Introduction aux modèles linéaires mixtes -

Motivations et objectifs.

- Principe général de l'ANOVA : Analyse de la variabilité des données en étudiant la répartition de la variance observée.

  ↳ Au départ on se pose la question de la comparaison de moyennes, mais cette comparaison nécessite une bonne estimation de la ou des variances.

  ↳ Si on met un modèle (gaussien) on peut tester la significativité de la différence de moyennes entre certaines catégories

- Modèles à effets aléatoires : souvent utilisés pour décrire les relations entre des variables réponse et des covariables pour des données structurées (groupées) selon un ou plusieurs facteurs : blocs, strates, analyse longitudinale ....

  ↳ On introduit des effets aléatoires pour des obs. du même groupe → modélise la covariance induite par le design.

Facteurs, niveaux, cellules, effets.

. On cherche à attribuer la variabilité à différentes caractéristiques des données.

Ex : Essai clinique, 3 drogues ABC, sur hommes/femmes, marié ou non.

↳ 3 classifications : drogue, sexe, statut marital

→ identifie 3 sources potentielles de variation.
= Facteurs

	Marié			Non Marié		
	A	B	C	A	B	C
♂						
♀						

↳ Ces facteurs ont ≠ modalités → niveaux.

↳ les sous ensembles de données qui correspondent aux répétitions à un crois^t de facteurs sont appelés cellules.

Objectifs de l'ANOVA : Quel est (sont) l'effet des différents niveaux des facteurs sur une variable d'intérêt ?

l'effet peut concerner
→ la moyenne
→ la variance.

- Effets fixes : . Ils sont attribuables à un ensemble fini de niveaux d'un facteur.
  . Ils sont présents parce que d'intérêt (en principe !).

- Effets Aléatoires : Attribuable à un ensemble infini de niveaux pour lesquels seul un sous ensemble a été échantillonné.

→ C'est aussi un choix de modélisation, donc discutable !
avant tout ?

Exemple de modèle à effets fixes.

- Agronomie : . rendement de 24 plants de tomates
  . 6 plants × 4 variétés d'intérêt.

y_{ij} = rendement du fruit de variété i provenant du plant j.

$i = 1,4 \quad j = 1,6.$

- On suppose que l'espérance du rendement dépend de la variété :

$$\mathbb{E}\, Y_{ij} = \mu_i$$

- On décompose μ_i t.q. :

$$\mathbb{E}\, Y_{ij} = \mu_i = \mu_. + \alpha_i$$

moyenne générale — effet dû à la variété i

→ j est l'indice de répétition

↳ Dans ce modèle μ_i est considéré comme une constante.

↳ On souhaite comparer les magnitudes $\mu_1 - \mu_2$, ... pour comparer les niveaux du facteur variété (et choisir la meilleure par ex).

↳ Si on ne s'intéresse qu'à ces 4 variétés → l'effet est fixe.

- A partir du modèle $\mathbb{E}\, Y_{ij} = \mu + \alpha_i$

On construit la déviation de Y_{ij} par rapport @ $\mathbb{E} Y_{ij}$

$$\underbrace{Y_{ij} - \mathbb{E}\, Y_{ij}}_{\text{Résidu}} = Y_{ij} - (\mu + \alpha_i)$$

Résidu : par construct° / définit° centré : $\mathbb{E}\, E_{ij} = 0$.

↳ Structure au 2nd ordre ?

- chq E_{ij} a m̂ variance $\mathbb{V}(E_{ij}) = \sigma_e^2$
- E_{ij} iid. $\operatorname{cov}(E_{ij}, E_{i'j'}) = 0$. sauf pour $i = i'$, $j = j'$

Vers un modèle à effets mixtes.

. Essai clinique pour tester l'efficacité d'un placébo vs 3 trait$^{\text{t}}$ pour réduire la pression artérielle.

. 24 hommes âgés de 40-45 ans gagnant des salaires entre 100 et 250 k$/an.
. 6 h$\bar{o}$ pour chq type de drogue.
. mesure de la pression sanguine après 30j.

→ Effet traitement : $i = 1,4$: ne sont pas issus d'une population + large de traitements.

j : hommes : $1,6$: indice de répétitions

. Supposons que les essais sont faits dans des hôpitaux ≠. choisis aléatoirement dans la ville de NYC avec 5 patients par traitement dans chq clinique (15 cliniques)

→ étude multicentre.

$i = 1,4$, $j = 1,15$ cliniques, $k = 1,5$ (individus).

$$\mathbb{E}(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

effet clinique fixe ?.

→ plutôt aléatoire car les cliniques choisis st un échantillon

→ $\mathbb{E}(Y_{ijk}) = \mu + \alpha_i$ mais comment modéliser β_j ?

Modèle à effets Aléatoires.

. Exemple du modèle Animal en Agronomie / Génétique

. éleveurs font appel à des banques de semences pour "produire" des bovins, en vue d'avoir un bon rendement.

Ex : production laitière.

. les semenciers disposent de 80 - 150 taureaux qui sont considérés comme un échantillon aléatoire d'une population (ex : Holstein).

. On enregistre le rendement laitier de leurs descendants après 3 ans.

$$\begin{cases} i = 1, I : & \text{effet "père"} \\ j = 1, n_i : & \text{nombre de descendants du père } j. \end{cases}$$

. On suppose que les demi sœurs partageant le même père vont "partager" une performance semblable.

↳ Connaissant leur père, l'espérance de leur rendement serait μ_i

$$\mathbb{E}(Y_{ij} \mid A_i) = \mu_i = \mu + A_i$$

Mais "marginalement" $\mathbb{E}(Y_{ij}) = \mu$

- On suppose que l'effet du père est une réalisation d'un processus aléatoire (en général Gaussien).

$$\mathbb{E}(Y_{ij} \mid A_i = a_i) = \mu + a_i$$

$$\begin{cases} A_i \sim \mathcal{N}(0, \sigma_A^2) \\ \mathrm{cov}(A_i, A_k) = 0 \end{cases}$$

(les pères ne sont pas apparentés).

- Dans ce modèle $(\mu + A_i)$ est aléatoire et on n'en observe qu'une réalisat° $\mu + a_i$.

- Marginalement $\mu = \mathbb{E}\left(\mathbb{E}(Y_{ij} \mid A_i)\right)$

espérance suivant les différents pères
↳ Variabilité inter.

espérance suivant les répétitions pour 1 père
↳ Variabilité "intra"

Rq : si A_i n'était pas centré, on pourrait mettre tout espérance dans μ.

. Dans le modèle $E_{ij} = Y_{ij} - \mathbb{E}(Y_{ij} \mid A_i)$
$$= Y_{ij} - (\mu + A_i)$$

$$Y_{ij} = \underbrace{\mu}_{\text{effet Fixe}} + \underbrace{A_i}_{\text{effet Aléatoire}} + \underbrace{E_{ij}}_{\text{erreur aléatoire}}.$$

erreur aléatoire = dispersion de Y_{ij} autour de son espérance sachant le père.

On suppose : $A_i \sim N(0, \sigma_A^2)$
$E_{ij} \sim N(0, \sigma_E^2)$
$A_i \perp\!\!\!\perp E_{ij}$.

VARIANCE COMPONENT du rendement Y.

. $\mathbb{V}(Y_{ij}) = \mathbb{V}(A_i) + \mathbb{V}(E_{ij}) = \sigma_A^2 + \sigma_E^2$

. $\sigma_A^2 = \mathrm{Cov}(Y_{ij}, Y_{ij'})$ = Covariance intra père entre le rendement de 2 vaches issues du m père.

. On peut s'intéresser au ratio
$\left[\sigma_A^2 / \sigma_A^2 + \sigma_E^2\right]$ = % de variance due au père
$\cong h^2$ (héritabilité).

en psychologie /educational training s'interprete comme une fiabilité.

Prédiction des effets Aléatoires

→ Comment quantifier la Contribution du père i ? (sur le rdt).

→ Comment sélectionner les "bons" pères ?

A_i = effet spécifique du père i.

→ La contribution exacte du père i est inconnue et différente d'une fille à une autre (transmission d'une moitié aléatoire de génome).

↳ Ce n'est pas un effet fixe contrôlé.

↳ On pourra quantifier $\mathbb{E}(A_i \mid Y) = \hat{A}_i$

= prédiction de l'effet aléatoire ↑ tts les données.

→ $\hat{\mu} + \hat{A}_i$ permet de prédire le rendement d'un individu connaissant ses parents

→ pour sélectionner i vs i' $\hat{A}_{i'} < \hat{A}_i$

Problématiques générales du modèle mixte.

1)- Estimer les composants de la variance.
↳ permet de quantifier l'importance des effets aléatoires par rapport à la residuelle.

Plusieurs méthodes :
- . méthode ANOVA, ML, REML.
- Algorithme EM pour ML.
- problématique equilibre / non equilibré.

2). Estimer les effets fixes. (GLS, ML, Algo. EM pour ML).

3). Predire les Effets Aleatoires (BL, BLUP).

4). Tests des facteurs (effets fixes, Aleatoires).