Université de Strasbourg
M1 - Magistère
Statistique

Ségolen Geffray
geffray@math.unistra.fr
Année 2019/2020

Sujet 9

• **Brefs rappels sur la régression linéaire**:
Lorsque $\left(X^{(1)},...,X^{(p)}\right)$ est un vecteur de covariables quantitatives, le modèle de régression linéaire correspondant à l'observation de $\left(Y,X^{(1)},...,X^{(p)}\right)$ s'écrit:

$$Y = \beta_0 + \sum_{k=1}^{p} \beta_k X^{(k)} + \varepsilon \tag{1}$$

où $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Var}(\varepsilon) < \infty$ et $\varepsilon \perp\!\!\!\perp \left(X^{(1)},...,X^{(p)}\right)$.
Le modèle de régression linéaire est dit gaussien lorsqu'on rajoute l'hypothèse $\varepsilon \sim \mathcal{N}(0,\sigma^2)$.
Le modèle de régression linéaire est dit homoscédastique lorsqu'on rajoute l'hypothèse que la variance de l'erreur résiduelle est constante d'un individu à l'autre.

• **Résidus du modèle de régression linéaire**:
Notons $\beta \in \mathbb{R}^{p+1}$ le vecteur des paramètres de régression à estimer dans le cadre d'un modèle de régression linéaire et $\widehat{\beta}$ son estimateur. Afin d'effectuer des diagnostics sur l'ajustement du modèle aux données, on utilise les résidus du modèle. On utilise souvent en pratique les résidus studentisés:

$$\widehat{\varepsilon}_i^{\text{stud}} := \frac{Y_i - \mathbb{X}_i.\widehat{\beta}}{\sqrt{\widehat{\sigma}^2_{(-i)}(1-h_i)}}$$

en notant $\mathbb{X}_i$ la $i^{\text{ème}}$ ligne de la matrice de *design* $\mathbb{X}$, h_i le $i^{\text{ème}}$ coefficient de la matrice des leviers $H = \mathbb{X}.(\mathbb{X}^t.\mathbb{X})^{-1}.\mathbb{X}^t$ et $\widehat{\sigma}^2_{(-i)}$ est l'estimateur de la variance résiduelle σ^2 calculé sans l'individu i. Sous les hypothèses $(H_1)-(H_7)$ du cours, ces résidus suivent une loi de Student $T(n-p-2)$ que l'on peut approximer par une loi $\mathcal{N}(0,1)$ dès que n est assez grand devant p, ce qui est de toute façon souhaitable. Ces résidus sont calculés par le logiciel `R` au moyen de la fonction `rstudent`.

• **Exercice**:
Notons β le vecteur des paramètres de régression à estimer dans le cadre d'un modèle d'analyse de la variance à un facteur et $\widehat{\beta}$ son estimateur.

1. Proposer des simulations de Monte-Carlo permettant d'évaluer le biais empirique de $\widehat{\beta}$ et son écart-type estimé.
   Le biais et l'écart-type estimés sont-ils sensibles à l'hypothèse de normalité? à l'hypothèse d'homoscédasticité? à la taille de l'échantillon? au *design*? au nombre de coefficients du modèle?

2. Proposer des simulations de Monte-Carlo permettant d'illustrer la convergence asymptotique de $\widehat{\beta}$ vers β. La convergence est-elle sensible à l'hypothèse de normalité? à l'hypothèse d'homoscédasticité? au *design*? au nombre de coefficients du modèle?

3. Le test de Anderson-Darling est un test de normalité. L'hypothèse nulle est H_0: "adéquation à la loi normale" tandis que l'hypothèse alternative est H_1: "non-adéquation à la loi normale". Ce test est implémenté dans le logiciel `R` au moyen de la fonction `ad.test` du package `nortest`.
   Présenter ce test. Proposer des simulations de Monte-Carlo permettant d'évaluer l'erreur empirique de $1^{\text{ère}}$ espèce et la puissance empirique de ce test. Les résultats obtenus sont-ils sensibles à l'hypothèse d'homoscédasticité? à la taille de l'échantillon? au *design*? au nombre de coefficients du modèle?