MATH 60604
Modélisation statistique
§ 6a - Inclusion d'effets de groupe dans la moyenne

Léo Belzile

HEC Montréal
Département de sciences de la décision

Inclusion d'effets de groupe

- Jusqu'à maintenant, nous avons seulement utilisé la structure de groupe dans la modélisation de la corrélation intra-groupe.
- On pourrait également inclure un effet de groupe dans la moyenne, ce qui revient à fixer une ordonnée à l'origine différente pour chaque groupe.
- Pour ce faire, on ajoute une variable catégorielle g dans le modèle linéaire, ce qui se traduit par l'ajout de $m-1$ indicatrices $\mathbf{1}_{\mathrm{g}=i}$ pour $i=1,\ldots,m-1$ s'il y a m groupes.

Équation avec effets de groupe

- Si on inclut dans le modèle les indicatrices pour les niveaux de g, alors

$$Y_{ij} = \beta_0 + \sum_{i=1}^{m-1} \beta_i \mathbf{1}_{\mathrm{g}=i} + \varepsilon_{ij},$$

  - l'ordonnée à l'origine de la catégorie de référence (groupe m) est β_0,
  - l'effet de groupe pour $\mathrm{g} = i$ est β_i $(i = 1, \ldots, m-1)$, et
  - « l'ordonnée à l'origine spécifique au groupe i » est $\beta_0 + \beta_i$.

Modèle linéaire avec effet de groupe - données `vengeance`

On considère une régression linéaire ordinaire pour `vengeance` avec un effet de groupe pour illustrer quelques subtilités.

- On désire inclure le fait que le désir de vengeance varie par individu.
- Nous avons seulement cinq observations par personne pour estimer l'effet groupe.
- Notre modèle ne prend pas en compte la corrélation intra-individu pour le moment.

Modèle avec effet groupe par individu

Code SAS pour ajuster un modèle linéaire avec REML

```
proc mixed data=vengeance method=reml;
class id;
model vengeance = id sexe age vc wom t / solution;
run;
```

En plus de la variable catégorielle `id`, le modèle comprend les même variables explicatives qu'avant. Chaque personne a sa propre « ordonnée à l'origine », avec `id=80` comme catégorie de référence.

Estimés des paramètres de la moyenne

Solution pour effets fixes

Effet	id	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|
Intercept		6.7425	0.2290	319	29.44	<.0001
id	1	-1.6400	0.3152	319	-5.20	<.0001
id	2	-3.8400	0.3152	319	-12.18	<.0001
id	3	-1.3200	0.3152	319	-4.19	<.0001
id	4	0.2000	0.3152	319	0.63	0.5262
id	5	-2.6000	0.3152	319	-8.25	<.0001
		⋮				
id	79	-0.6000	0.3152	319	-1.90	0.0578
id	80	0	.	.	.	.
sexe		0	.	.	.	.
age		0	.	.	.	.
vc		0	.	.	.	.
wom		0	.	.	.	.
t		-0.5675	0.01762	319	-32.21	<.0001

Tableau des tests-F

Tests des effets fixes de type 3				
Effet	DDL num.	DDL den.	Valeur F	Pr > F
id	75	319	3.77	<.0001
sexe	0	.	.	.
age	0	.	.	.
vc	0	.	.	.
wom	0	.	.	.
t	1	319	1037.49	<.0001

Il n'y **aucun** paramètre estimé ou test d'hypothèse pour les variables `sexe`, `age`, `vc` or `wom`, mais ces derniers sont disponibles pour la variable `t`. Parce que certaines covariables ne varient pas dans le temps, leur effet n'est pas identifiable (collinéarité parfaite). En retirant `id` du modèle, leur effet devient estimable (on obtient 75 `ddl` plutôt que 79 dans le tableau de tests-F).

Collinéarité

- En fait, une fois qu'on incorpore un effet individuel pour chaque personne, il n'est pas possible d'incorporer en même temps une variable explicative qui est fixe dans le temps pour une personne.
- Ici, les variables `sexe`, `age`, `vc` et `wom` sont fixes dans le temps pour chaque personne (`vc` et `wom` ont été mesurées uniquement dans le premier questionnaire).
- Ces variables sont donc déjà implicitement incorporées dans l'effet individuel. Techniquement, il y a ici colinéarité parfaite entre une variable fixe dans le temps et la variable catégorielle `id`.
- On peut donc prédire parfaitement la valeur de `sexe` uniquement à partir de `id` (idem pour les autres covariables fixes).
- Ainsi, on ne peut pas avoir simultanément un effet fixe pour chaque personne et incorporer des variables explicatives qui sont fixes pour les personnes.

Défis découlant de l'inclusion d'un effet groupe

- La variable groupe est catégorielle: s'il y a peu d'observations par groupe, on ne peut estimer de façon fiable les estimés des coefficients de l'effet groupe.
- Si le nombre de groupes m est grand par rapport à la taille totale de l'échantillon, il peut aussi y avoir trop de paramètres dans le modèle.
- On ne peut pas estimer l'effet de variables qui sont fixées au sein des groupes si un effet groupe est inclus.

Modèle avec effet de groupe et erreurs corrélées

Le modèle qu'on ajuste inclut seulement `id` et la vague du questionnaire `t` comme variables explicatives pour la moyenne, mais on inclut une structure autorégressive d'ordre un pour la corrélation intra-individu des erreurs ε.

Code SAS pour un modèle avec effet groupe et corrélation AR(1)

```
proc mixed data=vengeance method=reml;
class id tcat;
model vengeance = id t / solution;
repeated tcat / subject=id type=ar(1);
run;
```

Le paramètre de corrélation du modèle AR(1) est significativement non-nul (statistique du rapport de vraisemblance de $21{,}68$ de loi nulle χ^2_1, valeur-p négligeable). L'estimé de `t` est $-0{,}568$, presque le même que dans le modèle incluant `sexe`, age, `vc` et `wom` avec une structure AR(1) du chapitre précédent.

Remarque sur la comparaison de modèles

- Il faut faire attention à ne pas comparer les critères d'information du modèle avec `id` de ceux du modèle avec `sexe`, `age`, `vc` et `wom`, car nous avons utilisé la méthode d'estimation REML (option par défaut).
- Nous avons mentionné à la section précédente que le AIC et le BIC obtenus de la méthode REML, ne sont **pas comparables** si les variables explicatives de la moyenne (effets fixes) des modèles ne sont pas les mêmes.
- Si on veux comparer ces deux modèles, il faut plutôt utiliser la méthode d'estimation du maximum de vraisemblance `method=ml` dans l'appel à `proc mixed`.

Informations sur le modèle	
Table	WORK.VENGEANCE
Variable dépendante	vengeance
Structure de covariance	Diagonal
Méthode d'estimation	ML

Remarque sur la comparaison de modèles

On ajuste les deux modèles par maximum de vraisemblance avec des erreurs autocorrélées (modèle AR(1)).

Modèle	**AIC**	**BIC**	$\widehat{\rho}$ (valeur-p)
`sexe`, `age`, `vc`, `wom`, `t`	**666,1**	685,1	$0{,}48$ (10^{-20})
`id`, `t`	653,4	**851,1**	$-0{,}013$ $(0{,}83)$

- Le modèle préférable selon le AIC inclut `id`, mais le AIC tend à choisir des modèles plus compliqués.
- Le modèle préférable selon le BIC n'inclut pas l'effet de groupe, mais plutôt `sexe`, `age`, `vc` et `wom`.
- Si on inclut un effet groupe, la corrélation entre les erreurs semble superflue — l'estimé du coefficient de corrélation est même négatif, ce qui est contre-intuitif et suggère un modèle surparamétrisé.

Remarque sur la comparaison de modèles

- Le choix des covariables dépend des buts de l'étude. Si on est intéressé par les effets de certaines des variables `sexe`, `age`, `vc` et `wom`, alors n'a pas le choix: on choisit le modèle qui les inclut.
- Si c'est seulement l'effet du temps qui est d'intérêt, alors les deux modèles mènent à la même conclusion de toute façon.
- Souvent, l'optimisation échoue — ça devient difficile d'estimer des paramètres de covariance et un effet fixe de groupe
- Nous verrons plus loin qu'il est possible d'incorporer à la fois des variables fixes au niveau des groupes (des personnes dans notre exemple) et des effets groupes (`id` dans notre exemple) en utilisant des effets aléatoires.