
UNIVERSITÉ DES SCIENCES
(MONTPELLIER)

RAPPORT DU PROJET

# INTRODUCTION AUX MODÈLES LINÉAIRES GÉNÉRALISÉS MIXTES


Présenté par :
KHALIFI OUMAYMA
M2 MIND

*Encadré par :* M. JOSEPH SALMON

08/11/2020

**Partie théorique :**

**Définition :**

Un modèle linéaire mixte est un modèle pour lequel le modèle comprend à la fois des effets fixes et des effets aléatoires. Les MLM incluent des variables à effets fixes et aléatoires. Le mélange entre les deux types de facteurs dans un même modèle est à l'origine du nom. Les effets fixes décrivent les relations entre les covariables et la variables dépendante pour une population entière, les effets aléatoires sont spécifiques à l'échantillon.
En d'autres termes,
- *un effet aléatoire* est effet dont nous ne voulons pas généraliser les propriétés (les modalités ont été choisies de manière aléatoire dans quelque chose de plus grand).
-*un effet fixe* est un effet dont on veut généraliser les propriétés. Il s'agit de la variable manipulée. Les niveaux de cette variable ont été choisi de manière spécifique.
Il est important de comprendre qu'une variable peut être considérée comme un effet fixe ou un effet aléatoire en fonction de l'hypothèse qui va être testée.
IMPORTANT 1 : les effets aléatoires doivent nécessairement être des variables catégorielles. L'utilisation de nombres peut créer de l'ambiguïté. Il est donc préférable de s'assurer que le format de la variable aléatoire soit une variable catégorielle.
IMPORTANT 2 : six modalités de cette variable catégorielle semble être un minimum pour pouvoir estimer de manière pertinente une variance.

## les facteurs fixes :

Les facteurs fixes sont les variable pour lesquelles tous les niveaux (i.e., toutes les conditions) qui sont d'intérêt sont inclus dans l'étude. Les variables peuvent être qualitatives comme le genre, ou des variables d'échantillonnage de l'étude comme la région. Les niveaux sont choisis de sorte à représenter des conditions spécifiques qui peuvent être utilisées pour définir des contrastes.

## les facteurs aléatoires :

Les effets aléatoires sont échantillonnés de manière aléatoire dans l'ensemble des niveaux possibles de la population étudiée. On n'a pas tous les niveaux de la population, mais l'objectif est de pouvoir inférer les propriétés de l'échantillon à une population entière.

Contrairement aux facteurs fixes, les niveaux de la variables aléatoire ne représentent pas des conditions choisies de manière spécifique pour répondre aux objectifs de l'étude.

## les effets fixes vs les effets aléatoires :

Les effets fixes sont représentés par des coefficients de régression. Ces effets décrivent les relations entre la variable dépendante et les prédicteurs (des facteurs fixes ou des covariables continues). Ces effets permet d'identifier, par exemple, les différences de moyennes entre des groupes sur la VD pour les facteurs fixes, ou le lien entre une covariable continue et la variable dépendante. On fait l'hyptohèse que les effets fixes sont inconnus et que nous les estimons sur la base des données récoltées.

Les effets aléatoire sont spécifique à un niveau donné du facteur aléatoire. Ces effets représentent une déviation de la relation décrite par les effets fixes.

# 1 MODÈLES LINÉAIRES GÉNÉRALISÉS MIXTES :

Dans un modèle linéaire mixte normal, on pose :

$$\mathbf{y} = \mu + \varepsilon = \mathbf{X}\beta + \mathbf{Zu} + \varepsilon$$

où $\mu = E(y \mid u) = X\beta + Zu$ est la moyenne conditionnelle de $y$ sachant $u$, $Z$ est une matrice n×q d'incidence des effets aléatoires $u$
On suppose que $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{G})$ où $0$ est le vecteur nul de dimension q×1, $G$ est la matrice de variance-covariance de $u$. On suppose que $u$ et  sont indépendants. Les autres symboles sont comme dans le modèle linéaire généralisé où la fonction de lien serait la fonction identité. La variance de $y$ conditionnellement à $u$ est $Var(\mathbf{y} \mid \mathbf{u}) = \mathbf{R} = Var(\varepsilon)$, cette dernière pouvant prendre une forme bloc-diagonale si la variable aléatoire Y est mesurée à plusieurs reprises sur un même sujet.

Les modèles linéaires généralisés mixtes sont une généralisation à la fois des modèles linéaires généralisés et des modèles linéaires mixtes. Dans un tel modèle, on suppose comme dans ces derniers que :

$$\mathbf{y} = \mu + \varepsilon$$

et que $g(\mu) = X\beta + Zu$ où, comme dans un modèle linéaire généralisé, $\mathrm{g}(.)$ est une fonction de lien monotone, et comme dans un modèle linéaire mixte, $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{G})$. $\mu$ est la moyenne condtionnelle, $\mathrm{E}(\mathbf{y} \mid \mathbf{u})$ . On a encore $Var(\varepsilon) = \mathbf{R}$ .

# 2 L'anova vs le MLM :

Les modèles linéaires mixtes sont des alternatives au modèle linéaire, et en particulier aux anovas. Tout comme les anovas, les modèles linéaires mixtes permettent de déterminer si des variables catégorielles ont un impact sur une variable dépendante continue.

Tout comme les anovas, le résidu doit être distribué en suivant un distribution normale.

Les prédicteurs peuvent être de nature catégorielle, ordinale ou continue.

Le modèle linéaire classique est défini de la manière suivante :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Où Y représente les mesures observées sur la variable dépendante, 0 est l'intercept et 1 est le coefficient de régression. Dans ce cas,  représente le résidu, c'est-à-dire ce que nous ne sommes pas en mesure de modéliser. Si nous nous intéressons à un individu i, la prédiction pour cet individu i est

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

Cela signifie que, pour un individu i, la valeur qu'on peut prédire pour la variable dépendante correspond à l'intercept auquel on ajoute la valeur de l'individu i pour la variable indépendante multiplié

par le coefficient de régression. Entre cette valeur prédite de manière théorique et la valeur réellement obserbée, il y a une différence. La différence entre la valeur prédite et la valeur observée est le résidu et ce résidu n'est pas modélisé de manière explicite dans le modèle linéaire.

## 2.1 Le MLM vs l'anova : les différences

En plus du résidu, les variables aléatoires doivent aussi suivre une distribution multinormale.

En revanche :

*les variances peuvent être hétérogènes (et on peut le modéliser).

*Les résidus peuvent ne pas être indépendants (il peut y avoir des corrélations sur les résidus).

*Il est possible d'ajuster le modèle avec une large sélection de structure de covariance parcimonieuse (ce qui est plus efficace que d'estimer la structure complète de variance-covariance).

*On peut modéliser la distribution de la variable dépendante (qui n'est pas nécessairement continue).

*La sphéricité de la matrice de covariance n'est pas requise pour les modèles ayant des facteurs en mesure répétée. on peut suivre des individus dans le temps même si le nombre de mesures est différent (on n'enlève pas tout l'individu parce qu'on a une valeur manquante). Dans le modèle linéaire mixte, toutes les observations disponibles pour un individu sont utilisées dans l'analyse.

*lorsqu'on utilise une anova, les niveaux pour le facteur temps doivent être les mêmes. Dans le modèle linéaire mixte, les mesures de temps peuvent varier entre les participants. En revanche, il faut tout de même que la plupart des informations soit disponible à chaque mesure de temps.

*Le plan ne doit pas être équilibré, même pour les facteurs à mesure répétée. En d'autres termes, il peut y avoir des valeurs manquantes sur les facteurs à mesure répétée sans perdre des participants.

*On peut modéliser des plans expérimentaux pour lesquels certains facteurs sont partiellement emboîtés.

*En résumé, toutes ces différences entre le MLM et l'ANOVA permet de comprendre que les MLM sont bien plus flexibles que les anovas.

## 2.2 Les covariables :

*Les modèles mixtes permettent d'avoir à la fois des covariables au niveau individuel (tel que l'âge ou le sexe) et au niveau du cluster (telle que la taille du cluster) lorsqu'on ajuste l'effet aléatoire.

*On peut avoir des covariables qui varient en fonction du temps dans le modèle.

**Partie pratique :**

Pour rendre cela plus clair et concret, considérons un exemple d'un jeu de données simulé, constutué de 407 Médecins, de 8525 patient et de 6 prédicteurs ou variables explicatives à effets fixes : âge(en années), marié(0=non,1=oui), sexe(0=femme,1 :homme), nombre de globules rouge (GR) et globules blancs (WBC). La représentation matricielle de cet exemple est la suivante :

$$\overbrace{\mathbf{y}}^{8525\times1} = \underbrace{\mathbf{X}}_{8525\times6}\underbrace{\beta}_{6\times1} + \underbrace{\mathbf{Z}}_{8525\times407}\underbrace{u}_{407\times1} + \overbrace{\varepsilon}^{8525\times1}$$

$$\mathbf{y} = \begin{bmatrix} mobility \\ 2 \\ 2 \\ \dots \\ 3 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} Intercept & Age & Married & Sex & WBC & RBC \\ 1 & 64.97 & 0 & 1 & 6087 & 4.87 \\ 1 & 53.92 & 0 & 0 & 6700 & 4.68 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 56.07 & 0 & 1 & 6430 & 4.73 \end{bmatrix}$$

$$\beta = \begin{bmatrix} 4.782 \\ .025 \\ .011 \\ .012 \\ 0 \\ -.009 \end{bmatrix}$$

On simule les données aléatoirement sur python et on obtient :

| | Intercept | Age | Married | sex | WBC | RBC | y |
|---|---|---|---|---|---|---|---|
| **0** | 1 | 46.0 | 1 | 0 | 5991.528980 | 5.753987 | 2 |
| **1** | 1 | 62.0 | 1 | 0 | 5994.072082 | 3.175610 | 3 |
| **2** | 1 | 62.0 | 0 | 1 | 6018.001567 | 6.613776 | 2 |
| **3** | 1 | 53.0 | 0 | 0 | 5990.948873 | 4.675376 | 3 |
| **4** | 1 | 47.0 | 1 | 1 | 5985.482100 | 5.478123 | 2 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **8520** | 1 | 58.0 | 1 | 1 | 5988.672529 | 5.861227 | 3 |
| **8521** | 1 | 50.0 | 1 | 1 | 5993.302204 | 5.698575 | 2 |
| **8522** | 1 | 66.0 | 1 | 0 | 5979.884308 | 7.333334 | 3 |
| **8523** | 1 | 44.0 | 1 | 0 | 6006.765988 | 10.734307 | 3 |
| **8524** | 1 | 67.0 | 1 | 1 | 5999.967000 | 4.500191 | 2 |

8525 rows × 7 columns

FIGURE 1 –
Tableau des données simulées

Informations sur les données statistiques du tableau :

| | Intercept | Age | Married | sex | WBC | RBC | y |
|---|---|---|---|---|---|---|---|
| count | 8525.0 | 8525.000000 | 8525.000000 | 8525.000000 | 8525.000000 | 8525.000000 | 8525.000000 |
| mean | 1.0 | 54.189091 | 0.493372 | 0.499472 | 6000.635660 | 5.123284 | 2.498065 |
| std | 0.0 | 11.013202 | 0.499985 | 0.500029 | 8.987584 | 2.017213 | 0.500026 |
| min | 1.0 | 14.000000 | 0.000000 | 0.000000 | 5965.000000 | -2.000000 | 2.000000 |
| 25% | 1.0 | 47.000000 | 0.000000 | 0.000000 | 5995.000000 | 4.000000 | 2.000000 |
| 50% | 1.0 | 54.000000 | 0.000000 | 0.000000 | 6001.000000 | 5.000000 | 2.000000 |
| 75% | 1.0 | 62.000000 | 1.000000 | 1.000000 | 6007.000000 | 6.000000 | 3.000000 |
| max | 1.0 | 96.000000 | 1.000000 | 1.000000 | 6042.000000 | 13.000000 | 3.000000 |

FIGURE 2 –
Résumé statistique

GLM :

```
==============================================================================
Dep. Variable:                      y   No. Observations:                 8525
Model:                            GLM   Df Residuals:                     8519
Model Family:                 Poisson   Df Model:                            5
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -12373.
Date:                Sun, 08 Nov 2020   Deviance:                       858.05
Time:                        12:38:58   Pearson chi2:                     853.
No. Iterations:                     4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.5515      4.577      0.120      0.904      -8.420       9.523
sex            0.0029      0.014      0.208      0.835      -0.024       0.030
Age            0.0003      0.001      0.456      0.649      -0.001       0.002
Married        0.0048      0.014      0.348      0.728      -0.022       0.032
WBC         5.814e-05      0.001      0.076      0.939      -0.001       0.002
RBC           -0.0009      0.003     -0.253      0.800      -0.008       0.006
==============================================================================
```

FIGURE 3 –
Régression linéaire du modèle généralisé suivant une loi de poisson

MLM :

```
============================================================
Model:               MixedLM  Dependent Variable:  y
No. Observations:    8525     Method:              REML
No. Groups:          1        Scale:               0.2500
Min. group size:     8525     Log-Likelihood:      -6222.4240
Max. group size:     8525     Converged:           Yes
Mean group size:     8525.0
------------------------------------------------------------
                Coef.    Std.Err.     z    P>|z| [0.025 0.975]
------------------------------------------------------------
Intercept        2.444       0.503  4.860 0.000  1.458  3.430
sex              0.059       0.077  0.772 0.440 -0.091  0.210
Married         -0.048       0.077 -0.622 0.534 -0.199  0.103
sex:Married      0.038       0.109  0.348 0.728 -0.175  0.251
Age              0.001       0.001  0.787 0.431 -0.001  0.003
sex:Age         -0.001       0.001 -0.599 0.549 -0.004  0.002
Married:Age      0.001       0.001  0.885 0.376 -0.001  0.004
sex:Married:Age -0.001       0.002 -0.486 0.627 -0.005  0.003
Group Var        0.250 3587403.181
============================================================
```

FIGURE 4 –
Régression linéaire du modèle linéaire mixte

**Conclusion :**

Un modèle mixte est un modèle statistique qui comporte à la fois des effets fixes et des effets aléatoires. Ce type de modèle est utile dans une grande variété de domaines, tels que la physique, la biologie ou encore les sciences sociales. Les modèles mixtes sont particulièrement utiles dans les situations où des mesures répétées sont effectuées sur les mêmes variables . Ils sont souvent préférés à d'autres approches telle que l'ANOVA, dans le mesure où ils peuvent être utilisés dans le cas où le jeu de données présente des valeurs manquantes.