Chapitre 1. **Introduction**

Cours de modèle linéaire gaussien par S. Donnet

Executive Master Statistique et Big-Data

Août 2020

Avant de commencer

- Sophie Donnet (CR INRAE) : `sophie.donnet@inrae.fr`
- Cours de 15h sur le modèle linéaire gaussien.
- Accent mis sur pratique, preuves mathématiques réduites au minimum
- Toutes les expérimentations numériques seront faites sous R.

Jeu de données introductif

- Nous nous intéressons au jeu de données tiré du Economics Web Institute
- Dans ce jeu de données, on a relevé le salaire horaire de 534 personnes (en dollars) ainsi que des caractéristiques économiques :
  - l'occupation (divisée en 6 catégories, 1=Management, 2=Sales, 3=Clerical , 4=Service, 5=Professional, 6=Other),
  - le nombre d'années d'étude,
  - le nombre d'années d'expérience,
  - l'âge,
  - le sexe,
  - le statut marital...

Chargement du jeu de données

```
data1
```

```
## # A tibble: 534 x 12
##       ID  WAGE OCCUPATION SECTOR UNION EDUCATION EXPERIENCE
##    <int> <dbl> <fct>      <fct>  <fct>     <int>      <int> <
##  1     1  5.1  6          1      0             8         21
##  2     2  4.95 6          1      0             9         42
##  3     3  6.67 6          1      0            12          1
##  4     4  4    6          0      0            12          4
##  5     5  7.5  6          0      0            12         17
##  6     6 13.1  6          0      1            13          9
##  7     7  4.45 6          0      0            10         27
##  8     8 19.5  6          0      0            12          9
##  9     9 13.3  6          1      0            16         11
## 10    10  8.75 6          0      0            12          9
## # ... with 524 more rows, and 2 more variables: RACE <fct>, S
```

Objectifs

- *Objectifs de cette étude* :
  - Point de vue de l'économiste, démographe ...
  - Evaluer l'effet éventuel des caractéristiques socio-démographiques sur le salaire des employés.
- *Objectifs du cours* :
  - Point de vue du statisticien
  - Etudier les outils statistiques permettant une analyse rigoureuse des données (estimateurs, intervalles de confiance, tests statistiques, $\cdots$)

Etude descriptive des données

Etude de la distribution des salaires

```
library(dplyr)
data1 %>% select(WAGE)  %>% summary()
```

```
##       WAGE
##  Min.   : 1.000
##  1st Qu.: 5.250
##  Median : 7.780
##  Mean   : 9.024
##  3rd Qu.:11.250
##  Max.   :44.500
```

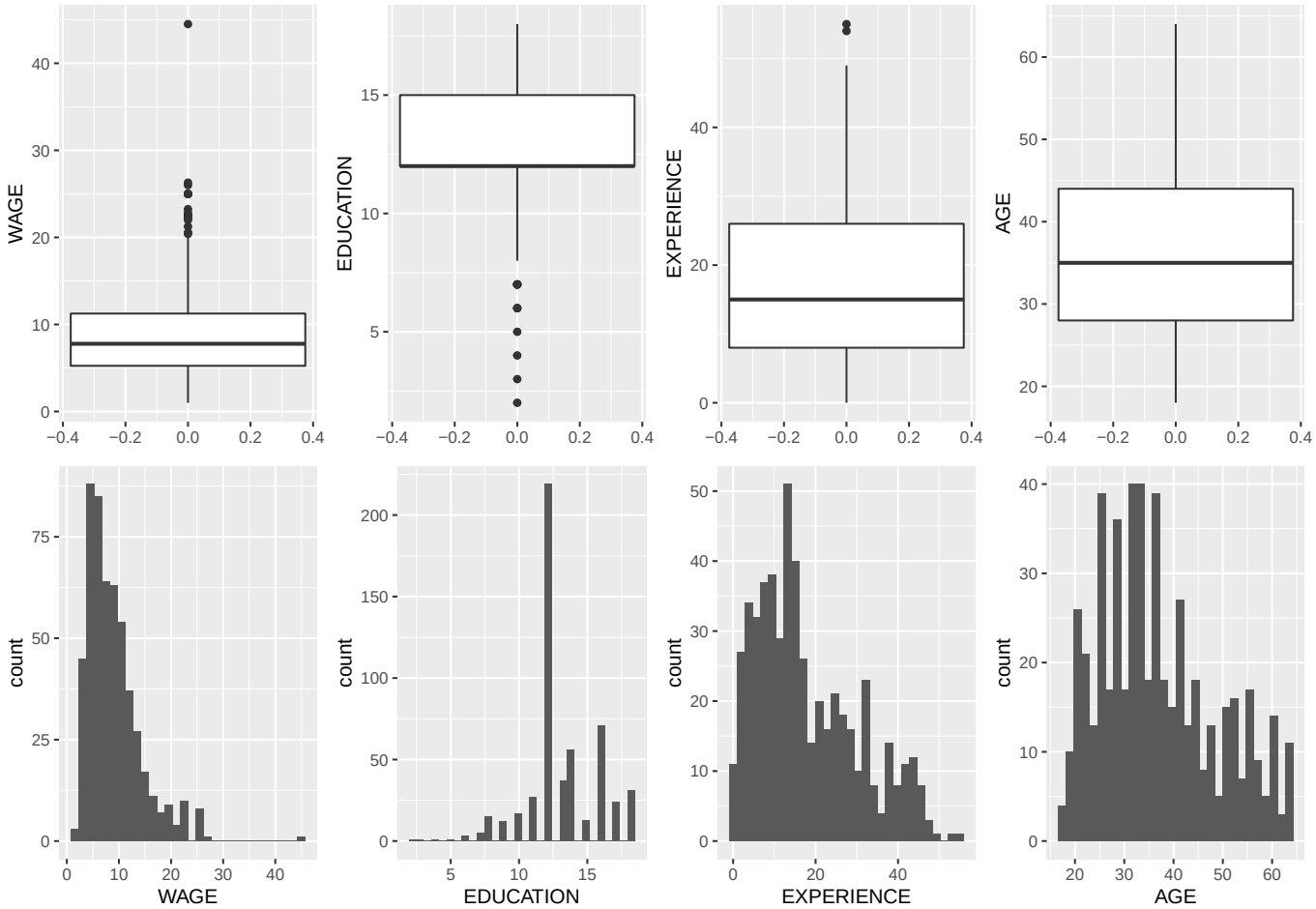
Etude descriptive des données

Etude de la distribution des types d'emploi

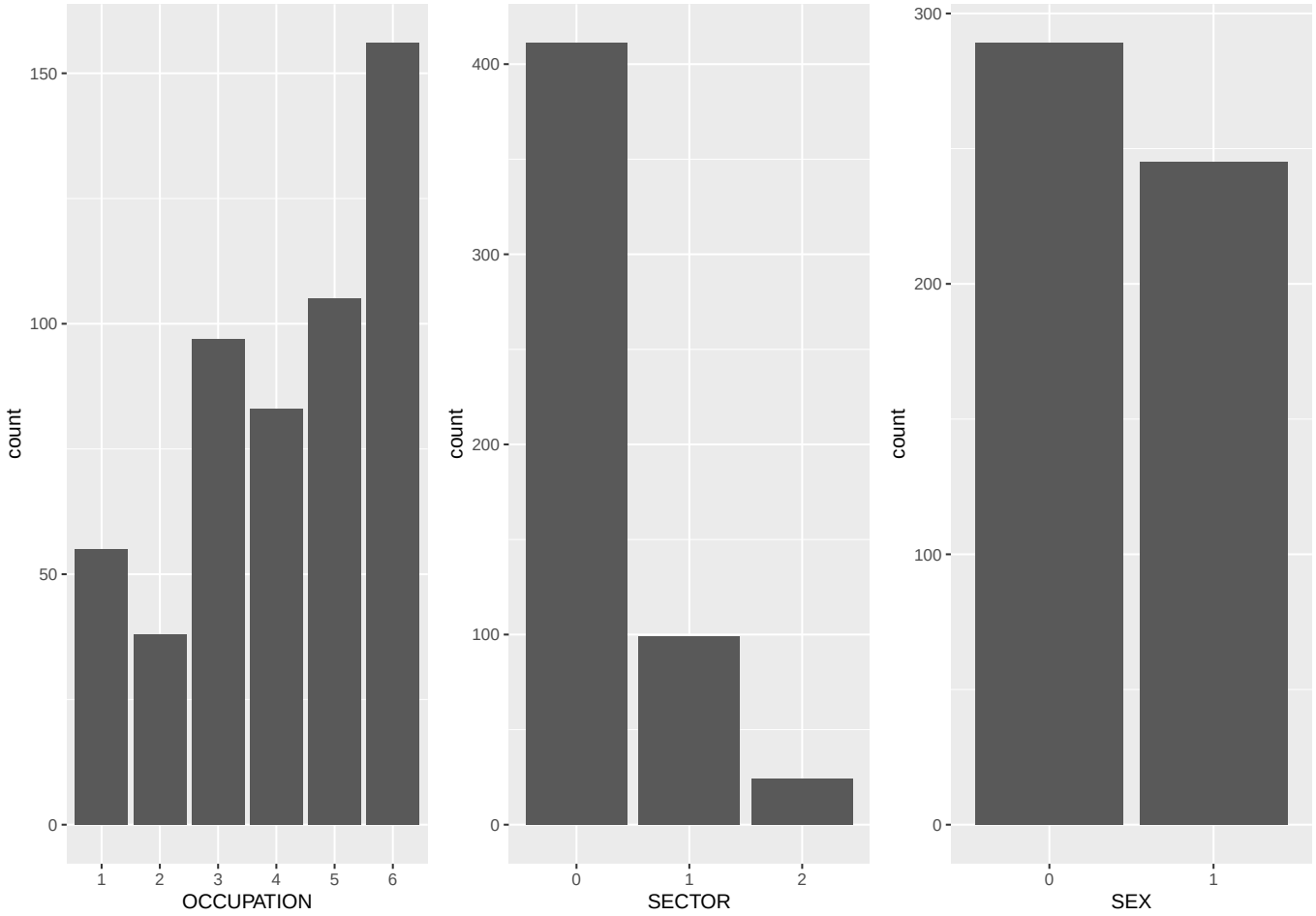
```
data1 %>% select(OCCUPATION)  %>% summary()
```

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

Représentation des variables quantitatives : histogrammes ou boxplot

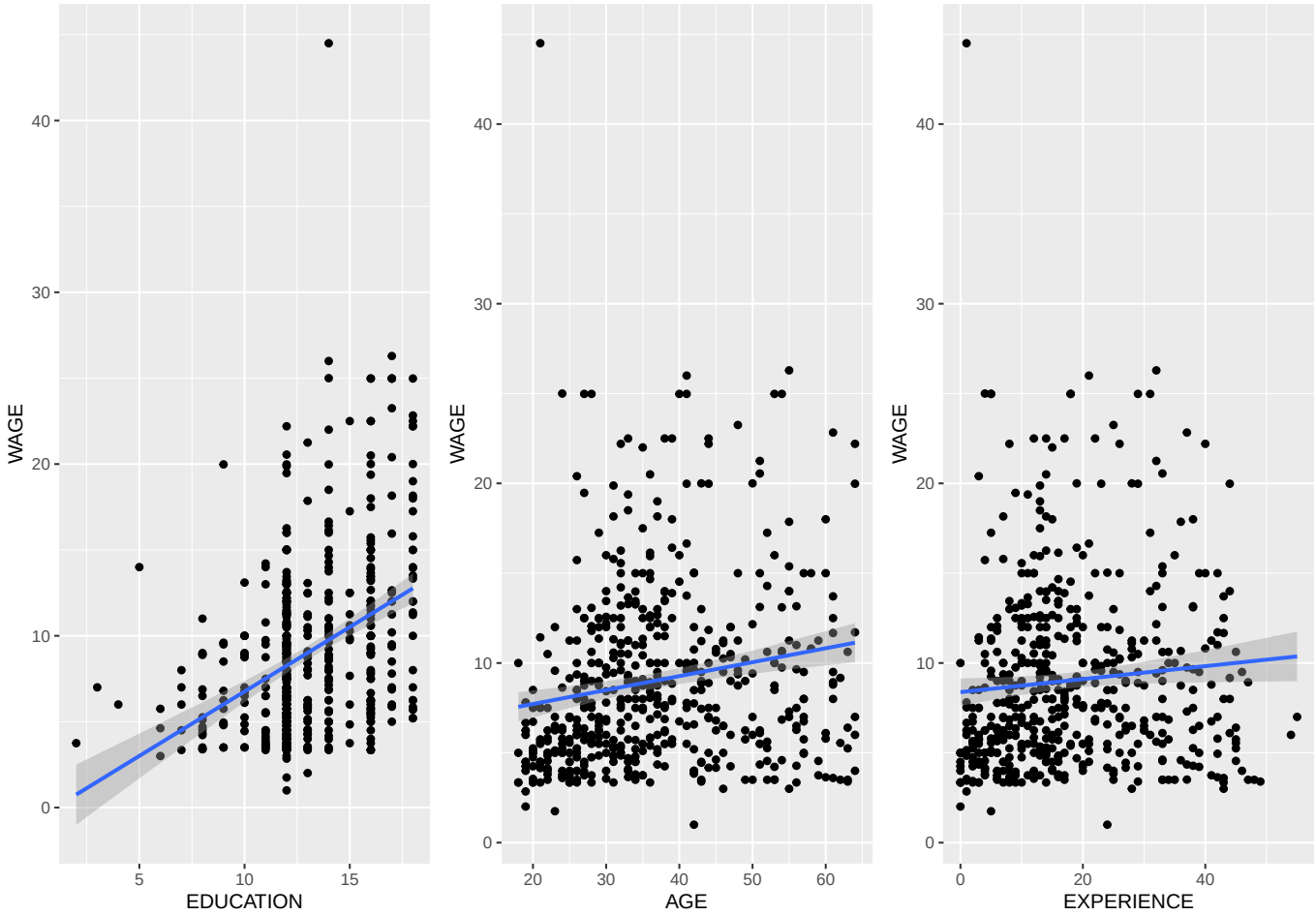


Représentation des variables qualitatives : diagrammes en batons



Relations entre variables quantitatives

Comprendre l'influence des variables quantitatives (Education, Age, Experience) sur le salaire (Wage).



Coefficient de corrélation linéaire

$$\rho_{XY} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i,j}(y_i - \bar{y})(x_j - \bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

où $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

- Par construction, $|\rho_{XY}| \leq 1$.
- Si les points sont parfaitement alignés alors $|\rho_{XY}| = 1$.

```
corr1 <- cor(data1$WAGE,data1$EDUCATION);
corr2 <- cor(data1$WAGE,data1$AGE);
corr3 <- cor(data1$WAGE,data1$EXPERIENCE);
print(c(corr1,corr2,corr3))
```

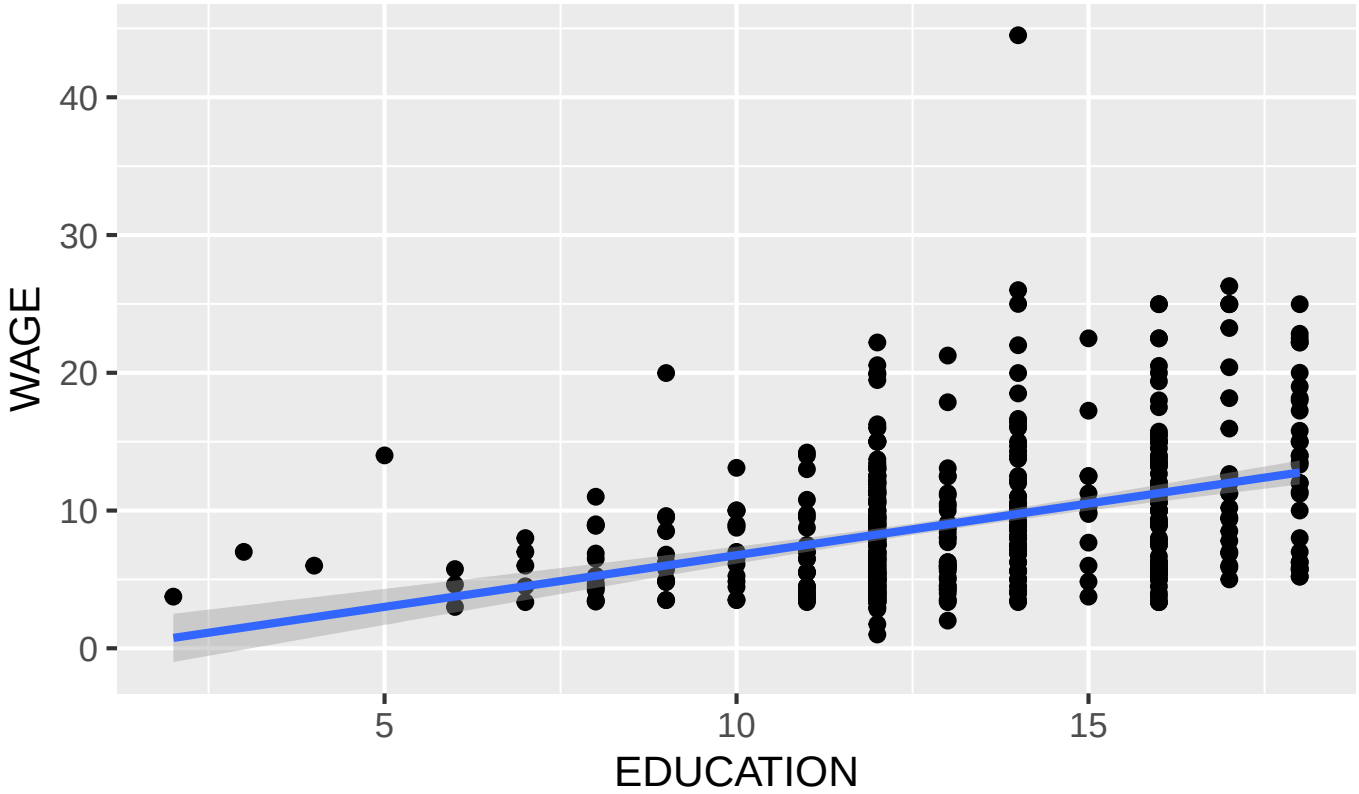
```
## [1] 0.38192207 0.17696688 0.08705953
```

Conclusions sur le ρ

Question : Est-ce grand ? petit ? Significatif ?

- Des tests statistiques. Par exemple tests non paramétriques reposant sur la méthode des rangs.
- Sous R, ces tests sont implémentés dans la fonction cor.test

Régression linéaire simple



Il existe a et b tels que pour l'individu i,

$$\begin{aligned} \text{Wage}_i &\approx a\,\text{Education}_i + b \\ \text{Wage}_i &= a\,\text{Education}_i + b + e_i \end{aligned}$$

où e_i est un terme d'erreur entre la droite de l'observation y_i.

Régression linéaire simple

- Modèle très simple,
- Variable d'intérêt Y (quantitative) expliquée par une variable quantitative x dite *explicative* (ou *covariable*).
- Pente (a) et ordonnée à l'origine (b) de la droite *estimées* à partir des observations pour "placer" convenablement la droite.

Dans ce cours (Chapitre 2) :

- On verra comment estimer ces paramètres, quelles sont les propriétés d'un tel estimateur.
- Pente significativement différente de 0 : par conséquence on va chercher à écrire des tests sur les paramètres du modèles.
- Prédiction du salaire attendu pour un x quelconque.

Régression linéaire multiple (Chapitre 3)

- Expliquer le salaire comme une combinaison linéaire des autres variables quantitatives : il existe f telle que pour tous les individus :

$$\begin{aligned}
\text{Wage}_i &\approx f(\text{Education}_i, \text{Age}_i, \text{Experience}_i) \\
\text{Wage}_i &= f(\text{Education}_i, \text{Age}_i, \text{Experience}_i) + e_i \\
\text{Wage}_i &= \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Age}_i + \beta_3 \text{Experience}_i + e_i \\
Y_i &= \beta_0 + \sum_{k=1}^{K} \beta_k x_i^k + e_i
\end{aligned}$$

où x_i^k est la valeur de la k-ième variable explicative de l'individu i.

- Les mêmes questions que précédemment se posent : significativité des β_k (tests statistiques), etc ...

Analyse de la variance (Chapitre 4)

- Prise en compte des variables non quantitatives (appelées *facteurs*)

Plan du cours

- Chapitre 2 : Régression linéaire simple
- Chapitre 3 : Régression linéaire multiple
- Chapitre 4 : Anova à un et deux facteurs. Ancova

Ressources

- Tous les documents sont en ligne sur Mycourse.
- Poly beaucoup plus complet que les slides
- Annexes avec rappels
- Tous les codes R du cours sont disponibles sous la forme des Rmd.

Biblio

- **Régression avec R** de Pierre-André Cornillon et Eric Matzner-Lober, paru chez Springer.
- **Le modèle linéaire par l'exemple**, Jean marc Azaïs, Jean-Marc Bardet, Dunod, 2005.
- `https://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf`
- **An Introduction to Statistical Learning with Applications in R**
- **Le modèle linéaire par l'exemple** de J.-M. Azais et J.-M. Bardet (Dunod)
- **Le modèle linéaire et ses extensions** de J.-J. Daudin (Ellipse)

Biblio (sur le web)

Vous trouverez aussi de très bonnes références sur le web. En voici une sélection.

- Jeux de données pour modèle linéaire : `http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html`
- `http://www.math.univ-toulouse.fr/~azais/styles/other/student/modlin.pdf`
- `https://perso.univ-rennes2.fr/system/files/users/fromont_m/Poly_Reg.pdf`
- `https://www.agroparistech.fr/IMG/pdf/ExemplesModeleLineaire-AgroParisTech.pdf`
- Cours de C Chouqet (Toulouse) : `https://www.math.univ-toulouse.fr/~barthe/M1modlin/poly.pdf`