Fondements statistiques de l’apprentissage automatique

Chapitre 4 : Analyse de covariance et modèles mixtes

Laurent Risser (CNRS) — Année 2025/2026

Nous avons étudié jusque là le lien entre :

- des variables d'entrée quantitatives X
- des variables quantitative de sortie Y

Nous allons maintenant traiter de modèles pour lesquels Y dépend aussi d'une variable qualitative, c'est d'une appartenance à un groupe.

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)
- des facteurs à effets aléatoires (qui interviennent sur la variance du modèle)

Nous allons dans ce cours nous restreindre à l'étude des modèles linéaires gaussiens mixtes :

$$Y = X\beta + \sum_{k=1}^{K} Z_k A_k + U$$

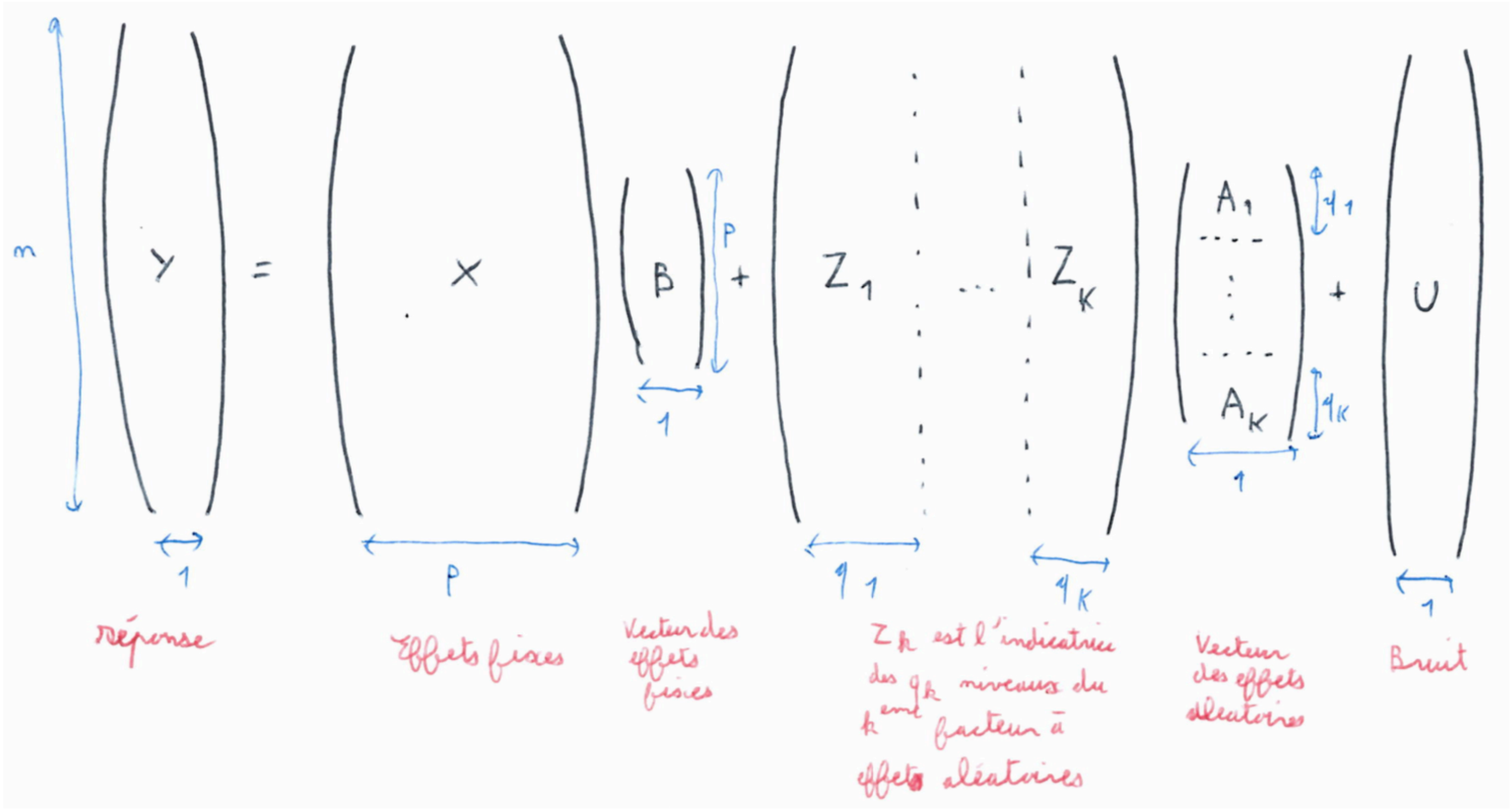
vecteur aléatoire réponse de $\mathbb{R}^n$.

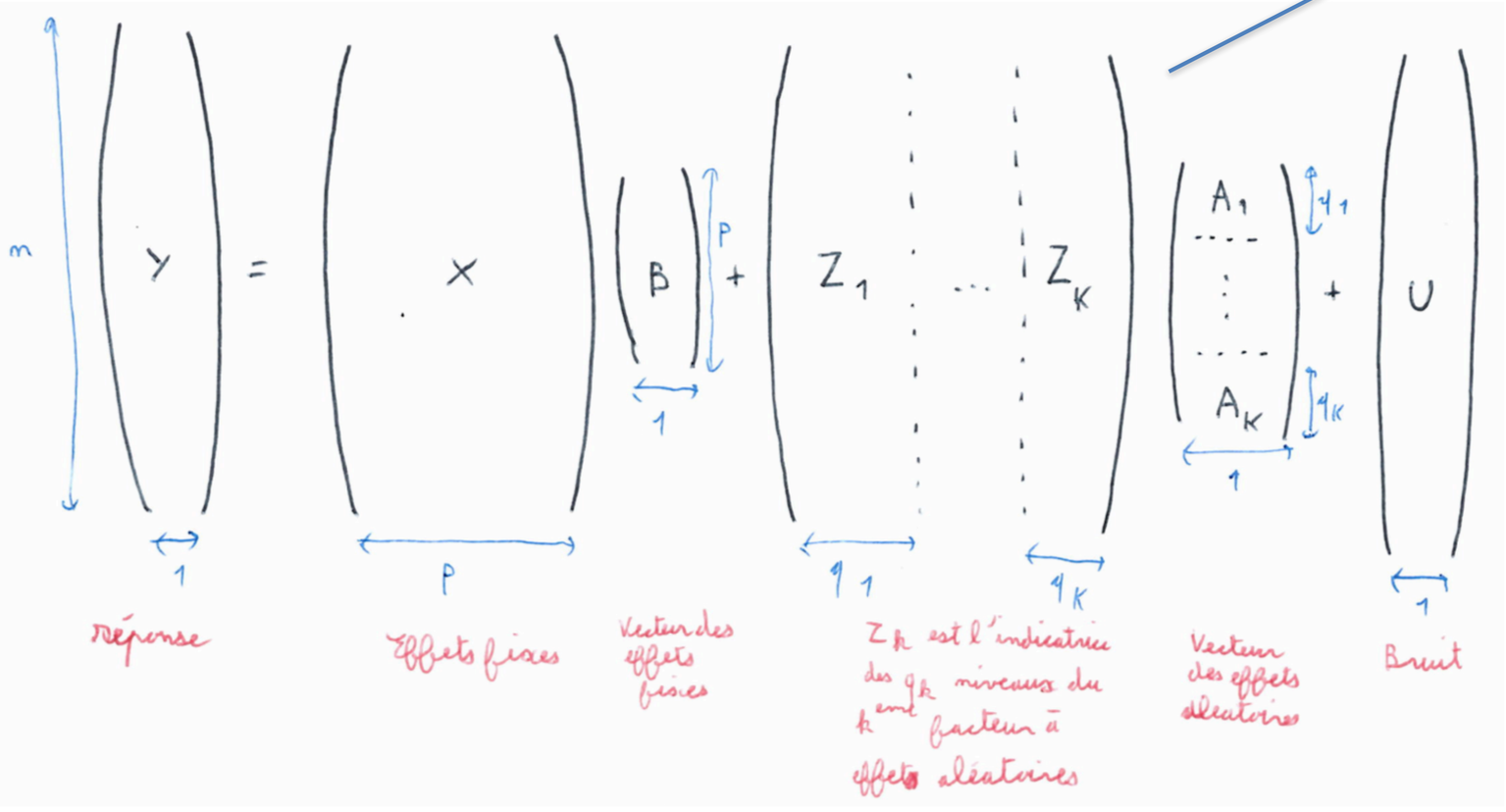
matrice $n \times p$ relative aux effets fixes du modèle

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 $n \times q_k$.

vecteur aléatoire des erreurs du modèle qui vérifie $U \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$.

… Avant de décrire ce que représente A_k , nous allons représenter le détail de ce modèle …

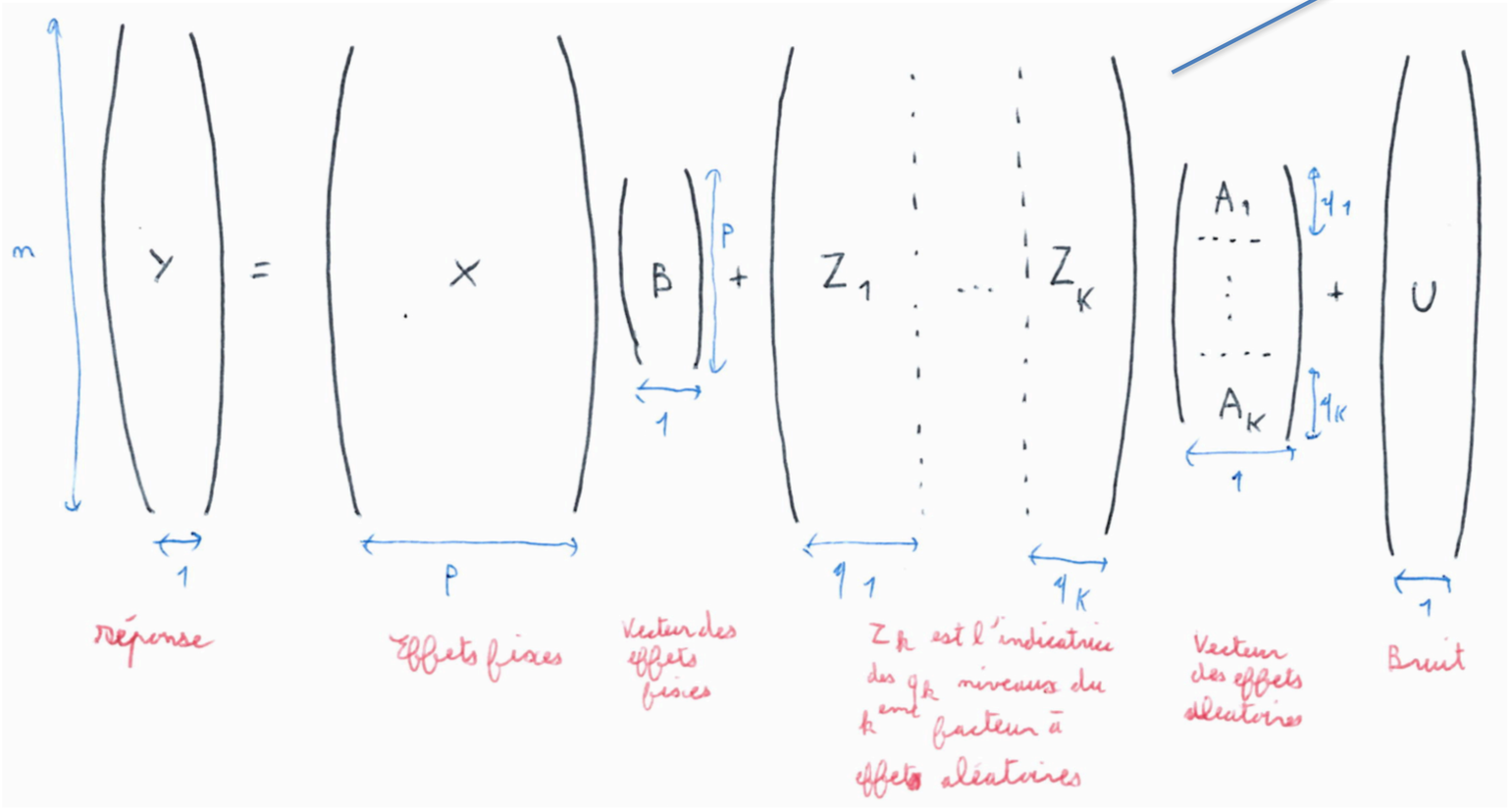

n
Y
=
X
P
β
P
1
+
Z 1
Z K
q 1
q K
A 1
A K
q 1
q K
1
+
U
1
réponse
Effets fixes
Vecteur des effets fixes
Z k est l'indicatrice des q k niveaux du k ème facteur à effets aléatoires
Vecteur des effets aléatoires
Bruit



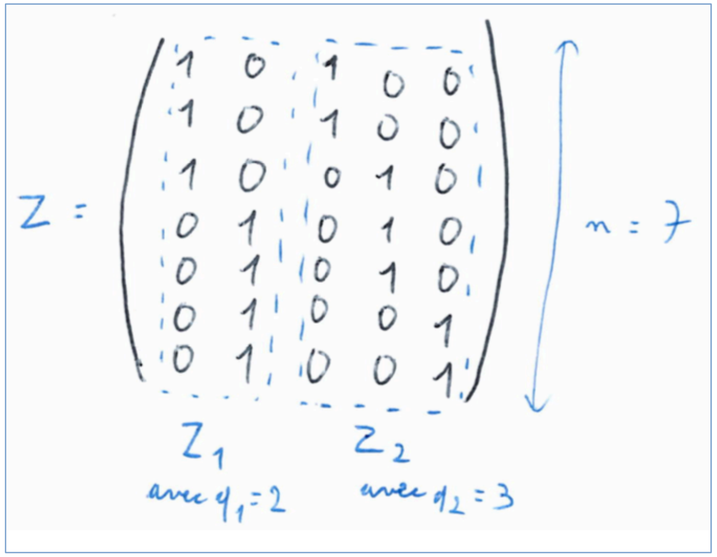
Exemple

$$Z = \left(\begin{array}{cc|ccc} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{array}\right) \quad n = 7$$

Z_1 avec $q_1 = 2$ $\qquad$ Z_2 avec $q_2 = 3$



Exemple



On note alors A_{kl} la variable aléatoire associée au niveau l du facteur k à effets aléatoires avec $l = 1,\ldots, q_k$.

Pour un facteur k :

- On note $A_k = \left(A_{k1}, \ldots, A_{k1}\right)^\mathsf{T}$ le vecteur colonne des A_{kl}
- On suppose $A_{kl} \sim \mathcal{N}(0,\sigma_k^2)$ pour tous les l du facteur k
- Les A_{kl} ne seront pas estimés. Les σ_k le seront par contre pour trouver β

Estimation de

$$\mathsf{V} = Var(\mathsf{Y}) = \sum_{k=1}^{K} \left(\sigma_k^2 Z_k Z_k^\intercal \right) + \sigma \, \mathsf{I}_n$$

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.

Estimation de 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*.

Test 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)

Analyse de covariance

L'analyse de covariance se situe dans un contexte où une variable quantitative Y est expliquée par plusieurs variables à la fois quantitatives et qualitatives. Le principe général est alors d'estimer des modèles *intra-groupes* et de tester des effets différentiels *inter-groupes* des paramètres des régressions.

Exemple : On s'intéresse à l'effet d'un médicament en fonction du dosage de plusieurs particules (variables quantitatives) et de l'origine géographique du patient (variable qualitative).

Nous nous intéressons ici au cas le plus simple où seulement une variable X parmi les explicatives est quantitative.

On considère une variable quantitative Y expliquée par une variable qualitative T à J niveaux et une variable quantitative (appelée encore covariable) X. Pour chaque niveau j de T , on observe n_j valeurs $x_{1j}, \ldots, x_{n_j j}$ de X et n_j valeurs $y_{1j}, \ldots, y_{n_j j}$ de Y. La taille de l'échantillon est alors $n = \sum_{j=1}^{J} n_j$.

Données du niveau *(groupe)* 1 : $\{x_{1,1}, y_{1,1}\}, \{x_{2,1}, y_{2,1}\}, \ldots, \{x_{n_1,2}, y_{n_1,2}\}$

Données du niveau *(groupe)* 2 : $\{x_{1,2}, y_{1,2}\}, \{x_{2,2}, y_{2,2}\}, \ldots, \{x_{n_2,2}, y_{n_2,2}\}$

…

Données du niveau *(groupe)* j : $\{x_{1,j}, y_{1,j}\}, \{x_{2,j}, y_{2,j}\}, \ldots, \{x_{n_j,j}, y_{n_j,j}\}$

…

Données du niveau *(groupe)* J : $\{x_{1,J}, y_{1,J}\}, \{x_{2,J}, y_{2,J}\}, \ldots, \{x_{n_J,J}, y_{n_J,J}\}$

Pour chaque $j = 1, \ldots, J$ et $i = 1, \ldots, n_j$, on suppose alors que :

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$$

où mes ε_{ij} sont i.i.d. de loi centrée et de variance σ^2. Pour la construction de tests, on suppose que $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

Il sera intéressant de noter que l'estimation des paramètres de ce modèle pourra se faire indépendamment dans chaque groupe/niveau à l'aide d'une régression linéaire simple. Les paramètres estimés vont par contre permettre d'effectuer des tests statistiques potentiellement informatif sur l'impact (ou non) de l'appartenance à un groupe.

Avant de tester l'impact potentiel des groupes, le modèle y $= \sum_{j=1}^{J} \left(\beta_{Oj} 1_j + \beta_{1j} x 1_j \right)$ + u va être ré-écrit de manière à mettre en lumière des différences entre chaque groupe et un groupe de référence, disons J.

On considère alors

- $\beta'_{0j} = \beta_{0j} - \beta_{0J}$
- $\beta'_{1j} = \beta_{1j} - \beta_{1J}$

Le modèle devient alors : $\mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1} \left(\beta'_{0j}\mathbf{1_j} + \beta'_{1j}\mathbf{x}.\mathbf{1_j} \right) + \mathbf{u}$

$$\mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1}\left(\beta'_{0j}\mathbf{1_j} + \beta'_{1j}\mathbf{x}.\mathbf{1_j}\right) + \mathbf{u}$$

Différentes hypothèses peuvent alors être testées en comparant ce modèle à un des modèles réduits suivants :

$$(i)\,\mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1}\left(\beta'_{0j}\mathbf{1_j}\right) + \mathbf{u}$$

$$(ii)\,\mathbf{y} = \beta_{0J}\mathbf{1} + \sum_{j=1}^{J-1}\left(\beta'_{0j}\mathbf{1_j} + \beta'_{1j}\mathbf{x}.\mathbf{1_j}\right) + \mathbf{u}$$

$$(iii)\,\mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1}\left(\beta'_{1j}\mathbf{x}.\mathbf{1_j}\right) + \mathbf{u}$$

$$\mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1}\left(\beta'_{0j}\mathbf{1_j} + \beta'_{1j}\mathbf{x}.\mathbf{1_j}\right) + \mathbf{u}$$

Différentes hypothèses peuvent alors être testées en comparant ce modèle à un des modèles réduits suivants :

$$(i)\ \mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1}\left(\beta'_{0j}\mathbf{1_j}\right) + \mathbf{u} \longrightarrow H_0^{(i)} : \beta_{11} = \ldots = \beta_{1J}$$, c'est à dire que les droites partagent la même pente

$$(ii)\ \mathbf{y} = \beta_{0J}\mathbf{1} + \sum_{j=1}^{J-1}\left(\beta'_{0j}\mathbf{1_j} + \beta'_{1j}\mathbf{x}.\mathbf{1_j}\right) + \mathbf{u} \longrightarrow H_0^{(ii)} : \beta_{1J}$$ est nul

$$(iii)\ \mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1}\left(\beta'_{1j}\mathbf{x}.\mathbf{1_j}\right) + \mathbf{u} \longrightarrow H_0^{(iii)} : \beta_{01} = \ldots = \beta_{0J}$$, c'est à dire que les droites partagent la même origine

Un test de Fisher teste alors l'égalité des variances des résidus entre les modèles comparées avec les β_{0j} et β_{1j} optimums.

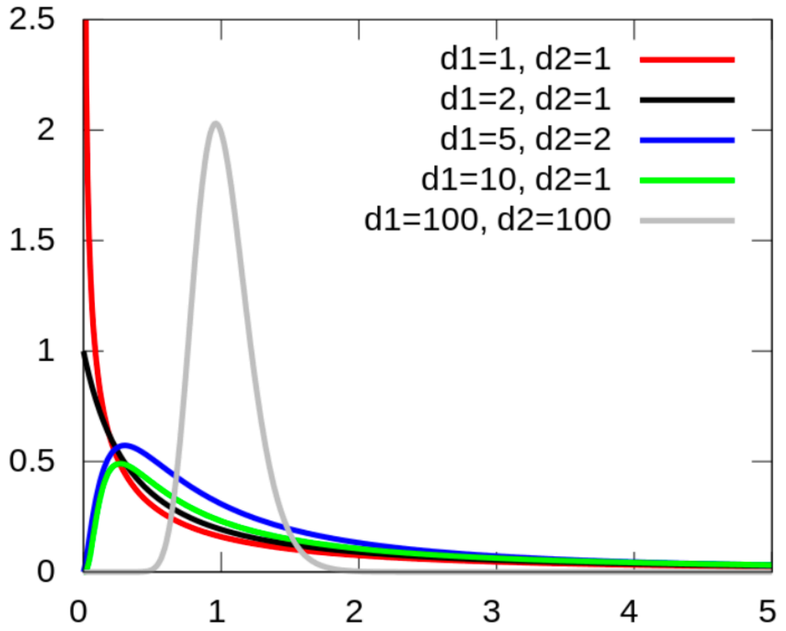
Test F de Fisher d'égalité des variances : teste l'hypothèse nulle que deux lois normales ont la même variance.

→ Ici nous supposerons que les erreurs obtenues avec les modèles (1) et (2) sont $\{e_i^{(1)}\}_{i=1,\dots,n_1}$ et $\{e_i^{(2)}\}_{i=1,\dots,n_2}$

→ On suppose $e^{(1)} \sim \mathcal{N}(0,\sigma_1^2)$ et $e^{(2)} \sim \mathcal{N}(0,\sigma_2^2)$

Alors :

- La statistique de test fait appel à la loi de Fisher F : $Z = \dfrac{S_{n_1}^2}{S_{n_2}^2} \sim F(n_1, n_2)$
- L'hypothèse testée est : $H_0 : \sigma_1^2 = \sigma_2^2$ et $H_1 : \sigma_1^2 \neq \sigma_2^2$



Loi de Fisher-Snedecor