Université de Montpellier
Département de Mathématiques Appliquées

Outstanding 2 : Régression linéaire

Atelier Projet — HAX916X

Réalisé par :
DIALLO Ousmane
ATTOUMANI Ibrahim

STATISTIQUE
SCIENCE DES DONNÉES
UNIVERSITÉ DE MONTPELLIER

L'INSTITUT agro Montpellier

Année Universitaire 2025 – 2026

Table des matières

1. Introduction

Commençons par un exemple afin de fixer les idées. On cherche à savoir s'il est possible d'expliquer le taux maximal d'ozone de la journée par la température T12 à midi. Les données sont :

Température à 12h	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
O_3 max	115.4	76.8	113.8	81.6	115.4	125.0	83.6	75.2	136.8	102.8

Table 1: 10 données journalières de température et d'ozone.

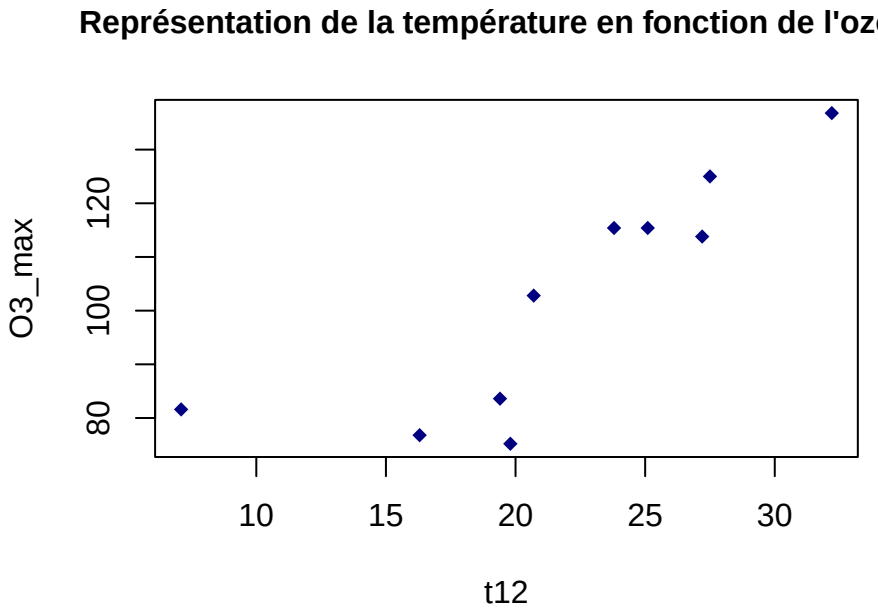
L'objectif de la régression est demodéliser la variable Ozone (O_3) en fonction de la température à 12h (T_{12}).

Avant toute analyse, il est intéressant de représenter les données:

Code R : Représentation de la température en fonction de l'ozone

```
# Création des vecteurs de données
t12 <- c(23.8, 16.3, 27.2, 7.1, 25.1, 27.5,
19.4, 19.8, 32.2, 20.7)
O3_max <- c(115.4, 76.8, 113.8, 81.6, 115.4, 125.0, 83.6, 75.2, 136.8, 102.8)

# Création du data frame
df <- data.frame(t12=t12, O3_max = O3_max)
titre <- "Représentation de la température en fonction de l'ozone"
# Affichage
plot(x = t12, y = O3_max, pch = 18, col="navy", lwd=3, main=titre, cex.main=1)
```



2. Modélisation

Définition 1 : Modèle de régression linéaire simple

Un modèle de régression linéaire simple est défini par une équation de la forme :

$$\forall i \in \{1, \dots, n\}, \quad y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

Les quantités ϵ_i viennent du fait que les points ne sont jamais parfaitement alignés sur une droite. On les appelle les erreurs (ou bruits) et elles sont supposées aléatoires.

On impose les hypothèses suivantes sur ces quantités:

$$(H) \begin{cases} (H_1) : \mathbb{E}[\epsilon_⊐] = 0, \quad \forall i \\ (H_2) : \mathbb{C}ov[\epsilon_i, \epsilon_j] = \delta_{ij}\sigma^2 \quad \forall (i,j) \end{cases}$$

Notons que le modèle de régression linéaire simple peut encore s'écrire de façon vectorielle :

$$Y = \beta_1 \mathbb{1} + \beta_2 X + \epsilon,$$

où:

- le vecteur à expliquer $Y = [y_1, \dots, y_n]^t$ est aléatoire de dimension n ,
- le vecteur $\mathbb{1} = [1, \dots, 1]^t$ est le vecteur de $\mathbb{R}^n$ dont les n composantes valent toutes 1,
- le vecteur explicatif $X = [x_1, \dots, x_n]^t$ est aléatoire de dimension n donné (non aléatoire),
- les coefficients β_1 et β_2 sont les paramètres inconnus (mais non aléatoires !) du modèle,
- le vecteur $\epsilon = [\epsilon_1, \dots, \epsilon_n]^t$ est aléatoire de dimension n.

Définition 2 : Estimateurs des Moindres Carrés Ordinaires

On appelle estimateurs des Moindres Carrés Ordinaires (MCO) $\hat{\beta}_1$ et $\hat{\beta}_2$ les valeurs minimisant la quantité :

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2.$$

Autrement dit, la droite des moindres carrés minimise la somme des carrés des distances verticales des points (x_i, y_i) du nuage à la droite ajustée $y = \hat{\beta}_1 + \hat{\beta}_2 x$.
Les estimateurs des MCO ont pour expressions :

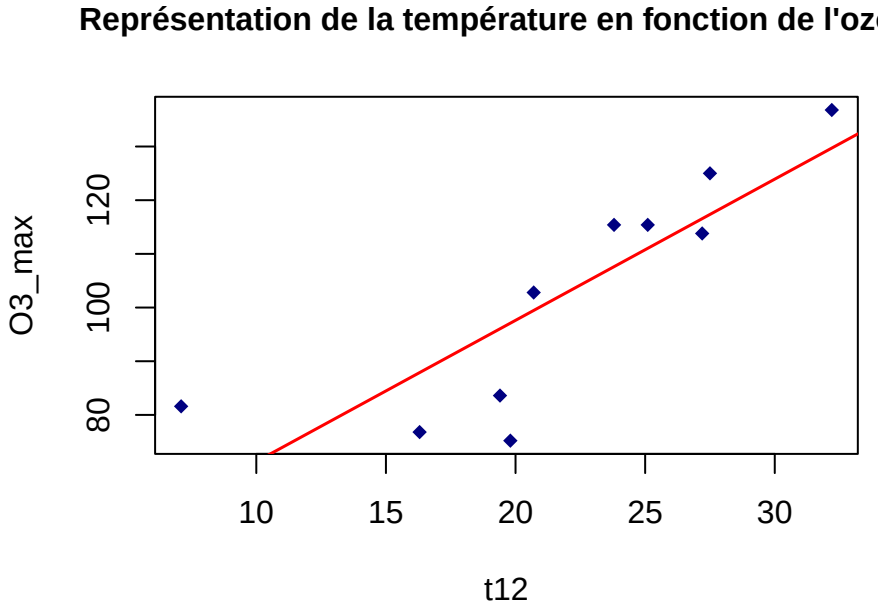
$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \qquad \hat{\beta}_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2} = \frac{\mathbb{C}ov(x,y)}{\sigma_x^2}$$

2.1. Application direct régression linéaire simple sous R

Dans cette partie, nous allons réaliser une application direct de regression linéaire simple de la température à 12h **t12** en fonction de l'ozone sous R. Nous utiliserons la fonction lm disponible par défaut dans R.

```
# creation du modèle
model <- lm(O3_max ~ t12, data=df)

# Affichage avec la droite de regression ajustee
plot(x = t12, y = O3_max, pch = 18, col="navy", lwd=3, main=titre, cex.main=1)
 # ajout de la droite de regression ajustee
abline(model, col = "red", lwd = 1.5)
```




```
# resume du modele
summary(model)
```

```
##
## Call:
## lm(formula = O3_max ~ t12, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.890  -9.001   3.856   7.514  17.919
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  45.0044    13.8050   3.260   0.0115 *
## t12           2.6306     0.6029   4.363   0.0024 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.67 on 8 degrees of freedom
## Multiple R-squared:  0.7041, Adjusted R-squared:  0.6671
## F-statistic: 19.03 on 1 and 8 DF,  p-value: 0.002403
```

Les sorties du logiciel donnent les valeurs estimées $\hat{\beta}_1 = 45.0044$ et $\hat{\beta}_2 = 2.6306$, ainsi que leurs écarts-types respectifs $\hat{\sigma}_1 = 13.8050$ et $\hat{\sigma}_2 = 0.6029$.

D'après le test d'hypothèse :

$$H_0 : \beta_i = 0 \quad \text{vs} \quad H_1 : \beta_i \neq 0, \qquad i = 1, 2,$$

on peut constater que les coefficients sont significatifs. En effet, les p-values $Pr(> |t|)$ des paramètres $\hat{\beta}_1$ et $\hat{\beta}_2$ valent respectivement 0.0115 et 0.0024, ce qui est strictement inférieur à 0.05 pour un test de niveau 5%. On rejete donc l'hypothèse (H_0) au profit de (H_1). Le modèle explique environ 67% de la variabité du modèle puisque le $R^2_{ajusted} \approx 0.67$,ce qui est plutôt satisfesant.

Interprétation Géométrique:

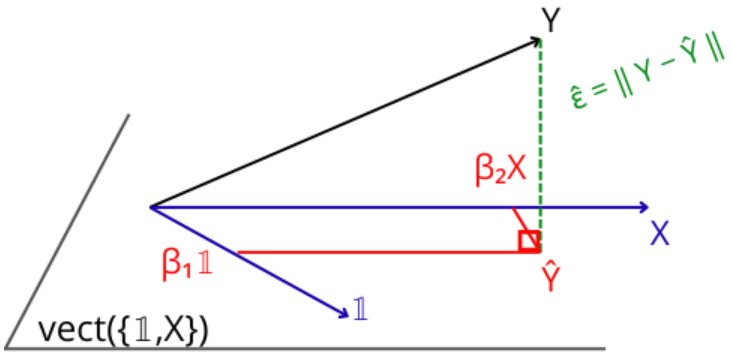


Figure 1: La méthode des moindres carrés

Définition 3 : Modèle de régression linéaire multiple

Un modèle de régression linéaire est défini par une équation de la forme :

$$Y = X\beta + \epsilon$$

où :

- Y est un vecteur aléatoire de dimension n,
- X est une matrice de taille $n \times p$ connue, appelée matrice du plan d'expérience,
- β est le vecteur de dimension p des paramètres inconnus du modèle,
- ϵ est le vecteur de dimension n des erreurs.

Les hypothèses concernant le modèle sont :

$$(H) : \begin{cases} (H_1) : \text{rg}(X) = p, \\ (H_2) : \mathbb{E}[\epsilon] = 0, \quad \text{Var}(\epsilon) = \sigma^2 I_n \end{cases}$$

L'hypothèse (H_2) signifie que les erreurs sont centrées, de même variance (homoscédasticité) et non corrélées entre elles.

2.1. Application : régression linéaire multiple sous R

Dans cette partie, nous allons réaliser une application de régression linéaire multiple de l'ozone **O3_max** en fonction de plusieurs variables explicatives (**t12**, **V**, **N12**) sous R. Nous utiliserons la fonction **lm()** disponible par défaut dans R.

```
# Création de nouvelles variables
V <- c(9.25, -6.15, -4.92, 11.57, -6.23, 2.76, 10.15, 13.5, 21.27, 13.79)
N12 <- c(5, 7, 6, 5, 2, 7, 4, 6, 1, 4)

# Ajout des nouvelles variables au dataframe existant
df <- cbind(df, V, N12)

# Ajustement du modèle linéaire multiple
fit <- lm(O3_max ~ t12 + V + N12, data = df)

# Résumé du modèle
summary(fit)
```

```
##
## Call:
## lm(formula = O3_max ~ t12 + V + N12, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.571  -5.936   1.292   7.070  15.596
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.12494   25.43617   2.324   0.0591 .
## t12          2.42787    0.71394   3.401   0.0145 *
## V           -0.01327    0.53141  -0.025   0.9809
## N12         -2.04114    2.72277  -0.750   0.4818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.88 on 6 degrees of freedom
## Multiple R-squared:  0.7337, Adjusted R-squared:  0.6006
## F-statistic: 5.511 on 3 and 6 DF,  p-value: 0.03693
```

Malgré l'ajout de deux variables explicatives, **V** et **N12**, le modèle de régression multiple ne s'est pas amélioré. En effet, le test d'hypothèse sur la nullité des coefficients montre que la plupart des paramètres β_i ne sont pas significatifs (on conserve H_0), sauf pour le coefficient de la variable **t12**.

Le modèle explique environ **60%** de la variabilité ($R^2_{ajusted} = 0.6006$), comparé au modèle précédent qui expliquait **67%**. Ainsi, l'ajout des variables **V** et **N12** n'a apporté que du bruit au modèle.