Méthodologie statistique

M2016/05

Les modèles multiniveaux

Pauline Givord – Marine Guillerm

Document de travail

Institut National de la Statistique et des Études Économiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail « Méthodologie Statistique »

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

M 2016/05

Les modèles multiniveaux

Pauline Givord * – Marine Guillerm **

La rédaction de ce document a été initiée suite à une sollicitation de la Division des Études Territoriales de l'Insee pour une étude sur l'impact du quartier d'habitation sur le retard scolaire. Nous remercions les participants du séminaire de méthodologie statistique de l'Insee du 12 juin 2015, ainsi que les nombreux relecteurs et tout particulièrement Cédric Afsa, Brigitte Baccaïni, Alexandre Cazenave-Lacroutz, Laurent Davezies, Jean-Michel Floch, Éric Lesage, Ronan Le Saout, Olivier Monso, Fabrice Murat et Olivier Sautory qui, par leur lecture attentive et leurs commentaires constructifs, ont permis d'améliorer significativement des versions antérieures de ce document.

Nous restons seules responsables des erreurs ou omissions qui pourraient y demeurer.

* DMCSI – DMS (Département des Méthodes Statistiques) – Division des méthodes appliquées de l'économétrie et de l'évaluation
18, bd Adolphe Pinard – 75675 PARIS CEDEX 14.

** Au moment de la rédaction de ce document de travail, DMCSI – DMS (Département des Méthodes Statistiques) – Division des méthodes appliquées de l'économétrie et de l'évaluation
18, bd Adolphe Pinard – 75675 PARIS CEDEX 14.

Direction de la méthodologie et de la coordination statistique et internationale -Département des Méthodes Statistiques - Timbre L101
18, bd Adolphe Pinard - 75675 PARIS CEDEX - France -
Tél. : 33 (1) 41 17 66 33 - Fax : 33 (1) 41 17 61 97 - CEDEX - E-mail :-DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

Les modèles multiniveaux

Pauline Givord Marine Guillerm

Résumé

Les modèles multiniveaux (aussi appelés modèles hiérarchiques ou modèles mixtes) ont été développés pour répondre aux problèmes spécifiques posés par des données structurées selon plusieurs niveaux, typiquement dans le cas où des individus partagent un environnement commun qui peut affecter le comportement étudié. C'est par exemple le cas d'élèves dans un même établissement scolaire, de salariés d'une même entreprise, de patients d'un même hôpital... Les questions classiques auxquelles tentent de répondre les modèles multiniveaux sont ainsi de mettre en évidence l'existence de ces "effets de contexte", de quantifier dans quelle mesure ils contribuent à expliquer l'hétérogénéité entre individus et/ou plus simplement d'estimer correctement l'effet des variables individuelles auxquelles on s'intéresse. Ce document en présente une première introduction pratique, les détails de leur implémentation concrète par les logiciels statistiques standards (Sas, R, Stata) et l'interprétation qui peut être faite des résultats obtenus par ces méthodes. Il présente deux exemples concrets correspondant à une variable d'intérêt respectivement continue et binaire.
Classification JEL : C23, C33, C25, C87.
Mots clés : Modèles multiniveaux, Modèles hiérarchiques, Modèles mixtes, Modèles à effets aléatoires, Modèles à effets fixes, Modèles binaires.

Abstract

Multilevel models (also called hierarchical or mixed models) have been developed to answer issues raised by data structured by several levels, typically when some individuals share a common context that may affect the considered behaviour. This is for instance the case for pupils in one school, employees in one firm, patients in one hospital... The classic questions that are adressed by multilevel models are to highlight the existence of these "contextual effects", to quantify in which measure they contribute to explain heterogeneity between individuals and/or simply obtain unbiased estimates of the impact of some individual variables we are interested in. This document presents a first practical introduction of these models. It insists on the details of their concrete implementation by usual statistical softwares (Sas, R, Stata) and on the interpretation that can be done of the results obtained by these methods. It shows two concrete examples corresponding on a variable of interest respectively continous and binary.
Classification JEL : C23, C33, C25, C87.
Keywords : Multilevel models, Hierarchical models, Mixed models, Random effects models, fixed-effects models, binary models.

Table des matières

1 Effets de contexte et modèles multiniveaux **4**
- 1.1 Pourquoi une modélisation multiniveaux ? . . . 4
- 1.2 Le modèle de base : le cas linéaire . . . 9
  - 1.2.1 Modèle à effets fixes . . . 12
  - 1.2.2 Modèle à effets aléatoires . . . 13
  - 1.2.3 Modèle à effets fixes ou modèle à effets aléatoires ? . . . 15
- 1.3 Variable d'intérêt binaire . . . 19
  - 1.3.1 Généralités . . . 19
  - 1.3.2 Estimation . . . 20
  - 1.3.3 Interprétation : les coefficients estimés . . . 20
  - 1.3.4 Interprétation : les effets marginaux . . . 21

2 Extensions **24**
- 2.1 Des effets variables selon les groupes . . . 24
- 2.2 Plus de deux niveaux . . . 25

3 En pratique **26**
- 3.1 Principes de l'estimation . . . 26
  - 3.1.1 Modèle à effets fixes . . . 26
  - 3.1.2 Modèle à effets aléatoires . . . 27
- 3.2 Dans les logiciels statistiques . . . 29
  - 3.2.1 Avec Sas . . . 29
  - 3.2.2 Avec R . . . 32
  - 3.2.3 Avec STATA . . . 33

4 Exemples **34**
- 4.1 Modélisation du salaire . . . 34
- 4.2 Cas binaire : modélisation du retard scolaire . . . 37

5 Conclusion **46**

A Détails des programmes utilisés pour les exemples **50**
- A.1 Cas d'une variable continue . . . 50
  - A.1.1 Programmes et sorties avec le logiciel Sas . . . 50
  - A.1.2 Programmes et sorties avec le logiciel R . . . 52
  - A.1.3 Programmes et sorties avec le logiciel Stata . . . 53
- A.2 Cas d'une variable binaire . . . 55
  - A.2.1 Programmes et sorties avec le logiciel SAS . . . 55
  - A.2.2 Programmes et sorties avec le logiciel R . . . 57
  - A.2.3 Programmes et sorties avec le logiciel STATA . . . 59

Introduction

Les modèles multiniveaux (parfois aussi désignés sous le terme de modèles hiérarchiques, ou modèles mixtes), ont été développés pour répondre aux problèmes spécifiques posés par des données structurées selon plusieurs niveaux, typiquement dans le cas où les individus partagent un environnement commun qui peut affecter le comportement auquel on s'intéresse. C'est par exemple le cas d'élèves dans une classe : au-delà des caractéristiques propres à chacun d'entre eux, ces élèves bénéficient de conditions d'apprentissage communes, et celles-ci peuvent également influer sur leurs résultats scolaires. De même, les patients d'un même hôpital, les habitants d'un même quartier, les salariés d'une entreprise, les enfants d'une même famille... partagent un environnement ou un ensemble de caractéristiques communes qui peuvent jouer dans la réalisation individuelle de tel ou tel événement. Certaines de ces caractéristiques communes peuvent être identifiées et mesurées, mais d'autres le sont plus difficilement. Dans l'exemple des élèves, on pourra en général avoir simplement des indications sur le niveau d'expérience des enseignants de cette classe, la localisation de l'établissement scolaire dans un quartier considéré comme "sensible", mais rarement sur le climat scolaire ou la cohésion de l'équipe pédagogique. Pour autant, ces dimensions peuvent être importantes pour l'analyse. Les modèles multiniveaux ont été développés pour traiter de cette structure des données. Ils sont très fréquemment utilisés en sciences de l'éducation (voir Bressoux, 2007 pour une introduction) et en économie de la santé (une brève synthèse est par exemple fournie par Chaix et Chauvin, 2002), mais correspondent aussi au cas des données de panel. Ils permettent de répondre à plusieurs types de questions. En considérant, par exemple, le cas de données sur les résultats scolaires d'élèves, regroupés au sein de classes, on pourra ainsi mettre simplement en évidence et quantifier les différences au niveau des classes (qu'on pourrait par exemple attribuer à un "effet maître"), et surtout étudier dans quelle mesure ces différences demeurent, une fois tenu compte de la composition sociale des élèves qui les composent. Quand l'analyse porte au contraire sur les déterminants individuels de la réussite scolaire des élèves, il peut s'avérer indispensable de tenir compte de l'existence de ces "effets de contexte" pour estimer sans biais l'impact propre des variables dont on souhaite mesurer l'effet.

Deux types de modèles statistiques sont classiquement utilisés pour répondre aux questions posées par la structure emboîtée des données. Le premier consiste à supposer que les effets de contexte inobservés ont une distribution qu'on peut bien approximer par une loi normale dont il s'agira d'estimer la dispersion, en même temps que l'effet des variables observables (qu'elles soient individuelles ou au niveau du groupe). Ce modèle "à effets aléatoires"[1] a l'avantage de permettre une modélisation explicite des effets de contexte, et fournit souvent des interprétations plus riches des résultats, mais au prix d'hypothèses statistiques fortes : en particulier, l'estimation suppose que ces effets de contexte ne sont pas corrélés aux variables explicatives. Cette hypothèse n'est pas toujours vérifiée, ce qui expose au risque d'obtenir des estimations biaisées. Pour cette raison, les économètres préfèrent souvent recourir à un modèle dit "à effets fixes", dans lequel on contrôle simplement de ces effets de contexte, sans tenter de les modéliser explicitement. Les estimations des variables individuelles sont *a priori* plus robustes, mais il peut être plus difficile d'exploiter les résultats, en particulier lorsqu'on s'intéresse à des variables discrètes. Le choix entre ces deux méthodes sera dépendant du type de questions auxquelles on souhaite répondre, mais surtout des données dont on dispose et des informations qu'elles contiennent, qui permettent, ou non, de se reposer sur telle ou telle hypothèse.

[1] Selon les disciplines, les mêmes termes peuvent désigner des choses différentes. On utilise ici les termes utilisés classiquement par les économètres. Une définition formelle est fournie plus bas.

Ce document propose une première introduction pratique à ces modèles : il s'agit d'abord de présenter les problématiques auxquelles ils tentent de répondre, les principaux modèles statistiques qui sont classiquement mobilisés, les critères qui peuvent amener à choisir entre l'une ou l'autre des modélisations, et l'interprétation qui peut être faite des résultats. La première partie illustre, à partir d'un modèle linéaire à deux niveaux, pourquoi il peut être nécessaire de tenir compte de cette structure dans les estimations et les deux principales spécifications retenues en général à cette fin (effets fixes ou effets aléatoires). La partie suivante présente des extensions importantes, en particulier lorsque la variable d'intérêt est binaire, mais également le cas où les données sont structurées selon trois niveaux (ou plus) pas nécessairement emboîtés, ou encore celui où la relation entre la variable d'intérêt et les variables explicatives peut varier d'un groupe à l'autre. La troisième partie détaille le principe de l'estimation et sa mise en œuvre dans les principaux logiciels statistiques, et la dernière illustre les différentes étapes de leur utilisation à partir de deux exemples concrets.

Dans l'ensemble du document, l'accent est mis sur la mise en œuvre en insistant sur ce que signifient les choix de modélisation et les hypothèses sous-jacentes, sans présenter explicitement les démonstrations. Le lecteur intéressé pourra se référer par exemple à Davezies (2011) pour une présentation détaillée des aspects statistiques. Pour aller au-delà de l'introduction présentée par ce document, il existe de très nombreux manuels, majoritairement anglophones[2], qui présentent l'usage de ces modèles. Snijders et Berkhof (2007) sont des auteurs de références, et présentent des applications dans de nombreux logiciels. L'ouvrage de Gelman et Hill (2007) fournit une introduction très claire à ces méthodes, avec une approche bayésienne et des exemples pratiques sur le logiciel R[3] .

1 Effets de contexte et modèles multiniveaux

1.1 Pourquoi une modélisation multiniveaux ?

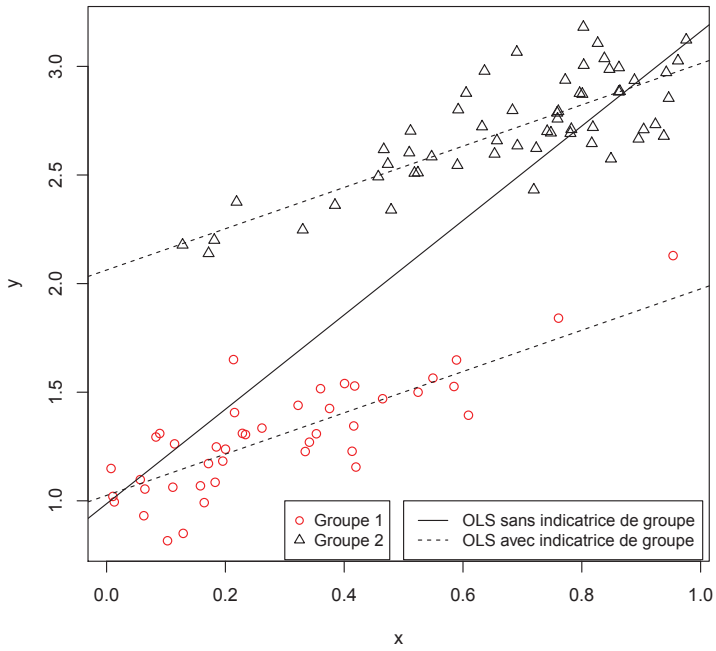
Les modèles multiniveaux répondent au cas où les données sont "structurées", au sens où les différentes observations (correspondant par exemple à des élèves, des patients, des salariés...) peuvent être regroupées au sein d'unités (par exemple, des classes, des médecins, des entreprises...) et partagent ainsi des caractéristiques communes. Les questions auxquelles on peut souhaiter répondre dans ce contexte sont multiples. On peut ainsi souhaiter simplement décrire la part de la variance des salaires qui s'explique par les spécificités des entreprises, ou analyser s'il existe des différences de niveau scolaire entre écoles, une fois contrôlé des caractéristiques socio-démographiques des élèves. On peut également souhaiter quantifier un éventuel "effet maître", ou au contraire mesurer l'effet de variables spécifiques aux élèves (l'origine sociale par exemple) sur la réussite scolaire. Dans tous les cas, ne pas tenir compte de la structure particulière des données peut conduire à des analyses faussées. En effet, l'existence de caractéristiques partagées, qui ne sont pas toutes observables, par l'ensemble des individus d'un groupe signifie

[2]Une exception étant le manuel de Courgeau (1994).

[3]Le manuel de Wang et al. (2011) est explicitement tourné vers le logiciel SAS, et celui de Rabe-Hesketh et Skrondal (2005) vers Stata. Enfin, de très nombreux documents pédagogiques - également en anglais- sont disponibles sur le site du Center for Multilevel Modelling de l'université de Bristol, librement accessible à l'adresse : `http://www.bristol.ac.uk/cmm/learning/`. Ce centre a développé et maintient un logiciel, MLWin, spécifiquement dédié aux modèles multiniveaux, dont les auteures de ce document de travail n'ont pas exploré les possibilités.

que les observations auront une structure particulière, et que l'inférence statistique classique risque de fournir des résultats biaisés.

Des estimateurs biaisés Le graphique 1 illustre par un exemple simple le problème que la présence de caractéristiques communes à l'ensemble des individus d'un groupe peut poser. Il présente des données simulées pour deux groupes qui se distinguent nettement. Cette situation pourrait correspondre, par exemple, à l'effet du niveau de diplôme (représenté en abscisse), sur le salaire (représenté en ordonnée), à partir d'observations obtenues dans deux entreprises différentes (repérées par des symboles différents). Dans cet exemple fictif, le rendement des diplômes est le même dans les deux entreprises : on peut de fait observer que les pentes des deux droites en pointillés (qui correspondent à deux régressions linéaires estimées séparément dans les deux groupes) sont identiques. Cependant, les deux groupes de points sont nettement séparés. Ne pas tenir compte de la spécificité de chacun des groupes peut conduire à fortement surestimer la relation causale entre les deux variables, comme illustré par la pente de la droite de régression linéaire "naïve" (en trait plein) qui est bien plus élevée que celles observées *au sein* de chaque groupe. Comment expliquer cet écart ? Un simple aperçu suggère que les deux entreprises sont différentes. Dans notre interprétation, on pourrait dire que l'une des entreprises emploie majoritairement des salariés très qualifiés, tandis qu'on trouve plutôt dans l'autre des salariés non qualifiés. Mais au-delà de cet écart de recrutement, la première entreprise a aussi des niveaux de salaires bien plus élevés, quel que soit le niveau de diplôme. Ces différences pourraient s'expliquer, par exemple, par le secteur d'activité (industrie versus banque/assurance, public ou privé) ou la composition en termes d'ancienneté, qui ne sont pas bien captés par la qualification. En termes statistiques, le fait que l'entreprise dans laquelle les salaires sont les plus élevés recrute aussi les plus diplômés signifie qu'il y a une corrélation entre la composante spécifique de l'entreprise (qui sera notée par la suite α_j) et le diplôme. La régression sur l'ensemble pourra donc fournir une estimation biaisée du rendement du diplôme *au sein* d'une entreprise.

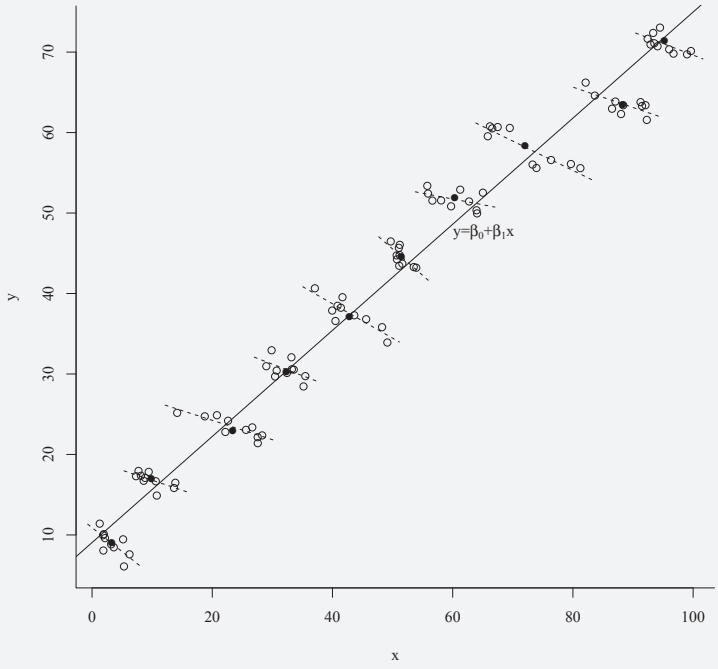


Graphique 1 – Biais de variable omise

Le biais d'agrégation ou erreur écologique

L'existence d'effets "de contexte", c'est-à-dire d'éléments inobservés communs à l'ensemble des unités d'un groupe, peut également induire des erreurs importantes dans l'interprétation de certaines corrélations. C'est en particulier le cas lorsqu'on tente d'inférer des causalités individuelles alors qu'on ne dispose que de données agrégées. Dans un exemple devenu classique, Robinson (1950) a illustré le risque de tirer des conclusions erronées à partir de données agrégées. À partir du recensement américain de 1930, il montrait ainsi que la corrélation entre taux d'illettrisme et la proportion de personnes noires était de 0,9 en utilisant des données agrégées au niveau des États américains, alors qu'elle se réduisait à 0,2 en utilisant les données individuelles. Robinson montre aussi que le biais d'agrégation peut dans certains cas inverser le sens de la corrélation : ainsi, en étudiant le lien entre illettrisme et cette fois le fait d'être né hors des États-Unis, il obtient une corrélation de 0,12 en utilisant des données individuelles, mais cette corrélation est négative de -0,53 en utilisant des données agrégées au niveau des États. Ce dernier résultat ne signifie pas que les personnes nées aux États-Unis étaient plus souvent illettrées que les autres, mais sans doute que les immigrés récents se regroupaient dans les États les plus riches et donc en moyenne les plus instruits (alors qu'à l'inverse dans l'exemple précédent, les afro-américains se concentraient en 1930 dans les États du sud moins avancés économiquement).

Ce biais est illustré dans l'exemple fictif du graphique ci-dessous. Dans chacun des groupes, il existe une corrélation négative entre la variable d'intérêt y et la variable x représentée en abscisse. Néanmoins, entre groupes, on observe une corrélation positive entre la valeur moyenne de ces deux variables au niveau de chacun des groupes.



Erreur écologique : illustration

L'erreur de raisonnement consistant à attribuer à un membre d'un groupe les caractéris-

tiques moyennes de ce groupe, ou plus généralement à passer d'une relation établie au niveau d'un agrégat d'individus à une relation au niveau des individus, est souvent appelée "erreur" écologique (en épidémiologie, le terme écologique s'applique à des études qui reposent uniquement sur des données agrégées, à l'échelle d'un pays par exemple). Elle est fréquemment évoquée en géographie, où on travaille souvent sur les caractéristiques d'unités spatiales qui sont des agrégats d'individus.

Une précision trompeuse Même lorsqu'il n'existe pas de corrélation entre les effets groupes et les variables explicatives, négliger ces termes communs aux individus des groupes dans l'estimation peut biaiser l'estimation de la précision des estimateurs. En effet, le modèle linéaire "standard" suppose que les observations sont indépendantes entre elles. Si ce n'est pas le cas, ne pas tenir compte de cette corrélation va conduire à une estimation biaisée de la vraie variance de l'estimateur.

En pratique, cela signifie qu'on risque de conclure à tort qu'une variable explicative a un effet non nul sur notre variable d'intérêt. Un ordre de grandeur est fourni par Moulton (1990). Dans un cadre simplifié, où en particulier le nombre d'individus par groupe est constant et égal à m, il montre que les écarts-types obtenus par les moindres carrés ordinaires sous-estiment les "vrais" écarts-types d'un facteur $1 + \rho(m-1)$ où ρ mesure la corrélation intra-groupe (dans quelle mesure les observations d'un même groupe se ressemblent). Cette formule ne peut pas s'appliquer à tous les cas, mais est présentée simplement à titre d'illustration : si chaque groupe ne comprend qu'un individu ($m = 1$) on retombe dans le cadre linéaire classique. En revanche, dès lors qu'il y a plusieurs individus par groupe, la sur-estimation de la précision par l'estimateur des moindres carrés ordinaires sera d'autant plus élevée que les individus d'un même groupe se ressemblent.

Plusieurs possibilités existent alors pour tenir compte de ce problème. La plus directe consiste à estimer la composante spécifique à chaque groupe, par exemple en ajoutant une indicatrice de groupe (on parle d'effets fixes). Cette composante spécifique est supposée capter l'ensemble des spécificités, observées ou inobservées, d'un groupe. On peut aussi choisir d'expliciter cette composante, en introduisant des covariables correspondant au groupe (dans notre exemple, le secteur de l'entreprise, sa taille, la structure de sa main d'œuvre) dont on souhaite contrôler l'effet. À noter qu'il n'est pas possible d'identifier simultanément l'effet de ces variables définies au niveau du groupe et un effet fixe : intuitivement, ce dernier capte l'ensemble des effets des caractéristiques, y compris observables, des groupes. Si on est prêt à supposer que, au-delà de ces caractéristiques observables au niveau du groupe, la part spécifique restante n'est pas corrélée avec les variables explicatives individuelles, on peut choisir d'estimer un modèle à "effets aléatoires". Celui-ci repose, outre sur cette hypothèse de non corrélation entre les termes inobservés des groupes et les variables explicatives, sur l'hypothèse que la distribution de ces termes suit une loi spécifique (en général une loi normale). Il est alors possible d'estimer les paramètres qui lui correspondent. Nous détaillons plus bas les deux types d'estimateurs, mais il faut être conscient d'ores et déjà que cette hypothèse d'indépendance est forte, et devra être justifiée au cas par cas. Recourir à l'estimation d'un modèle à effets aléatoires permet alors d'estimer non seulement l'impact des caractéristiques au niveau de l'individu et du groupe, mais également la dispersion des termes liés aux caractéristiques inobservées du groupe.

Il est donc en général indispensable d'utiliser des modèles tenant compte de la structure hié-

rarchique, ou "groupée" des données. Pour ce type de modèle, une question récurrente sera le choix retenu pour modéliser la partie inobservée commune à l'ensemble des observations d'un groupe, soit entre les modèles à effets fixes ou à effets aléatoires. On détaille dans la section suivante les principes de ces deux spécifications, leur estimation, et les critères permettant de choisir entre elles. Il est aussi utile d'insister sur le fait que ces modèles statistiques peuvent avoir comme objet de répondre à des questions différentes. La première, centrale en économétrie, est de mettre en évidence l'*effet causal d'une variable* (par exemple, l'effet propre du diplôme sur le salaire, ou le lien entre origine sociale et réussite des élèves...). Dans ce cadre, l'introduction de variables explicatives supplémentaires et la modélisation des composantes inobservées interviennent surtout pour "contrôler" d'une hétérogénéité qui pourrait biaiser l'effet auquel on s'intéresse, mais ces dimensions n'ont pas d'intérêt en elles-mêmes. Dans d'autres contextes, cette hétérogénéité est intéressante en tant que telle. Un exemple classique est lié à la notion de "valeur ajoutée" d'un établissement scolaire (voir une illustration pour les indicateurs de performance pour les lycées français dans Murat et al., 2014). Il s'agit ici d'évaluer si un établissement permet à ses élèves de faire mieux (ou moins bien) que ce qui est observé en moyenne pour des élèves comparables en termes de caractéristiques scolaires et socio-démographiques. Cette évaluation se fait à partir de l'estimation de l'"effet groupe" (ici, l'établissement) qui permet de déterminer où l'établissement se positionne en termes de réussite des élèves qu'il scolarise, par rapport à ce qui est prévisible compte tenu des caractéristiques de son recrutement.

Données de panel Les données de panel, pour lesquelles on dispose d'observations pour une même unité (individu, entreprise,...) à différentes dates constituent un cas particulier, important, de données structurées : les observations d'une même unité forment un groupe. De même que précédemment, deux types de variables peuvent avoir un effet sur la variable d'intérêt : celles constantes pour toutes les unités d'un même groupe (qu'on appellera par la suite variables de niveau 2) qui correspondent dans le cadre des données de panel aux variables intrinsèques de l'unité considérée fixes dans le temps, et les autres (variables dites de niveau 1) qui au contraire varient d'une observation à une autre au sein d'un groupe et correspondent ici aux variables non constantes dans le temps (et sont généralement indicées par t). En pratique, on a souvent beaucoup d'unités et peu d'observations temporelles pour chacune de ces unités. L'économètre habitué à l'utilisation des données de panel trouvera ici une proximité troublante aux discussions classiques dans l'économétrie des données de panel. Néanmoins, l'utilisation de ces modélisations dans d'autres cadres peut conduire à mettre l'accent sur des caractéristiques du modèle qui sont peu étudiées voire ignorées en économétrie des panels. Cela s'explique par l'usage différent qui peut être fait de données groupées. En général, l'utilisation de données individuelles en panel a comme objectif premier d'éliminer une éventuelle hétérogénéité individuelle pour estimer les relations causales entre des variables explicatives et la variable d'intérêt, et non d'analyser cette hétérogénéité *per se*.

1.2 Le modèle de base : le cas linéaire

En pratique, on part d'un modèle de régression classique dans lequel on cherche à modéliser une variable d'intérêt continue y (par exemple, les résultats scolaires, les dépenses de santé, le salaire...) en fonction de caractéristiques x. Ces caractéristiques peuvent être observées au niveau individuel (dans nos exemples respectivement l'élève, le patient, le salarié), mais aussi représenter des variables communes à des individus au sein d'un groupe (la classe, l'hôpital, l'entreprise). Cette structure emboîtée est reflétée par le terme de modèles multiniveaux, ou de

modèle hiérarchique. En pratique, on appelle variables de niveau 1 celles au niveau le plus fin, et niveau 2 les variables du groupe. On peut évidemment envisager plus de niveaux (par exemple, les classes sont regroupées au niveau d'un établissement scolaire...), mais on se contente à fin d'illustration d'un modèle à deux niveaux seulement[4]. En pratique, on écrit donc :

$$y_{ij} = \beta_0 + x_{ij}\beta + x_j\gamma + \alpha_j + \varepsilon_{ij} \quad j = 1,\dots,J \quad i = 1,\dots,n_j \tag{1}$$

Les doubles indices ij illustrent la structure multiniveaux du modèle[5]. Par convention, l'indice i désigne l'individu (niveau 1) tandis que l'indice j désigne le groupe (niveau 2)[6]. La variable d'intérêt y_{ij} peut par exemple désigner les notes de l'élève i faisant partie de la classe j, x_{ij} correspond aux covariables de niveau 1 : par exemple, l'âge, le sexe ou l'origine sociale de l'élève. x_j correspond aux covariables de niveau 2, donc dans notre exemple observées au niveau de la classe : l'enseignant en particulier, mais aussi la composition générale (la proportion de filles ou d'élèves issus de milieu familial favorisé par exemple). Ces covariables sont constantes pour toutes les unités d'un même groupe. Les paramètres d'intérêt principaux du modèle sont généralement les coefficients β, l'impact des caractéristiques individuelles sur la variable d'intérêt, mais on peut également souhaiter estimer les coefficients γ qui correspondent aux variables de contexte. Les termes $\alpha_j + \varepsilon_{ij}$ correspondent aux termes inobservés, qu'on choisit là aussi de décomposer en un terme strictement individuel ε_{ij} et un terme α_j commun à tous les individus d'un même groupe j qui résume l'effet des variables inobservées affectant simultanément tous les individus de ce groupe. On peut illustrer cette décomposition des termes inobservés dans le graphique 2. Dans cet exemple simplifié, on ignore l'effet de variables explicatives éventuelles (les points sont dispersés sur l'axe des abscisses pour permettre de lire le graphique). La moyenne sur l'ensemble de l'échantillon correspond à β_0. Pour chaque groupe, le terme α_j représente l'écart de la moyenne du groupe à cette moyenne générale. Enfin, au sein de chaque groupe, le terme ε_{ij} correspond à l'écart individuel à la moyenne du groupe $\beta_0 + \alpha_j$. L'équivalent pour un modèle avec une variable explicative est présenté sur le graphique 3.

Une première hypothèse importante, que l'on supposera vérifiée dans toute la suite, est que les termes d'erreur strictement individuels ne sont pas corrélés à l'ensemble des covariables du modèle, et sont indépendants entre eux.

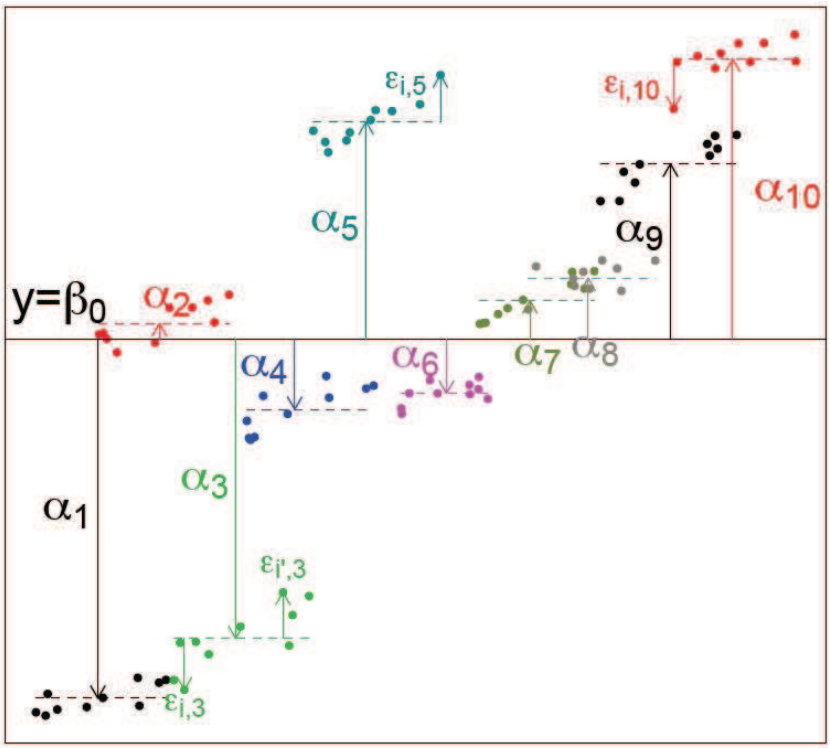
[4] L'extension à plus de deux niveaux est présentée au paragraphe 2.2.

[5] Ce type de modèle qui est un cas particulier de modèle à effets aléatoires est aussi appelé modèle à constante aléatoire (*random intercept model*). Il est parfois possible de le trouver présenté ainsi :

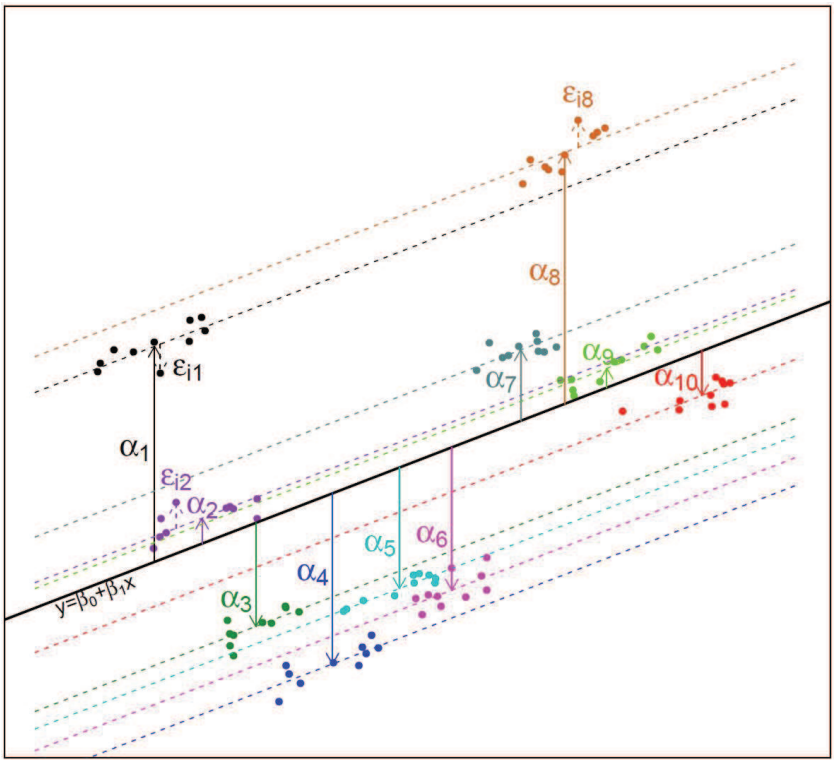
$$y_{ij} = \beta_{0j} + x_{ij}\beta + x_j\gamma + \varepsilon_{ij} \quad j = 1,\dots,J \quad i = 1,\dots,n_j \tag{2}$$

$$\text{avec } \beta_{0j} = \beta_0 + \alpha_j \tag{3}$$

[6] Dans le cas des données de panel, l'indice de niveau 1 correspond à la date d'observation (en général notée t), tandis que l'indice de niveau 2 correspond à l'unité observée sur différentes dates (souvent noté i).



Graphique 2 – Décomposition des erreurs : illustration



Graphique 3 – Décomposition des erreurs : illustration

Questions de sémantique

On trouve parfois le terme d'"effets fixes" pour désigner les paramètres β, tandis que les termes "effets aléatoires" correspondent aux effets α_j. C'est en particulier le cas dans les sorties des procédures Sas permettant de mettre en œuvre les modèles à effets aléatoires, parfois appelée "modèles mixtes" pour insister sur le fait que l'on autorise les coefficients de certaines variables à varier en fonction du groupe, tandis que les effets des autres variables sont considérés comme constants pour toutes les observations. Cette terminologie diffère cependant de la convention plus classique en économétrie dans laquelle on appelle "effets fixes" les effets groupes α_j lorsque leur distribution statistique n'est pas spécifiée explicitement, par opposition aux modèles à effets aléatoires où cette distribution est supposée normale et indépendante des observables. Nous suivons ici cette dernière convention, car nous insistons sur la comparaison entre ces deux spécifications. Il faut bien garder à l'esprit que dans une spécification à effets fixes, ces effets ne sont fixes que pour un groupe donné, mais sont évidemment variables d'un groupe à un autre (et pourraient donc, *stricto sensu*, être considérés comme les réalisations pour chacun des groupes d'une variable aléatoire même si on ne cherche pas à identifier la distribution de celle-ci).

1.2.1 Modèle à effets fixes

Le modèle à effets fixes consiste alors à considérer ces effets non observés spécifiques à chaque groupe comme autant de paramètres à estimer dans l'équation, sans faire plus d'hypothèses. Le modèle peut alors s'écrire :

$$y_{ij} = x_{ij}\beta + \sum_{k=1}^{J} \delta_k \mathbb{1}_{k=j} + \varepsilon_{ij} \quad j = 1,\dots,J \quad i = 1,\dots,n_j \tag{4}$$

où les $\mathbb{1}_{k=j}$ correspondent à des indicatrices de groupes (elles valent 1 lorsque k correspond au groupe j, 0 sinon). Comme souligné plus haut, dans ce cas il n'est pas possible d'identifier séparément l'effet de ces indicatrices de groupes et des variables décrivant ce groupe (de niveau 2, par exemple des caractéristiques de l'entreprise dans l'exemple du salaire). Les coefficients δ_j correspondent de fait à une agrégation de l'ensemble des caractéristiques de groupe, c'est-à-dire en reprenant les notations du modèle (1) on a $\delta_j = \alpha_j + x_j\gamma$: dans ce type de modèle, on tente bien de contrôler de ces effets groupe, mais sans tenter de les décomposer en fonction de caractéristiques observables.

En pratique, on peut estimer directement le modèle (4) comme une régression linéaire classique, c'est-à-dire une méthode de moindres carrés ordinaires de y sur x et les indicatrices de groupes. Si le nombre de groupes est élevé, le nombre de paramètres peut être conséquent et le modèle lourd à estimer. L'alternative classique consiste à transformer le modèle par différentiation pour "se débarrasser" des termes spécifiques au groupe.

$$y_{ij} - \bar{y}_{.j} = (x_{ij} - \bar{x}_{.j})\beta + \varepsilon_{ij} - \bar{\varepsilon}_j$$

c'est-à-dire qu'on estime le modèle en utilisant la variation individuelle d'une variable z à la moyenne observée sur l'ensemble du groupe $\bar{z}_{.j} = \frac{1}{n_j}\sum_{i\in j} z_{ij}$, où n_j correspond au nombre d'observations dans le groupe j.

Ce modèle est estimé par une régression linéaire classique de $y_{ij} - \bar{y}_{.j}$ sur $x_{ij} - \bar{x}_{.j}$. En pratique, la plupart des logiciels statistiques permettent d'estimer le modèle à effets fixes, sans nécessiter une transformation "à la main" des données (voir section 3.2). Il est préférable de recourir à ces procédures que de faire l'estimation à la main, car le calcul de la précision n'est pas immédiat[7].

Même si ces paramètres ne sont pas estimés directement par des indicatrices, il est possible d'obtenir une estimation des effets fixes à partir de ce modèle *within* (de manière standard l'exposant FE, pour *Fixed Effects* précise ici qu'il s'agit de l'estimation obtenue par effets fixes) :

$$\hat{\alpha}_j^{FE} = \bar{y}_j - \bar{x}_{.j}\hat{\beta}^{FE} \tag{5}$$

cependant, ces estimations peuvent être très imprécises si le nombre d'observations n_j par groupe est faible ou lorsque ces groupes sont constitués d'éléments très hétérogènes (au sens où la variance à l'intérieur de chaque groupe est élevée). Dans ce cas, la moyenne observée peut s'éloigner sensiblement de sa "vraie" valeur.

1.2.2 Modèle à effets aléatoires

Une alternative à l'estimation exhaustive des effets spécifiques groupes consiste à supposer que ces termes ont une distribution normale, dont on cherchera simplement à estimer la variance. Le modèle est ainsi plus parcimonieux. Il permet également d'isoler l'effet des variables observables décrivant le contexte (contrairement au modèle à effets fixes, dans lequel elles sont absorbées par l'effet fixe). En revanche, ceci se fait au prix de deux hypothèses supplémentaires. La première est que la "vraie" distribution des effets groupes inobservés suit une loi normale (ou tout au moins une loi paramétrique qu'on peut spécifier). La seconde est que ces effets groupes sont indépendants des variables explicatives. Si la première peut être simplement relâchée, il faut être conscient que cette seconde hypothèse est particulièrement forte et qu'elle peut ne pas être vérifiée. C'était par exemple le cas de notre graphique illustratif 1 où l'entreprise à la politique salariale la plus généreuse était aussi celle qui regroupait les salariés les plus diplômés. Ce modèle est très pauvre, puisqu'il ne comprend qu'une seule variable explicative. On pourrait supposer, par exemple, que l'ajout de variables supplémentaires (par exemple le secteur de l'entreprise) pourrait rendre moins invraisemblable l'hypothèse d'indépendance de la part inexpliquée et des variables observables.

Formellement, on retrouve donc l'équation 1 :

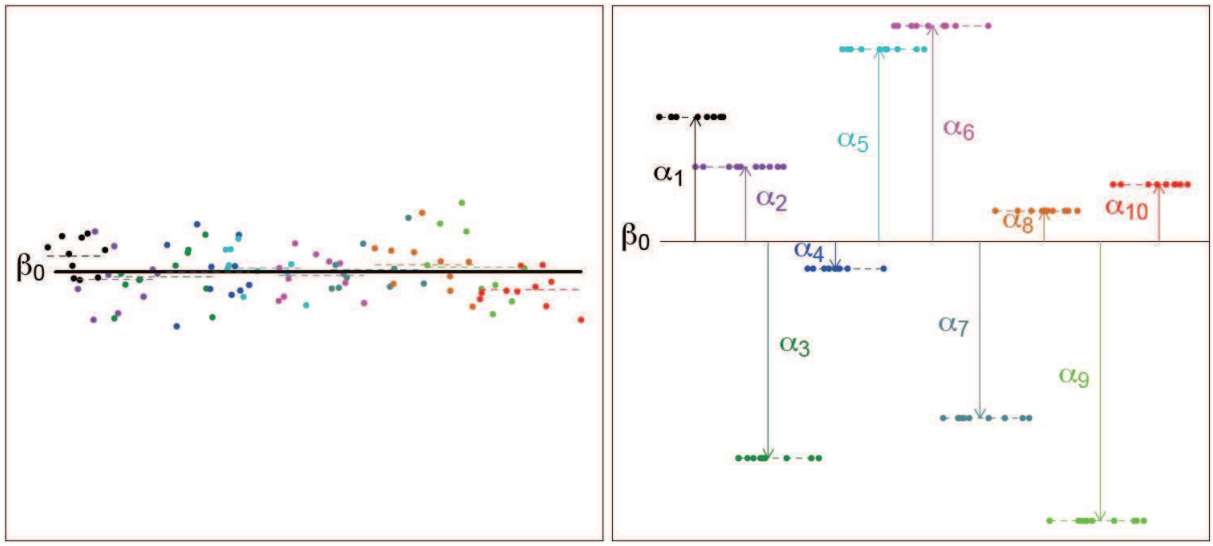
$$y_{ij} = \beta_0 + x_{ij}\beta + x_j\gamma + \alpha_j + \varepsilon_{ij} \quad j = 1,\ldots,J \quad i = 1,\ldots,n_j \tag{6}$$

avec les hypothèses notamment que $\varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2)$ et $\alpha_j \sim N(0,\sigma_\alpha^2)$.

Sous ces hypothèses, on peut estimer le modèle 1 (voir la section 3 pour des détails). On obtient donc une estimation des paramètres du modèle : les coefficients β_0, β et γ correspondant à la constante et aux variables explicatives, ainsi que les variances inter-groupe σ_α^2 et intra-groupe

[7] Si on souhaite faire la transformation à la main, il faut en effet prendre garde au calcul de la précision. Celui-ci prend en compte le nombre de degrés de liberté du modèle. Le nombre d'observations du modèle transformé est N et le nombre de paramètres qu'on estime est K (nombre de coefficients des variables explicatives). Une régression "à la main" de $y_{ij} - \bar{y}_{.j}$ sur $x_{ij} - \bar{x}_{.j}$ considérera que le nombre de degrés de liberté est $N - K$. Or le nombre de degrés de liberté "réel" du modèle (qui doit être utilisé pour le calcul des écarts-types) est $N - K - J$. La précision s'en trouverait alors surestimée, avec donc le risque de conclure à tort à la significativité d'une variable.

σ^2_ε.
En termes d'interprétation, on remarquera que si la corrélation entre deux observations individuelles de deux groupes différents est nulle, la corrélation entre deux observations individuelles au sein d'un même groupe (désignée sous le terme de *ICC* pour *Intra Class Correlation*) vaut $corr(y_{ij}, y_{i'j}) = \frac{\sigma^2_\alpha}{\sigma^2_\varepsilon + \sigma^2_\alpha}$ $(i \neq i')$. L'ICC correspond à la part de la variance expliquée par l'effet groupe et peut être utilisé comme un premier indicateur de l'utilité de recourir à une modélisation multiniveaux. Si la corrélation est proche de zéro, cela signifie que les observations à l'intérieur d'un même groupe ne sont pas plus semblables entre elles que des observations de groupes différents, et donc que recourir à une modélisation multiniveaux n'apporte pas beaucoup plus qu'un modèle linéaire classique. À l'inverse, lorsque cet indicateur est proche de 1, cela signifie que les unités au sein d'un même groupe sont très proches et se distinguent beaucoup de celles des autres groupes. Ces deux cas extrêmes sont illustrés dans le graphique 4, dans lequel on ignore à nouveau l'effet de variables explicatives éventuelles. Sur le graphique de gauche, il n'y a pas d'effet spécifique aux groupes ; à l'inverse dans le cas caricatural du graphique de droite, toute la variabilité observée entre les points provient d'effets liés au groupe. Évidemment, en général, la réalité se trouve entre ces deux extrêmes.



(a) *ICC* proche de zéro (b) $ICC = 1$

Graphique 4 – Illustration de deux cas extrêmes pour la valeur de l'ICC

Prédictions des effets groupes Comme dans le modèle à effets fixes, on peut obtenir à partir de ces paramètres une estimation du terme α_j, c'est-à-dire de la part inexpliquée spécifique à chaque groupe. L'estimateur le plus simple peut être obtenu, comme dans le cas des effets fixes (équation 5), en comparant pour chaque groupe les moyennes par groupe des valeurs observées de la variable d'intérêt et des prévisions obtenues à partir des variables observables. Cependant, on peut montrer que cet estimateur n'est pas optimal. Lorsque le nombre d'observations dans un groupe est faible, les moyennes empiriques calculées pour chaque groupe peuvent être éloignées de leur vraie valeur. Cela introduit de l'imprécision dans les estimations. On préfère donc utiliser un terme qui tient compte de ce risque pour estimer la meilleure prédiction linéaire sans biais (on parle de BLUP, pour *Best Linear Unbiased Predictor*) qui est fournie par (de manière standard, l'exposant RE, pour *Random Effects* précise qu'il s'agit ici de l'estimation obtenue par effets

aléatoires) :

$$\hat{\alpha}_j^{RE} = c_j(\bar{y}_j - \bar{x}_{.j}\hat{\beta}^{RE}) \tag{7}$$

où c_j vaut

$$c_j = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \frac{\hat{\sigma}_\varepsilon^2}{n_j}} \tag{8}$$

Ce terme c_j, toujours compris entre 0 et 1, est appelé facteur de contraction (*shrinkage factor*). Il se rapproche de 1 lorsque le nombre d'observations dans le groupe n_j est grand par rapport à la variance intra-groupe σ_ε^2 (ou plus précisément au rapport entre variance intra-groupe σ_ε^2 et la variance inter-groupe σ_α^2)[8]. Dans le cas contraire en revanche, lorsque le nombre d'observations par groupe n_j est faible ce facteur s'approche de 0 : l'intuition est que lorsque le nombre d'observations pour un groupe est faible, l'estimation de l'effet spécifique à ce groupe risque d'être très bruitée. On préfère donc pour ce groupe utiliser comme prévision le niveau moyen observé sur l'ensemble de l'échantillon.

Tests d'hypothèse La significativité des coefficients correspondant aux variables explicatives se fait classiquement par un test de Student. En revanche, il est *incorrect* d'utiliser ce test pour l'estimateur de la variance intra-groupe $\hat{\sigma}_\alpha^2$. Pour évaluer si la variance intra-groupe est significative, on utilise en général un test de rapport de vraisemblance (ce dernier repose sur le log du carré du ratio des log-vraisemblances des modèles respectivement avec et sans effets aléatoires, dont on peut montrer qu'il suit une loi du χ^2 à un degré de liberté). Ce test est directement fourni par la plupart des logiciels statistiques.

1.2.3 Modèle à effets fixes ou modèle à effets aléatoires ?

Questions d'usage Choisir entre les modèles à effets fixes et effets aléatoires est une question classique, à laquelle il n'existe pas de réponse univoque malgré une très vaste littérature. De manière caricaturale, on est tenté de dire que les économètres préfèrent les modèles à effets fixes et les statisticiens les modèles à effets aléatoires. Ces préférences reflètent aussi les usages différents qui peuvent être faits des modèles. Les micro-économètres tentent en général de mettre en évidence l'effet causal d'une variable explicative, et les effets de contexte ne sont considérés que comme des "nuisances" (en général appelées hétérogénéité inobservée) dont il faut se débarrasser. Les modèles à effets fixes sont plus adaptés à cet objectif que les modèles à effets aléatoires. Les modèles à effets fixes en particulier ne reposent pas sur l'indépendance entre ces effets non observés (de niveau 2) et des variables explicatives, contrairement aux modèles à effets aléatoires. Cette hypothèse est importante : si elle n'est pas vérifiée, les estimateurs à effets aléatoires pourront être sérieusement biaisés. Si on a de bonnes raisons de croire que cette hypothèse est vérifiée et que l'objet d'intérêt est d'abord la mise en évidence des effets de contexte, les modèles à effets aléatoires sont en théorie plus adaptés. Ils permettent d'obtenir des estimations plus précises, et de faire des prédictions pour des individus de groupes qui ne feraient pas partie de l'échantillon. Comme souligné par Clark et Linzer (2015), on peut voir la question du choix entre modèles à effets fixes ou effets aléatoires comme un arbitrage biais/variance classique. Les modèles à effets aléatoires sont plus précis mais ont un risque très élevé de fournir des estimateurs biaisés. Un modèle à effets fixes fournit certes un estimateur sans biais, c'est-à-dire que son espérance correspond à la vraie valeur du paramètre, mais l'estimation peut être en pratique très éloignée de cette dernière si l'estimateur est peu précis.

[8] En général, on s'attend à ce que la première soit supérieure à la seconde.

Questions de taille De manière pratique, le nombre de groupes, mais aussi le nombre d'observations par groupe, peuvent être des dimensions importantes à prendre en compte (même si on ne s'aventurera pas ici à donner des règles chiffrées définitives). Intuitivement, si le nombre de groupes est faible, l'estimation de la variabilité inter-groupe (le σ^2_α) sur laquelle repose le modèle à effets aléatoires sera imprécise. On trouvera par exemple dans Bryan et Jenkins (2015) une analyse des risques de l'utilisation de modèles à effets aléatoires pour les comparaisons d'enquêtes internationales : il s'agit typiquement d'un cas où on dispose de beaucoup d'observations par groupe, mais de peu de groupes. Ce n'est pas un problème *a priori* pour un modèle à effets fixes, puisqu'on ne cherche pas à estimer la distribution des effets groupes. Á l'inverse, comme déjà discuté plus haut, disposer d'un faible nombre d'observations par groupe peut conduire à privilégier l'estimation d'un modèle à effets aléatoires, car l'estimation des effets fixes sera médiocre. Comme souligné par Gelman et Hill (2007) en revanche, le fait d'avoir plusieurs groupes avec peu d'observations n'est pas un problème en soi : les effets spécifiques à ces groupes seront estimés avec une faible précision, mais ils apportent néanmoins de l'information pour permettre l'estimation du modèle sur l'ensemble de l'échantillon (pourvu qu'il y ait "assez" de groupes avec suffisamment d'observations). Maas et Hox (2005) et Bell et al. (2008) montrent sur des données simulées que les estimateurs à effets aléatoires seraient assez robustes même lorsque le nombre de groupes ne comportant qu'une seule observation (*singletons*) est grand, pour peu que le nombre de groupes soit suffisamment élevé. Ces résultats ne sont valables que dans le cas des simulations étudiées par ces auteurs et il est déconseillé d'en tirer des conclusions trop générales : la sensibilité des estimations pourra dépendre aussi de la variance intra-groupe par exemple. Notons enfin que dans le cas d'une variable d'intérêt binaire, les choses peuvent aussi être plus complexes car l'estimation repose sur des approximations numériques (voir partie 1.3).

Questions pratiques Enfin, les estimateurs à effets aléatoires ont des avantages "pratiques" pour la modélisation des effets de contexte qui expliquent aussi sans doute leur popularité. Le fait de supposer explicitement que les effets groupes suivent une distribution normale a un inconvénient, puisque rien ne garantit que la "vraie" distribution est bien approximée par une distribution normale. Mais cela constitue aussi un avantage de ces modèles. En effet, on estime explicitement les paramètres de cette distribution (en pratique, l'écart-type d'une loi normale). Cela permet d'étendre des prédictions de ces effets pour des observations ne faisant pas partie de l'échantillon initial : cela peut se faire par exemple en tirant aléatoirement une valeur pour un groupe à partir de cette loi normale, toujours sous l'hypothèse que ces effets de contexte sont indépendants des variables observables. Elle est aussi précieuse dans les modèles non linéaires, car elle permet d'estimer des effets marginaux d'une variable (voir section 1.3), ce qui n'est pas possible avec des effets fixes.

Test d'Hausman et modèles augmentés En résumé, si les modèles à effets aléatoires ont beaucoup d'avantages "pratiques", ils reposent aussi sur des hypothèses plus fortes. Si l'indépendance des effets groupes aux variables explicatives n'est pas vérifiée, les estimateurs pourront être biaisés. Le modèle à effets fixes permet de s'affranchir de cette hypothèse, mais a un autre inconvénient, il est moins efficace (c'est-à-dire moins précis). Pour arbitrer entre ces deux difficultés, on peut utiliser un test d'Hausman. Celui-ci consiste à comparer les estimateurs des coefficients des variables individuelles obtenus respectivement avec le modèle à effets aléatoires et le modèle à effets fixes, et de vérifier qu'ils ne sont pas trop différents. En pratique, il s'agit donc d'estimer les deux jeux de coefficients. La statistique de test correspond alors à :

$$\left(\hat{\beta}^{FE} - \hat{\beta}^{RE}\right)' \left[\hat{var}\hat{\beta}^{FE} - \hat{var}\hat{\beta}^{RE}\right]^{-1} \left(\hat{\beta}^{FE} - \hat{\beta}^{RE}\right)$$

L'hypothèse nulle consiste à supposer que les effets groupes α_j sont non corrélés aux variables explicatives individuelles. L'intuition du test est que si cette hypothèse n'est pas vérifiée, les estimateurs du modèle à effets fixes sont convergents alors que ceux du modèle à effets aléatoires ne le sont pas. Si la différence entre les deux estimateurs est élevée, il est probable que l'estimateur à effets aléatoires soit biaisé car l'hypothèse nulle n'est pas vérifiée. Techniquement, on peut montrer que la statistique de test suit une loi du χ^2 à K degrés de liberté (K étant le nombre de variables du vecteur x_{ij}). Si la différence entre les deux estimateurs est significative (c'est-à-dire si la valeur de la statistique de test est plus grande que la valeur qu'on s'attend à trouver avec une probabilité "raisonnable" si cette hypothèse était vérifiée), on rejette l'hypothèse nulle. Cela signifie qu'il faut préférer l'estimateur à effets fixes. Dans le cas contraire, on choisira le modèle à effets aléatoires, *a priori* plus précis.

De manière équivalente mais plus directe, on peut estimer un modèle qui prend en compte explicitement cette corrélation éventuelle entre les variables individuelles et les effets groupes. En pratique, cette corrélation peut se décliner sous la forme $\alpha_j = \bar{x}_{.j}\delta + u_j$, où δ est non nul et $\bar{x}_{.j}$ correspond à la moyenne des variables explicatives sur le groupe j ($\bar{x}_{.j} = 1/n_j \sum_{i \in j} x_{ij}$) , et u_j un terme d'erreur de moyenne nulle non corrélé à X (on pose en général $u_j \sim N(0, \sigma_u^2)$). On peut alors introduire ce terme dans l'estimation, en ajoutant au modèle les moyennes intra-groupes des variables individuelles. Le modèle devient donc :

$$y_{i,j} = \beta_0 + x_{ij}\beta + x_j\gamma + \underbrace{\bar{x}_{.j}\delta + u_j}_{\alpha_j} + \varepsilon_{ij} \tag{9}$$

Cette solution a été proposée par Mundlak (1978) pour les données de panel et porte donc son nom. Dans d'autres disciplines, on parle parfois de "modèles mixtes" car ils consistent à mélanger des effets aléatoires avec des effets fixes au niveau groupe. On peut montrer que le modèle enrichi de toutes ces variables conduit au même estimateur de β que le modèle à effets fixes (autrement dit on retrouve l'estimateur *within* sans biais). On peut également tester directement l'hypothèse d'indépendance de ces variables individuelles avec les effets aléatoires, soit le test $\delta = 0$[9].

Cependant, cette méthode pas plus que le test d'Hausman ne permet pas de tester la possible dépendance des variables observables définies au niveau du groupe (i.e. de niveau 2) avec les effets inobservés. Ces paramètres ne peuvent être identifiés dans le modèle augmenté car évidemment confondus avec leur moyenne au niveau groupe. Ils ne sont donc pas estimés dans ce modèle et ne peuvent dont être comparés avec les estimations obtenues par le modèle à effets aléatoires dans un test d'Hausman. Le fait qu'il n'existe pas de test statistique pour évaluer la pertinence de cette hypothèse importante d'indépendance ne doit pas dissuader de la discuter. Bien au contraire, il faudra évaluer, au cas par cas, si elle paraît crédible. Si ce n'est pas le cas, il

[9] Snijders et Bosker (2011a) montrent qu'on peut choisir d'utiliser les variables centrées $x_{ij} - \bar{x}_{.j}$ à la place des variables individuelles : cela n'a pas d'incidence sur la valeur estimée de β (qui correspond à l'impact de la variable *intra*, soit au sein des groupes), mais permet d'estimer le coefficient de la variable moyenne comme l'impact de cette même variable *inter*, i.e. entre les groupes (comment est-ce que les différences de moyennes entre les groupes s'expliquent par les différences de composition selon cette dimension). Si on n'utilise pas les variables centrées, l'impact *inter* correspond à $\beta + \delta$.

pourra être préférable de rechercher d'autres méthodes d'estimation que celles présentées dans ce document. La plus classique, si l'on s'intéresse spécifiquement à l'effet de ces variables, serait par exemple d'utiliser des variables instrumentales... mais cela demande de disposer de telles variables. Une autre solution peut être d'estimer une procédure en deux étapes. La première étape consiste à estimer un modèle avec des effets fixes, dont les estimations sont utilisées dans une seconde étape, sur données agrégées, dans laquelle on régresse ces effets fixes estimés sur la moyenne des variables observables de niveau 2 agrégées. Il faut prendre garde cependant que dans ce type de régression qui repose sur une variable estimée (et donc mesurée avec erreur), le calcul de précision est plus complexe. Comme souligné plus haut, les estimations des effets fixes obtenues en première étape risquent d'autant moins d'être très éloignées de leur vraie valeur que le nombre d'observations par groupe est élevé.

1.3 Variable d'intérêt binaire

Dans de nombreux cas, la variable d'intérêt qu'on souhaite modéliser est dichotomique : le retard scolaire d'un élève, la défaillance d'une entreprise... Si les données sont structurées de manière hiérarchique, l'ensemble des points discutés restent valables, mais la non linéarité demande néanmoins des méthodes spécifiques qui, de fait, sont plus complexes. On commence par rappeler brièvement les principes de l'estimation des modèles dichotomiques classiques, dont on trouvera une présentation plus détaillée par exemple dans Afsa (2015) ou Le Blanc et al. (2000), avant de détailler comment ces questions se combinent avec la modélisation de plusieurs niveaux.

1.3.1 Généralités

Rappelons que lorsque la variable d'intérêt y est binaire, on modélise la probabilité d'observer l'occurrence 1 de y conditionnellement aux covariables. Deux spécifications sont classiquement utilisées. Elles consistent toutes deux à retenir une transformation d'une fonction linéaire de ces covariables. Plus précisément (dans le cas simple où les données ne sont pas groupées), on suppose que :

$$P(y_i = 1|x_i) = G(\beta_0 + x_i\beta)$$

où G correspond à la fonction de répartition de la loi logistique standard (modèle logit) ou normale standard (modèle probit). Il peut être utile de rappeler que cette modélisation se rationalise en supposant l'existence d'une variable latente y_i^* qui est liée à notre variable dichotomique observée par :

$$\begin{aligned} y_i^* &= \beta_0 + x_i\beta + \varepsilon_i \\ y_i &= \begin{cases} 1 \text{ si } y_i^* \geq 0 \\ 0 \text{ sinon} \end{cases} \end{aligned}$$

qui fait apparaître le terme de variabilité individuelle ε_i, dont la loi, logistique ou normale, définit le type de modélisation.

Lorsque les données sont hiérarchisées, la modélisation multiniveaux découle directement de ce modèle binaire classique, en incluant comme covariables des observables correspondant au premier niveau et/ou au deuxième niveau (respectivement x_{ij} et x_j) et des effets inobservables spécifiques au deuxième niveau α_j, soit :

$$p_{ij} = P(y_{ij} = 1|x_{ij}, x_j, \alpha_j) = G(\beta_0 + x_{ij}\beta_1 + x_j\beta_2 + \alpha_j)$$

Ou, en faisant apparaître la variable latente sous-jacente, le modèle s'écrit aussi :

$$\begin{aligned} y_{ij}^* &= \beta_0 + x_{ij}\beta_1 + x_j\beta_2 + \alpha_j + \varepsilon_{ij} \\ y_{ij} &= \begin{cases} 1 \text{ si } y_{ij}^* \geq 0 \\ 0 \text{ sinon} \end{cases} \end{aligned}$$

De même que dans un modèle binaire classique, la loi du résidu individuel ε_{ij} est soit logistique, soit normale.
Comme dans le modèle linéaire présenté plus haut, le terme α_j correspond à un terme inobservé

commun à l'ensemble des observations du groupe j. De même que dans le cas linéaire, on fait la distinction entre les modèles dits "à effets aléatoires", dans lesquels on suppose que ces "effets groupes" suivent une loi normale, et surtout qu'ils sont indépendants des autres variables explicatives, et les modèles "à effets fixes" qui reposent sur des hypothèses moins strictes (et qui donc ont plus de chances d'être vérifiées). Néanmoins, la non linéarité peut apporter une complexité supplémentaire d'une part pour l'estimation de ces modèles, mais surtout pour leur interprétation.

1.3.2 Estimation

Rappelons que l'estimation des modèles logit et probit "simples" se fait classiquement en maximisant la vraisemblance du modèle, qui découle de la spécification retenue pour le résidu. Cependant, l'ajout des termes inobservés au niveau groupe rend les choses plus complexes. Même sous les hypothèses restrictives du modèle à effets aléatoires, sous lesquelles on suppose que les effets groupes suivent une loi normale indépendante des covariables observables, la solution du programme de maximisation de la vraisemblance n'a pas une forme analytique explicite en fonction des paramètres et ne peut qu'être approchée par des méthodes numériques (voir partie 3). Si ces hypothèses ne paraissent pas vraisemblables, l'estimation d'un modèle à effets fixes est possible mais également plus complexe. Estimer des effets fixes en ajoutant des indicatrices pour chaque groupe peut s'avérer une très mauvaise idée lorsqu'on dispose de peu d'observations par groupe. En effet, comme dans le modèle linéaire, les estimations de ces effets fixes seront de mauvaise qualité. Mais en plus, ces estimateurs biaisés "contamineront" l'estimation des coefficients des autres variables. Ce problème est connu sous le nom de *problème des paramètres incidents*. Le modèle n'étant pas linéaire, une simple différentiation ne suffit pas à faire "disparaître" les effets groupes. On peut montrer qu'il est possible sous certaines conditions d'estimer un modèle logistique à effets fixes, mais contrairement au cas linéaire, ce modèle ne permet pas d'obtenir des estimations des effets groupes : du fait de la non linéarité, il n'est pas possible de revenir à ces estimateurs. Cette limite a des conséquences sur l'interprétation des résultats.

1.3.3 Interprétation : les coefficients estimés

Les coefficients estimés des variables correspondent strictement à l'effet d'une variable sur la transformée de la probabilité $G^{-1}(p_{ij})$, ce qui ne fournit pas une interprétation très intuitive des résultats de l'estimation. Dans le cas d'un modèle logistique, il est cependant possible d'en obtenir une interprétation plus parlante, puisque l'exponentielle de chaque coefficient fournit une évaluation de l'effet d'une variable sur le rapport de chances (*odds ratio*) de notre variable d'intérêt : par exemple, la probabilité d'avoir déjà redoublé sur celle de ne pas avoir redoublé pour une fille comparé à ce même rapport de probabilités pour un garçon, les autres caractéristiques ayant été prises en compte.

Les valeurs des coefficients d'un modèle multiniveaux binaire sont à interpréter avec précaution car leur valeur dépend directement des choix de modélisation et notamment de la contrainte identifiante qui consiste à fixer la variance des résidus individuels. Elle vaut classiquement $\pi^2/3$

dans le modèle logit (variance d'une loi logit standard) et 1 dans le modèle probit[10]. De ce fait, si les deux modèles fournissent en général des résultats qualitativement semblables, les coefficients estimés par un logit seront plus grands (en valeur absolue) que ceux d'un probit, approximativement d'un coefficient multiplicatif $\sqrt{\pi^2/3} \approx 1,8$. Il faut insister sur le fait que dans ce type de modèle, la valeur précise de l'estimation du coefficient d'une variable n'a pas d'interprétation en soi. Ce sont les signes de ces coefficients, et leurs valeurs les uns par rapport aux autres, qui sont pertinents pour l'analyse.

Le fait de contraindre la valeur de la variance individuelle dans la modélisation multiniveaux modifie aussi les coefficients estimés par rapport à ceux estimés avec un modèle binaire classique. Rappelons que la variance d'une observation individuelle prise au hasard correspond à $var(y^*_{ij}) = \sigma^2_\varepsilon + \sigma^2_\alpha$, alors que dans un modèle binaire classique elle vaut σ^2_ε (i.e. 1 ou $\pi^2/3$ selon le modèle). En pratique, cela signifie que si on compare les résultats obtenus en spécifiant une structure groupée avec ceux obtenus par un modèle "simple" qui néglige ces effets, les coefficients correspondant aux différentes variables explicatives seront mécaniquement dilatés par un facteur $\sqrt{\frac{\sigma^2_\alpha+\sigma^2_\varepsilon}{\sigma^2_\varepsilon}}$ dans le modèle à effets aléatoires par rapport au modèle simple. Cette dilatation est mécanique et simplement induite par la contrainte imposée pour l'estimation du modèle sur la variance individuelle.

Du fait de la contrainte identifiante, il est également plus difficile d'interpréter les variations des coefficients lorsqu'une variable est incluse ou retirée du modèle. Dans un modèle binaire classique, les coefficients des variables ne changent pas lors de l'ajout d'une variable qui ne leur est pas corrélée. Ce n'est pas le cas dans un modèle multiniveaux. Par exemple, l'ajout d'une variable explicative de niveau 1 non corrélée à l'effet groupe fait augmenter mécaniquement la variance inter-classe et a aussi un impact sur les coefficients des autres variables explicatives. Pour des variables qui ne sont pas corrélées à cette variable (et donc pour lesquelles l'introduction de cette variable sur la modélisation est *a priori* neutre), on observera une augmentation des coefficients estimés, approximativement d'un facteur $\sqrt{\tilde{\sigma}^2_\alpha/\sigma^2_\alpha}$, où $\tilde{\sigma}^2_\alpha$ est la variance inter-classe du modèle augmenté. Ils sont en effet proportionnels à la variance totale du modèle auquel on a ajouté des variables explicatives supplémentaires. On se gardera ici encore de surinterpréter ces évolutions en termes de "pouvoir explicatif" de ces variables. La même prudence est de mise dans l'interprétation de la corrélation intra-classe (*ICC*).

1.3.4 Interprétation : les effets marginaux

Comme on l'a vu, la valeur du coefficient d'une variable est sensible à la modélisation retenue et aux autres variables incluses dans le modèle. Dans un modèle logit, son exponentielle

[10] En l'absence de cette contrainte, le modèle n'est pas identifiable : il n'y a pas unicité des coefficients. Le modèle pourrait en effet être réécrit sous différentes formes aboutissant à des jeux de coefficients différents :

$$y_{ij} = \mathbb{1}(\beta_0 + x_{ij}\beta_1 + x_j\beta_2 + \alpha_j + \varepsilon_{ij} \geq 0)$$

ou encore, en multipliant l'inégalité par un coefficient K strictement positif :

$$y_{ij} = \mathbb{1}(\tilde{\beta}_0 + x_{ij}\tilde{\beta}_1 + x_j\tilde{\beta}_2 + \tilde{\alpha}_j + \tilde{\varepsilon}_{ij} \geq 0) \quad \text{avec } (\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}_j, \tilde{\varepsilon}_{ij}) = (K\beta_0, K\beta_1, K\beta_2, K\alpha_j, K\varepsilon_{ij})$$

Ce qui conduit à deux jeux de coefficients différents. Fixer la variance inter-classe assure l'unicité des coefficients estimés.

correspond à l'effet de la variable sur le rapport des chances [11] mais pas directement sur cette probabilité. On ne pourra donc pas directement mesurer, par exemple, l'impact du fait d'être une fille ou un garçon sur la probabilité d'être en retard scolaire (le signe du coefficient indique tout de même si le fait d'être une fille augmente cette probabilité par rapport au fait d'être un garçon). Pour cela, on peut revenir aux probabilités prédites, ou plus précisément estimer l'effet marginal d'une variable explicative du modèle (comment cette probabilité prédite varie en fonction de cette variable explicative). Cependant, du fait de la non linéarité, l'effet marginal d'une variable n'est pas le même selon les valeurs prises par les autres covariables : par exemple, le fait d'être une fille n'aura pas le même effet sur la probabilité d'être en retard scolaire selon que l'élève vient d'un milieu favorisé ou non. Dans le cadre des modèles multiniveaux, l'effet marginal dépend en plus du groupe.

Formellement, l'effet marginal (souvent noté *PE* pour *Partial Effect*) d'une variable binaire x_k correspond à la différence entre la probabilité d'observer $y = 1$ selon que la valeur explicative x_k vaut 0 ou 1, les autres variables étant fixées :

$$PE_k = P(y_{ij} = 1|x_{ij}^{-k}, x_j^{-k}, x_{ij}^k = 1, \alpha_j) - P(y_{ij} = 1|x_{ij}^{-k}, x_j^{-k}, x_{ij}^k = 0, \alpha_j) \quad (10)$$

$$= G(\beta_0 + x_{ij}^{-k}\beta_1^{-k} + x_j^{-k}\beta_2^{-k} + \beta_{\mathbf{k}} + \alpha_j) - G(\beta_0 + x_{ij}^{-k}\beta_1^{-k} + x_j^{-k}\beta_2^{-k} + \alpha_j) \quad (11)$$

où x_{ij}^{-k} et x_j^{-k} correspondent à l'ensemble des variables observables respectivement de niveaux 1 et 2, à l'exclusion de la variable x_k à laquelle on s'intéresse (et qui peut être indistinctement de niveau 1 ou 2), et β_1^{-k} et β_2^{-k} les coefficients correspondants.

Ou pour une variable continue :

$$PE_k = \frac{\partial P(y_{ij} = 1|x_{ij}, x_j, \alpha_j)}{\partial x_k} = \beta_k g(\beta_0 + x_{ij}\beta_1 + x_j\beta_2 + \alpha_j) \quad (12)$$

où g correspond à la densité de la distribution choisie pour le résidu individuel (i.e. la dérivée de G).

On trouvera dans Afsa (2015) une discussion détaillée sur les modes de calcul des effets marginaux. Par rapport à un modèle binaire simple, la complexité supplémentaire vient du fait qu'il faut aussi tenir compte des effets groupes. Dans le cas du modèle à effets aléatoires en revanche, on peut estimer l'effet moyen d'une variable, défini en intégrant l'effet marginal de cette variable tel que décrit plus haut sur la distribution des effets groupes.

Formellement, l'effet marginal moyen (souvent noté *APE* pour *Average Partial Effect*) pour une variable x_k , les valeurs des autres variables étant fixées, correspond à :

$$APE_k = \int_\alpha PE_k(\alpha) \frac{1}{\sigma_\alpha} \varphi\left(\frac{\alpha}{\sigma_\alpha}\right) d\alpha$$

avec φ la densité d'une loi normale standard. Dans le cas général, cette intégrale n'a pas de forme analytique explicite.

En pratique, deux méthodes sont souvent utilisées pour la calculer. **La première méthode** consiste à définir pour le calcul une situation de référence : on fixe une valeur pour chaque variable, qui peut correspondre pour les variables continues à la valeur moyenne observée sur

[11] *odd ratio*, qui correspond au rapport de deux rapports : au dénominateur la probabilité d'observer la valeur un pour la variable d'intérêt sur la probabilité de ne pas l'observer pour une certaine valeur de la variable, et au numérateur ce même rapport pour la valeur de la variable augmentée d'une unité.

l'échantillon, ou, comme pour les variables catégorielles, être choisie *a priori*. Dans ce cas, l'interprétation des résultats sera valable pour les personnes présentant ces caractéristiques de référence. On peut alors calculer l'effet marginal pour un groupe donné, qui peut être exprimé par

$$PE_k(x_{R,1}^{-k}, x_{R,2}^{-k}, \alpha_j) = P(y_{ij} = 1 | x_{R,1}^{-k}, x_{R,2}^{-k}, x_k = 1, \alpha_j) - P(y_{ij} = 1 | x_{R,1}^{-k}, x_{R,2}^{-k}, x_k = 0, \alpha_j)$$

pour une variable x_k catégorielle, en notant $x_{R,1}^{-k}$ le vecteur des valeurs de référence pour les variables de niveau 1 et $x_{R,2}^{-k}$ pour celle de niveau 2 .

Ou pour une variable continue :

$$PE_k(x_{R,1}^{-k}, x_{R,2}^{-k}, \alpha_j) = \frac{\partial P(y_{ij} = 1 | x_{R,1}, x_{R,2}, \alpha_j)}{\partial x_k} \tag{13}$$

Pour estimer la moyenne sur l'ensemble des groupes, on peut procéder par simulation par une méthode de Monte-Carlo[12] : on génère M réalisations aléatoires d'une loi normale centrée et de variance $\hat{\sigma}_\alpha^2$, notées $a_1, a_2, \ldots, a_M$, et on calcule l'effet moyen comme la moyenne empirique de ces effets calculés sur l'ensemble des tirages aléatoires, soit :

$$APE_k(x_{R,1}^{-k}, x_{R,2}^{-k}) = \frac{1}{M} \sum_{m=1}^{M} PE_k(a_m)$$

qui, par exemple pour une variable continue, se développe ainsi :

$$APE_k(x_{R,1}^{-k}, x_{R,2}^{-k}) = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_k g(\hat{\beta}_0 + x_{R,1}\hat{\beta}_1 + x_{R,2}\hat{\beta}_2 + a_m)$$

La seconde méthode pour estimer l'effet marginal moyen est de prendre la moyenne sur l'ensemble de l'échantillon des effets. Cette méthode évite le choix, forcément arbitraire, d'une situation de référence, et permet d'obtenir un estimateur de l'effet moyen sur l'échantillon qui tienne compte des caractéristiques de cet échantillon. Plus précisément, on utilise, pour chaque observation, les valeurs observées des autres variables que celle dont on veut mesurer l'effet, et on prend la moyenne sur l'échantillon des effets marginaux "individuels" ainsi calculés. Cet effet marginal de la variable continue k pourra donc se calculer comme :

$$APE_k = \frac{1}{N} \sum_{j=1}^{J} \sum_{i \in j} \hat{\beta}_k g(\hat{\beta}_0 + x_{ij}\hat{\beta}_1 + x_j\hat{\beta}_2 + \hat{\alpha}_j)$$

Et si la variable x_k est binaire :

$$APE_k = \frac{1}{N} \sum_{j=1}^{J} \sum_{i \in j} G(\hat{\beta}_0 + x_{ij}^{-k}\hat{\beta}_1^{-k} + x_j^{-k}\hat{\beta}_2^{-k} + \hat{\beta}_\mathbf{k} + \hat{\alpha}_j) - G(\hat{\beta}_0 + x_{ij}^{-k}\hat{\beta}_1^{-k} + x_j^{-k}\hat{\beta}_2^{-k} + \hat{\alpha}_j) \tag{14}$$

Cependant, dans le cas d'un modèle à effets fixes, l'estimation du modèle se limite à l'effet des variables explicatives mais ne permet pas d'obtenir une estimation des termes α_j (et on ne fait par ailleurs pas d'*a priori* sur cette distribution des effets groupes). Il n'est donc pas possible

[12] C'est par exemple la procédure utilisée par le logiciel MLwin.

d'estimer cette intégrale et donc de calculer les effets moyens d'une variable sur l'échantillon[13]. En pratique, si on utilise ces modèles, on commentera les résultats en termes de rapport de chances (telle variable augmente de tant le rapport de la probabilité d'observer $y = 1$ sur celle d'observer $y = 0$).

2 Extensions

2.1 Des effets variables selon les groupes

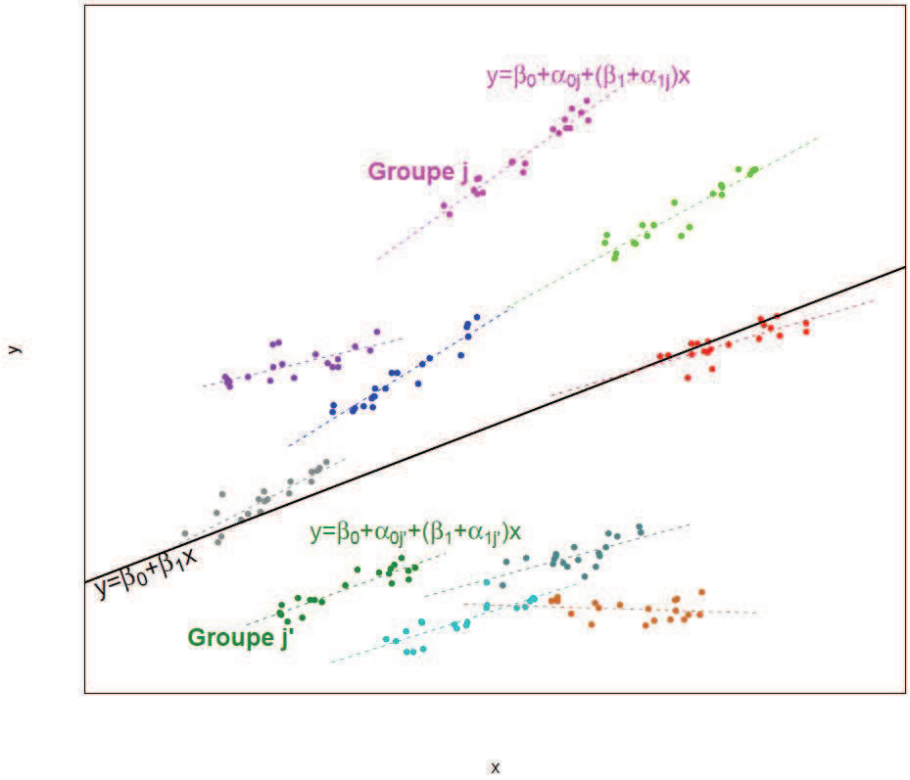
Jusqu'à présent on a introduit la variabilité liée au groupe uniquement dans le terme constant. Il est fréquent dans les modèles à effets aléatoires d'introduire également de la variabilité dans les coefficients correspondant aux variables explicatives individuelles. Par exemple, la pratique d'un enseignant (ce qu'on appelle "l'effet maître" en sciences de l'éducation) pourra être plus bénéfique pour les bons élèves que pour ceux en difficulté, ou inversement (voir par exemple Bressoux, 2007). Une première manière de modéliser cette hétérogénéité est d'introduire des termes d'interaction entre chaque groupe (l'enseignant dans notre exemple) et les variables individuelles (le niveau scolaire initial dans cet exemple). Cela correspond à étendre la modélisation à effets fixes à ce cas. Cependant, si le nombre d'observations par groupes est faible, les paramètres d'hétérogénéité risquent d'être mal estimés. Et par ailleurs, les résultats peuvent être difficiles à interpréter. C'est pourquoi dans cette littérature il est plutôt d'usage d'utiliser une spécification à effets aléatoires. Il est en effet très simple d'étendre le modèle à effets aléatoires présenté précédemment pour rendre compte de l'hétérogénéité des effets. Formellement, cela revient à écrire :

$$\begin{cases} y_{ij} = \beta_{0j} + x_{ij}\beta_{1j} + x_j\beta_2 + \varepsilon_{ij} \quad j = 1,\dots,J \quad i = 1,\dots,n_j \\ \beta_{0j} = \beta_0 + \alpha_{0j} \\ \beta_{1j} = \beta_1 + \alpha_{1j} \end{cases} \tag{15}$$

Ce modèle est illustré par le graphique 5. Le terme β_1 va correspondre à l'effet moyen des variables individuelles, tandis que α_{1j} correspond à l'écart de chaque groupe à cette relation moyenne. On se place dans le cadre d'un modèle à effets aléatoires, et les termes α_{0j} et α_{1j} sont supposés distribués selon une loi normale de moyenne nulle et de variance respectivement σ_0^2 et σ_1^2. Comme auparavant, on suppose que les termes aléatoires (ε_{ij}, α_{0j} et α_{1j}) sont indépendants entre des observations appartenant à des groupes distincts. En revanche, on peut supposer qu'il existe une corrélation entre les termes aléatoires α_{0j} et α_{1j} au sein d'un même groupe, et on note σ_{01} cette covariance. Une étape supplémentaire dans la complexification du modèle est d'expliquer ces pentes par des variables explicatives, ce qui correspondrait à des termes d'interactions entre des variables de niveau 1 et de niveau 2. On pourrait par exemple écrire : $\beta_{1j} = \beta_1 + \alpha_{1j} + z_j\gamma$.

Pour intéressantes que ces extensions soient en principe, la pratique peut se heurter à des problèmes d'estimation car le nombre de paramètres devient vite important (si on a k coefficients aléatoires, on doit estimer $k(k+1)/2+1$ paramètres qui correspondent à la matrice de

[13] Il est possible de calculer des effets marginaux en fixant comme "valeur de référence" une valeur nulle pour ces termes. Cette solution n'est cependant pas satisfaisante puisqu'elle fait l'impasse sur l'existence de ces effets groupes qui peuvent être importants (d'autant qu'ils comprennent dans ce cas les variables explicatives de niveau 2). Elle n'est par ailleurs pas vraiment cohérente : la raison principale qui justifie d'utiliser un modèle à effets fixes est qu'on pense probable une corrélation entre les variables observables et les effets groupes. Les négliger finalement pour calculer l'effet des variables n'est donc pas satisfaisant.



Graphique 5 – Illustration du modèle (15)

variance-covariance de ces termes aléatoires) ; si le nombre d'observations par groupe est faible, ou si les variables explicatives qu'on utilise varient peu au sein des groupes, l'estimation sera également peu fiable. Enfin et surtout, l'interprétation des coefficients peut être difficile. Même si l'estimation de ce type de modèles est techniquement possible, leur validité reste, ici encore, tributaire d'hypothèses d'indépendance entre effets aléatoires et variables observées qui sont très fortes... et ont donc beaucoup de risques de ne pas être vérifiées.

Le lecteur intéressé par ces développements pourra se référer au chapitre 13 de Gelman et Hill (2007).

2.2 Plus de deux niveaux

Le principe des modèles hiérarchiques s'étend très naturellement au cas où les données sont structurées selon plus de deux niveaux. On peut par exemple penser à des observations disponibles pour des élèves (niveau 1), compris dans une classe (niveau 2) au sein d'un établissement scolaire (niveau 3), chacun de ces niveaux apportant une dimension particulière à l'analyse. Un autre cas notable à trois niveaux est celui où on disposerait par exemple de données de panel (la date d'observation est le niveau 1) pour des élèves (niveau 2) dans certains établissements (niveau 3). La transposition de l'analyse précédente se fait en théorie sans difficulté. Formellement, cela peut s'écrire sous la forme :

$$y_{ijk} = \beta_0 + x_{ijk}\beta_1 + x_{jk}\beta_2 + x_k\beta_3 + \alpha_k + \alpha_{jk} + \varepsilon_{ijk} \tag{16}$$

L'indice i désigne les observations de niveau 1, j celles de niveau 2 et k celles de niveau 3.

Légèrement plus complexe est le cas où ces différents niveaux ne sont plus emboîtés. C'est le cas par exemple lorsqu'on dispose de données de panel sur les rémunérations de plusieurs salariés (voir par exemple Abowd et al., 1999). On peut souhaiter contrôler des effets individuels de la rémunération (on dispose de plusieurs observations répétées pour un même salarié), mais aussi de l'effet spécifique à l'entreprise qui l'emploie (à une date donnée, on dispose de l'ensemble des salariés de cette entreprise). Les données ne sont cependant pas parfaitement emboîtées : les salariés peuvent changer d'entreprise au cours du temps. De même, on peut penser qu'au-delà de l'effet spécifique à un établissement scolaire, le quartier dans lequel l'élève réside a aussi des caractéristiques inobservables qui influent sur sa réussite scolaire (voir par exemple Raudenbush, 1993). Ces deux effets de contexte sont intéressants et on peut souhaiter modéliser les deux simultanément, mais il ne s'agit souvent pas de niveaux parfaitement emboîtés, dans la mesure où il est assez fréquent que des élèves d'un même quartier fréquentent des établissements différents.

Il est tout à fait possible d'étendre les méthodes présentées jusqu'à présent à ces cas de figures, mais les modèles peuvent être plus complexes à estimer. La prise en compte de modèles avec deux effets fixes a été beaucoup développée dans le cas des données de panel, et il existe de très nombreux estimateurs. Parmi les plus récents, on citera Guimarães et Portugal (2010) qui offrent une procédure intensive en calculs, mais qui demande peu de ressources mémoires (ce qui peut être souhaitable lorsque le nombre des observations est très élevé), ou Gaure (2013b). Ces auteurs ont développé des packages d'implémentation de ces méthodes (respectivement `reg2hdfe` sous Stata, voir Guimarães (2009) et `lfe` sous R, voir Gaure (2013a)). L'estimation est délicate lorsque la variable est binaire, du fait du problème des paramètres incidents déjà cités. On trouvera une méthode par exemple dans Hospido (2012). Il est également possible d'estimer des modèles à doubles effets aléatoires. On en trouvera une description par exemple dans Snijders et Bosker (2011b). Cependant, si en principe cette estimation se fait très simplement dans la plupart des logiciels statistiques, comme souvent, la complexité a un prix en termes computationnels. Il est donc plus raisonnable dans une première approche de tester séparément les deux modèles. La décision d'utiliser les deux dimensions dépendra évidemment de la corrélation supposée entre les deux niveaux de regroupement. Si on a des raisons de supposer que ces effets sont corrélés, lorsqu'on recourt à une modélisation classique en ne retenant qu'un seul des niveaux de regroupement celui-ci risque de capter une partie de l'effet de l'autre niveau, négligé dans la spécification.

3 En pratique

3.1 Principes de l'estimation

3.1.1 Modèle à effets fixes

Lorsque la variable d'intérêt est continue le modèle à effets fixes correspond à un modèle linéaire standard (qui comprend les indicatrices de groupes comme variables explicatives) et peut donc être estimé par des méthodes de régressions linéaires classiques. Cependant, comme souligné plus haut, cette estimation repose en principe sur un grand nombre de paramètres, puisqu'il y a un effet fixe par groupe. On utilise donc plutôt des estimateurs différentiés qui permettent d'éliminer les indicatrices de groupe dans une première étape. L'estimation se fait ensuite simplement par une méthode de moindres carrés ordinaires, mais il faut ensuite corriger les estimateurs de la variance pour tenir compte de la transformation initiale (voir page 13). On

trouvera par exemple dans Duguet (1999) ou Wooldridge (2002) une description de ces modèles, classiques en économétrie des panels.

Lorsque la variable d'intérêt est binaire en revanche ces méthodes ne s'appliquent plus : une simple différentiation ne permet pas de faire disparaître les effets fixes. Du fait de la non linéarité, introduire des indicatrices par groupe peut sérieusement biaiser l'ensemble des estimateurs. Comme montré par Chamberlain (1980), sous l'hypothèse que la variable suit une loi logistique, il existe une transformation du modèle qui permet de faire disparaître les effets fixes. On peut en effet montrer que la probabilité qu'une observation ait une valeur $y_{ij} = 1$ *conditionnelle* au nombre d'observations dans son groupe qui ont ce même comportement, est indépendante de l'effet fixe α_j (pour plus de détails voir par exemple Davezies, 2011). On utilise donc une méthode de maximum de vraisemblance conditionnelle, raison pour laquelle ce modèle est appelé logit conditionnel. L'estimation ne porte que sur les groupes dans lesquels on observe à la fois des observations avec $y_{ij} = 0$ et $y_{ij} = 1$, ce qui nécessite donc d'avoir suffisamment de variabilité dans les données.

3.1.2 Modèle à effets aléatoires

Pour discuter de l'estimation du modèle à effets aléatoires, il est utile de remarquer que dans le cas où la variable est continue le modèle correspondant à l'équation 1 peut se réécrire sous forme matricielle (les observations étant ordonnées par groupes) :

$$Y = X\beta + U \tag{17}$$

avec X l'ensemble des variables observables (de niveau 1 et 2, y compris la constante), U le vecteur de résidu $U = Z\alpha + \varepsilon$ où Z est une matrice correspondant aux effets groupes[14] et le vecteur de résidu suit $U|X,Z \sim \mathcal{N}(0,V)$ où $V = \sigma_\varepsilon^2 I_N + \sigma_\alpha^2 Z'Z$ est une matrice bloc-diagonale dont le terme diagonal générique est constitué de la matrice carrée de taille n_j et de terme générique $\sigma_\alpha^2 + 1_{jj}\sigma_\varepsilon^2$ (1_{jj} représentant une indicatrice valant 1 sur la diagonale). Cette écriture permet de mettre en avant pourquoi il n'est pas efficace d'utiliser les moindres carrés ordinaires pour estimer les paramètres β correspondant aux variables observables : le modèle n'est pas homoscédastique, c'est-à-dire qu'il existe une corrélation dans les termes inobservés. L'estimateur des moindres carrés ordinaires est certes sans biais, mais il n'est pas le plus efficace, au sens où sa variance n'est pas la plus faible possible. Surtout, l'estimation de cette variance telle que fournie par la méthode des moindres carrés ordinaires, qui repose sur cette hypothèse d'homoscédasticité, sera biaisée puisqu'elle ignore la corrélation intra-groupe. Plusieurs types de méthodes peuvent être utilisées pour corriger ce problème.

La première famille de méthodes est celle des *moindres carrés quasi-généralisés*. Ils reposent sur le fait que le modèle (17) transformé par $V^{-1/2}$ est homoscédastique et peut donc s'estimer par les moindres carrés ordinaires (on parle alors de moindres carrés généralisés). Les variances σ_ε^2 et σ_α^2 sont cependant inconnues, et on utilise une méthode reposant sur une approximation de ces termes. En pratique, on procède de manière itérative en plusieurs étapes, qui permettent d'approcher σ_ε^2 et σ_α^2. En première étape, on peut par exemple initialiser la méthode

[14] Elle a donc N lignes et J colonnes, la colonne j étant constituée de 1 uniquement pour les observations correspondantes au groupe j, soit $(\underbrace{0\dots0}_{\sum_{k=0}^{n_{j-1}} n_k}, \underbrace{1\dots1}_{n_j}, \underbrace{0\dots0}_{\sum_{k=n_{j+1}}^{N} n_k})'$.

par une estimation par les moindres carrés ordinaires : les résidus obtenus permettent d'obtenir des premières approximations $\hat{\sigma}^2_{\varepsilon 0}$ et $\hat{\sigma}^2_{\alpha 0}$, à partir desquelles on peut calculer en utilisant la racine carrée de l'inverse de la matrice correspondante $\hat{V}_0$ le modèle transformé par les moindres carrés généralisés. Celui-ci permet d'obtenir une nouvelle approximation des variances résiduelles $\hat{\sigma}^2_{\varepsilon 1}$ et $\hat{\sigma}^2_{\alpha 1}$, et ainsi de suite (l'estimation s'arrête lorsque la différence entre les valeurs obtenues entre deux itérations successives est inférieure à un critère de précision que l'on définit)[15].

La seconde famille de méthodes correspond à des méthodes de maximum de vraisemblance : sous les hypothèses du modèle à effets aléatoires, qui supposent à la fois l'indépendance des variables observables et des termes inobservés, et la normalité de ces derniers, il est simple d'écrire la vraisemblance de chaque observation en fonction des paramètres que l'on souhaite estimer. Là encore, la méthode d'estimation est itérative. En pratique on préfère en général utiliser pour les modèles multiniveaux une méthode dérivée de la méthode du maximum de vraisemblance classique, le maximum de vraisemblance restreint (*restricted estimator maximum likelihood, REML*). Cette estimation par maximum de vraisemblance restreint consiste à ne plus estimer simultanément l'ensemble des paramètres comme dans l'estimateur de maximum de vraisemblance classique, mais à procéder en deux étapes. La première étape consiste à estimer les paramètres correspondant aux termes de variances résiduelles, dont l'estimation est ensuite utilisée dans une deuxième étape pour estimer les paramètres "fixes" correspondant aux observables $\hat{\beta}$. Cette méthode est en général jugée préférable, car elle produit des estimateurs plus fiables lorsque les effectifs par groupe sont faibles. En revanche, elle est plus complexe.

Enfin, les méthodes d'estimation bayésiennes sont également devenues courantes pour l'estimation des modèles multiniveaux. Elles consistent à partir d'une distribution *a priori* sur les valeurs des paramètres, qui sera corrigée en fonction de l'observation. Le résultat final correspond bien à l'idée qu'on se fait d'une estimation.

Lorsque la variable d'intérêt est binaire, ces différentes méthodes ne s'appliquent plus directement. La vraisemblance est beaucoup plus complexe à estimer car on ne peut pas la décomposer en fonction des différents paramètres comme dans le cas linéaire. Pour s'en rendre compte, rappelons la forme que prend la vraisemblance du modèle. Dans le cas d'un modèle logistique, pour l'ensemble des observations d'un groupe j, la vraisemblance des observations conditionnellement à l'ensemble des explicatives $\mathbf{x}$ et à l'effet groupe α_j s'écrit :

$$P(y_{1j} \dots y_{1n_j} | \mathbf{x}_j, \alpha_j) = \Pi_{i=1}^{n_j} \frac{\exp(\beta_0 + \beta_1 x_{ij} + \beta_2 x_j + \alpha_j)^{y_{ij}}}{1 + \exp(\beta_0 + \beta_1 x_{ij} + \beta_2 x_j + \alpha_j)} \tag{18}$$

Cependant, par définition le terme d'effet groupe α_j est inobservé. Sous l'hypothèse de normalité de ces effets, il est possible d'obtenir une expression de la vraisemblance conditionnelle uniquement aux variables observées :

$$P(y_{1j} \dots y_{1n_j} | \mathbf{x}_j) = \int P(y_{1j} \dots y_{1n_j} | \mathbf{x}_j, \alpha_j) \varphi(\alpha_j) d\alpha_j \tag{19}$$

[15] On peut aussi remarquer que si on s'intéresse exclusivement aux effets des variables explicatives et non aux effets groupes (c'est souvent le cas dans les données de panel), on peut utiliser des méthodes dites robustes à l'hétéroscédasticité par cluster (groupe), qui sont par exemple implémentées très simplement dans un logiciel comme Stata. Les estimateurs de la précision des estimateurs seront non biaisés à condition que le nombre de groupes soit élevé.

où φ correspond à la densité des termes inobservés (la loi normale sous les hypothèses classiques du modèle aléatoire), mais du fait de l'intégrale il n'existe pas de forme analytique simple de la dérivée de cette vraisemblance, et il est nécessaire de recourir à des méthodes numériques pour l'approximer. Les deux méthodes principales consistent soit à approcher la vraisemblance par une approximation linéaire, développée en série de Taylor (méthode de pseudo maximum de vraisemblance), soit à utiliser une approximation numérique de l'intégrale (on discrétise l'intégrale en une somme discrète). Des méthodes bayésiennes peuvent également être utilisées. Le fait que l'estimation repose sur des approximations numériques a une conséquence, puisque les résultats peuvent différer d'une méthode à une autre. Li et al. (2011) proposent une comparaison des différentes méthodes obtenues par la plupart des logiciels statistiques et montrent que, pour des données de taille suffisante, les résultats des estimations sont en général assez proches (mais certains sont plus performants en termes de temps de calcul ou de flexibilité) : les procédures GLIMMIX de SAS et *lm*4 de R apparaissent parmi les plus performantes (avec le [*R*]*IGLS* du logiciel spécialisé MLWin). Sur les petits échantillons en revanche, ils obtiennent des résultats plus fluctuants (en particulier pour les méthodes bayésiennes). Sur un échantillon comprenant de nombreux "singletons" et une forte variance intra, Rodríguez et Goldman (1995) obtiennent par exemple des écarts conséquents selon la méthode d'estimation.

3.2 Dans les logiciels statistiques

On présente ici des instructions permettant de mettre en œuvre des modèles à effets fixes ou aléatoires, dans les trois principaux logiciels statistiques (SAS, R et STATA). On trouvera en annexe l'application de ces programmes aux deux exemples développés dans la partie suivante.

3.2.1 Avec Sas

Effets fixes L'estimation des modèles à effets fixes peut se faire, pour les variables continues, avec la procédure `GLM`. La syntaxe est proche de la procédure `REG` :

```
PROC SORT DATA=table;BY ident_niv2;RUN;
PROC GLM DATA=table;
ABSORB ident_niv2;
CLASS  var_cat;
MODEL y = var_num var_cat ident_niv2 /SOLUTION;
RUN;
```

L'option `ABSORB`, qui nécessite que la table soit préalablement triée selon la variable identifiant le niveau 2 (ici `ident_niv2`), permet de calculer les variables transformées pour le modèle *within*, qui "élimine" les effets groupes. Si on souhaite plutôt estimer l'ensemble des effets groupes (plutôt dans le cas où on dispose de suffisamment d'observations par groupe, et surtout où ces groupes ne sont pas trop nombreux, car on peut rapidement rencontrer des problèmes de mémoire), on omettra cette option et on indiquera simplement la variable identifiant le niveau 2 dans l'option `CLASS`. Rappelons que cette dernière est utilisée pour introduire dans le modèle des variables catégorielles (la procédure crée directement le nombre d'indicatrices correspondant à chacune des modalités de la variable catégorielle).

Lorsque la variable d'intérêt est binaire, on utilisera la procédure `LOGISTIC`. La syntaxe est la même que pour un modèle binaire classique. La différence est l'instruction `STRATA` où on indique la variable identifiant le niveau 2. La syntaxe peut par exemple s'écrire :

```
PROC LOGISTIC DATA= table;
CLASS var_cat;
MODEL y(DESCENDING)=var_num var_cat ;
STRATA ident_niv2;
RUN;
```

Rappelons que l'estimation avec effets fixes ne peut se faire qu'avec une spécification logistique : l'option `STRATA` permet d'estimer un logit conditionnel, qui permet de se "débarrasser" des effets groupes dans l'estimation et donc d'éviter le problème des paramètres incidents. Du fait de la forme non linéaire cependant, il n'est pas possible à partir de ces estimations d'obtenir une approximation des effets groupes. Si on souhaite estimer un effet marginal d'une variable, cela ne pourra se faire que conditionnellement à un effet groupe (ce qui limite l'interprétation qu'on peut en faire).

Effets aléatoires Plusieurs procédures permettent d'estimer des modèles à effets aléatoires. La procédure `GLIMMIX` est la plus générale. Pour une variable d'intérêt continue, la syntaxe sera :

```
PROC GLIMMIX DATA= table ;
CLASS ident_niv2 var_cat;
MODEL y=var_num var_cat/ SOLUTION;
RANDOM INTERCEPT / SUBJECT = ident_niv2 SOLUTION;
RUN;
```

Lorsque la variable d'intérêt est binaire, on ajoute l'option `DIST=BINARY` dans l'instruction `MODEL` :

```
MODEL y (DESCENDING)=var_num var_cat/DIST=BINARY LINK=logit SOLUTION;
```

La distribution logistique est celle retenue par défaut sous SAS ; `LINK=probit` permet de spécifier une modélisation probit. Pour une variable y codée 0/1, on modélise la probabilité de succès $y = 1$, plutôt que $y = 0$ (par défaut), par l'option `DESCENDING` ou de manière équivalente par `event="1"`.

L'instruction `RANDOM` est celle qui permet de définir la forme des effets aléatoires. On indique dans l'instruction `SUBJECT=` la variable identifiant le niveau 2. Dans l'exemple présenté ci-dessus, l'effet aléatoire ne porte que sur la constante, indiqué par `INTERCEPT` (ou `INT`). Des modèles plus complexes pourraient inclure des effets aléatoires sur les coefficients des variables explicatives, ce qui se fait simplement en listant ces variables dans l'instruction `RANDOM` (avec les précautions nécessaires indiquées dans la section 2.1). Pour définir plusieurs niveaux de regroupement (par exemple le quartier en plus de l'école), deux instructions `RANDOM` seront nécessaires et s'écriront ainsi :
selon que les deux niveaux sont emboîtés :

```
RANDOM INTERCEPT / subject=ident_niv3;
RANDOM INTERCEPT / subject=ident_niv2 (ident_niv3);
```

Ou qu'ils ne sont pas emboîtés :

```
RANDOM INTERCEPT / subject=ident_niv2;
RANDOM INTERCEPT / subject=ident_niv3;
```

L'estimation des modèles est cependant très consommatrice en temps de calcul, *a fortiori* en présence de plusieurs niveaux de regroupement ou lorsque le nombre d'observations de niveau 2 est élevé. `SOLUTION` permet d'afficher les estimations des coefficients du modèle, mais on peut choisir de l'omettre ce qui peut permettre de gagner en temps de calcul. D'autres "astuces" sont décrites dans Kiernan et al. (2012) pour améliorer les temps de calcul parmi lesquelles on peut retenir :

- L'option `DDFM=` permet de spécifier le mode de calcul du nombre de degrés de liberté qui entre en compte dans le calcul des statistiques de test. La valeur par défaut `DDFM=CONTAINMENT` demande des temps de calcul importants en particulier lorsque le nombre d'observations de niveau 2 est élevé. Choisir `DDFM=BW` améliore très nettement les temps de calcul.

- Spécifier l'effet aléatoire par l'instruction `RANDOM _residual_ / ident_niv2 TYPE=CS` améliore très nettement les temps de calcul. Elle permet de spécifier que la matrice de variance-covariance est diagonale par blocs, chaque bloc ayant pour termes diagonaux $\sigma^2_\alpha + \sigma^2_\varepsilon$ et σ^2_ε en dehors de la diagonale. Attention, cette option n'est pas valable lorsque la variable est binaire.

Par défaut, l'estimation repose sur une méthode de pseudo-maximum de vraisemblance (`METHOD=RESL`). On peut choisir une approximation numérique avec `METHOD=QUAD (Qpoints=)` où `Qpoints` correspond au nombre de points utilisés pour approximer l'intégrale.

L'instruction `OUTPUT OUT=` permet de créer en sortie une table SAS, qui en plus des données initiales contient des variables utiles à l'analyse :

- Si le modèle est linéaire, on peut vouloir obtenir pour chaque individu l'estimation de la variable d'intérêt : $\hat{y}_{ij} = \hat{\beta}_0 + x_{ij}\hat{\beta} + x_j\hat{\gamma} + \hat{\alpha}_j$. On l'appelle par l'instruction `pred(blup)=`. `BLUP` indique que c'est la meilleure prédiction linéaire sans biais pour les effets aléatoires $\hat{\alpha}_j$ qui est utilisée (voir section 1.2.2).
  Par exemple la commande `OUTPUT OUT=table_ychap PRED(BLUP)= ychap` crée la table SAS `table_ychap` avec la variable `ychap` (en plus des variables initiales) qui correspond à la valeur prédite de y.

- Pour un modèle binaire, les probabilités prédites sont obtenues par l'instruction `pred(BLUP ILINK)=`. La fonction `ILINK` correspond à l'inverse de la fonction de lien, qui appliquée à $\hat{\beta}_0 + x_{ij}\hat{\beta} + x_j\hat{\gamma} + \hat{\alpha}_j$ fournit bien la probabilité prédite.
  Exemple de commande : `OUTPUT OUT=table_ychap PRED(BLUP ILINK)= pchap`

L'instruction `ODS OUTPUT` permet de créer des tables SAS utiles à l'analyse :

- `SOLUTIONR=` crée la table contenant les estimations des effets aléatoires (les $\hat{\alpha}_j$). L'option `SOLUTION` doit pour cela être spécifiée dans l'instruction `RANDOM`.

- `PARAMETERESTIMATES=` crée la table contenant les estimations des coefficients des covariables (les $\hat{\beta}$).

- `COVPARMS=` crée la table contenant l'estimation de la variance des effets aléatoires ($\hat{\sigma}^2_\alpha$) et individuels ($\hat{\sigma}^2_\varepsilon$, sauf dans le cas binaire où celle-ci est fixée).

Exemple de syntaxe :

```
ODS OUTPUT SOLUTIONR=alphachap PARAMETERESTIMATES=betachap
                                                COVPARMS=sigmachap
```

Enfin, le test de nullité de la variance inter-classes est obtenu par la commande `covtest 0;`.

3.2.2 Avec R

Effets fixes L'estimation d'un modèle linéaire à effets fixes se fait par exemple avec la procédure `plm` disponible dans le package du même nom. Elle estime le modèle *within*.

```
modelefe<-plm(formula=y~ var_num+factor(var_cat),index=c("ident_niv2"),
                                                              data=table)
```

Il est aussi possible d'estimer un simple modèle linéaire (procédure `lm`) en ajoutant autant d'indicatrices que de groupes, mais on peut être rapidement confronté à un problème d'allocation mémoire compte tenu du grand nombre de paramètres à estimer.

Lorsque la variable d'intérêt est binaire, on peut utiliser la procédure `clogit` du package `survival`. La syntaxe sera ici :

```
modelebinaire2<-clogit(ybinaire~var_num+factor(var_cat)+strata(ident_niv2),
                                                                  data=table)
```

Effets aléatoires Il existe plusieurs packages R permettant de mettre en œuvre des méthodes à effets aléatoires : les plus utilisés sont `lme4` et `multilevel` (Bliese, 2013). Dans le cas d'un modèle linéaire, on peut par exemple utiliser la fonction `lmer` du package `lme4` (voir Bates et al., 2014). Après avoir appelé cette librairie (`library(lme4)`), la syntaxe pour estimer un modèle avec un simple effet aléatoire sur la constante s'écrit :

```
modelecontinu<-lmer(y~var_num+factor(var_cat)+(1|ident_niv2),data=table)
```

Dans ce modèle très simple qui utilise la table `table` la variable dépendante `y` est régressée sur plusieurs variables séparées par des `+`. L'effet aléatoire correspondant au groupe `ident_niv2` s'écrit `(1|ident_niv2)`. Les variables explicatives peuvent être continues `var_num` ou catégorielles `var_cat`, ces dernières étant transformées en indicatrices par la fonction `factor`. Si on souhaite ajouter un terme aléatoire sur la pente d'une ou plusieurs variables `var_p` (incluses dans `var_num` et `var_cat`), on écrira simplement `(1+var_p | ident_niv2)` (avec les précautions nécessaires indiquées dans la section 2.1).

Dans le cas où la variable d'intérêt est binaire, on peut par exemple utiliser la fonction `gmlr` de ce même package `lme4`[16]. La syntaxe est alors :

```
modelebinaire<-glmer(ybinaire~var_num+var_cat+(1|ident_niv2),data=table,
               family=binomial(link=logit))
```

L'option `family=binomial` spécifie qu'on modélise une variable binaire, avec un modèle logit `(link=probit)` ou probit `(link=probit)`.

[16] On peut également utiliser la librairie nmle.

3.2.3 Avec STATA

Effets fixes On pourra spécifier un modèle linéaire à effets fixes par exemple par l'option `fe` dans la procédure `xtreg`. Il faut préciser que la variable de groupement est `ident_niv2`, par exemple en utilisant l'instruction `xtset`. Cette instruction ne fonctionne que pour une variable d'intérêt continue.

```
xtset ident_niv2
xtreg var_num i.var_cat, fe
```

Lorsque la variable d'intérêt est binaire, la procédure pour modéliser un logit conditionnel s'appelle comme sous R `clogit`. La syntaxe est alors :

```
clogit ybinaire var_num i.var_cat, group(ident_niv2)
```

Le suffixe `i.` indique qu'il est nécessaire de créer des indicatrices.

Effets aléatoires Pour un modèle à variable continue, on spécifiera un modèle à effets aléatoires par exemple par la procédure `xtmixed`.

```
xtmixed y var_num i.var_cat || ident_niv2:, reml
```

où ici aussi la variable modélisée est la variable `y`, avec des variables explicatives continues `var_num` ou catégorielles `var_cat`.

Pour les variables binaires, on peut utiliser la fonction `xtlogit` :

```
xtset ident_niv2
xtlogit ybinaire var_num i.var_cat, re
```

l'instruction `xtset` permet de définir la variable identifiant le niveau sur lequel porte l'effet aléatoire. Cela peut aussi se faire en une étape par :

```
 xtlogit ybinaire var_num i.var_cat i(ident_niv2), re
```

On peut ajouter comme option la méthode d'intégration par l'option `intmethod` et le nombre de points utilisés pour l'intégration numérique `intpoints(#)` (la valeur par défaut est 12, le maximum 196).

Deux autres procédures permettent de faire cette estimation : `xtmelogit` et `gllamm`. `gllamm` est la plus lente, mais peut donner des résultats moins sensibles aux choix de spécification. La syntaxe de cette dernière est :

```
gllamm ybinaire var_num i.var_cat, i(ident_niv2) link(logit) family(binom)
                                                          nip(#) adapt
```

outre les options essentielles, l'option `nip` demande de spécifier le nombre de points d'intégration (rappelons que l'estimation est approchée), `adapt` utilise la quadrature adaptative.

4 Exemples

4.1 Modélisation du salaire

Pour illustrer la mise en œuvre d'un modèle multiniveaux, on utilise les données issues d'un échantillon de 2 000 unités légales dont l'effectif est compris entre 20 et 100 salariés. Les données sont issues des déclarations annuelles de données sociales (DADS), on peut donc disposer, pour chaque salarié de la firme, du salaire horaire moyen sur l'année, de son âge, de son genre, de la qualification de son poste à un niveau fin.

On modélise alors le logarithme du salaire horaire (calculé à partir des périodes d'emplois obtenus sur l'année) sur un ensemble de caractéristiques. Les résultats, obtenus sur un échantillon réduit d'unités légales, ne sont évidemment présentés qu'à titre d'illustration des modèles multiniveaux, et ne constituent pas une étude en tant que telle.

Le tableau 1 présente les résultats obtenus en utilisant plusieurs spécifications. Par exemple, la première colonne présente les résultats obtenus sur un modèle à effets aléatoires (au niveau firme) sans covariable. La deuxième colonne correspond au même modèle mais des covariables ont été ajoutées. L'analyse des termes de covariance dans ce deuxième modèle montre que l'estimation de la variance des effets firme vaut : $\hat{\sigma}^2_\alpha = 0,02039$, tandis que celle des termes d'erreurs résiduels vaut $\hat{\sigma}^2_\varepsilon = 0,05078$. Ces termes permettent de mesurer l'ampleur des effets entreprise dans la variance des salaires, ce qui est synthétisé par l'ICC. Rappelons que celui-ci est par définition égal à $\frac{\sigma^2_\alpha}{\sigma^2_\alpha+\sigma^2_\varepsilon}$ soit ici 0,29. Cela signifie que, même une fois pris en compte les effets de composition en termes d'âge, de sexe et de qualification du poste occupé, environ 29% de la variance du salaire horaire s'explique par des effets entreprise[17]. On peut vérifier que cet effet aléatoire entreprise est bien significatif par un test du maximum de vraisemblance (attention, il est différent d'un test de Student qui repose sur le ratio de l'estimateur et de son écart-type). Dans notre exemple, la valeur de la statistique de test du χ^2 vaut 22 844, et la probabilité d'observer cette valeur sous l'hypothèse nulle que les effets firmes sont nuls est infinitésimale.

On peut aussi s'intéresser à mesurer l'ampleur des différences de salaire constatées selon les caractéristiques individuelles des salariés[18]. Selon ces estimations, dans cet échantillon de firmes les salariées perçoivent un salaire inférieur à celui des salariés de 7,1%, une fois tenu compte des effets de l'âge et de la qualification du poste occupé, ainsi que d'un éventuel effet spécifique à l'entreprise (voir colonne 2 du tableau). L'âge, qui est fortement corrélé avec l'expérience professionnelle (non disponible dans les fichiers), a également un effet positif qui s'atténue avec le temps.

Cependant, ces derniers coefficients peuvent être biaisés s'il existe, par exemple, une corrélation entre les caractéristiques individuelles des salariés et les effets entreprises : par exemple, si les entreprises dont les politiques salariales sont les plus généreuses ont un biais de recrutement en faveur des hommes expérimentés. Si on s'intéresse spécifiquement à l'effet de ces variables, on pourra donc préférer utiliser un modèle à effets fixes. Sur cet exemple, les estimations obtenues par le modèle à effets fixes apparaissent légèrement différentes de celles obtenues par le modèle précédent, même si l'ampleur des différences est faible : on obtient cette fois un écart

[17] À titre de comparaison, lorsque l'on se limite au modèle "vide" sans aucune variable individuelle, 33% de la variabilité des salaires correspond à de la variabilité entre les entreprises. L'écart substantiel entre ces deux grandeurs indique, sans surprise, des différences dans la composition de la main d'œuvre des différentes entreprises.

[18] On a centré ici l'ensemble des variables, ce qui permet d'avoir une interprétation directe de la constante comme valeur au point moyen de l'échantillon.

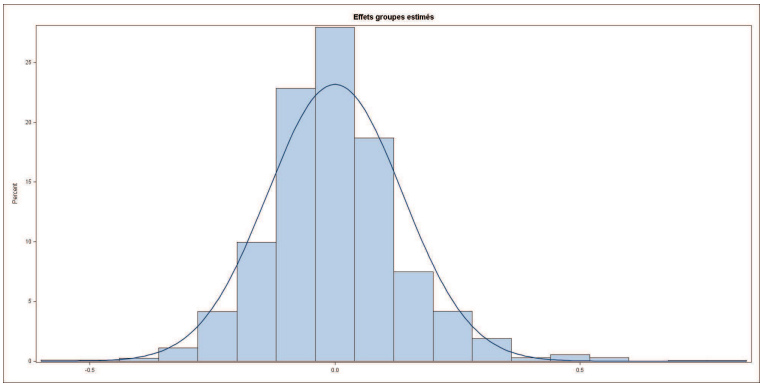
Tableau 1 – Modélisation du salaire horaire (en log), prise en compte d'effets firmes

	Effets aléatoires			Effets fixes
	(1)	(2)	(3)	(4)
Constante	2,353	1,809	1,425	
	(0,005)	(0,027)	(0,206)	
Femme		réf.	réf.	réf.
Homme		0,071	0,068	0,068
		(0,002)	(0,002)	(0,002)
Age		0,021	0,021	0,021
		$(4{,}37.10^{-4})$	$(4{,}39.10^{-4})$	$(4{,}39.10^{-4})$
Age^2		$-1{,}80.10^{-4}$	$-1{,}80.10^{-4}$	$-1{,}80.10^{-4}$
		$(5{,}43.10^{-6})$	$(5{,}450.10^{-4})$	$(5{,}45.10^{-6})$
Chefs d'entreprises		0,929	0,925	0,924
		(0,028)	(0,028)	(0,028)
Cadres		0,528	0,522	0,521
		(0,026)	(0,026)	(0,026)
Professions intermédiaires		0,128	0,123	0,123
		(0,026)	(0,026)	(0,026)
Employés		−0,094	−0,095	−0,095
		(0,026)	(0,026)	(0,026)
Ouvriers		−0,142	−0,146	−0,146
		(0,026)	(0,026)	(0,026)
Agriculteur		réf.	réf.	réf.
Moyennes firmes				
proportions d'hommes			0,134	
			(0,015)	
Age moyen			0,01	
			(0,005)	
Age^2 moyen			0,163	
			(0,271)	
Proportion de chefs d'entreprises			0,194	
			(0,194)	
Proportion de cadres			0,156	
			(0,193)	
Proportion de professions intermédiaires			0,088	
			(0,193)	
Proportion d'employés			0,058	
			(0,193)	
Proportion d'ouvriers			0,058	
			(0,193)	
Variance inter firmes σ^2_α	0,046	0,02	0,018	
	(0,002)	(0,001)	(0,018)	
Variance individuelle σ^2_ε	0,091	0,051	0,051	
	(0,000455)	(0,000253)	(0,051)	

Estimations sur un échantillon d'unités légales du logarithme du salaire horaire. Les modèles (1), (2), (3) correspondent à une spécification avec effets aléatoires, respectivement sans variables explicatives (modèle vide), avec variables individuelles seulement puis avec variables individuelles et leurs moyennes sur l'unité légale dans laquelle le salarié travaille (modèle de Mundlak).

de 7,1% [19]. Une manière classique et simple de tester l'existence d'une corrélation entre les variables observables individuelles et les effets groupes est, comme détaillé dans la section 1.2.3, d'ajouter les moyennes par établissement de chacune de ces variables (voir équation 9, colonne 3 du Tableau) : on parle souvent de spécification de Mundlak. La significativité statistique des estimateurs obtenus pour ces variables constitue alors un test de l'hypothèse d'indépendance. Dans notre exemple, on voit que certaines de ces variables sont significativement corrélées avec le salaire individuel. Leur prise en compte dans la modélisation modifie légèrement les résultats des estimations des variables à effets aléatoires. Les résultats sont alors identiques à ceux obtenus par effets fixes.

Il est possible d'estimer les effets entreprises, qui sont représentés dans le graphique 6. La distribution est normale, ce qui est tautologique avec la modélisation retenue.

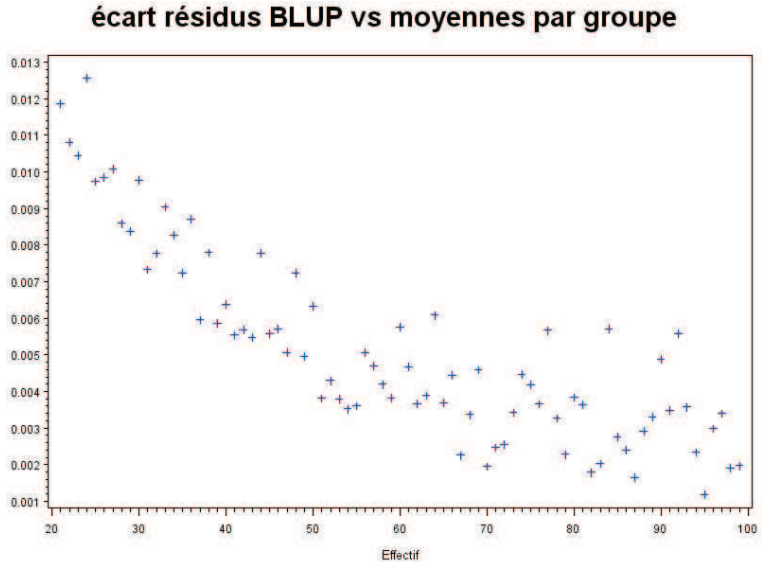


Graphique 6 – Distribution des effets entreprises estimés

Rappelons que l'estimation des effets groupes est la contrepartie empirique de l'estimateur dit "BLUP" (pour *best linear unbiased predictor*). Comme décrit dans la section 1.2.2, cette prédiction est fondée sur les résidus observés $u_{ij} = y_{ij} - X_{ij}\hat{\beta}$ (l'écart entre la valeur observée de la variable d'intérêt et la prédiction obtenue à partir des seules variables observables), mais ne correspond pas simplement à la moyenne de ceux-ci. L'estimation tient compte en effet du fait que lorsque les groupes comportent peu d'observations, cette moyenne offre une estimation peu précise et peut donc être éloignée de la vraie valeur. Un facteur de contraction est appliqué. Pour s'en rendre compte, on a calculé, à simple titre d'illustration (ce calcul n'ayant *a priori* pas d'intérêt en soi), l'écart (en valeur absolue) entre les effets entreprises fournis par l'estimation, et les moyennes par entreprise des résidus observés. Pour que le graphique soit plus lisible, on estime la moyenne de ces écarts par taille d'entreprise. Conformément à ce qui est attendu, on observe dans le graphique 7 que cet écart est décroissant (en fait, inversement proportionnel) avec le nombre d'observations par groupe - l'effectif de l'entreprise dans notre exemple.

Enfin, il est important de souligner que, dans toutes les modélisations étudiées ici, on suppose que les termes résiduels individuels, relatifs aux salariés, sont indépendants des effets entreprises, ainsi que des variables explicatives introduites dans le modèle. Ces termes résiduels correspondent à l'ensemble des déterminants du salaire qui ne sont pas observés dans les données (par exemple, les compétences professionnelles du salarié, qui ne sont qu'imparfaitement mesurées par la qualification du poste qu'il occupe). Ces déterminants peuvent être corrélés

[19] Notons que dans cet exemple, on dispose de groupes de taille non négligeable et homogène - par construction, l'échantillon porte sur des entreprises comprenant entre vingt et cent salariés.



On calcule la moyenne des écarts absolus entre les effets groupes estimés ("BLUP") et la moyenne par entreprise des résidus estimés. On utilise une échelle logarithmique dans un souci de présentation.

Graphique 7 – Ecart entre les effets entreprises empiriques et les moyennes des résidus en fonction du nombre d'observations par entreprise

avec les variables explicatives ainsi qu'avec les effets entreprise (si les salariés les plus qualifiés se regroupent dans les entreprises avec les politiques salariales les plus généreuses), ce qui peut donc conduire à biaiser les résultats.

4.2 Cas binaire : modélisation du retard scolaire

On propose ici d'illustrer les modèles multiniveaux binaires en s'intéressant aux effets de certaines caractéristiques du quartier des élèves sur le fait qu'ils soient ou non en retard scolaire à l'entrée en 6e. Le quartier d'habitation de l'élève est défini à partir des données carroyées : il correspond à un carreau de 200 mètres, à partir duquel sont définies un certain nombre de variables décrivant son voisinage (sur la couronne d'un kilomètre autour de ce carreau : part de locataires HLM, part de ménages de cinq personnes ou plus, etc.). La structure emboîtée des données (plusieurs élèves partagent le même voisinage) requiert une modélisation multiniveaux. La présentation met ici l'accent sur l'interprétation des estimations et ses limites. Elle insiste plus particulièrement sur ce qui est spécifique à la modélisation d'une variable binaire.

L'usage des modèles multiniveaux est courant en éducation. Au-delà des caractéristiques propres de l'élève, son contexte familial ou scolaire peut en effet avoir un effet sur ses performances scolaires. Son quartier d'habitation est également l'objet d'analyses. Les études montrent en général que les élèves vivant dans un environnement plutôt favorisé ont de meilleurs résultats scolaires. Ceci pourrait s'expliquer par des effets de pairs ou par l'influence des adultes du voisinage sur les aspirations scolaires des élèves (voir par exemple Owens, 2010 ou Vallet, 2005).

Les résultats des différentes estimations du modèle sont présentés dans le tableau 2 : modèle logit simple, modèle à effets aléatoires sur le quartier d'habitation de l'élève, modèle à effets fixes et modèle de Mundlak. Le retard scolaire d'élèves primo entrants en 6e en Île-de-France

en 2011 est expliqué par des caractéristiques propres de l'élève (sexe, milieu social...), des caractéristiques disponibles de l'école primaire[20] fréquentée l'année précédente (secteur, composition sociale des élèves...), et des caractéristiques de son quartier d'habitation. Les sorties de ces estimations avec les logiciels SAS, R et STATA sont présentées en annexe (paragraphe A.2).

Les variances des résidus Dans le modèle binaire, la variance des résidus individuels est fixée pour rendre le modèle identifiable (classiquement à $\pi^2/3$ qui correspond à la variance d'une loi logit standard). C'est une spécificité du modèle binaire par rapport au modèle linéaire. Comme dans le modèle linéaire, on peut calculer l'ICC qui mesure la part de la variance résiduelle due au niveau carreau d'habitation. Dans le modèle avec effets aléatoires (tableau 2), il vaut donc $\frac{0,1066}{0,1066+\pi^2/3} = 3,14\%$.

Les coefficients estimés du modèle logit simple et ceux du modèle à effets aléatoires sont proches. Ceci est lié au fait que la variance variance inter-classe apparaît faible. Même lorsque les variables vérifient les hypothèses du modèle et ne sont pas corrélées avec les effets quartier, on s'attend à ce que le coefficient du modèle à effets aléatoires corresponde à celui du modèle logit simple multiplié par $\sqrt{1+\frac{0,1066}{\pi^2/3}} \approx 1,016$ (cf encadré page 42) . On remarque également que les écarts-types des estimateurs sont également plus grands que ceux du modèle logit simple. Cela signifie que, à partir d'un modèle logit simple, on peut parfois être amené à conclure à tort à la significativité d'une variable.

L'interprétation des coefficients des variables change peu par rapport à un modèle logit simple. Le coefficient correspond strictement à l'effet sur le logit de la probabilité d'être en retard scolaire de l'augmentation d'une unité de la variable explicative. Un coefficient positif indique que la probabilité de retard scolaire croît avec la variable. Ainsi par exemple le fait d'être un garçon augmente les risques de retard scolaire à l'entrée en 6e, à autres caractéristiques observables fixées. Pour quantifier cet effet, on calcule les odd-ratios ou les effets marginaux qui ont une interprétation plus parlante que la valeur du coefficient elle-même.

[20] On pourrait ici envisager d'ajouter un effet spécifique à l'école en plus de celui de l'école primaire. Ce modèle à deux niveaux est plus complexe à estimer et nous ne le détaillons pas ici.

Tableau 2 – Modélisation du retard scolaire

Variable Modalités	Modèle logit simple	Modèle à effets aléatoires Coef.	 Effets marginaux (1)	 Effets marginaux (2)	 Odds ratio	Modèle à effets fixes	Modèle de Mundlak
Constante	$-2,16^{***}$ (0,1337)	$-2,18^{***}$ (0,1411)					$-2,13^{***}$ (0,1431)
Élève							
Sexe							
garçon	$0,27^{***}$ (0,0175)	$0,27^{***}$ (0,0176)	2,66	2,58	1,30	$0,29^{***}$ (0,020)	$0,28^{***}$ (0,0198)
fille	réf.	réf.				réf.	réf.
Nationalité							
étranger	$0,79^{***}$ (0,0313)	$0,80^{***}$ (0,0317)	9,70	9,70	2,22	$0,77^{***}$ (0,0368)	$0,76^{***}$ (0,0359)
français	réf.	réf.				réf.	réf.
Boursier							
Boursier	$0,29^{***}$ (0,0226)	$0,29^{***}$ (0,0228)	2,90	2,97	1,33	$0,20^{***}$ (0,0257)	$0,20^{***}$ (0,0255)
Non boursier	réf.	réf.				réf.	réf.
Milieu social							
Très favorisé (A)	$-1,20^{***}$ (0,0334)	$-1,20^{***}$ (0,0336)	-6,81	-9,13	0,30	$-1,11^{***}$ (0,0386)	$-1,12^{***}$ (0,0375)
favorisé (B)	$-0,35^{***}$ (0,0325)	$-0,34^{***}$ (0,0327)	-2,71	-3,07	0,71	$-0,33^{***}$ (0,0374)	$-0,32^{***}$ (0,0371)
Moyen (C)	réf.	réf.				réf.	réf.
Défavorisé (D)	$0,48^{***}$ (0,0211)	$0,48^{***}$ (0,0212)	5,25	4,92	1,62	$0,48^{***}$ (0,0245)	$0,47^{***}$ (0,0242)
voisinage (couronne de 1km)							
Part de ménages à bas revenus	$2,8.10^{-3}$ (0,0033)	$2,6.10^{-3}$ (0,0035)	0,024	0,025	1,0026		$4,4.10^{-3*}$ (0,0035)
Part d'étrangers	$-2,2.10^{-3}$ (0,0036)	$-3,4.10^{-3}$ (0,0038)	-0,031	-0,033	0,997		$3,8.10^{-3}$ (0,0041)
Part de locataires HLM	$8,9.10^{-5}$ (0,0006)	$1,0.10^{-4}$ (0,0007)	0,0009	0,0010	1,0001		0,0002 (0,0007)
Part de ménages de 5 pers ou +	$-8,0.10^{-3***}$ (0,00247)	$-8,5.10^{-3***}$ (0,0026)	-0,076	-0,082	0,99		$-8,0.10^{-3***}$ (0,0026)
Part de familles monoparentales	$0,030^{***}$ (0,0088)	$0,029^{***}$ (0,0093)	0,26	0,28	1,03		$0,028^{***}$ (0,0093)
% milieu A	$-8,5.10^{-3***}$ (0,0017)	$-8,4.10^{-3***}$ (0,0018)	-0,076	-0,082	0,99		$-5,6.10^{-3***}$ (0,0020)
% milieu B	$1,9.10^{-3}$ (0,0034)	$2,2.10^{-3}$ (0,0035)	0,020	0,021	1,0022		$3,1.10^{-3}$ (0,0036)
% milieu D	$-1,1.10^{-3}$ (0,0022)	$-6,8.10^{-4}$ (0,0023)	$-6,2.10^{-3}$	$-6,6.10^{-3}$	0,999		$6,9.10^{-4}$ (0,0025)
% de boursiers	$3,8.10^{-3*}$ (0,00376)	$4,0.10^{-3*}$ (0,0023)	0,036	0,038	1,004		$2,1.10^{-3}$ (0,0027)
École primaire							
% d'étrangers	$7,2.10^{-3***}$ (0,0014)	$7,7.10^{-3***}$ (0,0015)	0,070	0,075	1,008	$0,013^{***}$ (0,0020)	$0,013^{***}$ (0,0019)
% de boursiers	$-1,9.10^{-3**}$ (0,0009)	$-1,8.10^{-3*}$ (0,0009)	-0,016	-0,017	0,998	$-1,6.10^{-3}$ (0,0013)	$-1,5.10^{-3}$ (0,0013)
% d'élèves de milieu A	$-3,4.10^{-4}$ (0,0010)	$-2,1.10^{-4}$ (0,0011)	$-1,9.10^{-3}$	$-2,0.10^{-3}$	0,99979	$8,3.10^{-4}$ (0,0014)	$1,2.10^{-3}$ (0,0014)
% d'élèves de milieu B	$3,8.10^{-3**}$ (0,0015)	$3,8.10^{-3**}$ (0,0015)	0,034	0,037	1,0038	$0,0052^{**}$ (0,0022)	$5,6.10^{-3**}$ (0,0022)

.../...

Variable Modalités	Modèle logit simple	Modèle à effets aléatoires Coef.	 Effets marginaux (1)	 Effets marginaux (2)	 Odds ratio	Modèle à effets fixes	Modèle de Mundlak
% d'élèves de milieu D	$2,9.10^{-3***}$ (0,0009)	$3,1.10^{-3***}$ (0,0010)	0,028	0,030	1,003	$0,0042^{***}$ (0,0014)	$4,6.10^{-3***}$ (0,0013)
Public / privé							
Privé	$-0,060$ (0,0417)	$-0,051$ (0,042)	-0,45	-0,49	0,95	0,0041 (0,051)	$-0,026$ (0,050)
Public	réf.	réf.				réf.	réf.
Moyenne... (par voisinage) (Modèle de Mundlak)							
... de garçons							$-0,077^{*}$ (0,0430)
... d'élèves scolarisés dans le privé							$-0,058$ (0,0853)
... de boursiers							$0,46^{***}$ (0,0621)
... d'élèves de nationalité étrangère							$0,25^{***}$ (0,0951)
... d'élèves de milieu A							$-0,27^{***}$ (0,0648)
... d'élèves de milieu B							$-0,072$ (0,0748)
... d'élèves de milieu D							0,027 (0,0553)
... du pourcentage d'élèves étrangers dans l'école primaire							$-0,015^{***}$ (0,0032)
... du pourcentage d'élèves de milieu A dans l'école primaire							$-2,4.10^{-3}$ (0,0021)
... du pourcentage d'élèves de milieu B dans l'école primaire							$-3,2.10^{-3}$ (0,0031)
... du pourcentage d'élèves de milieu D dans l'école primaire							$-3,6.10^{-3*}$ (0,0020)
... du pourcentage d'élèves boursiers dans l'école primaire							$-3,4.10^{-3*}$ (0,0019)
Variance inter-classe		0,1066 (0,0128)					0,1026 (0,01276)

Champ : Primo-entrants en 6e en Île-de-France à la rentrée 2011.
Source : MENESR-DEPP, Système d'information Scolarité.
Note : *** :p<1%, ** :p<5%, * :p<10%. (1) calcul des effets marginaux pour un élève dans une situation de référence (fille non boursière scolarisée dans le secteur public, de milieu social moyen (C), et dont les autres caractéristiques (variables continues) correspondent à la moyenne sur l'ensemble des observations) ; (2) calcul des effets marginaux en moyenne sur l'échantillon.

Calcul des effets marginaux La valeur des coefficients estimés d'un modèle logit a une interprétation presque directe en termes de rapport des chances. Par exemple, pour la variable "sexe", l'odds-ratio correspond à $\exp(\hat{\beta}_{\text{garçon}} - \hat{\beta}_{\text{fille}}) = 1,30$ (cf tableau 2) ce qui signifie que la probabilité d'être en retard sur la probabilité de ne pas être en retard pour deux élèves ayant les mêmes caractéristiques observables (en particulier, résidant dans le *même voisinage*) est 1,30

fois plus élevée pour un garçon que pour une fille. Cependant, l'interprétation des rapports de chance n'est pas totalement intuitive et on préfère en général calculer les effets marginaux, qui correspondent à l'effet d'une variable sur la probabilité d'être en retard scolaire. Comme détaillé dans la partie 1.3.4, il existe deux façons de les calculer. La première demande de fixer une situation de référence. On la fixe ici au fait d'être une fille non boursière scolarisée dans le secteur public et de milieu social moyen. Les autres caractéristiques de son école primaire et celles de son voisinage d'habitation correspondent à la moyenne observée sur l'ensemble de l'échantillon. Les effets marginaux peuvent également être calculés sur l'ensemble des individus de l'échantillon : ils correspondent à l'évolution moyenne de la probabilité d'être en retard scolaire lorsqu'une variable augmente d'une unité (variable continue) ou que l'individu a une modalité plutôt qu'une autre. Le tableau 2 présente les valeurs estimées de ces effets marginaux selon ces deux méthodes (qui ici donnent des résultats proches). On observe qu'être un garçon augmente en moyenne de 2,6 points la probabilité d'être en retard pour un élève dont les caractéristiques correspondent par ailleurs à la situation de référence. Au-delà des variables individuelles, les caractéristiques de l'école élémentaire fréquentée auparavant ont une incidence statistiquement significative, ainsi que certaines caractéristiques observables du quartier. Par exemple, la probabilité d'être en retard scolaire diminue quand la proportion de ménages de milieu très favorisé (A) dans le quartier augmente. Cependant, la magnitude de ces effets est faible. Quand le pourcentage de ménages de milieu très favorisé augmente d'un point, la probabilité d'être en retard scolaire pour un élève dans la situation de référence définie diminue de 0,08 point de pourcentage.

Interprétation Pour tirer des conclusions en termes d'effets causaux des variables, des hypothèses d'exogénéité sont nécessaires. Pour interpréter par exemple l'effet de la part d'élèves de milieu très favorisé dans le quartier sur le retard scolaire comme un effet causal, il faut pouvoir supposer que cette variable n'est pas liée avec les caractéristiques de l'élève ou de son voisinage non incluses dans le modèle. Cette hypothèse peut être remise en cause en particulier si les parents ne choisissent pas leur lieu d'habitation au hasard. Si l'implication des parents a un effet positif sur la scolarité de leur enfant et que les parents les plus impliqués (caractéristiques non observées) choisissent en moyenne un lieu d'habitation où la part d'élèves de milieu très favorisé est élevée, le coefficient associé à la part d'élèves de milieu A dans le voisinage capte en fait l'effet positif de l'implication des parents sur la scolarité de l'élève.

Le modèle de Mundlak Le modèle de Mundlak consiste à inclure dans le modèle les moyennes de toutes les variables de niveau individuel et à tester la nullité des coefficients associés à ces variables (cf tableau 2). Un test de nullité globale de ces coefficients permet de tester indirectement l'hypothèse d'exogénéité du modèle à effets aléatoires. Ici, l'hypothèse de nullité est rejetée, ce qui suggère qu'il existe bien une corrélation entre l'effet quartier et les variables incluses dans le modèle (voir sortie SAS page 57). On observera par ailleurs que, contrairement au cas linéaire, le modèle à effets fixes et le modèle de Mundlak ne coïncident pas.

Effet de l'ajout de variables sur la valeur des coefficients et les termes de variances

Ajouter de nouvelles variables à un modèle peut avoir un impact sur les coefficients estimés pour d'autres variables explicatives, si ces dernières sont corrélées avec ces variables supplémentaires. Intuitivement, dans le "petit" modèle une partie de l'effet de la variable supplémentaire est capté par les variables avec lesquelles elle est corrélée. Dans un modèle linéaire, on pourra directement interpréter par ces corrélations entre variables les variations des coefficients estimés pour des variables explicatives lorsqu'on retire ou qu'on ajoute de nouvelles variables : par exemple, si on modélise le salaire, l'impact estimé du diplôme sera plus élevé si le modèle estimé ne comprend pas de variables correspondant à la qualification. Le fait d'occuper telle ou telle position sociale est en effet fortement corrélé au diplôme. En termes de variance, ajouter des variables individuelles non corrélées à l'effet groupe diminue la variance des résidus individuels et n'a pas d'effet sur la variance des effets groupes. La variance totale du modèle diminue donc. La part de la variance qui s'explique par le niveau groupe augmente. Il s'agit d'un cas d'école. En pratique, ces variations peuvent être moins mécaniques, notamment si les variables ajoutées sont corrélées avec les effets groupes.

Les choses sont différentes dans une modélisation binaire de type probit ou logit. La variance individuelle est fixée par la contrainte d'identification (donc à 1 dans le cas d'un probit, et $\frac{\pi^2}{3}$ pour un logit)et ne varie donc pas lorsqu'on introduit de nouvelles variables. Les coefficients estimés correspondant peuvent être modifiés *même lorsque les variables explicatives ne sont pas corrélées*. En effet, du fait de la contrainte d'identification nécessaire sur la variance du modèle, les coefficients seront tous multipliés par une constante différente pour chaque spécification. Ainsi, ajouter des effets aléatoires à un modèle logistique "simple" va multiplier les valeurs des coefficients estimés pour une variable de niveau individuel par $\sqrt{1+\frac{\sigma_\alpha^2}{\pi^2/3}}$, où σ_α^2 correspond à la variance des effets aléatoires groupe. Ajouter une variable explicative supplémentaire à un modèle à effets aléatoires va également avoir un impact mécanique sur les coefficients et la variance résiduelle. La variance individuelle étant fixée, elle ne peut pas diminuer suite à l'ajout d'une nouvelle variable individuelle. On va logiquement observer une diminution de sa part dans la variance résiduelle totale, qui va se matérialiser par une augmentation de la variance de niveau groupe, même lorsque cette variable n'est pas corrélée aux effets groupes. Pour toutes les autres variables, même lorsqu'elles ne sont pas corrélées avec cette nouvelle variable, le coefficient doit augmenter. Lorsque les hypothèses du modèle sont vérifiées (en particulier si elle n'est pas corrélée aux effets groupes), et que cette nouvelle variable est indépendante des autres variables explicatives, les coefficients correspondant à ces dernières sont multipliés par un facteur $\sqrt{\tilde{\sigma}_\alpha^2/\sigma_\alpha^2}$ avec $\tilde{\sigma}_\alpha^2$ la variance du modèle augmenté.

Cette propriété des modèles logistiques est illustrée à partir d'un modèle particulièrement simple, dans lequel on n'introduit que deux variables individuelles (sexe et boursier) pour expliquer le retard scolaire, en plus d'un éventuel effet du quartier de résidence. Les résultats sont présentés dans le tableau ci-dessous : le modèle 1 correspond au logit simple (sans effets aléatoires), les modèles 2 et 3 ajoutent un effet aléatoire voisinage sur la constante. Alors qu'on peut supposer que la variable sexe est non corrélée à un

quelconque effet du quartier, on remarque que le coefficient estimé dans le modèle 2 pour la variable sexe est pourtant supérieur à celui estimé dans le logit simple. De même, alors qu'il est peu probable que le sexe des élèves soit corrélé avec leur statut de boursier, on observe que la variance inter-classe augmente lorsqu'on ajoute la variable sexe (modèle 2, par comparaison au modèle 3) ainsi que le coefficient de la variable "boursier".

Impact de l'ajout de variables supplémentaires

Variable	Modalité	Modèle 1 logit simple	Modèle 2	Modèle 3
Constante		$-2,3038^{***}$ (0,0131)	$-2,3264^{***}$ (0,01377)	$-2,1931^{***}$ (0,01014)
Sexe	Garçon	$0,2476^{***}$ (0,0164)	$0,2493^{***}$ (0,01669)	
	Fille	réf.	réf.	
Boursier	Boursier	$0,8782^{***}$ (0,0192)	$0,8108^{***}$ (0,02005)	$0,8029^{***}$ (0,02001)
	Non boursier	réf.	réf.	réf.
σ^2_α			$0,2545^{***}$ (0,01389)	$0,2533^{***}$ (0,01385)

Nombre et tailles des groupes

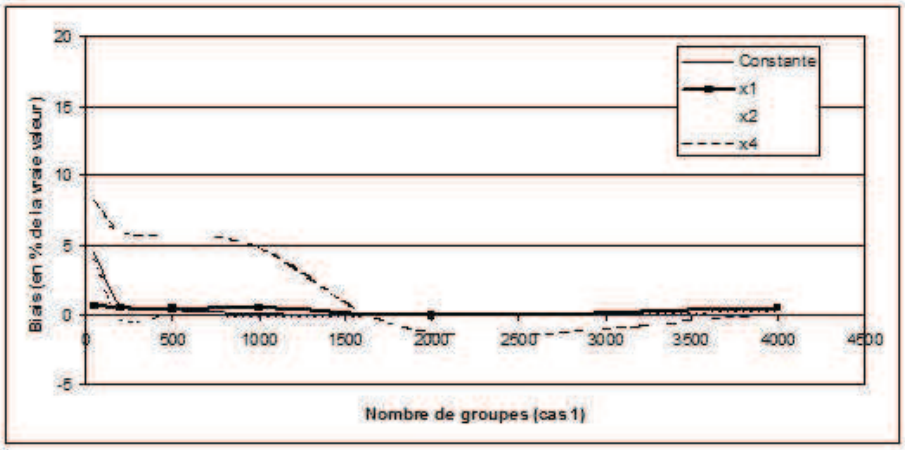
La modélisation des effets quartiers repose sur des voisinages pour lesquels on observe au moins un élève scolarisé en sixième. L'identification séparée entre effets quartiers et variables individuelles n'a de sens cependant que lorsqu'on observe plusieurs élèves de sixième dans le même voisinage. En général, le nombre d'élèves scolarisés est faible, et la proportion de "singletons" (un seul élève dans un voisinage) est élevée. Elle est ainsi d'un tiers environ sur l'Ile-de-France. Pour évaluer l'incidence possible sur les estimations, on a procédé à des simulations en faisant varier en particulier la taille de l'échantillon et l'ampleur des effets voisinages. Plus précisément, on tire dans un premier temps un échantillon d'un certain nombre de groupes (les voisinages), auxquels on affecte des observations individuelles (les élèves). On fait varier le nombre de groupes. Pour la distribution du nombre d'élèves, on distingue deux cas : le premier correspond à ce qui est observé en définissant comme voisinage en zones rurales l'ensemble de la commune, soit une moyenne de 7,3 élèves par groupes, avec 21% de singletons ; la seconde correspond à ce qui est observé en utilisant une définition beaucoup plus stricte du voisinage, soit un carreau de 200 m ce qui est possible en zone urbaine, soit une moyenne 2 élèves par groupe et 46% de singletons.

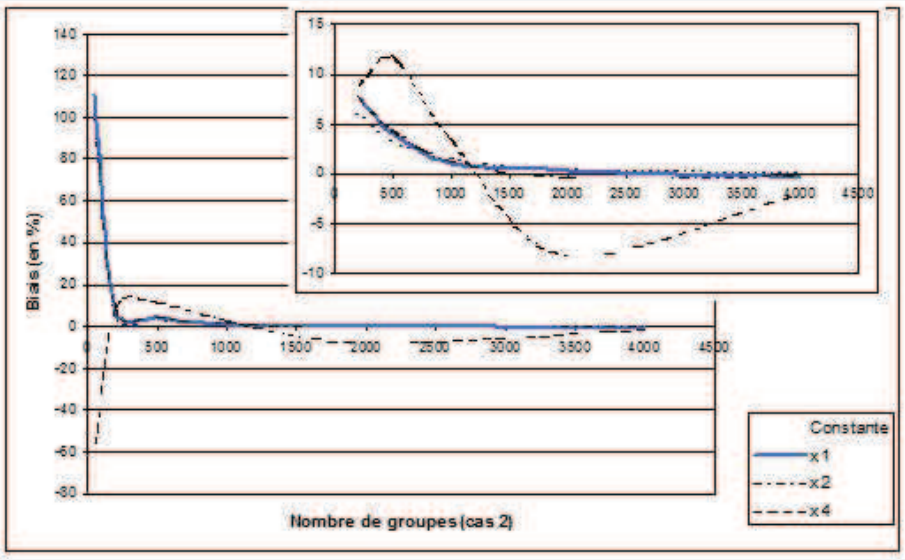
On a donc simulé des données à partir du modèle :

$$y_{ij} = -2 + x^1_{ij} + x^2_{ij} + x^4_j + \alpha_j + \varepsilon_{ij} \quad \text{et} \quad y_{ij} = \mathbb{1}(y^\star_{ij} \geq 0)$$

où α_j suit une normale centrée et de variance σ^2_α, ε_{ij} une loi logistique standard. L'ensemble des variables explicatives sont tirées d'une loi normale centrée réduite. Les variables x^1 et x^2 sont des variables de niveau individuel. La variable x^4 est une variable de niveau voisinage.

On teste plusieurs scénarios en faisant varier d'une part la variance inter-classes (de 0,05 à 0,3) et d'autre part le nombre de groupes (N=50, 100, 500, 1000, 2000 et 4000). Pour chaque scénario, 50 échantillons sont simulés, et on compare les estimations obtenues avec les vrais paramètres. Les résultats sont résumés dans les graphiques suivants :

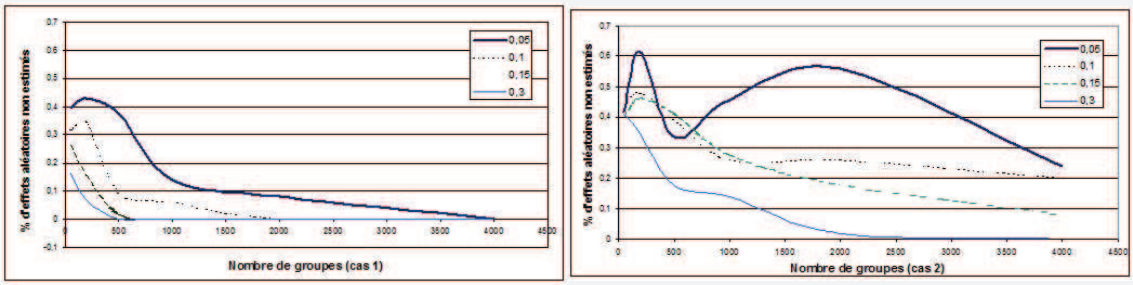




Conformément à l'intuition, les biais se réduisent avec le nombre de groupes. Ils sont très nettement plus élevés lorsque la taille des groupes est la plus faible (cas 2). Dans le cas 1 les biais observés, dans les cas envisagés dans ces simulations, ne dépassent pas 10%. Ils sont nettement plus élevés dans le cas 2, et peuvent attendre 100% lorsque le nombre de groupes est réduit.

À noter, lorsque le nombre de groupes est faible, on peut être confronté à des problèmes d'estimation des effets groupes. Les graphiques suivants présentent la proportion de cas où les estimations n'ont pu aboutir. Lorsque la variance des effets groupes est faible, il

est très fréquent de ne pouvoir obtenir d'estimations convergentes même pour des tailles d'échantillons "respectables".



5 Conclusion

Les modèles multiniveaux sont une manière de prendre en compte les effets liés à une structure emboîtée des données qui apparaît dans de nombreux contextes. L'une des caractéristiques de cette famille de modèles est l'accent mis sur l'explicitation dans l'estimation des effets de contexte qui sont souvent considérés comme l'un des sujets d'intérêt, et non simplement comme un paramètre de nuisance (pour l'estimation de la précision par exemple). Ces modèles peuvent être mobilisés pour répondre à plusieurs types de questions, ce qui explique sans doute leur popularité dans la littérature. Des exemples d'utilisation de ces modèles peuvent être trouvés, en se limitant à des références francophones récentes, dans Duclos et Murat (2014) qui tentent de mettre en évidence la valeur ajoutée des lycées, c'est-à-dire ce qui ne peut être expliqué dans la réussite scolaire au baccalauréat par les simples effets de composition initiale. L'économie de la santé est également un domaine dans lequel ces modèles trouvent naturellement leur place. Par exemple, Or et Renaud (2009) et Milcent et Rochut (2009) s'intéressent à l'hétérogénéité des pratiques entre les hôpitaux, Pilorge et al. (2013) à celles des médecins, tandis que Chaix et Chauvin (2005) tentent de mesurer l'effet du contexte de résidence dans les comportements de recours aux soins. Golaz et Bringé (2009) s'intéressent aux apports de ces modèles pour les études démographiques. Rappelons aussi qu'une très vaste littérature en particulier économique, qu'il serait illusoire de résumer, s'appuie sur l'utilisation de données de panel.

Il est désormais classique d'utiliser pour modéliser les effets de contexte soit une forme paramétrique explicite, les effets "aléatoires", soit une forme moins contrainte appelée "modélisation à effets fixes", même si, dans certaines disciplines, le terme multiniveaux se rapporte implicitement aux premiers modèles. Ce document rappelle les forces et les faiblesses de chacune de ces modélisations, sachant qu'il n'est pas possible *a priori* de fournir une réponse définitive. Les modèles à effets aléatoires reposent sur des hypothèses d'indépendance plus fortes, et risquent donc de fournir des estimateurs biaisés, mais les modèles à effets fixes peuvent être moins précis. Le choix entre l'un ou l'autre de ces modèles se fera au cas par cas, après une analyse approfondie de la vraisemblance des hypothèses retenues, et de la structure des données (nombre de groupes et/ou nombres d'observations par groupe en particulier). Il faut aussi souligner que, comme toujours, cette analyse approfondie est indispensable, complétée par des tests de robustesses (est-ce que les conclusions de l'étude sont modifiées lorsque la modélisation varie) et ne peut être remplacée par des règles générales.

Références

Abowd, J. M., F. Kramarz et D. N. Margolis. 1999, «High wage workers and high wage firms», *Econometrica*, vol. 67, n° 2, p. 251–333.

Afsa, C. 2015, «Le modèle logit. Théorie et application.», Document de travail "Méthodologie Statistique" - DMS Working Paper 2016-01, Institut National de la Statistique et des Études Économiques.

Bates, D., M. Machler, B. M. Bolker et S. C. Walker. 2014, «Fitting linear mixed-effects models using lme4», R Package.

Bell, B. A., J. M. Ferron et J. D. Kromrey. 2008, «Cluster size in multilevel models : the impact of sparse data structures on point and interval estimates in two-level models», dans *Proceedings of the Joint Statistical Meetings*, p. 1122–1129.

Bliese, P. 2013, «Multilevel modeling in R (2.5). a brief introduction to R, the multilevel package and the nlme package», R Package.

Bressoux, P. 2007, «L'apport des modèles multiniveaux à la recherche en éducation», *Éducation et didactique*, vol. 1, n° 2, p. 73–88.

Bryan, M. L. et S. P. Jenkins. 2015, «Multilevel modelling of country effects : A cautionary tale», *European Sociological Review*.

Chaix, B. et P. Chauvin. 2002, «L'apport des modèles multiniveaux dans l'analyse contextuelle en épidémiologie sociale : une revue de littérature», *Revue d'épidemiologie et de santé publique*, vol. 50, p. 489–499.

Chaix, B. et P. Chauvin. 2005, «Influence du contexte de résidence sur les comportements de recours aux soins. L'apport des méthodes d'analyse multiniveaux et spatiales», cahier de recherche 104.

Chamberlain, G. 1980, «Analysis of covariance with qualitative data», *The Review of Economic Studies*, vol. 47, n° 1, p. 225–238.

Clark, T. S. et D. A. Linzer. 2015, «Should I use fixed or random effects ?», *Political Science Research and Methods*, vol. 3, p. 399–408.

Courgeau, D. 1994, «Du groupe à l'individu : l'exemple des comportements migratoires», *Population*, vol. 49, n° 1, p. 7–25.

Davezies, L. 2011, «Modèles à effets fixes, à effets aléatoires, modèles mixtes ou multi-niveaux : propriétés et mises en œuvre des modélisations de l'hétérogénéité dans le cas de données groupées», Documents de Travail de la DESE G2011-03, Institut National de la Statistique et des Études Économiques.

Duclos, M. et F. Murat. 2014, «Comment évaluer la performance des lycées ? Un point sur la méthodologie des IVAL (Indices de Valeur Ajoutée des Lycées)», *Éducation et Formation*, vol. 85, p. 73–84.

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

Gaure, S. 2013a, «LFE : Linear group fixed effects», *The R Journal*, vol. 5, n° 2, p. 104–117. User documentation of the 'lfe' package.

Gaure, S. 2013b, «OLS with multiple high dimensional category variables», *Computational Statistics & Data Analysis*, vol. 66, n° C, p. 8–18.

Gelman, A. et J. Hill. 2007, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Analytical Methods for Social Research, Cambridge University Press.

Golaz, V. et A. Bringé. 2009, «Apports et enjeux de l'analyse multiniveau en démographie», dans *Actes des 10eme Journées de Méthodologie Statistique*.

Guimarães, P. 2009, «REG2HDFE : Stata module to estimate a Linear Regression Model with two High Dimensional Fixed Effects», Statistical Software Components, Boston College Department of Economics.

Guimarães, P. et P. Portugal. 2010, «A Simple Feasible Procedure to Fit Models with High-Dimensional Fixed Effects», *Stata Journal*, vol. 10, n° 4, p. 628–649.

Hospido, L. 2012, «Estimating nonlinear models with multiple fixed effects : A computational note», *Oxford Bulletin of Economics and Statistics*, vol. 74, n° 5, p. 760–775.

Kiernan, K., J. Tao et P. Gibbs. 2012, «Tips and Strategies for Mixed Modeling with SAS/STAT Procedures», *SAS Global Forum*, , n° 332.

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

Li, B., H. Lingsma, E. Steyerberg et E. Lesaffre. 2011, «Logistic random effects regression models : a comparison of statistical packages for binary and ordinal outcomes», *BMC Medical Research Methodology*, vol. 11, n° 1, 77.

Maas, C. J. et J. J. Hox. 2005, «Sufficient sample sizes for multilevel modeling.», *Methodology : European Journal of Research Methods for the Behavioral and Social Sciences*, vol. 1, n° 3, p. 86.

Milcent, C. et J. Rochut. 2009, «Tarification hospitalière et pratique médicale. La pratique de la césarienne en France», *Revue économique*, vol. 60, n° 2, p. 489–506.

Moulton, B. R. 1990, «An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit», *The Review of Economics and Statistics*, vol. 72, n° 2, p. 334–38.

Mundlak, Y. 1978, «On the Pooling of Time Series and Cross Section Data», *Econometrica*, vol. 46, n° 1, p. 69–85.

Murat, F., F. Evain et L. Evrard. 2014, «Trois indicateurs de résultats des lycées publics et privés sous contrat», cahier de recherche, Ministère de l'Éducation Nationale - Direction de l'évaluation, de la prospective et de la performance.

Or, Z. et T. Renaud. 2009, «Impact du volume d'activité sur les résultats de soins à l'hôpital en France», *Économie publique*, vol. 24-25, p. 187–219.

Owens, A. 2010, «Neighborhoods and schools as competing and reinforcing contexts for educational attainment», *Sociology of Education*, vol. 83, n° 4, p. 287–311.

Pilorge, C., C. Marbot et R. Legal. 2013, «Coût de l'ordonnance des médecins généralistes. Peut-on caractériser les pratiques de prescription ?», Dossier Solidarités et santé 44.

Rabe-Hesketh, S. et A. Skrondal. 2005, *Multilevel and Longitudinal analysis using Stata*, Stata Press, College Station.

Raudenbush, S. W. 1993, «A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research», *Journal of Educational Statistics*, vol. 18, n° 4, p. 321–349.

Robinson, W. S. 1950, «Ecological correlations and the behavior of individuals», *American Sociological Review*, vol. 15, n° 3, p. 351–357.

Rodríguez, G. et N. Goldman. 1995, «An assessment of estimation procedures for multilevel models with binary responses», *Journal of the Royal Statistical Society, Series A*, , n° 158.

Snijders, T. et R. Bosker. 2011a, *Multilevel Analysis : An Introduction to Basic and Advanced Multilevel Modeling*, SAGE Publications.

Snijders, T. et R. Bosker. 2011b, *Multilevel Analysis : An Introduction to Basic and Advanced Multilevel Modeling*, chap. Imperfect hierarchies, SAGE Publications.

Snijders, T. A. B. et J. Berkhof. 2007, *Handbook of Multilevel Analysis*, chap. Diagnostic Checks for Multilevel Models, Springer, New York.

Vallet, L.-A. 2005, «La mesure des effets de quartier/voisinage : Un objet important et difficile à la croisée des sciences sociales : Commentaire», *Revue économique*, vol. 56, n° 2, p. 363–369.

Wang, J., H. Xie, J. Fisher et H. Press. 2011, *Multilevel Models : Applications using SAS®*, De Gruyter.

Wooldridge, J. M. 2002, *Econometric Analysis of Cross Section and Panel Data*, *MIT Press Books*, vol. 1, The MIT Press.

A Détails des programmes utilisés pour les exemples

A.1 Cas d'une variable continue

On utilise la table `echant`. La variable à expliquer est `log_salh`, les variables explicatives `SEXE` (2 modalités), `CS` (6 modalités), `AGE` et `AGESQ`. Chaque entreprise est identifiée par `Siren_empl`.

A.1.1 Programmes et sorties avec le logiciel Sas

La mise en œuvre du modèle à effets fixes sous Sas est la suivante :

```
PROC GLM DATA=echant;
ABSORB siren_empl;
MODEL logsal_h=sexe1 age agesq  cs2-cs6 /SOLUTION;
RUN;
```

On a ici transformé les variables `sexe` et `cs` en indicatrices, mais on aurait aussi pu les inclure simplement dans le modèle avec la ligne supplémentaire `CLASS cs sexe;`. La référence est sexe=0 (femme), et cs=1. Attention, pour utiliser l'instruction `ABSORB` il faut avoir préalablement trié la table selon `siren_empl`. On obtient l'estimation de l'effet des variables explicatives dans :

Parameter	Estimate	Standard Error	t Value	Pr > |t|
sexe1	0.0681669161	0.00203654	33.47	<.0001
AGE	0.0208168638	0.00043916	47.40	<.0001
agesq	-.0001776561	0.00000545	-32.59	<.0001
cs2	0.9244590893	0.02785076	33.19	<.0001
cs3	0.5213805954	0.02604065	20.02	<.0001
cs4	0.1232177167	0.02594624	4.75	<.0001
cs5	-.0949826267	0.02587533	-3.67	0.0002
cs6	-.1456642789	0.02587891	-5.63	<.0001

La syntaxe de la procédure `GLIMMIX` pour un modèle à effets aléatoires est :

```
PROC GLIMMIX DATA=echant noclprint;
CLASS siren_empl;
RANDOM intercept/subject=siren_empl;
MODEL logsal_h=sexe1 age agesq  cs2-cs6 /SOLUTION;
covtest 0;
RUN;
```

On trouve les estimations des variances des termes d'erreur dans

Cov Parm	Subject	Estimate	Standard Error
Intercept	SIREN_EMPL	0.02039	0.000692
Residual		0.05078	0.000253

Pour tester si la covariance des effets groupes est significativement non nulle, on ajoute une ligne `covtest 0;` . On obtient ici :

Tests of Covariance Parameters
Based on the Restricted Likelihood

Label	DF	-2 Res Log Like	ChiSq	Pr > ChiSq	Note
Parameter list	1	16720	22844.6	<.0001	MI

MI: P-value based on a mixture of chi-squares.

Les résultats des estimations pour les effets des variables sont dans l'item "fixed effects" (conformément à la typologie utilisée parfois pour les modèles dits mixtes, où le terme "effets aléatoires" (*random effects*) est réservé aux effets groupes, et le coefficient des variables observables du modèle, qu'on ne fait pas varier en fonction du groupe, est appelé "effets fixes").

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	1.8085	0.02685	1997	67.35	<.0001
sexe1	0.07077	0.002016	80683	35.11	<.0001
AGE	0.02098	0.000437	80683	47.98	<.0001
agesq	-0.00018	5.432E-6	80683	-33.03	<.0001
cs2	0.9286	0.02762	80683	33.62	<.0001
cs3	0.5277	0.02579	80683	20.46	<.0001
cs4	0.1280	0.02570	80683	4.98	<.0001
cs5	-0.09357	0.02563	80683	-3.65	0.0003
cs6	-0.1422	0.02564	80683	-5.55	<.0001

On peut obtenir une estimation des effets groupes en ajoutant l'instruction :

```
ODS OUTPUT SOLUTIONR=estim ;
```

mais il faut ajouter l'instruction `solution` (ou simplement en abbrégé `s`) dans la ligne correspondant à l'instruction `RANDOM`.

```
RANDOM intercept / s subject=siren_empl  ;
```

Une autre solution, plus indirecte, est d'utiliser l'instruction `OUTPUT` :

```
OUTPUT OUT=pred     pred(blup)=p  resid(blup)=r ;
```

Ici on crée une table `pred` avec pour tous les individus de la table initiale, les variables `p`, qui correspond à la prédiction obtenue à partir des estimations du modèle, et `r`, les résidus individuels. La précision `blup` signifie que `p` correspond à $X_{ij}\hat{\beta}+\hat{\gamma}_j$, `r` à $y_{ij}-X_{ij}\hat{\beta}-\hat{\gamma}_j$. Si on indique au contraire `noblup`, ces deux termes ne tiendront pas compte de $\hat{\gamma}_j$ ($\hat{\gamma}_j = 0$).

A.1.2 Programmes et sorties avec le logiciel R

Pour le modèle à effets aléatoires, on a écrit simplement :

```
modelecontinu <- lmer(formula=logsal_h ~  AGE+ agesq + sexe1+cs2+cs3+cs4+cs5+cs6
                      + (1|SIREN_EMPL ),data = echant,REML=TRUE)
```

dont on obtient les principaux résultats avec l'instruction `summary` :

```
summary(modelecontinu)
```

On obtient les coefficients estimés pour les variables là aussi appelés "fixed effects" :

```
Fixed effects:
              Estimate Std. Error t value
(Intercept)  1.808e+00  2.685e-02   67.35
AGE           2.098e-02  4.373e-04   47.98
agesq        -1.794e-04  5.432e-06  -33.03
sexe1         7.077e-02  2.016e-03   35.11
cs2           9.286e-01  2.762e-02   33.62
cs3           5.277e-01  2.579e-02   20.46
cs4           1.280e-01  2.570e-02    4.98
cs5          -9.357e-02  2.563e-02   -3.65
cs6          -1.422e-01  2.564e-02   -5.55
```

La partie "Random effects" fournit la variance et l'écart-type pour les deux termes aléatoires :

```
Random effects:
 Groups     Name        Variance Std.Dev.
 SIREN_EMPL (Intercept) 0.02039  0.1428
 Residual               0.05078  0.2253
Number of obs: 82689, groups:  SIREN_EMPL, 1998
```

sur la ligne de `SIREN_EMPL (Intercept)`, on a donc à la fois σ^2_α et σ_α, et la ligne *Residual* σ^2_ε et σ_ε. En revanche, au moment de la rédaction de ce document la procédure ne fournit pas de test de significativité de ce modèle. Il faut donc faire le test "à la main". Pour cela, on estime un modèle sans effet aléatoire (donc un simple modèle linéaire), en utilisant la procédure `lm`.

```
modeleols<-lm(formula=logsal_h~AGE+agesq+sexe1+cs2+cs3+cs4+cs5+cs6,data=tab)
```

La statistique de test du modèle avec effets aléatoires et sans effets aléatoires correspond au double de la différence de la log-vraisemblance des deux modèles (enregistré dans `logLik`) :

```
LR1 <- 2*(logLik(modelecontinu)-logLik(modeleols))[1]
```

Sous l'hypothèse nulle d'absence d'effets aléatoires, cette statistique de test `LR1` doit suivre une loi du χ^2 à 1 degré de liberté. On peut alors obtenir la P-value du test avec l'instruction :

```
1-pchisq(LR1,1)
```

L'estimation du modèle à effets fixes se fait par exemple avec la procédure `plm` qui estime le modèle within.

```
modelefe<-plm(formula=logsal_h~AGE+agesq+sexe1+cs2+cs3+cs4+cs5+cs6
,index=c("SIREN_EMPL"),data=tab)
```

On peut visualiser les résultats avec l'instruction : `summary(modelefe)`

```
Unbalanced Panel: n=1998, T=2-99, N=82689

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-7.72000 -0.10900 -0.00768  0.10100  3.78000

Coefficients :
         Estimate  Std. Error  t-value  Pr(>|t|)
AGE    2.0817e-02  4.3916e-04  47.4020 < 2.2e-16 ***
agesq -1.7766e-04  5.4514e-06 -32.5890 < 2.2e-16 ***
sexe1  6.8167e-02  2.0365e-03  33.4719 < 2.2e-16 ***
cs2    9.2446e-01  2.7851e-02  33.1933 < 2.2e-16 ***
cs3    5.2138e-01  2.6041e-02  20.0218 < 2.2e-16 ***
cs4    1.2322e-01  2.5946e-02   4.7490 2.048e-06 ***
cs5   -9.4983e-02  2.5875e-02  -3.6708  0.000242 ***
cs6   -1.4566e-01  2.5879e-02  -5.6287 1.822e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    7383.1
Residual Sum of Squares: 4097.1
R-Squared:      0.44507
Adj. R-Squared: 0.43427
F-statistic: 8088.8 on 8 and 80683 DF, p-value: < 2.22e-16
```

Le modèle linéaire obtenu en introduisant simplement autant d'indicatrices qu'il y a d'unités légales doit théoriquement fournir des résultats identiques, mais il n'a pas convergé ici du fait du grand nombre de paramètres en jeu.

A.1.3 Programmes et sorties avec le logiciel Stata

La mise en œuvre du modèle à effets fixes sous STATA peut se faire en deux temps. On peut d'abord définir par une option `xtset` la variable qui sert à définir les groupes, mais celle-ci doit être numérique. Si ce n'est pas le cas comme dans notre exemple (`siren_empl` est une variable caractère), il faut la transformer en numérique par l'instruction `encode siren_empl, gen(id)` qui crée ici une nouvelle variable numérique `id` à partir de la variable caractère `siren_empl`. La syntaxe est ensuite :

```
xtset id
xtreg logsal_h sexe1 age agesq  cs2-cs6, fe
```

Stata fournit la sortie ci-dessous. On notera que la procédure fournit des estimateurs empiriques des *écarts-types* des termes d'erreurs estimés (et non leur variance).

```
------------------------------------------------------------------------------
    logsal_h |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sexe1 |   .0681669   .0020365    33.47   0.000     .0641753    .0721585
```

```
         age |   .0208169   .0004392    47.40   0.000     .0199561    .0216776
       agesq |  -.0001777   5.45e-06   -32.59   0.000    -.0001883    -.000167
         cs2 |   .9244591   .0278508    33.19   0.000     .8698718    .9790464
         cs3 |   .5213806   .0260407    20.02   0.000     .4703411    .5724201
         cs4 |   .1232177   .0259462     4.75   0.000     .0723633    .1740722
         cs5 |  -.0949826   .0258753    -3.67   0.000    -.1456981   -.0442671
         cs6 |  -.1456643   .0258789    -5.63   0.000    -.1963868   -.0949418
       _cons |   1.820749   .0268992    67.69   0.000     1.768027    1.873471
-------------+----------------------------------------------------------------
     sigma_u |   .1480486
     sigma_e |   .2253444
         rho |  .30149717   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(1997, 80683) =    17.97         Prob > F = 0.0000
```

Pour un modèle à effets aléatoires, on peut utiliser la procédure `xtmixed`.

```
xtmixed logsal_h sexe1 age agesq  cs2-cs6  || siren_empl:, reml
```

après des informations sur la convergence du modèle, on obtient en particulier l'estimation des coefficients des variables.

```
------------------------------------------------------------------------------
    logsal_h |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sexe1 |   .0707678   .0020156    35.11   0.000     .0668173    .0747182
         age |   .0209832   .0004373    47.98   0.000     .0201261    .0218402
       agesq |  -.0001794   5.43e-06   -33.03   0.000    -.0001901   -.0001688
         cs2 |   .9286182   .0276193    33.62   0.000     .8744853     .982751
         cs3 |    .527722   .0257946    20.46   0.000     .4771656    .5782785
         cs4 |   .1279704   .0257015     4.98   0.000     .0775963    .1783444
         cs5 |  -.0935711   .0256324    -3.65   0.000    -.1438097   -.0433325
         cs6 |  -.1421651   .0256375    -5.55   0.000    -.1924137   -.0919166
       _cons |   1.808458   .0268516    67.35   0.000      1.75583    1.861086
------------------------------------------------------------------------------
```

La procédure fournit également les estimations des écarts-types des termes aléatoires :

```
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
siren_empl: Identity         |
                   sd(_cons) |   .1427957   .0024235      .1381238    .1476257
-----------------------------+------------------------------------------------
                sd(Residual) |   .2253464    .000561      .2242495    .2264485
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) = 22844.64 Prob >= chibar2 = 0.0000
```

Il s'agit des estimations des écarts-types des termes d'erreurs : `sd(_cons)` correspond à σ_α et `sd(Residual)` à σ_ε dans les notations du document. La procédure `xtmixed` de STATA fournit les estimations des écarts-types, alors que `GLIMMIX` de SAS fournit les estimations des variances (soit les carrés des précédentes), et `lme` fournit les deux. La ligne sous le tableau correspond au test de nullité des termes d'effets groupes.

A.2 Cas d'une variable binaire

A.2.1 Programmes et sorties avec le logiciel SAS

La procédure `LOGISTIC`, classiquement utilisée sous SAS pour l'estimation d'un modèle binaire, est également utilisée pour les modèles binaires à effets fixes. La différence est l'instruction `STRATA` qui permet d'estimer un logit conditionnel :

```
ODS OUTPUT ParameterEstimates=beta_FE;
PROC LOGISTIC data = idf.idf NAMELEN=32;
STRATA local;
MODEL retard (DESCENDING) =
/* variables niveau établissement: */
petab_elc_conatio_999_pourcent petab_bourse_1_pourcent
petab_pcs_indicateur_A_pourcent petab_pcs_indicateur_B_pourcent
petab_pcs_indicateur_D_pourcent prive
/* variables individu */
boursier etranger garcon pcsA pcsB pcsD ;
RUN;
```

On obtient une sortie classique donnant les estimations des coefficients du modèle :

The LOGISTIC Procedure

Conditional Analysis

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
petab_elc_conatio_999_pourcent	1	0.0131	0.00200	42.7957	<.0001
petab_bourse_1_pourcent	1	-0.00164	0.00126	1.6938	0.1931
petab_pcs_indicateur_A_pourcent	1	0.000833	0.00143	0.3389	0.5605
petab_pcs_indicateur_B_pourcent	1	0.00529	0.00218	5.8638	0.0155
petab_pcs_indicateur_D_pourcent	1	0.00416	0.00135	9.5557	0.0020
prive	1	0.00405	0.0509	0.0063	0.9365
boursier	1	0.1963	0.0257	58.2999	<.0001
etranger	1	0.7653	0.0368	431.5055	<.0001
garcon	1	0.2900	0.0202	206.2960	<.0001
pcsA	1	-1.1152	0.0386	834.6247	<.0001
pcsB	1	-0.3299	0.0374	77.6502	<.0001
pcsD	1	0.4798	0.0245	383.2801	<.0001

Pour un modèle binaire à effets aléatoires, on utilise la même procédure que pour les modèles linéaires à effets aléatoires : la procédure `GLIMMIX`. On indique que le modèle est binaire en ajoutant `DIST=BINARY` dans l'instruction `MODEL`.

```
ODS OUTPUT SOLUTIONR=alphaj PARAMETERESTIMATES=beta COVPARMS=sigma;
PROC GLIMMIX DATA= idf.idf NOCLPRINT namelen=32 ;
CLASS local ;
MODEL retard (DESCENDING) = part_basrev_25_pourcent part_etranger_25_pourcent
part_lochlm_25_pourcent part_5pers_25_pourcent part_monop_25_pourcent
part_pcsa_25_pourcent part_pcsb_25_pourcent part_pcsd_25_pourcent
part_boursier_25_pourcent petab_elc_conatio_999_pourcent
petab_bourse_1_pourcent petab_pcs_indicateur_A_pourcent
petab_pcs_indicateur_B_pourcent petab_pcs_indicateur_D_pourcent
boursier etranger garcon pcsA pcsB pcsD prive
/ DIST=BINARY link=logit DDFM=BW SOLUTION;
RANDOM INTERCEPT / SUBJECT = local SOLUTION;
COVTEST 'Test de nullité de la variance inter-classe' 0;
OUTPUT OUT=prediction /* probabilité prédite: */ pred(blup ILINK)=pred
/* xbeta + alpha_j (pour le calcul des APE)*/ pred(blup NOILINK)=xbetapalpha;
RUN;
```

En sortie, SAS fournit l'estimation de la variance inter-classe σ^2_α. SAS ne fournit pas la variance intra-classe σ^2_ε car elle est fixée dans un modèle binaire (ici à $\pi^2/3$ car il s'agit d'une modélisation logit).

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	local	0.1066	0.01282

L'estimation des coefficients des variables explicatives :

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-2.1800	0.1411	32312	-15.45	<.0001
part_basrev_25_pourcent	0.002613	0.003481	32312	0.75	0.4529
part_etranger_25_pourcent	-0.00343	0.003767	32312	-0.91	0.3631
part_lochlm_25_pourcent	0.000101	0.000692	32312	0.15	0.8839
part_5pers_25_pourcent	-0.00845	0.002620	32312	-3.22	0.0013
part_monop_25_pourcent	0.02927	0.009293	32312	3.15	0.0016
part_pcsa_25_pourcent	-0.00841	0.001815	32312	-4.63	<.0001
part_pcsb_25_pourcent	0.002209	0.003482	32312	0.63	0.5259
part_pcsd_25_pourcent	-0.00068	0.002336	32312	-0.29	0.7698
part_boursier_25_pourcent	0.003959	0.002323	32312	1.70	0.0883
petab_elc_conatio_999_pourcent	0.007727	0.001466	103E3	5.27	<.0001
petab_bourse_1_pourcent	-0.00177	0.000921	103E3	-1.92	0.0553
petab_pcs_indicateur_A_pourcent	-0.00021	0.001054	103E3	-0.20	0.8441
petab_pcs_indicateur_B_pourcent	0.003791	0.001520	103E3	2.49	0.0126
petab_pcs_indicateur_D_pourcent	0.003057	0.000957	103E3	3.19	0.0014
boursier	0.2884	0.02280	103E3	12.65	<.0001
etranger	0.7969	0.03170	103E3	25.14	<.0001
garcon	0.2661	0.01761	103E3	15.11	<.0001
pcsA	-1.1984	0.03357	103E3	-35.70	<.0001

pcsB	-0.3447	0.03269	103E3	-10.54	<.0001
pcsD	0.4841	0.02122	103E3	22.82	<.0001
prive	-0.05149	0.04203	103E3	-1.23	0.2205

Le test de nullité de la variance inter-classe :

Tests of Covariance Parameters
Based on the Residual Pseudo-Likelihood

Label	DF	-2 Res Log P-Like	ChiSq	Pr > ChiSq	Note
Test de nullité de la variance inter-classe	1	724834	84.09	<.0001	MI

MI: P-value based on a mixture of chi-squares.

Pour le modèle de Mundlak, on ajoute simplement les moyennes par groupe au modèle à effets aléatoires. Pour tester la nullité jointe des coefficients de ces "nouvelles variables", on ajoute l'instruction `CONTRAST` :

```
CONTRAST 'Test de Mundlak'
moy_q_petab_elc_conatio_999_ 1, moy_q_petab_bourse_1_pourcen 1,
moy_q_petab_pcs_indicateur_A 1,moy_q_petab_pcs_indicateur_B 1 ,
moy_q_petab_pcs_indicateur_D 1,moy_q_boursier 1, moy_q_etranger 1,
moy_q_garcon 1, moy_q_pcsA 1, moy_q_pcsB 1, moy_q_pcsD 1, moy_q_prive 1;
```

La sortie associée :

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
Test de Mundlak	12	32300	11.37	<.0001

A.2.2 Programmes et sorties avec le logiciel R

L'estimation du modèle à effets fixes par un logit conditionnel se fait avec la fonction `clogit` du package `survival`.

```
modelebinFE<-clogit(retard~ petab_elc_conatio_999_pourcent + petab_bourse_1_pourcent
+petab_pcs_indicateur_A_pourcent+ petab_pcs_indicateur_B_pourcent
+ petab_pcs_indicateur_D_pourcent +prive+boursier +etranger
+ garcon +pcsA+ pcsB+ pcsD+strata(local),data =idf)
```

On obtient classiquement sous R les sorties par l'option `summary` :

```
summary(modelebinFE)

Call:
coxph(formula = Surv(rep(1, 138915L), retard) ~ petab_elc_conatio_999_pourcent +
    petab_bourse_1_pourcent + petab_pcs_indicateur_A_pourcent +
    petab_pcs_indicateur_B_pourcent + petab_pcs_indicateur_D_pourcent +
    prive + boursier + etranger + garcon + pcsA + pcsB + pcsD +
    strata(local), data = idf, method = "exact")
```

```
  n= 138915, number of events= 16645

                                      coef  exp(coef)   se(coef)       z Pr(>|z|)
petab_elc_conatio_999_pourcent   0.0129039  1.0129875  0.0019628   6.574 4.89e-11 ***
petab_bourse_1_pourcent         -0.0014827  0.9985184  0.0012385  -1.197  0.23125
petab_pcs_indicateur_A_pourcent  0.0009965  1.0009969  0.0014108   0.706  0.47999
petab_pcs_indicateur_B_pourcent  0.0049545  1.0049667  0.0021606   2.293  0.02184 *
petab_pcs_indicateur_D_pourcent  0.0040085  1.0040166  0.0013248   3.026  0.00248 **
prive                           -0.0087801  0.9912583  0.0506007  -0.174  0.86224
boursier                         0.2023555  1.2242831  0.0253067   7.996 1.33e-15 ***
etranger                         0.7701966  2.1601910  0.0361983  21.277  < 2e-16 ***
garcon                           0.2874918  1.3330796  0.0199171  14.434  < 2e-16 ***
pcsA                            -1.1132068  0.3285038  0.0383399 -29.035  < 2e-16 ***
pcsB                            -0.3250237  0.7225103  0.0369756  -8.790  < 2e-16 ***
pcsD                             0.4772691  1.6116671  0.0241572  19.757  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                 exp(coef) exp(-coef) lower .95 upper .95
petab_elc_conatio_999_pourcent      1.0130     0.9872    1.0091    1.0169
petab_bourse_1_pourcent             0.9985     1.0015    0.9961    1.0009
petab_pcs_indicateur_A_pourcent     1.0010     0.9990    0.9982    1.0038
petab_pcs_indicateur_B_pourcent     1.0050     0.9951    1.0007    1.0092
petab_pcs_indicateur_D_pourcent     1.0040     0.9960    1.0014    1.0066
prive                               0.9913     1.0088    0.8977    1.0946
boursier                            1.2243     0.8168    1.1650    1.2865
etranger                            2.1602     0.4629    2.0122    2.3190
garcon                              1.3331     0.7501    1.2820    1.3861
pcsA                                0.3285     3.0441    0.3047    0.3541
pcsB                                0.7225     1.3841    0.6720    0.7768
pcsD                                1.6117     0.6205    1.5371    1.6898

Rsquare= 0.026   (max possible= 0.302 )
Likelihood ratio test= 3658  on 12 df,   p=0
Wald test            = 3171  on 12 df,   p=0
Score (logrank) test = 3529  on 12 df,   p=0
```

Pour le modèle à effets aléatoires on peut utiliser la procédure `glmer`.

```
modelebinRE<-glmer(retard ~ part_basrev_25_pourcent+ part_etranger_25_pourcent
+ part_lochlm_25_pourcent +part_5pers_25_pourcent+ part_monop_25_pourcent
+part_pcsa_25_pourcent+ part_pcsb_25_pourcent +part_pcsd_25_pourcent+
part_boursier_25_pourcent+ petab_elc_conatio_999_pourcent+
petab_bourse_1_pourcent +petab_pcs_indicateur_A_pourcent+
petab_pcs_indicateur_B_pourcent + petab_pcs_indicateur_D_pourcent
+ boursier+ etranger+ garcon +pcsA +pcsB+ pcsD +prive
+(1|local),data=idf,family=binomial(link=logit))
```

dont les résultats en sortie sont :

```
     AIC      BIC   logLik deviance df.resid
 89691.0  89916.9 -44822.5  89645.0   135810

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.3847 -0.4026 -0.2911 -0.1667  7.5844
```

```
Random effects:
 Groups Name        Variance Std.Dev.
 local  (Intercept) 0.1339   0.3659
Number of obs: 135833, groups:  local, 32322

Fixed effects:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -2.2242855  0.1438766  -15.46  < 2e-16 ***
part_basrev_25_pourcent         0.0025013  0.0035503    0.70 0.481110
part_etranger_25_pourcent      -0.0035702  0.0038434   -0.93 0.352933
part_lochlm_25_pourcent         0.0001888  0.0007067    0.27 0.789412
part_5pers_25_pourcent         -0.0088131  0.0026709   -3.30 0.000968 ***
part_monop_25_pourcent          0.0289847  0.0094668    3.06 0.002201 **
part_pcsa_25_pourcent          -0.0085663  0.0018431   -4.65 3.36e-06 ***
part_pcsb_25_pourcent           0.0019798  0.0035342    0.56 0.575361
part_pcsd_25_pourcent          -0.0005068  0.0023771   -0.21 0.831183
part_boursier_25_pourcent       0.0039465  0.0023676    1.67 0.095535 .
petab_elc_conatio_999_pourcent  0.0080000  0.0014887    5.37 7.70e-08 ***
petab_bourse_1_pourcent        -0.0017065  0.0009330   -1.83 0.067395 .
petab_pcs_indicateur_A_pourcent -0.0001719  0.0010645   -0.16 0.871682
petab_pcs_indicateur_B_pourcent  0.0038374  0.0015362    2.50 0.012489 *
petab_pcs_indicateur_D_pourcent  0.0030813  0.0009695    3.18 0.001481 **
boursier                         0.2915786  0.0230134   12.67  < 2e-16 ***
etranger                         0.8117178  0.0321581   25.24  < 2e-16 ***
garcon                           0.2698082  0.0177569   15.19  < 2e-16 ***
pcsA                            -1.2101556  0.0336794  -35.93  < 2e-16 ***
pcsB                            -0.3523395  0.0328777  -10.72  < 2e-16 ***
pcsD                             0.4893750  0.0214048   22.86  < 2e-16 ***
prive                           -0.0532556  0.0423191   -1.26 0.208236
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A.2.3 Programmes et sorties avec le logiciel STATA

La mise en œuvre du modèle à effets fixes sous stata pour un modèle binaire se fait simplement en utilisant la procédure `clogit`. La syntaxe est alors la suivante :

```
clogit retard petab_elc_conatio_999_pourcent petab_bourse_1_pourcent
petab_pcs_indicateur_a_pourcent petab_pcs_indicateur_b_pourcent
petab_pcs_indicateur_d_pourcent prive boursier etranger
garcon pcsa pcsb pcsd,group(local)
```

On obtient les résultats :

```
Conditional (fixed-effects) logistic regression   Number of obs   =      69869
                                                  LR chi2(12)     =    3593.30
                                                  Prob > chi2     =     0.0000
Log likelihood = -22531.549                       Pseudo R2       =     0.0739

------------------------------------------------------------------------------
      retard |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
p~9_pourcent |    .013067   .0019975     6.54   0.000     .0091521    .0169819
p~1_pourcent |   -.001639   .0012593    -1.30   0.193    -.0041072    .0008293
```

```
p~a_pourcent |   .0008329   .0014309     0.58   0.560    -.0019715    .0036374
p~b_pourcent |   .0052886    .002184     2.42   0.015     .0010081    .0095691
p~d_pourcent |   .0041636   .0013469     3.09   0.002     .0015237    .0068035
       prive |   .0040542   .0509227     0.08   0.937    -.0957524    .1038609
    boursier |   .1963037   .0257096     7.64   0.000     .1459139    .2466936
    etranger |   .7652826   .0368408    20.77   0.000     .6930761    .8374892
      garcon |   .2900311   .0201929    14.36   0.000     .2504537    .3296085
        pcsa |  -1.115215   .0386023   -28.89   0.000    -1.190874   -1.039556
        pcsb |  -.3298518   .0374324    -8.81   0.000    -.4032179   -.2564856
        pcsd |   .4797527   .0245053    19.58   0.000     .4317233    .5277822
------------------------------------------------------------------------------
```

Pour le modèle à effets aléatoires, plusieurs procédures peuvent être utilisées. On décrit ici les résultats obtenus avec la procédure `xtlogit`. Comme précédemment, il faut définir d'abord la variable de groupe par l'instruction `xtset`, qui n'accepte qu'une variable numérique (qui peut être créée si nécessaire par `encode`).

```
encode local,gen(id)
xtset id
xtlogit retard part_basrev_25_pourcent part_etranger_25_pourcent
part_lochlm_25_pourcent part_5pers_25_pourcent part_monop_25_pourcent
part_pcsa_25_pourcent part_pcsb_25_pourcent part_pcsd_25_pourcent
part_boursier_25_pourcent petab_elc_conatio_999_pourcent
petab_bourse_1_pourcent petab_pcs_indicateur_a_pourcent
petab_pcs_indicateur_b_pourcent petab_pcs_indicateur_d_pourcent
boursier etranger garcon pcsa pcsb pcsd prive,re
```

On obtient la sortie suivante :

```
Fitting comparison model:

Iteration 0:   log likelihood = -49528.933
Iteration 1:   log likelihood = -45664.515
Iteration 2:   log likelihood = -44882.845
Iteration 3:   log likelihood = -44874.563
Iteration 4:   log likelihood = -44874.555
Iteration 5:   log likelihood = -44874.555

Fitting full model:

tau =  0.0     log likelihood = -44874.555
tau =  0.1     log likelihood = -44835.992
tau =  0.2     log likelihood =  -44895.28

Iteration 0:   log likelihood = -44835.992
Iteration 1:   log likelihood = -44826.022
Iteration 2:   log likelihood = -44825.991
Iteration 3:   log likelihood = -44825.991
```

```
Random-effects logistic regression              Number of obs      =    135833
Group variable: id                              Number of groups   =     32322

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       4.2
                                                               max =        57

                                                Wald chi2(21)      =   7106.21
Log likelihood  = -44825.991                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      retard |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
part_basre~t |   .0027511   .0035362     0.78   0.437    -.0041798     .009682
part_etran~t |  -.0035025   .0038287    -0.91   0.360    -.0110066    .0040016
part_lochl~t |   .0001364   .0007036     0.19   0.846    -.0012427    .0015154
part_5pers~t |  -.0085698   .0026609    -3.22   0.001     -.013785   -.0033547
part_monop~t |   .0295816   .0094291     3.14   0.002     .0111009    .0480622
part_pcsa_~t |  -.0085173   .0018385    -4.63   0.000    -.0121206   -.0049139
part_pcsb_~t |   .0022248   .0035258     0.63   0.528    -.0046857    .0091352
part_pcsd_~t |  -.0006903   .0023702    -0.29   0.771    -.0053358    .0039551
part_bours~t |   .0040871   .0023592     1.73   0.083    -.0005369     .008711
p~9_pourcent |   .0079401   .0014861     5.34   0.000     .0050275    .0108527
p~1_pourcent |  -.0017739   .0009319    -1.90   0.057    -.0036004    .0000525
p~a_pourcent |  -.0001908   .0010642    -0.18   0.858    -.0022767     .001895
p~b_pourcent |   .0038415   .0015352     2.50   0.012     .0008325    .0068505
p~d_pourcent |   .0031318   .0009684     3.23   0.001     .0012337    .0050298
    boursier |   .2929044   .0230167    12.73   0.000     .2477925    .3380162
    etranger |   .8122605   .0321223    25.29   0.000      .749302     .875219
      garcon |   .2701394   .0177667    15.20   0.000     .2353174    .3049615
        pcsa |  -1.206783   .0337084   -35.80   0.000     -1.27285   -1.140716
        pcsb |  -.3484459   .0329014   -10.59   0.000    -.4129315   -.2839603
        pcsd |    .492027   .0214226    22.97   0.000     .4500396    .5340145
       prive |  -.0506451   .0423172    -1.20   0.231    -.1335854    .0322952
       _cons |   -2.23078    .143383   -15.56   0.000    -2.511806   -1.949755
-------------+----------------------------------------------------------------
    /lnsig2u |  -2.085492   .1194734                     -2.319656   -1.851329
-------------+----------------------------------------------------------------
     sigma_u |   .3524854   .0210563                      .3135402    .3962681
         rho |   .0363919   .0041896                      .0290149    .0455565
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =    97.13 Prob >= chibar2 = 0.000
```

L'option `or` permet d'obtenir directement les valeurs des odds-ratios (ie l'exponentielle des coefficients). On peut également augmenter le nombre de points par l'instruction `intpoints()`. Les méthodes reposent en effet sur des approximations numériques d'intégrales et il peut être utile de vérifier que les estimations ne sont pas modifiées lorsqu'on augmente ce nombre.

La syntaxe de la procédure `xtmelogit` est identique à `xtmixed`, mais on précise ici le nombre de points utilisés pour l'approximation numérique de l'intégrale.

```
xtmelogit retard part_basrev_25_pourcent part_etranger_25_pourcent
part_lochlm_25_pourcent part_5pers_25_pourcent part_monop_25_pourcent
part_pcsa_25_pourcent part_pcsb_25_pourcent part_pcsd_25_pourcent
part_boursier_25_pourcent petab_elc_conatio_999_pourcent
petab_bourse_1_pourcent petab_pcs_indicateur_a_pourcent
petab_pcs_indicateur_b_pourcent petab_pcs_indicateur_d_pourcent
boursier etranger garcon pcsa pcsb pcsd prive|| local:,intpoints(30)
```

L'estimation est beaucoup plus lente.

```
Refining starting values:

Iteration 0:   log likelihood = -45739.291
Iteration 1:   log likelihood =  -44907.65
Iteration 2:   log likelihood = -44826.581

Performing gradient-based optimization:

Iteration 0:   log likelihood = -44826.581
Iteration 1:   log likelihood = -44825.991
Iteration 2:   log likelihood = -44825.991

Mixed-effects logistic regression               Number of obs      =    135833
Group variable: local                           Number of groups   =     32322

                                                Obs per group: min =         1
                                                               avg =       4.2
                                                               max =        57

Integration points =  30                        Wald chi2(21)      =   7106.22
Log likelihood = -44825.991                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      retard |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
part_basre~t |   .0027511   .0035362     0.78   0.437    -.0041798     .009682
part_etran~t |  -.0035025   .0038287    -0.91   0.360    -.0110066    .0040017
part_lochl~t |   .0001364   .0007036     0.19   0.846    -.0012427    .0015154
part_5pers~t |  -.0085699   .0026609    -3.22   0.001    -.0137851   -.0033547
part_monop~t |   .0295817   .0094291     3.14   0.002      .011101    .0480623
part_pcsa_~t |  -.0085172   .0018385    -4.63   0.000    -.0121206   -.0049139
part_pcsb_~t |   .0022248   .0035258     0.63   0.528    -.0046857    .0091352
part_pcsd_~t |  -.0006903   .0023701    -0.29   0.771    -.0053357    .0039551
part_bours~t |   .0040871   .0023592     1.73   0.083    -.0005369     .008711
p~9_pourcent |   .0079401   .0014861     5.34   0.000     .0050275    .0108527
p~1_pourcent |  -.0017739   .0009319    -1.90   0.057    -.0036004    .0000525
p~a_pourcent |  -.0001909   .0010642    -0.18   0.858    -.0022767     .001895
p~b_pourcent |   .0038415   .0015352     2.50   0.012     .0008325    .0068505
p~d_pourcent |   .0031317   .0009684     3.23   0.001     .0012336    .0050298
```

```
    boursier |   .2929043   .0230167    12.73   0.000     .2477924    .3380162
    etranger |   .8122603   .0321223    25.29   0.000     .7493018    .8752188
      garcon |   .2701394   .0177667    15.20   0.000     .2353174    .3049615
        pcsa |  -1.206784   .0337084   -35.80   0.000    -1.272851   -1.140717
        pcsb |  -.3484464   .0329014   -10.59   0.000     -.412932   -.2839608
        pcsd |   .4920269   .0214225    22.97   0.000     .4500394    .5340143
       prive |  -.0506451   .0423173    -1.20   0.231    -.1335854    .0322952
       _cons |   -2.23078   .1433811   -15.56   0.000    -2.511801   -1.949758
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
local: Identity              |
                  sd(_cons) |   .3524849   .0210563      .3135396    .3962676
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =    97.13 Prob>=chibar2 = 0.0000
```

Série des Documents de Travail « Méthodologie Statistique »

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

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

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

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

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

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

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

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

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

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

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

C. LAGARENNE, C. THIESSET

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS – K.MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS – K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR – R. LE SAOUT B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD – M. GUILLERM