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

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

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

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

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

Question 3

Voir Python.

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

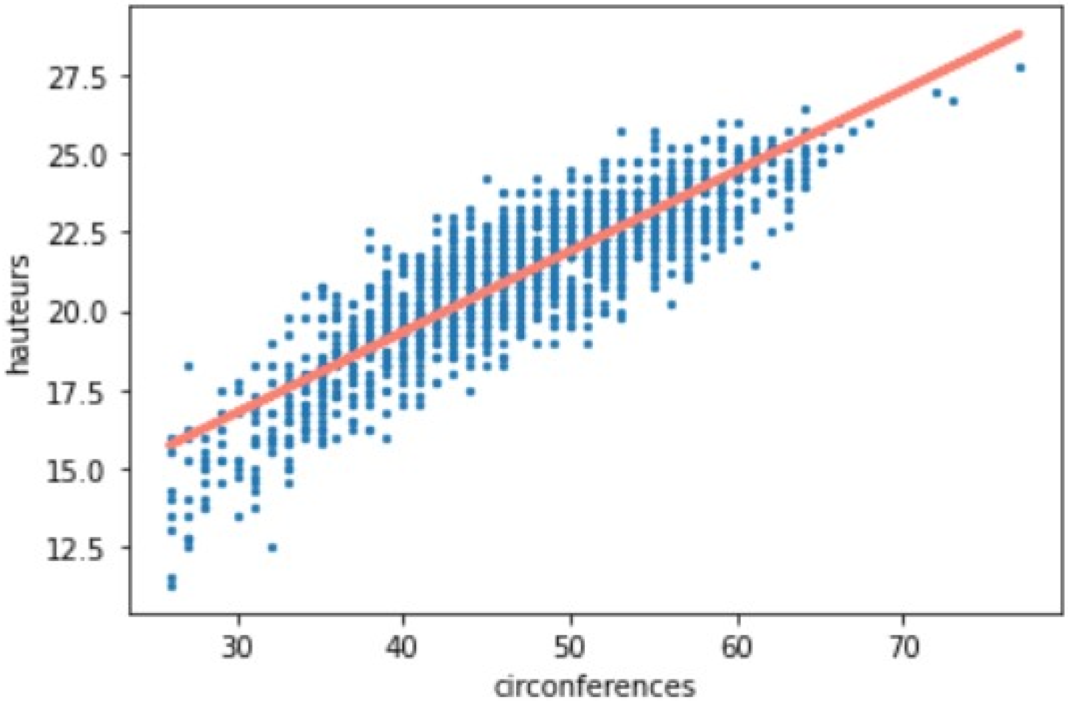

Figure 1: Regression simple

Question 4

Il semble raisonnable de dire que la circonférence (resp. la hauteur) d'un eucalyptus suit une loi normale. Ceci est visible sur le nuage de points car il y a plus de points au centre. 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 $\sum_{i=1}^{n} (Y - X\beta)^2 = (Y - X\beta)^t (Y - X\beta)$

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

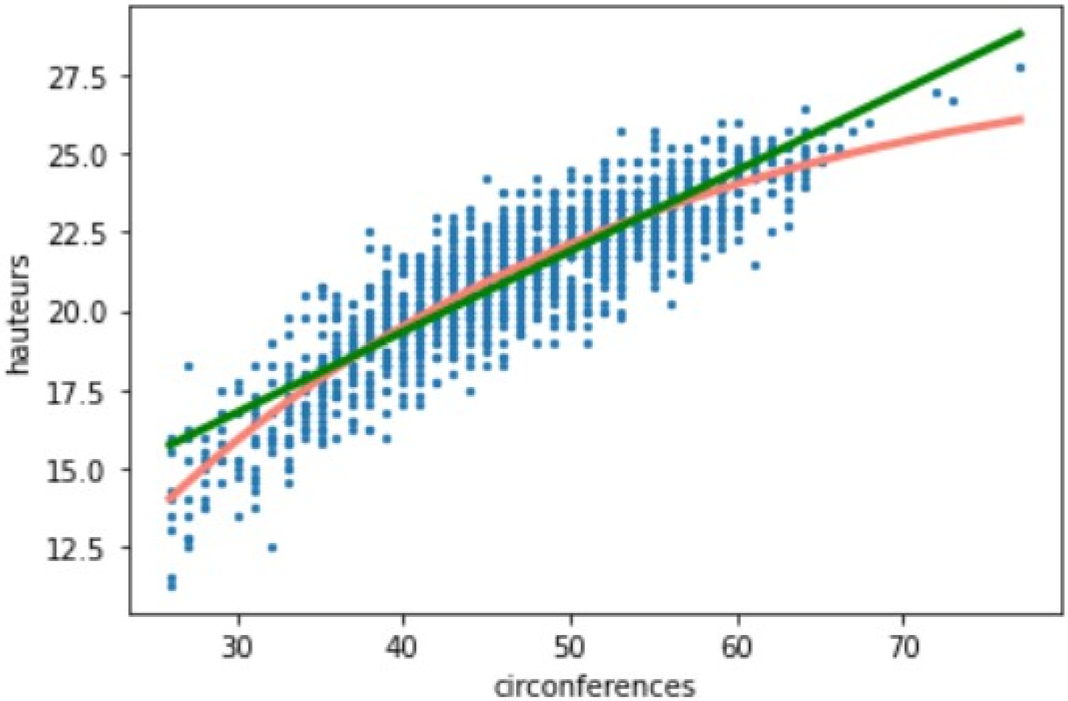

Figure 2: Regression multiple

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

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

Donc

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

Question 7

Voir Python.

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

Question 8

Les y_i sont identiques à ceux de la question 4. Ils suivent une loi normale.
??? pour le reste.

Test de Student

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

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