

Correction Lab 4

Guillaume Metzler

11/23/2021

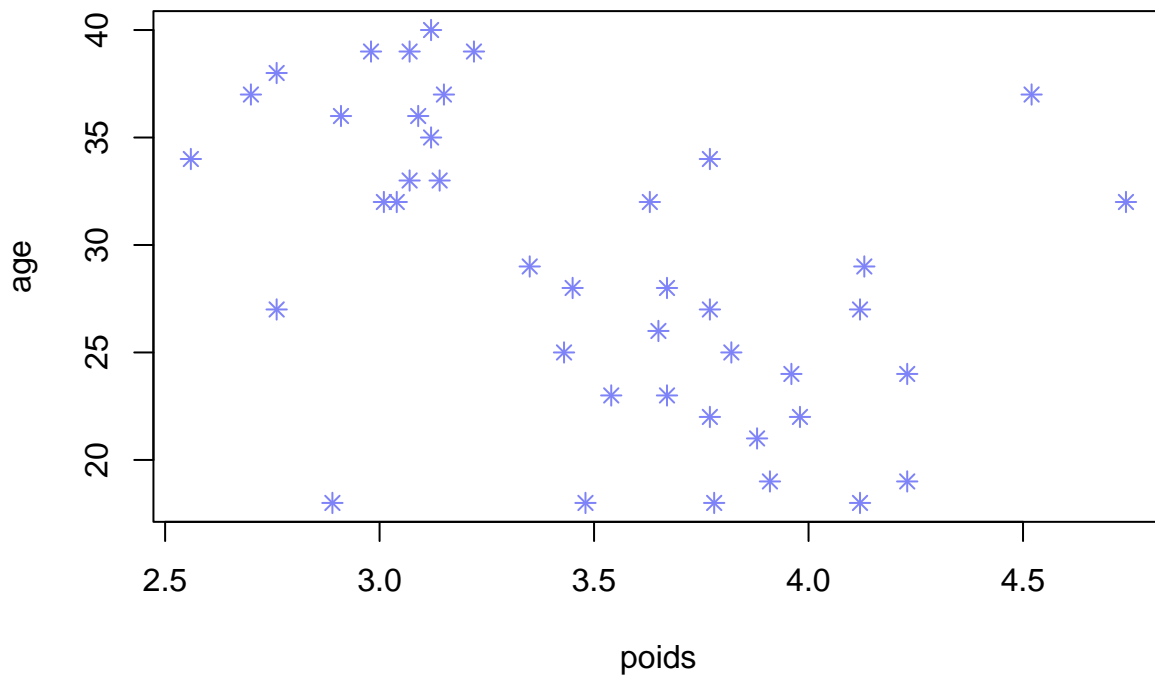
```
library(readxl)
data <- read_excel("enfants.xlsx")
colnames(data)= c("poids", "age")
```

Question a)

On va d'abord regarder ce que donne nos données, en faisant une représentation graphique sur laquelle le poids du nouveau-né est donné en fonction de l'âge de la mère le jour de l'accouchement.

```
plot(data, main = "Représentation graphique des données", pch = 8, col = "#7e83f7")
```

Représentation graphique des données



On peut observer une tendance qui est globalement négative, *{i.e.} le poids du nouveau-né est d'autant plus faible que l'âge de la mère est élevé au moment de la naissance de l'enfant.*

On cherche maintenant à construire un modèle qui approximera le mieux cette tendance que l'on observe les données, à l'aide d'un modèle linéaire simple. Des commandes permettent de faire cela simplement à l'aide du logiciel. C'est ce que nous allons faire dans un premier temps, puis nous chercherons à retrouver les paramètres du modèle.

Pour cela, on rappelle que l'on va chercher à estimer les paramètres à l'aide d'un modèle dit Gaussien, qui va prendre la forme suivante

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

où Y représente la variable réponse, c'est-à-dire le poids du nouveau-né, X la variable explicative, c'est-à-dire celle qui va nous servir à expliquer les valeurs de Y , c'est l'âge de la mère. Les paramètres a et b sont les paramètres de notre droite et ε est ce que l'on appelle un "bruit blanc", $\varepsilon_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ où la variance de la loi normale σ^2 est inconnue.

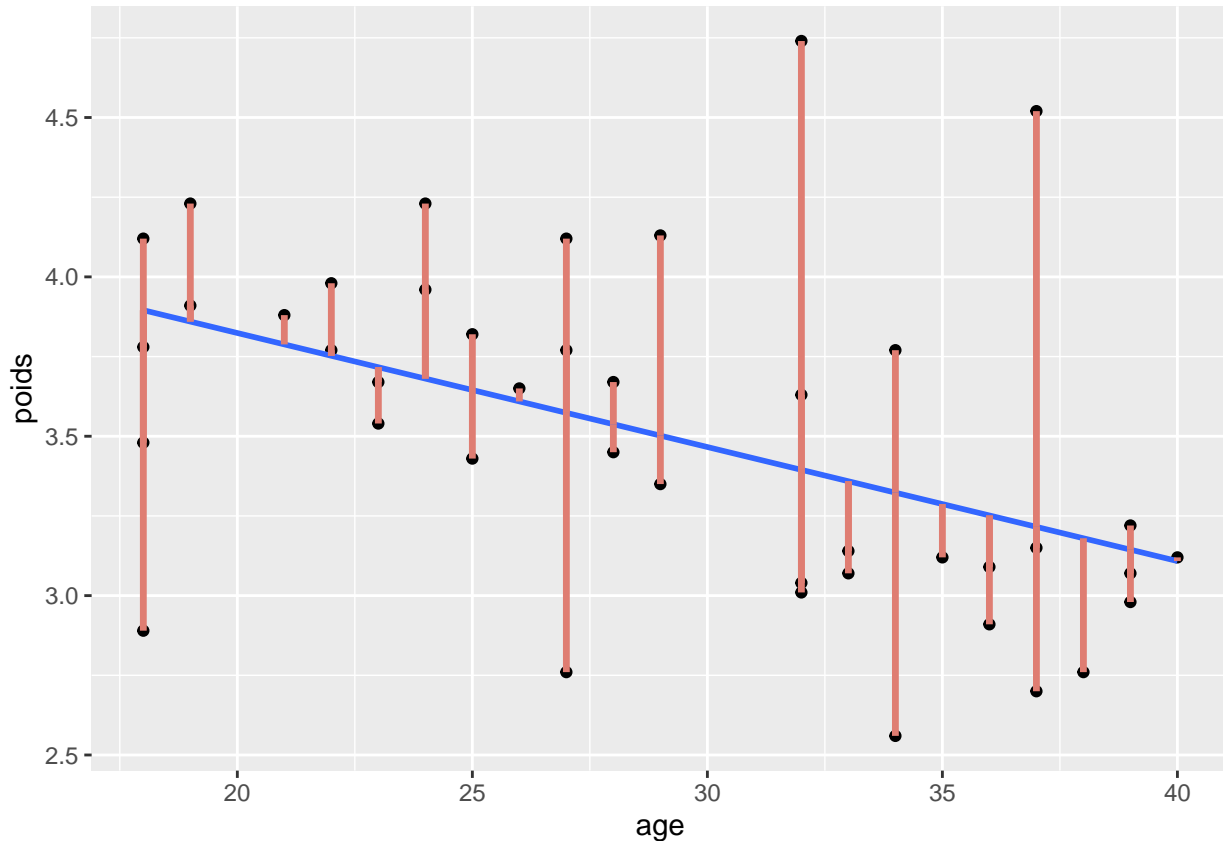
On peut alors estimer les paramètres de la droite qui estime le mieux le nuage de points de la façon suivante et ajouter cette droite sur notre graphique précédent

```
# Estimation des paramètres du modèle
my_lm <- lm(poids~age, data = data)
coeff <- my_lm$coefficients
names(coeff) = c("poids", "age")
coeff
```

```
##      poids      age
## 4.53948929 -0.03577658
```

```
# Représentation graphique de la droite de régression
library(ggplot2)
ggplot(data, aes(x=age, y=poids)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE) +
  geom_segment(aes(x = age, y = poids, xend = age,
                  yend = coeff[1] + coeff[2]*age, col = "Residuals"),
              col = "#DF7D72", lwd= 1.2, data = data)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



L'estimation effectuée nous donne les valeurs suivantes pour notre modèle

$$a = -0.0358 \quad \text{et} \quad b = 4.5395.$$

Ces coefficients sont obtenus par la méthode des "moindres carrés ordinaires (MCO)". Cela signifie qu'ils sont obtenus en résolvant le problème de minimisation suivant :

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

Question b)

Dans la question précédente, la valeur de b (ordonnée à l'origine) nous donne la masse de référence du nouveau si l'âge de la mère était de 0 hypothétiquement. La valeur de a nous donne alors une indication de la tendance sur l'évolution de la masse du nouveau né en fonction de l'âge de la mère. Dans le cas présent, la masse du nouveau né a tendance à décroître de 35.8g par année de la mère.

Question c)

On peut reprendre les coefficients obtenus à la question a) afin de déterminer le poids d'un nouveau-né dont la mère aurait 34 ans. Le poids du nouveau né \hat{y} serait alors égal à :

$$\hat{y} = a * 34 + b,$$

avec les valeurs de a et b estimées. Numériquement, nous obtenons la valeur suivante :

```
# Poids du nouveau-né d'une mère de 34 ans
```

```
poids_newborn <- coeff[1] + 34*coeff[2]  
poids_newborn
```

```
##      poids
```

```
## 3.323085
```

Le poids hypothétique du nouveau né serait alors de 3.323 kg.

Question d)

Dans le cas de la régression linéaire simple, le coefficient de détermination R^2 n'est rien d'autre que le carré de la valeur du coefficient de corrélation de Pearson ρ . Ainsi

$$R^2 = \rho^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^n (x_i - \bar{x})^2) (\sum_{i=1}^n (y_i - \bar{y})^2)}} \right)^2.$$

On rappelle que le coefficient de détermination permet de juger de la qualité du modèle de régression, afin de savoir si ce dernier a un sens d'un point de vue statistique. Il va chercher à regarder si la variable x permet bien d'expliquer les valeurs observées pour la variable y .

Regardons cela maintenant d'un point de vue numérique pour savoir ce qu'il en est.

```
# Coefficient de corrélation  
rho <- cor(data$poids, data$age)  
rho
```

```
## [1] -0.4720261
```

Ici on trouve un coefficient de corrélation qui est plutôt négatif, ce qui est cohérent avec la tendance observée dans les données mais aussi avec le coefficient directeur de la droite de notre modèle.

```
# Coefficient de détermination  
R_square <- rho^2  
R_square
```

```
## [1] 0.2228086
```

Le coefficient de détermination est relativement faible (proche de zéro) on peut donc se demander si notre modèle est vraiment significatif, s'il a un sens. C'est-ce que nous allons chercher à déterminer dans les prochaines questions.

Question e)

On cherche maintenant à savoir si le modèle est significatif ou non. Pour cela, on va procéder à un test statistique pour étudier si la valeur du coefficient de corrélation ρ est significative ou non.

Dans un modèle linéaire simple, tester la significativité du modèle (c'est-à-dire si les deux paramètres du modèle sont tous les deux non nuls), revient au même que de tester la significativité de la pente du modèle (c'est-à-dire le fait que le paramètre a du modèle est significativement différent de 0).

Pour faire cela, on étudie la quantité statistique $t_{\bar{a}}$ sous l'hypothèse H_0 : le coefficient a (la pente) est égal à 0

$$t_{\bar{a}} = \frac{\bar{a} - 0}{\sigma_{\bar{a}}},$$

où $\sigma_{\bar{a}}$ est l'écart-type de la distribution d'échantillonnage lié à l'estimateur de pente. Il nous reste donc à estimer la valeur de $\sigma_{\bar{a}}$.

Commençons d'abord par montrer que \hat{a} est un estimateur sans biais de a , c'est-à-dire que $\mathbb{E}[\hat{a}] = a$. On utilisera le fait que

- $\mathbb{E}[y_i] = ax_i + b$
- $\mathbb{E}[\bar{y}] = a\bar{x} + b$

$$\hat{\mathfrak{D}} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}[Y_i - \bar{Y}]}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (1)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x}) (\mathbb{E}[Y_i] - \mathbb{E}[\bar{Y}])}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (2)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x}) (ax_i + b - a\bar{x} - b)}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (3)$$

$$= a \frac{\sum_{i=1}^n (x_i - \bar{x}) (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (4)$$

$$= a. \quad (5)$$

Question g)

Question h)

On