

TP n° 2 : Régression linéaire simple 1

Exercice 1.

1. Reproduire et comprendre l'enjeu des commandes suivantes :

```
data(anscombe)
attach(anscombe)
anscombe
par(mfrow = c(2, 2))
plot(x1, y1, xlim = c(3, 19), ylim = c(3, 13))
plot(x2, y2, xlim = c(3, 19), ylim = c(3, 13))
plot(x3, y3, xlim = c(3, 19), ylim = c(3, 13))
plot(x4, y4, xlim = c(3, 19), ylim = c(3, 13))
```

2. Parmi les nuages de points affichés, est-ce qu'un ajustement par une droite est-il toujours judicieux ?

Exercice 2. Une entreprise fixe des prix différents pour un produit particulier dans huit régions différentes des États-Unis. Elle souhaite étudier la liaison éventuelle entre le nombre de ventes (variable Y) et le prix du produit (variable X). Pour les $n = 8$ régions, on observe les valeurs $(x_1, y_1), \dots, (x_n, y_n)$ de (X, Y) suivantes :

x_i	420	380	350	400	440	380	450	420
y_i	5.5	6.0	6.5	6.0	5.0	6.5	4.5	5.0

1. Représenter le nuage de points $\{(x_1, y_1), \dots, (x_n, y_n)\}$. À partir de celui-ci, expliquer pourquoi on peut envisager l'existence d'une liaison linéaire entre Y et X .
2. On adopte alors le modèle de *rls* : $Y = \beta_0 + \beta_1 X + \epsilon$. Les paramètres β_0 et β_1 sont des réels inconnus. On considère la forme matricielle usuelle : $Y = X\beta + \epsilon$, avec $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

Créer dans R la matrice X associée.

3. En posant $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, calculer $b = (X^t X)^{-1} X^t y$. Que représente b par rapport à β ?
4. Vérifier que l'on a $b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$, avec

$$b_1 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad b_0 = \bar{y} - b_1 \bar{x}.$$

Retrouver ces résultats numériques avec les commandes `lm` et `coef`.

- Tracer la droite de régression sur le nuage de points.
- Calculer "à la main" le coefficient de détermination et le coefficient de détermination ajusté. Est-ce que le modèle de *rls* est pertinent avec les données ?
- Retrouver les estimations précédentes avec la commande `summary`.

Exercice 3. On veut étudier la liaison éventuelle entre le taux de fibre oxydative (variable X) et la teneur en lipides dans la chair de lapins (variable Y). Pour $n = 9$ échantillons de chair de lapins, on observe les valeurs $(x_1, y_1), \dots, (x_n, y_n)$ de (X, Y) suivantes :

x_i	3	4	4	17	24	45	55	68	73
y_i	0.9	1.3	1.0	2.4	2.8	4.4	5.2	6.3	6.6

- Représenter le nuage de points $\{(x_1, y_1), \dots, (x_n, y_n)\}$. À partir de celui-ci, expliquer pourquoi on peut envisager l'existence d'une liaison linéaire entre Y et X .
- Donner l'équation de la droite de régression avec les commandes `lm` et `coef`. Vérifier que celle-ci passe par le point de coordonnées : $(\bar{x}, \bar{y})$. Tracer cette droite sur le nuage de points.
- Donner les valeurs de la droite de régression prises aux points d'abscisses $x_1, \dots, x_n$ avec la commande `fitted`.
- Donner la prédiction de Y lorsque $X = 28$.
- Donner les n écarts dont le i -ème est défini par la différence entre y_i et la valeur prédite de Y lorsque $X = x_i$, $i \in \{1, \dots, n\}$ avec la commande `residuals`. Comment s'appelle ces écarts ?
- Peut-on admettre que ces écarts sont les réalisations d'une *var* suivant une loi normale ? Est-ce que les hypothèses standards semblent être satisfaites ?

Exercice 4. On considère le jeu de données `airquality` disponible dans R.

- Charger les données et comprendre d'où elles émanent.
- Reproduire et comprendre l'enjeu des commandes suivantes :

```
Ozone = airquality$Ozone[!(is.na(airquality$Ozone))]
Vent = airquality$Wind[!(is.na(airquality$Ozone))]
plot(Vent, Ozone)
reg = lm(Ozone ~ Vent)
summary(reg)
```

Peut-on affirmer un lien linéaire fort entre `Ozone` et `Vent` ?

- Tracer la droite de régression sur le nuage de points.