Modèle Linéaire Gaussien
Introduction

Achille Thin

7 Décembre 2023

Executive Master Statistique et Big Data

Informations Pratiques

Contact

- Achille Thin (Maître de Conférences), thin.achille@gmail.com

Syllabus

- 15 heures de cours/TP.
- Matériel disponible sur `https://achillethin.github.io/teaching/`
- Évaluation finale : projet sur R.

Exemple introductif

Description

Données: salaire horaire en dollars (WAGE) de 534 personnes (États-Unis, 1985) accompagné de 10 caractéristiques économiques dont

- la catégorie professionnelle (OCCUPATION) : *Management* (1), *Sales* (2), *Clerical* (3), *Service* (4), *Professional* (5), *Other* (6),
- le nombre d'années d'étude (EDUCATION),
- le nombre d'années d'expérience (EXPERIENCE),
- l'âge (AGE),
- le genre (SEX),
- le statut marital (MARR).

Référence: Berndt, E. R. (1991). *The Practice of Econometrics : Classical and Contemporary*. (Source : `www.economicswebinstitute.org/glossary/wages.htm`)

Objectifs

Question : existe-t-il un effet des caractéristiques socio-démographiques sur le salaire des employé·e·s ?

Chargement des données

Chargement à partir d'un fichier Excel:

```
library(readxl)
donnees <- read_excel("../data/wagesmicrodata.xls")
```

Chargement à partir d'un fichier .csv:

```
donnees <- read.csv("../data/wagesmicrodata.csv",
    sep = ";", dec = ",")
```

```
  ID  WAGE OCCUPATION EDUCATION EXPERIENCE AGE
1  1  5.10          6         8         21  35
2  2  4.95          6         9         42  57
3  3  6.67          6        12          1  19
4  4  4.00          6        12          4  22
5  5  7.50          6        12         17  35
6  6 13.07          6        13          9  28
```

Étude descriptive des données : variables quantitatives

Variable quantitative: représente une « mesure », *e.g.*, WAGE, EDUCATION,...

Résumés: indicateurs numériques comme la moyenne ou les quartiles.

```
mean(donnees$WAGE)
```

```
[1] 9.0241
```

```
quantile(donnees$WAGE, c(0.25, 0.5, 0.75))
```

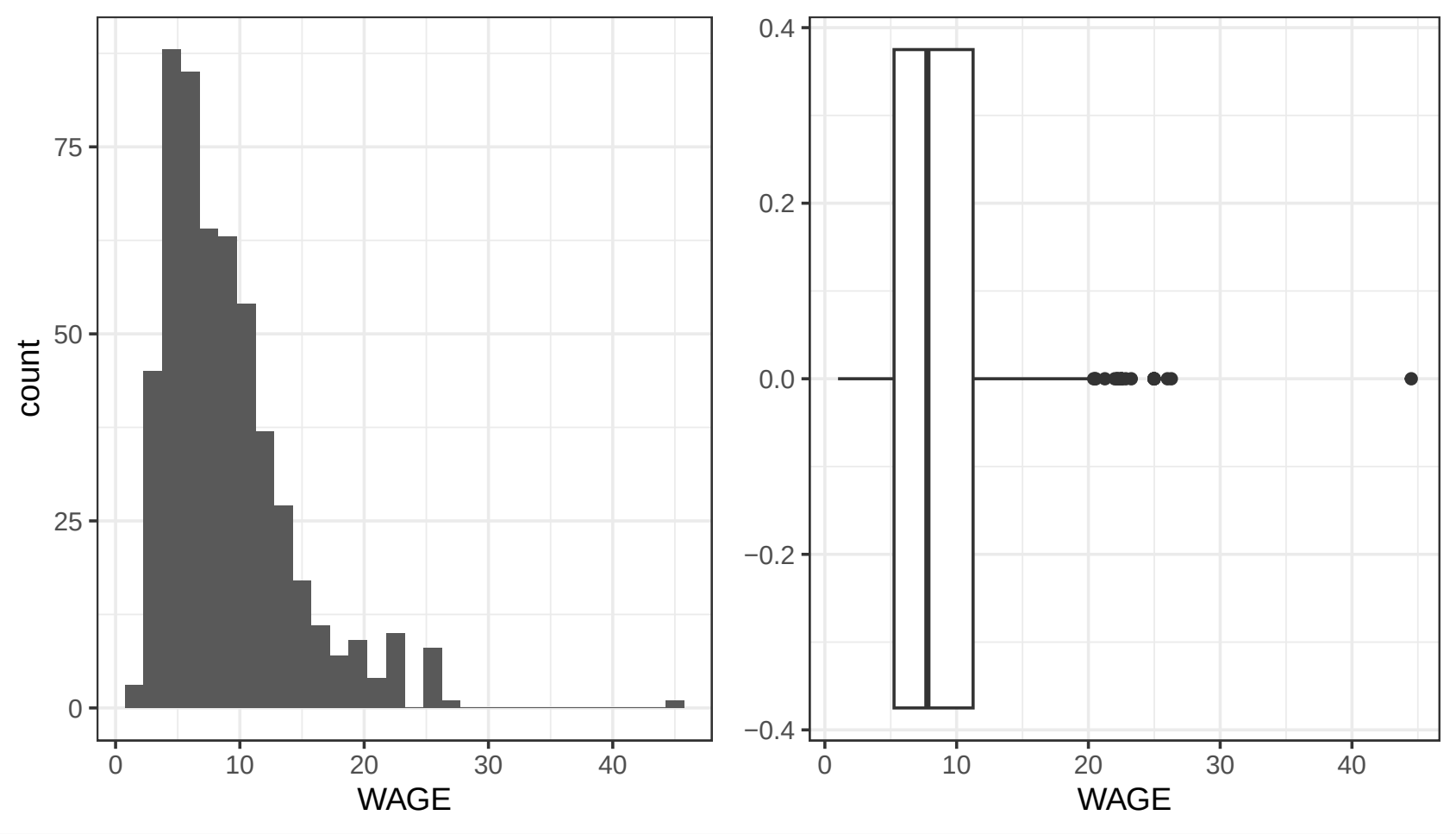
```
  25%   50%   75%
 5.25  7.78 11.25
```

```
summary(donnees$WAGE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    5.25    7.78    9.02   11.25   44.50
```

Étude descriptive des données : variables quantitatives

Représentations graphiques: histogramme (`hist`) ou box-plot (`boxplot`).



Étude descriptive des données : variables qualitatives

Variable qualitative: représente des « catégories », *e.g.*, OCCUPATION, SEX,...

Si les catégories sont représentées par des nombres, il faut le spécifier à R pour qu'il ne voit pas ces variables comme des variables quantitatives.

```
donnees$OCCUPATION <- as.factor(donnees$OCCUPATION)
```

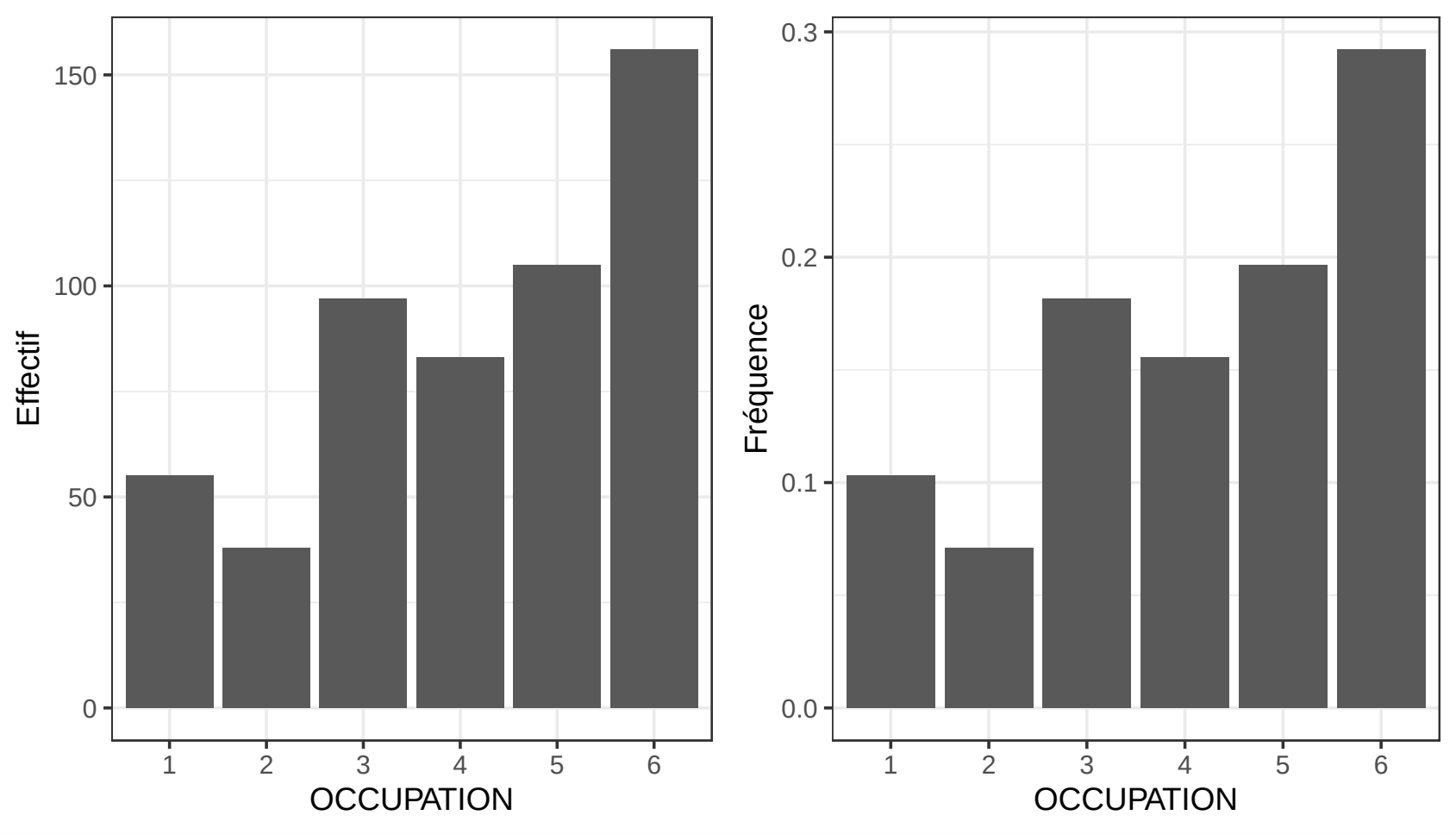
Résumés: effectifs ou fréquences de chacune des catégories (modalité).

```
summary(donnees$OCCUPATION)
```

```
  1   2   3   4   5   6
 55  38  97  83 105 156
```

Étude descriptive des données : variables qualitatives

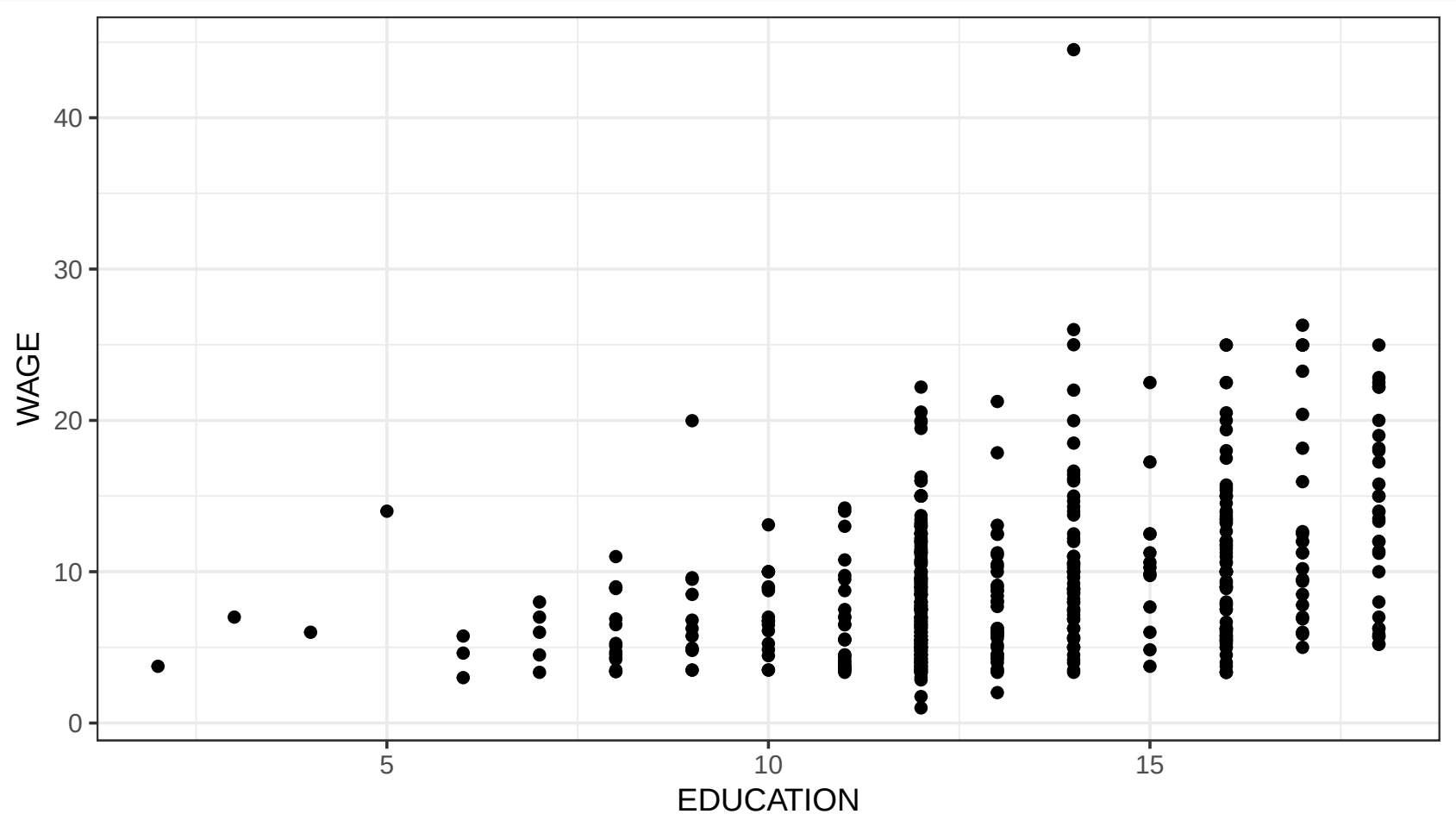
Représentation graphique: diagramme en bâtons (`barplot`).



Exercice

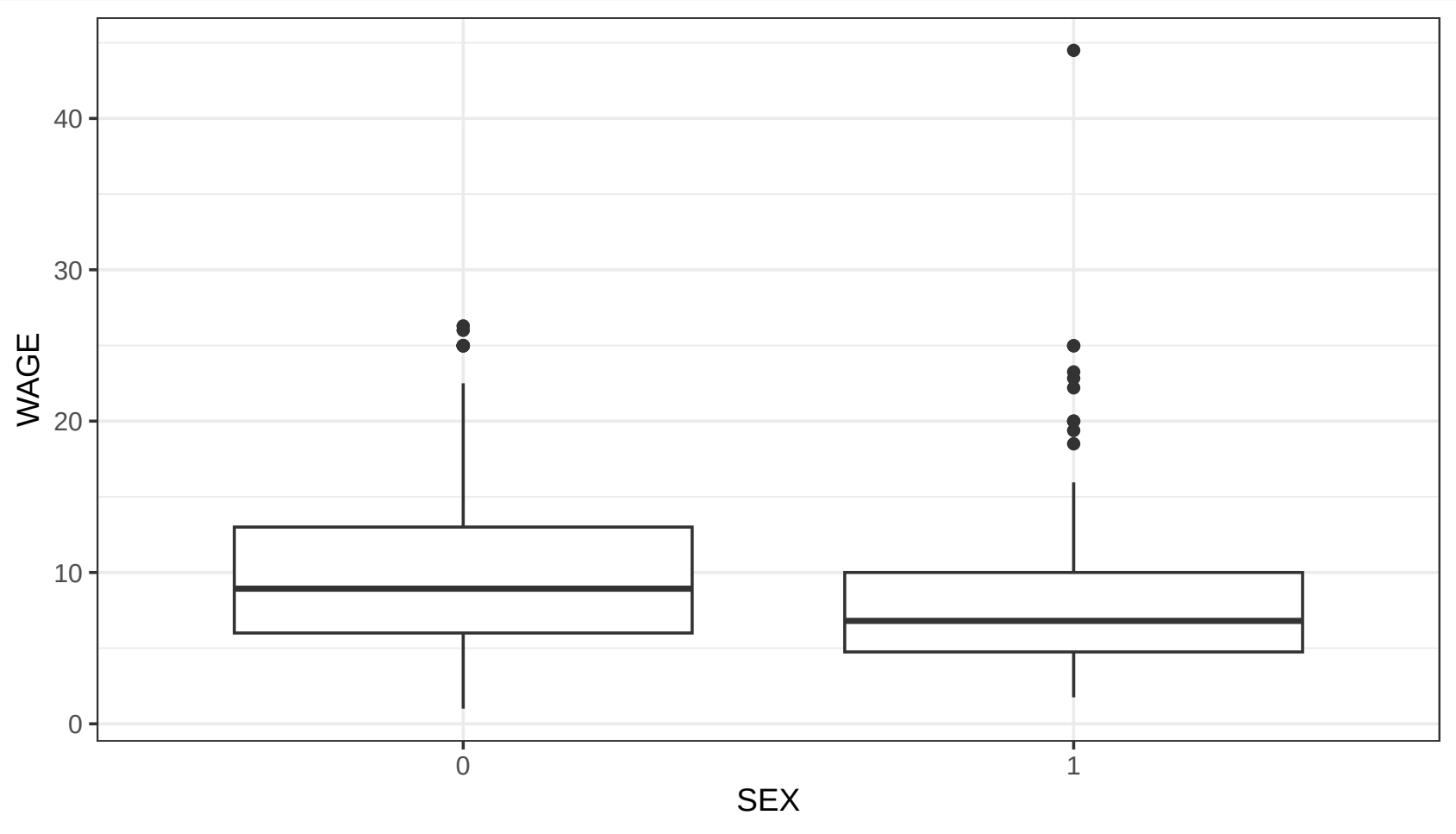
1. Identifier les variables quantitatives et qualitatives du jeu de données.
2. Faire une étude descriptive de l'ensemble des données.

Relation entre variables quantitatives



Question: peut-on résumer ce nuage de point par une droite et établie un lien entre le salaire et le nombre d'années d'étude ?

Relation entre variable quantitative et variables qualitatives



Question: est ce qu'il y a un lien entre le salaire et le genre ?

Cadre du cours

Démarche générale

- En statistique, on cherche à expliquer ou prédire, une variable d'intérêt y en fonction d'une autre variable x.
- On dipose de n valeurs $(x_1, \dots x_n)$ et $(y_1, \dots, y_n)$ pour *apprendre* la relation en x et y :

$$y \approx f(x)$$

- On ne sait pas *à l'avance*
  - si x explique correctement y (x est-elle en particulier suffisante pour prévoir y ?),
  - quelle est la forme de f.

Structure du cours

Modèle linéaire: modèle mathématique décrivant le lien entre une variable explicative **quantitative** (le salaire) et des variables explicatives (le nombre d'années d'expérience, le nombre d'années d'étude, le genre, ...) à l'aide d'une relation linéaire.

- Chapitre 1 – Régression linéaire simple : expliquer en fonction d'une variable quantitative.
- Chapitre 2 – Régression linéaire multiple : expliquer en fonction de plusieurs variables quantitatives.
- Chapitre 3 – Analyse de la variance : expliquer en fonction d'une ou plusieurs variables qualitatives.
- Chapitre 4 – Analyse de la covariance : expliquer en fonction d'une variable quantitative et d'une variable qualitative.

Références

Cours en ligne:

- S. Donnet : `https://sophiedonnet.github.io/linearmodel.html`
- C. Chouquet : `https://www.math.univ-toulouse.fr/~barthe/M1modlin/poly.pdf`

Livre:

- Le modèle linéaire par l'exemple, Jean-Marc Azaïs, Jean-Marc Bardet, Dunod, 2005.
- Le modèle linéaire et ses extensions, Liliane Bel et al., 2016.