

Modèles Linéaires

Corrections Séance 1 Licence 3 MIASHS (2021-2022)

Guillaume Metzler

Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC EA3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

1 Exercice 1

Graphes des données et hypothèses du modèle


On rappelle qu'un modèle linéaire gaussien dit simple est un modèle de la forme

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

- Y représente la taille de l'individu
- X représente son âge
- les coefficients a et b représentent respectivement le coefficient directeur de la droite (i.e. la pente) et l'ordonnée à l'origine de la droite
- ε est un terme d'erreur, il représentera plus tard les erreurs du modèle, un bruit (blanc).

On rappelle également les hypothèses du modèle linéaire simple gaussien

- $(y_i, x_i)_{i=1}^n$ doivent être *i.i.d.*, i.e. indépendantes et identiquement distribuées,
- $y_i \sim \mathcal{N}(ax_i + b, \sigma^2)$
- $\varepsilon_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$: hypothèse d'homoscédasticité.

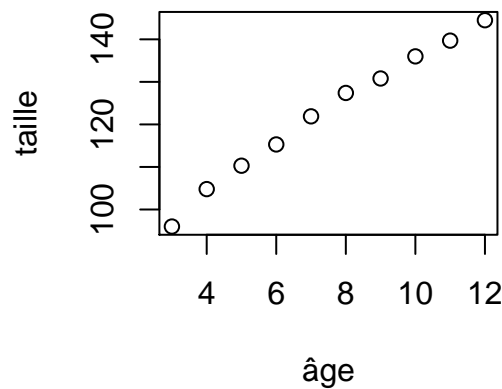
Représentons graphiquement les données à l'aide du code  suivant

```

# Un vecteur y qui représente la taille de l'individu
y = c(96,104.8,110.3,115.3,121.9,127.4,130.8,136,139.7,144.5)
# Un vecteur x qui représentera l'âge de l'individu
x = c(3:12)
# Mise sous format data.frame
data <- data.frame(age = x, taille = y)

# Graphe
plot(x,y, xlab = "âge", ylab = "taille")

```



Les points semblent distribués sur une droite, il paraît donc tout à fait pertinent de chercher à prédire la taille d'un individu en fonction de son âge.

Apprentissage de la droite de régression

On rappelle que la droite qui approxime le mieux le nuage de points est donné en résolvant le problème d'optimisation suivant :

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

Et les solutions sont données dans le résultat suivant

Proposition 1.1: Problème de régression linéaire

On considère le problème de régression linéaire gaussien de la forme

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

Les paramètres a et b sont solutions du problème d'optimisation

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

Dont les solutions sont donnés par

$$\hat{a} = \frac{Cov[X, Y]}{Var[X]} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad \hat{b} = \mathbb{E}[Y] - \mathbb{E}[X] \times \hat{a} = \bar{y} - \hat{a} \times \bar{x}.$$

Démonstration. La fonction L que l'on cherche à optimiser est une fonction convexe en les variables a et b , elle admet donc une unique solution. Cette solution est obtenue en résolvant l'équation d'Euler se présentant sous la forme d'un système linéaire

$$\frac{\partial L}{\partial a} = 0 \iff -2 \sum_{i=1}^n (y_i - ax_i - b)x_i = 0, \quad (1)$$

$$\frac{\partial L}{\partial b} = 0 \iff -2 \sum_{i=1}^n (y_i - ax_i - b) = 0. \quad (2)$$

On se concentre sur l'équation (2) pour le moment, on va la développer ce qui nous permet d'écrire

$$\begin{aligned} -2 \sum_{i=1}^n (y_i - ax_i - b) &= 0 \\ \iff \sum_{i=1}^n y_i - a \sum_{i=1}^n x_i - nb &= 0 \\ \iff \frac{1}{n} \sum_{i=1}^n y_i - a \frac{1}{n} \sum_{i=1}^n x_i &= b \\ \iff \bar{y} - a\bar{x} &= b \end{aligned}$$

On obtient une expression de l'estimateur \hat{b} de b , elle dépend de a dont on va pouvoir déterminer l'expression en injectant la valeur de b dans l'équation (1)

$$\begin{aligned}
& -2 \sum_{i=1}^n (y_i - ax_i - b)x_i &= 0 \\
\iff & \sum_{i=1}^n y_i x_i - a \sum_{i=1}^n x_i^2 - b \sum_{i=1}^n x_i &= 0 \\
\iff & \sum_{i=1}^n y_i x_i - a \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \left(\frac{1}{n} \sum_{i=1}^n y_i - a \frac{1}{n} \sum_{i=1}^n x_i \right) &= 0 \\
\iff & \sum_{i=1}^n y_i x_i - \frac{1}{n} (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i) - a \left(\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 \right) &= 0 \\
\iff & \frac{\sum_{i=1}^n y_i x_i - \frac{1}{n} (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} &= a \\
\iff & \frac{Cov[x, y]}{Var[x]} &= a
\end{aligned}$$

Ce qui achève la démonstration.

□


Ainsi les paramètres de la régression peuvent être calculés à partir des données comme suit

```
# Caclul des estimateurs de notre droite
# R calcule les covariances (et variances) débiaisées.
a_hat <- cov(x,y)/var(x)
b_hat <- mean(y) - a_hat*mean(x)
print(paste("Le coefficient directeur de ma droite est ", a_hat))

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

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

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

On peut apprendre les paramètres du modèle de la régression à l'aide de la fonction *lm* de .

```
mymodel <- lm(taille~age,data)
coeff <- mymodel$coefficients
coeff
```

## (Intercept)	age
## 83.52	5.22

Calcul et représentation des résidus

On se rappelle qu'un modèle linéaire gaussien est défini par

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

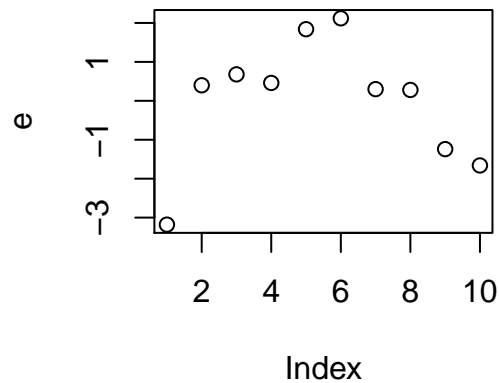
On se propose de calculer les erreurs d'estimation de notre modèle, i.e. on va déterminer les valeurs de ε_i pour tout i

$$\varepsilon_i = y_i - \hat{y}_i = y_i - (\hat{a}x_i + \hat{b}).$$

Commençons par regarder comment calculer manuellement les résidus. Il faut d'abord calculer les valeurs estimées par le modèle de régression

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

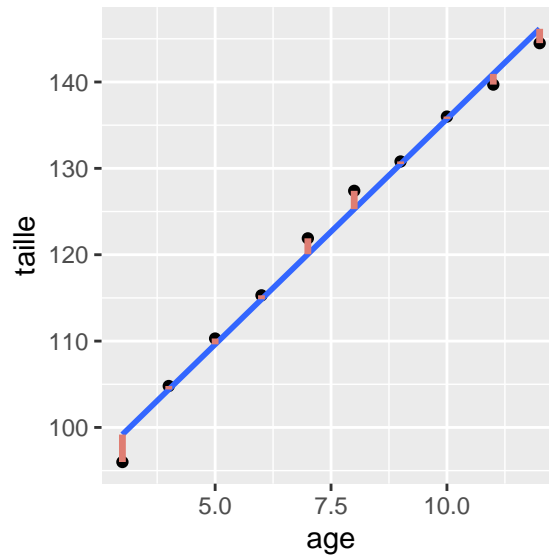
```
a_hat <- coeff[2]
b_hat <- coeff[1]
# Valeurs estimées par le modèle
y_hat <- a_hat*x + b_hat
# Résidus ou erreurs du modèle
e = y - y_hat
e_bis <- mymodel$residuals
plot(e)
```



On peut aussi visualiser graphiquement les résidus sur la droite de régression directement

```
# On peut extraire les coefficients de la régression comme suit

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



Enfin, on pourrait simplement représenter graphiquement la droite de régression comme suit

```
# Régression linéaire
mymodel <- lm(y~x)
plot(x,y)
abline(mymodel$coefficients, col="red")
```

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.

Pour ce qui est de l'interprétation du modèle, il faut se restreindre à l'échelle des valeurs observées sur notre échantillon.

2 Exercice 2

Représentation des données et estimation du modèle

On commence par entrer nos données

```
# 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)

data <- data.frame(tu = x, tn = y)

```

On garde la même typologie de modèle qu'à l'exercice précédent

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

dont on va chercher à estimer les paramètres a et b . On procédera exactement de la même façon qu'à l'exercice précédent en représentant aussi directement les résidus sur le graphique comprenant la droite de régression.

```

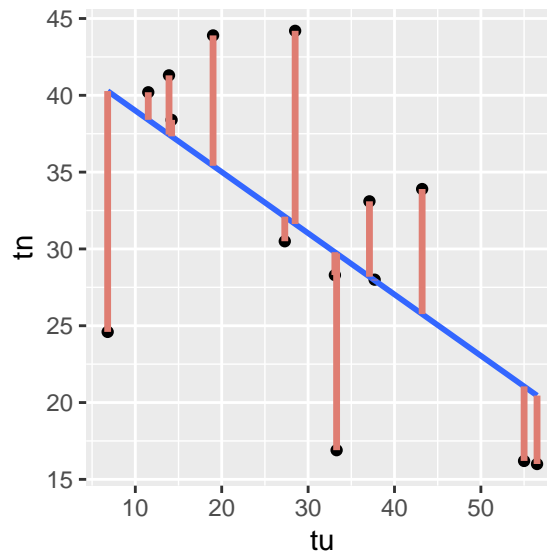
# Apprentissage du modèle
mymodel <- lm(tn~tu,data)
coeff <- mymodel$coefficients
coeff

## (Intercept)          tu
## 42.9905457 -0.3988675

# Représentation graphique de la droite de régression et résidus

library(ggplot2)
ggplot(data, aes(x=tu, y=tn)) +
  geom_point() +
  geom_smooth(method=lm,se=FALSE) +
  geom_segment(aes(x = tu, y = tn, xend = tu,
                  yend = coeff[1] + coeff[2]*tu, col = "Residuals"),
              col = "#DF7D72", lwd= 1.2, data = data)

```

Significativité du modèle

L'énoncé nous invite tout d'abord à calculer le carré des résidus, ce que nous pouvons faire simplement à l'aide de

```
# Extraction des résidus
e <- mymodel$residuals
# Calcul de la somme des carrés des erreurs
sse <- sum(e^2)
```

On peut montrer que la somme (ou la moyenne) des résidus ε_i est nulle. Ainsi la somme des carrés des résidus, i.e. $\sum_{i=1}^n \varepsilon_i^2$ représente, à un facteur multiplicatif près, un terme de variance des résidus.

Significativité du modèle 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

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

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

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 construire l'intervalle de confiance sur l'estimation du paramètre.

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

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

$$\text{où } \omega_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

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

$$\text{Var}[\hat{a}] = \mathbb{E}[(\hat{a} - \mathbb{E}[\hat{a}])^2],$$

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

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 :

$$\text{Var}[\hat{a}] = \frac{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

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.

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

Cette dernière peut également s'écrire

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

où ρ représente le coefficient de corrélation linéaire de Pearson du modèle, *i.e.*

$$\rho = \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}} \in [-1, 1].$$

Ce coefficient de corrélation permet de mesurer l'intensité de la liaison entre deux variables X et Y tout en indiquant l'impact de l'une des variables sur l'autre. Si $|\rho|$ est proche de 1, on dira que la corrélation entre les deux variables est *forte*, à l'inverse, si elle est proche de 0, elle sera *faible*.

De plus, une valeur **négative** de ρ signifie que, globalement, valeurs *croissantes* de X impliquent des valeurs *décroissantes* de Y (et réciproquement), *i.e.* la pente de la droite de régression sera **négative**. De même, une valeur positive de ρ signifie que des valeurs *croissantes* de X impliquent des valeurs *croissantes* de Y (et réciproquement), *i.e.* la pente de la droite de régression sera **positive**.

La statistique de test précédente 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.

Une autre façon de construire notre statistique de test est d'utiliser le fait qu'une loi de Student est la donnée d'une loi normale centrée et réduite quotientée par la racine carrée d'une loi du χ^2 divisé par son nombre de degré de liberté.

On connaît la distribution de notre estimateur $\hat{a} \sim \mathcal{N}\left(a, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$.

Donc

$$\frac{\hat{a} - a}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathcal{N}(0, 1).$$

De plus, nous avons

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2,$$

ainsi

$$\frac{\frac{\hat{a} - a}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}}{\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \cdot \frac{1}{n-2}}} \underset{\text{sous } H_0}{=} \frac{\hat{a}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim T_{n-2}.$$

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 $|t| \geq t_{1-\alpha/2}$, puis construire un intervalle de confiance de niveau $1 - \alpha$ sur la paramètre a

```
# Calcul de la statistique de test
n = length(data$tu)
a_hat <- coeff[2]

t = a_hat/(sqrt((sse/(n-2))/((n-1)*var(data$tu))))
t

##          tu
## -2.745884

# On compare cette valeur au quantile d'ordre 0.975 d'une loi
# de student à $n-2$ degrés de liberté.

abs(t) > qt(0.975,n-2)

##      tu
## TRUE
```

Nous sommes donc amenés à rejeter l'hypothèse H_0 et on peut dire que notre est donc bien significatif, même si le nuage de points laissé à penser le contraire.

Regardons maintenant l'intervalle de confiance, ce dernier est donné par

$$I_{1-\alpha} = \left[\hat{a} - t_{1-\alpha/2, n-2} \sqrt{\text{Var}[\hat{a}]}; \hat{a} + t_{1-\alpha/2, n-2} \sqrt{\text{Var}[\hat{a}]} \right]$$

où les différents valeurs sont données par

```
# estimateur a_hat
coeff[2]

##          tu
## -0.3988675

# quantile de la loi de student
qt(0.975,n-2)
```

```
## [1] 2.178813

# variance de l'estimateur a_hat
(sse/(n-2))/((n-1)*var(x))

## [1] 0.02110052
```