Régression linréaire pour mesurer la hauteur des eucalyptus

Charlote Ayrault - The ghost

30 avril 2022

Régression linéaire simple

Question 1

Pourquoi proposer un estimateur linéaire simple ?

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

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

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

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

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

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

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

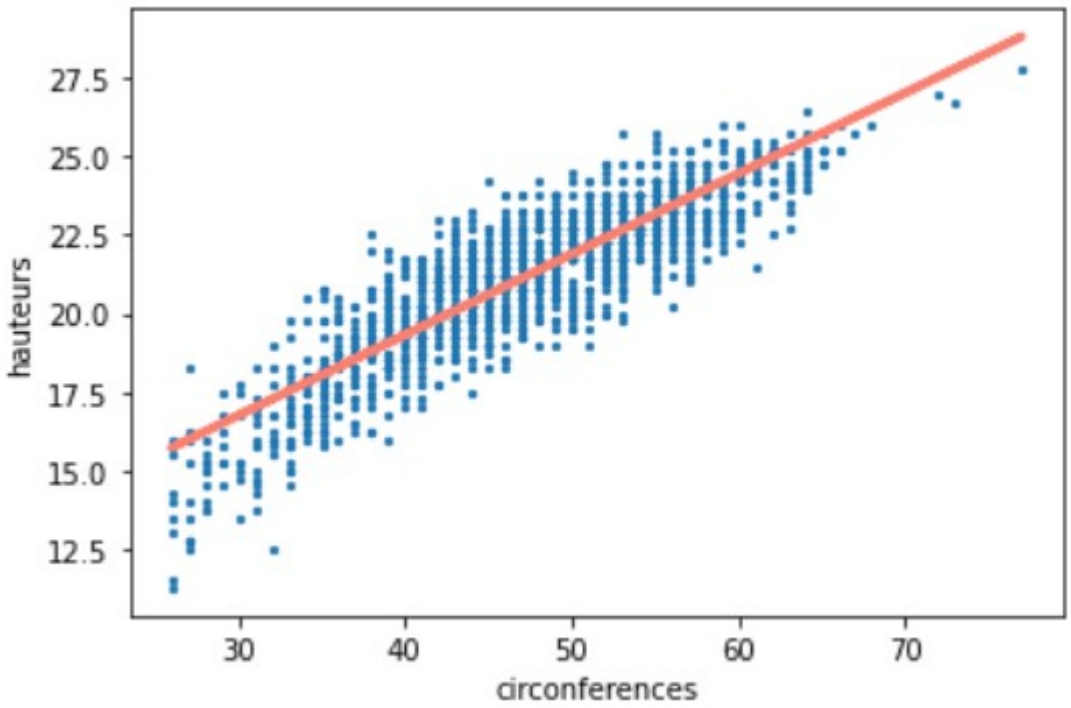

FIGURE 1 – Régression simple

Question 3

Programmer et tracer la droite de régression $y = \hat{\beta}_1 + \hat{\beta}_2 x$ ***?***

Voir la figure 1.
On a obtenu $\hat{\beta}_1 = 9.037475668452768$ et $\hat{\beta}_2 = 0.257137855007109$.
— Moyenne de espilon : -3.603610217941441e-11
— risque quadratique : 19.492804231375466

Question 4

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

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 aléatoires 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 $\sigma_{e_i} = {\hat{\beta}_2}^2 \sigma_x^2 + \sigma_y^2$.

Régression linéaire multiple

Question 5

Montrer que $X\hat{\beta} = P_F(Y)$***, où*** $P_F(Y)$ ***est la projection orthonogale de*** Y ***sur*** F***. En déduire :*** $\forall \theta \in \mathbb{R}^3, \langle Y - X\hat{\beta}, X\theta \rangle = 0$***.***

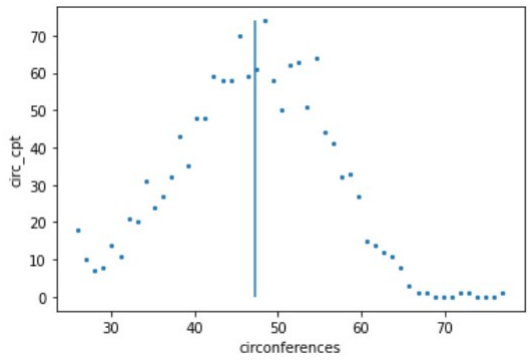

FIGURE 2 – Circonference

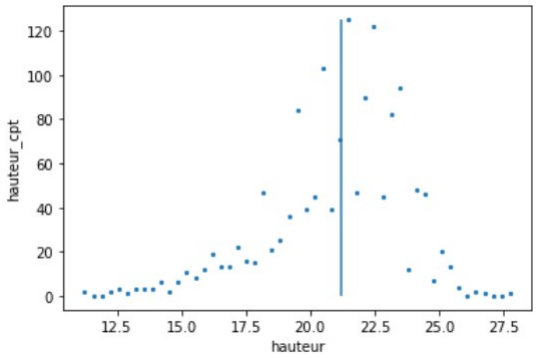

FIGURE 3 – Hauteur

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

Question 6

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

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 $\sum_{i=1}^{n} (Y - X\beta)^2 = (Y - X\beta)^t (Y - X\beta)$

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

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

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

Voir figure 4.

On a obtenu les valeurs suivantes :
— $\hat{\beta}_1 = -24.35200327$
— $\hat{\beta}_2 = -0.48294547$
— $\hat{\beta}_3 = 9.98688814$

On a également calculé :
— Moyenne de espilon : 1.0692449957862278e-13
— risque quadratique : 19.32298986873724

Les valeurs proches mais meilleures que celles de la régression simple à la question 3.

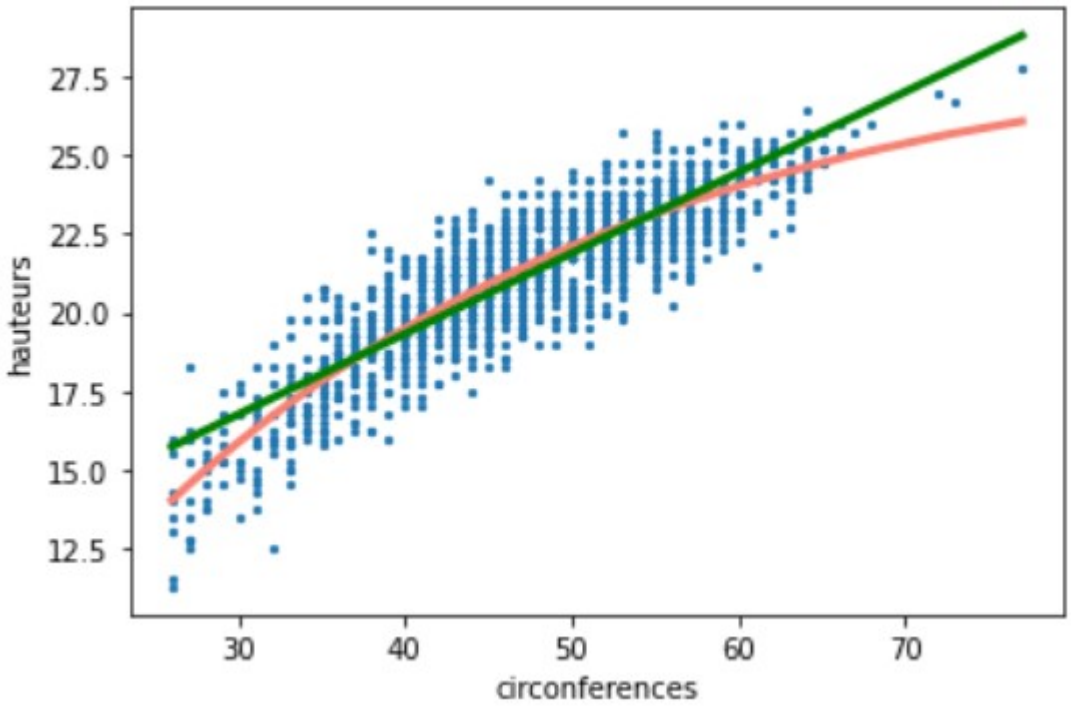

FIGURE 4 – Regression multiple

Question 8

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

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 sur).

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

Montrer qur T suit une loi de Student à $(n-3)$ degrés de liberté $\tau(n-3)$

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 $k = n-3$.

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

Question 10

En déduire une rocédure de test. L'implémenter sur les donnés. Quelle conclusion pouvez-vous en tirer ? Pourrait-on se passer de a composante linéaire en x de la régression ?

L'erreur de deviation standard de $Se(\hat{\beta}_3) = \sqrt{[(X^tX)^{-1}]_{3x3}\hat{\sigma}^2}$. la t-value pour le test est $\frac{\hat{\beta}_3 - 0}{Se(\hat{\beta}_3)}$ avec $\hat{\beta}_3 = 9.98688814$, $[(X^tX)^{-1}]_{3x3} =$, $\hat{\sigma}^2 =$ 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 regression simple associée ne donne pas le plus petit risque quadratique.

Question 11

Dans le cas de la régression linéaire simple, donner les intervalles de confiance a 95% et 99% pour β_1 et β_2. Les tracer en fonction de n pour les données fournies

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 $\hat{\beta}_1 \in [0.247496784300733, 0.2667789257134847]$.

On n'a pas pu les tracer en fonction de n car cela dépend des n données que l'on prend pour faire le calcul.

Estimateur de la variance

Question 12

Montrer que $AU + b \sim \mathcal{N}(Am + b, A\Sigma A^T)$.

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

Montrer que $Y - X\hat{\beta}$ peut s'écrire $P\epsilon$ où P est la matrice d'une projection orthogonale à préciser.

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

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

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

Question 14

Déterminer l'espérance et la matrice de variance de $Y - X\hat{\beta}$.

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

Question 15

En déduire que $\hat{\sigma}^2$ est un estimateur sans biais de σ^2.

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

Question 16

Montrer que $(n-3)\hat{\sigma}^2/\sigma^2 \sim \Xi(n-3)$ et $\hat{\sigma}^2$ indépendant de $\hat{\beta}$

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

Calculons $e^t e$

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

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 degrés de liberté $rang(I_n - H)$.

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

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

Best Linear Unbiased Estimator (BLUE)

Question 17

Interpréter la propriété $MSE_\lambda(\bar{\beta}) \leq MSE_\lambda(\tilde{\beta})$

Question 18

Montrer que, si $\tilde{\beta}$ est sans bias, alors $MSE_\lambda(\tilde{\beta}) = VAR[\lambda^T \tilde{\beta}]$. En déduire que $\bar{\beta}$ est le BLUE si et seulement si, pout tout estimateur linéaire sans biais $\tilde{\beta}$, $Var(\tilde{\beta}) - Var(\bar{\beta})$ est une matrice positive.

Question 19

En écrivant $\tilde{\beta} = \hat{\beta} + DY = ((X^T X)^{-1} X^T + D)Y$, montrer $DX = 0$, puis $Var(b\tilde{e}ta) - Var(\hat{\beta})$ est positive. Conclure.