Chapitre 5

Modèle linéaire mixte

On appelle modèle mixte un modèle statistique dans lequel on considère à la fois des facteurs à effets fixes (qui interviennent sur la moyenne dans différents groupes du modèle) et des facteurs à effets aléatoires (qui interviennent sur la variance du modèle). Un modèle est dit mixte lorsqu'il y a au moins un facteur de chaque nature. Dans le cadre de ce cours, nous ne considérons que des modèles linéaires gaussiens mixtes, mais la notion de modèle mixte se rencontre également dans d'autres contextes, notamment dans le modèle linéaire généralisé.

5.1 Écriture du modèle

Modèle

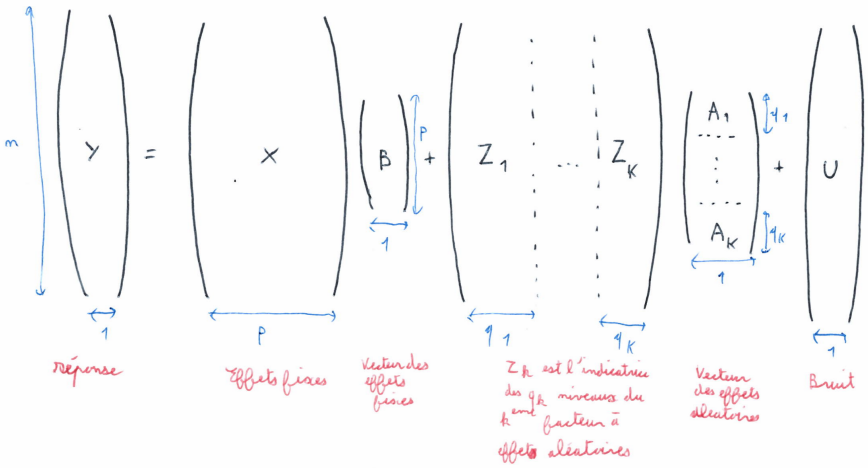
Un modèle linéaire gaussien mixte à n observations s'écrit sous la forme matricielle suivante :

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\beta + \sum_{k=1}^{K} \mathbf{Z_k}\mathbf{A_k} + U \\ &= \mathbf{X}\beta + \mathbf{ZA} + U\end{aligned}$$

où :

- $\mathbf{Y}$ est le vecteur aléatoire réponse de $\mathbb{R}^n$.
- $\mathbf{X}$ est la matrice $n \times p$ relative aux effets fixes du modèle, où p est le nombre total d'effets fixes pris en compte dans le modèle.
- β est le vecteur des p effets fixes β_j, $j = 1, \ldots, p$ à estimer.
- Z_k est la matrice des indicatrices (disposées en colonnes) des niveaux du $k$ième facteur à effets aléatoires ($k = 1, \ldots, K$). On note q_k le nombre de niveaux de ce facteur. $\mathbf{Z_k}$ est alors de dimension $n \times q_k$.
- On note A_{kl} la v.a.r. associée au $l$ème niveau du $k$ème facteur à effets aléatoires avec $l = 1, \ldots, q_k$. Pour tout l lié au facteur k, on suppose $A_{kl} \sim \mathcal{N}(0, \sigma_k^2)$.
- Pour un facteur k donnée, on note $A_k = (A_{k1}, \ldots, A_{kq_k})'$ le vecteur colonne des A_{kl}. On suppose que $A_k \sim \mathcal{N}(0, \sigma_k^2 \mathbf{I}_{q_k})$.
- Enfin, U est le vecteur aléatoire des erreurs du modèle qui vérifie $U \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$.

Le modèle peut alors être écrit sous la forme :



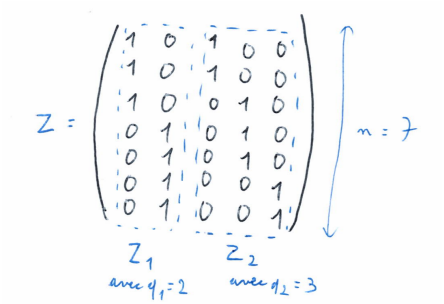
Moments du modèle

Il est évident de montrer que $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\beta$ d'après les hypothèses sur le modèle. On note aussi $\mathbf{V}$ la variance de $\mathbf{Y}$ qui se calcule par :

$$\begin{aligned}\mathbf{V} &= Var(\mathbf{Y}) \\ &= Var(\mathbf{ZA}) + Var(\mathbf{U}) \\ &= \sum_{k=1}^{K} \left(\sigma_k^2 \mathbf{Z_k} \mathbf{Z_k}'\right) + \sigma^2 \mathbf{I_n} \\ &= \mathbf{Z_k} \mathbf{G} \mathbf{Z_k}' + \sigma^2 \mathbf{I_n}\end{aligned}$$

où $\mathbf{G} = diag(\sigma_1^2 \mathbf{I_{q_1}}, \ldots, \sigma_K^2 \mathbf{I_{q_K}})$. On obtient alors $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\beta, \mathbf{V})$.

Les composantes de Y ne sont ainsi pas indépendantes au sein de chaque niveau l d'un facteur aléatoire k donné. Ceci est évident si on observe un exemple de ce à quoi peut resembler $\mathbf{Z}$:



Dans cet exemple, le vecteur $\mathbf{ZA}$ aura la forme $(\xi_{11} + \xi_{21}, \xi_{11} + \xi_{21}, \xi_{11} + \xi_{22}, \xi_{12} + \xi_{22}, \xi_{12} + \xi_{22}, \xi_{12} + \xi_{23}, \xi_{12} + \xi_{23})'$ où les $\xi_{kl} \sim \mathcal{N}(0, \sigma_k^2)$ ce qui induit les dépendances intra-niveau des Y.

5.2 Estimation des β

L'expression que l'on obtient dans le cas général pour $\hat{\beta}$ fait intervenir l'estimation de la matrice des variances-covariances $\mathbf{V}$ de $\mathbf{Y}$. Cette expression obtenue est fournie par la méthode des moindres carrés généralisés notée $GLSE(\beta)$ (pour Generalized Least Squares Estimator) :

$$\hat{\beta} = \arg\min_{\beta}(\mathbf{Y} - \mathbf{X}\hat{\beta})'\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

où $\hat{\mathbf{V}} = \sum_{k=1}^{K} \hat{\sigma}_k^2 \mathbf{Z_k}\mathbf{Z_k}' + \hat{\sigma}^2\mathbf{I_n}$ et les $\hat{\sigma}_k^2$ et $\hat{\sigma}^2$ sont les composantes de variances. On a alors :

$$\hat{\beta} = GLSE(\beta) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}$$

et il est nécessaire d'estimer les composantes de covariance, ce qui se fait typiquement par maximum de vraisemblance.

On remarquera que dans le cas équilibré où tous les q_k ont la même valeur, on a plus simplement :

$$\hat{\beta} = OLSE(\beta) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

où $OLSE$ signifie *Ordinary Least Squares Estimator*.

5.3 Estimation de V

Pour estimer $\mathbf{V}$ par maximum de vraisemblance, on note d'abord $\Psi = (\hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2, \hat{\sigma}^2)$ les paramètres à estimer dont dépend $\mathbf{V}$. La log-vraisemblance du modèle mixte gaussien s'écrit :

$$l(y, \beta, \mathbf{V}(\Psi)) = -\frac{1}{2}\log\left(\det\left(\mathbf{V}(\Psi)\right)\right) - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{V}(\Psi))^{-1}(\mathbf{Y} - \mathbf{X}\beta)$$

On en déduit le système de p équations :

$$\frac{\partial l}{\partial \beta} = \mathbf{X}'\mathbf{V}^{-1} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta$$

dont découlent les équations normales pour $\hat{\beta}$.

On remarque ensuite que :

$$\frac{\partial \mathbf{V}}{\partial \sigma_k^2} = \mathbf{Z_k}\mathbf{Z_k}'$$

On déduit alors que pour chaque σ_k^2 :

$$\frac{\partial l}{\partial \sigma_k^2} = -\frac{1}{2}tr(\mathbf{V}(\Psi)\mathbf{Z_k}\mathbf{Z_k}') + \frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{V}(\Psi))^{-1}\mathbf{Z_k}\mathbf{Z_k}'(\mathbf{V}(\Psi))^{-1}(\mathbf{Y} - \mathbf{X}\beta)$$

On obtient ainsi un système de $K + 1 + p$ équations non linéaires à $K + 1 + p$ inconnues que l'on résoud par une méthode numérique itérative. Ces procédures numériques fournissent en plus, à la convergence, la matrice des variances-covariances asymptotiques des estimateurs.

5.4 Tests de significativité des facteurs

Ces tests sont standards dans le cas équilibré (tous les q_k sont égaux), mais deviennent assez problématiques dans le cas déséquilibré. Dans le cas équilibré le test de Fisher sur les variances est en effet valable (comme dans ANOVA sous-section 4.3 ou ANCOVA sous-section 4.6). Il n'y a cependant pas de test exact, ni même de test asymptotique, qui permette de tester les effets, que ce soient les effets fixes ou les effets aléatoires, dans un modèle mixte avec un plan déséquilibré. Il existe seulement des tests approchés (dont on ne contrôle pas réellement le niveau, et encore moins la puissance)