Méthodologie statistique

M 2015/02

Les méthodes de PSEUDO-PANEL

Marine GUILLERM

Document de travail

Institut National de la Statistique et des Etudes Economiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail « Méthodologie Statistique »

de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2015/02

Les méthodes de PSEUDO-PANEL

Marine GUILLERM*

Ce document a bénéficié des commentaires, corrections et remarques de nombreuses personnes. Je les en remercie et tout particulièrement Pauline Givord pour son aide et son soutien tout au long de ce travail, ainsi que Simon Beck, Didier Blanchet, Richard Duhautois, Bertrand Garbinti, Stéphane Gregoir, Ronan Le Saout, Simon Quantin et Olivier Sautory. Je reste seule responsable des erreurs qui pourraient y demeurer.
Je remercie également Pierre Lamarche pour son aide sur l'utilisation des enquêtes Patrimoine.
.

*DMCSI
18, bd Adolphe Pinard - 75675 PARIS CEDEX 14

Direction de la méthodologie et de la coordination statistique et internationale -Département des Méthodes Statistiques - Timbre L101 - 18, bd Adolphe Pinard - 75675 PARIS CEDEX - France -
Tél. : 33 (1) 41 17 66 01 - Fax : 33 (1) 41 17 66 33 - CEDEX - E-mail : L001-dg@insee.fr - Site Web Insee : http://www.insee.fr

Ces documents de travail ne reflètent pas la position de l'Insee et n'engagent que leurs auteurs.
Working papers do not reflect the position of INSEE but only their author's views.

Les méthodes de pseudo-panel

Marine Guillerm *

Résumé

Les méthodes de pseudo-panel sont une alternative à l'utilisation de données de panel, lorsque seules des données en coupes répétées indépendantes sont disponibles. Elles sont très couramment utilisées pour des analyses en cycle de vie ou des estimations d'élasticités-prix ou revenus. Il s'agit de suivre, plutôt que des individus, des cohortes, c'est-à-dire des groupes stables d'individus. Sous certaines conditions de constitution des cohortes, les méthodes d'estimation classiques des données de panel peuvent être utilisées. Ce document détaille les étapes d'une estimation en pseudo-panel : de la constitution des cohortes jusqu'à l'estimation. Des exemples de programmes sous les logiciels SAS, R et STATA sont présentés, ainsi qu'une application sur les données des enquêtes Patrimoine.

Mots clés : Pseudo-panel, données groupées, modèles à effets fixes, données en coupes répétées

Abstract

Pseudo-panel methods are an alternative to the use of panel data when only repeated cross-sections are available. They are commonly used for life-cycle models, price elasticies or wage elasticities estimations. Pseudo-panel methods follow cohorts, meaning groups of individuals that are stable over time, rather than individuals. Under appropriate conditions in constructing these cohorts, classic panel data models can be used. This document explains in details pseudo-panel estimation : from the composition of the cohorts to the estimation. Examples of SAS, R and STATA programs are presented, and finally an application on successive waves of the Household Wealth Surveys.

Key words : Pseudo-panel, grouped data, fixed effect models, repeated cross-sections

*INSEE, Département des Méthodes Statistiques. marine.guillerm@insee.fr

Table des matières

Introduction

L'analyse économique des comportements se heurte généralement au fait que de nombreuses dimensions importantes pour l'analyse ne sont pas observables dans les données disponibles. Par exemple, les comportements de consommation dépendent de préférences individuelles qui ne sont qu'imparfaitement captées dans les données statistiques. Les estimations d'élasticités-revenu peuvent alors s'en trouver biaisées. Parfois, il est difficile de dissocier les effets de plusieurs variables alors même qu'elles sont observées simultanément. Ainsi, bien que l'âge et la génération soient couramment disponibles, il sera impossible à partir d'une source de données "en coupe" (à une date donnée) de distinguer ce qui relève de l'un ou de l'autre. C'est particulièrement dommageable pour des analyses en cycle de vie. Supposons que l'on s'intéresse aux carrières salariales au cours de la vie, que l'on tenterait de décrire à partir d'une seule enquête. Celle-ci permet bien d'observer des personnes à des âges différents, et donc à des moments successifs de leur vie professionnelle. Mais il ne sera pas possible de distinguer ce qui dans l'évolution observée du salaire s'explique effectivement par un effet d'âge (ou d'expérience professionnelle acquise) d'un effet génération. Ce dernier conditionne en partie le fait d'avoir fait des études plus ou moins longues, d'être entré sur le marché du travail à un moment plus ou moins propice... autant de facteurs qui peuvent aussi influer sur le salaire.

Il est classique d'utiliser des données de panel pour répondre à ces questions. À partir des observations répétées dans le temps d'unités identiques, on tente de neutraliser d'éventuelles spécificités individuelles. Cela se fait en général par l'introduction d'un "effet fixe" individuel censé capter ces spécificités. Le fait d'observer les mêmes variables à plusieurs dates peut être aussi un moyen de traiter en partie les problèmes d'identification décrits ci-dessus. L'âge varie avec le temps, contrairement à la génération ce qui permet de suivre une même génération à différents âges. Ces données sont cependant rares, souvent limitées à des échantillons de petite taille, et couvrent des périodes de temps réduites (ce qui diminue leur intérêt pour une analyse en cycle de vie par exemple). Elles sont en outre sujettes à des problèmes d'attrition ou de non-réponses : il est difficile de suivre les mêmes individus sur une longue période. Au fil du temps, la représentativité des données de panel peut devenir problématique.

Les méthodes de pseudo-panel peuvent constituer une manière de pallier l'absence de données de panel. Leur usage remonte à Deaton (1985) qui le premier a suggéré d'utiliser des méthodes de panel à partir de données en coupes répétées. L'avantage de ces données est qu'elles sont très souvent disponibles, et permettent de couvrir de longues périodes. En effet, de nombreuses enquêtes sont menées à des intervalles réguliers dans le temps. Elles constituent en général des données en coupes répétées indépendantes, au sens où elles portent sur des échantillons différents. Les méthodes de panel ne peuvent pas être directement appliquées, les individus observés changeant à chaque date. Même lorsque l'on dispose de sources exhaustives comme le recensement ou certaines données administratives, il n'est pas possible de suivre des personnes dans le temps par exemple pour des raisons de confidentialité. Cependant, à défaut de suivre des mêmes individus, on peut suivre des types d'individus, qu'on désigne généralement sous le terme de "cohortes" ou encore "cellules". Ces cohortes représentent des profils, identifiés par un ensemble de caractéristiques observées dans les données et stables dans le temps (comme la génération ou le sexe). Dans les estimations, on captera les spécificités inobservées qui pour-

raient biaiser les estimations par un effet fixe “cohorte”. Les pseudo-panels ont été utilisés pour modéliser des sujets aussi différents que l’investissement (Duhautois, 2001), la consommation (Gardes, 1999; Gardes et al., 2005; Marical et Calvet, 2011), ou encore l’évolution des comportements sur longue période, comme la carrière salariale (Koubi, 2003), l’activité féminine (Afsa et Buffeteau, 2005), le bonheur (Afsa et Marcus, 2008) ou le niveau de vie (Lelièvre et al., 2012). En pratique, la mise en œuvre de ces méthodes repose sur la manière de définir les cohortes. Dans le cas de modèles linéaires, les méthodes d’estimation classiques sur données de panel peuvent alors être adaptées assez simplement.

Ce document propose une introduction à ces techniques, en insistant sur les aspects pratiques. Après un bref rappel des estimations des modèles à effets fixes sur données de panel, il insiste sur les principes qui doivent guider le choix des critères définissant les cohortes. La seconde partie présente les techniques d’estimation, appuyées sur les procédures permettant de les appliquer dans les logiciels statistiques. Ces deux premières parties ne traitent que le cas des modèles linéaires. La troisième partie apporte des compléments techniques et évoque notamment l’extension aux modèles dichotomiques. Enfin, la dernière partie propose un exemple pratique tiré de l’exploitation des enquêtes Patrimoine.

1 Principe général : de l'effet fixe à l'effet cohorte

1.1 Pourquoi utiliser des données de panel, que faire en leur absence

Le point de départ des modèles de pseudo-panel sont les modèles linéaires à effets fixes, dont l'usage est classique lorsque l'on dispose de données de panel. Il est donc utile de les présenter (pour une présentation plus détaillée, voir par exemple Magnac, 2005). Typiquement, on souhaite modéliser l'influence d'une ou de variables explicatives sur une variable d'intérêt. On s'intéresse ici au cas où la variable d'intérêt est continue. Lorsqu'elle est discrète, il faut mobiliser des méthodes spécifiques (voir section 3.3). La difficulté de l'estimation de tels modèles vient en général du fait que tous les déterminants de cette variable d'intérêt ne sont pas observés. Si ces déterminants inobservés sont en partie corrélés aux variables explicatives du modèle, le risque existe d'attribuer à tort une partie de leur effet à ces variables explicatives. Par exemple, supposons qu'on souhaite étudier l'accumulation du patrimoine au cours du cycle de vie. Une analyse "naïve" peut consister, à partir d'observations à une date donnée, à étudier les différences de patrimoine selon l'âge. Cependant, de nombreuses autres caractéristiques individuelles peuvent expliquer les différences de patrimoine entre individus : la carrière salariale, le niveau d'étude, les ressources familiales, la plus ou moins grande propension à épargner... Certaines caractéristiques peuvent être corrélées à l'âge. Ce serait le cas par exemple si des générations ont connu des conditions d'entrée dans la vie active plus favorables. Ne pas tenir compte de ces déterminants risque de fournir des estimations biaisées de l'effet de l'âge sur le niveau de patrimoine. Une solution classique est d'introduire ces dimensions supplémentaires dans l'analyse (on "contrôle" de l'effet de ces variables), par un modèle linéaire. Cependant, si certains de ces déterminants sont couramment disponibles dans la plupart des enquêtes, tous ne le sont pas. On aura ainsi facilement une mesure de l'âge, du niveau d'étude ou du salaire actuel, mais il est moins fréquent de disposer d'indications précises sur l'ensemble de la carrière salariale, l'héritage dont les personnes interrogées ont pu bénéficier et encore moins s'ils sont plutôt "fourmi" ou "cigale" au sens d'une plus ou moins grande propension à épargner.

Une solution classique est alors d'utiliser des données de panel (c'est-à-dire des observations pour le même individu répétées dans le temps), qui permettent de contrôler de certains facteurs dont l'effet peut être supposé constant dans le temps. On ajoute alors un effet fixe individuel au modèle linéaire classique, censé capter l'effet des caractéristiques individuelles constantes dans le temps sur la variable d'intérêt :

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it} \quad i = 1,\ldots,N \quad t = 1,\ldots,T \tag{1}$$

où y_{it} est la variable d'intérêt (dans notre exemple, le patrimoine), x_{it} est un vecteur (ligne) de K variables explicatives observées sur l'individu i à la date t (dans notre exemple, l'âge, le niveau de salaire,...), β est l'effet de ces variables (soit un vecteur de paramètres de dimension K). α_i est l'effet fixe individuel. Il capte tous les déterminants de la variable d'intérêt fixes au cours du temps. En effet, seuls les paramètres associés à des variables non constantes dans le temps sont identifiables dès lors qu'on introduit un effet fixe dans le modèle. Par exemple, on ne pourra pas obtenir une estimation de l'effet intrinsèque du sexe si le modèle intègre un effet fixe. Enfin, ε_{it} est un terme résiduel, c'est-à-dire tout ce qui n'est pas pris en compte par le modèle (choc temporaire sur le patrimoine par exemple). Ignorer l'effet fixe dans l'estimation conduit à des

estimateurs biaisés de l'effet des variables explicatives considérées dès lors que ces variables sont corrélées à cet effet fixe.

Quand on dispose d'observations répétées, il est possible dans le cadre de ce modèle linéaire d'estimer l'impact des variables explicatives qui nous intéressent en neutralisant l'impact des effets fixes individuels. En pratique, cela se fait en utilisant non plus les variables en niveau mais des variables différentiées pour un même individu, de manière à faire disparaître l'effet fixe individuel. L'estimateur le plus couramment utilisé (car le plus efficace sous certaines hypothèses) est obtenu en procédant à une transformation "within" : on utilise à chaque date les observations centrées par rapport à la moyenne individuelle sur la période, c'est-à-dire les variables transformées $z_{it} - \bar{z}_i$, avec $\bar{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_{it}$ la moyenne des valeurs individuelles pour une variable z sur l'ensemble de la période d'observation. Une autre solution serait d'estimer directement les effets fixes comme des paramètres du modèle, mais cela suppose l'estimation d'un très grand nombre de paramètres (un effet fixe pour chacun des individus observés en plus des paramètres des variables explicatives), sans grand intérêt en terme d'interprétation [1].

Cet estimateur converge vers les vraies valeurs des paramètres d'intérêt dès lors que les variables explicatives ne sont pas corrélées aux termes résiduels restants. Dit autrement, il ne faut pas que les chocs individuels à chaque date, pour un individu donné, soient liés à la réalisation d'une des variables explicatives incluses dans le modèle [2].

Les méthodes de panel reposent cependant sur le fait d'observer les mêmes individus à des dates différentes, ce qui est rare. Dans de nombreux cas, on dispose de données en coupes transversales indépendantes répétées. Le principe des pseudo-panels est alors de suivre dans le temps non plus des individus, mais des cohortes, c'est-à-dire des groupes d'individus partageant un ensemble de caractéristiques fixes dans le temps. On va alors considérer le modèle au niveau de ces cohortes d'individus et non plus au niveau des individus qui les composent. En pratique, cela signifie qu'on remplace les variables observées par les moyennes de ces variables au sein de chaque cohorte. Ces données sont assimilées à des données de panel et, quand les conditions le permettent, les techniques d'estimation sur données de panel leur sont appliquées.

Les analyses en cycle de vie et les estimations d'élasticités-revenus ou d'élasticités-prix offrent de nombreux exemples d'utilisation des pseudo-panels. Ce sont spécifiquement des analyses où les données de panel sont rares. Les analyses en cycle de vie demandent de disposer de données sur longues périodes. Des séries de coupes transversales offrent plus souvent cet horizon temporel que des panels. Ceci peut justifier que même en présence de données de panel, on ait recours à des estimations par pseudo-panel. Par exemple, Antman et McKenzie (2005) disposent d'un panel rotatif pour analyser la mobilité salariale. Ne retenir à chaque trimestre que le cinquième des nouveaux entrants dans le panel leur permet de disposer de données sur longue période, alors qu'ils seraient limités à cinq trimestres avec le panel. De plus, contrairement aux panels, les pseudo-panels ne sont pas confrontés à l'attrition liée notamment à la difficulté de

1. D'autant moins que si on dispose de peu d'observations temporelles par individu, l'estimation des effets fixes est peu précise.

2. Rappelons que ce terme résiduel représente dans le modèle à effets fixes tous les facteurs individuels variables dans le temps que l'on n'observe pas.

suivre des ménages suite à un déménagement qui peut en outre être consécutif à une évolution salariale. Disposant des données de panel, Gardes et al. (2005) mènent des estimations d'élasticités-revenus sur données de panel et sur données de pseudo-panels. Ils montrent que les deux estimations sont proches.

Les modèles à effets aléatoires sont un autre type de modélisation classiquement mis en œuvre sur des données de panel. Contrairement aux modèles à effets fixes, ils reposent sur l'hypothèse que les caractéristiques individuelles inobservées ne sont pas corrélées aux covariables. Ces modèles permettent de tenir compte dans l'estimation avec des données de panel du fait que les résidus associés à un même individu aux différentes dates d'observation sont corrélés. Avec des coupes transversales indépendantes, il n'y a pas de corrélation entre les observations, chaque individu n'étant observé qu'une fois. Les modèles peuvent donc être estimés directement sur les données individuelles empilées. Les pseudo-panels n'ont pas d'utilité dans ce cas. On ne s'intéresse donc pas ici à l'estimation de tels modèles.

Formellement, on s'intéresse à $y^\star_{ct} = E(y_{it}|i \in c,t)$, espérance de la variable d'intérêt sur la cohorte c à la date t. On obtient en effet à partir du modèle précédent (en l'intégrant conditionnellement à la date et à la cohorte) :

$$y^\star_{ct} = x^\star_{ct}\beta + \alpha^\star_{ct} + \varepsilon^\star_{ct} \quad c = 1,\ldots,C \quad t = 1,\ldots,T \tag{2}$$

où pour chaque variable z, $z^\star_{ct} = E(z_{it}|i \in c,t)$.

Comme le modèle initial au niveau individuel, le modèle du pseudo-panel (2) est linéaire en ses paramètres, ce qui permet en principe d'appliquer les techniques d'estimation classiques des modèles de panel. Cependant, en pratique les choses sont un peu plus complexes.
Tout d'abord, les "vraies" valeurs $y^\star_{ct}$ et $x^\star_{ct}$ ne sont pas connues. On ne dispose que d'une estimation, leur contrepartie empirique au sein de la cohorte observée : $\bar{y}_{ct} = \frac{1}{n_{ct}} \sum_{i \in c,t} y_{it}$ et $\bar{x}_{ct} = \frac{1}{n_{ct}} \sum_{i \in c,t} x_{it}$ (c'est-à-dire, à chaque date, les moyennes des valeurs observées pour les individus de l'échantillon appartenant à cette cohorte). Pour bien comprendre la différence, il faut se rappeler qu'une cohorte est par définition constituée d'individus ayant le même profil. La "vraie" moyenne correspond à la moyenne calculée sur l'ensemble des personnes de la cohorte. À partir de données d'enquête cependant, on n'observera que quelques individus appartenant à cette cohorte. L'estimation à partir de ce (sous-)échantillon d'individus risque de ne pas correspondre exactement à la "vraie" valeur. Cette différence n'est pas anodine. En pratique, on estimera le modèle sur les contreparties empiriques $\bar{z}_{ct}$ et non sur les vraies valeurs $z^\star_{ct}$. On utilise donc des variables mesurées avec erreur, ce qui peut être problématique en terme de biais pour les estimateurs (pour plus de détails voir section 3.2). Le point positif est que cette moyenne empirique converge vers la vraie valeur : intuitivement, plus le nombre d'individus de la cohorte est grand dans l'échantillon d'observations, plus cette estimation sera proche de la vraie valeur et les estimateurs des valeurs moyennes suffisamment précis pour pouvoir négliger les erreurs de mesure dans le modèle économétrique.

La fluctuation d'échantillonnage des individus d'une cohorte d'une date à une autre constitue une deuxième difficulté pour l'estimation du modèle (2). À chaque date, les individus observés ne sont pas les mêmes. La moyenne des effets fixes $\bar{\alpha}_{ct}$ est ainsi susceptible de varier

au cours du temps, alors qu'elle est en théorie fixe dans le temps [3]. Ces remarques orientent les critères qu'on retiendra pour définir les cohortes d'individus.

1.2 Comment constituer les cohortes ?

En premier lieu, le critère de regroupement doit être observable pour l'ensemble des individus et former une partition de la population (chaque individu est classé dans exactement une cohorte). Au-delà de ces évidences, le critère de constitution des cohortes ne peut pas être choisi au hasard. Il doit viser à rendre plausible l'hypothèse que les termes de cohorte $\bar{\alpha}_{ct}$ sont effectivement fixes au cours du temps. Deux facteurs distincts peuvent remettre en cause cette hypothèse. Pour bien le comprendre, il est utile de distinguer les "vraies" cohortes des cohortes observées. Une "vraie" cohorte est l'ensemble des individus de la population de cette cohorte. Avec des données d'enquête, seul un échantillon est observé. La première source de variation de $\bar{\alpha}_{ct}$ est liée aux fluctuations d'échantillonnage : $\bar{\alpha}_{ct}$ correspond à la moyenne des effets fixes sur les observations de la cohorte c de l'échantillon disponible à la date t. Il s'agit d'un estimateur de la "vraie" valeur $\alpha^{\star}_{ct}$, inobservée. Même si la "vraie" cohorte est stable, les individus la représentant changent d'une date à une autre. $\alpha^{\star}_{ct}$ peut aussi varier si la "vraie" cohorte regroupe une population mouvante au cours du temps, notamment si le critère retenu ne correspond pas à une caractéristique stable des individus. Il s'agit de la deuxième source de variation possible de $\bar{\alpha}_{ct}$.

1.2.1 Un critère stable sur une population stable

Choisir un critère de constitution des cohortes de sorte à rendre $\alpha^{\star}_{ct}$ constant dans le temps permet d'éliminer dans une certaine mesure une des sources de variation de $\bar{\alpha}_{ct}$. $\alpha^{\star}_{ct}$ est fixe lorsque les "vraies" cohortes regroupent à chaque date les mêmes individus. Deux conditions sont requises : définir les cohortes sur une population stable et sur la base d'un critère stable (cela signifierait sinon que les personnes pourraient changer de profil au cours du temps).

L'année de naissance est évidemment un exemple de critère de regroupement qui correspond à une caractéristique stable des individus. Dans ce cas, on suit des générations d'individus. Ce critère est très fréquemment retenu dans les estimations par pseudo-panel. Le terme cohorte ne doit pas laisser penser que seul ce critère est valide (certains auteurs utilisent le terme "cellule"). D'autres regroupements sont possibles. Plusieurs critères peuvent aussi être combinés. Par exemple Bodier (1999) forme des cohortes par génération et diplôme de fin d'étude pour étudier les effets d'âge sur le niveau et la structure de consommation des ménages. Gardes et al. (2005) combinent âge et niveau d'éducation pour estimer par pseudo-panel des élasticités-revenus. À l'inverse, un critère de regroupement fondé sur le salaire ou la situation sur le marché du travail ne serait *a priori* pas pertinent : il est susceptible de changer pour une même personne au cours du temps.

3. En pratique, avec un modèle à effet fixe cohorte estimé sur des échantillons, cela signifie que cette partie variable liée à la fluctuation d'échantillonnage "passe" dans le terme résiduel de l'équation.

Mais cette condition de stabilité du critère au niveau individuel n'est pas suffisante. Il faut aussi que la population qu'elle représente (la cohorte) n'évolue pas elle-même dans le temps. Cette question est particulièrement cruciale lorsque l'on s'intéresse à des données d'enquêtes répétées, sur des échantillons différents. Dans une enquête, les individus d'un certain profil constituent un échantillon de l'ensemble de la cohorte d'intérêt. Mais dans certains cas, leur représentation dans le champ de l'enquête peut varier en fonction des critères retenus pour définir la cohorte. Supposons par exemple qu'on constitue des cohortes à partir de l'année de naissance. Selon la date de l'enquête, les différentes générations seront plus ou moins bien représentées : elles rentreront progressivement en fonction de l'âge minimum requis pour être enquêté (ou de la prise d'indépendance des jeunes pour des enquêtes ménages), tandis qu'à l'inverse les plus âgées en sortiront progressivement (décès, départs en institutions spécialisées si celles-ci sont hors champ d'enquête). Il faut faire attention à ces effets de composition pour l'analyse, s'ils sont liés à la variable d'intérêt. Supposons par exemple qu'on s'intéresse au profil de revenu de générations successives. L'espérance de vie et le revenu sont en partie corrélés (voir par exemple Blanpain, 2011). Aux âges avancés, les personnes aux revenus les plus élevés sont donc sur-représentées parmi les "survivants" d'une même génération. Une analyse par cohorte qui suivrait une génération laisserait penser que le revenu des individus de cette génération augmente avec l'âge, ce qui n'est probablement pas le cas. En pratique, une analyse au cas par cas est nécessaire pour évaluer si les cohortes représentent au cours du temps une population stable, quitte à restreindre le champ de l'analyse. Par exemple, pour une analyse sur les effets d'âge et de génération sur le niveau et la structure de la consommation, Bodier (1999) restreint la population d'étude aux 25-84 ans, considérant que les ménages constitués de personnes au-delà de ces limites risquent de plus de ne pas être représentatifs de l'ensemble des personnes de leur génération.

Il faut souligner que ce problème n'est pas spécifique aux pseudo-panels, mais il est particulièrement apparent lors du suivi sur de longues périodes pour lesquelles ces phénomènes d'entrée-sortie (naissances, décès, migrations, etc.) sont susceptibles d'apparaître. En revanche, contrairement aux données de panels classiques, on n'est pas confronté à des problèmes d'attrition liés à la difficulté de suivre des individus identiques au cours du temps (déménagement, refus de répondre à nouveau).

1.2.2 Former des cohortes de taille suffisante...

Le principe des pseudo-panels est de constituer des cohortes, autrement dit des profils, regroupant des individus dont les comportements puissent être considérés comme proches. Cette hypothèse sera d'autant plus plausible qu'on définira des profils précis. Néanmoins, en particulier avec des données d'enquête, ceci peut avoir un coût. Pour chaque vague d'enquête, les individus d'une cohorte ne constituent qu'un échantillon de toutes les personnes de cette cohorte à cette date. Si les critères de constitution de cohorte sont très précis, on n'observera que très peu d'individus de la "vraie" cohorte (celle de l'ensemble de la population). Les moyennes empiriques pour une cohorte des variables observées (les $\bar{y}_{ct}$ et $\bar{x}_{ct}$), estimées à partir des représentants de cette cohorte dans l'échantillon, risquent de fournir une estimation très imprécise des vraies valeurs moyennes de l'ensemble des personnes de cette cohorte type (les $y^{\star}_{ct}$ et

$x^{\star}_{ct}$). Dit autrement, le risque existe que l'erreur due à l'échantillonnage soit souvent très grande.

Par souci de simplicité, on peut négliger ces erreurs en remplaçant $y^{\star}_{ct}$ et $x^{\star}_{ct}$ par leur moyenne empirique intra-cohorte $\bar{y}_{ct}$ et $\bar{x}_{ct}$. Cette approximation a le mérite de la simplicité mais fournit cependant des estimateurs biaisés (voir section 3.2 et annexe C). Il est possible de limiter ce biais en formant des cohortes suffisamment grandes, l'amplitude des erreurs d'échantillonnage diminuant avec la taille des cohortes. Ceci permet également de limiter les variations temporelles de $\bar{\alpha}_{ct}$ liées à l'échantillonnage. En pratique, dans les études empiriques, il est généralement considéré que le seuil de 100 individus par cohorte est suffisant pour négliger les erreurs d'échantillonnage (et donc simplifier l'estimation). Ce choix s'appuie en particulier sur les études de Verbeek et Nijman (1992, 1993). À partir de données simulées, ces derniers concluent que l'hypothèse est raisonnable (au sens où le biais qui en résulte n'est "pas trop" élevé) pour des catégories regroupant 100 individus minimum. Cependant, ils préconisent des tailles deux fois supérieures pour réduire significativement les risques de biais.

1.2.3 ... tout en conservant de la variabilité

L'amplitude des erreurs de mesure et le biais qu'elles génèrent se réduisent à mesure que la taille des cohortes augmente. Cependant, la taille des cohortes n'est pas le seul paramètre à prendre en compte. De fait, il est assez simple de se rendre compte qu'à taille d'échantillon total fixée, constituer de larges cohortes signifie qu'on réduit le nombre d'observations utilisées pour le modèle de pseudo-panel. Supposons par exemple que le critère de constitution des cohortes soit l'année de naissance, mais que les données en coupes répétées contiennent à chaque date peu de personnes d'une génération. Pour réduire les fluctuations d'échantillonnage qui risquent d'en découler, une solution classique est d'augmenter la taille des cohortes en formant des générations définies plus largement (par exemple, par tranche de cinq ans). Mais dans ce cas, la variabilité des observations à une date donnée se réduit, le nombre final d'observations utiles diminuant. En outre, regrouper des générations proches mais différentes signifie aussi qu'on réduit la variabilité au cours du temps de ces moyennes. Ces deux éléments (nombre d'observations utilisées pour l'estimation, faible variabilité) sont deux facteurs qui classiquement réduisent la précision de l'estimateur final. Intuitivement, moins on a d'observations et moins l'estimation est précise. Mais il est aussi nécessaire d'observer des valeurs différentes des grandeurs d'intérêt, autrement dit que ces valeurs varient dans le temps, pour mesurer la force de leur corrélation. On est donc confronté à un classique arbitrage biais - variance : former des cohortes de grande taille permet de limiter le biais de l'estimateur, mais fait perdre de la variabilité, de nature à réduire la précision des estimateurs. Verbeek et Nijman (1992) montrent que le biais de l'estimateur within peut être élevé même si les cohortes sont de grande taille si la variabilité inter-temporelle est faible par rapport aux erreurs de mesure.

En résumé, un bon critère de regroupement doit : (1) être une caractéristique qui ne change pas au cours du temps au niveau individuel, et définir une (sous-)population stable, et résulter d'un arbitrage permettant de (2) former des cohortes suffisamment grandes tout en (3) ne faisant pas perdre trop de variabilité. Ces différentes contraintes limitent fortement le choix des critères de constitution des cohortes. En pratique, de nombreuses études utilisent l'année de naissance car ce critère répond à beaucoup de ces contraintes. Il est très souvent disponible dans

les données, et il est stable. En outre, selon la taille des échantillons des enquêtes en coupe, il est possible de jouer sur le regroupement de générations proches pour constituer des cohortes plus ou moins grandes. Enfin, il ne faut pas négliger que cette dimension a un intérêt en tant que tel dans de nombreuses études. L'effet cohorte s'interprète ainsi directement comme un effet génération, qu'il peut être intéressant d'étudier. Dans les analyses en cycle de vie en particulier, regrouper les individus par génération permet de garder de la variabilité sur la variable "âge".

2 L'estimation des modèles en pseudo-panels

2.1 Les différentes méthodes d'estimation

Lorsque le critère de constitution des cohortes a les qualités requises pour considérer le modèle (2) comme un modèle de panel à effets fixes, l'estimation des paramètres repose généralement sur les techniques classiques d'estimation sur données de panel. En pratique, le modèle estimé est donc :

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \bar{\alpha}_c + \bar{\varepsilon}_{ct} \quad t = 1,\ldots,T \quad c = 1,\ldots,C \tag{3}$$

On peut donc appliquer une transformation "within" évoquée plus haut, dans laquelle, pour chaque cohorte, on centre les différentes variables autour de la moyenne des valeurs observées pour cette cohorte sur l'ensemble des dates d'observation. En pratique, on régresse donc $\bar{y}_{ct} - \bar{y}_c$ sur $\bar{x}_{ct} - \bar{x}_c$, où pour chaque variable z, $\bar{z}_c = 1/T \sum_{t=1}^{T} \bar{z}_{ct}$. Formellement, l'estimateur within est :

$$\hat{\beta}_W = \left[\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)'(\bar{x}_{ct} - \bar{x}_c) \right]^{-1} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)'(\bar{y}_{ct} - \bar{y}_c) \tag{4}$$

On peut en déduire un estimateur de l'effet cohorte :

$$\hat{\alpha}_c = \bar{y}_c - \bar{x}_c \hat{\beta}_W \tag{5}$$

L'estimateur within peut être obtenu de manière équivalente, soit en incluant des indicatrices de cohortes, soit par instrumentation. Inclure les indicatrices de cohortes dans le modèle (3) permet d'obtenir directement des estimateurs des effets fixes [4] qui peuvent avoir un intérêt en tant que tel. Dans une analyse en cycle de vie dans laquelle les cohortes seraient constituées des générations, on estime ainsi directement l'effet génération. Attention cependant, l'estimation de ces effets fixes ne sera précise que si le nombre de périodes d'observation est suffisant.

Une méthode d'estimation alternative par instrumentation est proposée par Moffitt (1993). Il montre que l'estimateur within (4) du modèle en pseudo-panel correspond techniquement à l'estimateur des doubles moindres carrés sur les données individuelles (variables explicatives

4. L'estimation directe des effets fixes est déconseillée avec des données individuelles, car elle demande d'estimer un très grand nombre de paramètres. Dans le cadre des pseudo-panels, le nombre de cohortes est en général limité. Si chaque cohorte regroupe environ 100 individus, le nombre d'effets fixes à estimer dans le modèle de pseudo-panel est divisé d'autant par rapport au modèle de panel.

ainsi que des indicatrices de cohortes), dans lequel on utiliserait comme instrument l'ensemble des indicatrices de cohortes croisées avec les indicatrices de temps. La preuve formelle est fournie en annexe A. Pour en saisir l'intuition, rappelons que dans la première étape des doubles moindres carrés, on projette les variables explicatives sur les instruments. La projection de x_{it} sur les indicatrices cohorte $\times$ date d'observation correspond exactement à la moyenne empirique $\bar{x}_{ct}$, c étant la cohorte à laquelle appartient l'individu i. La deuxième étape consiste à remplacer dans le modèle initial les variables instrumentées par leur projection, soit ici à régresser y_{it} sur $\bar{x}_{ct}$ et les indicatrices de cohortes. On obtient le même estimateur que l'estimateur within (4).
Ceci peut simplifier l'estimation : on travaille directement sur les données individuelles. Cette analogie sert également de base à l'extension des pseudo-panels aux modèles dichotomiques (voir section 3.3). Un autre intérêt de cette approche est que d'autres types d'instruments plus parcimonieux peuvent être utilisés. Par exemple, si l'année de naissance est retenue, on peut utiliser une fonction de l'année de naissance (un polynôme par exemple) pour construire l'instrument plutôt que des indicatrices associées à une partition des années de naissance.
On peut d'ailleurs remarquer que cette approche permet de retrouver là encore les critères de regroupement des individus en cohortes [5]. Rappelons que deux conditions sont requises pour définir un bon instrument. Il doit d'abord être corrélé aux variables explicatives. Ici, cela renvoie au fait que la constitution des cohortes doit conserver suffisamment de variabilité pour permettre l'estimation du modèle agrégé au niveau des cohortes. Pour en comprendre l'intuition, on peut penser au cas extrême où ces indicatrices croisées cohorte $\times$ date seraient totalement indépendantes des variables explicatives du modèle : dit autrement, que la distribution de ces variables explicatives est identique à chaque date, et d'une cohorte à l'autre. Dans ce cas les moyennes empiriques de ces variables au niveau d'une date et d'une cohorte sont très proches, ce qui signifie qu'on ne pourra estimer le modèle. L'autre propriété d'un instrument valide est qu'il ne doit pas être corrélé avec les déterminants inobservés de la variable d'intérêt. Moffitt montre que cette propriété est vérifiée si les cohortes sont définies sur un critère stable et quand la taille des cohortes tend vers l'infini.

Au-delà de l'estimation proprement dite, plusieurs remarques s'imposent. Tout d'abord, sur le choix des variables explicatives. Rappelons que dans le modèle classique linéaire à effets fixes, seuls les paramètres associés à des variables non constantes dans le temps sont identifiables : l'effet fixe "absorbe" l'effet des variables constantes. L'agrégation en cohorte peut créer artificiellement de la variabilité et donner l'impression que les paramètres associés aux caractéristiques fixes sont identifiables. Par exemple, une variable constante au niveau individuel telle que l'indicatrice "être une femme" devient dans les données du pseudo-panel "la proportion de femmes dans la cohorte c à la date t". Les variations temporelles observées (normalement faibles) ne sont dues qu'à l'erreur d'échantillonnage. Il est donc préférable de ne pas introduire ce type de variables dans l'analyse, au risque d'augmenter artificiellement un terme résiduel.
L'échantillonnage peut introduire de l'hétéroscédasticité dans le modèle, en particulier si le nombre d'individus observés varie fortement d'une cohorte à une autre à une date donnée, ou d'une date à une autre pour une même cohorte. En effet, la précision de l'estimation dépendant

5. Pour plus de détail, voir Moffitt (1993) et Verbeek (2008).

directement de ce nombre, on risque d'introduire des termes d'erreur plus ou moins importants selon les cohortes, ce qui remet en cause l'hypothèse d'homoscédasticité (tous les termes individuels résiduels de l'équation ont une variance identique en particulier). Ceci peut donc conduire à des estimations biaisées de la précision des estimateurs par pseudo-panel. On propose dans la section 3.1 une extension permettant de prendre en compte ce problème.

Par ailleurs, les techniques d'estimation ci-dessus reposent implicitement sur l'hypothèse que l'on puisse négliger les erreurs d'échantillonnage. Quand ce n'est pas le cas, on pourra être amené à utiliser des techniques appropriées (voir section 3.2). Enfin, les modèles présentés jusqu'à présent ne sont adaptés qu'au cas où la variable d'intérêt est continue. Lorsqu'elle est discrète, il faut mettre en œuvre des techniques d'estimation spécifiques qui sont présentées à la section 3.3.

2.2 En pratique avec les logiciels statistiques

En pratique les estimations en pseudo-panel consistent (1) à calculer les moyennes intra-cohortes et (2) à calculer l'estimateur within sur ces moyennes. Des exemples d'estimation avec les logiciels SAS, R et STATA sont présentés ci-dessous. On peut les tester en utilisant des données simulées dont les programmes de construction sont présentés en annexe B.

2.2.1 Sous SAS

La table `donnees` empile 5 coupes transversales indépendantes. Les variables `x1`, `x2`, `y`, `cohorte` et `t` représentent respectivement les deux variables explicatives, la variable d'intérêt, l'identifiant de la cohorte à laquelle l'individu appartient et la date d'observation. La table regroupe les données de 50 000 individus répartis sur cinq dates d'observation (10 000 individus par date).

1. Création des données de pseudo-panel
   La première étape de l'estimation en pseudo-panel consiste à calculer les moyennes intra-cohortes par la procédure `MEANS` munie de l'option `MEAN`. L'option `ODS` permet de créer en sortie la table SAS `pseudo` contenant les moyennes intra-cohortes.

   ```
   ODS OUTPUT SUMMARY=pseudo;
   PROC MEANS DATA = donnees MEAN;
   VAR x1 x2 y;
   CLASS cohorte t;
   RUN;
   ```

   La table `pseudo` regroupe ainsi les données du pseudo-panel. Les noms des variables de cette table sont ceux spécifiés dans le paramètre `VAR` avec le suffixe `_mean`, soit ici : `x1_mean`, `x2_mean`, `y_mean`.
2. Estimation des paramètres
   Comme indiqué à la section 2.1, deux méthodes équivalentes permettent d'estimer les paramètres du modèle (2), soit en procédant à la transformation within, soit en intégrant les indicatrices de cohortes dans le modèle.

(a) L'estimateur within
L'estimation peut se faire en utilisant la macro SAS `bwithin` développée par Duguet (1999). Celle-ci calcule les variables différentiées (centrées autour de la moyenne intra), et estime le modèle linéaire sur ces variables ainsi que la précision associée.

```
%bwithin(TAB=pseudo,Y=y_mean,X=x1_mean x2_mean,I=cohorte,T=t,
COVA=c,RES=u,VIEUX=vi,TRANSF=moy);
```

Les arguments de cette macro sont :
- `TAB=` : table contenant les données
- `Y=` : variable d'intérêt
- `X=` : liste des covariables
- `I=` : l'identifiant des cohortes
- `T=` : la variable de date d'observation
- `COVA=` : nom de la table des résultats de l'estimation robuste
- `RES=` : nom de la table des résidus
- `VIEUX=` : nom de la table des résultats de l'estimation non robuste
- `TRANSF=` : nom de la table des données après transformation within

On pourrait vouloir estimer le modèle en procédant d'abord à une transformation within et en calculant l'estimateur des moindres carrés sur ces variables centrées. Mais il faut faire attention car, du fait que l'on travaille sur des variables transformées, l'estimateur standard de la variance fourni par la procédure des moindres carrés ordinaires ne correspond pas directement à l'estimateur sans biais de la variance du modèle within. Il la surestime. Il faut tenir compte d'un facteur multiplicatif $(CT-K)/(CT-C-K)$ où C est le nombre de cohortes, T le nombre de périodes et K le nombres de variables explicatives. La macro `bwithin` tient compte de cette difficulté. On privilégiera donc son utilisation.

(b) Estimateur en intégrant les effets fixes cohortes
Intégrer dans le modèle des indicatrices de cohortes permet d'estimer de manière équivalente le modèle. La procédure `GLM` permet de procéder à l'estimation sur les données de la table `pseudo`.

```
ODS OUTPUT PARAMETERESTIMATES=betachap;
PROC GLM DATA = pseudo;
CLASS cohorte;
MODEL y_mean = x1_mean x2_mean cohorte/SOLUTION NOINT;
QUIT;
```

- Le paramètre `CLASS` spécifie la variable désignant la cohorte.
- `SOLUTION` pour afficher les estimations des paramètres.
- `NOINT` pour spécifier un modèle sans constante.
- `ODS` permet de créer en sortie une table SAS `betachap` contenant les estimations des paramètres et leurs écarts-types, ainsi que les p-valeurs et les statistiques de test de nullité des coefficients à 5%.

Lorsque le nombre de cohortes est très élevé et que l'estimation des effets cohortes n'est pas souhaitée, il est possible d'utiliser la commande ABSORB.

```
ODS OUTPUT PARAMETERESTIMATES=betachap;
PROC GLM DATA = pseudo;
ABSORB cohorte;
MODEL y_mean = x1_mean x2_mean;
QUIT;
```

3. Estimation alternative par instrumentation

L'estimation en ayant recours à l'instrumentation ne requiert pas le calcul des moyennes intra-cohortes. On travaille directement sur les données individuelles initiales (table donnees).
La première étape consiste à dichotomiser les variables catégorielles cohorte et t et à créer les indicatrices de cohorte croisées avec celles des dates d'observation. La macro indicatrice suivante procède à la création dans la table donnees de 100 indicatrices de cohortes sous le nom indiccohorte&i et de 500 indicatrices croisant indicatrices de temps et de cohortes sous le nom indic&t_&i.

```
PROC SQL NOPRINT;
SELECT COUNT(distinct cohorte),COUNT(distinct t) INTO :Cb,:tempsb
FROM donnees;
QUIT;
%let C=%sysfunc(strip(&Cb.));
%let temps=%sysfunc(strip(&tempsb.));
/* On crée ainsi deux macro-variables contenant le nombre de cohortes
et le nombre de dates d'observation */

%macro indicatrice;
DATA donnees;
SET donnees;
%do i =1 %to &C.;
indiccohorte&i.=(cohorte=&i.);
%do t =1 %to &temps.;
indic&t._&i.=(cohorte=&i.)*(t=&t.);
%end;
%end;
RUN;
%mend;
%indicatrice;
```

Ensuite, la procédure SYSLIN appliquée directement aux données individuelles de la table donnees calcule les estimations des paramètres. L'option 2sls est requise pour spécifier une estimation en deux étapes.

```
%macro instrument;
ODS OUTPUT ParameterEstimates=betachapsyslin;
PROC SYSLIN DATA=donnees 2sls;
ENDOGENOUS x1 x2 ;
INSTRUMENTS
%do i =1 %to &C.;
%do t =1 %to &temps.;
indic&t._&i. %end; %end;;
MODEL y = x1 x2 %do i =1 %to &C.; indiccohorte&i. %end; / NOINT;
RUN;QUIT;
%mend;
%instrument;
```

2.2.2 Sous R

Le tableau de données `donnees` contient les données de 50 000 individus répartis sur cinq dates d'observation (10 000 individus par date). Le programme de constitution de cette table se trouve en annexe page 38.

1. Création des données de pseudo-panel
   Les instructions suivantes visent à construire la table de données `pseudo` qui regroupe les moyennes intra-cohortes.

   ```
   > pseudo=aggregate(cbind(donnees$x1,donnees$x2,donnees$y),
                      by=list(cohorte = donnees$cohorte,t=donnees$t),mean)
   > colnames(pseudo)=c("cohorte","t","x1mean","x2mean","ymean")
   ```

   La table `pseudo` contient les variables :
   – `ymean`, `x1mean`, `x2mean` : moyennes intra-cohortes
   – `cohorte` : identifiant de la cohorte
   – `t` : date d'observation
2. Estimation des paramètres
   (a) L'estimateur within
   Le package `plm` est adapté à l'estimation sur données de panel. L'instruction `plm` permet de calculer l'estimateur within.

   ```
   > library(plm)
   > estimw=plm(formula=ymean ~ x1mean + x2mean,
                              data=pseudo,index=c("cohorte","t"))
   > summary(estimw)
   ```

   Les arguments sont :
   – `formula` : spécifie le modèle
   – `data` : la table où se trouvent les données.
   – `index` : indique les variables identifiant les cohortes puis la date d'observation.

   (b) Intégrer les effets fixes cohortes
   On peut également avoir recours à l'estimation classique par moindres carrés par l'instruction `lm` en intégrant les indicatrices de cohortes.

```
> estim=lm(ymean ~ x1mean + x2mean + factor(cohorte)-1,data=pseudo)
> summary(estim)
```

Il peut être nécessaire d'indiquer par `factor` que la variable désignant les cohortes est catégorielle. `-1` permet d'indiquer que le modèle est sans constante.

3. Estimation alternative par instrumentation
   L'estimation par instrumentation ne requiert pas le calcul des moyennes intra-cohortes. On travaille directement sur les données individuelles de la table `donnees`. Pour procéder à cette estimation, on peut utiliser par exemple la fonction `ivreg()` du package `AER`.

```
> donnees$cohortet=with(donnees,paste(cohorte,t,sep="-"))
> ivreg(y ~ x1 + x2 + factor(cohorte)-1| factor(cohortet),data=donnees)
```

2.2.3 Sous STATA

Le tableau de données contient les données de 50 000 individus répartis sur 5 dates. Le programme de constitution de cette table se trouve en annexe page 40.

1. Création des données de pseudo-panel
   Les instructions suivantes visent à construire la table de données regroupant les moyennes intra-cohortes.

```
bysort cohorte t: egen x1mean=mean(x1)
bysort cohorte t: egen x2mean=mean(x2)
bysort cohorte t: egen ymean=mean(y)
bysort cohorte t: egen nobs=count(ident)
keep cohorte t x1mean x2mean ymean nobs
by cohorte t: keep if _n == 1
```

   La table `pseudo` contient les variables :
   – `ymean`, `x1mean`, `x2mean` : moyennes intra-cohortes
   – `cohorte` : identifiant de la cohorte
   – `t` : date d'observation
2. Estimation des paramètres

   (a) L'estimateur within
   La commande `xtreg` est adaptée à l'estimation sur données de panel. L'option `fe` (fixed effect) permet de calculer l'estimateur within.

```
xtset cohorte t
xtreg ymean x1mean x2mean , fe
```

   (b) Intégrer les effets fixes cohortes
   On peut également avoir recours à l'estimation classique par moindres carrés par l'instruction `regress` en intégrant les indicatrices de cohortes.

```
set matsize 600
regress ymean x1mean x2mean i.cohorte
```

   Il peut être nécessaire d'indiquer le préfixe `i.` pour que la variable soit considérée comme catégorielle.

3. Estimation alternative par instrumentation
   L'estimation par instrumentation ne requiert pas le calcul des moyennes intra-cohortes. On travaille directement sur les données individuelles de la table `donnees`. La commande `ivreg` permet de procéder à cette estimation.

```
set matsize 3000
forv i=1/10{
forv t=1/5{
gen indic'i'_'t'=(t=='t')*(cohorte=='i')
}
}

forv i=1/10{
gen varcohorte'i'=(cohorte=='i')
}

ivreg y (x1 x2=indic* ) varcohorte*,first
```

3 Compléments techniques

Cette section propose deux extensions au cadre classique qui traitent des difficultés techniques que peuvent soulever les estimations en pseudo-panel évoquées dans la section précédente : la prise en compte de (1) l'hétéroscédasticité des termes résiduels, et (2) celle des erreurs de mesure dans l'estimation. Enfin, on introduit les méthodes d'estimation de modèles binaires.

3.1 Hétéroscédasticité dans les pseudo-panels

En pratique, la taille des cohortes varie d'une cohorte à une autre et pour une même cohorte, d'une date à une autre. Cela n'est pas sans conséquence. Ces variations de taille sont susceptibles de créer de l'hétéroscédasticité dans le modèle (2). En présence d'hétéroscédasticité, l'estimateur within (4) est sans biais mais l'estimateur de sa précision est biaisé et par conséquent les statistiques de test sont invalides.

L'estimateur intra efficace est obtenu en pondérant les observations par la taille de la cohorte, soit en estimant par les moindres carrés le modèle suivant :

$$\sqrt{n_{ct}}\bar{y}_{ct} = \sqrt{n_{ct}}\bar{x}_{ct}\beta + \sum_{c=1}^{C} \sqrt{n_{ct}}\alpha_c \mathbb{1}_c + \sqrt{n_{ct}}\bar{\varepsilon}_{ct} \tag{6}$$

En pratique, sous le logiciel SAS, on a recours à l'instruction `weight` de la procédure `GLM` :

```
ODS OUTPUT PARAMETERESTIMATES=betachap_pond3;
PROC GLM DATA = pseudo;
CLASS cohorte;
MODEL y_mean = x1_mean x2_mean cohorte / SOLUTION NOINT;
WEIGHT nobs;
QUIT;
```

La variable `nobs` indique le nombre d'individus observés dans la cohorte. Elle est calculée directement dans la procédure `MEANS`.

Ce modèle nécessite l'estimation de $K+C$ paramètres. Cette estimation est facile à mettre en œuvre sauf si le nombre de cohortes est trop important, auquel cas, on souhaite en général procéder à une transformation within pour éliminer les effets fixes avant estimation. Mais dans ce modèle, une transformation within classique n'élimine pas les indicatrices de cohorte car le poids qui est affecté à chaque cohorte (n_{ct}) varie dans le temps. Gurgand (1997) montre que dans ce cas l'estimateur intra est :

$$\hat{\beta}_{WP} = \left(X'(WDW)^{-}X\right)^{-1}\left(X'(WDW)^{-}y\right) \tag{7}$$

où X matrice de dimension $CT \times K$ empile les vecteurs lignes $\bar{x}_{ct}$, y vecteur de dimension CT empile les valeurs $\bar{y}_{ct}$, $(WDW)^{-}$ est l'inverse généralisée de la matrice WDW, où W est la matrice within classique de dimension CT et D est la matrice diagonale dont les éléments diagonaux sont $1/n_{ct}$.

Sous SAS, un exemple de calcul de l'estimateur (7) :

```
PROC SORT DATA = pseudo;BY cohorte t;RUN;

PROC IML;
USE pseudo ;
READ all VAR{Nobs} INTO Nobs;
READ all VAR{x1_mean,x2_mean} INTO X;
READ all VAR{y_mean} INTO Y;

D=DIAG(1/Nobs);
/* Matrice within: */
W=I(%sysevalf(&C.*&temps.))- I(&C.)@ J(&temps.,&temps.,%sysevalf(1/&temps.));

/* Estimateur: */
betaWP=INV(T(X)*GINV(W*D*W)* X)*(T(X)*GINV(W*D*W)*Y) ;
PRINT betaWP;

QUIT;
```

Ce programme fait suite à ceux présentés au paragraphe 2.2.1, dans lesquels sont définies la table `pseudo` et les macro-variables `temps` et `C`.

3.2 Modèle à erreurs de mesure

3.2.1 Principe

Dans les pseudo-panels, les cohortes sont suivies au cours du temps à travers les moyennes intra-cohortes. Les coupes transversales indépendantes n'offrent qu'un estimateur de ces

moyennes qui sont observées avec erreur. Les estimations du modèle (2) proposées par Deaton (1985) reposent ainsi sur des modèles à erreurs de mesure qui prennent en compte ce problème. Il reprend la théorie développée par Fuller (1986) et l'adapte à l'estimation en pseudo-panels. Pour rappel, les "vraies" moyennes sont notées $y^{\star}_{ct}$ et $x^{\star}_{ct}$ et sont inobservées. $\bar{y}_{ct}$ et $\bar{x}_{ct}$ sont leurs estimateurs respectifs. Ils sont entachés d'erreurs de mesure (qui correspondent aux erreurs d'échantillonnage) notées u_{ct} et v_{ct} :

$$\bar{y}_{ct} = y^{\star}_{ct} + u_{ct} \tag{8}$$
$$\bar{x}_{ct} = x^{\star}_{ct} + v_{ct} \tag{9}$$

En remplaçant ces équations dans le modèle (2), on trouve :

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \alpha_c + \tilde{\varepsilon}_{ct} \quad c = 1,\dots,C \quad t = 1,\dots,T \tag{10}$$

avec $\tilde{\varepsilon}_{ct} = \varepsilon^{\star}_{ct} + u_{ct} - v_{ct}\beta$.

Les erreurs de mesure ne sont pas neutres pour l'estimation. Elles créent de la corrélation entre $\bar{x}_{ct}$ et le résidu $\tilde{\varepsilon}_{ct}$ (voir annexe C page 41). L'estimateur within (4) est donc biaisé.

3.2.2 Estimation du modèle

L'estimateur du paramètre β proposé par Verbeek et Nijman (1993) repose sur une spécification paramétrique de l'erreur de mesure et de sa corrélation avec la variable d'intérêt [6] (pour plus de détails, voir annexe C page 41). Il vaut :

$$\tilde{\beta} = \left(\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)'(\bar{x}_{ct} - \bar{x}_c) - \frac{T-1}{T} \frac{1}{n} \hat{\Sigma} \right)^{-1} \left(\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)'(\bar{y}_{ct} - \bar{y}_c) - \frac{T-1}{T} \frac{1}{n} \hat{\sigma} \right) \tag{12}$$

Σ et σ correspondent respectivement à la matrice de variance-covariance des erreurs de mesure de $x^{\star}_{ct}$ et à la covariance entre les erreurs de mesure de $x^{\star}_{ct}$ et $y^{\star}_{ct}$. Elles ne sont en général pas connues. Deaton propose de les estimer à partir des données individuelles :

$$\hat{\Sigma} = \frac{1}{CT} \sum_{t=1}^{T} \sum_{c=1}^{C} \hat{\Sigma}_{ct} \quad \text{avec } \hat{\Sigma}_{ct} = \frac{1}{n-1} \sum_{i \in c,t} (x_{it} - \bar{x}_{ct})'(x_{it} - \bar{x}_{ct}) \tag{13}$$

$$\hat{\sigma} = \frac{1}{CT} \sum_{t=1}^{T} \sum_{c=1}^{C} \hat{\sigma}_{ct} \quad \text{avec } \hat{\sigma}_{ct} = \frac{1}{n-1} \sum_{i \in c,t} (x_{it} - \bar{x}_{ct})'(y_{it} - \bar{y}_{ct}) \tag{14}$$

6. Cet estimateur diffère de l'estimateur proposé par Deaton (1985) :

$$\hat{\beta}_D = \left(\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)'(\bar{x}_{ct} - \bar{x}_c) - \frac{1}{n} \hat{\Sigma} \right)^{-1} \left(\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)'(\bar{y}_{ct} - \bar{y}_c) - \frac{1}{n} \hat{\sigma} \right) \tag{11}$$

L'estimateur de Verbeek et Nijman est préféré à celui proposé par Deaton. L'estimateur de Deaton n'est en effet pas convergent à T fixé. Il converge néanmoins vers l'estimateur de Verbeek et Nijman quand T tend vers l'infini. Lorsque le nombre de dates d'observations est faible (ce qui est fréquemment le cas), la différence entre les deux estimateurs peut être importante. De plus, la variance de l'estimateur de Deaton est plus élevée que celle de l'estimateur de Verbeek et Nijman.

3.2.3 Convergence des estimateurs

Plusieurs types de convergences peuvent être envisagées dans le cas des estimations en pseudo-panels, car plusieurs paramètres entrent en jeu : N le nombre d'individus observés à chaque date, C le nombre de cohortes constituées, n_{ct} la taille des cohortes formées et T le nombre de dates d'observation.

On peut s'intéresser à la convergence des estimateurs quand la taille des cohortes augmente. Intuitivement, les moyennes intra-cohortes sont des estimateurs des vraies moyennes intra-cohorte d'autant plus précis que la taille des cohortes est grande. Les erreurs de mesure deviennent négligeables et le modèle (2) peut être estimé par l'estimateur within classique présenté dans la section 2.1.

L'estimateur de Verbeek et Nijman a une variance plus élevée que l'estimateur within. On est donc confronté à un classique arbitrage biais-variance. L'estimateur within a un biais asymptotique quand la taille des cohortes est fixée mais une variance plus faible que l'estimateur de Verbeek et Nijman (pour plus de détails voir Verbeek et Nijman, 1993).

3.2.4 En pratique avec les logiciels statistiques

Le programme ci-dessous est un exemple d'estimation de modèle en tenant compte des erreurs de mesure. Il reprend les données simulées utilisées à la section 2.2 ainsi que la table `pseudo` et les macro variables créées.

1. Estimation de la matrice de variance-covariance des erreurs de mesure

```
PROC SORT DATA = donnees;BY cohorte t;RUN;

ODS OUTPUT COV=cov;
PROC CORR DATA = donnees COV VARDEF=df;
VAR y x1 x2;
BY cohorte t;
RUN;

PROC SORT DATA = cov;BY cohorte t;RUN;
```

La table `cov` regroupe pour chaque cohorte c à la date t les estimations $\hat{\Sigma}_{ct}$ et $\hat{\sigma}_{ct}$.

```
ODS OUTPUT SUMMARY=sigma;
PROC MEANS DATA = cov MEAN;
VAR x1 x2 y;
CLASS variable;
RUN;

DATA sigma;
SET sigma;
```

```
x1_meanb=x1_mean/%eval(&taille.-1);
x2_meanb=x2_mean/%eval(&taille.-1);
y_meanb=y_mean/%eval(&taille.-1);
KEEP variable x1_meanb x2_meanb y_meanb;
RUN;
```

2. Calcul de l'estimateur (12)

   On procède à une transformation within des données. La table `wpseudo` contient les moyennes intra-cohortes centrées.

```
PROC SORT DATA = pseudo;BY group;run;
PROC STANDARD DATA = pseudo OUT = wpseudo MEAN=0;
VAR y_mean x1_mean x2_mean;
BY cohorte ;
RUN;
```

   L'estimateur (12) est calculé sous SAS/IML :

```
DATA gsigma;
SET sigma;
WHERE variable IN ("x1","x2");
KEEP x1_meanb x2_meanb;
RUN;

DATA psigma;
SET sigma;
WHERE variable IN ("x1","x2");
KEEP y_meanb;
RUN;

PROC IML;

USE gsigma;
READ ALL INTO gsigma;
USE psigma;
READ ALL INTO psigma;

USE wpseudo;
READ all var {cohorte} into cohorte;
nbcohorte=NCOL(UNIQUE(cohorte));

READ ALL VAR {x1_mean x2_mean} INTO x;
READ ALL VAR {y_mean} INTO y;

xpx=1/(nbcohorte#&temps.)# T(x)*x- %SYSEVALF((&temps.-1)/&temps.)#gsigma;
```

```
xpy=1/(nbcohorte#&temps.)#T(x)*y-%SYSEVALF((&temps.-1)/&temps.)#psigma;

betachap=INV(xpx)*(xpy) ;

PRINT betachap;
QUIT;
```

3.3 Estimation de modèles dichotomiques

Les estimateurs précédents ne sont valables que pour des modèles linéaires. Ils sont adaptés au cas où la variable d'intérêt est continue. Lorsque cette variable d'intérêt est dichotomique (ie. binaire), il faut faire appel à des techniques d'estimation spécifiques. L'extension des méthodes d'estimations par pseudo-panel n'est pas immédiate.

Avant de détailler les stratégies qui peuvent être mises en œuvre dans ce cas, il est utile de rappeler le principe général de l'estimation dans le cas de la modélisation d'une variable binaire. Il est classique alors de supposer que cette variable dépend d'une variable latente $y^{\star}_{it}$ dont on suppose une dépendance linéaire en fonction des différentes caractéristiques [7] :

$$y^{\star}_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it} \tag{15}$$

$$y_{it} = \begin{cases} 1 & \text{si } y^{\star}_{it} \geq 0 \\ 0 & \text{sinon.} \end{cases} \tag{16}$$

La variable $y^{\star}_{it}$ n'est pas observée. On observe la variable binaire y_{it}. x_{it} est le vecteur des variables explicatives, α_i est l'effet fixe individuel (lié à la présence de caractéristiques inobservables) et ε_{it} est le terme d'erreur. On suppose en général que le terme d'erreur suit une loi normale (modèle probit) ou logistique (modèle logit). L'estimation de ce modèle paramétrique se fait alors par maximum de vraisemblance. Cependant, du fait de l'effet fixe individuel, l'estimation est plus compliquée. Même avec de "vraies" données de panel (c'est-à-dire lorsqu'on suit le même individu au cours du temps), aucune des deux méthodes utilisées dans le cas du modèle linéaire ne s'applique directement. Du fait de la non linéarité, procéder à une transformation within n'a pas de sens. Estimer le modèle en incluant les indicatrices individuelles n'est pas correct dès lors qu'on a, comme souvent, peu d'observations pour le même individu : les effets fixes individuels seront mal estimés mais surtout, du fait de la non linéarité, cette mauvaise estimation conduit à biaiser l'estimation des autres paramètres (on parle de problème des paramètres incidents, pour plus de détails voir par exemple Davezies (2011)).

Dans le cas spécifique des pseudo-panels, dans lequel on dispose non plus de données de panels mais d'une série de coupes transversales indépendantes, la difficulté supplémentaire de l'estimation des modèles dichotomiques vient de la non linéarité : il n'est plus possible de déduire une relation simple entre les moyennes intra-cohortes de la variable d'intérêt d'une part et des covariables d'autre part. D'autres stratégies d'estimation ont été proposées dans la littérature.

7. Pour une présentation des modèles dichotomiques classiques voir par exemple Le Blanc et al. (2000).

La méthode proposée par Moffitt (1993) repose sur le parallèle établi entre estimation en pseudo-panel et instrumentation (voir section 2.1). Pour rappel, dans le cas linéaire, l'estimation du modèle par pseudo-panel est équivalente à une régression sur les données individuelles, en incluant les indicatrices de cohortes et en instrumentant les autres covariables par les indicatrices de cohorte croisées avec les indicatrices de date d'observation. Cette même instrumentation peut être utilisée pour estimer le modèle (15). On est donc ramené à un problème d'estimation de modèle dichotomique à variables endogènes par instrumentation.

Le choix de l'instrument ou de manière équivalente du critère définissant les cohortes doit obéir aux mêmes règles que celles présentées à la section 1.2. Ensuite, l'estimation est assez simple dans le cas où les régresseurs endogènes sont continus (mais n'est pas adaptée à des régresseurs endogènes discrets). On calcule dans ce cas l'estimateur en deux étapes de Newey (1987) sous l'hypothèse supplémentaire de normalité des résidus, ce qui implique de mettre en œuvre un modèle probit.
En pratique, la procédure `ivprobit` sous STATA permet de mener ce type d'estimation. À notre connaissance aucune procédure équivalente n'existe sous les logiciels SAS et R.

Exemple de programmation sous STATA :
A partir des mêmes données simulées utilisées à la section 2.2.3, on construit la variable binaire `ybin` :

```
gen ybin=1*(0.2*x1+0.2*x2+alpha+rnormal(0,1)>2)
```

et on procède à l'estimation du modèle :

```
set matsize 1000
ivprobit ybin (x1 x2 = i.cohorte#i.t) i.cohorte,twostep
```

Une méthode alternative proposée par Collado (1998) consiste à utiliser une approximation pour les termes de cohortes (à partir des moyennes individuelles), et à traiter explicitement le problème d'erreur de mesure que cela implique. Elle est cependant bien plus complexe à mettre en œuvre. Nous n'en donnons ici que les grandes lignes.

Le point de départ de cette estimation est l'approche de Chamberlain (1984). On exprime la relation entre l'effet fixe individuel et les covariables :

$$\alpha_i = x_{i1}\lambda_1 + \ldots + x_{iT}\lambda_T + \theta_i \tag{17}$$

$$\text{avec } E(\theta_i | x_{i1}, \ldots, x_{iT}) = 0 \tag{18}$$

En substituant (17) dans (15), on obtient la forme réduite :

$$y^{\star}_{it} = x_{i1}\pi_{t1} + \ldots + x_{iT}\pi_{tT} + \theta_i + \varepsilon_{it} \quad i = 1, \ldots, N \quad t = 1, \ldots, T \tag{19}$$

avec $\pi_{ts} = \beta + \lambda_s$ si $s = t$ et $\pi_{ts} = \lambda_s$ sinon. Le terme d'erreur $\theta_i + \varepsilon_{it}$ n'est pas corrélé aux covariables.

Les deux estimateurs du paramètre β proposés par Collado sont calculés en deux étapes. La première étape commune aux deux estimateurs consiste à estimer les paramètres π_{ts}. Cette estimation n'est pas immédiate car les données ne permettent pas d'observer pour chaque individu toute la série des valeurs des covariables. On substitue donc aux valeurs individuelles les valeurs moyennes de la cohorte à laquelle appartient l'individu dans (19) :

$$y_{it}^{\star} = \bar{x}_{c1}\pi_{t1} + \ldots + \bar{x}_{cT}\pi_{tT} + \varepsilon_{it}^{\star} \quad t = 1, \ldots, T$$

Ceci introduit des erreurs de mesure dans l'équation et de la corrélation entre le terme d'erreur $\varepsilon_{it}^{\star}$ et les covariables. Sous une hypothèse de normalité des résidus, on peut exprimer l'espérance et la variance conditionnelles du résidu en fonction du vecteur des paramètres π_{st} et de la matrice de variance-covariance des erreurs de mesure (qui peut être estimée à partir des données individuelles). On en déduit alors une expression de la probabilité que y_{it} vale 1, permettant d'estimer (séparément à chaque date) les paramètres π_{st} par pseudo-maximum de vraisemblance.

Les deux estimateurs du paramètre β proposés se déduisent de l'estimateur des paramètres π_{st}. L'un est calculé par distance minimale, l'autre en procédant à une transformation within sur les données. L'estimateur within présente l'avantage d'être plus simple à calculer mais n'est pas efficace contrairement à l'estimateur par distance minimale.

4 Un exemple d'application des pseudo-panels : effet d'âge et de génération sur le niveau de patrimoine

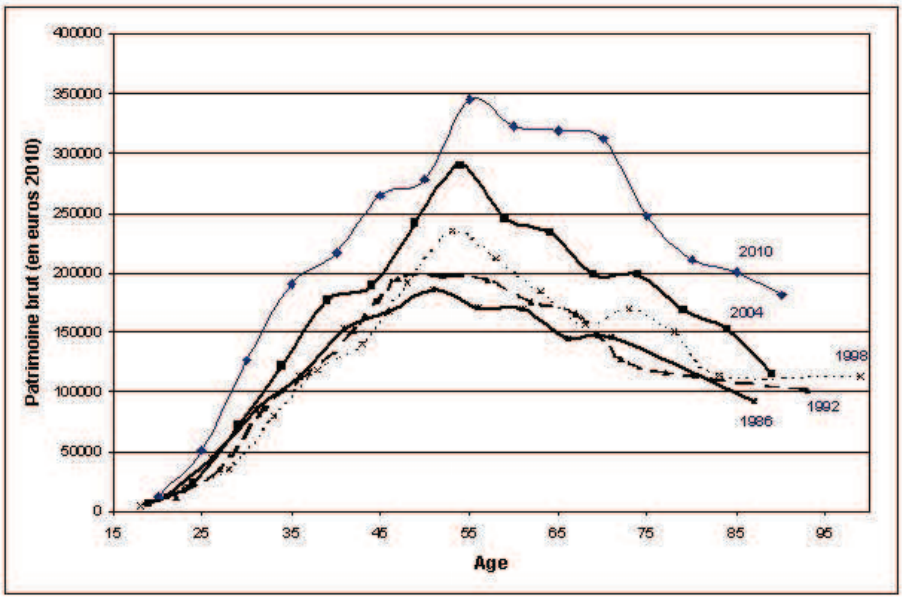
Nous proposons ici une application des méthodes de pseudo-panel à une analyse en cycle de vie du patrimoine détenu par les ménages. Cette application ne vise qu'à illustrer la mise en œuvre concrète de ces méthodes. On trouvera dans Lamarche et Salembier (2012) une analyse plus complète de cette question.

On utilise les différentes enquêtes Patrimoine. Cette enquête est menée tous les six ans depuis 1986 [8]. La dernière interrogation date de 2010. Nous disposons donc de cinq dates d'observation (1986, 1992, 1998, 2004 et 2010). Les ménages sont interrogés sur leur détention de biens immobiliers, financiers et professionnels. La somme de ces trois patrimoines constitue le patrimoine brut. Celui-ci est calculé en euros constants 2010.

Pour décrire très brièvement la problématique, il s'agit d'étudier les logiques d'épargne aux différents âges. Dans sa version initiale formulée par Modigliani (1986), la théorie du cycle de vie prévoit que les individus procèdent à une affectation intertemporelle de leurs revenus. Au cours de leur vie, ils connaissent trois périodes durant lesquelles leurs revenus, leurs comportements d'épargne et de consommation diffèrent. Le début de leur vie active est marqué par des revenus faibles et une désépargne. Ensuite, au cours de leur vie active, leur revenu augmentant, ils épargnent et se constituent un patrimoine, en prévision d'une baisse de revenu au moment de leur retraite. Le patrimoine suit ainsi une évolution en cloche. Il est difficile de tester cette hypothèse en estimant par exemple comment le patrimoine évolue avec l'âge, car cela demanderait

8. En 1986 et 1992, il s'agissait de l'enquête Actifs financiers.

de disposer du suivi des mêmes personnes sur très longue période, ce qui n'est pas possible. Comme on l'a déjà souligné, une estimation en coupe ne serait pas satisfaisante, car on ne peut pas distinguer les effets de l'âge de ceux de la génération. Les deux graphiques suivants permettent d'illustrer cette question. Chaque enquête Patrimoine permet de représenter l'évolution des patrimoines bruts moyens en fonction de l'âge (graphique 1). Les profils obtenus semblent conforter sans restriction la théorie du cycle de vie. On observe bien en effet une courbe en cloche, avec une croissance du patrimoine brut jusqu'à près de 60 ans et une baisse au-delà. Cependant, une partie de ce profil s'explique par le fait que l'on compare à chaque date des générations distinctes. Le contexte économique, l'âge d'entrée dans la vie active, la fiscalité sont autant de caractéristiques partagées par les individus d'une même génération qui peuvent avoir un effet sur le patrimoine accumulé et expliquer des différences de patrimoine à âge égal entre les différentes générations. Séparer ces deux effets demande un suivi sur longue période de ces générations.

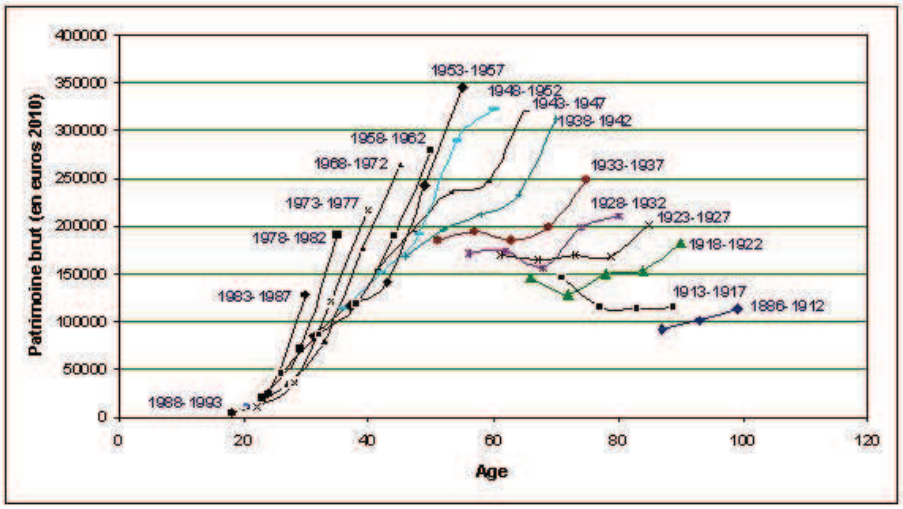


Graphique 1 – Patrimoine en fonction de l'âge

Pour tenter de capter cette dimension "génération", on empile donc toutes les enquêtes pour disposer de l'observation de personnes de générations identiques à des dates différentes (et donc des âges différents). On dispose donc de cinq observations, correspondant au patrimoine moyen à cinq âges différents, pour presque toutes les générations (sauf pour les plus jeunes ou les plus âgées). En principe, on pourrait représenter un profil pour toute génération, définie par l'année de naissance. En pratique cependant, on est confronté au problème que dans l'échantillon d'une enquête, le nombre d'individus d'une génération donnée n'est pas très élevé. Ces estimations risquent donc d'être très imprécises. Pour pallier ce problème, on définit des cohortes comme le regroupement de générations adjacentes (cinq sur le graphique 2).

On peut alors étudier, pour chaque cohorte, le profil d'accumulation du patrimoine par âge (graphique 2). Il est très différent de celui présenté en utilisant uniquement la dimension en coupe. Contrairement à ce que ce dernier suggère, le patrimoine continue de croître bien après 60 ans. Comme souligné par Lamarche et Salembier (2012), ce fait stylisé peut s'expliquer par

plusieurs facteurs. Même au-delà de la retraite les ménages peuvent être incités à épargner, dans l'idée de transmettre un patrimoine ou simplement pour constituer une épargne de précaution (liée aux risques de dépendance). Par ailleurs, les plus âgés peuvent renoncer à se séparer de leur patrimoine immobilier, souvent synonyme de déménagement, en raison de son coût particulièrement élevé (voir par exemple Angelini et Laferrère (2012)). Il faut aussi souligner que la croissance du patrimoine avec l'âge peut traduire en partie des changements de composition des générations observées aux âges extrêmes. Le champ de l'enquête ne porte que sur les ménages ordinaires et exclut donc les personnes âgées en maison de retraite. De plus, les ménages aisés ont une espérance de vie plus élevée que les autres (et aussi probablement un patrimoine supérieur).



Graphique 2 – Patrimoine, génération, âge

On peut également comparer sur le graphique 2 le patrimoine moyen des différentes cohortes au même âge. On observe des écarts parfois conséquents. L'écart vertical entre les courbes correspond à l'effet de génération, ainsi qu'à un effet date. Pour illustrer, supposons que ces effets dates, qui correspondent à l'augmentation au cours du temps des patrimoines (rappelons qu'on travaille en euros constants 2010 afin de ne pas intégrer l'inflation), soient négligeables [9]. Sous cette hypothèse, le graphique suggère qu'à âge égal, chaque génération a accumulé plus de patrimoine que la précédente. L'écart est particulièrement élevé entre les générations nées dans les années 50 et qui ont connu les Trente Glorieuses et les précédentes qui ont connu la guerre. La décroissance du patrimoine après 60 ans observée sur le graphique 1 tient ainsi certainement plus à des écarts de richesse importants entre ces deux générations qu'à une désépargne au moment de la retraite.

Pour avoir une vision plus synthétique de l'évolution du patrimoine en cycle de vie, on peut recourir à une modélisation économétrique en pseudo-panel. On part d'un modèle initial de la

9. Effet d'âge, de génération et de date ne sont pas identifiables en raison de la relation linéaire "année de naissance + âge = date". Voir par exemple Deaton et Paxson (1994) pour une analyse plus complète.

forme :

$$\log Pat_{it} = \beta_1 age_{it} + \beta_2 age_{it}^2 + \alpha_i + \varepsilon_{it} \quad i = 1,\ldots,N \quad t = 1,\ldots,T \tag{20}$$

$\log Pat_{it}$ est le logarithme du patrimoine de l'individu i à la date t, age_{it} son âge à la date t. On suppose ici que l'effet de l'âge sur le patrimoine est identique pour toutes les générations, et qu'il a un profil quadratique [10].

α_i est un effet fixe individuel. Il estime l'impact des caractéristiques fixes inobservées de l'individu i sur son patrimoine. Certaines sont individuelles (revenu permanent, profil épargnant), d'autres sont liées à sa génération. Formellement, on peut donc le décomposer en un effet génération α_g et une déviation individuelle à cet effet génération $v_i = \alpha_i - \alpha_g$. Chaque individu n'étant observé qu'une seule fois, il n'est pas possible d'estimer le modèle avec effet fixe individuel. Ne pas intégrer cette dimension (ou ne l'intégrer que partiellement via des indicatrices de cohortes) risque de fournir des estimateurs biaisés si ces déterminants inobservés v_i sont aussi corrélés à l'âge. En revanche, on peut utiliser des méthodes de pseudo-panel. L'année de naissance est ici un critère naturel pour regrouper les individus. Outre les avantages évoqués dans la section 1.2, il permet d'estimer directement un effet génération qui est intéressant en soit.

Le modèle en pseudo-panel qu'on estime en pratique s'écrit :

$$(\log Pat)_{gt} = \beta_1 age_{gt} + \beta_2 age_{gt}^2 + \alpha_g + \varepsilon_{gt} \quad g = 1,\ldots,G \quad t = 1,\ldots,T \tag{21}$$

où pour chaque variable z, $z_{gt} = E(z_{it}|i \in g,t)$. Ces valeurs ne sont pas observées. Elles sont estimées par les moyennes intra-cohortes $\bar{z}_{gt} = \frac{1}{n_{gt}} \sum_{i \in g,t} z_{it}$ calculées sur les données disponibles.

Deux remarques pratiques doivent être faites. La première porte sur la constitution de l'échantillon. L'estimation repose sur le fait que $\bar{\alpha}_{gt}$ est fixe dans le temps. Ceci peut être remis en cause. Comme discuté plus haut, pour les générations les plus âgées deux effets de composition peuvent jouer : les ménages les plus aisés ont en moyenne une longévité supérieure et l'enquête Patrimoine n'interroge pas les individus en maison de retraite. À l'autre extrême, l'enquête Patrimoine ne comprend que quelques ménages très jeunes, qui sont sans doute très spécifiques. Pour travailler sur une population stable, on se restreint aux ménages de plus de 26 ans et de moins de 80 ans [11]. La deuxième remarque porte sur la taille des cohortes. Les cohortes regroupent plusieurs générations successives. Restreindre le nombre de ces générations successives réduit le risque d'agréger des comportements hétérogènes mais au prix d'estimations reposant sur un nombre d'observations par cohortes très faibles : elles risquent donc d'être très imprécises. Pour illustrer cette question, on a estimé le modèle en utilisant des cohortes plus ou moins larges (trois, cinq et dix années) (tableau 2).

On présente dans le tableau 1 les résultats des estimations en pseudo-panel. À titre de comparaison, on présente également les résultats obtenus par une régression en coupe (on empile

10. L'accumulation du patrimoine avec l'âge entre les différentes générations ne diffère qu'en niveau. On pourrait complexifier le modèle en intégrant des termes d'interaction entre l'âge et la génération.

11. Par ailleurs, les moyennes étant sensibles aux valeurs extrêmes, certains ménages aux patrimoines très élevés ont été retirés de l'analyse. De même, on supprime les quelques observations correspondant à un patrimoine nul, car on utilise une modélisation en logarithme.

les données des cinq enquêtes successives) et les estimations en tenant compte des erreurs de mesure.

Tableau 1 – Estimation des effets de l'âge

		Estimations par pseudo-panel		
	Données en coupe	Générations de 3 ans	Générations de 5 ans	Générations de 10 ans
		Estimateur within		
Constante	$4,59^{***}$ (0,127)	$4,80^{***}$ (0,383)	$4,65^{***}$ (0,437)	$4,89^{***}$ (0,542)
âge	$0,223^{***}$ (0,0052)	$0,197^{***}$ (0,0142)	$0,199^{***}$ (0,0160)	$0,193^{***}$ (0,0212)
âge^2	$-0,0019^{***}$ (0,0000493)	$-0,00140^{***}$ (0,000135)	$-0,00136^{***}$ (0,000145)	$-0,00136^{***}$ (0,00020)
		Modèle à erreurs de mesure		
Constante		$4,63^{***}$ (0,279)	$5,05^{***}$ (0,307)	$5,63^{***}$ (0,398)
âge		$0,203^{***}$ (0,0104)	$0,187^{***}$ (0,0127)	$0,162^{***}$ (0,0172)
âge^2		$-0,00143^{***}$ (0,000092)	$-0,00128^{***}$ (0,00012)	$-0,00102^{***}$ (0,00016)

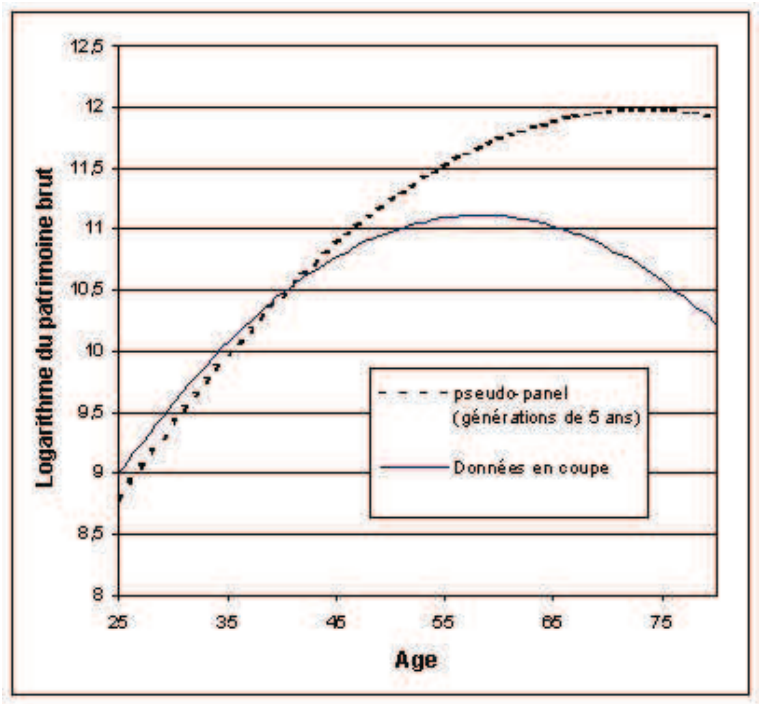
Remarque : La constante est calculée en prenant comme générations de référence : 1951-1953 pour les générations de 3 ans, 1953-1957 pour celles de 5 ans et 1953-1962 pour celles de 10 ans.
Les écarts-types ont été calculés par bootstrap pour le modèle à erreurs de mesure.

Le graphique 3 représente l'effet de l'âge sur le patrimoine tel qu'il est estimé en coupe d'une part et en pseudo-panel d'autre part [12]. Les deux estimations montrent une relation entre patrimoine et âge en forme de cloche. Sur données en coupe, on estime que le patrimoine commence à décroître à partir de l'âge de 58 ans. Dans l'estimation par pseudo-panel, cet âge est beaucoup plus avancé : il se situe à 70 ans. Lorsque l'on tient compte de l'effet génération, la baisse du patrimoine est ainsi beaucoup plus tardive qu'une coupe transversale le suggère.

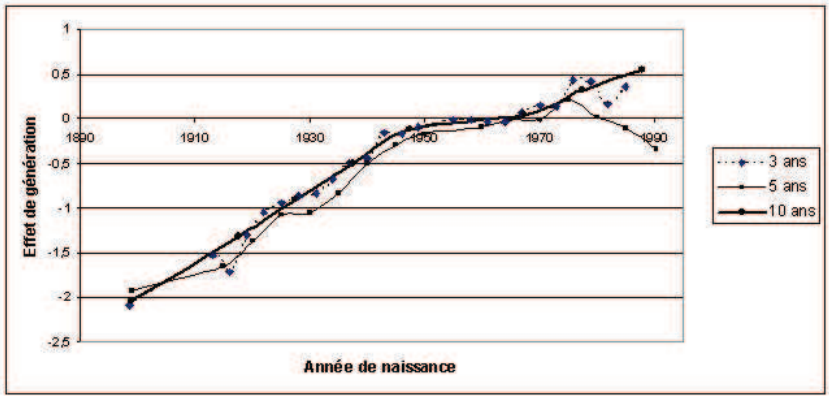
Le modèle étant log-linéaire, $100 \times [\exp(\alpha_g) - 1]$, où α_g est le coefficient associé à la génération g dans le modèle (tableau 3 et graphique 4), correspond à l'effet en pourcentage sur le patrimoine d'appartenir à la génération g plutôt qu'à la génération 1951-1953 (génération de référence). Par exemple, être né entre 1939 et 1941 plutôt qu'entre 1951 et 1953 a un effet négatif sur le patrimoine, estimé à $100 \times [\exp(-0.44) - 1] = -35,6\%$. On estime ainsi qu'entre les générations 1939-1941 et 1951-1953, le patrimoine a cru de 3,7 % en moyenne annuelle. Ensuite, la croissance a ralenti.

La sensibilité des estimations au critère de regroupement des cohortes n'apparaît pas trop élevée ici. On représente ci-dessous les effets générations, tels qu'on peut les estimer selon la largeur choisie pour définir les générations. Sans surprise, plus la largeur est élevée et plus

12. On représente donc le polynôme de degré deux : $\beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2$ dont les coefficients sont estimés sur données en coupe d'une part et par pseudo-panel d'autre part.



Graphique 3 – Patrimoine en fonction de l'âge tel qu'estimé par les modèles



Graphique 4 – Les effets génération (α_g)

le profil est lisse. On observe dans tous les cas, une augmentation importante du patrimoine des générations successives jusqu'à celles du baby boom, et une stagnation ensuite. Pour les générations les plus jeunes le diagnostic semble diverger selon le critère de regroupement, mais ces évolutions ne sont jamais significatives (cf tableaux 3). Cette incertitude tient au fait que les estimations sont effectuées sur des échantillons plus réduits (ces générations ne sont pas observées dans les enquêtes les plus anciennes...), comme illustré dans le tableau 2.
On remarque également que, comme attendu, la précision des estimateurs des coefficients β_1 et β_2 est plus importante pour des générations de trois ans que pour des générations de cinq ou dix ans.

Tableau 2 – Effectifs des cohortes

Générations de 3 ans

	1986	1992	1998	2004	2010
1886-1911	267				
1912-1914	191	124			
1915-1917	109	132			
1918-1920	179	268	153		
1921-1923	321	431	375		
1924-1926	278	502	397	228	
1927-1929	305	544	440	421	
1930-1932	301	498	469	444	336
1933-1935	282	522	512	468	555
1936-1938	287	426	456	413	593
1939-1941	284	430	488	445	569
1942-1944	317	481	502	392	704
1945-1947	372	614	654	467	804
1948-1950	408	727	728	562	894
1951-1953	391	683	680	570	838
1954-1956	373	731	626	554	756
1957-1959	292	704	652	560	774
1960-1962	77	569	582	544	723
1963-1965		407	582	552	743
1966-1968		124	465	506	654
1969-1971			463	511	599
1972-1974			132	426	541
1975-1977				367	414
1978-1980				112	396
1981-1983					290
1984-1986					85

Générations de 5 ans

	1986	1992	1998	2004	2010
1886-1912	344				
1913-1917	223	256			
1918-1922	391	551	395		
1923-1927	477	831	672	359	
1928-1932	516	861	767	734	336
1933-1937	476	787	804	744	964
1938-1942	478	742	815	707	954
1943-1947	588	944	993	734	1307
1948-1952	678	1181	1192	938	1457
1953-1957	615	1213	1068	964	1295
1958-1962	248	1020	1008	888	1233
1963-1967		531	915	877	1209
1968-1972			727	842	1000
1973-1977				643	742
1978-1982				112	598
1983-1987					173

Générations de 10 ans

	1986	1992	1998	2004	2010
1886-1912	344				
1913-1922	614	807	395		
1923-1932	993	1692	1439	1093	336
1933-1942	954	1529	1619	1451	1918
1943-1952	1266	2125	2185	1672	2764
1953-1962	863	2233	2076	1852	2528
1963-1972		531	1642	1719	2209
1973-1982				755	1340
1983-1993					173

Tableau 3 – Effets de génération estimés

Générations de 3 ans	
1886-1911	$-2,09^{***}$
	(0,302)
1912-1914	$-1,53^{***}$
	(0,279)
1915-1917	$-1,71^{***}$
	(0,308)
1918-1920	$-1,30^{***}$
	(0,214)
1921-1923	$-1,05^{***}$
	(0,170)
1924-1926	$-0,94^{***}$
	(0,158)
1927-1929	$-0,87^{***}$
	(0,146)
1930-1932	$-0,83^{***}$
	(0,140)
1933-1935	$-0,68^{***}$
	(0,132)
1936-1938	$-0,49^{***}$
	(0,131)
1939-1941	$-0,44^{***}$
	(0,127)
1942-1944	$-0,16$
	(0,122)
1945-1947	$-0,17$
	(0,114)
1948-1950	$-0,088$
	(0,109)
1951-1953	réf.
1954-1956	$-0,0078$
	(0,112)
1957-1959	$-0,014$
	(0,114)
1960-1962	$-0,042$
	(0,120)
1963-1965	$-0,035$
	(0,125)
1966-1968	$0,069$
	(0,136)
1969-1971	$0,14$
	(0,144)
1972-1974	$0,13$
	(0,163)
1975-1977	$0,43$
	(0,187)
1978-1980	$0,41$
	(0,223)
1981-1983	$0,16$
	(0,283)
1984-1986	$0,36$
	(0,493)

Générations de 5 ans	
1886-1912	$-1,93^{***}$
	(0,281)
1913-1917	$-1,65^{***}$
	(0,290)
1918-1922	$-1,37^{***}$
	(0,220)
1923-1927	$-1,08^{***}$
	(0,189)
1928-1932	$-1,04^{***}$
	(0,171)
1933-1937	$-0,84^{***}$
	(0,160)
1938-1942	$-0,51^{***}$
	(0,152)
1943-1947	$-0,29^{**}$
	(0,138)
1948-1952	$-0,17$
	(0,128)
1953-1957	réf.
1958-1962	$-0,089$
	(0,134)
1963-1967	$-0,0086$
	(0,144)
1968-1972	$-0,0089$
	(0,163)
1973-1977	$0,21$
	(0,197)
1978-1982	$0,0098$
	(0,239)
1983-1987	$-0,10$
	(0,324)
1988-1993	$-0,34$
	(0,590)

Générations de 10 ans	
1886-1912	$-2,03^{***}$
	(0,375)
1913-1922	$-1,31^{***}$
	(0,210)
1923-1932	$-0,90^{***}$
	(0,151)
1933-1942	$-0,50^{***}$
	(0,125)
1943-1952	$-0,12$
	(0,097)
1953-1962	réf.
1963-1972	$0,038$
	(0,107)
1973-1982	$0,32^{*}$
	(0,165)
1983-1993	$0,55$
	(0,485)

Références

Afsa, C. et S. Buffeteau. 2005, «L'évolution de l'activité féminine en France : une approche par pseudo-panel», *Document de travail de la Direction des Études et Synthèses Économiques*, vol. n° G2005/02.

Afsa, C. et V. Marcus. 2008, «Le bonheur attend-il le nombre des années ?», *France, portrait social*, p. 163–174.

Angelini, V. et A. Laferrère. 2012, «Residential mobility of the European elderly», *CESifo Economic Studies*, vol. 58, n° 3, p. 544–569.

Antman, F. et D. McKenzie. 2005, «Earnings mobility and measurement error : a pseudo-panel approach», Policy Research Working Paper Series 3745, The World Bank.

Blanpain, N. 2011, «L'espérance de vie s'accroît, les inégalités sociales face à la mort demeurent», *Insee Première*.

Bodier, M. 1999, «Les effets d'âge et de génération sur le niveau et la structure de la consommation», *Économie et statistique*, vol. 324-325, p. 163–180.

Chamberlain, G. 1984, «Panel data», dans *Handbook of Econometrics*, vol. 2, édité par Z. Griliches et M. D. Intriligator, chap. 22, Elsevier, p. 1247–1318.

Collado, M. D. 1998, «Estimating binary choice models from cohort data», *Investigaciones Economicas*, vol. 22, n° 2, p. 259–276.

Davezies, L. 2011, «Modèles à effets fixes, à effets aléatoires, modèles mixtes ou multi-niveaux : propriétés et mises en œuvre des modélisations de l'hétérogénéité dans le cas de données groupées», *Série des documents de travail de la Direction des Études et Synthèses Économiques*.

Deaton, A. 1985, «Panel data from time series of cross-sections», *Journal of Econometrics*, vol. 30, n° 1-2, p. 109–126.

Deaton, A. et C. Paxson. 1994, «Saving, growth, and aging in Taiwan», dans *Studies in the Economics of Aging*, NBER Chapters, National Bureau of Economic Research, Inc, p. 331–362.

Duguet, E. 1999, «Macro-commandes SAS pour l'économétrie des panels et des variables qualitatives», *Série des documents de travail de la Direction des Études et Synthèses Économiques*.

Duhautois, R. 2001, «Le ralentissement de l'investissement est plutôt le fait des petites entreprises tertiaires», *Économie et statistique*, vol. 341-342, p. 47–66.

Fuller, W. A. 1986, *Measurement Error Models*, John Wiley & Sons, Inc.

Gardes, F. 1999, «L'apport de l'économétrie des panels et des pseudo-panels à l'analyse de la consommation», *Économie et statistique*, vol. 324-325, p. 157–162.

Gardes, F., G. J. Duncan, P. Gaubert, M. Gurgand et C. Starzec. 2005, «Panel and pseudo-panel estimation of cross-sectional and time series elasticities of food consumption : The case of U.S. and Polish data», *Journal of Business and Economic Statistics*, vol. 23, p. 242–253.

Gurgand, M. 1997, *Éducation et efficacité de la production agricole*, thèse de doctorat, École des Hautes Études en Sciences Sociales.

Koubi, M. 2003, «Les carrières salariales par cohorte de 1967 à 2000», *Économie et statistique*, vol. 369-370, p. 149–170.

Lamarche, P. et L. Salembier. 2012, «Les déterminants du patrimoine : facteurs personnels et conjoncturels», dans *Les revenus et le patrimoine des ménages*, Insee Références Édition 2012.

Le Blanc, D., S. Lollivier, M. Marpsat et D. Verger. 2000, *L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT)*, n° 0001 dans Série des documents de travail "Méthodologie statistique" de l'Insee.

Lelièvre, M., O. Sautory et J. Pujol. 2012, «Niveau de vie par âge et génération entre 1996 et 2005», dans *Les revenus et le patrimoine des ménages*, Insee Références Édition 2012.

Magnac, T. 2005, «Économétrie linéaire des panels : une introduction», *Neuvièmes Journées de Méthodologie Statistique*. URL :
`http://jms.insee.fr/files/documents/2005/433_1-JMS2005_SESSION14_MAGNAC_ACTES.PDF` .

Marical, F. et L. Calvet. 2011, «Consommation de carburant : effets des prix à court et à long terme par type de population», *Économie et Statistique*, vol. 446, p. 25–44.

Modigliani, F. 1986, «Life cycle, individual thrift, and the wealth of nations», *The American Economic Review*, vol. 76, n° 3, p. 297–313.

Moffitt, R. 1993, «Identification and estimation of dynamic models with a time series of repeated cross-sections», *Journal of Econometrics*, vol. 59, n° 1-2, p. 99–123.

Newey, W. K. 1987, «Efficient estimation of limited dependent variable models with endogenous explanatory variables», *Journal of Econometrics*, vol. 36, n° 3, p. 231–250.

Verbeek, M. 2008, «Pseudo-panels and repeated cross-sections», dans *The Econometrics of Panel Data*, *Advanced Studies in Theoretical and Applied Econometrics*, vol. 46, édité par L. Mátyás et P. Sevestre, Springer Berlin Heidelberg, p. 369–383.

Verbeek, M. et T. Nijman. 1992, «Can cohort data be treated as genuine panel data ?», *Empirical Economics*, vol. 17, n° 1, p. 9–23.

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*, vol. 59, n° 1-2, p. 125–136.

Annexes

A Pseudo-panel et instrumentation

Moffitt (1993) montre que l'estimation par pseudo-panel et l'estimation en instrumentant par les indicatrices de cohortes croisées avec celles de la date d'observation fournissent le même estimateur.
Une estimation par les doubles moindres carrés suit les deux étapes suivantes :

1. Première étape : Projection des variables explicatives sur l'instrument.
   Si on décompose l'effet fixe individuel α_i en un effet fixe cohorte α_c et une déviation individuelle $v_i = \alpha_i - \alpha_c$, le modèle (1) se réécrit :

   $$y_{it} = x_{it}\beta + \alpha_c + v_i + \varepsilon_{it} \tag{22}$$

   x_{it} est potentiellement corrélé à v_i. Aussi x_{it} est instrumenté par les indicatrices de cohortes en interaction avec les indicatrices de temps. La première étape consiste à projeter x_{it} sur l'instrument. On montre que la valeur prédite de x_{it} dans cette régression correspond à la moyenne intra-cohorte $\bar{x}_{ct}$.
2. Deuxième étape
   La deuxième étape consiste à remplacer x_{it} par sa valeur prédite dans (22). On régresse ainsi y_{it} sur $\bar{x}_{ct}$ et les indicatrices de cohortes, ce qui fournit le même estimateur que l'estimateur within (4).

B Simulation de données de pseudo-panel

1 Sous SAS

1. 1re étape : Construction des données de panel

   Des données de panel sont simulées dans la table `panel`. 50 000 individus sont observés à 5 dates différentes. Les variables sont :
   – `ident` : l'identifiant de l'individu.
   – `t` : la date d'observation.
   – `z` : variable dont les quantiles vont servir à former les cohortes.
   – `alpha` : l'effet fixe individuel, tel que : $\alpha_i = 0,05\left[\sum_{t=1}^{5}(x_{it1}+x_{it2})+0,1\times u_i\right]$ où $u_i \sim N(0,1)$.
   – `x1`, `x2` : les covariables, corrélées à l'effet fixe `alpha`.
   – `y`, la variable d'intérêt, telle que $y_{it} = 0,8\times x_{it}^1 + 0,8\times x_{it}^2 + \alpha_i + 0,1\times\varepsilon_{it}$ où $\varepsilon_{it} \sim N(0,1)$.

   Définition des paramètres du programme :

```
option mprint macrogen;
 /* nombre d'individus: */
%let nobs=50000;
 /* nombre de dates d'observation: */
%let temps=5;
/* taille des cohortes formées: */
%let taille=100;
```

   Création de la table `panel` qui regroupe les observations de 50 000 individus à 5 dates :

```
DATA panel;
seed=20000;
DO ident = 1 TO &nobs.;
CALL RANNOR(seed,z);
DO t=1 TO &temps.;
OUTPUT;
END;
END;
RUN;
```

   Calcul des covariables `x1` et `x2`, de l'effet fixe `alpha` et de la variable d'intérêt `y` :

```
DATA panel;
SET panel;
x1 = (t=1)*(1+1*z+0.1*RANNOR(-1))
+ (t=2)*(1.5+2*z+0.1*RANNOR(-1))
+ (t=3)*(2+3*z+0.1*RANNOR(-1))
+ (t=4)*(3+1*z+0.1*RANNOR(-1))
+ (t=5)*(1.5+2.5*z+0.1*RANNOR(-1));
```

```
x2 = (t=1)*(2+3*z+0.1*RANNOR(-1))
+ (t=2)*(1+1.5*z+0.1*RANNOR(-1))
+ (t=3)*(2+2*z+0.1*RANNOR(-1))
+ (t=4)*(3+1*z+0.1*RANNOR(-1))
+ (t=5)*(2.5+2*z+0.1*RANNOR(-1));
alpha=(t=5)*0.05*(x1+LAG1(x1)+LAG2(x1)+LAG3(x1)+LAG4(x1)+
x2+LAG1(x2)+LAG2(x2)+LAG3(x2)+LAG4(x2)+0.1*RANNOR(-1));
IF alpha=. THEN alpha=0;
RUN;

PROC SORT DATA = panel (WHERE = (t=5))
OUT = alpha (KEEP = ident alpha);
BY ident;
RUN;

PROC SORT DATA = panel;BY ident;RUN;
DATA panel;
MERGE panel (DROP=alpha) alpha;
BY ident;
/* la variable d'intérêt:*/
y=0.8*x1+0.8*x2+alpha+0.1*RANNOR(-1);
run;
```

2. 2e étape : Regroupement des individus en cohortes

   On ordonne les individus par valeurs de la variable z croissantes. Les quantiles de cette variable délimitent les cohortes d'individus : les individus ayant les 100 valeurs les plus faibles de z sont dans la cohorte 1, les 100 suivants dans la cohorte 2, etc...

```
PROC SORT DATA = panel (WHERE=(t=1))
OUT=z;
BY z;
RUN;

DATA z (KEEP = ident cohorte);
SET z;
cohorte=FLOOR((_n_-1)/(&taille.*&temps.))+1;
run;
PROC SORT DATA = z;BY ident;RUN;
PROC SORT DATA = panel;BY ident;RUN;
DATA panel;
MERGE panel (drop=z) z;
BY ident;
RUN;
```

```
PROC DATASETS LIBRARY=work;
DELETE alpha z;
RUN;QUIT;
```

3. 3e étape : création de la table des coupes transversales

   Pour former la table `donnees` qui contient 5 coupes transversales indépendantes, on ne retient pour chaque individu de la table `panel` qu'une seule des 5 dates d'observation. La sélection se fait aléatoirement à partir de la variable `u`, simulée selon une loi uniforme. Les observations d'un individu sont triées selon les valeurs de `u` croissantes. Pour chaque individu, seule la première observation est retenue.

```
DATA panel;
SET panel;
u=RANUNI(-1);
PROC SORT;BY ident u;
RUN;

DATA donnees;
SET panel;
BY ident;
IF FIRST.ident;
DROP u;
RUN;
```

2 Sous R

Les mêmes données générées sous SAS sont générées ci-dessous sous R.

1. 1re étape : Construction de données de panel

   Définition des paramètres :

```
# Nombre d'observations
nobs=50000
# Nombre de dates d'observations
temps=5
# Taille des cohortes formées
taille=100
```

   Création des données de panel : 50 000 individus observés à 5 dates différentes :

```
ident=sort(rep(1:nobs,temps))
t  <- rep(1:temps,nobs)
```

   Calcul des covariables `x1` et `x2`, de l'effet fixe `alpha` et de la variable d'intérêt `y` :

```
z<-rnorm(nobs)

x1<-(t==1)*(1*rep(1,nobs*temps)+1*z[ident]+0.1*rnorm(nobs*temps))
   +(t==2)*(1.5*rep(1,nobs*temps)+2*z[ident]+0.1*rnorm(nobs*temps))
   +(t==3)*(2*rep(1,nobs*temps)+3*z[ident]+0.1*rnorm(nobs*temps))
   +(t==4)*(3*rep(1,nobs*temps)+1*z[ident]+0.1*rnorm(nobs*temps))
   +(t==5)*(1.5*rep(1,nobs*temps)+2.5*z[ident]+0.1*rnorm(nobs*temps))
x2<-(t==1)*(2*rep(1,nobs*temps)+3*z[ident]+0.1*rnorm(nobs*temps))
   +(t==2)*(1*rep(1,nobs*temps)+1.5*z[ident]+0.1*rnorm(nobs*temps))
   +(t==3)*(2*rep(1,nobs*temps)+2*z[ident]+0.1*rnorm(nobs*temps))
   +(t==4)*(3*rep(1,nobs*temps)+1*z[ident]+0.1*rnorm(nobs*temps))
   +(t==5)*(2.5*rep(1,nobs*temps)+2*z[ident]+0.1*rnorm(nobs*temps))

# l'effet fixe individuel
alpha<-0.05*(tapply(x1+x2,ident,sum)+0.1*rnorm(nobs))

# la variable d'intérêt
y<-0.8*x1+0.8*x2+alpha[ident]+0.1*rnorm(nobs*temps)

panel=data.frame(ident,t,y,x1,x2,alpha[ident],z[ident])
```

2. 2e étape : Regroupement des individus en cohortes

   On ordonne les individus par valeurs de la variable `z` croissantes. Les quantiles de cette variable délimitent les cohortes d'individus : les individus ayant les 100 valeurs les plus faibles de `z` sont dans la cohorte 1, les 100 suivants dans la cohorte 2, etc...

```
zb=data.frame(order(z),z[order(z)],1:nobs)

colnames(zb) = c("ident","z","numero")

cohorte=floor((zb$numero-1)/(taille))+1
zb=data.frame(zb,cohorte)
colnames(zb) = c("ident","z","numero","cohorte")

panel = data.frame( merge ( panel,zb, by = "ident") )
```

3. 3e étape : création de la table des coupes transversales

   Pour former la table `donnees` qui contient 5 coupes transversales indépendantes, une partie des données de la table `panel` est masquée : pour chaque individu, une seule date d'observation est retenue. La sélection se fait aléatoirement à partir de la variable `u`, simulée selon une loi uniforme. Les observations d'un individu sont triées selon les valeurs de `u` croissantes. Pour chaque individu, seule la première observation est retenue.

```
u=runif(nobs*temps)
panel=data.frame(panel,u)
panel=panel[order(panel$ident,panel$u),]
donnees = panel[seq(from=1,to=nobs*temps,by=5), ]
```

3 Sous STATA

```
clear
set obs 50000
gen z=rnormal(0,1)
expand 5
sort z
gen ident=floor((_n-1)/5)+1
gen t= mod(_n-1,5)+1

gen x1=(t==1)*(1+z+0.1*rnormal(0,1))+(t==2)*(1.5+2*z+0.1*rnormal(0,1))
+(t==3)*(2+3*z+0.1*rnormal(0,1))
+(t==4)*(3+z+0.1*rnormal(0,1))
+(t==5)*(1.5+2.5*z+0.1*rnormal(0,1))

gen x2=(t==1)*(2+3*z+0.1*rnormal(0,1))+(t==2)*(1+1.5*z+0.1*rnormal(0,1))
+(t==3)*(2+2*z+0.1*rnormal(0,1))
+(t==4)*(3+z+0.1*rnormal(0,1))
+(t==5)*(2.5+2*z+0.1*rnormal(0,1))

gen alpha=(t==5)*0.05*(x1+x1[_n-1]+x1[_n-2]+x1[_n-3]+x1[_n-4]+
x2+x2[_n-1]+x2[_n-2]+x2[_n-3]+x2[_n-4]+0.1*rnormal(0,1))

replace alpha = alpha[ident*5]

gen y = 0.8*x1+0.8*x2+alpha+0.1*rnormal(0,1)

sort z
gen cohorte=floor((ident-1)/(100))+1

gen u=runiform()
sort ident u

keep if mod(_n-1,5)+1 ==1
```

C Détails sur l'estimation des paramètres d'un modèle à erreurs de mesure

$\bar{x}_{ct}$ et $\bar{y}_{ct}$ sont des observations avec erreurs des "vraies" moyennes intra-cohortes $x^{\star}_{ct}$ et $y^{\star}_{ct}$. u_{ct} et v_{ct} sont les erreurs de mesure :

$$\bar{y}_{ct} = y^{\star}_{ct} + u_{ct} \tag{23}$$

$$\bar{x}_{ct} = x^{\star}_{ct} + v_{ct} \tag{24}$$

Elles sont supposées être normalement distribuées :

$$\begin{pmatrix} u_{ct} \\ v'_{ct} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} ; \frac{1}{n} \begin{pmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{pmatrix} \right) \tag{25}$$

Avec n la taille des cohortes.

En intégrant (23) et (24) dans le modèle (2), on obtient :

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \alpha_c + \tilde{\varepsilon}_{ct} \quad c = 1, \ldots, C \quad t = 1, \ldots, T \tag{26}$$

avec $\tilde{\varepsilon}_{ct} = \varepsilon^{\star}_{ct} + u_{ct} - v_{ct}\beta$. La corrélation entre ce résidu et les covariables vaut :

$$E(\bar{x}'_{ct}\tilde{\varepsilon}_{ct}) = \frac{1}{n}(\sigma - \Sigma\beta)$$

Elle n'est pas nulle en général. L'estimateur des moindres carrés de $\bar{y}_{ct}$ sur $\bar{x}_{ct}$ est donc biaisé.

Le modèle (26) est un modèle à effet fixe. Après transformation within, le modèle (26) devient :

$$\bar{y}_{ct} - \bar{y}_c = (\bar{x}_{ct} - \bar{x}_c)\beta + \tilde{\varepsilon}_{ct} - \bar{\tilde{\varepsilon}}_c \quad \text{où } \bar{\tilde{\varepsilon}}_c = \frac{1}{T}\sum_{t=1}^{T} \tilde{\varepsilon}_{ct} \tag{27}$$

On montre que :

$$E(\bar{x}_{ct} - \bar{x}_c)'(\bar{y}_{ct} - \bar{y}_c) = E(\bar{x}_{ct} - \bar{x}_c)'(\bar{x}_{ct} - \bar{x}_c)\beta + \frac{T-1}{T}\frac{1}{n}(\sigma - \Sigma\beta)$$

De cette équation, on déduit une expression de β :

$$\beta = \left[E(\bar{x}_{ct} - \bar{x}_c)'(\bar{x}_{ct} - \bar{x}_c) - \frac{T-1}{T}\frac{1}{n}\Sigma \right]^{-1} \left[E(\bar{x}_{ct} - \bar{x}_c)'(\bar{y}_{ct} - \bar{y}_c) - \frac{T-1}{T}\frac{1}{n}\sigma \right]$$

L'estimateur (12) est la contre-partie empirique de cette expression.

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD, X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS, K. MOUSSALLAM