HMMA307

Modèles à un facteur aléatoire

Cours: Joseph Salmon *Scribes: JOLY Julien - Mohamed Sahardid - Anas El Benna*

1 Modèles à deux facteurs avec intéractions

Modèle :

$$y_{i,j,k} = \mu^* + \alpha_i^* + \beta_j^* + \gamma_{i,j}^* + \varepsilon_{i,j,k}, \text{ pour } i = 1, \dots, I \text{ et } j = 1, \dots, J \ . \tag{1}$$

Remarque 1.1. *Ce modèle est compliqué si les classes sont déséquilibrées, on supposera donc dans la suite que*

$$\forall i \in [\![1, I]\!], j \in [\![1, J]\!], \quad n_{i,j} = K \ . \tag{2}$$

Notation :

$$\hat{y}_{i,j} = \frac{1}{K} \sum_{k=1}^{K} y_{i,j,k} = \overline{y}_{i,j:} \tag{3}$$

Contraintes sur l'estimateur : (XXX definition de l'estimateur manquante :)

$$\sum_{i=1}^{I} \alpha_i = 0 \tag{4}$$

$$\sum_{j=1}^{J} \beta_j = 0 \tag{5}$$

$$\forall i \in [\![1, I]\!], \quad \sum_{j=1}^{J} \gamma_{i,j} = 0 \tag{6}$$

$$\forall j \in [\![1, J]\!], \quad \sum_{i=1}^{I} \gamma_{i,j} = 0 \ . \tag{7}$$

Mise en forme matricielle : On pose : $n = I \times J \times K$. On peut donc écrire une matrice de design :

$$X = [\mathbb{1}_n, ..., \underbrace{\mathbb{1}_{C_1}, \dots, \mathbb{1}_{C_I}}_{Facteur\,1}, \underbrace{\mathbb{1}_{D_1}, \dots, \mathbb{1}_{D_K}}_{Facteur\,2}, \underbrace{\mathbb{1}_{C_1 \cap D_1}, \dots, \mathbb{1}_{C_I \cap D_J}}_{I.J\,termes}] \tag{8}$$

Nous avons donc $I \times J$ classes pour leurs intéractions. En effet :

$$\text{rang}(X) = 1 + (I-1) + (J-1) + (I-1)(J-1) = I \times J \ . \tag{9}$$

L'estimateur associé : Sous les contraintes (4), (5), (6) et (7) on obtient donc comme estimateur

$$\begin{aligned}\underset{(\mu,\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^I\times\mathbb{R}^J\times\mathbb{R}^{IJ}}{\arg\min} & \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{i,j,k}-\mu-\alpha_i-\beta_j-\gamma_{i,j})^2 \\ \text{s.c.} \quad & \sum_{i=1}^{I}\alpha_i=0 \ , \\ & \sum_{j=1}^{J}\beta_j=0 \ , \\ & \forall i\in[\![1,I]\!], \quad \sum_{j=1}^{J}\gamma_{i,j}=0 \\ & \forall j\in[\![1,J]\!], \quad \sum_{i=1}^{I}\gamma_{i,j}=0 \ .\end{aligned} \tag{10}$$

On peut alors former le Lagrangien du problème

$$\begin{aligned}\mathcal{L}(\mu,\alpha,\beta,\gamma,\lambda) =& \frac{1}{2}\sum_i\sum_j\sum_k(y_{i,j,k}-\alpha_i-\beta_j-\gamma i,j)^2 \\ & +\lambda_\alpha\sum_i\alpha_i+\lambda_\beta\sum_j\beta_j+\sum_i\nu_i\left(\sum_j\gamma_{i,j}\right)+\sum_j\eta_j\left(\sum_i\gamma_{i,j}\right) \ .\end{aligned} \tag{11}$$

Les conditions du premier ordre impliquent :

$$\begin{cases}\hat{\mu}=\overline{y}_n=\frac{1}{IJK}\sum_i\sum_j\sum_k y_{i,j,k} \\ \hat{\alpha}_i=\overline{y}_{i,:,:}-\overline{y}_n \\ \hat{\beta}_j=y_{:,j,:}-\overline{y}_n \\ \hat{\gamma}_{i,j}=\overline{y}_{i,j:}-\overline{y}_{i,:,:}-\overline{y}_{:,j,:}+\overline{y}_n \ .\end{cases} \ . \tag{12}$$

(XXX détails des calculs manquants.)

Test : Nous voulons tester l'hypothèse nulle :

$$(H_0): \text{“}\beta_1^*=\cdots=\beta_J^*=0'' \ . \tag{13}$$

Ce test permet de tester si le deuxième facteur a un effet ou non.

$$F_{obs}=\frac{\frac{1}{J-1}\sum_{j=1}^{J}(\overline{y}_{:,j,:}-y_n)^2}{\frac{1}{n-IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{y}_n-y_{i,j,k})^2} \ . \tag{14}$$

Sous H_0 : $F_{obs}\sim F_{n-IJ}^{J-1}$. (XXX test de Fisher) On peut fixer le niveau du test à α et considérer les quantiles $F_{n-IJ}^{J-1}(1-\alpha)$.

Remarque 1.2. *Par symétrie, en testant “$\alpha_1^*=\cdots=\alpha_I^*$”, le premier facteur n'a pas d'effet sous cette hypothèse.*

$$\overline{y}_k=\frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}y_{i,j,k} \ . \tag{15}$$

• Test :

(H_0) : "$\gamma_{i,j} = 0$" ; $\forall i, j$

$$F_{obs} = \frac{\frac{1}{(I-1)(J-1)} \sum_i \sum_j \hat{\gamma}_{i,j}^2}{\frac{1}{n-IJ} \sum_i \sum_j \sum_k (\overline{y}_n - y_{i,j,k})^2}$$

Sous H_0 : $F_{obs} \sim F_{n-IJ}^{(J-1)(I-1)}$

2 Modèle à un facteur aléatoire

Nous considérons ici les situations où par exemple les traitements sont des échantillons aléatoires provenant d'une large population de traitements. Cela peut sembler assez spécial à première vue, mais c'est en fait très naturel dans de nombreuses situations. Pensez par exemple à un échantillon aléatoire de classes scolaires qui a été tiré de toutes les classes scolaires d'un pays. Un autre exemple pourrait être des machines qui ont été choisies au hasard parmi une grande population de machines. Généralement, nous souhaitons faire une déclaration sur certaines propriétés de l'ensemble de la population et non sur celle des individus observés (ici : classes scolaires ou machines).

2.1 Modèle

$$y_{i,j} = \mu + A_j + \varepsilon_{i,j} \ . \tag{16}$$

- $\mu \in \mathbb{R}$: **effet fixe**
- $A_j \sim \mathcal{N}(0, \sigma_A^2)$ et *iid* , avec $j = 1, \ldots, J$ (les J niveaux) ; **effet aléatoire**
- $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ *iid*
- $\forall i, j$, $A_j \perp\!\!\!\perp \varepsilon_{i,j}$
- $n = \sum_{j=1}^{J} n_j$, où n_j est le nombre d'observations de modalité J.

Regardons maintenant l'espérance et la variance associées au modèle. On a :

$$\mathbb{E}[y_{i,j}] = \mathbb{E}[\mu + A_j + \varepsilon_{i,j}] = \mathbb{E}[\mu] = \mu \ . \tag{17}$$

Concernant la variance :

$$\begin{aligned} \mathrm{Var}[y_{i,j}] &= \mathrm{Var}[\mu + A_j + \varepsilon_{i,j}] \\ &= \mathrm{Var}[A_j + \varepsilon_{i,j}] \\ &= \mathrm{Var}[A_j] + \mathrm{Var}[\varepsilon_{i,j}] \\ &= \sigma_A^2 + \sigma_\varepsilon^2 \ . \end{aligned} \tag{18}$$

Nous considérons maintenant la covariance du modèle. Pour j fixé et $i \neq i'$:

$$\begin{aligned}\mathrm{Cov}(y_{i,j}; y_{i'j}) &= \mathrm{Cov}(\mu + A_j + \varepsilon_{i,j}, \mu + A_j + \varepsilon_{i'j}) && (19)\\ &= \mathrm{Cov}(A_j + \varepsilon_{i,j}, A_j + \epsilon_{i'j}) && (20)\\ &= \mathrm{Cov}(A_j, A_j) + \mathrm{Cov}(\varepsilon_{i,j}, \varepsilon_{i'j}) + \mathrm{Cov}(A_j, \varepsilon_{i,j}) + \mathrm{Cov}(A_j, \varepsilon_{i'j}) && (21)\\ &= \mathrm{Cov}(y_{i,j}; y_{i'j}) && (22)\\ &= \sigma_A^2 \ . && (23)\end{aligned}$$

Avec $j \neq j'$:

$$\mathrm{Cov}(y_{i,j}; y_{i'j}) = \mathrm{Cov}(A_j + \varepsilon_{i,j}, A_j + \epsilon_{i'j}) = 0 \ . \quad (24)$$

Remarque 2.1. *J niveaux modélisent (souvent) un effet aléatoire de niveaux qui font partie d'une plus grande population (dont on observe une sous-partie).*

Exemples :

- cadre médical : cohorte (sous-partie de la population), essais cliniques, etc.
- étude animale, végétale, etc.

Écriture matricielle :

On a le modèle suivant :

$$\begin{aligned} y &= \nu \mathbb{1}_n + \varepsilon && (25)\\ \varepsilon \in \mathbb{R}^n &\sim \mathcal{N}(0, \sigma^2 \,\mathrm{Id}_n) \ . && (26)\end{aligned}$$

Nous posons alors : $Z = [\mathbb{1}_{C_1}, \mathbb{1}_{C_2}, \ldots, \mathbb{1}_{C_J}] \in \mathbb{R}^{n \times J}$ avec $C_1 \cup C_2 \cup \cdots \cup C_n = [\![1, n]\!]$.

Remarquons également que Z est déterministe.

• $A = (A_1, \ldots, A_J)^\top \in \mathbb{R}^J$ et tel que $A \sim \mathcal{N}(0, \sigma_A^2 \,\mathrm{Id}_J)$.

On a alors :

$$\mathrm{Var}(ZA) = Z \,\mathrm{Var}(A) Z^\top = \sigma_A^2 (ZZ^\top) \ . \quad (27)$$

$$ZZ^\top = \sum_{j=1}^{J} \mathbb{1}_{C_j} \mathbb{1}_{C_j}^\top \in \mathbb{R}^{n \times n} \ . \quad (28)$$

Donc :

$$\mathrm{Var}(y) = \mathrm{Var}(ZA) + \mathrm{Var}(\epsilon) = \sigma_A^2 ZZ^\top + \sigma^2 \,\mathrm{Id} \ . \quad (29)$$

Exemple : Posons $J = 2$, $n_1 = 3$ et $n_2 = 2$, nous avons alors :

$$Z = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \; A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \text{ et } ZZ^\top = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}. \text{ Donc,}$$

$$\mathrm{Var}(y) = \begin{pmatrix} \sigma_A^2 & \sigma_A^2 & \sigma_A^2 & 0 & 0 \\ \sigma_A^2 & \sigma_A^2 & \sigma_A^2 & 0 & 0 \\ \sigma_A^2 & \sigma_A^2 & \sigma_A^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_A^2 & \sigma_A^2 \\ 0 & 0 & 0 & \sigma_A^2 & \sigma_A^2 \end{pmatrix} + \sigma_\epsilon^2 \,\mathrm{Id} = \begin{pmatrix} \sigma_A^2 + \sigma_\epsilon^2 & \sigma_A^2 + \sigma_\epsilon^2 & \sigma_A^2 + \sigma_\epsilon^2 & 0 & 0 \\ \sigma_A^2 + \sigma_\epsilon^2 & \sigma_A^2 + \sigma_\epsilon^2 & \sigma_A^2 + \sigma_\epsilon^2 & 0 & 0 \\ \sigma_A^2 + \sigma_\epsilon^2 & \sigma_A^2 + \sigma_\epsilon^2 & \sigma_A^2 + \sigma_\epsilon^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_A^2 + \sigma_\epsilon^2 & \sigma_A^2 + \sigma_\epsilon^2 \\ 0 & 0 & 0 & \sigma_A^2 + \sigma_\epsilon^2 & \sigma_A^2 + \sigma_\epsilon^2 \end{pmatrix} \tag{30}$$

Bibliographie

— "ANOVA : A Short Intro Using R", Lukas Meier, https://stat.ethz.ch/~meier/teaching/anova/

— "Modèles à effets aléatoires et modèles mixtes", Philippe Besse, https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modmixt6-modmixt.pdf