
Sujet 8

• **Brefs rappels sur la régression linéaire:**

Lorsque $(X^{(1)}, \dots, X^{(p)})$ est un vecteur de covariables quantitatives, le modèle de régression linéaire correspondant à l'observation de $(Y, X^{(1)}, \dots, X^{(p)})$ s'écrit:

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

où $\mathbb{E}[\varepsilon] = 0$, $\text{Var}(\varepsilon) < \infty$ et $\varepsilon \perp (X^{(1)}, \dots, X^{(p)})$.

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 homoscedastique 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 $\hat{\beta}$ 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:

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

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 $H = \mathbb{X} \cdot (\mathbb{X}^t \cdot \mathbb{X})^{-1} \cdot \mathbb{X}^t$ et $\hat{\sigma}_{(-i)}^2$ 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 $\hat{\beta}$ son estimateur.

1. Proposer des simulations de Monte-Carlo permettant d'évaluer le biais empirique de $\hat{\beta}$ 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 $\hat{\beta}$ 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 Shapiro-Wilk 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 `shapiro.test`. Présenter ce test. Proposer des simulations de Monte-Carlo permettant d'évaluer l'erreur empirique de 1^{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?