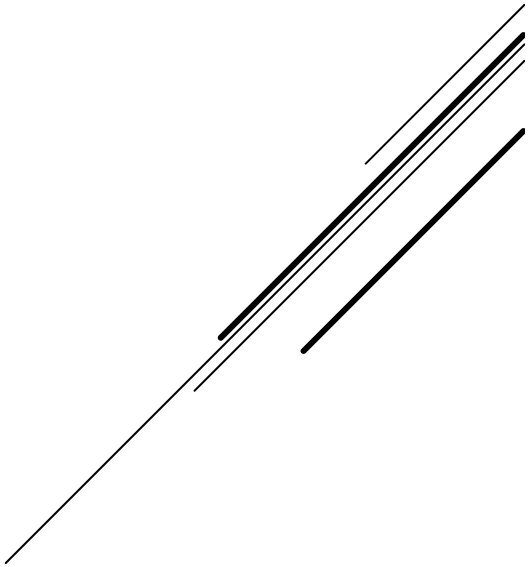
Alain Marchand

École de Relations industrielles

Université de Montréal

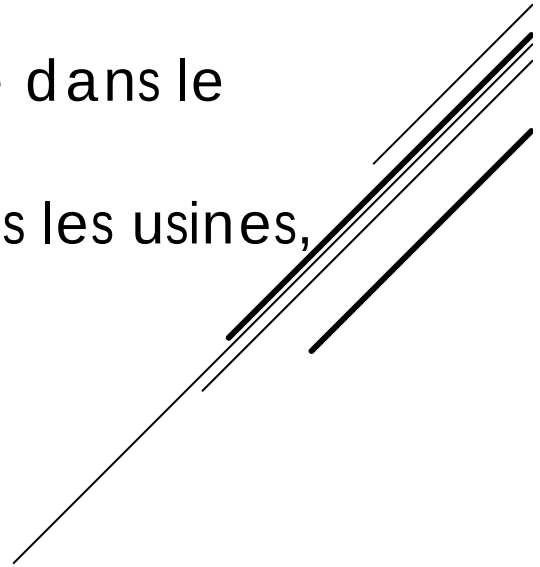
L'ANALYSE MULTI-NIVEAUX AVEC MLWIN



INTRODUCTION

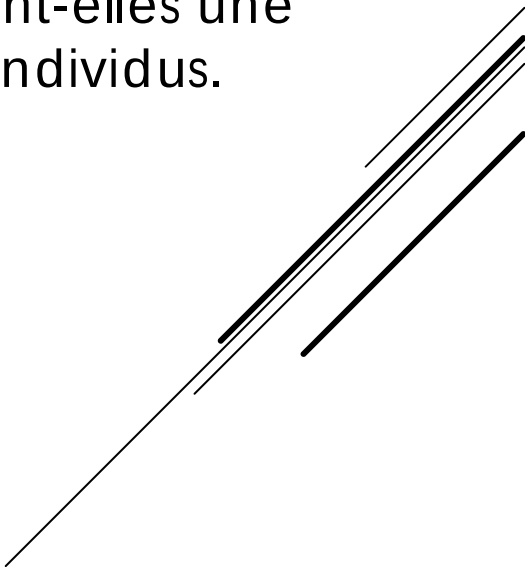
La recherche en sociologie implique souvent la présence de données hiérarchiques:

- Travailleurs nichés dans des usines
- Étudiants nichés dans des écoles
- Patients nichés dans des hôpitaux
- Longitudinale: même mesure répétée dans le temps
- Plus complexe: travailleurs nichés dans les usines, usines nichées dans industries



- Dans ce type de recherche il y a des variables qui sont définies au niveau des individus (travailleurs) et au niveau des groupes (usines)

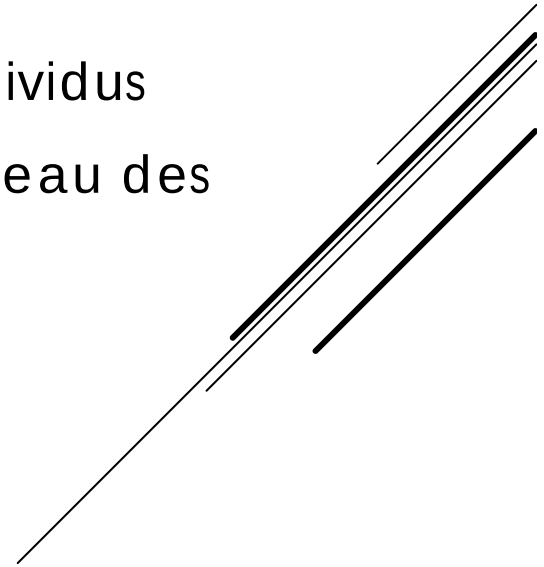
- La question de recherche se pose souvent ainsi: Comment les variables individuelles et de groupes influencent-elles une variable dépendante mesurée au niveau des individus.



LES APPROCHES ANTÉRIEURES

Les modèles linéaires traditionnels sont basés sur les moindres carrés ordinaires avec deux approches:

- Imputation des variables de groupes aux individus
- Agrégation des variables individuelles au niveau des groupes



CES DEUX APPROCHES SONT INSATISFAISANTES

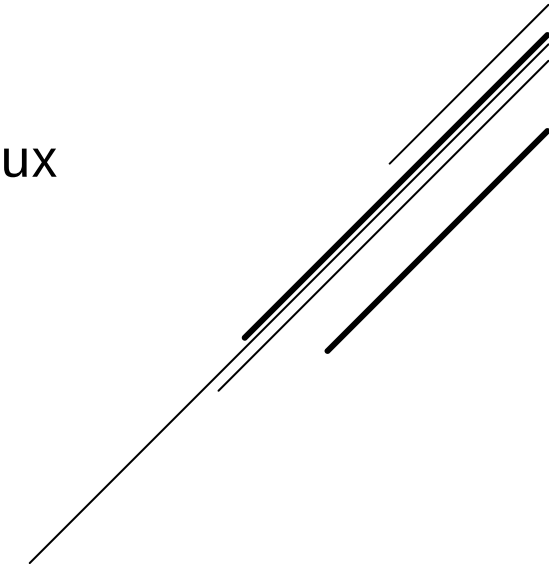
Approche imputation

- Violation du postulat d'indépendance.
- Covariation entre les observations partageant un même contexte.
- MCO (moindres carrés ordinaires) produisent estimés instables; erreurs types biaisées (i.e. trop grandes).

APPROCHE AGRÉGATION

- Perte de l'information concernant la variation intra groupe (80-92%)
- Produit généralement des corrélations plus fortes: Effet Robinson
- On ne peut plus conclure sur les individus sans commettre l'erreur écologique. Les variables individuelles ne sont plus de niveau individuel

Solution: les modèles multi-niveaux



BREF HISTORIQUE

- La formalisation des modèles multiniveaux est connue depuis au moins les années 60. (Elston et Grizzle, 1962: Biométrie).
- Ils apparaissent dans plusieurs disciplines:

Sociologie: modèles linéaires multi-niveaux ou hiérarchiques

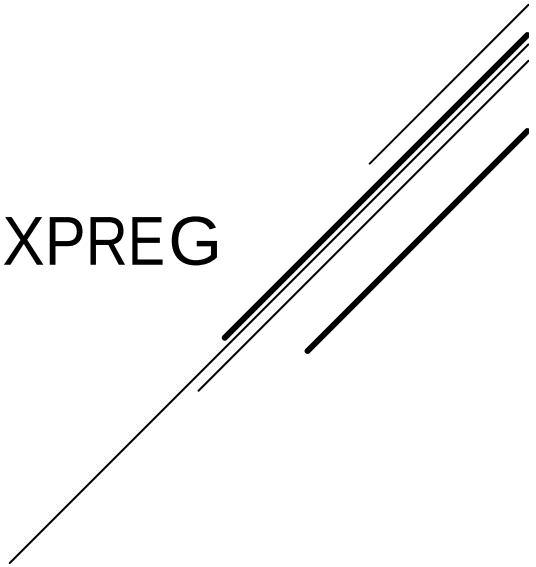
Biométrie: modèles à effets mixtes; modèles à effets aléatoires

Économétrie: modèles de régression à coefficients aléatoires

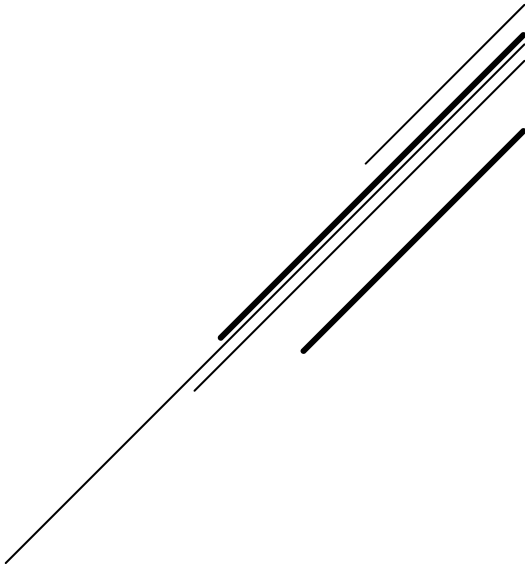
- Ce n'est que depuis une vingtaine d'années qu'il est techniquement possible – et plus facile -- d'estimer ces modèles

LES LOGICIELS DISPONIBLES

- HLM
- MLwiN (mon préféré)
- R
- Lisrel
- SPSS
- Stata xtreg
- BMDP (BMDP3V)
- SAS (MIXED)
- VARCL
- MIXREG, MIXOR, MIXNO, MIXPREG (gratuit)
- EGRET



MODÈLE POUR UNE VARIABLE DÉPENDANTE NORMALEMENT DISTRIBUÉE



Modèle à deux niveaux avec j=1...K niveau 2 et i=1...nj niveau 1

Niveau 1 : $Y_{ij} = \beta_{0j} + \epsilon_{ij}$

β_{0j} = moyenne dans le groupe j

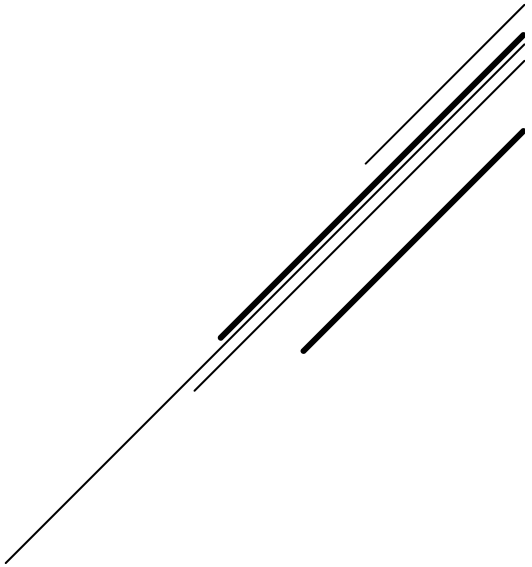
ϵ_{ij} = résidu pour individu i du groupe j

Niveau 2 : $\beta_{0j} = \gamma_{00} + \mu_{0j}$

γ_{00} = moyenne pour l'ensemble des groupes

μ_{0j} = résidu pour le groupe j

Alors: $Y_{ij} = \gamma_{00} + \mu_{0j} + \epsilon_{ij}$



Postulat:

μ_{0j} et ϵ_{ij} sont non-corrélées et distribuées normalement avec comme moyenne 0 et des variances σ^2_{μ}, σ^2_{ϵ} estimées par les données.

Corrélation intra-classe

$$\rho_i = \sigma^2_{\mu} / (\sigma^2_{\mu} + \sigma^2_{\epsilon})$$

ρ_i = proportion (%) de la variance de la variable dépendante Y_{ij} qui est entre les groupes.

Avec variables indépendantes

Niveau 1 : X_{pij} (p=1...,P) Niveau 2 : Z_{qj} (q=1...,Q)

Niveau 1 : $Y_{ij} = \beta_{0j} + \beta_{pj} X_{pij} + \epsilon_{ij}$

Niveau 2 : $\beta_{0j} = \gamma_{00} + \gamma_{0q} Z_{qj} + \mu_{0j}$

Si on pose que β_{0j} varie entre les groupes et que les pentes de niveau 1 β_{pj} sont constantes, soit $\beta_{pj} = \gamma_{p0}$

Alors : $Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + (\mu_{0j} + \epsilon_{ij})$

X_{pij} expliquent variation intra et inter groupes

Z_{qj} expliquent variation inter groupes

MODÈLE À COEFFICIENTS ALÉATOIRES

Pentes aléatoires au niveau 2

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + (\mu_{pj} X_{pij} + \mu_{0j} + \epsilon_{ij})$$

Pentes aléatoires niveau 2 avec interaction

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + \gamma_{0q} Z_{qj} X_{pij} + (\mu_{pj} X_{pij} + \mu_{0j} + \epsilon_{ij})$$

Pentes aléatoires niveau 1 et 2

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + \gamma_{0q} Z_{qj} X_{pij} + (\mu_{pj} X_{pij} + \epsilon_{ij} X_{pij} + \mu_{0j} + \epsilon_{ij})$$

Pentes aléatoires niveau 1 seulement

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + (\epsilon_{ij} X_{pij} + \mu_{0j} + \epsilon_{ij})$$

Exemple

Simard et Marchand (1997). Ergonomics, 40, 2: 172-188.

Effets des facteurs micro et macro organisationnels sur le niveau de prudence des équipes de travail.

Modèle de composition de la variance

$$\text{prudence}_{ij} = \gamma_{00} + \gamma_{p0}\,\text{micro}_{pij} + \gamma_{0q}\,\text{macro}_{qj} + (\mu_{0j} + \epsilon_{ij})$$

n_2= 97 établissements manufacturiers
n_1=1061 équipes de travail

IGLS

	Modèle 1		Modèle 2		Modèle 3		Modèle 4		Modèle 5	
	γ	T	γ	T	γ	T	γ	T	γ	T
Partie fixe Niveau 1										
Constante	74,71	130,75	75,00	164,76	74,97	135,15	74,95	164,87	75,04	16423
non-routine			-0,25	0,68			-0,27	0,74		
routine			0,01	0,02			0,01	0,02		
risques			-0,41	1,09			-0,36	0,94		
coop			4,65	12,74			4,71	12,87	4,69	12,89
cohesion			0,84	2,30			0,85	2,31	0,85	2,34
suparprev			2,91	7,54			2,92	7,36	3,03	7,93
supexp			0,65	1,77			0,67	1,83		
Niveau 2										
leadership					0,75	1,46	0,72	1,65		
organsst					-0,23	0,39	-0,62	1,24		
segsecond					-0,96	1,70	0,08	0,16		
instabilite					-0,35	0,71	0,21	0,51		
Partie aléatoire										
σ^2_{μ}	11,35	p=0,00	4,81	p=0,01	8,30	p=0,00	4,27	p=0,02	5,04	p=0,01
σ^2_{ϵ}	160,8	p=0,00	128,8	p=0,00	161,4	p=0,00	128,6	p=0,00	129,3	p=0,00
Statistiques										
Deviance	8448,95		8194,74		8443,11		8190,53		8199,97	
Chi-carré	-		254,21		5,84		258,42		248,98	
DL	-		7,00		4,00		11,00		3,00	
P	-		0,00		0,21		0,00		0	

Modèle5-modèle4: $\chi^2 = 9.44$ dl=8 p=.31

COMMENT CALCULER LES R^2

Modèle de composition de la variance

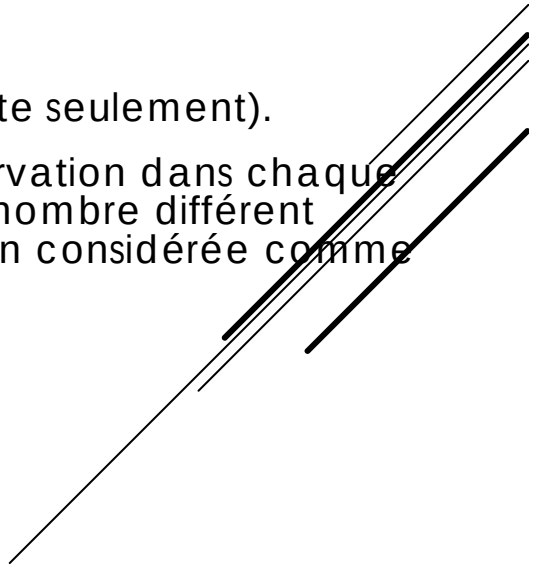
$$R^2_2 = 1 - \frac{(\sigma^2_{\epsilon 1}/n) + \sigma^2_{\mu 1}}{(\sigma^2_{\epsilon 0}/n) + \sigma^2_{\mu 0}}$$

$$R^2_1 = 1 - \frac{\sigma^2_{\epsilon 1} + \sigma^2_{\mu 1}}{\sigma^2_{\epsilon 0} + \sigma^2_{\mu 0}}$$

$\sigma^2_{\epsilon 1}$ et $\sigma^2_{\mu 1}$ = Modèle avec les variables indépendantes.

$\sigma^2_{\epsilon 0}$ et $\sigma^2_{\mu 0}$ = Modèle sans les variables indépendantes (constante seulement).

n = taille des groupes lorsque qu'il y a le même nombre d'observation dans chaque groupe (devis balancé). Lorsque le devis est non-balancé (nombre différent d'observations par groupe), on peut prendre une valeur de n considérée comme représentative ou encore la moyenne harmonique.



	R^2_2	R^2_1
modèle 2	0,36	0,22
modèle 3	0,11	0,01
modèle 4	0,38	0,23
modèle 5	0,35	0,22

R^2 POUR MODÈLES SIMARD ET MARCHAND (1997)

Comparaison de trois modèles de régression

Marchand (sous presse)

	Modèle imputation		Modèle agrégation		Modèle multi-niveaux	
Variables explicatives	**Estimé**	**T**	**Estimé**	**T**	**Estimé**	**T**
Niveau 1						
Constante	74,93	210,77	75,02	103,92	74,95	164,87
non-routine	-0,40	1,11	-1,81	1,66	-0,27	0,74
routine	-0,18	0,49	0,35	0,33	0,01	0,02
risques	-0,33	0,89	-0,79	0,65	-0,36	0,94
coop	4,74	12,91	7,58	6,20	4,71	12,87
cohesion	0,89	2,42	1,52	1,18	0,85	2,31
suparprev	2,89	7,39	2,19	2,08	2,92	7,36
supexp	0,66	1,81	1,72	1,52	0,67	1,83
Niveau 2						
leadership	0,89	2,31	0,76	1,19	0,72	1,65
organsst	-0,34	0,82	-1,05	1,53	-0,62	1,24
segsecond	0,13	0,32	0,31	0,46	0,08	0,16
instabilite	0,12	0,34	0,49	0,91	0,21	0,51
Variance résiduelle	134,09		33,78		5,04 (2) 129,30 (1)	

NOTE SUR LA TAILLE DE L'ÉCHANTILLON

- La taille de l'échantillon permettant d'évaluer le nombre de variables indépendantes pouvant être estimées simultanément est le **nombre d'observations pour le niveau supérieur** de la hiérarchie des données.
- Si les données sont hiérarchiques et n du niveau supérieur est < 30, il faut être très prudent.
- On retiendra le ratio minimal de **5 cas par variable**, idéalement entre 10 et 20.

CONCLUSION

- Les modèles multi-niveaux représentent un pas énorme car ils sont statistiquement exacts et ne perdent pas d ’information.
- On se retrouve beaucoup mieux équipé pour vérifier des hypothèses et des théories traitant entre autres de l'influence de variables de contexte, d'organisation ou d'environnement sur le comportement des individus ou des groupes

Autres applications possibles de l 'analyse multi-niveaux

- Dépendante nominale, multinominale et ordonnée
- Séries chronologiques
- Composante principale
- Factorielle exploratoire et confirmatoire
- Cheminent de la causalité
- Régression de survie
- Meta analyse
- Longitudinale pour VD dichotomique
- Multivarié VD dichotomiques
- Multivarié mixte: multinomiale et ordonnée avec une ou plusieurs variables continues
- Modèles Poisson, log-log, probit

Ressources sur l ’internet

Liste électronique de discussions:

MULTILEVEL@JISCMAIL.AC.UK

Corps du message: join multilevel prénom nom.

Site web de Joop Hox: http://joophox.net/

Site Web de Tom Snijders: https://www.stats.ox.ac.uk/~snijders/