Introduction

Cathy Maugis-Rabusseau

INSA Toulouse / IMT
GMM 116
cathy.maugis@insa-toulouse.fr

2023 - 2024

Outline

1. When the response variable is quantitative

2. When the response variable is qualitative

Example

- For 100 individuals, we have their height, weight, age and sex (75 men and 25 women). We also know whether they are smokers or not; whether they snore at night or not.

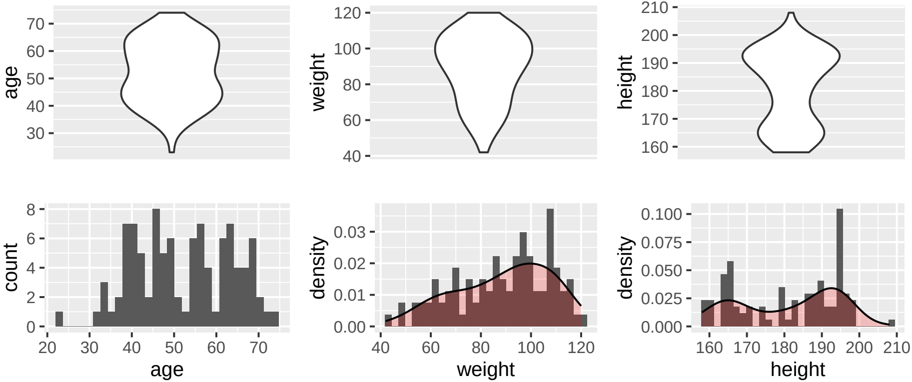
	age	weight	height	sex	snore	tobacco
1	47	71	158	M	N	Y
2	56	58	164	M	Y	N
3	46	116	208	M	N	Y
4	70	96	186	M	N	Y
5	51	91	195	M	Y	Y
6	46	88	188	F	N	N

- 3 quantitative variables and 3 qualitative variables

Description

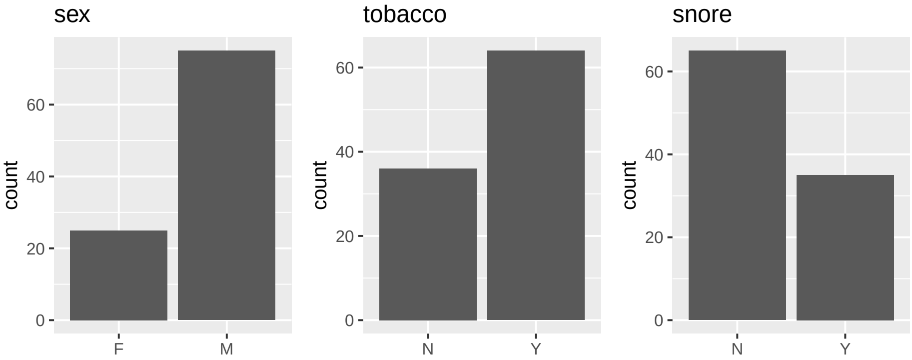
```
summary(don)
```

```
      age            weight           height       sex     snore    tobacco
 Min.   :23.00   Min.   : 42.00   Min.   :158.0   F:25    N:65     N:36
 1st Qu.:43.00   1st Qu.: 75.50   1st Qu.:166.0   M:75    Y:35     Y:64
 Median :52.00   Median : 92.00   Median :186.0
 Mean   :52.27   Mean   : 88.83   Mean   :181.1
 3rd Qu.:62.25   3rd Qu.:104.25   3rd Qu.:194.0
 Max.   :74.00   Max.   :120.00   Max.   :208.0
```



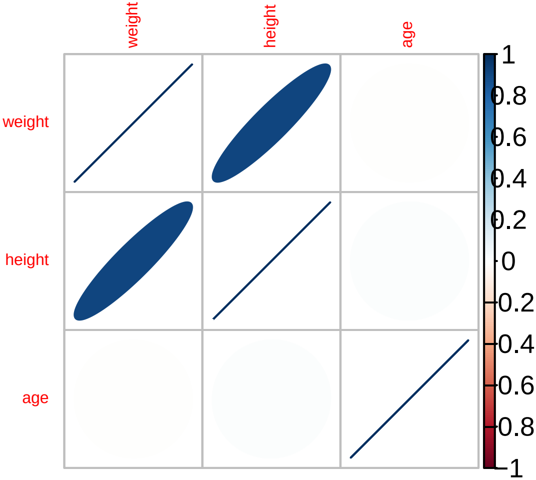
Description

Variable	Levels	Freq %
sex	Female	25
	Male	75
tobacco	Yes	64
	No	36
snore	Yes	35
	No	65



Explain weight $\sim$ height / age (linear regression)

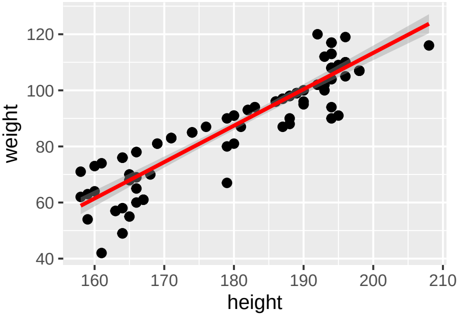
- Correlation between the quantitative variables:

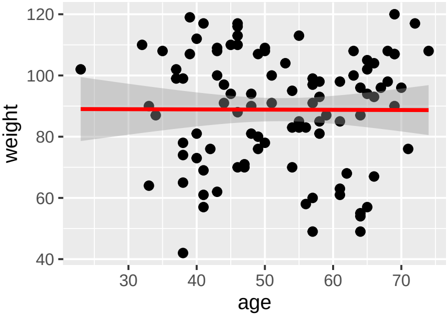


Explain weight $\sim$ height / age (linear regression)

- Pearson correlation coefficient:

	height	age
weight	0.92	-0.004
p-value	$< 2.2\ 10^{-16}$	0.9687





Simple linear regression

- Model:

$$weight_i = a + b \times height_i + \varepsilon_i,\ i = 1, \cdots, 100$$

where ε_i is the noise for the i-th observation

- Assumptions: $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d $\mathcal{N}(0, \sigma^2)$
Gaussian errors with the same unknown variance σ^2

- Matricial writing:

$$\underbrace{\begin{pmatrix} weight_1 \\ \vdots \\ weight_{100} \end{pmatrix}}_{weight} = \underbrace{\begin{pmatrix} 1 & height_1 \\ \vdots & \vdots \\ 1 & height_{100} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{100} \end{pmatrix}}_{\varepsilon}$$

$$\Leftrightarrow weight = X\theta + \varepsilon,\ \ \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$$

Least squares estimators

$$\hat{\theta} = (\hat{a}, \hat{b}) = \underset{(\alpha,\beta)}{\operatorname{argmin}} \sum_{i=1}^{100} (weight_i - \alpha - \beta \ height_i)^2$$
$$= \underset{(\alpha,\beta)}{\operatorname{argmin}} ||weight - \alpha \mathbb{1}_{100} - \beta \ height||^2.$$

```
reg1<-lm(weight~height,data=don)
summary(reg1)
```

```
Call:
lm(formula = weight ~ height, data = don)

Residuals:
     Min       1Q   Median       3Q      Max
-20.7482  -3.8787   0.6629   4.1182  17.0261

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -146.16586   10.35384  -14.12   <2e-16 ***
height         1.29760    0.05702   22.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.583 on 98 degrees of freedom
Multiple R-squared:  0.8409,    Adjusted R-squared:  0.8393
F-statistic: 517.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

Least squares estimations

- $\left(\widehat{b}\right)^{obs} = 1.298$: estimation of the linear regression slope
- $(\widehat{a})^{obs} =$ -146.166: estimation of the linear regression intercept
- $\left(\widehat{\sigma^2}\right)^{obs} = (7.583)^2$

The slope estimation 1.298 is significantly different from 0, showing that weight and height are significantly related

$\Rightarrow$ testing procedure to validate

Explain weight $\sim$ height and age (multiple linear regression)

- Model:

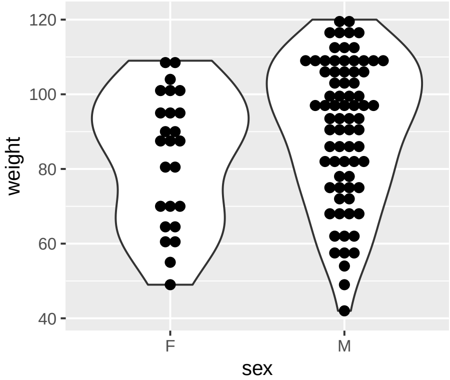
$$weight_i = \theta_0 + \theta_1 \times height_i + \theta_2 \times age_i + \varepsilon_i,\ \ i = 1, \ldots, 100$$

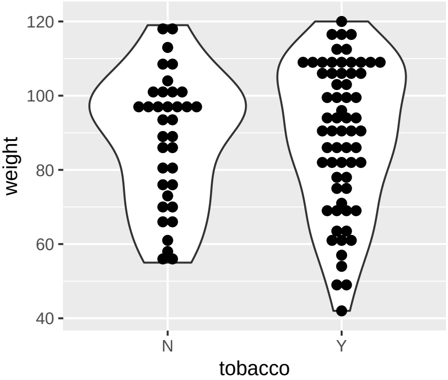
where ε_i are assumed i.i.d $\mathcal{N}(0, \sigma^2)$.

- Matricial writing:

$$\underbrace{\begin{pmatrix} weight_1 \\ \vdots \\ weight_{100} \end{pmatrix}}_{weight} = \underbrace{\begin{pmatrix} 1 & height_1 & age_1 \\ \vdots & \vdots & \vdots \\ 1 & height_{100} & age_{100} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{100} \end{pmatrix}}_{\varepsilon}$$

Explain weight $\sim$ sex / tobacco (Anova)





Explain weight $\sim$ sex (One-way Anova)

- Model per observation:

$$weight_i = \mu_1 \mathbb{1}_{sex_i=F} + \mu_2 \mathbb{1}_{sex_i=M} + \varepsilon_i \text{ where } \varepsilon_i \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

- Matricial writing:

$$\underbrace{\begin{pmatrix} weight_{11} \\ \vdots \\ weight_{1n_1} \\ weight_{21} \\ \vdots \\ weight_{2n_2} \end{pmatrix}}_{weight} = \underbrace{\begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \end{pmatrix}}_{\varepsilon},$$

where $weight_{i,j}$ = weight of the j-th individual with sex $i = F$ or M, $j \in \{1, \ldots, n_i\}$.

Explain weight ∼ sex (One-way Anova)

```r
anova1<-lm(weight~sex-1,data=don)
summary(anova1)
```

```
Call:
lm(formula = weight ~ sex - 1, data = don)

Residuals:
   Min     1Q Median     3Q    Max 
-48.77 -13.44   4.00  16.23  29.23 

Coefficients:
     Estimate Std. Error t value Pr(>|t|)    
sexF   83.000      3.741   22.19   <2e-16 ***
sexM   90.773      2.160   42.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.7 on 98 degrees of freedom
Multiple R-squared:  0.9584,	Adjusted R-squared:  0.9576 
F-statistic:  1129 on 2 and 98 DF,  p-value: < 2.2e-16
```

Explain weight $\sim$ sex and tobacco (Two-way Anova)

- Principal effect of factors sex and tobacco
  $+$ interaction of the two factors on weight

- Model (two-way anova with interaction):

$$weight_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

  where $weight_{ijk}$ = weight of the k-th individual with $sex = i \in \{F, M\}$ and $tobacco = j \in \{Y, N\}$, $k \in \{1, \ldots, n_{ij}\}$

- This model can also be written matricially

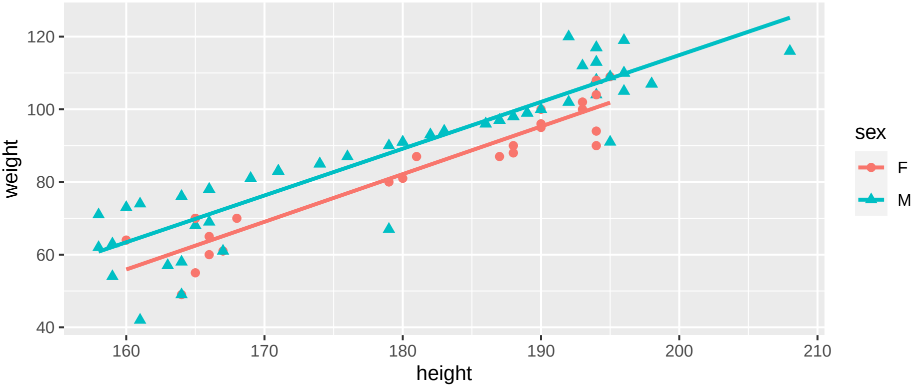
$$weight = X\theta + \varepsilon,\ \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$$

Explain weight $\sim$ sex and height (ANCOVA)

- Model:

$$\begin{cases} weight_{ij} = a_i + b_i \; height_{ij} + \varepsilon_{ij}, \; i \in \{F, M\} \text{ and } j = 1, \cdots, n_i \\ \\ \varepsilon_{ij} \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \end{cases}$$

where $weight_{ij}$ = weight of the j-th individual with sex i.



Conclusion

In the different examples (linear regression, anova, ancova), we have

- the same matricial model

$$Y = X\theta + \varepsilon$$

- the same assumptions on the errors $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$
- the least squares estimators

$\Rightarrow$ These different models are grouped together in the family of **general linear models**.

Outline

Binary response - Logistic regression

- The logistic regression allows to generalize the linear regression for a binary response
- Let $Y = (Y_1, \dots, Y_n)'$ where $Y_i \sim \mathrm{Ber}(\pi_i)$, $i \in \{1, \dots, n\}$.

Goal: Explain Y according to several regressors $z^{(1)}, \dots, z^{(m)}$

- **Example** : An insurance company seeks to detect fraud cases. It has n files for this. Each of these files is associated with the value 0 (for fraud case), 1 otherwise. After having selected the most interesting characteristics (household indebtedness, social environment, place of residence, ...), the company seeks to know to what extent these characteristics influence the probability of existence of a fraud. It hopes that in the future it will be able to detect any “sensitive” files.

Logistic regression

- With a linear model:

$$\mathbb{E}[Y_i] = \pi_i = a_1 z_i^{(1)} + a_2 z_i^{(2)} + \cdots + a_m z_i^{(m)},\ \ i = 1, \ldots, n.$$

But, since we want to model and predict probabilities, this approach seems not recommended insofar as certain predicted values could not belong to the interval $[0,1]$!

- **Logistic regression model** : $\forall i \in \{1, \ldots, n\}$,

$$g(\pi_i) = a_1 z_i^{(1)} + \cdots + a_m z_i^{(m)},\ \text{where } g(t) = \log\left(\frac{t}{1-t}\right)$$

$g :]0,1[\to \mathbb{R}$ is called a **link function**.

Generalized linear model

- More generally, it is possible to consider other probability distributions for the response variable Y and other link functions.
- We will see that it is possible to study all these models by the same framework: **the generalized linear model**
- But parameter estimation, confidence interval construction and testing procedures have to be modified from those of the linear model.

Objectives of this course

- Know how to choose a modeling adapted to the problem among linear models and generalized linear models.
- Know how to write modeling, estimate the parameters, construct confidence intervals and testing procedures.
- Know some procedures to choose the “most” explanatory variables and know how to simplify a model.
- ...