Correction Lab 4

Guillaume Metzler

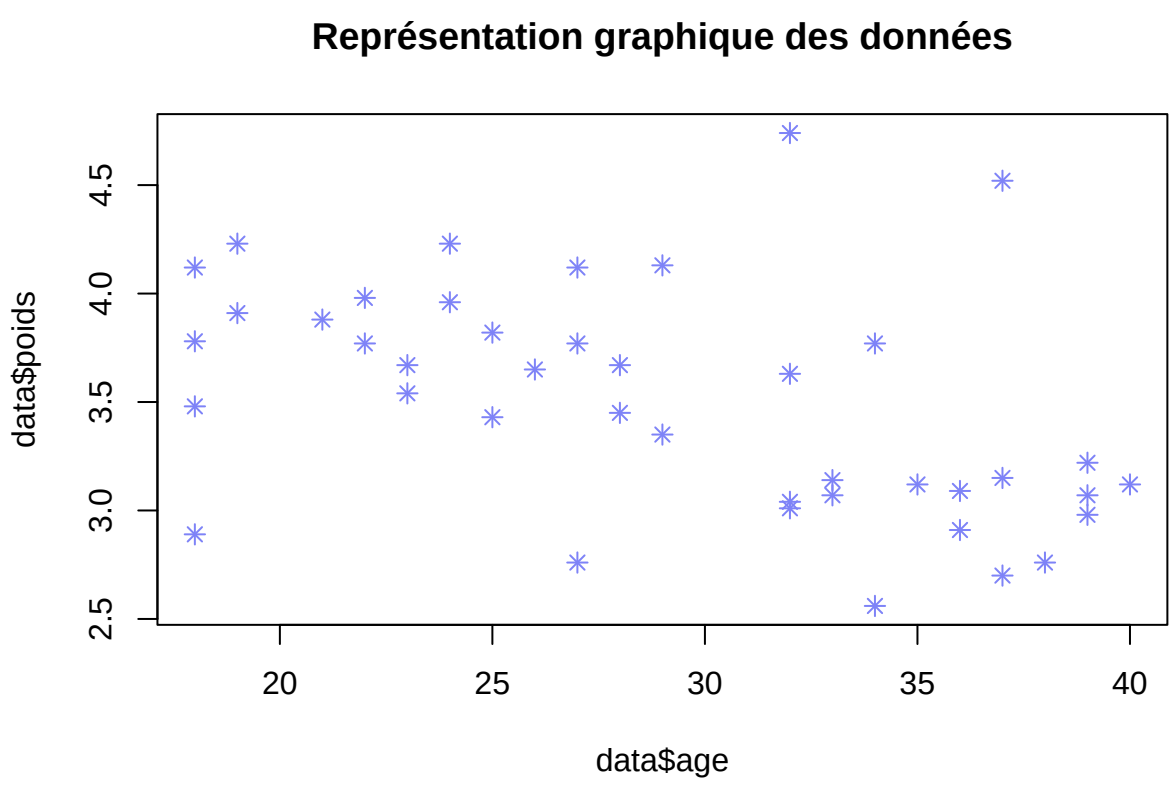
11/23/2021

```
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$age,data$poids, main = "Représentation graphique des données", pch = 8, col = "#7e83f7")
```



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 sint 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.

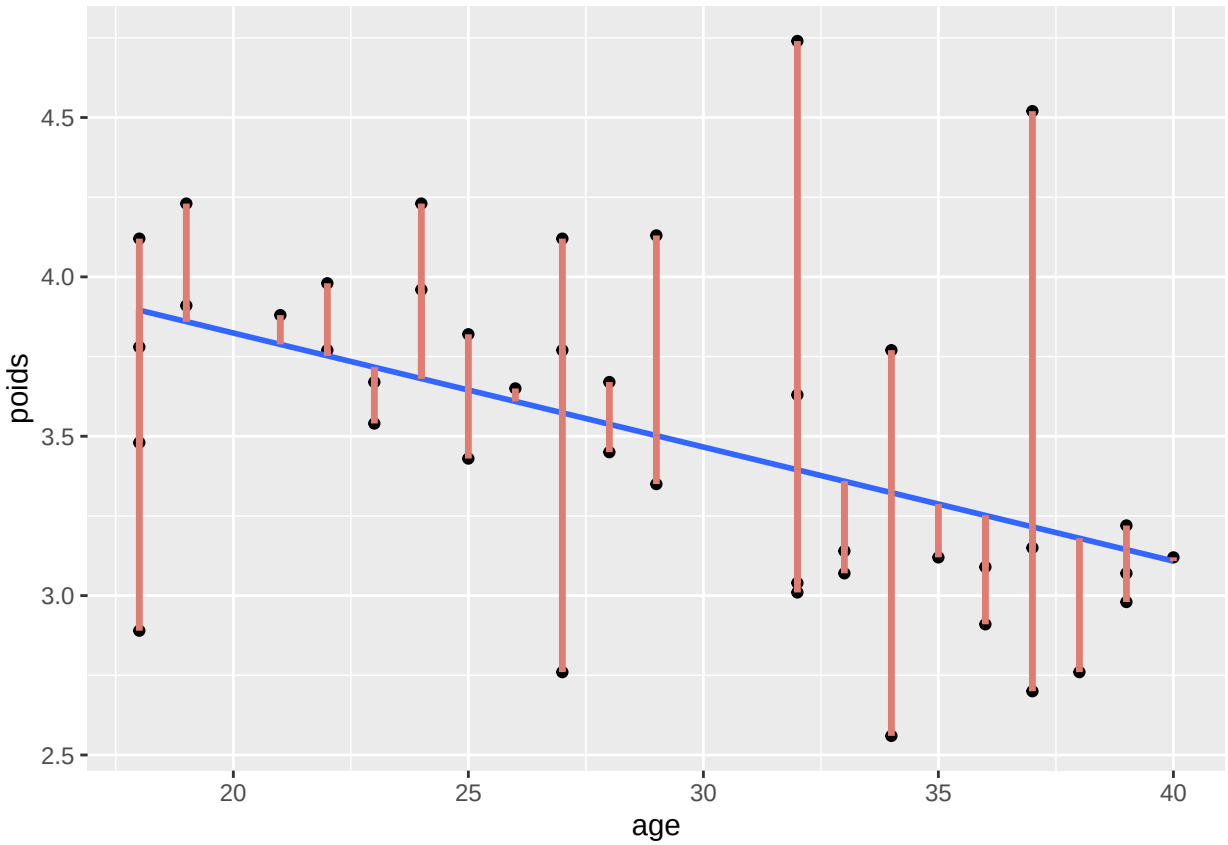
On peut alors estimer les paramètres de la droite qui estime le mieux le nuage de points de la façon suivante et ajouter cette droite sur notre graphique précédent

```
# 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

$$a = -0.0358 \quad et \quad b = 4.5395.$$

Ces coefficients sont obtenus par la méthodes des "moindres carrés ordinaires (MCO)". Cela signfie 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 (ordonné 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 à tendance à décroitre de 35.8g par année de la mère.

Question c)

On peut reprendre les coefficient 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 :

```r
# 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*

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

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 maitenant d'un point de vue numérique pour savoir ce qu'il en est.

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

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

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

```r
# 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 étudier 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

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

$$\begin{aligned}
\mathbb{E}[\hat{a}] &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})\mathbb{E}[y_i - \bar{y}]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\mathbb{E}[y_i] - \mathbb{E}[\bar{y}])}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(ax_i + b - a\bar{x} - b}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\
&= a\frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\
&= a.
\end{aligned}$$

On peut maintenant faire de même avec la variance de l'estimateur afin de déterminer son écart-type, ce qui nous servira à tester la significativité de la pente, mais aussi à constuire l'intervalle de confiance sur l'estimation du paramètre.

Pour cela on utilisera le fait que l'on peut écrire :

$$\hat{a} = a + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = a + \sum_{i=1}^{n}\omega_i x_i.$$

Cette relation découle des hypothèses du modèle de linéaire. Ce qui nous donne :

$$
\begin{aligned}
Var[\hat{a}] &= \mathbb{E}\left[(\hat{a} - \mathbb{E}[\hat{a}])^2\right], \\
&= \mathbb{E}\left[(a + \sum_{i=1}^{n} \omega_i x_i - a)^2\right], \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} \omega_i x_i\right)^2\right], \\
&= \mathbb{E}\left[\sum_{i=1}^{n} (\omega_i x_i)^2 + 2\sum_{i<i'}^{n} \omega_i \omega_{i'} x_i x_{i'}\right], \\
&= \sum_{i=1}^{n} \underbrace{\mathbb{E}\left[\omega_i^2\right]}_{=Var[\varepsilon_i]=\sigma^2} x_i^2 + 2\sum_{i<i'}^{n} \underbrace{\mathbb{E}\left[\omega_i \omega_{i'}\right]}_{=0} x_i x_{i'}, \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
\end{aligned}
$$

La deuxième somme est nulle, c'est l'hypothèse d'indépendance entre les bruits pour les différentes données.

Dans cette expression σ^2 reste inconnue, mais ce n'est pas grave, car on est en mesure de l'estimer ! En effet, on se rappelle qu'une estimation de σ^2, notée $\hat{\sigma}^2$ est très proche de la variance de nos résidus. Plus précisément :

$$
\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .
$$

Au final notre variance de l'estimateur $\hat{a}$ est alors donnée par :

$$
\frac{\frac{1}{n-2}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
$$

On peut alors définir notre statistique de test t par la relation habituelle "estimateur moins son espérance, le tout diviser par son écart-type", i.e.

$$
t = \frac{\hat{a} - a}{\sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \underset{sous\ H_0}{=} \frac{\hat{a}}{\sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}.
$$

Cette dernière peut également s'écrire

$$
t = \frac{\rho}{\sqrt{\frac{1-\rho^2}{n-2}}}.
$$

Cette statistique de test suit une loi de Student à $n-2$ degrés de liberté. Pourquoi $n-2$? Cela correspond tout simplement à la taille de l'échantillon moins le nombre de paramètres à estimer dans le modèle.

On va maintenant regarder si notre modèle est significatif, au risque de première espèce $\alpha = 5\%$, on va donc rejeter l'hypothèse H_0 si $|t| \geq t_{1-\alpha/2}$.

```
# Calcul de la statistique de test
n = length(data$poids)
t <- rho/sqrt((1-rho^2)/(n-2))

# Calcul de la valeur critique

t_crit <- qt(0.975,40)
abs(t) > t_crit
```

```
## [1] TRUE
```

```
# On rejette donc l'hypothèse selon laquelle notre coefficient directeur
# n'est pas significativement différent de zéro.

# Calcul de la p-value éventuellement

2*(1-pt(abs(t),40))
```

```
## [1] 0.001599646
```

```
# La p-value est bien inférieure à 0.05, on rejette donc l'hypothèse nulle.
# Notre modèle est donc ben significatif.
```

Question f)

La définition de l'intervalle de confiance découle directement de la définition de la statistique de test précédente. Ainsi un intervalle de confiance de niveau $1-\alpha$ pour l'estimation du paramètre a de la régression est donnée par :

$$\left[\hat{a} - t_{1-\alpha/2,n-2}\sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \hat{a} + t_{1-\alpha/2,n-2}\sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right].$$

```
# Calcul de l'écart-type associé à l'estimateur de a

a_hat <- coeff[2]
sigma_a = sqrt((1/(n-2))*sum(my_lm$residuals^2)/((n-1)*var(data$age)))

# Borne inférieure
print("La borne inférieure de notre intervalle est donnée par")
```

```
## [1] "La borne inférieure de notre intervalle est donnée par"
```

```
borne_inf <- a_hat - qt(0.975,40)*sigma_a
borne_inf
```

```
##         age
## -0.05712912
```

```
# Borne supérieure
print("La borne supérieure de notre intervalle est donnée par")
```

```
## [1] "La borne supérieure de notre intervalle est donnée par"
```

```r
borne_sup <- a_hat + qt(0.975,40)*sigma_a
borne_sup
```

```
##         age
## -0.01442405
```

Question h)

Bien que la corrélation entre les deux variables soit relativement faible, elle reste significative, comme le montre le calcul de la p-value.