Question 1

On voit clairement sur le nuage de points (circonference/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.

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

$$= \sum_{i=1}^n -Y_i - Y_i + 2\beta_1 + \beta_2 x_i + \beta_2 x_i = 2\sum_{i=1}^n -Y_i + \beta_2 x_i + \beta_1$$

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

$$= \sum_{i=1}^n -x_i Y_i + \beta_1 x_i - x_i Y_i + \beta_1 x_i + 2\beta_2 x_i^2 = 2\sum_{i=1}^n x_i(-Y_i + \beta_2 x_i + \beta_1)$$

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

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

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

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

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

Question 3

Voir Python.

On obtient $\beta_1 = 9.037475668452768$ et $\beta_2 = 0.257137855007109$.

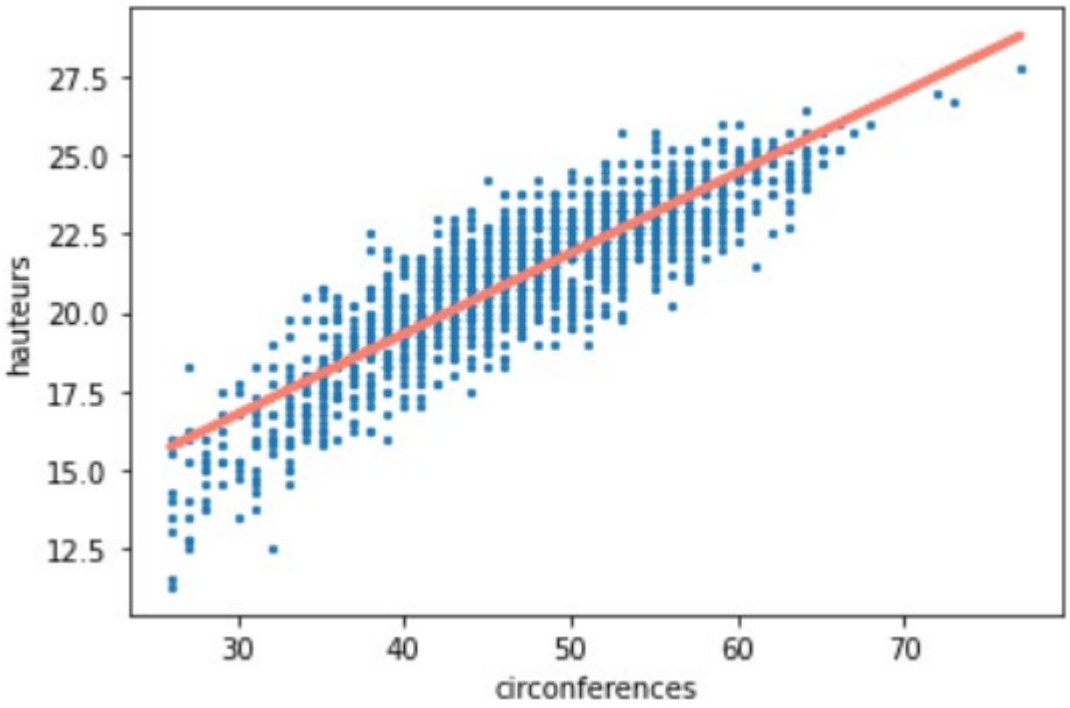

Figure 1: Regression simple

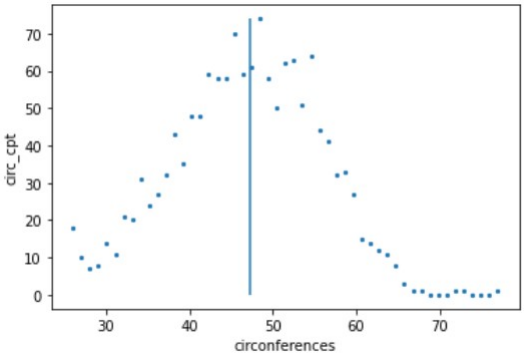

Figure 2: Circonference

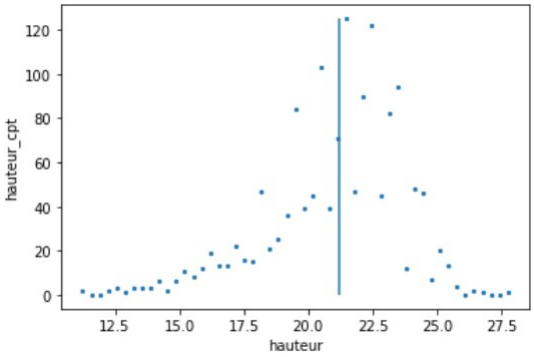

Figure 3: Hauteur

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 coincidence!

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 a $X\theta$. Le produit scalaire de 2 vecteurs orthogonaux est nul, donc $\forall\theta \in \mathbb{R}^3 < Y - X\hat{\beta}, X\theta >= 0$.

Question 6

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

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

et

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

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

$$-Y^tX + \hat{\beta}^tX^tX = 0$$

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

Donc

$$\hat{\beta} = (X^tX)^{-1}X^tY$$

Question 7

Voir Python.

$$\beta = [-24.35200327, -0.48294547, 9.98688814]$$

Question 8

On suppose que ϵ_i $\mathcal{N}(0, \sigma^2)$ et on a $Y_i = X_i\beta + \epsilon_i$. La loi suivit par Y_i dépend de la loi suivit par X_i. 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 sur).

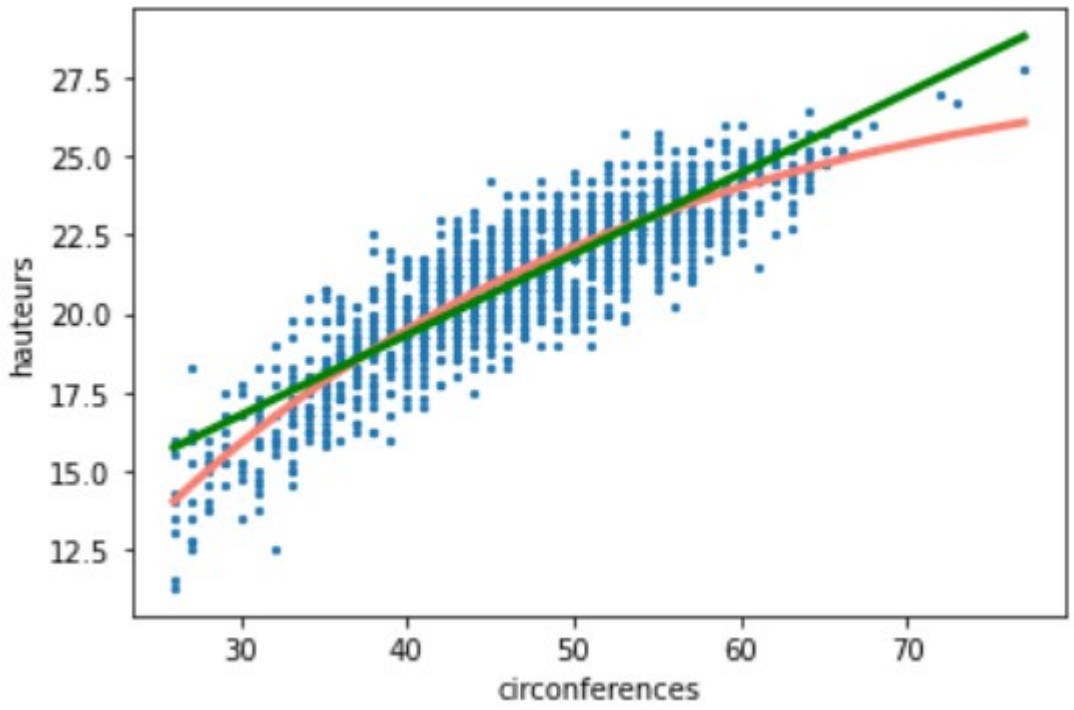

Figure 4: Regression multiple

Test de Student

Au vu du nuage de points, pourquoi ne pas prendre $Y = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4$? Bien meilleure courbe. Et surement meilleur si on prend un polynome d'ordre 10 ou 30.

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.

Cela parait évident que plus le polynome est grand plus la courbe sera précise. Si ce n'était pas la cas alors les valeurs des coefficients seraient nulles.

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 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 $k = n-3$.

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

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 $Y = X\beta + \epsilon$ et $X\hat{\beta} = P_F(Y)$ la projection orthonogale de Y sur F (voir question 5). Donc

$$Y - X\hat{\beta} = X\beta + \epsilon - P_F(Y)$$

$$\epsilon = P_F(Y) - X\beta$$

$P_F(Y) = P_F(X\beta + \epsilon) =$