

## Modèles Linéaires

### Correction Contrôle Continu 1 Licence 3 MIASHS (2021-2022)

Stéphane Chrétien & Guillaume Metzler

Université de Lyon, Université Lumière Lyon 2  
Laboratoire ERIC UR 3083, Lyon, France

[stephane.chretien@univ-lyon2.fr](mailto:stephane.chretien@univ-lyon2.fr) [guillaume.metzler@univ-lyon2.fr](mailto:guillaume.metzler@univ-lyon2.fr)

### Exercice 1 : Régression Linéaire Simple

Les données suivantes représentent le score en anglais sur 100 d'un échantillon de 12 personnes d'âge différent :

âge (années)	29	28	27	26	25	24	23	22	21	20	19	18
score	76.1	77	78.1	78.2	78.8	79.7	79.9	81.1	81.2	81.8	82.8	83.5

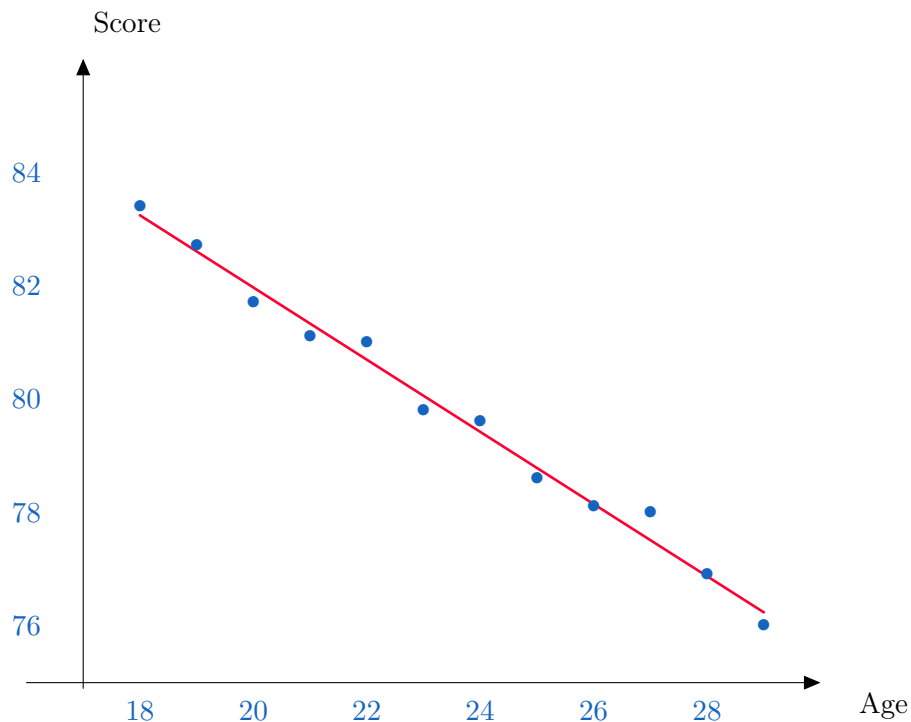
1. Tracer les points, la droite de régression et en déduire graphiquement l'équation de la droite de régression (on cherchera deux points visiblement proches de la vraie droite de régression ce qui permettra de calculer son équation, à une légère erreur près).

Commençons par représenter le nuage de points ainsi que la droite de régression et nous intéresserons à l'équation de la droite de la forme

$$y = \beta_0 + \beta_1 x$$

dans un deuxième temps.

On appelle  $Y$  la variable à estimer et  $X$  la variable explicative. Ici  $Y$  est le score et  $X$  représente l'âge des individus.



Pour déterminer l'équation de la droite, on va sélectionner deux points proches de la droite. Ici on va prendre les points

$$(25, 78.8) \quad \text{et} \quad (26, 78.2)$$

Ce qui conduit au système d'équations d'inconnus  $\beta_0$  et  $\beta_1$

$$78.8 = \beta_0 + 25\beta_1 \quad (1)$$

$$78.2 = \beta_0 + 26\beta_1 \quad (2)$$

En faisant l'opération (2)-(1), on trouve  $\beta_1 = -0.6$ .

En injectant le résultat dans l'équation (1), on trouve :


$$78.8 = \beta_0 - 25 \times 0.6, \quad \text{soit} \quad \beta_0 = 78.8 + 25 \times 0.6 = 93.8$$

2. Calculer une estimation des coefficients de la droite de régression pour ces données.

On rappelle que les coefficients de la droite de régression sont donnés par

$$\beta_1 = \frac{Cov[X, Y]}{Var[X]} \quad \text{et} \quad \beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X]$$

On rappelle aussi que la covariance peut s'exprimer comme *la moyenne des produits moins le produit des moyennes*.

Nous allons calculer ces quantités sous  directement.

```
y <- c(76.1,77,78.1,78.2,78.7,79.7,
       79.9,81.1,81.2,81.8,82.8,83.5)
x <- seq(29,18,-1)

n <- length(x)

# Moyenne de Y
my <- mean(y)

# Moyenne et variance de X
mx <- mean(x)
vx <- var(x)*(n-1)/n

# Moyenne des produits
mp <- mean(x*y)

# Produit des moyennes
pm <- mean(x)*mean(y)

# Covariance
covxy <- mp-pm

# Estimation des coefficients

beta1 <- covxy/vx
beta0 <- my - beta1*mx

beta1
## [1] -0.636014

beta0
## [1] 94.788
```

3. On se focalise maintenant sur le point de vue de la statistique inférentielle pour le modèle linéaire. Pour ce faire, on suppose

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

avec  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ .

(a) Quelle est la loi de  $Y_i$ ,  $i = 1, \dots, n$  ?

La variable aléatoire  $Y_i$  suit une loi normale de paramètres  $\beta_0 + \beta_1 X_i$  et  $\sigma^2$ , *i.e.*

$$\forall i = 1, \dots, n, \quad Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2).$$

Il a été vu au TD 1 que

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

De plus, la variance  $\sigma^2$  des résidus peut être estimée par

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

On en a déduit que

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \tau_{n-2}. \quad (3)$$

(b) Proposez un test afin de savoir si le coefficient est nul. Le résultat vous surprend-t-il ?

On va effectuer le test suivant :

$$H_0 : \beta_1 = 0 \quad \text{v.s.} \quad \beta_1 \neq 0.$$

Il suffit maintenant de calculer les différentes quantités qui entrent en jeu dans la statistique de test définie par l'équation (3).

```
data <- data.frame(y,x)
mymodel <- lm(y~x,data)

SCR = sum(mymodel$residuals^2)

# On peut alors calculer la statistique de test

t.beta1 = beta1/sqrt( (SCR/(n-2)) / (n*vx) )

t.beta1

## [1] -29.08731
```

On compare cette valeur (en valeur absolue !) au quantile d'ordre 0.975 d'une loi de Student à  $n - 2$  soit 10 degrés de liberté.

```
t.seuil = qt(0.975,10)

abs(t.beta1)>t.seuil

## [1] TRUE
```

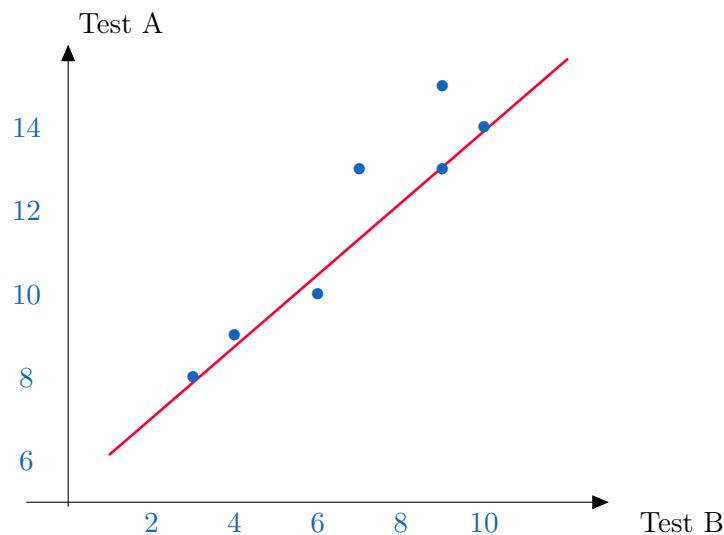
Ici on rejette sans surprise l'hypothèse  $H_0$  selon laquelle notre paramètre est nul. Ce résultat n'est pas surprenant étant donné l'ajustement de la droite à nos données.

## Exercice 2 : Droite de Régression et Points Atypiques

7 personnes sont inscrites à une formation à la conduite automobile. Au début de la formation, ces stagiaires subissent un test A notée sur 20. A la fin de la formation, elles subissent un test B de niveau identique. Les résultats sont donnés dans le tableau suivant :

Test A	3	4	6	7	9	10	9
Test B	8	9	10	13	15	14	13

1. Représenter le nuage de points.



2. La fonction "somme des carrés des erreurs" est donnée par

$$f(b) = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2, \quad (4)$$

comme fonction des coefficients  $b_0$  et  $b_1$ .

- (a) Déterminer la dérivée de cette fonction, lorsqu'elle est considérée comme fonction de  $b_1$  uniquement.

Le calcul de la dérivée nous donne :

$$\frac{\partial f}{\partial b_1}(b) = -2 \sum_{i=1}^n x_i (y_i - b_0 - b_1 x_i).$$

- (b) Déterminer la dérivée de cette fonction, lorsqu'elle est considérée comme fonction de  $b_0$  uniquement.

Le calcul de la dérivée nous donne :

$$\frac{\partial f}{\partial b_0}(b) = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i).$$

On peut faire le choix de réécrire les deux expressions précédentes en fonction de la moyenne des variables aléatoires  $X$  et  $Y$  ou encore de la moyenne du produit des deux variables aléatoires.

- (c) Déterminer les coefficients de la droite de régression. Commenter.

On va procéder comme à la question similaire de l'exercice précédent.

```
y <- c(8,9,10,13,15,14,13)
x <- c(3,4,6,7,9,10,9)

n <- length(x)

# Moyenne de Y
my <- mean(y)

# Moyenne et variance de X
mx <- mean(x)
vx <- var(x)*(n-1)/n
```

```

# Moyenne des produits
mp <- mean(x*y)

# Produit des moyennes
pm <- mean(x)*mean(y)

# Covariance
covxy <- mp-pm

# Estimation des coefficients

beta1 <- covxy/vx
beta0 <- my - beta1*mx

beta1
## [1] 0.95

beta0
## [1] 5.2

```

En représentant la droite de régression, on remarque que deux individus semblent se démarquer des autres.

3. Deux stagiaires semblent se distinguer des autres. Lesquels ?

Deux stagiaires se démarquent en effet, ce sont les stagiaires 4 et 5 qui se trouvent très éloignés de la droite de régression. On peut donc les considérer comme des points atypiques.

4. La "somme des carrés des erreurs" peut aussi se regarder comme fonction du vecteur  $y$  dont les composantes sont les  $y_i$ ,  $i = 1, \dots, n$ . On l'écrit dans le cas spécial où on prend  $b_0 = \hat{\beta}_0$  et  $b_1 = \hat{\beta}_1$  :

$$g(y) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2, \quad (5)$$

- (a) Calculer la dérivée de  $g$  lorsqu'elle est considérée comme fonction de  $y_{i_0}$  pour un indice  $i_0$  particulier.

Le calcul de la dérivée nous donne :

$$\frac{\partial g}{\partial y_{i_0}}(y) = 2(y_{i_0} - b_0 - b_1 x_{i_0}),$$

$$= 2(y_{i_0} - \hat{y}_{i_0}).$$

Cette dernière expression n'est rien d'autre que deux fois la valeur du résidu de la régression pour l'individu  $i_0$ .

- (b) Calculer les valeurs numériques de ces dérivées pour chaque  $i_0 = 1, \dots, n$ . Commenter.

Nous utilisons la remarque formulée à la question précédente et nous obtenons directement les résidus comme suit :

```
data <- data.frame(y,x)
mymodel <- lm(y~x,data)

res = mymodel$residuals

2*res
```

##	1	2	3	4	5
##	-1.000000e-01	-4.173745e-15	-1.800000e+00	2.300000e+00	2.500000e+00
##	6	7			
##	-1.400000e+00	-1.500000e+00			

Nous pouvons faire la même remarque qu'à la question précédente et on voit bien que les individus 4 et 5 sont ceux présentant un résidu élevé.