
PANEL SOCIO-ECONOMIQUE "LIEWEN ZU LËTZEBUERG"

LES GAINS D'ACTIVITE DES JEUNES
ADULTES EUROPEENS SONT-ILS
LIES A LA GENEROSITE
DES TRANFERTS SOCIAUX ?
UNE ANALYSE AU MOYEN DE
MODELES MULTI-NIVEAUX

par

Jean-Claude RAY

Présentation du programme PSELL 2

Avec le programme PSELL 2 développé par la Division "Population et Ménages" du CEPS/INSTEAD, le Grand-Duché de Luxembourg dispose d'un instrument exceptionnel permettant de connaître les conditions d'existence des personnes et des ménages qui y vivent depuis 1985 : le panel socio-économique "Liewen zu Lëtzebuerg" (PSELL).

Dans le cadre de ce programme, de nombreuses informations sont récoltées chaque année sur les principaux aspects de la vie de la population du pays :

- Conditions de logement, équipement et composition des ménages
- Principales dépenses
- Précarité
- Endettement
- Position scolaire des enfants
- Position socioprofessionnelle des adultes
- Revenus,...

En 1994, cette étude a fêté son dixième anniversaire. Sur le plan scientifique, cet événement représentait certainement un succès parce qu'il est très rare qu'un même programme de recherche puisse être développé sur une période aussi longue. Une large part de ce succès revient toutefois aux milliers de personnes qui, au fil des années, ont accepté de recevoir chez elles nos enquêteurs et de participer à ce vaste programme ; par leur contribution, elles ont permis de réunir un capital de connaissances inestimable, couvrant dix ans de la population de notre pays.

Les données récoltées ont déjà fait l'objet de nombreuses études publiées pour la plupart au CEPS/INSTEAD dans les séries suivantes :

- Documents PSELL (voir liste en annexe)
- Documents de recherche
- PSELL INFO
- ECOCEPS
- Population & Emploi

A partir de 1994, l'échantillon de l'étude a été rénové. Il compte désormais 8232 personnes réparties dans 2978 ménages (avant pondération). Cet échantillon évolue comme la population du pays. Il prend en compte les naissances, l'immigration, les mariages, les décès et l'émigration.

Pour plus d'informations :

Isabelle BOUVY

Tél. : (00352) 58 58 55-513

Fax : (00352) 58 55 60

e-mail : isabelle.bouvy@ceps.lu

Document produit par le

CEPS/INSTEAD

Centre d'études de populations, de pauvreté et de politiques socio-économiques

B.P. 48

L-4501 DIFFERDANGE

Président : Gaston SCHABER

Document PSELL n°132 - ISBN 2-87987-312-6

LYON, GATE, 16-17 mai 2002

Journées de l'AFSE « Ressources humaines »

Thème N°5 : « Relations rémunération-incitations-carrière »

Sous-thème « Protection sociale, retraites, avantages »

**Les gains d'activité des jeunes adultes européens sont-ils
liés à la générosité des transferts sociaux ?**

Une analyse au moyen de modèles multi-niveaux.

Jean-Claude RAY ¹

¹ Université Nancy 2 et CNRS (ESA 7003 « Emploi et Politiques Sociales »). Courriel : Jean-Claude.Ray@univ-nancy2.fr. L'auteur exprime sa reconnaissance à la Section 37 du Comité National et à la Direction du Département SHS du CNRS pour la délégation qui lui a été accordée, depuis septembre 2001, pour qu'il puisse mener à bien un projet de recherche intitulé « Au-delà du facteur 'pays'. Une analyse des comportements des jeunes adultes européens à l'aide de modèles multi-niveaux » ; le présent travail s'inscrit dans ce cadre, et sa présentation aux Journées de l'AFSE à Lyon les 16-17 mai 2002 vise notamment à bénéficier de remarques et suggestions pour la poursuite et l'approfondissement des premiers résultats obtenus.

L'auteur remercie par ailleurs vivement le CEPS/INSTEAD (G.D. de Luxembourg), et tout spécialement Pierre HAUSMAN et Uwe WARNER, pour l'aide matérielle apportée à la réalisation de cette recherche.

AVERTISSEMENT

Le présent texte analyse la relation entre les transferts sociaux et les gains des jeunes adultes européens pour l'ensemble des 16-30 ans. Il le fait sur la base de la quatrième vague de l'ECHP (European Community Household Panel 1997), et pour quatorze pays de l'Union. Les modèles employés sont des modèles multiniveaux, dont les résultats sont comparés à ceux que fournirait des modèles classiques de régression par les moindres carrés ordinaires.

La conclusion méthodologique principale est que, lorsque les variables explicatives d'un phénomène se situent à plusieurs niveaux différents, il faut recourir à des modèles multiniveaux ; sinon on risque, comme ici, de croire, au vu des coefficients de régression des moindres carrés ordinaires, que certaines variables (ici, les transferts sociaux) ont un lien statistiquement significatif avec la variable expliquée (ici, les gains d'activité marchande des jeunes adultes) alors qu'en réalité il n'en est rien : il s'agit d'un pur artefact statistique, dû à ce que l'on néglige de prendre en compte le fait que tous les jeunes adultes d'un même pays ne constituent pas des observations indépendantes les unes des autres mais ont en commun toutes les caractéristiques du pays (barèmes de prestations sociales, richesse du pays, taux de chômage, etc.).

Le présent texte trouve un prolongement dans le Document PSELL n°133, qui s'en distingue doublement : d'une part il prend en compte l'effet de sélection qui fait que certains jeunes adultes, ayant décidé de ne pas participer au marché du travail, n'ont pas de gains. D'autre part, afin de tenir compte de l'hétérogénéité, selon l'âge, du groupe formé par les jeunes adultes, l'analyse y est menée séparément pour trois tranches d'âges quadriennales : 18-21 ans, 22-25 ans, 26-29 ans.

Ce texte a fait l'objet d'une communication aux Journées " Ressources humaines " de l'AFSE (Association Française de Science Economique). Ces Journées ont été organisées par le GATE-CNRS à Lyon, les 16-17 mai 2002.

SOMMAIRE

| | | |
|----------------------|--|----------|
| INTRODUCTION | | 7 |
| CHAPITRE I | LA PROBLEMATIQUE ADOPTEE POUR CETTE ANALYSE DES GAINS D'ACTIVITE DES JEUNES ADULTES EUROPEENS | 11 |
| 1. | Une approche multi-pays des gains d'activité des jeunes adultes . | 13 |
| 2. | Notre approche « au-delà du facteur-pays » | 16 |
| CHAPITRE II | LE MODELE | 19 |
| 1. | Les prédicteurs des gains d'activité | 21 |
| 2. | Notre modèle à deux niveaux | 25 |
| | 2.1. <i>La non-indépendance des observations et ses conséquences</i> . | 25 |
| | 2.2. <i>Présentation de la structure-type des équations d'un modèle multi-niveaux</i> | 27 |
| | 2.3. <i>Exposé du modèle à deux niveaux</i> | 29 |
| CHAPITRE III | LES DONNEES | 31 |
| CHAPITRE IV | LES MODELES ESTIMES ET LEURS RESULTATS | 37 |
| CHAPITRE V | MODELES HIERARCHIQUES MULTI-NIVEAUX VERSUS REGRESSIONS MULTIPLES CLASSIQUES | 47 |
| CONCLUSION | | 51 |
| BIBLIOGRAPHIE | | 57 |
| ANNEXES | | 65 |
| | Annexe A Résultats obtenus avec SAS | 67 |
| | Annexe B Résultats de modèles obtenus avec HLM5, MlwiN et SAS | 69 |
| | Annexe C Résultats complémentaires obtenus avec HLM5, MlwiN et SAS . | 90 |

INTRODUCTION

Ce texte présente d'abord, et justifie, la problématique adoptée dans le cadre de cette analyse des gains d'activité des jeunes adultes européens. L'objectif est en effet de voir si on peut **identifier un rôle éventuel des transferts sociaux sur ces gains, lorsqu'on contrôle** à la fois les **déterminants individuels** que sont l'âge, le sexe et le capital humain, et les **déterminants collectifs** de ces gains (au niveau national, ici² : le taux de chômage, la générosité des transferts sociaux, etc.).

Mais, face à ces **déterminants se situant à deux niveaux distincts** (l'individu, le pays), encastrés l'un dans l'autre de façon hiérarchique, il ne serait pas techniquement satisfaisant d'utiliser **des régressions multiples classiques** ; en effet, parce que tous les ressortissants d'un même pays ont des caractéristiques communes (car ils sont notamment confrontés au même système de protection sociale et à une même conjoncture économique nationale), **les observations ne sont pas indépendantes les unes des autres**. Or cette indépendance constitue, on le sait, une des exigences essentielles pour que la régression multiple fournisse des estimateurs au moins aussi bons que d'autres techniques. On en conclut qu'il faut se tourner vers des méthodes prenant explicitement en compte cette non-indépendance, et on explique en quoi les **modèles hiérarchiques³ multi-niveaux** remplissent cette fonction.

Après avoir rappelé par quelles variables la **théorie économique** suggère d'expliquer les gains d'activité, et quel est le signe de chaque effet attendu, on en déduit un **modèle à deux niveaux** qu'il paraîtrait souhaitable de parvenir à estimer. On examine ensuite dans quelle mesure les **données** de la quatrième vague (1997) du Panel Communautaire de Ménages permettraient cette estimation. On met alors en œuvre un **modèle central**, entouré de **deux modèles simplistes** (servant de base de comparaison) et d'un **modèle simplifié**, limité aux variables présentant des effets significatifs. On les présente sous leur forme structurelle et on commente leurs résultats.

Enfin, ces résultats sont **mis en regard** de ceux qui auraient été produits si on avait appliqué la technique de **la régression multiple classique** ; ceci permet de prendre la mesure des différences de résultats, particulièrement en termes de significativité des effets.

² Nous nous en tenons, ici, au niveau supérieur que sont les **pays** ; mais nous comptons bien, dans un second temps, affiner l'analyse en ajoutant un ou plusieurs niveaux intermédiaires : d'une part les **régions**, d'autre part les **ménages** dans lesquels vivent les jeunes adultes.

³ Il s'agit de modèles **hiérarchiques** parce que ces modèles supposent un emboîtement parfait de chacun des niveaux dans le niveau supérieur ; ici, tout jeune adulte est considéré comme relevant d'un pays et d'un seul (ce qui fait certes l'impasse sur les mobilités, par exemple celles qui ont lieu dans le cadre d'ERASMUS, mais ce qui demeure une hypothèse raisonnable lorsqu'il s'agit d'expliquer des gains d'activité).

Car, au-delà des résultats, assez préliminaires, présentés ici quant aux déterminants individuels et macro des gains d'activité des jeunes adultes, la question à laquelle ce texte vise à apporter un début de réponse est celle que se pose **tout chercheur se trouvant face à un choix d'investissement de son temps** : les différences de résultats entre ceux obtenus par régressions multiples et ceux que produisent des modèles multi-niveaux sont-elles si importantes qu'elles justifient qu'on consacre du temps à **se familiariser avec cette nouvelle technique** (comme il a fallu le faire, dans les années 80, avec les modèles de durée, qui se sont imposés depuis comme incontournables en cas de durées fréquemment censurées) ? **Ou bien** s'agit-il d'un **phénomène de mode** ("un raffinement économétrique de plus", railleront certains), qui passera et qui ne mériterait donc pas le détour (productif) ?

Ce que nous montrons ici, c'est que, **au moins dans certaines configurations** (et celle que nous étudions⁴ nous semble typique à cet égard, avec quatorze pays et de nombreux individus par pays⁵), **vouloir ignorer la non-indépendance des observations conduirait à des résultats exagérant considérablement la significativité statistique des effets**, ce qui amènerait notamment à inférer qu'il existe, dans la population de référence, des influences (influences, ici, des transferts sociaux sur les gains d'activité), alors que ces effets, sinon n'existent pas, du moins ne peuvent pas être mis en évidence de façon indiscutable avec des instruments et des données du type de ceux que nous utilisons ici.

Notre **conclusion**, qui correspond à l'intuition mais qui la confirme preuves à l'appui, sera donc que, **au moins dans le type de configuration « peu de groupes, de nombreux individus par groupe », les modèles hiérarchiques multi-niveaux méritent vraiment qu'on les utilise.**

⁴ Notre présent travail n'a aucunement la prétention de démontrer, de façon générale, une quelconque supériorité éventuelle des modèles multi-niveaux. Il nous semble simplement qu'en prouvant, sur un cas particulier, que les conclusions peuvent diverger très sensiblement selon que l'on emploie des modèles classiques de régression multiple ou des modèles multi-niveaux, on peut utilement attirer l'attention des chercheurs concernés par ce type de problématique et par ce type de configurations de données, sur l'utilité que peut représenter un investissement intellectuel en matière de modèles multi-niveaux. Signalons, à toutes fins utiles, que l'autre configuration polaire ("de nombreux groupes, mais peu étoffés chacun", ce qui est le cas, par exemple, des études portant sur les élèves/classes/écoles) est celle qui semble avoir motivé, à l'origine, le développement des techniques des modèles multi-niveaux par des statisticiens et des sociologues spécialistes de l'éducation ; et ceci bien que, dans ce cas, le degré de non indépendance des observations soit bien moindre que dans notre cas ; a fortiori le recours aux modèles multi-niveaux s'impose-t-il donc lorsque les groupes sont peu nombreux et que la taille des groupes est grande.

⁵ Le fait de disposer de plus de 2500 jeunes adultes par pays, en moyenne, implique une dose considérable de dépendance des observations les unes vis-à-vis des autres ; cette situation contraste avec celle où on a un grand nombre de groupes – un millier d'écoles, par exemple – mais ne comprenant chacun qu'un nombre limité d'individus.

CHAPITRE I

**LA PROBLEMATIQUE ADOPTÉE
POUR CETTE ANALYSE DES GAINS
D'ACTIVITÉ DES JEUNES ADULTES EUROPÉENS**

Notre objectif est, nous venons de l'indiquer, d'examiner si les transferts sociaux influencent les comportements de gains des jeunes adultes, et ceci lorsqu'on contrôle à la fois les déterminants individuels et les déterminants nationaux de ces gains. Il est donc logique, du fait de l'inclusion de ces facteurs macro dans le modèle, d'adopter une approche multi-pays (Union Européenne, 1997). Mais on s'efforcera aussi d'aller au-delà du facteur-pays, comme nous allons l'expliquer.

1. UNE APPROCHE MULTI-PAYS DES GAINS D'ACTIVITÉ DES JEUNES ADULTES

La question de l'explication, au niveau individuel, des **gains d'activité**, est un thème fréquemment analysé par les économistes, notamment lorsqu'il s'agit d'essayer d'isoler les effets, sur le niveau et le profil de ces gains, des sources de capital humain que peuvent être la formation initiale et l'expérience (spécifique et/ou générale) acquise sur le lieu de travail [MINCER 1958], [WILLIS 1986], [JAROUSSE et MINGAT 1986], [GOUX et MAURIN 1994], [GUILLOTIN et SEVESTRE 1994], [CAHUC et ZYLBERBERG 1996].

Mais le traitement économétrique de cette question est, on le sait, suspendu à la question de **savoir s'il existe, dans le cas étudié, des biais de sélection**⁶, dont il faudrait alors tenir compte de façon appropriée [HECKMAN 1979] [SOLLOGOUB et ULRICH 1999]. Ces biais de sélection peuvent être liés notamment au chômage ou "à une inégale propension à se retirer du marché du travail (prolongation d'études, maintien ou retour au foyer,...)"⁷.

⁶ Comme le notent fort à propos BAUDELLOT et GOLLAC [1997], la notion même de biais de sélection et donc l'opportunité de sa correction "ne se justifient pleinement que dans une conception du marché du travail où les individus se verraient offrir des salaires et refuseraient des emplois si ces salaires sont inférieurs à leur 'salaire de réserve' ; [or] cette conception ne correspond pas toujours bien à la réalité empirique : certains individus au chômage ne se voient proposer aucun emploi pour des raisons non entièrement réductibles au salaire (localisation, pénibilité, conception des rôles sexuels,...)" (op.cit., encadré 2). On pourrait ajouter que, dans le cas des jeunes adultes, la correction des éventuels biais de sélection serait rendue plus difficile par l'absence d'une variable (disponible en revanche pour leurs aînés) souvent bien utile pour modéliser le biais de sélection : l'expérience de travail antérieure [RYAN 2001].

⁷ BAUDELLOT et GOLLAC [1997], p. 28.

Si, dans le cas qui nous occupe, de véritables biais de sélection sont identifiés au moyen de raisonnements convaincants, il faudra considérer sous un autre jour les résultats de nos modèles de régression (classiques ou de type modèles multi-niveaux) : à la manière de la façon dont BAUDELLOT et GOLLAC [1997] présentent leur analyse des salaires des trentenaires – la tranche d'âge située juste après celle de nos jeunes adultes –, **notre propos devra alors être considéré comme essentiellement descriptif** : en permettant de contrôler de multiples facteurs, l'**analyse de variance** évite d'attribuer à un facteur particulier (l'âge ou le sexe, par exemple) un excès de rôle qu'il doit en fait à sa corrélation avec d'autres facteurs, corrélation qui se traduit par des effets de structure ; sans y parvenir, certes (en raison de trop fortes corrélations entre certains facteurs mais aussi parce que nombreux sont les facteurs non observés⁸), on tendra donc vers l'identification des "effets purs" de chacun des facteurs.

Appliquée aux **jeunes adultes**, cette problématique du capital humain est compliquée par le fait qu'un certain nombre des individus concernés sont en cours d'études et par le fait qu'une proportion non négligeable de ceux-ci, mais aussi de ceux qui ont achevé leur formation initiale, ne travaille, volontairement ou non, pas du tout ou seulement à temps très partiel ; la valeur de la variable expliquée – les gains – est donc nulle pour un certain nombre de jeunes adultes, et très faible pour beaucoup d'autres. Par ailleurs l'existence de gains substantiels ou non résulte de choix multiples et enchevêtrés (poursuivre ou non ses études, continuer ou non à cohabiter avec ses parents, etc.) ; et ces choix sont parfois fortement contraints (par les possibilités de trouver un emploi et/ou un logement indépendant) ; enfin, l'acuité de ces choix est généralement plus grande à cet âge que plus tard. D'où l'intérêt (et la difficulté technique) de l'étude des choix d'activité, et donc des gains qui en résultent, en relation avec d'autres choix, cruciaux à cet âge (décohabitation parentale, poursuite d'études, activité ou vie au foyer, mais aussi choix de vie en couple, voire de fécondité, voire de mobilité résidentielle) [GOLDSCHIEDER 1989], [AQUILINO 1990 et 1991], [ROSENZWEIG et WOLPIN 1993 et 1994], [FESTY 1994], [CAUSSAT 1995], [VILLENEUVE-GOKALP 1997 et 2000], [CICHELLI et ERLICH 2000], [GALLAND 2000], [SOLAZ 2000], [AMROUNI 2001].

A ces spécificités viennent s'ajouter, pour ceux qui cherchent à évaluer l'éventuel impact des transferts sociaux et de la fiscalité personnelle sur les comportements individuels [WEI-YIN 2001], [WHITTINGTON et PETERS 1996], le fait que de **multiples types de transferts sociaux** sont théoriquement susceptibles d'influencer les choix des jeunes adultes et des ménages dans lesquels ils vivent : prestations familiales et de logement, certes, mais aussi bourses d'études, indemnisation du chômage, voire prestations d'assistance ; ce qui ne facilite pas l'identification du rôle spécifique éventuellement joué par **chacun de ces types de transferts sociaux**.

⁸ Citons, au titre des facteurs inobservés qui, corrélés à la fois avec les gains d'activité et avec la variable d'intérêt, peuvent créer un biais de sélection, des caractéristiques comme la motivation de l'individu, sa personnalité, son look, etc. ; toutes caractéristiques susceptibles de favoriser son embauche et ses gains mais qui peuvent aussi avoir joué sur le niveau de diplôme qu'il a atteint : ne pas introduire ces caractéristiques parce qu'elles sont indisponibles dans les données dont on dispose, c'est biaiser les coefficients afférents aux variables explicatives figurant dans le modèle. Certes, on peut tenter de remédier à ces effets de sélection avec des techniques comme la procédure de Heckman en deux étapes, les variables instrumentales ou les modèles à effets fixes. Mais les résultats économétriques ainsi obtenus sont très sensibles [HECKMAN et SMITH 1996] aux hypothèses faites quant à la distribution des caractéristiques inobservées, aux processus de sélection et au choix du groupe de référence. Cette fragilité devrait conduire [RYAN 2001] à ne pas écarter, pour les recherches évaluatives, le recours aux comparaisons internationales, méthode dite "expérimentale à un faible degré" parce qu'on manque de moyens de contrôler des facteurs comme les cultures nationales.

Pour ces diverses raisons, et en dépit de l'intérêt porté actuellement à la **dimension financière de l'entrée dans l'âge adulte** [EKERT-JAFFÉ, ARBONVILLE et WITWER 1995], [HERPIN et VERGER 1997], [PAUGAM et ZOYEM 1997], [THELOT et VILLAC 1998], [COMMISSARIAT GENERAL DU PLAN 2001], les analyses de l'influence éventuelle des transferts sociaux sur les comportements individuels des jeunes adultes ne sont pas légion, sans doute en partie parce qu'elles sont particulièrement délicates à mener ; ceci d'autant plus que, dans des systèmes de protection sociale fortement uniformes sur tout le territoire national, comme le sont le système français et ceux de la plupart des autres pays européens⁹, l'analyste est **en peine de trouver, s'il se cantonne à un seul pays, suffisamment de variabilité** dans les transferts sociaux pour pouvoir isoler leurs éventuels effets sur les gains d'activité¹⁰.

Certes, **au sein d'un pays donné**, la création de certaines prestations offre parfois une sorte d'expérience naturelle et, à condition que l'environnement conjoncturel ne se modifie pas trop vite, les panels permettent d'observer des modifications attribuables à des changements de barèmes. La solution alternative consiste à se tourner vers des **comparaisons entre pays**, en espérant que les différences sensibles, d'un pays à l'autre, de générosité et de barèmes des transferts sociaux aideront à identifier les effets qui font l'objet de l'étude. Toutefois, outre leur lourdeur, les comparaisons internationales sont délicates [COMMAILLE et DE SINGLY 1997] [PALIER 2001] et souffrent de multiples incertitudes, allant de l'arbitraire du choix des taux de conversion en unités monétaires communes (du moins avant l'avènement de l'euro) aux incompatibilités des nomenclatures – qui révèlent, en fait, des différences historiques et culturelles bien plus profondes, apparemment irréductibles. Et, quand on utilise, pour ces comparaisons internationales, des enquêtes auprès des ménages (et non pas les comptes de la protection sociale ou des cas-types), on rencontre un certain nombre de limites difficilement dépassables par l'utilisateur des données :

- manque de représentativité des échantillons nationaux (notamment pour certains segments de la population, dont, tout spécialement, les jeunes adultes) en dépit des redressements
- exclusion de certains aspects de la protection sociale (les services gratuits, de logement par exemple [CHAMBAZ 2001]) qui, inégalement développés selon les pays, faussent les comparaisons
- et, enfin, divergences de conceptions entre pays quant à des définitions comme celle du ménage (c'est ainsi que, dans la vague 1997 de notre source de données, pourtant réputée harmonisée au plan européen, on s'est étonné de ne trouver aucun jeune adulte suédois de 16 à 30 ans vivant dans le ménage de ses parents : c'est simplement que, pour la statistique suédoise, on forme son propre ménage dès lors qu'on est majeur, même si on vit chez ses parents).

⁹ Comme le note RYAN [2001], les barrières culturelles qui, en Europe, s'opposent à des expérimentations sociales du type de celles menées aux Etats-Unis (par affectation aléatoire d'individus – informés et consentants – à un groupe de contrôle ou à un groupe bénéficiant de transferts sociaux plus généreux) existent en revanche des deux côtés de l'Atlantique pour ce qui est de l'éducation.

¹⁰ Dans le cas français par exemple, les transferts sociaux qui ont des barèmes uniformes sur le territoire l'emportent considérablement, en termes de masses financières, sur ceux qui connaissent des barèmes diversifiés (selon les politiques des Conseils généraux, des Caisses d'allocations familiales, etc.).

2. NOTRE APPROCHE « AU-DELÀ DU FACTEUR-PAYS »

Du fait de cette irréductibilité apparente découlant des différences historiques et culturelles entre pays, lorsque l'on s'est décidé pour une approche multi-pays – comme nous l'avons fait ici –, on a généralement tendance à préférer une modélisation qui, par un usage intensif de **variables indicatrices représentant chaque pays** (y compris en croisant ces *dummies* avec d'autres variables explicatives d'intérêt), permet de mettre en lumière les spécificités nationales – en termes, certes, de niveaux de gains mais aussi d'importance d'un effet particulier, comme celui, par exemple, du sexe sur les gains.

Mais cette approche, si elle est pratique lorsque quelques pays seulement sont analysés, conduit vite, dès qu'on augmente le nombre de pays, à des modélisations très lourdes en termes de nombre de variables (car admettre que l'effet de telle variable peut différer selon les pays, c'est s'obliger à croiser cette variable avec les indicatrices de pays) et à des commentaires excessivement complexes. C'est typiquement le cas lorsqu'on se penche sur les Quinze pays membres de l'Union européenne¹¹ : comme **l'énumération des situations pays par pays est fastidieuse**, le commentateur des analyses statistiques en est réduit à quelques coups de projecteur sur telle et telle situations extrêmes ou à forger des typologies regroupant les pays, de façon souvent assez artificielle, en quelques groupes constitués chacun d'un tout petit nombre de pays.

L'idée, pour contourner cet obstacle de la complexité et de la lourdeur, est alors de ne plus donner aux indicatrices de pays l'exclusivité dans la représentation des différences nationales : on va les compléter par des variables permettant de situer chaque pays sur les mêmes échelles continues (le PIB/tête, par exemple, ou le ratio dépenses de protection sociale sur PIB)¹².

¹¹ Et ce le sera plus encore avec l'élargissement – qui a déjà commencé en termes de statistiques, avec l'association de certains pays candidats à la production de données selon les normes d'Eurostat.

¹² Ce choix, apparemment purement technique, a par ailleurs une certaine portée politique puisqu'il estompe les clivages nationaux traditionnels, et cela au profit de critères analogues à ceux dont on use déjà au sein de chaque pays c'est-à-dire quand l'intégration a permis de ne plus raisonner uniquement en termes de régions (qu'on se borne à comparer les une aux autres) mais aussi, et cette fois toutes régions confondues, en termes de revenu, d'âge, de générosité des politiques familiales, etc., ce qui postule une certaine homogénéité inter-régions. Autrement dit, au lieu de ne cesser de souligner des particularismes nationaux comme le font la plupart des études sur l'Europe, des modèles spécifiés comme nous venons de l'expliquer permettraient de gommer un peu la dimension 'pays' ; en effet, sur des échelles quantitatives du type de celles que nous proposons, chaque pays peut certes se situer, mais il peut aussi rentrer dans une comparaison sans qu'on fasse nécessairement référence à son nom (ex. : sur le continuum de l'échelle du PIB/tête, on repérera non seulement des pays en tête et d'autres en queue, mais on pourra dire de combien les uns et les autres, en tant que groupes, s'écartent de la moyenne communautaire). D'abord technique, notre choix rejoint donc les vœux des responsables européens, soucieux qu'ils sont que, dans les études sur l'Europe, le critère national cesse d'être omniprésent.

On passera donc **d'une modélisation fondamentalement qualitative** (consistant à recourir exclusivement aux variables indicatrices des pays), à **une approche faisant appel à des variables quantitatives** constituant des dénominateurs communs¹³ aux différents pays (PIB/tête, taux de chômage, générosité des transferts sociaux, etc.) [RAY 1997]. Ce qui n'empêchera d'ailleurs pas de lire aussi les résultats des modèles en termes de situation pays par pays¹⁴, mais ce qui n'en fait plus l'unique mode de commentaire des analyses multi-pays. En outre, en remplaçant en partie la **boîte noire** que constitue la variable 'pays' par des variables macro pertinentes, on rendra possible le test d'hypothèses concernant l'existence d'**effets d'interaction**¹⁵ entre niveaux (micro et macro).

¹³ Ce que l'on peut reprocher, en effet, à une approche utilisant exclusivement, au niveau macro, des indicatrices de pays, ce n'est pas de postuler l'incommensurabilité des différences entre pays ; car la comparaison des coefficients de régression des indicatrices nationales donne, justement, la mesure des différences de valeur, d'un pays à l'autre, de la variable expliquée ; et les différences des coefficients des variables explicatives croisées avec les indicatrices de pays mesurent les différences d'effet qu'ont, selon les pays, ces variables explicatives sur la variable expliquée. Mais n'utiliser que la variable 'pays' pour tenir compte des différences entre pays, ce serait s'interdire de séparer les rôles des diverses dimensions, pas nécessairement homothétiques entre pays, que recouvre la variable nationale ; autrement dit, on constate des différences de niveau moyen de la variable expliquée et d'action des X sur Y selon les pays, mais on ne sait pas à quoi les attribuer vraiment : à des différences de richesse entre pays ? A des différences de générosité globale de la protection sociale ? A des divergences dans la générosité de tel type de transferts sociaux ?

¹⁴ Avec la procédure PROC MIXED de SAS, il suffit de spécifier l'option SOLUTION dans l'instruction RANDOM pour obtenir, pour chaque pays, une estimation des coefficients de régression qui lui sont propres.

¹⁵ "**Effets d'interaction**" est, on le sait, l'expression consacrée pour désigner une action conjuguée de deux facteurs, non pas l'un sur l'autre et réciproquement, mais le niveau de l'un renforçant ou affaiblissant l'impact de l'autre sur la variable dépendante. En régression multiple traditionnelle, ces deux facteurs caractérisent l'individu (ex.: le fait de fumer aura un effet d'aggravation de l'impact de la pilule sur les risques cardio-vasculaires, et réciproquement). En matière de modèles multi-niveaux aussi, on peut ajouter, aux effets principaux de chaque variable, des effets d'interaction, et cela à la fois entre variables au niveau individuel et entre variables au niveau des groupes ; mais, par construction, ces modèles incorporent des effets d'interaction entre niveaux. C'est ainsi qu'on se demandera non seulement, par exemple, si le sexe (variable de niveau individuel) influence, ceteris paribus, les gains des jeunes adultes, et si le niveau des transferts sociaux (variable de niveau pays) a un impact direct sur ces gains, mais on aura aussi les moyens de tester, et de façon techniquement valable (ce qui ne serait pas le cas avec les moindres carrés ordinaires), si le niveau des transferts sociaux renforce ou affaiblit le rôle du sexe sur les gains d'activité. On notera au passage que l'interprétation symétrique (à savoir : le rôle du sexe de l'individu sur le niveau des transferts sociaux dans le pays), bien que techniquement possible, n'aurait, ici, pas beaucoup de sens ; ce qui pousserait, si sa longueur n'était pas rédhibitoire, à préférer une expression asymétrique (comme "effets de l'appartenance à un groupe sur l'impact des facteurs individuels") au terme "interaction", qui donne, à tort ici, l'image de la réciprocité des actions.

CHAPITRE II

LE MODELE

Après avoir rappelé par quelles variables la **théorie économique** suggère d'expliquer les gains d'activité, et quel est le signe de chaque effet attendu, on présente le **modèle à deux niveaux** que l'on s'efforcera ensuite d'estimer au mieux, en fonction des données disponibles.

1. LES PRÉDICTEURS DES GAINS D'ACTIVITÉ

1. Au niveau individuel, les prédicteurs classiques sont ceux qui expliquent les deux termes de ce produit que sont les gains d'activité : les **heures de travail** accomplies et le **taux horaire de rémunération**. Il ne faudrait pas oublier cependant qu'un effet revenu peut jouer aussi, conduisant à inclure, parmi les variables explicatives des gains du jeune adulte, ses **autres sources de revenus** et, plus largement les divers types de revenus dont disposent les membres du ménage abritant le jeune adulte ou lui assurant à distance un certain niveau de vie.

Parmi ces autres revenus, on trouve les transferts sociaux, dont la prise en compte n'est pas simple, car elle pose un **problème d'endogénéité** : le montant des prestations sous conditions de ressources va en effet dépendre, en partie, des gains d'activité du jeune adulte (phénomène de causalité réciproque) ; ce qui rend problématique l'inclusion naïve, au titre des facteurs explicatifs des gains (avec des coefficients attendus négatifs, en raison de l'effet revenu), des montants de transferts sociaux perçus par le jeune adulte ou son ménage. Ce n'est pas un des moindres avantages de la prise en considération au niveau collectif, de ces transferts, que d'éviter ce biais d'endogénéité. Considérons, par exemple, le niveau national moyen des prestations familiales : il constitue un indicateur de ce qu'une famille résidant dans le pays en question peut espérer toucher ; c'est un indicateur certes assez grossier, puisqu'il ignore la situation particulière de l'allocataire et de sa famille ; or, selon que la famille sera ou non monoparentale, selon le nombre de ses enfants et selon leur(s) âge(s), selon la façon dont les revenus de la famille se situent vis-à-vis des différents plafonds de ressources, etc., le montant des prestations familiales espéré pourra beaucoup varier. Mais cet indicateur a l'avantage d'être corrélé à ce qu'une famille peut espérer toucher, et non pas à ce qu'elle touche de façon endogène, c'est-à-dire compte tenu des choix d'activité faits par ses membres lorsqu'ils prennent en considération, entre autres, le barème des prestations familiales du pays en question. Autrement dit, notre indicateur dépend certes, de diverses manières¹⁶, des choix d'activité des familles du pays, mais il ne dépend guère des choix particuliers d'activité faits par les membres de la famille du jeune adulte dont on essaie d'expliquer les gains.

¹⁶ A la dépendance directe (par l'effet mécanique de distribution de davantage d'allocations ; par exemple, plus forte est la proportion des mères qui travaillent, plus fort sera le montant moyen des prestations visant à permettre de mieux concilier vie familiale et vie professionnelle) s'ajoute en effet la dépendance par le relais des choix de politique familiale, qui sont eux-mêmes doublement influencés par l'importance de l'activité marchande des jeunes mères de famille : plus de revenus du travail distribués, c'est plus de cotisations et d'impôts, ce qui peut faciliter le financement d'une politique familiale généreuse ; et généreuse, elle doit l'être pour offrir des places de crèche bon marché, des allocations parentales d'éducation, des remboursements de frais de garde à domicile, etc., si on ne veut pas que l'activité des mères se traduise par une baisse trop prononcée de la fécondité.

Mais, avant de passer à ces transferts sociaux considérés au niveau collectif, rappelons les déterminants individuels fondamentaux des gains d'activité : du côté de l'offre de travail (en termes d'heures) comme du côté du taux horaire, **l'âge** joue un rôle capital (sans être soupçonné d'endogénéité, lui) : à mesure que le jeune passe du statut d'adolescent scolarisé à celui d'adulte inséré dans la vie active, heures de travail et taux horaires de rémunération croissent, gonflant les gains à la fois progressivement et avec, du moins au niveau individuel, les sauts soudains associés au passage de petits travaux ponctuels à l'occupation d'un emploi... et à sa perte éventuelle¹⁷.

La seconde variable démographique (exogène, elle aussi¹⁸) susceptible d'expliquer les gains, est **le sexe** du jeune adulte : on peut en effet supposer que, selon son genre, le jeune adulte aura une offre de travail différente (notamment, pour les jeunes femmes, dans les pays traditionnels, en fonction de leurs projets matrimoniaux et familiaux) et, par ailleurs, on ne peut exclure une différenciation des taux horaires de gains selon le sexe (tournant à la discrimination si ces écarts persistent quand on contrôle des facteurs comme l'âge, le niveau et le type de diplôme, l'expérience professionnelle, etc.).

A ces deux variables démographiques s'ajoute le **capital humain**. Comme la source de données que nous utiliserons nous obligera à le mesurer de façon extrêmement rudimentaire, nous n'allons pas raffiner ici l'analyse théorique de son rôle ; son impact attendu est positif, sans qu'on sache dire si le fait d'avoir bénéficié d'une éducation supérieure majore le taux horaire (et la probabilité de trouver un emploi), par rapport au cas du titulaire d'une éducation de niveau intermédiaire, davantage ou moins (et de combien) que ce n'est le cas lorsque l'on compare les avantages supplémentaires attendus d'une éducation de niveau intermédiaire par rapport à une éducation de niveau faible : tout, ici, est dépendant des nomenclatures choisies, et on sait combien est problématique la fusion des nomenclatures nationales de diplômes dans une nomenclature européenne unique.

Même si la littérature théorique et empirique a pu mettre en avant de nombreux autres facteurs de différenciation des gains d'activité (le secteur, notamment), nous nous en tiendrons aux trois facteurs que nous venons de mentionner (sexe, âge, éducation), la plupart de ces autres facteurs étant sans objet pour une fraction importante des jeunes adultes (les scolaires, les chômeurs, les mères au foyer, etc.) que nous étudions.

¹⁷ La question empirique est ici celle de l'**allure du profil d'évolution des gains** à mesure que le jeune adulte prend de l'âge. La théorie économique prédit, certes, les profils de gains sur toute la carrière, mais, si elle considère qu'en début de celles-ci, les gains vont croissant, elle ne précise pas le rythme de cette croissance ; cette question ressortit dès lors aux analyses empiriques. Si ce profil est quasi-linéaire, on pourra avantageusement remplacer le jeu de variables indicatrices des années d'âge par la variable continue 'âge', ce qui allégera d'autant plus le modèle qu'on aura souhaité croiser cette variable avec d'autres pour rendre compte d'éventuelles différences (entre pays, entre sexes, etc.) des effets de l'âge sur les gains. Le graphique proposé en annexe, qui représente les profils, selon l'âge, des moyennes (pays par pays) des gains personnels d'activité des jeunes adultes européens permet de conclure que, entre 17 ans et 30 ans ces profils peuvent être raisonnablement approximatés par des droites (de pentes variables, remarquablement groupées et plus faibles pour les quatre pays du Sud de l'Europe, ce qui n'est pas une surprise étant donné le niveau moyen des gains à 30 ans, presque deux fois moindres que ceux de la Belgique et des Pays-Bas).

¹⁸ Du moins, on le suppose.

2. Au niveau collectif, les prédicteurs des gains d'activité doivent concerner les unités qui vont constituer les groupes du modèle. Il s'agit ici des pays. Et ces prédicteurs, au-delà de leur action directe sur les gains des jeunes adultes, doivent pouvoir rendre compte du fait que, d'un pays à l'autre, l'impact, sur les gains, des prédicteurs individuels peut différer (ce sont les effets d'interaction discutés supra).

a) Pour satisfaire à la première exigence, deux possibilités complémentaires s'offrent :

- prendre des **indicateurs macro classiques**, qu'il s'agisse du PIB/tête, du taux de chômage (des jeunes, de préférence), des dépenses de protection sociale (entendues globalement ou par grande fonction : aides aux familles, allocations de logement, bourses scolaires, indemnisation du chômage, prestations d'assistance – en excluant donc les retraites, les soins médicaux et les prestations d'invalidité, qui concernent moins¹⁹ les jeunes adultes que la population générale)
- construire, **pour les jeunes adultes de l'échantillon**, des **indicateurs de valeurs moyennes** (de tel type de prestation, par exemple), et des **indicateurs de fréquence** (le pourcentage de chômeurs, par exemple).

Ce second type d'indicateurs présente l'avantage de mieux cibler la population des jeunes adultes que ne le font la plupart des indicateurs macro disponibles ; ainsi, par exemple, EUROSTAT publie certes, pour chacun des Quinze, un taux de chômage des jeunes, mais il s'agit des 18-25 ans (alors que nous étudions les 16-30 ans). Le revers de la médaille est la fragilité des mesures ne reposant que sur des échantillons plus petits que ceux (de l'enquête Emploi par exemple) qui servent à l'établissement des statistiques officielles.

b) La seconde exigence impose de faire appel à la théorie. D'abord, pays par pays, quels prédicteurs nationaux sont susceptibles d'avoir une action directe sur les gains des jeunes adultes ? Ensuite, quels facteurs peuvent rendre compte du fait que, d'un pays à l'autre, l'impact, sur les gains, des prédicteurs individuels peut différer (ce sont les effets d'interaction discutés supra).

Au titre des premiers, nous avons déjà cité les **transferts sociaux**, qui constituent notre variable d'intérêt. Précisons que, si on veut tenter d'apprécier l'impact spécifique de chaque grand type de transferts sociaux, il faut s'attendre à un problème aigu de multicollinéarité, du fait que certains pays ont des prestations sociales relativement généreuses pour tous les risques (chômage, famille, etc.) tandis que d'autres (le Sud, pour simplifier) ont une protection sociale plus spartiate. Comme le suggèrent les Analyses en Composantes Principales que nous avons faites (ACP qui font ressortir nettement un premier axe s'identifiant au degré global de générosité de la protection sociale du pays), il serait opportun d'introduire un **indicateur général**, de type transferts sociaux moyens ou ratio dépenses sociales/PIB (voire PIB/tête, si la richesse du pays n'est pas systématiquement liée à la générosité de sa protection sociale), et de le compléter par des **indicateurs spécifiques** rendant compte, à générosité globale donnée, de l'effort plus spécialement fait en direction des familles, des chômeurs, du logement, des bourses, des pauvres, etc.

¹⁹ "Moins" mais pas nécessairement "pas", car certains membres des familles des jeunes adultes perçoivent ces types de revenus.

Aux transferts sociaux il convient d'ajouter, en tant que variables de contrôle :

- un **indicateur des difficultés à trouver du travail** (ce qui fait peser une contrainte sur les possibilités de gains)
- un indicateur de la fréquence avec laquelle les jeunes adultes **sont étudiants** (encore que, dans certains pays comme l'Allemagne et l'Autriche, études ne riment pas toujours avec absence ou faiblesse des gains d'activité)
- des indicateurs renseignant sur la proportion des **jeunes adultes qui ont terminé leur scolarité en obtenant un diplôme de tel ou tel niveau**, car plus le niveau moyen d'éducation dans le pays est faible plus les gains d'activité risquent de l'être, en raison d'une faible productivité générale du travail
- un indicateur de la fréquence avec laquelle les jeunes adultes **ont un conjoint** (on ignore l'effet de cette variable collective sur les gains individuels, d'autant que, au niveau individuel l'effet net n'est pas prévisible : on peut s'attendre en effet à la fois à un effet positif sur les gains – s'il faut gagner sa vie et celle de son conjoint – et à un effet négatif – si les gains du conjoint permettent de vivre à deux)
- un indicateur de la fréquence des cas où les jeunes adultes **vivent chez leurs parents** (a priori, l'effet de cette variable collective sur les gains individuels devrait être plutôt négatif, sauf dans les cas où, à l'instar de ce qui se passe en Espagne par exemple, la cohabitation avec les parents se poursuit, pour les jeunes adultes, jusqu'à des âges avancés, en raison des grandes difficultés à trouver et financer un logement indépendant)
- et un indicateur du montant moyen des **transferts privés reçus** par le ménage où vit le jeune adulte : dans un pays où la solidarité privée (essentiellement familiale, entre générations) est forte, les gains des jeunes adultes n'ont pas autant besoin d'être élevés que dans les pays où il est de tradition que les jeunes adultes se débrouillent davantage par eux-mêmes, quitte à recourir à la solidarité publique.

Ceci dit, quels sont les facteurs nationaux qui sont susceptibles de modifier, d'un pays à l'autre, l'impact, sur les gains, de ces trois facteurs individuels que sont le sexe, l'âge et le niveau d'éducation ? De ce côté-là, la théorie économique n'est pas très diserte, et il nous faut donc recourir au simple raisonnement.

On supposera ici que **l'effet de l'âge sur les gains est influencé** (négativement) par le **taux de chômage** et (positivement) par la **durée moyenne des études**²⁰ **dans le pays** (on suppose ici que les pays où l'investissement en capital humain est plus important parviennent à des niveaux de productivité du travail qui autorisent des gains d'activité croissant, avec l'âge, plus vite qu'ailleurs).

²⁰ La durée moyenne des études traduit, au niveau national, la variable de capital humain individuel. On pourrait penser que ces variables posent un problème technique particulier (au-delà de la question de sa disponibilité dans la source statistique utilisée) ; en effet, une fraction substantielle des jeunes adultes étudiés sont en cours d'études, et, pour eux, leur niveau d'éducation actuel ne reflète pas nécessairement ce qu'il sera plus tard. Il semble que cette censure des durées d'études ne soit pas trop gênante ici, puisque les possibilités de gains des jeunes adultes sont souvent fonction, à un âge donné, du niveau d'éducation acquis jusque-là (beaucoup d'étudiants, en France, font des petits boulots payés au SMIC) ; il est vrai qu'un élève-ingénieur trouvera sans doute, pour mille raisons, des stages mieux rémunérés que son conserit titulaire, comme lui, d'un baccalauréat mais s'étant arrêté là.

Et on fera l'hypothèse que l'effet, sur les gains, du fait d'être une femme, est influencé (négativement) par une certaine conception, traditionnelle, du rôle de la femme ; conception que nous envisageons d'approximer par un indicateur objectif : **la part des femmes qui, dans chaque pays, se déclarent "au foyer"**.

2. NOTRE MODÈLE À DEUX NIVEAUX

Après le rappel des conséquences de la non-indépendance des observations et une brève présentation de la structure des équations d'un modèle multi-niveaux, viendra l'exposé du modèle qu'on s'efforcera par la suite d'estimer.

2.1. La non-indépendance des observations et ses conséquences

Dès qu'on étudie des **individus qui appartiennent à divers groupes** (différents pays, ici), les observations dont on dispose ont **peu de chances d'être indépendantes les unes des autres** ; et plus les groupes sont de grande taille, plus est accentué ce phénomène de non indépendance puisque de très nombreux individus partagent les mêmes caractéristiques. Or les méthodes classiques de régression postulent l'indépendance des observations. Cette hypothèse est cruciale parce que, si les erreurs ne peuvent pas être considérées comme non corrélées les unes aux autres, la matrice des variances-covariances des erreurs (qui entre dans le calcul de la matrice des variances-covariances des estimateurs de la régression multiple classique) devient bien plus complexe qu'elle ne l'est lorsqu'il y a indépendance des erreurs ; dans ce dernier cas, en effet, tous les termes de covariance s'annulent²¹.

De même, quand on écrit la fonction de vraisemblance d'un échantillon, on le fait en multipliant les probabilités individuelles d'observer telle valeur de la variable dépendante ; mais ce simple produit de probabilités conditionnelles aux variables indépendantes n'est pertinent que si les diverses observations ont des probabilités qui ne dépendent pas les unes des autres. Or les conséquences de la non prise en compte de la non indépendance des observations peuvent être graves puisque, si cette dépendance ne biaise pas les estimateurs, elle affecte en revanche leurs écarts-types, qui se trouvent être sous-estimés ; d'où des tests statistiques conduisant à trouver, à tort, que des effets existent alors qu'on a trop peu d'informations réellement indépendantes pour pouvoir l'assurer.

²¹ Cette matrice se ramène donc à une diagonale de termes non nuls, qui représente les variances des erreurs ; s'il y a, de surcroît, homoscedasticité, chacune de ces variances devient égale à un terme unique, ce qui permet une simplification radicale de la matrice complexe initiale : elle se ramène à un scalaire (la variance des erreurs, supposée constante) multipliant la matrice unité.

Et, naturellement, ce phénomène s'accroît avec la taille moyenne des groupes, puisque de plus en plus d'individus partagent alors les mêmes informations. D'où le recours aux modèles multi-niveaux, dont on espère²² qu'ils s'appliquent bien aux différents cas de données hiérarchiques.

2.2. Présentation de la structure-type des équations d'un modèle multi-niveaux

Afin de tenir compte correctement de la dépendance des observations, les unes vis-à-vis des autres, au sein de chaque groupe (ce qu'on appelle la corrélation intra-classe), on commence par écrire le modèle en reconnaissant explicitement l'existence de plusieurs niveaux, ici deux seulement : le niveau individuel (niveau 1) et le niveau national (niveau 2).

Au niveau 1 on fait figurer dans le membre de droite de l'équation expliquant les gains du jeune adulte les seules variables individuelles (ici, le sexe, l'âge et le niveau d'éducation) et un terme d'erreur. La particularité de cette équation est que les coefficients de certaines variables individuelles (celles que l'on veut considérer comme n'ayant pas des effets fixes mais des effets aléatoires) sont considérés comme des variables aléatoires.

Au niveau 2 on entreprend d'expliquer chacun de ces coefficients. On le fait par autant d'équations (plus une, pour la "constante" aléatoire de l'équation de niveau 1) qu'il y a de variables explicatives que l'on veut considérer comme ayant des effets aléatoires. Ces diverses équations de niveau 2 peuvent ne comporter qu'une constante (fixe) ou voir s'y ajouter une ou plusieurs variables explicatives de niveau 2, caractérisant donc chacun des groupes (et ayant la même valeur pour tous les individus appartenant à un même groupe). Ces variables explicatives de niveau 2 peuvent être communes à plusieurs équations de niveau 2 ou spécifiques à certaines d'entre elles.

²² Les manuels de statistique qui se préoccupent d'aider le lecteur dans le choix de méthodes appropriées à son problème et à ses données insistent souvent, et à raison, sur la **nature des variables** en cause : ainsi l'étude de l'intensité de la liaison entre deux variables, par exemple, se fera-t-elle avec des instruments analytiques et graphiques différents selon que les variables d'intérêt sont continues, discrètes, binaires, qualitatives ordonnées ou purement nominales ; et il en va de même pour les méthodes multivariées, la nature des variables dictant alors le choix entre ACP et AFCM, le choix de la distance en classification, le critère de segmentation, etc. D'autres caractéristiques des données peuvent jouer aussi, comme la **taille de l'échantillon** (notamment pour savoir quel test appliquer, autour d'un seuil de l'ordre de 30 pour le choix entre un Student et un test reposant sur la loi de Gauss, ou autour d'un seuil de 2000 pour le choix entre le critère de normalité W de Shapiro-Wilk et le D de Kolmogorov-Smirnov, par exemple). Mais, pour les **données à plusieurs dimensions**, les critères sont plus complexes. Ainsi, en cas de données de panel, croisant des dimensions individu et temps, les méthodes statistiques différeront sensiblement selon que l'on dispose de quelques individus (des firmes, par exemple) suivis sur de nombreuses années, ou de nombreux individus suivis uniquement sur quelques vagues (cas fréquent des panels de ménages, à la notable exception de panels à forte longévité, comme le PSID, plus que trentenaire) : ainsi GREENE [1992] structure-t-il son chapitre intitulé "Models That Use Both Cross-Section and Time-Series Data" en opposant, sans critère précis de distinction reposant sur des nombres-seuils d'individus et de vagues, les "Models of Several Time Series" au cas des "Longitudinal Data", chaque type de données étant justiciable de méthodes économétriques particulières ; et c'est dans cette seconde section qu'il traite des "Random Effects [Models]" (qui s'appliquent lorsque les effets spécifiques à chaque individu peuvent être considérés moins en eux-mêmes que comme donnant une idée des effets pour la population dont ont été tirés les individus) alors que c'est dans la première qu'il consacre un paragraphe au cas du "Random Coefficients Model" (qui s'applique lorsqu'on peut supposer qu'une partie des différences entre les valeurs des paramètres pour les quelques individus étudiés s'explique par des différences réelles, ce qui limite le rôle du terme d'erreur). De même, **en cas de niveaux multiples**, il serait utile de savoir non seulement si on peut valablement appliquer tel ou tel type de modèles multi-niveaux au cas où les groupes sont de petite taille (en moyenne, comme lorsqu'on étudie des couples ou des frères et sœurs), et/ou certains groupes seulement le sont (lorsque telle école ne compte que quelques élèves) ou sont peu nombreux, mais aussi si les méthodes changent lorsque, à nombre total d'individus donné, la configuration des données s'infléchit (moins de groupes, mais plus nombreux chacun, en moyenne ; ou plus de groupes, mais comprenant chacun moins de membres) ; or les réponses à ce type d'interrogations essentielles pour l'utilisateur sont difficiles à trouver.

La présentation la plus naturelle d'un modèle multi-niveaux consiste donc à écrire ce modèle niveau par niveau ; c'est la **forme structurelle**. Mais, par remplacement, dans l'équation de premier niveau, des coefficients aléatoires par les équations qui les expliquent au second niveau, on peut en déduire, comme nous allons le voir, une équation unique, qui constitue la **forme réduite**²³ du modèle.

Pour illustrer ces deux modes d'écriture, on se limitera ici à trois facteurs explicatifs : un facteur de niveau 1, appelé X, et deux facteurs de niveau 2, appelés Z, qui expliquent l'un (Z1) la variable aléatoire β_{0j} (c'est la "constante" du niveau 1), et l'autre (Z2) la variable aléatoire β_{1j} (c'est le coefficient de X). Il y a J groupes, indicés par la lettre j, et N individus, indicés à la fois par la lettre i et par la lettre j (pour marquer leur appartenance au groupe j).

Niveau 1 :

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \quad \text{où } r_{ij} \approx N(0, \sigma^2)$$

Niveau 2 : il y a, à ce niveau 2, autant d'équations qu'il y a, au niveau 1, de coefficients aléatoires :

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z1_j + u_{0j} \quad \text{où } u_{0j} \approx N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z2_j + u_{1j} \quad \text{où } u_{1j} \approx N(0, \tau_{11})$$

$$\text{et } \text{cov}(u_{0j}, u_{1j}) = \tau_{01} \quad (= \tau_{10})$$

Pour passer de la forme structurelle ci-dessus (trois équations) à la **forme réduite** (une seule équation, servant à l'estimation), il suffit de substituer aux coefficients aléatoires (β_{0j} et β_{1j}) de l'équation de niveau 1 leurs valeurs telles que spécifiées dans les équations de niveau 2.

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z1_j + u_{0j} + (\gamma_{10} + \gamma_{11}Z2_j + u_{1j})X_{ij} + r_{ij}$$

On arrange ensuite les termes de façon à regrouper les trois termes aléatoires à la fin de l'équation :

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z1_j + \gamma_{10}X_{ij} + \gamma_{11}Z2_jX_{ij} + u_{0j} + u_{1j}X_{ij} + r_{ij}$$

Le terme d'erreur est, on le voit, composé et dépendant de la variable explicative de premier niveau, ce qui disqualifie les moindres carrés ordinaires comme méthode d'estimation et oblige à se tourner vers d'autres méthodes, et notamment les moindres carrés généralisés²⁴.

Après ce bref exposé de la structure typique des équations d'un modèle multi-niveaux, venons-en à l'exposé de notre modèle.

²³ Certains auteurs utilisent à ce propos l'expression "combined model" [Annie QU].

²⁴ Pour en savoir plus, voir BRYK et RAUDENBUSH [1992], ou HOX [1995] ou KREFT et DE LEEUW [1998] ou GOLDSTEIN [1999] ou encore SNIJDERS et BOSKERS [1999].

2.3. Exposé du modèle à deux niveaux

Niveau 1 : une équation unique :

$$G_{ij} = \beta_{0j} + \beta_{1j} AGE_{ij} + \beta_{2j} SEXE_{ij} + \beta_{3j} EDUC_{ij} + r_{ij}$$

où $r_{ij} \approx N(0, \sigma^2)$

où : G_{ij} représente les gains d'activité du jeune adulte i appartenant au pays j

AGE_{ij} représente l'âge du jeune adulte en question

$FEMME_{ij}$ est l'indicatrice du sexe du jeune adulte

$EDUC_{ij}$ représente le niveau d'éducation du jeune adulte

r_{ij} est le terme d'erreur afférent au jeune adulte concerné ; ce terme est supposé distribué normalement, avec une espérance nulle et une variance égale à σ^2

Enfin, les quatre coefficients β sont tous considérés, ici, comme des variables aléatoires²⁵.

Ecrit sous sa forme structurelle, le modèle comprend donc, au second niveau, quatre équations, une par coefficient β à expliquer :

Niveau 2, première équation :

$$\begin{aligned} \beta_{0j} = & \gamma_{0,0} + \gamma_{0,1} GENEROSITE_j + \gamma_{0,2} ALLOC_CHOM_j + \gamma_{0,3} ALLOC_FAM_j \\ & + \gamma_{0,4} ALLOC_LOGT_j + \gamma_{0,5} ALLOC_EDUC_j + \gamma_{0,6} ALLOC_ASSIST_j \\ & + \gamma_{0,7} SOLIDARITE_PRIVEE_j + \gamma_{0,8} RICHESSE_j + \gamma_{0,9} TAUX_CHOM_j \\ & + \gamma_{0,10} FREQ_PARENTS_j + \gamma_{0,11} FREQ_CONJOINT_j + \gamma_{0,12} FREQ_EDUC3_j \\ & + \gamma_{0,13} FREQ_EDUC2_j + \gamma_{0,14} FREQ_EDUC1_j + \gamma_{0,15} FREQ_ETUDIANTS_j \\ & + u_{0j} \end{aligned}$$

où $u_{0j} \approx N(0, \tau_{00})$

où :

$GENEROSITE_j$ représente la générosité globale de la protection sociale dans le pays j (ou, mieux, la générosité globale de la protection sociale s'adressant spécifiquement aux ménages comprenant au moins un jeune adulte ; le caractère souhaitable de cette restriction de la population concernée s'étend aux autres indicateurs, sauf $RICHESSE_j$)

$ALLOC_CHOM_j$ est un indicateur sans dimension représentant la générosité relative des allocations de chômage dans le pays j (pour cela on rapporte le montant moyen des allocations

²⁵ A priori on considère que les effets de chacun de ces facteurs peuvent différer selon les pays ; ce n'est que si les résultats de l'estimation de leur variabilité selon les pays montre qu'elle est faible que l'on pourra modifier la spécification et les considérer comme des effets fixes, c'est-à-dire comme identiques d'un pays à l'autre.

de chômage au montant moyen de l'ensemble des prestations sociales)
ALLOC_FAM_j : idem pour les prestations reçues au titre de la famille
ALLOC_LOGT_j : idem pour les allocations de logement
ALLOC_EDUC_j : idem pour les bourses scolaires
ALLOC_ASSIST_j : idem pour les prestations d'assistance
SOLIDARITE_PRIVEE_j représente l'ampleur moyenne des transferts privés reçus par les ménages du pays j
RICHESS_j caractérise le pays j dans lequel réside le jeune adulte
TX_CHOM_j est le taux de chômage dans le pays j dans lequel réside le jeune adulte
FREQ_PARENTS_j est la fréquence, pour le pays j dans lequel réside le jeune adulte, de cohabitation des jeunes adultes avec leurs parents
FREQ_CONJOINT_j est la fréquence, pour le pays j dans lequel réside le jeune adulte, de la cohabitation des jeunes adultes avec un conjoint
FREQ_EDUC3_j est la fréquence, pour le pays j dans lequel réside le jeune adulte, des jeunes adultes qui ont atteint un niveau d'éducation supérieure
FREQ_EDUC2_j est la fréquence, pour le pays j dans lequel réside le jeune adulte, des jeunes adultes qui ont atteint un niveau d'éducation intermédiaire
FREQ_EDUC1_j est la fréquence, pour le pays j dans lequel réside le jeune adulte, des jeunes adultes qui ont un niveau faible d'éducation
FREQ_ETUDIANTS_j est la fréquence, pour le pays j dans lequel réside le jeune adulte, des jeunes adultes qui sont étudiants.

Niveau 2, deuxième équation :

$$\beta_{1j} = \gamma_{10} + \gamma_{11} \text{TX_CHOM}_j + \gamma_{12} \text{DUREE_ETUDES}_j + u_{1j} \\ \text{où } u_{1j} \approx N(0, \tau_{11})$$

où DUREE_ETUDES_j est la durée moyenne des études dans le pays j dans lequel réside le jeune adulte.

Niveau 2, troisième équation :

$$\beta_{2j} = \gamma_{20} + \gamma_{21} \text{FREQ_AU_FOYER}_j + u_{2j} \quad \text{où } u_{2j} \approx N(0, \tau_{22})$$

où FREQ_AU_FOYER_j est la fréquence, pour le pays j dans lequel réside le jeune adulte, des jeunes adultes qui déclarent être au foyer.

Niveau 2, quatrième équation :

$$\beta_{3j} = \gamma_{30} + u_{3j} \quad \text{où } u_{3j} \approx N(0, \tau_{33})$$

Spécifié de la sorte, le coefficient β_{3j} de la variable $EDUC_{ij}$ est certes aléatoire, mais il n'est pas expliqué par une ou plusieurs variables de second niveau.

Il reste à préciser la notation adoptée pour les covariances des termes d'erreur, dont les valeurs feront par la suite l'objet d'hypothèses pour réduire le nombre de paramètres à estimer au total.

$$\text{cov}(u_{0j}, u_{1j}) = \tau_{01}, \text{cov}(u_{0j}, u_{2j}) = \tau_{02}, \text{cov}(u_{0j}, u_{3j}) = \tau_{03}$$

$$\text{cov}(u_{1j}, u_{2j}) = \tau_{12}, \text{cov}(u_{1j}, u_{3j}) = \tau_{13}$$

$$\text{cov}(u_{2j}, u_{3j}) = \tau_{23},$$

Soit, au total, 22 paramètres de type γ , et 11 variances ou covariances à estimer (le paramètre σ^2 et 10 τ).

L'examen des caractéristiques de la base de données dont on dispose va nous permettre d'apprécier dans quelle mesure il est possible d'estimer un tel modèle à partir de ces données.

CHAPITRE III

LES DONNEES

Nous conjuguons l'usage de deux sources de données :

- le **SESPROS** (Système Européen de Statistiques intégrées de PROtection Sociale, *ESSPROS* en anglais) nous fournit les **indicateurs macro** (publications papier et site Web d'EUROSTAT) issus des comptes nationaux harmonisés
- l'**ECHP** (European Community Household Panel, appelé en France '**panel communautaire de ménages**').

Bien qu'on dispose maintenant de cinq vagues (1994, 1995, 1996, 1997 et, depuis deux mois, 1998), nous travaillerons ici exclusivement sur des données en coupe, relatives à la quatrième vague (**enquête 1997**, avec revenus de l'année 1996), qui inclut les Quinze pays de l'Union Européenne (sauf le Luxembourg, dont les données de l'enquête PSELL sont en phase finale d'adaptation aux normes ECHP). Cette vague comprend 138345 individus de 16 ans et plus, qui sont membres de 66137 ménages, ce qui fournit un échantillon de **36006 jeunes de 16 à 30 ans**. Le choix de cette vague tient à ce que les vagues antérieures incluent moins de pays (la Suède n'a rejoint l'ECHP que depuis 1997, la Finlande depuis 1996, l'Autriche depuis 1995) et les révisions substantielles [EUROSTAT 2001c] qui ont été apportées en décembre 2001 aux données de 1997 incitent à ne pas faire confiance d'emblée aux données de la cinquième vague.

L'ECHP, malgré ses limites (voir CHAMBAZ 2000, encadré 1, page 2), constitue une **source assez homogène**, et donc précieuse à ce titre, pour les pays de l'Union européenne. Il s'agit en outre d'une **source relativement récente**, et qui **détaille revenus et composition familiale**, ce qui est crucial pour le repérage des jeunes adultes et de leurs gains d'activité, pour l'attribution à chacun d'eux de caractéristiques de son ménage mais aussi pour l'identification (en combinant les informations des bases Individus, Registre et Relations) de l'existence, dans le ménage où vit le jeune adulte, d'un éventuel conjoint et d'un ou deux des parents du jeune adulte.

La taille assez substantielle des échantillons nationaux (voir en Annexe le tableau 1, colonnes 2 à 4, et colonne 6) autorise l'obtention d'effectifs non négligeables à l'issue de la sélection de la tranche d'âge (de 16 à 30 ans) sur laquelle porte l'étude : pour 1997, on dénombre 36006 jeunes adultes.

Cette taille permet aussi de construire des variables contextuelles (par exemple un indicateur de la générosité de la politique familiale nationale vis-à-vis des jeunes adultes, ou un taux de chômage des jeunes adultes) constituant, du fait de leur ciblage sur la population étudiée, des solutions alternatives intéressantes par rapport aux indicateurs macro issus des comptes nationaux.

Bien que portant sur une vague antérieure (la troisième vague, enquête en 1996, revenus de 1995, Luxembourg inclus mais Suède non incluse), l'étude très documentée de Christine CHAMBAZ intitulée "Les jeunes adultes en Europe" [CHAMBAZ 2000] est extrêmement précieuse comme outil de cadrage pour notre propre analyse, et nous suggérons vivement au lecteur intéressé de s'y reporter²⁶.

S'agissant spécifiquement des jeunes adultes, un des problèmes majeurs est constitué par l'écart sensible entre leur nombre tel qu'il ressort de la base Registre (qui liste les membres de chaque ménage) et le nombre de jeunes adultes présents dans la base Individus, censée inclure tous les individus de 16 ans et plus : en 1997, le taux de perte pour les 16-30 ans est cinq fois supérieur à celui enregistré pour les 30 ans et plus (10,9% versus 2,0%).

Ces pertes varient sensiblement **selon les pays** : de 2% seulement pour l'Autriche à 30% pour la Suède (voir en Annexe le tableau 1, colonne 7). Mais elles sont également très différentes **selon les âges**. Ainsi, à 17 ans, c'est presque 10% des jeunes adultes recensés comme membres de ménages ECHP97 qui manquent à l'appel parmi les questionnaires individuels²⁷, et cette proportion, bien que déclinant progressivement à mesure que l'âge croît, reste longtemps élevée (7,7% à 20 ans, 4% à 25 ans, 3% à 30 ans, à comparer à 2% environ à compter de 34 ans).

²⁶ Nous nous permettons d'attirer l'attention des chercheurs intéressés par le recours à cette base de données récente, très riche et sans équivalent, sur le fait que son usage n'est pas immédiat. En effet, de multiples et souvent longues étapes séparent le moment où on dispose des données EUROSTAT sur CD-ROM et le moment où on peut les exploiter avec un minimum de confiance. Il y a d'abord la nécessité de convertir en Standards de pouvoir d'achat (SPA) les données monétaires (parce qu'elles sont toutes fournies en monnaie nationale). Cette nécessité résulte du fait que choisir l'euro pour convertir toutes les sommes en une même monnaie ne tiendrait pas compte des inégalités, parfois sensibles, du niveau général des prix entre pays, comme le souligne ABRAMOVICI [2002], qui montre que la dépense de protection sociale par habitant au Danemark, en 1999, serait de 88% supérieure à celle de l'Italie si on les exprimait en euros, mais seulement de 35% en termes de SPA. Cette conversion en SPA, apparemment facile, nécessite cependant quelques précautions, à cause d'erreurs dans les notes de la documentation EUROSTAT indiquant, variable par variable, les unités de mesure (on s'aperçoit ainsi que, pour l'Italie, alors que la plupart des montants sont dits exprimés en milliers de lires, ceux qui sont censés ne pas l'être le sont quand même). Puis il faut passer, pour certains revenus (en France mais aussi en Finlande pour certains éléments de revenu) du brut au net, prendre la mesure des conséquences des méthodes (actuellement en cours de changement) de correction pour tenir compte des revenus individuels qui font défaut pour certains ménages, établir un système de pondérations adapté aux comparaisons internationales et l'appliquer à la sous-population qu'on étudie, vérifier que chacune des variables utilisées ne comporte pas uniquement des manquants pour tel pays (ce qui est le cas, en 1997, pour les revenus d'assistance au Royaume-Uni et pour les transferts privés reçus par les ménages en Suède), etc. Tout ceci sans parler des incohérences, sans doute difficilement évitables complètement, au niveau des données individuelles ; ainsi, des vérifications minutieuses nous ont permis de voir, par exemple, que tel ménage anglais serait constitué uniquement de deux enfants de 4 et 7 ans, et que telle Hollandaise aurait deux conjoints, l'un de son âge, l'autre ayant un an, etc. Certaines des incohérences que nous avons ainsi repérées ont d'ailleurs été corrigées par Eurostat à la faveur d'une révision des données (en l'occurrence celle intervenue entre la version Mai 2001 et la version Décembre 2001 des données de la vague 1997).

²⁷ Cette absence de certains questionnaires individuels pour des membres de ménages répertoriés prend une forme extrême quand, pour certains ménages, on ne dispose d'aucun questionnaire individuel. C'est le cas pour 32 ménages (incluant 94 individus) dans la version Mai 2001 de ECHP97, et encore pour 33 ménages (90 individus) dans la version révisée de Décembre 2001 : trois ménages finlandais sont portés disparus mais on a récupéré deux ménages anglais pour lesquels on n'a aucun questionnaire individuel. Cette absence totale de questionnaires individuels pour certains ménages nous fait perdre 26 jeunes adultes : 16 Espagnols, 6 Portugais, 3 Anglais et un Finlandais. On regrettera particulièrement la perte d'un ménage espagnol et d'un ménage portugais, qui comportent chacun quatre jeunes adultes vivant avec leurs deux parents : aucun membre du ménage n'a rempli de questionnaire individuel.

Les 16 ans posent un problème de frontière, sans doute en partie dû à la date d'enquête : si on trouve, parmi les questionnaires individuels (théoriquement remplis uniquement par les 16 ans et plus), un certain nombre de jeunes Autrichiens de 15 ans, on se heurte en revanche à un taux de perte considérable pour bien des pays, allant jusqu'à l'absence complète de jeunes adultes de 16 ans (sans anomalie, pour autant, à 17 ans). En moyenne, ce sont 83% des 16 ans qui sont manquants.

Si nous avons opté au départ pour une **tranche d'âge large** (16-30 ans), en dépit de son évidente hétérogénéité, c'est pour pouvoir ensuite étudier séparément des sous-populations plus homogènes, qu'il s'agisse de groupes d'âge ou de sous-ensembles obtenus par croisement de critères, et ceci en nous calant sur les champs choisis par d'autres auteurs. Ainsi avons-nous repris nos modèles 16-30 ans en les estimant cette fois sur chacun des trois groupes d'âge distingués par CHAMBAZ [2000] (18-21, 22-25, 26-29) puis sur chacun des quatre groupes de HERPIN et VERGER [1997], qui croisent le fait de vivre ou non chez ses parents avec le fait d'être ou non étudiant, le tout pour la tranche d'âge 18-28 ans²⁸.

²⁸ Faute de place nous n'exposons pas ici ces divers modèles.

CHAPITRE IV

LES MODELES ESTIMES ET LEURS RESULTATS

Compte tenu des problèmes de disponibilité des données dans la vague 1997 de l'ECHP, il nous a fallu renoncer à certaines des hypothèses que nous voulions tester.

Du côté du niveau d'études individuel, il a fallu se contenter des trois indicatrices correspondant aux niveaux faible, intermédiaire et supérieur²⁹.

Du côté des transferts sociaux, on a donné la préférence à la conjonction d'un indicateur macro de dépenses de protection sociale qui englobe tous les transferts perçus par les tous ménages, avec deux indicateurs ciblés sur les ménages incluant au moins un jeune adulte : les prestations Chômage (moyennes par pays, calculées sur l'échantillon) et les Prestations Famille. Les autres types de transferts publics posaient en effet problème, soit que des fonctions fort différentes fussent agrégées (ex.: logement et exclusion dans les données SESPROS dont on disposait), soit que la comparabilité internationale fût mal assurée (cas du logement, avec l'exclusion des prestations non monétaires, sous forme de loyers faibles), soit que tel ou tel agrégat par fonction fût manquant pour certains pays (cas de l'assistance pour le Royaume-Uni). Et, du côté des transferts privés reçus, les inclure dans le modèle eût conduit à perdre le cas suédois.

Quant aux autres indicateurs par pays, on a essentiellement retenu des moyennes et des proportions issues de l'échantillon, les indicateurs issus des comptes ou de grandes enquêtes ne ciblant guère ou pas la population qui est l'objet de notre étude.

Sur la base de ces variables, on a estimé un certain nombre de modèles³⁰, et quatre en particulier, que nous présentons maintenant (les statistiques descriptives correspondant aux modèles estimés sont fournies en annexe, dans le tableau 2).

1) Un premier modèle multi-niveaux, appelé A, n'inclut **aucune variable explicative** ; sa fonction est justement de permettre d'apprécier ce qu'on gagnera en introduisant des facteurs explicatifs. Il n'y a donc qu'un seul coefficient aléatoire, constitué par la constante de l'équation de niveau 1. Ce coefficient fait l'objet de l'unique équation de niveau 2. Ce modèle permet, à travers le calcul du coefficient de corrélation intra-groupe, de situer la variabilité des gains respectivement attribuable aux différences entre pays et à l'hétérogénéité au sein de chaque pays. Le tableau 3, en annexe, qui donne les résultats de quatre modèles multi-niveaux, mentionne la valeur 5,4% pour le coefficient de corrélation intra-groupe. Ce chiffre indique que les gains sont certes différents entre pays (leur variabilité est donnée par le coefficient estimé pour le terme d'erreur de l'équation dont la constante aléatoire est la variable expliquée) mais que l'essentiel de la variabilité est interne aux pays, ce qui rend

²⁹ La documentation EUROSTAT (DOC.PAN166/01, mai 2001, p.350] mentionne que ce niveau 3 correspond aux échelons 5 à 7 de l'ISCED, tandis que le niveau 2 (lycée) correspond à l'échelon 3, tandis que le niveau 1, ici omis et servant donc de niveau de référence pour l'interprétation des paramètres estimés, correspond aux échelons 0, 1 et 2. Les jeunes adultes encore scolarisés ont un code "sans objet" pour ces trois variables. Selon EUROSTAT, "cette information n'est collectée qu'à l'entrée dans le panel mais [sic] elle n'est plus actualisée depuis la vague 5 (1998)".

³⁰ Nous avons adopté, suivant en cela les recommandations des manuels traitant des modèles multi-niveaux (HOX, SINGER, SNIJDERS et BOSKER, etc.), une démarche de **modélisation ascendante**, consistant à partir d'un modèle extrêmement simple, que l'on complexifie ensuite progressivement. Nous avons procédé de façon systématique, de manière à tenter d'identifier les facteurs explicatifs appropriés à chacun des deux niveaux : ainsi chacun de nos quatre X a-t-il été introduit séparément, en sus de la constante, et ceci d'abord sans chercher à expliquer son coefficient aléatoire puis en introduisant les Z un à un pour chaque X. La création d'une macro SAS a facilité l'extraction des informations essentielles, au sein des copieux résultats de chacun des trente modèles concernés (tables de l'ODS, Output Delivery System), et leur comparaison par inclusion dans une base unique. Ces résultats préliminaires ne sont pas reproduits ici faute de place.

intéressante une tentative d'introduction de facteurs explicatifs individuels, ayant des coefficients aléatoires mais expliqués eux-mêmes par des facteurs macro ou de contexte ou non expliqués par de tels facteurs.

2) C'est justement ce que fait le **deuxième modèle multi-niveaux, appelé B**. Il **ajoute quatre variables individuelles**³¹ : l'âge, le sexe et deux niveaux d'éducation (intermédiaire et supérieur) mais les coefficients de ces X, quoique aléatoires et donnant donc lieu à cinq équations au niveau 2 (une par X, plus une pour la constante aléatoire), ne sont **pas expliqués par des variables Z**, de niveau 2, variables qui ne seront introduites que dans les deux modèles suivants (C et D).

L'estimation de ce modèle requiert une **hypothèse quant à la structure de la matrice des variances-covariances des termes d'erreur** car, alors que le modèle A ne faisait intervenir que le calcul de deux variances (celles des termes d'erreur de l'équation de niveau 1 et celle de l'unique équation de niveau 2), on se trouve ici en présence de cinq termes d'erreur pour les équations de niveau 2, pouvant former 10 paires distinctes, ce qui fait qu'il y a donc 10 covariances à estimer, en plus des 6 variances (constante + 4 X + résidus). Le plus simple est de postuler que les cinq termes d'erreurs ne sont pas liés les uns aux autres, ce qui revient à supposer que les dix covariances sont nulles ; on obtient alors un modèle dit de composantes de la variance (*'Variance Components'*), qui n'exige d'estimer que les six variances.

Mais cette hypothèse de nullité simultanée de toutes les covariances est bien hardie, et c'est pourquoi il vaudrait mieux commencer par ne pas faire d'hypothèses particulières sur les valeurs des variances-covariances (option TYPE=UN de PROC MIXED de SAS, pour *'Unstructured'*), examiner les résultats obtenus et ne passer au modèle simplifié de type Composantes de la variance que si les probabilités associées à l'estimation de chacune des covariances excèdent 10% (même si ces tests Z ne sont pas totalement fiables hors du cas d'échantillons de très grande taille). Ou alors, si le nombre de paramètres à estimer est jugé trop important, on peut le réduire en faisant par exemple l'hypothèse de l'égalité de toutes les covariances (ce qui réduit à un le nombre de paramètres à estimer en plus des variances mais ce qui nous paraît complètement irréaliste en cas de variables à métriques diverses, comme c'est le cas ici) ; conjuguée à l'hypothèse d'égalité de toutes les variances (on en est loin, ici), l'unicité des covariances donne cette structure très particulière appelée *'Compound Symmetry'*. On lui préférera, dans notre cas, sa version modifiée pour tenir compte de l'hétérogénéité : la structure dénommée *'Heterogeneous Compound Symmetry'* permet d'avoir des variances différentes les unes des autres et des covariances également différentes les unes des autres, mais ces dernières sont toutes censées être égales au produit des racines des variances correspondantes et d'un coefficient unique, appelé ρ (cf. [SAS 1999], p.2138-2140) ; aux variances toutes différentes ne s'ajoute alors qu'un seul paramètre supplémentaire à estimer, ρ .

³¹ On aurait également pu n'introduire que des variables de niveau 2, dont la fonction serait alors d'expliquer au mieux la variabilité de la constante aléatoire de l'équation de niveau 1, c'est-à-dire la diversité des gains moyens par pays.

Nous avons mis en œuvre ces différents hypothèses permettant de ne pas s'en tenir à des covariances supposées nulles. Malheureusement, à part dans des cas très simples (ex.: uniquement le facteur âge et la constante), jamais nous ne sommes parvenu³² à obtenir des estimations des covariances avec des matrices de type 'Unstructured' ou de type 'Heterogeneous Compound Symmetry'.

Tous les modèles que nous commentons ici reposent donc, faute de mieux, sur **l'hypothèse de covariances nulles**. Mais il semble bien que cette hypothèse ne soit pas trop restrictive car une seule des covariances (celle liant les erreurs associées à la constante aléatoire et au coefficient aléatoire de la variable de sexe) s'avère significativement non nulle. Du moins est-ce la conclusion que nous tirons de la mise en œuvre des modèles présentés ci-dessus au moyen d'**autres logiciels que SAS**, à savoir les deux principaux logiciels spécialisés en matière de modèles multi-niveaux : **HLM5.02** [RAUDENBUSH et alii, 2000], et **MlwiN** version 1.10.0006 [RASBASH et alii, 2000]. Nous faisons figurer en annexe B les résultats que nous avons obtenus en recourant à ces deux logiciels (qui utilisent des paramétrisations différentes entre elles et différentes de celle de SAS, et qui, surtout, n'emploient pas les mêmes méthodes et/ou algorithmes d'estimation, ce qui ne permet pas de retrouver exactement les mêmes résultats). Ces deux logiciels offrant d'intéressantes possibilités complémentaires, nous en avons profité pour examiner, grâce à eux (Annexe C), diverses questions moins aisément³³ traitables avec SAS, comme celle de l'impact, sur les résultats, de certaines **observations influentes** (en l'occurrence, le cas de deux jeunes adultes néerlandais) ou comme celle du test de **l'hypothèse de normalité des** (différents types de) **résidus** ou de **l'hypothèse de l'homogénéité, entre pays, de la variance au niveau 1**.

Ceci dit, quels sont les **résultats du modèle B** ? On constate d'abord une notable réduction de la composante résiduelle (qui passe de 45 à 32 millions), qui traduit le fait qu'en introduisant des X on laisse inexplicée une moins grande portion de la variabilité interne aux pays.

Toutes les composantes de la matrice de variances-covariances sont clairement significativement différentes de zéro au seuil de 1 à 2%. Mais peut-on interpréter leurs tailles relatives ? Si c'était admissible, on concluerait que la forte composante relative à l'indicatrice de niveau supérieur d'éducation (4 millions) semblerait montrer que la diversité des gains au sein de chaque pays s'explique pour une part non négligeable par les différences entre les gains des jeunes adultes n'ayant qu'un niveau faible d'éducation (c'est ici la situation implicite de référence) et ceux des jeunes adultes qui ont fait des études supérieures. Et ce serait naturellement moins vrai pour les jeunes adultes n'ayant qu'un niveau intermédiaire, car ils sont plus proches de ceux qui ont un niveau faible.

³² Et ceci en dépit d'essais multiples incluant bien sûr les options de critère, absolu ou relatif, de convergence (et de valeur de ce critère), mais aussi en changeant, via l'instruction PARMS, les limites imposées aux valeurs estimées des paramètres de la matrice des variances-covariances. Et en dépit de nombreuses tentatives pour recalibrer, par des changements de variables appropriées (en divisant les gains par 1000 et en divisant aussi – ce qui est moins intuitif – chaque X par un scalaire approprié), les ordres de grandeur des paramètres estimés pour les variances (car SAS attire l'attention, page 2172, sur les problèmes rencontrés par l'algorithme de Newton-Raphson en cas de valeurs des paramètres ayant des ordres de grandeur trop différents : dans ce cas, l'estimation jointe de deux paramètres bute, pour un de ceux-ci, sur les limites internes de tolérance posées par PROC MIXED). Puisque, après des essais concluants quant à l'obtention d'ordres de grandeurs similaires pour les variances (partant de valeurs dans un rapport de 1 à 1000, cf. tableau 3, modèle B, 31228 pour l'âge versus 31907793 pour les résidus, on est parvenu à six variances se situant toutes entre 13 et 52), on se heurte toujours au message "Out of memory" (alors que plusieurs Go sont disponibles), c'est que le problème vient d'une autre cause ou que, en rendant les variances comparables, on a modifié aussi les covariances mais sans les rapprocher assez des variances... Toute suggestion serait très bienvenue ...

³³ En effet, SAS ne semble pas offrir, dans PROC MIXED, d'options comparables à celles qu'offre PROC REG en matière de statistiques d'influence.

Quant aux différences selon le sexe, elles n'apparaîtraient³⁴ pas comme jouant un rôle considérable une fois que l'on contrôle les différences d'âge et d'éducation.

Quant aux effets fixes, tous significatifs, ils s'interprètent ici comme la partie fixe du coefficient de chaque X, autour de laquelle fluctuent de façon aléatoire les coefficients de chaque pays : une année d'âge apporterait, en moyenne pour tous les âges et tous les pays, 677 SPA par an, tandis que le fait d'être une femme diminuerait les gains des jeunes adultes, en moyenne pour tous les pays, de 2204 SPA par an une fois que l'on contrôle les différences d'âge et d'éducation ; et toujours en moyenne pour tous les pays, on gagnerait 975 SPA de plus par an en passant d'une éducation faible à une éducation de niveau intermédiaire, et 3190 SPA/an pour ceux ayant une éducation supérieure.

3) Quant au **troisième modèle multi-niveaux, appelé C, il ajoute** aux quatre variables individuelles du modèle B (l'âge, le sexe et deux niveaux d'éducation) **des variables de niveau 2**, pour tenter d'expliquer la variabilité des coefficients aléatoires de ces X ; ces variables sont à la fois **nos variables d'intérêt (les transferts sociaux) et variables de contrôle visant à essayer de tenir compte de la grande hétérogénéité** de la population des jeunes adultes. On retrouve donc, comme avec le modèle B, cinq équations au niveau 2, et donc autant de paramètres de variance-covariance que précédemment. Mais il faut estimer plus de coefficients d'effets fixes puisque à chaque variable Z introduite dans une équation de niveau 2 correspond un paramètre supplémentaire, s'ajoutant aux cinq constantes estimées ici comme dans le cadre du modèle B.

³⁴ On pourrait songer à **comparer entre elles les composantes de la variance** associées aux coefficients des trois variables indicatrices (sexe, éducation supérieure et éducation intermédiaire) parce qu'il s'agit de trois variables sans unité. En revanche, la valeur relative, par rapport aux autres composantes, de celle qui a trait à l'âge n'a pas de sens, puisqu'elle dépend des unités employées pour mesurer l'âge (c'est ce qu'illustrent nos essais pratiqués avec l'âge exprimé en mois : la composante de la covariance fond alors littéralement, divisée qu'elle est par $12^2=144$). Il est bien dommage qu'on ne puisse donc pas situer, sur la base de la simple comparaison des composantes de la variance, l'importance du rôle de l'âge, d'autant qu'on le sait central (à quoi servent alors les valeurs de la variable RATIO fournies par PROC MIXED, en dehors de celle relative à la constante, qui rapportent chaque composante à la variance résiduelle ?).

Pour effectuer cette comparaison de l'importance respective, pour expliquer la variance globale, de chacun des facteurs de niveau 1, on pourrait être tenté de standardiser ces facteurs pour les rendre comparables. On obtient les résultats suivants : les valeurs des tests Z et des p changent, mais peu ; le coefficient de corrélation intraclasse passe de 26,8% à 5,75% ($=0,04457/(0,04457+0,7302)$). Les variances s'effondrent comme prévu, mais, en outre, les rapports entre composantes sont bousculés (ce qui se comprend parce qu'on divise chaque X par une grandeur différente, son écart-type) : ainsi la composante educ_sup, qui valait 3,73 fois la composante educ_mid, ne la vaut plus que 1,76 fois. Et la composante résiduelle, 1022 fois plus grande que la composante AGE, ne l'est plus que 55 fois, ce qui rapproche certes les ordres de grandeur.

En fait, **standardiser les facteurs ne permet pas de juger de leur importance relative**, comme le soulignent WILLETT et al. [1998], et comme l'a confirmé un débat entre statisticiens praticiens, en avril 2002, au sein de la liste de discussion MULTILEVEL ; en effet, les coefficients de régression standardisés dépendent de la variance de Y dans l'échantillon, ce qui les rend inutiles pour la comparaison des effets à partir soit de deux échantillons issus de la même population (car Y peut être dispersé différemment dans les deux échantillons), soit, a fortiori, à partir d'échantillons issus de populations différentes.

Les résultats de ce modèle, traduits sous forme structurelle, sont les suivants :

Niveau 1 : une équation unique :

$$G_{ij} = \beta_{0j} + \beta_{1j} \text{AGE}_{ij} + \beta_{2j} \text{SEXE}_{ij} + \beta_{3j} \text{EDUCMID}_{ij} + \beta_{4j} \text{EDUCSUP}_{ij} + r_{ij}$$

où $r_{ij} \approx N(0 ; 31907606)$

Niveau 2, première équation, expliquant la variabilité de la constante aléatoire (entre parenthèses, les écarts-types estimés) :

$$\beta_{0j} = -2714 - 0,967 \text{TRF_SOCX}_j + 0,125 \text{ALLOC_CHOM}_j - 1,120 \text{ALLOC_FAM}_j$$

(5863,1) (0,886) (1,905) (1,573)

$$- 7694,6 \text{TAUX_CHOM}_j - 3455,7 \text{FREQ_ETUDIANTS}_j + u_{0j}$$

(33468) (19831)

où $u_{0j} \approx N(0 , 13548961)$

Autrement dit, quand on contrôle l'âge, le sexe et l'éducation, les gains sont nuls en moyenne pour les différents pays puisque aucune variable n'est significative pour expliquer une constante aléatoire qui différerait de zéro. C'est dire que les transferts sociaux qui sont notre variable d'intérêt – et qui ne figurent que dans cette équation – n'ont pas d'action décelable (du moins par cette méthode et avec les données et indicateurs utilisés), et ceci ni quand on les considère tous, ni quand on cherche à isoler l'effet spécifique des transferts Famille et des transferts Chômage.

Niveau 2, deuxième équation (expliquant le coefficient aléatoire de l'âge) :

$$\beta_{1j} = 716,1 + 626,3 \text{PROPORT_ETUDIANTS}_j - 2062,7 \text{TX_CHOM}_j + u_{1j}$$

(206,6) (857,6) (1399,6)

où $u_{1j} \approx N(0 , 32464)$

Ceci montre que l'âge a un effet significatif mais que cet effet ne diffère, selon les pays, ni en fonction de la proportion d'étudiants, ni en fonction (mais c'est moins net) du taux de chômage.

Niveau 2, troisième équation (expliquant le coefficient aléatoire du sexe) :

$$\beta_{2j} = -3463,7 + 11665 \text{ TX_CHOM}_j + u_{2j}$$

(388,0) (3269,3)

$$\text{où } u_{2j} \approx N(0, 219368)$$

Ceci montre que le fait d'être une jeune adulte a un effet sur les gains qui dépend du niveau du taux de chômage des jeunes adultes : si ce dernier était égal à 10%, le coefficient β_{2j} vaudrait 2297, signifiant que, ceteris paribus, les jeunes adultes de sexe féminin gagneraient 2297 SPA de moins par an que leurs homologues masculins ; mais avec un taux de chômage de 30% parmi les jeunes adultes, cet écart s'annulerait pratiquement (-35 SPA) et, au-delà, se transformerait en avantage pour les jeunes femmes.

Niveau 2, quatrième équation (expliquant le coefficient aléatoire de l'éducation intermédiaire) :

$$\beta_{3j} = 964,2 + u_{3j} \quad \text{où } u_{3j} \approx N(0, 1065098)$$

(296,3)

L'effet, sur les gains des jeunes adultes, du fait d'avoir une éducation de niveau intermédiaire est clair. Mais, puisque nos tentatives d'expliquer cet effet par diverses caractéristiques nationales n'ont débouché sur aucun coefficient significatif, nous devons conclure que cet effet ne dépend pas des caractéristiques des pays que nous avons retenues ; ce qui irait plutôt dans le sens de l'idée d'une certaine homogénéité, sinon des critères de rémunération dans les pays européens, du moins du rôle, sur les gains, du fait d'avoir une éducation de niveau intermédiaire au lieu d'une éducation de niveau faible.

Niveau 2, cinquième équation (expliquant le coefficient aléatoire de l'éducation supérieure) :

$$\beta_{4j} = 3175,4 + u_{4j} \quad \text{où } u_{4j} \approx N(0, 3972936)$$

(568,9)

Cette dernière équation montre que l'effet, sur les gains des jeunes adultes, du fait d'avoir une éducation de niveau supérieur, et non pas faible, est très net. Mais il ne dépend pas, ici, de caractéristiques des pays, ce qui, comme pour l'éducation de niveau intermédiaire, irait plutôt dans le sens de l'idée d'une certaine homogénéité quant au rôle, dans les divers pays européens, du rôle joué, sur les gains, par le fait d'avoir une éducation de niveau supérieur au lieu d'une éducation de niveau faible.

Quant à la qualité de l'ajustement global du modèle aux données, il est difficile ici de se prononcer (en comparant les valeurs du critère d'information de AKAIKE ou en opérant un test de différence des valeurs de $-2RLL$ (*Restricted/residual Log-Likelihood*) entre deux modèles car ce test, qui n'est d'ailleurs qu'approximatif ([SINGER 1998, p.339], n'est utilisable que pour des modèles ayant en commun leur partie fixe, c'est-à-dire les Z et les Z croisés avec les X dont ils influencent les coefficients, i.e. les termes de l'instruction MODEL de SAS/PROC MIXED) ; or, ici, aucun de nos modèles A, B, C et D ne partage avec un autre sa partie fixe (constituée par la liste des variables figurant dans la partie inférieure du tableau 3).

4) Enfin, le **quatrième modèle multi-niveaux, appelé D, conserve les quatre variables individuelles** des modèles B et C (l'âge, le sexe et deux niveaux d'éducation) et **limite les variables de niveau 2 à celles qui se sont révélées significatives dans le modèle complet (C)**. Ceci élimine nos variables d'intérêt (les transferts sociaux) et la plupart des variables de contrôle visant à essayer de tenir compte de la grande hétérogénéité de la population des jeunes adultes : seule garde un rôle explicatif la variable de niveau 2 qu'est le taux de chômage des jeunes adultes, qui a un impact sur le rôle du sexe sur les gains d'activité. Comme avec les modèles B et C, on a donc cinq équations au niveau 2, et donc autant de paramètres de variance-covariance que précédemment. Mais il faut estimer moins de coefficients d'effets fixes que précédemment.

Les résultats de ce modèle simplifié, traduits sous forme structurelle, sont les suivants :

Niveau 1 : une équation unique :

$$G_{ij} = \beta_{0j} + \beta_{1j} AGE_{ij} + \beta_{2j} SEXE_{ij} + \beta_{3j} EDUCMID_{ij} + \beta_{4j} EDUCSUP_{ij} + r_{ij}$$

$$\text{où } r_{ij} \approx N(0 ; 31907872)$$

Niveau 2, première équation, expliquant la variabilité de la constante aléatoire (entre parenthèses, les écarts-types estimés) :

$$\beta_{0j} = -10481 \quad + u_{0j}$$

(966,5)

$$\text{où } u_{0j} \approx N(0 , 11945750)$$

Si ce coefficient est fortement négatif et significatif, c'est qu'il correspond aux gains à...0 ans, pour un homme n'ayant qu'une éducation de faible niveau. A mesure que l'âge croît, les gains sont censés croître (cf. l'équation suivante), et pour un jeune adulte de 16 ans (et donc n'ayant pas encore d'éducation intermédiaire ou supérieure), l'effet d'âge donne $16 * 676,7 = 10827$, ce qui, ajouté à -10481 , donne $346,2$ SPA/an en moyenne.

Niveau 2, deuxième équation (expliquant le coefficient aléatoire de l'âge) :

$$\beta_{1j} = \begin{matrix} 676,7 \\ (48,9) \end{matrix} + u_{1j}$$

$$\text{où } u_{1j} \approx N(0, 31238)$$

Ceci montre que l'âge a un effet significatif mais que cet effet ne diffère pas selon les caractéristiques des pays (du moins celles que nous avons tenté, mais en vain, de prendre en compte ici).

Niveau 2, troisième équation (expliquant le coefficient aléatoire du sexe) :

$$\beta_{2j} = \begin{matrix} -3463,8 \\ (391,1) \end{matrix} + \begin{matrix} 11666 \\ (3295,9) \end{matrix} \text{TX_CHOM}_j + u_{2j}$$

$$\text{où } u_{2j} \approx N(0, 223877)$$

Les valeurs des coefficients estimés pour cette troisième équation sont quasiment identiques à celles du modèle plus complet C. Elle montre que le fait d'être une jeune adulte a un effet sur les gains qui dépend du niveau du taux de chômage des jeunes adultes.

Niveau 2, quatrième équation (expliquant le coefficient aléatoire de l'éducation intermédiaire) :

$$\beta_{3j} = \begin{matrix} 976,0 \\ (297,0) \end{matrix} + u_{3j} \quad \text{où } u_{3j} \approx N(0, 1062511)$$

Les valeurs des coefficients estimés pour cette quatrième équation sont, elles aussi, quasiment identiques à celle du modèle plus complet C. Elle montre que l'effet, sur les gains des jeunes adultes, du fait d'avoir une éducation de niveau intermédiaire est clair, par rapport au fait d'avoir une éducation de niveau faible.

Niveau 2, cinquième équation (expliquant le coefficient aléatoire de l'éducation supérieure) :

$$\beta_{4j} = \begin{matrix} 3191,1 \\ (566,9) \end{matrix} + u_{4j} \quad \text{où } u_{4j} \approx N(0, 4004142)$$

Le commentaire est ici le même que précédemment.

CHAPITRE V

**MODELES HIERARCHIQUES MULTI-NIVEAUX
VERSUS REGRESSIONS MULTIPLES CLASSIQUES**

Les résultats ainsi obtenus sont **mis en regard** de ce qu'on aurait eu si on avait appliqué la technique de **la régression multiple classique** ; ceci permet de prendre la mesure des différences de résultats, particulièrement en termes de significativité des effets.

La partie gauche du tableau 4 de l'annexe met côte à côte **le modèle multi-niveaux C** (correspondant aux variables susceptibles d'avoir un impact sur les gains) et son homologue C', qui reprend les mêmes variables. On relève que la multicolinéarité affecte un nombre non négligeable de variables, avec des VIF (*Variance inflation factor*) de 13,8 pour l'âge, de 6,6 pour le sexe et de 7,5 pour la croisée du sexe avec le taux de chômage, mais surtout de 80 à plus de 100 pour deux variables de contexte (le taux de chômage et la proportion d'étudiants) et pour les deux variables croisées qu'elles forment avec l'âge ; sans doute ces phénomènes de multicolinéarité sont-ils en partie dus à la construction même des variables croisées.

En revanche, aucune des trois variables de transferts sociaux (globaux, chômage, famille) ne présente de VIF excédant 3,5 et leurs seuils de significativité sont inférieurs à 1%. On aurait alors tendance à conclure à un effet très net, sur les gains, de la générosité des transferts sociaux de type Famille et Chômage.

Or, ce que montrent les résultats du modèle multi-niveaux C, c'est qu'en dehors de la variable croisant le sexe avec le taux de chômage, aucune variable macro ou de contexte n'est significative, alors que chacune des cinq variables de niveau individuel l'est.

Le tableau 4 de l'annexe met ensuite en parallèle, dans sa partie droite, **le modèle multi-niveaux D** (correspondant aux seules variables à effets significativement différents de zéro) et son homologue D', qui reprend les mêmes variables, en y ajoutant cependant le taux de chômage afin que la variable croisant le sexe avec le taux de chômage soit accompagnée de ses deux variables composantes.

Cette fois, dans le modèle de régression linéaire (D'), la multicolinéarité n'affecte plus que la variable de sexe et sa croisée avec le taux de chômage (VIF de 6,6 et 7,5 respectivement). Par rapport au modèle précédent, la perte de R² est très faible (passage de 31,35% à 30,34%).

Comme dans le modèle multi-niveaux D, tous les coefficients sont significatifs. La comparaison des valeurs estimées pour les coefficients des variables explicatives, qui ne peut être stricte du fait de la présence d'une variable supplémentaire dans le modèle D', n'exhibe pas d'écart gigantesques. Si on fait tourner le même modèle sans le taux de chômage (modèle D'', ne figurant pas dans le tableau 4 faute de place et dont le R² vaut 27,79%), les deux listes de variables sont alors identiques ; la comparaison des coefficients estimés montre alors à la fois des similitudes importantes (pour la constante, l'âge et les deux variables indicatrices du niveau d'éducation) et deux différences troublantes, qui impliquent justement la variable de sexe et sa croisée avec le taux de chômage :

constante : -11347 pour le modèle D'' contre -10481 pour D (multi-niveaux)
âge : 707,6 pour le modèle D'' contre 676,7 pour D
femme : 41,9 pour le modèle D'' (mais valeur hautement non significative : p=73%)
contre -3463,8 pour D (p < 0,01%)
éduc. intermed. : 839,6 pour le modèle D'' contre 976,0 pour D
éduc. sup. : 3285,3 pour le modèle D'' contre 3191,1 pour D
croisée femme*chômage : -18605 pour le modèle D'' contre 11666 pour D.

Ici, l'opposition du signe est inquiétante, s'agissant de valeurs très significatives puisque associées à des p très faibles (moins de 0,1% dans les deux cas) ; et ceci d'autant plus que le coefficient standardisé de cette variable est du même ordre de grandeur que celui de la variable d'éducation supérieure (-0,19 vs 0,16) et vient, en termes de valeur absolue, juste derrière le coefficient standardisé de l'âge (égal à 0,42) ; et d'autant plus, encore, que le modèle (D') incluant la variable de chômage présente, pour le coefficient de croisée femme*chômage, un coefficient (significatif à $p > 0,01\%$) valant 12130, c'est-à-dire assez voisin du coefficient du modèle multi-niveaux (11666) et de même signe.

CONCLUSION

Comme on a pu le remarquer tout au long de ce texte, qui ne passe pas sous silence les incertitudes de la technique employée mais souligne au contraire un certain nombre de ses limites et multiplie les interrogations, nous avons encore beaucoup à apprendre³⁵ pour découvrir toutes les possibilités des modèles multi-niveaux et pour en maîtriser les finesses d'interprétation.

Par ailleurs il est clair que, à ce stade, tout à fait exploratoire, on ne saurait prétendre à aucune conclusion solide sur le fond, et on s'en gardera donc.

En revanche, les **pistes** sont nombreuses pour nos travaux ultérieurs, de la recherche de **meilleurs indicateurs de niveau pays** (susceptibles d'expliquer une partie de la variabilité considérable qui demeure au niveau individuel et entre pays) au **passage au longitudinal**, avec des modèles de type "courbes de croissance" (*'growth curve'*) traités par les modèles multi-niveaux (instruction REPEATED de SAS, module HMLM de HLM5, section "Repeated Measures" de MlwiN) puisque les observations pour chaque vague successive et pour un même individu sont "nichées" dans l'unité représentée par l'individu concerné, ce qui ajoute un niveau hiérarchique à l'analyse ; cela permettrait de faire le lien avec les techniques de l'économétrie des données de panel et de voir les avantages et inconvénients respectifs de ces techniques et des modèles multi-niveaux³⁶.

Notre intention serait aussi de tenter une **analyse substituant aux pays les régions d'Europe** (au niveau NUTS2 de la nomenclature), si nous arrivons à trouver suffisamment d'indicateurs de caractéristiques régionales. Et, en ce qui concerne la technique des modèles multi-niveaux, nous voudrions utiliser notamment les possibilités de **représentations graphiques** qu'offrent les logiciels spécialisés que sont MlwiN et HLM, qui devraient, à notre sens, permettre de mieux faire le parallèle avec les résultats d'analyses de types "ordonnées à l'origine et pentes en tant que résultats" (*'slopes and intercepts as outcomes'*), analyses plus proches de ce que donneraient des modèles de régression classique faisant usage de jeux d'indicatrices par pays, en tant que telles et croisées avec divers autres facteurs explicatifs³⁷.

³⁵ Avant de mettre en œuvre, ici, des modèles multi-niveaux sur données à la fois réelles et concernant notre propre problématique, nous avons fait nos premières armes sur des données fictives construites par nos soins, puis nous avons répliqué, avec des variantes, les modèles présentés dans l'article de J.SINGER (sur les données HSB, et sur le petit panel de WILLETT, dont J.SINGER nous a aimablement communiqué les données) et dans le livre de KREFT et DE LEEUW (sur les données NELS88) ; ce qui est instructif et ménage parfois des surprises.

³⁶ On se bornera ici à relever l'intérêt, maintes fois signalé dans la littérature spécialisée, représenté par les modèles multi-niveaux dès lors que les **données de panel ne sont pas parfaitement balancées**, ce qui est notamment le cas lorsque certains individus ne sont pas présents à certaines vagues. Au moins dans le cas où on peut faire l'hypothèse que les **valeurs manquantes** le sont de façon aléatoire, les modèles multi-niveaux offrent des solutions permettant d'obtenir des estimateurs efficaces, utilisant toute l'information disponible au lieu d'exclure les individus absents lors d'une des vagues au moins (ce que font la plupart des autres méthodes statistiques) [RASBASH et al. p.129].

³⁷ Faute de temps et de place, nous n'avons pas présenté ici les résultats des régressions traditionnelles que nous avons menées à titre exploratoire, avant de nous lancer dans les modèles multi-niveaux, en utilisant des *'dummies'* de pays (analyse de covariance).

Enfin, parce que modéliser directement les gains est, à l'évidence, inapproprié s'agissant d'une variable marquée à la fois par de nombreuses valeurs nulles (surtout pour les jeunes adultes) et par une distribution non gaussienne, nous devons explorer en détail les possibilités offertes par les logiciels existants (non seulement SAS, HLM5 et MlwiN mais peut-être aussi GLLAMM de STATA, aML, Splus, MIXOR/MIXREG, etc.) pour transposer aux modèles multiniveaux la procédure classique en deux étapes du modèle de HECKMAN : estimation de la probabilité d'avoir des gains non nuls, puis estimation du logarithme³⁸ du montant des gains lorsqu'ils ne sont pas nuls, avec, parmi les explicatives, l'inverse du ratio de MILL issu du Probit. Pour ce que nous en savons actuellement (sans en avoir la pratique pour l'instant), chaque logiciel offre la possibilité d'estimer des modèles de type Probit, mais des zones d'ombre demeurent (concernant, par exemple, la possibilité de pondérer les observations dans PROC NLMIXED de SAS, alors que la macro SAS GLIMMIX le permet).

Telles sont quelques unes des pistes sur lesquelles nous envisageons de nous engager maintenant. Nous serions très reconnaissant à quiconque pourrait attirer notre attention sur des erreurs que nous aurions commises dans la mise en œuvre de nos modèles multi-niveaux et dans l'interprétation de leurs résultats, et/ou nous suggérer des pistes d'améliorations et/ou répondre aux questions soulevées ici et là dans ce texte.

Avant de conclure, et puisque un de nos objectifs est de permettre à l'utilisateur de modèles de covariance de se faire une idée de l'intérêt du passage à des modèles multi-niveaux, nous voudrions attirer l'attention sur les avantages et inconvénients respectifs³⁹ des modèles multi-niveaux et des **modèles de covariance (à coefficients fixes)**. Ces derniers représentent en effet, avec l'introduction (qu'ils opèrent) de dummies par pays et de variables croisant ces dummies avec les principales variables explicatives du modèle (dont on soupçonne qu'elles ont, sur Y, un impact différent selon les pays), une solution attractive, quoique vite lourde. Leur usage se justifie surtout lorsque le nombre de modalités de la variable qualitative (ici, le pays) est restreint (c'est le cas du sexe, par exemple : deux modalités) et lorsqu'on doit considérer que les groupes ainsi formés sont uniques dans leur genre (ainsi, il est difficile de concevoir que les pays européens constituent un échantillon aléatoire issu de la population de tous les pays du monde ; a fortiori, il est impossible de considérer les deux sexes comme deux modalités parmi d'autres, auxquelles on entendrait extrapoler les résultats obtenus sur des échantillons seulement composés d'hommes et de femmes). Mais, dans le cas qui nous intéresse, ne peut-on considérer que, en dépit des spécificités nationales, les quatorze pays sous revue constituent un échantillon tiré d'une population plus large, constituée par exemple de tous les pays européens de l'Ouest (on ajouterait alors le Luxembourg, la Suisse, l'Islande et la Norvège) et d'au moins certains pays de l'ancien bloc soviétique, à commencer par les candidats les mieux placés pour l'adhésion à l'Union européenne ? Dans ce cas, recourir à des coefficients aléatoires (et donc à des modèles multi-niveaux) peut se justifier, surtout si on s'intéresse (ce qui est notre cas) à l'impact, sur l'ampleur des effets des variables individuelles sur les gains, de facteurs nationaux.

³⁸ Choisir le **logarithme** des gains comme variable dépendante est certes contraignant pour l'estimation (puisque cela oblige à écarter tous les gains nuls ou à adopter une procédure en deux étapes qui modélise la probabilité d'apparition de ceux-ci, avant d'expliquer les seuls gains non nuls) mais cela évite d'obtenir, en prévision, des gains négatifs –comme c'est le cas ici pour certains d'entre eux.

³⁹ Sur ce point, voir, entre autres, SNIJDERS et BOSKER p.43 et suivantes.

Il est un dernier argument en faveur des modèles multi-niveaux, qui importe peu pour le cas d'espèce mais peut être crucial pour des analyses dont certains groupes de niveau supérieur ne comportent que peu d'unités de niveau 1 : à la différence des modèles de covariance, qui juxtaposent des estimations groupe par groupe (qui restent cloisonnées en quelque sorte), les modèles multi-niveaux permettent de s'appuyer, pour améliorer l'estimation des coefficients propres aux groupes à petits effectifs, sur les coefficients estimés à partir des autres groupes (c'est ce qu'on appelle le processus de rétrécissement ou de contraction – *'shrinkage to the mean'* – qui voit l'estimateur des moindres carrés pour tel groupe être d'autant plus poussé vers le coefficient d'ensemble que la taille du groupe est faible et fragilise donc l'estimation⁴⁰) ; du moins ce procédé n'est-il valable que si les hypothèses faites sur les distributions des coefficients aléatoires sont tenables : les distributions des termes d'erreur de chaque équation de niveau 2 doivent être indépendantes (ce qui suppose des covariances nulles) et identiques (i.e. suivant la même loi statistique, pas nécessairement Normale⁴¹, nécessairement centrées et de mêmes variances⁴²). Ce qui, dans notre cas, voudrait dire que, une fois que l'on prend en compte les Z, et donc les différences d'effets des X (selon les pays) qui peuvent tenir au niveau différent des Z selon les pays, la part inexpliquée des effets est générées par des mécanismes voisins d'un pays à l'autre, ce qui rend ces derniers comparables – *exchangeable groups, exchangeable effects*. Dans ce cas, l'avantage des modèles multi-niveaux sur les modèles de covariance est net car ces derniers, en donnant à l'effet (de chaque variable) propre à chaque pays le statut de paramètre à estimer, multiplie le nombre de paramètres ce qui, pour les groupes de petite taille, conduit à trop de paramètres (*overfitting*), et donc à de forts écarts types pour ceux-ci ; ce qui n'est pas le cas, justement, avec les modèles multi-niveaux, où les termes d'erreur ne sont pas des paramètres à estimer ; toutefois, si tous les groupes sont, comme c'est le cas ici, de grande taille, cet avantage ne joue presque plus.

Si le présent texte est, dans sa version actuelle, d'un quelconque apport, c'est sans doute sur le terrain méthodologique et pédagogique qu'il faut surtout chercher ce dernier : d'une part, ce texte a essayé de convaincre que, au moins dans le type de configurations « peu de groupes, beaucoup d'individus par groupe », les modèles hiérarchiques multi-niveaux méritent qu'on les utilise ; d'autre part, ce texte s'est efforcé d'illustrer de façon concrète et intuitive quelques uns des principaux concepts et instruments de la modélisation multi-niveaux.

⁴⁰ La statistique bayésienne distingue en effet la connaissance qu'on a sur les effets de groupes à partir uniquement de l'information concernant l'appartenance du groupe à telle population – *prior knowledge* – et la connaissance qu'on en a quand on se fonde, en plus, sur l'information que recèlent les observations propres à ce groupe – *posterior knowledge*. L'estimateur empirique de Bayes – appelé *posterior mean* –, qui est utilisé par les modèles multi-niveaux pour mesurer les effets aléatoires, relève de cette logique.

⁴¹ Selon SNIJDERS et BOSKER (page 44), si les coefficients aléatoires, ainsi que les résidus de premier et de deuxième niveau, ne sont pas approximativement **gaussiens**, les résultats obtenus peuvent être peu fiables ; c'est ce qui peut se produire, expliquent ces auteurs, si, par exemple, il y a plus de groupes extrêmes (sous l'angle des valeurs des u_{kj}) qu'on ne devrait en avoir si le terme d'erreur de niveau 2 représentant l'effet de groupe sur tel coefficient était distribué normalement et avec la même variance que les autres.

⁴² Cette exigence de similitude des variances semble s'appliquer aux u_{kj} – autant de variables aléatoires d'indice j que de pays – relatifs à un coefficient aléatoire d'indice k donné (et donc à la même variable X_k), mais pas entre coefficients, leurs métriques pouvant différer.

Pour aller plus loin et de façon plus rigoureuse, le lecteur intéressé par ce type de modèle aura l'embaras du choix tant sont nombreuses les introductions à cette technique statistique : au sein de la bibliographie ci-après, nous recommandons cependant tout particulièrement le manuel de SNIJDERS et BOSKER [1999], remarquablement clair, et celui de KREFT et DE LEEUW [1998], même s'il est désormais dépassé pour ce qui est de l'apprentissage concomitant de MlwiN (maintenant que la version Windows a supplanté le mode Commandes utilisé dans ce livre). S'agissant des manuels de l'utilisateur des logiciels SAS, HLM et MlwiN, on ne saurait être aussi laudatif ; fort heureusement, pour SAS PROC MIXED, il existe un article remarquable, qu'on ne saurait trop conseiller de travailler en détail [SINGER 1998].

BIBLIOGRAPHIE

RÉFÉRENCES

- ABRAMOVICI Gérard, "La protection sociale en Europe", Statistiques en bref, Population et conditions sociales, Thème 3, N°1/2002, EUROSTAT, 2002, 7 p.
- AMROUNI Isabelle et Anne-Catherine RASTIER, "Les CAF et les jeunes adultes", Dossier d'étude N°18, CNAF, mars 2001, 90 p.
- AMROUNI Isabelle, "20 ans, un âge charnière...", RECHERCHES et PREVISIONS, CNAF, N°40, 1995, pp.7-17.
- AQUILINO William, "The Likelihood of Parent-Adult Child Coresidence: Effects of Family Structure and Parental Characteristics", JOURNAL OF MARRIAGE AND THE FAMILY, N°52, 1990, pp. 405-419
- AQUILINO William, "Family Structure and Home Leaving: A Further Specification of the Relationship", JOURNAL OF MARRIAGE AND THE FAMILY, N°53, 1991, pp. 999-1010.
- BAUDELLOT Christian et GLAUDE Michel : "Les diplômés se dévaluent-ils en se multipliant ?", ECONOMIE et STATISTIQUE, N°225, 1989.
- BAUDELLOT Christian et GOLLAC Michel : "Le salaire du trentenaire : question d'âge ou de génération", in ECONOMIE et STATISTIQUE, numéro spécial "Les trajectoires des jeunes : distances et dépendances entre générations", N°304-305, 4/5 1997, pp. 17-35.
- BELLOCO Rino, "Analysis of Longitudinal Data in STATA, Splus and SAS", Karolinska Intitutet, Stockholm, 2001, 32 p.
- BERGER Frédéric, "Habiter ou ne plus habiter chez ses parents", POPULATION ET EMPLOI, STATEC-CEPS/Instead-IGSS, N°2, 1998, pp. 1-12.
- BROWNING Martin, BOURGUIGNON François, CHIAPPORI Pierre-André, LECHENE Valérie, "Income and Outcomes: A Structural Model of Intrahousehold Allocation", JOURNAL OF POLITICAL ECONOMY, 1994, volume 102, N°6, pp. 1067-1096.
- BRYK Anthony S. et RAUDENBUSH Stephen W., "Hierarchical Linear Models: Applications and Data Analysis Methods", Sage, 1992
- CAHUC Pierre et ZYLBERBERG L., "Economie du travail, la formation des salaires et les déterminants du chômage", De Boeck Université, 1996.
- CARD David, "The Causal Effect of Education on Earnings", chapitre 30 du Handbook of Labor Economics, ASHENFELTER O. et CARD D. ed., 1999, Elsevier, pp.1801-1863.
- CAUSSAT Laurent, "Les chemins vers l'indépendance financière", ECONOMIE et STATISTIQUE, N°283-284, 1995, pp. 127-136.
- CHAMBAZ Christine, "Les jeunes adultes en Europe", Etudes et Résultats, DREES, N°90, novembre 2000, 12 p.
- CHAMBAZ Christine, "Construction, usages et portée des indicateurs statistiques", in "Comparer les systèmes de protection sociale en Europe", première partie de l'ouvrage "La protection sociale en Europe. Le temps des réformes", Christine DANIEL et Bruno PALIER (eds), La Documentation française, 2001, pp. 47-54.
- CHIAPPORI Pierre-André, "Collective Labor Supply and Welfare", JOURNAL OF POLITICAL ECONOMY, 1992, volume 100 N°3, pp. 437-467.
- CICCHELLI Vincenzo et ERLICH Valérie, "Se construire comme jeune adulte. Autonomie et autonomisation des étudiants par rapport à leurs familles", RECHERCHES et PREVISIONS, N°60, juin 2000, pp. 61-78.

COMMAILLE Jacques et DE SINGLY François, "Les règles de la méthode comparative dans le domaine de la famille. Le sens d'une comparaison", in "La question familiale en Europe", COMMAILLE Jacques et DE SINGLY François (ed.), pp. 7-30, L'Harmattan, 1997.

COMMISSARIAT GENERAL DU PLAN, "Jeunesse, le devoir d'avenir", Rapport de la commission présidée par Dominique CHARVET, La Documentation française, 2001.

COURGEAU Daniel, "Le départ de chez les parents : une analyse démographique sur le long terme", ECONOMIE et STATISTIQUE, N°337-338, 7/8 2000, pp. 37-60.

CURRAN Patrick J., "Introduction to Hierarchical Linear Models of Individual Growth: An Applied Example using the SAS Data System", 1998, University of North Carolina, Chapel Hill, 20 p.

EKERT-JAFFÉ Olivia, Denise ARBONVILLE et Jérôme WITTWER, "Ce que coûtent les jeunes de 18 à 25 ans", Rapport pour le Haut Conseil de la Population et de la Famille, Dossiers et Recherches N°47, INED, 1995, 48 p.

EUROSTAT 1996, "ECHP – 1997, Wave 4. Variable list ", DOC.PAN81/97, December 1996, 73 p.

EUROSTAT 2000a, "ECHP Classifications used in questionnaires", DOC.PAN73/00, May 2000, 25 p.

EUROSTAT 2000b, "Imputation of Income in the ECHP", DOC.PAN164/00, November 2000, 92 p.

EUROSTAT 2000c, "Construction of weights in the ECHP", DOC.PAN165/00, November 2000, 18 p.

EUROSTAT 2001a, "ECHP UDB Manual. European Community Household Panel. Longitudinal Users'Database. Waves 1 to 4. Survey years 1994 to 1997", DOC.PAN168/01, May 2001 (et version actualisée en décembre 2001), 18 pages

EUROSTAT 2001b, "ECHP UDB Construction of variables. From ECHP questions to UDB variables", DOC.PAN167/01, May 2001, 195 pages

EUROSTAT 2001c, "Data Alert N°1 on the ECHP Users'Database. Waves 1 to 4 (1994 to 1997)", DocPAN174/01, June 2001, 2 p.

EUROSTAT 2001d, "ECHP UDB Description of variables. Data Dictionary, Codebook and Differences between Countries and Waves", DOC.PAN166/01, May 2001 (et version actualisée en décembre 2001), 484 pages

EUROSTAT 2002a, "Adjusting Household Income for Within Household Non-response", DOC.PAN175/2002-01, January 2002

EUROSTAT 2002b, "EUROSTAT Yearbook 2001. The statistical guide to Europe. Data 1989-99", 2001. Sur le site Web d'EUROSTAT (<http://europa.eu.int/comm/eurostat>) on peut télécharger "100 basic indicators from Eurostat Yearbook 2001" (56 p.).

FERNANDEZ CORDON Juan Antonio, "Youth residential independence and autonomy: A comparative study", JOURNAL OF FAMILY ISSUES, N°18, 1997, pp. 567-607.

FESTY Patrick (ed.), "L'enfant dans la famille. Vingt ans de changements", numéro spécial de POPULATION, novembre-décembre 1994, 49^e année, N°6, 400 pages

FLIPO Anne, "Les comportements matrimoniaux de fait", Document de travail G2000/06 de la Direction des Etudes et Synthèses Economiques, INSEE, juin 2000, 45 p.

GACCIONE Peter, BLANCHARD M.S., "Nonlinear Mixed effects Models, a Tool for Analyzing Repeated Measurements. A Brief Tutorial Using SAS Software" (SAS NLINMIX macro), 6 p.

- GALLAND Olivier, "Entrer dans la vie adulte : des étapes toujours plus tardives mais resserrées", in *ECONOMIE et STATISTIQUE*, numéro spécial "Jeunes : l'âge des indépendances", N°337-338 7/8, 2000, pp. 13-36.
- GOLDSCHIEDER Frances, "Pathways to Independent Living in Early Adulthood: Marriage, Semiautonomy, and Premarital Residential Independence", *DEMOGRAPHY*, Volume 26, N°4, 1989, pp. 597-614.
- GOLDSCHIEDER Frances, "Recent Changes in US Young Adult Living Arrangements in Comparative Perspective", in "Still in the Nest", CHERLIN, SCABINI et ROSSI ed., *JOURNAL of FAMILY ISSUES*, 1997.
- GOLDSTEIN Harvey, "Multilevel Statistical Models", Edward Arnold, 1994, édition Internet, avril 1999.
- GOUX Dominique et MAURIN Eric, "Education, expérience et salaire : tendances récentes et évolutions de long terme", *ECONOMIE et PREVISION*, N°116, 1994, pp. 155-178.
- GREENE William H., "Econometric Analysis", MacMillan, 2^d édition, 1992, 791 p.
- GUILLOTIN Yves et SEVESTRE Jacques, "Estimations de fonctions de gains sur données de panel : endogénéité du capital humain et effets de la sélection", *ECONOMIE et PREVISION*, N°116, 1994, pp. 119-136.
- HAUSMAN Pierre, "Le mode de vie des jeunes adultes : cohabitation avec les parents et départ du foyer parental", Document PSELL N°90, CEPS/INSTEAD, 1996, 27 p.
- HECK Ronald et THOMAS Scott L., "An Introduction to Multilevel Modeling Techniques", Lawrence Erlbaum Associates Publishers, 2000
- HECKMAN James J. "Sample Selection Bias as a Specification Error", *ECONOMETRICA*, volume 47, N°1, 1979
- HECKMAN James J. et SMITH Jeffrey A., "Experimental and Nonexperimental Evaluation", in "International Handbook of Labour Market Policy and Policy Evaluation", SCHMID G., O'REILLY J et SCHÖMANN eds, Edward Elgar, 1996, pp. 37-88.
- HECKMAN James J. et SMITH Jeffrey A., "The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies", *ECONOMIC JOURNAL*, Volume 109, 1999, pp. 313-348.
- HEDEKER Donald, "Mixed Pattern-Mixture Models for Missing Data", University of Illinois, Chicago, 38 p.
- HERPIN Nicolas et VERGER Daniel, "Les étudiants, les autres jeunes, leur famille et la pauvreté", *ECONOMIE et STATISTIQUE* N°308-309-310, 8/9/10 1997, pp. 211-227.
- HOX Joop J., "Applied Multilevel Analysis", TT-Publikaties, Amsterdam, 1995, 118 p.
- IACOVU M., "Young People in Europe: Two Models of Household Formation", Working Paper N°98-13, Institute for Social and Economic Research, 1998
- IVERSEN Gudmund R., "Contextual Analysis", Series Quantitative Applications in the Social Sciences, N°81, Sage, 1991, 84 p.
- JAROUSSE Jean-Pierre et MINGAT Alain, "Un réexamen du modèle de gains de Mincer", *REVUE ECONOMIQUE*, volume 37, 1986, pp. 999-1031.
- KREFT Ita G.G. et DE LEEUW Jan, "Introducing Multilevel Modeling", Sage, 1998, 149 p.
- LITTELL Ramon C., MILLIKEN George A., STROUP Walter W. et WOLFINGER Russell D., "SAS System for Mixed Models", SAS Institute, 1996, 633 p.

MINCER J. "Investment in Human Capital and Personal Income Distribution", JOURNAL OF POLITICAL ECONOMY, vol. LXVI, N°4, 1958, pp. 281-302.

PALIER Bruno, "Les différents modèles de protection sociale et leur impact sur les réformes nationales", in "Comparer les systèmes de protection sociale en Europe", première partie de l'ouvrage "La protection sociale en Europe. Le temps des réformes", Christine DANIEL et Bruno PALIER (ed.), La Documentation française, 2001, pp. 33-46.

PAUGAM Serge et ZOYEM Jean-Paul, "Le soutien financier de la famille : une forme essentielle de la solidarité", ECONOMIE et STATISTIQUE N°308-309-310, 8/9/10, 1997, pp. 187-210.

PENARD Thierry et SOLLOGOUB Michel, "Les politiques françaises d'emploi en faveur des jeunes : une évaluation économétrique", Laboratoire de Microéconomie Appliquée, Paris I, 1994.

PERROT Anne et ZYLBERBERG André, "Salaire d'efficience et dualisme du marché du travail", REVUE ECONOMIQUE, volume 40, N°1, 1989, pp. 5-20.

QU Annie, "Comparison of PROC MIXED in SAS and HLM for Hierarchical Linear Models", 1997, 25 p.

RASBASH Jon, BROWNE William, GOLDSTEIN Harvey, YANG Min, PLEWIS Ian, HEALY Michael, WOODHOUSE Geoff, DRAPER David, LANGFORD Ian, LEWIS Toby, "A user's guide to MlwiN", version 2.1a, Multilevel Models Project, Institute of Education (IOE), University of London, novembre 2000, 278 p.

RAUDENBUSH Stephen W., BRYK Anthony S., CHEONG Yuk Fai, CONGDON Richard T. Jr., "HLM5. Hierarchical Linear and Nonlinear Modeling", SSI 5Scientific Software International), 2000, 305 p.

RAY Jean-Claude, "The Role of Demographic and Economic Characteristics of Twelve Member States of the European Union in the Diversity of State Support for Families", JOURNAL OF EUROPEAN SOCIAL POLICY, Volume 7, N°1, 1997, pp. 5-16.

RAY Jean-Claude, "Le rôle de la baisse des prestations Famille dans l'évolution du niveau de vie des familles quand les enfants grandissent", in "Comparer les systèmes de protection sociale en Europe du Nord et en France. Rencontres de Copenhague", DREES, Collection MiRe, Denis BOUGET et Bruno PALIER (ed.), Maison des Sciences de l'Homme Ange-Guépin, Nantes, volume 4, tome 1, 1999, pp. 343-372

ROSENZWEIG Mark et WOLPIN Kenneth, "Intergenerational Support and the Life-Cycle Incomes of Young Men and Their Parents: Human Capital Investments, Coresidence, and Intergenerational Financial Transfers", JOURNAL OF LABOR ECONOMICS, 1993, volume 11 N°1, pp. 84-112.

ROSENZWEIG Mark et WOLPIN Kenneth, "Parental and Public Transfers to Young Women and Their Children", AMERICAN ECONOMIC REVIEW, 1994, volume 84-85, pp. 1195-1212.

RYAN Paul, "The School-to-Work Transition: A Cross National Perspective", JOURNAL OF ECONOMIC LITERATURE, vol. XXXIX, 2001, pp. 34-92.

SA Carla et PORTELA Miguel : "Working and Studying: what explains youngsters' decisions?", Working Paper LIS, CEPS/INSTEAD, avril 1999, 19 p.

SAS/STAT User's Guide, version 8, 3 volumes, SAS Institute, 3809 p.

SINGER Judith D., "Fitting multilevel models using SAS", Multilevel Modelling Newsletter, vol.10, N°2, novembre 1998, pp. 5-9.

SINGER Judith D., "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models", Journal of Educational and Behavioral Statistics, hiver 1998, vol.24, N°4, pp. 323-355

SNIJDERS T.A.B. et BOSKERS Roel J., "Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling", SAGE, 1999, 266 p.

SOLAZ Anne, "Une réflexion économique sur le lien famille-chômage et la constitution du couple", in RECHERCHES et PREVISIONS, N°60, 2000, pp. 19-34.

SOLLOGOUB Michel et ULRICH Valérie, "Les jeunes en apprentissage ou en lycée professionnel : une mesure quantitative et qualitative de leur insertion sur le marché du travail", ECONOMIE et STATISTIQUE, N°323, 1999, pp. 31-52.

THELOT Claude et VILLAC Michel, "Politique familiale. Bilan et perspectives", Rapport à la ministre de l'emploi et de la solidarité et au ministre de l'économie, des finances et de l'industrie, La Documentation française, mai 1998, 251 p.

VERBEKE G. et MOLENBERGHS G., "Linear Mixed Models for Longitudinal Data", Springer, 1997, 600 p.

VILLENEUVE-GOKALP Catherine, "Le départ de chez les parents : définition d'un processus complexe", ECONOMIE et STATISTIQUE, N°304-305, 1997, pp. 149-162.

VILLENEUVE-GOKALP Catherine, "Les jeunes partent toujours au même âge de chez leurs parents", ECONOMIE et STATISTIQUE, N°337-338, 2000, pp. 61-80.

WEI-YIN Hu, "Welfare and Family Stability. Do Benefits Affect When Children Leave the Nest?", THE JOURNAL OF HUMAN RESOURCES, XXXVI-2, 2001, pp. 274-303.

WHITTINGTON Leslie et PETERS H. Elizabeth, "Economic Incentives for Financial and Residential Independence", DEMOGRAPHY, 1996, volume 33 N°1, pp. 82-97.

WILLETT John B., SINGER Judith D. et MARTIN Nina C., "The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations", DEVELOPMENT AND PSYCHOPATHOLOGY, 1998, volume 10, pp. 395-426.

WILLIS R. "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Function", chapitre 10, "Handbook of Labor Economics", volume 1, ASHENFELTER O. et LAYARD R. eds, North Holland, 1986.

ANNEXES

ANNEXE A : Résultats obtenus avec SAS

Graphique 1 : Profils, selon l'âge, des moyennes, par pays, des gains personnels d'activité des jeunes adultes européens (en SPA 1997 par an)

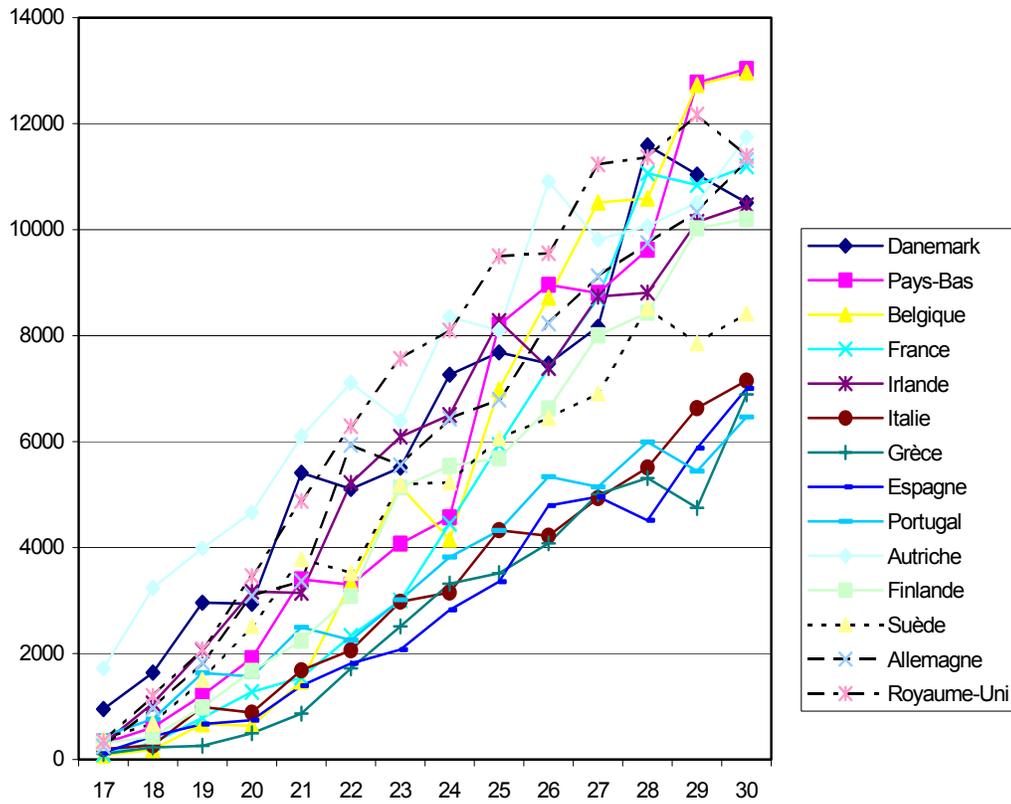


Tableau 1 : Effectifs de jeunes adultes par pays

| Pays | Nombre de ménages ECHP97 colonne [2] | Nombre d'individus ECHP97 colonne [3] | Nombre de jeunes adultes membres des ménages ECHP97 colonne [4] | % de jeunes adultes parmi les individus ECHP97 (=[4]/[3]) | Nombre de jeunes adultes ayant rempli un questionnaire individuel colonne [6] | Taux de perte parmi les jeunes adultes (=1-[6]/[4]) |
|--------------------|---|--|--|--|--|--|
| Danemark | 2745 | 6204 | 1288 | 20,8% | 1115 | 13,4% |
| Pays-Bas | 5049 | 12584 | 2438 | 19,4% | 1979 | 18,9% |
| Belgique | 3039 | 7916 | 1473 | 18,6% | 1247 | 15,3% |
| France | 6176 | 15758 | 3522 | 22,4% | 3263 | 7,4% |
| Irlande | 2945 | 9952 | 2539 | 25,5% | 2116 | 16,7% |
| Italie | 6713 | 20074 | 4867 | 24,2% | 4580 | 5,9% |
| Grèce | 4604 | 13491 | 2962 | 22,0% | 2743 | 7,4% |
| Espagne | 5794 | 18167 | 4645 | 25,6% | 4483 | 3,5% |
| Portugal | 4802 | 14428 | 3374 | 23,4% | 3078 | 8,8% |
| Autriche | 3142 | 8733 | 1867 | 21,4% | 1831 | 1,9% |
| Finlande | 4105 | 10887 | 2229 | 20,5% | 1985 | 10,9% |
| Suède | 5891 | 13361 | 2951 | 22,1% | 2079 | 29,5% |
| Allemagne | 6163 | 15942 | 3490 | 21,9% | 3090 | 11,5% |
| Royaume-Uni | 4965 | 12397 | 2711 | 21,9% | 2378 | 12,3% |

Tableau 2 : Statistiques descriptives concernant les variables des modèles estimés

| Variables | Minimum | Maximum | Moyenne | Ecart-type |
|--|---------|---------------------|---------|------------|
| <i>gains des jeunes adultes, en 1996</i> ⁺ | 0 | 199594 [°] | 5107 | 6902 |
| <i>âge</i> | 16 | 30 | 23,7 | 4,10 |
| <i>femme</i> | 0 | 1 | 0,50 | 0,50 |
| <i>éducation niveau intermédiaire</i> | 0 | 1 | 0,33 | 0,47 |
| <i>éducation niveau supérieur</i> | 0 | 1 | 0,14 | 0,35 |
| transferts sociaux / tête, en 1997 ⁺ | 2523 | 6614 | 5096 | 1123 |
| prestations chômage/ménage où vit un JA ⁺ | 40,89 | 2509 | 744 | 474 |
| prestations famille par ménage où vit un JA ⁺ | 35,32 | 3327 | 948 | 752 |
| taux chômage jeunes adultes | 0,05 | 0,18 | 0,12 | 0,05 |
| proportion d'étudiants parmi JA | 0,11 | 0,41 | 0,28 | 0,08 |
| âge * proportion étudiants | 1,89 | 12,23 | 6,56 | 2,18 |
| âge * taux chômage JA | 0,73 | 5,44 | 2,80 | 1,29 |
| femme * taux chômage JA | 0 | 0,18 | 0,06 | 0,07 |

⁺ : En SPA par an.

[°] : Le 99^e centile de la distribution des gains vaut 24512 SPA, soit environ 2000 SPA/mois. Au-delà du 20^e plus gros gain, on se situe en dessous de 5000 SPA/mois. Parmi les dix gains les plus élevés (qui s'échelonnent sans saut véritable entre pratiquement 200000 et 100000 SPA/an), on trouve des jeunes adultes de sept pays différents : âgés de 28 ans en moyenne, ces cinq hommes et ces cinq femmes vivent, pour six d'entre eux, avec un conjoint, et deux résident chez leurs parents.

Source : ECHP 1997, vague 4 (EUROSTAT)

14 pays (= les Quinze – Luxembourg), 35906 jeunes adultes (16-30 ans), données pondérées.

En italique, les variables individuelles (35906 jeunes adultes)

En gras, la variable issue des comptes de la protection sociale (SESROS/EUROSTAT), 14 valeurs (une par pays)

En caractères normaux, les montants moyens et les proportions calculés sur l'échantillon, 14 valeurs (une par pays)

En petits caractères, les variables croisées.

Tableau 3 : Résultats des modèles multi-niveaux

| | modèle A | | | | modèle B | | | | modèle C | | | | modèle D | | | |
|---------------------------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| Composantes de la variance | τ | | p(Z) (%) | |
| constante aléatoire | 2567156 | | 0,51 | | 11677962 | | 0,78 | | 13548961 | | 2,95 | | 11945750 | | 0,87 | |
| âge | - | | - | | 31228 | | 0,77 | | 32464 | | 1,28 | | 31238 | | 0,77 | |
| femme | - | | - | | 516143 | | 1,48 | | 219368 | | 4,95 | | 223877 | | 4,88 | |
| éducation niveau intermédiaire | - | | - | | 1065098 | | 1,52 | | 1056253 | | 1,54 | | 1062511 | | 1,52 | |
| éducation niveau supérieur | - | | - | | 3972936 | | 1,62 | | 4032199 | | 1,63 | | 4004142 | | 1,61 | |
| résidu | 44947062 | | <0,01 | | 31907793 | | <0,01 | | 31907606 | | <0,01 | | 31907872 | | <0,01 | |
| corrélation intra-classe ρ° | 5,4 % | | | | 26,8 % | | | | 29,8 % | | | | 27,2 % | | | |
| Effets fixes | β | σ | t | p(t) (%) | β | σ | t | p(t) (%) | β | σ | t | p(t) (%) | β | σ | t | p(t) (%) |
| constante | 5200,6 | 431,8 | 12,0 | <0,01 | -10491,0 | 956,6 | -11,0 | <0,01 | -2714 | 5863,1 | -0,46 | 65,58 | -10481 | 966,5 | -10,8 | <0,01 |
| âge | - | - | - | - | 676,5 | 48,9 | 13,8 | <0,01 | 716,1 | 206,6 | 3,47 | 0,05 | 676,7 | 48,9 | 13,8 | <0,01 |
| femme | - | - | - | - | -2204,1 | 211,5 | -10,4 | <0,01 | -3463,7 | 388,0 | -8,93 | <0,01 | -3463,8 | 391,0 | -8,9 | <0,01 |
| éducation niveau intermédiaire | - | - | - | - | 974,7 | 297,3 | 3,3 | 0,10 | 964,2 | 296,3 | 3,25 | 0,11 | 976,0 | 297,0 | 3,3 | 0,01 |
| éducation niveau supérieur | - | - | - | - | 3190,1 | 564,9 | 5,7 | <0,01 | 3175,4 | 568,9 | 5,58 | <0,01 | 3191,1 | 566,9 | 5,6 | <0,01 |
| transferts sociaux / tête | - | - | - | - | - | - | - | - | -0,967 | 0,886 | -1,09 | 30,67 | - | - | - | - |
| prest.chômage/ménage où JA | - | - | - | - | - | - | - | - | 0,125 | 1,905 | 0,07 | 94,94 | - | - | - | - |
| prest.famille/ménage où JA | - | - | - | - | - | - | - | - | -1,120 | 1,573 | -0,71 | 49,66 | - | - | - | - |
| taux chômage jeunes adultes | - | - | - | - | - | - | - | - | -7694,6 | 33468 | -0,23 | 82,39 | - | - | - | - |
| proportion d'étudiants parmi JA | - | - | - | - | - | - | - | - | -3455,7 | 19831 | -0,17 | 86,60 | - | - | - | - |
| âge * proportion étudiants | - | - | - | - | - | - | - | - | 626,3 | 857,6 | 0,73 | 46,52 | - | - | - | - |
| âge * taux chômage JA | - | - | - | - | - | - | - | - | -2062,7 | 1399,6 | -1,47 | 14,06 | - | - | - | - |
| femme * taux chômage JA | - | - | - | - | - | - | - | - | 11665 | 3269,3 | 3,57 | 0,04 | 11666 | 3295,9 | 3,5 | 0,04 |
| Critère d'ajustement : - 2 ResLL* | 758382,2 | | | | 746198,0 | | | | 746084,0 | | | | 746171,1 | | | |
| Critère d'ajustement : AIC** | 758382,2 | | | | 746210,0 | | | | 746096,0 | | | | 746183,1 | | | |

Source : ECHP 1997, vague 4 (EUROSTAT)

Variable expliquée : gains des jeunes adultes en 1996, en SPA par an.

14 pays (= les Quinze – Luxembourg), 35906 jeunes adultes (16-30 ans), 4580 observations au maximum par pays, données pondérées

Méthode d'estimation : maximum de vraisemblance restreint (REML). Méthode pour le calcul des degrés de liberté : Between-Within. SAS/PROC MIXED

Structure choisie pour la matrice de covariance : Composantes de la variance ('Variance Components')

^o Le coefficient de corrélation intra-classe ρ est calculé ainsi (cf. [SINGER 1998]) : τ (constante) / [τ (constante) + τ (résidus)]. Souvent τ (résidus) est noté σ^2 .

* ResLL = *Restricted (ou Residual) Log-Likelihood*, i.e. logarithme de la vraisemblance calculée selon la méthode REML.

** AIC = *AKAIKE Information Criterion*, critère d'information d'AKAIKE, qui pénalise l'ajout de variables explicatives supplémentaires.

Tableau 4 : Parallèle entre les résultats des modèles multi-niveaux (MMN) et ceux de régressions multiples classiques (REG)

| | modèle C (multiniveaux) | | | | modèle C' (régression classique) | | | | | modèle D (multiniveaux) | | | | modèle D' (régression classique) | | | | | |
|--|--------------------------------|----------|----------|----------|----------------------------------|----------|-------|----------|-------|-------------------------|----------|----------|----------|----------------------------------|----------|-------|----------|------|--|
| Composantes de la variance (modèles multi-niveaux) | τ | | p(Z) (%) | | | | | | | τ | | p(Z) (%) | | | | | | | |
| | constante aléatoire | 13548961 | | 2,95 | | | | | | | 11945750 | | 0,87 | | | | | | |
| | âge | 32464 | | 1,28 | | | | | | | 31238 | | 0,77 | | | | | | |
| | femme | 219368 | | 4,95 | | | | | | | 223877 | | 4,88 | | | | | | |
| | éducation niveau intermédiaire | 1056253 | | 1,54 | | | | | | | 1062511 | | 1,52 | | | | | | |
| | éducation niveau supérieur | 4032199 | | 1,63 | | | | | | | 4004142 | | 1,61 | | | | | | |
| | résidu | 31907606 | | <0,01 | | | | | | | 31907872 | | <0,01 | | | | | | |
| corrélation intra-classe ρ | 29,8 % | | | | | | | | | 27,2 % | | | | | | | | | |
| Effets fixes (MMN) ou paramètres estimés (REG) | β | σ | t | p(t) (%) | β | σ | t | p(t) (%) | VIF | β | σ | t | p(t) (%) | β | σ | t | p(t) (%) | VIF | |
| constante | -2714 | 5863,1 | -0,46 | 65,58 | -15624 | 680,94 | -22,3 | <0,01 | 0 | -10481 | 966,5 | -10,8 | <0,01 | -7576,6 | 213,7 | -35,5 | <0,01 | 0 | |
| âge | 716,1 | 206,6 | 3,47 | 0,05 | 950,0 | 27,3 | 34,7 | <0,01 | 13,8 | 676,7 | 48,9 | 13,8 | <0,01 | 705,1 | 8,1 | 87,1 | <0,01 | 1,19 | |
| femme | -3463,7 | 388,0 | -8,93 | <0,01 | -3608,7 | 154,5 | -23,4 | <0,01 | 6,6 | -3463,8 | 391,0 | -8,9 | <0,01 | -3607,4 | 155,6 | -23,2 | <0,01 | 6,56 | |
| éducation niveau intermédiaire | 964,2 | 296,3 | 3,25 | 0,11 | 776,4 | 72,1 | 10,8 | <0,01 | 1,3 | 976,0 | 297,0 | 3,3 | 0,01 | 818,9 | 72,0 | 11,4 | <0,01 | 1,25 | |
| éducation niveau supérieur | 3175,4 | 568,9 | 5,58 | <0,01 | 2892,3 | 97,6 | 29,6 | <0,01 | 1,3 | 3191,1 | 566,9 | 5,6 | <0,01 | 3063,4 | 96,3 | 31,8 | <0,01 | 1,21 | |
| transferts sociaux / tête | -0,967 | 0,886 | -1,09 | 30,67 | 0,339 | 0,035 | 9,6 | <0,01 | 1,7 | - | - | - | - | - | - | - | - | - | |
| prest.chômage/ménage où JA | 0,125 | 1,905 | 0,07 | 94,94 | 0,288 | 0,101 | 2,9 | 0,42 | 2,5 | - | - | - | - | - | - | - | - | - | |
| prest.famille/ménage où JA | -1,120 | 1,573 | -0,71 | 49,66 | 0,356 | 0,075 | 4,7 | <0,01 | 3,5 | - | - | - | - | - | - | - | - | - | |
| taux chômage jeunes adultes | -7694,6 | 33468 | -0,23 | 82,39 | 27636 | 5496,9 | 5,0 | <0,01 | 83,9 | - | - | - | - | -30902 | 853,5 | -36,2 | <0,01 | 1,99 | |
| proportion d'étudiants parmi JA | -3455,7 | 19831 | -0,17 | 86,60 | -3797,5 | 3440,0 | -1,1 | 26,96 | 79,7 | - | - | - | - | - | - | - | - | - | |
| âge * proportion étudiants | 626,3 | 857,6 | 0,73 | 46,52 | -59,9 | 141,7 | -0,4 | 67,26 | 104,5 | - | - | - | - | - | - | - | - | - | |
| âge * taux chômage JA | -2062,7 | 1399,6 | -1,47 | 14,06 | -1960,3 | 223,8 | -8,8 | <0,01 | 91,5 | - | - | - | - | - | - | - | - | - | |
| femme * taux chômage JA | 11665 | 3269,3 | 3,57 | 0,04 | 12149 | 1200,8 | 10,1 | <0,01 | 7,5 | 11666 | 3295,9 | 3,5 | 0,04 | 12130 | 109,5 | 10,0 | <0,01 | 7,50 | |
| Critère d'ajustement MMN : - 2 ResLL | 746084,0 | | | | - | | | | | 746171,1 | | | | - | | | | | |
| Critère d'ajustement REG : R ² | | | | | 31,35% | | | | | - | | | | 30,34% | | | | | |

Source : ECHP 1997, vague 4 (EUROSTAT)

Variable expliquée : gains des jeunes adultes en 1996, en SPA par an.

14 pays (= les Quinze – Luxembourg), 35906 jeunes adultes (16-30 ans), 4580 observations au maximum par pays, données pondérées

Annexe B : Résultats de modèles multi-niveaux obtenus avec HLM5, MlwiN et SAS

Les **paramétrisations** de SAS/PROC MIXED, de HLM5 et de MlwiN diffèrent parfois sensiblement, ce qui ne facilite pas le passage d'un logiciel à l'autre. C'est ainsi que, dans son écran d'affichage des équations, MlwiN inclut le terme d'erreur individuel dans la première équation de niveau 2 (celle relative à l'explication de la constante aléatoire du niveau 1) au lieu de le faire figurer, comme dans SAS et HLM5, en queue d'équation de niveau 1 ; plus perturbant, MlwiN, au lieu d'inclure dans les équations de niveau 2, comme le fait HLM, les variables explicatives de type Z, dont c'est la place naturelle (ce qui rend très lisible la présentation de l'écran d'équations du logiciel HLM), place ces variables au niveau 1, et cela isolément si elles expliquent la constante aléatoire ou, si elles expliquent le coefficient aléatoire de tel X, de façon croisée avec cet X. Certes on retrouve ici le type de spécification exigé par l'instruction MODEL de SAS/PROC MIXED, mais SAS présente ce mode d'écriture comme résultant de la combinaison d'une équation de niveau 1 qui n'inclut aucun Z et d'autant d'équations de niveau 2 qu'il y a, au niveau 1, de coefficients considérés comme aléatoires (et expliqués, ou non, par des Z, qu'on voit donc bien apparaître là où ils interviennent).

Afin d'illustrer les différences de présentation des équations et des résultats de ces trois logiciels majeurs, on a choisi de les présenter d'abord (comme le fait QU [1997] pour son parallèle SAS/HLM4.01 entre deux modèles très simples) **sous la forme où ils sont fournis**, tout en les assortissant de quelques commentaires (mis en italique). Mais nous allons aussi en fournir un **commentaire synthétique**.

Le même modèle va être mis en œuvre ici avec HLM5, puis MlwiN puis SAS/PROC MIXED : il s'agit du **meilleur modèle** auquel nous sommes parvenu **après élimination des facteurs de contrôle apparaissant comme sans effet significatif**. En cela, il se distingue de notre modèle C (dont les résultats ont été présentés dans le corps du texte) qui, pour démontrer leur absence d'effet, avait maintenu, quoique qu'elles apparussent dépourvues d'effets significatifs, les variables d'intérêt que sont, dans ce travail, d'une part la fraction des jeunes adultes qui sont étudiants, et, d'autre part, le montant national annuel moyen (en SPA¹) des prestations Famille par ménage où vit un jeune adulte : ici, nous avons retiré ces variables (et, pour ne pas accroître encore la complexité du modèle, nous n'avons **pas introduit d'effet d'interaction au niveau 1**, se traduisant par des variables croisant les variables explicatives individuelles, par exemple le sexe et l'âge ; on peut pourtant se demander si l'effet du sexe sur les gains n'est pas fonction de l'âge ou, réciproquement, si l'effet de l'âge n'est pas fonction du sexe).

¹ Parce qu'un bogue – non signalé - dans le format choisi par HLM5 pour le fichier des résidus le rend illisible si certaines valeurs sont "trop grandes" (ex.: effet inférieur à -999 ou supérieur à 9999, covariance inférieure à -99999 ou variance ou covariance excédant 999999 : les nombres sont alors collés les uns aux autres dans le fichier texte, voire se superposent partiellement), nous avons refait tourner nos modèles sous SAS, HLM5 et MlwiN après avoir exprimé tous les montants financiers (i.e. les gains et les divers transferts sociaux) en **milliers** de SPA (et non plus en SPA). Cela améliore la lisibilité des résultats chiffrés (parce que leur présentation par ces trois logiciels ignore les séparateurs de milliers) mais, sans doute parce qu'il s'agit de méthodes itératives, les résultats obtenus ne sont pas prédictibles comme le seraient ceux d'une régression multiple dont Y et certains X seraient divisés par 1000 (i.e. coefficients de régression et écarts-types estimés divisés exactement par 1000 pour ces X-là, mêmes t et mêmes p, même R²) : ici, les modifications de résultats sont seulement approximatives, sans changement de grande ampleur cependant ; quant aux estimations des variances-covariances, elles sont grosso modo divisées par 1000², soit un million.

Ce retrait des variables ETUDIANTS et PRESTFAMILLE caractérisait aussi notre modèle D, similaire par conséquent, dans son esprit, à notre "meilleur modèle" présenté dans cette annexe. Mais, en adoptant la même procédure de modélisation ascendante sans, cette fois, contraindre toutes les covariances à être nulles (ce qui était le cas avec le modèle D), on aboutit à des résultats parfois sensiblement différents de ceux du modèle D ; ce qui interroge, voire inquiète, quant à la robustesse des coefficients estimés. En effet, comme on va le voir, certains facteurs de niveau 2, qui ne passaient pas, dans le modèle D, le seuil de significativité, le franchissent ici nettement : il en va ainsi de la proportion de chômeurs parmi les jeunes adultes, qui explique ici non seulement le coefficient aléatoire de FEMME, comme dans le modèle D, quoique avec un coefficient sensiblement différent, mais aussi la constante aléatoire. C'est également le cas du montant moyen des allocations chômage, qui explique ici le coefficient aléatoire de EDUC_SUP, et du montant moyen des transferts sociaux, qui explique significativement le coefficient aléatoire de l'âge.

L'autre caractéristique de ce modèle, et c'est une autre différence avec le modèle C, est qu'il ne suppose rien quant aux valeurs des **covariances des termes d'erreurs des effets aléatoires** : elles ne sont **pas contraintes à être nulles** comme c'était le cas avec les modèles présentés dans le corps du texte. Cette caractéristique oblige à utiliser, dans SAS/PROC MIXED, la **méthode MIVQUE0**, puisque nous n'arrivons pas à obtenir de résultat avec les méthodes ML ou REML (maximum de vraisemblance simple ou restreint) ; cette méthode, peu explicitée dans la documentation SAS (qui renvoie seulement à des articles de statistique pure), a la particularité de ne pas être itérative. LITTELL et al. [1996, p.499] recommandent son emploi pour les bases de données de grande taille (ce qui nous semble être le cas ici) ou pour des situations dans lesquelles les procédures REML et ML ne parviennent pas à la convergence (ce qui est aussi notre cas). Effectivement, cette méthode nous permet d'obtenir des résultats, mais certains indices nous laissent penser qu'on ne doit pas leur accorder une confiance aveugle : d'une part, des essais menés systématiquement sur des modèles pour lesquels REML et ML fonctionnent nous ont montré que les résultats obtenus avec MIVQUE0 s'écartaient parfois sensiblement de ceux affichés par REML et ML ; d'autre part, lorsqu'on n'a plus de point de comparaison possible avec ML et REML (parce que ces procédures avortent), on constate, chez MIVQUE0 de curieux phénomènes (ainsi plusieurs valeurs du test Z sont identiques – et égales à 133,85 – comme si on atteignait là une sorte de limite) et la divergence avec les résultats issus des logiciels MLwiN et HLM peut être très sensible, notamment en ce qui concerne les valeurs estimées des variances-covariances. Dans le doute, on adoptera donc une attitude prudente, consistant à ne mettre en avant aucun résultat de fond issu de cette méthode.

Enfin, le modèle que nous commentons ici se différencie aussi du modèle C par le fait que, pour donner un sens immédiat aux coefficients, on a **centré les Z sur la moyenne communautaire** (via une simple moyenne arithmétique des Z par pays, Luxembourg non inclus), et on a **translaté l'âge en lui ôtant 16 années**, ce qui dégonfle la constante du modèle puisqu'elle représente désormais les gains moyens pour un garçon de 16 ans n'ayant pas (ou pas encore) accédé au niveau lycée, au lieu de mesurer ses gains à la naissance, rétropolés à partir de l'évolution, supposée linéaire, des gains moyens (à sexe et niveau d'éducation donnés) entre 16 et 30 ans.

La proximité des résultats des modèles estimés par HLM et par MlwiN (désignons-les respectivement par les lettres E et F) permet d'en faire un **commentaire unique** ; en revanche, les résultats de SAS/PROC MIXED (modèle G), quoique souvent voisins, diffèrent parfois sensiblement et seront donc fournis à titre simplement illustratif.

Que constatons-nous donc au vu des résultats de HLM et de MlwiN ?

D'abord, la **constante aléatoire** de niveau 1 est exclusivement expliquée par le taux de chômage des jeunes adultes ; sa constante fixe est en effet trouvée non significative. Quant à la variance du terme d'erreur de cette constante aléatoire, elle apparaît non significativement différente de 0 ; ce qui signifie que, lorsqu'on contrôle ces facteurs individuels et le taux de chômage national, la moyenne des gains des jeunes adultes par pays ne diffère plus sensiblement d'un pays à l'autre (autrement dit, les différences de gains moyens tiennent essentiellement à la diversité des structures par âge, sexe et niveau d'éducation et à des taux de chômage différents ; si ces caractéristiques étaient identiques dans toute l'Union européenne, il ne resterait plus guère de différences de gains moyens entre pays).

Quant aux quatre variables de niveau 1 (caractéristiques individuelles d'âge, de sexe et de niveaux d'éducation), elles ont effectivement des effets aléatoires, i.e. significativement variables selon les pays. Trois effets sur quatre (et non plus un seul, à la différence du modèle D) sont trouvés dépendre de facteurs de niveau pays, ce qui module l'ampleur de l'effet selon la valeur de ces variables, pays par pays.

Ainsi l'effet de l'**âge** est-il égal à environ² $675^3 + 0,06 \text{ TRFSOCX}$, ce qui correspond à une majoration moyenne de gains, par année d'âge, de 773 SPA/an (=675+98) pour le pays le plus généreux⁴ (où TRFSOCX=1628 de plus que la moyenne communautaire) et à une majoration de 527 seulement (=675-148) pour le moins généreux (TRFSOCX=-2463) ; on se trouve typiquement là dans une situation où, quoique significatif, l'effet d'une variable Z n'a qu'un impact relatif assez modeste sur l'effet de la variable individuelle (âge). Quant au fond, on notera que l'effet des transferts sociaux est positif (ils amplifient donc l'effet de l'âge sur les gains), ce qui accrédirait l'idée qu'il n'y a pas substituabilité entre les efforts faits par l'individu pour gagner lui-même sa vie et le soutien financier de la collectivité nationale à son égard ; du moins à condition que ce signe positif ne soit pas dû à un phénomène de multicolinéarité entre la possibilité de financer des transferts sociaux généreux et une productivité nationale autorisant des gains substantiels...

² Le calcul de l'effet prédit, pour tel groupe, d'une variable individuelle est, on le sait, moins simple en principe que ce que nous écrivons ici, puisqu'il est égal à une moyenne arithmétique de l'effet estimé par les moindres carrés sur ce seul groupe et de l'effet estimé (par la méthode empirique de BAYES, qui produit des 'moyennes postérieures') pour l'ensemble de la population, ces deux éléments étant pondérés par un facteur donnant d'autant plus de poids au second terme que le nombre d'individus de l'unité en question est petit (ce qui rend fragile l'estimation sur la base de ces seuls individus). Mais, ici, chaque pays a un très grand nombre d'individus, ce qui fait que l'effet bayésien est quasiment oublié puisque le facteur de pondération de l'estimateur des moindres carrés est presque égal à un (ce n'est autre que la fiabilité – *reliability* – de la moyenne du groupe en question, égale à la variance vraie de cette moyenne sur la variance mesurée à partir des seuls individus observés : comme ils sont très nombreux, le ratio de ces deux variances avoisine 1).

³ Ce chiffre est une moyenne grossière et arrondie des valeurs estimées (avec un bon degré de précision dans chaque cas) par les logiciels HLM5 et MlwiN. Il ne s'agit ici, en effet, que d'avoir un ordre de grandeur.

⁴ Compte tenu de l'ambition de cet article (à savoir, aller au-delà du facteur 'pays'), nous nous sommes bien gardé de publier des **résultats pays par pays**, bien que les modèles multi-niveaux permettent – cf., infra, l'extrait de sorties SAS, donnant l'estimation des effets aléatoires pays par pays – de telles analyses (qui sont facilitées quand on n'a, comme ici, qu'un petit nombre d'unités de niveau 2 : avec 14 pays, on peut en effet examiner en détail les résultats pays par pays, alors qu'avec un millier d'écoles on ne peut prêter véritablement attention qu'à celles qui sont très atypiques). Ceci dit, pour donner une idée de l'ampleur des effets, comme nous le faisons ici, il est commode d'utiliser les valeurs moyennes ou extrêmes des Z observés dans l'Union Européenne, même si on ne nomme pas le pays concerné.

Quant à l'effet du **sexe** sur les gains des jeunes adultes, il est égal à environ $-2240 + 13130 * \text{CHOMEURS}$, ce qui correspond, pour les jeunes femmes du pays ayant le plus fort taux de chômage (à savoir $\text{CHOMEURS}=7,54$ points de pourcentage de plus que la moyenne communautaire) à des gains inférieurs de 1249 SPA/an ($=991-2240$) à ceux de leurs compatriotes masculins, mais seulement de 796 ($=796-2240$) pour le pays ayant le taux de chômage le plus faible ($\text{CHOMEURS}=6,07$ points de pourcentage de moins que la moyenne communautaire) : à en croire ces chiffres, l'impact du taux de chômage serait donc très sensible ; quant à son signe, il signifie qu'un taux de chômage faible réduirait plutôt, ceteris paribus, les différences de gains hommes/femmes aux âges jeunes (peut-être davantage en raison d'une moindre inégalité, alors, du volume des heures travaillées selon le sexe, qu'en raison d'une inégalité sexuelle des taux de salaire, à qualification égale ; mais on ne dispose pas ici, en la matière, d'éléments pour aller au-delà de cette simple conjecture).

Quant à l'effet de l'**éducation** sur les gains des jeunes adultes, il est égal à environ 1000 SPA/an de majoration pour les jeunes disposant d'une éducation de niveau lycée ou baccalauréat par rapport à ceux qui sont moins éduqués. Mais l'effet d'une éducation supérieure au baccalauréat est plus complexe puisqu'il dépend du niveau moyen des allocations de chômage versées aux jeunes adultes dans le pays en question : plus fortes sont ces allocations, plus important semble être l'effet : environ $3260 + 1,255 \text{ ALLOCHOM}$ SPA/an de plus pour une éducation supérieure par rapport à une éducation de niveau inférieure ou égale au collège ; ce qui donnerait des gains accrus de $3260+1987=5247$ SPA/an dans le pays ayant les montants moyens d'allocations de chômage les plus élevés, soit 1582 SPA/an de plus que la moyenne communautaire ; et majorés de 1111 seulement ($=3260-2149$) dans le pays ayant les montants moyens d'allocations de chômage les plus faibles, soit 885 SPA/an de moins que la moyenne communautaire. Mais peut-être sommes-nous là en présence non pas d'une relation de comportement (exemple : l'investissement en capital humain pourrait être mieux rémunéré dans les pays où les cotisations d'assurance-chômage grèveraient moins les comptes des entreprises) mais d'une relation presque comptable – soulevant d'ailleurs un problème d'endogénéité - dans la mesure où les allocations de chômage relevant de la logique d'assurance sont fonction du niveau des gains passés.

Avant de passer aux résultats de chacun des logiciels, soulignons que **HLM5 et MlwiN ne sont pas autonomes** : ils requièrent l'usage, en amont et en aval des estimations, de logiciels de traitement de données (comme SAS ou SPSS). C'est en effet en dehors de ces logiciels spécialisés qu'il faut préparer les données (en amont des analyses), et cela selon des spécifications bien précises (un fichier unique pour MlwiN, incluant une variable 'constante' – encore qu'on puisse la créer au sein de MlwiN – et trié à l'avance ; mais deux fichiers distincts pour HLM, reliés par l'identifiant de pays) ; et, en aval de l'estimation des modèles, pour un certain nombre de tâches d'utilisation des sous-produits des modèles (ex.: graphiques de résidus dans le cas de HLM5), chaque utilisateur est invité à recourir à son logiciel préféré – SAS par exemple –, pour traiter les données créées par HLM5 dans un fichier qu'il faudra préalablement soumettre à SAS (ou SPSS, ou SYSTAT) pour que soit créée une base de données adaptée au logiciel en question.

1. HLM5

On présente ici successivement l'image de l'écran de choix du modèle à estimer puis le listing des résultats (dans lequel on a fait figurer en plus petits caractères ce qui concerne les estimations robustes puisque, dans le cas de nos modèles, HLM5 fait suivre leur exposé d'une note indiquant qu'elles sont inutiles compte tenu de ce que nos données ne comprennent pas un nombre d'unités de second niveau – les pays – moyennement grand ou vraiment grand).

Il s'agit uniquement de résultats concernant le module HLM2, qui est le module adapté au cas de modèles à deux niveaux seulement, et où Y est continu et distribué normalement.

On ne présente ici ni le mode de constitution des fichiers de données (qui est assez particulier, avec le concept intéressant de SSM – *Sufficient Statistics Matrices*) ni les écrans de spécifications de base ou de spécifications optionnelles (ex.: écran au moyen duquel on spécifie les contraintes qu'on veut éventuellement imposer aux coefficients fixes).

On ne reproduit pas non plus les quatre fichiers de résultats associés au listing central : un fichier créant (sous SAS, SPSS ou SYSTAT) une base de données contenant les divers types de résidus (attention aux nombres collés à cause du format choisi par HLM : le fichier devient illisible), un fichier texte contenant les tau (et, en principe, leur matrice de variances-covariances, mais je n'ai pas réussi à l'obtenir) et deux fichiers texte contenant les gammas et leur matrice de variances-covariances (un fichier pour les résultats standards, un autre pour les résultats robustes).

```
Program:                HLM 5 Hierarchical Linear and Nonlinear Modeling
Authors:               Stephen Raudenbush, Tony Bryk, & Richard Congdon
Publisher:             Scientific Software International, Inc. (c) 2000
                                techsupport@ssicentral.com
                                www.ssicentral.com
```

```
-----
Module:                HLM2.EXE (5.02.20321.1)
Date:                  20 April 2002, Saturday
Time:                  12:23:28
-----
```

SPECIFICATIONS FOR THIS HLM2 RUN

```
-----
Problem Title:        MODELE AVEC TOUTES COV, POUR COMPARAISON AVEC SAS MIVQUE0 ET ML
```

```
The data source for this run = HLM5_JA1.SSM
The command file for this run = D:\_CD_echp\ephemere\ttesCOV_avec_cste.hlm
Output file name       = D:\_CD_echp\ephemere\hlm2_ttesCOV_avec_cste.txt
The maximum number of level-2 units = 14
The maximum number of iterations = 100
Method of estimation:  restricted maximum likelihood
```

Weighting Specification

```
-----
                                Weight
                                Variable
Level 1      Weighting?  Name      Normalized?
Level 1      yes        POIDS    yes
Level 2      no                 no
Generalizations are at level-1
```

```
The outcome variable is      GAINS
```

The model specified for the fixed effects was:

| Level-1 Coefficients | Level-2 Predictors |
|-------------------------|-----------------------|
| INTRCPT1, B0 | INTRCPT2, G00 |
| FEMME slope, B1 | CHOMEURS, G01 |
| EDUC_SUP slope, B2 | INTRCPT2, G10 |
| EDUC_MID slope, B3 | CHOMEURS, G11 |
| AGE_16 slope, B4 | INTRCPT2, G20 |
| | ALLOCHOM, G21 |
| | INTRCPT2, G30 |
| | INTRCPT2, G40 |
| | TRFSOCKX, G41 |

The model specified for the covariance components was:

Sigma squared (constant across level-2 units)

Tau dimensions

INTRCPT1

FEMME slope

EDUC_SUP slope

EDUC_MID slope

AGE_16 slope

Summary of the model specified (in equation format)

Level-1 Model

$$Y = B0 + B1*(FEMME) + B2*(EDUC_SUP) + B3*(EDUC_MID) + B4*(AGE_16) + R$$

Level-2 Model

$$B0 = G00 + G01*(CHOMEURS) + U0$$

$$B1 = G10 + G11*(CHOMEURS) + U1$$

$$B2 = G20 + G21*(ALLOCHOM) + U2$$

$$B3 = G30 + U3$$

$$B4 = G40 + G41*(TRFSOCKX) + U4$$

=====

Level-1 OLS regressions

| Level-2 Unit | INTRCPT1 | FEMME slope | EDUC_SUP slope | EDUC_MID slope | AGE_16 slope |
|--------------|-------------|-------------|----------------|----------------|--------------|
| 1 | 1373.07690 | -2891.29028 | 4968.81982 | 1543.88818 | 689.41272 |
| 2 | 1733.84009 | -2927.10815 | 2763.07544 | 1515.07043 | 665.92145 |
| 3 | -681.35138 | -2077.34595 | 1997.77930 | -283.31393 | 996.32642 |
| 4 | -934.19720 | -2688.73511 | 3702.99048 | 582.49310 | 976.57971 |
| 6 | -2336.65088 | -1447.74622 | 4087.72681 | 2047.62476 | 896.08124 |
| 7 | 1203.46570 | -3235.69775 | 2476.34863 | 776.50684 | 855.71649 |
| 8 | 794.14801 | -2284.11157 | 7595.00928 | 3590.29883 | 587.88892 |
| 9 | -261.83542 | -1532.44116 | -56.97269 | 155.79456 | 553.69031 |
| 10 | -384.64258 | -1586.07043 | 1811.63049 | 667.36749 | 456.98746 |
| 11 | 300.11172 | -1626.67627 | 1783.28345 | -494.48453 | 446.82245 |
| 12 | 739.45856 | -1206.49524 | 8201.90332 | -493.80841 | 434.47153 |
| 13 | 3049.42065 | -3651.15186 | 3278.68750 | 2640.66626 | 568.77307 |
| 14 | 27.52871 | -2504.10620 | 3177.32422 | 844.61627 | 693.35919 |
| 15 | 365.75287 | -1550.92920 | 1119.42126 | 1394.49548 | 565.36151 |

The average OLS level-1 coefficient for INTRCPT1 = 356.29470
 The average OLS level-1 coefficient for FEMME = -2229.27896
 The average OLS level-1 coefficient for EDUC_SUP = 3350.50195
 The average OLS level-1 coefficient for EDUC_MID = 1034.80110
 The average OLS level-1 coefficient for AGE_16 = 670.52803

Least Squares Estimates

sigma_squared = 32626138.88080

The outcome variable is GAINS

Least-squares estimates of fixed effects

| Fixed Effect | Coefficient | Standard Error | T-ratio | d.f. | P-value |
|------------------------|---------------|----------------|---------|-------|---------|
| For INTRCPT1, B0 | | | | | |
| INTRCPT2, G00 | 438.342532 | 72.138137 | 6.076 | 35897 | 0.000 |
| CHOMEURS, G01 | -26093.681031 | 895.060193 | -29.153 | 35897 | 0.000 |
| For FEMME slope, B1 | | | | | |
| INTRCPT2, G10 | -2332.170013 | 62.177303 | -37.508 | 35897 | 0.000 |
| CHOMEURS, G11 | 12028.306643 | 1200.119039 | 10.023 | 35897 | 0.000 |
| For EDUC_SUP slope, B2 | | | | | |
| INTRCPT2, G20 | 3317.890600 | 103.206549 | 32.148 | 35897 | 0.000 |
| ALLOCHOM, G21 | 0.762580 | 0.172732 | 4.415 | 35897 | 0.000 |
| For EDUC_MID slope, B3 | | | | | |
| INTRCPT2, G30 | 697.428978 | 71.693736 | 9.728 | 35897 | 0.000 |
| For AGE_16 slope, B4 | | | | | |
| INTRCPT2, G40 | 688.998522 | 8.060489 | 85.478 | 35897 | 0.000 |
| TRFSOCX, G41 | 0.076528 | 0.003504 | 21.838 | 35897 | 0.000 |

The outcome variable is GAINS [ce qui suit est mis en petits caractères car nos données n'ont pas besoin d'estimation robuste des écarts-types]

Least-squares estimates of fixed effects

(with **robust** standard errors)

| Fixed Effect | Coefficient | Standard Error | T-ratio | d.f. | P-value |
|------------------------|---------------|----------------|---------|-------|---------|
| For INTRCPT1, B0 | | | | | |
| INTRCPT2, G00 | 438.342532 | 406.771610 | 1.078 | 35897 | 0.282 |
| CHOMEURS, G01 | -26093.681031 | 3016.742943 | -8.650 | 35897 | 0.000 |
| For FEMME slope, B1 | | | | | |
| INTRCPT2, G10 | -2332.170013 | 116.535582 | -20.013 | 35897 | 0.000 |
| CHOMEURS, G11 | 12028.306643 | 1993.449971 | 6.034 | 35897 | 0.000 |
| For EDUC_SUP slope, B2 | | | | | |
| INTRCPT2, G20 | 3317.890600 | 348.652694 | 9.516 | 35897 | 0.000 |
| ALLOCHOM, G21 | 0.762580 | 0.537308 | 1.419 | 35897 | 0.156 |
| For EDUC_MID slope, B3 | | | | | |
| INTRCPT2, G30 | 697.428978 | 271.327679 | 2.570 | 35897 | 0.010 |
| For AGE_16 slope, B4 | | | | | |
| INTRCPT2, G40 | 688.998522 | 43.443413 | 15.860 | 35897 | 0.000 |
| TRFSOCX, G41 | 0.076528 | 0.011717 | 6.531 | 35897 | 0.000 |

The robust standard errors are appropriate for datasets having a moderate to large number of level 2 units. These data do not meet this criterion.

The least-squares likelihood value = -361509.433293

Deviance = 723018.86659

Number of estimated parameters = 1

STARTING VALUES

sigma(0)_squared = 31857200.63556

Tau(0)

| | | | | | |
|-------------|---------------|---------------|---------------|---------------|---------------|
| INTRCPT1,B0 | 1188323.36239 | -207854.63054 | -564807.06820 | 99708.01200 | -137237.15532 |
| FEMME,B1 | -207854.63054 | 219847.56295 | 529075.31424 | -190142.06826 | 2981.31540 |
| EDUC_SUP,B2 | -564807.06820 | 529075.31424 | 4340652.68417 | 411969.90577 | 72260.99426 |
| EDUC_MID,B3 | 99708.01200 | -190142.06826 | 411969.90577 | 1343355.50821 | -20469.30943 |
| AGE_16,B4 | -137237.15532 | 2981.31540 | 72260.99426 | -20469.30943 | 22960.57886 |

+++++

The outcome variable is GAINS

Estimation of fixed effects
(Based on starting values of covariance components)

| Fixed Effect | Coefficient | Standard Error | T-ratio | Approx. d.f. | P-value |
|------------------------|---------------|----------------|---------|--------------|---------|
| For INTRCPT1, B0 | | | | | |
| INTRCPT2, G00 | 367.654020 | 309.863805 | 1.187 | 12 | 0.259 |
| CHOMEURS, G01 | -22360.334838 | 4031.384365 | -5.547 | 12 | 0.000 |
| For FEMME slope, B1 | | | | | |
| INTRCPT2, G10 | -2263.700696 | 147.158750 | -15.383 | 12 | 0.000 |
| CHOMEURS, G11 | 12998.719398 | 2472.327431 | 5.258 | 12 | 0.000 |
| For EDUC_SUP slope, B2 | | | | | |
| INTRCPT2, G20 | 3238.736472 | 585.481580 | 5.532 | 12 | 0.000 |
| ALLOCHOM, G21 | 1.280432 | 0.712143 | 1.798 | 12 | 0.097 |
| For EDUC_MID slope, B3 | | | | | |
| INTRCPT2, G30 | 985.034273 | 328.475152 | 2.999 | 13 | 0.011 |
| For AGE_16 slope, B4 | | | | | |
| INTRCPT2, G40 | 673.890477 | 42.493582 | 15.859 | 12 | 0.000 |
| TRFSOCX, G41 | 0.062399 | 0.013063 | 4.777 | 12 | 0.000 |

The value of the likelihood function at iteration 1 = -3.611463E+005
The value of the likelihood function at iteration 2 = -3.611454E+005
The value of the likelihood function at iteration 3 = -3.611449E+005
The value of the likelihood function at iteration 4 = -3.611446E+005
The value of the likelihood function at iteration 5 = -3.611439E+005

.
.
The value of the likelihood function at iteration 17 = -3.611431E+005
The value of the likelihood function at iteration 18 = -3.611431E+005
The value of the likelihood function at iteration 19 = -3.611431E+005
The value of the likelihood function at iteration 20 = -3.611429E+005
Iterations stopped due to small change in likelihood function

***** ITERATION 21 *****

Sigma_squared = 31854585.26541

Tau

| | | | | | |
|-------------|---------------|---------------|---------------|---------------|---------------|
| INTRCPT1,B0 | 1219466.20600 | -304207.40346 | -552492.36393 | 12649.97296 | -147492.38754 |
| FEMME,B1 | -304207.40346 | 257793.41237 | 371356.72305 | -273057.28652 | 6031.73569 |
| EDUC_SUP,B2 | -552492.36393 | 371356.72305 | 3450209.24773 | 273606.00405 | 25780.91851 |
| EDUC_MID,B3 | 12649.97296 | -273057.28652 | 273606.00405 | 1184513.85674 | 11092.64314 |
| AGE_16,B4 | -147492.38754 | 6031.73569 | 25780.91851 | 11092.64314 | 24798.14753 |

Tau (as correlations)

| | | | | | |
|-------------|--------|--------|--------|--------|--------|
| INTRCPT1,B0 | 1.000 | -0.543 | -0.269 | 0.011 | -0.848 |
| FEMME,B1 | -0.543 | 1.000 | 0.394 | -0.494 | 0.075 |
| EDUC_SUP,B2 | -0.269 | 0.394 | 1.000 | 0.135 | 0.088 |
| EDUC_MID,B3 | 0.011 | -0.494 | 0.135 | 1.000 | 0.065 |
| AGE_16,B4 | -0.848 | 0.075 | 0.088 | 0.065 | 1.000 |

Random level-1 coefficient **Reliability** estimate

| | |
|--------------|-------|
| INTRCPT1, B0 | 0.884 |
| FEMME, B1 | 0.717 |
| EDUC_SUP, B2 | 0.873 |
| EDUC_MID, B3 | 0.870 |
| AGE_16, B4 | 0.911 |

The value of the likelihood function at iteration 21 = -3.611429E+005

The outcome variable is GAINS

Final estimation of fixed effects:

| Fixed Effect | Coefficient | Standard Error | T-ratio | Approx. d.f. | P-value |
|------------------------|---------------|----------------|---------|--------------|---------|
| For INTRCPT1, B0 | | | | | |
| INTRCPT2, G00 | 372.358726 | 313.482564 | 1.188 | 12 | 0.258 |
| CHOMEURS, G01 | -24647.185539 | 3625.803001 | -6.798 | 12 | 0.000 |
| For FEMME slope, B1 | | | | | |
| INTRCPT2, G10 | -2261.695890 | 153.930530 | -14.693 | 12 | 0.000 |
| CHOMEURS, G11 | 13271.556867 | 2702.379595 | 4.911 | 12 | 0.000 |
| For EDUC_SUP slope, B2 | | | | | |
| INTRCPT2, G20 | 3210.525675 | 529.232309 | 6.066 | 12 | 0.000 |
| ALLOCHOM, G21 | 1.364661 | 0.734160 | 1.859 | 12 | 0.087 |
| For EDUC_MID slope, B3 | | | | | |
| INTRCPT2, G30 | 971.261658 | 310.074169 | 3.132 | 13 | 0.008 |
| For AGE_16 slope, B4 | | | | | |
| INTRCPT2, G40 | 674.950786 | 43.999174 | 15.340 | 12 | 0.000 |
| TRFSOCX, G41 | 0.039233 | 0.009843 | 3.986 | 12 | 0.002 |

The outcome variable is GAINS

Final estimation of fixed effects
(with **robust** standard errors)

| Fixed Effect | Coefficient | Standard Error | T-ratio | Approx. d.f. | P-value |
|------------------------|---------------|----------------|---------|--------------|---------|
| For INTRCPT1, B0 | | | | | |
| INTRCPT2, G00 | 372.358726 | 296.950777 | 1.254 | 12 | 0.234 |
| CHOMEURS, G01 | -24647.185539 | 3367.791034 | -7.319 | 12 | 0.000 |
| For FEMME slope, B1 | | | | | |
| INTRCPT2, G10 | -2261.695890 | 141.068142 | -16.033 | 12 | 0.000 |
| CHOMEURS, G11 | 13271.556867 | 2373.958655 | 5.590 | 12 | 0.000 |
| For EDUC_SUP slope, B2 | | | | | |
| INTRCPT2, G20 | 3210.525675 | 493.039388 | 6.512 | 12 | 0.000 |
| ALLOCHOM, G21 | 1.364661 | 0.550297 | 2.480 | 12 | 0.029 |
| For EDUC_MID slope, B3 | | | | | |
| INTRCPT2, G30 | 971.261658 | 296.176416 | 3.279 | 13 | 0.006 |
| For AGE_16 slope, B4 | | | | | |
| INTRCPT2, G40 | 674.950786 | 42.067770 | 16.044 | 12 | 0.000 |
| TRFSOCX, G41 | 0.039233 | 0.008012 | 4.897 | 12 | 0.000 |

The **robust** standard errors are appropriate for datasets having a moderate to large number of level 2 units. These data do **not** meet this criterion.

Final estimation of variance components: [les écarts-types estimés des composantes de la variance ne sont pas fournis ici, pas plus que les estimations des covariances et de leurs écarts-types. Ces informations se trouvent en principe dans le fichier annexe TAUVC.dat, mais nous ne les y avons pas vues]

| Random Effect | | Standard Deviation | Variance Component | df | Chi-square | P-value |
|-----------------|----|--------------------|--------------------|----|------------|---------|
| INTRCPT1, | U0 | 1104.29444 | 1219466.20600 | 12 | 262.39511 | 0.000 |
| FEMME slope, | U1 | 507.73360 | 257793.41237 | 12 | 46.60145 | 0.000 |
| EDUC_SUP slope, | U2 | 1857.47389 | 3450209.24773 | 12 | 113.27798 | 0.000 |
| EDUC_MID slope, | U3 | 1088.35374 | 1184513.85674 | 13 | 188.35067 | 0.000 |
| AGE_16 slope, | U4 | 157.47428 | 24798.14753 | 12 | 362.69652 | 0.000 |
| level-1, | R | 5643.98665 | 31854585.26541 | | | |

Statistics for current covariance components model

Deviance = 722285.751685

Number of estimated parameters = 16

Test of homogeneity of level-1 variance

Chi-square statistic = 2824.29313

Number of degrees of freedom = 13

P-value = 0.000

+++++

Exploratory Analysis: estimated level-2 coefficients and their standard errors obtained by regressing EB residuals on level-2 predictors selected for possible inclusion in subsequent HLM runs

| Level-1 Coefficient | Potential Level-2 Predictors | | | | |
|---------------------|------------------------------|----------|-----------|-----------|---------|
| | ALLOCHOM | ALLOCFAM | ETUDIANT | TRFSOXC | |
| INTRCPT1,B0 | | | | | |
| Coefficient | -0.101 | -0.315 | 59.427 | -0.163 | |
| Standard Error | 0.424 | 0.312 | 4027.404 | 0.204 | |
| t value | -0.239 | -1.009 | 0.015 | -0.799 | |
| | ALLOCHOM | ALLOCFAM | ETUDIANT | TRFSOXC | |
| FEMME,B1 | | | | | |
| Coefficient | -0.235 | -0.188 | 1051.843 | -0.094 | |
| Standard Error | 0.179 | 0.136 | 1786.086 | 0.090 | |
| t value | -1.316 | -1.383 | 0.589 | -1.046 | |
| | CHOMEURS | ALLOCFAM | ETUDIANT | TRFSOXC | |
| EDUC_SUP,B2 | | | | | |
| Coefficient | -16548.053 | 0.194 | -8706.040 | -0.520 | |
| Standard Error | 9751.955 | 0.531 | 6115.049 | 0.309 | |
| t value | -1.697 | 0.366 | -1.424 | -1.683 | |
| | CHOMEURS | ALLOCHOM | ALLOCFAM | ETUDIANT | TRFSOXC |
| EDUC_MID,B3 | | | | | |
| Coefficient | -6033.015 | 0.856 | 0.631 | -4578.577 | 0.211 |
| Standard Error | 6328.419 | 0.342 | 0.266 | 3771.266 | 0.198 |
| t value | -0.953 | 2.502 | 2.372 | -1.214 | 1.066 |
| | CHOMEURS | ALLOCHOM | ALLOCFAM | ETUDIANT | |
| AGE_16,B4 | | | | | |
| Coefficient | -989.031 | 0.037 | 0.075 | -148.274 | |
| Standard Error | 918.407 | 0.061 | 0.042 | 583.984 | |
| t value | -1.077 | 0.613 | 1.791 | -0.254 | |

A residual file, called residus_ttesCOV_avec_cste, has been created. Note, some statistics could not be computed and a value of -99 has been entered. These should be recoded to 'missing values' before any analyses are performed.

tauvc.dat, containing tau has been created.

gamvc.dat, containing the variance-covariance matrix of gamma has been created.

gamvcr.dat, containing the robust variance-covariance matrix of gamma has been created.

The above files have been created with a (nE15.7,1X) format.

WHLM: hlm2 SSM File: HLM5_JA1.SSM Command File: ttesCOV_avec_cste.hlm

File Basic Specifications Optional Specifications Run Analysis Help

| Level-2 vars | LEVEL 1 MODEL (bold: group-mean centering; bold italic: grand-mean centering) |
|---------------------|---|
| LEVEL-1 VARS | GAINS = $\beta_0 + \beta_1(\text{FEMME}) + \beta_2(\text{EDUC_SUP}) + \beta_3(\text{EDUC_MID}) + \beta_4(\text{AGE_16}) + r$ |
| INTRCPT1 | LEVEL 2 MODEL (bold italic: grand-mean centering) |
| GAINS | <input checked="" type="checkbox"/> Error term for currently selected level-2 equation |
| FEMME | <input checked="" type="checkbox"/> $\beta_0 = \gamma_{00} + \gamma_{01}(\text{CHOMEURS}) + u_0$ |
| EDUC_SUP | <input type="checkbox"/> $\beta_1 = \gamma_{10} + \gamma_{11}(\text{CHOMEURS}) + u_1$ |
| EDUC_MID | <input type="checkbox"/> $\beta_2 = \gamma_{20} + \gamma_{21}(\text{ALLOCHOM}) + u_2$ |
| AGE_16 | <input type="checkbox"/> $\beta_3 = \gamma_{30} + u_3$ |
| SEXÉ | <input type="checkbox"/> $\beta_4 = \gamma_{40} + \gamma_{41}(\text{TRFSOCX}) + u_4$ |
| EDUC | |
| IDJEUNE | |

2. MLwiN

On présente ici successivement l'image de l'écran de choix du modèle à estimer puis des copies d'écran des résultats (puisque MLwiN n'offre malheureusement aucun fichier de résultat ou listing rappelant les options du modèle et fournissant les résultats sous une forme susceptible d'être arrangée au moyen d'un traitement de texte).

On ne présente pas ici le mode de constitution des fichiers de données (qui est assez astreignant, les noms de variables devant être tapés les uns après les autres, ainsi que toutes les modalités des variables qualitatives ; et si on se trompe, on est bon pour tout recommencer) ni les nombreux écrans de spécifications (ex.: fenêtre concernant les pondérations, dont le mécanisme de transformation en pondérations standardisées est mystérieux : on s'attendrait à ce que les poids standardisés soient liés linéairement aux poids bruts ; or ce n'est pas le cas : la standardisation serait-elle opérée pays par pays, sans le dire ?).

MLwiN - D:_CD_echp\ephemere\MLwiN_eq_ttesCOV_avec_cste.ws

File Edit Options Model Estimation Data Manipulation Basic Statistics Graphs Window Help

Start More Stop RIGLS Estimation control..

Equations

$gains_{ij} \sim N(XB, \Omega)$
 $gains_{ij} = \beta_{0ij}cste + \beta_{1j}age_16_{ij} + \beta_{2j}femme_{ij} + \beta_{3j}educ_mid_{ij} + \beta_{4j}educ_sup_{ij} + \beta_{5j}chomeur_j + \beta_{6j}agetrf_{ij} + \beta_{7j}femchom_{ij} + \beta_{8j}supalch_{ij}$
 $\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$
 $\beta_{1j} = \beta_1 + u_{1j}$
 $\beta_{2j} = \beta_2 + u_{2j}$
 $\beta_{3j} = \beta_3 + u_{3j}$
 $\beta_{4j} = \beta_4 + u_{4j}$

Weights

| Level | Raw weights in | Standardised weight to |
|---------------------|-----------------|------------------------|
| 2: idcode =codepays | [equal weights] | c199 |
| 1: idcode =idjeune | poids | c200 |

Weighting mode

Off
 Use raw weights
 Use standardised weights

NOTE : sandwich estimators will be used for standard errors

Done Help

$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & & & & \\ \sigma_{u10} & \sigma_{u1}^2 & & & \\ \sigma_{u20} & \sigma_{u21} & \sigma_{u2}^2 & & \\ \sigma_{u30} & \sigma_{u31} & \sigma_{u32} & \sigma_{u3}^2 & \\ \sigma_{u40} & \sigma_{u41} & \sigma_{u42} & \sigma_{u43} & \sigma_{u4}^2 \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} \sigma_{e0}^2 \end{bmatrix}$$

-2*loglikelihood(IGLS) = 719460,300(35906 of 35906 cases in use)

Fonts Subs Name + - Add Term Estimates Nonlinear Help Clear

random 0% fixed 0% iteration 11

MLwiN - D:\CD_echp\ephemere\MLwiN_eq_ttesCOV_avec_cste.ws

File Edit Options Model Estimation Data Manipulation Basic Statistics Graphs Window Help

Start More Stop RIGLS Estimation control..

Equations

gains_{ij} ~ N(XB, Ω)

$$\text{gains}_{ij} = \beta_{0i} \text{cste} + \beta_{1i} \text{age}_{16ij} + \beta_{2i} \text{femme}_{ij} + \beta_{3i} \text{educ}_{\text{mid}ij} + \beta_{4i} \text{educ}_{\text{sup}ij} + -23802,350(3464,413) \text{chomeur}_{ij} + 0,042(0,009) \text{agetrf}_{ij} + 13000,930(2687,577) \text{femchom}_{ij} + 1,146(0,667) \text{supalch}_{ij}$$

$$\beta_{0i} = 363,204(296,336) + u_{0i} + e_{0i}$$

$$\beta_{1i} = 670,796(43,263) + u_{1i}$$

$$\beta_{2i} = -2220,463(149,299) + u_{2i}$$

$$\beta_{3i} = 1023,907(319,434) + u_{3i}$$

$$\beta_{4i} = 3306,992(561,067) + u_{4i}$$

$$\begin{bmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \\ u_{3i} \\ u_{4i} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 1155146,000(464323,100) & & & & \\ -144616,800(62589,290) & 25136,980(9898,880) & & & \\ -247473,200(180359,900) & -2029,090(24104,520) & 261416,400(117226,400) & & \\ 102437,000(354814,000) & -1847,348(51689,000) & -208991,700(186383,900) & 1346316,000(539463,700) & \\ -461536,500(633512,000) & 11519,450(90788,210) & 538233,400(343914,900) & 335085,300(677991,300) & 4184237,000(1664583,000) \end{bmatrix}$$

$$[e_{0i}] \sim N(0, \Omega_e) : \Omega_e = [28765900,000(214895,400)]$$

-2*loglikelihood(IGLS) = 719460,300(35906 of 35906 cases in use)

iteration 10

Démarrer | JC_RAY_communic... | hlm2_ttesCOV_ave... | MLwiN - D:\... | Document - WordPad | 19:55

3. SAS PROC MIXED

On présente successivement le code SAS puis le listing des résultats, et ceci d'abord sans imposer au modèle une structure de variances-covariances de type "composantes de la variance", impliquant la nullité de toutes les covariances (ce choix d'une matrice non structurée impose de recourir à la méthode MIVQUE0 puisque les méthodes ML et REML ne donnent pas de résultats d'estimations dans ce cas). Puis, afin d'obtenir des résultats par la méthode REML, on impose que les covariances soient toutes nulles.

3.1 Modèle ne contraignant pas les covariances à être toutes nulles

```
proc mixed data=fmr.gains41 noITprint noCLprint COVtest METHOD=MIVQUE0 ;
class pays;
model JA_gainsactiv = C_chomeur
                    age_moins_16 age_moins_16*C_TRF_SOCX
                    femme femme*C_chomeur
                    educ_mid
educ_sup educ_sup*C_allochom / solution notest DDFM=BW;
random intercept age_moins_16 femme educ_mid educ_sup / subject=pays TYPE=UN;
weight poiJA_UE;
run;title;
```

The Mixed Procedure

Model Information

| | |
|---------------------------|----------------|
| Data Set | FMR.GAINS41 |
| Dependent Variable | JA_gainsactiv |
| Weight Variable | poiJA_UE |
| Covariance Structure | Unstructured |
| Subject Effect | pays |
| Estimation Method | MIVQUE0 |
| Residual Variance Method | Profile |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Between-Within |

Dimensions

| | |
|--------------------------|-------|
| Covariance Parameters | 16 |
| Columns in X | 9 |
| Columns in Z Per Subject | 5 |
| Subjects | 14 |
| Max Obs Per Subject | 4580 |
| Observations Used | 35906 |
| Observations Not Used | 0 |
| Total Observations | 35906 |

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate | Standard Error | Z Value | Pr Z |
|----------|---------|----------|----------------|---------|--------|
| UN(1,1) | pays | 1581891 | 1808503 | 0.87 | 0.1909 |
| UN(2,1) | pays | -141703 | 93912 | -1.51 | 0.1313 |
| UN(2,2) | pays | 135845 | 1014.90 | 133.85 | <.0001 |
| UN(3,1) | pays | -261310 | 724990 | -0.36 | 0.7185 |
| UN(3,2) | pays | 18600 | 81177 | 0.23 | 0.8188 |
| UN(3,3) | pays | 220163 | 244890 | 0.90 | 0.1843 |
| UN(4,1) | pays | -776935 | 1149466 | -0.68 | 0.4991 |
| UN(4,2) | pays | 54166 | 404.68 | 133.85 | <.0001 |
| UN(4,3) | pays | 106406 | 468850 | 0.23 | 0.8205 |
| UN(4,4) | pays | 725072 | 744172 | 0.97 | 0.1649 |
| UN(5,1) | pays | -1128884 | 923095 | -1.22 | 0.2214 |
| UN(5,2) | pays | 121703 | 909.25 | 133.85 | <.0001 |
| UN(5,3) | pays | -58022 | 663671 | -0.09 | 0.9303 |
| UN(5,4) | pays | 784438 | 576581 | 1.36 | 0.1737 |
| UN(5,5) | pays | 1482472 | 11076 | 133.85 | <.0001 |
| Residual | | 32052008 | 239460 | 133.85 | <.0001 |

Fit Statistics

| | |
|--------------------------|----------|
| -2 Res Log Likelihood | 746154.7 |
| AIC (smaller is better) | 746186.7 |
| AICC (smaller is better) | 746186.7 |
| BIC (smaller is better) | 746196.9 |

Null Model Likelihood Ratio Test

| DF | Chi-Square | Pr > ChiSq |
|----|------------|------------|
| 15 | 655.90 | <.0001 |

Solution for Fixed Effects
Standard

| Effect | Estimate | Error | DF | t Value | Pr > t |
|----------------------|----------|---------|------|---------|---------|
| Intercept | 331.05 | 352.43 | 12 | 0.94 | 0.3661 |
| C_chomeur | -25059 | 5356.30 | 12 | -4.68 | 0.0005 |
| age_moins_16 | 678.97 | 99.2526 | 36E3 | 6.84 | <.0001 |
| age_moins*C_TRF_SOCX | .07643 | 0.06827 | 36E3 | 1.12 | 0.2630 |
| femme | -2212.20 | 150.45 | 36E3 | -14.70 | <.0001 |
| C_chomeur*femme | 10876 | 2913.98 | 36E3 | 3.73 | 0.0002 |
| educ_mid | 957.01 | 251.37 | 36E3 | 3.81 | 0.0001 |
| educ_sup | 2972.83 | 363.56 | 36E3 | 8.18 | <.0001 |
| educ_sup*C_allochom | 1.0216 | 0.3673 | 36E3 | 2.78 | 0.0054 |

3.2 Modèle contraignant les covariances à être toutes nulles

```
proc mixed data=fmr.gains41 noITprint noCLprint COVtest METHOD=REML ;
class pays;
model JA_gainsactiv = C_chomeur
                    age_moins_16 age_moins_16*C_TRF_SOCX
                    femme femme*C_chomeur
                    educ_mid
                    educ_sup educ_sup*C_allochom / solution notest DDFM=BW;
random intercept age_moins_16 femme educ_mid educ_sup / solution subject=pays
TYPE=VC;
weight poiJA_UE;
run;title;
```

The Mixed Procedure

Model Information

| | |
|---------------------------|---------------------|
| Data Set | FMR.GAINS41 |
| Dependent Variable | JA_gainsactiv |
| Weight Variable | poiJA_UE |
| Covariance Structure | Variance Components |
| Subject Effect | pays |
| Estimation Method | REML |
| Residual Variance Method | Profile |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Between-Within |

Dimensions

| | |
|--------------------------|-------|
| Covariance Parameters | 6 |
| Columns in X | 9 |
| Columns in Z Per Subject | 5 |
| Subjects | 14 |
| Max Obs Per Subject | 4580 |
| Observations Used | 35906 |
| Observations Not Used | 0 |
| Total Observations | 35906 |

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate | Standard Error | Z Value | Pr Z |
|--------------|---------|----------|----------------|---------|--------|
| Intercept | pays | 967929 | 435505 | 2.22 | 0.0131 |
| age_moins_16 | pays | 20479 | 9165.33 | 2.23 | 0.0127 |
| femme | pays | 209495 | 128266 | 1.63 | 0.0512 |
| educ_mid | pays | 1109908 | 512260 | 2.17 | 0.0151 |
| educ_sup | pays | 3338525 | 1726481 | 1.93 | 0.0266 |
| Residual | | 31907701 | 238387 | 133.85 | <.0001 |

Fit Statistics

| | |
|--------------------------|----------|
| -2 Res Log Likelihood | 746113.2 |
| AIC (smaller is better) | 746125.2 |
| AICC (smaller is better) | 746125.2 |
| BIC (smaller is better) | 746129.0 |

Solution for Fixed Effects

| Effect | Estimate | Standard Error | DF | t Value | Pr > t |
|----------------------|----------|----------------|------|---------|---------|
| Intercept | 343.10 | 282.72 | 12 | 1.21 | 0.2483 |
| C_chomeur | -16917 | 6339.78 | 12 | -2.67 | 0.0205 |
| age_moins_16 | 675.14 | 40.2955 | 36E3 | 16.75 | <.0001 |
| age_moins*C_TRF_SOCX | 0.08176 | 0.02933 | 36E3 | 2.79 | 0.0053 |
| femme | -2226.20 | 148.59 | 36E3 | -14.98 | <.0001 |
| C_chomeur*femme | 11566 | 3207.41 | 36E3 | 3.61 | 0.0003 |
| educ_mid | 985.75 | 302.64 | 36E3 | 3.26 | 0.0011 |
| educ_sup | 3209.97 | 523.20 | 36E3 | 6.14 | <.0001 |
| educ_sup*C_allochom | 1.4528 | 0.8008 | 36E3 | 1.81 | 0.0697 |

Solution for Random Effects

| Effect | pays | Estimate | Std Err Pred | DF | t Value | Pr > t |
|--------------|------|----------|--------------|------|---------|---------|
| Intercept | AL | 446.86 | 379.36 | 36E3 | 1.18 | 0.2388 |
| age_moins_16 | AL | -64.7536 | 53.2562 | 36E3 | -1.22 | 0.2240 |
| femme | AL | -263.42 | 215.32 | 36E3 | -1.22 | 0.2212 |
| educ_mid | AL | 527.61 | 340.38 | 36E3 | 1.55 | 0.1211 |
| educ_sup | AL | 1402.06 | 618.19 | 36E3 | 2.27 | 0.0233 |
| Intercept | AU | 1242.38 | 562.37 | 36E3 | 2.21 | 0.0272 |
| age_moins_16 | AU | -132.19 | 66.4170 | 36E3 | -1.99 | 0.0466 |
| femme | AU | -366.56 | 330.47 | 36E3 | -1.11 | 0.2674 |
| educ_mid | AU | 1297.63 | 509.00 | 36E3 | 2.55 | 0.0108 |
| educ_sup | AU | 226.56 | 1046.01 | 36E3 | 0.22 | 0.8285 |
| (...) | | | | | | |
| Intercept | UK | 53.0423 | 442.77 | 36E3 | 0.12 | 0.9046 |
| age_moins_16 | UK | 168.64 | 44.5577 | 36E3 | 3.78 | 0.0002 |
| femme | UK | -410.66 | 247.39 | 36E3 | -1.66 | 0.0969 |
| educ_mid | UK | -191.70 | 374.08 | 36E3 | -0.51 | 0.6083 |
| educ_sup | UK | 307.01 | 772.40 | 36E3 | 0.40 | 0.6910 |

Annexe C : Résultats complémentaires obtenus avec HLM5, MlwiN et SAS

Contenu de cette annexe :

1. L'impact, sur les résultats, de certaines observations influentes (en l'occurrence, le cas de deux jeunes adultes néerlandais)
2. Tests de l'hypothèse de normalité des (différents types de) résidus
3. Le test de l'hypothèse de l'homogénéité, entre pays, de la variance au niveau 1 et la modélisation de l'hétérogénéité des variances.

1. Repérage d'observations influentes, leur impact sur les résultats, les nouveaux résultats une fois ces observations écartées

C'est ici MlwiN qui est utilisé, à cause de ses facilités graphiques. MlwiN propose en effet des tests graphiques permettant de voir si l'hypothèse de normalité des résidus est tenable au niveau 1 et au niveau 2.

1.1 Au niveau des individus, on peut se servir des **35906 résidus de niveau 1** (ou de leur version standardisée) pour repérer des **points écartés** du reste du nuage (*outliers*). En l'occurrence, deux individus sont situés complètement à l'écart des autres. MlwiN permet de les identifier commodément grâce à la souris : en cliquant sur un de ces points, on affiche en effet leur identifiant individuel et leur pays (cf., dans la copie d'écran infra, la petite fenêtre concernant le point ayant le plus gros résidu).

Grâce aux options offertes par MlwiN, il est alors assez aisé de **faire tourner à nouveau le modèle sans ces deux individus atypiques** puisqu'ils ont beaucoup d'influence sur les valeurs des estimateurs, nous disent les statistiques d'influence. De fait, en ôtant les individus 27541 et 28323 (un homme et une femme des Pays-Bas), les coefficients estimés bougent souvent d'environ 1%, et jusqu'à plus de 3% (pour le coefficient de la variable CHOMEURS dans l'équation du coefficient aléatoire de FEMME : on passe de 12600 à 13000) ; et les variances changent nettement parfois, celle de la constante aléatoire chutant de 15%, celle du coefficient aléatoire de FEMME de 8%, etc.

1.2 Au niveau des pays, on pourrait se servir des **résidus de niveau 2** pour mener des analyses semblables à celles menées au niveau 1. De fait, le Danemark (pays #2) apparaît comme ayant un effet de levier (*leverage*) sensiblement plus important (0,10) que les autres pays, mais les Pays-Bas (pays #3) ont des résidus moyens si extrêmes que, malgré un moindre effet de levier (<0,08) ils ont une influence (i.e. un indicateur combinant les deux dimensions précédentes) bien plus forte que les autres pays (0,7 contre 0 à 0,4 pour les autres pays). Mais, vu le petit nombre de pays dont nous disposons, nous sommes moins enclin à écarter un pays que ne peuvent l'être les chercheurs qui travaillent sur un millier d'écoles.

2. Tests de l'hypothèse de normalité des (différents types de) résidus

2.1. MIwiN propose des tests graphiques pour voir si l'hypothèse de normalité des résidus est tenable au niveau 1 et au niveau 2.

2.1.1 Au niveau des individus, on peut obtenir l'histogramme des **35906 résidus de niveau 1** (ou de leur version standardisée) afin d'examiner l'allure de cette distribution. On peut aussi (cf. copie d'écran infra) obtenir une droite de Henry : ici il apparaît que si la distribution est relativement gaussienne en général, elle ne l'est plus du tout pour les gains élevés, surtout pour deux cas atypiques.

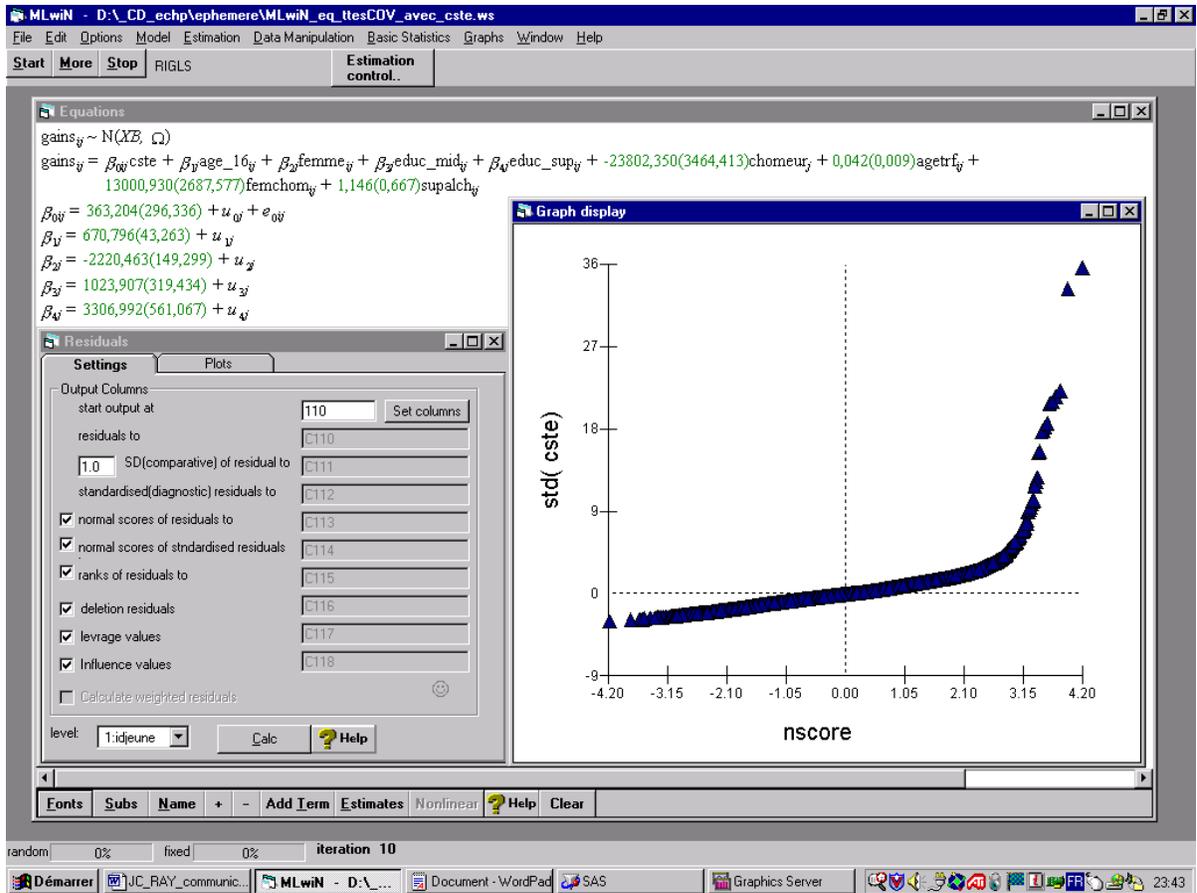
On peut aussi croiser ces résidus avec telle ou telle variable de niveau 1, l'âge par exemple, pour voir si la **variance des résidus de niveau 1 est homogène** ; en l'occurrence, comme le montre la copie d'écran suivante, la dispersion des résidus standardisés de niveau 1 croît nettement avec l'âge.

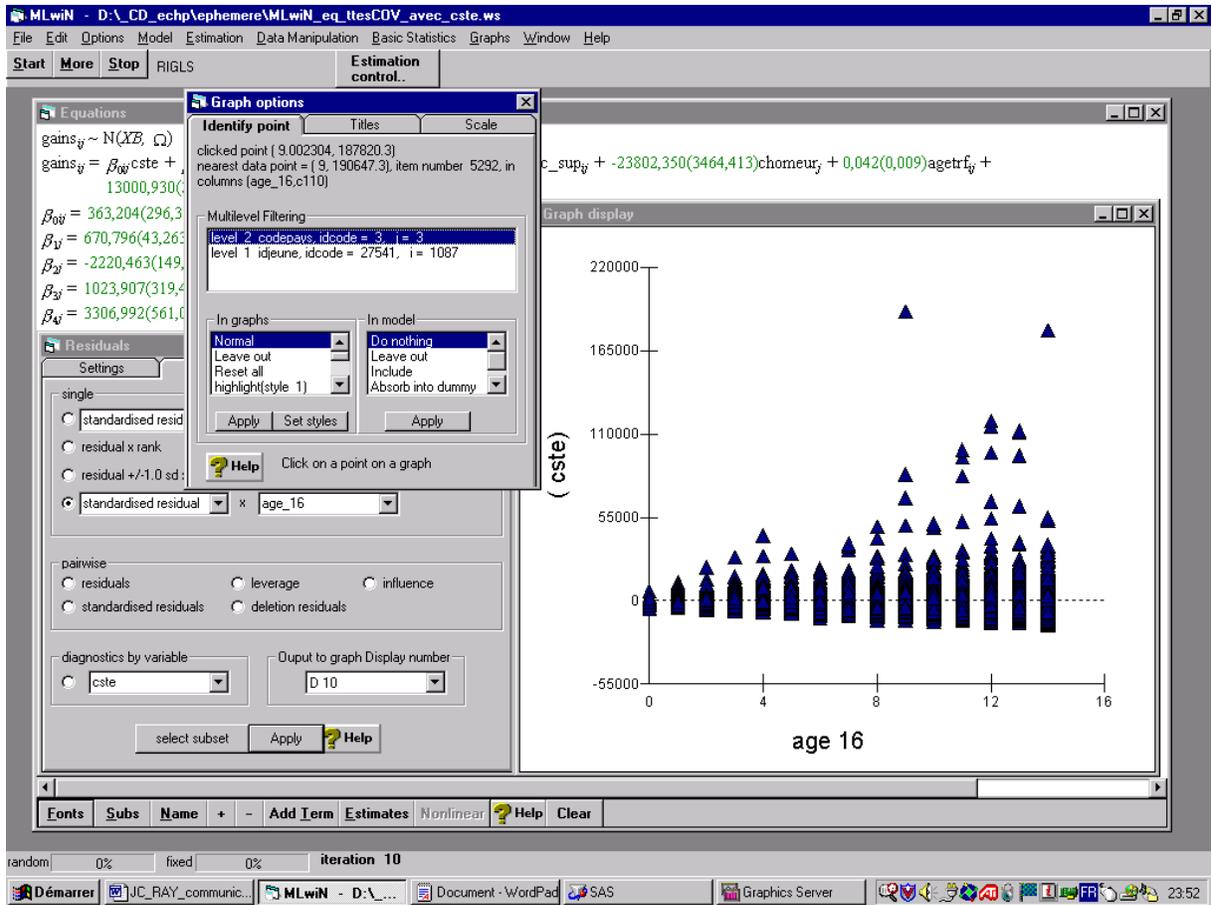
2.1.2. Au niveau des pays, on n'a plus une distribution unique comme au niveau 1 mais autant de distributions que de facteurs X (+ la constante) à coefficient supposé aléatoire. MIwiN affiche pour chaque distribution un graphique comparant les centiles (interpolés) de la distribution empirique de chaque résidu avec les centiles de la distribution normale. On s'aperçoit (cf. copie d'écran infra) que, ici, l'écart entre la droite de Henry et la distribution empirique est sensible, notamment pour le coefficient de FEMME.

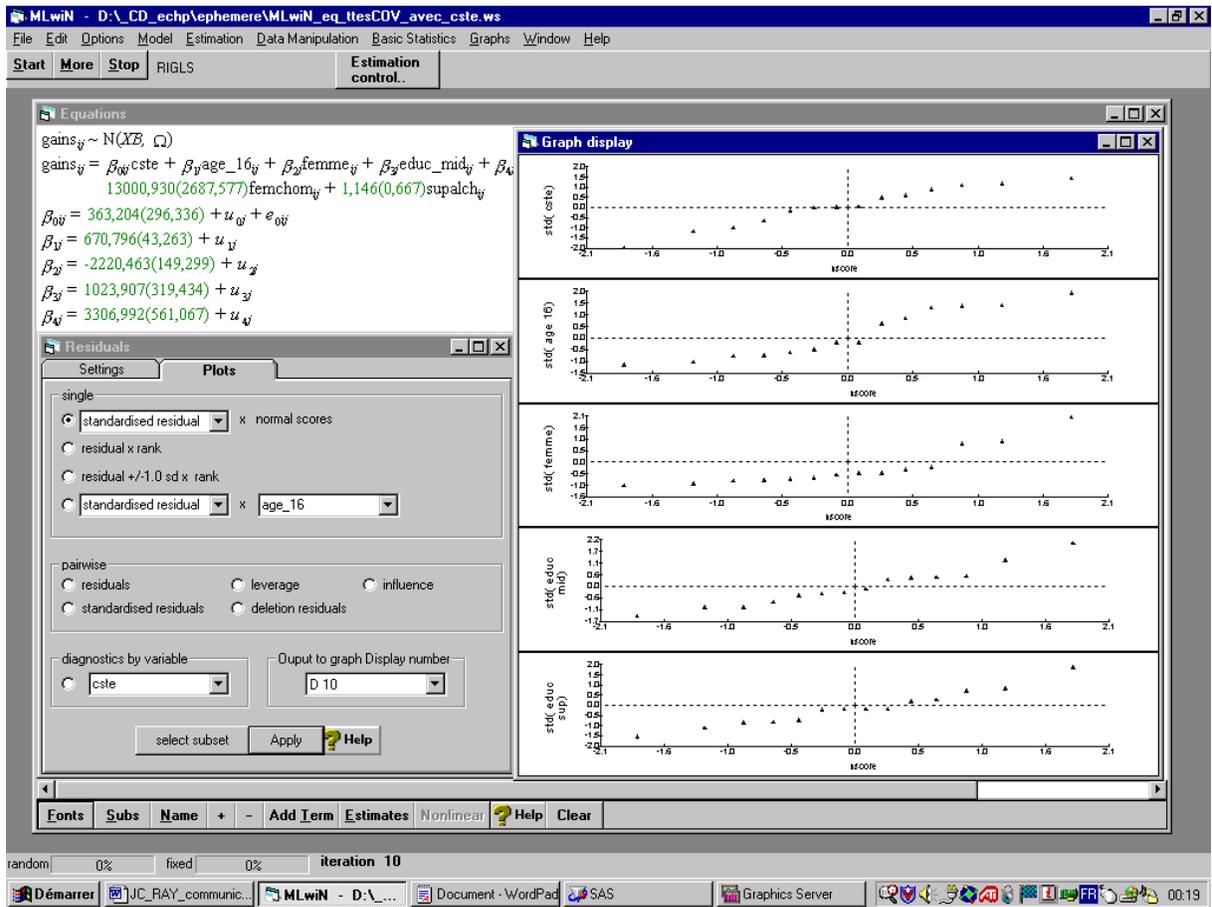
2.2. HLM5 ne semble pas proposer grand chose d'autre qu'un fichier des résidus, qu'il faut traiter avec un autre logiciel, SAS par exemple. Ce fichier de résidus étant inutilisable (à cause d'un bogue) si nous ne transformons pas les montants financiers en milliers de SPA, c'est ce que nous avons fait sous SAS avant de construire un nouveau fichier SSM sous HLM5, de faire tourner le modèle multi-niveaux analogue au précédent et de récupérer un fichier de résidus effectivement exploitable sous SAS.

Avec, notamment, PROC GPLOT et PROC UNIVARIATE, SAS permet en effet d'analyser les résidus créés par HLM5, après standardisation éventuelle via PROC STANDARD. Ici, le test de SHAPIRO-WILK offert par PROC UNIVARIATE, appliqué à la variable MDRSVAR du fichier de résidus de HLM5 (variable qui représente l'écart-type résiduel par pays) donne une probabilité de 91,6% (on devrait donc conclure que MDRSVAR suit une loi de Gauss). En fait, le nombre d'observations étant tout petit (uniquement 14 groupes de niveau 2), il est possible que, même avec une valeur de p aussi grande, la distribution des écarts-types résiduels ne soit pas gaussienne.

2.3. SAS propose naturellement, pour chaque modèle estimé par PROC MIXED, ses propres fichiers de résidus (option OUTPRED=nom-de-fichier de l'instruction MODEL).







3. Le test de l'hypothèse de l'homogénéité, entre pays, de la variance au niveau 1, et la modélisation de l'hétérogénéité des variances

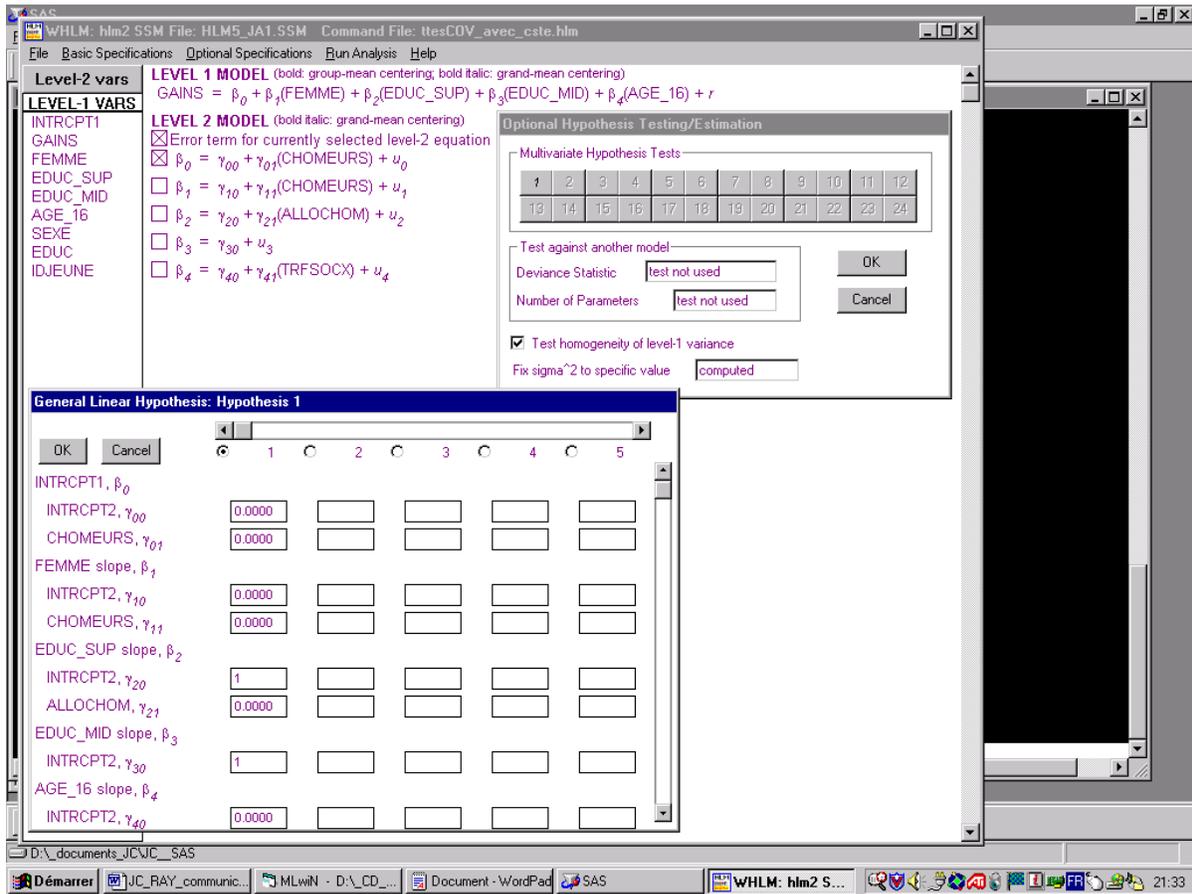
C'est ici HLM5 qui est utilisé.

Pour obtenir le test d'homogénéité des variances de niveau 1, il suffit (voir, infra, la copie d'écran) de cocher la case "Test homogeneity of level-1 variance" dans l'écran "Optional hypothesis test/estimation".

On obtient ceci :

```
Test of homogeneity of level-1 variance
-----
Chi-square statistic      = 2824.29313
Number of degrees of freedom = 13
P-value                  = 0.000
```

*La conclusion est ici que l'on peut, sans grand risque de se tromper, rejeter l'hypothèse nulle. La variance résiduelle (i.e. la variance des gains au sein de chaque pays une fois que l'on a contrôlé les effets des X) est donc significativement différente selon les pays. On pourrait donc utilement tenter de **modéliser l'hétérogénéité des variances par pays**, en les expliquant par des variables de niveau 1, pour voir, par exemple, si les gains des jeunes garçons (quand on contrôle âge et éducation) ne seraient pas non seulement supérieurs en moyenne à ceux des jeunes filles/femmes, mais également plus (ou moins) dispersés. Pour tester cela, on pourrait modéliser la variance résiduelle comme étant fonction d'une variable de niveau 1 (par exemple le sexe) et d'une constante. En l'occurrence, HLM5 propose pour cela une fonction exponentielle, c'est-à-dire un modèle log-linéaire (puisque'il s'agit de modéliser une grandeur nécessairement positive, la variance). Mais cette modélisation n'est pas possible avec HLM5 lorsque les données sont pondérées.*



Si le fichier utilisé n'était pas pondéré, il suffirait d'ouvrir la fenêtre "Heterogeneous sigma^2" via le menu "Optional hypothesis test/estimation", et de déplacer (par double clic) la variable FEMME de la partie gauche à la partie droite.

Heterogeneous sigma²: Predictors of level-1 variance

Double-click to move variables in and out of Predictors column

| Possible choices | Predictors of level-1 variance |
|---|--------------------------------|
| EDUC_SUP EDUC_MID AGE_16 SEXÉ EDUC IDJEUNE | FEMME |

OK

Iteration Control

| | | | |
|--------------------|--------------|--------------------|--------------|
| Macro Iterations | 100 | Micro Iterations | 100 |
| Stopping Criterion | 0.0001000000 | Stopping Criterion | 0.0000010000 |

Pour un **test d'égalité de coefficients fixes** (par contraste avec MlwiN, HLM5 ne semble pas offrir la possibilité de tester des combinaisons linéaires de coefficients, et donc, par exemple, l'égalité à 0 ou à 1 d'une somme de deux coefficients), il faut cliquer sur la case numérotée 1 de "Multivariate hypothesis tests", ce qui ouvre la fenêtre située en bas à gauche de la copie d'écran supra ; on remplace alors par un 1 les 0 qui figurent en regard des deux coefficients fixes dont on veut tester l'égalité. On obtient alors ceci :

Results of General Linear Hypothesis Testing

| | Coefficients | Contrast |
|------------------------|---------------|----------|
| For INTRCPT1, B0 | | |
| INTRCPT2, G00 | 372.358726 | 0.000 |
| CHOMEURS, G01 | -24647.185539 | 0.000 |
| For FEMME slope, B1 | | |
| INTRCPT2, G10 | -2261.695890 | 0.000 |
| CHOMEURS, G11 | 13271.556867 | 0.000 |
| For EDUC_SUP slope, B2 | | |
| INTRCPT2, G20 | 3210.525675 | 1.000 |
| ALLOCHOM, G21 | 1.364661 | 0.000 |
| For EDUC_MID slope, B3 | | |
| INTRCPT2, G30 | 971.261658 | 1.000 |
| For AGE_16 slope, B4 | | |
| INTRCPT2, G40 | 674.950786 | 0.000 |
| TRFSOCX, G41 | 0.039233 | 0.000 |

Chi-square statistic = 40.794634

Degrees of freedom = 1

P-value = 0.000001

Conclusion : G20 et G30 sont donc significativement différents.

Liste des publications du panel socio-économique des ménages "Liewen zu Lëtzebuerg"

- **Méthodologie générale et répertoire des variables** - Année d'enquête: 1985 (Première vague). Document PSELL n°1. P. DICKES, P. HAUSMAN, A. KERGER -1987.
- **L'état de la pauvreté au Grand-Duché de Luxembourg**. Document PSELL n° 2. F.COURTOIS, P.HAUSMAN (1987).
- **Description des niveaux de vie et de bien-être économique dans les ménages résidant au Luxembourg - Année 1985-1987**. Série "Niveau de vie " 1. Document PSELL n°3 P. HAUSMAN (1987).
- **Niveaux de vie et de bien-être économique des ménages en 1985: principaux résultats en 1985**. Série "Niveau de vie " 2. Document PSELL n°4. P. HAUSMAN (1987).
- **Un indicateur pour mesurer la pauvreté objective: théorie et application dans la première vague du panel socio-économique Luxembourgeois - 1985**. Document PSELL n°5, P. DICKES (1987).
- **Un indicateur pour mesurer la pauvreté subjective. Théorie et application dans la première vague du panel socio-économique luxembourgeois : année d'enquête 1985** (première vague). Document PSELL n°6 P. DICKES (1987).
- **Pratique de l'échelonnement multidimensionnel**. Document PSELL n°7 P. DICKES, J. TOURNOIS (1989).
- **Pauvreté et conditions d'existence: théories, modèles et mesures**. Document PSELL n°8. P.DICKES (1989).
- **Logistique & documentation - Principes d'organisation de la documentation dans le panel**. Document PSELL n°9. J. TOURNOIS (1988).
- **Documentation transversale des variables 1985: première vague**. Document PSELL n°10. J.TOURNOIS (1988).
- **Evolution d'un groupe de ménages pauvres entre 1985 et 1987**. Document PSELL n°11. A. WAGNER (1989).
- **Description statistique des variables du questionnaire -1986-** (deuxième vague). Document PSELL n°12. A.KERGER, R. DE WEVER (1988).
- **Activité féminine, isolement et prestations familiales: un premier parallèle Luxembourg- Lorraine**. Annexes. Document PSELL n° 13. J.C.RAY, B.JEANDIDIER, S.CARVOYEUR (1990).
- **Le mode d'échantillonnage du panel "Liewen zu Lëtzebuerg" - Bilan des deux premières vagues**. Document PSELL n°14. P.HAUSMAN (1990).
- **Analyse des données irlandaises (enquête pilote -1987) pour construire une échelle de pauvreté**. Document PSELL n°15. P.DICKES (1988).
- **La collecte des données en 1986 - Elaboration du questionnaire, déroulement de l'enquête, opérations de chiffrement**. Document PSELL n°16. A.KERGER (1989).
- **Organisation der Daten des Luxemburger Haushaltspansels. (Eingabe, Speicherung und Analyse von Paneldaten)**. Document PSELL n°17. G.SCHMAUS (1990) - (version anglaise: 17a).
- **MNDr, partition valuée selon la méthode de Roubens et Libert**. Document PSELL n°18. B.GAILLY (1989).
- **Dispositif des pondérations individuelles et des pondérations des ménages en 1985 et 1986**. Document PSELL n°19. B.GAILLY, P.HAUSMAN (1990).
- **Les modes de garde des jeunes enfants**. Document PSELL n° 20. A.AUBRUN, P.HAUSMAN (1990).
- **Les indicateurs sociaux de pauvreté: Tableaux de base et documentation**. Document PSELL n°21. P.HAUSMAN (1990).
- **Les personnes âgées et/ou retraitées au Luxembourg: leur environnement familial et leurs réseaux de solidarité**. Document PSELL n° 22. P.HAUSMAN, G.SCHABER (1991).
- **Examen des effets du phénomène d'attrition sur l'étude des revenus et de l'emploi - Années de références: 1985, 1986 et 1987**. Document PSELL n° 23 P.HAUSMAN, B.GAILLY (1990).
- **La constitution des fichiers de référence, nécessaire à l'étude du phénomène d'attrition**. Document PSELL n°24. R.DE WEVER (1990).

- **Dispositif des pondérations individuelles et des pondérations des ménages en 1985 et 1987.** Document PSELL n°25. B.GAILLY, P.HAUSMAN (1990).
- **Bilan de l'attrition au cours des trois premières vagues d'enquêtes: 1985/1986/1987.** Document PSELL n°26. B.GAILLY, P.HAUSMAN (1990).
- **Imputation des revenus manquants dans le panel socio-économique luxembourgeois.** Document PSELL n°27. P.HAUSMAN (1990).
- **"PSELLDOC" Système documentaire pour le panel Luxembourgeois.** Document PSELL n°28. J.J.WESTER, avec la collaboration de A.AUBRUN (1990).
- **Le déroulement de la collecte en 1987. Elaboration du questionnaire, déroulement de l'enquête, opération de chiffrage.** Document PSELL n°29. A.KERGER (1990).
- **La production des données: Vague 1988-1990.** Document PSELL n° 30. A.KERGER (1990).
- **Description statistique des variables du questionnaire -1987- Troisième vague.** Document PSELL n°31. R.DE WEVER, A.KERGER (1991).
- **Les ménages de retraités et les ménages d'actifs - Comparaison des niveaux de vie et des niveaux de dépenses.** Série "Niveau de vie " 3. Document PSELL n° 34. P.HAUSMAN, A.AUBRUN, A.KERGER (1991).
- **Situation der Arbeitslosen in Luxemburg.** Document PSELL n° 35. G.SCHMAUS(1991).
- **La recherche sur la pauvreté au Grand-Duché de Luxembourg.** Document PSELL 36. A.WAGNER (1990).
- **La distribution des revenus entre ménages en 1986 - Une comparaison de statistiques Luxembourg - Lorraine - Belgique.** Document PSELL n° 38. B.DELVAUX (1991).
- **Efficacité de la sécurité sociale dans la lutte contre la pauvreté.** Document PSELL n°39. P.HAUSMAN (1990).
- **Dispositif des pondérations individuelles et des pondérations des ménages de 1985 à 1988.** Document PSELL n°40. B.GAILLY (1991).
- **Projet de recommandation du conseil portant sur les critères communs relatifs à des ressources et prestations suffisantes dans les systèmes de protection sociale.** Rapport préparatoire relatif au dispositif de suivi. Document PSELL n° 42. J.C.BROWN, A.WAGNER avec la collaboration de: P. HAUSMAN, A.KERGER, G.MENARD (1991).
- **La loi sur le Revenu Minimum Garanti. Quelques avis du public.** Document PSELL n° 43. A.WAGNER (1992).
- **L'endettement des ménages au Luxembourg 1985 -1989.** Série "Mode de vie" - 1. Document PSELL n°45. B.GAILLY, P.HAUSMAN (1992).
- **Les femmes au Grand-Duché de Luxembourg: Démographie-Famille I.** Document PSELL n°46. P.HAUSMAN avec la collaboration de M. Langers (STATEC) et du Ministère de la Famille et de la Solidarité (1992).
- **Les femmes au Grand-Duché de Luxembourg: Revenus-Conditions de vie II.** Document PSELL n°47. P.HAUSMAN, J.VECERNIK avec la collaboration du Ministère de la Famille et de la Solidarité (1992).
- **Dispositif des pondérations individuelles et des pondérations des ménages de 1985 à 1989.** Document PSELL n°48. B.GAILLY (1992).
- **Luxembourg, 1985 à 1989, une vague d'endettement.** Série "Mode de vie" - 2. Document PSELL n°49. B.GAILLY, P.HAUSMAN (1993).
- **Les Effets de la Protection Sociale dans la Communauté Européenne.Contribution au programme "Convergence en matière de Protection Sociale"** (Commission C.E., DG V). Document PSELL n°50. P.HAUSMAN (1993) .
- **Situation démographique de l'Europe des Douze - EUR12.** Document PSELL n°51. P.HAUSMAN (1993).
- **Les phénomènes associés au vieillissement de la population.** Document PSELL N°52. P.HAUSMAN (1993).
- **Dispositif des pondérations individuelles et des pondérations des ménages de 1985 à 1990.** Document PSELL n°53. B.GAILLY (1993).
- **Insérer des nouveaux membres dans un panel longitudinal de ménages et d'individus: simulations.** Document PSELL n°54. B.GAILLY (CEPS/Insteat) avec la collaboration de P.LAVALLÉE (Statistics-Canada) (1993). (Version anglaise disponible).
- **Dispositif des pondérations individuelles et des pondérations des ménages de 1985 à 1991.** Tome II. Document PSELL n°55. B.GAILLY (1993).
- **Evolution générale du revenu des ménages: 1978-1990.** Série "Niveau de vie" 4. Document PSELL n°56. P.HAUSMAN (1994).

- **1985-1990. Endettement et risques de surendettement.** Série "Mode de vie" - 3. Document PSELL n°57. B.GAILLY (1994).
- **L'intégration sociale des personnes âgées au Grand-Duché de Luxembourg.** Document PSELL n°58. G.SCHABER, P.BOUSCH (1993).
- **Insérer un échantillon complémentaire dans un panel longitudinal de ménages et d'individus: simulations.** (2e partie). Document PSELL n°59. B.GAILLY, P.LAVALLÉE (1994).
- **Revenus et endettement.** Série "Mode de vie" - 4. Document PSELL n°60. B.GAILLY (1994).
- **Mesure de l'efficacité des transferts sociaux selon une approche en termes de sécurisation du niveau de vie des ménages: une analyse statique puis dynamique, appliquée au cas du Luxembourg et de la Lorraine.** Document PSELL n°61. B.JEANDIDIER & N.POUSSING (1994).
- **Budget temps des femmes: l'opinion des femmes.** Document PSELL n°62. A.AUBRUN (1994).
- **Dispositif des pondérations des individus et des ménages de 1985 à 1992.** Document PSELL n°63. B.GAILLY (1994).
- **1985-1991. L'endettement au Luxembourg. Diffusion et concentration.** Série "Mode de vie" - 5. Document PSELL n°64. B.GAILLY (1994).
- **La variance des estimateurs d'un panel ménage. La méthode des groupes aléatoires appliquée au panel luxembourgeois.** Document PSELL n°65. M.RIEBSCHLÄGER. (1994).
- **Etude comparative de l'efficacité de la protection sociale.** Document PSELL n°66. P.HAUSMAN (1994). (Version anglaise disponible).
- **Les politiques économiques et sociales et les personnes âgées au Grand-Duché de Luxembourg** (Rapport national pour la Commission des Communautés Européennes - Novembre 1992). Document PSELL n°67. G.SCHABER, P.BOUSCH.
- **La mise en oeuvre des politiques sociales et économiques en faveur des personnes âgées au Grand-Duché de Luxembourg** (Rapport national pour la Commission des Communautés Européennes et l'Observatoire Européen portant sur le vieillissement démographique et les personnes âgées). Document PSELL n°68. G.SCHABER, P.BOUSCH.
- **Actives, mais à quel prix?** Document PSELL n°69. B.LEJEALLE (1994).
- **Les Luxembourgeoises moins actives que leurs homologues européennes. Etude comparative de la structure des ménages et de l'activité féminine au Luxembourg et dans six autres pays européens -(Suède, Pays-Bas, Belgique, Allemagne de l'Ouest, Allemagne de l'Est, Danemark et Grande-Bretagne).** Document PSELL n°70. B.LEJEALLE (1994).
- **Les conditions de travail des personnes actives.** Document PSELL n°71. B.GAILLY (1995).
- **Etre au chômage au Luxembourg.** Document PSELL n°72. B.LEJEALLE (1995).
- **Les femmes au Grand-Duché de Luxembourg: 4. Place et rôle de la femme dans la société.** Document PSELL n°73. A.AUBRUN (1995).
- **Les femmes au Grand-Duché de Luxembourg: 3. Revenus-Conditions de vie.** Document PSELL n°74. P.HAUSMAN, en collaboration: J.VECERNIK et Ministère de la Famille et de la Solidarité (1995).
- **1985-1992. L'endettement au Luxembourg.** Série "Mode de vie" - 6. Document PSELL n°75. B.GAILLY (1995).
- **Les femmes au Grand-Duché de Luxembourg: 5. Encadrement institutionnel de la femme luxembourgeoise : Conditions juridiques - Politiques visant la famille - Mesures relatives à la formation et à l'emploi - Services d'aide.** Document PSELL n°76. M. PELS (1995).
- **Le revenu des ménages. Evolution de 1985 à 1992.** Série "Niveau de vie " 5.Document PSELL n°77. P.HAUSMAN (1995).
- **Les femmes au Grand-Duché de Luxembourg: 6. Les familles monoparentales au Luxembourg ou élever seule son enfant au Luxembourg.** Document PSELL n°78. B.LEJEALLE (1995).
- **L'évolution du niveau de vie des personnes au Grand-Duché de Luxembourg entre 1985 et 1992.** Série "Niveau de vie " 6. Document PSELL n°79. P.HAUSMAN (1995).
- **Le suivi des diplômés du BAC technique E.C.G. : Promotions 1987 à 1994.** Document PSELL n°80. P.HAUSMAN (1995).
- **Les demandeurs d'une admission en maison de soins. Les demandeurs d'une allocation de soins. Année 1994.** Document PSELL N° 81 A.KERGER en collaboration : N.ARENDT, L. FALCHERO et M. MERTENS. Publication à diffusion restreinte.(1995)
- **Le maintien à domicile des personnes âgées - Analyse du fichier des clients de longue durée du Service d'Aides familiales/Aides seniors des régions nord, ouest et est..** Document PSELL n° 82. A.KERGER. (1995).

- **Les dépenses : Structure et poids dans le revenu des ménages.** Série "*Mode de vie*" 7. Document PSELL n° 83. P.HAUSMAN. (1996).
- **Répertoire des poids individuels et des poids des ménages Psell 1985-1993.** Document PSELL n°84. B.GAILLY (1996).
- **Répertoire des poids individuels et des poids des ménages Psell 1985-1994.** Document PSELL n°85. B.GAILLY (1996)
- **Procédure de pondération de l'étude relative aux employées privées.** Document PSELL n°86. B.GAILLY (1996).
- **Revenu du travail des jeunes en 1993. Equivalents à ceux de leurs aînés ?** Document PSELL n°87. B.GAILLY (1996)
- **Les employées de statut privé occupées au Luxembourg.** Enquête réalisée pour la Chambre des Employées Privées 1995-1996. Document PSELL n° 88. (Document interne) A.AUBRUN, B.LEJEALLE, P.HAUSMAN. (1996)
- **Les mariés de l'an 1989. Pondération de la cohorte de 1991 à 1994.** Document PSELL n°89. B.GAILLY.
- **Le mode de vie des jeunes adultes : cohabitation avec les parents et départ du foyer parental.** Document PSELL N°90. Série Mode de vie n° 8 P.HAUSMAN. (1996)
- **Les mesures dans le domaine de l'emploi en faveur des groupes de personnes particulièrement désavantagées sur le marché du travail.** Rapport sur le séminaire de suivi sur les décisions du Conseil de l'Union Européenne dans le domaine de l'emploi (Sommet d'ESSEN 1994). Document PSELL n° 91. A.WAGNER, G.SCHABER. (1996)
- **Les comportements de consommation au Luxembourg.** Impact des caractéristiques socio-économiques des ménages selon différents niveaux d'agrégation des dépenses. Paru dans les Cahiers Economiques du STATEC. Document PSELL n° 92 (Enquête Budgets des Ménages).Mode de vie n° 9 M. ZANARDELLI avec la participation du STATEC (diffusion interne). (1996)
- **Les femmes et le chômage en 1994.** Enquête Forces et Travail 1994. Document PSELL n° 93 B.LEJEALLE avec la participation du STATEC. (1996)
- **Fondements méthodologiques de l'échantillon du PSELL n°2.** Document PSELL n°94. B.GAILLY. (1996)
- **La politique familial au Luxembourg - évolution au cours des cinq dernières années.** Document PSELL n°95. M. BORSENBERGER, M. PELS. (1996)
- **Entre famille et activité professionnelle. Mode d'organisation des employées privées.** Document PSELL n°96. P.HAUSMAN, B.LEJEALLE.(1996)
- **L'évolution de l'habitat au Luxembourg. 1985-1994.** Document PSELL n°97.Série "*Mode de vie*" n°10. B.GAILLY. (1996).
- **La mesure de la dépendance. Potentialités et limites du CTMSP pour son application au Grand-Duché de Luxembourg. Compte rendu de l'expérience-pilote réalisée au cours des mois de mai et juin 1996.** Document PSELL n°98. A.KERGER. (1996)
- **Les comportements de consommation au Luxembourg. Une typologie des ménages.** Document PSELL n°99. Série "*Mode de vie*" n°11. M.ZANARDELLI. (1996).
- **Recueil d'études sociales 1996.** Document PSELL n°100.
- **Compenser l'absence de nouveaux immigrants dans le PSELL.1 ?** Document PSELL n°101. B.GAILLY (1997)
- **Représentativité et pondération des échantillons du PSELL2. 1994-1995.** Document PSELL n°102. B.GAILLY (1997)
- **Propriétaire ou locataire : quelles habitations ? 1985-1994.** Document PSELL n°103. B.GAILLY (1997)
- **Femmes au foyer.** Document PSELL n°104. B.LEJEALLE (1997)
- **L'endettement des ménages au Luxembourg en 1994.** Document PSELL n°105. Série "*Mode de vie n°13*" B.GAILLY (1997)
- **Habiter au Luxembourg. Les travaux dans l'habitation. 1985-1994.** Document PSELL n°106 Série "*Mode de vie n°14*" B.GAILLY. (1997)
- **Bacheliers, bachelières de la filière ECG.** Document PSELL n°107. Enquête réalisée en collaboration avec le SCRIPT/Ministère de l'Education Nationale et de la Formation Professionnelle. B. LEJEALLE. (1997).
- **Les Luxembourgeois se sentent-ils en bonne santé et que font-ils pour la préserver ?** Document PSELL n°108 A. AUBRUN (1997)
- **L'emploi du temps des femmes : un partage entre famille, ménage et activité professionnelle.** Document PSELL n°109. B.LEJEALLE (1997)

- **Représentativité et pondération des échantillons du PSELL 2.** 1994-1996. Document PSELL n°110, B.GAILLY (1998)
- **L'endettement des ménages au Luxembourg en 1996.** Document PSELL n°111, B.GAILLY (1998)
- **Eléments d'analyse économique des choix éducatifs au Luxembourg.** Document PSELL n°112, C.KLEIN (1998)
- **La discrimination salariale entre hommes et femmes employés privés.** Document PSELL n°113, P.HAUSMAN, J.LANGERS, B.LEJEALLE, (interne) (1998)
- **Les perspectives familiales : les femmes peuvent-elles choisir librement entre leur vie familiale et leur vie professionnelle ? Envisagent-elles de concilier les deux ? Comment ?** Document PSELL n°114, A.AUBRUN (1998)
- **Revenu disponible et niveau de vie des ménages en 1996 et quelques aspects d'évolution : 1985-1996.** Document PSELL n°115, F.BERGER, P.HAUSMAN, (1998) (à paraître)
- **Habiter ou ne plus habiter chez ses parents.** Document PSELL n°116, F.BERGER, (1998)
- **Représentativité et pondérations des échantillons du Psell 2. 1994-1997.** Document PSELL n°117, B.GAILLY, (1998)
- **Entre activité professionnelle, activité familiale : les choix des femmes luxembourgeoises.** Document PSELL n°118, B.LEJEALLE (1999)
- **Les femmes du secteur des banques et des assurances.** Document PSELL n°119, B.LEJEALLE, (1999)
- **La garde des enfants au Luxembourg.** Document PSELL n°120, M. BORSENBURGER, B.LEJEALLE, (2000)
- **Dans quelle mesure les transferts de politique familiale et sociale réduisent-ils la fréquence et l'intensité de la pauvreté des enfants. Une comparaison France – Luxembourg – Etats-Unis,** Document PSELL n°121, B. JEANDIDIER en collaboration avec P.HAUSMAN, K. VLEMINCKX, R.DE WEVER, M.ZANARDELLI.
- **Situation de l'emploi à Dudelange. Analyse descriptive et dynamique. Rapport détaillé.** Document PSELL n°122, M.BORSENBURGER, P.BOUSCH, (2000)
- **Situation de l'emploi à Dudelange. Analyse descriptive et dynamique. Rapport de synthèse et tableau de bord.** Document PSELL n°123, M.BORSENBURGER, P.BOUSCH, (2000)
- **Les différences salariales en 1995.** Document PSELL n°124, B.LEJEALLE (2001)
- **Passé professionnel et durée de chômage en Europe. Mise en lumière des effets de structure et d'hétérogénéité,** Document PSELL n°125, M.ZANARDELLI, A.REINSTADLER, P.HAUSMAN, J.C.RAY, (2001)
- **A la recherche des déterminants de la durée du chômage au Luxembourg,** Document PSELL n°126, J.BROSIUS, (2001)
- **Evolution des configurations familiales des ménages au Luxembourg,** Document PSELL n°127, M. KUEPIE, (2002)
- **La dynamique de la pauvreté au Grand-Duché de Luxembourg,** Document PSELL n°128, F. BERGER, (2002)
- **Formation initiale, formation professionnelle et profession,** Document PSELL n°129, F. BERGER, (2002)
- **La dynamique de pauvreté en Europe : une approche en termes d'événements d'emploi et/ou de vie familiale,** Document PSELL n°130, F. BERGER (CEPS/INSTEAD), C. BOURREAU-DUBOIS (EPS-ADEPS), B. JEANDIDIER (EPS-ADEPS), (2002).
- **Rendement moyen de l'éducation et l'effet des interruptions involontaires des carrières professionnelles sur le rendement de l'éducation,** Document PSELL n°131, C. KLEIN, (2002).

(Octobre 2002)

