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.

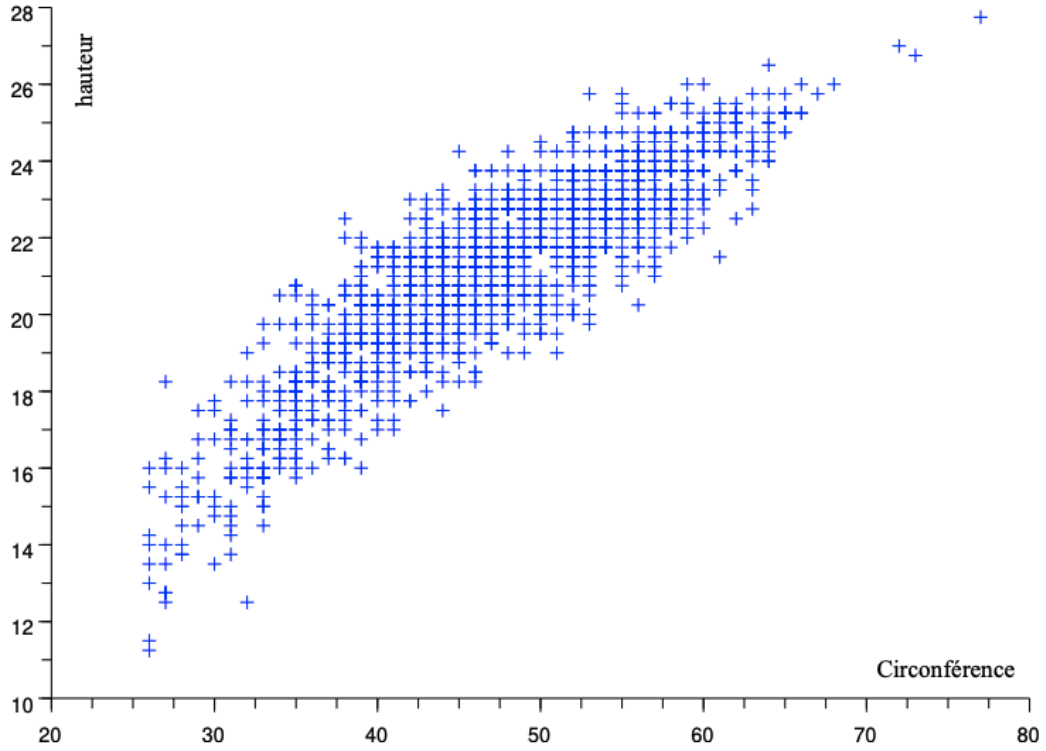

Figure 1 – *Données de hauteurs d'eucalyptus (m) en fonction de leur circonférence (cm).*

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 $X^T X$. *Solution. La norme au carré du vecteur des 1 c'est n. La norme au carré des $\sqrt{x_i}$ c'est juste la somme au carré des x_i. On complète par symétrie.*

D'où $X^TX = \begin{bmatrix} 1429 & 67660 & 9792 \\ 67660 & 3306000 & 471200 \\ 9792 & 471200 & 67660 \end{bmatrix}$

2. Que vaut la circonférence moyenne empirique $\overline{x}$? *Solution. C'est 67660/1429 = 47,3 cm*

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. *Solution. On lit* $\hat{\beta} = \begin{bmatrix} -16.8 \\ -0.30 \\ 7.62 \end{bmatrix}$, *puis on met des petits points et on interpole.*

4. Vérifier que pour tout modèle linéaire identifiable, $Y^TX\hat{\beta} = \|X\hat{\beta}\|^2$. *Solution. Il suffit d'écrire :*

$$\begin{aligned} \|X\hat{\beta}\|^2 &= \hat{\beta}^TX^TX\hat{\beta} \\ &= ((X^TX)^{-1}X^TY)^TX^TX\hat{\beta} \\ &= Y^TX(X^TX)^{-1}X^TX\hat{\beta} = Y^TX\hat{\beta}\,. \end{aligned}$$

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% : $q = 1.96$. *Solution. En utilisant la question 4, on obtient*

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n-3} = \frac{\|Y\|^2 - \|X\hat{\beta}\|^2}{n-3}\,.$$

Mais à nouveau $\|X\hat{\beta}\|^2 = \hat{\beta}^TX^TY$ *qui se calcule avec les données : c'est* $-16.8 \times 30310 - 0.3 \times 1462000 + 7.62 \times 209700 = 650106$. *Au final on a* $\hat{\sigma}^2 = \frac{651900-650106}{1426} = 1.26$. *On a que* $\hat{\sigma}^2/\sigma^2 \sim \frac{\chi^2(n-p)}{n-p} \simeq 1 + \frac{\sqrt{2}}{\sqrt{n-p}}\mathcal{N}(0,1)$ *donc on prend l'intervalle bilatère* $\hat{\sigma}^2[(1 \mp \sqrt{2} \times 1.96/\sqrt{1426})^{-1}] \simeq \hat{\sigma}^2[0.932, 1.079] = [1.174, 1.360]$.

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. *Solution. On a classiquement l'intervalle* $\left[\hat{\beta}_3 \pm q_{1-\alpha/2}^{(n-p)}\hat{\sigma}\sqrt{(X^TX)^{-1}_{3,3}}\right]$ *qui ici vaut* $[7.62 - 1.96 \times \sqrt{1.26} \times \sqrt{0.411}, 7.62 + 1.96 \times \sqrt{1.26} \times \sqrt{0.411}] = [6.21, 9.03]$.

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. *Solution. On sait que sous* H_0, $T = \frac{\hat{\beta}_2}{\hat{\sigma}\sqrt{(X^TX)^{-1}_{2,2}}} \sim \mathcal{T}(n-p) \simeq \mathcal{N}(0,1)$. *Calculons la statistique* T *ici, elle donne*

$$T = \frac{-0.3}{\sqrt{1.26}\sqrt{0.002}} = -5.98\,.$$

On a $|T| > q$ *(et même vénère), on rejette absolument* H_0. *Interprétation : la dépendance de* y *en* x *(et pas que en* $\sqrt{x}$*) est donc primordiale pour le modèle.*

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, $Z = (z_1, z_2, \dots, z_n)^T \in \mathbb{R}^n$, et ε un bruit centré.

1. Donner une condition nécéssaire et suffisante explicite sur le vecteur Z pour que le modèle soit identifiable. Interpréter. *Solution. On a, dans notre cas $X \in \mathbb{R}^{n\times p}$ qui s'écrit $X = [e \,|\, Z]$ avec $p = 2$. X est de rang plein ssi Z n'est pas colinéaire à e, c'est-à-dire de coordonnées non toutes égales. Intuitivement, pour déterminer l'équation de la droite de régression, il faut au moins deux points d'abcisses distinctes.*

On se place sous la condition d'identifiabilité de la question 1. On introduit la covariance empirique entre Y et Z définie par $C(Y,Z) := \frac{1}{n}\sum_{i=1}^n z_i Y_i - \overline{Z} \times \overline{Y}$, 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. *Solution. On applique la formule explicite du cours. Notons que l'EMC est toujours défini même si le modèle n'est pas gaussien. On a*
$$X^T X = \begin{pmatrix} e^T e & e^T Z \\ e^T Z & Z^T Z \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n z_i \\ \sum_{i=1}^n z_i & \sum_{i=1}^n z_i^2 \end{pmatrix}$$
et
$$(X^T X)^{-1} = \frac{1}{n^2 V(Z)} \begin{pmatrix} \sum_{i=1}^n z_i^2 & -\sum_{i=1}^n z_i \\ -\sum_{i=1}^n z_i & n \end{pmatrix}$$
et de plus
$$X^T Y = \begin{pmatrix} e^T Y \\ Z^T Y \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n z_i Y_i \end{pmatrix}.$$
On obtient finalement
$$\hat{\theta}_{MC} = (X^T X)^{-1} X^T Y = \begin{pmatrix} e^T Y \\ Z^T Y \end{pmatrix} = \frac{1}{n^2 V(Z)} \begin{pmatrix} (\sum_{i=1}^n z_i^2)(\sum_{i=1}^n Y_i) - (\sum_{i=1}^n z_i)(\sum_{i=1}^n z_i Y_i) \\ n\sum_{i=1}^n z_i Y_i - (\sum_{i=1}^n z_i)(\sum_{i=1}^n Y_i) \end{pmatrix}$$
$$= \frac{1}{V(Z)} \begin{pmatrix} (V(Z) + \overline{Z}^2)\overline{Y} - \overline{Z}(\sum_{i=1}^n z_i Y_i) \\ C(Y,Z) \end{pmatrix}$$
$$= \begin{pmatrix} \overline{Y} - \overline{Z}(\sum_{i=1}^n z_i Y_i - \overline{ZY}))/V(Z) \\ C(Y,Z)/V(Z) \end{pmatrix} = \begin{pmatrix} \hat{\theta}_0 = \overline{Y} - \overline{Z}\hat{\theta}_1 \\ \hat{\theta}_1 = C(Y,Z)/V(Z) \end{pmatrix}$$

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 $C = (X^T X)^{-1} X^T + D$ avec $D \in \mathbb{R}^{p\times n}$.

1. Montrer que $DX = 0$. *Solution. Le biais doit être nul et on a $\mathbb{E}_\theta[\widetilde{\theta}] = \theta + DX\theta$. On doit avoir $DX\theta = 0$ et ce pour tout θ, donc $DX = 0$.*

2. Montrer que $\mathrm{Var}(\widetilde{\theta}) = \mathrm{Var}(\hat{\theta}_{MC}) + \sigma^2 DD^T$. Conclure. *Solution. On a*

$$\begin{aligned}\mathrm{Var}_\theta(\widetilde{\theta}) &= C\mathrm{Var}_\theta(Y)C^T = ((X^TX)^{-1}X^T + D)\sigma^2 I_p((X^TX)^{-1}X^T + D)^T \\ &= \sigma^2((X^TX)^{-1}X^T + D)(X(X^TX)^{-1} + D^T) \\ &= \mathrm{Var}(\hat{\theta}_{MC}) + \sigma^2 DD^T + \underbrace{\sigma^2(X^TX)^{-1}X^TD^T}_{=0} + \underbrace{\sigma^2 DX(X^TX)^{-1}}_{=0}.\end{aligned}$$

On conclut car $0 \preceq DD^T$.