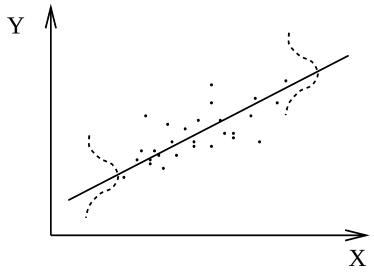
Module 10 :

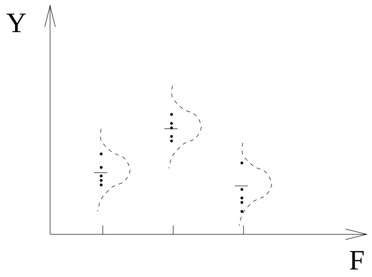
Modèle linéaire généralisé

Voici le plan du module.

Un exemple est d'abord présenté, il montre la manière dont on peut utiliser la méthodologie du modèle linéaire généralisé, et justifie la nécessité d'une telle méthodologie. Des éléments de théorie sont ensuite présentés, dans le cadre de données binomiales. Des outils de validation de modèle sont proposés, ainsi qu'une stratégie de choix de modèle. Le modèle multinomial est abordé ainsi que le cas d'autres types de lois.

Le modèle linéaire - Rappel





Y : V.A. quantitative continue

Postulats :

-
-
-
-

- Deux exemples de modèle linéaire :
  - une régression linéaire simple, $Y_i = \alpha + \beta X_i + \varepsilon_i$,
  - une ANOVA à un facteur, $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$.
    Ces deux exemples relèvent du modèle linéaire.
- Discuter des **postulats** sur ces deux exemples :
  - **adéquation du choix de la partie systématique du modèle**, $E(\varepsilon) = 0$,
  - **indépendance** entre toutes les observations,
  - **homogénéité des variances**,
  - **normalité des erreurs**.
- Retrouver $Y = X\theta + \varepsilon$, pour chacun des deux exemples.

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ y_{31} \\ \vdots \\ y_{3n_3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_{3n_3} \end{pmatrix}$$

- On peut généraliser à l'explication d'une V.A.[1] Y par un ensemble de facteurs et de régresseurs (facteurs et régresseurs seront appelés variables explicatives, sans distinction).

1. variable aléatoire

Le modèle linéaire - Rappel

Données : graphiques

↓

Modèle linéaire

Y = X Θ + ε → anova / régression

↓

loi : normale

↓

Estimation : moindres carrés

$$\widehat{\theta} = (X'X)^{-1}X'Y$$

$$\widehat{\theta} \underset{(postulats)}{\sim} \boxed{\mathcal{N}(\theta, \sigma^2 (X'X)^{-1})}$$

$$\widehat{Y} = X\widehat{\theta} = \underbrace{X(X'X)^{-1}X'}_{H}Y$$

Dans le schéma d'analyse, la première étape consiste à visualiser les données. Une étude graphique est toujours riche d'enseignements, et nécessaire avant une étape de modélisation.

Le modèle linéaire regroupe dans un même formalisme analyse de variance, régression, et des modèles mixant facteurs et régresseurs dans la partie systématique du modèle, comme l'analyse de covariance.

La **loi des erreurs** ε est supposée **normale** (cela implique la normalité de la variable aléatoire à expliquer Y, pour chaque traitement, c'est à dire pour chaque combinaison de niveaux de facteurs, dans le cas d'une ANOVA). Y est donc une variable quantitative prenant ses valeurs dans l'ensemble des nombres réels. Les **erreurs** ε sont supposées de variance constante.

Le vecteur θ des paramètres est estimé par le **critère des moindres carrés**. $\widehat{\theta}$ minimise $\sum(y_i - \widehat{y_i})^2$. $\widehat{\theta}$ s'obtient directement, par simple calcul matriciel. Les postulats assurent à $\widehat{\boldsymbol{\theta}}$ de bonnes propriétés : **normalité, estimateur sans biais, de variance minimum**. Ces propriétés permettent de faire des tests. (σ^2 est la vraie valeur inconnue de la variance, qui est supposée constante).

La matrice H est un projecteur, elle envoie Y sur $\widehat{Y}$, elle s'appelle "matrice chapeau" ou "hat matrix" . Elle rappelle que l'estimation d'un modèle linéaire se fait par projection orthogonale. Elle est utile dans les méthodes de diagnostic, car on trouve sur sa diagonale, le vecteur "leverage" permettant de détecter des observations particulièrement influentes, car ayant des valeurs prises par les variables explicatives très écartées des autres.

Rappeler que le schéma d'analyse présenté ici est incomplet. La phase d'estimation terminée, l'analyse n'est pas terminée, il manque les phases de présentation numérique et graphique des résultats, et la phase de validation du modèle, qui peut, si le modèle étudié est jugé inadéquat, nous ramener à la case "modèle".

Le modèle linéaire - Rappel

$$Y = \underbrace{X\theta}_{\text{partie systématique}} + \underbrace{\varepsilon}_{\text{partie aléatoire}}$$

ANOVA :

source	d.l.	SC	CM	F	pvalue
Résiduelle			$\widehat{\sigma}^2$		

On appelle $X\theta$ partie systématique du modèle (ou encore partie déterministe) et ε partie aléatoire ou résiduelle. Les termes de la partie systématique du modèle sont les facteurs, régresseurs et termes d'interaction portés en colonne dans la matrice d'incidence X.

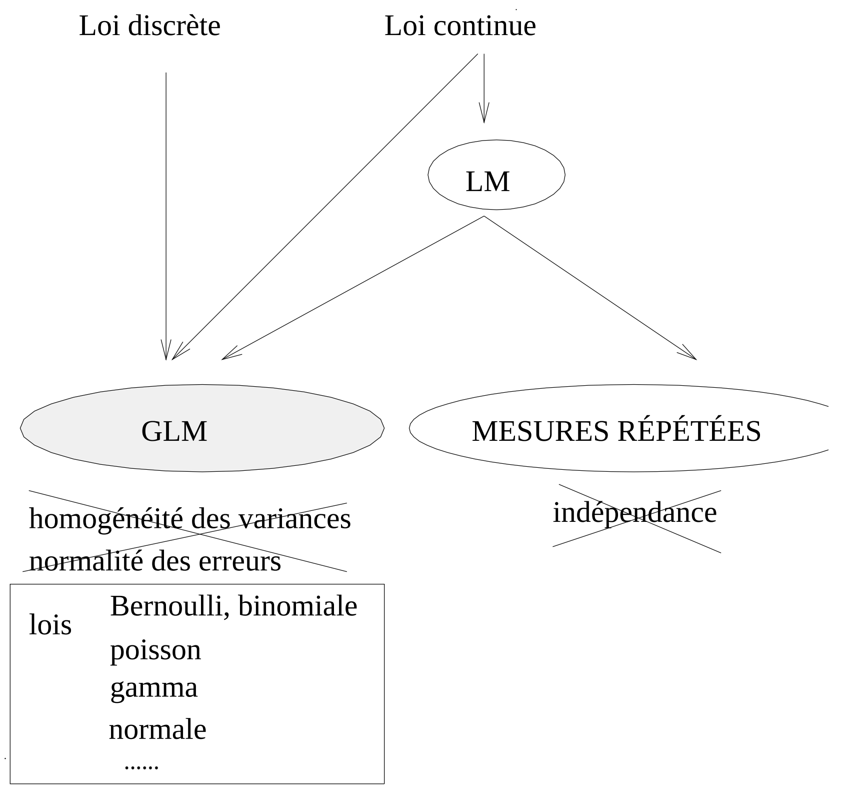
Les résultats sont souvent présentés dans une table d'ANOVA. Les **sources de variation** sont listées, ce sont les termes du modèle. A chaque source de variation, sont associés un nombre de degrés de libertés, une somme de carrés d'écarts, un carré moyen, une statistique de Fisher, et la p-value associée.

La colonne SC contient des **sommes de carrés d'écarts** expliquées par les différentes sources de variation. Par exemple, dans le cas d'une analyse de variance à un facteur, la somme de carrés d'écarts associée à ce facteur est la somme des carrés des écarts entre les moyennes de chacun des niveaux du facteur, et la moyenne générale (pondérée par le nombre d'observations ayant la modalité). La **somme de carrés résiduelle** est la somme des carrés des écarts entre chaque observation et la valeur estimée correspondante.

Le **carré moyen** est égal à la somme de carrés par unité de degrés de liberté. Le **carré moyen résiduel** est l'estimation de la variance résiduelle $\widehat{\sigma}^2$. Les **carrés moyens suivent des lois de $\boldsymbol{\chi^2}$ indépendantes**. La **statistique de Fisher** est le rapport entre le carré moyen associé au terme et le carré moyen résiduel. C'est le rapport de deux statistiques indépendantes (si le plan est orthogonal), qui suivent des lois de χ^2. Sous l'hypothèse nulle d'absence d'effet du terme testé, il suit une loi de Fisher. L'effet est déclaré significatif si la statistique calculée est supérieure au $(1-\alpha)$ quantile de la loi de Fisher dont les paramètres sont le degré de liberté associé au terme, et le degré de liberté résiduel. Un effet est déclaré significatif, si la variabilité associée à cet effet est significativement plus grande que la variabilité "de base", c'est à dire la variabilité résiduelle $\widehat{\sigma}^2$, celle qui ne peut être imputée à aucun facteur de variation introduit dans le modèle. La p-value affichée correspond à la probabilité de déclarer à tort un effet significatif. Plus cette p-value est faible, plus la preuve de l'effet est forte.

Le modèle linéaire - Rappel

Les limites du modèle linéaire?



Le modèle linéaire suppose vérifiés 4 postulats : le bon choix de la partie systématique du modèle, l'indépendance entre les observations, l'homogénéité des variances, la normalité des erreurs. Il suppose aussi, mais c'est implicite dans le postulat de normalité, que la variable à expliquer est quantitative continue et prend ses valeurs sur tout l'ensemble des nombres réels (si l'on imposait des contraintes sur $X\theta$, on ne serait plus dans le cadre du modèle linéaire).

Le **GLM** est une extension du LM (modèle linéaire), qui **permet de lever les postulats d'homogénéité des variances**, de **normalité des erreurs**, de prendre en compte des **contraintes sur le domaine de variation de l'espérance** de l'observation. L'indépendance entre les observations reste nécessaire. Ce sont les méthodes qui relèvent des données répétées qui permettent de lever cette hypothèse d'indépendance.

Les limites du modèle linéaire?
C'est l'occasion de discuter avec les stagiaires de leurs propres exemples. Il s'agit d'identifier des exemples où les postulats d'homogénéité des variances et de normalité sont des obstacles (pas forcément insurmontables) à l'utilisation du modèle linéaire.

→ de nombreux exemples où la variable observée initialement est une variable binaire (données 0/1, proportions).
→ variable à expliquer discrète, qualitative, ordonnée ou non ordonnée.
→ hétérogénéité des variances.