Séance 1

Guillaume Metzler

2/3/2022

Abstract

On va traiter les deux premiers exercices de la fiche de TD 1 qui portent sur la régression linéaire simple.

Exercice 1

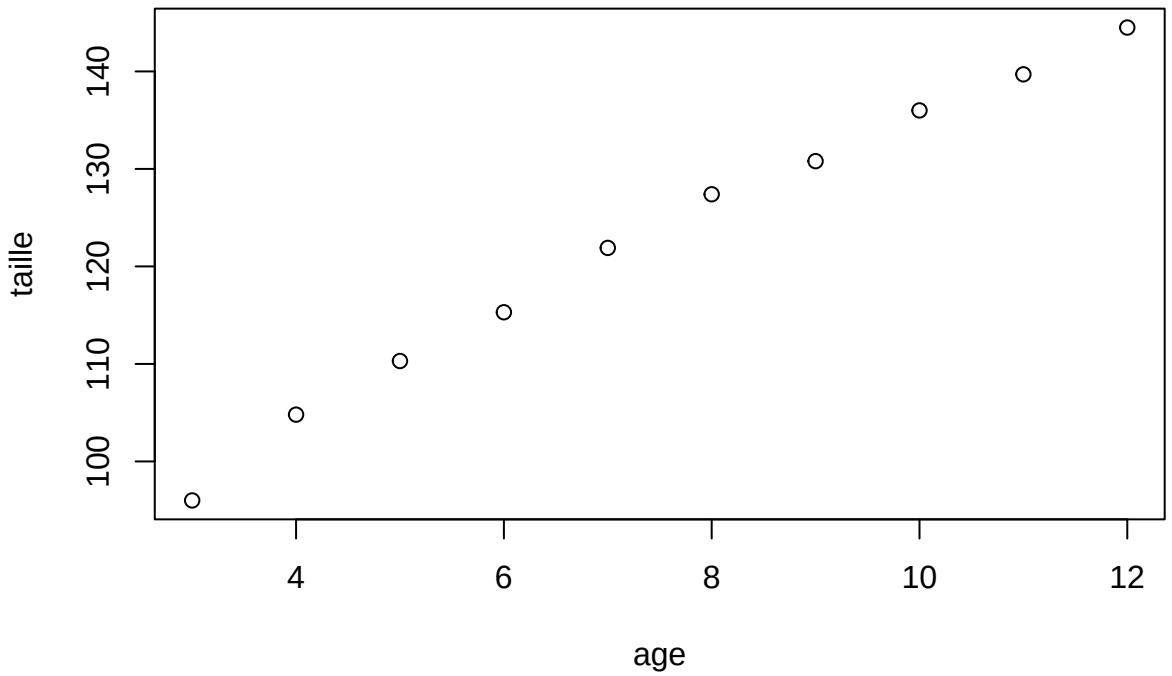
On s'intéresse à un modèle linéaire gaussien simple

$$Y = aX + b + \varepsilon$$

a : coefficient de la droite b : ordonnée à l'orgine $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ est le bruit présent dans les données. Ce bruit est supposé indépendant entre les individus et la variance est la même (homoscédasticité).

Question 1

```
taille <- c(96, 104.8,110.3,115.3,121.9,127.4,130.8,136,139.7,144.5)
age <- c(3:12)
plot(age,taille)
```



Les poins sont alignés, ce qui justifie bien l'utilisation d'un modèle linéaire (gaussien)

Question 2 :

On peut estimer les paramètres de la droite de régression par méthode des moindres carrés ordinaires, et on obtient les estimations suivantes des paramètres a et b

$$\hat{a} = \frac{Cov[X,Y]}{Var[X]} \underset{ech}{=} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

et

$$\hat{b} = \mathbb{E}[Y] - \hat{a}\mathbb{E}[X] \underset{ech}{=} \frac{1}{n}\sum_{i=1}^{n} x_i - \hat{a} \times \frac{1}{n}\sum_{i=1}^{n} y_i.$$

On rappelle que les valeurs de a et b sont obtenus en résolvant le problème suivant :

$$\min_{a,b} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min_{a,b} \sum_{i=1}^{n}(y_i - (ax_i + b_i))^2.$$

Vous avez pu voir que cela vous ammène à résoudre un système linéaire de deux équations à deux inconnues.

Regardons comment les estimer sur R

```
a_hat = cov(age,taille)/var(age)
b_hat = mean(taille) - a_hat*mean(age)

print(paste("Le coefficient directeur de la droite est ",a_hat))
```

```
## [1] "Le coefficient directeur de la droite est  5.22"
```

```
print(paste("L'ordonnée à l'origine de la droite est ",b_hat))
```

```
## [1] "L'ordonnée à l'origine de la droite est  83.52"
```

Limites du modèles

Le modèle n'est valable que sur l'espaces des valeurs observées pour l'âge.
Une taille théorique de 83.52cm à la naissance.

```
# on cherche à prédire la taille en fonction de l'âge.
mymodel <- lm(taille~age)
mymodel$coefficients
```

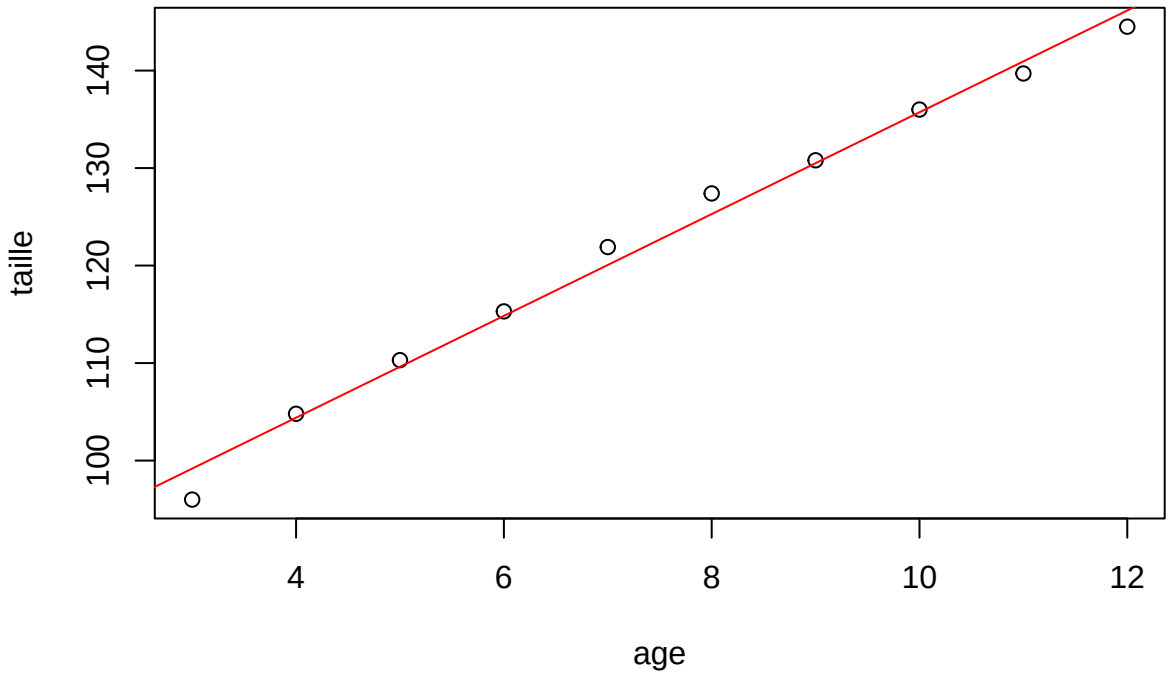
```
## (Intercept)         age
##       83.52        5.22
```

On a un modèle qui prédit la taille en fonction de l'âge par la relation

$$\text{taille} = 5.22 \times \text{âge} + 83.52.$$

Ajout de la droite de régression sur la graphe précédent

```
plot(age, taille)
abline(mymodel$coefficients, col ="red")
```



Question 3

On rappelle que les résidus de la régression $(\varepsilon_i)_{i=1}^n$ son définis par

$$\varepsilon_i = y_i - \hat{y}_i,$$

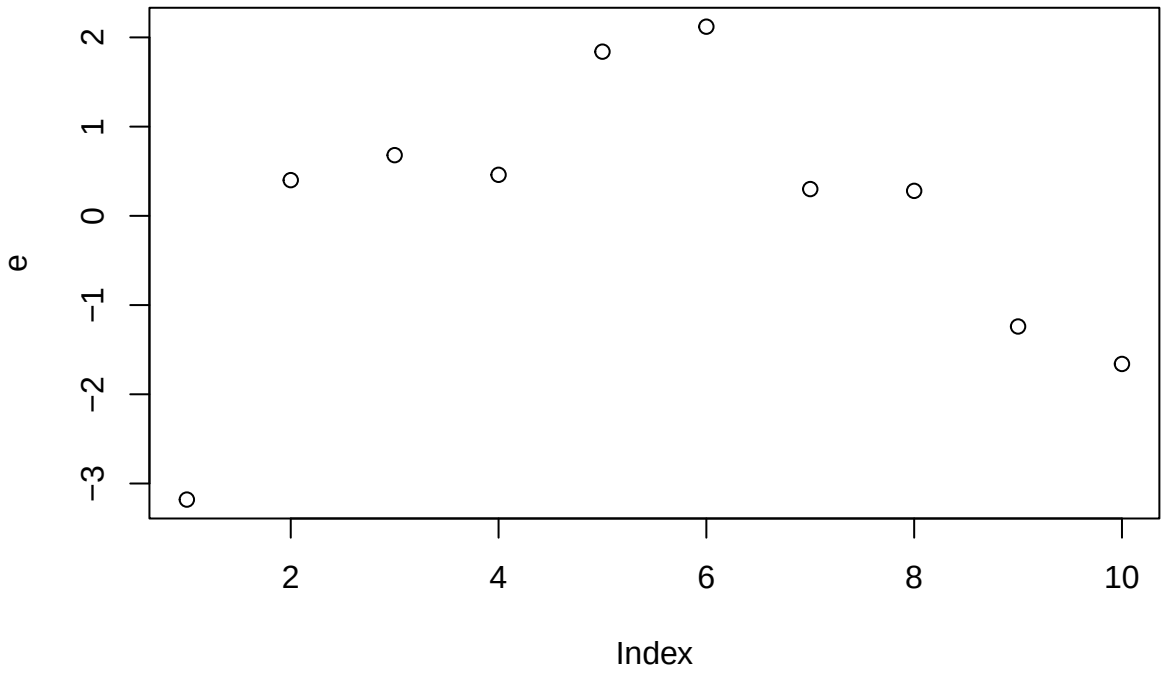
où $\hat{y}_i = \hat{a} \times x_i + \hat{b}$.

```
a_hat <- mymodel$coefficients[2]
b_hat <- mymodel$coefficients[1]

# Calcul des valeurs prédites
y_hat <- a_hat*age + b_hat

# Calcul des résidus
e = taille - y_hat

# Graphe des résidus
plot(e)
```



Le modèle n'est ici a priori pas valable, les résidus ne sont pas normalement distribués : absence de symétrie, il y a plus de valeurs positives.

Exerice 2 :

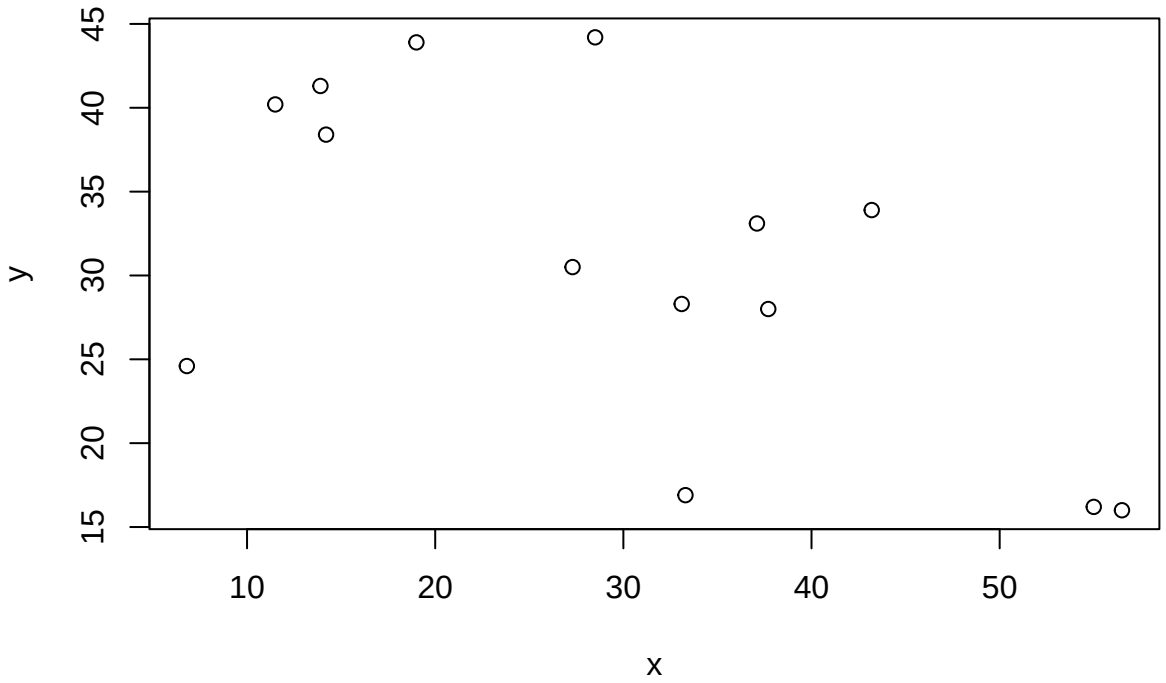
On fait l'hypothèsed d'un modèle de régression linéaire simple de la forme

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Question 1

```
# taux de natalité
y <- c(16.2,30.5,16.9,16,40.2,38.4,41.3,43.9,28.3,33.9,44.2,24.6,28,33.1)
# taux d'urbanisation
x <- c(55,27.3,33.3,56.5,11.5,14.2,13.9,19,33.1,43.2,28.5,6.8,37.7,37.1)
plot(x,y)
```



Chercher à estimer le taux de natalité uniquement en fonction du taux d'urbanisation risque de ne pas être pertinent car les points ne sont pas alignés.

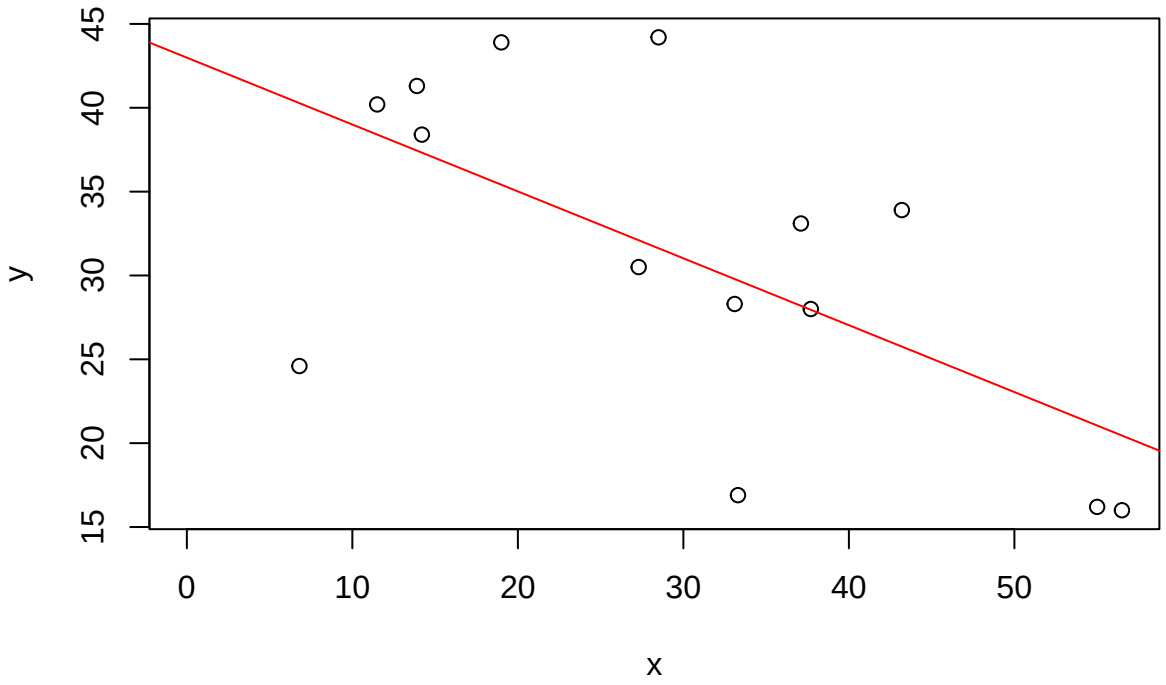
```
# Régression linéaire
mymodel <- lm(y~x)
mymodel$coefficients
```

```
## (Intercept)           x
##  42.9905457  -0.3988675
```

Cette fois-ci, on a un modèle de la forme

$$y_i = -0.40 \times x_i + 42.99.$$

```
# Représentation graphique
plot(x,y, xlim=c(0,max(x)))
abline(mymodel$coefficients, col = "red")
```



Question 3

On nous demande de déterminer :

$$sse = \sum_{i=1}^{n} \varepsilon_i^2.$$

```
# On peut obtenir les résidus directement comme suit
e <- mymodel$residuals

# Calculer la somme du carré des résidus
sse <- sum(e^2)

# Rappel :

# e^2 : calcul le carré des composantes du vecteur e
# sum(e) : fait la somme des éléments du vecteur e
```

Question 4 :

Je vous rappelle que l'on considère un modèle de la forme

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon.$$

On cherche à déterminer si β_1 est différent de 0. On va essayer de construire le test qui nous permet de vérifier cela.

$$\hat{\beta_1} \sim \mathcal{N}\left(\beta_i, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

On montre que l'on peut estimer σ^2 par la relation

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2.$$

On va considérer une autre variable aléatoire

$$\frac{(n-2)\hat{\sigma}^2}{\sigma} \sim \mathcal{X}^2_{n-2}.$$

In fine, on va définir notre statistique de test qui suit une loi de Student à $n-2$ de degrés de liberté. Pour rappel une loi de Student

$$T_{n-2} = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\mathcal{X}^{\in}_{\setminus -\in}}{(n-2)}}}.$$

Construisons cette variable

$$T_{n-2} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}}{\sqrt{\frac{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}}{n-2}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}.$$