

Question 1

On voit clairement sur le nuage de points (circonférence/hauteur) que cela suit une droite. On essaye de trouver les valeurs de la droites qui minimisent le risque quadratique.

Question 2

Pour minimiser la fonction $\varphi(\beta_1, \beta_2)$ il faut trouver dériver la fonction par rapport à β_1 et β_2 et trouver les valeurs qui annulent les 2 dérivées.

$$\begin{aligned}\frac{\partial \varphi(\beta_1, \beta_2)}{\partial \beta_1} &= \frac{\sum_{i=1}^n (Y_i - \beta_1 - \beta_2 x_i)^2}{\beta_1} = \frac{\sum_{i=1}^n Y_i^2 - \beta_1 Y_i - \beta_2 x_i Y_i - \beta_1 Y_i + \beta_1^2 + \beta_1 \beta_2 x_i - \beta_2 x_i Y_i + \beta_2 x_i \beta_1 + \beta_2^2 x_i^2}{\beta_1} \\ &= \sum_{i=1}^n -Y_i - Y_i + 2\beta_1 + \beta_2 x_i + \beta_2 x_i = 2 \sum_{i=1}^n -Y_i + \beta_2 x_i + \beta_1 \\ \frac{\partial \varphi(\beta_1, \beta_2)}{\partial \beta_2} &= \frac{\sum_{i=1}^n (Y_i - \beta_1 - \beta_2 x_i)^2}{\beta_2} = \frac{\sum_{i=1}^n Y_i^2 - \beta_1 Y_i - \beta_2 x_i Y_i - \beta_1 Y_i + \beta_1^2 + \beta_1 \beta_2 x_i - \beta_2 x_i Y_i + \beta_2 x_i \beta_1 + \beta_2^2 x_i^2}{\beta_2} \\ &= \sum_{i=1}^n -x_i Y_i + \beta_1 x_i - x_i Y_i + \beta_1 x_i + 2\beta_2 x_i^2 = 2 \sum_{i=1}^n x_i (-Y_i + \beta_2 x_i + \beta_1)\end{aligned}$$

On cherche $\hat{\beta}_1$ et $\hat{\beta}_2$ les valeurs qui annulent le système

$$\begin{cases} \sum_{i=1}^n x_i (-Y_i + \beta_2 x_i + \beta_1) = 0 \\ \sum_{i=1}^n -Y_i + \beta_2 x_i + \beta_1 = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^n -Y_i x_i + \sum_{i=1}^n \beta_2 x_i^2 + \sum_{i=1}^n \beta_1 x_i = 0 & (1) \\ \sum_{i=1}^n -Y_i + \sum_{i=1}^n \beta_2 x_i + \sum_{i=1}^n \beta_1 = 0 & (2) \end{cases}$$

$$\begin{cases} \sum_{i=1}^n -Y_i x_i + \beta_2 \sum_{i=1}^n x_i^2 + \beta_1 \sum_{i=1}^n x_i = 0 & (1) \\ \sum_{i=1}^n -Y_i + \beta_2 \sum_{i=1}^n x_i + n\beta_1 = 0 & (2) \end{cases}$$

On fait (3) = n(1) - (2) $\sum_{i=1}^n x_i$

$$\begin{cases} n \sum_{i=1}^n -Y_i x_i + n\beta_2 \sum_{i=1}^n x_i^2 + \sum_{i=1}^n Y_i \sum_{i=1}^n x_i - \beta_2 (\sum_{i=1}^n x_i)^2 = 0 & (3) \\ \sum_{i=1}^n -Y_i + n\beta_2 \sum_{i=1}^n x_i + n\beta_1 = 0 & (2) \end{cases}$$

$$\begin{cases} -n \sum_{i=1}^n Y_i x_i + \sum_{i=1}^n Y_i \sum_{i=1}^n x_i = \beta_2 (\sum_{i=1}^n x_i)^2 - n\beta_2 \sum_{i=1}^n x_i^2 & (3) \\ \sum_{i=1}^n -Y_i + n\beta_2 \sum_{i=1}^n x_i + n\beta_1 = 0 & (2) \end{cases}$$

$$\begin{cases} \beta_2 = \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i - n \sum_{i=1}^n Y_i x_i}{(\sum_{i=1}^n x_i)^2 - n \sum_{i=1}^n x_i^2} & (3) \\ \beta_1 = \frac{1}{n} (\sum_{i=1}^n Y_i - n\beta_2 \sum_{i=1}^n x_i) & (2) \end{cases}$$

$$\begin{cases} \beta_2 = \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i - n \sum_{i=1}^n Y_i x_i}{(\sum_{i=1}^n x_i)^2 - n \sum_{i=1}^n x_i^2} & (3) \\ \beta_1 = \frac{\sum_{i=1}^n x_i \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n x_i^2 \sum_{i=1}^n Y_i}{(\sum_{i=1}^n x_i)^2 - n \sum_{i=1}^n x_i^2} & (2) \end{cases}$$

Question 3

Voir Python.

On obtient $\hat{\beta}_1 = 9.037475668452768$ et $\hat{\beta}_2 = 0.257137855007109$.

- Moyenne de epsilon: -3.603610217941441e-11
- risque quadratique: 19.492804231375466

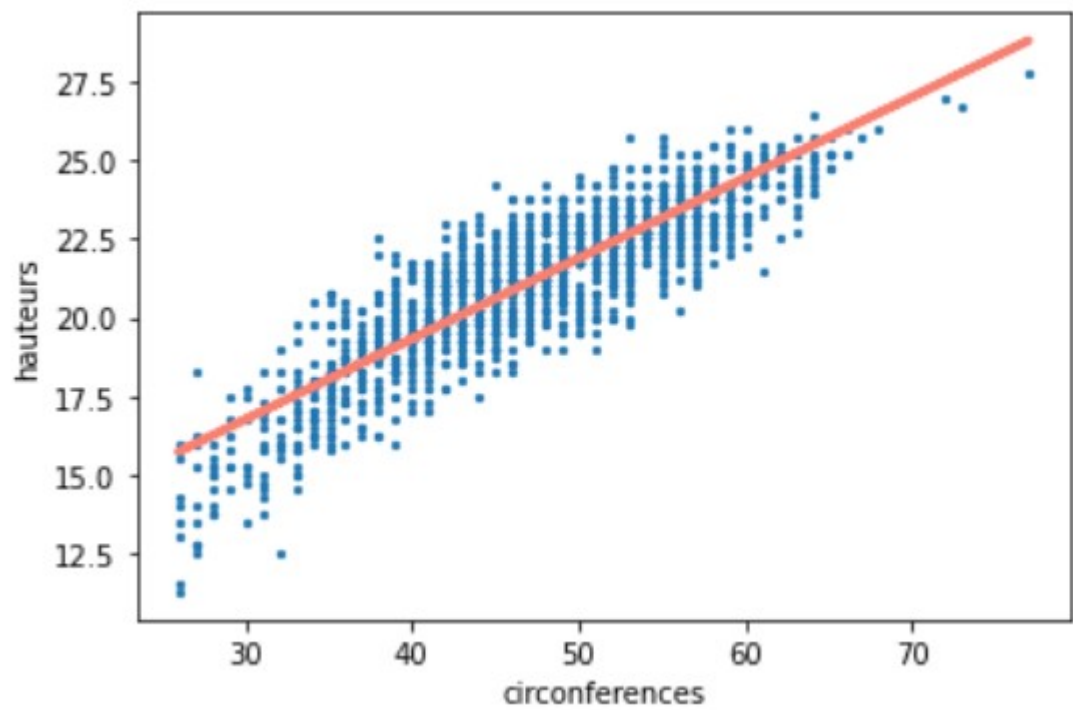


Figure 1: Regression simple

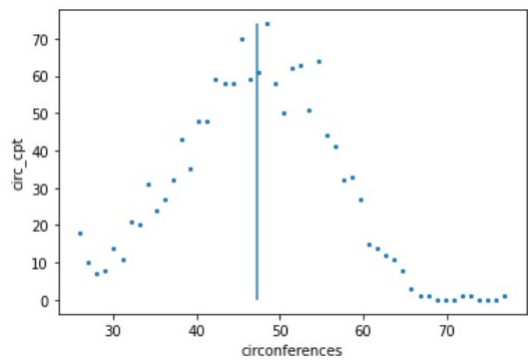


Figure 2: Circonférence

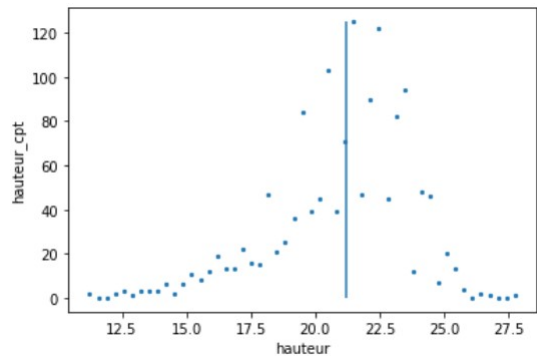


Figure 3: Hauteur

Question 4

Comme le montre les figures 2 et 3, il semble raisonnable de dire que la circonférence (resp. la hauteur) d'un eucalyptus suit une loi normale. Maintenant on n'a qu'un seul échantillon, donc cela pourrait être une pure coïncidence!

Comme les 2 variables aléatoire suivent une loi normale, elle sont indépendentes et identiquement distribués.

Si X suit une loi normale $\mathcal{N}(m, \sigma^2)$ et $Y = AX + b$ alors, Y suit une loi normale $\mathcal{N}(am + b, a^2\sigma^2)$

Si X (resp. Y) suit une loi normale $\mathcal{N}(m_x, \sigma_x^2)$ (resp. $\mathcal{N}(m_y, \sigma_y^2)$) alors $X + Y$ suit une loi normale $\mathcal{N}(m_x + m_y, \sigma_x^2 + \sigma_y^2)$.

Dans notre cas on a $e_i = Y_i - \hat{\beta}_1 + \hat{\beta}_2 x_i$. Donc e_i suit une loi normale $\mathcal{N}(-\hat{\beta}_1 - \hat{\beta}_2 m_x + m_y, \hat{\beta}_2^2 \sigma_x^2 + \sigma_y^2)$.

On a par définition $m_y = \hat{\beta}_1 + \hat{\beta}_2 m_x$. Donc $E(e_i) = 0$ et $\sigma_{e_i} = \hat{\beta}_2^2 \sigma_x^2 + \sigma_y^2$.

Question 5

En cherchant à minimiser $\|Y - X\beta\|^2$, on cherche à trouver l'élément de F le plus proche de Y au sens de la distance euclidienne. Il s'agit de la projection orthogonale de Y sur F . Comme $z \in F$, si et seulement si $z = X\beta$, on cherche $\hat{\beta}$ tel que $X\hat{\beta} = P_F(Y)$.

Comme $X\hat{\beta} = P_F(Y)$, on a $Y - X\hat{\beta} = Y - P_F(Y)$ qui est un vecteur orthogonal à X et par conséquent aussi à $X\theta$. Le produit scalaire de 2 vecteurs orthogonaux est nul, donc $\forall \theta \in \mathbb{R}^3 \langle Y - X\hat{\beta}, X\theta \rangle = 0$.

Question 6

Pour trouver le minimum par rapport à β , il suffit de dériver l'expression par rapport à β et annuler l'expression. On remarque que $\sum_{i=1}^n (Y - X\beta)^2 = (Y - X\beta)^t (Y - X\beta)$

$$(Y - X\beta)^t (Y - X\beta) = (Y^t - \beta^t X^t)(Y - X\beta) = Y^t Y - Y^t X\beta - \beta^t X^t Y + \beta^t X^t X\beta$$

et

$$\frac{\partial (Y - X\beta)^t (Y - X\beta)}{\partial \beta} = -Y^t X + \beta^t X^t X$$

On cherche $\hat{\beta}$ tel que

$$\begin{aligned} -Y^t X + \hat{\beta}^t X^t X &= 0 \\ (\hat{\beta}^t X^t X)^t &= (-Y^t X)^t \end{aligned}$$

Donc

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Question 7

Voir Python.

$$\hat{\beta} = [-24.35200327, -0.48294547, 9.98688814]$$

- Moyenne de epsilon: 1.0692449957862278e-13
- risque quadratique: 19.32298986873724

Valeurs proches et meilleures que celles de la régression simple à la question 3.

Question 8

On suppose que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et on a $Y_i = X_i \beta + \epsilon_i$. La loi suivit par Y_i dépend de la loi suivit par X_i et β . Donc si on suppose que X_i suit une loi exponentielle, il y a de grande chance que Y_i suive aussi une loi exponentielle.

Maintenant, sur le seul échantillon fourni, on a montré à la question 4 que X_i suit certainement une loi normale (mais avec un seul échantillon on ne peut pas être sûr).

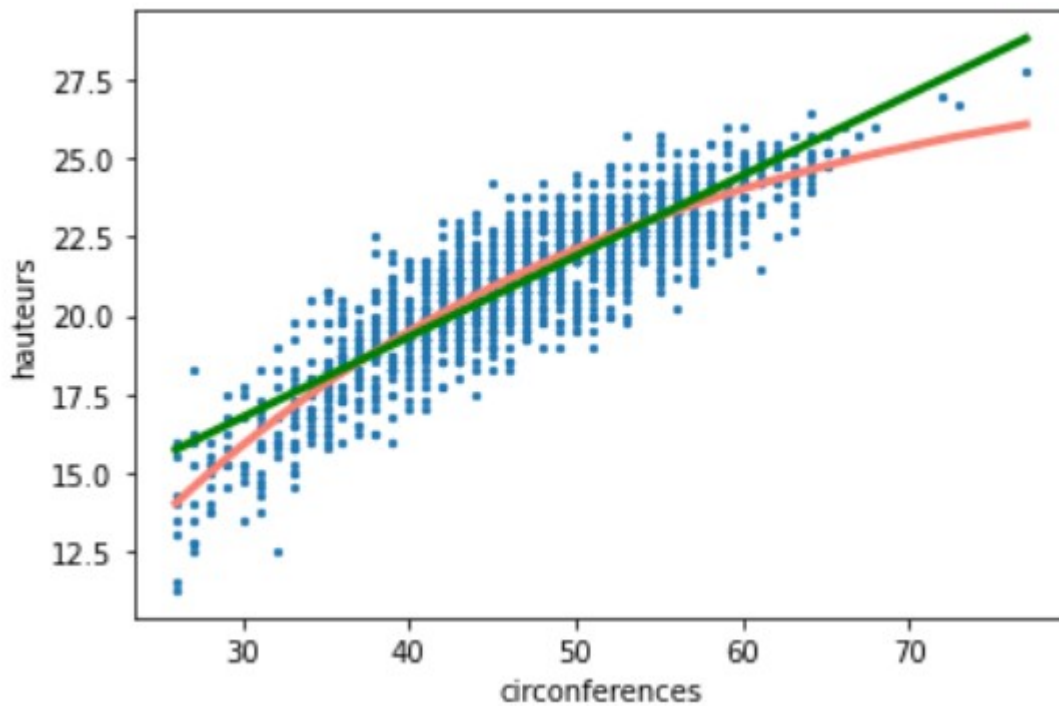


Figure 4: Regression multiple

Test de Student

Pourquoi on cherche à se demander si $\beta_3 = 0$. On fait la supposition ici que les β_1 et β_2 sont identiques pour les 2 types de régression pour l'échantillon donné. Ce qui n'est pas le cas.

Sur l'échantillon donné la régression multiple est meilleure.

Question 9

Soient Z une variable aléatoire de loi normale centrée et réduite et U une variable indépendante de Z et distribuée suivant la loi du chi-deux à k degrés de liberté. Par définition la variable $T = \frac{Z}{\sqrt{U/k}}$ suit une loi de Student à k degrés de liberté.

Prenons $U = (n-3)\hat{\sigma}^2/\sigma^2$. On sait que U suit une loi de chi-deux à $(n-3)$ degrés de liberté (voir question précédente) et $Z = \frac{\hat{\beta}_3}{\sigma m_3}$ suit une loi normale centrée et réduite et $k = n-3$.

$$\frac{Z}{\sqrt{\frac{U}{n-3}}} = \frac{\frac{\hat{\beta}_3}{\sigma m_3}}{\sqrt{\frac{(n-3)\hat{\sigma}^2/\sigma^2}{n-3}}} = \frac{\hat{\beta}_3}{\sigma m_3 \frac{\hat{\sigma}}{\sigma}} = \frac{\hat{\beta}_3}{m_3 \hat{\sigma}} = T$$

Question 10

L'erreur de deviation standard de $Se(\hat{\beta}_3) = \sqrt{[(X^t X)^{-1}]_{3 \times 3} \hat{\sigma}^2}$. la t-value pour le test est $\frac{\hat{\beta}_3 - 0}{Se(\hat{\beta}_3)}$ avec $\hat{\beta}_3 = 9.98688814$, $[(X^t X)^{-1}]_{3 \times 3} =$, $\hat{\sigma}^2 =$ et $Se(\hat{\beta}_3) =$.

Les coefficients $\hat{\beta}_1$ $\hat{\beta}_2$ de la régression linéaire simple et multiple sont différents donc on ne peut pas se passer de la composante \sqrt{x} dans la régression multiple car dans ce cas la régression simple associée ne donne pas le plus petit risque quadratique.

Question 11

Les intervalles de confiance sont

$$\hat{\beta}_0 \pm t_{\alpha/2; n-2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)}$$

$$\hat{\beta}_1 \pm t_{\alpha/2; n-2} \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Voir programme python.

Intervalle de confiance à 95% de $\hat{\beta}_0 = -0.14857451813593908$, celui de $\hat{\beta}_1 = -0.007332356088614205$.
Donc, $\hat{\beta}_0 \in [8.888901150316945, 9.186050186588824]$ et $\hat{\beta}_1 \in [0.24980549891849463, 0.26447021109572305]$.

Intervalle de confiance à 99% de $\hat{\beta}_0 = -0.19535568338512507$, celui de $\hat{\beta}_1 = -0.009641070706375857$.
Donc, $\hat{\beta}_0 \in [8.84211998506776, 9.23283135183801]$ et $\hat{\beta}_1 \in [0.247496784300733, 0.2667789257134847]$.

Estimateur de la variance**Question 12**

Par linéarité de l'espérance on a $E[AX + b] = AE[X] + b$. Pour la variance on a

$$VAR(AX + b) = VAR(AX) = E[(AX)(AX)^t] = E[AXX^tA^t] = AE[XX^t]A^t = AVAR(X)A^t$$

Question 13

On a $\hat{\beta} = (X^tX)^{-1}X^tY$ et $Y = X\beta + \epsilon$ donc

$$\begin{aligned} Y - X\hat{\beta} &= Y - X(X^tX)^{-1}X^tY = (I_n - X(X^tX)^{-1}X^t)Y = (I_n - X(X^tX)^{-1}X^t)(X\beta + \epsilon) \\ &= X\beta - X(X^tX)^{-1}X^tX\beta + (I_n - X(X^tX)^{-1}X^t)\epsilon = X\beta - X\beta = (I_n - X(X^tX)^{-1}X^t)\epsilon \end{aligned}$$

Notons $H = X(X^tX)^{-1}X^t$, on a donc $P = (I_n - H)$.

Question 14

On a $E(Y - X\hat{\beta}) = E(P\epsilon) = PE(\epsilon) = P \cdot 0 = 0$.

Question 15

Il faut montrer que $E(\hat{\sigma}^2) = \sigma^2$.

Question 16

$$(n-3) \frac{\hat{\sigma}^2}{\sigma^2} = \frac{\|Y - X\hat{\beta}\|^2}{\sigma^2} = \frac{\sum_{i=1}^n e_i^2}{\sigma^2} = \frac{e^te}{\sigma^2}$$

Calculons e^te

$$e^te = (P\epsilon)^t(P\epsilon) = \epsilon^t(I_n - H)^t(I_n - H)\epsilon = \epsilon^t(I_n - H)\epsilon$$

Donc on a

$$(n-3) \frac{\hat{\sigma}^2}{\sigma^2} = \frac{\epsilon^t(I_n - H)\epsilon}{\sigma^2} = \frac{\epsilon^t}{\sigma} (I_n - H) \frac{\epsilon}{\sigma}$$

En utilisant le théorème de Fisher Cochran, la formule ci-dessus a une distribution χ^2 avec un degré de liberté $\text{rang}(I_n - H)$.

$$\text{rang}(I_n - H) = \text{tr}(I_n - H) = n - \text{tr}(H) = \text{tr}(X(X^tX)^{-1}X^t) = \text{tr}(I_3) = 3$$

Donc $(n-3) \frac{\hat{\sigma}^2}{\sigma^2}$ a une distribution $\chi^2(n-3)$.

Best Linear Unbiased Estimator (BLUE)**Question 17****Question 18****Question 19**