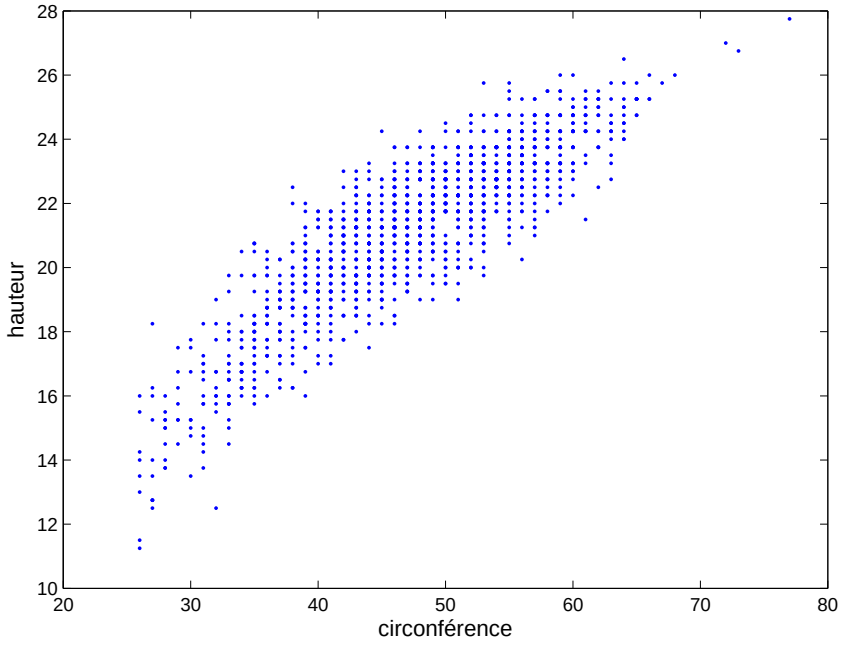
Université Paris-Sud 2014-2015
Projet en modélisation L3 MINT

Régression linéaire pour mesurer la hauteur des eucalyptus

Le graphique ci-dessous montre les mesures prises sur environ 1400 eucalyptus. On en a mesuré la hauteur ainsi que la circonférence à 1,30m du sol. On cherche à trouver la relation entre ces deux quantités. En effet, mesurer la hauteur d'un arbre n'est pas toujours simple, et cela peut permettre d'avoir une valeur approximative de la hauteur par la simple mesure de la circonférence.



Pour chaque arbre i, on notera Y_i la hauteur et x_i la circonférence. On notera également x le vecteur colonne contenant les (x_i). On se placera dans $\mathbb{R}^n$ muni du produit scalaire usuel : $\langle u, v\rangle = \sum_{i=1}^{n} u_i v_i$ et de la norme associée. Les données représentées ci-dessus se trouvent dans le fichier `Eucalyptus.mat`.

1 Régression linéaire simple

Le modèle de régression linéaire simple consiste à supposer que la variable Y s'exprime linéairement en fonction de x, à une petite variation aléatoire près, notée ε et appelée "erreur" ou "bruit" :

$$\forall i \in \{1, \dots, n\}, \qquad Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

Les variables (x_i) et (Y_i) sont observées, tandis que les réels β_1 et β_2 sont inconnus. On cherche donc une droite d'équation $y = \beta_1 + \beta_2 x$ qui traduit la liaison entre Y et x.

On appelle estimateurs des moindres carrés les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ obtenus en minimisant la fonction

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

Question 1. Pourquoi proposer cet estimateur ?

Question 2. Comment minimise-t-on une fonction de deux variables ? Trouver $\hat{\beta}_1$ et $\hat{\beta}_2$.

Question 3. Programmer et tracer la droite de régression $y = \hat{\beta}_1 + \hat{\beta}_2 x$.

On suppose maintenant que les variables aléatoires ε_i sont indépendantes identiquement distribuées, de moyenne nulle et de variance σ^2.

Question 4. Que pensez-vous de ces hypothèses ? Comment peut-on estimer ce paramètre de variance σ^2 ?

2 Régression linéaire multiple

Le modèle précédent est un peu simpliste. Au vu du nuage de points, on suppose maintenant que les données vérifient le modèle suivant :

$$\forall i \in \{1, \dots, n\}, \qquad Y_i = \beta_1 + \beta_2 x_i + \beta_3 \sqrt{x_i} + \varepsilon_i.$$

On peut réécrire :

$$Y = X\beta + \varepsilon$$

avec Y et ε des vecteurs colonnes $n \times 1$, X une matrice $n \times 3$ et $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$.

Comme précédemment, pour estimer β, on cherche à minimiser

$$\sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 x_i - \beta_3 \sqrt{x_i})^2 = \|Y - X\beta\|^2.$$

Notons F l'espace vectoriel engendré par les trois colonnes de X : $F = \mathrm{Vect}(1, x, \sqrt{x})$. Ainsi un vecteur z appartient à F si et seulement s'il existe β tel que $z = X\beta$. Minimiser $\|Y - X\beta\|^2$

revient alors à minimiser la distance entre Y et les points de F. Cette distance est minimum en la projection orthogonale de Y sur F : $X\hat{\beta} = P_F(Y)$. Par définition le projeté orthogonal de Y sur F vérifie la propriété suivante : $Y - P_F(Y)$ est orthogonal à tout vecteur z de F. C'est-à-dire

$$\forall \theta \in \mathbb{R}^3 \qquad \langle Y - X\hat{\beta}, X\theta \rangle = 0.$$

Question 5. Montrer que $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Question 6. Programmer et tracer la courbe de régression $y = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 \sqrt{x}$.

On suppose maintenant que les variables aléatoires ε_i suivent une loi gaussiennes $\mathcal{N}(0, \sigma^2)$.

Question 7. Quel est alors la loi des Y_i ? Montrer que $\hat{\beta}$ est l'estimateur du maximum de vraisemblance. Calculer la loi des $\hat{\beta}_j$.

En ce qui concerne l'estimation de σ^2, on peut montrer que

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

est un estimateur sans biais de σ^2, tel que $(n-3)\hat{\sigma}^2/\sigma^2$ suit une loi du chi-deux à $(n-3)$ degrés de liberté, et tel que les variables $\hat{\sigma}^2$ et $\hat{\beta}$ sont indépendantes.

3 Test de Student

On se demande maintenant lequel des deux modèles (régression simple ou régression multiple) correspond le mieux aux données. Cela revient ici à se demander si $\beta_3 = 0$. On cherche donc à effectuer le test :

$$H_0 : \beta_3 = 0 \qquad \text{vs} \qquad H_1 : \beta_3 \neq 0.$$

Pour cela, on va bien sûr utiliser la valeur $\hat{\beta}_3$ obtenue, et observer si elle est proche de 0. Plus précisément, on introduit la statistique de test suivante

$$T = \frac{\hat{\beta}_3}{m_3 \hat{\sigma}}$$

où $m_3 = \sqrt{[(X^T X)^{-1}]_{3,3}}$.

Question 8. Montrer que T suit une loi de Student à $(n-3)$ degrés de liberté $\mathcal{T}(n-3)$ (la loi de Student à d degrés de liberté est la loi d'une variable normale centrée réduite divisée par la racine carrée d'une variable du chi-deux à d degrés de liberté, les deux variables étant indépendantes).

Question 9. En déduire une procédure de test. L'implémenter sur les données.