5. Modèles linéaire et linéaire gaussien : définitions, estimateur des moindres carrés, test de Student

Objectifs : Savoir faire mener les calculs de la régression linéaire, interpréter les résultats, donner des intervalles de confiance sur les coefficients, tester la nullité d'un coefficient. L'exercice 5.1 est à faire pendant le TD, les autres sont à chercher de votre côté.

Exercice 5.1 (Les eucalyptus)**.** On souhaite expliquer la hauteur y (en mètres) d'un eucalyptus en fonction de sa circonférence x (en centimètres) à 1 mètre 30 du sol, et de la racine carrée de celle-ci. On a relevé $n = 1429$ mesures de couples (x_i, y_i), le nuage de points étant représenté sur la figure 1 ci-contre.

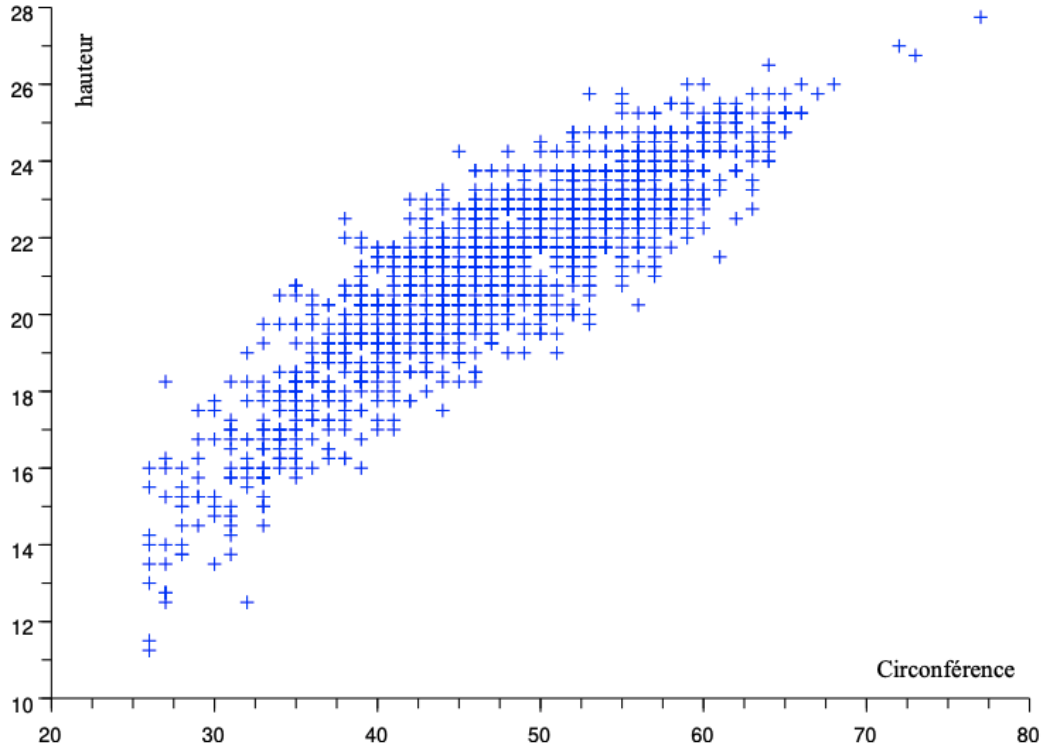

Figure 1 – *Données de hauteurs d'eucalyptus (m) en fonction de leur circonférence (cm).*

On propose donc le modèle linéaire suivant : pour tout $1 \leq i \leq n$, $Y_i = \beta_1 + \beta_2 X_i + \beta_3 \sqrt{X_i} + \varepsilon_i$, où les ε_i sont gaussiennes i.i.d. $\mathcal{N}(0, \sigma^2)$. On pose

$$X = \begin{bmatrix} 1 & x_1 & \sqrt{x_1} \\ \vdots & \vdots & \vdots \\ 1 & x_n & \sqrt{x_n} \end{bmatrix} \quad \text{et} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Nous avons observé

$$X^T X = \begin{bmatrix} ? & ? & 9792 \\ ? & 3306000 & ? \\ ? & 471200 & 67660 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 30310 \\ 1462000 \\ 209700 \end{bmatrix} \quad \text{et} \quad Y^T Y = 651900.$$

1. Compléter la matrice $X^T X$.
2. Que vaut la circonférence moyenne empirique $\overline{x}$?

3. Les calculs donnent en arrondissant :

$$(X^TX)^{-1} = \begin{bmatrix} 4.646 & 0.101 & -1.379 \\ 0.101 & 0.002 & -0.030 \\ -1.379 & -0.030 & 0.411 \end{bmatrix} \quad \text{et} \quad (X^TX)^{-1}X^TY = \begin{bmatrix} -16.8 \\ -0.30 \\ 7.62 \end{bmatrix}.$$

   Que vaut ici l'estimateur des moindres carrés $\hat{\beta}$? Représenter la courbe de régression obtenue sur la Figure 1 via le calcul de quelques points.

4. Vérifier que pour tout modèle linéaire identifiable, $Y^TX\hat{\beta} = \|X\hat{\beta}\|^2$.
5. En déduire la valeur de l'estimateur de σ^2 débiaisé et en donner une intervalle de confiance de niveau 95%. On utilisera l'approximation suivante : quand m est grand, une variable $\chi^2(m)$ est proche d'une variable $\mathcal{N}(m, 2m)$. On donne le quantile gaussien standard à 97.5% : $q = 1.96$.
6. Donner un intervalle de confiance pour β_3 de probabilité de couverture 95%. On approchera la loi $\mathcal{T}(m)$ par la loi $\mathcal{N}(0,1)$, quand m est grand.
7. Tester l'hypothèse $\beta_2 = 0$ au niveau de risque 10%. On fera les mêmes approximations que précédemment. On donne le quantile gaussien standard à 95% : $q = 1.645$. Interpréter.

Exercice 5.2 (Exemple de régression linéaire à la main)**.** On considère le modèle linéaire qui s'écrit matriciellement $Y = \theta_0 e + \theta_1 Z + \varepsilon$, avec $Y \in \mathbb{R}^n$, e le vecteur de $\mathbb{R}^n$ dont les coordonnées valent toutes 1, $Z = (z_1, z_2, \ldots, z_n)^T \in \mathbb{R}^n$, et ε un bruit centré.

1. Donner une condition nécéssaire et suffisante explicite sur le vecteur Z pour que le modèle soit identifiable. Interpréter.

On se place sous la condition d'identifiabilité de la question 1. On introduit la covariance empirique entre Y et Z définie par $C(Y,Z) := \frac{1}{n}\sum_{i=1}^n z_iY_i - \overline{Z} \times \overline{Y}$, ainsi que la variance empirique de Z définie par $V(Z) := \frac{1}{n}\sum_{i=1}^n z_i^2 - (\frac{1}{n}\sum_{i=1}^n z_i)^2$.

2. Calculer à la main l'estimateur des moindres carrés $\hat{\theta}_{MC} = (\hat{\theta}_0, \hat{\theta}_1)^T$. On écrira $\hat{\theta}_1$ en fonction de C et V, puis $\hat{\theta}_0$ en fonction de θ_0.
3. Montrer que le point moyen, de coordonnées $(\bar{Z}, \bar{Y})$, appartient à la droite de régression obtenue.

Exercice 5.3 (Théorème de Gauss-Markov)**.** On note $\preceq$ la relation d'ordre dans $S_p(\mathbb{R})$ définie par :

$$A \preceq B \iff B - A \in S_p^+(\mathbb{R}).$$

Pour $\hat{\theta}_1$ et $\hat{\theta}_2$ deux estimateurs sans biais de $\theta \in \mathbb{R}^p$, on dira que $\hat{\theta}_1$ *est meilleur que* $\hat{\theta}_2$ si $\mathrm{Var}(\hat{\theta}_1) \preceq \mathrm{Var}(\hat{\theta}_2)$.

Le but de cet exercice est de démontrer le *théorème de Gauss-Markov*, dont l'énoncé est le suivant : *dans un modèle linéaire identifiable, parmi tous les estimateurs de θ linéaires en Y et sans biais, l'estimateur des moindres carrés est le meilleur (au sens de l'ordre $\preceq$).*

On considère donc $\widetilde{\theta}$ un autre estimateur de θ, linéaire en Y et sans biais. On l'écrit $\widetilde{\theta} = CY$ avec $C = (X^TX)^{-1}X^T + D$ avec $D \in \mathbb{R}^{p\times n}$.

1. Montrer que $DX = 0$.
2. Montrer que $\mathrm{Var}(\widetilde{\theta}) = \mathrm{Var}(\hat{\theta}_{MC}) + \sigma^2 DD^T$. Conclure.