Simple Statistics, Linear and Generalized Linear Models

Rajaonarifara Elinambinina

PhD Student: Sorbonne Universté

January 8, 2020

Outline

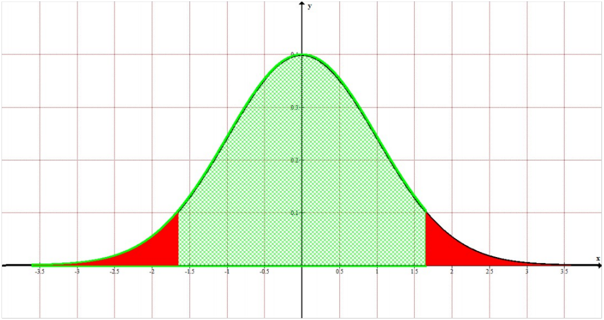
Basic statistics

A statistical variable is either quantitative or qualitative

- **quantitative variable**: numerical variable: discrete or continuous
  - Discrete distribution: countable
  - Continuous: not countable
- **qualitative variable**: not numerical variable: nominal or ordinal(cathegories)
  - ordinal: can be ranked
  - nominal: cannot be ranked

- Example of discrete distribution
  - **Binomial distribution**
    - binary data:there are only two issues (0 or 1)
    - repeat the experiment several times
  - **Poisson distribution**
    - Count of something in a limited time or space: positive
    - The probability of realising the event is small
    - The variance is equal to the mean
  - **Binomial Negative**
    - Number of experiment needed to get specific number of success

- Example of continuous distribution
  - **Uniform distribution**
  - **Normal distribution**:
    - Symmetrical with respect to the mean: the most important distribution in statistics

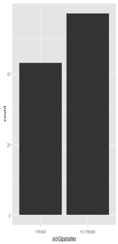


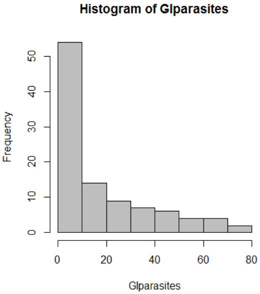
Data Analysis

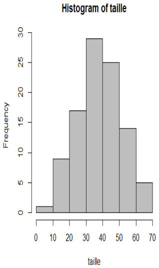
- Explore Data
- Build models
- Validate model
- Select the best model

Explore data

- Statistical parameters
  - central tendancy: mean, mode, median, . . .
  - dispersion: variance, standard deviation,. . .
- graph
  - histogram
  - boxplot







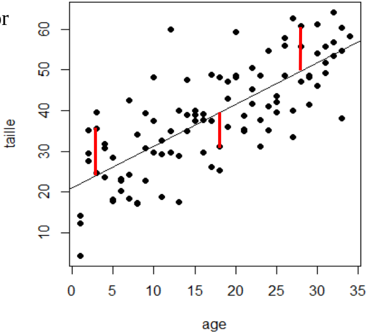
Build models

- Identify the response variable and the explanatory variables
- Identify the characteristic of the variables
- Choose the model to be used
  - univariate linear model: simple linear regression
    - response variable continuous(eg: normally distributed)
    - one explanatory variable
  - multivariate linear model: multiple linear regression
    - response variable continuous
    - several explanatory variables
  - generalized linear model
    - binary data
    - count data

Univariate linear model: simple linear regression

- Quantify the relationship between the response variable and each explanatory variable
- Linear relationship: $y = a + bx + \epsilon$
  - y: response variable, x: explanatory variable
  - a: intercept, b: slope, ϵ: Error or residual
- Minimize the error

Taille=20+1.15 *Age(Months)+Error



Univariate Linear model

- Rcommand: **lm(response_variable∼ explanatory_variable)**
- **R-squared**: a statistical metric used to measure how much of the variation in outcome can be explained by the variation in the independent variables

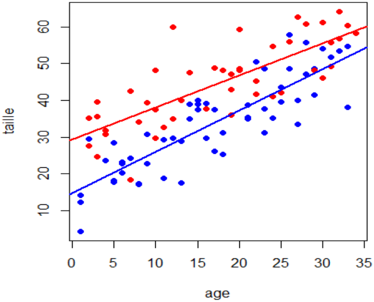
```
Call:
lm(formula = taille ~ age, data = lemur.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.559  -5.655   0.519   7.776  17.097 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  24.0529     1.9950  12.057  < 2e-16 ***
age           0.8944     0.1014   8.824 4.29e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.754 on 98 degrees of freedom
Multiple R-squared:  0.4428,	Adjusted R-squared:  0.4371 
F-statistic: 77.87 on 1 and 98 DF,  p-value: 4.29e-14
```

Multivariate linear model: Multiple linear regression

- Quantify the relationship between the response variable and a set of explanatory variables
- Relationship: $y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n + \epsilon$
  - y: response variable
  - $x = (x_1, x_2, \ldots, x_n)$: explanatory variables
- Rcommand: **lm(response_variable∼ explanatory_variable$_1$ + explanatory_variable$_2$)**
  - Three types of relationship between two explanatory variables A and B:
    - ★ A+B: Effect of each variable
    - ★ A∗B: Effect of each variable and their interaction
    - ★ A:B: Effect of the interaction of the variables

Multivariate linear model

Taille= 15+1.15*Age(Months)+15*Sexe(Female)+Error



Multilinear model

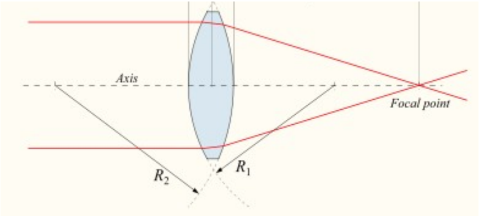
```
Call:
lm(formula = taille ~ age + sexe + GIparasites + malaria, data = lemur.dat

Residuals:
     Min       1Q   Median       3Q      Max
-12.3696  -4.2168   0.0111   3.8716   9.9466

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.37448    1.40044  17.405  < 2e-16 ***
age          0.87527    0.05423  16.141  < 2e-16 ***
sexeMale    10.20143    1.04410   9.771 5.11e-16 ***
GIparasites -0.30170    0.02601 -11.598  < 2e-16 ***
malariaOui  -0.10413    1.04603  -0.100    0.921
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.203 on 95 degrees of freedom
Multiple R-squared:  0.8463,    Adjusted R-squared:  0.8399
F-statistic: 130.8 on 4 and 95 DF,  p-value: < 2.2e-16
```

Generalized linear model

- Extend the linear model framework by using a linear predictor and a link function
- link function: describe the relationship betweeen the linear combination of the explanatory variables and the mean of the response variable
- Rcommand: **glm(response_variable∼ explanatory_variable,family= family_distribution)**

Most common family function :

Gaussian : Identity
Binomial : logit
Poisson : log
Neg binomial : log

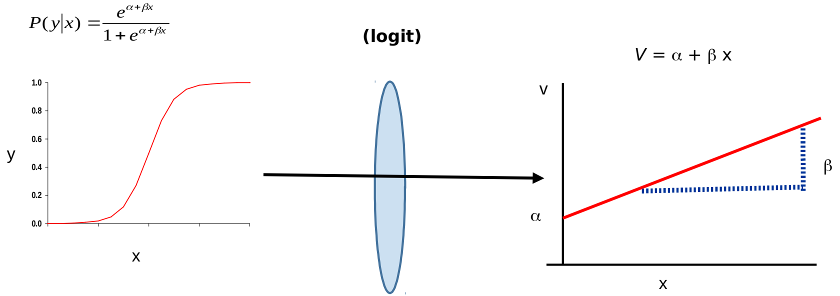


Binary data: Binomial

$$log(\frac{p}{1-p}) = \alpha + \beta x \quad (1)$$

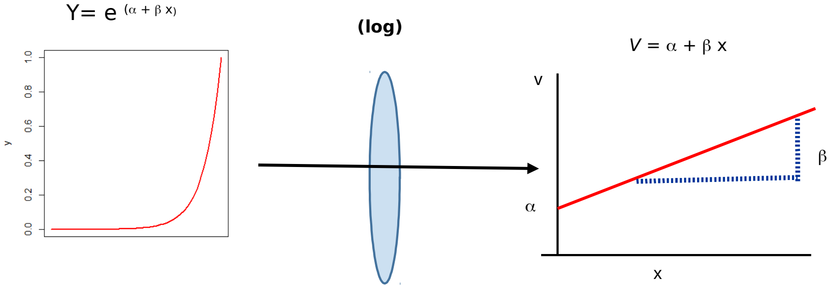
Rcommand:

glm(responsevariable$\sim$ *explanatory* variable,family="binomial")



Count data: Poisson or Negative binomial

$$log(y) = \alpha + \beta x \quad (2)$$



Selection of the best model

Choose the best model that fit our data by selecting the set of predictors that best explains the response variable(backward, forward, stepwise)
Based on AIC: the AIC is a measure of how well a model fits a dataset

- drop1
- add1
- step

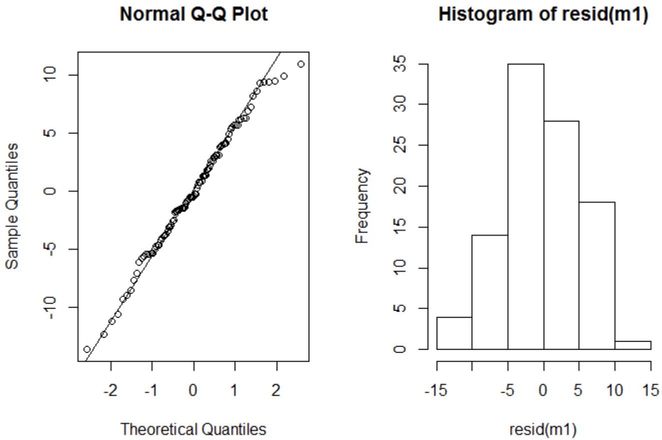
Model validation

Check that models assumption are not violated

- Residuals should be normally distributed with a mean of 0 and variance σ.
- homoscedasticity(the error should be the same accross all value of the explanatory variable, scatter plot of predicted value versus residual, there should be no clear pattern in the distribution)

Normality of the residuals

Make a QQplot to test the normality



Homoscedasticity

Plot the residuals vs fitted values