Modèles Linéaires (LM)

Définition

Les **modèles linéaires** (LM) englobent :

- **les modèles de régression linéaire**

  $\hookrightarrow$ les covariables sont des régresseurs : des variables quantitatives

  Ex : On explique la teneur en ozone par la température, la nébulosité

- **les modèles d'analyse de la variance** (ANOVA)

  $\hookrightarrow$ les covariables sont des facteurs : des variables qualitatives

  Ex : On explique la teneur en ozone par la direction du vent ou la saison

$\Longrightarrow$ Tout dépend donc de **la matrice X des covariables ...**
... mais Y n'a pas changé : on cherche à expliquer une mesure numérique !

$\Longrightarrow$ On peut aussi mélanger les 2 natures de covariables : **les modèles d'analyse de covariance** (ANCOVA)

Modèle Linéaire || Modèle Linéaire Général ||
Modèle Linéaire Généralisé

- **Régresseur** : valeur numérique $\Rightarrow$ une infinité de valeurs possibles, souvent toutes différentes dans l'échantillon

  $\hookrightarrow$ une valeur de Y différente pour chaque valeur de X

- **Facteur** : niveau $\Rightarrow$ un nombre fini, faible, de niveaux possibles

  $\hookrightarrow$ répétition des valeurs observées de Y dans chaque niveau

Écriture du modèle

- **Modèle de régression**

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$\hookrightarrow$ on cherche la part de variabilité de Y induite par les changements de valeurs du régresseur x

$\hookrightarrow$ 1 paramètre **par régresseur**

$\hookrightarrow$ signification des paramètres :
β_0 : intercept - niveau moyen de y quand x vaut 0
β_1 : pente - impact d'un accroissement de 1 de x sur la variable réponse

$\hookrightarrow$ forme de X : ...

Modèle d'ANOVA

Puisque plusieurs observations de Y pour chaque niveau de x, on note $y_{j,k}$: la réponse de l'individu k dans le niveau j.

$$Y_{j,k} = \beta_0 + \beta_j + \varepsilon_{j,k}$$

$\hookrightarrow$ on cherche la part de variabilité de Y induite par les différents niveaux du facteur x

$\hookrightarrow$ 1 paramètre **par niveau du facteur**

$\hookrightarrow$ signification des paramètres :
β_0 : intercept - valeur moyenne des valeurs y observées
β_j : impact du niveau j de x sur la variable réponse

$\hookrightarrow$ forme de X : ...

Les contraintes dans les modèles d'ANOVA

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

$\hookrightarrow$ pour que $X'X$ soit inversible, il faut que X soit de plein rang (colonne) :

$\hookrightarrow$ la dimension de l'espace vectoriel engendré par les colonnes de X est égal à p (le nombre de colonnes de X)

- **Pas d'intercept**

$$\beta_0 = 0$$

$\hookrightarrow$ $\forall j \in \{1, .., J\}$ β_j : niveau moyen de la réponse dans le niveau j

$\hookrightarrow$ forme de X : ...

- **Choix d'une référence** : un niveau du facteur sert de référence

$$\beta_1 = 0$$

$\hookrightarrow$ le niveau 1 sert de référence

$\hookrightarrow$ β_0 niveau moyen de la réponse dans le niveau 1

$\hookrightarrow$ $\forall j \in \{2, .., J\}$ β_j : écart de la réponse du niveau j au niveau 1

$\hookrightarrow$ forme de X : ...

$\hookrightarrow$ Dans R : contr.treatment

- **Somme des paramètres**

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

$\hookrightarrow$ β_0 moyenne des paramètres

$\hookrightarrow$ $\forall j \in \{1, .., J-1\}$ $\beta_0 + \beta_j$: niveau moyen de la réponse du niveau j

$\hookrightarrow$ $\beta_0 - \sum_{j=1}^{J-1} \beta_j$: niveau moyen de la réponse du niveau J

$\hookrightarrow$ Dans R : contr.sum

ANOVA à 2 facteurs

- Considérons 2 facteurs F_1 (à J_1 modalités) et F_2 (à J_2 modalités).
  $\hookrightarrow$ plusieurs observations de Y à chaque croisement d'une modalité de F_1 et d'une modalité de F_2
- y_{ijk} : observation de l'individu numéro k au croisement du niveau i de F_1 et le niveau j de F_2
- **Modèle** :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

$\hookrightarrow$ on cherche la part de variabilité de Y induite par chacun des niveaux du facteur F_1 d'une part, et chacun des niveaux du facteur F_2 d'autre part

$\hookrightarrow$ 1+J_1+J_2 paramètres

$\hookrightarrow$ signification des paramètres :
β_0 : intercept - valeur moyenne des valeurs y observées
α_i : impact du niveau i de F_1 sur la variable réponse
β_j : impact du niveau j de F_2 sur la variable réponse

$\hookrightarrow$ forme de X : ...

- **Contraintes**

  $\hookrightarrow$ Choix d'une **référence** : niveau 1 de F_1 ($\alpha_1 = 0$) et niveau 1 de F_2 ($\beta_1 = 0$)

  $\hookrightarrow \forall i \in \{2, .., J_1\} \quad \alpha_i$: écart de la réponse du niveau i au niveau 1 de F_1
  $\hookrightarrow \forall j \in \{2, .., J_2\} \quad \beta_j$: écart de la réponse du niveau j au niveau 1 de F_2

- **Additivité des effets principaux**

  $\hookrightarrow \alpha_i$: effet du niveau i de F_1 quelque soit le niveau j de F_2
  $\hookrightarrow \beta_j$: effet du niveau j de F_2 quelque soit le niveau i de F_1

- **Interaction**

  $\hookrightarrow$ On ajoute un effet du niveau i en interaction avec le niveau j : γ_{ij}

  $\hookrightarrow$ Le niveau i de F_1 a un effet différent selon le niveau j de F_2 et réciproquement

  $\hookrightarrow$ Ex : l'orientation du vent n'a pas le même effet avec ou sans pluie

- **Modèle**

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

$\hookrightarrow$ Forme de X : ...

$\hookrightarrow$ Dans R : l'interaction est désignée par $F_1 : F_2$

$\Rightarrow$ Le modèle avec interaction : $Y \sim F_1 + F_2 + F_1 : F_2$ ou $Y \sim F_1 * F_2$

- **Contraintes**

$$\forall i \in \{1, .., J_2\} \quad \gamma_{i1} = 0$$
$$\forall j \in \{1, .., J_2\} \quad \gamma_{1j} = 0$$

Modèle d'ANCOVA

La partie explicative du modèle mélange : régresseurs et facteurs.

- X : le régresseur
- F : le facteur

- **Modèle sans interaction** : $Y \sim X + F$

$$Y_{jk} = \mu + \beta_j + \delta x_{jk} + \varepsilon_{jk}$$

$\hookrightarrow$ forme de X : ...

$\hookrightarrow$ l'effet du facteur n'a un impact que sur l'intercept

- **Modèle avec interaction** : $Y \sim X * F$

$$Y_{jk} = \mu + \beta_j + (\delta + \delta_j)x_{jk} + \varepsilon_{jk}$$

$\hookrightarrow$ forme de X : ...

$\hookrightarrow$ l'effet du facteur a un impact sur l'intercept et sur la pente