Chapitre 5. **Resumé du cours**

Cours de modèle linéaire gaussien par S. Donnet

Executive Master Statistique et Big-Data

Août 2020

Contexte

- On cherche à comprendre la **variabilité d'une variable d'intérêt** y **en fonction de covariables** qualitatives / et ou quantitatives.

Contexte

- On cherche à comprendre la **variabilité d'une variable d'intérêt** y **en fonction de covariables** qualitatives / et ou quantitatives.
- y est toujours **quantitative**.

Contexte

- On cherche à comprendre la **variabilité d'une variable d'intérêt** y **en fonction de covariables** qualitatives / et ou quantitatives.
- y est toujours **quantitative**.
- Pour apprendre la relation entre y et ces variables, on dispose d'observations de la variable y : $y_1, \ldots, y_n$

Contexte

- On cherche à comprendre la **variabilité d'une variable d'intérêt** y **en fonction de covariables** qualitatives / et ou quantitatives.
- y est toujours **quantitative**.
- Pour apprendre la relation entre y et ces variables, on dispose d'observations de la variable y : $y_1, \ldots, y_n$
- Pour chaque observation i ; on observe aussi les covariables / facteurs.

Modélisation

Dans tous les cas, on **fait l'hypothèse que** les $(y_i)_i$ sont la réalisation de variables aléatoires indépendantes (Y_i) telles que

$$Y_i = f(\text{covariables}_i, \text{facteurs}_i; \beta) + \varepsilon_i$$

avec

- $\varepsilon_i \sim_{i.i.d;} \mathcal{N}(0, \sigma^2)$
- $\beta \in \mathbb{R}^d$
- **Cas particulier** modèle linéaire (en les paramètres).

$$f(\text{covariables}_i, \text{facteurs}_i; \beta) = \sum_{k=1}^{d} \beta_k x_i^k$$

Estimation de β et σ^2

- On cherche β de façon à ajuster au mieux le modèle à nos observations.
- Ajustement mesuré par perte L^2 (**estimateur des moindres carrés**).

$$\boxed{\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - f(\text{covariables}_i, \text{facteurs}_i; \beta))^2}$$

- σ^2 mesure la variance des erreurs.

$$\boxed{\hat{\sigma}^2 = \frac{1}{n - \bullet} \sum_{i=1}^{n} \left(Y_i - f(\text{covariables}_i, \text{facteurs}_i; \hat{\beta})\right)^2}$$

Régression : covariables quantitatives

- x_i^j : j-ième covariable pour l'observation i. p covariables
- $$\begin{cases} Y_i & = & \sum_{j=1}^{p} \beta_j x_i^j + \varepsilon_i \\ \varepsilon_i & \sim & {}_{i.i.d.}\mathcal{N}(0, \sigma^2) \end{cases}$$
- En général $x_i^1 = 1$ pour tout i : premier paramètre intercepte.
- X matrice de taille $n \times p$: $X_{ij} = x_i^j$, $Y = (Y_1, \dots, Y_n)'$
- $$\hat{\beta} = (X'X)^{-1}X'Y$$
- $x_i = (x_i^1, \dots, x_i^p)$.

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - x_i\hat{\beta})^2$$

Anova (I)

Anova à un facteur

- Le facteur a I modalités
- Re-numérotation des données : y_{ij} j-ième observation dans la modalité i.
- $$\begin{cases} Y_{ij} & = & \mu_i + \varepsilon_{ij} & \text{mod. régulier} \\ & = & \mu + \alpha_i + \varepsilon_{ij} & \text{mod. singulier} \\ \varepsilon_{ij} & \sim & i.i.d. \mathcal{N}(0, \sigma^2) & \end{cases}$$
- $\widehat{\mu_i} = \overline{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$
- Estimation des $(\mu, \alpha_1, \ldots, \alpha_I)$ sous contrainte. Sous **R**,

$$\alpha_1 = 0 \quad \Rightarrow \quad \mu = \mu_1, \alpha_i = \mu_i - \mu.$$

- $$\hat{\sigma}^2 = \frac{1}{n - I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2$$

Anova (II)

Anova à deux facteurs

- Re-numérotation des données : y_{ijk} k-ième observation dans la modalité i du 1er facteur et j du 2ème facteur.
- $$\left\{\begin{array}{lcll} Y_{ijk} & = & \mu_{ij} + \varepsilon_{ijk} & \text{mod. régulier} \\ & = & \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} & \text{mod. singulier} \\ \varepsilon_{ijk} & \sim & {}_{i.i.d.}\mathcal{N}(0, \sigma^2) & \end{array}\right.$$
- $\widehat{\mu_{ij}} = \overline{Y}_{ij\bullet} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk}$
- Estimation des $(\mu, \alpha_i, \beta_j, \gamma_{ij})$ sous contraintes. Sous **R**,
  $$\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0$$
- $$\hat{\sigma}^2 = \frac{1}{n - IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y}_{ij\bullet})^2$$

Ancova

Variables qualitatives et quantitatives.

Ancova à 1 facteur et une covariable

- Re-numérotation des données : y_{ij} j-ième observation dans la modalité i du facteur.
- $$\left\{ \begin{array}{ccccll} Y_{ij} & = & \mu + \beta_i & + & \alpha x_{ij} + \gamma_i x_{ij} & +\varepsilon_{ij} \quad \text{mod. singulier} \\ & = & b_i & + & a_i x_{ij} & +\varepsilon_{ij} \quad \text{mod. régulier} \\ \varepsilon_{ij} & \sim & i.i.d. & & \mathcal{N}(0, \sigma^2) & \end{array} \right.$$
- $\begin{pmatrix} \hat{b}_i \\ \hat{a}_i \end{pmatrix}$: EMC de la régression linéaire simple pour chaque modalité.
- Estimation des $(\mu, \alpha, \beta_i, \gamma_i)$ sous contraintes. Sous **R**, $\beta_1 = \gamma_1 = 0$
- $$\hat{\sigma}^2 = \frac{1}{n - 2I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \hat{a}_i x_{ij} - \hat{b}_i)^2$$

Validation des postulats

Avant toute inférence, vérification que le postulat “les observations sont les réalisations du modèle linéaire gaussien” est raisonnable

- Lecture des graphes des résidus
- Détection d'un oubli de tendance dans le modèle
- Détection d'une hétéroscédasticité
- Détection de points aberrants

Si postulats semblent vérifiés, on continue l'analyse

Intervalles de confiances / tests sur les paramètres

- Pour tous les modèles, on peut construire des IC sur les paramètres β_j...
  Du type :

$$\left[\widehat{\beta}_j - q\hat{\sigma}_{\widehat{\beta}_j}, \widehat{\beta}_j + q\hat{\sigma}_{\widehat{\beta}_j}\right]$$

  où
  - q est le quantile d'une loi de Student à $n - \{\text{nb paramètres}\}$
  - $\hat{\sigma}_{\widehat{\beta}_j}$ est l'écart type estimé de $\hat{\beta}_j$ (σ ayant été remplacé par $\hat{\sigma}$).
- De la même façon on construit des tests de nullité d'un paramètre.

Tests d'un sous modèle

Définition d'un sous modèle tel que $[X^{(0)}] \subset [X]$

- **Exemple 1** (anova) : $\alpha_2 = \alpha_3, \cdots = \alpha_I = 0$. Pas d'effet du facteur
- **Exemple 2** (ancova) : $\gamma_2 = \gamma_3, \cdots = \gamma_I = 0$. Pas d'interaction entre le facteur et la covariable.

On définit SCR la somme des carrés résiduelle $\sum_{\bullet=1}^{\cdots}(y_\bullet - \hat{y}_\bullet)^2$

$$F = \frac{(\text{SCR}_0 - \text{SCR })/(d - d_0)}{\text{SCR }/(n-d)}$$

Alors, sous l'hypothèse que les données sont réalisations du modèle le plus simple, on a

$$F \sim \mathcal{F}(d - d_0, n - d)$$

On rejette cette hypothèse si $f > q^{1-\alpha}_{\mathcal{F}(d-d_{\mathbf{0}},n-d)}$.