivent varier dans la dimension temporelle et ceci de façon non colinéaire entre elles.

L'estimateur de la covariance ou within est donné par :

$$\hat{\beta}^{(w)} = (X'WX)^{-1} \ X'WY$$

et sa variance est:

$$V\hat{\beta}^{(w)} = \sigma_u^2 \ (X'WX)^{-1}$$

Comme c'est un estimateur MCO dans un modèle linéaire homoscédastique[5], cet estimateur est convergent vers β, asymptotiquement normal quand le nombre d'observations N tend vers l'infini, et il est l'estimateur de meilleur précision dans l'ensemble des estimateurs linéaires en Y.

En effet, on notera que sous de faibles conditions, la loi des grands nombres nous donne :

$$\underset{N\to\infty}{p\lim} \frac{X'WX}{N} = E((x_i - x_{i.})'(x_i - x_{i.})) = R$$

la limite en probabilité de la matrice de variance-covariance du vecteur des écarts des variables à leur moyennes individuelles, d'élément, $x_{it} - x_{i.}$. On supposera que R est de plein rang par extension de l'hypothèse H_4. On a donc :

$$\begin{aligned}
\underset{N\to\infty}{plim}\, \hat{\beta}^{(w)} &= \underset{N\to\infty}{plim} (X'WX)^{-1} \ X'(WX\beta + WU) \\
&= \beta + \underset{N\to\infty}{p\lim} \left(\frac{X'WX}{N} \right)^{-1} \frac{X'WU}{N} \\
&= \beta + R^{-1} \ \underset{N\to\infty}{p\lim} \frac{X'WU}{N}
\end{aligned}$$

Or par l'hypothèse H_1, il n'y a pas de corrélation entre les perturbations et les variables explicatives et donc, sous des conditions faibles, une loi des grands nombres nous donne :

$$\underset{N\to\infty}{p\lim} \frac{X'WU}{N} = E((x_i - x_{i.})'u_i) = 0$$

ce qui montre la convergence de l'estimateur $\overline{\beta}^{(w)}$ vers β.

La propriété de normalité asymptotique s'appuie sur le même développement en notant que:

---

[5]Cela n'est pas évident à première vue puisque:

$$E(WUU'W \mid X) = \sigma_u^2 W$$

Cela néanmoins ne fait que retraduire le fait que le modèle est projeté et W est bien l'opérateur identité dans l'espace projeté.

$$\overline{\beta}^{(w)} - \beta = \left(\frac{X'WX}{N} \right)^{-1} \frac{X'WU}{N}$$

et en utilisant un théorème central limite:

$$\sqrt{N} \frac{X'WU}{N} \underset{N\to\infty}{\overset{L}{?}} \mathbf{N}(0, E((x_i - x_{i.})' u_i^2 (x_i - x_{i.}))$$

Or:

$$\begin{aligned} E((x_i - x_{i.})' u_i^2 (x_i - x_{i.})) &= E((x_i - x_{i.})' E(u_i^2 \mid x_i)(x_i - x_{i.})) \\ &= \sigma_u^2 E((x_i - x_{i.})'(x_i - x_{i.})) \\ &= \sigma_u^2 . R \end{aligned}$$

et donc:

$$\sqrt{N}(\hat{\beta}^{(w)} - \beta) \underset{N\to\infty}{\overset{L}{?}} \mathbf{N}(0, \sigma_u^2 R^{-1})$$

Dans certains cas, on peut chercher à obtenir une estimation des effets individuels et il faut utiliser pour cela la dimension inter-individuelle. La manière la plus simple de procéder est d'utiliser le modèle interindividuel puisque c'est une équation d'analyse de variance. L'estimateur de $\widetilde{\alpha}_i$ est alors $y_{i.}$ et on en déduit alors un estimateur de α_i :

$$\hat{\alpha}_i = y_i - x_i \hat{\beta}^{(w)}$$

On notera que cette estimation est effectuée individu par individu. Elle n'est donc pas convergente quand le nombre d'individus augmente mais seulement si le nombre de périodes T tend vers l'infini.

1.5. Le modèle à effets individuels

Les enquêtes dont sont dérivées les données de panel comportent en général des milliers d'individus. Comme on ne s'intéresse pas ou peu, à la mesure des effets individuels mais que l'on cherche simplement à les contrôler, l'idée est de considérer que ces effets sont des perturbations. Ils ne sont donc plus traités comme des paramètres comme dans le modèle à effets fixes.

1.5.1. L'écriture du modèle

Le modèle général à effets aléatoires individuels (EI) s'écrit comme :

$$y_{it} = x_{it}\beta + \alpha_i + u_{it}$$

où α_i est un terme d'hétérogénéité individuelle inobservable constante dans le temps alors que u_{it} est un terme d'hétérogénéité individuelle inobservable variable au cours du temps. On inclut toujours parmi les régresseurs des indicatrices temporelles. Pour simplifier la suite de la discussion, on distinguera deux types de variables explicatives, distinction qui rappellera la discussion sur l'identification dans le modèle à effets fixes. D'une part, des variables, $x_{it}^{(1)}$, varient dans la dimension intra-individuelle. Comme dans la section précédente, on supposera que, sous forme matricielle:

$$H_4 \ : \ X_1' W X_1 \text{ est de plein rang}$$

où l'opérateur W est l'opérateur intraindividuel. En second lieu, des variables $x_i^{(2)}$, comme la constante, ne varient pas dans la dimension intra-individuelle. On a donc alors sous forme

matricielle :

$$WX_2 = 0 \;\; BX_2 = X_2$$

La forme générale du modèle à effets individuels (EI) est :

$$y_{it} = x_{it}^{(1)}\beta_1 + x_i^{(2)}\beta_2 + \alpha_i + u_{it}$$

On ne retiendra pour l'instant comme hypothèse de spécification que les hypothèses H_1 à H_2 exposées dans la section précédente en modifiant la première hypothèse pour tenir compte du caractère aléatoire de l'effet individuel. En notant

$$x_i = (x_{i1}^{(1)}, ..., x_{iT}^{(1)}, x_i^{(2)})$$

on écrit donc :

$$H_1 \; : \; Ex_i^{'}u_{it} = 0, Eu_{it} = 0$$

$$H_2^a \; : \; u_i \text{ et } u_j \text{ sont indépendants, } H_2^b \; : \; E(u_i u_i^{'} \mid x_i) = \sigma_u^2 . I_T$$

$$H_3 \; : \; E\alpha_i = 0, E\alpha_i u_{it} = 0$$

où H_1 et H_2 s'interprètent comme auparavant et où H_3 précise que la structure globale d'hétérogénéité dépend d'un facteur individuel constant seulement, α_i, et de T facteurs orthogonaux, u_{it}.

Quelle est donc la différence entre le modèle à effets fixes et le modèle à effets individuels? Si on ne fait pas d'autres hypothèses, il n'y en a aucune et c'est ce que nous montrerons en premier lieu en exposant le modèle dit de Mundlak. Mais les différences entre les deux modèles peuvent apparaitre plus flagrantes à un utilisateur qui voudrait estimer le modèle (EI) par MCO sans prendre le temps de réfléchir aux hypothèses nécessaires pour que cette procédure ait de bonnes propriétés. Dans ce cas, on a besoin, en effet, de préciser la structure de corrélation entre la partie de la perturbation qu'est l'effet individuel, α_i, et les variables explicatives, x_{it}. L'estimateur ``naif'' des MCO n'est convergent que si cette corrélation est nulle.

S'interroger sur la corrélation entre effets individuels et variables explicatives est sans doute la meilleure façon de comprendre que la différence que l'on mettait en valeur autrefois, entre des modèles qui traitent les effets individuels comme des paramètres ou comme des perturbations, est moins importante qu'on ne le croyait. Dans le modèle à effets fixes, la corrélation des effets individuels et des variables explicatives est quelconque puisque les effets individuels sont des paramètres associés à des variables explicatives. Or on sait que les estimateurs des MCO sont convergents quelle que soit la corrélation entre les variables explicatives du modèle. Ce n'est plus vrai quand les effets individuels sont aléatoires et donc ``inclus'' dans les perturbations. Il faut maintenant que la corrélation entre effets individuels et variables explicatives soit nulle pour satisfaire la condition d'exogénéité des variables explicatives. Car ce n'est que dans ce cas qu'on obtient la propriété de convergence de l'estimateur des MCO (voir Robin, 2000).

1.5.2. Le modèle de Mundlak

Pour estimer le modèle (EI) il nous faut préciser les hypothèses sur la corrélation entre les variables explicatives et les effets individuels aléatoires. Dans le modèle de Mundlak, on cherche à contrôler cette corrélation. Ainsi, on peut toujours écrire la projection linéaire des effets individuels sur les moyennes individuelles des variables explicatives et sur les variables fixes au cours du temps :

$$\alpha_i = x_{i.}^{(1)}\theta_1 + x_i^{(2)}\theta_2 + \nu_i = z_i\theta + \nu_i$$

de telle façon que $E(z_i' v_i) = 0$. On a défini $x_{i.}^{(1)}$ comme la moyenne individuelle des variables explicatives, $x_{it}^{(1)}$, qui varient au cours du temps et z_i comme l'assemblage des $x_{i.}^{(1)}$ et $x_i^{(2)}$. L'équation précédente est une simple technique de projection linéaire. En effet, il suffit pour cela de s'apercevoir que l'hypothèse $E(z_i' v_i) = 0$ est équivalente à :

$$E(z_i'(\alpha_i - z_i\theta)) = 0$$

$$\theta = \left[E(z_i' z_i)\right]^{-1} E(z_i' \alpha_i)$$

et qu'elle n'est donc en réalité qu'un moyen de définir le paramètre θ :[6]

On peut alors réécrire le modèle (EI) comme:

$$\begin{aligned} y_{it} &= x_{it}^{(1)}\beta_1 + x_i^{(2)}\beta_2 + x_{i.}^{(1)}\theta_1 + x_i^{(2)}\theta_2 + v_i + u_{it} \\ &= (x_{it}^{(1)} - x_{i.}^{(1)})\beta_1 + x_{i.}^{(1)}(\beta_1 + \theta_1) + x_i^{(2)}(\beta_2 + \theta_2) + v_i + u_{it} \\ &= (x_{it}^{(1)} - x_{i.}^{(1)})\beta_1 + x_{i.}^{(1)}\gamma_1 + x_i^{(2)}\gamma_2 + v_i + u_{it} \end{aligned}$$

Deux conséquences se dégagent immédiatement de cette écriture. D'abord, sans autres hypothèses sur θ_1 et θ_2, il n'y a aucune relation entre les paramètres β_1 d'une part, et les paramètres γ_1 et γ_2 d'autre part. Deuxièmement, le modèle s'écrit de façon matricielle comme le modèle (EM) :

$$Y = WX_1\beta_1 + BX_1\gamma_1 + BX_2\gamma_2 + \tilde{\varepsilon}$$

où on a redéfini le terme d'erreur $\tilde{\varepsilon}$ de manière appropriée. Ses éléments sont $v_i + u_{it}$. En projetant sur les dimensions inter et intra-individuelles, on obtient la forme équivalente :

$$\begin{cases} WY = WX_1\beta_1 + W\tilde{\varepsilon} \\ BY = BX_1\gamma_1 + BX_2\gamma_2 + B\tilde{\varepsilon} \end{cases}$$

ce qui revient à séparer les deux groupes de régresseurs orthogonaux, le premier groupe étant constitué par WX_1 et le deuxième par BX_1 et BX_2. Remarquons que :

$$B\tilde{\varepsilon} = \{v_i + \frac{1}{T}\sum_{t=1}^{T} u_{it}\}_{i,t} \quad W\tilde{\varepsilon} = \{u_{it} - \frac{1}{T}\sum_{t=1}^{T} u_{it}\}_{i,t}$$

où, entre accolades apparaissent les éléments des vecteurs ainsi définis. On notera de plus que $W\tilde{\varepsilon} = W\varepsilon$ puisque la différence entre $\tilde{\varepsilon}$ et ε n'est qu'un terme individuel, $\alpha_i - v_i$.

L'hypothèse H_1 d'absence de corrélation entre les chocs u_{it} et les variables explicatives se traduit par :

$$E(X_1' W\varepsilon) = 0$$

[6]S'il y a multicolinéarité entre les w_i , cette inverse n'existe pas. Comme c'est un problème d'identification, il existe alors plusieurs valeurs de θ qui justifie l'écriture de la projection linéaire. On en considère alors une d'entre elles.

et on obtient alors le résultat suivant :

Proposition: *L'estimateur des moindres carrés ordinaires de* β_1 *dans le modèle de Mundlak est l'estimateur de la covariance ou estimateur within :*

$$\hat{\beta}_1^{(M)} = \hat{\beta}_1^{(w)}$$

Preuve : Comme il n'y a aucune relation entre les paramètres apparaissant dans la dimension inter-individuelle et dans la dimension intra-individuelle, l'estimateur MCO de β_1 est égal à l'estimateur MCO du modèle projeté dans la dimension intra-individuelle. Cet estimateur a les propriétés usuelles de l'estimateur MCO puisque la condition d'absence de corrélation entre erreurs et régresseurs est vérifiée et puisque le modèle projeté est homoscédastique de par les hypothèses H_2. ?

L'estimation de ce modèle, ou au moins des coefficients des variables qui varient au cours du temps est donc identique à l'estimation d'un modèle à effets fixes puisque c'est le même estimateur, celui de la covariance, qui hérite des propriétés des MCO, dans les deux modèles.

Pour les autres coefficients, il suffit de considérer le modèle dans sa dimension inter-individuelle et l'écrire sous la forme:

$$y_{i.} = x_{i.}^{(1)}\gamma_1 + x_i^{(2)}\gamma_2 + v_i + \frac{1}{T}\sum_{t=1}^{T} u_{it}$$

L'hypothèse H_1 suppose l'absence de corrélation entre u_{it} et les variables explicatives et par construction, v_i, n'est pas corrélé avec les variables explicatives ($E(z_i' v_i) = 0$). Sous la condition d'absence de multicolinéarité entre variables explicatives[7], l'estimateur MCO de ce modèle projeté dans la dimension inter-individuelle est donc convergent quand le nombre d'individus N tend vers l'infini. Il est appelé estimateur inter-individuel ou estimateur ``between''. Pour simplifier, on fait de plus une hypothèse d'homoscédasticité des effets individuels, v_i. On remplace alors H_3 par H_3':

$$H_3'^{(a)} \ : \ v_i \text{ et } v_j \text{ sont indépendants}$$

$$H_3'^{(b)} \ : \ E(z_i' v_i) = 0, \quad E(v_i u_{it} \mid z_i) = 0, \quad E(v_i^2 \mid z_i) = \sigma_v^2$$

où la première partie de $H_3'^{(b)}$ est vraie par construction de v_i et où la seconde partie de $H_3'^{(b)}$ est une hypothèse d'homoscédasticité simplificatrice. On obtient alors que:

$$E\left[\left(v_i + \frac{1}{T}\sum_{t=1}^{T} u_{it}\right)^2 \mid z_i\right] = \sigma_v^2 + \frac{\sigma_u^2}{T}$$

et le modèle dans sa dimension inter-individuelle est homoscédastique. L'estimateur between, noté $\hat{\gamma}^{(b)}$, est convergent, asymptotiquement normal et efficace dans l'ensemble des estimateurs linéaires en BY. On remarquera que cet estimateur s'obtient par l'estimation par les MCO du modèle projeté dans la dimension inter-individuelle :

$$BY = BX_1\gamma_1 + BX_2\gamma_2 + B\tilde{\varepsilon}$$

[7]Ce problème n'est pas abstrait puisque, par exemple, les moyennes individuelles des variables indicatrices temporelles sont toutes égales à $1/T$. Les coefficients des indicatrices temporelles sont donc absorbés par la constante du modèle dans sa dimension interindividuelle. Cela ne fait que traduire le fait que les coefficients temporels ne varient pas dans la dimension inter-individuelle et ne varient que dans la dimension intra-individuelle. Mais on adapte le modèle à un tel cas en enlevant les variables qui créent la multicolinéarité.

En utilisant la propriété de Frisch-Waugh, on obtient donc l'estimateur between de γ_1 en projetant sur l'espace orthogonal à BX_2, projection dont l'opérateur est noté M_{BX_2}. Ainsi, l'estimateur between de γ_1 est :

$$\hat{\gamma}_1^{(b)} = (X_1' B M_{BX_2} B X_1)^{-1} X_1' B M_{BX_2} B Y$$

Finalement, on peut recombiner les deux étapes d'estimation en une seule étape puisque les régressions within et between sont indépendantes. Les deux espaces de projection sont orthogonaux. La procédure de Mundlak revient alors à régresser la variable dépendante sur les écarts des variables explicatives, x_{it}, à leurs moyennes individuelles, sur les moyennes individuelles de ces variables et sur les variables constantes au cours du temps comme le montre l'équation (EM). Les coefficients des premières variables sont les estimateurs within de β_1 et les coefficients des autres variables sont les estimateurs between des paramètres γ.

Néanmoins, ces derniers paramètres γ n'ont aucune signification économique. Ils sont sommes de paramètres structurels β_1 et β_2 et de paramètres de contrôle, θ_1 et θ_2. On a d'ailleurs suivi en cela une procédure générale en économétrie en incluant dans ce modèle à effets individuels, des termes qui contrôlent des corrélations gênantes. Mais en supposant que ces paramètres de contrôle, θ_1 et θ_2 ne sont pas nuls, on perd la capacité d'identifier β_2 et on perd le lien qui existerait entre l'estimateur "within" de β_1 dans la dimension intra-individuelle et l'estimateur "between" de β_1 dans la dimension inter-individuelle.

Il est difficile de surmonter le problème d'identification de β_2 à moins de recourir à des méthodes à variables instrumentales ce que nous ferons plus bas. On supposera directement que la condition d'orthogonalité est satisfaite. Il n'y a pas de corrélation entre les effets individuels et les variables $x_i^{(2)}$, $\theta_2 = 0$ et $E(x_i^{(2)'} v_i) = 0$. Donc $\gamma_2 = \beta_2$. Par exemple, dans une équation de taux de salaire, on supposera que les effets individuels affectant les taux de salaire, ne sont pas corrélés avec l'éducation. Cette hypothèse a été fortement contestée dans la littérature empirique mais ce débat se rattache plutôt au problème d'endogénéité qui n'est pas propre aux données de panel.

Par contre, comme on dispose d'un estimateur de β_1 dans la dimension intra-individuelle et d'un estimateur de $\beta_1 + \theta_1$ dans la dimension inter-individuelle, il est facile de recombiner les deux informations pour obtenir un estimateur de θ_1 :

$$\hat{\theta}_1 = \hat{\gamma}_1^{(b)} - \hat{\beta}_1^{(w)}$$

La matrice de variance-covariance asymptotique de cet estimateur est la somme des matrices de variance-covariance asymptotiques de ces deux estimateurs puisqu'ils sont obtenus dans les dimensions inter et intra-individuelles qui sont orthogonales. On peut donc tester l'hypothèse de nullité de ce paramètre par un test de Wald. Il y a une procédure pratique plus simple puisque l'équation (eagm3) peut se réécrire en fonction de $x_{it}^{(1)}$, $x_{i.}^{(1)}$, $x_i^{(2)}$ et en utilisant $\gamma_2 = \beta_2$ comme :

$$y_{it} = x_{it}^{(1)} \beta_1 + x_{i.}^{(1)} (\gamma_1 - \beta_1) + x_i^{(2)} \beta_2 + v_i + u_{it}$$

On a par définition $\theta_1 = \gamma_1 - \beta_1$. Comme l'équation précédente n'est qu'une réécriture de (EM), les estimateurs MCO de cette équation sont $\hat{\beta}_1^{(w)}$, $\hat{\theta}_1$ et $\hat{\beta}_2^{(b)} = \hat{\gamma}_2^{(b)}$. Le test de H_0 : $\theta_1 = 0$ est

donc le test de Wald de nullité d'un sous-vecteur des paramètres dans cette régression, plus pratique à réaliser par les logiciels habituels.[8]

La signification de l'hypothèse nulle, $\theta_1 = 0$, est la suivante. Par définition de θ_1 que l'on a vue plus haut et sous l'hypothèse maintenue que $\theta_2 = 0$, on a :

$$\theta_1 = 0 \Leftrightarrow E(x_i^{(1)'}\alpha_i) = 0$$

Les effets individuels ne seraient donc pas corrélés avec les moyennes des variables explicatives. En d'autres termes, il n'y aurait pas de corrélation entre les effets individuels qui affectent la variable dépendante et des effets individuels qui affectent les variables explicatives. Si on ne peut pas rejeter cette hypothèse d'absence de corrélation, le vrai modèle (EEC) s'écrit maintenant :

$$y_{it} = x_{it}^{(1)}\beta_1 + x_i^{(2)}\beta_2 + v_i + u_{it}$$

sous les hypothèses H_1, H_2 et $H_3^{'}$. La perturbation $v_i + u_{it}$ n'est pas corrélée, par construction, aux variables explicatives et l'estimateur des MCO est maintenant convergent en utilisant les propriétés du modèle linéaire. Or l'estimateur de la covariance reste convergent et l'estimateur ``between'', $\hat{\gamma}_1^{(b)}$ est maintenant aussi convergent vers β_1 puisque $\theta_1 = 0$. On a donc une multiplicité d'estimateurs convergents dont on doit comparer la précision. Mais dans ce cas, les deux estimations dans les dimensions inter et intra-individuelles sont liées et il y a des gains de précision à attendre de l'estimation conjointe dans les deux dimensions, c'est à dire de l'estimation par moindres carrés généralisés de la dernière équation. Ce modèle est un des premiers modèles qui a été proposé pour analyser les données de panel (Balestra et Nerlove, 1966) et c'est ce modèle à effets aléatoires non corrélés que nous exposons maintenant.

1.6. Le modèle à effets aléatoires non corrélés

Ce modèle se définit par l'équation (EEC) ci-dessus sous les hypothéses H_1, H_2 à $H_3^{'}$. Si on note la perturbation ``agrégée'' comme:

$$\varepsilon_{it} = v_i + u_{it}$$

On dit dans ce cas que le terme aléatoire a une structure à un facteur, ici le facteur individuel. En effet, une manière équivalente d'écrire les hypothèses H_2 et $H_3^{'}$ est d'écrire la matrice de moment du second ordre du vecteur $\varepsilon_i = (\varepsilon_{i1},..,\varepsilon_{iT})$ sous la forme suivante :

$$E(\varepsilon_i\varepsilon_i^{'} \mid x_i) = \Sigma = \begin{pmatrix} \sigma_v^2 + \sigma_u^2 & \sigma_v^2 & \dots & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 + \sigma_u^2 & \sigma_v^2 & \vdots \\ \vdots & \sigma_v^2 & \ddots & \sigma_v^2 \\ \sigma_v^2 & \dots & \sigma_v^2 & \sigma_v^2 + \sigma_u^2 \end{pmatrix}$$

La matrice de moment du second ordre de la perturbation ε, en écriture matricielle, est alors une matrice bloc-diagonale dont les blocs sont identiques et égaux à Σ. A cause de l'écriture de Σ, elle est donc la somme de deux termes : une matrice diagonale, $\sigma_u^2.Id_{NT}$ où la matrice Id_{NT} est la matrice carrée identité de taille NT et qui est la matrice relative aux perturbations u_{it} ; une

[8]On trouvera aussi dans la littérature une procédure de test dite de Hausman (Hausman et Taylor, 1981) mais la statistique utilisée pour cette procédure est *numériquement* égale à la statistique de Wald que nous présentons ici (Arellano, 1993). De plus, cette dernière s'étend plus facilement au cas hétéroscédastique.

matrice $\sigma_v^2 M$, où M est une matrice bloc-diagonale de dimension (NT, NT) et dont les blocs sont des matrices d'éléments égaux à 1 et qui est la matrice qui concerne les effets individuels v_i .

On peut simplifier cette écriture en réécrivant M . Considérons un vecteur Y quelconque de R^{NT} . Alors le vecteur MY a pour élément générique $\sum_{t=1}^{T} y_{it}$, ce qui prouve que $M = T.B.$ La matrice de moment du second ordre des perturbations est donc :

$$\begin{aligned} E(\varepsilon\varepsilon' \mid X) &= \sigma_u^2 Id + T\sigma_v^2 B \\ &= \sigma_u^2 W + (\sigma_u^2 + T\sigma_v^2) B \\ &= \sigma_u^2 (W + \frac{1}{\rho^2} B) \end{aligned}$$

en utilisant $W + B = Id$ et où $\rho = \frac{1}{\sqrt{T}} (\frac{\sigma_u^2}{\frac{\sigma_u^2}{T} + \sigma_v^2})^{1/2}$ varie entre 0 et 1. On remarquera que ce coefficient ρ tend vers 0 quand T tend vers l'infini. Son carré est proportionnel au rapport de la variance des perturbations dans la dimension intra-individuelle, σ_u^2, et de la variance dans la dimension inter-individuelle, $\frac{\sigma_u^2}{T} + \sigma_v^2$ que l'on a déterminée plus haut.

Le modèle à effets aléatoires (EEC) sous les hypothèses H_1, H_2 et $H_3^{'}$ est donc un modèle linéaire général:

$$Y = X_1 \beta_1 + BX_2 \gamma_2 + \varepsilon$$

$$E(X'\varepsilon) = 0$$

$$E(\varepsilon\varepsilon' \mid X) = \sigma_u^2 (W + \frac{1}{\rho^2} B) = \sigma_u^2 \Omega$$

On dispose déjà d'un certain nombre d'estimateurs convergents des coefficients de ce modèle: l'estimateur de la covariance qui est l'estimateur MCO du modèle projeté dans la dimension intraindividuelle, l'estimateur between qui est l'estimateur MCO du modèle projeté dans la dimension interindividuelle, l'estimateur MCO lui-même. Mais, l'estimateur de meilleure précision est celui des moindres carrés généralisés que nous voyons maintenant.

1.6.1. L'estimateur des moindres carrés généralisés

Le traitement du modèle linéaire général se déroule en deux temps. On suppose dans ce paragraphe que **l'on connait la forme de l'hétéroscédasticité et donc le paramètre** ρ . On sphéricise alors le modèle pour le rendre homoscédastique en le prémultipliant par une matrice H telle que :

$$H.\Omega.H' = I$$

Il est facile de vérifier qu'ici, $H = W + \rho B$ et le modèle sphéricisé prend donc la forme :

$$(W + \rho B)Y = (W + \rho B) X_1 \beta_1 + \rho B X_2 \beta_2 + H\varepsilon$$

Si on ne s'intéresse qu'à l'estimateur de β_1, on utilise le théoréme de Frish-Waugh en prémultipliant cette équation par le projecteur sur l'espace orthogonal aux variables BX_2, que

l'on note M_{BX_2}. On remarquera que $M_{BX_2}W = W$ puisque BX_2 ne varie que dans la dimension inter-individuelle. Le modèle se réécrit :

$$(W + \rho M_{BX_2}B)Y = (W + \rho M_{BX_2}B)X_1\beta_1 + M_{BX_2}H\varepsilon$$

et l'estimateur des MCG est donc :

$$\hat{\beta}_1^{(MCG)} = (X_1'(W + \rho^2 BM_{BX_2}B)X_1)^{-1}X_1'(W + \rho^2 BM_{BX_2}B)Y$$

On peut utiliser l'expression de l'estimateur de la covariance :

$$(X_1'WX_1)\hat{\beta}_1^{(w)} = X_1'WY$$

et de l'estimateur between:

$$(X_1'BM_{BX_2}BX_1)\hat{\beta}_1^{(b)} = X_1'BM_{BX_2}BY$$

pour obtenir que :

$$\hat{\beta}_1^{(MCG)} = \Lambda.\hat{\beta}_1^{(w)} + (I - \Lambda)\hat{\beta}_1^{(b)}$$

L'estimateur des MCG est donc une moyenne pondérée des estimateurs within et between et combine donc l'information dans les dimensions intra et interindividuelles. De plus, cette pondération dépend de T. En effet, comme ρ tend vers 0 quand T tend vers l'infini, l'expression de $\hat{\beta}_1^{(MCG)}$ montre que numériquement l'estimateur des MCG tend vers l'estimateur de la covariance quand T tend vers l'infini. Cela semble naturel puisque le poids de la dimension intra-individuelle est proportionnel à T.

Pour finir, on remarquera que **l'on ne connait pas** ρ**.** On adopte donc la méthode d'estimation dite des **moindres carrés quasi-généralisés.**

1.6.2. L'estimateur des MCQG

On cherche d'abord un estimateur convergent de ρ, $\hat{\rho}$ et on remplace dans l'expression de l'estimateur MCG le paramètre inconnu par son estimateur :

$$\hat{\beta}_1^{(MCQG)} = (X_1'(W + \hat{\rho}^2 BM_{BX_2}B)X_1)^{-1}X_1'(W + \hat{\rho}^2 BM_{BX_2}B)Y$$

Pour construire un estimateur convergent de ρ ou de ρ^2, on utilise les deux estimateurs, within et between, que l'on sait convergents. En effet, dans la dimension intra-individuelle, la variance des perturbations est donnée par :

$$E(W\varepsilon\varepsilon'W \mid X) = \sigma_u^2 W$$

ce qui permet de construire un estimateur convergent de σ_u^2 à partir des résidus de l'estimation de la covariance[9]:

$$\hat{\sigma}_u^2 = \frac{\hat{\varepsilon}^{(w)'}W\hat{\varepsilon}^{(w)}}{N(T-1)}$$

où $\hat{\varepsilon}^{(w)} = Y - X_1\hat{\beta}_1^{(w)}$.

[9]On notera le facteur $T-1$ au dénominateur. Si l'estimation est faite dans SAS par exemple, il faudra faire attention au fait que l'estimateur utilise un facteur T au dénominateur comme nous le verrons dans l'application empirique.

Dans la dimension inter-individuelle, on a de façon similaire :

$$E(B\varepsilon\varepsilon' B \mid X) = \frac{\sigma_u^2}{T\rho^2} B$$

ce qui permet de construire un estimateur convergent de $\frac{\sigma_u^2}{T\rho^2}$ à partir des résidus de l'estimation between:

$$\overset{\wedge}{\left(\frac{\sigma_u^2}{T\rho^2}\right)} = \frac{\hat{\varepsilon}^{(b)'} B \hat{\varepsilon}^{(b)}}{NT}$$

où $\hat{\varepsilon}^{(b)} = Y - X\hat{\beta}^{(b)}$.[10]

L'estimateur convergent de ρ^2 est alors donné par :

$$\hat{\rho}^2 = \frac{\hat{\sigma}_u^2}{\overset{\wedge}{\left(\frac{\sigma_u^2}{\rho^2}\right)}}$$

Remarquons finalement que cet estimateur ne domine pas forcément -- en termes de précision -- d'autres estimateurs comme les estimateurs within et between quand le nombre d'observations reste petit (50 ou 100 observations individuelles) puisque ses propriétés sont des propriétés asymptotiques.

1.6.3. Méthodes à variables instrumentales

Nous venons de voir comment l'hypothèse d'absence de corrélation entre variables explicatives et effets individuels permettait de construire un estimateur MCQG plus précis que l'estimateur within qui n'est efficace que sous l'hypothèse de corrélation non nulle entre variables explicatives et effets individuel.

Bien entendu, toutes les situations intermédiaires sont possibles, situations où certaines variables explicatives sont corrélées avec les effets individuels (l'éducation par exemple dans une équation de salaire) alors que d'autres ne le sont pas (l'âge par exemple). Une procédure d'estimation pourrait s'inspirer des procédures précédentes. On inclurait dans l'équation à estimer, les moyennes des variables explicatives corrélées aux effets individuels et on estimerait le modèle par une méthode de MCQG. En effet, dans ce cas, les projections dans les dimensions within et between ne résultent pas en des modèles indépendants puisque les coefficients des variables explicatives non corrélées aux effets individuels sont les mêmes dans les deux dimensions. Les méthodes within et between sont donc génériquement moins précises asymptotiquement que l'estimation MCQG.

On présente ici une autre méthode due à Hausman et Taylor (1981) et qui est une méthode à variables instrumentales dont le principe nous servira dans la partie suivante. Supposons donc que les variables explicatives, variables ou non au cours du temps, se scindent en deux ensembles de variables: celles qui sont corrélées aux effets individuels et celles qui ne le sont pas :

$$x_{it}^{(1)} = (s_{it}^{(1)}, z_{it}^{(1)}) \quad x_i^{(2)} = (s_i^{(2)}, z_i^{(2)})$$

[10]Là aussi, il faut faire attention au logiciel utilisé. Dans SAS par exemple, l'estimateur between sera calculé à partir des N données individuelles et non de NT données. Il faudra donc le corriger d'un facteur T comme nous le verrons dans l'application empirique.

sous l'hypothèse H_5 :

$$E(s_{it}^{(1)'}\alpha_i) = h^{(1)} \quad E(s_i^{(2)'}\alpha_i) = h^{(2)}$$

$$E(z_{it}^{(1)'}\alpha_i) = 0 \quad E(z_i^{(2)'}\alpha_i) = 0$$

La méthode d'Hausman et Taylor consiste alors à estimer l'équation :

$$y_{it} = x_{it}^{(1)}\beta_1 + x_i^{(2)}\beta_2 + \alpha_i + u_{it}$$

par une méthode à variables instrumentales où les variables instrumentales sont les variables non corrélées avec les perturbations $\alpha_i + u_{it}$ sous les hypothèses H_1 et H_5 :

- les variables $z_{it}^{(1)}$ et $z_i^{(2)}$
- les moyennes des variables individuelles $z_{it}^{(1)}$, $z_{i.}^{(1)}$ et les écarts des variables explicatives $s_{it}^{(1)}$ à leurs moyennes individuelles, $s_{it}^{(1)} - s_{i.}^{(1)}$ puisque la corrélation entre $s_{it}^{(1)}$ et α_i est constante au cours du temps.

La condition d'ordre habituelle pour l'identification nous dit que le nombre de variables instrumentales doit être supérieur au nombre de variables explicatives endogènes. On doit donc supposer que le nombre de variables dans $z_{i.}^{(1)}$ doit être supérieur ou égal au nombre de variables dans $s_i^{(2)}$. L'estimation peut dans ce cas s'effectuer par doubles moindres carrés où on remplacera les variables explicatives endogènes $s_{it}^{(1)}$ et $s_i^{(2)}$ par leurs prédicteurs linéaires en fonction des variables instrumentales. Il y a deux moyens d'améliorer la précision de cette estimation. D'abord, si le modèle est sur-identifié, l'estimateur efficace est celui des triples moindres carrés où on prendra en compte la corrélation au cours du temps entre les perturbations, $\alpha_i + u_{it}$, donnée par les hypothèses H_2 et H_3. Ceci revient à prendre en compte explicitement que le modèle s'écrit pour chaque individu sous la forme d'un système à T équations simultanées. Ensuite, Amemiya et MaCurdy (1986) remarquent que d'autres instruments valides sont les variables explicatives $z_{it'}^{(1)}$ à **toutes les dates** t' sous les hypothèses H_1 et H_5. De façon similaire, Breusch, Mizon et Schmidt (1989) remarquent que les écarts aux moyennes individuelles des variables explicatives $s_{it'}^{(2)} - s_i^{(2)}$ à toutes les dates t' sont aussi des instruments valides. On améliore ainsi l'efficacité de l'estimation comme l'ont montré, dans un cadre réunifié, Arellano et Bover (1995).

1.7. Résumé et conclusion

On résume ici la construction de l'estimation de modèles à effets individuels du type :

$$y_{it} = x_{it}^{(1)}\beta_1 + x_i^{(2)}\beta_2 + \alpha_i + u_{it}$$

Sans hypothèse supplémentaires, le modèle le plus général est donné par le modèle de Mundlak :

$$y_{it} = x_{it}^{(1)}\beta_1 + x_{i.}^{(1)}\theta_1 + x_i^{(2)}\gamma_2 + \nu_i + u_{it}$$

sous les hypothèses H_1, H_2 et $H_3^{'}$. On inclut les moyennes individuelles des variables dans la régression pour tenir compte des corrélations entre effets individuels α_i et variables explicatives.

Ce modèle s'estime par moindres carrés ordinaires et c'est donc ce modèle qui doit constituer la première étape des estimations sur données de panel.

Pour gagner en précision, il est alors utile de savoir si on peut utiliser aussi la dimension interindividuelle pour estimer β_1. Ceci passe alors par le test de:

$$H_0 \ : \ \theta_1 = 0$$

dans la régression précédente. Si cette procédure permet de rejeter l'hypothèse nulle, l'estimateur de la covariance obtenu dans le modèle de Mundlak est le meilleur estimateur linéaire convergent de β_1. Si on ne peut rejeter cette hypothèse, les effets individuels sont dits non corrélés -- *implicitement* ``non corrélés avec les variables explicatives". Il est alors utile de recourir à l'estimation du modèle par une méthode de moindres carrés généralisés ou par une méthode des doubles ou triples moindres carrés, comme on l'a vu plus haut. Les estimateurs, au moins quand la dimension individuelle N est grande, seront plus précis.

On voit donc que la différence entre modèle à effets fixes -- $\theta_1 \neq 0$ -- et à effets aléatoires -- $\theta_1 = 0$ -- ne tient pas à une différence conceptuelle très profonde entre ce qui a un caractère fixe et ce qui a un caractère aléatoire. Elle tient beaucoup plus simplement à une hypothèse sur l'absence de corrélation entre effets individuels et variables explicatives. En conclusion, toutes ces méthodes s'appliquent au cas où les effets individuels sont aléatoires et on préférera l'appellation ``modèles à effets individuels" aux appellations ``modèles à effets fixes" ou ``modèles à effets aléatoires" qui tendent à obscurcir les différences.

Il y a de nombreuses extensions à ce modèle. On peut d'abord considérer que les effets individuels n'affectent pas seulement les constantes de l'équation d'intérêt mais aussi les ``pentes", c'est-à-dire les coefficients des variables explicatives. On écrira par exemple dans le cas d'un seul régresseur:

$$y_{it} = \beta_i x_{it} + \alpha_i + u_{it}$$

où β_i est un coefficient aléatoire (voir Hsiao, 1985). Dans ce cas, on complétera le modèle par la donnée de :

$$\beta_i = \beta + \zeta_i$$

où $E(\zeta_i \mid x_i) = 0$ par exemple, pour réécrire :

$$\begin{aligned} y_{it} &= \beta x_{it} + \alpha_i + \zeta_i x_{it} + u_{it} \\ &= \beta x_{it} + \alpha_i + \tilde{u}_{it} \end{aligned}$$

On retrouve ainsi le modèle précédent à la différence près que les perturbations individuelles α_i et les perturbations individuelles-temporelles $\tilde{u}_{it}$ sont maintenant hétéroscédastiques. En effet, l'hypothèse H_1 est vérifiée puisque :

$$\begin{aligned} E(x_i' \tilde{u}_{it}) &= E(x_i'(\zeta_i x_{it} + u_{it})) \\ &= E(x_i' E(\zeta_i \mid x_i) x_{it}) + E(x_i' u_{it}) = 0 \end{aligned}$$

Or on sait que l'hétéroscédasticité n'affecte pas les propriétés de convergence des estimateurs dans les modèles linéaires. Ils perdent néanmoins leurs propriétés d'optimalité et il faut prendre garde au calcul de leur matrice de variance-covariance. Mais les techniques développées dans cette section restent conceptuellement valides (Arellano, 1993).

Un autre exemple d'hétéroscédasticité vient de la présence de certaines formes de corrélation sérielle dans les perturbations individuelles-temporelles u_{it} de ce modèle. On modifiera ainsi l'hypothèse H_2 pour supposer par exemple que les perturbations sont un processus moyenne-mobile d'ordre 1 :

$$u_{it} = \nu_{it} + \varphi \nu_{it-1}$$

Dans ce cas, la décomposition que nous avons faite plus haut de la matrice des moments du second ordre de $E(\varepsilon\varepsilon' \mid X)$ n'est plus valide même si les estimateurs que nous avons vu restent convergents. Pour calculer leurs écart-types, il faut prendre en compte cette hétéroscédasticité. Pour obtenir des estimateurs plus précis, on pourra aussi adopter la démarche générale des moindres carrés quasi-généralisés.

Il est néanmoins plus intéressant dans la plupart des applications économiques d'étudier des modèles dynamiques. Les modèles que nous avons vu sont des modèles dits statiques puisque toute dynamique est absente. On ne fait que relier une variable y_{it} à des variables x_{it} à une date donnée et l'hypothèse H_1 nous empêche de considérer des modèles où on aurait un feedback des perturbations u_{it} sur la trajectoire future des variables explicatives x_{it}. Dans les modèles que nous avons vu, l'estimateur de référence est l'estimateur de la covariance ou estimateur within. Il est en effet robuste à la présence de corrélation entre les effets individuels et les variables explicatives. Comme nous allons le voir ceci n'est vrai que s'il y a stricte exogénéité des x_{it} par rapport aux perturbations u_{it} à toutes les dates, c'est-à-dire l'hypothèse H_1 :

$$\forall t,t'; E(x_{it'}^{'} u_{it}) = 0$$

C'est cette hypothèse-là que nous allons relâcher maintenant en commençant par étudier le cas le plus parlant où on inclut l'endogène retardée, y_{it-1}, parmi les variables explicatives. L'hypothèse de stricte exogénéité devient alors intenable puisque par définition du modèle, l'endogène retardée est corrélée aux chocs à la période $t-1$.

2. Les modèles dynamiques

Dans les modèles dynamiques de données de panel, les propriétés des estimateurs que nous venons de voir sont fort différentes. On montre en particulier ici que l'estimateur de la covariance, pourtant le plus robuste dans les modèles statiques, n'est pas convergent en général (Nickell, 1981). On fait ressortir aussi une autre différence avec le modèle statique qui est le problème des conditions initiales du processus. Ceci reste une introduction brève à l'estimation de ces modèles qui ont connu et connaissent encore un essor important. Pour simplifier, on considérera de manière principale des modèles autorégressifs d'ordre 1 du type :

$$y_{it} = \alpha y_{it-1} + x_{it}\beta + \nu_i + u_{it}$$

où comme précédemment on supposera que des indicatrices temporelles font partie des variables x_{it}. On peut obtenir un tel modèle à partir d'un modèle économique à coûts d'ajustement comme celui que l'on a évoqué dans l'introduction. Ces décisions des entreprises quant à leur main d'oeuvre portent sur les embauches et licenciements en fonction des salaires et des effectifs de la période précédente compte-tenu des départs volontaires et en retraites. Dans un modèle simplifié, on aura donc

$$\Delta n_{it} = \gamma n_{it-1} + \beta w_{it} + \nu_i + u_{it}$$

où Δn_{it} représente la variation d'emploi dans l'entreprise i au temps t, n_{it-1} sont les effectifs de

la date précédente et w_{it}, les salaires. Ce modèle peut ainsi se réécrire sous la forme précédente comme :

$$n_{it} = (1+\gamma)n_{it-1} + \beta w_{it} + v_i + u_{it}$$

On notera au passage que le cas intéressant est celui pour lequel $\gamma \neq 0$ puisque dans le cas où $\gamma = 0$, le modèle en premières différences est un modèle statique pour lequel les procédures précédentes s'appliquent.

On peut aussi obtenir un modèle dynamique à partir d'un modèle statique quand les perturbations ont une structure autorégressive. Supposons par exemple que :

$$y_{it} = x_{it}\beta + v_i + u_{it}$$

où $u_{it} = \alpha u_{it-1} + \eta_{it}$ est un processus autorégressif d'ordre 1. Les chocs individuel-temporels, comme un choc sur la demande adressée à une entreprise, s'atténuent au cours du temps. Un calcul élémentaire donne :

$$y_{it} - \alpha y_{it-1} = (x_{it} - \alpha x_{it-1})\beta + (1-\alpha)v_i + \eta_{it}$$

et on obtient à nouveau un modèle dynamique[11].

Dans cette section, nous montrons d'abord que l'estimateur de la covariance est biaisé dans les panels courts et dans le cas où il n'y a pas de variables explicatives x_{it}. Puis nous décrivons le principe d'estimation convergente. Nous revenons ensuite sur le traitement des conditions initiales pour montrer comment dans certains cas on peut améliorer la précision des estimateurs. Nous finissons cette section en discutant du cas général.

2.1. Estimateur de la covariance

Pour simplifier l'argument qui est général, on considère le modèle autorégressif sans variables explicatives :

$$y_{it} = \alpha y_{it-1} + v_i + u_{it}$$

où $\alpha \neq 1$ puisque sinon le modèle peut se réécrire sous forme statique comme on l'a vu plus haut. Comme dans les sections précédentes, on supposera que l'échantillon est cylindré et les observations pour tout $i = 1,...,N$ et $t = 0,...,T$ sont disponibles[12]. On peut toujours développer l'expression de y_{it} en fonction de son passé en remplaçant de façon itérative, les variables retardées par leur expression donnée par le modèle ci-dessus:

(EQ_P) :

$$y_{it} = \alpha^t y_{i0} + \frac{1-\alpha^t}{1-\alpha} v_i + u_{it} + \alpha u_{it-1} + ... + \alpha^{t-1} u_{i1}$$

C'est cette expression dont il faut se souvenir pour retrouver les résultats qui suivent dans ce type

[11]On remarquera que cette équation diffère de la précédente par l'inclusion des variables explicatives à la période $t-1$ et t et par le lien entre leurs coefficients. D'abord, les deux modèles ne peuvent se distinguer que si les variables x_{it} varient au cours du temps. Ensuite, on n'a jamais précisé la nature des variables x_{it} qui pourraient inclure des variables décalées dans le temps dans la première version du modèle. D'autre part, la contrainte entre coefficients peut être prise en compte en utilisant des moindres carrés sous contrainte.

[12]Puisque le modèle est dynamique, il faut en effet une observation supplémentaire dans la dimension temporelle, y_{i0}, pour obtenir le même nombre d'équations que dans la partie précédente (T).

de modèle dynamique. A partir de cette expression, on voit que l'on caractérise complètement la loi des variables dépendantes en fonction des lois de perturbations u_{it}, des effets individuels, v_i, et de la condition initiale du processus y_{i0}. Pour simplifier, on fera aussi l'hypothèse que la série y_{it} est asymptotiquement stationnaire[13] et donc que $|\alpha|<1$. On écarte ainsi les phénomènes "explosifs", la variance augmentant avec le temps par exemple. Les études de telles séries sont encore rares même si de nombreuses procédures de tests de racine unité ($\alpha=1$) dans les données de panel ont été proposées pendant les années 90. L'extension aux séries telles que $|\alpha|\geq 1$ des résultats que nous allons voir sont possibles -- et faciles quand T est fixe puisque les théorèmes asymptotiques standard s'appliquent -- mais cela nécessite des ajustements qui nuisent au confort de lecture.

Précisons maintenant les hypothèses sur ces lois qui sont habituelles dans les modèles dynamiques et qui copient partiellement, en les adaptant, les hypothèses posées dans le modèle statique. On supposera que les erreurs u_{it} sont centrées et non corrélées à la valeur initiale du processus :

$$H_1^{'} \ : \ Eu_{it}=0, E(u_{it}y_{i0})=0$$

qu'elles sont homoscédastiques et ne sont pas corrélées au cours du temps ou entre individus (H_2). On suppose aussi que les effets individuels v_i sont centrés, orthogonaux aux chocs u_{it}, indépendants entre individus et homoscédastiques ($H_3^{'}$). Ainsi, il ne restera qu'à spécifier les relations entre effets individuels, v_i, et condition initiale y_{i0} pour caractériser complètement la loi de la variable dépendante. Nous n'en avons pas besoin tout de suite pour examiner le comportement de l'estimateur de la covariance.

Il s'obtient en estimant la régression entre écarts des variables à leur moyenne individuelle :

$$y_{it}-y_{i.}=\alpha(y_{it-1}-y_{i.}^0)+(u_{it}-u_{i.})$$

où il faut faire attention au fait que le modèle est dynamique pour définir :

$$y_{i.}=\frac{1}{T}\sum_{t=1}^{T}y_{it} \quad y_{i.}^0=\frac{1}{T}\sum_{t=0}^{T-1}y_{it}$$

$$u_{i.}=\frac{1}{T}\sum_{t=1}^{T}u_{it}$$

Puisqu'il n'y a qu'une variable explicative, l'estimateur de la covariance est l'estimateur d'une régression simple donc :

[13]Tous les moments, Ey_{it}, $E(y_{it}y_{it-k})$ admettent des limites bornées quand t tend vers l'infini. Ceci ne veut pas dire que y_{it} est stationnaire puisque sans hypothèses supplémentaires sur y_{i0} (voir *infra*), $E(y_{it}y_{it-k})$ par exemple dépend de t.

$$\hat{\alpha} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - y_{i.})(y_{it-1} - y_{i.}^{0})}{\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it-1} - y_{i.}^{0}\right)^2}$$

$$= \alpha + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it} - u_{i.})(y_{it-1} - y_{i.}^{0})}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it-1} - y_{i.}^{0}\right)^2}$$

$$= \alpha + \frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T} u_{it} y_{it-1} - T u_{i.} y_{i.}^{0}\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it-1} - y_{i.}^{0}\right)^2}$$

où la première égalité s'obtient en utilisant l'équation donnée par le modèle et la deuxième équation s'obtient en utilisant les deux écritures d'une covariance empirique.
On voit donc que montrer que cet estimateur est biaisé asymptotiquement revient à montrer que la limite en probabilité de la fraction est un scalaire $\alpha_b \neq 0$. On traite les termes de cette fraction les uns après les autres.

En premier lieu, les hypothèses du modèle permettent d'utiliser une loi des grands nombres pour montrer que le dénominateur du deuxième terme a pour propriété :

$$\underset{N\to\infty}{p\lim}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it-1} - y_{i.}^{0}\right)^2 = \sum_{t=1}^{T} E\left(y_{it-1} - y_{i.}^{0}\right)^2$$

$$= T\left(\frac{1}{T}\sum_{t=1}^{T} E\left(y_{it-1} - y_{i.}^{0}\right)^2\right) = T.A(T) > 0$$

qui est positive car la variable dépendante est variable au cours du temps. De plus, on peut montrer que l'hypothèse que la série y_{it} est asymptotiquement stationnaire fait que:

$$\lim_{T\to\infty} A(T) = A_{\infty} = \frac{\sigma_u^2}{1-\alpha^2}$$

En deuxième lieu, la première partie du numérateur peut être évaluée puisque sous les conditions vérifiées ici à cause de H_1' à H_3', de la loi des grands nombres, l'expression:

$$\underset{N\to\infty}{p\lim}\frac{1}{N}\sum_{i=1}^{N} u_{it} y_{it-1} = E u_{it} y_{it-1} = 0$$

est nulle. En effet, l'équation (EQ_P) montre que la variable y_{it-1} ne dépend que des chocs avant $t-1$, u_{it-1},..., u_{it-k}, de y_{i0} et de v_i qui ne sont pas corrélés avec u_{it} par les hypothèse H_1' à H_3.

Le biais ne peut venir que du dernier terme et donc :

$$\alpha_b \neq 0 \Leftrightarrow \underset{N\to\infty}{p\lim}\frac{1}{N}\sum_{i=1}^{N} u_{i.} y_{i.}^{0} = E u_{i.} y_{i.}^{0} \neq 0$$

Il suffit donc de calculer $Eu_{i.}y_{i.}^{0}$. Or en utilisant l'équation (EQ_P) on obtient :

$$Ty_{i.}^{0} = \sum_{t=1}^{T-1} \frac{1-\alpha^{T-t}}{1-\alpha} u_{it} + \left(\frac{T-1}{1-\alpha} - \alpha \frac{1-\alpha^{T-1}}{(1-\alpha)^2} \right) v_i + \frac{1-\alpha^{T}}{1-\alpha} y_{i0}$$

Donc :

$$T^2 E(u_{i.}y_{i.}) = \sum_{t=1}^{T-1} \frac{1-\alpha^{T-t}}{1-\alpha} \sigma_u^2 = \sigma_u^2 \left(\frac{T-1}{1-\alpha} + \frac{\alpha^{T}-\alpha}{(1-\alpha)^2} \right)$$

En reconsidérant les différents termes dans l'expression de l'estimateur de la covariance, le biais s'exprime alors comme :

$$\alpha_b = -\frac{\sigma_u^2}{T^2.A(T)} \left(\frac{T-1}{1-\alpha} + \frac{\alpha^{T}-\alpha}{(1-\alpha)^2} \right)$$

qui est strictement négatif pour tout $|\alpha| < 1$.

L'estimateur de la covariance est ainsi génériquement biaisé. Par exemple, si on suppose que le vrai modèle n'est pas dynamique ($\alpha = 0$), l'estimateur de la covariance du terme dynamique convergera pourtant vers la quantité $-\frac{\sigma_u^2(T-1)}{T^2 A(T)}$. Ce n'est que dans le cas **où le nombre de périodes devient très grand ($T \to \infty$) que l'estimateur de la covariance est convergent** puisque le biais est équivalent à $\frac{1}{T}$**.**

Le biais vient du fait que l'élimination de l'effet fixe par écart à la moyenne individuelle introduit de la corrélation entre perturbations, $u_{i.}$, et variables explicatives, $y_{i.}$. Les variables explicatives sont donc endogènes mais il est difficile de penser à des méthodes simples de variables instrumentales. Mais il est vrai aussi que le biais est calculable et on peut utiliser à une méthode en deux étapes pour recouvrer la vraie valeur de α (Sevestre et Trognon, 1983, Kiviet, 1995) et ces procédures semblent avoir des propriétés satisfaisantes quand N et T ont des valeurs semblables (panel de pays). On envisage ici une autre transformation qui permette d'éliminer l'effet individuel.

2.2. Principe d'estimation convergente

Au lieu de retirer la moyenne individuelle à toutes les observations, on utilise maintenant l'opérateur de première différence défini par :

$$\Delta y_{it} = y_{it} - y_{it-1} = \alpha(y_{it-1} - y_{it-2}) + (u_{it} - u_{it-1})$$

Néanmoins, les hypothèses du modèle linéaire ne sont pas vérifiées puisque par (EQ_P) la variable y_{it-1} est corrélée avec u_{it-1}. Il y a corrélation entre perturbations et variables explicatives et endogénéité des régresseurs.

On recherche ici des variables instrumentales. On pourrait avoir recours à des variables extérieures z_{it} qui respecteraient les deux conditions de validité de variables instrumentales. Premièrement, elles seraient corrélées avec les variables explicatives $(y_{it-1} - y_{it-2})$ et deuxièmement, elles ne seraient pas corrélées avec les perturbations.

On dispose pourtant déjà d'instruments potentiels que sont les variables dépendantes à d'autres dates. C'est une spécificité des données de panel que de pouvoir fournir des instruments de façon interne au modèle. En premier lieu, on utilise (EQ_P) pour montrer que :

$$E(y_{i\tau}(u_{it} - u_{it-1})) = \sigma_u^2(\alpha^{\tau-t}\mathbf{1}\{\tau \geq t\} - \alpha^{\tau+1-t}\mathbf{1}\{\tau \geq t-1\})$$

où $\mathbf{1}$ {.} est la fonction indicatrice. Les variables $y_{it-1},.,y_{iT}$ ne peuvent donc pas être des instruments valides puisque ces variables sont corrélées avec les perturbations de l'équation en premières différences. Par contre, les variables passées $y_{it-2},.,y_{i1}$ ne sont pas corrélées avec les perturbations.

De plus, on vérifie en utilisant (EQ_P) que de façon générique:

$$E(y_{it-1} - y_{it-2})y_{it-2} \neq 0$$

et :

$$E(y_{it-1} - y_{it-2})(y_{it-2} - y_{it-3}) \neq 0$$

Les méthodes d'estimation à variables instrumentales de l'équation en premières différences et qui sont basées sur les instruments que sont, soit la variable dépendante retardée de deux périodes, soit la première différence de la variable dépendante retardée de deux périodes, ont été proposées par Anderson et Hsiao (1982). Cependant, pour des raisons de précision, Arellano et Bond (1991) recommande l'utilisation de la première méthode. On dit qu'on instrumente les premières différences par les variables en niveau. On remarquera que, pour mettre en oeuvre cette procédure, le nombre de périodes est nécessairement au moins égal à trois ($T = 2$ et $t = 0,1,2$) puisque c'est seulement dans ce cas que la condition d'ordre est vérifié. Cette méthode s'apparente donc à une méthode de doubles moindres carrés.

On peut néanmoins gagner facilement en précision par rapport à cette méthode de doubles moindres carrés. En effet, on remarquera que l'équation en premières différences peut aussi s'écrire comme un système à $T-1$ équations constitué par les équations en différences pour chaque période de $t = 2$ à $t = T$:

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta u_{it}$$

où l'instrument utilisé est y_{it-2}. On remarquera ensuite que sous les hypothèses ci-dessus, il y a des corrélations entre les équations correspondant à des périodes différentes :

$$E(\Delta u_{it}\Delta u_{it-1}) = -\sigma_u^2$$

On gagnera donc en précision en utilisant une méthode de triples moindres carrés qui tient compte de ces corrélations.
Par extension, on pourrait utiliser d'autres instruments dans le passé. Toute variable y_{ik} dont le retard est au moins de deux périodes, $k \leq t-2$ est un instrument valide sous réserve qu'il ne soit pas redondant avec les autres instruments ("condition de rang"). Deux directions sont possibles: d'abord, choisir un nombre K de retards pour les variables instrumentales, $y_{it-2},.,y_{it-1-K}$ et construire l'estimateur à variables instrumentales correspondant. Cette procédure implique qu'il faille disposer de $K+2$ périodes d'observation au moins. Cela est donc très coûteux en termes d'information et donc de précision des estimateurs et ce n'est pas une direction recommandable.

Par contre, on peut aussi utiliser les méthodes généralisées de moment et leur congruence avec des méthodes de triples moindres carrés en utilisant des instruments dont la liste dépend de la période considérée. Par exemple, quant $t=2$, le seul instrument possible est y_{i0} mais pour $t>2$, on peut utiliser $y_{it-2},.,y_{i0}$. Cette méthode est donc asymptotiquement plus efficace que toutes les méthodes précédentes puisque le nombre d'équations de moment que l'on utilise est plus important (Arellano et Bond, 1991). Pour la résumer, on considère le système formé par les $T-1$ équations en premières différences. A la période t on utilise les instruments $y_{it-2},.,y_{i0}$ pour prédire Δy_{it-1} et pour construire l'estimateur des doubles moindres carrés de α.[14] Cet estimateur étant convergent, on peut donc construire les résidus d'estimation. On en déduit un estimateur de la matrice de variance-covariance des perturbations que l'on utilise pour finir par construire l'estimateur des triples moindres carrés en estimant le système dans son entier par moindres carrés quasi-généralisés. Sous les hypothèses d'homoscédasticité des résidus H_2, cette méthode est équivalente à une méthode généralisée de moments, qui s'appuie sur l'absence de corrélations entre Δu_{it} et $y_{it-2},.,y_{i0}$, comme l'a montré Wooldridge, 1996. Elle est donc efficace asymptotiquement (N grand et T petit) mais cela ne veut pas dire qu'à distance finie, l'estimateur des triples moindres carrés soit toujours de meilleur précision que l'estimateur des doubles moindres carrés[15] Les lois asymptotiques ne sont que des approximations. C'est le problème des instruments faibles que l'on ne fait qu'évoquer ici. On comparera donc systématiquement les deux estimateurs qui sont tous les deux convergents.

Il y a d'autres façons de ``gagner'' de la précision. D'abord, puisque l'estimation par variables instrumentales de l'équation en premières différences fait ``perdre'' au moins les deux premières périodes, il semble que, dans certains cas, des gains de précision importants puissent être obtenus en spécifiant de manière plus détaillée les hypothèses sur les conditions initiales. C'est en particulier le cas quand le coefficient α est proche de 1. Par exemple quand on essaie d'estimer des fonctions de production sur données de panel, les estimateurs que nous venons de voir sont en général peu précis et mal identifiés. Ensuite, si on prend à cœur les conditions données par H_1, H_2 et H_3, on peut en déduire des conditions de moment additionnelles (Ahn et Schmidt, 1995) mais elles ont le défaut de ne pas être linéaires en $\mathfrak{C}$. On étudie maintenant la spécification des conditions initiales qui, dans un cas que nous étudierons, permet de résoudre les deux problèmes en une seule fois.

2.3. Les conditions initiales

L'équation (EQ_P) se réécrit :

$$y_{it}=\alpha^t y_{i0}+v_i\frac{1-\alpha^t}{1-\alpha}+\varepsilon_{it}$$

où ε_{it} est un bruit autorégressif :

$$\varepsilon_{it}=\sum_{\tau=0}^{t-1}\alpha^\tau u_{it-\tau}=\alpha\varepsilon_{it-1}+u_{it}$$

Le processus est ainsi donné en fonction des valeurs initiales y_{i0} et d'un effet individuel v_i. On précise maintenant quelles sont les hypothèses stochastiques portant sur ces conditions initiales.

[14]C'est alors un système de régressions simultanées sous la contrainte que le même paramètre α apparaît dans toutes les équations.

[15]On aura pris soin de calculer la matrice de variance-covariance de cet estimateur des doubles moindres carrés de manière correcte car il y a deux difficultés ``inhabituelles'': il y a corrélation entre équations et le même paramètre α apparaît dans toutes les équations.

L'hypothèse la plus simple mais la moins crédible, serait de supposer que les variables y_{i0} sont des constantes fixes non aléatoires. Si on reste dans le cas où les variables sont aléatoires, ceci est équivalent à faire l'hypothèse que :

$$E(y_{i0}v_i) = 0$$

Dans ce cas, l'équation précédente pourrait être utilisée pour estimer α puisque le régresseur y_{i0} serait indépendant des erreurs v_i et ε_{it}. Ce serait un modèle statique où les coefficients des variables explicatives seraient variables au cours du temps (α^t) et les perturbations suivraient un processus autorégressif d'ordre 1. Les méthodes de la première partie pourraient alors être adaptées.

Mais l'hypothèse d'absence de corrélation entre condition initiale et effets individuels, est critiquable. Elle suppose que le processus a été commencé à la date 0 à partir d'un point exogène et sans rapport avec les effets individuels qui gouvernent la dynamique ultérieure du processus. Peu de données économiques satisfont *a priori* cette hypothèse. Par exemple, dans le cas des équations de salaire, cette hypothèse signifierait que les effets individuels intervenant dans la formation des salaires après la date 0 ne seraient pas corrélés avec le salaire initial ce qui n'est pas très convaincant.

On considère donc, dans le modèle général que la variable y_{i0} est indépendamment distribuée entre individus et homoscédastique mais qu'elle est corrélée avec les effets individuels v_i :

$$Ey_{i0}^2 = \sigma_0^2 < +\infty \quad Ey_{i0}v_i = \alpha_0$$

On pourrait alors estimer le modèle par la méthode du maximum de vraisemblance en spécifiant que les divers aléas suivent des lois normales (Bhargava et Sargan, 1983, Hsiao, 1986) ou du pseudo-maximum de vraisemblance (voir Gouriéroux et Monfort, 1989) puisque tous les moments du second ordre des variables sont maintenant spécifiés . Les estimateurs obtenus seront de meilleure précision que les estimateurs obtenus par des méthodes de variables instrumentales vus dans la section précédente. On utilise en effet maintenant l'information à toutes les dates au faible prix d'introduire les deux paramètres supplémentaires σ_0 et α_0 (Blundell et Smith, 1991).

Le coût de mise en oeuvre de cette méthode par les logiciels standards est plus important surtout quand on réintroduit des variables explicatives à côté de la variable endogène retardée. Il est plus intéressant, quand on a des problèmes de précision pour l'estimateur à variables instrumentales, d'explorer un cas particulier de telles conditions initiales que l'on obtient quand le processus est stationnaire. Comme on a supposé que le processus est asymptotiquement stationnaire ($|\alpha| < 1$), il suffit de considérer que le modèle dynamique a commencé suffisamment loin dans le passé ($t = -\infty$). Dans ce cas, on peut écrire en utilisant l'équation dynamique et $|\alpha| < 1$:

$$y_{i0} = \frac{v_i}{1-\alpha} + \sum_{\tau=0}^{-\infty} \alpha^\tau u_{i\tau} = \frac{v_i}{1-\alpha} + \varepsilon_{i0}$$

où ε_{i0} est un processus stationnaire autorégressif d'ordre 1 et de coefficient α . On remarquera que cette hypothèse fait peser des contraintes sur les paramètres précédents σ_0 et α_0. On obtient la variable, y_{it}, qui est donnée par :

$$y_{it} = \frac{v_i}{1-\alpha} + \varepsilon_{it}$$

qui est bien stationnaire au second ordre. C'est donc un modèle dynamique que l'on a réécrit comme un modèle statique (Blundell et Bond, 1998). On remarquera dans ce cas que la première

différence de la variable dépendante ne dépend pas des effets individuels :

$$\Delta y_{it} = \varepsilon_{it} - \varepsilon_{it-1}$$

Des retards de ces variables fournissent donc de bons candidats pour des instruments de l'équation de départ :

$$y_{it} = \alpha y_{it-1} + v_i + u_{it}$$

puisque :

$$E(\Delta y_{it-1} u_{it}) = 0$$

$$\begin{aligned} E(\Delta y_{it-1}.y_{it-1}) &= E(\Delta \varepsilon_{it-1}.\varepsilon_{it-1}) \\ &= (1-\alpha) E\varepsilon_{it-1}^2 \neq 0 \end{aligned}$$

Le principe de la méthode d'estimation proposée par Arellano et Bover (1995) est alors simple à expliquer[16]. On utilise en même temps deux équations, celle en niveau instrumentée par la première différence retardée et celle en première différence instrumentée par l'endogène retardée deux fois comme dans la section précédente. On obtient donc non seulement un système de $T-1$ équations comme dans la partie précédente mais un système à $2(T-1)$ équations. On conçoit que cette procédure soit plus efficace que la précédente et on montre que c'est surtout le cas quand α est proche de 1 (Blundell et Bond, 1998). On peut alors répéter la méthode évoquée dans la section précédente. On construit les prédicteurs des variables à droite de chaque équation en utilisant les instruments correspondants à chaque équation. On estime ce système comme un système de régressions simultanées en prenant garde au fait d'imposer la contrainte que le même paramètre α apparaît dans toutes les équations. En estimant les corrélations entre équations, on peut alors aussi construire l'estimateur des triples moindres carrés mais celui-ci n'est pas, cette fois-ci, équivalent à l'estimateur par la méthode généralisée des moments car les conditions de Wooldridge (1996) ne sont pas satisfaites.

2.4. Application aux modèles apparemment statiques

Muni de ces outils, on peut maintenant revenir aux modèles statiques du type :

$$y_{it} = x_{it}\beta + v_i + u_{it}$$

où on fera l'hypothèse que les variables x_{it} sont non seulement corrélées avec l'effet individuel ($Ex_{it}v_i \neq 0$) mais ne sont aussi que prédéterminées ou faiblement exogènes et non plus fortement exogènes :

$$\begin{cases} Ex_{it}u_{is} \neq 0 & \text{si } s < t \\ Ex_{it}u_{is} = 0 & \text{si } s \geq t \end{cases}$$

C'est le cas si les chocs contemporains sur la variable dépendante affectent aussi les variables explicatives dans le futur mais ne sont pas corrélés avec leur passé et présent. Il y a donc bien des effets de feed-back des chocs sur les variables dépendantes sur les variables explicatives. Dans une équation de salaire, l'expérience professionnelle dépendra par exemple des chocs passés sur les salaires. En effet, l'expérience professionnelle n'est que la somme des participations passées sur le marché du travail. Ces variables passées de participation sont sensibles au niveau des salaires si les agents arbitrent entre travail et non-travail en comparant leur salaire et leur coût d'opportunité.

[16]En fait, l'hypothèse de stationnarité implique ces résultats mais n'est pas nécessaire (Blundell et Bond, 1998). On peut en effet ajouter un bruit individuel à y_{i0} pourvu qu'il soit orthogonal à l'effet individuel.

L'estimateur de la covariance dans ce modèle apparemment statique est non convergent pour la même raison que dans les modèles dynamiques. Il y a corrélation entre régresseurs et erreurs puisque la technique de la covariance a pour défaut de moyenner les explicatives sur toute la période. La moyenne individuelle des variables explicatives est alors corrélée à la moyenne individuelle des chocs.

Le principe de la méthode est alors le même que dans la section précédente. La première idée d'une procédure convergente est comme précédemment de considérer le modèle en premières différences :

$$\Delta y_{it} = (\Delta x_{it}).\beta + \Delta u_{it}$$

Néanmoins, il y a corrélation entre régresseur et erreur si la corrélation entre x_{it} et u_{it-1} est non nulle comme on en a fait l'hypothèse. On considère donc que les instruments valides sont $x_{it-1},...x_{i1}$ et $y_{it-2},...y_{i1}$. On applique alors, par exemple, une méthode à variables instrumentales en retenant x_{it-1} et y_{it-2}, ou une méthode où les instruments différent entre les équations, $x_{it-1},...x_{i1}$ et $y_{it-2},...y_{i1}$. Des hypothèses de stationnarité comme dans la sous-section précédente permettent d'améliorer la précision des estimateurs.

2.5. Modèle dynamique général

Un tel modèle est donné par :

$$y_{it} = \alpha y_{it-1} + x_{it}\beta + v_i + u_{it}$$

où les variables explicatives sont corrélées aux effets individuels et sont faiblement exogènes. On peut aussi considérer que les perturbations u_{it} ont une structure de moyenne mobile, au moins d'ordre 1 :

$$u_{it} = \eta_{it} - \theta\eta_{it-1}$$

Il n'y a pas de recette toute faite pour l'estimation d'un tel modèle. On poursuit la recherche de spécification en utilisant les techniques classiques de tests de variables instrumentales . L'élimination de l'effet individuel se fait en général en considérant la première différence. Puis on recherche les instruments parmi les retards de la variable endogène et des variables explicatives. On peut aussi imposer des conditions de stationnarité.

On pourrait finalement considérer que les coefficients dépendent du temps (Holtz-Eakin, Newey et Rosen, 1988) ou que le coefficient de la variable retardée est un coefficient individuel (Pesaran et Smith, 1995). Dans ces derniers cas, les estimateurs usuels doivent être adaptés car ils ne sont pas généralement convergents. Mais ces exemples dépassent le cadre de cette note et on consultera Arellano et Honoré (2000) pour traiter de ces extensions.

3. Cylindrage des données et sélection

Les données de panel ne sont pas souvent cylindrées comme nous l'avons supposé jusqu'à maintenant. Certains individus sont enquêtés à certaines périodes seulement ou ils peuvent refuser de répondre à partir d'une certaine date. On dira alors qu'il y a attrition -- l'image de la guerre d'usure entre l'enquêteur et l'enquêté dans laquelle celui-ci a le dernier mot. Cela pose un certain nombre de problèmes. Le plus facile à traiter est celui de la précision des estimateurs. En cylindrant un fichier de données, on perd beaucoup d'information et donc de la précision dans les estimateurs. En adaptant les méthodes à des données non cylindrées, on accroît donc la précision

des méthodes d'estimation et la puissance des procédures de tests. Un autre problème est celui de la sélection des échantillons. L'attrition des individus est parfois en rapport avec le phénomène que l'on veut étudier. Par exemple, comme l'enquête Emploi est une enquête par logements, on perdra systématiquement d'une enquête à l'autre tous les individus qui déménagent. Or la sélection peut être en rapport avec le phénomène que l'on veut étudier, des épisodes de chômage par exemple. On dira, dans ce cas, que la sélection est endogène par opposition au cas où les sorties de l'enquête sont sans rapport avec le phénomène à étudier (mais peut être en rapport avec les explicatives) et donc aléatoires ou exogènes.

On traite d'abord de la sélection exogène en montrant comment l'estimateur de la covariance, dans un modèle statique, peut être adapté au cas des données non cylindrées. L'extension au cas dynamique se fait de la même façon. Puis on montre le problème que pose la sélection endogène en ne le résolvant que dans un cas particulier. La méthode générale nécessite en effet le traitement de données de panel de modèles à variables discrètes que ne peut couvrir cette note et qui reste d'ailleurs un sujet qui ne fait que commencer à être exploré dans la littérature économétrique.

3.1. Sélection exogène

Comme dans la section 2, on veut estimer le modèle à hétérogénéité inobservable corrélée :

$$y_{it} = x_{it}\beta + v_i + u_{it}$$

On suppose que l'on dispose d'observations (y_{it}, x_{it}) pour un individu i au temps t si une variable p_{it} prend la valeur 1. Sinon $p_{it} = 0$ et (y_{it}, x_{it}) n'est pas observé[17]. La variable discrète p_{it} décrit donc le schéma d'observation ou la sélection de l'échantillon.

3.1.1. Sélection strictement exogène ou échantillonnage aléatoire

On considère que $z_{it} = (y_{it}, x_{it})$ et $z_i = (z_{i1},.,z_{iT})$. On dira que la sélection est strictement exogène ou ignorable au second ordre, si et seulement si :

$$E\left(z_i' z_i \mid \{p_{it}\}_{t=1,.,T}\right) = E(z_i' z_i)$$

Les moments du second ordre des variables sont donc les mêmes dans l'échantillon observable et dans la population totale[18]. On peut donc estimer, à partir de l'échantillon observé, les paramètres de population puisque les paramètres des modèles linéaires ne dépendent que des moments du second ordre des variables.

Dans ce cas, l'estimateur de la covariance garde ses propriétés. En effet, si on suppose que les variables x_{it} sont non corrélées aux perturbations dans la population, elles ne le seront pas non plus dans l'échantillon observé :

$$E\left[x_{it'}'(u_{it} - u_{i.})\right] = 0 \Rightarrow E(x_{it'}'(u_{it} - u_{i.}) \mid p_{it} = p_{it'} = 1) = 0$$

puisque ces espérances peuvent se construire linéairement à partir de $E(z_i' z_i)$ et où on construit

[17]On pourra compliquer le problème à loisir en supposant que certaines variables sont observées d'autres non, et en multipliant le nombre de ces indicateurs.

[18]On peut considérer une hypothèse plus forte qui est l'exogénéité en loi. La loi conditionnelle de z_i ne dépendra pas de la sélection p_i. Mais l'hypothèse faite ici est suffisante tant que le modèle reste linéaire.

les moyennes individuelles des variables $z_{it} = (y_{it}, x_{it})$ comme :

$$z_{i.} = \frac{1}{\sum_{t=1}^{T} p_{it}} \sum_{t=1}^{T} p_{it} z_{it}$$

où les valeurs inobservées ne sont pas prises en compte. On construit alors l'estimateur de la covariance de la manière habituelle en régressant les écarts à la moyenne individuelle de la variable dépendante sur les écarts aux moyennes individuelles des variables indépendantes pour obtenir :

$$\hat{\beta}_{\text{cov}} = \left(\sum_{i=1,t=1}^{n,T} p_{it} (x_{it} - x_{i.})'(x_{it} - x_{i.}) \right)^{-1} \sum_{i=1,t=1}^{n,T} p_{it} (x_{it} - x_{i.})'(y_{it} - y_{i.})$$

En premier lieu, les individus pour lesquels on ne dispose que d'une observation ne contribuent d'aucune façon à la construction de cet estimateur et peuvent donc être écartés. Comme, par construction :

$$y_{it} - y_{i.} = (x_{it} - x_{i.})\beta + (u_{it} - u_{i.}) \text{ si } p_{it} = 1$$

et comme il n'y a pas de corrélation entre $(x_{it} - x_{i.})$ et $(u_{it} - u_{i.})$ quand $p_{it} = 1$ grâce à l'ignorabilité de la sélection, cet estimateur est convergent. Si les perturbations sont homoscédastiques, sa matrice de variance covariance est :

$$\sigma^2 \left(\sum_{i=1,t=1}^{n,T} p_{it} (x_{it} - x_{i.})'(x_{it} - x_{i.}) \right)^{-1}$$

La matrice de variance-covariance asymptotique quand $n \to \infty$ est donc :

$$\sigma^2 \left(\sum_{t=1}^{T} E(p_{it} (x_{it} - x_{i.})'(x_{it} - x_{i.}) \right)^{-1}$$

Il est facile de se rendre compte que cette matrice de variance-covariance est plus petite -- dans le sens des matrices définies positives -- que la matrice de variance covariance de l'estimateur de la covariance obtenu à partir de l'échantillon cylindré :

$$\sigma^2 \left(\sum_{t=1}^{T} E(\left[\prod_{t=1}^{T} p_{it} \right] (x_{it} - x_{i.})'(x_{it} - x_{i.}) \right)^{-1}$$

puisque $\prod_{t=1}^{T} p_{it} \leq p_{it}$. On notera finalement que **l'estimateur de la covariance sur données incomplètes n'est pas forcément celui de variance minimale** puisque les propriétés d'orthogonalité entre dimension inter-individuelle et intra-individuelle sont perdues (Wansbeek et Kapteyn, 1989).

3.1.2. Sélection strictement exogène conditionnellement aux variables indépendantes

On dira que la sélection est strictement exogène conditionnellement aux variables explicatives si et seulement si :

$$E\left(y_i \mid \{x_{it}, p_{it}\}_{t=1,.,T}\right) = E\left(y_i \mid \{x_{it}\}_{t=1,.,T}\right)$$

On pourra aussi dire que cette sélection est ignorable pour $E\left(y_i \mid \{x_{it}\}_{t=1,..,T}\right)$. L'espérance conditionnelle des variables dépendantes reste la même dans l'échantillon observable et dans la population totale même si la sélection de l'échantillon peut être fonction des variables x. On ne demande donc plus une propriété de représentativité de l'échantillon et on peut donc avoir des échantillons sous ou sur-pondérés pour certaines sous-populations à condition que ce mode d'échantillonnage soit basé sur des variables exogènes.

Pour estimer un tel modèle, il faut renforcer les hypothèses du modèle de départ. On supposera maintenant que :

$$E\left(u_{it} \mid \{x_{it}\}_{t=1,..,T}\right) = 0$$

Cette hypothèse est beaucoup plus forte que la non corrélation. C'est en quelque sorte le prix à payer pour traiter la sélection exogène conditionnellement aux variables explicatives. Dans ce cas, on a toujours :

$$\begin{aligned} E(x_{it'}^{'} u_{it} \mid p_{it} = p_{it'} = 1) &= E(x_{it'}^{'} E\left(u_{it} \mid p_{it} = p_{it'} = 1, x_{it}\right)) \\ &= E(x_{it'}^{'} E\left(u_{it} \mid x_{it}\right)) = 0 \end{aligned}$$

On retrouve ainsi l'hypothèse sous jacente au modèle standard dans l'échantillon séléctionné et les méthodes précédentes s'appliquent en prenant garde à d'éventuels problèmes d'hétéroscédasticité qui disparaissent si on fait l'hypothèse additionnelle :

$$E\left(y_i^{'} y_i \mid \{x_{it}, p_{it}\}_{t=1,..,T}\right) = E\left(y_i^{'} y_i \mid \{x_{it}\}_{t=1,..,T}\right) = \Omega$$

On pourra donc là aussi utiliser les méthodes précédentes.

3.2. Sélection endogène

On se place maintenant dans le cas où la sélection est endogène :

$$E(v_i + u_{it} \mid \{x_{it}, p_{it}\}_{t=1,..,T}) \neq E(v_i + u_{it} \mid \{x_{it}\}_{t=1,..,T})$$

Dans ce cas général, ce sont les procédures habituelles employées dans le cas d'échantillon sélectionné et dites à la Heckman qui doivent être étendues au cas des panels. Ce sont des procédures qui ne commencent à être explorées que depuis peu (Arellano et Honoré, 2000, pour un survey). Elles demandent des techniques plus sophistiquées faisant appel à la modélisation de variables discrètes ou à variation limitée en données de panel qui ne sont pas l'objet de cette note.

Il y a pourtant des cas particuliers qui, faute de mieux, peuvent être exploités. Supposons ainsi que le modèle de sélection est donné par le modèle latent suivant :

$$p_{it} = 1 \text{ ssi } \zeta_i + \kappa_{it} > 0$$

et que les termes κ_{it} sont indépendants au cours du temps. Supposons aussi que cette sélection est telle que la moyenne conditionnelle des chocs u_{it} dans l'équation d'intérêt est constante au cours du temps pour chaque individu :

$$\forall i; \;\; E(u_{it} \mid \{x_{it}\}_{t=1,..,T}, p_{it} = 1, \zeta_i, v_i) = \varphi(\zeta_i)$$

Ceci veut dire que l'endogénéité de la sélection ne vient que d'un terme individuel ζ_i qui ne varie pas au cours du temps et ne dépend donc pas des variables explicatives x_{it}. C'est une hypothèse forte puisqu'on peut penser que la variable p_{it} est affectée par x_{it} et non pas seulement par un

terme individuel comme $x_{i.}$, mais ce cadre permet néanmoins de prendre en compte une corrélation constante entre u_{it} et κ_{it}. Sous cette hypothèse, on aura donc pour chaque individu:

$$E(y_{it} \mid \{x_{it}\}_{t=1,..,T}, p_{it} = 1, \zeta_i, v_i) = x_{it}\beta + v_i + \varphi(\zeta_i)$$

Alors, on peut réécrire le modèle original sur l'échantillon sélectionné comme :

$$y_{it} = x_{it}\beta + \bar{v}_i + \eta_{it} \text{ si } p_{it} = 1$$

en définissant :

$$\eta_{it} = y_{it} - E(y_{it} \mid \{x_{it}\}_{t=1,..,T}, p_{it} = 1, \zeta_i, v_i)$$

qui satisfait donc à l'hypothèse :

$$E(\eta_{it} \mid \{x_{it}\}_{t=1,..,T}, p_{it} = 1, \zeta_i) = 0$$

Dans ce cas, la sélection a été ``absorbée'' par l'effet individuel v_i et les méthodes usuelles s'appliquent, en prenant garde à l'hétéroscédasticité éventuelle.

4. Pseudo-panels

Les techniques développées pour les données de panel peuvent s'appliquer dans le cas où on ne suit pas les mêmes individus au cours du temps mais où on dispose d'enquêtes répétées sur la même population au cours du temps. L'exemple le plus célèbre est l'enquête de Budget des Ménages en Angleterre (FES) qui est répétée annuellement depuis 1972 et qui a été l'objet de nombreux articles. L'inconvénient de la manipulation d'un tel corps de données est de ne pas pouvoir suivre les mêmes individus. Mais cet inconvénient a pour contrepartie un avantage qui est de pouvoir, par construction, négliger l'attrition puisque les individus sont continûment ré-échantillonnés.

L'idée de cette technique est de constituer des cohortes d'individus homogènes quant à certaines caractéristiques. Les données agrégées par cohortes sont alors similaires à des données de panel. Dans la population, on pourra, par exemple, constituer des groupes par année de naissance, sexe et niveau d'éducation. L'individu au sens statistique de ces pseudo-panels sera une cohorte d'éducation et de sexe donnés et dont l'âge $a + k$ croit avec les dates d'enquêtes $t + k$. Les variables dépendantes et indépendantes seront les moyennes des variables d'intérêt dans chaque cohorte à toutes les dates d'enquêtes.

Il y a deux points importants. D'abord, il faut prendre garde à la construction des cohortes. Construire des cohortes par statut d'emploi n'aurait guère d'intérêt puisque la composition des cohortes changerait au cours du temps. Cela donnerait lieu à **des biais de composition**. Ensuite, les moyennes des variables par cohortes sont des estimateurs avec erreurs de mesure, des espérances de chaque cohorte puisqu'il y a systématiquement ré-échantillonnage. Ces erreurs de mesure ne sont négligeables que si la taille de la cohorte est grande. Il y a donc ainsi arbitrage à exercer entre tailles des cohortes et nombre de cohortes que l'on étudiera plus bas.

4.1. Le modèle à erreurs de mesure

Ce modèle est dû à Deaton (1985). On considère le modèle individuel statique à effets fixes[19] où pour simplifier les variables (y_{it}, x_{it}) sont supposées indépendamment distribuées entre individus et :

$$y_{it} = x_{it}\beta + \alpha_i + u_{it}$$

[19]L'extension à des modèles dynamiques suit le même principe (Collado, 1997).

sous les hypothèses H_1 et H_2 et donc sous les mêmes notations que dans les sections précédentes :

$$Ex_i' u_i = 0 \quad Eu_{it} = 0$$

$$E\left(u_i u_i' \mid x_i\right) = \sigma_u^2 I$$

On fera l'hypothèse que l'on observe l'individu i à la période t. Le cas intéressant à traiter est le modèle à effets fixes où on ne fait pas l'hypothèse d'absence de corrélation entre variables explicatives et effet individuel. En effet, si l'effet individuel est aléatoire et si cette corrélation est nulle, la perturbation totale, $\alpha_i + u_{it}$, n'est pas corrélée aux variables explicatives et on peut estimer β par moindres carrés ordinaires en rassemblant les diverses coupes transversales. On ne peut pas identifier la variance de l'effet individuel et l'estimateur de β est moins efficace que si l'on disposait de données de panel mais ce sont des conséquences directes de l'information moindre apportée par des pseudo-panels. Par contre, si les effets individuels et variables explicatives sont corrélées, ce n'est plus seulement un problème de précision mais aussi un problème de biais.

Pour résoudre ce problème, on considère une partition de la population en cohortes, $c \in \{1,.,C\}$. C'est une partition qui vaut pour toutes les enquêtes et certaines cohortes - les plus jeunes et les plus âgées -- ne seront observées que pour une période plus courte que la période totale. Nous sommes donc généralement dans le cas de panels non cylindrés mais nous pouvons toujours cylindrer l'échantillon pour simplifier l'argument. On supposera aussi pour simplifier les calculs que le nombre d'individus de la cohorte c au temps t est constant et égal à n. Le nombre total d'individus enquêtés est donc $N = cn$ à T périodes. Même si le nombre de périodes est supposé fixe, les propriétés asymptotiques peuvent être obtenues en faisant tendre soit la taille de la cohorte n, soit le nombre de cohortes, C, vers l'infini. C'est ainsi que l'on pose l'arbitrage entre taille de cohortes et nombre de cohortes. Par exemple, pour une enquête de 5000 individus, des cohortes de taille 200 donne un nombre de cohortes relativement petit et égal à 25.

Pour se ramener à un modèle de panel, on notera pour les variables $z_{it} = (y_{it}, x_{it})$:

$$\hat{z}_{ct} = \frac{1}{n} \sum_{i \in c} z_{it}$$

Ces moyennes empiriques sont les contreparties observables des quantités inobservables :

$$z_{ct}^* = E\left(z_{it} \mid i \in c, t\right)$$

L'idée de Deaton est d'agréger le modèle individuel par cohorte en utilisant les quantités inobservables :

$$y_{ct}^* = x_{ct}^* \beta + \alpha_c + u_{ct}^*$$

On notera puisque l'agrégation ne fait pas nécessairement disparaître les chocs temporels $E\left(u_{it} \mid i \in C, t\right)$ qui peuvent être propres à une cohorte. On a fait l'hypothèse implicite que la quantité :

$$E\left(\alpha_i \mid i \in C, t\right) = \alpha_c$$

est constante au cours du temps. La composition de la cohorte est donc supposée stationnaire c'est à dire que les flux démographiques entrants et sortants de la cohorte sont supposées exogènes ce qui pour des cohortes rassemblant des générations est approximativement vrai tant que leur âge n'est pas trop élevé.

Le modèle (mag) ne peut être estimé puisque les quantités y_{ct}^* et x_{ct}^* ne sont pas observables. On dispose néanmoins de mesures de ces quantités qui sont $\hat{y}_{ct}$ et $\hat{x}_{ct}$. On écrit :

$$\begin{cases} \hat{y}_{ct} = y_{ct}^* + \varepsilon_{ct} \\ \hat{x}_{ct} = x_{ct}^* + \eta_{ct} \end{cases}$$

et les erreurs de mesure sont centrées :

$$E(\varepsilon_{ct}) = E(\eta_{ct}) = 0$$

Si on fait l'hypothèse d'homoscédasticité, la variabilité intra-cohortes est la même pour toutes les cohortes. On obtient donc les variances des erreurs de mesure :

$$V\begin{pmatrix} \varepsilon_{ct} \\ \eta_{ct} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \sigma_{yy} & \sigma_{yx}' \\ \sigma_{yx} & \sigma_{xx} \end{pmatrix}$$

où σ_{yy} dénote la variabilité intracohortes de la variable y. Grâce aux hypothèses du modèle, on peut facilement construire des estimateurs sans biais des éléments de la matrice de variance-covariance, par exemple :

$$\hat{\sigma}_{xx} = \frac{1}{CT}\sum_{c,t}\left[\frac{1}{n-1}\sum_{i\in c,t}(x_{it} - \hat{x}_{ct})^2\right]$$

Le modèle peut alors se réécrire comme :

$$\begin{aligned} \hat{y}_{ct} &= \hat{x}_{ct}\beta + \alpha_c + u_{ct}^* + \varepsilon_{ct} - \eta_{ct}\beta \\ &= \hat{x}_{ct}\beta + \alpha_c + \tilde{u}_{ct} \end{aligned}$$

qui est un modèle à effets fixes mais où la corrélation entre les régresseurs $\hat{x}_{ct}$ et la perturbation $\tilde{u}_{ct}$ à la même date est non nulle :

$$E(\hat{x}_{ct}'\tilde{u}_{ct}) = \frac{1}{n}(\sigma_{xy} - \sigma_{xx}\beta)$$

en utilisant les notations ci-dessus. Or, on sait que dans les modèles à effets fixes, ce modèle peut se réécrire dans les dimensions between et within :

$$\hat{y}_{ct} - \bar{y}_c = (\hat{x}_{ct} - \bar{x}_c)\beta + \tilde{u}_{ct} - \bar{u}_c$$

$$\bar{y}_c = \bar{x}_c\beta + \alpha_c + \bar{u}_c = \tilde{\alpha}_c + \bar{u}_c$$

où $\bar{y}_c$, $\bar{x}_c$ désignent les moyennes par cohortes des variables, par exemple :

$$\bar{x}_c = \frac{1}{T}\sum_{t=1}^{T}\hat{x}_{ct}$$

et où les effets des variables explicatives dans la dimension between ont été absorbés dans l'effet cohorte $\tilde{\alpha}_c$, en l'absence d'a priori sur les corrélations entre $\bar{x}_c$ et α_c. L'estimateur obtenu dans la dimension within est donc efficace mais il faut prendre garde aux erreurs de mesure puisque :

$$E((\hat{x}_{ct} - \bar{x}_c)'(\tilde{u}_{ct} - \bar{u}_c) = \frac{T-1}{T}\frac{1}{n}(\sigma_{xy} - \sigma_{xx}\beta)$$

où le terme $\frac{T-1}{T}$ vient du fait que l'on considère la dimension within.

4.2. Estimation

Le modèle est donc un modèle à erreur de mesure sur les variables dans lequel la matrice de variance-covariance de ces erreurs est connue. Soient les différents moments empiriques :

$$\hat{\Omega} = \frac{1}{CT}\sum_{c,t}(\hat{x}_{ct} - \bar{x}_c)'(\hat{x}_{ct} - \bar{x}_c)$$

$$\hat{\omega} = \frac{1}{CT}\sum_{c,t}(\hat{x}_{ct} - \bar{x}_c)'(\hat{y}_{ct} - \bar{y}_c)$$

Pour s'abstraire du problème d'erreur de mesure, supposons d'abord que le nombre d'observations par cohortes est tel que $n \to \infty$. Les erreurs d'échantillonnage et de mesure disparaissent et on a :

$$E(\hat{x}_{ct} - \bar{x}_c)'(\tilde{u}_{ct} - \bar{u}_c) = 0$$

On obtient donc un modèle de panel usuel et l'estimateur within est donné par :

$$\hat{\beta}_W = \hat{\Omega}^{-1}\hat{\omega}$$

Cet estimateur hérite des propriétés de l'estimateur de la covariance et il est donc sans biais et convergent quand le nombre de cohortes C ou le nombre de périodes T tend vers l'infini.

On peut maintenant supposer au contraire que n est fixe et il faut donc traiter du problème d'erreurs de mesure. Or le paramètre β dans le modèle projeté dans la dimension within est donné par la solution de :

$$\begin{aligned} E(\hat{x}_{ct} - \bar{x}_c)'(\hat{y}_{ct} - \bar{y}_c) &= E(\hat{x}_{ct} - \bar{x}_c)'(\hat{x}_{ct} - \bar{x}_c)\beta + E(\hat{x}_{ct} - \bar{x}_c)'(\tilde{u}_{ct} - \bar{u}_c) \\ &= E(\hat{x}_{ct} - \bar{x}_c)'(\hat{x}_{ct} - \bar{x}_c)\beta + \frac{T-1}{T}\frac{1}{n}(\sigma_{xy} - \sigma_{xx}\beta) \end{aligned}$$

et β est donc égal à :

$$\left(E(\hat{x}_{ct} - \bar{x}_c)'(\hat{x}_{ct} - \bar{x}_c) - \frac{T-1}{T}\frac{1}{n}\sigma_{xx}\right)^{-1}\left(E(\hat{x}_{ct} - \bar{x}_c)'(\hat{y}_{ct} - \bar{y}_c) - \frac{T-1}{T}\frac{1}{n}\sigma_{xy}\right)$$

Un estimateur convergent et efficace, quand C ou T tend vers l'infini, est :

$$\hat{\beta}_{EM} = \left(\hat{\Omega} - \frac{T-1}{T}\frac{1}{n}\hat{\sigma}_{xx}\right)^{-1}\left(\hat{\omega} - \frac{T-1}{T}\frac{1}{n}\hat{\sigma}_{xy}\right)$$

qui cette fois-ci prend en compte le problème d'erreur de mesure. Les deux estimateurs $\hat{\beta}_{EM}$ et $\hat{\beta}_W$ coïncident quand le nombre d'individus par cohorte, n, tend vers l'infini mais en pratique il semble qu'il faille un nombre n relativement grand ($n = 200$ est suggéré par Verbeek et Nijman, 1993).

On peut maintenant revenir sur l'arbitrage entre taille des cohortes, n, et nombre de cohortes, C, quand on suppose que la dimension temporelle est fixée T. La précision de l'estimateur $\hat{\beta}_{EM}$ dépend de la magnitude de l'estimateur du terme :

$$E(\hat{x}_{ct} - \bar{x}_c)'(\hat{x}_{ct} - \bar{x}_c) - \frac{T-1}{T}\frac{1}{n}\sigma_{xx} = E(x^*_{ct} - \bar{x}^*_c)'(x^*_{ct} - \bar{x}^*_c)$$

qui est la variance intercohortes des variables explicatives. On conçoit donc bien que quand les cohortes sont trop petites, cet estimateur va être dominé par les erreurs de mesure et peut avoir

un comportement très erratique. Mais il n'y a pas de recette toute faite puisque son comportement dépend de la quantité inconnue qui est la variance intercohortes des variables explicatives donnée par l'expression précédente. Par exemple, si des variables de secteur ont été inclues dans des équations de salaire et si la structure par âge des secteurs dont relativement similaires, il sera difficile d'obtenir des estimateurs précis des coefficients des secteurs.

5. Application empirique[20]

Pour notre application, nous avons exploité 10% du fichier apparié DADS-EDP (cf. Annexe 1), ce qui correspond à un échantillonnage aléatoire uniforme de la population salariée au 2500ème environ. Nous nous limitons également à la période postérieure à 1991 principalement pour éviter des problèmes liés à l'absence des années 1981, 1983 et 1990 dans le panel. Enfin, nous limitons l'étude aux seuls hommes ayant travaillé à temps complet.

5.1. Sélection et nettoyage des données

Le fichier contient initialement 41 913 observations correspondant à 6 707 hommes ayant occupé un emploi entre 1991 et 1998. Chaque observation est définie par un identifiant individuel (le NIR) et l'année de l'observation. Au cours de l'année, on dispose ainsi de la rémunération versée par l'entreprise au salarié, du type de l'emploi occupé(temps complet ou partiel), du nombre de jours rémunérés, du secteur et de la taille de l'entreprise employeur, du lieu de travail de l'individu, de la date de début et de fin d'emploi (éventuellement censuré à droite car on ne peut pas connaître la date de fin d'un emploi toujours occupé) et du diplôme grâce à l'EDP. En fait, nous allons nous restreindre par la suite à l'étude des salariés pour lesquels on connaît le diplôme. Or, les derniers renseignements concernant le diplôme inclus dans l'EDP datent de 1990 (année du recensement)[21].

Sur les 6 707 individus échantillonnés, 1132 ont un diplôme manquant. Dans l'EDP, un individu peut aussi ne pas avoir fini ses études au moment où il est interrogé sur son diplôme. Pour éviter ce problème, nous excluons de notre analyse tous les salariés âgés de moins de 24 ans au moment de la déclaration de leur diplôme. Cette déclaration ayant eu lieu pour la dernière fois en 1990, aucun salarié né après 1966 n'appartient à notre échantillon. On exclut ainsi 3872 individus.

Ensuite, l'étude ne portera que sur les salariés à temps complet et nous ne disposons pas d'informations relatives aux heures avant 1994. Nous éliminons ainsi du fichier tous les emplois qui ne sont pas à temps complet. Après ces premiers contrôles, nous disposons donc d'un fichier contenant 15937 observations correspondantes à 2673 individus. Dans le cadre de cette application, on va aussi considérer, pour simplifier, que l'individu n'a qu'un emploi au cours de l'année. Si un individu a eu plusieurs emplois, on conserve celui qui a duré le plus de jours. Le salaire correspondant à cette observation correspond au salaire journalier versé à l'individu pour cet emploi. En ne conservant qu'un emploi par an pour chaque salarié, on obtient un fichier contenant 14 856 observations, le nombre d'individus restant inchangé par construction.

Le panel n'est pas pour l'instant nettoyé: nous allons enlever un certain nombre d'observations parce que les variables correspondantes sont clairement fausses. On contrôle aussi les valeurs aberrantes des variables continues en enlevant les observations correspondant à des valeurs extrêmes puisque celles-ci correspondent sans doute à des erreurs de collecte ou de saisie. Dans le cas d'un panel cylindré, cette exclusion est loin d'être anodine. Enlever une observation parce qu'elle semble aberrante (la raison pourrait en être une erreur de codage) implique d'enlever toutes les observations relatives à un individu. Dans la pratique, on choisit d'enlever les observations au dessous du premier centile et au dessus du dernier centile de la distribution des salaires réels annualisés et différenciés entre hommes et femmes. Le tableau 1 donne les centiles de la distribution des salaires des hommes. Il faut noter que l'on restreint ainsi les salaires

[20] Cette section est le résultat d'un travail mené principalement par Sébastien Roux.
[21] L'EDP ne sera pas enrichi de la vague du recensement de 1999 avant 2002.

annuels à ne pas être inférieurs à 17451 francs. Ces valeurs sont faibles, notamment en comparaison du SMIC annuel. Il s'agit probablement de valeurs aberrantes liées à des erreurs de codage sur la nature à temps partiel ou à temps complet du poste occupé mais le propos n'est pas ici d'enlever toutes les valeurs aberrantes, ce qui reviendrait à faire des hypothèses très fortes sur les données, mais d'obtenir des estimations plus robustes. En enlevant les observations correspondant à des valeurs extrêmes, on conserve 14 560 observations correspondant à 2631 individus.

Tableau 1 : Centiles de la distribution des salaires réels (francs 80) annualisés des hommes

	1er centile	5ème centile	1er décile	1er quartile	Médiane	3ème quartile	9ème décile	95ème centile	99ème centile
Hommes	17451	31007	35113	42901	81306	81306	122350	165881	302398

Le tableau 2 décrit la structure des données en fonction du diplôme et de l'année. Ce tableau illustre différents phénomènes. D'une part le nombre d'observations par année décroît. Ceci est d'abord dû à l'absence d'entrées dans le panel puisque nous ne pouvons connaître le diplôme des jeunes entrants sur le marché du travail. Ensuite, cela traduit les départs naturels vers d'autres états comme le statut d'indépendant ou la retraite. On observe également que la part des diplômés ne cesse d'augmenter avec le temps. Cela correspond aux effets de sortie déjà mentionnés.

Tableau 2 : Evolution du niveau d'éducation sur la période couverte par le panel

	Sans aucun diplôme	CEP	BEPC	Bac général	CAP-BEP	Bac Pro, F,G, ouH	BTS- DEUG	2ème, 3ème cycle	Effectif Total
1991	21,0	28,5	6,2	17,3	10,3	9,1	4,4	3,3	2074
1992	20,8	27,9	6,2	17,9	10,4	9,0	4,6	3,3	2026
1993	19,4	27,1	6,1	19,3	11,1	9,0	4,8	3,5	1852
1994	20,0	25,3	6,4	20,1	11,7	8,9	4,1	3,5	1778
1995	19,3	24,7	6,2	21,0	11,3	8,9	4,7	4,0	1797
1996	19,2	24,1	6,5	21,4	11,1	9,3	4,5	4,1	1746
1997	19,5	23,4	6,3	21,6	11,0	9,5	4,7	4,0	1655
1998	19,3	21,7	6,5	22,2	11,6	9,7	5,1	3,9	1632
Marge	19,9	25,5	6,3	20,0	11,0	9,2	4,6	3,7	14560

Le tableau 3 présente l'évolution de la composition des travailleurs par région d'emploi. Celle-ci a assez peu évoluéau cours du temps. Ce tableau permet de soulever un problème dans les données, à savoir la chute brutale du nombre de travailleurs dans la région Méditerranée en 1993. Cette chute est probablement due à un trou de collecte dans les DADS important cette année là. Les tableaux 4 et 5 présentent les évolutions de la composition des travailleurs par taille d'établissement et grand secteur d'activité. Alors que la part de salariés travaillant dans le commerce diminue de 1991 à 1998, la part de ceux travaillant dans les services augmente sur la même période.

Tableau 3 : Evolution de la composition par zone d'emploi

	Région Parisienne	Bassin Parisien	Nord	Est	Ouest	Sud-Ouest	Centre-Est	Méditerranée	Total
1991	21,1	19,2	7,7	9,9	12,5	8,4	13,5	7,8	2074
1992	21,1	18,8	7,7	9,9	13,2	8,4	13,4	7,6	2026
1993	21,5	21,2	7,9	10,9	14,2	8,5	12,3	3,6	1852
1994	20,6	20,3	7,5	10,6	13,3	7,7	12,3	7,7	1778
1995	20,5	19,9	7,4	10,7	13,2	8,4	12,0	7,9	1797
1996	20,3	19,6	7,7	9,9	13,6	8,1	11,7	8,9	1746
1997	20,7	19,2	8,0	9,6	14,1	8,3	11,7	8,5	1655
1998	20,5	18,9	7,4	9,7	14,2	8,3	12,0	9,0	1632
Marge	20,8	19,6	7,6	10,2	13,5	8,3	12,4	7,6	14560

Tableau 4 : Evolution de la composition du panel par taille de l'établissement employeur

	Moins de 9 postes	Entre 10 et 49 postes	Entre 50 et 199 postes	Entre 200 et 999 postes	Plus de 1000 postes	Total
1991	19,4	28,7	21,8	17,4	12,6	2074
1992	19,6	28,1	21,9	18,2	12,2	2026
1993	19,8	26,7	22,5	18,8	12,2	1852
1994	20,5	27,0	22,9	17,6	12,1	1778
1995	23,0	27,4	21,7	17,4	10,6	1797
1996	22,7	25,7	22,4	17,5	11,7	1746
1997	21,7	25,9	22,5	18,0	11,9	1655
1998	24,8	24,8	22,1	17,1	11,2	1632
Marge	21,3	26,9	22,2	17,8	11,9	14560

Tableau 5 : Evolution de la composition du panel par grand secteur d'activité

	Industrie	Construction	Commerce	Services	Total
1991	39,6	11,4	17,2	31,8	2074
1992	40,6	11,0	16,4	32,0	2026
1993	42,5	11,3	15,4	30,7	1852
1994	41,2	11,1	15,4	32,3	1778
1995	39,1	11,6	15,8	33,6	1797
1996	40,3	10,9	15,5	33,3	1746
1997	39,2	10,9	15,8	34,0	1655
1998	38,7	11,5	16,3	33,5	1632
Marge	40,2	11,2	16,0	32,6	14560

Toutes ces évolutions ne permettent pas de distinguer les effets d'entrée et de sortie de ceux liés aux évolutions des situations individuelles. Pour contrôler des effets d'entrée et de sortie, il faut considérer une population stable sur le temps. Différents critères peuvent être utilisés pour échantillonner les individus. Nous n'en considèrerons que deux dans cette application : conserver l'ensemble des observations ou conserver les observations des seuls individus présents sans discontinuité de 1991 à 1998. Ce dernier mode d'échantillonnage correspond au cylindrage des données. Un individu pour lequel il manque une observation soit parce qu'il n'était pas employé cette année-là , soit parce que son identifiant a été mal codé, disparaît du panel cylindré. Ce panel compte 7504 observations pour 938 individus. Le tableau 6 présente le niveau d'éducation des individus échantillonnés dans le panel cylindré. Les chiffres sont du même ordre de grandeur que ceux du tableau 3, avec toutefois une représentation légèrement plus faible des moins diplômés et des plus diplômés.

Tableau 6 : Répartition des diplômes du panel cylindré.

	Sans aucun diplôme	CEP	BEPC	Bac général	CAP-BEP	Bac Pro, F,G, ouH	BTS-DEUG	2ème, 3ème cycle	Effectif Total
Marge	19,6	24,7	6,2	21,4	10,5	9,4	4,5	3,7	938

L'application portera d'abord sur le panel cylindré qui comprend des observations sur les hommes salariés à temps complet dans le secteur privé et présent dans le panel entre 1991 à 1998. Ils sont 938. Nous considérerons ensuite les estimations sur le panel non cylindré. Le nombre d'observations prises en compte sera alors de 14560 pour 2631 individus.

5.2. Estimations du modèle statique

Nous cherchons donc à expliquer les salaires en fonction de variables explicatives. Nous nous sommes limités dans cette application à utiliser l'expérience professionnelle, le diplôme et la région d'emploi. Nous n'introduisons pas d'autres variables disponibles dans le fichier telles que

l'ancienneté ou le secteur car elles sont susceptibles d'être endogènes (surtout l'ancienneté). Le lecteur intéressé par des études empiriques sur ce sujet pourront se référer à Lollivier et Payen (1990), Margolis (1996) ou Abowd, Kramarz et Margolis, (1999). Dans la suite, nous commenterons essentiellement les résultats portant sur le rendement de l'expérience et montrerons dans quelle mesure l'estimation tenant compte d'une hétérogénéité inobservable fixe change les résultats. Comme dans la section 1, nous avons choisi de présenter d'abord les résultats d'estimation du modèle de Mundlak sur le panel cylindré. Nous chercherons à exploiter les caractéristiques des données pour améliorer la précision en estimant le modèle par moindres carrés quasi-généralisés. Puis nous comparerons ces résultats à ceux que l'on obtient par MCO. Nous conclurons en présentant les estimations à partir du panel non cylindré.

5.2.1. Mise en œuvre du modèle de Mundlak

Comme nous l'avons vu dans la partie théorique, l'estimation du modèle de Mundlak permet de tenir compte de l'hétérogénéité fixe inobservable. Il s'agit ici simplement d'introduire parmi les variables explicatives les moyennes individuelles des variables (signalées par un « m » en tête) évoluant avec le temps. Cette estimation, comme nous l'avons vu, fournit des estimés égaux aux coefficients estimés dans le modèle à effets fixes (within). La procédure de test d'Hausman qui permet de choisir entre modèle à effets fixes et à erreurs composées, se fonde également sur une statistique qui peut être tirée de cette estimation : il correspond au test de l'égalité jointe à 0 des coefficients correspondant aux moyennes temporelles des variables explicatives introduites dans le modèle.

Les tables qui suivent présentent les estimations des coefficients correspondant à chaque variable. Comme nous l'avons vu, les coefficients estimés par la méthode within sont aussi ceux des variables ne commençant pas par « m » sauf le diplôme. Celui-ci étant une caractéristique fixe de l'individu, il ne peut pas apparaître dans le modèle within. L'année n'étant pas une caractéristique individuelle propre n'apparaîtra pas dans le modèle between.

On obtient un rendement de l'expérience élevé, le coefficient correspondant au terme linéaire est égal à 0,154 et celui correspondant au terme quadratique à -0,0038. Avec cette spécification, un individu atteint son maximum de salaire au bout de 20 ans d'expérience (0,154/(2*0,0038)).

Le programme

```
/*construction des moyennes par individu pour l'estimateur du modèle de Mundlak */

proc summary data=final3 nway;
class nni;
id cyl9198  dipl1-dipl8 tagentr1-tagentr6;
var lsnre tait01-tait11 reg1-reg5 reg7-reg9 sect1-sect15 expr expr2 expp expp2
   anc anc2;
output out=finbetw mean=mlsnre mtait01-mtait11 mreg1-mreg5 mreg7-mreg9 msect1-msect15
        mexpr mexpr2 mexpp mexpp2 manc manc2;
run;


 /*introduction dans les données par années des moyennes individuelles*/

data finmund;
merge final3 finbetw;
by nni;run;


 /*estimation du modèle*/
```

```
proc reg data=finmund(where=(cyl9198=1)) ;
model lsnre=expr expr2 mexpr mexpr2 an92-an98
        dipl1-dipl7
        reg1-reg4 reg7-reg9
        mreg1-mreg4 mreg7-mreg9 ;
 /* test de la nullite jointe des coefficients */
test mexpr,mexpr2,mreg1,mreg2,mreg3,mreg4,mreg7,mreg8,mreg9;
test mexpr,mexpr2;
test mreg1,mreg2,mreg3,mreg4,mreg7,mreg8,mreg9;
run;
```

On notera qu'il n'y a pas de variable indicatrice correspondant à la région numéro 6. La variable indicatrice de référence est la région 5.

La sortie SAS

The REG Procedure
Model: MODEL1
Dependent Variable: LSNRE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	32	597.14721	18.66085	164.01	<.0001
Error	7471	850.05562	0.11378		
Corrected Total	7503	1447.20283			

F-Value correspond à une statistique de Fischer, permettant de tester si le modèle a un pouvoir explicatif. Cette statistique est égale au rapport de la variance de la variable dépendante sur la variance des résidus. Elle suit une loi de Fischer de degré de liberté égal au nombre de variables explicatives introduites dans le modèle. Il est très rare que ce test soit accepté, c'est à dire que le modèle n'ait aucun pouvoir explicatif (dans ce dernier cas, il vaut mieux changer d'approche).

Root MSE	0.33731	R-Square	0.4126
Dependent Mean	4.13080	Adj R-Sq	0.4101
Coeff Var	8.16582		

Root MSE correspond à l'écart-type des résidus, R-square est le R carré, Adj R-sq est le R carré corrigé du nombre de degrés de liberté.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.82341	0.10064	47.93	<.0001
EXPR	1	0.15354	0.02860	5.37	<.0001
EXPR2	1	-0.00380	0.00131	-2.91	0.0036
mexpr	1	-0.13666	0.02878	-4.75	<.0001
mexpr2	1	0.00542	0.00164	3.30	0.0010
AN92	1	-0.12114	0.03165	-3.83	0.0001

AN93	1	-0.24687	0.05747	-4.30	<.0001
AN94	1	-0.40478	0.08407	-4.82	<.0001
AN95	1	-0.50360	0.11115	-4.53	<.0001
AN96	1	-0.63351	0.13841	-4.58	<.0001
AN97	1	-0.76695	0.16558	-4.63	<.0001
AN98	1	-0.86778	0.19296	-4.50	<.0001
DIPL1	1	-0.85959	0.02379	-36.13	<.0001
DIPL2	1	-0.80549	0.02397	-33.61	<.0001
DIPL3	1	-0.58857	0.02696	-21.83	<.0001
DIPL4	1	-0.73046	0.02326	-31.40	<.0001
DIPL5	1	-0.40419	0.02483	-16.28	<.0001
DIPL6	1	-0.43633	0.02513	-17.37	<.0001
DIPL7	1	-0.24260	0.02776	-8.74	<.0001
REG1	1	0.16592	0.07212	2.30	0.0214
REG2	1	0.03859	0.08312	0.46	0.6424
REG3	1	0.14709	0.10746	1.37	0.1711
REG4	1	0.07757	0.10576	0.73	0.4633
REG7	1	0.11600	0.13158	0.88	0.3780
REG8	1	0.08268	0.09587	0.86	0.3885
REG9	1	0.15334	0.09597	1.60	0.1101
mreg1	1	0.15086	0.07347	2.05	0.0401
mreg2	1	0.01248	0.08420	0.15	0.8821
mreg3	1	-0.09661	0.10877	-0.89	0.3745
mreg4	1	-0.00193	0.10691	-0.02	0.9856
mreg7	1	-0.12843	0.13276	-0.97	0.3334
mreg8	1	0.04591	0.09706	0.47	0.6363
mreg9	1	-0.05515	0.09919	-0.56	0.5782

On peut tester la nullité de la corrélation (partielle) avec l'effet fixe pour chaque variable à partir de l'estimation présentée ci-dessus. Par exemple, le terme linéaire de l'expérience est corrélé avec l'effet fixe. En effet, le coefficient estimé sur la moyenne temporelle de l'expérience « mexpr » est significativement différent de 0. C'est également le cas pour la variable moyenne correspondant à l'habitation en Ile de France (mreg1).

Quant à lui, le test d'Hausman se fonde sur la statistique correspondant à la nullité jointe de ces coefficients. Il teste l'existence d'une hétérogénéité fixe inobservable corrélée aux variables explicatives. Ce test est initié par la commande dans la procédure :

test mexpr, mexpr2, mreg1, mreg2, mreg3, mreg4, mreg7 ,mreg8, mreg9

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable LSNRE

Source	DF	Mean Square	F Value	Pr > F
Numerator	9	0.55633	4.89	<.0001
Denominator	7471	0.11378		

Ce test teste la nullité globale des coefficients mexpr, mexpr2, mreg1 à mreg9. La valeur de la statistique de test (correspondant à un test de Fisher à 2 degrés de liberté) permet de rejeter très clairement l'hypothèse nulle : les coefficients correspondant aux moyennes des variables explicatives sont différents de 0.

Le test suivant teste la nullité globale des coefficients mexpr et mexpr2, la statistique de test (correspondant à un test de Fisher à 2 degrés de liberté), rejette très clairement l'hypothèse nulle

: les coefficients correspondant à l'expérience moyenne sont différents de 0. Ce test est initié par la commande dans la procédure : test mexpr,mexpr2

```
                          The REG Procedure
                            Model: MODEL1

                Test 2 Results for Dependent Variable LSNRE

                                      Mean
Source                  DF          Square     F Value     Pr > F

Numerator                2         1.57225       13.82     <.0001
Denominator           7471         0.11378
```

Enfin, on peut tester la nullité des coefficients régionaux.

```
                          The REG Procedure
                            Model: MODEL1

                Test 3 Results for Dependent Variable LSNRE

                                      Mean
Source                  DF          Square     F Value     Pr > F

Numerator                7         0.27450        2.41     0.0182
Denominator           7471         0.11378
```

Dans ce dernier cas, on observe que le test de non corrélation entre l'effet fixe et la région n'est pas rejeté au seuil de 1%. Il l'est en revanche au seuil de 5%. Cette absence de corrélation va nous permettre d'utiliser les moindres carrés généralisés pour augmenter l'efficacité des estimations. Dans le cas général, il est plutôt déconseillé d'accepter une hypothèse au seuil d'1% alors qu'elle est rejetée au seuil de 5%. Ce n'est qu'à titre illustratif et afin de montrer la possibilité de gagner de l'efficacité sur les estimateurs que nous acceptons ici l'hypothèse d'absence de corrélation entre la variable indiquant la région et l'effet fixe.

5.2.2. Les moindres carrés quasi-généralisés

Une fois estimé le modèle de Mundlak, il est possible d'améliorer les estimations en cas d'absence de corrélation entre certaines variables explicatives et l'effet fixe individuel. C'est ce que nous exploiterons ici en supposant que les effets régionaux ne sont pas corrélés aux effets individuels. On applique alors la méthode des moindres carrés quasi-généralisé. L'implémentation de la méthode est la suivante:

- Estimer les modèles between et within pour obtenir une estimation de rho (cf. section 2.3.2).
- Estimer le modèle sphéricisé: on obtient alors des estimateurs plus efficaces.

Le programme

On estime d'abord le modèle within :

```
/* estimation de rho dans le cas où l'on cherche à estimer un estimateur MCQG */

/*transformation des données en leur écart à la moyenne temporelle*/
```

```
proc standard data=final3 out=finwith mean=0;
by nni ;
var lsnre tait01-tait11 reg1-reg5 reg7-reg9 sect1-sect15 expr expr2 expp expp2;
run;

 /*dans ce cas, les variables correspondant à l'anciennete ne sont pas significativement
   differentes de zero cf Test de Hausman */

  /* estimateur within */

proc reg data=finwith(where=(cyl9198=1)) outest=with ;
model lsnre=expr expr2 an92-an98
        reg1-reg4 reg7-reg9 ;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: LSNRE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	12.74418	0.79651	45.37	<.0001
Error	7487	131.44515	0.01756		
Corrected Total	7503	144.18933			

Root MSE	0.13250	R-Square	0.0884
Dependent Mean	-8.522E-18	Adj R-Sq	0.0864
Coeff Var	-1.55481E18		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.44308	0.03806	11.64	<.0001
EXPR	1	0.15354	0.01124	13.66	<.0001
EXPR2	1	-0.00380	0.00051308	-7.41	<.0001
AN92	1	-0.12114	0.01243	-9.74	<.0001
AN93	1	-0.24687	0.02257	-10.94	<.0001
AN94	1	-0.40478	0.03302	-12.26	<.0001
AN95	1	-0.50360	0.04366	-11.53	<.0001
AN96	1	-0.63351	0.05437	-11.65	<.0001
AN97	1	-0.76695	0.06504	-11.79	<.0001
AN98	1	-0.86778	0.07580	-11.45	<.0001
REG1	1	0.16592	0.02833	5.86	<.0001
REG2	1	0.03859	0.03265	1.18	0.2372
REG3	1	0.14709	0.04221	3.48	0.0005
REG4	1	0.07757	0.04155	1.87	0.0619
REG7	1	0.11600	0.05168	2.24	0.0248
REG8	1	0.08268	0.03766	2.20	0.0282
REG9	1	0.15334	0.03770	4.07	<.0001

On remarque au passage que les coefficients du modèle within sont strictement égaux à ceux obtenus à partir du modèle de Mundlak pour les variables évoluant avec le temps. On vérifie ainsi

la théorie que l'on applique. C'est également un moyen de vérifier les programmes !!!

Dans cette estimation within, la seule chose qui nous intéresse ici est l'écart-type des résidus du modèle estimé, soit 0,13250. On récupère cette valeur dans la table créée par la procédure REG de SAS contenant les estimations du modèle, ici with.

On estime maintenant l'estimateur between :

```
  /* estimateur between */

proc reg data=finbetw(where=(cyl9198=1)) outest=betw;
model mlsnre=mexpr mexpr2
        dipl1-dipl4 dipl6-dipl8
        mreg1-mreg4 mreg7-mreg9 ;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: mlsnre

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	73.05038	4.56565	46.81	<.0001
Error	921	89.82631	0.09753		
Corrected Total	937	162.87669			

Root MSE	0.31230	R-Square	0.4485
Dependent Mean	4.13080	Adj R-Sq	0.4389
Coeff Var	7.56027		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.97615	0.07064	56.28	<.0001
mexpr	1	0.01688	0.00843	2.00	0.0456
mexpr2	1	0.00161	0.00261	0.62	0.5369
DIPL1	1	-0.45541	0.04002	-11.38	<.0001
DIPL2	1	-0.40130	0.03853	-10.42	<.0001
DIPL3	1	-0.18438	0.05217	-3.53	0.0004
DIPL4	1	-0.32627	0.03986	-8.19	<.0001
DIPL6	1	-0.03214	0.04638	-0.69	0.4885
DIPL7	1	0.16159	0.05906	2.74	0.0063
DIPL8	1	0.40419	0.06501	6.22	<.0001
mreg1	1	0.31679	0.03668	8.64	<.0001
mreg2	1	0.05107	0.03520	1.45	0.1471
mreg3	1	0.05048	0.04414	1.14	0.2531
mreg4	1	0.07563	0.04080	1.85	0.0641
mreg7	1	-0.01242	0.04631	-0.27	0.7886
mreg8	1	0.12858	0.03962	3.25	0.0012
mreg9	1	0.09819	0.06561	1.50	0.1349

Là encore, la seule chose qui nous intéresse dans cette estimation est l'écart-type des résidus du modèle, soit 0,31230. Cette valeur est récupérée dans la table betw.

```
/*calcul et stockage de rho*/
data statt;
merge with(keep=_rmse_ obs=1 rename=(_rmse_=with))
   betw(keep=_rmse_ obs=1 rename=(_rmse_=betw));
         run;

data rho;
if _n_=1 then
set statt;
set nbper;
rho=with/betw/sqrt(nbper-1); /* utilisation de la formule de la section 2.3.2. les écarts-types sont
utilisés ici*/
run;
```

On obtient avec ces données rho = 0.16. On sphéricise le modèle pour estimer le modèle par la méthode des moindres carrés quasi-généralisés.

```
  /*spericisation des variables pour les quelles le test d'Hausman passe et de la variable
    à expliquer */

data finmcqg;
if _n_=1 then set rho;
merge finwith finbetw(keep=nni mlsnre mexpr mexpr2
        dipl1-dipl4 dipl6-dipl8
        mreg1-mreg4 mreg7-mreg9);
by nni;
array vvar(9) lsnre reg1-reg4 reg6-reg9;
array vmvar(9) mlsnre mreg1-mreg4 mreg6-mreg9;
array vrvar(9) rlsnre rreg1-rreg4 rreg6-rreg9;
do i=1 to 9;
  vrvar(i)=vvar(i)+rho*vmvar(i);
end;
run;

proc reg data=finmcqg(where=(cyl9198=1)) ;
model rlsnre=expr expr2 mexpr mexpr2 an92-an98
        dipl1-dipl7
        rreg1-rreg4 rreg7-rreg9;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: rlsnre

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	25	27.09939	1.08398	53.83	<.0001
Error	7478	150.59757	0.02014		
Corrected Total	7503	177.69696			

Root MSE	0.14191	R-Square	0.1525
Dependent Mean	0.66242	Adj R-Sq	0.1497
Coeff Var	21.42321		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.14903	0.04228	27.17	<.0001
EXPR	1	0.15219	0.01199	12.69	<.0001
EXPR2	1	-0.00389	0.00054877	-7.09	<.0001
mexpr	1	0.00272	0.00135	2.01	0.0442
mexpr2	1	0.00026071	0.00041879	0.62	0.5336
AN92	1	-0.11939	0.01327	-9.00	<.0001
AN93	1	-0.24320	0.02409	-10.09	<.0001
AN94	1	-0.39937	0.03523	-11.34	<.0001
AN95	1	-0.49643	0.04658	-10.66	<.0001
AN96	1	-0.62480	0.05801	-10.77	<.0001
AN97	1	-0.75669	0.06940	-10.90	<.0001
AN98	1	-0.85588	0.08088	-10.58	<.0001
DIPL1	1	-0.14588	0.00988	-14.76	<.0001
DIPL2	1	-0.13610	0.00999	-13.63	<.0001
DIPL3	1	-0.10210	0.01124	-9.08	<.0001
DIPL4	1	-0.12422	0.00969	-12.82	<.0001
DIPL5	1	-0.06799	0.01041	-6.53	<.0001
DIPL6	1	-0.07581	0.01048	-7.24	<.0001
DIPL7	1	-0.04438	0.01163	-3.82	0.0001
rreg1	1	0.20695	0.02267	9.13	<.0001
rreg2	1	0.03694	0.02455	1.50	0.1324
rreg3	1	0.09308	0.03147	2.96	0.0031
rreg4	1	0.07096	0.02997	2.37	0.0179
rreg7	1	0.01878	0.03531	0.53	0.5947
rreg8	1	0.09904	0.02813	3.52	0.0004
rreg9	1	0.14614	0.03185	4.59	<.0001

Ces estimations par MCQG n'apportent que des changements marginaux dans les valeurs estimées des rendements de l'expérience par rapport au modèle de Mundlak. Les valeurs du rendement de l'expérience linéaire sont très proches : 0,15354 pour Mundlak et 0,15219 pour MCQG. En revanche, l'écart-type de cette estimation est plus faible pour le modèle MCQG, 0,01199 contre 0,02860 pour Mundlak. Il est divisé par plus de deux. On observe le même phénomène pour le terme quadratique de l'expérience. Les valeurs des coefficients liés à l'expérience ayant très peu bougé, le niveau de l'expérience pour lequel le salaire est maximum est également inchangé. Mais l'estimation est plus précise.

5.2.3. L'estimateur MCO est biaisé

L'estimateur par les moindres carrés ordinaires ne peut pas être convergent puisque nous avons montré que les effets individuels inobservables étaient corrélés avec l'expérience professionnelle au moins (voir test d'Hausman). Les variables explicatives sont donc endogènes et les estimateurs MCO sont biaisés. C'est pourquoi on peut ne pas s'étonner de l'ampleur de la différence entre les coefficients estimés dans le modèle de Mundlak et ceux obtenus par MCO. Une estimation par les MCO donne comme coefficients pour l'expérience et l'expérience carrée respectivement 0,023 et 0,00034 (au lieu de 0,154 et –0,0038 !). Pour le terme linéaire, il s'agit d'un faible coefficient par rapport aux résultats obtenus dans d'autres études avec la même spécification (Abowd, Kramarz

et Margolis, 1999). Le coefficient correspondant au terme quadratique n'est pas significativement négatif, en contradiction avec l'ensemble des études sur le sujet.

Le programme

```
/* mco simples sur le panel cylindré*/

proc reg data=final3(where=(cyl9198=1)) ;
model lsnre=expr expr2 an92-an98
      dipl1-dipl4 dipl6-dipl8
      reg1-reg4 reg7-reg9 ;
run;
```

```
                                The REG Procedure
                                  Model: MODEL1
                              Dependent Variable: LSNRE

                                Analysis of Variance

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

Model                      23          592.14028       25.74523     225.22    <.0001
Error                    7480          855.06255        0.11431
Corrected Total          7503         1447.20283


Corrected total

                 Root MSE            0.33810    R-Square     0.4092
                 Dependent Mean      4.13080    Adj R-Sq     0.4073
                 Coeff Var           8.18491


                              Parameter Estimates

                                Parameter       Standard
Variable        DF               Estimate          Error    t Value    Pr > |t|

Intercept        1                3.96874        0.02515     157.80      <.0001
EXPR             1                0.02311        0.00259       8.92      <.0001
EXPR2            1            -0.00033805     0.00079329      -0.43      0.6700
AN92             1                -0.00137        0.01564      -0.09      0.9301
AN93             1                -0.00634        0.01570      -0.40      0.6861
AN94             1                -0.04631        0.01578      -2.93      0.0033
AN95             1                -0.02724        0.01588      -1.72      0.0863
AN96             1                -0.04040        0.01599      -2.53      0.0116
AN97             1                -0.05856        0.01612      -3.63      0.0003
AN98             1                -0.04323        0.01627      -2.66      0.0079
DIPL1            1                -0.45871        0.01530     -29.99      <.0001
DIPL2            1                -0.40538        0.01472     -27.54      <.0001
DIPL3            1                -0.18678        0.01996      -9.36      <.0001
DIPL4            1                -0.33093        0.01522     -21.75      <.0001
DIPL6            1                -0.03203        0.01773      -1.81      0.0709
DIPL7            1                 0.16296        0.02259       7.21      <.0001
DIPL8            1                 0.41487        0.02474      16.77      <.0001
REG1             1                 0.30786        0.01372      22.45      <.0001
REG2             1                 0.04856        0.01327       3.66      0.0003
```

REG3	1	0.05234	0.01668	3.14	0.0017
REG4	1	0.07440	0.01544	4.82	<.0001
REG7	1	-0.01272	0.01755	-0.72	0.4687
REG8	1	0.12605	0.01496	8.43	<.0001
REG9	1	0.10899	0.02385	4.57	<.0001

5.2.4. Estimations à partir du panel non cylindré

Pour illustrer l'importance de la sélection des données, nous présentons ici les estimations effectuées sur le panel non cylindré. Dans ce cas, il y a beaucoup plus d'observations. On s'attend alors à ce que les coefficients soient estimés plus précisément. Néanmoins, les résultats sont très différents de ceux obtenus à partir du panel cylindré. En effet, on peut mener l'estimation des rendements de l'expérience sur le panel non cylindrépar la méthode de Mundlak. Cela consiste à calculer les moyennes individuelles des variables même si les trajectoires sont incomplètes et à régresser la variable dépendante sur les variables explicatives et ces moyennes individuelles des variables. Cela donne le résultat suivant : le coefficient du terme linéaire est égal à 0,03, tandis que celui correspondant au terme quadratique est, cette fois, significativement négatif, égal à -0,0048. Ces résultats sont très différents de ceux obtenus à partir du modèle cylindré(0,154 et -.0038). A titre d'illustration, avec ces résultats, un individu atteint son maximum de salaire après 3 ans d'expérience. Ce résultat est peu crédible.

Le programme

```
proc reg data=finmund ;
model lsnre=expr expr2 mexpr mexpr2 an92-an98
        dipl1-dipl7
        reg1-reg4 reg7-reg9
        mreg1-mreg4 mreg7-mreg9 ;
 /* test de la nullite jointe des coefficients */
test mexpr,mexpr2,mreg1,mreg2,mreg3,mreg4,mreg7,mreg8,mreg9;
test mexpr,mexpr2;
test mreg1,mreg2,mreg3,mreg4,mreg7,mreg8,mreg9;
run;
```

La sortie SAS

The REG Procedure
Model: MODEL1
Dependent Variable: LSNRE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	32	1134.29011	35.44657	232.65	<.0001
Error	14527	2213.34231	0.15236		
Corrected Total	14559	3347.63242			

Root MSE	0.39033	R-Square	0.3388
Dependent Mean	4.10953	Adj R-Sq	0.3374
Coeff Var	9.49826		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.42323	0.02495	177.31	<.0001
EXPR	1	0.03013	0.00529	5.70	<.0001
EXPR2	1	-0.00484	0.00115	-4.20	<.0001
mexpr	1	-0.02542	0.00556	-4.57	<.0001
mexpr2	1	0.00998	0.00132	7.58	<.0001
AN92	1	0.00255	0.01259	0.20	0.8395
AN93	1	0.00340	0.01383	0.25	0.8059
AN94	1	-0.01741	0.01541	-1.13	0.2588
AN95	1	-0.00929	0.01747	-0.53	0.5949
AN96	1	-0.02232	0.01987	-1.12	0.2614
AN97	1	-0.02317	0.02270	-1.02	0.3074
AN98	1	-0.00500	0.02560	-0.20	0.8450
DIPL1	1	-0.81492	0.01921	-42.42	<.0001
DIPL2	1	-0.77164	0.01920	-40.19	<.0001
DIPL3	1	-0.58496	0.02183	-26.79	<.0001
DIPL4	1	-0.68996	0.01896	-36.39	<.0001
DIPL5	1	-0.36330	0.02007	-18.10	<.0001
DIPL6	1	-0.41558	0.02047	-20.30	<.0001
DIPL7	1	-0.23988	0.02283	-10.51	<.0001
REG1	1	0.17293	0.05288	3.27	0.0011
REG2	1	0.10846	0.05890	1.84	0.0656
REG3	1	0.10740	0.07633	1.41	0.1594
REG4	1	0.12693	0.08102	1.57	0.1172
REG7	1	0.04556	0.07634	0.60	0.5506
REG8	1	0.06529	0.06967	0.94	0.3487
REG9	1	0.12459	0.07101	1.75	0.0793
mreg1	1	0.15990	0.05418	2.95	0.0032
mreg2	1	-0.04352	0.06005	-0.72	0.4686
mreg3	1	-0.05831	0.07778	-0.75	0.4535
mreg4	1	-0.06289	0.08217	-0.77	0.4440
mreg7	1	-0.02659	0.07772	-0.34	0.7323
mreg8	1	0.06904	0.07086	0.97	0.3299
mreg9	1	-0.03123	0.07261	-0.43	0.6672

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable LSNRE

Source	DF	Mean Square	F Value	Pr > F
Numerator	9	1.64306	10.78	<.0001
Denominator	14527	0.15236		

The REG Procedure
Model: MODEL1

Test 2 Results for Dependent Variable LSNRE

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	4.44527	29.18	<.0001
Denominator	14527	0.15236		

The REG Procedure
Model: MODEL1

Test 3 Results for Dependent Variable LSNRE

Source	DF	Mean Square	F Value	Pr > F
Numerator	7	0.78184	5.13	<.0001
Denominator	14527	0.15236		

On remarquera d'abord que l'absence de corrélation de variables explicatives avec l'hétérogénéité fixe inobservable est toujours rejetée. Mais l'important n'est pas là . La différence observée entre les estimations du modèle de Mundlak sur le panel cylindré et sur le panel non cylindré est imputable à la sélection des données : dans le panel cylindré, on conserve les individus présents à temps complet 8 années consécutives, c'est à dire particulièrement bien insérés sur le marché du travail. Cette différence illustre la sensibilité des résultats au phénomène de sélection endogène : cylindrer le panel ne résout pas le problème. En effet, en cylindrant le panel, on sélectionne des individus ayant des carrières stables et donc plutôt mieux rémunérés. Ne pas le cylindrer fait l'hypothèse implicite que les phénomènes d'entrées-sorties ne sont pas liés aux salaires. On fait là aussi des hypothèses très fortes sur les comportements des agents économiques. Mais que les deux estimations sont différentes ne fait que révéler le problème. Elles ne permettent pas de le traiter. Il faudrait disposer pour cela de méthodes sur les données qualitatives de panel. Par contre, si on avait obtenu des résultats similaires sur le panel cylindré et sur le panel non cylindré, cela voudrait dire que ces résultats sont robustes au cylindrage et que les biais de sélection sont peu importants.

D'autre part, on pourrait aussi se demander quelles sont les propriétés dynamiques de la variable de salaire. Certaine statistiques descriptives portant sur les premières différences et sur les matrices de variance-covariance temporelle du (log) salaire et de sa première différence est proposée dans l'annexe 2. On y voit que l'hypothèse de présence d'hétérogénéité inobservable et d'indépendance temporelle entre les perturbations individuelles-temporelles peut être acceptée.

Annexe 1 : Description des données

Les données utilisées dans cette application sont issues de l'appariement des DADS (Déclarations Administratives de Données Sociales) et de l'EDP (Echantillon Démographique Permanent). Les DADS sont issues de fichiers administratifs collectés auprès des entreprises et donnant des renseignements sur l'emploi et le salaire de chaque employé. L'Insee a pu obtenir l'autorisation d'exploiter à des fins d'étude les données correspondant aux individus nés le mois d'octobre des années paires, ce qui correspond à 1/25ème de la population. Le panel permet de suivre ainsi la carrière salariale et professionnelle dans le secteur privé depuis 1976 jusqu'en 1998 de tous les individus nés en Octobre d'une année paire (sauf en 1981, 1983 et 1990, les DADS n'ayant pas été collectées ces années-là par l'Insee pour cause de recensement)[22].

L'échantillon démographique permanent suit les individus nés au cours des premiers jours du mois d'octobre. Il contient les données issus des recensements de 1968, 1975, 1982 et 1990 ainsi que certains renseignements d'état civil (mariage, naissance d'enfants, décès ...). Il s'agit grosso-modo d'un échantillon au 1/125ème de la population contenant les variables du recensement, on dispose ainsi du diplôme, variable non disponible dans les DADS.

Le rapprochement de ces deux sources est partiellement possible car les individus peuvent être identifiés dans les deux fichiers par le NIR (disponible dans les DADS et reconstruit dans l'EDP). Pour les individus nés hors de France, l'identifiant ne correspond pas à celui utilisé dans les DADS. Pour cette raison, nous ne pouvons pas les inclure dans l'analyse. Cela revient à exclure approximativement 8% des individus. Etant donnée la nature de l'échantillonnage, dans l'appariement vont se trouver les individus nés en France au cours des premiers jours du mois d'octobre d'une année paire.

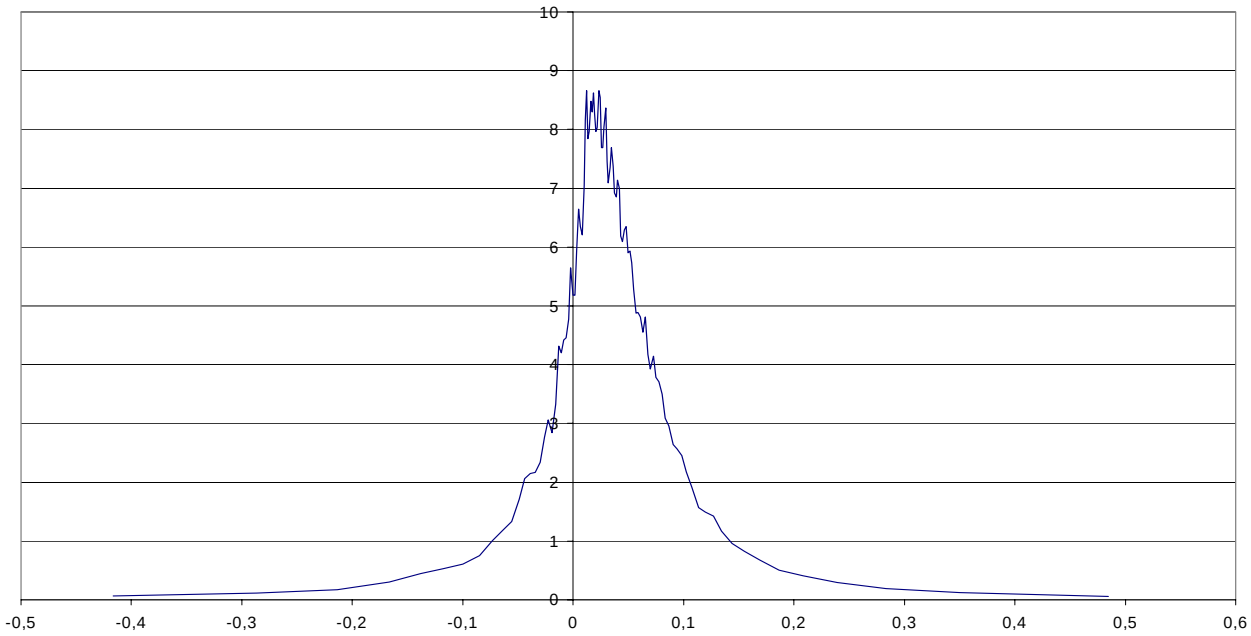
[22] Pour une présentation plus complète du panel, cf. Refonte du panel DADS (Document de travail de la DSDS).

Annexe 2 : Quelques statistiques exploratoires.

Y-a-t-il une rigidité nominale des salaires ?

Avec le panel cylindré, nous pouvons déjà tester un certain nombre de faits constatés empiriquement dans d'autres études (cf. Card et Hyslop, 1996). L'un deux consiste en l'existence d'une rigidité nominale des salaires se traduisant par un point d'accumulation à 0 de la distribution des évolutions de salaires. Le graphique 1 reproduit cette distribution : ce fait n'est pas constaté sur les données françaises. On peut expliquer cette différence de deux manières. D'une part, on travaille ici sur des données françaises, alors que la rigidité nominale a été constatée sur des données américaines. Les différences de marché du travail entre France et Etats-Unis peuvent placer la rigidité à un autre niveau que nominale. Notamment en France l'existence du Smic et le système de convention collective peuvent avoir pour conséquence que les salaires et leurs évolutions ne se fixent pas de la même façon dans les deux pays. L'autre possibilité d'explication tient au fait que nous travaillons ici sur des données administratives où nous disposons des valeurs précises des salaires, alors que la plupart des autres études se fondent sur des enquêtes annuelles auprès des ménages où le salaire déclaré est souvent approché à un chiffre rond. Il suffit qu'un nombre de gens suffisant aient des évolutions de salaire trop faibles pour modifier le chiffre rond pour qu'une masse d'individus aient une évolution nominale de leurs salaire apparemment nulle.

Graphique 1 : Distribution des évolutions de salaire



La structure dynamique des salaires: quelques éléments d'introduction.

Nous effectuons ici des traitements statistiques similaires à ceux effectués par Abowd et Card, 1989, pour rendre compte de la structure dynamique des salaires, c'est à dire pour observer à quels modèles statistiques ils se conforment le mieux. Pour cela, nous estimons la matrice de variance covariance des salaires sur la période considérée et testons dans quelle mesure cette matrice de variance-covariance est proche de celle générée par tel ou tel type de processus statistique.

La matrice de variance-covariance des niveaux de salaire met en évidence leur très forte persistance. La corrélation entre deux salaires à des dates différentes n'est jamais inférieure à 0,8. Ceci montre bien qu'une modélisation avec effets individuels est justifiée. En présence d'effets individuels, cette corrélation est en effet bornée inférieurement comme ici. De plus, la part de la variance expliquée par l'effet individuel est alors d'environ 80%.

Matrice de variance covariance du log des salaires nominaux

	LSN91	LSN92	LSN93	LSN94	LSN95	LSN96	LSN97	LSN98
LSN91	0,18	0,16	0,17	0,16	0,16	0,16	0,16	0,16
LSN92	0,93 0,0001	0,17	0,17	0,17	0,16	0,17	0,17	0,17
LSN93	0,88 0,0001	0,91 0,0001	0,20	0,18	0,17	0,17	0,17	0,17
LSN94	0,89 0,0001	0,93 0,0001	0,90 0,0001	0,19	0,18	0,18	0,18	0,18
LSN95	0,88 0,0001	0,90 0,0001	0,88 0,0001	0,92 0,0001	0,19	0,18	0,18	0,18
LSN96	0,86 0,0001	0,88 0,0001	0,86 0,0001	0,90 0,0001	0,89 0,0001	0,20	0,18	0,18
LSN97	0,87 0,0001	0,90 0,0001	0,88 0,0001	0,92 0,0001	0,92 0,0001	0,91 0,0001	0,19	0,18
LSN98	0,84 0,0001	0,88 0,0001	0,85 0,0001	0,90 0,0001	0,90 0,0001	0,88 0,0001	0,93 0,0001	0,20

Lecture : Dans la diagonale et le triangle supérieur de la matrice est reproduit l'équivalent de la matrice de variance-covariance des salaires. Dans le triangle inférieur est reproduit l'équivalent de la matrice de corrélation, avec les probabilités de non-significativitécorrespondantes.

champ : Hommes salariés présents dans le panel cylindréde 91 à 98.

La matrice de variance-covariance des évolutions de salaire ne montre pas une telle persistance du niveau des évolutions. Au contraire, les évolutions d'une année sur l'autre semblent être corrélées négativement, les corrélations au cours des années suivantes étant plus faibles encore. Ces corrélations négatives peuvent être liées à des problèmes d'erreur de mesure. En effet, si le

salaire est mal mesuré au cours d'une année, par exemple sous-estimé, alors l'évolution de salaire est sous-estimée l'année de l'erreur de mesure et surestimée l'année suivante. Pour cette raison, les évolutions sont négativement corrélées en cas d'erreur de mesure. Dans les DADS, les erreurs de mesure peuvent venir du fait suivant : si les rémunérations annuelles sont observées de manière très fiable, c'est moins le cas de la période correspondant à cette rémunération. Les salaires annuels, c'est à dire ramenés à cette période, peuvent donc être mesurés avec une certaine erreur.

On s'aperçoit aussi que les corrélations à des ordres supérieurs ou égaux à 2 sont non significatifs. Il n'y a donc pas d'effets individuels permanents dans les taux de croissance. Comme cette auto-corrélation entre les différences n'est significative qu'à l'ordre 1, il est crédible que pour les variables en niveau l'hypothèse d'absence d'autocorrélation temporelle soit acceptable. Mais il faudrait le tester formellement.

Matrices de covariance et de corrélation des évolutions de salaire.

	DLSN92	DLSN93	DLSN94	DLSN95	DLSN96	DLSN97	DLSN98
DLSN92	0,02	-0,01	0,00	0,00	0,00	0,00	0,00
DLSN93	-0,19 0,0001	0,03	-0,02	0,00	0,00	0,00	0,00
DLSN94	0,00 0,9971	-0,63 0,0001	0,04	-0,01	0,00	0,00	0,00
DLSN95	-0,06 0,0823	0,00 0,8798	-0,31 0,0001	0,03	-0,02	0,00	0,00
DLSN96	-0,01 0,653	0,00 0,986	0,01 0,7372	-0,46 0,0001	0,04	-0,02	0,00
DLSN97	0,05 0,0924	-0,02 0,536	0,00 0,9327	0,04 0,2588	-0,61 0,0001	0,04	-0,01
DLSN98	0,02 0,4646	-0,02 0,4663	0,08 0,0216	0,00 0,9371	-0,03 0,3193	-0,27 0,0001	0,03

Lecture : Dans la diagonale et le triangle supérieur de la matrice est reproduit l'équivalent de la matrice de variance-covariance des salaires. Dans le triangle inférieur est reproduit l'équivalent de la matrice de corrélation, avec les probabilités de non-significativité correspondantes.

champ : Hommes salariés présents dans le panel cylindré de 91 à 98.

Annexe 3

La propriété de Frisch-Waugh

Supposons que l'on veuille estimer le modèle linéaire suivant :

$$y = x_1\beta_1 + x_2\beta_2 + u$$

La propriété de Frisch-Waugh peut être résumée par le fait que les estimateurs des MCO, $\hat{\beta}_1$ et $\hat{\beta}_2$ peuvent s'obtenir en deux étapes :

1) $\hat{\beta}_2$ s'obtient par la régression de $M_{X_1}Y$ sur $M_{X_1}X_2$ où M_{X_1} dénote le projecteur orthogonal sur l'espace orthogonal à l'espace engendré par les variables X_1.

2) $\hat{\beta}_1$ s'obtient par la régression des résidus $Y - X_2\hat{\beta}_2$ sur X_1.

Preuve : En multipliant l'équation :

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{U}$$

par M_{X_1} et en utilisant la condition $M_{X_1}\hat{U} = \hat{U}$ puisque les résidus sont par construction orthogonaux à X_2 :

$$M_{X_1}Y = M_{X_1}X_2\hat{\beta}_2 + \hat{U}$$

En utilisant la condition $X_2'\hat{U} = 0,$ on montre la première assertion.

La régression de $Y - X_2\hat{\beta}_2$ sur X_1 peut donc aussi s'écrire :

$$Y - X_2\hat{\beta}_2 = X_1\hat{\beta}_1 + \hat{U}$$

On utilise alors la condition $X_1'\hat{U} = 0$ pour montrer la deuxième assertion.

Bibliograophie

[1] Abowd, J.M., et D., Card, (1989), "On the Covariance Structure of Earnings and Hours Changes", *Econometrica*, 57:411-45.

[2] Abowd, J.M., F., Kramarz et D.N., Margolis, (1999), "High Wage Workers and High Wage Firms", *Econometrica*, 67:251-333.

[3] Ahn, S.C., P., Schmidt, (1995), "Efficient Estimation of Models for Dynamic Panel Data", *Journal of Econometrics,* 68:5-27.

[4] Ahn, S.C., P., Schmidt, (1997), "Efficient Estimation of Models for Dynamic Panel Data: Alternative Assumptions and Simplified Estimation", *Journal of Econometrics, 68:5-27.*

[5] Amemiya, T., et T.E., MaCurdy, (1986), "Instrumental-Variable Estimation of an Error-Components Model", *Econometrica*, 54: 869-80.

[6] Anderson, T.W, et C., Hsiao, (1982), "Formulation and Estimation of Dynamic Models Using Panel Data", *Journal of Econometrics*, 18: 47-82.

[8] 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: 277-97.

[9] Arellano, M. et O., Bover, (1995), "Another Look at the Instrumental Variable Estimation of Error-Components Models", *Journal of Econometrics,* 68:29-51.

[10] Arellano, M. et B., Honoré, (2001), "Panel Data Models: Some Recent Developments", in eds Heckman J.J et E. Leamer, *Handbook of Econometrics*, V, North Holland, Amsterdam, à paraî tre.

[11] Balestra, P., et M., Nerlove, (1966), "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: the Demand for Natural Gas", *Econometrica*, 34: 585-612.

[12] Baltagi, B.H., (1995), *Econometric Analysis of Panel Data,* Wiley: London.

[13] Bhargava, A., et J.D., Sargan, (1983), "Estimating Dynamic Random Effects Model from Panel Data Covering Short Time Periods", *Econometrica,* 51: 1635-59.

[14] Blundell, R. et S., Bond, (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models", *Journal of Econometrics,* 87:115-143.

[15] Blundell, R., et R., Smith, (1991), "Conditions initiales et estimation efficace dans les modèles dynamiques sur données de panel", *Annales d'Economie et Statistiques,* 20/21:109-124.

[16] Breusch, T.S, G.E., Mizon et P., Schmidt, (1989), "Efficient Estimation Using Panel Data", *Econometrica*, 57: 695-700.

[17] Card David et Dean Hyslop, (1995), " Does inflation grease the wheels of the labor market ? " NBER Working Paper n° 5538.

[18] Collado, M.D., (1997), "Estimating Dynamic Models from Time-Series of Independent Cross-Sections", *Journal of Econometrics,* 82:37-62.

[19] Deaton, A., (1985), "Panel Data from Time Series of Cross Sections", *Journal of Econometrics,* 30:109-26.

[20] Dormont, C., (1989), *Introduction à l'économétrie des données de panel,* Editions du CNRS: Paris.

[21] Gouriéroux, A. et A., Monfort, (1989), *Statistique des modèles économétriques*, Economica: Paris

[22] Hahn, J., (1997), "Efficient Estimation of Panel Data Models", *Journal of Econometrics,* 79:1-21.

[23] Hausman, J.A., et T., Taylor, (1981), "Panel Data And Unobservable Fixed Effects", *Econometrica*, 49: 1377-98.

[24] Holtz-Eakin, D., W., Newey, et H.S., Rosen, (1988), "Estimating Vector Autoregressions with Panel Data", *Econometrica,* 56: 1371-95.

[25] Hsiao, C., (1986), *Analysis of Panel Data,* Cambridge U.P: Cambridge.

[26] Keane, M.P., et D.E., Runkle, (1992), "On the Estimation of Panel Data Models with Serial Correlation when Instruments are not Strictly Exogenous", *Journal of Business and Economic Statistics*, 10: 1-9.

[27] Kiviet, J., (1995), "On Bias, Inconsistency and Efficiency of Various Estimators in Dynamic Panel Data Models", *Journal of Econometrics,* 68: 53-78.

[28] Lollivier, S. et J.F. Payen, (1990), "L'hétérogénéité des carrières individuelles mesurées sur données de panel", *Economie et Prévision,* 92-93:89-95.

[29] Margolis, D.N., (1996), "Cohort Effects and Returns to Seniority in France", *Annales d'Economie et de Statistique*, 41-42:443-64.

[30] Moulton, B.R., (1987), "Diagnostics for Group Effects in Regression Analysis", *Journal of Business and Economic Statistics*, 5: 275-82.

[31] Mundlak, Y., (1978), "On the Pooling of Time Series and Cross-Section Data", *Econometrica*, 46: 69-85.

[32] Nickell, S., (1981), "Biases in Dynamic Models with Fixed Effects", *Econometrica*, 49: 1417-26.

[33] Robin, J.M., (2000), "Modèles structurels et variables explicatives endogènes", Méthodologie Statistique, DSDS-INSEE: Paris.

[34] Sevestre, P., et A., Trognon, (1985), "A Note on Autoregressive Error Components Models", *Journal of Econometrics, 28:231-245.*

[35] Verbeek, M., et T., Nijman, (1992), "Incomplete Panels and Selection Bias", in L., Matyas et P., Sevestre, *The Econometrics of Panel Data,* Kluwer: Amsterdam, 263-302.

[36] Verbeek, M., et T., Nijman, (1993), "Minimum MSE Estimation of a Regression Model with Fixed Effects from a Series of Cross-Sections", *Journal of Econometrics,* 59:125-36.

[37] Wansbeek, T., et A., Kapteyn, (1989), "Estimation of the Error-Components Model with Incomplete Panels", *Journal of Econometrics,* 41: 341-61.

[39] Wooldridge, J.M., (1996), "Estimating Systems of Equations with Different Instruments for Different Equations", *Journal of Econometrics,* 74:387-405.