Les modèles multiniveaux : principes et pratiques

Pauline Givord et Marine Guillerm (DMCSI)

Séminaire de Méthodologie Statistique

12 Juin 2015

Plan

Plan

Introduction

- Première présentation générale sur les modèles multiniveaux (pourquoi, comment ?) avant deux "vraies" applications sur deux champs où ces méthodes sont classiques (éducation et santé)
- Investissement initié à la division MAEE par une demande de la Division Etudes Territoriales, sur une étude en collaboration avec la DEPP sur l'impact du quartier de résidence sur le retard scolaire
- Rédaction d'un mode d'emploi pratique de ces méthodes, à destination des chargés d'études, à partir de cet exemple
- Voir Davezies (2011) pour une présentation plus complète de la théorie

Plan

Quelles questions, quelles données ?

- Dans de nombreux cas, on est face à des données "groupées" :
  - élèves dans une classe (elles mêmes dans une école / dans une académie...)
  - patients dans un hôpital
  - habitants d'un quartier
  - salariés dans une entreprise
  - panel (plusieurs observations temporelles pour des mêmes unités)
- Remarque : on peut avoir des structures plus complexes (plusieurs niveaux : patients/médecin/hôpital ; niveaux pas parfaitement emboîtés : élèves /quartier et école)
- Cette structure peut avoir une incidence sur les estimations... et/ou est en elle-même intéressante à étudier

Problématiques

Types de questions auxquelles on peut souhaiter répondre :

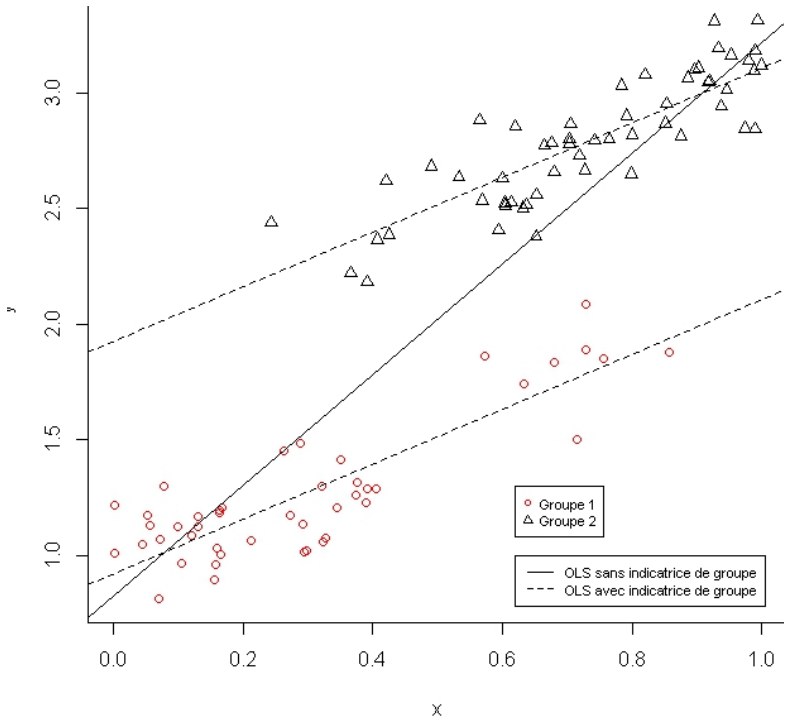
- Est-ce qu'il y a des différences de niveau de santé selon les bassins de vie ? de réussite scolaire selon les quartiers de résidence ?
- Quelle part de la variabilité des salaires s'explique par des effets "entreprises" ?
- Reste-t-il des différences de niveau de santé entre régions, une fois contrôlé des caractéristiques socio-démographiques des habitants ?
- Peut-on estimer un "effet maître" ?
- Peut-on quantifier l'effet de certaines caractéristiques observables des élèves/du quartier/de l'école sur la réussite scolaire ?

Plusieurs niveaux d'analyse

- Il peut être intéressant de conduire l'analyse aux deux niveaux (individuel et groupe)
- Dans l'exemple du lien entre quartiers et retard en sixième
  - Analyse inter groupe : quelle est la variabilité entre quartiers
  - Analyse intra groupe : quelles sont les corrélations entre l'origine sociale et la réussite scolaire, une fois tenu compte de l'influence du quartier
- Remarque : ce n'est pas toujours le cas (exemple des données de panel : analyse inter moins pertinente)
- Dans tous les cas on peut aboutir à des analyses fausses si on ne tient pas compte de la structure des données

Exemple de risque de biais 1

- Négliger l'effet groupe peut conduire à des résultats biaisés

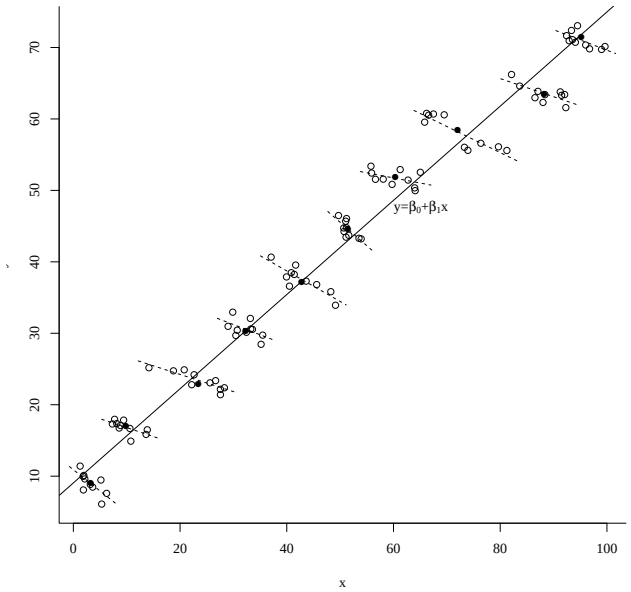


- droite trait plein : régression sur les données individuelles (on néglige la composante "groupe")
- estimation biaisée de l'effet de x sur y
- interprétation : les deux groupes sont différents en termes de composition

Exemple de risque de biais 2 : Biais d'agrégation

- Raisonner sur des données agrégées : ne renseigne pas sur les liens entre les variables au niveau individuel
- Exemple "classique" : Robinson (1950) à partir de données issues du recensement Etatsunien de 1930
  - forte corrélation positive entre taux d'illettrisme moyen par Etat et la proportion d'afro-américains dans l'Etat
  - la corrélation est divisée par 5 lorsqu'elle est estimée sur des données individuelles
  - interprétation : en 1930 la population noire était concentrée dans les Etats du Sud les plus pauvres (et donc dans l'ensemble les moins éduqués)
- dans certains cas, le biais d'agrégation peut conduire à inverser les conclusions

Biais d'agrégation : exemple (fictif)



- la corrélation entre la valeur moyenne des points par groupe (les points noirs) est positive
- alors qu'au sein de chaque groupe, la corrélation entre x et y est négative

- on parle d'"erreur écologique" quand on attribue (à tort) aux individus d'un groupe les comportements moyens du groupe auquel il appartient
- Remarque : cela ne signifie pas que l'analyse au niveau du groupe n'est pas intéressante et informative, mais qu'elle **ne doit pas être utilisée pour inférer une relation causale individuelle**

Exemple de risque de biais 3 : Sur-estimation de la précision

Aspect plus "technique" mais important :

- Les estimateurs classiques font l'hypothèse que les observations sont indépendantes entre elles
- Ce n'est pas le cas si il existe un terme inobservé commun à toutes les unités du groupe,
- La précision sera **sur-estimée**
- On risque de conclure qu'une variable a un impact "significativement non nul" alors que ce n'est pas le cas

Plan

La décomposition des erreurs

- Les modèles multiniveaux sont une manière de tenir compte de la structure groupée des données
- En pratique, ils consistent à décomposer l'erreur du modèle en deux termes
- Formellement dans le plus simple (sans variables explicatives) :

$$y_{ij} = \beta_0 + \alpha_j + \epsilon_{ij}$$

- où :
  1. α_j correspond à un terme groupe (commun à tous les individus du groupe)
  2. ϵ_{ij} représente un écart strictement individuel à la moyenne

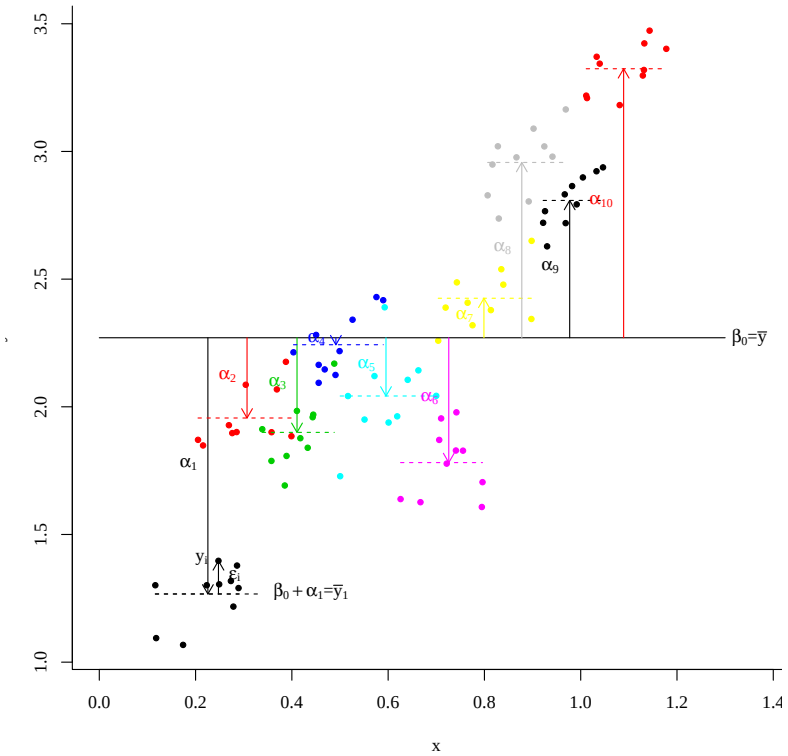
La décomposition des erreurs

- Les modèles multiniveaux sont une manière de tenir compte de la structure groupée des données
- En pratique, ils consistent à décomposer l'erreur du modèle en deux termes
- Formellement dans le plus simple (sans variables explicatives) :

$$y_{ij} = \beta_0 + \alpha_j + x_{ij}\beta + \epsilon_{ij}$$

- où :
  1. α_j correspond à un terme groupe (commun à tous les individus du groupe)
  2. ϵ_{ij} représente un écart strictement individuel à la moyenne
- Remarque : évidemment, en général on rajoute des variables explicatives !

La décomposition des erreurs : illustration



- β_0 est la moyenne de y sur l'ensemble de la population.
- α_j **l'effet du groupe : écart de la moyenne du groupe** j **($\beta_0 + \alpha_j$) par rapport à la moyenne totale** β_0**.**
- ϵ_{ij} est le terme résiduel individuel : écart de la valeur observée pour i par rapport à la moyenne de son groupe $\beta_0 + \alpha_j$

Comment estimer ces modèles ? $\rightarrow$ on distingue classiquement les modèles à effets fixes / à effets aléatoires

Le modèle à effets fixes

- Le plus souple *a priori* : on fait très peu d'hypothèses sur les effets groupes α_j
- En termes d'estimation :
  - soit on les estime directement (une indicatrice par groupe... donc beaucoup d'estimateurs)
  - soit on s'en "débarrasse" par exemple en estimant les écarts à la moyenne :

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

- Le **plus sûr pour estimer correctement (sans biais) les liens entre les variables individuelles et la variable d'intérêt**

Le modèle à effets fixes - limites

- Plus difficile d'étudier les effets de contexte (distribution des α_j), simplement considérés ici comme des variables de nuisance $\rightarrow$ donc difficile de caractériser les différences entre les groupes (analyse "inter")
- On ne peut pas répondre aux questions :
  - Quelle part de la variance des salaires peut s'expliquer par un "effet entreprise" ?
  - Quelles caractéristiques observables du quartier sont corrélées avec la réussite scolaire ?
- Ces limites expliquent le succès d'une spécification plus explicite des effets de contextes : les modèles à effets aléatoires

Le modèle à effets aléatoires

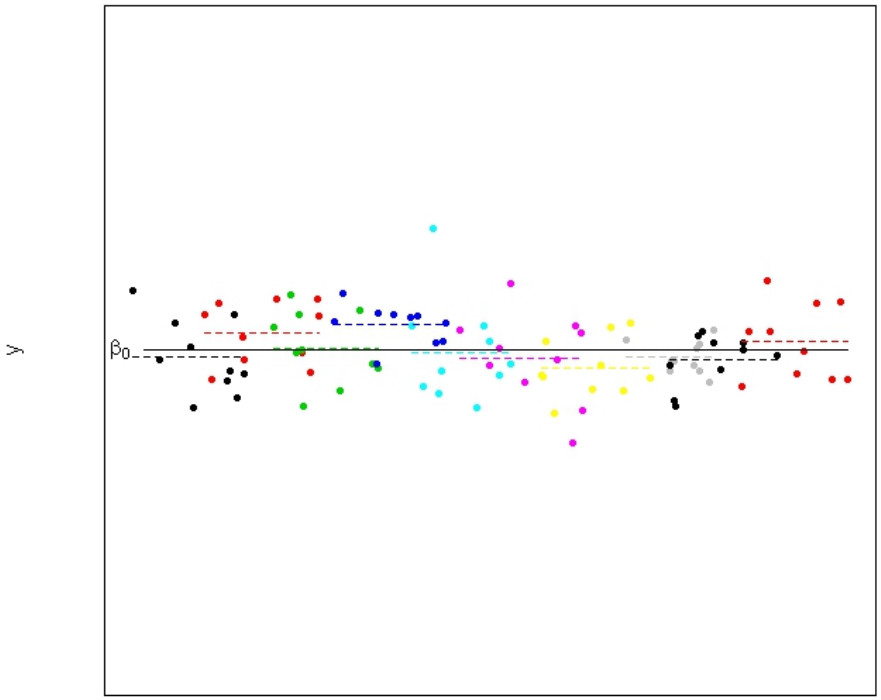
Modèle vide (sans variables explicatives) :

$$y_{ij} = \beta_0 + \alpha_j + \varepsilon_{ij} \qquad \alpha_j \sim N(0, \sigma^2_\alpha)$$
$$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$$

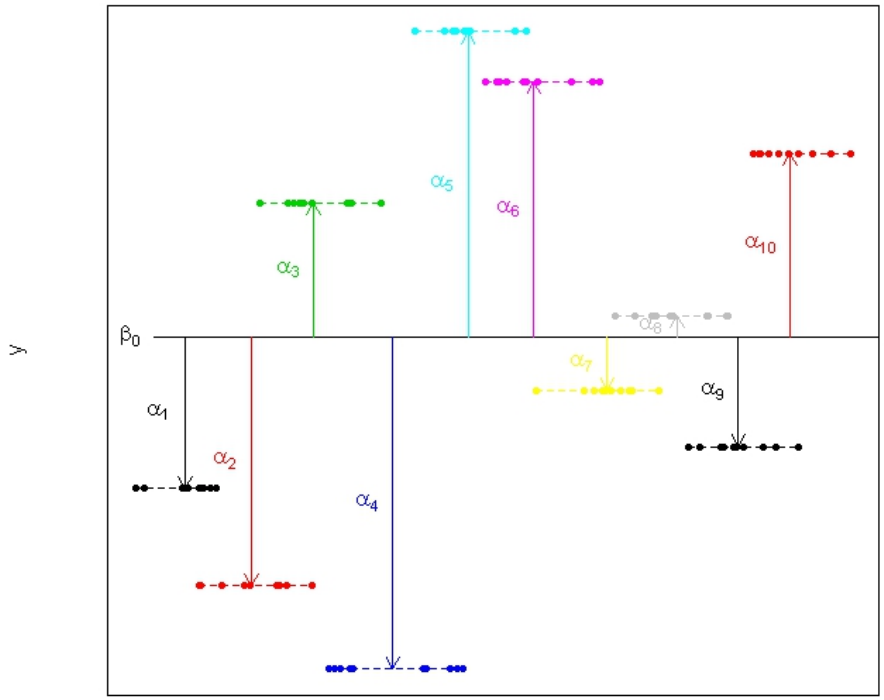
- Une modélisation plus parcimonieuse
- La variance totale de y se décompose :
  - σ^2_α, la variance inter-groupe : variabilité de y d'un groupe à un autre.
  - σ^2_ε, variance intra-groupe : variabilité de y entre individus à l'intérieur d'un même groupe.
  - Coefficient de corrélation intra-classe $\rho = \sigma^2_\alpha / (\sigma^2_\alpha + \sigma^2_\varepsilon)$
    - Correspond à la corrélation entre deux individus d'un même groupe : $\rho = corr(y_{ij}, y_{i'j})$.
    - Part de la variance de y "expliquée" par le groupe.

Illustrations : différentes valeurs de ρ

ρ proche de zéro



$\rho = 1$



Le retard scolaire

- Étude du retard scolaire à l'entrée en 6e, spécificité : variable binaire.
- Prise en compte de l'effet quartier d'habitation de l'élève.
- Le modèle logit à effets aléatoires

$$Retard_{ij} = \mathbb{1}(\beta_0 + \alpha_j + \varepsilon_{ij} \geq 0) \quad \alpha_j \sim N(0, \sigma_\alpha^2)$$
$$\varepsilon_{ij} \sim logit$$

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta_0 + \alpha_j \qquad p_{ij} = \text{probabilité d'être en retard}$$

- Résultats :
  $\sigma_\alpha^2 = 0,436$ et $\sigma_\varepsilon^2 = \pi^2/3 = 3,28$ (fixé dans un modèle logit pour rendre le modèle identifiable)
  $\rightarrow \rho = 11,7\%$.

Modèles à effets aléatoires avec des covariables

On souhaite en général inclure des variables pour :

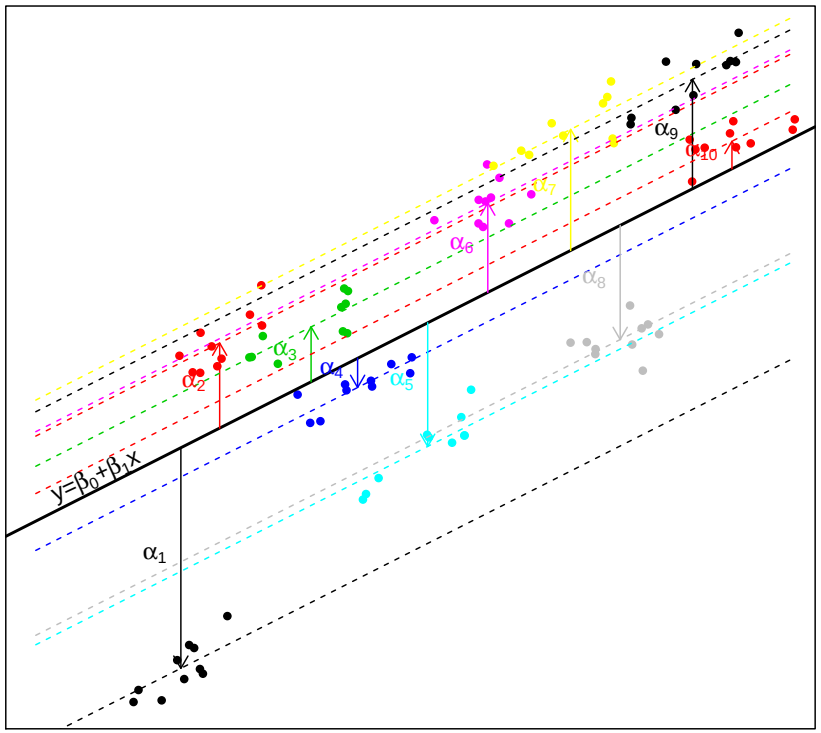
- Estimer l'effet d'une caractéristique individuelle ou du groupe sur y.
- Contrôler les estimations : les groupes sont rarement formés aléatoirement, leurs différences sont aussi le résultat d'effets de composition.
- Le modèle

$$y_{ij} = \beta_0 + x_{ij}\beta_1 + \alpha_j + \varepsilon_{ij} \qquad \alpha_j \sim N(0, \sigma^2_\alpha)$$
$$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$$

  - x_{ij} vecteur de variables de niveau individu et/ou groupe.
  - Les paramètres du modèle : β_0, β_1, σ^2_α et σ^2_ε.

La forme du modèle

- Relation linéaire générale entre y et x : $\beta_0 + \beta_1 x$.
- Une constante $\beta_0 + \alpha_j$ qui varie d'un groupe à un autre. $\rightarrow$ Modèle à "constante aléatoire".
  $\rightarrow$ Chaque groupe a sa propre droite, parallèle à la relation générale.
  $\rightarrow$ α_j est "l'effet" sur y d'être dans le groupe j, contrôlé de x.
  $\rightarrow$ L'effet de la variable x sur y est le même d'un groupe à un autre.



Interpréter les paramètres du modèle

- β_0 et β_1 s'interprètent de la même manière que dans un modèle classique. La modélisation multiniveaux évite de sur-estimer la précision de leur estimation.
- σ^2_α variance résiduelle au niveau groupe, σ^2_ε variance résiduelle au niveau individuel.
  $\rightarrow$ On peut tester la nullité de σ^2_α.
- Retard scolaire avec effet du quartier d'habitation de l'élève

Variable	Modalité	Logit simple	logit à effets aléatoires	logit à effets fixes
Constante		$-1,9406^{***}$ (0,0156)	$-2,0096^{***}$ (0,01756)	
Catégorie sociale des parents	Très favorisée	$-1,4463^{***}$ (0,0323)	$-1,4416^{***}$ (0,03269)	$-1,1619^{***}$ (0,0389)
	Favorisée	$-0,3988^{***}$ (0,0321)	$-0,3992^{***}$ (0,03257)	$-0,3368^{***}$ (0,0376)
	Moyen	réf.	réf.	réf.
	Défavorisée	$0,6085^{***}$ (0,0201)	$0,6112^{***}$ (0,02067)	$0,5185^{***}$ (0,0246)
Ecole primaire	Privée	$-0,174^{***}$ (0,0367)	$-0,1664^{***}$ (0,03729)	$-0,1065^{**}$ (0,0436)
	Publique	réf.	réf.	réf.
σ^2_α		-	0,1550 (0,01611)	
ρ			4,5%	

Estimer les α_j

- Pour certaines problématiques, les effets groupes α_j peuvent avoir un intérêt.

  Par exemple : estimer les effets école ou faire des prédictions.
- Calcul des effets groupe :

$$\hat{\alpha}_j = c_j \times (\bar{y}_j - \hat{\beta}_0 - \bar{x}_j \hat{\beta}_1) \qquad \text{avec } c_j = \frac{\hat{\sigma}^2_\alpha}{\hat{\sigma}^2_\alpha + \hat{\sigma}^2_\varepsilon / n_j}$$

  $0 \leq c_j \leq 1$: facteur de contraction.
  $\rightarrow$ La contraction rapproche la droite du groupe j de la droite générale.
- La contraction est d'autant plus importante que le groupe est de petite taille.
  $\rightarrow$ Quand les effectifs sont faibles, les estimations risquent quand même de ne pas être très fiables.

Les hypothèses

- Enjeu : Avoir une interprétation causale des paramètres estimés.
- H1 : Exogénéité des covariables $cov(\varepsilon_{ij}, x_{ij}^k) = 0 \quad k = 1, \ldots, p$
  $\rightarrow$ Les caractéristiques individuelles non observées ayant un effet sur la variable d'intérêt ne sont pas corrélées aux covariables.
- H2 : Hypothèse sur les effets aléatoires
  $cov(\alpha_j, x_{ij}^k) = 0 \quad k = 1, \ldots, p.$
  $\rightarrow$ Les caractéristiques inobservées du groupe ayant un effet sur la variable d'intérêt ne sont pas corrélées aux variables incluses dans le modèle.
- H2 est spécifique à la modélisation à effets aléatoires. Dans le modèle à effets fixes, pas d'hypothèse sur les effets groupe...
- Mais H1 doit être vérifiée qu'on soit dans le cadre effets aléatoires ou effets fixes.

Les hypothèses

- Les hypothèses H1 et H2 sont des hypothèses fortes.
- Les individus sont rarement répartis aléatoirement entre les groupes.
- Exemple : estimation de l'effet de la taille de la classe sur les résultats scolaires :
  - De nombreuses caractéristiques de l'élève et de sa classe ont un effet sur les résultats scolaires.
  - Certaines sont corrélées à la taille de la classe, par exemple si les classes les plus petites sont formées pour les élèves les plus en difficultés scolaires.
  - Inclure le niveau scolaire des élèves en début d'année permet de mieux contrôler les estimations.

Le test d'Hausman

- On distingue les variables qui caractérisent l'individu, notées x_{ij} ; de celles qui caractérisent son groupe, notées x_j

$$y_{ij} = \beta_0 + x_{ij}\beta_1^1 + x_j\beta_1^2 + \alpha_j + \varepsilon_{ij}$$

- Principe :
  - On estime β_1^1 par un modèle à effets aléatoires $\rightarrow \hat{\beta}_{1,RE}^1$; et par un modèle à effets fixes $\rightarrow \hat{\beta}_{1,FE}^1$
  - $\hat{\beta}_{1,RE}^1$ est sans biais seulement si H2 est vérifiée
  - $\hat{\beta}_{1,FE}^1$ est convergent que H2 soit vérifiée ou pas
  - Tester l'égalité des deux estimateurs permet de vérifier l'hypothèse H2.
- Remarque : Ne permet de tester l'hypothèse que pour les coefficients des variables de niveau individuel

Le modèle de Mundlak

$$y_{ij} = \beta_0 + x_{ij}\beta_1^1 + x_j\beta_1^2 + \alpha_j + \varepsilon_{ij}$$

- Si on pense que α_j est corrélée à x_{ij} :

$$\alpha_j = \bar{x}_{.j}\delta + u_j \quad \text{avec } \bar{x}_{.j} = 1/n_j \sum_{i \in j} x_{ij}$$

  $\rightarrow \bar{x}_{.j}$ apparaît comme une variable omise.
- Modèle de Mundlak : ajouter les moyennes par groupe des variables individuelles

$$y_{ij} = \beta_0 + x_{ij}\beta_1^1 + x_j\beta_1^2 + \underbrace{\bar{x}_{.j}\delta + u_j}_{\alpha_j} + \varepsilon_{ij}$$

- $\delta \neq 0$ suggère que H2 n'était pas vérifiée.
- Remarque : le modèle de Mundlak résout le problème de corrélation pour les variables individuelles mais pas pour les variables de niveau groupe dont l'estimation du coefficient peut donc être biaisée.

Plan

Conclusion

- Deux modélisations qui répondent à l'analyse de données groupées : modèles à effets fixes et modèles à effets aléatoires
- Ne permettent pas de répondre aux mêmes questions, ne reposent pas sur les mêmes hypothèses
- Les modèles à effets aléatoires offrent une analyse plus riche...
- Mais attention à l'interprétation, une réflexion au cas par cas est nécessaire en fonction des données et des questions auxquelles on souhaite répondre