

# **Fondements statistiques de l'apprentissage automatique**

## **Chapitre 4 : Analyse de covariance et modèles mixtes**

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.**

par exemple la volonté d'acheter des produits dépend aussi du pays de la personne (pcq ya toute la culture en arrière), ou les maladies, ça dépend aussi du métabolisme dcp si la personne est d'un groupe de risque, etc

# Modèle linéaire mixte

On appelle modèle mixte un modèle statistique dans lequel on considère à la fois

explicação desses dois  
na página 5

- 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)

era pra ser gaussien

Nous allons dans ce cours nous restreindre à l'étude des modèles linéaires gaulliens 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

*ex de facteurs: homme/femme, pays, etc  
alors, le premier, par ex, a deux niveaux (H/F), le deuxième dépend de tes données, etc*

*Z<sub>k</sub> est la matrice des indicatrices (disposées en colonnes) des niveaux du k<sup>ième</sup> facteur à effets aléatoires ( $k = 1, \dots, K$ ). On note  $q_k$  le nombre de niveaux de ce facteur. Z<sub>k</sub> est alors de dimension  $n \times q_k$ .*

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

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

$$y = X \beta + Z_1 q_1 + \dots + Z_K q_K + A_1 \gamma_1 + \dots + A_K \gamma_K + u$$

Diagram illustrating the components of a mixed model equation:

- Réponse**: The dependent variable  $y$  is represented by a vertical vector with dimension  $m$ .
- Effets fixes**: The matrix  $X$  represents fixed effects, with dimension  $P$ .
- Vecteur des effets fixes**: The vector  $\beta$  represents the fixed effect parameters.
- Bruit**: The error term  $u$  represents the noise or random error.
- Indicateurs**: The matrices  $Z_1, \dots, Z_K$  are indicator matrices that identify the levels of the random factors.
- Niveaux**: The vectors  $q_1, \dots, q_K$  represent the levels of the random factors.
- Effets aléatoires**: The vectors  $\gamma_1, \dots, \gamma_K$  represent the random effect parameters.
- Dimensions**: Dimensions are indicated by arrows:  $m$  for the response vector,  $P$  for the fixed effects matrix, and  $q_k$  for the indicator vectors.

# FS-AA - chapitre 4 : Analyse de covariance et modèles mixtes → 4.2 Modèles mixtes

## FIXED EFFECTS

### Definition

Fixed effects are variables whose levels are specifically chosen and of direct interest. You want to estimate and compare the effect of each level.

### Key Characteristics

- > Levels are not random; they are predefined.
- > You care about the individual effect of each level.
- > Results are generalizable only to those levels.

### Example

Suppose you're studying the effect of three teaching methods (A, B, and C) on student performance.

- > You deliberately chose these methods.
- > You want to compare their specific effects.
- > These methods are fixed effects.

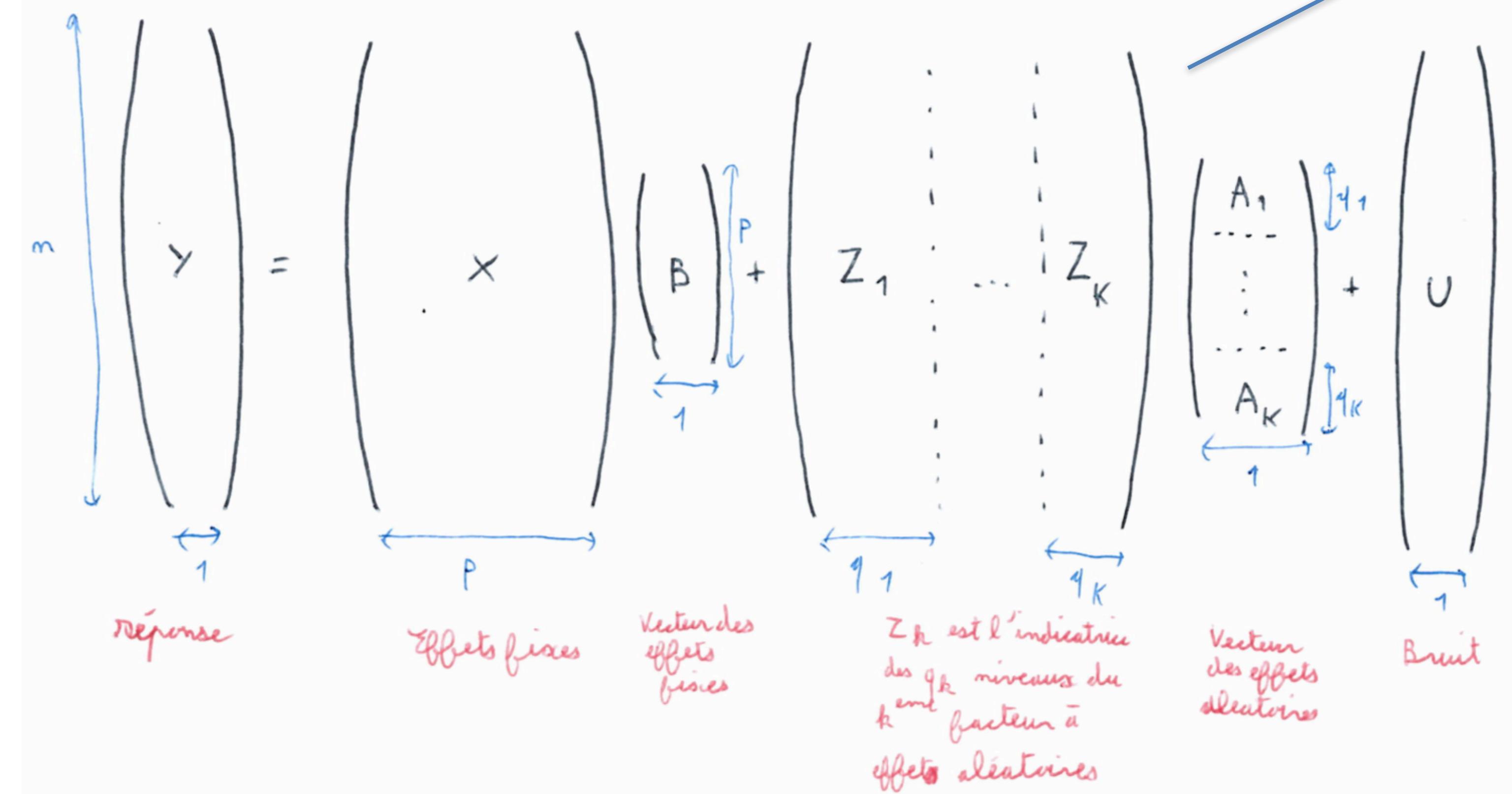
### Interpretation

You might say: "Method A improves scores by 5 points more than Method B."

Exemple

$$Z = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \quad n = 7$$

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



## RANDOM EFFECTS

### Definition

Random effects are variables whose levels are random samples from a larger population. You're not interested in the specific levels, but in the variation they introduce.

### Key Characteristics

- > Levels are randomly selected.
- > You care about the variance they contribute, not their individual effects.
- > Results are generalizable to the population from which the levels were drawn.

### Example

Suppose you're studying student performance across 10 randomly selected schools.

- > You don't care about School 1 vs. School 2.
- > You care about how much school-to-school variation affects performance.
- > Schools are random effects.

### Interpretation

You might say: "School-level differences account for 20% of the total variance in student scores."

## FS-AA - chapitre 4 : Analyse de covariance et modèles mixtes → 4.2 Modèles mixtes

por ex, estudando a nota q, no sul do Brasil, dão pra bebida (y) em função da intensidade da erva (x), mas pode variar com dois fatores de efeitos aleatórios: sexo e estado. Daí o primeiro efeito tem dois níveis (Masculino/Feminino) e o segundo 3 (PR, SC e RS). Daí, a matriz Z das indicatrices poderia ser como essa da direita

$$Z = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \quad n = 7$$

$Z_1$   
avec  $q_1 = 2$

$Z_2$   
avec  $q_2 = 3$

$$\begin{matrix} m \\ \downarrow \\ y \end{matrix} = \begin{matrix} \times \\ P \\ \leftrightarrow \\ \text{réponse} \end{matrix} = \begin{matrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{matrix} \times \begin{matrix} B \\ \uparrow \\ P \\ \leftrightarrow \\ \text{effets fixes} \end{matrix} + \begin{matrix} Z_1 \\ \vdots \\ \dots \\ \vdots \\ \dots \\ \vdots \\ Z_K \end{matrix} + \begin{matrix} A_1 \\ \vdots \\ \dots \\ \vdots \\ A_K \\ \uparrow \\ q_1 \\ \dots \\ q_K \end{matrix} + \begin{matrix} u \\ \uparrow \\ \text{Bruit} \end{matrix}$$

Z<sub>k</sub> est l'indicatrice des  $q_k$  niveaux du k<sup>ème</sup> facteur à effets aléatoires  
 Vecteur des effets aléatoires

Vecteur des effets fixes

Exemple

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

$$A_{11} = \text{Masc}, A_{23} = \text{RS}, A_{21} = \text{PR}$$

Pour un facteur  $k$ :

- On note  $A_k = (A_{k1}, \dots, A_{k1})^T$  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  $\sigma_k$  le seront par contre pour trouver  $\beta$

Estimation de

$$\mathbf{V} = \text{Var}(\mathbf{Y}) = \sum_{k=1}^K \left( \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k^\top \right) + \sigma^2 \mathbf{I}_n$$

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

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

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  $\sigma_k^2$  :

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

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  $\beta$ 

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

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

où  $\hat{\mathbf{V}} = \sum_{k=1}^K \hat{\sigma}_k^2 \mathbf{Z}_k \mathbf{Z}_k'$  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  $x_{1j}, \dots, x_{n_j j}$  de  $X$  et  $n_j$  valeurs  $y_{1j}, \dots, y_{n_j j}$  de  $Y$ . La taille de l'échantillon est alors  $n = \sum_{j=1}^J n_j$ .

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

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

...

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

...

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

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

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

où mes  $\varepsilon_{ij}$  sont i.i.d. de loi centrée et de variance  $\sigma^2$ . 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.

$1_j$  est le vecteur colonne de taille n des indicatrices du fait d'être dans la classe j.  
Son ième élément vaut 1 si la ième observation est dans la classe j, et 0 sinon.

ent esse cara vira, essencialmente, um vetor coluna de tamanho n (número de observações) em q cada elemento é o  $\beta_0 + \beta_1 * x$  usando os parâmetros do grupo ao qual o x daquela linha, ou seja, ao qual a observação pertence

Avant de tester l'impact potentiel des groupes, le modèle  $y = \sum_{j=1}^J (\beta_{0j} 1_j + \beta_{1j} x 1_j) + u$  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 :  $\mathbf{y} = \beta_{0J} \mathbf{1} + \beta_{1J} \mathbf{x} + \sum_{j=1}^{J-1} (\beta'_{0j} \mathbf{1}_j + \beta'_{1j} \mathbf{x} \cdot \mathbf{1}_j) + \mathbf{u}$

perceba q esse é idêntico ao de cima, pq se vc somar todas as matrizes sem ser o u, vc tem a soma dos parâmetros pro nível J com a diferença dos parâmetros do nível daquela observação pro J, ou seja, vc tem o  $\beta_0 + \beta_1 * x$  usando os parâmetros correspondentes ao nível da feature ao qual a observação pertence. Por ex: se a linha 1 for uma observação do estado do RS, ent vão ser usados os betas do RS

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

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

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

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

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

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

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

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

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

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

Un test de Fisher teste alors l'égalité des variances des résidus entre les modèles comparés avec les  $\beta_{0j}$  et  $\beta_{1j}$  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  $F$  :  $Z = \frac{S_{n_1}^2}{S_{n_2}^2} \sim F(n_1, n_2)$
- L'hypothèse testée est :  $H_0 : \sigma_1^2 = \sigma_2^2$  et  $H_1 : \sigma_1^2 \neq \sigma_2^2$

