

Técnicas de Aprendizaje Estadístico

Introduction to R

Mathias Bourel

IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

March 12, 2019

¿What is R?

- R is the most powerful and most widely used statistical software
- It is a language and environment for statistical computing and graphics.
- It is an integrated suite of software facilities for data manipulation, calculation and graphical display.
- It provides a wide variety of classical and modern statistical techniques
- It is a collaborative project, where all users can upload their own codes and packages.
- It is **free**
- It runs on Windows, Mac and Linux
- It have many tools for visualization

You can download R at www.cran-project and R-studio at www.rstudio.com

Many tutorials: Introduction to R, R RefCard, etc.

Packages and Library

```
# list all available packages
> library()
# load packages
> library(MASS)
> library(rpart)
> library(e1071)
> library(ks)
# documentation on package
> library(help = rpart)
# download and install packages from CRAN or local files
> install.packages (...)
```

Basic commands

```
#load the package MASS (functions and datasets of the book
"Modern Applied Statistics with S").
> library(MASS)
# get help
> help.start()
> help(mvrnorm)
> ?lm
# load the data from the local directory
> read.table("miejemplo.txt", sep= ",", header=T)
> read.csv("miejemplo2.csv", sep= ",", header=T)
> write.table()
# quit R
> q()
```

Basic commands

```
#generate a sequence
> a = seq(0,10,0.5)
#generate samples from a specified distribution.
> set.seed(2019)
> x = rnorm(10, mean=0, sd=1) # samples from N(0,1)
> x = rnorm(100, mean=-2, sd=4) # samples from N(0,1)
> x = runif(50, min=0, max=1) # sample from Unif(0,1)
> x = runif(30, min=-2, max=2) # sample from Unif(-2,2)
#generate data from multivariate normal distribution
> mean1 = c(-2,2)
> cov1 = matrix(c(1,0.5,0.5,1),nrow=2)
> data = mvrnorm(15,mean1,cov1)
```

Basic commands

```
> x = seq(0,10,0.5)
> y = seq(5,15,0.5)
> plot(x,y)
> lines(x,y)
> abline(a=0,b=1)
# Histograms
>?hist
>hist(x)
#Boxplots
> barplot(x)
# plot density
> plot(density(x))
>a=rnorm(50,mean=0,sd=1)
>plot(density(a))
>curve(dnorm,col='red',add=T)
```

Playing with data I

```
# load data
> data(iris)
# show data
> iris
# make the data columns available by name
> attach(iris)
> Sepal.Length
#compute mean and variance, summary
> mean(Sepal.Length)
> var(Sepal.Length)
> summary(iris)
```

Playing with data II

```
> data(cars)
> cars
> attach(cars)
> model= lm(dist ~ speed)
> summary(model)
> plot(cars, main="Stopping Distance versus Speed")
> lines(speed, fitted(mymodel), type="l", lty=1, col=2)
> detach(cars)
```


Writing a function

```
> cuentas = function(x, y){  
>   a = x+2*y  
>   b=x-y  
>   c=x/y  
>   final=c(mean(a),sd(a),mean(b),sd(b),mean(c),sd(c))  
>   return(final)  
> }  
> x <- runif(50,0,1)  
> y <- rnorm(50,0,3)  
> result=cuentas(x,y)  
> print(result)
```

Writing a function - Ifelse

```
>number=14
>if(number%% 2==0) print("the number is even") else print("the number is odd")

> z=c(3,2,0,6,1,0)
> for(i in 1:length(z)) {
> if (z[i]==0) z[i]=500
> }

> z=c(3,2,0,6,1,0)
> for(i in 1:length(z)) {
> if (z[i]%%2) z[i]=10 else (z[i]=9)
> }

> w=c(3,2,0,6,1,0)
> ifelse(w==0, 500, w)

> neww=cbind(w, "change"=ifelse(w==0, 500, w))
```