

**Direction des Études et Synthèses Économiques**

**G 2011 / 03**

**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**

**Laurent DAVEZIES**

**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  
de la Direction des Études et Synthèses Économiques*

**G 2011 / 03**

**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**

**Laurent DAVEZIES \***

JANVIER 2011

Je remercie mes anciens collègues de la Direction de l'évaluation, de la prospective et de la performance au Ministère de l'éducation nationale, qui, par leurs interrogations récurrentes, ont contribué à initier et améliorer la rédaction de ce document.

Je remercie également Romain AEBERHARDT, Cédric AFSA, Magali BEFFY, Chantal BRUTEL, Nila CECI-RENAUD, Vincent COTTET, Pauline GIVORD et Fabienne ROSENWALD pour leurs encouragements.

---

\* Faisait partie du Département des Études Économiques d'Ensemble - Division « Redistribution et Politiques Sociales » au moment de la rédaction de ce document.

**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**

**Résumé**

Ce document présente et discute les modélisations possibles de l'hétérogénéité sur données individuelles pouvant faire l'objet de regroupements, par exemple des données de résultats scolaires pour des élèves regroupés par classes ou par établissements. Sans aborder l'ensemble des modélisations possibles, il s'agit essentiellement de discuter, dans le cadre d'un modèle linéaire, de la pertinence d'une modélisation à effets fixes ou à effets aléatoires en fonction de la problématique traitée, et des constatations empiriques issues des données traitées. Les hypothèses statistiques des modèles sont discutées ainsi que les propriétés des estimateurs qui en découlent. Pour chacun des modèles, des exemples de programmes SAS sont présentés. Les tests d'Hausman et l'usage que l'empiriste peut en faire pour asseoir son choix de modélisation sont exposés. Au-delà du modèle linéaire, les modèles concernant les variables binaires sont présentés dans un dernier chapitre.

**Mots-clés** : effets fixes, effets aléatoires, test d'Hausman, modèle multiniveaux, modèle mixte

---

**Fixed effects models, random effects models,  
mixed models or multilevel models :  
properties and implementation of modeling  
of the heterogeneity in presence of clustered data**

**Abstract**

This document presents the different ways to model heterogeneity in case of clustering, such pupils achievement in classroom or schools. In the linear framework, it essentially discusses the pertinence of fixed or random effects assumptions depending upon the goal pursued and the empirical evidence displayed by data. Statistical assumptions of the different models are introduced and successively discussed, as well as the properties of the estimators that are derived. For each model, SAS code is provided. Hausman tests, and their use to choose models, are explained. Beyond the linear framework, binary models are presented in the last chapter.

**Keywords:** fixed effect, random effect, Hausman test, multilevel model, mixed model

**Classification JEL** : C01, I21

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	La problématique : l'exemple des données scolaires . . . . .	5
1.2	Organisation du document . . . . .	6
<b>2</b>	<b>Rappels d'économétrie linéaire</b>	<b>9</b>
2.1	Qu'est ce qu'un modèle linéaire ? . . . . .	9
2.2	Le cadre . . . . .	9
2.3	Définition, conditions d'existence et d'unicité de l'estimateur des Moindres Carrés Ordinaires (MCO). . . . .	10
2.4	Quand l'estimateur des MCO est-il sans biais ? . . . . .	12
2.5	Quelle est la précision de l'estimateur des MCO ? . . . . .	14
2.6	Comment estimer la précision de l'estimateur des MCO ? . . . . .	16
2.7	L'estimateur des moindres carrés ordinaires est-il convergent ? . . . . .	17
2.8	Quelle est la loi asymptotique de l'estimateur ? . . . . .	18
2.9	Quelle est la loi à distance finie de l'estimateur si les résidus sont normaux ? . . . .	20
2.10	Comparaison avec l'estimateur du Maximum de Vraisemblance (MV) . . . . .	23
2.11	Programmation sous SAS . . . . .	24
<b>3</b>	<b>Le modèle à effets fixes</b>	<b>25</b>
3.1	Le modèle théorique . . . . .	25
3.1.1	Le modèle théorique de base . . . . .	25
3.2	Identification et définition de l'estimateur . . . . .	26
3.2.1	Identification du modèle . . . . .	26
3.2.2	Définition de l'estimateur des Moindres Carrés Contraints (MCC) . . . . .	28
3.2.3	Remarque sur le multiplicateur de Lagrange dans l'estimation directe . . . . .	28
3.3	Estimation par projection sur la dimension intra-classe . . . . .	30
3.3.1	Le théorème de Frisch-Waugh . . . . .	30
3.3.2	Le théorème de Frisch-Waugh appliqué au modèle à effets fixes . . . . .	31
3.4	Propriétés de l'estimateur des MCC . . . . .	33
3.4.1	L'estimateur est-il sans biais ? . . . . .	33
3.4.2	Quelle est la précision de l'estimateur ? . . . . .	34
3.4.3	Comment estimer la précision de l'estimateur ? . . . . .	37
3.4.4	L'estimateur est-il convergent ? . . . . .	38
3.4.5	Quelles sont les propriétés asymptotiques de l'estimateur ? . . . . .	39
3.4.6	Quelle est la loi à distance finie de l'estimateur si les résidus sont normaux ? . . . .	41
3.4.7	Estimation de l'intervalle de confiance des estimateurs dans l'estimation en deux étapes . . . . .	43
3.5	La prise en compte des variables constantes au sein des classes . . . . .	43
3.6	Programmation sous SAS... . . . .	44
3.6.1	... par l'estimateur "Within" . . . . .	44
3.6.2	... ou directement grâce à la proc reg . . . . .	50

<b>4</b>	<b>Le modèle à erreurs composées (autrement appelé "modèle multi-niveaux" ou modèle "mixte")</b>	<b>51</b>
4.1	Le modèle théorique de base . . . . .	51
4.2	Interprétation des paramètres estimés . . . . .	52
4.2.1	Le cas des variables constantes au sein de la classe . . . . .	53
4.2.2	Le cas des variables non constantes au sein de la classe . . . . .	53
4.3	Le lien avec les différents modèles de la littérature . . . . .	53
4.4	L'estimation des $\beta$ . . . . .	54
4.4.1	Propriétés de l'estimateur des MCO "naïf" . . . . .	54
4.4.2	Propriétés de l'estimateur à effets fixes . . . . .	54
4.4.3	L'estimateur des Moindres Carrés Généralisés (MCG) . . . . .	55
4.4.4	Quelques estimateurs de $\sigma^2$ et $\sigma_\gamma^2$ . . . . .	56
4.4.5	Estimateurs des Moindres Carrés Quasi-Généralisés (MCQG) . . . . .	58
4.4.6	Estimateur du Maximum de Vraisemblance (MV) et estimateur du Maximum de Vraisemblance Restreint (MVR) sous hypothèses de normalité des résidus	60
4.4.7	Comparaison de l'estimateur du MV et de l'estimateur du MVR . . . . .	65
4.5	Estimateurs des $\gamma$ . . . . .	67
4.6	Les tests . . . . .	69
4.6.1	Les tests sur $\beta$ . . . . .	69
4.6.2	Tests sur $\sigma_\gamma^2$ . . . . .	69
4.6.3	Les tests sur $\gamma$ . . . . .	70
4.7	Programmation sous SAS... . . . .	70
4.7.1	... des MCQG . . . . .	70
4.7.2	... ou des autres estimateurs . . . . .	71
<b>5</b>	<b>Quel modèle choisir ?</b>	<b>72</b>
5.1	Discussion théorique . . . . .	72
5.1.1	Un cadre commun de comparaison des modèles . . . . .	72
5.1.2	Le modèle à effets fixes est un sur-modèle du modèle à effets aléatoires . . .	73
5.1.3	Test d'Hausman . . . . .	74
5.2	Le modèle de Mundlak . . . . .	75
5.2.1	Les propriétés du modèle de Mundlak . . . . .	75
5.2.2	Implémentation du test d'Hausman et amélioration des estimations . . . . .	78
5.3	Programmation sous SAS... . . . .	80
5.3.1	... du modèle de Mundlak . . . . .	80
5.3.2	... et du test d'Hausman . . . . .	80
<b>6</b>	<b>Extension au cas d'une variable dépendante dichotomique</b>	<b>82</b>
6.1	Rappel sur les modèles à variables qualitatives . . . . .	82
6.1.1	Le modèle linéaire de probabilité . . . . .	83
6.1.2	L'estimateur du score maximum de Manski . . . . .	83
6.1.3	Les modèles Probit et Logit . . . . .	85
6.2	Problème de biais du Logit et du Probit à effets fixes . . . . .	88
6.2.1	Illustration du problème des paramètres "incidentes" . . . . .	89
6.3	Le Logit conditionnel . . . . .	90
6.3.1	Principe du maximum de vraisemblance conditionnelle . . . . .	90
6.3.2	Application aux modèles Logit . . . . .	91
6.4	Les modèles à effets aléatoires . . . . .	93
6.5	Programmation sous SAS... . . . .	94
6.5.1	... du Logit conditionnel . . . . .	94
6.5.2	... des modèles à effets aléatoires . . . . .	94

# Chapitre 1

## Introduction

Ce document présente les principales méthodes dont dispose l'économètre pour modéliser l'hétérogénéité entre unités micro-économiques, lorsque ces unités micro-économiques appartiennent à des groupements. On parle parfois de modélisation "multiniveaux" pour qualifier les méthodes économétriques applicables à ces données. Ce document peut donc être également vu comme introduction à l'analyse multiniveaux.

### 1.1 La problématique : l'exemple des données scolaires

Les données groupées se rencontrent dans de nombreux domaines : les ménages constituent un regroupement de données individuelles dans certains fichiers d'enquête, des quartiers ou des zones géographiques constituent des regroupements de ménages ou d'individus, des entreprises sont des regroupements de salariés. Dans ce document, on se référera de manière systématique à un domaine particulier, celui de l'économie de l'éducation, dans lequel les données sur les élèves sont en général disponibles sous forme regroupée par classe ou par établissement, pour des raisons de coût de collecte<sup>1</sup>. La question-type que se posent les études en économie de l'éducation est de modéliser les comportements ou les performances des élèves. On cherche ainsi à rendre compte de leur hétérogénéité et à l'expliquer par leurs caractéristiques individuelles. Or, parmi ces caractéristiques individuelles, certaines sont partagées par les élèves d'une même classe (par exemple le fait d'être scolarisé en ZEP ou dans le privé, le fait que l'instituteur ou le professeur ait telle caractéristique etc.). Ne pas prendre en compte la classe ou l'établissement pour expliquer l'hétérogénéité des élèves peut conduire à deux types de problèmes : ceci peut conduire à des estimations biaisées, et dont la précision sera surestimée.

Expliquons brièvement l'origine de ces deux problèmes.

Le biais des estimations relève du "mécanisme de la variable omise". Imaginons que les résultats des élèves soient entièrement déterminés par deux facteurs : leur niveau initial et l'établissement fréquenté. Si l'établissement fréquenté n'est pas pris en compte dans l'analyse, comme le niveau initial et l'efficacité de l'établissement sont vraisemblablement corrélés, l'importance du niveau initial des élèves sera mal estimée. Ceci est illustré sur la figure 1.1 page 6. Sur cette figure les élèves de niveau initial plus élevé sont regroupés dans des classes ou des établissements plus performants. Une régression naïve qui ne prend pas ce facteur en compte tendra, dans ce cas, à surestimer l'impact du niveau initial sur le niveau final. Le problème est loin d'être neutre en terme de politiques publiques.

S'agissant de la surestimation de la précision des coefficients, elle se comprend bien sur le cas suivant. Imaginons que "toutes choses égales par ailleurs", les établissements privés soient moins performants que les établissements publics et que l'on estime cette efficacité sur un fichier de 200 classes, comprenant en tout 4 000 élèves. Dans le cas où la modélisation se fait en régressant le niveau des élèves sur leurs caractéristiques individuelles et sur une variable public/privé, l'ordre de

---

1. Dans le cas d'une enquête par sondage, même dans le cas où tous les élèves d'une classe ne sont pas interrogés, plusieurs élèves de la même classe sont généralement sollicités de manière à diminuer les coûts de collecte. Dans le cas de données issues de fichiers administratifs, les élèves d'une même classe sont généralement tous présents dans le fichier.

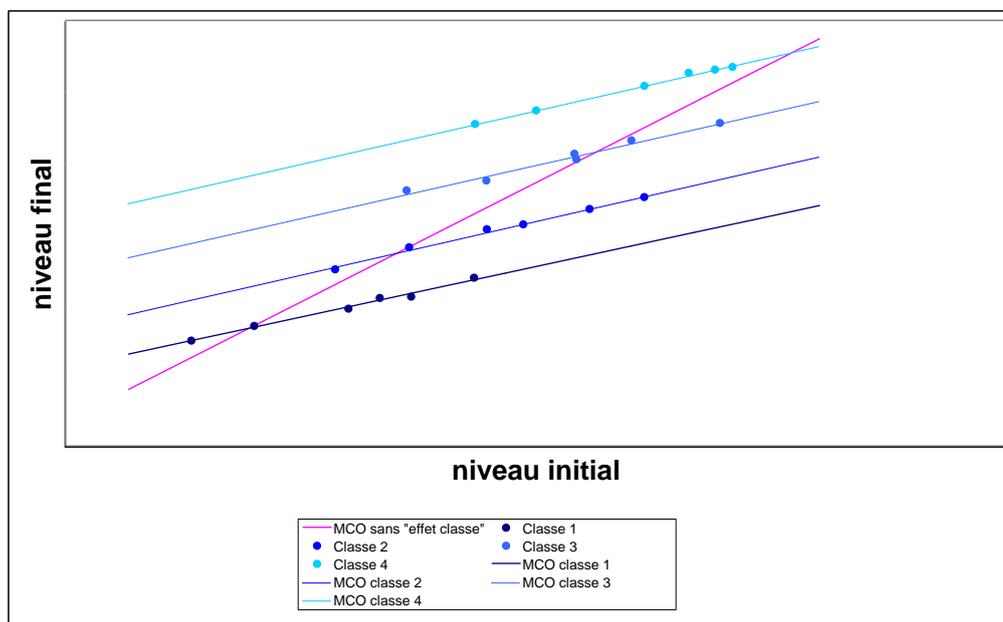


FIGURE 1.1 – Régressions classe par classe et régression sans effets "classe"

grandeur de l'écart-type du coefficient associé au secteur sera proportionnel à  $\frac{1}{\sqrt{4\ 000}}$ . Or on comprend assez intuitivement que dans la mesure où les élèves sont répartis dans seulement 200 classes qui relèvent soit du public soit du privé, cet ordre de grandeur devrait plutôt être proportionnel à  $\frac{1}{\sqrt{200}}$ . La surestimation de précision est nette car tous les élèves d'une même classe sont scolarisés dans le même secteur, mais elle existe également pour des caractéristiques individuelles (comme les caractéristiques parentales) dans la mesure où ces variables individuelles sont corrélées avec la classe fréquentée.

## 1.2 Organisation du document

Ce document présente donc les méthodes disponibles pour éviter ces deux conséquences du regroupement des données individuelles. En continuant de se référer à l'exemple des données scolaires, il existe deux manières de prendre en compte un effet spécifique à la classe : soit en considérant que la classe est une variable explicative au même titre que les autres covariables (on neutralise alors les effets des variables observées et inobservées partagés par les élèves d'une même classe), soit en considérant que la classe introduit de la corrélation entre les résidus de chaque élève de la même classe (on prend en compte l'existence de variables inobservées constantes au sein de la classe dont la distribution n'est pas corrélée avec les covariables).

Ces modèles prenant en compte l'appartenance à la classe sont des modèles proches du modèle usuel de régression linéaire. Dans un premier chapitre, nous rappelons donc les propriétés du modèle linéaire, puis dans un deuxième et un troisième chapitres, nous étudions comment peut être adapté ce modèle pour prendre en compte l'appartenance à la classe. Deux grands types de modèles seront alors présentés : les modèles à effets fixes et les modèles à effets aléatoires. La programmation sous SAS est rapidement présentée en fin de chaque chapitre. Nous étudions également les liens qui existent entre modèles à effets fixes et modèles à effets aléatoires et comment des tests de sélection de modèles peuvent être mis en œuvre dans un quatrième et un cinquième chapitres. Enfin dans

un sixième chapitre, nous traitons des difficultés spécifiques pour étendre ces modèles au cas de variables qualitatives.

Le lecteur familier de l'économétrie des panels retrouvera nombre de résultats connus, les dimensions individuelles et temporelles d'un panel étant transposées en dimension inter-classe et intra-classe. Ce document n'a donc pas la prétention de constituer une innovation mais plutôt de fournir à un empiriste travaillant sur des données du type "élèves-classes", un aperçu des problèmes et des solutions qui l'attendent.

Par ailleurs, ce document peut être lu de différentes manières : l'ensemble des démonstrations a été fourni par souci d'exhaustivité mais leur lecture n'est pas nécessaire pour appliquer les résultats énoncés. Afin de faciliter une telle lecture sélective, les démonstrations sont systématiquement encadrées par les symboles  $\square$  et  $\blacksquare$  en début et en fin de démonstration.

Pour conclure cette introduction, on indique que ce document ne traite pas le cas des modèles prenant en compte une hétérogénéité sur les rendements des variables explicatives. Cependant, les questions que l'empiriste doit se poser pour modéliser une telle hétérogénéité sont identiques à celles traitées dans les six chapitres de ce document. En particulier, l'arbitrage entre le risque d'endogénéité et la modélisation de l'hétérogénéité doit se faire suivant les mêmes critères que ceux discutés dans ce document. Ainsi, même si ce document n'a pas la prétention de traiter des modèles de régressions multiniveaux dans toute leur généralité, il peut être lu comme une introduction aux enjeux de modélisation dans les régressions multiniveaux.

# Notations

---

$n$	le nombre d'observations mobilisées pour les estimations
$i$	indice identifiant l'élève
$j$	indice identifiant la classe
$K$	le nombre de covariables utilisées (on ne compte pas la constante)
$C$	le nombre de classes différentes mobilisées pour les estimations
$n_j$	le nombre d'observation dans la classe $j$ , $n = \sum_{j=1}^C n_j$
$y_i$ ou $y_{ij}$	la valeur de la variable dépendante pour l'élève $i$ (dans la classe $j$ )
$\mathbf{y}$	le vecteur colonne de taille $n$ formé en "empilant" les $y_i$ , $i = 1, \dots, n$ (ou les
	$y_{i,j}, i = 1, \dots, n_j, j = 1, \dots, C) \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$
$x_{k,i}$ ou $x_{k,i,j}$	la valeur de la $k^{ieme}$ covariable pour l'individu $i$ (de la classe $j$ )
$\mathbf{x}_k$	le vecteur colonne de taille $n$ formé en "empilant" les $x_{k,i}$ , $i = 1, \dots, n$ (ou les
	$y_{i,j}, i = 1, \dots, n_j, j = 1, \dots, C) \quad \mathbf{x}_k = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,n} \end{bmatrix}$
$\mathbf{x}_0$	le vecteur colonne de taille $n$ dont toutes les composantes sont égales à 1
$X_i$ ou $X_{i,j}$	le vecteur ligne formé en "empilant" la constante et les $x_{k,i}$ , $k = 1, \dots, K$ (ou les $x_{k,i,j}$ , $k = 1, \dots, K$ )
$\mathbf{X}$	la matrice de taille $n \times (K + 1)$ obtenue en concaténant les vecteurs $\mathbf{x}_k$ , $k = 0, 1, \dots, K$ .
$\varepsilon_i$	hétérogénéité individuelle inobservée de l'individu $i$ , réel inconnu propre à l'individu $i$
$\epsilon$	hétérogénéité individuelle inobservée, vecteur inconnu obtenu en empilant les $\varepsilon_i$ , $i = 1, \dots, n$
$\ \cdot\ $	la norme euclidienne, pour un vecteur $\mathbf{u}$ de taille $n$ , $\ \mathbf{u}\  = \sqrt{\sum_{i=1}^n u_i^2}$
$U_n \xrightarrow[n \rightarrow +\infty]{p} U$	la suite de variables aléatoires $U_n$ tend en probabilité vers $U$ quand $n$ tend vers $+\infty$ c'est à dire que pour tout réel $a > 0$ , $P( U_n - U  > a)$ tend vers 0 quand $n$ tend vers $+\infty$
$U_n \xrightarrow[n \rightarrow +\infty]{loi} U$	la suite de variables aléatoires $U_n$ tend en loi vers $U$ quand $n$ tend vers $+\infty$ c'est-à-dire que la fonction de répartition de $U_n$ converge simplement vers la fonction de répartition de $U$ : pour tout réel $u$ tel que $P(U < u) = P(U \leq u)$ on a $P(U_n \leq u)$ qui tend vers $P(U \leq u)$ quand $n$ tend vers $+\infty$

Pour simplifier l'écriture, pour des propositions relatives à l'ensemble des individus, on utilisera souvent  $y$ ,  $X$ ,  $\epsilon$ ,  $x_k$  au lieu de  $y_i$ ,  $X_i$ ,  $\epsilon_i$ ,  $x_{ki}$ .

## Chapitre 2

# Rappels d'économétrie linéaire

### 2.1 Qu'est ce qu'un modèle linéaire ?

Dans tout ce document, nous parlerons de modèle économétrique pour désigner le fait qu'une variable  $y$  (par exemple la note scolaire d'un enfant) est engendrée par des variables observées  $X$  (le sexe, l'âge, l'origine social de cet enfant...) et de variables inobservées  $\varepsilon$  (par exemple, des événements biographiques de cet enfant ou -sous réserve que cela ait un sens- ses "capacités cognitives"...), soit  $y = f(X, \varepsilon)$ . Le recours à une telle abstraction permet au chercheur en sciences sociales d'interpréter des co-variations entre la variable  $y$  et une variable  $x$  particulière, les autres variables  $X$  et  $\varepsilon$  étant fixées. Estimer un tel modèle revient à déterminer la fonction  $f$ . Cela permet de décomposer des inégalités de manière descriptive, mais cela permet surtout de mesurer les rendements propres à chaque covariable. Dans un tel cadre, la covariation de  $y$  et d'une variable  $x$  "toutes choses (observées et inobservées) égales par ailleurs" définit l'impact causal de  $x$  sur  $y$ .

Cependant, l'empiriste ne peut raisonner que toutes choses observées égales par ailleurs. Pour pouvoir identifier et estimer  $f$  et donc les impacts causaux, il est obligé de faire des hypothèses sur les observées  $y, X$  et les inobservées  $\varepsilon$  :

- analytiquement, sur la relation entre  $y, X$  et  $\varepsilon$  : est-ce que  $f(X, \varepsilon)$  est de la forme  $g(X) + \varepsilon$  ? de la forme  $\varepsilon h(X)$  ? de la forme  $\phi(\psi(X) + \varepsilon)$  ? Etc.
- et statistiquement, sur la distribution jointe de  $X$  et  $\varepsilon$  : est-ce que les observées et les inobservées sont corrélées ? indépendantes ? Existe-t-il des covariables n'intervenant pas dans la fonction  $f$ , qui sont décorréliées des inobservées ? Etc.

La classe des modèles linéaires est l'ensemble des modèles tels que  $f(X, \varepsilon) = X\beta + \varepsilon$ . Les variables observées et inobservées se combinent de manière additive pour générer la variable  $y$ . Dans un tel cadre, déterminer  $f$  revient à déterminer le vecteur  $\beta$ .

### 2.2 Le cadre

On cherche par la suite à évaluer l'influence de différents facteurs sur une variable  $y$  continue. Le cas des variables dépendantes dichotomiques est traité dans le chapitre 6 de ce document. L'exemple type sur lequel nous appuierons cette présentation est celui où  $y$  est la note à un examen. Les facteurs pouvant influencer cette note sont multiples : il peut s'agir du sexe du candidat, de son passé scolaire, du capital culturel de ses parents, de leur capital économique, du type d'établissement fréquenté, etc. Ces facteurs étant corrélés les uns aux autres, il est nécessaire d'effectuer des analyses "toutes choses égales par ailleurs" pour évaluer l'impact relatif de chacun de ces facteurs. Ces facteurs peuvent être appréhendés au moyen de deux types de variables : des variables continues (par exemple le revenu des parents ou son logarithme, la taille de la classe fréquentée, ...), des variables indicatrices qui prennent des valeurs 0 ou 1 selon que l'événement considéré (être une fille, avoir un père cadre, avoir des parents gagnant entre 20 000 et 30 000 euros par an, etc.).

### Le modèle

On considère un vecteur de  $n$  observations  $\mathbf{y} = (y_1, y_2, \dots, y_n)'$  dont on cherche à expliquer la variabilité par un jeu de  $K$  covariables  $\mathbf{x}_k = (x_{k,1}, x_{k,2}, \dots, x_{k,n})'$  pour  $k = 1, \dots, K$ . On note  $\mathbf{X}$  la matrice des covariables de taille  $n \times (K + 1)$  obtenue en concaténant le vecteur constant de taille  $n$  et les covariables  $\mathbf{x}_k, k = 1, \dots, K$ . Le modèle linéaire consiste à se placer sous l'hypothèse suivante :

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Le paramètre  $\beta = (\beta_0, \beta_1, \dots, \beta_K)'$  est un vecteur de taille  $K + 1$  inconnu. La composante  $\beta_k, k = 1, \dots, K$  représente une "covariance nette" entre  $\mathbf{y}$  et  $\mathbf{x}_k$  qui rend compte de la variabilité conjointe de  $\mathbf{y}$  et de  $\mathbf{x}_k$  pour des valeurs de  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{k-1}, \mathbf{x}_{k+1}, \dots, \mathbf{x}_K$  données.

Le vecteur  $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$  est un vecteur inconnu : il représente l'effet des variables inobservées. Il permet de rendre compte de la part de variabilité des  $y$  qui n'est pas liée aux variables  $X$ , on le désigne classiquement par **hétérogénéité inobservée**. Une autre désignation classique de ce vecteur est celle de vecteur des résidus. Pour l'instant, on ne postule aucune hypothèse sur la loi de ce vecteur  $\epsilon$ . Une des questions centrales de ce chapitre est d'expliciter quelles sont les hypothèses que doit vérifier l'hétérogénéité inobservée pour obtenir de "bonnes" propriétés pour un estimateur de  $\beta$ . Il s'agit donc en creux de clarifier quelles seront les propriétés qui seront perdues par l'estimateur lorsque l'hétérogénéité  $\epsilon$  ne vérifiera plus certaines hypothèses usuelles.

Dans le cas où la variable  $\mathbf{x}_k$  est continue, le coefficient  $\beta_k$  s'interprète comme l'accroissement de  $y$  lorsque  $\mathbf{x}_k$  croît d'une unité (à  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{k-1}, \mathbf{x}_{k+1}, \dots, \mathbf{x}_K$  données). Dans le cas où la covariable est discrète, le coefficient  $\beta_k$  s'interprète comme l'accroissement de  $y$  associé au fait d'être dans la situation décrite par la covariable contre le fait d'être dans la situation de référence.

## 2.3 Définition, conditions d'existence et d'unicité de l'estimateur des Moindres Carrés Ordinaires (MCO).

L'estimateur étudié dans ce premier chapitre existe de manière univoque sous la condition essentielle d'absence de colinéarité entre les covariables. L'existence d'une telle colinéarité signifie intuitivement que certaines covariables sont "redondantes". Dans un tel cas, il n'est pas surprenant que se pose un problème de définition des estimateurs puisque le modèle n'est pas identifiable : il est impossible de séparer les effets de covariables "redondantes".

### L'hypothèse de non-colinéarité des covariables

$(H_1)$  : La matrice  $\mathbf{X}$  est de plein rang colonne (i.e. il n'existe pas de relation linéaire entre les variables  $\mathbf{x}_k$ , c'est-à-dire qu'il n'existe pas de vecteur  $\lambda$  non nul tel que  $\mathbf{X}\lambda = 0$ ).

**Remarque** : Si l'hypothèse  $(H_1)$  n'est pas vérifiée le paramètre  $\beta$  n'est pas identifiable : il existe  $\beta' = \beta + \lambda \neq \beta$  tel que  $\mathbf{X}\beta' = \mathbf{X}\beta$ . Dans un tel cas, les données n'apportent aucune information pour choisir la "vraie" valeur de  $\beta$ .

### Estimateur des MCO

Sous les conditions  $(H_1)$ , l'estimateur des moindres carrés ordinaires défini par :

$$\hat{\beta}_{MCO} = \text{Argmin} \|\mathbf{y} - \mathbf{X}\beta\|^2$$

existe et est unique. Son expression est alors :

$$\hat{\beta}_{MCO} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

**Remarque** : L'estimateur des moindres carrés ordinaires doit son nom au fait que  $\|\mathbf{y} - \mathbf{X}\beta\|^2 = \sum_{i=1}^n \epsilon_i^2$ .

Il correspond donc au  $\beta$  qui minimise la somme des carrés des résidus.

Preuve :

□

La condition nécessaire de minimisation du premier ordre (annulation du gradient) s'écrit :

$$-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = 0$$

La condition suffisante de minimisation du second ordre (hessienne définie positive) s'écrit :

$$\text{Pour tout vecteur } \lambda \text{ non nul de taille } n, \lambda' \mathbf{X}' \mathbf{X} \lambda > 0$$

La condition du second ordre est vérifiée grâce à l'hypothèse  $(H_1)$  : pour tout vecteur  $\lambda$  non nul de taille  $n$ ,  $\lambda' \mathbf{X}' \mathbf{X} \lambda = \|\mathbf{X}\lambda\|^2 > 0$

La condition du premier ordre s'écrit donc  $(\mathbf{X}'\mathbf{X})\beta = \mathbf{X}'\mathbf{y}$ . La matrice  $\mathbf{X}'\mathbf{X}$  est inversible car  $\mathbf{X}'\mathbf{X}\lambda = 0 \Rightarrow \lambda' \mathbf{X}' \mathbf{X} \lambda = 0 \Rightarrow \|\mathbf{X}\lambda\|^2 = 0 \Rightarrow \mathbf{X}\lambda = 0 \Rightarrow \lambda = 0$   
( $H_1$ )

On en déduit que  $\widehat{\beta}_{MCO} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ .

■

**Voici quelques cas (plus ou moins embêtants) où  $(H_1)$  peut être mise en défaut.**

– Le cas (peu embêtant) des variables qualitatives : si parmi les covariables  $\mathbf{x}_k$ , il y a des variables qualitatives, il est nécessaire d'"exclure la situation de référence". Par exemple, toujours dans le cas où on cherche à expliquer la variabilité d'une note à un examen, on peut penser que le sexe du candidat joue un rôle. On pourrait être tenté d'introduire dans la régression les covariables suivantes :  $\mathbb{1}_{\{\text{le candidat est une fille}\}}$  et  $\mathbb{1}_{\{\text{le candidat est un garçon}\}}$ . Mais cela pose un problème car :

$$\mathbb{1}_{\{\text{le candidat est une fille}\}} + \mathbb{1}_{\{\text{le candidat est un garçon}\}} = 1$$

Dans ce cas, on a :

$$\begin{aligned} \beta_0 + \beta_1 \mathbb{1}_{\{\text{le candidat est une fille}\}} + \beta_2 \mathbb{1}_{\{\text{le candidat est un garçon}\}} = \\ \beta'_0 + \beta'_1 \mathbb{1}_{\{\text{le candidat est une fille}\}} + \beta'_2 \mathbb{1}_{\{\text{le candidat est un garçon}\}} \end{aligned}$$

dès que :

$$\begin{cases} \beta_0 = \beta'_0 + c \\ \beta_1 = \beta'_1 - c \\ \beta_2 = \beta'_2 - c \end{cases}$$

Il faut donc éliminer du modèle une des deux variables (par exemple  $\mathbb{1}_{\{\text{le candidat est un garçon}\}}$ ). Le coefficient  $\beta_1$  associé à la variable  $\mathbb{1}_{\{\text{le candidat est une fille}\}}$  s'interprète alors comme le nombre de points en plus (ou en moins) sur la note à l'examen associé au fait d'être une fille par rapport à la situation de référence "être un garçon"<sup>1</sup>. Dans le cas où une variable qualitative est appréhendée au moyen de trois indicatrices (ou plus), par exemple  $\mathbf{x}_k = \mathbb{1}_{\{\text{famille française}\}}$ ,  $\mathbf{x}_l = \mathbb{1}_{\{\text{famille immigrée}\}}$ ,  $\mathbf{x}_m = \mathbb{1}_{\{\text{famille mixte}\}}$ , il faut de la même manière éliminer une des indicatrices de la régression (mettons  $\mathbf{x}_k = \mathbb{1}_{\{\text{famille française}\}}$ ) et le coefficient  $\beta_l$  (respectivement  $\beta_m$ ) s'interprète toujours comme l'écart de note à l'examen d'un candidat issu d'une famille immigrée (respectivement d'une famille mixte) par rapport à un candidat issu d'une famille française. La grandeur  $\beta_l - \beta_m$  s'interprète comme l'écart de note à l'examen d'un candidat issu d'une famille immigrée par rapport à un candidat issu d'une famille mixte. Un "bon" estimateur ne devrait donc pas dépendre de la modalité de référence choisie ; c'est le cas de l'estimateur des MCO. Cela ne veut pas dire pour autant que la situation de référence peut être choisie à la légère (notamment pour des questions de présentation des résultats). Choisir comme situation de référence une situation rare ou extrême conduit à estimer des effets par rapport à une situation exceptionnelle et donc nuit à l'interprétation. Choisir comme situation de référence une situation fréquente offre des estimations et des tests de significativité plus faciles à interpréter.

---

1. Éliminer de la régression une variable correspondant à la situation de référence revient à fixer la contrainte identifiante  $\beta_2 = 0$ . Dans certaines procédures de SAS utilisant l'instruction CLASS, la contrainte  $\sum \beta_1 \mathbb{1}_{\{\text{le candidat est une fille}\}} + \beta_2 \mathbb{1}_{\{\text{le candidat est un garçon}\}} = 0$  est utilisée, les coefficients  $\beta_1$  et  $\beta_2$  s'interprètent alors comme des écarts de la note  $y$  par rapport à la "situation moyenne". Cette "situation moyenne" n'étant pas d'une interprétation très naturelle, nous privilégions l'approche traditionnelle de l'exclusion d'une modalité de référence.

- Le cas de l'existence d'une relation "mécanique" entre les covariables : on peut légitimement postuler les hypothèses suivantes : la note à l'examen dépend de l'âge de l'individu au moment de l'épreuve (un élève "en retard" aura sûrement une moins bonne note qu'un élève en avance), de la génération à laquelle il appartient (un élève qui appartient à une génération qui a connu des mouvements sociaux importants arrive probablement moins bien préparé aux examens) et de la session concernée (les jurys peuvent être plus ou moins sévères d'une année sur l'autre, les sujets peuvent être plus ou moins difficiles). Ces trois effets qui peuvent contribuer à expliquer une part de la variance de la note obtenue à l'examen ne sont pas identifiables car on a la relation mécanique suivante : "session"="génération"+"âge". Comme précédemment, on est obligé d'éliminer du modèle une des trois variables, mais cette fois le problème est plus embêtant : il faut renoncer à identifier un des effets et l'estimation des deux autres effets est susceptible d'être biaisée, ou spécifier des contraintes identifiantes supplémentaires (par exemple sur la tendance des effets "session"). C'est un problème classique des modèles "âge-période-cohorte".
- Le cas d'une relation de linéarité liée à la taille finie de l'échantillon : plus généralement, même s'il n'existe pas une relation mécanique entre les covariables, il est possible que cela soit le cas au sein de l'échantillon dont dispose l'économètre. Si on dispose d'un échantillon d'enfants de différentes nationalités de naissance et que l'on s'intéresse à l'effet différencié de l'origine sociale sur la réussite scolaire, lorsque la taille de l'échantillon est modeste, il est possible que tous les enfants de telle nationalité de naissance soient également des enfants d'ouvriers. Pour tester si la position sociale a le même effet sur la réussite scolaire pour les différentes nationalités, l'économètre peut alors vouloir introduire des indicatrices croisant position sociale des parents et nationalité de naissance : ces indicatrices seront nulles pour certaines situations "rares". De même, si l'économètre veut tester l'influence de la nationalité de naissance sur la réussite scolaire au sein des enfants de cadres, sur un petit échantillon certaines nationalités de naissance ne seront pas représentées parmi les enfants de cadres. Il faut dans ce cas se tourner vers des données plus adaptées pour traiter ces questions.

## 2.4 Quand l'estimateur des MCO est-il sans biais ?

L'hypothèse d'exogénéité (ou de nullité de l'espérance conditionnelle de l'hétérogénéité inobservée)

$(H_2)$  : L'espérance des résidus conditionnellement aux covariables est nulle ce qui s'écrit  $E[\epsilon|\mathbf{X}] = 0$

**Remarque :** L'hypothèse  $(H_2)$  signifie intuitivement que pour tout individu  $i$ , lorsque ses  $x_{ki}$  changent, son  $\varepsilon_i$  n'a pas de raisons de changer (ni les  $\varepsilon_{i'}$  des autres individus). C'est une condition très forte, dont dépend la validité du modèle et la qualité des estimations. Cette hypothèse a deux conséquences immédiates : l'espérance des résidus est nulle (car  $E(\epsilon) = E(E[\epsilon|X]) = 0$ ) et les distributions des covariables et de des inobservées sont décorréelées (car  $Cov(\varepsilon_i, x_{ki}) = E(\varepsilon_i x_{ki}) = E(E[\varepsilon_i | x_{ki}]) = 0$ ). Une covariable  $x_k$  qui est corrélée avec les résidus  $\epsilon$  est dite endogène.

### Biais des MCO

Sous les conditions  $(H_1)$  et  $(H_2)$ , l'estimateur des MCO est sans biais, c'est-à-dire que :

$$E(\hat{\beta}_{MCO}) = \beta$$

Preuve :

□

$$\begin{aligned} E(\widehat{\beta}_{MCO}) &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] \\ &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon)\right] \\ &= \beta + E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\right] \\ &= \beta + E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\epsilon|\mathbf{X}]\right] \\ &= \beta \end{aligned}$$

■

L'hypothèse  $(H_2)$  permet même d'obtenir un résultat plus fort :  $E(\widehat{\beta}_{MCO}|\mathbf{X}) = \beta$ , ce qui signifie que les estimations sont sans biais même dans le cas où on travaille sur un échantillon tel que la distribution des  $X$  ne correspond à la distribution des  $X$  dans la population d'inférence. Une version plus faible de l'hypothèse  $(H_2)$  consiste à supposer seulement que  $cov(\epsilon, X) = 0$  sur la population d'inférence, dans ce cas les estimations sur un échantillon "non représentatif" (en terme de distribution de  $X$ ) peuvent être biaisées.

Lorsque l'hypothèse  $(H_2)$  semble peu crédible, un moyen usuel pour obtenir des estimations sans biais (asymptotiquement) consiste à recourir à des techniques de variables instrumentales. Pour une présentation détaillée des variables instrumentales, nous renvoyons le lecteur à d'autres documents (par exemple celui de Jean-Marc Robin : Modèles structurels et variables explicatives endogènes, Document de travail INSEE-UMS numéro 2002). Notons simplement qu'un ensemble de variables instrumentales  $\mathbf{W}$  permettra d'obtenir une estimation (asymptotiquement) sans biais de  $\beta$  si les variables  $\mathbf{W}$  sont corrélées aux variables  $\mathbf{X}$  et si  $E[\epsilon|\mathbf{W}] = 0$  (ou si  $cov(\epsilon, W) = 0$ ).

**Voici quelques cas où l'hypothèse  $(H_2)$  est mise en défaut :**

– Endogénéité liée aux "variables omises" et à l'"hétérogénéité inobservée"

Supposons que la réussite à un examen dépende du diplôme de la mère et des revenus de la famille, le modèle s'écrit :

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \epsilon \\ &= \beta_0 + \beta_1\mathbf{x}_1 + \sum_{k=2}^K \beta_k\mathbf{x}_k + \epsilon \end{aligned}$$

Où  $y$  est la note obtenue à l'examen,  $x_1$  est le revenu et où  $x_k, k = 2, \dots, K$  sont  $K - 1$  variables indicatrices des niveaux de diplôme de la mère. Supposons de plus que le revenu ne soit pas disponible dans les données exploitées (parce qu'une question sur les revenus n'a pas été incluse dans le questionnaire d'enquête par exemple). Pour estimer les  $\beta_k, k = 2, \dots, K$ , on peut être tenté de calculer l'estimateur des moindres carrés issus de la régression de  $y$  sur les  $K - 1$  niveaux de diplôme de la mère et donc poser :

$$\mathbf{y} = \beta_0 + \sum_{k=2}^K \beta_k\mathbf{x}_k + \mathbf{v}$$

avec  $E[\mathbf{v}|\mathbf{x}_k, k = 2, \dots, K] = 0$

Si cela était vrai, alors l'estimateur des MCO des  $\beta_k, k = 2, \dots, K$  serait sans biais et la covariance des résidus  $v$  et de  $x_k$  ( $k = 2, \dots, K$ ) serait nulle. Ce n'est en général pas le cas, car :

$$Cov(v, x_k) = Cov(\beta_1 x_1 + \epsilon, x_k) = \beta_1 Cov(x_1, x_k)$$

Le fait de ne pas intégrer la variable revenu dans la régression biaise les estimations concernant l'effet du diplôme de la mère à deux conditions : la première est que le revenu joue effectivement un rôle dans la réussite du candidat ( $\beta_1 \neq 0$ ), la seconde est que le diplôme de la mère et le revenu de la famille soient corrélés ( $Cov(x_1, x_k) \neq 0$ ). Dans ce cas précis, ces deux conditions ont toutes les chances d'être réunies. Intuitivement, en l'absence de contrôle du revenu, une partie de l'effet du revenu est capté par l'effet du niveau de diplôme de la mère.

Un autre exemple est donné par les résultats contradictoires des études sur la taille des classes (le lecteur pourra se référer au rapport du Haut conseil de l'évaluation de l'école

rédigé sur ce sujet par Denis Meuret en 2001, à l'étude de Thomas Piketty et Mathieu Valdenaire "L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français", les dossiers de la DEP (numéro 173) en 2006, et la réponse de Meuret dans l'Université Syndicaliste sous le titre "Sa méthode et la nôtre..." en 2004). Pour étudier l'effet de la taille des classes sur les acquis des élèves, les données disponibles permettent en général de régresser une mesure des compétences des élèves sur la taille de la classe (ou la taille de la classe fréquentée une année précédente) et sur des covariables susceptibles d'avoir un effet sur les acquis scolaires. Le problème est que les élèves ne sont pas répartis au hasard dans les classes. Il est fort possible que les élèves pour lesquels l'équipe pédagogique envisage le plus de difficultés scolaires à venir soient affectés dans les classes à faible effectif. Les élèves sont donc triés sur des critères difficilement observables qui ne peuvent être contrôlés dans la régression. Séparer l'effet de sélection (le tri des élèves sur des caractéristiques difficilement observables) et l'effet de la taille des classes demande donc une réflexion spécifique.

De même, si on cherche à étudier les "rendements" respectifs de différents établissements scolaires, il est probable que les élèves scolarisés dans le même établissement partagent un certain nombre de caractéristiques inobservées communes qui influent simultanément sur l'établissement fréquenté et les résultats scolaires. L'estimation du "rendement" de l'établissement est alors biaisé par l'effet des caractéristiques inobservées partagées par les élèves de l'établissement.

- Endogénéité liée à la "simultanéité"

L'hypothèse d'exogénéité peut également être violée du fait d'une simultanéité : imaginons que les résultats scolaires  $y_1$  dépendent de covariables  $X$  et de l'estime de soi  $y_2$  que peut avoir un élève, imaginons également que les résultats scolaires participent à la construction de l'estime de soi. Dans ce cas, on aura :

$$\mathbf{y}_1 = \mathbf{X}\beta_1 + \mathbf{y}_2\alpha_2 + \epsilon_1$$

$$\mathbf{y}_2 = \mathbf{X}\beta_2 + \mathbf{y}_1\alpha_1 + \epsilon_2$$

On suppose pour simplifier que  $E(\epsilon_1\epsilon_2) = cov(\epsilon_1, \epsilon_2) = 0$ . Supposons l'hypothèse d'exogénéité  $E(\epsilon_1|\mathbf{X}, \mathbf{y}_2) = 0$  valable. Dans ce cas  $cov(\epsilon_1, \mathbf{y}_2) = \alpha_1 cov(\epsilon_1, \mathbf{y}_1) + cov(\epsilon_1, \epsilon_2) = \alpha_1\alpha_2 cov(\epsilon_1, \mathbf{y}_2) + \alpha_1\sigma_1^2 + cov(\epsilon_1, \epsilon_2)$  donc  $cov(\epsilon_1, \mathbf{y}_2) = \frac{\alpha_1\sigma_1^2}{1-\alpha_1\alpha_2}$  qui est en général différent de 0 si  $\alpha_1 \neq 0$ , ce qui est en contradiction avec l'hypothèse d'exogénéité. Ainsi, dès lors que les résultats scolaires jouent sur l'estime de soi, cette dernière variable ne peut pas être considérée comme exogène pour expliquer les résultats scolaires. Dans ce cas, l'estimation MCO de la première équation est biaisée.

- Endogénéité liée aux erreurs de mesure

L'hypothèse d'exogénéité est également violée dans le cas où les variables exogènes sont entachées d'erreur de mesure. Supposons que le modèle  $\mathbf{y} = \mathbf{X}\beta + \epsilon$  vérifie les hypothèses  $(H_1)$  et  $(H_2)$ . Cependant l'analyste n'observe pas  $\mathbf{X}$  mais seulement une mesure imparfaite  $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{H}$ . Dans ce cas, on a :  $\mathbf{y} = \tilde{\mathbf{X}}\beta + \tilde{\epsilon}$  avec  $\tilde{\epsilon} = \epsilon - \mathbf{H}\beta$ . Si on suppose que l'erreur de mesure  $\mathbf{H}$  est "classique", c'est-à-dire non corrélée à  $X$  et  $\epsilon$  et d'espérance nulle, alors  $E(\tilde{\mathbf{X}}'\tilde{\epsilon}) = -\mathbf{H}'\mathbf{H}\beta \neq 0$  si  $\beta \neq 0$ . Ainsi, estimer  $\beta$  au moyen d'une régression MCO de  $\mathbf{y}$  sur  $\tilde{\mathbf{X}}$  conduit à des estimations biaisées.

## 2.5 Quelle est la précision de l'estimateur des MCO ?

### L'hypothèse d'homoscédasticité

$(H_3)$  : La variance des résidus conditionnellement aux covariables est telle que :

$$V[\epsilon|\mathbf{X}] = \sigma^2 I$$

**Remarque :** Sous l'hypothèse  $(H_2)$ , l'hypothèse  $(H_3)$  implique que  $V[\epsilon] = \sigma^2 I$  car  $V[\epsilon] = E[V[\epsilon|\mathbf{X}]] + V[E[\epsilon|\mathbf{X}]] = E[\sigma^2 I] + V[0]$ . L'hypothèse d'homoscédasticité est une hypothèse très forte qui signifie en particulier que :

- ◇  $cov(\epsilon_i, \epsilon_j) = 0$  : les résidus entre deux individus ne sont pas corrélés.

◊  $V(\varepsilon_i | X_i)$  ne dépend pas des covariables  $x_k, k = 1, \dots, K$ . La connaissance des covariables pour un individu n'apporte aucune information sur la variance de son résidu.

◊  $V(\varepsilon_i) = V(\varepsilon_j)$  ce qui signifie que la variance des résidus ne dépend pas de l'individu.

$(H_3)$  peut ne pas être vérifiée, soit parce que  $V[\varepsilon | \mathbf{X}]$  n'est pas diagonale, soit parce que  $V[\varepsilon | \mathbf{X}]$  dépend de  $\mathbf{X}$ , soit parce que  $V[\varepsilon | \mathbf{X}]$  est diagonale mais que les éléments diagonaux de la matrice ne sont pas égaux, soit pour une de ces deux raisons. Dans le cas où  $V[\varepsilon | \mathbf{X}]$  est diagonale mais que les éléments diagonaux de la matrice ne sont pas égaux, on parlera d'hétéroscédasticité. Dans le cas où la matrice n'est pas diagonale, on parlera de corrélation des résidus. Il existe de nombreux autres cas où l'hypothèse d'hétéroscédasticité est relâchée, c'est le cas des modèles "multiniveaux" ou à coefficients aléatoires présentés au chapitre 4.

### Variance de l'estimateur des MCO

Sous les hypothèses  $(H_1)$ ,  $(H_2)$  et  $(H_3)$ , la variance de l'estimateur des moindres carrés ordinaires conditionnellement aux covariables est :

$$V(\widehat{\beta}_{MCO} | \mathbf{X}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

La variance de l'estimateur des moindres carrés ordinaires est donnée par :

$$V(\widehat{\beta}_{MCO}) = \sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}]$$

Preuve :

□

On rappelle que :

$$E(\widehat{\beta}_{MCO} | \mathbf{X}) = \beta$$

Donc :

$$\begin{aligned} V[\widehat{\beta}_{MCO} | \mathbf{X}] &= E\left[(\widehat{\beta}_{MCO} - \beta)(\widehat{\beta}_{MCO} - \beta)' | \mathbf{X}\right] \\ &= E\left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \varepsilon \varepsilon' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X}\right] \\ &= E\left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \sigma^2 I \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X}\right] \\ &= \sigma^2 E\left[(\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X}\right] \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Et :

$$\begin{aligned} V(\widehat{\beta}_{MCO}) &= E\left[V[\widehat{\beta}_{MCO} | \mathbf{X}]\right] + V\left[E[\widehat{\beta}_{MCO} | \mathbf{X}]\right] \\ &= \sigma^2 E\left[(\mathbf{X}'\mathbf{X})^{-1}\right] + V(\beta) \\ &= \sigma^2 E\left[(\mathbf{X}'\mathbf{X})^{-1}\right] \end{aligned}$$

■

À ce stade, la précision de l'estimation dépend donc de deux grandeurs :  $\sigma^2$  qui représente la variance non liée aux covariables et  $E[(\mathbf{X}'\mathbf{X})^{-1}]$  qui dépend de la manière dont les covariables sont liées. Plus les covariables seront liées entre elles, plus la précision des estimations des coefficients correspondants sera faible. Il peut sembler surprenant que la précision des estimations ne dépende pas explicitement du nombre d'observations  $n$ . En réalité, la précision dépend bien du nombre d'observations au travers de la matrice  $\mathbf{X}'\mathbf{X} = [\sum_{i=1}^n x_{k,i} x_{j,i}]_{0 \leq k, j \leq K}$ . Comme nous allons le voir, l'écart-type de l'estimateur décroît en fait à la vitesse  $\frac{1}{\sqrt{n}}$ .

**Cas où l'hypothèse d'homoscédasticité n'est pas vérifiée.** Dans ce cas, on a :

$$V(\widehat{\beta}_{MCO}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'E[\epsilon\epsilon'|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

La précision en général dépend de toutes les covariances  $cov(\varepsilon_i, \varepsilon_{i'}|\mathbf{X})$ , soit  $\frac{n(n+1)}{2}$  termes. Des restrictions sur la structure de variance covariance des résidus sont donc en général postulées pour identifier et estimer  $\mathbf{X}'E[\epsilon\epsilon'|\mathbf{X}]\mathbf{X}$ . On suppose ainsi souvent que certaines covariances sont nulles. Par exemple : pour une estimation robuste à l'hétéroscédasticité (cf. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity" publié par White en 1980 dans la revue *Econometrica*), on suppose que les covariances des résidus entre deux observations sont nulles mais que les variances peuvent changer d'une observation à l'autre ; dans un panel, même si des inobservées peuvent affecter un même individu à différentes dates et/ou tous les individus à une certaine date en cas de chocs conjoncturels, les inobservées affectant un individu à une certaine date sont supposées décorréélées des inobservées affectant un autre individu à un autre moment. Une autre approche consiste à paramétrer la structure de covariance de résidus (par exemple : sur des observations répétées dans le temps sur un pays, on peut supposer que  $cov(\varepsilon_t, \varepsilon_{t'}|\mathbf{X}) = \rho(\mathbf{X})^{|t-t'|}$ ).

## 2.6 Comment estimer la précision de l'estimateur des MCO ?

On a vu que la variance de l'estimateur des moindres carrés dépend du paramètre inconnu  $\sigma^2$ . Il est donc nécessaire de trouver un estimateur de ce paramètre pour pouvoir estimer la variance de  $\widehat{\beta}_{MCO}$ . Il est possible d'obtenir un estimateur sans biais de  $\sigma^2$  sans faire d'hypothèses supplémentaires (à  $(H_1)$ ,  $(H_2)$  et  $(H_3)$ ). Par ailleurs, le paramètre  $\sigma^2$  représente la quantité de variance qui n'est pas liée aux covariables, c'est donc un paramètre indispensable à estimer dans une analyse de la variance.

### Estimation de la dispersion des inobservées

*Sous les hypothèses  $(H_1)$ ,  $(H_2)$  et  $(H_3)$ , l'estimateur*

$$\widehat{\sigma^2} = \frac{\widehat{\epsilon}'\widehat{\epsilon}}{n - K - 1}$$

(où  $\widehat{\epsilon} = \mathbf{y} - \mathbf{X}\widehat{\beta}$ ) est un estimateur sans biais du paramètre  $\sigma^2$ .

Preuve :

□

On note  $Tr(\cdot)$  l'opérateur trace qui à toute matrice associe la somme de ses éléments diagonaux :  $Tr(A) = \sum_l a_{l,l}$ .

On rappelle que pour toutes matrices  $A$  de taille  $l \times m$  et  $B$  de taille  $m \times l$  on a :  $Tr(AB) = Tr(BA)$ .

En notant  $\mathbf{M}$  la matrice  $\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ , on a :

$$\widehat{\epsilon} = \mathbf{y} - \mathbf{X}\widehat{\beta} = \mathbf{M}\mathbf{y} = \mathbf{M}(\mathbf{X}\beta + \epsilon) = \mathbf{M}\epsilon$$

La matrice  $\mathbf{M}$  vérifie  $\mathbf{M}^2 = \mathbf{M}$  et  $\mathbf{M}' = \mathbf{M}$  (cela caractérise les matrices de projection orthogonale,  $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  est la matrice de projection orthogonale sur l'espace engendré par  $\mathbf{X}$ ,  $\mathbf{M}$  est la matrice de projection orthogonale sur l'orthogonal à cet espace).

Donc :

$$\begin{aligned}
E(\widehat{\epsilon}'\widehat{\epsilon}|\mathbf{X}) &= E(\text{Tr}(\epsilon'\mathbf{M}\epsilon)|\mathbf{X}) \\
&= E(\text{Tr}(\mathbf{M}\epsilon\epsilon')|\mathbf{X}) \\
&= \text{Tr}(E(\mathbf{M}\epsilon\epsilon')|\mathbf{X}) \\
&= \text{Tr}(\mathbf{M}E(\epsilon\epsilon')|\mathbf{X}) \\
&= \text{Tr}(\mathbf{M}(\sigma^2\mathbf{I}_n)) \\
&= \sigma^2\text{Tr}(\mathbf{M}) \\
&= \sigma^2\text{Tr}(\mathbf{I}_n - \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}) \\
&= \sigma^2\left[\text{Tr}(\mathbf{I}_n) - \text{Tr}(\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})\right] \\
&= \sigma^2\left[n - \text{Tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X})\right] \\
&= \sigma^2[n - \text{Tr}(\mathbf{I}_{K+1})] \\
&= \sigma^2(n - K - 1)
\end{aligned}$$

■

### Estimation de la variance de l'estimateur des MCO

Sous les hypothèses  $(H_1)$ ,  $(H_2)$  et  $(H_3)$ ,

$$\widehat{V}(\widehat{\beta}_{MCO}) = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

est un estimateur sans biais de  $V(\widehat{\beta}_{MCO})$ .

Preuve :

□

On a bien  $E(\widehat{\sigma}^2) = \sigma^2$ , de plus on a  $E(\widehat{\sigma}^2|\mathbf{X}) = \sigma^2$ , ce qui permet d'écrire :

$$\begin{aligned}
E\left[\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\right] &= E\left[E\left[\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}\right]\right] \\
&= E\left[E\left[\widehat{\sigma}^2|\mathbf{X}\right](\mathbf{X}'\mathbf{X})^{-1}\right] \\
&= E\left[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right] \\
&= \sigma^2 E\left[(\mathbf{X}'\mathbf{X})^{-1}\right] \\
&= V(\widehat{\beta}_{MCO}) \\
&= V(\widehat{\beta}_{MCO})
\end{aligned}$$

■

## 2.7 L'estimateur des moindres carrés ordinaires est-il convergent ?

### Convergence en probabilité de l'estimateur des MCO

Sous les hypothèses  $(H_1)$ ,  $(H_2)$  et  $(H_3)$ , si :

◇  $\frac{1}{n}\mathbf{X}'\mathbf{X}$  tend en probabilité vers une matrice  $\mathbf{A}$  inversible quand  $n$  tend vers  $+\infty$

◇  $\frac{1}{n}\mathbf{X}'\epsilon$  tend en probabilité vers 0 quand  $n$  tend vers  $+\infty$

alors  $\widehat{\beta}_{MCO} \rightarrow \beta$  en probabilité

$\frac{1}{n}\mathbf{X}'\mathbf{X}$  tend en probabilité vers  $E[(x_{0,i}, x_{1,i}, \dots, x_{K,i})'(x_{0,i}, x_{1,i}, \dots, x_{K,i})]$  si on peut appliquer une loi des grands nombres à  $x_k x_l, 0 \leq k, l \leq K$  (on suppose alors que les  $x_{k,i} x_{l,i}$  suivent la même loi pour tout individu  $i$ ). De même,  $\frac{1}{n}\mathbf{X}'\epsilon$  tend en probabilité vers 0 si on peut appliquer une loi des grands nombres à  $(x_{0,i}, x_{1,i}, \dots, x_{K,i})'\epsilon_i$  car  $E[(x_{0,i}, x_{1,i}, \dots, x_{K,i})'\epsilon_i] = 0$  d'après  $(H_2)$ .

Les conditions nécessaires pour pouvoir appliquer ces lois des grands nombres sont de deux types : des conditions relatives à la dépendance des covariables entre deux individus  $i$  et des conditions sur les moments de ces covariables. Ces conditions sont sensiblement les mêmes que celles qui permettent d'appliquer des théorèmes de limite centrale qui permettront de caractériser la loi asymptotique de l'estimateur des moindres carrés.

Preuve :

□

$$\widehat{\beta}_{MCO} - \beta = \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \left( \frac{1}{n} \mathbf{X}' \epsilon \right)$$

D'après le lemme de Slutsky :  $\left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \rightarrow \mathbf{A}^{-1}$  en probabilité.

Et comme  $\left( \frac{1}{n} \mathbf{X}' \epsilon \right) \rightarrow 0$  en probabilité,  $\left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \left( \frac{1}{n} \mathbf{X}' \epsilon \right) \rightarrow 0$  en probabilité.

■

## 2.8 Quelle est la loi asymptotique de l'estimateur ?

On va donner ici un premier jeu d'hypothèses permettant de dériver la loi asymptotique de l'estimateur des moindres carrés ordinaires. On privilégie ici le fait de ne pas faire d'hypothèse sur la forme de la loi des résidus (on ne supposera pas que les résidus sont normaux), ce qui se paye par la nécessité de faire des hypothèses (relativement faibles cependant) sur la loi des covariables. Les résultats présentés ici ne sont donc exploitables qu'en présence d'un nombre d'observations suffisant (plusieurs centaines).

### Hypothèse sur la structure de dépendance des observations

$(H_a)$  : Les observations  $(y_i, X_i)_{i=1, \dots, n}$  sont indépendantes et identiquement distribuées.  $E[X_1' X_1]$  existe et est inversible.  $E(|x_{k,1} x_{l,1}|) < \infty$  pour tout  $0 \leq k, l \leq K$ .

### Normalité asymptotique de l'estimateur

Sous les hypothèses  $(H_a)$ ,  $(H_1)$ ,  $(H_2)$  et  $(H_3)$ , on a les propriétés suivantes :

◊ L'estimateur  $\widehat{\beta}_{MCO}$  est convergent :

$$\widehat{\beta}_{MCO} \xrightarrow[n \rightarrow +\infty]{p} \beta$$

◊ L'estimateur des moindres carrés ordinaires est asymptotiquement normal :

$$\sqrt{n} \left( \widehat{\beta}_{MCO} - \beta \right) \xrightarrow[n \rightarrow +\infty]{loi} N \left( 0, \sigma^2 E[X_1' X_1]^{-1} \right)$$

◊ L'estimateur  $\widehat{\sigma}^2$  est convergent :

$$\widehat{\sigma}^2 \xrightarrow[n \rightarrow +\infty]{p} \sigma^2$$

◊ On peut donc construire un estimateur convergent de la variance asymptotique de  $\sqrt{n} \left( \widehat{\beta}_{MCO} - \beta \right)$  :

$$n \widehat{\sigma}^2 \left( \mathbf{X}' \mathbf{X} \right)^{-1} \xrightarrow[n \rightarrow +\infty]{p} \sigma^2 E[X_1' X_1]^{-1}$$

◊ Et enfin :

$$\frac{1}{\sqrt{\widehat{\sigma}^2}} \left( \mathbf{X}' \mathbf{X} \right)^{\frac{1}{2}} \left( \widehat{\beta}_{MCO} - \beta \right) \xrightarrow[n \rightarrow +\infty]{loi} N(0, \mathbf{I}_{K+1})$$

Preuve :

□

On a toujours :

$$\widehat{\beta}_{MCO} - \beta = \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{X}' \epsilon$$

- $E(|x_{k,1}x_{l,1}|) < \infty$  pour tout  $0 \leq k, l \leq K$ , d'après la loi (faible) des grands nombres :  $\frac{1}{n} (\mathbf{X}'\mathbf{X}) = \frac{1}{n} \sum_i X'_i X_i \xrightarrow[n \rightarrow +\infty]{p} E[X'_1 X_1]$  inversible.
- Par ailleurs,  $V(X'_1 \varepsilon_1) = E(X'_1 E[\varepsilon_1^2 | X_1] X_1) = \sigma^2 A < +\infty$  donc d'après la loi forte des grands nombres  $\frac{1}{n} \mathbf{X}'\varepsilon = \frac{1}{n} \sum_i X'_i \varepsilon_i \xrightarrow[n \rightarrow +\infty]{p} E[X'_1 \varepsilon_1] = E[X'_1 E(\varepsilon_1 | X_1)] = 0$ . On en déduit donc la convergence en probabilité de  $\widehat{\beta}_{MCO}$  vers  $\beta$ .
- De plus le théorème de la limite centrale nous assure que :

$$\frac{1}{\sqrt{n}} \mathbf{X}'\varepsilon \xrightarrow[n \rightarrow +\infty]{loi} N(0, \sigma^2 A)$$

On a donc la normalité asymptotique de  $\widehat{\beta}_{MCO}$  :

$$\sqrt{n} (\widehat{\beta}_{MCO} - \beta) = \left( \frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{X}'\varepsilon \xrightarrow[n \rightarrow +\infty]{loi} N(0, \sigma^2 A^{-1} A A^{-1}) = N(0, \sigma^2 A^{-1})$$

- L'estimateur  $\widehat{\sigma}^2$  s'écrit :

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{n}{n-(K+1)} \left[ \frac{1}{n} (\mathbf{X}(\widehat{\beta}_{MCO} - \beta) + \varepsilon)' (\mathbf{X}(\widehat{\beta}_{MCO} - \beta) + \varepsilon) \right] \\ &= \frac{n}{n-(K+1)} \left[ (\widehat{\beta}_{MCO} - \beta)' \left( \frac{1}{n} \mathbf{X}'\mathbf{X} \right) (\widehat{\beta}_{MCO} - \beta) + \left( \frac{1}{n} \mathbf{X}'\varepsilon \right)' (\widehat{\beta}_{MCO} - \beta) + (\widehat{\beta}_{MCO} - \beta)' \left( \frac{1}{n} \mathbf{X}'\varepsilon \right) + \left( \frac{1}{n} \varepsilon'\varepsilon \right) \right] \end{aligned}$$

On peut appliquer la loi faible des grands nombres au dernier terme car  $\frac{1}{n} \varepsilon'\varepsilon = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2$ , les  $\varepsilon_i^2$  sont i.i.d. et  $E[|\varepsilon_1^2|] = V[\varepsilon_1] = \sigma^2 < +\infty$ .

On a donc :

$$\begin{aligned} \frac{1}{n} \varepsilon'\varepsilon &\xrightarrow[n \rightarrow +\infty]{p} \sigma^2 \\ \widehat{\beta}_{MCO} - \beta &\xrightarrow[n \rightarrow +\infty]{p} 0 \\ \frac{1}{n} \mathbf{X}'\varepsilon &\xrightarrow[n \rightarrow +\infty]{p} 0 \\ \frac{1}{n} \mathbf{X}'\mathbf{X} &\xrightarrow[n \rightarrow +\infty]{p} E[X'_1 X_1] \\ \frac{n}{n-(K+1)} &\xrightarrow[n \rightarrow +\infty]{} 1 \end{aligned}$$

Ce qui implique que :

$$\begin{aligned} \widehat{\sigma}^2 &\xrightarrow[n \rightarrow +\infty]{p} \sigma^2 \\ \frac{1}{\sqrt{\widehat{\sigma}^2}} \left( \frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{1/2} &\xrightarrow[n \rightarrow +\infty]{p} \frac{1}{\sqrt{\sigma^2}} A^{1/2} \end{aligned}$$

Par conséquent :

$$\frac{1}{\sqrt{\widehat{\sigma}^2}} (\mathbf{X}'\mathbf{X})^{\frac{1}{2}} (\widehat{\beta}_{MCO} - \beta) \xrightarrow[n \rightarrow +\infty]{loi} N(0, \mathbf{I}_{K+1})$$

■

Le fait que la loi asymptotique de  $\sqrt{\frac{1}{\widehat{\sigma}^2}} (\mathbf{X}'\mathbf{X})^{\frac{1}{2}} (\widehat{\beta}_{MCO} - \beta)$  ne dépende pas des paramètres estimés permet de construire des tests asymptotiques de l'hypothèse  $\beta = b$ . En effet, pour tout  $b \neq \beta$ ,  $\sqrt{\frac{1}{\widehat{\sigma}^2}} (\mathbf{X}'\mathbf{X})^{\frac{1}{2}} (\widehat{\beta}_{MCO} - b) \xrightarrow[n \rightarrow +\infty]{p.s.} +\infty$ . Dans le cas où l'hypothèse est fautive, la statistique calculée est donc "très grande". Dans le cas où l'hypothèse est vraie, la statistique calculée a peu de chance d'être grande car elle suit une loi normale. On peut de la même manière construire des tests sur les composantes de  $\beta$  ou sur une combinaison linéaire de ces composantes du type  $L\beta = Lb$  (où  $L$  est une matrice de taille  $(l \times (K+1))$  de rang  $l$ ). En effet, en appliquant la méthode "Delta" on a :

$$\frac{1}{\sqrt{\widehat{\sigma}^2}} (L'(\mathbf{X}'\mathbf{X})^{-1}L)^{-\frac{1}{2}} (L'\widehat{\beta}_{MCO} - L'b) \xrightarrow[n \rightarrow +\infty]{loi} N(0, \mathbf{I}_l)$$

Dans le cas où  $l \geq 2$  (ce qui revient à tester conjointement plusieurs hypothèses), il est alors nécessaire de raisonner sur le comportement asymptotique suivant :

$$\frac{1}{\sigma^2} \left( L' \widehat{\beta}_{MCO} - L' \beta \right)' (L' (\mathbf{X}' \mathbf{X})^{-1} L)^{-1} \left( L' \widehat{\beta}_{MCO} - L' \beta \right) \xrightarrow[n \rightarrow +\infty]{\text{loi}} \chi_l^2$$

Pour tester la significativité du coefficient  $\beta_k$ , on choisira donc  $L = [0, 0, \dots, 0, 1, 0, \dots, 0]$ , le vecteur ligne nul sauf sur la  $k$ ème composante. On pourra soit comparer  $\frac{\widehat{\beta}_k}{\sqrt{\widehat{\sigma}^2 [(\mathbf{X}' \mathbf{X})^{-1}]_{(k,k)}}}$  avec la distribution de la loi normale centrée réduite (pour un test de niveau 5%, cette statistique doit donc être comprise entre -1.96 et 1.96 pour accepter l'hypothèse  $\beta_k = 0$ ), soit de manière équivalente comparer  $\frac{\widehat{\beta}_k^2}{\widehat{\sigma}^2 [(\mathbf{X}' \mathbf{X})^{-1}]_{(k,k)}}$  avec la distribution de la loi du  $\chi_1^2$  ( pour un test de niveau 5%, cette statistique doit donc être inférieure à  $3.84 = (1.96)^2$  pour accepter l'hypothèse  $\beta_k = 0$ ).

## 2.9 Quelle est la loi à distance finie de l'estimateur si les résidus sont normaux ?

Les propriétés développées dans le paragraphe précédent sont valides dans un cadre asymptotique, c'est-à-dire pour un nombre d'observations suffisamment grand. Dans le cas où ces observations sont peu nombreuses, il est nécessaire de renforcer les hypothèses du modèle pour pouvoir caractériser la loi de l'estimateur et construire des tests à distance finie. Les hypothèses supplémentaires portent sur la loi des résidus. La philosophie du modèle change également, on va raisonner conditionnellement aux covariables, c'est-à-dire qu'on ne fait plus aucune hypothèse sur la loi des observations  $\mathbf{X}$  (l'hypothèse que les  $\mathbf{x}_i$  sont iid n'est plus nécessaire), la manière dont elles nous sont "données" n'est plus importante car on n'a pas à construire un raisonnement "à la limite" quand la taille de l'échantillon augmente.

### L'hypothèse de normalité des résidus

$(H_n)$  : Les résidus sont gaussiens, de moyenne nulle, identiquement distribués et indépendants des  $\mathbf{X}$ .

Cette hypothèse est très forte. Jusqu'à présent, seules des hypothèses sur les premiers moments (espérance, variance) avaient été nécessaires. Pour raisonner à distance finie, on réduit fortement l'ensemble de lois auquel peut appartenir la loi des résidus. Il existe un "très grand nombre" de lois qui vérifient les conditions  $(H_2)$  et  $(H_3)$  sans pour autant vérifier  $(H_n)$ . Par contre, l'hypothèse  $(H_n)$  implique les hypothèses  $(H_2)$  et  $(H_3)$  car sous  $(H_n)$ ,  $E[\epsilon|\mathbf{X}] = E(\epsilon) = 0$  et  $V[\epsilon|\mathbf{X}] = E(\epsilon) = \sigma^2 \mathbf{I}_n$ .

### Propriétés à distance finie de l'estimateur sous l'hypothèse de normalité des résidus

Sous les hypothèses  $(H_1)$  et  $(H_n)$ , on peut caractériser la loi des estimateurs conditionnellement aux covariables :

- ◇  $\widehat{\beta}_{MCO} | \mathbf{X} \rightsquigarrow N \left( \beta, \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} \right)$ . Conditionnellement aux covariables, l'estimateur des moindres carrés ordinaires est normalement distribué, de moyenne  $\beta$  et de variance  $\sigma^2 (\mathbf{X}' \mathbf{X})^{-1}$ .
- ◇  $\frac{n-(K+1)}{\widehat{\sigma}^2} \widehat{\sigma}_{MCO}^2 \rightsquigarrow \chi_{n-(K+1)}^2$ . À une constante de normalisation (inconnue) près, l'estimateur  $\widehat{\sigma}^2$  est distribué comme une loi du  $\chi_{n-(K+1)}^2$ .
- ◇  $\widehat{\beta}_{MCO} \perp \widehat{\sigma}_{MCO}^2$ . Les lois des estimateurs  $\widehat{\beta}_{MCO}$  et  $\widehat{\sigma}_{MCO}^2$  sont indépendantes.
- ◇  $\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{\sigma}^2 [(\mathbf{X}' \mathbf{X})^{-1}]_{(k,k)}}} \rightsquigarrow \text{Student}(n - (K + 1))$ . La loi de l'estimateur de la  $k$ ème composante

du vecteur  $\beta$  centré sur la valeur du paramètre et normalisé par l'estimation de son écart-type suit une loi de Student à  $(n-(K+1))$  degrés de liberté.

◊ L'estimateur des moindres carrés  $(\widehat{\beta}_{MCO}, \widehat{\sigma^2}_{MCO})$  est optimal parmi les estimateurs sans biais de  $(\beta, \sigma^2)$ , c'est-à-dire que pour tout estimateur  $(\widetilde{\beta}, \widetilde{\sigma^2})$  tel que pour tout  $(\beta, \sigma^2) \in \mathbb{R} \times \mathbb{R}^{+*}$ ,  $E[(\widetilde{\beta}, \widetilde{\sigma^2})] = (\beta, \sigma^2)$ , on a :

$$V \begin{bmatrix} \widehat{\beta}_{MCO} \\ \widehat{\sigma^2}_{MCO} \end{bmatrix} \ll V \begin{bmatrix} \widetilde{\beta} \\ \widetilde{\sigma^2} \end{bmatrix}$$

Preuve :

□

$$\begin{pmatrix} \widehat{\beta}_{MCO} - \beta \\ \widehat{\sigma^2}_{MCO} - \sigma^2 \end{pmatrix} = \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \\ \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \end{pmatrix} \epsilon$$

Comme  $\epsilon$  est un vecteur gaussien indépendant des covariables  $\mathbf{X}$ , la loi conditionnelle à  $\mathbf{X}$  de toute combinaison linéaire des  $\epsilon_i$  est un vecteur gaussien. On sait déjà que :

$$E \begin{pmatrix} \widehat{\beta}_{MCO} - \beta \\ \widehat{\sigma^2}_{MCO} - \sigma^2 \end{pmatrix} | \mathbf{X} = 0, V(\widehat{\beta}_{MCO} | \mathbf{X}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}, V(\widehat{\sigma^2}_{MCO} | \mathbf{X}) = \sigma^2 (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}')$$

De plus,

$$\begin{aligned} E \left[ \begin{pmatrix} \widehat{\beta}_{MCO} - \beta \\ \widehat{\sigma^2}_{MCO} - \sigma^2 \end{pmatrix} \epsilon' | \mathbf{X} \right] &= E \left[ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \epsilon' \epsilon (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') | \mathbf{X} \right] \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \sigma^2 \mathbf{I} (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' - \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \\ &= 0 \end{aligned}$$

Donc :

$$\begin{pmatrix} \widehat{\beta}_{MCO} \\ \widehat{\sigma^2}_{MCO} \end{pmatrix} | \mathbf{X} \rightsquigarrow N \left( \begin{pmatrix} \beta \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} & 0 \\ 0 & \sigma^2 (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \end{pmatrix} \right)$$

À partir de ce résultat, on déduit :

- La loi de  $\widehat{\beta}_{MCO} | \mathbf{X}$  :

$$\text{On a } \widehat{\beta}_{MCO} | \mathbf{X} \rightsquigarrow N(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$$

- La loi de  $\frac{n-(K+1)}{\sigma^2} \widehat{\sigma^2}$

Par hypothèse, on a  $\frac{\epsilon}{\sigma} \rightsquigarrow N(0, \mathbf{I}_n)$ , comme  $(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}')$  est un projecteur orthogonal de rang  $n - (K + 1)$ ,  $\frac{n-(K+1)}{\sigma^2} \widehat{\sigma^2} = \frac{1}{\sigma^2} \widehat{\epsilon}' \widehat{\epsilon} = \frac{1}{\sigma^2} \epsilon' (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \epsilon \rightsquigarrow \chi_{n-(K+1)}^2$

- L'indépendance de  $\widehat{\beta}_{MCO}$  et de  $\frac{n-(K+1)}{\sigma^2} \widehat{\sigma^2}$

$\begin{pmatrix} \widehat{\beta}_{MCO} \\ \widehat{\sigma^2}_{MCO} \end{pmatrix} | \mathbf{X}$  est un vecteur gaussien et  $Cov(\widehat{\beta}_{MCO}, \widehat{\sigma^2}_{MCO} | \mathbf{X}) = 0$ , on a donc  $\widehat{\beta}_{MCO} \perp \widehat{\sigma^2}_{MCO} | \mathbf{X}$ .

Comme de plus,  $\widehat{\sigma^2}$  est une fonction de  $\widehat{\epsilon}$  on a  $\widehat{\beta}_{MCO} \perp \widehat{\sigma^2} | \mathbf{X}$ .

Remarquons aussi que la loi de  $\widehat{\sigma^2}$  est indépendante de  $\mathbf{X}$ .

On a donc pour toutes fonctions  $f$  et  $g$  :

$$\begin{aligned} E(f(\widehat{\beta}_{MCO})g(\widehat{\sigma^2})) &= E \left( E \left[ f(\widehat{\beta}_{MCO})g(\widehat{\sigma^2}) | \mathbf{X} \right] \right) \\ &= E \left( E \left[ f(\widehat{\beta}_{MCO}) | \mathbf{X} \right] E \left[ g(\widehat{\sigma^2}) | \mathbf{X} \right] \right) \\ &= E \left( E \left[ f(\widehat{\beta}_{MCO}) | \mathbf{X} \right] E \left[ g(\widehat{\sigma^2}) \right] \right) \\ &= E \left( E \left[ f(\widehat{\beta}_{MCO}) | \mathbf{X} \right] \right) E \left( g(\widehat{\sigma^2}) \right) \\ &= E(f(\widehat{\beta}_{MCO})) E(g(\widehat{\sigma^2})) \end{aligned}$$

Donc  $\widehat{\beta}_{MCO} \perp \widehat{\sigma^2}$ .

– La loi de  $\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{\sigma^2}[(\mathbf{X}'\mathbf{X})^{-1}]_{(k,k)}}}$

$$\begin{aligned} \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\sigma^2 [(\mathbf{X}'\mathbf{X})^{-1}]_{(k,k)}}} &\rightsquigarrow N(0, 1) \\ \frac{n - (K + 1) \widehat{\sigma^2}}{\sigma^2} &\rightsquigarrow \chi_{n-(K+1)}^2 \\ \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\sigma^2 [(\mathbf{X}'\mathbf{X})^{-1}]_{(k,k)}}} &\parallel \frac{n - (K + 1) \widehat{\sigma^2}}{\sigma^2} \end{aligned}$$

Donc, par définition de la loi de Student,

$$\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{\sigma^2} [(\mathbf{X}'\mathbf{X})^{-1}]_{(k,k)}}} \rightsquigarrow Student(n - (K + 1))$$

– L'optimalité de l'estimateur

On se contentera ici de donner les grandes lignes de la démonstration.

On va raisonner dans un premier temps conditionnellement à  $\mathbf{X}$ . On suppose donc que  $E[\widehat{\beta}, \widehat{\sigma^2} | \mathbf{X}] = (\beta, \sigma^2)$ .

Une statistique exhaustive est une statistique qui "contient" toute l'information disponible dans l'échantillon susceptible d'être utilisée pour identifier les paramètres. Mathématiquement,  $(\widehat{\beta}_{MCO}, \widehat{\sigma^2}_{MCO})$  est une statistique exhaustive si la densité des observations s'écrit sous la forme :  $f_{Y|X}(\mathbf{y}) = g_{\beta, \sigma, X}(\widehat{\beta}_{MCO}, \widehat{\sigma^2}_{MCO}) \cdot h(\mathbf{y})$ .

C'est le cas car :

$$\begin{aligned} f_{Y|X}(\mathbf{y}) &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2} \cdot (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2} \cdot (\mathbf{y} - \mathbf{X}\widehat{\beta}_{MCO} + \mathbf{X}\widehat{\beta}_{MCO} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\widehat{\beta}_{MCO} + \mathbf{X}\widehat{\beta}_{MCO} - \mathbf{X}\beta)} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2} \cdot (\widehat{\epsilon}'\widehat{\epsilon} + (\widehat{\beta}_{MCO} - \beta)'(\mathbf{X}'\mathbf{X})(\widehat{\beta}_{MCO} - \beta))} \\ &= f_X(\mathbf{X}) \cdot (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2} \cdot ((n - (K + 1))\widehat{\sigma^2}_{MCO} + (\widehat{\beta}_{MCO} - \beta)'(\mathbf{X}'\mathbf{X})(\widehat{\beta}_{MCO} - \beta))} \end{aligned}$$

Une statistique exhaustive minimale est une statistique qui contient "toute l'information disponible pour l'identification" (c'est une statistique exhaustive), mais qui "ne contient que l'information nécessaire à l'identification" (elle peut se déduire de toute statistique exhaustive). Mathématiquement, cela s'écrit de la manière suivante : Une statistique  $T^*$  est dite exhaustive minimale si pour toute statistique exhaustive  $T$  il existe une fonction  $h$  (mesurable) telle que  $T^* = h(T)$ .

Une statistique complète est une statistique  $T$  qui vérifie la condition suivante :

$$[\forall (\beta, \sigma^2) \in (\mathbb{R}^{K+1}, \mathbb{R}^{+*}) \quad E[g(T)] = 0] \Rightarrow [g = 0 \text{ (presque partout)}]$$

On admet le résultat suivant : si la loi des observations peut s'écrire sous la forme :

$$f_{Y|X}(\mathbf{y}) = g_X(\mathbf{y}) \cdot K_X(\lambda_0, \lambda_1, \dots, \lambda_{K+1}) \cdot e^{\sum_{j=0}^{K+1} \lambda_j \cdot T_{X,j}(\mathbf{y})} \text{ avec } (\lambda_0, \lambda_1, \dots, \lambda_{K+1}) \in (\mathbb{R}^{K+1}, \mathbb{R}^{-*})$$

alors  $(T_{X,0}, T_{X,1}, \dots, T_{X,K+1})$  est une statistique exhaustive minimale et complète (pour le paramètre  $(\lambda_0, \lambda_1, \dots, \lambda_{K+1})$ ).

La loi que nous considérons peut bien se mettre sous une telle forme :

$$\begin{aligned} f_{Y|X}(\mathbf{y}) &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2} \cdot (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2} (\beta' \mathbf{X}' \mathbf{X} \beta)} \cdot e^{-\frac{1}{2\sigma^2} (\mathbf{y}' \mathbf{y} - 2\beta' \mathbf{X}' \mathbf{y})} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2} (\beta' \mathbf{X}' \mathbf{X} \beta)} \cdot e^{-\frac{1}{2\sigma^2} (\sum_{j=0}^{K+1} \lambda_j \cdot T_{X,j}(\mathbf{y}))} \\ &\text{avec } \lambda_k = -\frac{\beta_k}{2 \cdot \sigma^2} \text{ et } T_{X,k}(\mathbf{y}) = \mathbf{x}_k \mathbf{y} \text{ pour } k = 0, \dots, K \\ &\lambda_k = -\frac{1}{2 \cdot \sigma^2} \text{ et } T_{X,K+1}(\mathbf{y}) = \mathbf{y}' \mathbf{y} \end{aligned}$$

De plus, on vérifie aisément que si  $(T_{X,0}, T_{X,1}, \dots, T_{X,K+1})$  est une statistique exhaustive minimale et complète pour le paramètre  $(\lambda_0, \lambda_1, \dots, \lambda_{K+1})$  alors  $v(T_{X,0}, T_{X,1}, \dots, T_{X,K+1})$  est une statistique exhaustive minimale et complète pour le paramètre  $u(\lambda_0, \lambda_1, \dots, \lambda_{K+1})$ , pour  $u$  et  $v$  deux fonctions bijectives (bimesurables).

En prenant  $u(\lambda_0, \lambda_1, \dots, \lambda_{K+1}) = -\frac{1}{2\lambda_{K+1}}(\lambda_0, \lambda_1, \dots, \lambda_K, 1)$

$$\text{et } v(T_{X,0}, T_{X,1}, \dots, T_{X,K+1}) = \left[ \begin{array}{c} (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} T_{X,0} \\ \vdots \\ T_{X,K} \end{pmatrix} \\ \frac{1}{n-(K+1)} \left[ T_{X,K+1} - \begin{pmatrix} T_{X,0} \\ \vdots \\ T_{X,K} \end{pmatrix}' (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} T_{X,0} \\ \vdots \\ T_{X,K} \end{pmatrix} \right] \end{array} \right],$$

on déduit que  $(\widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2)$  est une statistique exhaustive minimale et complète pour  $(\beta, \sigma^2)$ .

Le théorème de Lehmann-Scheffé permet de conclure : si  $\tilde{\beta}, \tilde{\sigma}^2$  est un estimateur tel que  $E[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}] = (\beta, \sigma^2)$ ,  $E[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}, \widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2]$  est également un estimateur vérifiant :

$E[E[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}, \widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2] | \mathbf{X}] = (\beta, \sigma^2)$ . Donc la complétude de  $(\widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2)$  implique que  $E[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}, \widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2] = (\widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2)$ . Comme  $V[E[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}, \widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2] | \mathbf{X}] \ll V[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}]$  on a  $V[\widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2 | \mathbf{X}] \ll V[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}]$ . En intégrant par rapport à la loi de  $\mathbf{X}$  on a  $V[\widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2] \ll V[\tilde{\beta}, \tilde{\sigma}^2]$ .

Pour se passer de l'hypothèse  $E[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}] = (\beta, \sigma^2)$ , on applique le résultat obtenu à l'estimateur  $(\tilde{\beta}, \tilde{\sigma}^2) - E[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}] + (\beta, \sigma^2)$ . On en déduit que  $V[\widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2 | \mathbf{X}] \ll V[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}]$ . Donc,  $V[\widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2] = E[V[\widehat{\beta}_{MCO}, \widehat{\sigma}_{MCO}^2 | \mathbf{X}]] \ll E[V[\tilde{\beta}, \tilde{\sigma}^2 | \mathbf{X}]] \ll V[\tilde{\beta}, \tilde{\sigma}^2]$  on obtient le résultat désiré.

■

## 2.10 Comparaison avec l'estimateur du Maximum de Vraisemblance (MV)

**L'estimateur du maximum de vraisemblance** : l'estimateur du maximum de vraisemblance est obtenu comme son nom l'indique en trouvant les paramètres  $\beta$  et  $\sigma^2$  maximisant la vraisemblance de la loi des observations  $(y_i, X_i)_{i=1, \dots, n}$ .

### Expression de l'estimateur du MV

Sous les hypothèses  $(H_1)$  et  $(H_n)$ , l'estimateur des moindres carrés de  $\beta$  et  $\sigma^2$  est lié à l'estimateur du maximum de vraisemblance de la façon suivante :

$$\begin{aligned} \widehat{\beta}_{MCO} &= \widehat{\beta}_{MV} \\ \widehat{\sigma}_{MCO}^2 &= \frac{n}{n - (K + 1)} \widehat{\sigma}_{MV}^2 \end{aligned}$$

Preuve :

□

La loi du couple  $((y_i)_{i=1, \dots, n}, (X_i)_{i=1, \dots, n})$  se décompose en un produit de la loi de  $(X_i)_{i=1, \dots, n}$  et de la loi de  $(y_i)_{i=1, \dots, n} | (X_i)_{i=1, \dots, n}$ .

$$\mathcal{L}_{(y_i)_{i=1, \dots, n}, (X_i)_{i=1, \dots, n}}(\mathbf{y}, \mathbf{X}, \beta, \sigma^2) = \mathcal{L}_{(X_i)_{i=1, \dots, n}}(\mathbf{X}) * \mathcal{L}_{(y_i)_{i=1, \dots, n} | (X_i)_{i=1, \dots, n} = \mathbf{X}}(\mathbf{y}, \beta, \sigma^2)$$

Les valeurs de  $\beta$  et de  $\sigma^2$  qui maximisent la vraisemblance de  $((y_i)_{i=1,\dots,n}, (X_i)_{i=1,\dots,n})$  sont donc celles qui maximisent la quantité  $\mathcal{L}_{(y_i)_{i=1,\dots,n} | (X_i)_{i=1,\dots,n}=\mathbf{X}}(\mathbf{y}, \beta, \sigma^2)$ . Sous l'hypothèse de normalité des résidus, cette dernière quantité s'écrit :

$$\mathcal{L}_{Y|X=x}(\mathbf{y}, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}$$

Maximiser cette quantité en  $\beta$  revient à minimiser  $(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$  et donc l'estimateur des moindres carrés  $\widehat{\beta}_{MCO}$  coïncide avec l'estimateur du maximum de vraisemblance  $\widehat{\beta}_{MV}$ . Concernant l'estimation de  $\sigma^2$ , le calcul donne  $\widehat{\sigma^2}_{MV} = \frac{1}{n} (\mathbf{y} - \mathbf{X}\widehat{\beta}_{MV})' (\mathbf{y} - \mathbf{X}\widehat{\beta}_{MV}) = \frac{1}{n} \widehat{\epsilon}'\widehat{\epsilon}$ .

■

**Remarques :** L'estimateur du maximum de vraisemblance de  $\sigma^2$  sous-estime le paramètre car  $E[\widehat{\sigma^2}_{MV}] = \frac{n-(K+1)}{n}\sigma^2$ , ce biais est négligeable lorsque le nombre d'individus observés est grand devant le nombre de covariables ( $n \gg K$ ). Bien que biaisé à distance finie, cet estimateur est asymptotiquement sans biais.

Par ailleurs, on constate que l'estimateur des moindres carrés et celui du maximum de vraisemblance coïncident en ce qui concerne l'estimation de  $\beta$ , et deviennent équivalents lorsque  $n$  devient grand en ce qui concerne l'estimation de  $\sigma^2$ . On en déduit que les propriétés asymptotiques de ces deux estimateurs seront les mêmes.

Sous des conditions assez générales de régularité, l'estimateur du maximum de vraisemblance est un estimateur asymptotiquement optimal des paramètres, parmi la classe des estimateurs asymptotiquement sans biais, c'est-à-dire que :

$$\forall (\widehat{\beta}, \widehat{\sigma^2}) \text{ tels que } \forall (\beta, \sigma^2) \in (\mathbb{R}^{K+1}, \mathbb{R}^{+*}) \left\{ \begin{array}{l} E[\widehat{\beta}] \xrightarrow[n \rightarrow +\infty]{} \beta \\ E[\widehat{\sigma^2}] \xrightarrow[n \rightarrow +\infty]{} \sigma^2 \\ \lim n E \left[ \begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{\sigma^2} - \sigma^2 \end{pmatrix} \begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{\sigma^2} - \sigma^2 \end{pmatrix}' \right] \text{ existe} \end{array} \right.$$

Si on note  $\leq$  l'opérateur de comparaison des matrices symétriques ( $A \leq B$  signifie alors que  $B - A$  est une matrice symétrique positive), on a :

$$\lim n E \left[ \begin{pmatrix} \widehat{\beta}_{MV} - \beta \\ \widehat{\sigma^2}_{MV} - \sigma^2 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_{MV} - \beta \\ \widehat{\sigma^2}_{MV} - \sigma^2 \end{pmatrix}' \right] \leq \lim n E \left[ \begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{\sigma^2} - \sigma^2 \end{pmatrix} \begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{\sigma^2} - \sigma^2 \end{pmatrix}' \right]$$

En particulier, on a :

$$\begin{aligned} \lim n V(\widehat{\beta}_{MV}) &\leq \lim n V(\widehat{\beta}) \\ \lim n V(\widehat{\sigma^2}_{MV}) &\leq \lim n V(\widehat{\sigma^2}) \end{aligned}$$

On retrouve donc là dans un cadre asymptotique, la propriété à distance finie mise en évidence sur l'estimateur des moindres carrés ordinaires.

## 2.11 Programmation sous SAS

L'estimateur des MCO ainsi que toutes les statistiques de tests usuelles peuvent se calculer sous SAS grâce à la proc `reg`.

```
proc reg data=table;
  model Y= X1 X2 ... XK;
run; quit;
```

# Chapitre 3

## Le modèle à effets fixes

### 3.1 Le modèle théorique

#### 3.1.1 Le modèle théorique de base

Le modèle à effets fixes consiste à se placer sous l'hypothèse suivante :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

Le paramètre  $\beta = (\beta_0, \beta_1, \dots, \beta_K)$  est un vecteur de taille  $K + 1$  inconnu, représentant les covariances "nettes" entre  $\mathbf{y}$  et les  $\mathbf{x}_k$ .  $\mathbf{Z}$  est la matrice de taille  $n \times C$  permettant de regrouper les individus par classe. Elle est constituée en concaténant les  $c$  vecteurs-colonnes  $\mathbf{z}_j$  indiquant quels individus appartiennent à la classe  $j$ . Le paramètre  $\gamma = (\gamma_1, \dots, \gamma_C)$  représente donc les différents effets "classe". Le nombre d'élèves observés dans la classe  $j$  est noté  $n_j$ . Le vecteur  $\epsilon = (\epsilon_{ij})_{1 \leq i \leq n_j, 1 \leq j \leq C}$  est un vecteur inconnu de perturbation : il s'agit de la part de variabilité des  $y$  qui n'est pas liée aux variables  $X$  et  $Z$ .

À première vue, cela ressemble fort au modèle linéaire classique que nous avons présenté dans la première partie. Trois types de difficultés apparaissent néanmoins : tout d'abord, il apparaît des relations de colinéarité entre les covariables de  $\mathbf{X}$  et les covariables de  $\mathbf{Z}$ . D'autre part, le nombre d'élèves par classe étant borné (inférieur à 25 élèves dans le premier degré, à 40 dans le second degré), le nombre de classes présentes dans le fichier est très important ce qui pose des problèmes pratiques de calcul. Enfin, le nombre de classes est sensiblement proportionnel au nombre d'observations ce qui nécessite d'étudier attentivement le comportement asymptotique des estimateurs ; si le nombre d'observations augmente, le nombre de classes présentes dans le fichier augmente également et donc le nombre de coefficients à estimer aussi.

#### **La colinéarité entre constante et effets "classe"**

La matrice  $[\mathbf{X}|\mathbf{Z}]$  n'est pas de plein rang. En effet, on ne peut différencier les effets "classe" de l'effet constant sans imposer de restriction supplémentaire (car  $\mathbf{x}_0 = \sum_{j=1}^c \mathbf{z}_j$ ). Trois méthodes sont alors possibles : soit on élimine la constante  $\mathbf{x}_0$ , soit on élimine une des variables  $Z$ , soit on impose que la moyenne des effets "classe" est nulle.

Dans le premier cas (élimination de  $\mathbf{x}_0$ ), le coefficient  $\gamma_j$  associé à chaque classe représentera le "y moyen" de la personne de référence modulé d'un effet "classe". Algébriquement, cela revient à imposer comme contrainte identifiante sur le modèle  $\beta_0 = 0$ .

Dans le deuxième cas (élimination de  $\mathbf{z}_{j_0}$ ), le coefficient  $\gamma_j$  associé à chaque classe s'interprète comme le gain (ou la perte) à être scolarisé dans la classe  $j$  plutôt que dans la classe  $j_0$  ; le coefficient  $\beta_0$  s'interprète comme le "y moyen" de la personne de référence dans la classe  $j_0$ . Algébriquement, cela revient à imposer comme contrainte identifiante sur le modèle  $\gamma_{j_0} = 0$ .

Dans le troisième cas (moyenne des effets "classe" nulle), le coefficient  $\gamma_j$  associé à chaque classe s'interprète comme le gain (ou la perte) à être scolarisé dans la classe  $j$  plutôt que dans "la classe moyenne" ; le coefficient  $\beta_0$  s'interprète comme le "y moyen" de la personne de référence dans "la classe moyenne". Algébriquement cela revient à imposer comme contrainte identifiante sur le modèle  $\frac{1}{c} \sum_{j=1}^c \gamma_j = 0$ .

La troisième méthode d'estimation est sans doute la plus naturelle à interpréter. Mais il faut noter que les coefficients associés aux variables  $\mathbf{X}$  autre que la constante ne sont pas affectés par la méthode et que les coefficients  $\beta_0$  et  $\gamma_j$  peuvent facilement se déduire d'une méthode à l'autre (par soustraction du coefficient contraint à 0).

### Le cas des variables constantes au sein de chaque classe

Cette modélisation pose un autre problème lorsque certaines variables sont constantes pour les élèves d'une même classe (par exemple le secteur de scolarisation, l'appartenance académique, des caractéristiques du professeur, etc.) : on ne peut pas identifier directement les effets associés à ces variables en présence d'effets "classe", ces variables doivent donc être retirées de la régression. C'est ce que nous supposerons par la suite. En revanche, dans une deuxième étape on peut étudier les corrélations entre ces variables et les effets "classe" estimés, même si la prudence est recommandée concernant une interprétation en terme d'impact causal.

### L'optique asymptotique

Enfin, et c'est un point sur lequel nous reviendrons plus en détail, il faut bien avoir à l'esprit que dans une optique asymptotique, on ne peut considérer que le nombre de classes reste fixe alors que le nombre d'élèves sur lesquels porte l'estimation devient très grand. Plus on dispose d'une information riche (on observe beaucoup d'élèves), plus le nombre de paramètres à estimer devient grand (car le nombre de classes augmente), ne pas prendre en compte une telle dépendance peut amener à surestimer la précision des estimateurs.

## 3.2 Identification et définition de l'estimateur

### 3.2.1 Identification du modèle

La prise en compte des remarques concernant la colinéarité des covariables peut se formaliser de la manière suivante :

#### Formalisation et notations

On considère le modèle suivant :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

On ne suppose plus que  $\mathbf{A} = [\mathbf{X}, \mathbf{Z}]$  est de plein rang colonne car  $\mathbf{x}_0 = \sum_{j=1}^C \mathbf{z}_j$ . Si on note

$$e = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ -1 \\ \vdots \\ -1 \end{pmatrix}, \text{ on a alors } \mathbf{A}e = 0.$$

On suppose néanmoins que les vecteurs  $[(\mathbf{x}_k)_{k=1, \dots, K} (\mathbf{z}_j)_{j=1, \dots, C}]$  sont libres ce qui signifie en particulier qu'aucune variable  $\mathbf{x}_k$  ne peut être systématiquement constante pour les élèves d'une même classe (on exclut donc les variables comme le secteur de scolarisation ou l'académie).

Le rang de la matrice  $\mathbf{A}$  est donc  $K + C$  (et non  $K + C + 1$ ); les paramètres ne sont plus identifiables car  $\mathbf{A} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \mathbf{A} \left[ \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \lambda e \right]$ . Ainsi pour assurer l'identification, il faut donc ajouter une contrainte sur les paramètres. Les contraintes les plus naturelles sont  $\beta_0 = 0$ ,  $\gamma_{j_0} = 0$  ou  $\sum_{j=1}^C \gamma_j = 0$ , elles peuvent toutes s'écrire sous la forme  $R' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = 0$ .

De manière plus générale, du point de vue de l'identification des paramètres, il importe juste que

$R$  vérifie  $R = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \\ * \\ \vdots \\ * \end{pmatrix}$  (ce qui signifie intuitivement que  $R$  ne fait pas porter de contraintes sur

les paramètres  $(\beta_1, \beta_2, \dots, \beta_K)$  et  $R'e \neq 0$  (dans le cas contraire la contrainte est mal choisie car elle n'apporte pas d'information).

L'interprétation des paramètres estimés dépend du type de contrainte prise en compte.

		Interprétation dans le cas où $Y$ est une mesure de la réussite scolaire
$\beta_0 = 0$	$R = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$	$\gamma_j$ représente la réussite moyenne dans la classe une fois contrôlé les caractéristiques $X$ des élèves.
$\gamma_{j_0} = 0$	$R = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$	$\beta_0$ représente la réussite moyenne dans la classe de la classe $j_0$ une fois contrôlé les caractéristiques $X$ des élèves et $\gamma_j$ représente la "sur" ou la "sous" réussite de la classe $j$ par rapport à la classe $j_0$ .
$\sum_{j=1}^C \gamma_j = 0$	$R = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$	$\beta_0$ représente la réussite moyenne des élèves à $X$ donné et en contrôlant des effets "classe", $\gamma_j$ représente la "sur" ou la "sous" réussite de la classe $j$ par rapport à la classe moyenne.

Finalement, cela nous conduit à formuler une première hypothèse nécessaire pour assurer l'identification.

### Hypothèse de rang et normalisation des effets fixes

$$(H'_1) : \mathbf{A} \text{ est de rang } K + C, R = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \\ * \\ \vdots \\ * \end{pmatrix} \text{ et } R' = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ -1 \\ \vdots \\ -1 \end{pmatrix} \neq 0.$$

**Remarque :** par rapport à la présentation du modèle linéaire "de base" cf.2, cette hypothèse joue le même rôle que  $(H_1)$  : assurer l'identification du modèle.

### 3.2.2 Définition de l'estimateur des Moindres Carrés Contraints (MCC)

#### Estimateur des Moindres Carrés Contraints (MCC)

Sous l'hypothèse  $(H'_1)$ , l'estimateur des moindres carrés contraints défini par :

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \underset{(\beta', \gamma') \in \mathcal{C}}{\text{Argmin}} \quad (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma)' (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma)$$

avec  $\mathcal{C} = \left\{ (\beta', \gamma') \in \mathbb{R}^{K+C+1} \mid R' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = 0 \right\}$  existe et est unique.

Son expression est :

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \vdots \\ & & 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{A}'\mathbf{A} & R \\ R' & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{A}'\mathbf{y} \\ 0 \end{pmatrix}$$

Preuve :

□ On cherche donc à minimiser la quantité suivante :  $\frac{1}{2} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma)' (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma)$  sous contrainte que  $R' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = 0$ .

Cela implique d'annuler le gradient du lagrangien  $L(\beta, \gamma, l) = \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma)' (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma) + lR' \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$ . Comme  $\frac{\partial L}{\partial(\beta, \gamma)} = -\mathbf{A}' (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma) + lR$  et  $\frac{\partial L}{\partial l} = R' \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$ , on en déduit que :

$$\begin{pmatrix} \mathbf{A}'\mathbf{A} & R \\ R' & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \\ \hat{l} \end{pmatrix} = \begin{pmatrix} \mathbf{A}'\mathbf{y} \\ 0 \end{pmatrix}$$

Pour pouvoir conclure, il faut s'assurer que la matrice  $\begin{pmatrix} \mathbf{A}'\mathbf{A} & R \\ R' & 0 \end{pmatrix}$  est inversible : soit un vecteur  $\mathbf{v}$  tel que  $\begin{pmatrix} \mathbf{A}'\mathbf{A} & R \\ R' & 0 \end{pmatrix} \mathbf{v} = \begin{pmatrix} \mathbf{A}'\mathbf{A} & R \\ R' & 0 \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix} = 0$ . On a donc :

$$\begin{cases} \mathbf{A}'\mathbf{A}\mathbf{v}_1 + R\mathbf{v}_2 = 0 \\ R'\mathbf{v}_1 = 0 \end{cases}$$

Si on pré-multiplie la première ligne par  $\mathbf{v}'_1$ , comme  $\mathbf{v}_1 \in \text{Ker}(R')$ , on obtient donc que  $\mathbf{v}'_1 \mathbf{A}'\mathbf{A}\mathbf{v}_1 = 0$  ce qui implique que  $\mathbf{v}_1 \in \text{Ker}(\mathbf{A})$ , la première partie de l'hypothèse  $(H'_1)$  permet de conclure que  $\mathbf{v}_1 = 0$ . On en déduit alors que  $R\mathbf{v}_2 = 0$ , comme  $R$  est non nul et que  $\mathbf{v}_2$  est un réel, on en déduit que  $\mathbf{v}_2 = 0$ . La matrice  $\begin{pmatrix} \mathbf{A}'\mathbf{A} & R \\ R' & 0 \end{pmatrix}$  est donc inversible.

■

### 3.2.3 Remarque sur le multiplicateur de Lagrange dans l'estimation directe

La plupart des logiciels de statistique permettent de calculer directement des estimateurs des moindres carrés contraints. Certains logiciels ont besoin d'estimer dans un premier temps le modèle sans contrainte, ce qui dans notre cas pose un problème (le modèle non contraint n'étant pas identifiable). Cependant la proc `reg` de SAS ne souffre pas d'un tel défaut. Une programmation "directe" peut donc être utilisée lorsque le nombre de classes n'est pas trop grand. Cette méthode présente l'avantage de calculer directement les écarts-types et les statistiques de test des estimateurs (cf. 3.4 pour la définition des statistiques de test). Le détail de la programmation sera présenté en 3.6.2, mais nous allons présenter une propriété intéressante de l'estimateur obtenu dans la mesure où celle-ci peut (dans notre cas particulier) constituer une vérification des programmes.

### Propriété du multiplicateur de Lagrange

On considère le modèle suivant :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon \quad \text{et} \quad R' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = 0$$

On note  $\mathcal{C} = \left\{ (\beta', \gamma') \in \mathbb{R}^{K+C+1} \mid R' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = 0 \right\}$ .

$$(H'_1) : \mathbf{A} = [\mathbf{X}, \mathbf{Z}] \text{ est de rang } K + C, R = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \\ * \\ \vdots \\ * \end{pmatrix} \text{ et } R' \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ -1 \\ \vdots \\ -1 \end{pmatrix} \neq 0.$$

Sous l'hypothèse  $(H'_1)$ , on a alors la propriété suivante :

$$\begin{aligned} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} &= \underset{(\beta', \gamma') \in \mathcal{C}}{\text{Argmin}} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma)' (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma) \\ &\Leftrightarrow \\ \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \\ 0 \end{pmatrix} &= \underset{(\beta', \gamma', l) \in \mathbb{R}^{K+C+2}}{\text{Argmin}} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma)' (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma) - 2lR' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} \end{aligned}$$

**Remarque :** Le multiplicateur de Lagrange  $l$  représente le coût en terme de la variance expliquée, d'avoir imposé la contrainte  $R' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = 0$ . Dans le cas qui nous intéresse, ce coût est nul, la contrainte étant uniquement imposée pour éliminer les solutions ne correspondant pas à la référence que l'on s'est fixée. Ce paramètre est donc nul ou numériquement très faible (cela peut constituer une vérification de la programmation).

Cette propriété traduit mathématiquement le fait que la contrainte imposée n'est pas "informative". Elle joue uniquement un rôle d'identification afin d'imposer une solution unique au problème (seule l'interprétation des paramètres estimés change). Elle ne modifie donc pas la qualité globale du prédicteur  $\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma}$ . Si la contrainte avait réellement apporté de l'information (par exemple parce qu'on a un modèle théorique qui nous indique que les paramètres doivent vérifier une relation particulière), le multiplicateur de Lagrange  $l$  serait non nul, ce qui conduirait à un prédicteur d'une autre qualité (une meilleure qualité si la contrainte est prise en compte à juste titre, une moins bonne si la contrainte est prise en compte à mauvais escient). Dans un tel cas, le carré de l'estimateur du multiplicateur de Lagrange suit une loi Beta sous l'hypothèse de normalité des résidus (cf. "A Note on the Role of Independence in t Statistics Constructed From Linear Statistics in Regression Models" de La Motte publié dans The American Statistician en 1994).

Preuve :

□

Les conditions d'optimisation du premier ordre sont :  $\mathbf{A}'\mathbf{A} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} + \hat{l}R = \mathbf{A}'\mathbf{y}$  et  $R' \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = 0$ .

Si on note  $e$  le vecteur  $\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ -1 \\ \vdots \\ -1 \end{pmatrix}$ , on a  $e'\mathbf{A}' = 0$  et  $e'R \neq 0$ . La nullité de  $\hat{\lambda}$  s'obtient alors en pré-

multipliant par  $e'$  l'équation  $\mathbf{A}'\mathbf{A} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} + \hat{l}R = \mathbf{A}'\mathbf{y}$ .

■

Cependant, l'estimateur est rarement calculé sous cette forme pour des raisons pratiques : si le nombre de classes dans le fichier est très grand, il devient très lourd d'inverser "brutalement" la matrice  $\begin{pmatrix} \mathbf{A}'\mathbf{A} & R \\ R' & 0 \end{pmatrix}$ . Une méthode de calcul tirant parti de la structure de la matrice à inverser pour réduire la taille du problème est précieuse dans ce cas.

D'autre part, à partir de la forme précédente de l'estimateur, il n'est pas aisé d'étudier les propriétés asymptotiques de l'estimateur à partir de son expression précédente.

Il est plus aisé de calculer et d'étudier cet estimateur en "décomposant les problèmes" et en traitant d'une part ce qui relève de  $\beta$  et d'autre part ce qui relève de  $\hat{\gamma}$ . C'est ce que nous permet de faire une projection sur la dimension intra-classe.

### 3.3 Estimation par projection sur la dimension intra-classe

L'estimateur des moindres carrés avec prise en compte d'un effet "classe" pose deux types de difficultés par rapport au modèle linéaire classique : d'une part des difficultés pratiques car le nombre de classes peut être très grand ce qui alourdit les calculs, d'autre part des difficultés théoriques car dans une approche asymptotique, au fur et à mesure que le nombre d'élèves augmente, le nombre de classes observées augmente aussi alors que le nombre d'élèves dans une classe donnée restera borné (les effets "classe" ne pourront jamais être connus avec une très grande précision).

Le théorème de Frisch-Waugh permet de surmonter ces deux types de difficultés, en séparant l'estimation des paramètres  $\beta$  de celle du paramètre  $\gamma$ . D'un point de vue pratique l'intérêt d'une telle décomposition est de réduire la dimension du problème. D'un point de vue théorique, cette décomposition permet d'étudier séparément les propriétés asymptotiques et à distance finie des estimateurs de  $\beta$  et de  $\gamma$ .

#### 3.3.1 Le théorème de Frisch-Waugh

Le théorème de Frisch-Waugh permet d'explicitier la dépendance entre deux sous-ensembles de variables explicatives. Nous allons d'abord en présenter une version dans un cadre général. Puis nous l'appliquerons pour estimer le modèle à effets fixes.

**Théorème de Frisch-Waugh sous l'hypothèse ( $H_1$ )**

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

( $H_1$ ) : La matrice  $\mathbf{X}$  est de plein rang colonne.

On peut décomposer  $\mathbf{X}$  et  $\beta$  selon deux sous-ensembles de variables  $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$  avec  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ ,  $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$  et  $\hat{\beta}_{MCO} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ . Si on note  $M_{\mathbf{X}_2}$  la matrice  $1 - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$  alors :

$$\begin{aligned} \hat{\beta}_1 &= ((M_{\mathbf{X}_2}\mathbf{X}_1)'(M_{\mathbf{X}_2}\mathbf{X}_1))^{-1}(M_{\mathbf{X}_2}\mathbf{X}_1)'(M_{\mathbf{X}_2}\mathbf{y}) \\ \hat{\beta}_2 &= (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'(\mathbf{y} - \mathbf{X}_1\hat{\beta}_1) \end{aligned}$$

La première équation s'interprète comme le fait qu'on peut estimer  $\hat{\beta}_1$  uniquement en régressant  $M_{\mathbf{X}_2}\mathbf{X}_1$  (c'est-à-dire la projection de  $\mathbf{X}_1$  sur l'espace orthogonal à  $\mathbf{X}_2$  ou autrement dit les résidus estimés par la régression de  $\mathbf{X}_1$  sur  $\mathbf{X}_2$ ) sur  $M_{\mathbf{X}_2}\mathbf{y}$  (la projection de  $\mathbf{y}$  sur le même espace). La deuxième équation s'interprète comme le fait qu'on peut estimer  $\hat{\beta}_2$  uniquement en régressant sur  $\mathbf{X}_2$  la quantité  $\mathbf{y} - \mathbf{X}_1\hat{\beta}_1$ , c'est-à-dire la part de  $\mathbf{y}$  non expliquée par  $\mathbf{X}_2$ .

Preuve :

□

La condition  $\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\beta}_{MCO}) = 0$  se décline en :

$$\mathbf{X}'_1(\mathbf{y} - \mathbf{X}_1\widehat{\beta}_1 - \mathbf{X}_2\widehat{\beta}_2) = 0$$

$$\mathbf{X}'_2(\mathbf{y} - \mathbf{X}_1\widehat{\beta}_1 - \mathbf{X}_2\widehat{\beta}_2) = 0$$

La deuxième équation implique  $\widehat{\beta}_2 = (\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}_2(\mathbf{y} - \mathbf{X}_1\widehat{\beta}_1)$ .

On injecte l'expression de  $\widehat{\beta}_2$  ainsi obtenu dans la première équation et on obtient :

$$\mathbf{X}'_1(\mathbf{y} - \mathbf{X}_1\widehat{\beta}_1 - \mathbf{X}_2(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}_2(\mathbf{y} - \mathbf{X}_1\widehat{\beta}_1)) = 0$$

$$\text{soit } \mathbf{X}'_1M_{\mathbf{X}_2}(\mathbf{y} - \mathbf{X}_1\widehat{\beta}_1) = 0$$

Comme  $M_{\mathbf{X}_2}^2 = M_{\mathbf{X}_2} = M'_{\mathbf{X}_2}$  cela peut encore se réécrire :

$$(M_{\mathbf{X}_2}\mathbf{X}_1)'(M_{\mathbf{X}_2}\mathbf{y} - M_{\mathbf{X}_2}\mathbf{X}_1\widehat{\beta}_1) = 0$$

■

Dans la formulation classique du théorème de Frisch-Waugh, on suppose que la matrice des variables explicatives est de plein rang. Ce n'est pas le cas dans le modèle à effets fixes que l'on cherche à estimer, ce qui implique une petite adaptation.

### 3.3.2 Le théorème de Frisch-Waugh appliqué au modèle à effets fixes

**Théorème de Frisch-Waugh appliqué à l'estimation des effets intra-classe**

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

La matrice de projection sur l'espace orthogonal aux indicatrices de classes  $M_{\mathbf{Z}}$  est l'opérateur de différence à la moyenne de la classe, cet opérateur est généralement noté  $W$  (pour opérateur "Within") :

$$M_{\mathbf{Z}}\mathbf{U} = \left( u_{i,j} - \frac{1}{n_j} \sum_{k \in j} u_{k,j} \right)_{j=1 \dots C, i=1 \dots n_j} = W\mathbf{U}$$

Comme cela est fait dans le théorème de Frisch-Waugh, on peut définir des estimateurs "Within"  $\widehat{\beta}_W$  et  $\widehat{\gamma}_W$  :

$$\widehat{\beta}_W = \left( \begin{array}{c} 0 \\ ((W\mathbf{X}_{-0})'(W\mathbf{X}_{-0}))^{-1} (W\mathbf{X}_{-0})'(W\mathbf{y}) \end{array} \right) = \left( \begin{array}{c} 0 \\ (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1} (\mathbf{X}'_{-0}W\mathbf{y}) \end{array} \right)$$

$$\widehat{\gamma}_W = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\beta}_W)$$

Les estimateurs "Within" coïncident avec l'estimateur des moindres carrés obtenu sous la contrainte identifiante  $\beta_0 = 0$ . On peut retrouver par combinaison linéaire les estimateurs pour les contraintes  $\gamma_{j_0} = 0$  et  $\sum_{j=1}^C \gamma_j = 0$ .

Contrainte	Estimateur
$\beta_0 = 0$	$(\widehat{\beta}'_{\beta_0=0}, \widehat{\gamma}'_{\beta_0=0}) = (\widehat{\beta}'_W, \widehat{\gamma}'_W)$
$\gamma_{j_0} = 0$	$(\widehat{\beta}'_{\gamma_{j_0}=0}, \widehat{\gamma}'_{\gamma_{j_0}=0}) =$ $(\widehat{\beta}'_W + (\widehat{\gamma}_{Wj_0}, 0, \dots, 0), \widehat{\gamma}'_W - (\widehat{\gamma}_{Wj_0}, \widehat{\gamma}_{Wj_0}, \dots, \widehat{\gamma}_{Wj_0}))$
$\sum_{j=1}^C \gamma_j = 0$	$(\widehat{\beta}'_{\sum_{j=1}^C \gamma_j=0}, \widehat{\gamma}'_{\sum_{j=1}^C \gamma_j=0}) =$ $(\widehat{\beta}'_W + (\frac{1}{C} \sum_{j=1}^C \widehat{\gamma}_{Wj}, 0, \dots, 0), \widehat{\gamma}'_W - (\frac{1}{C} \sum_{j=1}^C \widehat{\gamma}_{Wj}, \dots, \frac{1}{C} \sum_{j=1}^C \widehat{\gamma}_{Wj}))$

Preuve :

□

Le calcul permet de vérifier la propriété concernant  $M_{\mathbf{Z}}$ .

$$\text{On a } \mathbf{Z}'\mathbf{U} = \begin{pmatrix} \sum_{k \in 1} u_{k,1} \\ \sum_{k \in 2} u_{k,2} \\ \vdots \\ \sum_{k \in C} u_{k,C} \end{pmatrix} \text{ et } (\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} \frac{1}{n_1} & & (0) \\ & \ddots & \\ (0) & & \frac{1}{n_C} \end{bmatrix}$$

Donc :

$$(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{U} = \begin{pmatrix} \frac{1}{n_1} \sum_{k \in 1} u_{k,1} \\ \frac{1}{n_2} \sum_{k \in 2} u_{k,2} \\ \vdots \\ \frac{1}{n_C} \sum_{k \in C} u_{k,C} \end{pmatrix}$$

$$\text{et } \mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{U} = \begin{pmatrix} \frac{1}{n_1} \sum_{k \in 1} u_{k,1} \\ \vdots \\ \frac{1}{n_1} \sum_{k \in 1} u_{k,1} \\ \frac{1}{n_2} \sum_{k \in 2} u_{k,2} \\ \vdots \\ \frac{1}{n_2} \sum_{k \in 2} u_{k,2} \\ \vdots \\ \frac{1}{n_C} \sum_{k \in C} u_{k,C} \\ \vdots \\ \frac{1}{n_C} \sum_{k \in C} u_{k,C} \end{pmatrix}$$

finalemt

$$M_{\mathbf{Z}}\mathbf{U} = \left( u_{i,j} - \frac{1}{n_j} \sum_{k \in j} u_{k,j} \right)_{j=1 \dots C, i=1 \dots n_j}$$

L'estimateur des moindres carrés du modèle sous la contrainte identifiante  $\beta_0 = 0$  s'écrit :

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} 0 \\ \hat{\beta}_{-0} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} 0 \\ ([\mathbf{X}_{-0}|\mathbf{Z}]' [\mathbf{X}_{-0}|\mathbf{Z}])^{-1} [\mathbf{X}_{-0}|\mathbf{Z}]' \mathbf{y} \end{pmatrix}$$

avec  $\hat{\beta}_{-0} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_K \end{pmatrix}$  et  $\mathbf{X}_{-0}$  la matrice des covariables privée de la colonne constante  $\mathbf{x}_0$ .

Il suffit alors d'appliquer le théorème de Frisch-Waugh au modèle  $\mathbf{y} = \mathbf{X}_{-0}\beta_{-0} + \mathbf{Z}\gamma + \epsilon$  pour conclure que  $(\hat{\beta}'_{\beta_0=0}, \hat{\gamma}'_{\beta_0=0}) = (\hat{\beta}'_W, \hat{\gamma}'_W)$ .

Pour retrouver les estimateurs correspondant aux autres contraintes, il suffit de vérifier que les expressions proposées minimisent également la fonction objectif, tout en respectant la contrainte choisie, l'unicité de l'estimateur associé à chaque contrainte permet de conclure.

■

Dans le cas où les effets "classe" ne sont pas le sujet d'analyse en temps que tel, il peut être pertinent de les prendre en compte dans l'analyse des effets individuels sans pour autant les estimer. Pour cela, il suffit alors de raisonner en "écart à la moyenne au sein de la classe", les calculs sont alors grandement simplifiés et allégés. D'autre part, même lorsqu'on est intéressé par le fait d'estimer les effets "classe", il n'est pas toujours possible de procéder à une estimation "directe" (cf. 3.2.3) car les calculs deviennent trop lourds pour un micro-ordinateur classique.

On procède alors de la manière suivante :

### Estimation des paramètres du modèle à effets fixes

Étape 1 : Calcul pour chaque individu des écarts de  $Y$  et  $X_{-0}$  à la moyenne de sa classe. On récupère ainsi  $W\mathbf{y} = \left( y_{i,j} - \frac{1}{n_j} \sum_{k=1}^{n_j} y_{k,j} \right)_{j=1 \dots C, i=1 \dots n_j}$  et  $W\mathbf{X}_{-0} = \left( X_{-0 \ i,j} - \frac{1}{n_j} \sum_{k=1}^{n_j} X_{-0 \ k,j} \right)_{j=1 \dots C, i=1 \dots n_j}$ .

Étape 2 : On régresse  $W\mathbf{y}$  sur  $W\mathbf{X}_{-0}$ , on récupère ainsi  $\hat{\beta}_W$ .

Étape 3<sup>1</sup> : On calcule pour chaque individu l'écart entre la variable  $y$  et sa prédiction obtenue sans prendre en compte l'appartenance à la classe  $Y - \hat{\beta}_W X_{-0}$

Étape 4<sup>2</sup> : On calcule classe par classe la moyenne des écarts  $Y - \hat{\beta}_W X_{-0}$ , on récupère ainsi les  $\hat{\gamma}_W$

Étape 5<sup>3</sup> : On normalise éventuellement  $\hat{\beta}_W$  et  $\hat{\gamma}_W$  pour obtenir les estimateurs sous contrainte  $\sum_{j=1}^C \gamma_j = 0$  ou  $\gamma_{j_0} = 0$ .

Étape 6 : Reste à calculer les écarts-types et les statistiques de test, cf.3.4.7

Dans certains cas (lorsque le nombre de classes n'est pas trop grand), il est possible d'estimer directement les paramètres du modèle avec la proc `reg` de SAS car cette procédure permet de prendre en compte des contraintes linéaires sur les paramètres. C'est naturellement bien plus confortable pour l'utilisateur.

## 3.4 Propriétés de l'estimateur des MCC

### 3.4.1 L'estimateur est-il sans biais ?

Comme dans le chapitre 2, nous allons étudier sous quelles hypothèses les estimateurs sont sans biais, et quelle est leur précision dans le cas homoscédastique. On repose donc le même type d'hypothèse que celles énoncées précédemment.

#### Hypothèse d'exogénéité

$(H_2)$  : L'espérance des résidus conditionnellement aux covariables est nulle ce qui s'écrit :

$$E[\epsilon | \mathbf{X}, \mathbf{Z}] = 0$$

On rappelle que cette hypothèse est loin d'être triviale et que lorsqu'elle est violée, on est par définition dans un cas d'endogénéité des régresseurs (cf. 2).

#### Biais de l'estimateur

Sous les conditions  $(H_1)$  et  $(H_2)$ , l'estimateur des MCC est sans biais, c'est-à-dire que :

$$E \begin{pmatrix} \hat{\beta}_{MCC} \\ \hat{\gamma}_{MCC} \end{pmatrix} = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$$

Preuve :

□

Commençons par calculer l'espérance des estimateurs "Within" (qui correspond à la contrainte  $\beta_0 = 0$ ), les autres cas se déduiront par combinaison linéaire.

- 
1. Étape nécessaire uniquement si on cherche à estimer  $\beta_0$  et  $\gamma$ .
  2. Idem.
  3. Idem.

$$\begin{aligned}
E(\widehat{\beta}_W) &= E \left[ \begin{array}{c} 0 \\ (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0} W \mathbf{y} \end{array} \right] \\
&= E \left[ \begin{array}{c} 0 \\ (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0} W (\mathbf{X} \beta + \mathbf{Z} \gamma + \epsilon) \end{array} \right] \\
&= \beta - \begin{pmatrix} \beta_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + E \left[ \begin{array}{c} 0 \\ (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0} W \epsilon \end{array} \right] \quad \text{car } W \mathbf{Z} = 0 \text{ et } W \mathbf{x}_0 = 0 \\
&= \beta - \begin{pmatrix} \beta_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + E \left[ \begin{array}{c} 0 \\ (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0} W E[\epsilon | \mathbf{X}, \mathbf{Z}] \end{array} \right] \\
&= \beta - \begin{pmatrix} \beta_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
E(\widehat{\gamma}_W) &= E \left[ (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' (\mathbf{y} - \mathbf{X}' \widehat{\beta}_W) \right] \\
&= \gamma + E \left[ (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} (\beta - \widehat{\beta}_W) + (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \epsilon \right] \\
&= \gamma + E \left[ (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} E(\beta - \widehat{\beta}_W | \mathbf{X}, \mathbf{Z}) + (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' E(\epsilon | \mathbf{X}, \mathbf{Z}) \right] \\
&= \gamma + E \left[ (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} \begin{pmatrix} \beta_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' E(\epsilon | \mathbf{X}, \mathbf{Z}) \right] \\
&= \gamma + \begin{pmatrix} \beta_0 \\ \beta_0 \\ \vdots \\ \beta_0 \end{pmatrix}
\end{aligned}$$

On en conclut que dans le cas où la contrainte est  $\beta_0 = 0$ , l'estimateur de moindres carrés contraints (qui coïncide avec l'estimateur "Within") est sans biais.

Pour conclure dans le cas général, il suffit de remarquer que si on note  $\begin{pmatrix} \beta_{\beta_0=0} \\ \gamma_{\beta_0=0} \end{pmatrix}$  la paramétrisation du modèle dans lequel on fixe la constante à 0, les autres paramétrisations du modèle (selon les différentes contraintes identifiantes retenues) s'écrivent sous la forme  $\begin{pmatrix} \beta \\ \gamma \end{pmatrix} =$

$$\begin{pmatrix} \beta_{\beta_0=0} \\ \gamma_{\beta_0=0} \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \text{ avec } \lambda = \gamma_{\beta_0=0, j_0} \text{ ou encore } \frac{1}{C} \sum_{j=1}^C \gamma_{\beta_0=0, j}. \text{ On vérifie bien qu'à chaque}$$

fois l'estimateur de  $\lambda$  utilisé pour changer de contrainte est bien sans biais ( $E(\widehat{\gamma}_{W_{j_0}}) = \gamma_{\beta_0=0, j_0}$  et  $E \left[ \frac{1}{C} \sum_{j=1}^C \widehat{\gamma}_{W_j} \right] = \frac{1}{C} \sum_{j=1}^C \gamma_{\beta_0=0, j}$ ).

### 3.4.2 Quelle est la précision de l'estimateur ?

Après avoir étudié le biais de l'estimateur, on cherche naturellement à caractériser sa variance. On se place généralement sous l'hypothèse d'homoscédasticité.

### Hypothèse d'homoscédasticité

( $H'_3$ ) : La variance des résidus conditionnellement aux covariables est telle que :

$$V[\epsilon|\mathbf{X}, \mathbf{Z}] = \sigma^2 I$$

### Variance des estimateurs

Sous les hypothèses ( $H'_1$ ), ( $H'_2$ ) et ( $H'_3$ ), les variances des estimateurs "Within" (contrainte identifiante  $\beta_0 = 0$ ) de  $\beta$  et  $\gamma$  conditionnellement aux covariables sont :

$$V(\widehat{\beta}_W|\mathbf{X}, \mathbf{Z}) = \sigma^2 \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & & \\ 0 & & & (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \end{bmatrix}$$

$$V(\widehat{\gamma}_W|\mathbf{X}, \mathbf{Z}) = \sigma^2 \left[ (\mathbf{Z}'\mathbf{Z})^{-1} + (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X}_{-0} (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0} \mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \right]$$

Les variances des estimateurs "Within" sont donc :

$$V(\widehat{\beta}_W) = V(\widehat{\beta}_{\beta_0=0}) = \sigma^2 \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & & E[(\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1}] \\ 0 & & & \end{bmatrix}$$

$$V(\widehat{\gamma}_W) = V(\widehat{\gamma}_{\beta_0=0}) = \sigma^2 E \left[ (\mathbf{Z}'\mathbf{Z})^{-1} + (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X}_{-0} (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0} \mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \right]$$

Les variances des estimateurs des moindres carrés sous la contrainte identifiante  $\gamma_{j_0} = 0$  sont de la forme :

$$V(\widehat{\beta}_{\gamma_{j_0}=0}) = \begin{bmatrix} v_{j_0, j_0} & * & * & * \\ * & & & \\ * & \sigma^2 E[(\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1}] & & \\ * & & & \end{bmatrix} \quad \text{avec } v_{j_0, j_0} = [V(\widehat{\gamma}_W)]_{j_0, j_0}$$

$$V(\widehat{\gamma}_{\gamma_{j_0}=0}) = \left( I_C - \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix} \right) V(\widehat{\gamma}_W) \left( I_C - \begin{bmatrix} 0 & \dots & 0 \\ \vdots & & \\ 0 & \dots & 0 \\ 1 & \dots & 1 \\ 0 & \dots & 0 \\ \vdots & & \\ 0 & \dots & 0 \end{bmatrix} \right)$$

Les variances des estimateurs des moindres carrés sous la contrainte identifiante  $\sum_{j=1}^C \gamma_j = 0$  sont de la forme :

$$V(\widehat{\beta}_{\sum_{j=1}^C \gamma_j=0}) = \begin{bmatrix} \bar{v} & * & * & * \\ * & & & \\ * & \sigma^2 E[(\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1}] & & \\ * & & & \end{bmatrix} \quad \text{avec } \bar{v} = \frac{1}{C^2} \sum_{1 \leq i, j \leq C} [V(\widehat{\gamma}_W)]_{i, j}$$

$$V(\widehat{\gamma}_{\sum_{j=1}^C \gamma_j=0}) = \left( I_C - \frac{1}{C} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix} \right) V(\widehat{\gamma}_W) \left( I_C - \frac{1}{C} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix} \right)$$

**Remarque :** La variance des  $\hat{\gamma}_j^W$  est toujours supérieure ou égale à  $\frac{\sigma^2}{n_j}$ . En effet,  $\left[ (\mathbf{Z}'\mathbf{Z})^{-1} \right]_{j,j} = \frac{1}{n_j}$  et le deuxième terme apparaissant dans l'expression de la variance est positif. On remarque de plus que ce deuxième terme sera d'autant plus grand que les  $\mathbf{X}$  seront corrélés aux  $\mathbf{Z}$ . Tout cela traduit deux idées : tout d'abord que la précision de l'estimation d'un effet "classe" est toujours limitée par le nombre nécessairement restreint d'élèves présents dans la classe (premier terme en  $\frac{\sigma^2}{n_j}$ ), d'autre part que les effets "classe" sont d'autant plus difficiles à estimer précisément que les classes sont différentes en terme de composition (au sens des variables  $\mathbf{X}$ ).

Preuve :

□

$V(\hat{\beta}_W | \mathbf{X}, \mathbf{Z})$  se calcule facilement à partir de  $\hat{\beta}_W = \begin{pmatrix} 0 \\ (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0} W \end{pmatrix} \mathbf{y}$ , sachant que  $V(\mathbf{y} | \mathbf{X}, \mathbf{Z}) = \sigma^2 I_n$ .

$V(\hat{\gamma}_W | \mathbf{X}, \mathbf{Z}) = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' V(\mathbf{y} - \mathbf{X} \hat{\beta}_W | \mathbf{X}, \mathbf{Z}) \mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1}$  et comme  $\mathbf{X} \hat{\beta}_W = \mathbf{X}_{-0} (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0} W \mathbf{y}$  on a :

$V(\hat{\gamma}_W | \mathbf{X}, \mathbf{Z}) = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' [I_n - \mathbf{X}_{-0} (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0} W] V(\mathbf{y} | \mathbf{X}, \mathbf{Z}) [I_n - W \mathbf{X}_{-0} (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0}] \mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1}$ .

On utilise ensuite le fait que  $W\mathbf{Z} = 0$  et  $\mathbf{Z}'W = 0$  pour obtenir le résultat annoncé.

Comme  $E(\hat{\beta}_W | \mathbf{X}, \mathbf{Z}) = \begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$  et  $E(\hat{\gamma}_W | \mathbf{X}, \mathbf{Z}) = \gamma + \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_0 \end{pmatrix}$  ne dépendent pas de  $\mathbf{X}$  et de  $\mathbf{Z}$ ,

on a  $V(\hat{\beta}_W) = E[V(\hat{\beta}_W | \mathbf{X}, \mathbf{Z})]$  et  $V(\hat{\gamma}_W) = E[V(\hat{\gamma}_W | \mathbf{X}, \mathbf{Z})]$ .

Comme  $\hat{\beta}_{\gamma_{j_0}=0} = \hat{\beta}_W + \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \hat{\gamma}_{W_{j_0}}$ , on a :

$$V(\hat{\beta}_{\gamma_{j_0}=0}) = V(\hat{\beta}_W) + \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} V(\hat{\gamma}_{W_{j_0}}) \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} +$$

$$\left[ E(\hat{\beta}_W \hat{\gamma}_{W_{j_0}}) - E(\hat{\beta}_W) E(\hat{\gamma}_{W_{j_0}}) \right] \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \left[ E(\hat{\gamma}_{W_{j_0}} \hat{\beta}'_W) - E(\hat{\gamma}_{W_{j_0}}) E(\hat{\beta}'_W) \right].$$

La première composante de  $\hat{\beta}_W$  étant toujours nulle,  $V(\hat{\beta}_{\gamma_{j_0}=0}) = \begin{pmatrix} V(\hat{\gamma}_{W_{j_0}}) & * \\ * & 0 \end{pmatrix} + V(\hat{\beta}_W)$ , et comme  $V(\hat{\gamma}_{W_{j_0}}) = [V(\hat{\gamma}_W)]_{j_0, j_0}$ , on obtient le résultat.

Le calcul de  $V(\hat{\gamma}_{\gamma_{j_0}=0})$  est immédiat à partir de  $\hat{\gamma}_{\gamma_{j_0}=0} = \left( I_C - \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix} \right) \hat{\gamma}_W$ .

Le calcul de  $V(\hat{\beta}_{\sum_{j=1}^C \gamma_j=0})$  suit exactement le même cheminement que celui de  $V(\hat{\beta}_{\gamma_{j_0}=0})$  une

fois remarqué qu'on a  $\hat{\beta}_{\sum_{j=1}^C \gamma_j=0} = \hat{\beta}_W + \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \frac{1}{C} \sum_{j=1}^C \hat{\gamma}_j$ .

Enfin, l'expression de  $V(\hat{\gamma}_{\sum_{j=1}^C \gamma_j=0})$  se déduit du fait que  $\hat{\gamma}_{\sum_{j=1}^C \gamma_j=0} = \left( I_C - \frac{1}{C} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix} \right) \hat{\gamma}_W$ .

■

### 3.4.3 Comment estimer la précision de l'estimateur ?

Comme dans le cas de l'estimateur des moindres carrés ordinaires dont les propriétés ont été rappelées en 2, la variance des estimateurs dépend de  $\sigma^2$ . Pour estimer la variance des estimateurs du modèle à effets fixes, on construit donc un estimateur de  $\sigma^2$ .

#### Estimation de la dispersion des inobservées

Sous les hypothèses  $(H'_1)$ ,  $(H'_2)$  et  $(H'_3)$ , l'estimateur :

$$\widehat{\sigma}^2 = \frac{(\widehat{W}\epsilon)'(\widehat{W}\epsilon)}{n - K - C} \quad \text{avec} \quad \widehat{W}\epsilon = W\mathbf{y} - W\mathbf{X}\widehat{\beta}_W$$

est un estimateur sans biais du paramètre  $\sigma^2$ .

Preuve :

□

On a :

$$\begin{aligned} \mathbf{y} - \mathbf{X}\widehat{\beta}_W &= \left[ I_n - \mathbf{X} \begin{pmatrix} 0 \\ (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}\mathbf{X}'_{-0}W \end{pmatrix} \right] \mathbf{y} \\ &= \mathbf{x}_0\beta_0 + \mathbf{Z}\gamma + \left[ I_n - \mathbf{X} \begin{pmatrix} 0 \\ (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}\mathbf{X}'_{-0}W \end{pmatrix} \right] \epsilon \quad \text{car } W\mathbf{Z} = 0 \text{ et } W\mathbf{x}_0 = 0 \end{aligned}$$

On en déduit :

$$\begin{aligned} E \left[ \widehat{W}\epsilon' \widehat{W}\epsilon \mid \mathbf{X}, \mathbf{Z} \right] &= E \left[ \epsilon' \left\{ W - W\mathbf{X} \begin{pmatrix} 0 \\ (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}\mathbf{X}'_{-0}W \end{pmatrix} \right\} \epsilon \mid \mathbf{X}, \mathbf{Z} \right] \\ &= E \left[ \text{Tr} \left( \epsilon \epsilon' \left\{ W - W\mathbf{X} \begin{pmatrix} 0 \\ (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}\mathbf{X}'_{-0}W \end{pmatrix} \right\} \right) \mid \mathbf{X}, \mathbf{Z} \right] \\ &= \text{Tr} \left( E \left[ \epsilon \epsilon' \mid \mathbf{X}, \mathbf{Z} \right] \left\{ W - W\mathbf{X} \begin{pmatrix} 0 \\ (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}\mathbf{X}'_{-0}W \end{pmatrix} \right\} \right) \\ &= \sigma^2 \text{Tr} \left( W - W\mathbf{X} \begin{pmatrix} 0 \\ (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}\mathbf{X}'_{-0}W \end{pmatrix} \right) \\ &= \sigma^2 \left[ \text{Tr}(W) - \text{Tr} \left\{ \begin{pmatrix} 0 \\ (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}\mathbf{X}'_{-0}W \end{pmatrix} W\mathbf{X} \right\} \right] \\ &= \sigma^2 \left[ \text{Tr}(W) - \text{Tr} \left\{ \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} & & & \\ & & & \\ & & I_K & \\ & & & \end{bmatrix} \right\} \right] \end{aligned}$$

On conclut en remarquant que  $\text{Tr}(W) = n - C$  et  $\text{Tr}(I_K) = K$ .

■

Une fois construit un estimateur de  $\sigma^2$ , on en déduit simplement des estimateurs de la variance des estimateurs des paramètres du modèle.

#### Estimation de la variance des estimateurs

Sous les hypothèses  $(H'_1)$ ,  $(H'_2)$  et  $(H'_3)$ , les estimateurs :

$$\widehat{V} \left( \widehat{\beta}_{\beta_0=0} \right) = \widehat{V} \left( \widehat{\beta}_W \right) = \widehat{\sigma}^2 \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 \\ \vdots \\ (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1} \\ 0 \end{bmatrix}$$

$$\widehat{V}(\widehat{\gamma}_{\beta_0=0}) = \widehat{V}(\widehat{\gamma}_W) = \widehat{\sigma}^2 \left[ (\mathbf{Z}'\mathbf{Z})^{-1} + (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X}_{-0} (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0}\mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \right]$$

sont des estimateurs sans biais des variances de  $\widehat{\beta}$  et  $\widehat{\gamma}$  si la contrainte identifiante est  $\beta_0 = 0$ .  
On en déduit les estimateurs de variance suivants si la contrainte identifiante est  $\gamma_{j_0=0}$  :

$$\widehat{V}(\widehat{\beta}_0) = \widehat{v}_{j_0, j_0} \left[ \widehat{V}(\widehat{\gamma}_W) \right]_{j_0, j_0}$$

$$\widehat{V} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_K \end{pmatrix} = \widehat{\sigma}^2 (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}$$

$$\widehat{V}(\widehat{\gamma}) = \left( I_C - \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix} \right) \widehat{V}(\widehat{\gamma}_W) \begin{pmatrix} I_C - \begin{bmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ 1 & \dots & 1 \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix} \end{pmatrix}$$

... et les estimateurs de variance suivants si la contrainte identifiante est  $\sum_{j=1}^C \gamma_j = 0$  :

$$\widehat{V}(\widehat{\beta}_0) = \widehat{v} = \frac{1}{C^2} \sum_{1 \leq i, j \leq C} \left[ \widehat{V}(\widehat{\gamma}_W) \right]_{i, j}$$

$$\widehat{V} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_K \end{pmatrix} = \widehat{\sigma}^2 (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}$$

$$\widehat{V}(\widehat{\gamma}) = \left( I_C - \frac{1}{C} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix} \right) \widehat{V}(\widehat{\gamma}_W) \left( I_C - \frac{1}{C} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix} \right)$$

Preuve :

□

C'est une conséquence directe du fait que  $E[\widehat{\sigma}^2 | \mathbf{X}, \mathbf{Z}] = \sigma^2$ .

■

### 3.4.4 L'estimateur est-il convergent ?

Concernant les propriétés asymptotiques, il y a deux points importants à avoir à l'esprit : le premier concerne le caractère non convergent de  $\widehat{\gamma}_W$  et le second la dépendance de la convergence de  $\widehat{\beta}_W$  par rapport au nombre d'élèves mais également de classes.

Pour toute classe  $j$ , l'estimateur  $\widehat{\gamma}_j$  n'est pas convergent : une fois qu'on a pris en compte tous les élèves de la classe  $j$  pour estimer l'effet "classe" on ne peut plus mobiliser d'information, dans tous les cas l'effet "classe" n'est donc connu qu'à une marge (irréductible) d'incertitude près.

Par ailleurs, la vitesse de convergence de l'estimateur  $\widehat{\beta}$  ne dépend plus uniquement du nombre d'élèves mais également du nombre de classes : en augmentant le nombre d'élèves pris en compte dans l'analyse, on augmente également le nombre de classes, une part de l'information apportée par l'observation des élèves d'une classe  $j$  ne peut être utilisée pour estimer  $\beta$  car il faut également estimer  $\gamma_j$ .

## Convergence en probabilité des estimateurs

Sous les hypothèses  $(H'_1)$ ,  $(H'_2)$  et  $(H'_3)$

si

◇  $\frac{1}{n-C} \mathbf{X}' W \mathbf{X}$  tend en probabilité vers une matrice  $\Sigma$  inversible quand  $n$  tend vers  $+\infty$

◇  $\frac{1}{n-C} \mathbf{X}' W \epsilon$  tend en probabilité vers 0 quand  $n$  tend vers  $+\infty$

Alors  $\widehat{\beta}_W$  tend vers  $\begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$  en probabilité mais  $\widehat{\gamma}_W$  ne tend pas vers  $\gamma + \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_0 \end{pmatrix}$  en probabilité.

### Remarque :

Dans le cas qui nous intéresse où les unités sont  $n$  élèves répartis en  $C$  classes, il est assez raisonnable de supposer que lorsque  $n$  tend vers  $+\infty$ , le nombre moyen d'élèves observés par classe  $\frac{n}{C}$  tend vers une constante  $c > 1$ . Les hypothèses sont alors équivalentes à  $\frac{1}{n} \mathbf{X}' W \mathbf{X}$  tend en probabilité vers  $\tilde{\Sigma} = (1 - \frac{1}{c})\Sigma$  inversible et  $\frac{1}{n} \mathbf{X}' W \epsilon$  tend en loi vers  $N(0, \tilde{\Sigma})$  ou encore à  $\frac{1}{C} \mathbf{X}' W \mathbf{X}$  tend en probabilité vers  $\tilde{\Sigma} = (1 - c)\Sigma$  inversible et  $\frac{1}{C} \mathbf{X}' W \epsilon$  tend en loi vers  $N(0, \tilde{\Sigma})$ . Le choix d'exprimer cette dernière hypothèse en utilisant la vitesse de convergence  $\frac{1}{n-C}$  répond surtout au besoin de simplifier la présentation par la suite : en effet le nombre de degrés de liberté du modèle qui intervient dans la construction de la statistique de test se comporte en  $n - C - K \sim n - C$ .

Par ailleurs, cette hypothèse peut paraître assez abstraite mais elle peut être vue comme une conséquence de la loi des grands nombres et du théorème central limite. On donnera donc par la suite un jeu d'hypothèses moins abstraites mais plus fortes qui permettront de caractériser plus précisément la loi asymptotique de  $\widehat{\beta}_W$ .

Preuve :

□

Concernant la convergence de  $\widehat{\beta}_W$ , il suffit de remarquer que :

$$\widehat{\beta}_W = \begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} + \begin{pmatrix} 0 \\ \left( \frac{1}{n-C} \mathbf{X}'_{-0} W \mathbf{X}_{-0} \right)^{-1} \left( \frac{1}{n-C} \mathbf{X}'_{-0} W \epsilon \right) \end{pmatrix}$$

$\left( \frac{1}{n-C} \mathbf{X}'_{-0} W \mathbf{X}_{-0} \right)^{-1}$  qui tend en probabilité vers  $\Sigma^{-1}$  et finalement  $\left( \frac{1}{n-C} \mathbf{X}'_{-0} W \mathbf{X}_{-0} \right)^{-1} \left( \frac{1}{n-C} \mathbf{X}'_{-0} W \epsilon \right)$  tend en probabilité vers 0 d'après le théorème de Slutsky.

Concernant la non convergence de  $\widehat{\gamma}_W$ , il suffit de s'intéresser à une classe  $j$  particulière :

$$\widehat{\gamma}_{Wj} = \gamma_j + \beta_0 + \frac{1}{n_j} \sum_{i \in j} \mathbf{X}_i \left( \begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} - \widehat{\beta}_W \right) + \frac{1}{n_j} \sum_{i \in j} \epsilon_i. \text{ Comme } \widehat{\beta}_W \text{ converge (en probabilité)}$$

vers  $\begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$ ,  $\widehat{\gamma}_{Wj}$  converge vers  $\gamma_j + \beta_0$  si et seulement si  $\frac{1}{n_j} \sum_{i \in j} \epsilon_i$  converge vers 0. Cela n'est pas vérifié puisque  $n_j$  est nécessairement borné par le nombre total d'élèves de la classe  $j$ .

■

### 3.4.5 Quelles sont les propriétés asymptotiques de l'estimateur ?

Pour caractériser plus précisément la loi de l'estimateur  $\widehat{\beta}_W$ , nous allons appliquer des théorèmes de convergence "au niveau des classes". Il est donc nécessaire d'introduire des notations qui nous permettent de regrouper pour une même classe les différentes observations au niveau élève.

On note donc  $\tilde{\mathbf{y}}_j$ ,  $\tilde{\mathbf{X}}_{-0j}$  et  $\tilde{\epsilon}_j$ , les vecteurs obtenus en "empilant" les observations de la classe  $j$ . On note  $W_j$ , l'opérateur de comparaison à la moyenne au sein de la classe  $j$ . Son expression

matricielle est  $I_{n_j} - \frac{1}{n_j} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix}$ .

### Hypothèse sur la structure de dépendance des observations

*Hypothèse ( $H'_{as}$ ) :*

◊ Les nombres d'élèves observés par classe  $n_j$  sont indépendants et identiquement distribués. On suppose de plus que  $1 \leq n_1 \leq n_{sup} < +\infty$  presque sûrement et que  $P(n_1 = 1) < 1$ . Dans ces conditions, on a  $1 < E(n_1) = \bar{n} < +\infty$ .

◊ Les variables  $(\tilde{\mathbf{y}}_j, \tilde{\mathbf{X}}_{-0j})_{j=1\dots C}$  sont indépendantes et identiquement distribuées conditionnellement à  $n_j$ .

◊  $E \left[ \tilde{\mathbf{X}}'_{-01} W_1 \tilde{\mathbf{X}}_{-01} \right]$  existe et est inversible.

◊  $E \left[ \left| \left( \tilde{\mathbf{X}}'_{-01} W_1 \tilde{\mathbf{X}}_{-01} \right)_{(k,l)} \right| \right] < +\infty$  pour tout  $0 \leq l, k \leq K$ .

### Normalité asymptotique des estimateurs

Sous les hypothèses ( $H'_1$ ), ( $H'_2$ ), ( $H'_3$ ) et ( $H'_{as}$ ), on a les propriétés suivantes :

◊ L'estimateur  $\hat{\beta}_W$  est convergent :

$$\hat{\beta}_W \xrightarrow[n \rightarrow +\infty]{p} \begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$$

◊ L'estimateur "Within" est asymptotiquement normal :

$$\sqrt{n-C} \left( \hat{\beta}_W - \begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} \right) \xrightarrow[n \rightarrow +\infty]{loi} \mathcal{N} \left( 0, \sigma^2 \begin{bmatrix} 0 & & 0 \\ 0 & (\bar{n}-1)E \left[ \tilde{\mathbf{X}}'_{-01} W_1 \tilde{\mathbf{X}}_{-01} \right]^{-1} & \end{bmatrix} \right)$$

◊ L'estimateur  $\hat{\sigma}^2$  est convergent :

$$\hat{\sigma}^2 \xrightarrow[n \rightarrow +\infty]{p} \sigma^2$$

◊ On peut donc construire un estimateur convergent de la variance asymptotique de  $\sqrt{n-C} \left( \hat{\beta}_W - \begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} \right)$  :

$$(n-C)\hat{\sigma}^2 \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{X}'_{-0} W \mathbf{X}_{-0} \end{pmatrix}^{-1} \xrightarrow[n \rightarrow +\infty]{p} \sigma^2 \begin{bmatrix} 0 & 0 \\ 0 & (\bar{n}-1)E \left[ \tilde{\mathbf{X}}'_{-01} W_1 \tilde{\mathbf{X}}_{-01} \right]^{-1} \end{bmatrix}$$

◊ Et enfin :

$$\frac{1}{\sqrt{\hat{\sigma}^2}} \begin{pmatrix} 0 & 0 \\ 0 & (\mathbf{X}'_{-0} W \mathbf{X}_{-0})^{\frac{1}{2}} \end{pmatrix} \left( \hat{\beta}_W - \begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} \right) \xrightarrow[n \rightarrow +\infty]{loi} \mathcal{N} \left( 0, \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{I}_K \end{pmatrix} \right)$$

Preuve :

□

Concernant la convergence de  $\widehat{\beta}_W$ , il suffit de montrer que  $\frac{1}{n-C}\mathbf{X}'_{-0}W\mathbf{X}_{-0}$  et  $\frac{1}{n-C}\mathbf{X}'_{-0}W\epsilon$  tendent (respectivement) en probabilité vers une matrice inversible  $\Sigma$  et 0.

$\frac{1}{n-C}\mathbf{X}'_{-0}W\mathbf{X}_{-0} = \frac{C}{n-C} \left( \frac{1}{C} \sum_{j=1}^C \tilde{\mathbf{X}}'_{-0j} W_j \tilde{\mathbf{X}}_{-0j} \right)$ , la loi des grands nombres implique que  $\frac{1}{C} \sum_{j=1}^C \tilde{\mathbf{X}}'_{-0j} W_j \tilde{\mathbf{X}}_{-0j} \xrightarrow{p} E \left[ \tilde{\mathbf{X}}'_{-01} W_1 \tilde{\mathbf{X}}_{-01} \right]$  et  $\frac{n}{C} \xrightarrow{p} \bar{n}$  donc que  $\frac{1}{n-C}\mathbf{X}'_{-0}W\mathbf{X}_{-0} \xrightarrow{p} \frac{1}{\bar{n}-1} E \left[ \tilde{\mathbf{X}}'_{-01} W_1 \tilde{\mathbf{X}}_{-01} \right] = \Sigma$ .

De même  $\frac{1}{n-C}\mathbf{X}'_{-0}W\epsilon \xrightarrow{p} \frac{1}{\bar{n}-1} E \left[ \tilde{\mathbf{X}}'_{-01} W_1 \tilde{\epsilon}_1 \right] = 0$  car  $E[\epsilon|\mathbf{X}] = 0$ .

La normalité asymptotique de  $\widehat{\beta}_W$  vient du fait que d'après le théorème de la limite centrée :

$$\frac{1}{\sqrt{n-C}}\mathbf{X}'_{-0}W\epsilon \xrightarrow{loi} N(0, \sigma^2\Sigma) \text{ on en déduit que } \sqrt{n-C} \left( \beta_W - \begin{pmatrix} 0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} \right) \xrightarrow{loi} N(0, \sigma^2\Sigma^{-1}\Sigma\Sigma^{-1}).$$

La convergence de  $\widehat{\sigma}^2$  s'obtient à partir de :

$$\widehat{\sigma}^2 = \frac{n-C}{n-K-C} \left[ \frac{1}{n-C}\epsilon'W\epsilon - \frac{1}{n-C}\epsilon'W\mathbf{X}_{-0} (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0}W\epsilon \right]$$

Or  $\frac{1}{n-C}\epsilon'W\epsilon = \frac{C}{n-C} \frac{1}{C} \sum_{j=1}^C \tilde{\epsilon}'_j W_j \tilde{\epsilon}_j \xrightarrow{p} \frac{1}{\bar{n}-1} E(\tilde{\epsilon}'_1 W_1 \tilde{\epsilon}_1)$

et  $\frac{1}{n-C}\epsilon'W\mathbf{X}_{-0} (\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1} \mathbf{X}'_{-0}W\epsilon = \frac{1}{n-C}\epsilon'W\mathbf{X}_{-0} \left( \frac{1}{n-C}\mathbf{X}'_{-0}W\mathbf{X}_{-0} \right)^{-1} \frac{1}{n-C}\mathbf{X}'_{-0}W\epsilon \xrightarrow{p} 0$ .

Pour conclure, il suffit de noter que  $\frac{n-C}{n-K-C} \rightarrow 1$  et que  $E[\tilde{\epsilon}'_1 W_1 \tilde{\epsilon}_1] = E[Tr(\tilde{\epsilon}'_1 W_1 \tilde{\epsilon}_1)] = E[Tr(\tilde{\epsilon}_1 \tilde{\epsilon}'_1 W_1)] = \sigma^2 E[Tr(W_1)] = \sigma^2 E[n_1 - 1] = \sigma^2(\bar{n} - 1)$ .

On en déduit que  $\widehat{\sigma}^2 \left( \frac{1}{n-C}\mathbf{X}'_{-0}W\mathbf{X}_{-0} \right)^{-1} \xrightarrow{p} \sigma^2\Sigma^{-1}$ .

Ce qui permet de conclure.

■

### 3.4.6 Quelle est la loi à distance finie de l'estimateur si les résidus sont normaux ?

L'estimateur des moindres carrés contraint par la condition  $\beta_0 = 0$  coïncide avec l'estimateur des MCO du modèle sans constante. Il hérite donc de toutes ses propriétés à distance finie qui ont été développées dans le chapitre 2.

Comme dans les rappels du chapitre 2, on va donc se placer sous l'hypothèse de normalité des résidus pour caractériser les propriétés de l'estimateur à distance finie.

#### Hypothèse de normalité des résidus

*Hypothèse ( $H'_n$ ) : Les résidus sont gaussiens, de moyenne nulle, identiquement distribués et indépendants des  $\mathbf{X}$  et des  $\mathbf{Z}$*

Sous cette hypothèse, la loi des estimateurs peut facilement être caractérisée conditionnellement aux covariables. On peut dans ce cas construire des tests à distance finie.

#### Propriétés à distance finie des estimateurs sous l'hypothèse de normalité des résidus

Sous les hypothèses ( $H'_1$ ) et ( $H'_n$ ), on a les propriétés suivantes :

Soit  $\begin{pmatrix} \widehat{\beta}_{MCC} \\ \widehat{\gamma}_{MCC} \end{pmatrix} = \text{Argmin} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma)'(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma)$  et

$$\text{s.c. } R' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = 0$$

$$\widehat{\sigma}^2_{MCC} = \frac{1}{n-K-C-1+\dim(\text{Im}(R'))} (\mathbf{y} - \mathbf{X}\widehat{\beta}_{MCC} - \mathbf{Z}\widehat{\gamma}_{MCC})'(\mathbf{y} - \mathbf{X}\widehat{\beta}_{MCC} - \mathbf{Z}\widehat{\gamma}_{MCC})$$

Sous les hypothèses  $(H'_1)$  et  $(H'_n)$ , on peut caractériser la loi des estimateurs conditionnellement aux covariables :

$$\diamond \begin{pmatrix} \hat{\beta}_{MCC} \\ \hat{\gamma}_{MCC} \end{pmatrix} | \mathbf{X}, \mathbf{Z} \rightsquigarrow \mathcal{N} \left( \begin{pmatrix} \beta \\ \gamma \end{pmatrix}, \sigma^2 V \right) \text{ avec}$$

$$V = \begin{pmatrix} 1 & & (0) & 0 \\ & \ddots & & \vdots \\ (0) & & 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{A}'\mathbf{A} & R \\ R' & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{A}'\mathbf{A} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{A}'\mathbf{A} & R \\ R' & 0 \end{pmatrix}^{-1} \begin{pmatrix} 1 & & (0) \\ & \ddots & \\ (0) & & 1 \\ 0 & \dots & 0 \end{pmatrix}$$

Conditionnellement aux covariables, l'estimateur des moindres carrés contraints est normalement distribué, de moyenne  $\begin{pmatrix} \beta \\ \gamma \end{pmatrix}$  et de variance  $\sigma^2 V$ .

$$\diamond \frac{n-(K+C)}{\sigma^2} \widehat{\sigma}_{MCC}^2 | \mathbf{X}, \mathbf{Z} \rightsquigarrow \chi_{n-(K+C)}^2$$

À une constante de normalisation (inconnue) près, l'estimateur  $\widehat{\sigma}_{MCC}^2$  est distribué comme une loi du  $\chi_{n-(K+C)}^2$ .

$\diamond \begin{pmatrix} \hat{\beta}_{MCC} \\ \hat{\gamma}_{MCC} \end{pmatrix} \perp \widehat{\sigma}_{MCC}^2 | \mathbf{X}, \mathbf{Z}$  Les lois des estimateurs  $\begin{pmatrix} \hat{\beta}_{MCC} \\ \hat{\gamma}_{MCC} \end{pmatrix}$  et  $\widehat{\sigma}_{MCC}^2$  sont indépendantes conditionnellement à  $\mathbf{X}$  et  $\mathbf{Z}$ .

$\diamond \frac{\hat{\beta}_k - \beta_k}{\sqrt{\widehat{\sigma}_{MCC}^2 V_{(k,k)}}} | \mathbf{X}, \mathbf{Z} \rightsquigarrow \text{Student}(n - (K + C))$  sauf pour  $k = 0$  dans le cas où on a choisi la contrainte  $\beta_0 = 0$ .

La loi de l'estimateur de la  $k$ ième composante du vecteur  $\beta$  centré sur la valeur du paramètre et normalisé par l'estimation de son écart-type suit une loi de Student à  $(n - (K + C))$  degrés de liberté.

$\diamond \frac{\hat{\gamma}_j - \gamma_j}{\sqrt{\widehat{\sigma}_{MCC}^2 V_{(K+1+j, K+1+j)}}} | \mathbf{X}, \mathbf{Z} \rightsquigarrow \text{Student}(n - (K + C))$  sauf pour  $j_0 = 0$  dans le cas où on a choisi la contrainte  $\gamma_{j_0} = 0$ .

La loi de l'estimateur de la  $j$ ième composante du vecteur  $\gamma$  centré sur la valeur du paramètre et normalisé par l'estimation de son écart-type suit une loi de Student à  $(n - (K + C))$  degrés de liberté.

$\diamond$  L'estimateur des moindres carrés contraints  $(\hat{\beta}_{MCC}, \hat{\gamma}_{MCC}, \widehat{\sigma}_{MCC}^2)$  est optimal parmi les estimateurs sans biais de  $(\beta, \gamma, \sigma^2)$ , c'est-à-dire que pour tout estimateur  $(\tilde{\beta}, \tilde{\gamma}, \tilde{\sigma}^2)$  tel que pour tout  $(\beta, \gamma, \sigma^2) \in \mathbb{R}^{K+1+C} \times \mathbb{R}^+*$ ,  $E[(\tilde{\beta}, \tilde{\gamma}, \tilde{\sigma}^2)] = (\beta, \gamma, \sigma^2)$  et  $R' \begin{pmatrix} \tilde{\beta} \\ \tilde{\gamma} \end{pmatrix} = 0$  on a :

$$V \begin{bmatrix} \hat{\beta}_{MCC} \\ \hat{\gamma}_{MCC} \\ \widehat{\sigma}_{MCC}^2 \end{bmatrix} \ll V \begin{bmatrix} \tilde{\beta} \\ \tilde{\gamma} \\ \tilde{\sigma}^2 \end{bmatrix}$$

### Remarque :

On a donné une expression très générale de la matrice  $V$ . En pratique, la covariance entre  $\hat{\beta}$  et  $\hat{\gamma}$  est rarement un objet d'intérêt, on utilise donc les expressions calculées en 3.4.3 pour calculer les termes  $V_{(k,k)}$  et  $V_{(K+1+j, K+1+j)}$  afin de calculer les statistiques de Student. Le calcul pratique est détaillé en 3.4.7.

Preuve :

□

Le résultat est une conséquence directe des propriétés de l'estimateurs de moindres carrés dans le cas où on fixe comme contrainte  $\beta_0 = 0$  : l'estimateur des moindres carrés contraints coïncide alors avec l'estimateur des MCO du modèle sans constante. Le même raisonnement vaut dans le cas où la contrainte est  $\gamma_{j_0} = 0$ . Dans le cas où la contrainte est  $\sum_{j=1}^C \gamma_j = 0$ , il suffit de remarquer que l'estimateur de  $\sigma$  ne dépend pas de la contrainte retenue et que les estimateurs de  $\beta$  et  $\gamma$  sont des combinaisons linéaires des estimateurs obtenus sous les autres contraintes. Le fait qu'une transformation linéaire d'un vecteur gaussien soit un vecteur gaussien permet d'achever la preuve.

■

### 3.4.7 Estimation de l'intervalle de confiance des estimateurs dans l'estimation en deux étapes

Si la décomposition de Frisch-Waugh est très utile en pratique pour calculer les estimateurs de  $\beta$  et de  $\gamma$ , il est en revanche nécessaire de "travailler" un peu plus pour récupérer les statistiques de tests. Concernant les écarts-types de l'estimateur  $\widehat{\beta}_W$ , il faut corriger ceux calculés usuellement lors de la régression "Within" pour tenir compte du fait qu'en raisonnant en écart par rapport à la moyenne de classe, on a éliminé  $C$  covariables. Concernant les écarts-types de  $\widehat{\gamma}_W$ , il faut les recalculer intégralement.

#### Test de nullité des $\widehat{\beta}_W$

Après avoir calculé pour chaque individu, l'écart de la variable  $y$  et des covariables  $X$  à la moyenne de sa classe, les estimations de  $\beta$  s'obtiennent par une régression linéaire. L'estimation de  $\sigma^2$  est calculée en divisant la somme des carrés des résidus par la quantité  $n - (K + 1)$  comme dans le cas standard (cf. chapitre 2). En réalité, pour le modèle que l'on estime, ce nombre devrait être  $n - (K + C)$ . Il faut donc corriger l'estimation de  $\sigma^2$  par un facteur  $\frac{n-(K+1)}{n-(K+C)}$ . De même, la statistique de Student doit être corrigée d'un facteur  $\frac{\sqrt{n-(K+C)}}{\sqrt{n-(K+1)}}$ . Enfin le calcul de la p-value doit également intégrer le bon nombre de degrés de liberté :  $n - (K + C)$ . Toutes ces opérations sont assez facilement exécutables en SAS (cf. 3.6.1), dans la mesure où les résultats de la régression peuvent être récupérés dans une table.

#### Test de nullité des $\widehat{\gamma}_W$

Les estimations de  $\gamma$  s'obtiennent facilement comme moyenne au sein de la classe des quantités  $\mathbf{y} - \mathbf{X}\widehat{\beta}$ . Les statistiques de test sont un peu plus délicates à récupérer : on récupère les quantités  $\widehat{\sigma}^2$  et  $\mathbf{X}'_{-0}W\mathbf{X}_{-0}$  à l'issue de la régression qui permet d'estimer  $\beta$ . De même la quantité  $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}_{-0}$  est facilement calculable car il s'agit des moyennes par classe des variables  $X$ . Enfin la quantité  $(\mathbf{Z}'\mathbf{Z})^{-1}$  est elle-même facilement calculable. Pour calculer la statistique de Student, il est nécessaire de calculer la diagonale de la matrice  $\widehat{\sigma}^2 \left[ (\mathbf{Z}'\mathbf{Z})^{-1} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}_{-0}(\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}\mathbf{X}'_{-0}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \right]$  ; or, cette matrice peut rapidement atteindre une taille trop importante pour être calculée à partir des ressources informatiques habituellement mises à disposition (par exemple sur un fichier administratif d'examen comme le brevet ou le bac, le nombre de classes atteint quelques dizaines de milliers ce qui conduit à une matrice de  $10^8$  éléments environ). Mais une programmation économe qui s'applique à ne pas calculer les termes non diagonaux de cette matrice permet d'obtenir des estimations de manière immédiate (on divise le nombre de paramètres à calculer et à stocker en mémoire par  $C$ ). La programmation que nous avons retenue nécessite néanmoins de pouvoir disposer d'un langage de programmation de calcul matriciel (module IML de SAS par exemple). Une fois calculée la variance de chaque  $\widehat{\gamma}_j$ , les statistiques de Student et les p-values correspondantes se calculent très facilement.

#### Test de nullité des paramètres dans le cas des contraintes identifiantes $\gamma_{j_0} = 0$ et $\sum_{j=1}^C \gamma_j = 0$

Les tests de Student peuvent également être calculés dans le cas où on impose comme contrainte  $\gamma_{j_0} = 0$  ou encore  $\sum_{j=1}^C \gamma_j = 0$ . Là encore, la seule précaution à prendre consiste à éviter autant que possible de manipuler de trop grosses matrices : une programmation qui se restreint à des matrices de taille  $K \times C$  nous a permis de ne pas rencontrer de difficultés informatiques (y compris sur de "gros" fichiers).

## 3.5 La prise en compte des variables constantes au sein des classes

On a jusqu'à maintenant supposé que parmi les variables  $X$ , il n'y avait pas de variables ayant la même valeur pour tous les élèves d'une même classe.

Supposons maintenant que l'output  $y_{ij}$  de l'élève  $i$  dans la classe  $j$  dépend de variables "individuelles"  $X_{ij}^1$  qui varient d'un élève à l'autre au sein d'une même classe mais également de variables

$X_j^2$  communes aux élèves d'une même classe, et enfin d'un effet spécifique de la classe  $\delta_j$  qui reflète l'influence de l'ensemble des variables communes aux élèves de la classe  $j$  non prises en compte par  $X^2$  ainsi que d'un facteur  $\varepsilon_{ij}$  spécifique à l'élève au sein de sa classe qui reflète l'influence des variables "individuelles" non prises en compte par le modèle.

$$y_{ij} = X_{ij}^1\beta_1 + X_j^2\beta_2 + \delta_j + \varepsilon_{ij}$$

Sous l'hypothèse d'exogénéité ( $H'_2$ ), le paramètre  $\beta_1$  peut être estimé de manière convergente. On peut également estimer sans biais (mais pas de manière convergente) le paramètre  $\gamma_j = X_j^2\beta_2 + \delta_j$ . En revanche, on ne peut estimer simplement  $\beta_2$  sans hypothèses supplémentaires contraignantes sur la distribution des variables inobservées ( $\delta_j$ ) et la distribution des variables observées au niveau de la classe ( $X_j^2$ ). En effet, si les variables inobservées induisant des différences de réussite scolaire entre les différentes classes sont corrélées aux observées, alors la régression linéaire de  $\hat{\gamma}$  sur  $X^2$  donnera des estimations biaisées de  $\beta_2$ .

Par exemple, imaginons qu'une faible taille de classe a un effet positif sur la réussite scolaire ( $\beta_2 > 0$ ) mais pour cette raison, les élèves scolarisés dans de telles classes sont souvent ceux que le directeur d'école identifie comme présentant des risques d'échec élevés ( $\delta_j < 0$ ). En effectuant, la régression linéaire de  $\hat{\gamma}$  sur  $X^2$ , on ne peut obtenir des estimations sans biais de  $\beta_2$  que si une hypothèse d'exogénéité au "niveau classe" est vérifiée ( $E(\delta|X^2) = 0$  ou au moins  $Cov(\delta, X^2) = 0$ ). Si on garde cette limite en tête, effectuer une régression linéaire de  $\hat{\gamma}$  sur  $X^2$  peut néanmoins être intéressant d'un point de vue descriptif pour relier le degré de réussite des classes et certaines variables observables  $X^2$ .

Lors de la mise en oeuvre d'une telle régression, la précision de l'estimation  $\hat{\gamma}_j$  varie d'une classe à l'autre (elle dépend du nombre d'élèves dans la classe et de la distribution des variables  $X^1$  dans les classes). Il est alors judicieux d'utiliser une matrice de variance-covariance robuste à l'hétéroscédasticité (matrice de White) pour obtenir des intervalles de confiance qui ne soient pas basés sur l'hypothèse que chaque  $\hat{\gamma}_j$  apporte la même quantité d'information. Par exemple, dans la proc `reg` de SAS, on utilisera l'option `ACOV`.

## 3.6 Programmation sous SAS...

### 3.6.1 ... par l'estimateur "Within"

Voici un exemple de code qui permet de calculer l'estimateur  $\hat{\beta}_W$ . La table de départ est nommée `TABLE` la variable expliquée `Y`, les covariables `X1`, `X2`, ..., `XK` et la variable `Z` est une variable correspondant à un identifiant de classe.

```
proc summary data=TABLE;
  var Y X1 X2 ... XK;
  class Z;
  output out=between(where=(STAT='MEAN' and TYPE=1) rename = (FREQ = nb));
  /*<====nb est le nombre d'élèves par classe*/
run;

proc sort data=TABLE out=within;
  by Z;
run;

proc sql;
  create table within as
    select t1.X1 - t2.X1 as w_X1,
           t1.X2 - t2.X2 as w_X2,
           ...
           t1.XK - t2.XK as w_XK,
           t1.Y-t2.Y as w_Y
    from within as t1, between as t2
    where t1.Z = t2.Z;
quit;

title 'Les effets intra-classe calculés avec le within (mauvais E-C)';run;
```

```

proc reg data=within outest=temp_param edf outseb covout;          /*<==== Options qui
    permettent de récupérer des valeurs nécessaires pour la suite*/
    model w_Y = w_X1 w_X2 ... w_XK /noint;
run;
    quit;

```

Pour faire de l'inférence sur  $\hat{\beta}_W$ , il faut corriger les écarts-types, les statistiques de Student et les p-values. Cette correction dépend du nombre de classes distinctes `NB_Z` (c'est le nombre d'observations de la table que nous avons nommée `between` ci-dessus) et du nombre de degrés de liberté `NB_DEG_L` utilisés dans la régression Within (c'est le nombre `_EDF_` dans la table que nous avons nommée `temp_param` ci-dessus).

```

data temp_param2;set temp_param (where=( _TYPE_ in ( 'PARMS', 'SEB' )));
    if _type_='SEB' then do;
        w_X1=sqrt(NB_DEG_L)/sqrt(NB_DEG_L-NB_Z)*w_X1;
        w_X2=sqrt(NB_DEG_L)/sqrt(NB_DEG_L-NB_Z)*w_X2;
        ...
        w_XK=sqrt(NB_DEG_L)/sqrt(NB_DEG_L-NB_Z)*w_XK;
    end;
run;

proc transpose data=temp_param2(keep=w_X1 w_X2—w_XK) out=temp_param3;run;

data paramesti(rename=( _NAME_=Variable COL1=parameter_estimate Col2=standard_error
    ));
    set temp_param3(keep=_NAME_ COL1 COL2 where=(substr( _NAME_,1,1)='w' and _NAME_
        ne "w_Y" ));
    tvalue=col1/col2;
    p_value=2*(1-probt(abs(tvalue),NB_DEG_L-NB_Z));
run;

title 'Les effets intra-classe calculés avec le within (bons E-C)';run;
proc print data=paramesti;run;

```

Le calcul des  $\hat{\gamma}$  et de leurs écarts-types associés peut ensuite s'effectuer selon la contrainte identifiante choisie. Dans tous les cas, il faut calculer les  $Y - X\hat{\beta}_W$ .

```

\%macro calcul;
    data _null_;
        set temp_param(where=( _TYPE_='PARMS' ));
        call symput("beta_X1",w_X1);
        call symput("beta_X2",w_X2);
        ...
        call symput("beta_XK",w_XK);
    run;

    data _TABLE_2;
        set TABLE;
        Y_Xbeta=Y-&beta_X1.*X1-&beta_X2.*X2-...-&beta_XK.*XK;
    run;
\%mend;

\%calcul;

```

1. Sous l'hypothèse identifiante  $\beta_0 = 0$  :

Pour calculer les estimateurs  $\hat{\gamma}$ , il suffit alors de faire la moyenne par classe `Z` des  $Y - X\hat{\beta}_W$  calculés dans la table `_TABLE_2`.

```

proc summary data=_TABLE_2 nway;
    var Y_Xbeta;
    class Z;
    output out=_effet_fixe(where=( _STAT_='MEAN' ));
run;

```

Pour calculer la précision des estimateurs  $\hat{\gamma}$ . Il faut utiliser la matrice de variance-covariance de  $\hat{\beta}_W$  (présente dans la table `temp_param`), les moyennes des `X` par classe et le nombre

d'élèves par classe (informations présentes dans la table `between`). Il faut également calculer une estimation de  $\hat{\sigma}^2$ , pour cela on corrige le "root means square error" (i.e. la moyenne quadratique des résidus estimés) dans la table `temp_param`, qu'on élève au carré et qu'on corrige par le bon nombre de degrés de liberté.

```
data mat1(keep=w_X1 w_X2 ... w_XK);
    set temp_param(where=(TYPE="COV" ));
run;

data mat2;
    set between(keep=X1 X2 ... XK);
run;

data mat3;
    set between(keep=nb);
run;

data sigma2(keep=sigma2);
    set temp_param(keep=TYPE_RMSE where=(TYPE="PARMS"));
    sigma2=_RMSE*_RMSE*_NB_DEG_L/(NB_DEG_L-NB_Z);
run;
```

Grâce à un peu de calcul matriciel (`proc IML`), on calcule la variance de  $\hat{\gamma}$ , puis on en déduit la valeur des écarts-types, statistiques de Student et p-values associés.

```
proc iml;
    use mat1;
        read all into mat1;
    use mat2;
        read all into mat2;
    use mat3;
        read all into mat3;
    use sigma2;
        read all into sigma2;
    diagvargw=sigma2*1/mat3+NB_DEG_L*1/(NB_DEG_L-NB_Z)*((mat2*mat1)#mat2)[, +];
    create diagvargw from diagvargw;append from diagvargw;
    close;
quit;run;

data gamma(drop=TYPE_FREQ_STAT);
    set _effet_fixe(rename=(Y_Xbeta=parameter_estimate));
    set diagvargw(rename=(COL1=variance));
    standard_error=sqrt(variance);
    tvalue=parameter_estimate/standard_error;
    p_value=2*(1-probt(abs(tvalue),NB_DEG_L-NB_Z));
run;

title "Les effets fixes (contante nulle)";run;
proc print data=gamma;
run;
```

## 2. Sous l'hypothèse identifiante $\sum \gamma = 0$ :

Cette fois on centre la moyenne par classe des  $Y - X\hat{\beta}_W$ , pour obtenir les estimateurs de  $\hat{\gamma}$ . L'estimateur de  $\beta_0$  est alors la moyenne des moyennes par classe des  $Y - X\hat{\beta}_W$ .

```
proc summary data=_TABLE_2 nway;
    var Y_Xbeta;
    class Z;
    output out=_effet_fixe(where=(STAT="MEAN" ));
run;

proc standard data=_effet_fixe out=_effet_fixe_centre mean=0;run;

proc summary data=_effet_fixe nway;
    var Y_Xbeta;
    output out=_effet_fixe_cst(where=(STAT="MEAN" ));
run;
```

Le calcul de la précision de  $\hat{\gamma}$  et de  $\hat{\beta}_0$  fait intervenir la matrice de variance-covariance de  $\hat{\beta}_W$ , les moyennes de  $X$  par classe ainsi que le nombre d'élèves par classe, il faut également calculer l'estimateur  $\hat{\sigma}^2$  à partir des sorties de la régression de  $WY$  sur  $WX$ .

```

data mat1(keep=w_X1 w_X2 ... w_XK);
  set temp_param(where=(TYPE="COV"));
run;

data mat2;
  set between(keep=X1 X2 ... XK);
run;

data mat3;
  set between(keep=nb);
run;

data sigma2(keep=sigma2);
  set temp_param(keep=TYPE_RMSE_where=(TYPE="PARAMS"));
  sigma2=_RMSE*_RMSE*_NB_DEG_L/(NB_DEG_L-NB_Z);
run;

proc iml;
  use mat1;
  read all into mat1;
  use mat2;
  read all into mat2;
  use mat3;
  read all into mat3;
  use sigma2;
  read all into sigma2;
  mat4=(I(nrow(mat2))-1/nrow(mat2))*J(nrow(mat2),nrow(mat2),1);
  diagvargamma=
    sigma2*((1-2/(nrow(mat2)))*1/mat3+1/(nrow(mat2))**2*((1/mat3)[+,+])*J(
      nrow(mat2),1,1))
    +NB_DEG_L*1/(NB_DEG_L-NB_Z)*((mat4*mat2*mat1)#(mat4*mat2))[+,+];
  varbeta0=sigma2*1/(nrow(mat2))**2*(1/mat3)[+,+]+
    1/(nrow(mat2))**2*mat2[+,]*mat1*((mat2[+,])');
  create diagvargamma from diagvargamma;append from diagvargamma;
  create varbeta0 from varbeta0;append from varbeta0;
  close;
  quit;
run;

data gamma(drop=_TYPE_FREQ_STAT);
  set _effet_fixe_centre(rename=(Y_Xbeta=parameter_estimate));
  set diagvargamma(rename=(COL1=variance));
  standard_error=sqrt(variance);
  tvalue=parameter_estimate/standard_error;
  p_value=2*(1-probt(abs(tvalue),NB_DEG_L-NB_Z));
run;

data beta0;
  set _effet_fixe_cst(rename=(Y_Xbeta=parameter_estimate_STAT=Variable));
  drop=_TYPE_FREQ;
  set varbeta0(rename=(COL1=variance));
  variable="CST";
  standard_error=sqrt(variance);
  tvalue=parameter_estimate/standard_error;
  p_value=2*(1-probt(abs(tvalue),NB_DEG_L-NB_Z));
run;

title "Les effets fixes centrés (somme des effets fixes nulle)";run;
proc print data=gamma;run;

title "La constante (somme des effets fixes nulle)";run;
proc print data=beta0;run;

```

### 3. Sous l'hypothèse identifiante $\gamma_{j0} = 0$ :

Cette fois on normalise la moyenne par classe des  $Y - X\hat{\beta}_W$  par la moyenne calculée sur la

classe  $j_0$ , pour obtenir les estimateurs de  $\hat{\gamma}$ . L'estimateur de  $\beta_0$  est la moyenne des  $Y - X\hat{\beta}_w$  sur la classe  $j_0$ .

```

data _effet_fixe;
    set _effet_fixe;
    indice=_n_;
run;

data ref;
    set _effet_fixe (where=(Z="J0"));
run;

data gammaref;
    set ref (keep=Y_Xbeta);
run;

data indiceref;
    set ref (keep=indice);
run;

%macro normalisation;
    data _null_;
        set gammaref;
        call symput ('gammaref',Y_Xbeta);
    run;

    data _effet_fixe_j0;
        set _effet_fixe;
        Y_Xbeta=Y_Xbeta-&gammaref.;
    run;
%mend;

%normalisation;

data mat1(keep=w_X1 w_X2 ... w_XK);
    set temp_param(where=( _TYPE_="COV" ) );
run;

data mat2;
    set between(keep=X1 X2 ... XK);
run;

data mat3;
    set between(keep=nb);
run;

data sigma2(keep=sigma2);
    set temp_param(keep= TYPE_RMSE where=( TYPE_="PARMS"));
    sigma2=_RMSE*_RMSE*_NB_DEG_L/(NB_DEG_L-NB_Z);
run;

proc iml;
    use mat1;
    read all into mat1;
    use mat2;
    read all into mat2;
    use mat3;
    read all into mat3;
    use sigma2;
    read all into sigma2;
    use indiceref;
    read all into indiceref;
    use gammaref;
    read all into gammaref;
    if indiceref=1 then do;
        mat4=I(nrow(mat2))-
            (J(nrow(mat2),1,1) || J(nrow(mat2),nrow(mat2)-1,0));
        diagvargamma=
            sigma2*(1/mat3+1/mat3[indiceref,1]*(-1//J(nrow(mat2)-1,1,1)))
            +NB_DEG_L*1/(NB_DEG_L-NB_Z)*((mat4*mat2*mat1)#(mat4*mat2))
    end;

```

```

varbeta0=
    [,+];
    (sigma2*1/mat3+NB_DEG_L*1/(NB_DEG_L-NB_Z)*((mat2*mat1)#mat2)
    [,+]) [indiceref,1];
end;
if indiceref=NB_Z then do;
mat4=I(nrow(mat2))-
    (J(nrow(mat2),nrow(mat2)-1,0)||J(nrow(mat2),1,1));
diagvargamma=
    sigma2*(1/mat3+1/mat3[indiceref,1]*(J(nrow(mat2)-1,1,1)//-1))
    +NB_DEG_L*1/(NB_DEG_L-NB_Z)*((mat4*mat2*mat1)#(mat4*mat2))
    [,+];
varbeta0=
    (sigma2*1/mat3+NB_DEG_L*1/(NB_DEG_L-NB_Z)*((mat2*mat1)#mat2)
    [,+]) [indiceref,1];
end;
if indiceref<NB_Z and indiceref>1 then do;
mat4=I(nrow(mat2))-
    (
    J(nrow(mat2),indiceref-1,0)
    ||J(nrow(mat2),1,1)
    ||J(nrow(mat2),nrow(mat2)-indiceref,0)
    );
diagvargamma=
    sigma2*(1/mat3+1/mat3[indiceref,1]*(J(indiceref-1,1,1)//-1//J(
    nrow(mat2)-indiceref,1,1)))
    +NB_DEG_L*1/(NB_DEG_L-NB_Z)*((mat4*mat2*mat1)#(mat4*mat2))
    [,+];
varbeta0=
    (sigma2*1/mat3+NB_DEG_L*1/(NB_DEG_L-NB_Z)*((mat2*mat1)#mat2)
    [,+]) [indiceref,1];
end;

create diagvargamma from diagvargamma;append from diagvargamma;
create varbeta0 from varbeta0;append from varbeta0;
close;quit;
run;

data gamma(drop=TYPE FREQ STAT);
set _effet_fixe_j0 (rename=(Y_Xbeta=parameter_estimate) drop=indice);
set diagvargamma(rename=(COL1=variance));
standard_error=sqrt(variance);
if Z ne "J0" then do;
tvalue=parameter_estimate/standard_error;
p_value=2*(1-probt(abs(tvalue),NB_DEG_L-NB_Z));
end;
if Z="J0" then do;
tvalue=.;
p_value=.;
end;
run;

data beta0(drop=Z TYPE FREQ);
set ref(rename=(Y_Xbeta=parameter_estimate STAT=variable) drop=indice);
set varbeta0(rename=(COL1=variance));
variable="CST";
standard_error=sqrt(variance);
tvalue=parameter_estimate/standard_error;
p_value=2*(1-probt(abs(tvalue),NB_DEG_L-NB_Z));
run;

title "Les effets fixes (reference Z = J0)";run;
proc print data=gamma3;run;

title "La constante (reference Z = J0)";run;
proc print data=beta0;run;

```

### 3.6.2 ... ou directement grâce à la proc reg

Quand cela est possible, la manière la plus directe de programmer consiste à estimer en une étape l'ensemble des effets liés aux caractéristiques purement individuelles de l'élève et les effets "classe" par l'estimateur des MCO contraint. La principale difficulté est qu'il faut avoir dichotomisé les indicatrices de classes ( $Z$ ) dans la `proc reg`. Un moyen de faire cela, quand le nombre de classes est très élevé, est d'utiliser la `proc glmmod` et un peu de programmation de macro langage. On suppose dans tout ce qui suit que les variables qualitatives ( $X$ ) ont été dichotomisées préalablement.

```
proc glmmod data=TABLE outparm=param outdesign=design noprint;
  class Z; model Y=X1 X2 ... XK Z;
run;
```

1. Sous l'hypothèse identifiante  $\beta_0 = 0$  :

Si la dernière colonne de la table `design` est `colNNN`, il suffit de coder :

```
proc reg data=design outest=sortie;
  model Y=col2 - - colNNN/noint;
run;quit;
```

2. Sous l'hypothèse identifiante  $\sum \gamma_j = 0$  :

Si  $nnn = K + 2$  et si la dernière colonne de la table `design` est `colNNN`, il suffit de coder :

```
%macro regression_effet_fixe;
  proc reg data=design outest=sortie;
    model Y=col2 - - colNNN;
    restrict colnn %do i=%eval(nnn+1) %to NNN;+col&i. %end; ;
  run;quit;
%amend;

%regression_effet_fixe;
```

3. Sous l'hypothèse identifiante  $\gamma_{j0} = 0$  :

Si la colonne de la table `design` correspondant à  $j_0$  est `colkkk`, si la dernière colonne de la table `design` est `colNNN`, et si  $jjj = kkk - 1$  et  $lll = kkk + 1$ , il suffit de taper :

```
proc reg data=design outest=sortie;
  model Y=col2 - - coljjj collll - - colNNN;
run;quit;
```

ou encore :

```
proc reg data=design outest=sortie;
  model Y=col2 - - colNNN;
  restrict colkkk;
run;quit;
```

## Chapitre 4

# Le modèle à erreurs composées (autrement appelé "modèle multi-niveaux" ou modèle "mixte")

### 4.1 Le modèle théorique de base

Le modèle à erreurs composées diffère du modèle à effets fixes sur le fait que les effets "classe"  $\gamma$  sont supposés être "aléatoires" alors qu'ils sont "fixes" dans le modèle à effets fixes. En réalité, il nous semble que cette manière d'exprimer les choses génère beaucoup de confusion et tend à rendre les enjeux qui sous-tendent les choix de modèle difficilement compréhensibles.

Pour bien comprendre ce qui se cache derrière cette distinction entre effets aléatoires et effets fixes, commençons par raisonner sur le modèle linéaire de base présenté dans le chapitre 2. Les résidus représentent l'effet des variables inobservables sur  $\mathbf{y}$ . Pour pouvoir identifier les effets des variables  $\mathbf{X}$ , il est nécessaire de supposer que ces variables inobservables ne sont pas corrélées avec les variables  $\mathbf{X}$ . On présente souvent, probablement à des fins pédagogiques, le modèle linéaire en disant que les variables  $\mathbf{X}$  sont connues ou "fixes" et que les résidus sont aléatoires. En réalité ce qui importe (et probablement ce que la formulation précédente est censée traduire), c'est qu'à caractéristiques observables  $\mathbf{X}$  données, les effets liés aux caractéristiques inobservables soient d'espérance nulle (voir l'hypothèse  $(H_2)$ ). Le débat entre effets aléatoires et effets fixes revient à supposer que les effets "classe" sont ou non d'espérance nulle à  $\mathbf{X}$  donné. Dans le cas où ces effets sont d'espérance nulle conditionnellement à  $\mathbf{X}$ , il n'est pas nécessaire de les prendre en compte dans les régresseurs mais ils doivent être pris en compte dans la structure des résidus pour calculer correctement la précision des estimateurs.

Considérons donc le modèle suivant :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

où  $\mathbf{Z}$  est la matrice d'appartenance aux différentes classes et  $\mathbf{X}$  est l'ensemble des autres covariables (dans ce chapitre, et contrairement à l'hypothèse maintenue tout au long du chapitre 3, des variables constantes au sein des classes comme le secteur de scolarisation ou l'académie peuvent être présentes).

Si on considère que les effets "classe"  $\gamma_j$  sont d'espérance nulle conditionnellement aux variables  $\mathbf{X}$  et  $\mathbf{Z}$ , alors le modèle peut se réécrire :

$$\mathbf{y} = \mathbf{X}\beta + v$$

avec  $v = \mathbf{Z}\gamma + \epsilon$ ,  $E[v|\mathbf{X}, \mathbf{Z}] = 0$  et  $V[v|\mathbf{X}, \mathbf{Z}] = \sigma_\gamma^2 \mathbf{Z}'\mathbf{Z} + \sigma^2 I_n$ . Ainsi, pour un individu  $i$  de la classe  $j$ , on a  $E[v_{ij}|\mathbf{X}, \mathbf{Z}] = 0$  et  $V[v_{ij}|\mathbf{X}, \mathbf{Z}] = \sigma_\gamma^2 + \sigma^2$ , pour deux individus  $i$  et  $i'$  de la même classe  $j$ ,  $Cov[v_{ij}v_{i'j}|\mathbf{X}, \mathbf{Z}] = \sigma_\gamma^2$  et enfin pour deux individus  $i$  et  $i'$  de deux classes différentes  $j$  et  $j'$ ,  $Cov[v_{ij}v_{i'j'}|\mathbf{X}, \mathbf{Z}] = 0$ . Le terme d'effet aléatoire vient du fait que formellement les égalités précédentes sont vérifiées si on suppose que les  $\gamma_j$  sont des variables aléatoires i.i.d. d'espérance nulle et de variance  $\sigma_\gamma^2$  (conditionnellement à  $\mathbf{X}$  et  $\mathbf{Z}$ ). Encore une fois, la chose la plus importante

à retenir n'est pas le caractère aléatoire des coefficients  $\gamma$ , mais bien les hypothèses d'orthogonalité entre les covariables  $\mathbf{X}$  et les effets "classe". Remarquons également que cette hypothèse n'était pas nécessaire pour estimer le modèle à effets fixes.

On constate alors que si les hypothèses ( $H_1$ ) et ( $H_2$ ) du modèle linéaire du chapitre 2 sont toujours vérifiées, alors il n'en est pas de même pour l'hypothèse d'homoscédasticité ( $H_3$ ). La première partie de ce chapitre s'attache donc d'une part à décrire les propriétés des estimateurs présentés dans les chapitres précédents et leur défauts lorsqu'ils sont appliqués à ce modèle, et d'autre part à présenter les autres estimateurs classiquement utilisés pour estimer de tels modèles.

On note donc les différentes hypothèses faites jusqu'à présent de la manière suivante :

### Le modèle et les hypothèses de moments

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon = \mathbf{X}\beta + v$$

*Hypothèse ( $H_1'$ )* : La matrice  $\mathbf{X}$  est de plein rang colonne.

*Hypothèse ( $H_2'$ )* : L'espérance des effets "classe"  $\gamma$  est nulle conditionnellement aux covariables,  $E[\gamma|\mathbf{X}, \mathbf{Z}] = 0$  et l'espérance des résidus "individuels"  $\epsilon$  est nulle conditionnellement aux covariables et à l'effet "classe"  $E[\epsilon|\mathbf{X}, \mathbf{Z}, \gamma] = 0$ . Cette hypothèse implique en particulier que l'espérance des résidus  $v$  est nulle conditionnellement aux covariables,  $E[v|\mathbf{X}, \mathbf{Z}] = 0$

*Hypothèse ( $H_3'$ )* : La variance des résidus conditionnellement aux covariables est telle que :

$$V[\gamma|\mathbf{X}, \mathbf{Z}] = \sigma_\gamma^2 I_C$$

$$V[\epsilon|\mathbf{X}, \mathbf{Z}, \gamma] = \sigma^2 I_n$$

Cette hypothèse implique en particulier que la variance des résidus  $v$  est de la forme suivante :

$$V[v|\mathbf{X}, \mathbf{Z}] = \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 I_n$$

Remarque : La première hypothèse est courante, elle permet d'identifier séparément les différentes composantes de  $\beta$ . La seconde est également tout à fait conforme à la version présentée dans les rappels du chapitre 2, en particulier on suppose que les autres variables que  $X$  qui influent sur  $y$  ne sont pas corrélées à  $X$ . Dans le cas qui nous intéresse, où des élèves appartiennent à une même classe, on suppose donc que les inobservables communes au sein de la classe ne sont pas corrélées aux covariables  $\mathbf{X}$ . Enfin, la troisième hypothèse traduit le fait que pour deux individus distincts un certain nombre de variables inobservées (orthogonales aux  $\mathbf{X}$ ) sont partagées du fait qu'ils sont dans la même classe.

## 4.2 Interprétation des paramètres estimés

Dans le modèle à effets fixes présenté dans le chapitre 3, nous avons indiqué que les effets des variables constantes au sein de la classe (par exemple le secteur public ou privé, la taille de classe, le pourcentage d'élèves défavorisés dans la classe, etc.) ne pouvaient pas être identifiés et estimés de manière simple. Dans le modèle à effets aléatoires, les variables constantes au sein des classes peuvent être intégrées dans la régression et les coefficients associés peuvent être estimés. Néanmoins cela n'implique pas que ces coefficients aient une interprétation économique ou sociologique simple. Sous certaines conditions, les estimations de l'effet des variables individuelles (variables qui varient d'un élève à l'autre au sein d'une classe) peuvent également être biaisées. Nous allons maintenant discuter ce problème en détail.

On postule le processus générateur de données suivant pour un individu  $i$  dans la classe  $j$  :

$$y_{ij} = X_{ij}^1 \beta_1 + X_{ij}^1 \beta_2^1 + X_{ij}^2 \beta_2^2 + \gamma_j + \epsilon_{ij}$$

où  $y_{ij}$  est la variable d'output (par exemple une note),  $X_{ij}^1$  un premier groupe de covariables qui varient d'un élève à l'autre au sein de la classe (par exemple : l'origine sociale de l'élève, son sexe, son mois de naissance etc.),  $X_{.j}^1 = \frac{1}{n_j} \sum_{i \in j} X_{ij}^1$  est l'ensemble des moyennes par classe des covariables, enfin  $X_{.j}^2$  est un ensemble d'autres covariables qui ne varient pas d'un élève à l'autre au

sein de la classe (secteur public ou privé, établissement fréquenté, caractéristiques du professeur, etc.). La quantité  $\gamma_j$  représente un effet spécifique de la classe  $j$  (c'est-à-dire l'effet de l'ensemble des variables communes au sein de la classe autres que les  $X_j^1$  et  $X_j^2$ ), et la quantité  $\varepsilon_{ij}$  représente l'effet des variables inobservées (ou non prises en compte dans le modèle).

On suppose que dans cette formulation les paramètres  $\beta_1$ ,  $\beta_2^1$  et  $\beta_2^2$  représentent les coefficients structurels du modèle, c'est-à-dire l'impact causal des différentes covariables. Ces coefficients sont identifiables et peuvent être estimés sans biais par régression si  $E(\gamma_j + \varepsilon_{ij} | X_{ij}^1, X_j^2) = 0$  (cf. les rappels du chapitre 2). Il faut donc qu'en plus de la condition classique d'exogénéité  $E(\varepsilon_{ij} | X_{ij}^1, X_j^2) = 0$ , la distribution des effets spécifiques à la classe vérifie  $E(\gamma_j | X_{ij}^1, X_j^2) = 0$ . Cela signifie en particulier que les variables qui jouent sur le niveau des élèves d'une classe et qui n'ont pas été intégrées dans le modèle ne sont pas corrélées aux variables  $X_j^2$  et  $X_j^1$ .

#### 4.2.1 Le cas des variables constantes au sein de la classe

Or, il semble assez illusoire d'espérer intégrer dans le modèle l'ensemble des variables corrélées simultanément au niveau des élèves de la classe et aux variables observables ( $X_j^1$  et  $X_j^2$ ). En effet, de nombreux facteurs enfouis dans les biographies des familles jouent simultanément sur la classe fréquentée par l'enfant et sa réussite scolaire. Par exemple, les déterminants du lieu d'habitation ou de scolarisation (difficilement observables) sont corrélés avec les caractéristiques sociales et scolaires des élèves.

Il semble donc plus raisonnable de penser que  $E(\gamma_j | X_{ij}^1, X_j^2) = X_j^1 \lambda_2^1 + X_j^2 \lambda_2^2$ . Les coefficients  $\lambda_2^1$  et  $\lambda_2^2$  n'ont pas d'interprétation structurelle, mais rendent compte d'une corrélation entre le niveau des élèves et les caractéristiques de la classe (à caractéristiques individuelles  $X^1$  contrôlées, et une fois que l'effet causal des caractéristiques de la classe a été pris en compte).

Dans ce cas, les estimateurs de  $\beta_2^1$  et  $\beta_2^2$  obtenus par régression seront biaisés :  $E(\widehat{\beta}_2^1) = \beta_2^1 + \lambda_2^1$  et  $E(\widehat{\beta}_2^2) = \beta_2^2 + \lambda_2^2$ . L'impact causal des variables constantes au sein des classes ne peut donc pas être estimé par régression.

#### 4.2.2 Le cas des variables non constantes au sein de la classe

Remarquons également que même si l'impact des  $X_j^1$  est nul, alors pour estimer sans biais l'effet des caractéristiques individuelles  $X_{ij}^1$ , il faut bien intégrer les moyennes de classe  $X_j^1$  à la régression (car les effets spécifiques à la classe peuvent être corrélés aux  $X_{ij}^1$  via les  $X_j^1$ ). Or, l'ensemble des moyennes individuelles  $X_j^1$  est rarement intégré dans les modèles à effet "classe" aléatoire comme les modèles multi-niveaux. Nous verrons plus loin (cf. chapitre 5) comment cette propriété peut être exploitée pour construire une règle de choix de modèle.

### 4.3 Le lien avec les différents modèles de la littérature

Le modèle que nous décrivons dans ce chapitre est dénommé de différente manière : modèle à effets aléatoires pour les économètres, modèle multi-niveaux ou modèle HLM (hierarchical linear model) pour les sociologues. Ces modèles ont été introduits en sociologie de l'éducation par Raudenbush & Brik. Parmi les nombreuses publications de ces auteurs, le lecteur intéressé pourra se reporter au livre "Hierarchical Linear Models in Social and Behavioral Research : Applications and Data Analysis Methods" publié en 1992 et réédité en 2002. Ce type de modèle est également très utilisé par les agronomes sous le nom de modèle mixte.

Nous nous attachons maintenant à montrer en quoi l'approche multi-niveaux<sup>1</sup> rentre bien dans le cadre que nous présentons et quelles sont les hypothèses d'identification et d'estimation nécessaires (rarement présentées explicitement dans la littérature empirique).

1. On se contentera d'un modèle à deux niveaux seulement, nous laissons au lecteur le soin de généraliser à un nombre supérieur de niveaux.

## 4.4 L'estimation des $\beta$

### 4.4.1 Propriétés de l'estimateur des MCO "naïf"

Un estimateur sans biais mais des tests incorrects

Sous les hypothèses  $(H_1'')$ ,  $(H_2'')$  et  $(H_3'')$ , l'estimateur des MCO de la régression de  $\mathbf{y}$  sur  $\mathbf{X}$  existe, est unique, est sans biais mais sa variance est :

$$V(\widehat{\beta}_{MCO}|\mathbf{X}, \mathbf{Z}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma_\gamma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \neq \lambda(\mathbf{X}'\mathbf{X})^{-1}$$

$$V(\widehat{\beta}_{MCO}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma_\gamma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Z}\mathbf{Z}'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \neq \lambda(\mathbf{X}'\mathbf{X})^{-1}$$

Les tests asymptotiques habituels qui se fondent sur l'hypothèse  $E(v|\mathbf{X}) = \sigma^2I$  sont biaisés. Les tests habituels de Student qui se fondent sur l'hypothèse  $v|\mathbf{X} \rightsquigarrow N(0, \sigma^2I)$  sont biaisés.

Remarque : Il apparaît donc que l'avantage (potentiel) d'un modèle multi-niveaux par rapport à un modèle linéaire "standard" porte seulement sur la précision de l'estimation des paramètres  $\beta$ , sur l'estimation de cette précision et non sur des considérations de biais des estimateurs. Pour l'estimateur des MCO, le terme de variance supplémentaire qui apparaît du fait de la présence de "chocs communs" au niveau de la "classe" sera d'autant plus important que ces chocs sont variables ( $\sigma_\gamma^2$  est grand) ou que les covariables  $\mathbf{X}$  sont corrélées aux variables de "classe"  $\mathbf{Z}$  (le terme  $\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$  pouvant s'interpréter comme la part de variance de  $\mathbf{X}$  inter classe, dans le cas où les classes sont de même taille). La non-prise en compte des effets "classes" induit alors une sous-estimation de la précision des estimateurs, ce qui peut conduire à des diagnostics erronés concernant la significativité de tel ou tel effet de  $\mathbf{x}$  sur  $\mathbf{y}$ .

Preuve :

□

On rappelle que l'expression de l'estimateur des moindres carrés est :  $\widehat{\beta}_{MCO} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ .

On en déduit que  $E(\widehat{\beta}_{MCO}|\mathbf{X}, \mathbf{Z}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + E(v|\mathbf{X}, \mathbf{Z})) = \beta$  et que :

$$\begin{aligned} V(\widehat{\beta}_{MCO}|\mathbf{X}, \mathbf{Z}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(v|\mathbf{X}, \mathbf{Z})\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2I_n + \sigma_\gamma^2\mathbf{Z}\mathbf{Z}')\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma_\gamma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

■

### 4.4.2 Propriétés de l'estimateur à effets fixes

Un estimateur sans biais et des tests corrects

On suppose dans un premier temps que les covariables  $\mathbf{X}_{-0}$  varient toutes au sein des différentes classes (sont donc exclues les variables du type "ZEP", "école privée", "académie de Paris" etc.). Sous les hypothèses  $(H_1'')$ ,  $(H_2'')$  et  $(H_3'')$ , les estimateurs de  $\beta$  et  $\gamma$  des moindres carrés du modèle à effets fixes contraint par  $\sum \gamma = 0$  existent, sont sans biais et leur variances sont de la forme :

$$V(\widehat{\beta}|\mathbf{X}, \mathbf{Z}, \gamma) = \begin{bmatrix} \bar{v} & * & * & * \\ * & \sigma^2(\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1} & & \\ * & & & \\ * & & & \end{bmatrix} \quad \text{avec } \bar{v} = \frac{1}{C^2} \sum_{1 \leq j', j \leq C} [V(\widehat{\gamma}_W|\mathbf{X}, \mathbf{Z}, \gamma)]_{j', j}$$

$$V(\widehat{\gamma}|\mathbf{X}, \mathbf{Z}, \gamma) = \left( I_C - \frac{1}{C} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix} \right) V(\widehat{\gamma}_W|\mathbf{X}, \mathbf{Z}, \gamma) \left( I_C - \frac{1}{C} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix} \right)$$

avec

$$V(\widehat{\gamma}_W|\mathbf{X}, \mathbf{Z}, \gamma) = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}_{-0}(\mathbf{X}'_{-0}W\mathbf{X}_{-0})^{-1}\mathbf{X}'_{-0}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$$

Les tests habituels (présentés dans le chapitre 3) de l'estimateur des moindres carrés contraints restent valables.

Dans le cas où des covariables  $\mathbf{X}$  sont constantes au sein des classes elles peuvent être éliminées de la matrice  $\mathbf{X}_{-0}$ , et le résultat reste vrai pour les composantes de  $\beta$  associées à des covariables qui varient au sein des classes.

**Remarque :** On conditionne les expressions par  $\gamma$  car l'estimateur du modèle à effets fixes ne fait aucune hypothèse sur la distribution des  $\gamma$ , les tests usuels sont donc menés dans un cadre conditionnel. Il est alors complètement naturel que l'estimateur des moindres carrés contraints conserve toutes ces propriétés, le modèle à effets fixes est un sur-modèle du modèle à effets aléatoires dans le sens où les hypothèses  $(H_1'')$ ,  $(H_2'')$  et  $(H_3'')$  sont plus restrictives que les hypothèses du modèle à effets fixes. En utilisant un estimateur à effets fixes, on introduit certes dans la régression des régresseurs supplémentaires, mais cela ne biaise pas les estimations des paramètres. De plus, le fait que ces régresseurs supplémentaires soient introduits permet de retrouver l'homoscédasticité sur les résidus et donc les estimations de variances sont également sans biais. En revanche, cette adjonction de régresseurs inutiles dégrade la variance des estimateurs et on va voir que l'estimateur du modèle à effets fixes perd ses propriétés d'optimalité (cf. 4.4.3). Pour résumer, si le "vrai" modèle est un modèle à erreurs composées et qu'on procède à une estimation en considérant que l'on a affaire à un modèle à effets fixes, les estimateurs obtenus sont sans biais, on est capable d'estimer correctement leur précision, le seul inconvénient est que ce ne sont pas les estimateurs les plus précis possibles (car on n'a pas exploité l'hypothèse sur la distribution des  $\gamma$  en raisonnant conditionnellement à ceux-ci). Après avoir présenté les différents estimateurs du modèles à erreurs composées (cf. 4.4.3 et 4.4.6), nous reviendrons en détail sur cette discussion (cf. 5) en considérant au contraire le cas où le "vrai" modèle est un modèle à effets fixes, et que l'on procède à une estimation en considérant que l'on a affaire à un modèle à erreurs composées.

Preuve :

□

Les propriétés de l'estimateur à effets fixes sont une conséquence directe des propriétés établies dans le chapitre 3.

■

### 4.4.3 L'estimateur des Moindres Carrés Généralisés (MCG)

La différence entre le modèle à erreurs composées et le modèle linéaire simple, provient uniquement de la forme de la matrice de variance-covariance des résidus. Si les résidus étaient homoscédastiques, alors l'estimateur de moindres carrés ordinaires serait sans biais et optimal, on disposerait de tests non biaisés des estimations. Le principe de construction de l'estimateur des moindres carrés généralisés et de l'estimateur des moindres carrés quasi-généralisés est donc de transformer le modèle afin de retrouver une structure de variance-covariance des résidus qui soit homoscédastique.

Le modèle à erreurs composées s'écrit  $\mathbf{y} = \mathbf{X}\beta + v$  avec  $E(v|\mathbf{X}, \mathbf{Z}) = 0$  et  $V(v|\mathbf{X}, \mathbf{Z}) = \sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}'$ . Comme  $\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}'$  est une matrice symétrique définie positive, il existe une unique matrice  $M$  inversible et symétrique telle que  $(M^{-1})^2 = \sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}'$ ; on note  $(\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-\frac{1}{2}}$  cette matrice.

Le modèle peut alors se réécrire de la manière suivante :

$$(\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-\frac{1}{2}} \mathbf{y} = (\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-\frac{1}{2}} \mathbf{X}\beta + (\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-\frac{1}{2}} v$$

avec  $E((\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-\frac{1}{2}} v|\mathbf{X}, \mathbf{Z}) = 0$  et  $V((\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-\frac{1}{2}} v|\mathbf{X}, \mathbf{Z}) = I$ . On se retrouve alors pour le modèle réécrit dans le cadre du modèle linéaire présenté dans le chapitre 2. On en déduit alors la proposition suivante :

#### Propriétés de l'estimateur des MCG

Sous les hypothèses  $(H_1'')$ ,  $(H_2'')$  et  $(H_3'')$ , l'estimateur de  $\beta$  des moindres carrés généralisés défini par :

$$\hat{\beta}_{MCG} = \left( \mathbf{X}' (\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' (\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{y}$$

existe, est unique, sans biais, et de variance  $V(\widehat{\beta}_{MCG}|\mathbf{X}, \mathbf{Z}) = (\mathbf{X}'(\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{X})^{-1}$ . Sous l'hypothèse supplémentaire de normalité de résidus, l'estimateur de  $\beta$  des moindres carrés généralisés est optimal parmi les estimateurs linéaires sans biais de  $\beta$ .

**Remarque :** Cet estimateur a surtout un intérêt théorique comme meilleur estimateur possible de  $\beta$ . Mais il n'est d'aucune utilité pratique dans la mesure où  $\sigma_\gamma^2$  et  $\sigma^2$  sont inconnus. On peut néanmoins remarquer que :

- La matrice  $(\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1}$  est égale à  $\frac{1}{\sigma^2} I - \frac{\sigma_\gamma^2}{\sigma^2} \mathbf{Z} (\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'$ . Par rapport au modèle linéaire homoscédastique, il apparaît donc des termes supplémentaires dans l'expression de l'estimateur et de sa variance liés à l'expression  $\frac{\sigma_\gamma^2}{\sigma^2} \mathbf{Z} (\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'$ ; ces termes sont nuls si  $\sigma_\gamma^2 = 0$  (absence d'effets "classe").
- Pour un modèle à erreurs composées, l'estimateur des MCO n'est plus l'estimateur optimal des  $\beta$ . Il est moins précis que l'estimateur des MCG.
- Pour un modèle à erreurs composées, l'estimateur "Within" n'est plus l'estimateur optimal des  $\beta$ . Il est moins précis que l'estimateur des MCG. Dans le cas où le nombre d'élèves par classe est constant, la matrice de variance-covariance de l'estimateur des MCG a pour expression  $\sigma^2 \left( \mathbf{X}'\mathbf{W}\mathbf{X} + \frac{\sigma^2}{\sigma^2 + n_1 \sigma_\gamma^2} \mathbf{X}\mathbf{B}\mathbf{X} \right)^{-1}$  avec  $B = I - W = \mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'$ . La perte d'efficacité de l'estimateur "Within" par rapport à l'estimateur des MCG est donc d'autant plus importante que la variance inter-classe des variables observables est grande devant la variance intra-classe ( $\mathbf{X}\mathbf{B}\mathbf{X} \gg \mathbf{X}\mathbf{W}\mathbf{X}$ ) et que la variance intra-classe des variables inobservables est grande devant la variance inter-classe ( $\sigma^2 \gg n_1 \sigma_\gamma^2$ ).

Preuve :

□

Les propriétés de l'estimateur de MCG sont une conséquence directe des propriétés de l'estimateur des MCO présentées dans le chapitre 2.

■

À partir de l'expression théorique de l'estimateur des MCG, une idée relativement naturelle pour obtenir une estimation des paramètres  $\beta$  est de remplacer la matrice  $(\sigma_\gamma^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 I)$  par une estimation, on obtient alors un estimateur des moindres carrés quasi-généralisés (MCQG). Pour cela on utilise des estimateurs convergents de  $\sigma_\gamma^2$  et  $\sigma^2$ . Plusieurs stratégies sont possibles pour estimer ces deux quantités mais nous nous contenterons de mentionner deux approches. La première est celle qui est classiquement utilisée en présence d'un panel cylindré (panel dans lequel tout les individus ont répondu à toutes les vagues d'enquête) et qui peut très facilement être appliquée dans le cas où le nombre d'élèves observés par classe est constant. La seconde est moins classique mais s'adapte bien au cas où le nombre d'élèves observés diffère d'une classe à l'autre et est déjà pré-programmé dans la procédure MIXED de SAS.

#### 4.4.4 Quelques estimateurs de $\sigma^2$ et $\sigma_\gamma^2$

**Estimateurs de  $\sigma^2$  et  $\sigma_\gamma^2$  basés sur les projections inter-classe et intra-classe**

On note  $W\mathbf{y}$  et  $W\mathbf{X}$  les écarts à la moyenne de classe des quantités  $\mathbf{y}$  et  $\mathbf{X}$ , et  $B\mathbf{y}$  et  $B\mathbf{X}$  les moyennes de classe. D'un point de vue matriciel  $W = I - B$  et  $B = \mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'$ . On note  $\widehat{Wv}$  (respectivement  $\widehat{Bv}$ ) les résidus estimés de la régression de  $W\mathbf{y}$  sur  $W\mathbf{X}$  (respectivement de  $B\mathbf{y}$  sur  $B\mathbf{X}$ ).

Sous les hypothèses  $(H_1'')$ ,  $(H_2'')$ ,  $(H_3'')$ , les estimateurs

$$\widehat{\sigma^2} = \frac{(\widehat{Wv})'(\widehat{Wv})}{n - K_W - C}$$

$$\widehat{\sigma_\gamma^2} = \frac{(\widehat{Bv})'(\widehat{Bv}) - (\widehat{Wv})'(\widehat{Wv}) \frac{(C - (K+1))}{n - K - C}}{n - \text{Tr}(\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1} \mathbf{X}'\mathbf{Z})}$$

(où  $K_W$  désigne le nombre de covariables  $\mathbf{X}$  non constantes au sein des classes) sont des estimateurs sans biais des paramètres  $\sigma^2$  et  $\sigma_\gamma^2$ . Ces estimateurs convergent de plus à la vitesse  $\sqrt{n}$ .

Si, de plus, le nombre d'élèves observés dans chaque classe est constant, alors ces estimateurs ont pour expression :

$$\widehat{\sigma^2} = \frac{(\widehat{Wv})'(\widehat{Wv})}{n - K_W - C}$$

$$\widehat{\sigma_\gamma^2} = \frac{(\widehat{Bv})'(\widehat{Bv})}{n_1(C - (K_W + 1))} - \frac{1}{n_1} \frac{(\widehat{Wv})'(\widehat{Wv})}{n - K_W - C} = \frac{(\widehat{Bv})'(\widehat{Bv})}{n_1(C - n_1(K + 1))} - \frac{1}{n_1} \widehat{\sigma^2}$$

**Remarque :** Dans le cas où le nombre d'élèves par classe est constant, la variance estimée des résidus fournie par les sorties de régressions sont  $\frac{(\widehat{Wv})'(\widehat{Wv})}{n - K_W - 1}$  et  $\frac{(\widehat{Bv})'(\widehat{Bv})}{n_1(C - (K_W + 1))}$ . Après avoir corrigé la variance estimée des résidus "Within" par le facteur  $\frac{n - K_W - 1}{n - K_W - C}$  (cf. chapitre 3), les estimateurs  $\widehat{\sigma^2}$  et  $\widehat{\sigma_\gamma^2}$  se calculent très facilement. Malheureusement, cette méthode d'estimation est bien moins pratique dans le cas où le nombre d'élèves par classe n'est pas constant (même s'il est possible de récupérer dans une table SAS la matrice  $(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}$ , le calcul de  $Tr(\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z})$  doit être programmé en partie "à la main").

Preuve :

□

Le modèle après transformation "Within" s'écrit :  $\mathbf{W}\mathbf{y} = \mathbf{W}\mathbf{X}\beta + \mathbf{W}\mathbf{v}$  avec  $\mathbf{W} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$

On note  $\widehat{Wv} = \mathbf{W}\mathbf{y} - \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}\mathbf{y}$ , avec  $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}$  une pseudo-inverse de  $\mathbf{X}'\mathbf{W}\mathbf{X}$  (une pseudo-inverse  $A^{-}$  de la matrice  $A$  vérifie  $AA^{-}A = A$  et  $A^{-}AA^{-} = A^{-}$ , dans le chapitre 3 la pseudo-inverse utilisée était construite par élimination de la variable constante  $\mathbf{x}_0$  de la liste des  $\mathbf{X}$ ). On note également  $M = \mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}$ , on a alors  $\mathbf{W}M = M$ ,  $\mathbf{W}\mathbf{Z} = 0$ .

On en déduit que :

$$\begin{aligned} E\left((\widehat{Wv})'(\widehat{Wv})|\mathbf{X}, \mathbf{Z}\right) &= E(v'\mathbf{W}M\mathbf{W}v|\mathbf{X}, \mathbf{Z}) \\ &= E[Tr(v'\mathbf{W}M\mathbf{W}v)|\mathbf{X}, \mathbf{Z}] \\ &= E[Tr(M\mathbf{W}v v'\mathbf{W})|\mathbf{X}, \mathbf{Z}] \\ &= Tr[ME(\mathbf{W}v v'\mathbf{W}|\mathbf{X}, \mathbf{Z})] \\ &= Tr[ME(\mathbf{W}(\sigma^2\mathbf{I} + \sigma_\gamma^2\mathbf{Z}\mathbf{Z}')\mathbf{W}|\mathbf{X}, \mathbf{Z})] \\ &= Tr[M\mathbf{W}\sigma^2] \\ &= \sigma^2 Tr[M] \\ &= \sigma^2 \left(Tr[\mathbf{W}] - Tr[\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}]\right) \end{aligned}$$

$\mathbf{W}$  et  $\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}$  sont des projecteurs orthogonaux de rang  $n - C$  et  $K_W$ . La trace d'un projecteur orthogonal étant égale à son rang, on a donc  $E\left((\widehat{Wv})'(\widehat{Wv})|\mathbf{X}, \mathbf{Z}\right) = \sigma^2(n - C - K_W)$ .

La vitesse de convergence de cet estimateur est en  $\sqrt{n - C}$ , dans le cas qui nous intéresse, il est logique de supposer que  $\frac{C}{n} \rightarrow cst$  quand  $n \rightarrow \infty$ , la vitesse de convergence de l'estimateur est alors bien en  $\sqrt{n}$ .

On note  $N$  la matrice de projection orthogonale sur  $\mathbf{B}\mathbf{X}^\perp$ ,  $N = \mathbf{I} - \mathbf{B}\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}$ .

On a :

$$\begin{aligned} E\left((\widehat{Bv})'(\widehat{Bv})|\mathbf{X}, \mathbf{Z}\right) &= E(v'\mathbf{B}'N\mathbf{B}v|\mathbf{X}, \mathbf{Z}) \\ &= Tr[\mathbf{B}'N\mathbf{B}E(vv'|\mathbf{X}, \mathbf{Z})] \\ &= \sigma^2 Tr(\mathbf{B}'N\mathbf{B}) + \sigma_\gamma^2 Tr(\mathbf{B}'N\mathbf{B}\mathbf{Z}\mathbf{Z}') \\ &= \sigma^2 Tr(N\mathbf{B}\mathbf{B}') + \sigma_\gamma^2 Tr(\mathbf{Z}'\mathbf{B}'N\mathbf{B}\mathbf{Z}) \end{aligned}$$

Comme  $\mathbf{B}' = \mathbf{B}$ ,  $\mathbf{B}^2 = \mathbf{B}$ ,  $Tr(N\mathbf{B}\mathbf{B}') = Tr(\mathbf{B}) - Tr(\mathbf{B}\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}) = C - (K + 1)$ .

Comme  $\mathbf{B}\mathbf{Z} = \mathbf{Z}$ ,  $Tr(\mathbf{Z}'\mathbf{B}'N\mathbf{B}\mathbf{Z}) = Tr(\mathbf{Z}'\mathbf{Z}) - Tr(\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{B}'\mathbf{B}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}) = Tr(\mathbf{Z}'\mathbf{Z}) - Tr(\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z})$ .

Dans le cas où le nombre d'élèves par classe est constant, on a :  $(\mathbf{Z}'\mathbf{Z})^{-1} = n_1\mathbf{I}$  donc  $Tr(\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}) = n_1 Tr((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}) = n_1 Tr(\mathbf{X}'\mathbf{B}\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}) = n_1(K_W + 1)$ .

■

La "PROC MIXED" de SAS utilise d'autres estimateurs de  $\sigma^2$  et  $\sigma_\gamma^2$ .

### Estimateurs de $\sigma^2$ et $\sigma_\gamma^2$ de la première étape de la proc MIXED de SAS

Soit  $M$  la matrice de projection  $I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Les estimateurs ci-dessous sont sans biais et convergents à la vitesse  $\sqrt{n}$ .

$$\widehat{\sigma^2} = \frac{\text{Tr}(M\mathbf{Z}\mathbf{Z}'M\mathbf{Z}\mathbf{Z}')\mathbf{y}'M\mathbf{y} - \text{Tr}(M\mathbf{Z}\mathbf{Z}')\mathbf{y}'M\mathbf{Z}\mathbf{Z}'M\mathbf{y}}{\text{Tr}(M)\text{Tr}(M\mathbf{Z}\mathbf{Z}'M\mathbf{Z}\mathbf{Z}') - \text{Tr}(M\mathbf{Z}\mathbf{Z}')\text{Tr}(M\mathbf{Z}\mathbf{Z}')}$$

$$\widehat{\sigma_\gamma^2} = \frac{\text{Tr}(M)\mathbf{y}'M\mathbf{Z}\mathbf{Z}'M\mathbf{y} - \text{Tr}(M\mathbf{Z}\mathbf{Z}')\mathbf{y}'M\mathbf{y}}{\text{Tr}(M)\text{Tr}(M\mathbf{Z}\mathbf{Z}'M\mathbf{Z}\mathbf{Z}') - \text{Tr}(M\mathbf{Z}\mathbf{Z}')\text{Tr}(M\mathbf{Z}\mathbf{Z}')}$$

Preuve :

□

Il suffit de remarquer que :

$$E(\mathbf{y}'M\mathbf{y}|\mathbf{X}, \mathbf{Z}) = E(\mathbf{v}'M\mathbf{v}|\mathbf{X}, \mathbf{Z}) = E(\text{Tr}(\mathbf{v}'M\mathbf{v})|\mathbf{X}, \mathbf{Z}) = E(\text{Tr}(M\mathbf{v}\mathbf{v}')|\mathbf{X}, \mathbf{Z}) = \text{Tr}(ME(\mathbf{v}\mathbf{v}'|\mathbf{X}, \mathbf{Z})) = \text{Tr}(M(\sigma_\gamma^2\mathbf{Z}\mathbf{Z}' + \sigma^2I)) = \sigma_\gamma^2\text{Tr}(M\mathbf{Z}\mathbf{Z}') + \sigma^2\text{Tr}(M)$$

et que :

$$E(\mathbf{y}'M\mathbf{Z}\mathbf{Z}'M\mathbf{y}|\mathbf{X}, \mathbf{Z}) = E(\mathbf{v}'M\mathbf{Z}\mathbf{Z}'M\mathbf{v}|\mathbf{X}, \mathbf{Z}) = E(\text{Tr}(\mathbf{v}'M\mathbf{Z}\mathbf{Z}'M\mathbf{v})|\mathbf{X}, \mathbf{Z}) = E(\text{Tr}(M\mathbf{Z}\mathbf{Z}'M\mathbf{v}\mathbf{v}')|\mathbf{X}, \mathbf{Z}) = \text{Tr}(M\mathbf{Z}\mathbf{Z}'ME(\mathbf{v}\mathbf{v}'|\mathbf{X}, \mathbf{Z})) = \sigma_\gamma^2\text{Tr}(M\mathbf{Z}\mathbf{Z}'M\mathbf{Z}\mathbf{Z}') + \sigma^2\text{Tr}(M\mathbf{Z}\mathbf{Z}')$$

On en déduit alors que  $E(\widehat{\sigma^2}|\mathbf{X}, \mathbf{Z}) = \sigma^2$  et  $E(\widehat{\sigma_\gamma^2}|\mathbf{X}, \mathbf{Z}) = \sigma_\gamma^2$  dès que  $\text{Tr}(M)\text{Tr}(M\mathbf{Z}\mathbf{Z}'M\mathbf{Z}\mathbf{Z}') - \text{Tr}(M\mathbf{Z}\mathbf{Z}')\text{Tr}(M\mathbf{Z}\mathbf{Z}') \neq 0$

L'inégalité de Cauchy-Schwarz assure que  $\text{Tr}(M\mathbf{Z}\mathbf{Z}') = \text{Tr}(MM\mathbf{Z}\mathbf{Z}') \leq \sqrt{\text{Tr}(M^2)\text{Tr}(M\mathbf{Z}\mathbf{Z}'M\mathbf{Z}\mathbf{Z}')}.$  Comme  $\text{Tr}(M^2) = \text{Tr}(M) = \text{rg}(M) = n - (K + 1)$  on en déduit que la quantité  $\text{Tr}(M)\text{Tr}(M\mathbf{Z}\mathbf{Z}'M\mathbf{Z}\mathbf{Z}') - \text{Tr}(M\mathbf{Z}\mathbf{Z}')\text{Tr}(M\mathbf{Z}\mathbf{Z}')$  est nulle si et seulement si  $M$  et  $M\mathbf{Z}\mathbf{Z}'$  sont colinéaires donc seulement si  $\text{rg}(M\mathbf{Z}\mathbf{Z}') = n - (K + 1)$  ou  $M\mathbf{Z}\mathbf{Z}' = 0$ . Or  $C = \text{rg}(\mathbf{Z}') \geq \text{rg}(M\mathbf{Z}\mathbf{Z}')$ , le dénominateur des estimateurs ne peut être nul dès que le nombre d'observations est supérieur au nombre de paramètres à estimer ( $n > C + K + 1$ ) et qu'on n'est pas dans le cas dégénéré où  $\mathbf{Z}$  est une combinaison linéaire de  $\mathbf{X}$  (ce qui voudrait dire que l'appartenance à la classe s'explique entièrement par une combinaison linéaire des caractéristiques  $\mathbf{X}$ !).

■

#### 4.4.5 Estimateurs des Moindres Carrés Quasi-Généralisés (MCQG)

On dispose à présent d'estimateurs convergents de  $\sigma^2$  et  $\sigma_\gamma^2$ . L'estimateur des moindres carrés quasi-généralisés (MCQG) suit la même démarche que l'estimateur des MCG, mais les quantités inconnues  $\sigma^2$  et  $\sigma_\gamma^2$  sont remplacées par leurs estimations. L'estimateur des MCQG vérifie asymptotiquement les mêmes propriétés que l'estimateur des MCG (cela est dû au fait que les estimations des paramètres se rapprochant de plus en plus des "vraies" valeurs quand  $n$  est très grand, l'estimation de première étape est donc sans influence sur le comportement des estimateurs). Par contre, le fait de remplacer des paramètres inconnus par des estimations dégrade les propriétés de l'estimateur à distance finie, il faut donc se placer sous des hypothèses asymptotiques. Comme dans le cas du modèle à effets fixes, il est plus facile de raisonner en faisant tendre le nombre de classes vers  $+\infty$ , plutôt que le nombre d'élèves. On note donc  $\tilde{\mathbf{y}}_j$ ,  $\tilde{\mathbf{X}}_{-0j}$  et  $\tilde{\epsilon}_j$ , les vecteurs obtenus en "empilant" les observations de la classe  $j$ .

#### Hypothèse

Hypothèse ( $H''_{as}$ ) :

◊ Les nombres d'élèves observés par classe  $n_j$  sont indépendants et identiquement distribués. On suppose de plus que  $1 \leq n_1 \leq n_{sup} < +\infty$  presque sûrement et que  $P(n_1 = 1) < 1$ . Dans ces conditions, on a  $1 < E(n_1) = \bar{n} < +\infty$ .

◊ Les variables  $(\tilde{\mathbf{y}}_j, \tilde{\mathbf{X}}_j)_{j=1\dots C}$  sont indépendantes et identiquement distribuées conditionnellement à  $n_j$ .

◊  $E \left[ \tilde{\mathbf{X}}_1' (\sigma^2 I_{n_1} + \sigma_\gamma^2 J_{n_1})^{-1} \tilde{\mathbf{X}}_1 \right]$  existe et est inversible ( $J_{n_1}$  est une matrice carrée de taille  $n_1$ , remplie de 1).

$$\diamond E \left[ \left| \left( \tilde{\mathbf{X}}_1' (\sigma^2 I_{n_1} + \sigma_\gamma^2 J_{n_1})^{-1} \tilde{\mathbf{X}}_1 \right)_{(k,l)} \right| \right] < +\infty \text{ pour tout } 0 \leq l, k \leq K.$$

### Normalité asymptotique des estimateurs des MCQG

Sous les hypothèses  $(H_1'')$ ,  $(H_2'')$ ,  $(H_3'')$  et  $(H_{as}'')$ , l'estimateur de  $\beta$  des moindres carrés quasi-généralisés défini par :

$$\hat{\beta}_{MCQG} = \left( \mathbf{X}' (\widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' (\widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{y}$$

existe et est unique.

Bien que biaisé à distance finie, il vérifie les propriétés suivantes :

$$\begin{aligned} \hat{\beta}_{MCQG} &\xrightarrow[n \rightarrow +\infty]{p} \beta \\ \sqrt{n} (\hat{\beta}_{MCQG} - \beta) &\xrightarrow[n \rightarrow +\infty]{loi} N \left( 0, \bar{n} E \left[ \tilde{\mathbf{X}}_1' (\sigma^2 I_{n_1} + \sigma_\gamma^2 J_{n_1})^{-1} \tilde{\mathbf{X}}_1 \right]^{-1} \right) \\ \left( \mathbf{X}' (\widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{X} \right)^{\frac{1}{2}} (\hat{\beta}_{MCQG} - \beta) &\xrightarrow[n \rightarrow +\infty]{loi} N(0, I_{K+1}) \end{aligned}$$

Sous l'hypothèse supplémentaire de normalité des résidus, l'estimateur de  $\beta$  des moindres carrés quasi-généralisés est asymptotiquement optimal parmi les estimateurs asymptotiquement sans biais de  $\beta$ .

**Remarque :** L'estimateur des moindres carrés quasi-généralisés n'a donc pas de bonnes propriétés à distance finie, mais c'est le meilleur estimateur (sous les hypothèses du modèle) dès que le nombre d'observations peut être considéré comme très grand. En pratique, cette dernière condition est très généralement acceptable (travail sur des fichiers d'enquête de plusieurs milliers d'élèves ou sur des fichiers administratifs de taille beaucoup plus importante).

On peut de plus construire des tests asymptotiques car la matrice  $\left( \mathbf{X}' (\widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{X} \right)$  fournit un estimateur asymptotique de la variance de  $\hat{\beta}_{MCQG}$ . Ainsi, un test de l'hypothèse  $\beta_k = 0$  pourra être mené en considérant la quantité  $\frac{\hat{\beta}_k}{\sqrt{\left[ \mathbf{X}' (\widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{X} \right]_{(k,k)}}}$  qui converge (en loi) vers

une loi normale centrée réduite sous l'hypothèse nulle, et qui devient très grande (presque sûrement) sous l'hypothèse alternative.

Preuve :

□

$$\text{On a } \hat{\beta}_{MCQG} = \beta + \left( \mathbf{X}' (\widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' (\widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} v.$$

$$\text{donc } \sqrt{C} (\hat{\beta}_{MCQG} - \beta) = \left[ \frac{1}{C} \mathbf{X}' (\widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{X} \right]^{-1} \frac{1}{\sqrt{C}} \mathbf{X}' (\widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} v$$

$$\frac{1}{C} \mathbf{X}' (\widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{X} = \frac{1}{\sigma^2} \frac{1}{C} \sum_{j=1}^C \left( \tilde{\mathbf{X}}_j' \tilde{\mathbf{X}}_j - \frac{\sigma_\gamma^2}{\sigma^2 + n_j \sigma_\gamma^2} \tilde{\mathbf{X}}_j' J_{n_j} \tilde{\mathbf{X}}_j \right)$$

Comme :

$$\frac{1}{C} \sum_{j=1}^C \frac{1}{\widehat{\sigma}^2 + n_j \widehat{\sigma}_\gamma^2} \tilde{\mathbf{X}}_j' J_{n_j} \tilde{\mathbf{X}}_j - \frac{1}{C} \sum_{j=1}^C \frac{1}{\sigma^2 + n_j \sigma_\gamma^2} \tilde{\mathbf{X}}_j' J_{n_j} \tilde{\mathbf{X}}_j = o_p(1) \frac{1}{C} \sum_{j=1}^C (1 + n_j) \tilde{\mathbf{X}}_j' J_{n_j} \tilde{\mathbf{X}}_j = o_p(1)$$

Et :

$$\frac{1}{C} \sum_{j=1}^C \frac{1}{\widehat{\sigma}^2 + n_j \widehat{\sigma}_\gamma^2} \tilde{\mathbf{X}}_j' J_{n_j} \tilde{\mathbf{X}}_j - E \left( \frac{1}{\sigma^2 + n_1 \sigma_\gamma^2} \tilde{\mathbf{X}}_1' J_{n_1} \tilde{\mathbf{X}}_1 \right) = o_p(1) \text{ ainsi que } \frac{1}{C} \sum_{j=1}^C \tilde{\mathbf{X}}_j' \tilde{\mathbf{X}}_j - E \left( \tilde{\mathbf{X}}_1' \tilde{\mathbf{X}}_1 \right) = o_p(1)$$

On a :

$$\frac{1}{C} \mathbf{X}' \left( \widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}' \right)^{-1} \mathbf{X} \xrightarrow{C \rightarrow +\infty} \frac{1}{\sigma^2} E \left( \tilde{\mathbf{X}}_1' \tilde{\mathbf{X}}_1 - \frac{\sigma_\gamma^2}{\sigma^2 + n_1 \sigma_\gamma^2} \tilde{\mathbf{X}}_1' J_{n_1} \tilde{\mathbf{X}}_1 \right) = E \left[ \tilde{\mathbf{X}}_1' \left( \sigma^2 I_{n_1} + \sigma_\gamma^2 J_{n_1} \right)^{-1} \tilde{\mathbf{X}}_1 \right]$$

De même :

$$\frac{1}{\sqrt{C}} \mathbf{X}' \left( \widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}' \right)^{-1} \mathbf{v} = \frac{1}{\sqrt{C}} \frac{1}{\sigma^2} \sum_{j=1}^C \left( \tilde{\mathbf{X}}_j' \tilde{v}_j - \frac{\widehat{\sigma}_\gamma^2}{\widehat{\sigma}^2 + n_j \widehat{\sigma}_\gamma^2} \tilde{\mathbf{X}}_j' J_{n_j} \tilde{v}_j \right)$$

Et par le même raisonnement que précédemment, on a :

$$\frac{1}{\sqrt{C}} \mathbf{X}' \left( \widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}' \right)^{-1} \mathbf{v} \xrightarrow{C \rightarrow +\infty} N \left( 0, E \left[ \tilde{\mathbf{X}}_1' \left( \sigma^2 I_{n_1} + \sigma_\gamma^2 J_{n_1} \right)^{-1} \tilde{\mathbf{X}}_1 \right] \right)$$

Finalement :

$$\sqrt{C} \left( \widehat{\beta}_{MCQG} - \beta \right) \xrightarrow{C \rightarrow +\infty} N \left( 0, E \left[ \tilde{\mathbf{X}}_1' \left( \sigma^2 I_{n_1} + \sigma_\gamma^2 J_{n_1} \right)^{-1} \tilde{\mathbf{X}}_1 \right]^{-1} \right)$$

Et donc :

$$\sqrt{n} \left( \widehat{\beta}_{MCQG} - \beta \right) \xrightarrow{n \rightarrow +\infty} N \left( 0, \bar{n} E \left[ \tilde{\mathbf{X}}_1' \left( \sigma^2 I_{n_1} + \sigma_\gamma^2 J_{n_1} \right)^{-1} \tilde{\mathbf{X}}_1 \right]^{-1} \right)$$

$$\left( \mathbf{X}' \left( \widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}' \right)^{-1} \mathbf{X} \right)^{\frac{1}{2}} \left( \widehat{\beta}_{MCQG} - \beta \right) \xrightarrow{n \rightarrow +\infty} N \left( 0, I_{K+1} \right)$$

Enfin, la propriété d'optimalité est assurée par le fait que l'estimateur des MCQG est asymptotiquement équivalent à l'estimateur des MCG.

■

#### 4.4.6 Estimateur du Maximum de Vraisemblance (MV) et estimateur du Maximum de Vraisemblance Restreint (MVR) sous hypothèses de normalité des résidus

Dans le cas où les résidus sont supposés appartenir à une famille de distribution paramétrée, les estimateurs basés sur la maximisation de la vraisemblance des observations possèdent des propriétés asymptotiques intéressantes. Il est néanmoins nécessaire de faire deux hypothèses techniques supplémentaires : l'hypothèse que le nombre d'élèves par classe est borné (condition vérifiée, on peut raisonnablement considérer que les classes comprennent toutes moins de 60 élèves) et l'hypothèse que les éléments de la matrice  $n \left( \mathbf{X}' (\sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}')^{-1} \mathbf{X} \right)^{-1}$  restent bornés quand  $n$  tend vers  $+\infty$ .

Dans le cas qui nous intéresse, nous allons donc faire l'hypothèse supplémentaire ( $H_n''$ ) suivante :

##### Hypothèse de normalité des résidus

( $H_n''$ ) : La distribution des effets "classe" et de l'hétérogénéité individuelle est normale conditionnellement aux variables  $\mathbf{X}$  et  $\mathbf{Z}$  :

$$\epsilon | \mathbf{X}, \mathbf{Z}, \gamma \rightsquigarrow N(0, \sigma^2 I_N)$$

$$\gamma | \mathbf{X}, \mathbf{Z} \rightsquigarrow N(0, \sigma_\gamma^2 I_C)$$

Dans ces conditions :

$$v | \mathbf{X}, \mathbf{Z} \rightsquigarrow N(0, V) \text{ avec } V = \sigma^2 I + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}'$$

**Remarque :** la première partie de l'hypothèse ( $H_n''$ ) est une hypothèse forte qui implique en particulier les hypothèses ( $H_2''$ ) et ( $H_3''$ ). Néanmoins cette restriction permet de se placer dans un cadre paramétrique classique et donc dans un cadre théorique bien connu. Cependant, pour dériver les comportements asymptotiques des estimateurs, il faut également contrôler le comportement de

$(\mathbf{X}'\widehat{V}^{-1}\mathbf{X})^{-1}$ , qui dépend de la suite des réalisations de  $X$ . L'hypothèse  $(H''_{as})$  sera donc maintenue à cette fin.

### Normalité asymptotique des estimateurs

Sous les hypothèses  $(H''_1)$ ,  $(H''_{as})$  et  $(H''_n)$ , l'estimateur du maximum de vraisemblance des paramètres  $\beta$ ,  $\sigma_\gamma^2$  et  $\sigma^2$  est défini par les trois équations :

$$\widehat{\beta} = (\mathbf{X}'\widehat{V}^{-1}\mathbf{X})^{-1} \mathbf{X}'\widehat{V}^{-1}\mathbf{y}$$

$$Tr(\widehat{V}^{-1}) = \mathbf{y}'\widehat{P}^2\mathbf{y}$$

$$Tr(\widehat{V}^{-1}\mathbf{Z}\mathbf{Z}') = \mathbf{y}'\widehat{P}\mathbf{Z}\mathbf{Z}'\widehat{P}\mathbf{y}$$

avec  $\widehat{V} = \widehat{\sigma}_\gamma^2\mathbf{Z}\mathbf{Z}' + \widehat{\sigma}^2I$  et  $\widehat{P} = \widehat{V}^{-1} - \widehat{V}^{-1}\mathbf{X}(\mathbf{X}'\widehat{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{V}^{-1}$

Pour  $\sigma^2 > 0$  et  $\sigma_\gamma^2 > 0$ , cet estimateur est asymptotiquement sans biais, et asymptotiquement optimal parmi les estimateurs asymptotiquement sans biais.

De plus pour  $\sigma^2 > 0$  et  $\sigma_\gamma^2 > 0$  :

$$\sqrt{n} \left( \begin{pmatrix} \widehat{\beta} \\ \widehat{\sigma}_\gamma^2 \\ \widehat{\sigma}^2 \end{pmatrix} - \begin{pmatrix} \beta \\ \sigma_\gamma^2 \\ \sigma^2 \end{pmatrix} \right) \rightarrow \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, V_{as} \right)$$

où la variance asymptotique de l'estimateur du maximum de vraisemblance est :

$$V_{as} = \lim_{n \rightarrow +\infty} \begin{bmatrix} n(\mathbf{X}'V^{-1}\mathbf{X})^{-1} & & 0 \\ 0 & 2n \begin{bmatrix} Tr(V^{-1}\mathbf{Z}\mathbf{Z}'V^{-1}\mathbf{Z}\mathbf{Z}') & Tr(V^{-1}\mathbf{Z}\mathbf{Z}'V^{-1}) \\ Tr(V^{-1}\mathbf{Z}\mathbf{Z}'V^{-1}) & Tr(V^{-2}) \end{bmatrix}^{-1} & \end{bmatrix}$$

avec  $V = \sigma_\gamma^2\mathbf{Z}\mathbf{Z}' + \sigma^2I$

**Remarque :** L'estimateur du maximum de vraisemblance doit ses propriétés à des développements de certaines fonctionnelles autour des paramètres (ici  $\beta, \sigma^2, \sigma_\gamma^2$ ). Pour pouvoir faire ces développements de manière standard, il faut que le voisinage des paramètres soit suffisamment "régulier" : c'est le cas lorsque  $\sigma^2 > 0$  et  $\sigma_\gamma^2 > 0$ . Dans le cas contraire, ( $\sigma_\gamma^2 = 0$ ),  $\sigma_\gamma^2$  ne peut pas être approché par valeurs inférieures et on sort du cadre habituel des théorèmes de convergence de l'estimateur du maximum de vraisemblance. C'est pour cette raison que les propriétés sont énoncées sous la condition que  $\sigma^2 > 0$  et  $\sigma_\gamma^2 > 0$ . Nous évoquerons dans la suite (cf. 4.6.2) comment tester de manière rigoureuse l'hypothèse  $\sigma_\gamma^2 = 0$ .

Preuve :

□

La log-vraisemblance conditionnelle du modèle s'écrit :

$$\log \mathcal{L}_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}(\beta, \sigma^2, \sigma_\gamma^2) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log[\det(\sigma^2I + \sigma_\gamma^2\mathbf{Z}\mathbf{Z}')] - \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' (\sigma^2I + \sigma_\gamma^2\mathbf{Z}\mathbf{Z}')^{-1} (\mathbf{y} - \mathbf{X}\beta)$$

Les conditions du premier ordre s'écrivent :  $\frac{\partial \log \mathcal{L}}{\partial \beta} = 0$ ,  $\frac{\partial \log \mathcal{L}}{\partial \sigma^2} = 0$ ,  $\frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} = 0$ .

On obtient ainsi trois types d'équations :

$$\mathbf{X}'(\widehat{\sigma}^2I + \widehat{\sigma}_\gamma^2\mathbf{Z}\mathbf{Z}')^{-1}(\mathbf{y} - \mathbf{X}\widehat{\beta}) = 0$$

$$-\frac{1}{2} Tr \left[ (\widehat{\sigma}^2I + \widehat{\sigma}_\gamma^2\mathbf{Z}\mathbf{Z}')^{-1} \right] + \frac{1}{2} (\mathbf{y} - \mathbf{X}\widehat{\beta})' (\widehat{\sigma}^2I + \widehat{\sigma}_\gamma^2\mathbf{Z}\mathbf{Z}')^{-2} (\mathbf{y} - \mathbf{X}\widehat{\beta}) = 0$$

$$-\frac{1}{2}Tr \left[ \left( \widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}' \right)^{-1} \mathbf{Z}\mathbf{Z}' \right] + \frac{1}{2} (\mathbf{y} - \mathbf{X}\widehat{\beta})' \left( \widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}' \right)^{-1} \mathbf{Z}\mathbf{Z}' \left( \widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}' \right)^{-1} (\mathbf{y} - \mathbf{X}\widehat{\beta}) = 0$$

On en déduit que  $\widehat{\beta} = \left( \mathbf{X}'\widehat{V}^{-1}\mathbf{X} \right)^{-1} \mathbf{X}'\widehat{V}^{-1}\mathbf{y}$ , expression qu'on peut injecter dans les deux équations précédentes pour obtenir :

$$Tr \left( \widehat{V}^{-1} \right) = \mathbf{y}'\widehat{P}^2\mathbf{y}$$

$$Tr \left( \widehat{V}^{-1}\mathbf{Z}\mathbf{Z}' \right) = \mathbf{y}'\widehat{P}\mathbf{Z}\mathbf{Z}'\widehat{P}\mathbf{y}$$

avec :  $\widehat{V} = \widehat{\sigma}_\gamma^2 \mathbf{Z}\mathbf{Z}' + \widehat{\sigma}^2 I$  et  $\widehat{P} = \widehat{V}^{-1} - \widehat{V}^{-1}\mathbf{X} \left( \mathbf{X}'\widehat{V}^{-1}\mathbf{X} \right)^{-1} \mathbf{X}'\widehat{V}^{-1}$

On admet<sup>2</sup> ici le comportement asymptotiquement normal des estimateurs et le fait que la matrice de variance asymptotique est égale à l'inverse de la matrice d'information multiplié par  $n$ . La matrice d'information a pour expression :

$$E \left( \begin{array}{ccc} \frac{\partial \log \mathcal{L}}{\partial \beta} \frac{\partial \log \mathcal{L}'}{\partial \beta} & \frac{\partial \log \mathcal{L}}{\partial \beta} \frac{\partial \log \mathcal{L}}{\partial \sigma^2} & \frac{\partial \log \mathcal{L}}{\partial \beta} \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \\ \frac{\partial \log \mathcal{L}}{\partial \sigma^2} \frac{\partial \log \mathcal{L}'}{\partial \beta} & \frac{\partial \log \mathcal{L}}{\partial \sigma^2} \frac{\partial \log \mathcal{L}}{\partial \sigma^2} & \frac{\partial \log \mathcal{L}}{\partial \sigma^2} \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \\ \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \frac{\partial \log \mathcal{L}'}{\partial \beta} & \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \frac{\partial \log \mathcal{L}}{\partial \sigma^2} & \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \end{array} \right)$$

De plus, les dérivées secondes de la log-vraisemblance permettent de calculer plus facilement la matrice d'information car on a :

$$E \left( \begin{array}{ccc} \frac{\partial \log \mathcal{L}}{\partial \beta} \frac{\partial \log \mathcal{L}'}{\partial \beta} & \frac{\partial \log \mathcal{L}}{\partial \beta} \frac{\partial \log \mathcal{L}}{\partial \sigma^2} & \frac{\partial \log \mathcal{L}}{\partial \beta} \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \\ \frac{\partial \log \mathcal{L}}{\partial \sigma^2} \frac{\partial \log \mathcal{L}'}{\partial \beta} & \frac{\partial \log \mathcal{L}}{\partial \sigma^2} \frac{\partial \log \mathcal{L}}{\partial \sigma^2} & \frac{\partial \log \mathcal{L}}{\partial \sigma^2} \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \\ \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \frac{\partial \log \mathcal{L}'}{\partial \beta} & \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \frac{\partial \log \mathcal{L}}{\partial \sigma^2} & \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \frac{\partial \log \mathcal{L}}{\partial \sigma_\gamma^2} \end{array} \right) = -E \left( \begin{array}{ccc} \frac{\partial^2 \log \mathcal{L}}{\partial \beta \partial \beta'} & \frac{\partial^2 \log \mathcal{L}}{\partial \beta \partial \sigma^2} & \frac{\partial^2 \log \mathcal{L}}{\partial \beta \partial \sigma_\gamma^2} \\ \frac{\partial^2 \log \mathcal{L}}{\partial \sigma^2 \partial \beta'} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma^2 \partial \sigma^2} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma^2 \partial \sigma_\gamma^2} \\ \frac{\partial^2 \log \mathcal{L}}{\partial \sigma_\gamma^2 \partial \beta'} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma_\gamma^2 \partial \sigma^2} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma_\gamma^2 \partial \sigma_\gamma^2} \end{array} \right)$$

$$E \left( \begin{array}{ccc} \frac{\partial^2 \log \mathcal{L}}{\partial \beta \partial \beta'} & \frac{\partial^2 \log \mathcal{L}}{\partial \beta \partial \sigma^2} & \frac{\partial^2 \log \mathcal{L}}{\partial \beta \partial \sigma_\gamma^2} \\ \frac{\partial^2 \log \mathcal{L}}{\partial \sigma^2 \partial \beta'} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma^2 \partial \sigma^2} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma^2 \partial \sigma_\gamma^2} \\ \frac{\partial^2 \log \mathcal{L}}{\partial \sigma_\gamma^2 \partial \beta'} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma_\gamma^2 \partial \sigma^2} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma_\gamma^2 \partial \sigma_\gamma^2} \end{array} \right) =$$

$$E \left( \begin{array}{ccc} -\mathbf{x}'\mathbf{v}^{-1}\mathbf{x} & -\mathbf{x}'\mathbf{v}^{-2}(\mathbf{y} - \mathbf{x}\beta) & -\mathbf{x}'\mathbf{v}^{-1}\mathbf{z}\mathbf{z}'\mathbf{v}^{-1}(\mathbf{y} - \mathbf{x}\beta) \\ -(\mathbf{y} - \mathbf{x}\beta)'\mathbf{v}^{-2}\mathbf{x} & \frac{1}{2}Tr(\mathbf{v}^{-2}) - (\mathbf{y} - \mathbf{x}\beta)'\mathbf{v}^{-3}(\mathbf{y} - \mathbf{x}\beta) & \frac{1}{2}Tr(\mathbf{v}^{-2}\mathbf{z}\mathbf{z}') - (\mathbf{y} - \mathbf{x}\beta)'\mathbf{v}^{-2}\mathbf{z}\mathbf{z}'\mathbf{v}^{-1}(\mathbf{y} - \mathbf{x}\beta) \\ -(\mathbf{y} - \mathbf{x}\beta)'\mathbf{v}^{-1}\mathbf{z}\mathbf{z}'\mathbf{v}^{-1}\mathbf{x} & \frac{1}{2}Tr(\mathbf{v}^{-2}\mathbf{z}\mathbf{z}') - (\mathbf{y} - \mathbf{x}\beta)'\mathbf{v}^{-2}\mathbf{z}\mathbf{z}'\mathbf{v}^{-1}(\mathbf{y} - \mathbf{x}\beta) & \frac{1}{2}Tr(\mathbf{v}^{-1}\mathbf{z}\mathbf{z}'\mathbf{v}^{-1}\mathbf{z}\mathbf{z}') - (\mathbf{y} - \mathbf{x}\beta)'\mathbf{v}^{-1}\mathbf{z}\mathbf{z}'\mathbf{v}^{-1}\mathbf{z}\mathbf{z}'\mathbf{v}^{-1}(\mathbf{y} - \mathbf{x}\beta) \end{array} \right)$$

Comme  $E(\mathbf{y} - \mathbf{X}\beta|\mathbf{X}\mathbf{Z}) = 0$  et  $E((\mathbf{y} - \mathbf{X}\beta)'M(\mathbf{y} - \mathbf{X}\beta)|\mathbf{X}\mathbf{Z}) = Tr(MV)$  pour toutes matrice  $M$  fonction de  $\mathbf{X}$  et  $\mathbf{Z}$ , on a :

$$-E \left( \begin{array}{ccc} \frac{\partial^2 \log \mathcal{L}}{\partial \beta \partial \beta'} & \frac{\partial^2 \log \mathcal{L}}{\partial \beta \partial \sigma^2} & \frac{\partial^2 \log \mathcal{L}}{\partial \beta \partial \sigma_\gamma^2} \\ \frac{\partial^2 \log \mathcal{L}}{\partial \sigma^2 \partial \beta'} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma^2 \partial \sigma^2} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma^2 \partial \sigma_\gamma^2} \\ \frac{\partial^2 \log \mathcal{L}}{\partial \sigma_\gamma^2 \partial \beta'} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma_\gamma^2 \partial \sigma^2} & \frac{\partial^2 \log \mathcal{L}}{\partial \sigma_\gamma^2 \partial \sigma_\gamma^2} \end{array} \right) = E \left( \begin{array}{ccc} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & 0 & 0 \\ 0 & \frac{1}{2}Tr(\mathbf{V}^{-2}) & \frac{1}{2}Tr(\mathbf{V}^{-2}\mathbf{Z}\mathbf{Z}') \\ 0 & \frac{1}{2}Tr(\mathbf{V}^{-2}\mathbf{Z}\mathbf{Z}') & \frac{1}{2}Tr(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}') \end{array} \right)$$

■

Nous allons présenter un estimateur légèrement différent de l'estimateur du maximum de vraisemblance : l'estimateur "restricted maximum likelihood" que nous pourrions traduire par maximum de vraisemblance restreint. Il a déjà été rappelé (cf. chapitre 2) que dans le cas du modèle

2. Ce résultat a été établi par Hartley et Rao en 1967 pour les paramètres  $(\beta, \sigma, \frac{\sigma^2}{\sigma_\gamma^2})$ . On peut alors retrouver la propriété sur les paramètres  $(\beta, \sigma^2, \sigma_\gamma^2)$  car  $(\beta, \sigma, \frac{\sigma^2}{\sigma_\gamma^2}) \rightarrow (\beta, \sigma, \sigma_\gamma^2)$  est une fonction suffisamment régulière (différentiable) sur  $\mathbb{R}^{K+1} \times \mathbb{R}^{+*} \times \mathbb{R}^{+*}$

linéaire "simple", l'estimateur du maximum de vraisemblance de  $\sigma^2$  était biaisé à distance finie d'un facteur multiplicatif  $\frac{K+1}{n-K-1}$ , à la différence de l'estimateur des moindres carrés ordinaires. Cette mauvaise propriété à distance finie de l'estimateur du maximum de vraisemblance provient du fait que le nombre de degrés de liberté du problème est mal pris en compte ( $n$  au lieu de  $n - K - 1$ ). Dans le cas plus complexe qui nous intéresse ici, l'estimateur du maximum de vraisemblance présente le même défaut. C'est pour cela que des estimateurs alternatifs sont parfois utilisés (en particulier dans les procédures de SAS).

### Propriétés de l'estimateur du MVR

Sous les hypothèses  $(H_1'')$  et  $(H_n'')$ , l'estimateur du maximum de vraisemblance restreint de  $\sigma_\gamma^2$  et de  $\sigma^2$  est construit sur la vraisemblance de  $T\mathbf{y}$ , qui sont  $n - K - 1$  combinaisons linéaires des observations  $\mathbf{y}$  telles que  $\text{rg}(T) = n - K - 1$  et  $T\mathbf{X} = 0$  :

$$T\mathbf{y}|\mathbf{X}, \mathbf{Z} \sim \mathcal{N}(0, T(\sigma_\gamma^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 I)T')$$

$$\left(\widehat{\sigma_\gamma^2}, \widehat{\sigma^2}\right) = \arg \max_{\sigma_\gamma^2, \sigma^2} \mathcal{L}_{TY|X,Z}(\sigma_\gamma^2, \sigma^2)$$

Dans un second temps, l'estimateur du maximum de vraisemblance restreint de  $\beta$  est construit sur la vraisemblance de  $\mathbf{y}$  dans laquelle les quantités  $\sigma_\gamma^2$  et  $\sigma^2$  sont remplacées par leurs estimations :

$$\widehat{\beta} = \arg \max_{\beta} \mathcal{L}_{Y|X,Z}(\beta, \widehat{\sigma_\gamma^2}, \widehat{\sigma^2})$$

La solution de ce programme ne dépend pas du choix de  $T$  et elle est définie par les trois équations :

$$\widehat{\beta} = \left(\mathbf{X}'\widehat{V}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\widehat{V}^{-1}\mathbf{y}$$

$$\text{Tr}(\widehat{P}) = \mathbf{y}'\widehat{P}^2\mathbf{y}$$

$$\text{Tr}(\widehat{P}\mathbf{Z}\mathbf{Z}') = \mathbf{y}'\widehat{P}\mathbf{Z}\mathbf{Z}'\widehat{P}\mathbf{y}$$

avec :  $\widehat{V} = \widehat{\sigma_\gamma^2}\mathbf{Z}\mathbf{Z}' + \widehat{\sigma^2}I$  et  $\widehat{P} = \widehat{V}^{-1} - \widehat{V}^{-1}\mathbf{X}\left(\mathbf{X}'\widehat{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{V}^{-1}$

Pour  $\sigma^2 > 0$  et  $\sigma_\gamma^2 > 0$ , cet estimateur est asymptotiquement sans biais, et asymptotiquement optimal parmi les estimateurs asymptotiquement sans biais.

De plus pour  $\sigma^2 > 0$  et  $\sigma_\gamma^2 > 0$  :

$$\sqrt{n}(\widehat{\beta} - \beta) \longrightarrow \mathcal{N}(0, V_{as,\beta})$$

$$\sqrt{n}\left(\left(\begin{array}{c} \widehat{\sigma_\gamma^2} \\ \widehat{\sigma^2} \end{array}\right) - \left(\begin{array}{c} \sigma_\gamma^2 \\ \sigma^2 \end{array}\right)\right) \longrightarrow \mathcal{N}\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), V_{as,\sigma}\right)$$

où la variance asymptotique des estimateurs du maximum de vraisemblance restreint est :

$$V_{as,\beta} = \lim_{n \rightarrow +\infty} n(\mathbf{X}'V^{-1}\mathbf{X})^{-1}$$

$$V_{as,\sigma} = \lim_{n \rightarrow +\infty} 2n \begin{bmatrix} \text{Tr}(P\mathbf{Z}\mathbf{Z}'P\mathbf{Z}\mathbf{Z}') & \text{Tr}(P\mathbf{Z}\mathbf{Z}'P) \\ \text{Tr}(P\mathbf{Z}\mathbf{Z}'P) & \text{Tr}(P^2) \end{bmatrix}^{-1}$$

avec  $V = \sigma_\gamma^2\mathbf{Z}\mathbf{Z}' + \sigma^2I$  et  $P = V^{-1} - V^{-1}\mathbf{X}\left(\mathbf{X}'V^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'V^{-1}$

**Remarque :** l'expression de l'estimateur (et *a fortiori* sa loi) ne dépend pas du choix de  $T$ . Cela peut sembler surprenant, il est en fait logique que les estimateurs de  $\sigma_\gamma^2$  et  $\sigma^2$  ne dépendent pas de l'application linéaire particulière utilisée pour projeter les observations sur l'espace orthogonal à  $\mathbf{X}$ .

Preuve :

□

La vraisemblance (conditionnelle à  $\mathbf{X}, \mathbf{Z}$ ) de  $T\mathbf{y}$  s'écrit :

$$\mathcal{L}_{TY|X,Z}(\sigma_\gamma^2, \sigma^2) = [\det(T(\sigma_\gamma^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 I)T')]^{-\frac{1}{2}} * (2\pi)^{-\frac{n-K-1}{2}} * e^{-\frac{1}{2}\mathbf{y}'T'(T(\sigma_\gamma^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 I)T')^{-1}T\mathbf{y}}$$

Les valeurs qui maximisent la vraisemblance maximisent également la log-vraisemblance :

$$\log(\mathcal{L}_{TY|X,Z}(\sigma_\gamma^2, \sigma^2)) = -\frac{1}{2}\det(T(\sigma_\gamma^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 I)T') - \frac{1}{2}\mathbf{y}'T'(T(\sigma_\gamma^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 I)T')^{-1}T\mathbf{y} + cst$$

Les conditions du premier ordre sont alors :

$$\frac{\partial \log(\mathcal{L}_{TY|X,Z}(\widehat{\sigma}_\gamma^2, \widehat{\sigma}^2))}{\partial \sigma^2} = 0 \Leftrightarrow -\frac{1}{2}Tr(T'(T\widehat{V}T')^{-1}T) + \frac{1}{2}\mathbf{y}'T'(T\widehat{V}T')^{-1}TT'(T\widehat{V}T')^{-1}T\mathbf{y} = 0$$

avec  $\widehat{V} = \widehat{\sigma}^2 I + \widehat{\sigma}_\gamma^2$ .

Comme  $\widehat{V}$  est symétrique et définie positive, il existe une unique matrice symétrique définie positive, notée  $\widehat{V}^{\frac{1}{2}}$ , telle que  $(\widehat{V}^{\frac{1}{2}})^2 = \widehat{V}$ . On note  $\widehat{V}^{-\frac{1}{2}}$  l'inverse de  $\widehat{V}^{\frac{1}{2}}$ .

La matrice  $\widehat{M} = \widehat{V}^{\frac{1}{2}}T'(T\widehat{V}T')^{-1}T\widehat{V}^{\frac{1}{2}}$  vérifie alors :  $\widehat{M}^2 = \widehat{M}$ ,  $\widehat{M}' = \widehat{M}$ ,  $Tr(\widehat{M}) = n - K - 1 = n - rg(\widehat{V}^{-\frac{1}{2}}\mathbf{X})$  et  $\widehat{M}\widehat{V}^{-\frac{1}{2}}\mathbf{X} = \widehat{V}^{\frac{1}{2}}T'(T\widehat{V}T')^{-1}T\mathbf{X} = 0$ .  $\widehat{M}$  est donc le projecteur orthogonal sur l'orthogonal de  $\widehat{V}^{-\frac{1}{2}}\mathbf{X}$ , il peut donc s'écrire sous la forme  $\widehat{M} = I - \widehat{V}^{-\frac{1}{2}}\mathbf{X}(\mathbf{X}'\widehat{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{V}^{-\frac{1}{2}}$ .

Finalement :

$$\frac{\partial \log(\mathcal{L}_{TY|X,Z}(\widehat{\sigma}_\gamma^2, \widehat{\sigma}^2))}{\partial \sigma^2} = 0 \Leftrightarrow Tr(\widehat{P}) = \mathbf{y}'\widehat{P}^2\mathbf{y}$$

De même :

$$\frac{\partial \log(\mathcal{L}_{TY|X,Z}(\widehat{\sigma}_\gamma^2, \widehat{\sigma}^2))}{\partial \sigma_\gamma^2} = 0 \Leftrightarrow -\frac{1}{2}Tr(T'(T\widehat{V}T')^{-1}T\mathbf{Z}\mathbf{Z}') + \frac{1}{2}\mathbf{y}'T'(T\widehat{V}T')^{-1}T\mathbf{Z}\mathbf{Z}'T'(T\widehat{V}T')^{-1}T\mathbf{y} = 0$$

$$\frac{\partial \log(\mathcal{L}_{TY|X,Z}(\widehat{\sigma}_\gamma^2, \widehat{\sigma}^2))}{\partial \sigma_\gamma^2} = 0 \Leftrightarrow -\frac{1}{2}Tr(\widehat{P}\mathbf{Z}\mathbf{Z}') + \frac{1}{2}\mathbf{y}'\widehat{P}\mathbf{Z}\mathbf{Z}'\widehat{P}\mathbf{y} = 0$$

A  $\widehat{\sigma}^2$  et  $\widehat{\sigma}_\gamma^2$  fixé, la maximisation de la vraisemblance de  $\mathbf{y}|\mathbf{X}, \mathbf{Z}$  en  $\beta$  implique :

$$\widehat{\beta} = (\mathbf{X}'\widehat{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{V}^{-1}\mathbf{y}$$

La matrice de variance-covariance asymptotique des estimateurs peut se déduire de la matrice d'information :

$$E \left( \begin{array}{cc} \frac{\partial^2 \log \mathcal{L}_{TY|XZ}}{\partial \sigma^2 \partial \sigma^2} & \frac{\partial^2 \log \mathcal{L}_{TY|XZ}}{\partial \sigma^2 \partial \sigma_\gamma^2} \\ \frac{\partial^2 \log \mathcal{L}_{TY|XZ}}{\partial \sigma^2 \partial \sigma_\gamma^2} & \frac{\partial^2 \log \mathcal{L}_{TY|XZ}}{\partial \sigma_\gamma^2 \partial \sigma_\gamma^2} \end{array} \right) =$$

$$E \left( \begin{array}{cc} -\frac{1}{2}Tr(P^2) + \mathbf{y}'P^3\mathbf{y} & -\frac{1}{2}Tr(P\mathbf{Z}\mathbf{Z}'P) + \frac{1}{2}\mathbf{y}'[P^2\mathbf{Z}\mathbf{Z}'P + P\mathbf{Z}\mathbf{Z}'P^2]\mathbf{y} \\ -\frac{1}{2}Tr(P\mathbf{Z}\mathbf{Z}'P) + \frac{1}{2}\mathbf{y}'[P^2\mathbf{Z}\mathbf{Z}'P + P\mathbf{Z}\mathbf{Z}'P^2]\mathbf{y} & -\frac{1}{2}Tr(P\mathbf{Z}\mathbf{Z}'P\mathbf{Z}\mathbf{Z}') + \mathbf{y}'P\mathbf{Z}\mathbf{Z}'P\mathbf{Z}\mathbf{Z}'P\mathbf{y} \end{array} \right)$$

Comme  $E(\mathbf{y}'P^3\mathbf{y}) = E(E[Tr(P^3\mathbf{y}\mathbf{y}')|\mathbf{X}, \mathbf{Z}]) = E(Tr[P^3(\mathbf{X}\beta\beta'\mathbf{X}' + V)])$ ,  $PX = 0$  et  $Tr(P^3V) = Tr(P^2)$ , on a bien  $E(-\frac{1}{2}Tr(P^2) + \mathbf{y}'P^3\mathbf{y}) = \frac{Tr(P^2)}{2}$ . De même :  $Tr(P^2\mathbf{Z}\mathbf{Z}'PV) = Tr(P\mathbf{Z}\mathbf{Z}'P)$ ,  $Tr(P\mathbf{Z}\mathbf{Z}'P^2V) = Tr(P\mathbf{Z}\mathbf{Z}'P)$  et  $Tr(P\mathbf{Z}\mathbf{Z}'P\mathbf{Z}\mathbf{Z}'PV) = Tr(P\mathbf{Z}\mathbf{Z}'P\mathbf{Z}\mathbf{Z}')$

La variance asymptotique de  $\widehat{\beta}$  se déduit du fait que  $\widehat{\beta}$  est un estimateur des MCQG, dont on a étudié la loi limite précédemment.

■

#### 4.4.7 Comparaison de l'estimateur du MV et de l'estimateur du MVR

Les deux estimateurs présentés dans le paragraphe précédent ont les mêmes propriétés asymptotiques. Ils diffèrent donc par leur comportement à distance finie. Ainsi, si l'on se pose la question de savoir quel estimateur est préférable, il faut se concentrer sur les propriétés à distance finie (espérance, variance des estimations à taille d'échantillon donnée). Ces estimateurs n'étant définis qu'implicitement comme solution d'équations non linéaires, il est difficile de les calculer explicitement en fonction des observations. Dès lors, la comparaison entre les deux estimateurs est délicate. D'un point de vue algébrique, dans le cas le plus général, il n'y a pas de raisons admises par la communauté des économètres statisticiens pour préférer l'un à l'autre. Des comparaisons sur données simulées pourraient être envisagées mais là encore il n'existe pas de consensus net sur cette question à notre connaissance. Par contre, dans certains cas particuliers, les calculs peuvent être menés explicitement et l'estimateur MVR apparaît alors préférable à l'estimateur MV. Nous allons donc développer deux exemples particuliers, celui du modèle linéaire gaussien en l'absence d'effets "établissement" et celui où les covariables se réduisent à la constante.

##### Comparaison sous l'hypothèse $\sigma_\gamma^2 = 0$

Dans le sous-modèle particulier où  $\sigma_\gamma^2 = 0$  (cas correspondant au modèle linéaire exposé au chapitre 2), les estimateurs du maximum de vraisemblance et du maximum de vraisemblance restreint ont pour expression :

$$\widehat{\beta}_{MV} = \widehat{\beta}_{MVR} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\widehat{\sigma}_{MV}^2 = \frac{\widehat{v}'\widehat{v}}{n} \quad \text{et} \quad \widehat{\sigma}_{MVR}^2 = \frac{\widehat{v}'\widehat{v}}{n-K-1}$$

Dans ce cas particulier, l'estimateur du maximum de vraisemblance restreint est donc égal à l'estimateur des moindres carrés ordinaires. L'espérance des estimateurs précédents est alors :

$$E(\widehat{\beta}_{MV}) = E(\widehat{\beta}_{MVR}) = \beta$$

$$E[\widehat{\sigma}_{MV}^2] = \frac{n-K-1}{n}\sigma^2 \quad \text{et} \quad E[\widehat{\sigma}_{MVR}^2] = \sigma^2$$

L'estimateur du maximum de vraisemblance restreint estime donc sans biais à distance finie le paramètre  $\sigma^2$ . À distance finie, dans ce cas particulier, il est donc préférable à l'estimateur du maximum de vraisemblance.

Preuve :

□

L'expression et les propriétés de l'estimateur du maximum de vraisemblance ont déjà été présentées au chapitre 2.

Les équations définissant l'estimateur du maximum de vraisemblance restreint sont :

$$\widehat{\beta}_{MVR} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$Tr\left(\frac{1}{\widehat{\sigma}_{MVR}^2}M\right) = \left(\frac{1}{\widehat{\sigma}_{MVR}^2}\right)^2 \mathbf{y}'M^2\mathbf{y} \quad \text{avec} \quad M = I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

$M$  est le projecteur orthogonal sur l'espace orthogonal à  $\mathbf{X}$ , on a donc  $Tr(M) = n - K - 1$  et  $\widehat{v} = M\mathbf{y}$ . Cela implique donc que :  $\widehat{\sigma}_{MVR}^2 = \frac{\widehat{v}'\widehat{v}}{n-K-1}$

L'estimateur du maximum de vraisemblance restreint de  $\sigma^2$  correspond donc à l'estimateur des MCO dont les propriétés ont été exposées dans le chapitre 2.

■

##### Comparaison dans le modèle sans covariables $\mathbf{X}$

Dans le sous-modèle où  $\mathbf{X} = \mathbf{x}_0$  (cas où seule la constante est introduite dans le modèle) et où le nombre  $n_1$  d'élèves par classe est constant, les estimateurs du maximum de vraisemblance et du maximum de vraisemblance restreint ont pour expression :

$$\widehat{\beta}_{MV} = \widehat{\beta}_{MVR} = \frac{1}{n}\mathbf{x}'_0\mathbf{y} = \frac{1}{n}\sum_{i=1}^n y_{i,j}$$

$$\widehat{\sigma_{MVR}^2} = \widehat{\sigma_{MV}^2} = \frac{1}{n(n_1 - 1)} [n_1 \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{Z}\mathbf{Z}'\mathbf{y}]$$

$$\widehat{\sigma_{\gamma MV}^2} = \frac{1}{n(n_1 - 1)} [\mathbf{y}'\mathbf{Z}\mathbf{Z}'\mathbf{y} - \mathbf{y}'\mathbf{y}] - \frac{1}{n^2} \mathbf{y}'\mathbf{x}_0\mathbf{x}_0'\mathbf{y}$$

$$\widehat{\sigma_{\gamma MVR}^2} = \frac{1}{n(n_1 - 1)} [n_1 \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{Z}\mathbf{Z}'\mathbf{y}]$$

Les espérances des estimateurs précédents sont :

$$E(\widehat{\beta}_{MV}) = E(\widehat{\beta}_{MVR}) = \beta$$

$$E(\widehat{\sigma_{MV}^2}) = E(\widehat{\sigma_{MVR}^2}) = \sigma^2$$

$$E(\widehat{\sigma_{\gamma MV}^2}) = \sigma_\gamma^2 - \frac{1}{n}(\sigma^2 + n_1\sigma_\gamma^2) \quad \text{et} \quad E(\widehat{\sigma_{\gamma MVR}^2}) = \sigma_\gamma^2$$

Contrairement à l'estimateur du maximum de vraisemblance, l'estimateur du maximum de vraisemblance restreint estime sans biais le paramètre  $\sigma_\gamma^2$ . Dans ce cas particulier, l'estimateur du maximum de vraisemblance restreint est donc préférable à l'estimateur du maximum de vraisemblance à distance finie.

Preuve :

□

Concernant les estimateurs de  $\beta$ , le calcul donne :

$$\widehat{V}^{-1} = \frac{1}{\sigma^2} I - \frac{\sigma_\gamma^2}{\sigma^2} \frac{1}{\sigma^2 + n_1\sigma_\gamma^2} \mathbf{Z}\mathbf{Z}'$$

Comme  $\mathbf{x}_0'\mathbf{Z}\mathbf{Z}' = n_1\mathbf{x}_0'$ ,

$$\mathbf{x}_0'\widehat{V}^{-1}\mathbf{x}_0 = \frac{n}{\sigma^2 + n_1\sigma_\gamma^2} \quad \text{et} \quad \mathbf{x}_0'\widehat{V}^{-1}\mathbf{y} = (\widehat{\sigma^2} + n_1\widehat{\sigma_\gamma^2})\mathbf{x}_0'\mathbf{y}$$

Donc

$$\widehat{\beta}_{MV} = \widehat{\beta}_{MVR} = \frac{1}{n} \mathbf{x}_0'\mathbf{y} = \frac{1}{n} \sum_{i=1}^n y_{i,j}$$

Pour l'estimateur du maximum de vraisemblance des paramètres  $\sigma^2$  et  $\sigma_\gamma^2$ , on peut adopter les étapes de calcul suivantes :

$$\widehat{P} = \widehat{V}^{-1} - \widehat{V}^{-1}\mathbf{x}_0 \left( \mathbf{x}_0'\widehat{V}^{-1}\mathbf{x}_0 \right)^{-1} \mathbf{x}_0'\widehat{V}^{-1} = \frac{1}{\sigma^2} I - \frac{\sigma_\gamma^2}{\sigma^2} \frac{1}{\sigma^2 + n_1\sigma_\gamma^2} \mathbf{Z}\mathbf{Z}' - \frac{1}{n} \frac{1}{\sigma^2 + n_1\sigma_\gamma^2} \mathbf{x}_0\mathbf{x}_0' \quad \text{car} \quad \mathbf{x}_0\mathbf{x}_0'\mathbf{x}_0\mathbf{x}_0' = n\mathbf{x}_0\mathbf{x}_0', \quad \mathbf{Z}\mathbf{Z}'\mathbf{x}_0\mathbf{x}_0' = \mathbf{x}_0\mathbf{x}_0'\mathbf{Z}\mathbf{Z}' = n_1\mathbf{x}_0\mathbf{x}_0' \quad \text{et} \quad \mathbf{Z}\mathbf{Z}'\mathbf{Z}\mathbf{Z}' = n_1\mathbf{Z}\mathbf{Z}'.$$

$$\mathbf{y}'\widehat{P}^2\mathbf{y} = \left( \frac{1}{\sigma^2} \right)^2 \mathbf{y}'\mathbf{y} - \widehat{\sigma_\gamma^2} \left( \frac{1}{\sigma^2} \right)^2 \left[ \frac{1}{\sigma^2 + n_1\sigma_\gamma^2} + \frac{\widehat{\sigma^2}}{(\sigma^2 + n_1\widehat{\sigma_\gamma^2})^2} \right] \mathbf{y}'\mathbf{Z}\mathbf{Z}'\mathbf{y} - \frac{1}{n} \frac{1}{(\sigma^2 + n_1\widehat{\sigma_\gamma^2})^2} \mathbf{y}'\mathbf{x}_0\mathbf{x}_0'\mathbf{y}$$

$$\mathbf{y}'\widehat{P}\mathbf{Z}\mathbf{Z}'\widehat{P}\mathbf{y} = \left( \frac{1}{\sigma^2} \right)^2 \mathbf{y}'\mathbf{Z}\mathbf{Z}'\mathbf{y} - n_1\widehat{\sigma_\gamma^2} \left( \frac{1}{\sigma^2} \right)^2 \left[ \frac{1}{\sigma^2 + n_1\sigma_\gamma^2} + \frac{\widehat{\sigma^2}}{(\sigma^2 + n_1\widehat{\sigma_\gamma^2})^2} \right] \mathbf{y}'\mathbf{Z}\mathbf{Z}'\mathbf{y} - n_1 \frac{1}{n} \frac{1}{(\sigma^2 + n_1\widehat{\sigma_\gamma^2})^2} \mathbf{y}'\mathbf{x}_0\mathbf{x}_0'\mathbf{y} - \frac{1}{(\sigma^2 + n_1\widehat{\sigma_\gamma^2})^2} \mathbf{y}'\mathbf{Z}\mathbf{Z}'\mathbf{y} - n_1 \frac{1}{n} \frac{1}{(\sigma^2 + n_1\widehat{\sigma_\gamma^2})^2} \mathbf{y}'\mathbf{x}_0\mathbf{x}_0'\mathbf{y}$$

$$Tr(\widehat{V}^{-1}) = \frac{n}{\sigma^2} \left( 1 - \frac{\widehat{\sigma_\gamma^2}}{\sigma^2 + n_1\widehat{\sigma_\gamma^2}} \right)$$

$$Tr(\widehat{V}^{-1}\mathbf{Z}\mathbf{Z}') = \frac{n}{\sigma^2 + n_1\widehat{\sigma_\gamma^2}} = \frac{n}{\sigma^2} \left( 1 - \frac{n_1\widehat{\sigma_\gamma^2}}{\sigma^2 + n_1\widehat{\sigma_\gamma^2}} \right)$$

En considérant l'égalité  $n_1 Tr(\widehat{V}^{-1}) - Tr(\widehat{V}^{-1}\mathbf{Z}\mathbf{Z}') = n_1 \mathbf{y}'\widehat{P}^2\mathbf{y} - \mathbf{y}'\widehat{P}\mathbf{Z}\mathbf{Z}'\widehat{P}\mathbf{y}$ , on peut alors en déduire l'expression de l'estimateur du maximum de vraisemblance de  $\sigma^2$  :

$$\widehat{\sigma_{MV}^2} = \frac{1}{n(n_1 - 1)} [n_1 \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{Z}\mathbf{Z}'\mathbf{y}]$$

Puis en considérant l'égalité  $\mathbf{y}'\widehat{P}\mathbf{Z}\mathbf{Z}'\widehat{P}\mathbf{y} = Tr(\widehat{V}^{-1}\mathbf{Z}\mathbf{Z}')$ , on déduit que :

$$\widehat{\sigma_{\gamma MV}^2} = \frac{1}{n(n_1 - 1)} [\mathbf{y}'\mathbf{Z}\mathbf{Z}'\mathbf{y} - \mathbf{y}'\mathbf{y}] - \frac{1}{n^2} \mathbf{y}'\mathbf{x}_0\mathbf{x}_0'\mathbf{y}$$

Concernant l'estimateur du maximum de vraisemblance restreint de  $\sigma^2$  et de  $\sigma_\gamma^2$ , les calculs donnent :  $Tr(\hat{P}) = \frac{n}{\sigma^2} \left(1 - \frac{1}{n} \frac{\widehat{\sigma^2 + n\sigma_\gamma^2}}{\widehat{\sigma^2 + n_1\sigma_\gamma^2}}\right)$

$$Tr(\hat{PZZ}') = \frac{n}{\sigma^2} \left(1 - \frac{n_1}{n} \frac{\widehat{\sigma^2 + n\sigma_\gamma^2}}{\widehat{\sigma^2 + n_1\sigma_\gamma^2}}\right) = \frac{n-n_1}{\widehat{\sigma^2 + n_1\sigma_\gamma^2}}$$

En considérant l'égalité  $n_1 Tr(\hat{P}) - Tr(\hat{PZZ}') = n_1 \mathbf{y}' \hat{P}^2 \mathbf{y} - \mathbf{y}' \hat{PZZ}' \hat{P} \mathbf{y}$ , on peut alors en déduire l'expression de l'estimateur du maximum de vraisemblance restreint de  $\sigma^2$  :

$$\widehat{\sigma_{MVR}^2} = \frac{1}{n(n_1 - 1)} [n_1 \mathbf{y}' \mathbf{y} - \mathbf{y}' \mathbf{ZZ}' \mathbf{y}]$$

En considérant l'égalité  $\mathbf{y}' \hat{PZZ}' \hat{P} \mathbf{y} = Tr(\hat{PZZ}')$ , on déduit que :

$$\widehat{\sigma_{\gamma MVR}^2} = \frac{n-1}{n(n_1-1)(n-n_1) \mathbf{y}' \mathbf{ZZ}' \mathbf{y}} - \frac{1}{n(n-n_1)} \mathbf{y}' \mathbf{x}_0 \mathbf{x}_0' \mathbf{y} - \frac{1}{n(n_1-1)} \mathbf{y}' \mathbf{y}$$

Enfin comme  $E(\mathbf{x}'_0 \mathbf{y}) = n\beta$ ,  $E(\mathbf{y}' \mathbf{y}) = n(\sigma^2 + \sigma_\gamma^2) + n\beta^2$ ,  $E(\mathbf{y}' \mathbf{ZZ}' \mathbf{y}) = n(\sigma^2 + n_1\sigma_\gamma^2) + n\beta^2$ ,  $E(\mathbf{y}' \mathbf{x}_0 \mathbf{x}_0' \mathbf{y}) = n(\sigma^2 + n_1\sigma_\gamma^2) + n^2\beta^2$ , on a :

$$E(\widehat{\beta}_{MV}) = E(\widehat{\beta}_{MVR}) = \beta$$

$$E(\widehat{\sigma_{MV}^2}) = E(\widehat{\sigma_{MVR}^2}) = \sigma^2$$

$$E(\widehat{\sigma_{\gamma MV}^2}) = \sigma_\gamma^2 - \frac{1}{n} (\sigma^2 + n_1\sigma_\gamma^2)$$

$$E(\widehat{\sigma_{\gamma MVR}^2}) = \sigma_\gamma^2$$

■

## 4.5 Estimateurs des $\gamma$

Nous allons présenter rapidement quelques estimateurs de  $\gamma$ . Les estimateurs présentés précédemment dans ces pages étaient soit sans biais à distance finie, soit asymptotiquement sans biais. Dans le modèle à effets aléatoires, le nombre d'élèves par classe étant borné, on ne peut obtenir une estimation convergente, il restera toujours une variabilité de l'estimateur lié au nombre limité d'élèves dans chaque classe. Sous l'hypothèse de normalité des effets "classe"  $\gamma$ , on peut préférer privilégier la minimisation de l'écart quadratique moyen ( $E((\widehat{\gamma} - \gamma)'(\widehat{\gamma} - \gamma) | \mathbf{X}, \mathbf{Z})$ ) à l'absence de biais ( $E(\widehat{\gamma} | \gamma, \mathbf{X}, \mathbf{Z})$ ). Il convient alors de noter que **ces estimateurs sont biaisés**, dans le sens où, pour un  $\gamma$  donné,  $E(\widehat{\gamma} | \gamma) \neq \gamma$  ou  $E(\widehat{\gamma} | \mathbf{X}, \mathbf{Z}, \gamma) \neq \gamma$ .

### Estimateur de $\gamma$

Sous les hypothèses ( $H_1''$ ), ( $H_2''$ ) et ( $H_3''$ ), l'estimateur de  $\gamma$  de variance minimale parmi les estimateurs sans biais est :

$$\widehat{\gamma} = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' (\mathbf{y} - \mathbf{X}\beta)$$

Cet estimateur est simplement la moyenne par classe des écarts entre  $Y$  et le prédicteur  $X\beta$ .

Sous les hypothèses ( $H_1''$ ) et ( $H_n''$ ), l'estimateur de  $\gamma$  qui minimise l'écart quadratique moyen est :

$$\widehat{\gamma} = \sigma_\gamma^2 (\sigma_\gamma^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 \mathbf{I})^{-1} \mathbf{Z}' (\mathbf{y} - \mathbf{X}\beta)$$

Ce dernier estimateur n'est pas sans biais au sens fréquentiste du terme puisque  $E(\widehat{\gamma} | \gamma, \mathbf{X}, \mathbf{Z}) \neq \gamma$  et que  $E(\widehat{\gamma} | \gamma) \neq \gamma$ . Cet estimateur est qualifié de "BLUP" (best linear unbiased predictor), mais le qualificatif "unbiased" est ici à entendre comme  $E(\widehat{\gamma}) = E(\gamma) = 0$ , c'est-à-dire que c'est un estimateur de  $\gamma$  qui est d'espérance nulle. Cet estimateur peut par ailleurs être vu comme un estimateur bayésien car  $\sigma_\gamma^2 (\sigma_\gamma^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 \mathbf{I})^{-1} \mathbf{Z}' (\mathbf{y} - \mathbf{X}\beta) = E(\gamma | \mathbf{y}, \mathbf{X}, \mathbf{Z})$ .

Ces estimateurs dépendent de  $\beta$ , de  $\sigma^2$  et de  $\sigma_\gamma^2$  qui ne sont pas connus. À partir d'une estimation convergente de  $\sigma^2$  et  $\sigma_\gamma^2$ , on peut donc construire les estimateurs suivants :

– Si on privilégie l'absence de biais asymptotique

$$E \left( \widehat{\beta}, \widehat{\gamma} | \beta, \gamma, \mathbf{X}, \mathbf{Z} \right) \rightarrow (\beta, \gamma)$$

on obtient comme estimateur :

$$\begin{aligned} \widehat{\beta} &= \left( \mathbf{X}' \widehat{\mathbf{V}}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \widehat{\mathbf{V}}^{-1} \mathbf{y} \\ \widehat{\gamma} &= \left( \mathbf{Z}' \mathbf{Z} \right)^{-1} \mathbf{Z}' \left( \mathbf{y} - \mathbf{X} \widehat{\beta} \right) \end{aligned}$$

– Si on privilégie la minimisation de l'écart quadratique moyen

$$\text{Argmin} E \left[ \begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{\gamma} - \gamma \end{pmatrix} \begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{\gamma} - \gamma \end{pmatrix}' \mid \mathbf{y}, \mathbf{X}, \mathbf{Z} \right]$$

on obtient l'estimateur "E-BLUP" pour empirical best linear unbiased predictor :

$$\begin{aligned} \widehat{\beta} &= \left( \mathbf{X}' \widehat{\mathbf{V}}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \widehat{\mathbf{V}}^{-1} \mathbf{y} \\ \widehat{\gamma} &= \widehat{\sigma}^2_{\gamma} \left( \widehat{\sigma}^2_{\gamma} \mathbf{Z}' \mathbf{Z} + \sigma^2 I \right)^{-1} \mathbf{Z}' \left( \mathbf{y} - \mathbf{X} \widehat{\beta} \right) \end{aligned}$$

Dans les deux cas, les estimateurs de  $\beta$  seront asymptotiquement sans biais.

**Remarque :** Pour une classe donnée, les deux estimateurs de  $\gamma$  présentés ci-dessus seront d'autant plus proches que le nombre d'élèves dans la classe est grand. Cette propriété peut être vue comme un cas particulier du fait qu'estimateur fréquentiste et estimateur bayésien sont asymptotiquement équivalents dans un cadre standard (en particulier le nombre de paramètres ne dépend pas de la taille de l'échantillon).

Preuve :

□

Commençons par l'estimateur qui privilégie l'absence de biais : l'estimateur de  $\beta$  est un estimateur des MCQG, ses propriétés asymptotiques ont été étudiées précédemment. Concernant l'estimateur de  $\gamma$ , on a :  $E(\widehat{\gamma} | \gamma, \mathbf{X}, \mathbf{Z}) = \left( \mathbf{Z}' \mathbf{Z} \right)^{-1} \mathbf{Z}' E \left( \mathbf{y} - \mathbf{X} \widehat{\beta} | \mathbf{X}, \mathbf{Z} \right) = \left( \mathbf{Z}' \mathbf{Z} \right)^{-1} \mathbf{Z}' \mathbf{Z} \gamma + \mathbf{X} \left( E(\beta - \widehat{\beta} | \mathbf{X}, \mathbf{Z}) \right) + E(\epsilon | \mathbf{X}, \mathbf{Z}) = \gamma + \mathbf{X} \left( E(\beta - \widehat{\beta} | \mathbf{X}, \mathbf{Z}) \right) \rightarrow \gamma$ .

Concernant l'estimateur qui privilégie la minimisation de l'écart quadratique moyen, on peut remarquer que la distribution de  $\gamma | \mathbf{y}, \mathbf{X}, \mathbf{Z}$  est un vecteur gaussien d'espérance  $\sigma_{\gamma}^2 \left( \sigma_{\gamma}^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 I_C \right)^{-1} \mathbf{Z}' \left( \mathbf{y} - \mathbf{X} \beta \right)$  et de variance  $\sigma_{\gamma}^2 \sigma^2 \left( \sigma_{\gamma}^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 I_C \right)^{-1} = \left( \frac{\mathbf{Z}' \mathbf{Z}}{\sigma^2} + \frac{I_C}{\sigma_{\gamma}^2} \right)^{-1}$  :

$$\mathcal{L}_{\gamma | Y=\mathbf{y}, X=\mathbf{X}, Z=\mathbf{Z}}(g) = \frac{\mathcal{L}_{Y | \gamma, X=\mathbf{X}, Z=\mathbf{Z}}(\mathbf{y}) \mathcal{L}_{\gamma | X=\mathbf{X}, Z=\mathbf{Z}}(g)}{\mathcal{L}_{Y | X=\mathbf{X}, Z=\mathbf{Z}}(\mathbf{y})}$$

Or  $\mathbf{y} | \gamma, \mathbf{X}, \mathbf{Z}$  est un vecteur gaussien d'espérance  $\mathbf{X} \beta + \mathbf{Z} \gamma$  et de variance  $\sigma^2 I_n$  ;  $\gamma | \mathbf{X}, \mathbf{Z}$  est un vecteur gaussien d'espérance nulle et de variance  $\sigma_{\gamma}^2 I_C$  et  $\mathbf{y} | \mathbf{X}, \mathbf{Z}$  est un vecteur gaussien d'espérance  $\mathbf{X} \beta$  et de variance  $\sigma_{\gamma}^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 I_n$ .

$$\mathcal{L}_{\gamma | Y=\mathbf{y}, X=\mathbf{X}, Z=\mathbf{Z}}(g) = \sqrt{2\pi}^{-C} \frac{\sqrt{\det(\sigma_{\gamma}^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 I_n)^n}}{\sqrt{\sigma_{\gamma}^{2C} \sqrt{\sigma^2}^n}}$$

$$\exp \left[ -\frac{1}{2} \left\{ \frac{(\mathbf{y} - \mathbf{X} \beta - \mathbf{Z} g)' (\mathbf{y} - \mathbf{X} \beta - \mathbf{Z} g)}{\sigma^2} + \frac{g' g}{\sigma_{\gamma}^2} - (\mathbf{y} - \mathbf{X} \beta)' \left[ \sigma_{\gamma}^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 I_n \right]^{-1} (\mathbf{y} - \mathbf{X} \beta) \right\} \right].$$

En remarquant que  $\det(\sigma_{\gamma}^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 I_n) = \sigma^{2n} \sigma_{\gamma}^{2C} \prod_{c=1}^C \left( \frac{1}{\sigma_{\gamma}^2} + \frac{n_c}{\sigma^2} \right)$  et que  $\left[ \sigma_{\gamma}^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 I_n \right]^{-1} = \frac{1}{\sigma^2} I_n - \frac{\sigma^2}{\sigma^2} \mathbf{Z} \left( \sigma^2 I_C + \sigma_{\gamma}^2 \mathbf{Z}' \mathbf{Z} \right)^{-1} \mathbf{Z}'$ , il vient :

$$\mathcal{L}_{\gamma | Y=\mathbf{y}, X=\mathbf{X}, Z=\mathbf{Z}}(g) = \frac{1}{\sqrt{2\pi \det \left( \frac{\mathbf{Z}' \mathbf{Z}}{\sigma^2} + \frac{I_C}{\sigma_{\gamma}^2} \right)^C}}$$

$$\exp \left[ -\frac{1}{2} \left( g - \sigma_{\gamma}^2 \left( \sigma_{\gamma}^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 I_C \right)^{-1} \mathbf{Z}' \left( \mathbf{y} - \mathbf{X} \beta \right) \right)' \left( \frac{\mathbf{Z}' \mathbf{Z}}{\sigma^2} + \frac{I_C}{\sigma_{\gamma}^2} \right) \left( g - \sigma_{\gamma}^2 \left( \sigma_{\gamma}^2 \mathbf{Z}' \mathbf{Z} + \sigma^2 I_C \right)^{-1} \mathbf{Z}' \left( \mathbf{y} - \mathbf{X} \beta \right) \right) \right]$$

Donc  $\hat{\gamma} = E(\gamma|\mathbf{y}, \mathbf{X}, \mathbf{Z}) = \sigma_\gamma^2 (\sigma_\gamma^2 \mathbf{Z}'\mathbf{Z} + \sigma^2 I_C)^{-1} \mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta)$  minimise l'écart quadratique  $E((\hat{\gamma} - \gamma)'(\hat{\gamma} - \gamma)|\mathbf{y}, \mathbf{X}, \mathbf{Z})$ .

L'estimateur BLUP n'est pas calculable, car il dépend des paramètres inconnus  $\sigma_\gamma^2, \sigma^2$ . Cependant, comme des estimateurs convergents de ces quantités sont disponibles, il suffit de les remplacer par leurs estimations pour obtenir un estimateur ayant les mêmes propriétés asymptotiques.

Concernant le biais de l'estimateur "BLUP" de  $\gamma$ , on a :

$$E(\hat{\gamma}|\mathbf{X}, \mathbf{Z}, \gamma) = E\left(\sigma_\gamma^2 (\sigma_\gamma^2 \mathbf{Z}'\mathbf{Z} + \sigma^2 I)^{-1} \mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\beta})|\mathbf{X}, \mathbf{Z}, \gamma\right)$$

$$E(\hat{\gamma}|\mathbf{X}, \mathbf{Z}, \gamma) = \sigma_\gamma^2 (\sigma_\gamma^2 \mathbf{Z}'\mathbf{Z} + \sigma^2 I)^{-1} \mathbf{Z}'\left(\mathbf{Z} - \mathbf{X}(\mathbf{X}'V^{-1}\mathbf{X})^{-1} \mathbf{X}'V^{-1}\mathbf{Z}\right)\gamma$$

$$E(\hat{\gamma}|\mathbf{X}, \mathbf{Z}, \gamma) = \sigma_\gamma^2 (\sigma_\gamma^2 \mathbf{Z}'\mathbf{Z} + \sigma^2 I)^{-1} \mathbf{Z}'\mathbf{Z}\gamma - \sigma_\gamma^2 (\sigma_\gamma^2 \mathbf{Z}'\mathbf{Z} + \sigma^2 I)^{-1} \mathbf{Z}'\mathbf{X}(\mathbf{X}'V^{-1}\mathbf{X})^{-1} \mathbf{X}'V^{-1}\mathbf{Z}\gamma$$

$$E(\hat{\gamma}|\mathbf{X}, \mathbf{Z}, \gamma) = \gamma - \sigma^2 (\sigma_\gamma^2 \mathbf{Z}'\mathbf{Z} + \sigma^2 I)^{-1} \gamma - \sigma_\gamma^2 \mathbf{Z}'V^{-1}\mathbf{X}(\mathbf{X}'V^{-1}\mathbf{X})^{-1} \mathbf{X}'V^{-1}\mathbf{Z}\gamma$$

Comme  $\sigma^2 (\sigma_\gamma^2 \mathbf{Z}'\mathbf{Z} + \sigma^2 I)^{-1} + \sigma_\gamma^2 \mathbf{Z}'V^{-1}\mathbf{X}(\mathbf{X}'V^{-1}\mathbf{X})^{-1} \mathbf{X}'V^{-1}\mathbf{Z}$  n'a pas de raison d'être nul, on n'a pas  $E(\hat{\gamma}|\mathbf{X}, \mathbf{Z}, \gamma) = \gamma$ .

De même,  $E\left(\sigma^2 (\sigma_\gamma^2 \mathbf{Z}'\mathbf{Z} + \sigma^2 I)^{-1} + \sigma_\gamma^2 \mathbf{Z}'V^{-1}\mathbf{X}(\mathbf{X}'V^{-1}\mathbf{X})^{-1} \mathbf{X}'V^{-1}\mathbf{Z}|\gamma\right)$  n'a pas non plus de raison d'être nul, donc *a priori* on a  $E(\hat{\gamma}|\gamma) \neq \gamma$ .

L'estimateur "BLUP" de  $\gamma$  n'est donc pas sans biais (que l'on raisonne conditionnellement ou non aux covariables  $\mathbf{X}$  et  $\mathbf{Z}$ ).

Pour les mêmes raisons (sans parler du fait que les estimations de  $\sigma^2$  et  $\sigma_\gamma^2$  rendent l'examen du biais plus ardu), l'estimateur "E-BLUP" de  $\gamma$  est lui aussi *a priori* biaisé.

■

## 4.6 Les tests

### 4.6.1 Les tests sur $\beta$

Que l'on considère l'estimateur des MCQG, celui du maximum de vraisemblance ou celui des du maximum de vraisemblance restreint, l'expression de l'estimateur de  $\beta$  vérifie<sup>3</sup> :

$$\left(\mathbf{X}'\left(\widehat{\sigma^2}I + \widehat{\sigma_\gamma^2}\mathbf{Z}\mathbf{Z}'\right)^{-1}\mathbf{X}\right)^{\frac{1}{2}}\left(\hat{\beta} - \beta\right) \xrightarrow[n \rightarrow +\infty]{loi} N(0, I_{K+1})$$

À partir du moment où le nombre d'individus est suffisamment grand, des tests du type " $\beta_k = 0$ " s'obtiennent très classiquement en comparant la statistique de Student avec les seuils de la loi normale.

Les sorties de SAS donnent les p-value correspondantes.

### 4.6.2 Tests sur $\sigma_\gamma^2$

Les tests statistiques sont généralement développés dans un cadre où la valeur testée n'est pas "à la borne" : par exemple pour les tests de Student, l'hypothèse testée est  $\beta_k = 0$  mais rien n'empêche  $\beta_k$  d'être supérieur ou inférieur à 0. Or dans le cadre du modèle estimé, il est particulièrement pertinent de tester l'existence d'effets "classes", ce qui peut se traduire par l'hypothèse  $\sigma_\gamma^2 = 0$ . Dans ce cas, le test mis en oeuvre est un test "à la borne" car par nature  $\sigma_\gamma^2 \geq 0$ .

Comme on l'a vu, sous l'hypothèse  $\sigma_\gamma^2 > 0$ , la loi limite de l'estimateur est :

$$\mathcal{N}\left(\sigma_\gamma, V_{as, \sigma_\gamma^2}\right)$$

---

3. Sous les hypothèses des théorèmes énoncés précédemment

Pour effectuer le test  $\sigma_\gamma^2 = 0$ , il faut alors calculer la loi limite de  $\widehat{\sigma}_\gamma^2$  sous cette hypothèse. Dans ce cas, on peut alors montrer<sup>4</sup> que la loi limite de l'estimateur est un mélange de la loi limite précédente et d'un Dirac en 0. Ce qui revient au résultat intuitif simple que le test doit être mené comme un test unilatéral, ainsi dans le cas d'un test asymptotique de niveau 5 %, la statistique de Student doit être comparée à la valeur 1,64 et non 1,96. C'est ce que font par défaut les logiciels.

### 4.6.3 Les tests sur $\gamma$

Concernant les estimateurs de  $\gamma$ , ils sont tous fondés sur la statistique  $\mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\beta}) = \mathbf{Z}'\mathbf{X}(\beta - \widehat{\beta}) + \mathbf{Z}'\mathbf{Z}\gamma + \mathbf{Z}'\epsilon$ . Or  $(\beta - \widehat{\beta})$  est asymptotiquement convergent à la vitesse  $\sqrt{n}$  (ou  $\sqrt{C}$ , ce qui revient au même car le nombre d'élèves par classe est borné). Sous les hypothèses de normalité des résidus  $\mathbf{Z}'\epsilon$  est un vecteur gaussien.

Si  $\widehat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\beta})$ , alors sous les hypothèses de normalité :

$$\widehat{\gamma}_j - \gamma_j | \mathbf{X}, \mathbf{Z} \xrightarrow[n \rightarrow +\infty]{loi} \mathcal{N}(0, \sigma^2 n_j)$$

De plus, les différentes composantes de  $\widehat{\gamma}$  sont asymptotiquement indépendantes entre elles et sont indépendantes de  $\widehat{\beta}$ , les tests usuels peuvent donc facilement s'implémenter.

Si  $\widehat{\gamma} = \widehat{\sigma}_\gamma^2 \left( \widehat{\sigma}_\gamma^2 \mathbf{Z}'\mathbf{Z} + \widehat{\sigma}^2 I \right)^{-1} \mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\beta})$ , il est toujours possible de calculer la variance asymptotique de  $\widehat{\gamma} | \mathbf{X}, \mathbf{Z}$ , il faut cependant veiller au fait que les tests standards fondés sur les statistiques de type Student ou Fischer ne sont pas directement applicables car l'estimateur de  $\gamma$  est biaisé (au sens fréquentiste du terme), pour ce qui concerne l'inférence on utilisera donc plutôt l'estimateur usuel  $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\beta})$ .

## 4.7 Programmation sous SAS...

### 4.7.1 ... des MCQG

Une des manières simples de programmer les différents estimateurs du modèle à effets aléatoires consiste à utiliser la `proc mixed` de SAS. Le choix de l'estimateur retenu se paramètre au moyen de l'option `method`; pour les MCQG, il faut écrire `method=MIVQUE0`. Par ailleurs, si on ne s'intéresse qu'aux estimateurs de  $\beta$ , le modèle peut s'écrire de manière équivalente :

$$\mathbf{y} = \mathbf{X}\beta + v \quad V(v | \mathbf{X}, \mathbf{Z}) = \sigma_\gamma^2 \mathbf{Z}'\mathbf{Z} + \sigma^2 I$$

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon \quad V((\gamma', \epsilon') | \mathbf{X}, \mathbf{Z}) = \begin{bmatrix} \sigma_\gamma^2 I_C & 0 \\ 0 & \sigma^2 I_N \end{bmatrix}$$

et il y a donc deux moyens de programmer le même estimateur, un premier en utilisant l'instruction `random` pour spécifier un effet aléatoire, un deuxième (qui donne lieu à des calculs plus rapides) en utilisant l'instruction `repeated/subject= type=` pour préciser la forme bloc-diagonale de la matrice de variance-covariance (la `proc mixed` permet de spécifier une grande variété de formes de matrice de variance-covariance, et même de faire de l'inférence robuste à la forme de corrélation des inobservées intra-classe grâce à l'option `empirical`). L'option `covtest` permet de tester la nullité des paramètres  $\sigma_\gamma^2$  et  $\sigma^2$ , l'option `solution` après l'instruction `random` permet de demander le calcul<sup>5</sup> des  $\widehat{\gamma} = \widehat{\sigma}_\gamma^2 \left( \widehat{\sigma}_\gamma^2 \mathbf{Z}'\mathbf{Z} + \widehat{\sigma}^2 I \right)^{-1} \mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\beta})$  (cela peut prendre du temps).

```
proc mixed data=TABLE method=mivque0 covtest;
  classes Z;
  model Y=X1 X2 ... XK/solution;
  random Z/solution;
run;
```

ou

4. cf. Self et Liang (1987), pour plus de détails sur le sujet

5. Pour calculer les estimateurs  $\widehat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\beta})$ , qui nous semblent plus pertinents dans un cadre de statistique inférentielle classique, il faut procéder comme dans le chapitre précédent en calculant les moyennes par classe des quantités  $\mathbf{y} - \mathbf{X}\widehat{\beta}$ .

```
proc mixed data=TABLE method=mivque0 covtest;  
  class Z;  
  model Y=X1 X2 ... XK/solution;  
  repeated /subject=Z type=cs;  
run;
```

#### 4.7.2 ... ou des autres estimateurs

Les autres estimateurs se calculent en changeant simplement la valeur de `method=`, pour l'estimateur du maximum de vraisemblance on utilisera `method=ML`, pour celui du maximum de vraisemblance restreint on utilisera `method=REML` (à défaut de préciser la méthode d'estimation, c'est le REML que SAS calcule).

# Chapitre 5

## Quel modèle choisir ?

Dans les chapitres précédents, nous avons présenté deux types de modèles, les modèles à "effets fixes" et les modèles à "effets aléatoires". L'objectif de ce chapitre est de clarifier les enjeux d'une telle distinction. Pour cela, nous allons nous placer dans le cadre le plus général possible et décliner les propriétés des estimateurs selon les spécifications retenues. Dans un deuxième temps, nous donnerons un test statistique de choix de modèle, afin de retenir la spécification la plus cohérente avec les données.

### 5.1 Discussion théorique

#### 5.1.1 Un cadre commun de comparaison des modèles

Revenons au cas le plus général qui soit. On postule que notre output  $y_{ij}$  de l'élève  $i$  dans la classe  $j$  dépend de variables "individuelles"  $X_{ij}^1$ , de la composition de la classe selon ces variables  $X_j^1 = \frac{1}{n_j} \sum_{k=1}^{n_j} X_{kj}^1$  et de variables de "classe"  $X_j^2$  et d'un effet spécifique de la classe  $\gamma_j$  qui reflète l'influence de l'ensemble des variables inobservées communes aux élèves de la classe  $j$ . Parmi ces variables inobservées par le statisticien, il peut y avoir des "leviers" du décideur comme la qualité du professeur ou les conditions matérielles de la classe aussi bien que des variables relevant d'une sélection difficilement contrôlable par le décideur comme le fait que les élèves scolarisés au même endroit partagent des conditions d'existence communes au-delà de celles prises en compte dans la régression par les variables  $X_{ij}^1$ .

$$y_{ij} = X_{ij}^1 \beta_1 + X_j^1 \beta_2^1 + X_j^2 \beta_2^2 + \gamma_j + \varepsilon_{ij} \quad (1)$$

Par ailleurs, les  $\gamma_j$  peuvent être corrélés aux covariables. La régression linéaire de  $\gamma_j$  sur les covariables est alors :

$$EL(\gamma_j | X_{ij}^1, X_j^1, X_j^2) = X_j^1 \lambda_2^1 + X_j^2 \lambda_2^2 \quad (2)$$

Contrairement à l'équation (1), l'équation (2) n'a pas d'interprétation structurelle, elle signifie juste que parmi les variables inobservées communes au sein de la classe, certaines sont liées aux variables  $X^1$  ou  $X^2$ . Par exemple, une faible taille de classe a un effet positif sur la réussite scolaire ( $\beta_2^2 > 0$ ) mais pour cette raison, les élèves scolarisés dans de telles classes sont souvent ceux que le directeur d'école identifie comme présentant des risques d'échec élevés ( $\lambda_2^2 < 0$ ). De même, la composition de la classe peut exercer un effet direct via les effets de pairs ( $\beta_2^1 \neq 0$ ) mais on peut également penser que toutes variables  $X_{ij}^1$  égales par ailleurs, les caractéristiques inobservées des élèves sont corrélées à la composition sociale de leur classe ( $\lambda_2^1 \neq 0$ ).

#### Comparaison d'estimateurs

On suppose que le processus de génération des données est le suivant :

$$y_{ij} = X_{ij}^1 \beta_1 + X_j^1 \beta_2^1 + X_j^2 \beta_2^2 + \gamma_j + \varepsilon_{ij} \quad (1)$$

$$EL(\gamma_j | X_{ij}^1, X_j^1, X_j^2) = X_j^1 \lambda_2^1 + X_j^2 \lambda_2^2 \quad (2)$$

Pour estimer les paramètres de l'équation (1), on peut recourir à différentes estimations.

**Modèle à effets fixes** : on estime par MCO ou par estimation "Within" le modèle suivant :

$$y_{ij} = X_{ij}^1 \beta_1 + \gamma_j + \varepsilon_{ij}$$

Dans ce cas :

- ◊  $\widehat{\beta}_1$  estime sans biais  $\beta_1$
- ◊  $\widehat{\gamma}_j$  estime sans biais  $X_{.j}^1 \beta_2^1 + X_j^2 \beta_2^2 + \gamma_j$
- ◊ La régression de  $\widehat{\gamma}_j$  sur  $X_{.j}^1$  et  $X_j^2$  estime sans biais  $\beta_2^1 + \lambda_2^1$  et  $\beta_2^2 + \lambda_2^2$

**Modèle à effets aléatoires "complet" (ou modèle de Mundlak)** : on estime par MCQG ou MV ou MVR le modèle suivant :

$$y_{ij} = X_{ij}^1 \beta_1 + X_{.j}^1 \beta_2^1 + X_j^2 \beta_2^2 + v_{ij}$$

avec  $v_{ij} = \gamma_j + \varepsilon_{ij}$

Dans ce cas :

- ◊  $\widehat{\beta}_1$  estime asymptotiquement sans biais  $\beta_1$
- ◊  $\widehat{\beta}_2^1$  estime asymptotiquement sans biais  $\beta_2^1 + \lambda_2^1$ , c'est donc un estimateur biaisé de  $\beta_2^1$  dès que  $\lambda_2^1 \neq 0$
- ◊  $\widehat{\beta}_2^2$  estime asymptotiquement sans biais  $\beta_2^2 + \lambda_2^2$ , c'est donc un estimateur biaisé de  $\beta_2^2$  dès que  $\lambda_2^2 \neq 0$
- ◊  $\widehat{\gamma}_j$  est un estimateur biaisé et asymptotiquement biaisé de  $\gamma_j - X_{.j}^1 \lambda_2^1 - X_j^2 \lambda_2^2$  et de  $\gamma_j$

**Modèle à effets aléatoires "incomplet" (omission des moyennes de classe)** : on estime par MCQG ou MV ou MVR le modèle suivant :

$$y_{ij} = X_{ij}^1 \beta_1 + X_j^2 \beta_2^2 + v_{ij}$$

avec  $v_{ij} = \gamma_j + \varepsilon_{ij}$

- ◊  $\widehat{\beta}_1$  est un estimateur biaisé de  $\beta_1$  dès que  $\beta_2^1 + \lambda_2^1 \neq 0$
- ◊  $\widehat{\beta}_2^2$  est un estimateur biaisé de  $\beta_2^2 + \lambda_2^2$  dès que  $\beta_2^1 + \lambda_2^1 \neq 0$  et que  $\text{cov}(X_{.j}^1, X_j^2) \neq 0$
- ◊  $\widehat{\gamma}_j$  est un estimateur biaisé et asymptotiquement biaisé de  $\gamma_j - X_j^2 \lambda_2^2$  et de  $\gamma_j$

**En conclusion**, sans recourir à des stratégies de variables instrumentales :

- ◊ Dès que  $\lambda_2^1 \neq 0$  et/ou dès que  $\lambda_2^2 \neq 0$  (c'est-à-dire dès que les effets "classe" sont corrélés à des variables ayant un impact sur  $y_{ij}$ ), on ne peut estimer sans biais (ou asymptotiquement sans biais) que les quantités  $\beta_1, \beta_2^1 + \lambda_2^1$  et  $\beta_2^2 + \lambda_2^2$ . Les quantités  $\beta_2^1$  et  $\beta_2^2$  ne peuvent s'estimer sans biais (ou asymptotiquement sans biais) que si l'on suppose une absence de corrélation entre les distributions des effets "classe"  $\gamma_j$  et les autres variables.
- ◊ Pour obtenir des estimations asymptotiquement sans biais de  $\beta_1$  dans un modèle où on introduit des effets aléatoires au niveau de la classe (comme dans les modèles multi-niveaux), il est indispensable d'introduire les moyennes par classe pour les covariables telles que  $\lambda_2^1 \neq 0$ , même si les moyennes de classe n'ont pas d'impact causal sur les résultats ( $\beta_2^1 = 0$ ).

### 5.1.2 Le modèle à effets fixes est un sur-modèle du modèle à effets aléatoires

Dans le cadre de comparaison commun exposé dans le paragraphe précédent, il apparaît de manière assez claire que la différence entre modèle à effets fixes et modèle à effets aléatoires tient essentiellement à la nature de la liaison postulée entre les effets "classes"  $\gamma_j$  et les covariables  $X$ .

Dans le modèle à effets fixes, aucune hypothèse n'est faite sur la distribution des effets "classes" conditionnellement aux  $X$ . Dans le modèle à effets aléatoires, une hypothèse structurelle supplémentaire est faite (parfois implicitement) :  $E(\gamma|\mathbf{X}) = 0$  et  $V(\gamma|\mathbf{X}) = \sigma_\gamma^2$  ou même l'hypothèse paramétrique  $\gamma|\mathbf{X} \sim \mathcal{N}(0, \sigma_\gamma^2)$ . Ces deux hypothèses sont des hypothèses d'indépendance (soit sur les premiers moments, soit sur la distribution) des effets "classe" par rapport aux covariables  $\mathbf{X}$ .

Dans le cas où cette hypothèse d'indépendance est vérifiée, le modèle à effet aléatoire permet d'obtenir des estimateurs de  $\beta_1$  plus précis que les estimateurs "à effets fixes" (on exploite de l'information supplémentaire quant à la distribution jointe des  $\gamma$  et des  $\mathbf{X}$ ). Dans le cas où cette hypothèse d'indépendance n'est pas vérifiée, le modèle à effets aléatoires conduit à des estimations biaisées de  $\beta_1$  à la différence du modèle à effets fixes qui permet d'obtenir des estimations sans biais dans les deux cas de figure.

On est donc dans un cas d'arbitrage classique en économétrie : faire des hypothèses qui permettent de gagner en efficacité (si elles sont justes) mais qui risquent de générer des biais dans l'estimation (si elles sont fausses). Il faut choisir entre robustesse et efficacité, entre biais et variance.

Les tests d'Hausman permettent de construire une procédure de test qui permet de fonder ce choix sur les données.

### 5.1.3 Test d'Hausman

#### Un test de modèle

On se place sous les hypothèses maintenues suivantes :

$$\mathbf{y} = \beta_0 + \mathbf{X}^1\beta_1 + \mathbf{X}^2\beta_2 + \mathbf{Z}\gamma + \varepsilon$$

Avec :  $E(\varepsilon|\mathbf{X}, \mathbf{Z}) = 0$ ,  $V(\varepsilon|\mathbf{X}, \mathbf{Z}) = \sigma^2$  et  $V(\gamma|\mathbf{X}, \mathbf{Z}) = \sigma_\gamma^2$

Les variables  $X^1$  sont les variables individuelles qui varient au sein des classes. Les variables  $X^2$  sont celles qui sont constantes au sein des classes (parmi elles, il peut y avoir éventuellement des moyennes par classe de certaines variables  $X^1$ ).

On teste l'hypothèse  $H_0$  suivante :

$$H_0 : EL(\gamma|\mathbf{X}) = 0$$

On note  $\hat{\beta}_1^{EF}$  l'estimateur de  $\beta_1$  du modèle à effets fixes (estimateur "Within") et  $\hat{\beta}_1^{EA}$  l'estimateur de  $\beta_1$  du modèle à effets aléatoires (estimateur MCQG, MV ou MVR).

Sous  $H_0$  on a :

$$\left(\hat{\beta}_1^{EF} - \hat{\beta}_1^{EA}\right)' \left(V\left(\hat{\beta}_1^{EF}\right) - V\left(\hat{\beta}_1^{EA}\right)\right)^{-1} \left(\hat{\beta}_1^{EF} - \hat{\beta}_1^{EA}\right) \xrightarrow[n \rightarrow +\infty]{loi} \chi^2 \left(\text{rg}\left(V\left(\hat{\beta}_1^{EF}\right) - V\left(\hat{\beta}_1^{EA}\right)\right)\right)$$

Et sous l'hypothèse alternative, on a :

$$\left(\hat{\beta}_1^{EF} - \hat{\beta}_1^{EA}\right)' \left(V\left(\hat{\beta}_1^{EF}\right) - V\left(\hat{\beta}_1^{EA}\right)\right)^{-1} \left(\hat{\beta}_1^{EF} - \hat{\beta}_1^{EA}\right) \xrightarrow[n \rightarrow +\infty]{proba} +\infty$$

Les résultats de convergence restent valides si on remplace les quantités  $V\left(\hat{\beta}_1^{EF}\right)$  et  $V\left(\hat{\beta}_1^{EA}\right)$  par des estimateurs convergents à la vitesse  $\frac{1}{n}$ .

**Remarque :** les tests d'Hausman peuvent se dériver dans une grande diversité de cas, le point central du raisonnement consistant à comparer un estimateur efficace sous l'hypothèse  $H_0$  à un estimateur convergeant aussi bien sous l'hypothèse nulle que sous l'hypothèse alternative.

Preuve :

□

La démonstration repose sur le fait que sous  $H_0$ , l'estimateur du modèle à effets aléatoires est asymptotiquement optimal. La quantité  $f(\lambda) = \lim nV\left(\lambda\hat{\beta}_1^{EF} + (1-\lambda)\hat{\beta}_1^{EA}\right)$  est donc minimale pour  $\lambda = 0$ .  $\left.\frac{\partial f}{\partial \lambda}\right|_{\lambda=0} = 0$  implique que  $\lim nCov\left(\hat{\beta}_1^{EF}, \hat{\beta}_1^{EA}\right) = \lim nV\left(\hat{\beta}_1^{EA}\right)$ .

Les résultats usuels de convergence sur les estimateurs présentés assurent que :

$$\sqrt{n} \begin{pmatrix} \widehat{\beta}_1^{EF} - \beta_1 \\ \widehat{\beta}_1^{EA} - \beta_1 \end{pmatrix} \xrightarrow[n \rightarrow +\infty]{loi} \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right)$$

Avec  $\Sigma = \lim n \begin{bmatrix} V(\widehat{\beta}_1^{EF}) & V(\widehat{\beta}_1^{EA}) \\ V(\widehat{\beta}_1^{EA}) & V(\widehat{\beta}_1^{EA}) \end{bmatrix}$

On en déduit que :

$$\sqrt{n} (\widehat{\beta}_1^{EF} - \widehat{\beta}_1^{EA}) \xrightarrow[n \rightarrow +\infty]{loi} \mathcal{N} \left( 0, \lim n (V(\widehat{\beta}_1^{EF}) - V(\widehat{\beta}_1^{EA})) \right)$$

Et finalement :

$$(\widehat{\beta}_1^{EF} - \widehat{\beta}_1^{EA})' (V(\widehat{\beta}_1^{EF}) - V(\widehat{\beta}_1^{EA}))^{-1} (\widehat{\beta}_1^{EF} - \widehat{\beta}_1^{EA}) \xrightarrow[n \rightarrow +\infty]{loi} \chi^2 (rg(V(\widehat{\beta}_1^{EF}) - V(\widehat{\beta}_1^{EA})))$$

■

Tel qu'il est formulé ici, le test d'Hausman nécessite de faire deux estimations différentes, puis de combiner les résultats de celles-ci pour construire la statistique de test.

Si le test d'Hausman rejette l'hypothèse  $H_0$  de non corrélation entre les  $\gamma$  et les  $X$ , une possibilité consiste à enrichir les variables de niveau classe prises en compte dans le modèle. Plus le nombre de variables  $X^2$  est important, plus on contrôle des caractéristiques observables de la classe, plus la partie "résiduelle" de l'effet "classe" a des chances d'être orthogonale aux variables  $X^1$ . Mais il y a de fortes chances pour qu'en pratique cela ne soit pas suffisant. On peut alors introduire dans le modèle les moyennes de classes de certaines variables  $X^1$  (par exemple : pourcentage de filles dans la classe, d'enfants de cadres ou d'ouvriers, etc.).

Une autre possibilité est d'introduire dans le modèle les moyennes de classe de l'ensemble de variable  $X^1$ . La partie résiduelle de l'effet "classe" qui n'aurait pas été contrôlée par cette introduction des moyennes est alors par nature orthogonale aux  $X^1$ . C'est sur cette propriété que le modèle de Mundlak est construit. Il permet en outre de généraliser et de mettre en œuvre plus simplement le test d'Hausman (au moyen d'une seule estimation).

## 5.2 Le modèle de Mundlak

Nous avons vu dans la partie précédente que le modèle à effets aléatoires n'est qu'un sous-modèle du modèle à effets fixes : on spécifie une forme particulière de l'effet "classe". Le modèle de Mundlak est le modèle à effets aléatoires "le plus riche possible", c'est-à-dire le modèle dans lequel on a introduit toutes les moyennes de classe. L'intérêt d'une telle introduction est que la partie résiduelle de l'effet "classe" qui n'a pas été captée par les variables introduites dans le modèle est alors par nature orthogonale à la distribution des variables individuelles. Nous allons voir qu'alors les estimations des effets des variables individuelles coïncident entre le modèle de Mundlak et le modèle à effets fixes. Une manière de vulgariser cela consiste à dire que le modèle à effets aléatoires a été suffisamment enrichi pour "rejoindre" le sur-modèle qu'est le modèle à effets fixes. Une manière de tester le sous-modèle est alors de procéder à des tests standards de nullité des coefficients (test de Wald).

### 5.2.1 Les propriétés du modèle de Mundlak

Le modèle de Mundlak est un modèle à effets aléatoires dans lequel les moyennes par classe des covariables  $X^1$  ont été introduites. Les propriétés présentées dans les chapitres précédents nous assurent que si "le vrai modèle" suit bien une telle spécification, l'estimateur des MCO est sans biais et les estimateurs des MCQG, du MV ou du MVR sont asymptotiquement sans biais (à condition d'introduire les moyennes de classes parmi les régresseurs). Plus remarquable encore, les estimations des paramètres associés aux variables "individuelles" coïncident avec l'estimateur "Within". Ce modèle est donc très proche du modèle à effets fixes en terme d'interprétation.

#### Propriétés des estimateurs dans le modèle de Mundlak

Le modèle de Mundlak est un modèle à effets aléatoires particulier de la forme suivante :

$$y_{i,j} = X_{ij}^1 \beta_1 + \delta_j + \varepsilon_{ij}$$

Avec :  $\delta_j = X_j^1 \beta_2^1 + X_j^2 \beta_2^2 + \gamma_j$ ,  $E(\varepsilon|\mathbf{X}, \mathbf{Z}) = 0$ ,  $E(\gamma|\mathbf{X}, \mathbf{Z}) = 0$ ,  $V(\varepsilon|\mathbf{X}, \mathbf{Z}) = \sigma^2$  et  $V(\gamma|\mathbf{X}, \mathbf{Z}) = \sigma_\gamma^2$

Ou encore :

$$y_{i,j} = X_{ij}^1 \beta_1 + X_j^1 \beta_2^1 + X_j^2 \beta_2^2 + \gamma_j + \varepsilon_{ij}$$

Avec :  $E(\varepsilon|\mathbf{X}, \mathbf{Z}) = 0$ ,  $E(\gamma|\mathbf{X}, \mathbf{Z}) = 0$ ,  $V(\varepsilon|\mathbf{X}, \mathbf{Z}) = \sigma^2$  et  $V(\gamma|\mathbf{X}, \mathbf{Z}) = \sigma_\gamma^2$

Les estimateurs MCO, MCQG, MV ou MVR de  $\beta_1$  coïncident avec l'estimateur "Within". Rappelons ici que les modèles à effets aléatoires font généralement des hypothèses sur la distribution jointe des effets "classe" et des covariables, à la différence des modèles à effets fixes. Bien que le modèle de Mundlak soit un modèle à effets aléatoires, les estimations gardent la robustesse d'un modèle à effets fixes car l'éventuelle corrélation entre covariables et effets "classe" a été contrôlée par la prise en compte des moyennes de classe des covariables.

Les estimateurs MCO, MCQG, MV ou MVR de  $\beta_2^1$  sont égaux à la différence entre l'estimation obtenue par régression (pondérée) des moyennes par classe de  $Y$  sur les moyennes par classe de  $X^1$  et sur les  $X^2$  et l'estimation obtenue par régression "Within". Si les effets "classe"  $\delta_j$  sont décorrélés des variables  $X^1$ , alors la régression "Between" (régression "au niveau de la classe" des moyennes de  $Y$  sur les moyennes de  $X$  qui exploite la variabilité inter-classe) et la régression "Within" (qui exploite la variabilité intra-classe) doivent donner des résultats proches et donc l'estimation de  $\beta_2^1$  doit être proche de 0.

Les estimateurs MCO, MCQG, MV ou MVR de  $\beta_2^2$  sont égaux à l'estimation obtenue par régression (pondérée) des moyennes par classe de  $Y$  sur les moyennes par classe de  $X^1$  et sur les variables  $X^2$ . On constate donc que toute l'information utilisée pour estimer  $\beta_2^2$  est comprise dans la variabilité inter-classe, la variabilité intra-classe n'apporte aucune information.

Preuve :

□

On note  $\mathbf{Z}$  les indicatrices de classe. Soit  $B = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$  et  $W = I - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$  les projections inter-classe ("Between") et intra-classe ("Within"), il s'agit des opérateurs qui calculent les moyennes de classe et les écarts à la moyenne de classe. On a comme relation  $B\mathbf{Z} = \mathbf{Z}$  et  $W\mathbf{Z} = 0$ . Les variables de niveau classe se décomposent sous la forme  $\mathbf{X}^2 = \mathbf{Z}M$ , donc  $B\mathbf{X}^2 = \mathbf{X}^2$  et  $W\mathbf{X}^2 = 0$ . Enfin  $\mathbf{X} = [\mathbf{X}^1|\mathbf{X}^2]$  est l'ensemble des variables  $\mathbf{X}^1$  et  $\mathbf{X}^2$ .

Le modèle de Mundlak consiste à régresser  $\mathbf{y}$  sur  $\mathbf{X}^1, B\mathbf{X}^1, \mathbf{X}^2$  avec une matrice de variance des résidus de la forme de  $\sigma_\gamma^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 I$ . Les estimateurs MCO, MCQG, ML ou REML des paramètres sont de la forme :

$$\hat{\beta} = \left[ \begin{pmatrix} \mathbf{X}^1' \\ \mathbf{X}^1' B \\ \mathbf{X}^{II'} \end{pmatrix} \hat{V}^{-1} (\mathbf{X}^1 | B\mathbf{X}^1 | \mathbf{X}^2) \right]^{-1} \begin{pmatrix} \mathbf{X}^1' \\ \mathbf{X}^1' B \\ \mathbf{X}^{II'} \end{pmatrix} \hat{V}^{-1} \mathbf{y}$$

Avec :  $\hat{V}^{-1} = \frac{1}{\sigma^2} I - \frac{\sigma_\gamma^2}{\sigma^2} \mathbf{Z} (\widehat{\sigma^2} I + \widehat{\sigma_\gamma^2} \mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'$ , (pour l'estimateur MCO on a  $\widehat{\sigma_\gamma^2} = 0$ ).

En utilisant le fait que  $B\mathbf{X}^2 = \mathbf{X}^2$  et que  $B\hat{V}^{-1} = B\hat{V}^{-1}B = \hat{V}^{-1}B$ , on peut écrire :

$$\begin{pmatrix} \mathbf{X}^1' \\ \mathbf{X}^1' B \\ \mathbf{X}^{II'} \end{pmatrix} \hat{V}^{-1} (\mathbf{X}^1 | B\mathbf{X}^1 | \mathbf{X}^2) = \begin{bmatrix} \mathbf{X}^1' \hat{V}^{-1} \mathbf{X}^1 & \mathbf{X}^1' B \hat{V}^{-1} B \mathbf{X}^1 & \mathbf{X}^1' B \hat{V}^{-1} \mathbf{X}^2 \\ \mathbf{X}^1' B \hat{V}^{-1} \mathbf{X}^1 & \mathbf{X}^1' B \hat{V}^{-1} B \mathbf{X}^1 & \mathbf{X}^1' B \hat{V}^{-1} \mathbf{X}^2 \\ \mathbf{X}^{II'} B \hat{V}^{-1} \mathbf{X}^1 & \mathbf{X}^{II'} B \hat{V}^{-1} B \mathbf{X}^1 & \mathbf{X}^{II'} B \hat{V}^{-1} \mathbf{X}^2 \end{bmatrix}$$

Comme  $W\widehat{V}^{-1}W = (I - B)\widehat{V}^{-1}(I - B) = \widehat{V}^{-1} - B\widehat{V}^{-1}B$ , on a :

$$\begin{bmatrix} \mathbf{x}^{1'}\widehat{v}^{-1}\mathbf{x}^1 & \mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}^1 & \mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}^2 \\ \mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}^1 & \mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}^1 & \mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}^2 \\ \mathbf{x}^{II'}B\widehat{v}^{-1}B\mathbf{x}^1 & \mathbf{x}^{II'}B\widehat{v}^{-1}B\mathbf{x}^1 & \mathbf{x}^{II'}B\widehat{v}^{-1}B\mathbf{x}^2 \end{bmatrix} \begin{bmatrix} (\mathbf{x}^{1'}W\widehat{v}^{-1}W\mathbf{x}^1)^{-1} & -(\mathbf{x}^{1'}W\widehat{v}^{-1}W\mathbf{x}^1)^{-1} & 0 \\ -(\mathbf{x}^{1'}W\widehat{v}^{-1}W\mathbf{x}^1)^{-1} & (\mathbf{x}^{1'}W\widehat{v}^{-1}W\mathbf{x}^1)^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} I & -I & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

D'autre part :

$$\begin{bmatrix} \mathbf{x}^{1'}\widehat{v}^{-1}\mathbf{x}^1 & \mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}^1 & \mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}^2 \\ \mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}^1 & \mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}^1 & \mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}^2 \\ \mathbf{x}^{II'}B\widehat{v}^{-1}B\mathbf{x}^1 & \mathbf{x}^{II'}B\widehat{v}^{-1}B\mathbf{x}^1 & \mathbf{x}^{II'}B\widehat{v}^{-1}B\mathbf{x}^2 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & [\mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}]^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & I & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$$

Donc l'inverse de  $\begin{pmatrix} \mathbf{X}^{1'} \\ \mathbf{X}^{1'}B \\ \mathbf{X}^{II'} \end{pmatrix} \widehat{V}^{-1} (\mathbf{X}^1|B\mathbf{X}^1|\mathbf{X}^2)$  est :

$$\begin{bmatrix} (\mathbf{x}^{1'}W\widehat{v}^{-1}W\mathbf{x}^1)^{-1} & -(\mathbf{x}^{1'}W\widehat{v}^{-1}W\mathbf{x}^1)^{-1} & 0 \\ -(\mathbf{x}^{1'}W\widehat{v}^{-1}W\mathbf{x}^1)^{-1} & (\mathbf{x}^{1'}W\widehat{v}^{-1}W\mathbf{x}^1)^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & [\mathbf{x}^{1'}B\widehat{v}^{-1}B\mathbf{x}]^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Les estimateurs sont donc de la forme :

$$\widehat{\beta} = \begin{pmatrix} (\mathbf{X}^{1'}W\widehat{v}^{-1}W\mathbf{X}^1)^{-1} \mathbf{X}^{1'}W\widehat{v}^{-1}\mathbf{y} \\ -(\mathbf{X}^{1'}W\widehat{v}^{-1}W\mathbf{X}^1)^{-1} \mathbf{X}^{1'}W\widehat{v}^{-1}\mathbf{y} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ [\mathbf{X}^{1'}B\widehat{v}^{-1}B\mathbf{X}]^{-1} \mathbf{X}^{1'}B\widehat{v}^{-1}\mathbf{y} \\ 0 \end{pmatrix}$$

Finalement, comme  $W\widehat{V}^{-1}W = W\widehat{V}^{-1} = \frac{1}{\sigma^2}W$  et que  $B\widehat{V}^{-1}B = B\widehat{V}^{-1}$

$$\widehat{\beta} = \begin{pmatrix} (\mathbf{X}^{1'}W\mathbf{X}^1)^{-1} \mathbf{X}^{1'}W\mathbf{y} \\ -(\mathbf{X}^{1'}W\mathbf{X}^1)^{-1} \mathbf{X}^{1'}W\mathbf{y} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ [\mathbf{X}^{1'}B\widehat{v}^{-1}B\mathbf{X}]^{-1} \mathbf{X}^{1'}B\widehat{v}^{-1}B\mathbf{y} \\ 0 \end{pmatrix}$$

Les estimations des paramètres associés aux variables individuelles dans le modèle de Mundlak coïncident avec les estimations "Within", on a ainsi la réponse à la question posée. Mais on peut aller un peu plus loin dans l'analyse.

En remarquant que  $B\widehat{V}^{-1}B = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \left[ \frac{1}{\sigma^2}I - \frac{\sigma_\gamma^2}{\sigma^2}\mathbf{Z}(\widehat{\sigma^2}I + \widehat{\sigma_\gamma^2}\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \right] \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \frac{1}{\sigma^2}\mathbf{Z} \left[ (\mathbf{Z}'\mathbf{Z})^{-1} - \widehat{\sigma_\gamma^2}(\widehat{\sigma^2}I + \widehat{\sigma_\gamma^2}\mathbf{Z}'\mathbf{Z})^{-1} \right] \mathbf{Z}' = \frac{1}{\sigma^2}\mathbf{Z} \left[ \widehat{\sigma^2}(\mathbf{Z}'\mathbf{Z})^{-1} + \widehat{\sigma_\gamma^2}I - \widehat{\sigma_\gamma^2}I \right] (\widehat{\sigma^2}I + \widehat{\sigma_\gamma^2}\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} (\widehat{\sigma^2}(\mathbf{Z}'\mathbf{Z})^{-1} + \widehat{\sigma_\gamma^2}I)^{-1} (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}$

On peut interpréter les quantités  $[\mathbf{X}^{1'}B\widehat{v}^{-1}B\mathbf{X}]^{-1} \mathbf{X}^{1'}B\widehat{v}^{-1}B\mathbf{y}$  comme les estimations en travaillant sur un fichier de classes pour lequel toutes les moyennes par classe ont été calculées et en effectuant une régression pondérée de la moyenne de  $\mathbf{y}$  sur les  $\mathbf{X}^2$  et les moyennes de  $\mathbf{X}^1$ . Les poids sont de la forme  $\frac{1}{\sqrt{\widehat{\sigma_\gamma^2} + \frac{\sigma^2}{n_j}}}$ , où  $n_j$  est la taille de la classe  $j$  (on accorde plus d'importance aux

grandes classes). On remarque que dans le cas des MCO, le poids est alors proportionnel à  $\sqrt{n_j}$  ce qui revient à faire la régression "Between" sur le fichier individuel sans pondérer.

Les coefficients associés aux variables  $B\mathbf{X}^1$  dans le modèle de Mundlak sont estimés comme différence entre la régression "Within" et la régression "Between" pondérée, les coefficients associés aux variables  $\mathbf{X}^2$  dans le modèle de Mundlak sont estimés uniquement au moyen de la régression "Between" pondérée.

■

## 5.2.2 Implémentation du test d'Hausman et amélioration des estimations

Le test d'Hausman présenté plus haut permet de tester la propriété d'orthogonalité entre la distribution des effets "classe" et l'ensemble des covariables introduites dans le modèle. C'est une hypothèse nécessaire dans le cas où on décide de recourir à un modèle à effets aléatoires. Si celle-ci est rejetée, on peut soit avoir recours à une estimation du modèle à effets fixes, soit ajouter des variables définies au niveau de la classe dans la régression pour rendre l'hypothèse plus crédible et avant de la tester à nouveau. Mais le test d'Hausman tel qu'il a été présenté nécessite pour chaque test deux estimations et une programmation spécifique. Nous allons présenter une méthode qui permet de l'implémenter de manière particulièrement simple avec SAS (et la plupart des logiciels statistiques).

### Le test d'Hausman comme un test de Wald dans le modèle de Mundlak

Le modèle de Mundlak est de la forme :

$$y_{i,j} = X_{ij}^1 \beta_1 + X_{ij}^1 \beta_2^1 + X_j^2 \beta_2^2 + \gamma_j + \varepsilon_{ij}$$

Avec :  $E(\varepsilon|\mathbf{X}, \mathbf{Z}) = 0$ ,  $E(\gamma|\mathbf{X}, \mathbf{Z}) = 0$ ,  $V(\varepsilon|\mathbf{X}, \mathbf{Z}) = \sigma^2$  et  $V(\gamma|\mathbf{X}, \mathbf{Z}) = \sigma_\gamma^2$

Pour les estimateurs MCO, MCQG, MV ou MVR de  $\beta_2^1$ , la statistique de Wald (qui permet de tester la nullité globale de  $\beta_2$ ) est numériquement égale à la statistique d'Hausman.

$$\widehat{\beta}_2^1{}' V(\widehat{\beta}_2^1)^{-1} \widehat{\beta}_2^1 = \left( \widehat{\beta}_1^{EF} - \widehat{\beta}_1^{EA} \right)' \left( V(\widehat{\beta}_1^{EF}) - V(\widehat{\beta}_1^{EA}) \right)^{-1} \left( \widehat{\beta}_1^{EF} - \widehat{\beta}_1^{EA} \right)$$

Mais les variances des différents estimateurs ne sont pas connues et doivent être elles-mêmes estimées de manière convergente.

Les tests associés aux estimations MCQG, MV et MVR s'appuient sur une estimation convergente de la variance de  $\beta_2$  :

$$\widehat{\beta}_2^1{}' \widehat{V}(\widehat{\beta}_2^1)^{-1} \widehat{\beta}_2^1$$

Sous l'hypothèse  $H_0$  ( $\beta_2 = 0$ ), cette statistique tend en loi vers un  $\chi^2$  ( $rg(V(\widehat{\beta}_2^1))$ )

Sous l'hypothèse alternative ( $\beta_2 \neq 0$ ), cette statistique tend presque sûrement vers  $+\infty$ .

Par contre, le test associé aux estimations MCO s'appuie sur un estimateur non convergent de la variance de  $\beta_2$ , il est donc biaisé car la statistique ne tend pas sous  $H_0$  vers une distribution connue.

Preuve :

□

On a déjà vu précédemment que l'estimateur  $\widehat{\beta}_2^1$  se décompose comme la différence de deux termes ; le premier est obtenu par régression "Between" (éventuellement pondérée si le nombre d'élèves par classe varie) et le second par régression "Within".

$$\widehat{\beta}_2^1 = \widehat{\beta}^B - \widehat{\beta}^W$$

avec :

$$\widehat{\beta}^B = \begin{bmatrix} \mathbf{X}^1{}' B \widehat{V}^{-1} B \mathbf{X}^1 - \mathbf{X}^1{}' B \widehat{V}^{-1} \mathbf{X}^2 \left( \mathbf{X}^{II'} \widehat{V}^{-1} \mathbf{X}^2 \right)^{-1} \mathbf{X}^{II'} \widehat{V}^{-1} B \mathbf{X}^1 \\ \mathbf{X}^1{}' B \widehat{V}^{-1} B \mathbf{y} - \mathbf{X}^1{}' B \widehat{V}^{-1} \mathbf{X}^2 \left( \mathbf{X}^{II'} \widehat{V}^{-1} \mathbf{X}^2 \right)^{-1} \mathbf{X}^{II'} \widehat{V}^{-1} B \mathbf{y} \end{bmatrix}^{-1}$$

$$\widehat{V}(\widehat{\beta}^B)^{-1} = \mathbf{X}^1{}' B \widehat{V}^{-1} B \mathbf{X}^1 - \mathbf{X}^1{}' B \widehat{V}^{-1} \mathbf{X}^2 \left( \mathbf{X}^{II'} \widehat{V}^{-1} \mathbf{X}^2 \right)^{-1} \mathbf{X}^{II'} \widehat{V}^{-1} B \mathbf{X}^1$$

$$\begin{aligned}\hat{\beta}^W &= (\mathbf{X}^{1'W}\mathbf{X}^1)^{-1}(\mathbf{X}^{1'W}\mathbf{y}) = (\mathbf{X}^{1'W}\hat{V}W\mathbf{X}^1)^{-1}(\mathbf{X}^{1'W}\hat{V}W\mathbf{y}) \\ \hat{V}(\hat{\beta}^W)^{-1} &= \mathbf{X}^{1'W}\hat{V}W\mathbf{X}^1\end{aligned}$$

Or :

$$\begin{aligned}\hat{\beta}^{EA} &= \left[ \mathbf{X}^{1'}\hat{V}^{-1}\mathbf{X}^1 - \mathbf{X}^{1'}\hat{V}^{-1}\mathbf{X}^2 (\mathbf{X}^{II'}\hat{V}^{-1}\mathbf{X}^2)^{-1} \mathbf{X}^{II'}\hat{V}^{-1}\mathbf{X}^1 \right]^{-1} \\ &\quad \left[ \mathbf{X}^{1'}\hat{V}^{-1}\mathbf{y} - \mathbf{X}^{1'}\hat{V}^{-1}\mathbf{X}^2 (\mathbf{X}^{II'}\hat{V}^{-1}\mathbf{X}^2)^{-1} \mathbf{X}^{II'}\hat{V}^{-1}\mathbf{y} \right] \\ \hat{V}(\hat{\beta}^{EA})^{-1} &= \mathbf{X}^{1'}\hat{V}^{-1}\mathbf{X}^1 - \mathbf{X}^{1'}\hat{V}^{-1}\mathbf{X}^2 (\mathbf{X}^{2'}\hat{V}^{-1}\mathbf{X}^2)^{-1} \mathbf{X}^{2'}\hat{V}^{-1}\mathbf{X}^1\end{aligned}$$

Donc :

$$\begin{aligned}\hat{V}(\hat{\beta}^{EA})^{-1} &= \hat{V}(\hat{\beta}^W)^{-1} + \hat{V}(\hat{\beta}^B)^{-1} \\ \hat{V}(\hat{\beta}^{EA})^{-1}\hat{\beta}^{EA} &= \hat{V}(\hat{\beta}^W)^{-1}\hat{\beta}^W + \hat{V}(\hat{\beta}^B)^{-1}\hat{\beta}^B\end{aligned}$$

De plus :

$$\begin{aligned}\hat{\beta}^{EF} &= \hat{\beta}^W \\ \hat{V}(\hat{\beta}^{EF}) &= \hat{V}(\hat{\beta}^W)\end{aligned}$$

On en déduit alors que :

$$\begin{aligned}\hat{V}(\hat{\beta}^{EA})^{-1}(\hat{\beta}^{EF} - \hat{\beta}^{EA}) &= \hat{V}(\hat{\beta}^{EA})^{-1}(\hat{\beta}^W - \hat{\beta}^{EA}) \\ &= \left[ \hat{V}(\hat{\beta}^{EA})^{-1} - \hat{V}(\hat{\beta}^W)^{-1} \right] \hat{\beta}^W - \hat{V}(\hat{\beta}^B)^{-1}\hat{\beta}^B \\ &= \hat{V}(\hat{\beta}^B)^{-1}(\hat{\beta}^W - \hat{\beta}^B) \\ &= -\hat{V}(\hat{\beta}^B)^{-1}\hat{\beta}_2^B\end{aligned}$$

$$\begin{aligned}&\hat{V}(\hat{\beta}^{EA}) \left[ \hat{V}(\hat{\beta}^W) - \hat{V}(\hat{\beta}^{EA}) \right]^{-1} \hat{V}(\hat{\beta}^{EA}) \\ &= \left[ \left( \hat{V}(\hat{\beta}^W)^{-1} + \hat{V}(\hat{\beta}^B)^{-1} \right) \hat{V}(\hat{\beta}^W) \left( \hat{V}(\hat{\beta}^W)^{-1} + \hat{V}(\hat{\beta}^B)^{-1} \right) - \left( \hat{V}(\hat{\beta}^W)^{-1} + \hat{V}(\hat{\beta}^B)^{-1} \right) \right]^{-1} \\ &= \left[ \hat{V}(\hat{\beta}^B)^{-1} + \hat{V}(\hat{\beta}^B)^{-1} \hat{V}(\hat{\beta}^W) \hat{V}(\hat{\beta}^B)^{-1} \right]^{-1} \\ &= \hat{V}(\hat{\beta}^B) \left[ \hat{V}(\hat{\beta}^B) + \hat{V}(\hat{\beta}^W) \right]^{-1} \hat{V}(\hat{\beta}^B)\end{aligned}$$

Ce qui permet de conclure.

■

Pour choisir entre un modèle à effets fixes et un modèle à effets aléatoires, il suffit donc d'estimer un modèle de Mundlak et de tester la nullité jointe des coefficients associés aux moyennes par classe des covariables. Ce test doit être mené en permettant des corrélations entre les inobservées des élèves dans une même classe. Il ne faut donc pas utiliser les tests usuels de Fisher des MCO. Cependant, les logiciels permettent facilement de faire de tels tests en introduisant des "clusters" (en français des effets de grappes) dans les inobservées, ou d'estimer directement la régression par MCQG, maximum de vraisemblance ou maximum de vraisemblance restreint.

Si la nullité jointe des coefficients est rejetée, alors il existe vraisemblablement une corrélation entre les inobservées constantes au sein de la classe et les covariables  $\mathbf{X}^1$ , dans ce cas le "vrai" modèle est un modèle à effets fixes. Sans recours à des hypothèses de type "variables instrumentales", les seuls coefficients qui peuvent être estimés sans biais sont ceux associés aux variables  $\mathbf{X}^1$ . Or l'estimation du modèle de Mundlak fournit directement un estimateur convergent de ces coefficients.

Si la nullité jointe des coefficients est acceptée, alors il n'y a pas de corrélation entre les inobservées constantes au sein de la classe et les covariables  $\mathbf{X}^1$ . Dans ce cas, le modèle à effets aléatoires, bien que plus restrictif que le modèle à effets fixes, ne peut pas être rejeté sur la base du test d'Hausman. Dans un souci de précision, il est alors possible d'estimer les coefficients  $\beta^1$  associés aux covariables  $\mathbf{X}^1$  sans introduire dans la régression les moyennes de classe des covariables.

## 5.3 Programmation sous SAS...

### 5.3.1 ... du modèle de Mundlak

L'estimation du modèle de Mundlak se programme de la même manière que dans le chapitre précédent, il faut juste ajouter les moyennes par classe pour les variables qui changent au sein de la classe (on les note  $X_{11}, X_{12}, \dots, X_{1K1}$  dans l'exemple ci-dessous par opposition aux variables  $X_{21}, X_{22}, \dots, X_{2K2}$  qui sont constantes au sein de la classe).

Il faut donc calculer pour chaque observation, la moyenne des covariables par classe, par exemple de la manière suivante :

```
proc summary data=TABLE;
  var X11 X12 ... X1K1;
  class Z;
  output out=between(where=(STAT='MEAN' and TYPE=1));
run;

proc sort data=TABLE;
  by Z;
run;

proc sql;
  create table TABMUND as
  select
    t1.Y as Y,
    t1.X11 as X11,
    t1.X12 as X12,
    ...
    t1.X1K1 as X1K1,
    t1.X21 as X21,
    t1.X22 as X22,
    ...
    t1.X2K2 as X2K2,
    t2.X11 as B_X11,
    t2.X12 as B_X12,
    ...
    t2.X1K1 as B_X1K1,
    t1.Z as Z
  from TABLE as t1, between as t2
  where t1.Z = t2.Z;
quit;
```

Il ne reste plus qu'à estimer le modèle à effets aléatoires (par exemple, par MCQG comme ci-dessous)

```
proc mixed data=TABMUND method=mivque0;
  class Z;
  model Y=X11 X12 ... X1K1 X21 X22 ... X2K2 B_X11 B_X12 ... B_X1K1/solution;
  repeated /subject=Z type=cs;
run;
```

### 5.3.2 ... et du test d'Hausman

Pour implémenter le test d'Hausman, il suffit d'estimer le modèle de Mundlak en précisant qu'on désire faire un test joint sur la nullité des coefficients associés aux moyennes par classe  $B_{X11}, B_{X12}, \dots, B_{X1K1}$  des covariables  $X_{11}, X_{12}, \dots, X_{1K1}$ .

```
proc mixed data=TABMUND method=mivque0;
  class Z;
  model Y=X11 X12 ... X1K1 X21 X22 ... X2K2 B_X11 B_X12 ... B_X1K1/solution;
```

```
contrast 'Test de nullité globale des moyennes par classe'  
B_X11 1,  
B_X12 1,  
...  
B_X1K1 1;  
repeated /subject=Z type=cs;  
run;
```

## Chapitre 6

# Extension au cas d'une variable dépendante dichotomique

On se place maintenant dans le cas où la variable d'intérêt  $y$  est une variable dichotomique prenant les valeurs 0 ou 1. Le cas typique concernant les problématiques relatives à l'éducation est le redoublement, le fait d'obtenir ou non un diplôme ou le choix de s'inscrire (ou le fait d'être admis) dans une filière particulière.

Dans la première partie de ce chapitre, nous présentons quelques résultats concernant les modèles dichotomiques dans le cas où il n'y a pas d'effets "classe". Ces modèles, bien qu'identifiés sous des hypothèses relativement faibles sont souvent estimés dans un cadre paramétrique par maximum de vraisemblance. Dans une deuxième partie, nous montrons comment, dans le cadre du maximum de vraisemblance, l'introduction d'effets fixes au niveau de la classe conduit à des estimateurs biaisés en raison du problème dit "des paramètres incidents". Dans une troisième partie, nous montrons comment ce problème de paramètres incidents peut être contourné en utilisant un maximum de vraisemblance conditionnel lorsque la distribution des inobservées est logistique. Enfin, nous présentons les modèles à effets aléatoires très utilisés en pratique, mais au prix d'hypothèses fortes sur la structure des effets "classes".

### 6.1 Rappel sur les modèles à variables qualitatives

La première partie de cette section présente le modèle linéaire de probabilité, c'est le modèle le plus simple, qui consiste simplement à ignorer que la variable expliquée est dichotomique et à la traiter comme une variable continue. Cette modélisation pose cependant problème concernant la structure des inobservées et la cohérence des prédictions obtenues.

La deuxième partie de cette section présente l'estimateur du score maximum de Manski, il s'agit essentiellement de mettre en évidence des hypothèses relativement faibles sous lesquelles un modèle dichotomique peut être identifié et estimé. Dans un tel modèle, les observées et les inobservées ont un impact causal sur une variable latente. Quand cette variable latente est au dessus (respectivement au dessous) d'un certain seuil, la variable expliquée vaut 1 (respectivement 0). L'identification provient d'une hypothèse concernant la variabilité d'une des variables explicatives  $\mathbf{X}$  et un quantile conditionnel des inobservées. L'intérêt d'un tel estimateur est d'être semi-paramétrique : l'ensemble des lois des inobservées étant alors de dimension infinie. On obtient cependant des estimateurs convergents dans le cadre standard (c'est-à-dire sans effets "classe"). Le calcul et l'inférence concernant cet estimateur restent cependant difficile à mettre en œuvre.

Enfin, la troisième partie de cette section présente en détail les modèles paramétriques de type logit ou probit estimés par maximum de vraisemblance. Ces modèles paramétriques restreignent les lois des inobservées à un espace de dimension finie (dimension 1 en l'occurrence pour le probit et la logit). Cette forte hypothèse de structure permet d'obtenir des estimateurs relativement simples à calculer ainsi que de construire des tests.

### 6.1.1 Le modèle linéaire de probabilité

On peut alors penser assez naturellement à un "modèle linéaire de probabilité" en faisant comme si la variable dépendante était continue :

#### Un modèle simple mais incohérent

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

avec  $\mathbf{y} \in \{0, 1\}$  et  $E(\epsilon|\mathbf{X}) = 0$

Ce modèle présente un certain nombre d'inconvénients :

- Il est intrinsèquement hétéroscédastique (puisque  $V(\epsilon|\mathbf{X}) = \text{diag}(\mathbf{X}\beta(1 - \mathbf{X}\beta))$ ). Pour effectuer des tests sur la valeur de  $\beta$ , il faut donc recourir aux MCQG.
- La probabilité  $P(\mathbf{y} = 1|\mathbf{X}) = E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\beta$  n'est pas bornée par le modèle, cela peut devenir particulièrement problématique si le support de  $\mathbf{X}$  est large. En pratique, il peut ainsi arriver que le prédicteur  $\mathbf{X}\hat{\beta}$  n'appartienne pas à l'intervalle  $[0, 1]$  pour certains individus.

**Remarque :** s'il peut être très facile d'estimer des modèles linéaires de probabilité à titre exploratoire, il faut bien garder en tête que les tests de Fischer sont biaisés pour une régression des moindres carrés et que la variable  $\mathbf{y}$  est bornée, ce qui nécessite souvent de recourir à une modélisation qui prenne en compte le fait que la variable  $\mathbf{y}$  soit dichotomique.

### 6.1.2 L'estimateur du score maximum de Manski

Manski a étudié le modèle suivant :

#### Un estimateur semi-paramétrique

$$y_i = \mathbb{1}_{\{X_i\beta + \epsilon_i > 0\}}$$

avec  $F_{-\epsilon|X}$  la fonction de répartition de  $-\epsilon_i|X$ .

Si on suppose de plus que :

- les covariables  $X$  ne sont pas reliées par une combinaison linéaire (hypothèse qui implique qu'avec suffisamment d'observations la matrice  $\mathbf{X}$  est de plein rang colonne).
- $F_{-\epsilon|X}(0) = \frac{1}{2}$ , c'est-à-dire que la médiane de  $\epsilon$  conditionnellement à  $X$  est nulle.
- il existe une covariable  $x_k$  particulière : un instrument à large support (c'est-à-dire tel que  $\forall(a, b) \in \mathbb{R}^2, P(x_k \in [a, b] | (x_{k'})_{k' \neq k}) > 0$ ) dont le coefficient  $\beta_k$  associé est non nul.

Dans ces conditions :

- $\beta$  est identifiable à un facteur d'échelle près (i.e. on peut identifier  $\frac{\beta}{\|\beta\|}$ ).
- L'estimateur  $\hat{\beta} = \min \sum_{i=1}^n |y_i - \mathbb{1}_{\{X_i\beta > 0\}}|$  est convergent (vers  $\frac{\beta}{\|\beta\|}$ )

**Remarque :** tous les modèles binaires nécessitent une normalisation, car le caractère binaire de la variable expliquée n'introduit pas assez de variabilité pour identifier la dispersion de la variable sous-jacente ; en effet, la variable latente  $\{2X_i\beta + 2\epsilon_i$  induit le même résultat que la variable latente  $X_i\beta + \epsilon_i$ . Cette normalisation peut soit porter sur la dispersion des résidus (par exemple : dans les modèles Probit et Logit, on fait en pratique l'hypothèse que  $V(\epsilon) = 1$ ), soit porter sur la normalisation d'une des composantes de  $\beta$  (par exemple, on fixe la valeur de la constante à 1), soit porter sur la norme de  $\beta$  (c'est ce qu'on fait implicitement ici en discutant l'identification de  $\frac{\beta}{\|\beta\|}$ ).

L'hypothèse de large support est assez restrictive en pratique car elle implique qu'il existe des individus dans la population pour lesquels la probabilité  $P(y = 1|X)$  tend vers 0, et d'autres pour lesquels elle tend vers 1. Cette hypothèse n'est pas crédible pour modéliser des événements rares (dont on peut penser que la probabilité de réalisation pour tous les individus est inférieure à  $\bar{p} < 1$ ), ou des événements fréquents (la probabilité de réalisation pour tous les individus est supérieure à  $\underline{p} > 0$ ).

L'estimateur du score maximum est surtout intéressant d'un point de vue théorique puisqu'il prouve que l'identification et la production d'un estimateur convergent ne reposent pas sur une loi particulière de  $\epsilon$  ( $F$  n'a pas besoin d'être explicitement connue ou estimée), ni sur l'indépendance de  $\epsilon$  et de  $X$  (par exemple,  $V(\epsilon|X)$  peut dépendre de  $X$ ). En pratique, il est souvent difficile à mettre en oeuvre car la fonction à minimiser n'est pas régulière, et sa convergence asymptotique n'est pas standard (convergence en  $n^{\frac{1}{3}}$  vers une distribution non normale). En pratique, pour mener des estimations semi-paramétriques, on utilisera plutôt des méthodes proposées par Horowitz ("A Smoothed Maximum Score Estimator for the Binary Response Model", *Econometrica* 1992) ou Klein et Spady ("An Efficient Semiparametric Estimator for Binary Response Models", *Econometrica* 1993).

Preuve :

□

On se contentera ici de donner l'idée de la preuve.

On suppose sans perte de généralité que notre instrument à large support est en dernière position dans la liste des covariables ( $k = K$ ) et que la constante est en première position :  $X_0$ . On distingue les  $K$  premières composantes de  $\beta$  de la composante associée à l'instrument à large support :  $\beta = (\beta_{-K}, \beta_K) \in \mathbb{R}^{K+1}$ . De même, on distinguera les  $K$  premières composantes d'un vecteur  $b \in \mathbb{R}^{K+1}$  de la dernière en adoptant la notation  $b = (b_{-K}, b_K)$ . On suppose également sans perte de généralité que  $\beta_K > 0$  (le cas  $\beta_K < 0$  se démontrant de manière identique). On remarque alors que si  $X\beta \geq 0$  (respectivement  $X\beta < 0$ ), alors  $P(y = 1|X) \geq \frac{1}{2}$  (respectivement  $P(y = 1|X) < \frac{1}{2}$ ).  $P(y = 1|X)$  est identifiable (si on avait un très grand nombre de données, il suffirait de faire des moyennes à  $X$  donné). Donc  $\beta$  sera identifiable à un facteur d'échelle près si  $\forall b \in \mathbb{R}^{K+1}$  tel que  $b_K \neq 0$ ,  $\|b\| = 1$ ,  $b \neq \beta$  on a  $\mathbb{1}_{\{Xb \geq 0\}} \neq \mathbb{1}_{\{X\beta \geq 0\}}$  pour un ensemble de valeurs de  $X$  qui ne soit pas de mesure nulle, c'est-à-dire qu'on a  $P(\mathbb{1}_{\{Xb \geq 0\}} \neq \mathbb{1}_{\{X\beta \geq 0\}}) > 0$ .

On a alors :

$$\begin{aligned} P(\mathbb{1}_{\{Xb \geq 0\}} \neq \mathbb{1}_{\{X\beta \geq 0\}}) &= P(Xb < 0 \leq X\beta) + P(X\beta < 0 \leq Xb) \\ &\geq P(Xb < 0 < X\beta) + P(X\beta < 0 < Xb) \end{aligned}$$

Distinguons deux cas selon la valeur de  $b_K$  :

- Si  $b_K < 0$  :  $P(Xb < 0 < X\beta) = P\left(-\frac{X_{-K}b_{-K}}{b_K} < x_K \cap -\frac{X_{-K}\beta_{-K}}{\beta_K} < x_K\right) = P\left(\max\left(-\frac{X_{-K}b_{-K}}{b_K} < x_K, -\frac{X_{-K}\beta_{-K}}{\beta_K} < x_K\right) < x_K\right)$ . Comme  $x_K$  est à large support, il existe de "grandes" valeurs de  $x_K$  qui assurent que  $P(\mathbb{1}_{\{Xb \geq 0\}} \neq \mathbb{1}_{\{X\beta \geq 0\}}) > 0$
- Si  $b_K > 0$  : on a  $P(Xb < 0 < X\beta) = P\left(-\frac{X_{-K}\beta_{-K}}{\beta_K} < x_K < -\frac{X_{-K}b_{-K}}{b_K}\right)$  et  $P(X\beta < 0 < Xb) = P\left(-\frac{X_{-K}b_{-K}}{b_K} < x_K < -\frac{X_{-K}\beta_{-K}}{\beta_K}\right)$ . Comme  $x_K$  est continue (conditionnellement à  $X_{-K}$ ), on en conclut que les quantités  $P\left(-\frac{X_{-K}\beta_{-K}}{\beta_K} < x_K < -\frac{X_{-K}b_{-K}}{b_K}\right)$  et  $P\left(-\frac{X_{-K}b_{-K}}{b_K} < x_K < -\frac{X_{-K}\beta_{-K}}{\beta_K}\right)$  ne peuvent être simultanément nulles que si  $P\left(\frac{X_{-K}\beta_{-K}}{\beta_K} = \frac{X_{-K}b_{-K}}{b_K}\right) = 1$ , ce qui contredit le fait que les covariables ne sont pas reliées par une combinaison linéaire. Donc  $P(\mathbb{1}_{\{Xb \geq 0\}} \neq \mathbb{1}_{\{X\beta \geq 0\}}) > 0$ .

On en conclut que  $\beta$  est bien identifié à un facteur d'échelle près.

Les théorèmes classiques sur les estimateurs obtenus par maximisation d'une fonction assurent la convergence sous les conditions suivantes :

- $\forall b \in \mathbb{R}^{K+1}$  tel que  $b_K \neq 0$  et  $\|b\| = 1$  on a :  $E[|y - \mathbb{1}_{\{X\beta > 0\}}|] > E[|y - \mathbb{1}_{\{Xb > 0\}}|]$
- La fonction  $b \mapsto E[|y - \mathbb{1}_{\{Xb\}}|]$  est continue sur  $\{b \in \mathbb{R}^{K+1}, b_K \neq 0, \|b\| = 1\}$
- $\sup_{b \in \mathbb{R}^{K+1}, b_K \neq 0, \|b\|=1} \left| \frac{1}{n} \sum_{i=1}^n |y - \mathbb{1}_{\{Xb \geq 0\}}| - E[|y - \mathbb{1}_{\{Xb \geq 0\}}|] \right| \xrightarrow[n \rightarrow +\infty]{P} 0$

Comme (cf. la partie de la preuve sur l'identification)  $P(\mathbb{1}_{\{X\beta > 0\}} \neq \mathbb{1}_{\{Xb > 0\}}) > 0$  et que  $E[|y - \mathbb{1}_{\{X\beta > 0\}}| - |y - \mathbb{1}_{\{Xb > 0\}}|] = E[|\mathbb{1}_{\{X\beta > 0\}} - \mathbb{1}_{\{Xb > 0\}}|] = P(\mathbb{1}_{\{X\beta > 0\}} \neq \mathbb{1}_{\{Xb > 0\}})$ , la première condition est vérifiée.

Pour la deuxième condition, on peut supposer sans perte de généralité que  $b_K > 0$ , dans ce cas :  $E[|y - \mathbb{1}_{\{Xb > 0\}}|] = P(y = 1, Xb \leq 0) + P(y = 0, Xb > 0)$ . La continuité en  $b$  du terme  $P(y = 1, Xb \leq 0)$  est assuré par le fait que  $P(y = 1, Xb \leq 0) = E\left(F(X\beta)\mathbb{1}_{\{x_K \leq -\frac{X_{-K}b_{-K}}{b_K}\}}\right)$ , qui

est une fonction continue en  $b$  et bornée. La continuité de  $P(y = 0, Xb > 0)$  se prouve de la même manière.

Enfin, la troisième condition est la plus difficile à montrer. En effet, si de manière immédiate, la loi des grands nombres implique que pour tout  $b$  fixé  $\frac{1}{n} \sum_{i=1}^n |y - \mathbb{1}_{\{Xb \geq 0\}}| \xrightarrow[n \rightarrow +\infty]{p} E[|y - \mathbb{1}_{\{Xb \geq 0\}}|]$ , obtenir une convergence uniforme pour  $b \in \mathbb{R}^{K+1}$  est plus difficile. Comme  $\frac{1}{n} \sum_{i=1}^n |y - \mathbb{1}_{\{Xb \geq 0\}}| = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{y_i=1, X_i b \geq 0\}} + \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{y_i=0, X_i b < 0\}}$  et que  $E[|y - \mathbb{1}_{\{Xb \geq 0\}}|] = P(y_i = 1, X_i b \geq 0) + P(y_i = 0, X_i b < 0)$ , il s'agit de généraliser à un cadre multi-dimensionnel le théorème de Glivenko-Cantelli qui assure que pour des variables  $z$  réelles et identiquement distribuées  $\sup_{a \in \mathbb{R}} |P(z_1 \leq a) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{z_i \leq a\}}| \xrightarrow[n \rightarrow +\infty]{p} 0$ . Rao a établi le résultat qui nous intéresse en 1962.

■

### 6.1.3 Les modèles Probit et Logit

La partie précédente nous a permis de mettre en évidence que sous des hypothèses distributionnelles relativement faibles concernant la loi de la variable latente (la distribution de  $\epsilon|X$  n'est pas précisée), les paramètres  $\beta$  sont identifiables et estimables de manière convergente. Faire des hypothèses plus restrictives sur la distribution des résidus présente des avantages pratiques : le calcul des estimateurs est plus simple, la loi asymptotique des estimateurs est plus précise (on abandonne un résultat robuste qui ne nécessitait pas de faire d'hypothèse distributionnelle pour un résultat plus précis si l'hypothèse distributionnelle est retenue à raison), il n'est plus nécessaire d'avoir un instrument à large support (instrument souvent manquant en pratique). Si on suppose que la distribution de  $\epsilon|X$  est connue, on se retrouve alors dans le cadre général du maximum de vraisemblance, ce qui permet d'obtenir les lois asymptotiques des estimateurs et des tests asymptotiques. Pour toutes ces raisons, les modèles paramétriques (c'est-à-dire les modèles qui supposent la loi de  $\epsilon|X$  connue) sont beaucoup plus utilisés dans la pratique que le modèle semi-paramétrique<sup>1</sup> de Manski.

#### L'estimateur du maximum de vraisemblance

$$y_i = \mathbb{1}_{\{X_i \beta + \epsilon_i > 0\}}$$

avec  $F$  la fonction de répartition de  $-\epsilon_i|X_i$ .

Si on suppose de plus que :

- Les  $y_i, x_{1i}, x_{2i}, \dots, x_{Ki}$  sont des vecteurs aléatoires indépendants et identiquement distribués
- La matrice  $\mathbf{X}$  est de plein rang colonne
- $\epsilon$  est indépendant de  $X$  (donc  $F$  ne dépend pas de  $X_i$ )
- $F$  est connue, à valeur dans  $]0, 1[$  et trois fois continûment dérivable sur  $\mathbb{R}$
- La fonction  $(\ln(F(x)))'' = \frac{F''F - F'^2}{F^2}$  est négative sur  $\mathbb{R}$
- La fonction  $(\ln(F(x)))''' = \frac{F'''F^2 - 3F''F'F + 2F'^3}{F^3}$  est bornée sur  $\mathbb{R}$
- Pour tout  $(k_1, k_2, k_3) \in \{1, \dots, K\}^3$ , les moments suivants existent :  $E[|x_{k_1}|] < +\infty, E[|x_{k_1} x_{k_2}|] < +\infty, E[|x_{k_1} x_{k_2} x_{k_3}|] < +\infty$
- La quantité  $E_0 = E \left[ \frac{F'(X_1 \beta)^2}{F(X_1 \beta)(1-F(X_1 \beta))} X_1' X_1 \right]$  existe et est inversible.

Alors l'information de Fisher définie par  $I_0 = E \left[ -\frac{\partial^2 [y_1 \ln(F(X_1 \beta)) + (1-y_1) \ln(1-F(X_1 \beta))]}{\partial \beta \partial \beta'} \right]$  existe et est égale à  $E_0$ .

L'estimateur du maximum de vraisemblance défini par :

$$\hat{\beta} = \underset{\beta}{\text{Argmax}} \sum_i y_i \ln(F(X_i \beta)) + (1 - y_i) \ln(1 - F(X_i \beta))$$

1. On qualifie le modèle du score maximum de Manski de semi-paramétrique car bien que relâchant les hypothèses paramétriques sur la loi de  $\epsilon|X$ , la loi conditionnelle de  $y|X$  dépend d'une combinaison linéaire  $X\beta$  des covariables  $X$ . Et le paramètre d'intérêt  $\beta$  est donc de dimension finie (à la différence des modèles non-paramétriques dans lesquels le paramètre d'intérêt n'appartient pas à un espace de dimension finie).

est bien défini, converge vers  $\beta$  en probabilité et sa loi asymptotique est :

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow[n \rightarrow +\infty]{\text{en loi}} \mathcal{N}(0, I_0^{-1})$$

- Dans le cas d'un modèle Logit, on a :  $F(u) = \frac{e^u}{1+e^u}$  et  $I_0 = E \left[ \frac{e^{X\beta}}{(1+e^{X\beta})^2} X'X \right]$

- Dans le cas d'un modèle Probit, on a :  $F(u) = \Phi(u) = \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$  et  $I_0 = E \left[ \frac{1}{2\pi} \frac{e^{-X_1\beta}}{\Phi(X_1\beta)(1-\Phi(X_1\beta))} X_1'X_1 \right]$

On peut également remplacer  $I_0$  par sa contrepartie empirique  $\frac{1}{n} \sum_i \frac{F'(X_i\widehat{\beta})^2}{F(X_i\widehat{\beta})(1-F(X_i\widehat{\beta}))} X_i'X_i$ , pour construire des tests multivariés ou univariés :

$$n(\widehat{\beta} - b)' \widehat{I}_0(\widehat{\beta} - b) \xrightarrow[n \rightarrow +\infty]{\text{loi}} \chi^2(K+1) \text{ si } b = \beta$$

$$n(\widehat{\beta} - b)' \widehat{I}_0(\widehat{\beta} - b) \xrightarrow[n \rightarrow +\infty]{\text{P}} +\infty \text{ si } b \neq \beta$$

$$\sqrt{n} \sqrt{\widehat{I}_{0(k,k)}}(\widehat{\beta}_k - b_k) \xrightarrow[n \rightarrow +\infty]{\text{loi}} \mathcal{N}(0, 1) \text{ si } b_k \neq \beta_k$$

$$\sqrt{n} \sqrt{\widehat{I}_{0(k,k)}}(\widehat{\beta}_k - b_k) \xrightarrow[n \rightarrow \pm\infty]{\text{P}} +\infty \text{ si } b_k \neq \beta_k$$

**Remarque :** la convergence des paramètres est ici à prendre au sens où si une partie bornée  $\Theta$  de  $\mathbb{R}^{K+1}$  contient  $\beta$ , alors l'estimateur obtenu par maximisation de la vraisemblance sur  $\Theta$  est convergent et asymptotiquement normal. Pour comprendre d'où vient cette restriction, le lecteur intéressé pourra considérer le modèle simple sans covariable  $x$  :  $P(Y = 1) = F(\beta_0)$ . Ce modèle devient de plus en plus difficile à discriminer du modèle  $P(Y = 1) = 1$  quand  $\beta_0$  devient très grand (respectivement du modèle  $P(Y = 1) = 0$  quand  $\beta_0$  devient très petit). Le fait d'être obligé de se restreindre à n'importe quelle partie bornée de  $\mathbb{R}^{K+1}$  n'est pas un problème en pratique, l'ensemble des nombres manipulés par n'importe quel ordinateur étant borné, tout "fonctionnera" correctement à partir du moment où les composantes de  $\beta$  ne sont pas trop "grandes" pour pouvoir être représentées par l'ordinateur utilisé, avant d'en arriver à ce stade on se retrouverait dans le cas où toutes les observations de l'échantillon seraient telles que  $Y = 1$  ou  $Y = 0$ .

Preuve :

□

- Le résultat concernant l'information de Fisher est assuré par le fait que :

$$\begin{aligned} I_0 &= E \left[ \left\{ E(y_1|X_1) \left( \frac{F''(X_1\beta)}{F(X_1\beta)} - \frac{F'(X_1\beta)^2}{F(X_1\beta)^2} \right) - (1 - E(y_1|X_1)) \left( \frac{F''(X_1\beta)}{1-F(X_1\beta)} + \frac{F'(X_1\beta)^2}{(1-F(X_1\beta))^2} \right) \right\} X_1'X_1 \right] \\ &= E \left[ \frac{F'(X_1\beta)^2}{F(X_1\beta)(1-F(X_1\beta))} X_1'X_1 \right] = E_0 \end{aligned}$$

De plus  $y^2 = y$ ,  $y(1-y) = 0$  et  $(1-y)^2 = 1-y$ ,  $E_0$  peut également s'interpréter de la manière suivante :

$$\begin{aligned} E_0 &= E \left[ \frac{F'(X_1\beta)^2}{F(X_1\beta)(1-F(X_1\beta))} X_1'X_1 \right] = \\ &E \left[ \left( \frac{\partial [y_1 \ln(F(X_1\beta)) + (1-y_1) \ln(1-F(X_1\beta))]}{\partial \beta} \right) \left( \frac{\partial [y_1 \ln(F(X_1\beta)) + (1-y_1) \ln(1-F(X_1\beta))]}{\partial \beta} \right)' \right] \end{aligned}$$

- Vérifions maintenant que  $\beta$  est identifiable. Pour assurer l'identification de  $\beta$ , il suffit de montrer que presque sûrement  $F(X_1\beta) = F(X_1\beta') \Rightarrow \beta = \beta'$ . Comme  $F$  est continue et strictement croissante, elle est inversible donc  $F(X_1\beta) = F(X_1\beta') \Rightarrow X_1\beta = X_1\beta'$ . Le fait que la matrice  $\mathbf{X}$  soit de plein rang colonne assure qu'il n'existe pas de relation linéaire presque sûre entre les  $X_1$ . Donc  $\beta$  est identifiable.

- Il faut ensuite s'assurer que le maximum est unique.

L'estimateur vérifie les  $K+1$  équations suivantes :

$$\forall k = 0, \dots, K \quad \frac{1}{n} \sum_i y_i \frac{F'(X_i\widehat{\beta})}{F(X_i\widehat{\beta})} x_{ik} - (1-y_i) \frac{F'(X_i\widehat{\beta})}{1-F(X_i\widehat{\beta})} x_{ik} = 0$$

Ce qui peut se réécrire de manière matricielle  $f_n(\beta) = \frac{1}{n} C(\widehat{\beta})' \mathbf{X} = 0$  avec  $C(\widehat{\beta})$  un vecteur colonne de composante  $y_i \frac{F'(X_i\widehat{\beta})}{F(X_i\widehat{\beta})} - (1-y_i) \frac{F'(X_i\widehat{\beta})}{1-F(X_i\widehat{\beta})}$ .

La solution de l'équation  $C(b)'X$  est unique si tous les éléments diagonaux de  $H(b) = \text{Diag}\left(\frac{F''(X_i b)F(X_i b) - F'^2(X_i b)}{F^2(X_i b)}\right)$  sont négatifs : en effet, la hessienne de la log-vraisemblance au point  $b$  est  $X'H(b)X$ . Comme  $H(b)$  peut s'écrire  $-D(b)^2$  où  $D(b)$  est une matrice diagonale à coefficients strictement positifs, le fait que  $X'H(b)X$  soit définie négative est équivalente au fait que  $D(b)X$  soit de plein rang colonne. Comme  $D(b)$  est de rang  $n$ , cela est équivalent au fait que  $X$  soit de plein rang colonne.

– Examinons maintenant la convergence en probabilité de l'estimateur.

Soit  $\Theta$  une partie bornée de  $\mathbb{R}^{K+1}$ . Il existe  $A \in \mathbb{R}$ , tel que pour tout  $b \in \Theta$ ,  $\max(b_0, b_1, \dots, b_K) < A$ .

Pour alléger les notations, on note :

$$M_n(b) = \frac{1}{n} \sum y_i \ln(F(X_i b)) + (1 - y_i) \ln(1 - F(X_i b))$$

$$M(b) = E [y_1 \ln(F(X_1 b)) + (1 - y_1) \ln(1 - F(X_1 b))]$$

Pour assurer la convergence en probabilité de  $\hat{\beta}$ , il suffit de montrer que :

$$- \sup_{b \in \Theta} |M_n(b) - M(b)| \xrightarrow[n \rightarrow +\infty]{\text{proba}} 0$$

$$- \forall \alpha > 0, \sup_{b, |b - \beta| \geq \alpha} M(b) < E_0 = M(\beta)$$

La première condition est la plus difficile à prouver.

La fonction  $b \mapsto |y \ln(F(Xb)) + (1 - y) \ln(1 - F(Xb))|$  est continue pour toute valeur de  $y$  et  $X$  et est dominée par  $|\ln(F(0)) + Xb \left[ y \frac{F'(0)}{F(0)} - (1 - y) \frac{F'(0)}{1 - F(0)} \right]| \leq |\ln(F(0))| +$

$A|X| \frac{F'(0)}{\max(F(0), 1 - F(0))}$  qui est intégrable ; les théorèmes usuels assurent alors que  $\sup_{b \in \Theta} y \ln(F(Xb)) + (1 - y) \ln(1 - F(Xb))$  est une fonction continue et bornée de  $y, X$ , pour toute boule fermée  $\Theta^*$  centrée sur un point  $b^*$  de  $\Theta$  et de rayon  $r$ .

On en déduit que :

$$h_n(\Theta^*) = \sup_{b \in \Theta^*} (M_n(b) - M(b)) \leq$$

$$\frac{1}{n} \sum \sup_{b \in \Theta^*} \{y_i \ln(F(X_i b)) + (1 - y_i) \ln(1 - F(X_i b))\} - E [inf_{b \in \Theta^*} \{y_1 \ln(F(X_1 b)) + (1 - y_1) \ln(1 - F(X_1 b))\}]$$

Le lemme de Portemanteau assure que :

$$\frac{1}{n} \sum \sup_{b \in \Theta^*} \{y_i \ln(F(X_i b)) + (1 - y_i) \ln(1 - F(X_i b))\} \rightarrow$$

$$E [\sup_{b \in \Theta^*} \{y_1 \ln(F(X_1 b)) + (1 - y_1) \ln(1 - F(X_1 b))\}]$$

On en déduit que pour  $n$  suffisamment grand,

$$h_n(\Theta^*) \leq 2E [\sup_{b \in \Theta^*} \{y_1 \ln(F(X_1 b)) + (1 - y_1) \ln(1 - F(X_1 b))\} - inf_{b \in \Theta^*} \{y_1 \ln(F(X_1 b)) + (1 - y_1) \ln(1 - F(X_1 b))\}]$$

Le théorème de convergence dominée assure que le membre de droite de l'inégalité tend vers 0 quand le rayon  $r$  de la boule  $\Theta^*$  tend vers 0.

Donc  $\forall b^* \in \Theta, \forall \eta > 0, \exists r > 0$  tel que  $h_n(\Theta^*) \leq 2\eta$  pour  $n$  suffisamment grand.

Comme  $\Theta$  est borné, on peut le recouvrir avec un ensemble fini  $\mathcal{J}$  de  $\Theta^*$  (autrement dit un nombre fini de  $b^*, r$ ) vérifiant cette dernière propriété. Comme  $h_n(\Theta) = \sup_{b \in \Theta} (M_n(b) - M(b))$  est égal à  $\sup_{\Theta^* \in \mathcal{J}} h_n(\Theta^*)$ , on en déduit que :

$\forall \eta > 0$ , pour  $n$  suffisamment grand,  $h_n(\Theta) \leq 2\eta$ . De la même manière, pour  $n$  suffisamment grand  $h_n(\Theta) \geq 2\eta$  (en raisonnant sur les quantités  $-y_i \ln(F(X_i b)) - (1 - y_i) \ln(1 - F(X_i b))$ ).

On en déduit la première partie de la condition suffisante pour assurer la convergence en probabilité.

La deuxième partie de la condition est une conséquence immédiate de :

- Pour tout  $b \in \Theta$ ,  $M(b) = E [F(X_1 b) \ln(F(X_1 b)) + (1 - F(X_1 b)) \ln(1 - F(X_1 b))]$
- Pour  $a, b$  dans l'intervalle  $]0, 1[$  avec  $a \neq b$ ,  $a \ln(b) + (1 - a) \ln(1 - b) < a \ln(a) + (1 - a) \ln(1 - a)$

Par définition de l'estimateur,  $M_n(\hat{\beta}) \geq M_n(\beta)$ , on a donc :

$$M(\beta) - M(\hat{\beta}) \leq M_n(\hat{\beta}) - M(\hat{\beta}) + M(\beta) - M_n(\beta) \leq 2 \sup_{b \in \Theta} |M_n(b) - M(b)|$$

Le terme de gauche tend vers 0 en probabilité d'après la première condition.

Considérons maintenant un  $\alpha \geq 0$  tel que  $|\hat{\beta} - \beta| > \alpha$ . D'après la deuxième condition, il existe  $\eta > 0$  tel que  $M(\hat{\beta}) < M(\beta) - \eta$ . On en déduit que :

$$\forall \alpha > 0, \exists \eta > 0, P(|\hat{\beta} - \beta| > \alpha) \leq P(M(\beta) - M(\hat{\beta}) > \eta) \leq P(2 \sup_{b \in \Theta} |M_n(b) - M(b)| > \eta)$$

Le dernier terme tend vers 0 quel que soit  $\eta$  donc  $\widehat{\beta}$  tend vers  $\beta$  en probabilité.

– Enfin concluons sur la normalité asymptotique.

Un développement classique nous assure qu'il existe  $\tilde{\beta}$  un vecteur compris dans le segment  $[\beta, \widehat{\beta}]$ , tel que pour chacune des  $K + 1$  équations définissant l'estimateur :

$$\begin{aligned} 0 &= \frac{1}{n} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \widehat{\beta}) \right) \right]' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \widehat{\beta}) \right) \right]' \right\} x_{ik} \\ &= \frac{1}{n} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \beta) \right) \right]' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \beta) \right) \right]' \right\} x_{ik} \\ &+ \frac{1}{n} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \beta) \right) \right]'' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \beta) \right) \right]'' \right\} x_{ik} X_i (\widehat{\beta} - \beta) \\ &+ \frac{1}{2n} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \tilde{\beta}) \right) \right]''' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \tilde{\beta}) \right) \right]''' \right\} x_{ik} (\widehat{\beta} - \beta)' X_i' X_i (\widehat{\beta} - \beta) \end{aligned}$$

La quantité  $\frac{1}{2n} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \tilde{\beta}) \right) \right]''' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \tilde{\beta}) \right) \right]''' \right\}$  est bornée et comme  $\widehat{\beta} - \beta$  tend vers 0 en probabilité, que les moments d'ordre 3 des covariables sont finis  $R_n^k = \frac{1}{2n} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \tilde{\beta}) \right) \right]''' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \tilde{\beta}) \right) \right]''' \right\} x_{ik} (\widehat{\beta} - \beta)' X_i' X_i$  tend en probabilité vers 0 pour tout  $k \in \{0, \dots, K\}$ . La matrice carrée  $R_n$  constituée des  $K + 1$  vecteurs lignes  $R_n^k$ , tend donc en probabilité vers 0.

La quantité  $\frac{1}{n} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \beta) \right) \right]'' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \beta) \right) \right]'' \right\} X_i' X_i$  (obtenue en considérant les  $K + 1$  équations estimantes) tend en probabilité vers  $-I_0$  d'après la loi des grands nombres.  $S_n = \frac{1}{n} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \beta) \right) \right]'' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \beta) \right) \right]'' \right\} X_i' X_i + I_0$  est donc une matrice tendant en probabilité vers 0.

Ainsi :

$$-\frac{1}{n} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \beta) \right) \right]' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \beta) \right) \right]' \right\} X_i = (-I_0 + S_n + R_n) (\widehat{\beta} - \beta)$$

La convergence vers 0 en probabilité de  $S_n + R_n$  et l'inversibilité de  $I_0$  nous assure que  $-I_0 + S_n + R_n$  est inversible avec une probabilité tendant vers 1.

En prémultipliant les deux termes de l'équation par  $T_n = \sqrt{n}(-I_0 + S_n + R_n)^{-1}$  si  $-I_0 + S_n + R_n$  est inversible et  $T_n = 0$  sinon, on obtient :

$$-T_n \frac{1}{n} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \beta) \right) \right]' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \beta) \right) \right]' \right\} X_i = \sqrt{n} (\widehat{\beta} - \beta) + U_n$$

Avec  $U_n$  qui tend vers 0 en probabilité.

Comme  $\frac{1}{\sqrt{n}} \sum_i \left\{ y_i \left[ \ln \left( F(X_i \beta) \right) \right]' + (1 - y_i) \left[ \ln \left( 1 - F(X_i \beta) \right) \right]' \right\} X_i$  tend en loi vers une gaussienne d'espérance nulle (car  $E(y_1 | X_1) = F(X_1 \beta)$ ) et de matrice de variance-covariance  $E_0$ . Donc :

$$\sqrt{n} (\widehat{\beta} - \beta) \xrightarrow[\text{en loi}]{n \rightarrow +\infty} \mathcal{N}(0, I_0^{-1} E_0 I_0^{-1})$$

■

## 6.2 Problème de biais du Logit et du Probit à effets fixes

Les modèles Logit et Probit sont très largement utilisés pour analyser des phénomènes tels que le redoublement, le fait qu'un élève donne une bonne ou une mauvaise réponse à une question dans un QCM, etc. Il existe par ailleurs de nombreuses raisons de vouloir contrôler un effet "classe" lors de l'analyse : le professeur peut être plus ou moins sévère dans ses jugements concernant

l'orientation ou la qualité des réponses de ses élèves, les élèves d'une même classe peuvent partager des caractéristiques communes, inobservées du sociologue ou du statisticien, qui influencent la variable expliquée. Par analogie avec le modèle linéaire, une idée assez naturelle est alors d'ajouter des indicatrices de classes dans l'analyse. C'est néanmoins une très mauvaise idée quand le nombre d'élèves observés dans chaque classe est faible. En effet, les modèles Logit et Probit sont des estimateurs du maximum de vraisemblance, dont les propriétés sont bonnes lorsque la taille de l'échantillon devient très grand devant le nombre de paramètres. Si on rajoute un paramètre pour chaque classe et qu'on a des données sur 5 élèves pour chaque classe, alors la taille de l'échantillon ne pourra jamais être plus de 5 fois plus importante que le nombre de paramètres. Les estimations des paramètres autres que les effets "classe" peuvent alors être biaisées de manière importante. La mauvaise estimation des effets "classe" vient contaminer l'estimation des autres coefficients. C'est ce qu'on appelle le problème des paramètres incidents. Dans le cas du modèle linéaire présenté au chapitre 3 ce problème n'apparaissait pas car les projections inter-classe et intra-classe sont orthogonales, autrement dit toute l'information contenue dans les données permettant d'identifier les effets "classe" est contenue dans la moyenne des variables par classe et toute l'information contenue dans les données permettant d'identifier les effets des covariables est contenue dans l'écart des variables individuelles à la moyenne de classe. Ces deux statistiques (moyenne de classe et écart à la moyenne de classe) sont complètement "libres" l'une de l'autre dans un modèle linéaire. En dehors du cas linéaire, le problème des paramètres incidents se pose.

Pour illustrer ce problème, on se contentera d'un exemple développé par Andersen en 1973 dans "conditional inference and models for measuring" publié dans Mentalhygiejnisk Forlag et repris par Abrevaya en 1997 dans "The equivalence of two estimators for the fixed effects logit model" publié en 1997 dans Economics Letters.

### 6.2.1 Illustration du problème des paramètres "incidents"

On se place dans l'hypothèse où l'output d'une variable binaire  $y$  (par exemple donner ou non la bonne réponse à un exercice scolaire) dépend d'une covariable individuelle également binaire (par exemple le sexe) et où il existe des caractéristiques inobservées qui influent sur les résultats des élèves d'une même classe (on peut penser à la qualité pédagogique du professeur qui n'est pas directement observable).

#### Un contre-exemple dû à Chamberlain

*Des données sont issues du modèle suivant :*

$$y_{ij} = \mathbb{1}_{\{x_{ij}\beta + \gamma_j + \epsilon_i > 0\}}$$

avec  $P(-\epsilon_i \leq u | X, \gamma) = \frac{e^u}{1+e^u}$  et les  $\gamma_j$  indépendants et identiquement distribués.

On observe un garçon ( $x_{ij} = 0$ ) et une fille ( $x_{i'j} = 1$ ) dans chaque classe  $j$ .

Dans ces conditions, l'estimateur du maximum de vraisemblance de  $\beta$  converge vers  $2\beta$

**Remarque :** Le fait de supposer que les  $\gamma_i$  sont indépendants et identiquement distribués ne signifie pas qu'ils ne sont pas corrélés à d'autres variables (observées ou non). Cette hypothèse concernant les  $\gamma_i$  signifie seulement que l'observation des résultats d'une classe n'apporte pas d'information sur les résultats d'une autre classe.

Preuve :

□

On suppose sans perte de généralité que les observations relatives aux garçons sont indicées par 1 et celles des filles par 2 au sein de chaque classe  $j$ .

Il y a quatre types de classes :

- Les classes telles que le garçon et la fille observés aient un  $y$  à 0 ; dans ce cas, on dira que  $j \in E_{00} = \{j | y_{1j} = y_{2j} = 0\}$
- Les classes telles que le garçon observé ait un  $y$  à 0 et la fille observée ait un  $y$  à 1 ; dans ce cas, on dira que  $j \in E_{01} = \{j | y_{1j} = 0, y_{2j} = 1\}$

- Les classes telles que le garçon observé ait un  $y$  à 1 et la fille observée ait un  $y$  à 0 ; dans ce cas, on dira que  $j \in E_{10} = \{j | y_{1j} = 1, y_{2j} = 0\}$
- Les classes telles que le garçon et la fille observés aient un  $y$  à 1 ; dans ce cas, on dira que  $j \in E_{11} = \{j | y_{1j} = y_{2j} = 1\}$

La log-vraisemblance de la classe  $j$  s'écrit :

$$\mathcal{L}_j = \ln \left( \frac{e^{y_{1j}\gamma_j}}{1+e^{\gamma_j}} \right) + \ln \left( \frac{e^{y_{2j}(\beta+\gamma_j)}}{1+e^{\beta+\gamma_j}} \right) = (y_{1j} + y_{2j})\gamma_j + y_{2j}\beta - \ln(1 + e^{\gamma_j}) - \ln(1 + e^{\beta+\gamma_j})$$

Pour  $j \in E_{00}$ , il apparaît alors que  $\mathcal{L}_j \leq 0 = \lim_{\gamma_j \rightarrow -\infty} -\ln(1 + e^{\gamma_j}) - \ln(1 + e^{\beta+\gamma_j})$ . Ainsi les classes dans lesquelles tout les individus observés ont une valeur de  $y$  nulle auront comme estimateur  $\hat{\gamma}_j = -\infty$ . De plus, ces classes n'apportent aucune information sur le paramètre  $\beta$ , puisque la contribution à la vraisemblance de ces classes sera nulle.

Il en est de même pour les classes dans lesquelles tous les individus observés ont une valeur de  $y$  égale à 1 (dans ce cas  $j \in E_{11}$ ). La contribution à la vraisemblance est maximale et nulle quelque soit  $\beta$  pour  $\hat{\gamma}_j = +\infty$  (car  $\ln(1 + e^u) = u + m(u)$  avec  $\lim_{u \rightarrow +\infty} m(u) = 0$ ).

Pour les classes  $j \in E_{01} \cup E_{10}$ , les conditions de maximisation du premier ordre concernant les paramètres  $\gamma_j$  sont :

$$y_{1j} + y_{2j} = \frac{e^{\hat{\gamma}_j}}{1+e^{\hat{\gamma}_j}} + \frac{e^{\hat{\beta}+\hat{\gamma}_j}}{1+e^{\hat{\beta}+\hat{\gamma}_j}}$$

Ce qui conduit aux estimateurs suivants concernant  $\gamma_j$  :

- $\hat{\gamma}_j = -\infty$  si  $j \in E_{00}$
- $\hat{\gamma}_j = +\infty$  si  $j \in E_{11}$
- $\hat{\gamma}_j = -\frac{\hat{\beta}}{2}$  si  $j \in E_{01} \cup E_{10}$

Les seuls termes de la vraisemblance globale qui apportent une information sur  $\beta$  sont donc ceux correspondant aux classes  $E_{01}$  et  $E_{10}$ . La condition du premier ordre concernant le paramètre  $\beta$  est donc :

$$\text{card}(E_{01}) = \sum_{j \in E_{01} \cup E_{10}} \frac{e^{\hat{\beta}+\hat{\gamma}_j}}{1+e^{\hat{\beta}+\hat{\gamma}_j}} = (\text{card}(E_{01}) + \text{card}(E_{10})) \frac{e^{\frac{\hat{\beta}}{2}}}{1+e^{\frac{\hat{\beta}}{2}}}$$

$$\text{Donc } \hat{\beta} = 2 \ln \left( \frac{\text{card}(E_{01})}{\text{card}(E_{10})} \right)$$

Si on suppose de plus que les  $\gamma_j$  sont les réalisations de variables aléatoires indépendantes et identiquement distribuées dont la fonction de répartition est  $F$ , alors :  $E(\mathbb{1}_{\{j \in E_{01}\}}) = e^\beta \int \frac{1}{(1+e^u)(1+e^{\beta+u})} dF(u) = e^\beta E(\mathbb{1}_{\{j \in E_{10}\}})$

La loi des grands nombres nous assure alors que :  $\frac{\text{card}(E_{01})}{\text{card}(E_{10})} \xrightarrow[n \rightarrow +\infty]{\text{proba}} e^\beta$

Finalement, comme la fonction exponentielle est continue :  $\hat{\beta} \xrightarrow[n \rightarrow +\infty]{\text{proba}} 2\beta$

■

Il existe différentes stratégies face au problème des paramètres incidents. Nous nous contenterons de présenter deux approches : celle du maximum de vraisemblance conditionnelle et celle d'une spécification à la Mundlak.

## 6.3 Le Logit conditionnel

### 6.3.1 Principe du maximum de vraisemblance conditionnelle

Le problème des paramètres incidents vient du fait que la vraisemblance du modèle dépend de deux ensembles de paramètres :

- d'une part des paramètres relatifs aux covariables individuelles (les  $\beta$ ). La taille de cet ensemble est fixe lorsque la taille de l'échantillon augmente.
- d'autre part des paramètres relatifs à l'hétérogénéité entre classes (les effets "classes"  $\gamma$ ). La taille de cet ensemble augmente avec la taille de l'échantillon (quand la taille d'échantillon tend vers  $+\infty$ , le nombre de classes sélectionnées augmente mécaniquement car le nombre d'élèves par classe est borné).

Le principe du maximum de vraisemblance conditionnelle consiste à trouver une statistique qui permet de "séparer" les problèmes pour se ramener à une estimation des  $\beta$  indépendante de celle des  $\gamma$ .

#### Le maximum de vraisemblance conditionnelle

Soit un modèle dont la vraisemblance  $\mathcal{L}_{\mathbf{y}, \mathbf{X}}(\beta, \gamma)$  dépend d'un paramètre  $\beta$  de taille fixe et d'un paramètre  $\gamma$  tel que  $\frac{\dim(\gamma)}{n}$  ne tende pas vers 0 avec la taille  $n$  de l'échantillon.

Un tel modèle est soumis au problème des paramètres incidents, une estimation  $\beta$  par maximisation de la vraisemblance par rapport à  $\beta$  et  $\gamma$  conduira à des estimations biaisées.

S'il existe une statistique  $S$  telle que la vraisemblance conditionnelle de  $y$  par rapport à  $X, S$  ne dépende pas de  $\gamma^2$  :

$$\mathcal{L}_{\mathbf{y}|\mathbf{X}, S}(\beta, \gamma) = f(\mathbf{y}, \mathbf{X}, S, \beta)$$

Alors l'estimateur de  $\beta$  obtenu par maximisation de  $f$  est bien défini, asymptotiquement sans biais, convergent et asymptotiquement normal de loi limite :

$$\sqrt{n} \left( \hat{\beta} - \beta \right) \xrightarrow[n \rightarrow +\infty]{\text{loi}} \mathcal{N}(0, V)$$

dès lors que la fonction  $f$  et la loi de  $y$  et  $X$  vérifient des conditions de régularité suffisante, par exemple :

- $f$  est trois fois continûment différentiable par rapport à  $\beta$  et  $\frac{\partial^3 \ln f(\mathbf{y}, \mathbf{X}, S, \beta)}{\partial \beta^3}$  est majorée par  $g(\mathbf{y}, \mathbf{X}, S)$  tel que  $E(g(\mathbf{y}, \mathbf{X}, S)) < +\infty$
- $E \left[ \left( \frac{\partial \ln f(\mathbf{y}, \mathbf{X}, S, \beta)}{\partial \beta} \right) \left( \frac{\partial \ln f(\mathbf{y}, \mathbf{X}, S, \beta)}{\partial \beta} \right)' \right]$  existe et  $E \left[ \frac{\partial^2 \ln f(\mathbf{y}, \mathbf{X}, S, \beta)}{\partial \beta \partial \beta'} \right]$  existe et est inversible

La mise en oeuvre de l'estimation et des tests asymptotiques est alors identique à celle adoptée dans le cas du maximum de vraisemblance.

**Remarque :** dans le cas du modèle linéaire exposé au chapitre 3, si on fait l'hypothèse que les résidus  $\epsilon$  sont gaussiens, alors la régression "Within" est équivalente à la maximisation de la vraisemblance conditionnelle à  $S = (B\mathbf{X}, B\mathbf{y})$ ; une fois fixées les moyennes par classe, les variations de  $y$  et  $X$  n'apportent aucune information sur  $\gamma$ .

Preuve :

□

Il s'agit d'une conséquence directe des théorèmes généraux du maximum de vraisemblance appliqué au modèle conditionnel  $\mathbf{y}|\mathbf{X}, S$

■

### 6.3.2 Application aux modèles Logit

Pour pouvoir appliquer la stratégie d'estimation présentée dans le paragraphe précédent, il faut donc trouver une statistique "libre" par rapport à  $\gamma$ . Dans le cas où le modèle de départ est un modèle logit, alors l'existence d'une telle statistique est assurée. À l'inverse, Magnac a montré que tant qu'on observe exactement deux élèves par classe dans l'échantillon, d'autres modèles que le modèle logit admettent une telle statistique libre, mais que dès qu'on observe des classes avec trois élèves dans l'échantillon, le seul modèle qui convient pour une telle approche est le modèle logit<sup>3</sup> (cf. "Panel Binary Variables and Sufficiency : Generalizing Conditional Logit" publié en 2004 dans *Econometrica*). Cela signifie que l'approche par maximum de vraisemblance conditionnelle repose essentiellement sur la loi des résidus (contrairement aux modèles sans effets "classe" comme le montre l'estimateur du score maximum de Manski).

#### Le logit conditionnel

On suppose que :

$$P(y_{ij} = 1|\mathbf{X}) = F(X_i\beta + \gamma_j)$$

avec  $F(u) = \frac{e^u}{1+e^u}$ .

2. Une telle statistique  $S$  est dite "exhaustive" par rapport à  $\gamma$  car elle contient toute l'information concernant ce paramètre, et la statistique  $\mathbf{y}|\mathbf{X}, S$  est dite "libre" par rapport à  $\gamma$  car elle ne dépend pas de ce paramètre.

3. Si on ne croit pas à une fonction de lien logistique, il est possible d'élargir la classe des modèles en raisonnant sur l'ensemble des couples d'élèves au sein de chaque classe, on se ramène ainsi au cas où on a deux élèves par classe. Nous laissons de côté cette approche, qui est plus compliquée à implémenter, pour plus de précision le lecteur intéressé pourra se reporter aux travaux cités de Magnac.

Pour une classe  $j$ , on note  $n_j$  le nombre d'élèves observés,  $m_j$  le nombre d'élèves observés ayant un  $y$  prenant la valeur 1 (donc  $n_j - m_j$  représente le nombre d'élèves observés ayant la valeur 0). Si  $n_j - 1 \geq m_j \geq 1$  (i.e. tous les élèves observés dans la classe n'ont pas la même valeur pour  $y$ ), on a alors pour  $\delta_1 \in \{0, 1\}, \delta_2 \in \{0, 1\}, \dots, \delta_{n_j} \in \{0, 1\}$  tels que  $\sum_{i=1}^{n_j} \delta_i = m_j$  :

$$P \left( y_{1j} = \delta_1, y_{2j} = \delta_2, \dots, y_{n_j j} = \delta_{n_j} | \mathbf{X}, \sum_{i=1}^{n_j} y_{ij} = \sum_{i=1}^{n_j} \delta_i = m_j \right) = \frac{\exp \left( \sum_{i=1}^{n_j} y_{ij} \cdot X_i \beta \right)}{\sum_{(\tilde{y}_1, \dots, \tilde{y}_{n_j}) \in B(n_j, m_j)} \exp \left( \sum_{i=1}^{n_j} \tilde{y}_{ij} \cdot X_i \beta \right)}$$

avec  $B(n_j, m_j)$  qui représente l'ensemble des  $n_j$ -uplets ayant  $m_j$  1 et  $n_j - m_j$  0 :

$$B(n_j, m_j) = \left\{ (\tilde{y}_1, \dots, \tilde{y}_{n_j}) \in \{0, 1\}^{n_j} \mid \sum_{i=1}^{n_j} \tilde{y}_i = m_j \right\}$$

La vraisemblance conditionnelle par rapport à la statistique  $S = (n_j, \sum_{i=1}^{n_j} y_i = m_j)$  de la classe  $j$  ne dépend donc plus des paramètres  $\gamma$  mais dépend des paramètres  $\beta$ , on peut donc maximiser cette vraisemblance conditionnelle pour obtenir des estimateurs asymptotiquement convergents et normaux de  $\beta$  (sous les hypothèses que  $(n_j, m_j, \gamma_j)_{j=1, \dots, C}$  sont i.i.d. et que les  $(\varepsilon_{ij}, X_{ij})_{i=1, \dots, n_j}$  sont i.i.d. conditionnellement à  $n_j, \gamma_j$ , que  $P(n_j > 1) > 0$  et qu'il existe  $\bar{n}$  tel que  $P(\bar{n} > n_j) = 1$ ).

**Remarque :** les classes dans lesquelles tout les élèves prennent la valeur 0, ou prennent la valeur 1 ne contribuent pas à l'estimation de  $\beta$ . Cela provient du fait qu'en l'absence de variation au sein d'une classe, il est impossible d'extraire la moindre information permettant de mettre en évidence une covariation intra-classe entre  $x$  et  $y$ .

Un tel modèle s'étend facilement au cas où  $y$  est une variable multinomiale ( $y \in \{0, 1, \dots, L\}$ ).

Dans ce cas, en prenant 0 comme modalité de référence pour la variable  $y$ , on a  $L$  effets fixes pour chaque classe et  $P(y_{1j} = l_1, y_{2j} = l_2, \dots, y_{n_j j} = k_{n_j} | \mathbf{X}, \gamma) = \prod_{i=1}^{n_j} \frac{\sum_{l=1}^L \mathbb{1}_{\{l=l_i\}} \exp(X_{ij} \beta_l + \gamma_{lj})}{1 + \sum_{l=1}^L \exp(X_{ij} \beta_l + \gamma_{lj})}$  ;

la statistique de conditionnement est  $S = (s_0 = \sum_{i \in j} \mathbb{1}_{\{y_i=0\}}, s_1 = \sum_{i \in j} \mathbb{1}_{\{y_i=1\}}, \dots, s_L = \sum_{i \in j} \mathbb{1}_{\{y_i=L\}})$  ; et la vraisemblance conditionnelle s'écrit :  $P(y_{1j} = l_1, y_{2j} = l_2, \dots, y_{n_j j} = k_{n_j} | S, \mathbf{X}, \gamma) =$

$$\frac{\exp \left( \sum_{l=1}^L \sum_{i=1}^{n_j} \mathbb{1}_{\{l=l_i\}} X_{ij} \beta_l \right)}{\sum_{(\tilde{y}_1, \dots, \tilde{y}_{n_j}) \in B(S)} \exp \left( \sum_{l=1}^L \sum_{i=1}^{n_j} \mathbb{1}_{\{l=\tilde{y}_i\}} X_{ij} \beta_l \right)},$$

où  $B(S)$  est l'ensemble des  $n_j$  uplets contenant  $s_0$  0,

$s_1$  1, ...,  $s_L$  L.

Preuve :

□ Au niveau de la classe  $j$ , on a :

$$P(y_{1j} = \delta_1, y_{2j} = \delta_2, \dots, y_{n_j j} = \delta_{n_j} | \mathbf{X}) = \prod_{i=1}^{n_j} \frac{\exp(\delta_i(X_i \beta + \gamma_j))}{1 + \exp(X_i \beta + \gamma_j)}$$

$$= \exp \left( \sum_{i=1}^{n_j} \delta_i X_i \beta \right) \cdot \frac{\exp \left( \left( \sum_{i=1}^{n_j} \delta_i \right) \gamma_j \right)}{\prod_{i=1}^{n_j} 1 + \exp(X_i \beta + \gamma_j)}$$

La probabilité écrite sous cette forme, il apparaît que le premier terme ne dépend pas de gamma et que le deuxième terme dépend de  $y$  uniquement via la statistique  $S$ , on en déduit donc que :  $P(y_{1j} = \delta_1, y_{2j} = \delta_2, \dots, y_{n_j j} = \delta_{n_j} | \mathbf{X}, \sum_{i=1}^{n_j} y_{ij} = \sum_{i=1}^{n_j} \delta_i = m_j)$  ne dépend pas de  $\gamma$  :

$$P \left( y_{1j} = \delta_1, y_{2j} = \delta_2, \dots, y_{n_j j} = \delta_{n_j} | \mathbf{X}, \sum_{i=1}^{n_j} y_{ij} = \sum_{i=1}^{n_j} \delta_i = m_j \right)$$

$$\begin{aligned}
&= \frac{P(y_{1j} = \delta_1, y_{2j} = \delta_2, \dots, y_{n_j j} = \delta_{n_j} | \mathbf{X})}{\sum_{(\tilde{y}_1, \dots, \tilde{y}_{n_j}) \in B(n_j, m_j)} P(y_{1j} = \tilde{y}_{1j}, y_{2j} = \tilde{y}_{2j}, \dots, y_{n_j j} = \tilde{y}_{n_j j} | \mathbf{X})} \\
&= \frac{\exp\left(\sum_{i=1}^{n_j} y_{ij} \cdot X_i \beta\right)}{\sum_{(\tilde{y}_1, \dots, \tilde{y}_{n_j}) \in B(n_j, m_j)} \exp\left(\sum_{i=1}^{n_j} \tilde{y}_{ij} \cdot X_i \beta\right)}
\end{aligned}$$

■

## 6.4 Les modèles à effets aléatoires

Dans le cas du modèle linéaire, on a vu que les estimateurs du modèle de Mundlak étaient les mêmes que les estimateurs du modèle à effets fixes. Pour contourner le problème des paramètres incidents dans le cas d'un modèle non linéaire, on peut être tenté d'adopter une approche du type Mundlak.

C'est une solution empirique très pratique, mais dont il ne faut pas surestimer la cohérence théorique. Par exemple, dans l'approche de Mundlak, on introduit les moyennes des covariables par classe pour prendre en compte le fait qu'une part de l'hétérogénéité constante au sein de la classe est corrélée aux covariables, cette modélisation dépend donc du nombre d'élèves observés dans chaque classe, ce qui pose un problème de cohérence du modèle. Dans le cas du modèle linéaire, ce problème de cohérence ne se répercute pas sur l'estimation des coefficients des covariables qui varient au sein de la classe, les estimateurs convergent vers les paramètres du modèle à effets fixes. Dans le cas d'un modèle non linéaire, l'estimation des coefficients associés aux covariables qui varient au sein de la classe converge vers une valeur qui dépend de la modélisation de l'hétérogénéité et donc du nombre d'élèves observés par classe. Face à ce problème, on peut adopter une modélisation de l'hétérogénéité plus flexible comme le propose Chamberlain<sup>4</sup>, non pas en introduisant dans le modèle les moyennes des covariables mais les covariables de tous les élèves de la classe que l'économètre observe, le modèle de Mundlak est alors un sous-modèle du modèle de Chamberlain. Mais cela ne lève pas l'objection de fond sur le caractère "ad-hoc" de la modélisation de l'hétérogénéité constante au sein de la classe, modélisation "ad-hoc" qui conditionne les résultats de convergence.

### Le modèle théorique, estimation et tests

On suppose que :

$$P(y_{ij} = 1 | \mathbf{X}) = F(X_i \beta + \gamma_j)$$

avec  $F$  une fonction de répartition trois fois continûment dérivable.

On suppose que les  $n_j$  sont i.i.d., que  $P(n_j > 1) > 0$  et qu'il existe  $\bar{n}$  tel que  $P(\bar{n} > n_j) = 1$

Si de plus la loi de  $\gamma_j | X_{1j}, \dots, X_{n_j j}$  est de la forme :

$$\gamma_j | X_{1j}, \dots, X_{n_j j}, n_j \sim \mathcal{N}(\bar{X}_{.j} \lambda, \sigma_\gamma^2)$$

Alors la contribution à la vraisemblance des observations de la  $j$ ème classe est :

$$\mathcal{L} = \frac{1}{\sigma_\gamma} \int \prod_{i=1}^{n_j} F(X_{ij} \beta + \bar{X}_{.j} \lambda + u) \phi\left(\frac{u}{\sigma_\gamma}\right) du$$

Les estimateurs du maximum de vraisemblance de  $\theta_0 = (\beta, \lambda, \sigma^2)$  convergent et sont asymptotiquement normaux, la matrice de variance-covariance asymptotique  $V = E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta'}\right)^{-1} E\left(\frac{\partial \ln \mathcal{L}}{\partial \theta} \frac{\partial \ln \mathcal{L}}{\partial \theta'}\right)$   
 $E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta'}\right)^{-1} = -E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta'}\right)^{-1} = E\left(\frac{\partial \ln \mathcal{L}}{\partial \theta} \frac{\partial \ln \mathcal{L}}{\partial \theta'}\right)^{-1}$ .

4. Chamberlain a proposé ce modèle dans un cadre d'économétrie des panels, en plus des covariables contemporaines, il ajoute comme régresseurs supplémentaires les covariables à chaque période. Il n'y a pas d'ambiguïté sur ce qu'est pour chaque individu la valeur de ses covariables aux différentes dates, il n'en est pas de même dans notre cas : comment ordonner de manière commune les élèves des différentes classes ?

Les tests de Fisher et de Student usuels sont valides.

**Remarque :** les différentes expressions de la matrice de variance-covariance proviennent du fait que sous les hypothèses du modèle  $E\left(\frac{\partial \ln \mathcal{L}}{\partial \theta} \frac{\partial \ln \mathcal{L}}{\partial \theta'}\right) = -E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta'}\right)$ .

Preuve :

□

On procède en deux étapes : on prouve d'abord la convergence en probabilité et ensuite on dérive le comportement asymptotique.

La convergence en probabilité est assurée par deux propriétés : la loi des grands nombres implique que  $\frac{1}{C} \sum_{j=1}^C \ln \mathcal{L}(\mathbf{y}, \mathbf{X}, \theta)$  converge en probabilité vers  $E(\ln \mathcal{L}(\mathbf{y}, \mathbf{X}, \theta))$  pour tout  $\theta$ , c'est donc également le cas uniformément sur un compact  $\Theta$ , et par ailleurs la fonction  $\theta \mapsto E(\ln \mathcal{L}(\mathbf{y}, \mathbf{X}, \theta))$  est continue et a un unique maximum atteint en  $\theta_0$ . Ces deux conditions suffisent à assurer la convergence en probabilité de  $\hat{\theta}$  vers  $\theta_0$ .

Pour la normalité asymptotique, une méthode élémentaire consiste à effectuer un développement de Taylor à partir des équations d'annulation du score.

■

## 6.5 Programmation sous SAS...

### 6.5.1 ... du Logit conditionnel

La `proc logistic` permet d'estimer simplement un logit conditionnel grâce à l'instruction `strata`.

```
proc logistic data=TABLE;
  strata Z;
  model Y=X1 X2 ... XK;
run;
```

### 6.5.2 ... des modèles à effets aléatoires

Les modèles non linéaires à effets aléatoires peuvent s'estimer sous SAS grâce à la `proc nlmixed`, sous l'hypothèse de distribution normale des effets aléatoires. Pour que cette procédure puisse être utilisée pour de nombreux modèles, il faut programmer la vraisemblance conditionnelle à  $\gamma$  de chaque observation dans le corps de la procédure (cela revient à dire qu'on considère que  $\gamma$  est un vecteur de réels). Pour indiquer que la vraisemblance doit être intégrée par rapport à la distribution des effets aléatoires  $\gamma$ , on utilise l'instruction `random` suivi de l'option `subject`. Il y a plusieurs moyens de calculer et de spécifier la vraisemblance, dans le cas d'un modèle binaire, un moyen simple est de calculer la probabilité  $p$  associée à la réalisation de l'événement  $y = 1$ , puis de déclarer que la variable suit une loi de Bernoulli de paramètre  $p$  (ou autrement une binomiale  $(1, p)$ ) au moyen de l'instruction `binomial(1, p)`. De la même manière, la procédure permet de maximiser n'importe quelle vraisemblance (on déclare alors la log-vraisemblance au moyen de l'instruction `general`).

D'autre part ces estimations peuvent être longues à tourner, il est recommandé d'initialiser les paramètres avec des valeurs susceptibles d'être proches des "vrais" paramètres : par exemple en estimant dans une première étape le même modèle mais en omettant les effets aléatoires  $\gamma$  dans la spécification, on peut ainsi récupérer les estimations dans une table qu'on utilise pour initialiser la maximisation de la vraisemblance du modèle à effets aléatoires.

La procédure évalue par défaut la vraisemblance par une approximation d'une intégrale par une somme finie en utilisant des points et des poids de quadrature, par défaut elle détermine le nombre de points de quadrature et effectue un certain nombre de corrections à l'approximation la plus simple. Ces opérations sont cependant consommatrices en temps de calcul, il est souvent plus efficace de préciser qu'on renonce aux corrections de l'approximation (option `noad`) et de fixer soi-même le nombre de points de quadrature (option `qpoints=`); une dizaine de points de quadrature semble suffisante en pratique.

Dans l'exemple de programme ci-dessous, on effectue donc deux estimations d'un modèle logit, le premier sans effets aléatoires afin d'initialiser les paramètres du modèle plus lourd à estimer, le modèle logit à effets aléatoires.

```

ods html file='Logit_sans_effets_aleatoires.xls' style=minimal;
ods output parameterestimates=PARAM;
proc nlmixed data=TABLE;

    /*Déclaration et initialisation des paramètres*/
    parms CST=0 BETA_X1=0 BETA_X2=0 BETA_X3=0 ... BETA_XK=0;

    /*Construction de l'index X*beta*/
    index=CST+BETA_X1*X1+BETA_X2*X2+...+BETA_XK*XK;

    /*Construction de la probabilité, pour un probit  $p=\text{probnorm}(\text{index})$ */
    p=exp(index)/(1+exp(index));

    /*La variable Y suit une Bernoulli de paramètre p*/
    model Y~binomial(1,p);
run;
ods html close;

proc nlmixed data=TABLE noad qpoints=10;

    /*Un paramètre de plus à déclarer : SIGMA2GAMMA, les autres dans la table PARAM
    */
    parms SIGMA2GAMMA=0.5 /data=PARAM;

    /*On ajoute GAMMA dans l'index*/
    index=CST+BETA_X1*X1+BETA_X2*X2+...+BETA_XK*XK+GAMMA;

    /*Comme précédemment, le calcul de la proba p et Y suit une Bernoulli de
    paramètre p*/
    p=exp(index)/(1+exp(index));
    model Y~binomial(1,p);

    /*Et ici on spécifie la distribution des effets aléatoires*/
    random GAMMA normal(0,SIGMA2GAMMA) subject=Z;
run;

```

Pour estimer un modèle à la Mundlak, il suffit juste d'utiliser le programme précédent en incluant les moyennes par classe des covariables qui varient au sein des classes.

G 9001	J. FAYOLLE et M. FLEURBAEY Accumulation, profitabilité et endettement des entreprises		Macro-economic import functions with imperfect competition - An application to the E.C. Trade		françaises : une évaluation empirique des théories de la structure optimale du capital	G 9412	J. BOURDIEU - B. CŒURÉ - B. COLIN-SEDILLOT Investissement, incertitude et irréversibilité Quelques développements récents de la théorie de l'investissement
G 9002	H. ROUSSE Détection et effets de la multicollinéarité dans les modèles linéaires ordinaires - Un prolongement de la réflexion de BELSLEY, KUH et WELSCH	G 9203	I. STAPIC Les échanges internationaux de services de la France dans le cadre des négociations multilatérales du GATT Juin 1992 (1ère version) Novembre 1992 (version finale)	G 9312	L. BLOCH - B. CŒURÉ Q de Tobin marginal et transmission des chocs financiers	G 9413	B. DORMONT - M. PAUCHET L'évaluation de l'élasticité emploi-salaire dépend-elle des structures de qualification ?
G 9003	P. RALLE et J. TOUJAS-BERNATE Indexation des salaires : la rupture de 1983	G 9204	P. SEVESTRE L'économétrie sur données individuelles-temporelles. Une note introductive	G 9313	Equipes Amadeus (INSEE), Banque de France, Métrix (DP) Présentation des propriétés des principaux modèles macroéconomiques du Service Public	G 9414	I. KABLA Le Choix de breveter une invention
G 9004	D. GUELLEC et P. RALLE Compétitivité, croissance et innovation de produit	G 9205	H. ERKEL-ROUSSE Le commerce extérieur et l'environnement international dans le modèle AMADEUS (réestimation 1992)	G 9314	B. CREPON - E. DUGUET Research & Development, competition and innovation	G 9501	J. BOURDIEU - B. CŒURÉ - B. SEDILLOT Irreversible Investment and Uncertainty : When is there a Value of Waiting ?
G 9005	P. RALLE et J. TOUJAS-BERNATE Les conséquences de la désindexation. Analyse dans une maquette prix-salaires	G 9206	N. GREENAN et D. GUELLEC Coordination within the firm and endogenous growth	G 9315	B. DORMONT Quelle est l'influence du coût du travail sur l'emploi ?	G 9502	L. BLOCH - B. CŒURÉ Imperfections du marché du crédit, investissement des entreprises et cycle économique
G 9101	Equipe AMADEUS Le modèle AMADEUS - Première partie - Présentation générale	G 9207	A. MAGNIER et J. TOUJAS-BERNATE Technology and trade : empirical evidences for the major five industrialized countries	G 9316	D. BLANCHET - C. BROUSSE Deux études sur l'âge de la retraite	G 9503	D. GOUX - E. MAURIN Les transformations de la demande de travail par qualification en France Une étude sur la période 1970-1993
G 9102	J.L. BRILLET Le modèle AMADEUS - Deuxième partie - Propriétés variantielles	G 9208	B. CREPON, E. DUGUET, D. ENCAOUA et P. MOHNEN Cooperative, non cooperative R & D and optimal patent life	G 9317	D. BLANCHET Répartition du travail dans une population hétérogène : deux notes	G 9504	N. GREENAN Technologie, changement organisationnel, qualifications et emploi : une étude empirique sur l'industrie manufacturière
G 9103	D. GUELLEC et P. RALLE Endogenous growth and product innovation	G 9209	B. CREPON et E. DUGUET Research and development, competition and innovation : an application of pseudo maximum likelihood methods to Poisson models with heterogeneity	G 9318	D. EYSSARTIER - N. PONTY AMADEUS - an annual macro-economic model for the medium and long term	G 9505	D. GOUX - E. MAURIN Persistence des hiérarchies sectorielles de salaires: un réexamen sur données françaises
G 9104	H. ROUSSE Le modèle AMADEUS - Troisième partie - Le commerce extérieur et l'environnement international	G 9301	J. TOUJAS-BERNATE Commerce international et concurrence imparfaite : développements récents et implications pour la politique commerciale	G 9319	G. CETTE - Ph. CUNÉO - D. EYSSARTIER - J. GAUTIÉ Les effets sur l'emploi d'un abaissement du coût du travail des jeunes	G 9505	D. GOUX - E. MAURIN Bis Persistence of inter-industry wages differentials: a reexamination on matched worker-firm panel data
G 9105	H. ROUSSE Effets de demande et d'offre dans les résultats du commerce extérieur manufacturé de la France au cours des deux dernières décennies	G 9302	Ch. CASES Durées de chômage et comportements d'offre de travail : une revue de la littérature	G 9401	D. BLANCHET Les structures par âge importent-elles ?	G 9506	S. JACOBZONE Les liens entre RMI et chômage, une mise en perspective <i>NON PARU - article sorti dans Economie et Prévision n° 122 (1996) - pages 95 à 113</i>
G 9106	B. CREPON Innovation, taille et concentration : causalités et dynamiques	G 9303	H. ERKEL-ROUSSE Union économique et monétaire : le débat économique	G 9402	J. GAUTIÉ Le chômage des jeunes en France : problème de formation ou phénomène de file d'attente ? Quelques éléments du débat	G 9507	G. CETTE - S. MAHFOUZ Le partage primaire du revenu Constat descriptif sur longue période
G 9107	B. AMABLE et D. GUELLEC Un panorama des théories de la croissance endogène	G 9304	N. GREENAN - D. GUELLEC / G. BROUSSAUDIER - L. MIOTTI Innovation organisationnelle, dynamisme technologique et performances des entreprises	G 9403	P. QUIRION Les déchets en France : éléments statistiques et économiques	G 9601	Banque de France - CEPREMAP - Direction de la Prévision - Erasme - INSEE - OFCE Structures et propriétés de cinq modèles macro-économiques français
G 9108	M. GLAUDE et M. MOUTARDIER Une évaluation du coût direct de l'enfant de 1979 à 1989	G 9305	P. JAILLARD Le traité de Maastricht : présentation juridique et historique	G 9404	D. LADIRAY - M. GRUN-REHOMME Lissage par moyennes mobiles - Le problème des extrémités de série	G 9602	Rapport d'activité de la DESE de l'année 1995
G 9109	P. RALLE et alii France - Allemagne : performances économiques comparées	G 9306	J.L. BRILLET Micro-DMS : présentation et propriétés	G 9405	V. MAILLARD Théorie et pratique de la correction des effets de jours ouvrables	G 9603	J. BOURDIEU - A. DRAZNIKES L'octroi de crédit aux PME : une analyse à partir d'informations bancaires
G 9110	J.L. BRILLET Micro-DMS <b>NON PARU</b>	G 9307	J.L. BRILLET Micro-DMS - variantes : les tableaux	G 9406	F. ROSENWALD La décision d'investir	G 9604	A. TOPIOL-BENSAÏD Les implantations japonaises en France
G 9111	A. MAGNIER Effets accélérateur et multiplicateur en France depuis 1970 : quelques résultats empiriques	G 9308	S. JACOBZONE Les grands réseaux publics français dans une perspective européenne	G 9407	S. JACOBZONE Les apports de l'économie industrielle pour définir la stratégie économique de l'hôpital public	G 9605	P. GENIER - S. JACOBZONE Comportements de prévention, consommation d'alcool et tabagie : peut-on parler d'une gestion globale du capital santé ? <i>Une modélisation microéconométrique empirique</i>
G 9112	B. CREPON et G. DUREAU Investissement en recherche-développement : analyse de causalités dans un modèle d'accélérateur généralisé	G 9309	L. BLOCH - B. CŒURE Profitabilité de l'investissement productif et transmission des chocs financiers	G 9408	L. BLOCH, J. BOURDIEU, B. COLIN-SEDILLOT, G. LONGUEVILLE Du défaut de paiement au dépôt de bilan : les banquiers face aux PME en difficulté	G 9606	C. DOZ - F. LENGART Factor analysis and unobserved component models: an application to the study of French business surveys
G 9113	J.L. BRILLET, H. ERKEL-ROUSSE, J. TOUJAS-BERNATE "France-Allemagne Couplées" - Deux économies vues par une maquette macro-économétrique	G 9310	J. BOURDIEU - B. COLIN-SEDILLOT Les théories sur la structure optimal du capital : quelques points de repère	G 9409	D. EYSSARTIER, P. MAIRE Impacts macro-économiques de mesures d'aide au logement - quelques éléments d'évaluation	G 9607	N. GREENAN - D. GUELLEC La théorie coopérative de la firme
G 9201	W.J. ADAMS, B. CREPON, D. ENCAOUA Choix technologiques et stratégies de dissuasion d'entrée	G 9311	J. BOURDIEU - B. COLIN-SEDILLOT Les décisions de financement des entreprises	G 9410	F. ROSENWALD Suivi conjoncturel de l'investissement		
G 9202	J. OLIVEIRA-MARTINS, J. TOUJAS-BERNATE			G 9411	C. DEFEUILLEY - Ph. QUIRION Les déchets d'emballages ménagers : une analyse économique des politiques française et allemande		

G 9608	N. GREENAN - D. GUELLEC Technological innovation and employment reallocation
G 9609	Ph. COUR - F. RUPPRECHT L'intégration asymétrique au sein du continent américain : un essai de modélisation
G 9610	S. DUCHENE - G. FORGEOT - A. JACQUOT Analyse des évolutions récentes de la productivité apparente du travail
G 9611	X. BONNET - S. MAHFOUZ The influence of different specifications of wages-prices spirals on the measure of the NAIRU : the case of France
G 9612	PH. COUR - E. DUBOIS, S. MAHFOUZ, J. PISANI-FERRY The cost of fiscal retrenchment revisited: how strong is the evidence ?
G 9613	A. JACQUOT Les flexions des taux d'activité sont-elles seulement conjoncturelles ?
G 9614	ZHANG Yingxiang - SONG Xueqing Lexique macroéconomique Français-Chinois
G 9701	J.L. SCHNEIDER La taxe professionnelle : éléments de cadrage économique
G 9702	J.L. SCHNEIDER Transition et stabilité politique d'un système redistributif
G 9703	D. GOUX - E. MAURIN Train or Pay: Does it Reduce Inequalities to Encourage Firms to Train their Workers?
G 9704	P. GENIER Deux contributions sur dépendance et équité
G 9705	E. DUGUET - N. IUNG R & D Investment, Patent Life and Patent Value An Econometric Analysis at the Firm Level
G 9706	M. HOUEBINE - A. TOPIOL-BENSAÏD Les entreprises internationales en France : une analyse à partir de données individuelles
G 9707	M. HOUEBINE Polarisation des activités et spécialisation des départements en France
G 9708	E. DUGUET - N. GREENAN Le biais technologique : une analyse sur données individuelles
G 9709	J.L. BRILLET Analyzing a small French ECM Model
G 9710	J.L. BRILLET Formalizing the transition process : scenarios for capital accumulation
G 9711	G. FORGEOT - J. GAUTIÉ Insertion professionnelle des jeunes et processus de déclassement
G 9712	E. DUBOIS High Real Interest Rates: the Consequence of a Saving Investment Disequilibrium or of an insufficient Credibility of Monetary Authorities?
G 9713	Bilan des activités de la Direction des Etudes et Synthèses Economiques - 1996
G 9714	F. LEQUILLER Does the French Consumer Price Index Overstate Inflation?
G 9715	X. BONNET Peut-on mettre en évidence les rigidités à la baisse des salaires nominaux ? Une étude sur quelques grands pays de l'OCDE
G 9716	N. IUNG - F. RUPPRECHT Productivité de la recherche et rendements d'échelle dans le secteur pharmaceutique français
G 9717	E. DUGUET - I. KABLA Appropriation strategy and the motivations to use the patent system in France - An econometric analysis at the firm level
G 9718	L.P. PELÉ - P. RALLE Âge de la retraite : les aspects incitatifs du régime général
G 9719	ZHANG Yingxiang - SONG Xueqing Lexique macroéconomique français-chinois, chinois-français
G 9720	M. HOUEBINE - J.L. SCHNEIDER Mesurer l'influence de la fiscalité sur la localisation des entreprises
G 9721	A. MOURougANE Crédibilité, indépendance et politique monétaire Une revue de la littérature
G 9722	P. AUGERAUD - L. BRIOT Les données comptables d'entreprises Le système intermédiaire d'entreprises Passage des données individuelles aux données sectorielles
G 9723	P. AUGERAUD - J.E. CHAPRON Using Business Accounts for Compiling National Accounts: the French Experience
G 9724	P. AUGERAUD Les comptes d'entreprise par activités - Le passage aux comptes - De la comptabilité d'entreprise à la comptabilité nationale - A <i>paraître</i>
G 9801	H. MICHAUDON - C. PRIGENT Présentation du modèle AMADEUS
G 9802	J. ACCARDO Une étude de comptabilité générationnelle pour la France en 1996
G 9803	X. BONNET - S. DUCHÊNE Apports et limites de la modélisation « Real Business Cycles »
G 9804	C. BARLET - C. DUGUET - D. ENCAOUA - J. PRADEL The Commercial Success of Innovations An econometric analysis at the firm level in French manufacturing
G 9805	P. CAHUC - Ch. GIANELLA - D. GOUX - A. ZILBERBERG Equalizing Wage Differences and Bargaining Power - Evidence from a Panel of French Firms
G 9806	J. ACCARDO - M. JLASSI La productivité globale des facteurs entre 1975 et 1996
G 9807	Bilan des activités de la Direction des Etudes et Synthèses Economiques - 1997

G 9808	A. MOURougANE Can a Conservative Governor Conduct an Accommodative Monetary Policy ?
G 9809	X. BONNET - E. DUBOIS - L. FAUVET Asymétrie des inflations relatives et menus costs : tests sur l'inflation française
G 9810	E. DUGUET - N. IUNG Sales and Advertising with Spillovers at the firm level: Estimation of a Dynamic Structural Model on Panel Data
G 9811	J.P. BERTHIER Congestion urbaine : un modèle de trafic de pointe à courbe débit-vitesse et demande élastique
G 9812	C. PRIGENT La part des salaires dans la valeur ajoutée : une approche macroéconomique
G 9813	A.Th. AERTS L'évolution de la part des salaires dans la valeur ajoutée en France reflète-t-elle les évolutions individuelles sur la période 1979-1994 ?
G 9814	B. SALANIÉ Guide pratique des séries non-stationnaires
G 9901	S. DUCHÊNE - A. JACQUOT Une croissance plus riche en emplois depuis le début de la décennie ? Une analyse en comparaison internationale
G 9902	Ch. COLIN Modélisation des carrières dans Destinie
G 9903	Ch. COLIN Evolution de la dispersion des salaires : un essai de prospective par microsimulation
G 9904	B. CREPON - N. IUNG Innovation, emploi et performances
G 9905	B. CREPON - Ch. GIANELLA Wages inequalities in France 1969-1992 An application of quantile regression techniques
G 9906	C. BONNET - R. MAHIEU Microsimulation techniques applied to inter-generational transfers - Pensions in a dynamic framework: the case of France
G 9907	F. ROSENWALD L'impact des contraintes financières dans la décision d'investissement
G 9908	Bilan des activités de la DESE - 1998
G 9909	J.P. ZOYEM Contrat d'insertion et sortie du RMI Evaluation des effets d'une politique sociale
G 9910	Ch. COLIN - FI. LEGROS - R. MAHIEU Bilans contributifs comparés des régimes de retraite du secteur privé et de la fonction publique
G 9911	G. LAROQUE - B. SALANIÉ Une décomposition du non-emploi en France
G 9912	B. SALANIÉ Une maquette analytique de long terme du marché du travail
G 9912 Bis	Ch. GIANELLA Une estimation de l'élasticité de l'emploi peu qualifié à son coût
G 9913	Division « Redistribution et Politiques Sociales » Le modèle de microsimulation dynamique DESTINIE
G 9914	E. DUGUET Macro-commandes SAS pour l'économétrie des panels et des variables qualitatives
G 9915	R. DUHAUTOIS Evolution des flux d'emplois en France entre 1990 et 1996 : une étude empirique à partir du fichier des bénéficiaires réels normaux (BRN)
G 9916	J.Y. FOURNIER Extraction du cycle des affaires : la méthode de Baxter et King
G 9917	B. CRÉPON - R. DESPLATZ - J. MAIRESSE Estimating price cost margins, scale economies and workers' bargaining power at the firm level
G 9918	Ch. GIANELLA - Ph. LAGARDE Productivity of hours in the aggregate production function: an evaluation on a panel of French firms from the manufacturing sector
G 9919	S. AUDRIC - P. GIVORD - C. PROST Evolution de l'emploi et des coûts par qualification entre 1982 et 1996
G 2000/01	R. MAHIEU Les déterminants des dépenses de santé : une approche macroéconomique
G 2000/02	C. ALLARD-PRIGENT - H. GUILMEAU - A. QUINET The real exchange rate as the relative price of nontradables in terms of tradables: theoretical investigation and empirical study on French data
G 2000/03	J.-Y. FOURNIER L'approximation du filtre passe-bande proposée par Christiano et Fitzgerald
G 2000/04	Bilan des activités de la DESE - 1999
G 2000/05	B. CREPON - F. ROSENWALD Investissement et contraintes de financement : le poids du cycle Une estimation sur données françaises
G 2000/06	A. FLIPO Les comportements matrimoniaux de fait
G 2000/07	R. MAHIEU - B. SÉDILLOT Microsimulations of the retirement decision: a supply side approach
G 2000/08	C. AUDENIS - C. PROST Déficit conjoncturel : une prise en compte des conjonctures passées
G 2000/09	R. MAHIEU - B. SÉDILLOT Equivalent patrimonial de la rente et souscription de retraite complémentaire
G 2000/10	R. DUHAUTOIS Ralentissement de l'investissement : petites ou grandes entreprises ? industrie ou tertiaire ?
G 2000/11	G. LAROQUE - B. SALANIÉ Temps partiel féminin et incitations financières à l'emploi
G2000/12	Ch. GIANELLA Local unemployment and wages
G2000/13	B. CREPON - Th. HECKEL - Informatisation en France : une évaluation à partir de données individuelles

G2001/01	- Computerization in France: an evaluation based on individual company data F. LEQUILLER - La nouvelle économie et la mesure de la croissance du PIB - The new economy and the measurement of GDP growth	G2002/01	F. MAGNIEN - J.-L. TAVERNIER - D. THESMAR Les statistiques internationales de PIB par habitant en standard de pouvoir d'achat : une analyse des résultats	G2002/16	F. MAUREL - S. GREGOIR Les indices de compétitivité des pays : interprétation et limites	G2004/06	M. DUÉE L'impact du chômage des parents sur le devenir scolaire des enfants
G2001/02	S. AUDRIC La reprise de la croissance de l'emploi profite-t-elle aussi aux non-diplômés ?	G2002/02	Bilan des activités de la DESE - 2001	G2003/01	N. RIEDINGER - E. HAUVY Le coût de dépollution atmosphérique pour les entreprises françaises : Une estimation à partir de données individuelles	G2004/07	P. AUBERT - E. CAROLI - M. ROGER New Technologies, Workplace Organisation and the Age Structure of the Workforce: Firm-Level Evidence
G2001/03	I. BRAUN-LEMAIRE Evolution et répartition du surplus de productivité	G2002/03	B. SÉDILLOT - E. WALRAET La cessation d'activité au sein des couples : y a-t-il interdépendance des choix ?	G2003/02	P. BISCOURP et F. KRAMARZ Création d'emplois, destruction d'emplois et internationalisation des entreprises industrielles françaises : une analyse sur la période 1986-1992	G2004/08	E. DUGUET - C. LELARGE Les brevets accroissent-ils les incitations privées à innover ? Un examen microéconométrique
G2001/04	A. BEAUDU - Th. HECKEL Le canal du crédit fonctionne-t-il en Europe ? Une étude de l'hétérogénéité des comportements d'investissement à partir de données de bilan agrégées	G2002/04	G. BRILHAULT - Rétropolation des séries de FBCF et calcul du capital fixe en SEC-95 dans les comptes nationaux français - Retropolation of the investment series (GFCF) and estimation of fixed capital stocks on the ESA-95 basis for the French balance sheets	G2003/03	Bilan des activités de la DESE - 2002	G2004/09	S. RASPILLER - P. SILLARD Affiliating versus Subcontracting: the Case of Multinationals
G2001/05	C. AUDENIS - P. BISCOURP - N. FOURCADE - O. LOISEL Testing the augmented Solow growth model : An empirical reassessment using panel data	G2002/05	P. BISCOURP - B. CRÉPON - T. HECKEL - N. RIEDINGER How do firms respond to cheaper computers? Microeconomic evidence for France based on a production function approach	G2003/04	P.-O. BEFFY - J. DERUYON - N. FOURCADE - S. GREGOIR - N. LAÏB - B. MONFORT Évolutions démographiques et croissance : une projection macro-économique à l'horizon 2020	G2004/10	J. BOISSINOT - C. L'ANGEVIN - B. MONFORT Public Debt Sustainability: Some Results on the French Case
G2001/06	R. MAHIEU - B. SÉDILLOT Départ à la retraite, irréversibilité et incertitude	G2002/06	C. AUDENIS - J. DERUYON - N. FOURCADE L'impact des nouvelles technologies de l'information et de la communication sur l'économie française - un bouclage macro-économique	G2003/05	P. AUBERT La situation des salariés de plus de cinquante ans dans le secteur privé	G2004/11	S. ANANIAN - P. AUBERT Travailleurs âgés, nouvelles technologies et changements organisationnels : un réexamen à partir de l'enquête « REPONSE »
G2001/07	Bilan des activités de la DESE - 2000	G2002/07	J. BARDAJI - B. SÉDILLOT - E. WALRAET Évaluation de trois réformes du Régime Général d'assurance vieillesse à l'aide du modèle de microsimulation DESTINIE	G2003/06	P. AUBERT - B. CRÉPON Age, salaire et productivité La productivité des salariés décline-t-elle en fin de carrière ?	G2004/12	X. BONNET - H. PONCET Structures de revenus et propensions différentes à consommer - Vers une équation de consommation des ménages plus robuste en prévision pour la France
G2001/08	J. Ph. GAUDEMET Les dispositifs d'acquisition à titre facultatif d'annuités viagères de retraite	G2002/08	J.-P. BERTHIER Réflexions sur les différentes notions de volume dans les comptes nationaux : comptes aux prix d'une année fixe ou aux prix de l'année précédente, séries chaînées	G2003/07	H. BARON - P.O. BEFFY - N. FOURCADE - R. MAHIEU Le ralentissement de la productivité du travail au cours des années 1990	G2004/13	C. PICART Évaluer la rentabilité des sociétés non financières
G2001/09	B. CRÉPON - Ch. GIANELLA Fiscalité, coût d'usage du capital et demande de facteurs : une analyse sur données individuelles	G2002/09	F. HILD Les soldes d'opinion résumant-ils au mieux les réponses des entreprises aux enquêtes de conjoncture ?	G2003/08	P.-O. BEFFY - B. MONFORT Patrimoine des ménages, dynamique d'allocation et comportement de consommation	G2004/14	J. BARDAJI - B. SÉDILLOT - E. WALRAET Les retraites du secteur public : projections à l'horizon 2040 à l'aide du modèle de microsimulation DESTINIE
G2001/10	B. CRÉPON - R. DESPLATZ Evaluation des effets des dispositifs d'allègements de charges sociales sur les bas salaires	G2002/10	I. ROBERT-BOBÉE Les comportements démographiques dans le modèle de microsimulation Destinie - Une comparaison des estimations issues des enquêtes Jeunes et Carrières 1997 et Histoire Familiale 1999	G2003/09	P. BISCOURP - N. FOURCADE Peut-on mettre en évidence l'existence de rigidités à la baisse des salaires à partir de données individuelles ? Le cas de la France à la fin des années 90	G2005/01	S. BUFFETEAU - P. GODEFROY Conditions de départ en retraite selon l'âge de fin d'études : analyse prospective pour les générations 1945 à 1974
G2001/11	J.-Y. FOURNIER Comparaison des salaires des secteurs public et privé	G2002/11	J.-P. ZOYEM La dynamique des bas revenus : une analyse des entrées-sorties de pauvreté	G2003/10	M. LECLAIR - P. PETIT Présence syndicale dans les firmes : quel impact sur les inégalités salariales entre les hommes et les femmes ?	G2005/02	C. AFSA - S. BUFFETEAU L'évolution de l'activité féminine en France : une approche par pseudo-panel
G2001/12	J.-P. BERTHIER - C. JAULENT R. CONVENEVOLE - S. PISANI Une méthodologie de comparaison entre consommations intermédiaires de source fiscale et de comptabilité nationale	G2002/12	F. HILD Prévisions d'inflation pour la France	G2003/11	P.-O. BEFFY - X. BONNET - M. DARRACQ-PARIES - B. MONFORT MZE: a small macro-model for the euro area	G2005/03	P. AUBERT - P. SILLARD Délocalisations et réductions d'effectifs dans l'industrie française
G2001/13	P. BISCOURP - Ch. GIANELLA Substitution and complementarity between capital, skilled and less skilled workers: an analysis at the firm level in the French manufacturing industry	G2002/13	M. LECLAIR Réduction du temps de travail et tensions sur les facteurs de production	G2004/01	P. AUBERT - M. LECLAIR La compétitivité exprimée dans les enquêtes trimestrielles sur la situation et les perspectives dans l'industrie	G2005/04	M. LECLAIR - S. ROUX Mesure et utilisation des emplois instables dans les entreprises
G2001/14	I. ROBERT-BOBÉE Modelling demographic behaviours in the French microsimulation model Destinie: An analysis of future change in completed fertility	G2002/14	E. WALRAET - A. VINCENT - Analyse de la redistribution intragénérationnelle dans le système de retraite des salariés du privé - Une approche par microsimulation - Intragenerational distributional analysis in the french private sector pension scheme - A microsimulation approach	G2004/02	M. DUÉE - C. REBILLARD La dépendance des personnes âgées : une projection à long terme	G2005/05	C. L'ANGEVIN - S. SERRAVALLE Performances à l'exportation de la France et de l'Allemagne - Une analyse par secteur et destination géographique
G2001/15	J.-P. ZOYEM Diagnostic sur la pauvreté et calendrier de revenus : le cas du "Panel européen des ménages"	G2002/15	P. CHONE - D. LE BLANC - I. ROBERT-BOBÉE Offre de travail féminine et garde des jeunes enfants	G2004/03	S. RASPILLER - N. RIEDINGER Régulation environnementale et choix de localisation des groupes français	G2005/06	Bilan des activités de la Direction des Études et Synthèses Économiques - 2004
G2001/16	J.-Y. FOURNIER - P. GIVORD La réduction des taux d'activité aux âges extrêmes, une spécificité française ?			G2004/04	A. NABOULET - S. RASPILLER Les déterminants de la décision d'investir : une approche par les perceptions subjectives des firmes	G2005/07	S. RASPILLER La concurrence fiscale : principaux enseignements de l'analyse économique
G2001/17	C. AUDENIS - P. BISCOURP - N. RIEDINGER Existe-t-il une asymétrie dans la transmission du prix du brut aux prix des carburants ?			G2004/05	N. RAGACHE La déclaration des enfants par les couples non mariés est-elle fiscalement optimale ?	G2005/08	C. L'ANGEVIN - N. LAÏB Éducation et croissance en France et dans un panel de 21 pays de l'OCDE

G2005/10	P.-O. BEFFY - C. L'ANGEVIN Chômage et boucle prix-salaires : apport d'un modèle « qualifiés/peu qualifiés »
G2005/11	B. HEITZ A two-states Markov-switching model of inflation in France and the USA: credible target VS inflation spiral
G2005/12	O. BIAU - H. ERKEL-ROUSSE - N. FERRARI Réponses individuelles aux enquêtes de conjoncture et prévision macroéconomiques : Exemple de la prévision de la production manufacturière
G2005/13	P. AUBERT - D. BLANCHET - D. BLAU The labour market after age 50: some elements of a Franco-American comparison
G2005/14	D. BLANCHET - T. DEBRAND - P. DOURGNON - P. POLLET L'enquête SHARE : présentation et premiers résultats de l'édition française
G2005/15	M. DUÉE La modélisation des comportements démogra- phiques dans le modèle de microsimulation DESTINIE
G2005/16	H. RAOUI - S. ROUX Étude de simulation sur la participation versée aux salariés par les entreprises
G2006/01	C. BONNET - S. BUFFETEAU - P. GODEFROY Disparités de retraite de droit direct entre hommes et femmes : quelles évolutions ?
G2006/02	C. PICART Les gazelles en France
G2006/03	P. AUBERT - B. CRÉPON - P. ZAMORA Le rendement apparent de la formation continue dans les entreprises : effets sur la productivité et les salaires
G2006/04	J.-F. OUVREARD - R. RATHELOT Demographic change and unemployment: what do macroeconomic models predict?
G2006/05	D. BLANCHET - J.-F. OUVREARD Indicateurs d'engagements implicites des systèmes de retraite : chiffrages, propriétés analytiques et réactions à des chocs démographiques types
G2006/06	G. BIAU - O. BIAU - L. ROUVIERE Nonparametric Forecasting of the Manufacturing Output Growth with Firm-level Survey Data
G2006/07	C. AFSA - P. GIVORD Le rôle des conditions de travail dans les absences pour maladie
G2006/08	P. SILLARD - C. L'ANGEVIN - S. SERRAVALLE Performances comparées à l'exportation de la France et de ses principaux partenaires Une analyse structurelle sur 12 ans
G2006/09	X. BOUTIN - S. QUANTIN Une méthodologie d'évaluation comptable du coût du capital des entreprises françaises : 1984- 2002
G2006/10	C. AFSA L'estimation d'un coût implicite de la pénibilité du travail chez les travailleurs âgés
G2006/11	C. LELARGE Les entreprises (industrielles) françaises sont- elles à la frontière technologique ?
G2006/12	O. BIAU - N. FERRARI Théorie de l'opinion Faut-il pondérer les réponses individuelles ?
G2006/13	A. KOUBI - S. ROUX Une réinterprétation de la relation entre productivité et inégalités salariales dans les entreprises
G2006/14	R. RATHELOT - P. SILLARD The impact of local taxes on plants location decision
G2006/15	L. GONZALEZ - C. PICART Diversification, recentrage et poids des activités de support dans les groupes (1993-2000)
G2007/01	D. SRAER Allègements de cotisations patronales et dynamique salariale
G2007/02	V. ALBOUY - L. LEQUIEN Les rendements non monétaires de l'éducation : le cas de la santé
G2007/03	D. BLANCHET - T. DEBRAND Aspiration à la retraite, santé et satisfaction au travail : une comparaison européenne
G2007/04	M. BARLET - L. CRUSSON Quel impact des variations du prix du pétrole sur la croissance française ?
G2007/05	C. PICART Flux d'emploi et de main-d'œuvre en France : un réexamen
G2007/06	V. ALBOUY - C. TAVAN Massification et démocratisation de l'enseignement supérieur en France
G2007/07	T. LE BARBANCHON The Changing response to oil price shocks in France : a DSGE type approach
G2007/08	T. CHANEY - D. SRAER - D. THESMAR Collateral Value and Corporate Investment Evidence from the French Real Estate Market
G2007/09	J. BOISSINOT Consumption over the Life Cycle: Facts for France
G2007/10	C. AFSA Interpréter les variables de satisfaction : l'exemple de la durée du travail
G2007/11	R. RATHELOT - P. SILLARD Zones Franches Urbaines : quels effets sur l'emploi salarié et les créations d'établissements ?
G2007/12	V. ALBOUY - B. CRÉPON Aléa moral en santé : une évaluation dans le cadre du modèle causal de Rubin
G2008/01	C. PICART Les PME françaises : rentables mais peu dynamiques
G2008/02	P. BISCOURP - X. BOUTIN - T. VERGÉ The Effects of Retail Regulations on Prices Evidence from the Loi Galland
G2008/03	Y. BARBESOL - A. BRIANT Economies d'agglomération et productivité des

G2009/09	G. LALANNE - E. POULIQUEN - O. SIMON Prix du pétrole et croissance potentielle à long terme
G2009/10	D. BLANCHET - J. LE CACHEUX - V. MARCUS Adjusted net savings and other approaches to sustainability: some theoretical background
G2009/11	V. BELLAMY - G. CONSALES - M. FESSEAU - S. LE LAIDIER - É. RAYNAUD Une décomposition du compte des ménages de la comptabilité nationale par catégorie de ménage en 2003
G2009/12	J. BARDAJI - F. TALLET Detecting Economic Regimes in France: a Qualitative Markov-Switching Indicator Using Mixed Frequency Data
G2009/13	R. AEBERHARDT - D. FOUGÈRE - R. RATHELOT Discrimination à l'embauche : comment exploiter les procédures de <i>testing</i> ?
G2009/14	Y. BARBESOL - P. GIVORD - S. QUANTIN Partage de la valeur ajoutée, approche par données microéconomiques
G2009/15	I. BUONO - G. LALANNE The Effect of the Uruguay round on the Intensive and Extensive Margins of Trade
G2010/01	C. MINODIER Avantages comparés des séries des premières valeurs publiées et des séries des valeurs révisées - Un exercice de prévision en temps réel de la croissance trimestrielle du PIB en France
G2010/02	V. ALBOUY - L. DAVEZIES - T. DEBRAND Health Expenditure Models: a Comparison of Five Specifications using Panel Data
G2010/03	C. KLEIN - O. SIMON Le modèle MÉSANGE réestimé en base 2000 Tome 1 – Version avec volumes à prix constants
G2010/04	M.-É. CLERC - É. COUDIN L'IPC, miroir de l'évolution du coût de la vie en France ? Ce qu'apporte l'analyse des courbes d'Engel
G2010/05	N. CECI-RENAUD - P.-A. CHEVALIER Les seuils de 10, 20 et 50 salariés : impact sur la taille des entreprises françaises
G2010/06	R. AEBERHARDT - J. POUGET National Origin Differences in Wages and Hierarchical Positions - Evidence on French Full- Time Male Workers from a matched Employer- Employee Dataset
G2010/07	S. BLASCO - P. GIVORD Les trajectoires professionnelles en début de vie active : quel impact des contrats temporaires ?
G2010/08	P. GIVORD Méthodes économétriques pour l'évaluation de politiques publiques
G2010/09	P.-Y. CABANNES - V. LAPÈGUE - E. POULIQUEN - M. BEFFY - M. GAINI Quelle croissance de moyen terme après la crise ?
G2010/10	I. BUONO - G. LALANNE La réaction des entreprises françaises à la baisse des tarifs douaniers étrangers
G2008/04	D. BLANCHET - F. LE GALLO Les projections démographiques : principaux mécanismes et retour sur l'expérience française
G2008/05	D. BLANCHET - F. TOUTLEMONDE Évolutions démographiques et déformation du cycle de vie active : quelles relations ?
G2008/06	M. BARLET - D. BLANCHET - L. CRUSSON Internationalisation et flux d'emplois : que dit une approche comptable ?
G2008/07	C. LELARGE - D. SRAER - D. THESMAR Entrepreneurship and Credit Constraints - Evidence from a French Loan Guarantee Program
G2008/08	X. BOUTIN - L. JANIN Are Prices Really Affected by Mergers?
G2008/09	M. BARLET - A. BRIANT - L. CRUSSON Concentration géographique dans l'industrie manufacturière et dans les services en France : une approche par un indicateur en continu
G2008/10	M. BEFFY - É. COUDIN - R. RATHELOT Who is confronted to insecure labor market histories? Some evidence based on the French labor market transition
G2008/11	M. ROGER - E. WALRAET Social Security and Well-Being of the Elderly: the Case of France
G2008/12	C. AFSA Analyser les composantes du bien-être et de son évolution Une approche empirique sur données individuelles
G2008/13	M. BARLET - D. BLANCHET - T. LE BARBANCHON Microsimuler le marché du travail : un prototype
G2009/01	P.-A. PIONNIER Le partage de la valeur ajoutée en France, 1949-2007
G2009/02	Laurent CLAVEL - Christelle MINODIER A Monthly Indicator of the French Business Climate
G2009/03	H. ERKEL-ROUSSE - C. MINODIER Do Business Tendency Surveys in Industry and Services Help in Forecasting GDP Growth? A Real-Time Analysis on French Data
G2009/04	P. GIVORD - L. WILNER Les contrats temporaires : trappe ou marchepied vers l'emploi stable ?
G2009/05	G. LALANNE - P.-A. PIONNIER - O. SIMON Le partage des fruits de la croissance de 1950 à 2008 : une approche par les comptes de surplus
G2009/06	L. DAVEZIES - X. D'HAULTFOEUILLE Faut-il pondérer?... Ou l'éternelle question de l'économètre confronté à des données d'enquête
G2009/07	S. QUANTIN - S. RASPILLER - S. SERRAVALLE Commerce intragroupe, fiscalité et prix de transferts : une analyse sur données françaises
G2009/08	M. CLERC - V. MARCUS Élasticités-prix des consommations énergétiques des ménages

- G2010/11 R. RATHELOT - P. SILLARD  
L'apport des méthodes à noyaux pour mesurer la concentration géographique - Application à la concentration des immigrés en France de 1968 à 1999
- G2010/12 M. BARATON - M. BEFFY - D. FOUGÈRE  
Une évaluation de l'effet de la réforme de 2003 sur les départs en retraite - Le cas des enseignants du second degré public
- G2010/13 D. BLANCHET - S. BUFFETEAU - E. CRENNER  
S. LE MINEZ  
Le modèle de microsimulation Destinie 2 : principales caractéristiques et premiers résultats
- G2010/14 D. BLANCHET - E. CRENNER  
Le bloc retraites du modèle Destinie 2 : guide de l'utilisateur
- G2010/15 M. BARLET - L. CRUSSON - S. DUPUCH - F. PUECH  
Des services échangés aux services échangeables : une application sur données françaises
- G2010/16 M. BEFFY - T. KAMIONKA  
Public-private wage gaps: is civil-servant human capital sector-specific?
- G2010/17 P.-Y. CABANNES - H. ERKEL-ROUSSE - G. LALANNE - O. MONSO - E. POULIQUEN  
Le modèle Mésange réestimé en base 2000  
Tome 2 - Version avec volumes à prix chaînés
- G2010/18 R. AEBERHARDT - L. DAVEZIES  
Conditional Logit with one Binary Covariate: Link between the Static and Dynamic Cases
- G2011/01 T. LE BARBANCHON - B. OURLIAC - O. SIMON  
Les marchés du travail français et américain face aux chocs conjoncturels des années 1986 à 2007 : une modélisation DSGE
- G2011/02 C. MARBOT  
Une évaluation de la réduction d'impôt pour l'emploi de salariés à domicile
- G2011/03 L. DAVEZIES  
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