Statistical learning

Introduction to R

V. Lefieux

1 First steps

1.1 Assignments

```
a <- 1
a
```

```
## [1] 1
```

```
a=1
a
```

```
## [1] 1
```

```
a=1/0
a
```

```
## [1] Inf
```

```
a=NaN
a
```

```
## [1] NaN
```

```
a=pi
a # Some text
```

```
## [1] 3.141593
```

1.2 Environment

```
ls()
```

```
## [1] "a"
```

```
rm(a)
ls()
```

```
## character(0)
```

It's possible to change the working directory using `setwd(dir=file.choose())`.

1.3 Helps

```
help(rm)
?rm
help(graphics)
```

2 Data structures

2.1 Vectors

```
x=2:9
x
```

```
## [1] 2 3 4 5 6 7 8 9
```

```
is.vector(x)
```

```
## [1] TRUE
```

```
x[2]
```

```
## [1] 3
```

```
x[c(2,2,1,3)]
```

```
## [1] 3 3 2 4
```

```
x[-1]
```

```
## [1] 3 4 5 6 7 8 9
```

```
x[-c(1,2)]
```

```
## [1] 4 5 6 7 8 9
```

```
seq(2,9,by=2)
```

```
## [1] 2 4 6 8
```

```
y=c(1,20,25)
z=c(x,y)
z^2
```

```
##  [1]   4   9  16  25  36  49  64  81   1 400 625
```

```
z*z
```

```
##  [1]   4   9  16  25  36  49  64  81   1 400 625
```

```
rep(x,3)
```

```
##  [1] 2 3 4 5 6 7 8 9 2 3 4 5 6 7 8 9 2 3 4 5 6 7 8 9
```

```
rep(x,each=3)
```

```
##  [1] 2 2 2 3 3 3 4 4 4 5 5 5 6 6 6 7 7 7 8 8 8 9 9 9
```

```
is.finite(c(-Inf,-1,2))
```

```
## [1] FALSE  TRUE  TRUE
```

```r
is.nan(c(NaN,1,10))
```

```
## [1]  TRUE FALSE FALSE
```

```r
x=-10:10
x
```

```
##  [1] -10  -9  -8  -7  -6  -5  -4  -3  -2  -1   0   1   2   3   4   5   6
## [18]   7   8   9  10
```

```r
y=seq(-10,10,2)^2
y
```

```
##  [1] 100  64  36  16   4   0   4  16  36  64 100
```

```r
which((y<2)|(y>20))
```

```
## [1]  1  2  3  6  9 10 11
```

```r
x[which((y<2)|(y>20))]
```

```
## [1] -10  -9  -8  -5  -2  -1   0
```

```r
which((y>2)&(y<20))
```

```
## [1] 4 5 7 8
```

```r
y[which((y>2)&(y<20))]
```

```
## [1] 16  4  4 16
```

```r
which(y==max(y))
```

```
## [1]  1 11
```

```r
order(y,decreasing=FALSE)
```

```
##  [1]  6  5  7  4  8  3  9  2 10  1 11
```

```r
y[order(y,decreasing=FALSE)]
```

```
##  [1]   0   4   4  16  16  36  36  64  64 100 100
```

2.2 Matrixes

```r
x=1:12
dim(x)=c(3,4)
x
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    4    7   10
## [2,]    2    5    8   11
## [3,]    3    6    9   12
```

```r
x[,3]
```

```
## [1] 7 8 9
```

```r
x[2,]
```

```
## [1]  2  5  8 11
```

```r
apply(x,1,mean)
```

```
## [1] 5.5 6.5 7.5
```

```r
apply(x,2,max)
```

```
## [1]  3  6  9 12
```

```r
x*x
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1   16   49  100
## [2,]    4   25   64  121
## [3,]    9   36   81  144
```

```r
y=matrix(1:12,nrow=3,byrow=TRUE)
t(y)
```

```
##      [,1] [,2] [,3]
## [1,]    1    5    9
## [2,]    2    6   10
## [3,]    3    7   11
## [4,]    4    8   12
```

```r
z=matrix(1:4,nrow=2,byrow=TRUE)
z
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

```r
x=matrix(1:4,nrow=2,byrow=FALSE)
x
```

```
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
```

```r
x^2
```

```
##      [,1] [,2]
## [1,]    1    9
## [2,]    4   16
```

```r
x*x
```

```
##      [,1] [,2]
## [1,]    1    9
## [2,]    4   16
```

```r
x%*%x
```

```
##      [,1] [,2]
## [1,]    7   15
## [2,]   10   22
```

2.3 Factors

```r
x=c("blue","red","yellow","yellow","yellow","red")
is.factor(x)
```

```
## [1] FALSE
```

```r
y=factor(x)
is.factor(y)
```

```
## [1] TRUE
```

```r
levels(y)
```

```
## [1] "blue"   "red"    "yellow"
```

2.4 Data frames

```r
x=c("blue","red","yellow","yellow","yellow","red")
y=1:6
z=data.frame(couleur=x,y=y)
z
```

```
##   couleur y
## 1    blue 1
## 2     red 2
## 3  yellow 3
## 4  yellow 4
## 5  yellow 5
## 6     red 6
```

```r
is.data.frame(z)
```

```
## [1] TRUE
```

3 Data import

The file *data_ozone.txt* contains 112 records from Rennes-summer 2001 (source: Laboratoire de mathématiques appliquées de l'Agrocampus Ouest).

Name	Description
obs	Observation number
MaxO3	Daily maximum of ozone concentration ($\mu\, gr/m^3$)
T9, T12, T15	Temperature (at 09:00, 12:00 and 15:00)
Ne9, Ne12, Ne15	Cloud cover (at 09:00, 12:00 and 15:00)
Vx9, Vx12, Vx15	East-west component of the wind (at 09:00, 12:00 and 15:00)
MaxO3v	Daily maximum of ozone concentration for the day before
wind	Wind direction at 12:00
rain	Rainy or dry day

A first solution:

```r
ozone=read.table("/Users/vincentlefieux/Documents/Data/data_ozone.txt",
                 header=TRUE,sep=";",dec=",")
```

Another solution:

```r
setwd("/Users/vincentlefieux/Documents/Data/")
ozone=read.table("data_ozone.txt",header=TRUE,sep=";",dec=",")
```

Or `ozone=read.table("data_ozone.txt",header=TRUE,sep=";",dec=",")` if *data_ozone.txt* is in the working directory.

For Windows: `ozone=read.table("c:/Documents and settings/vincentlefieux/Data/data_ozone.txt")`

```r
rownames(ozone)
```

```
##   [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"
##  [12] "12"  "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"
##  [23] "23"  "24"  "25"  "26"  "27"  "28"  "29"  "30"  "31"  "32"  "33"
##  [34] "34"  "35"  "36"  "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"
##  [45] "45"  "46"  "47"  "48"  "49"  "50"  "51"  "52"  "53"  "54"  "55"
##  [56] "56"  "57"  "58"  "59"  "60"  "61"  "62"  "63"  "64"  "65"  "66"
##  [67] "67"  "68"  "69"  "70"  "71"  "72"  "73"  "74"  "75"  "76"  "77"
##  [78] "78"  "79"  "80"  "81"  "82"  "83"  "84"  "85"  "86"  "87"  "88"
##  [89] "89"  "90"  "91"  "92"  "93"  "94"  "95"  "96"  "97"  "98"  "99"
## [100] "100" "101" "102" "103" "104" "105" "106" "107" "108" "109" "110"
## [111] "111" "112"
```

```r
colnames(ozone)
```

```
##  [1] "obs"    "maxO3"  "T9"     "T12"    "T15"    "Ne9"    "Ne12"
##  [8] "Ne15"   "Vx9"    "Vx12"   "Vx15"   "maxO3v" "wind"   "rain"
```

```r
head(ozone)
```

```
##   obs maxO3   T9  T12  T15 Ne9 Ne12 Ne15     Vx9    Vx12    Vx15 maxO3v
## 1 601    87 15.6 18.5 18.4   4    4    8  0.6946 -1.7101 -0.6946     84
## 2 602    82 17.0 18.4 17.7   5    5    7 -4.3301 -4.0000 -3.0000     87
## 3 603    92 15.3 17.6 19.5   2    5    4  2.9544  1.8794  0.5209     82
## 4 604   114 16.2 19.7 22.5   1    1    0  0.9848  0.3473 -0.1736     92
## 5 605    94 17.4 20.5 20.4   8    8    7 -0.5000 -2.9544 -4.3301    114
## 6 606    80 17.7 19.8 18.3   6    6    7 -5.6382 -5.0000 -6.0000     94
##    wind rain
## 1 North  Dry
## 2 North  Dry
## 3  East  Dry
## 4 North  Dry
## 5  West  Dry
## 6  West Rain
```

```r
tail(ozone)
```

```
##     obs maxO3   T9  T12  T15 Ne9 Ne12 Ne15     Vx9    Vx12    Vx15 maxO3v
## 107 924    76 13.3 17.7 17.7   7    7    6 -0.9397 -0.7660 -0.5000     92
## 108 925    84 13.3 17.7 17.8   3    5    6  0.0000 -1.0000 -1.2856     76
## 109 927    77 16.2 20.8 22.1   6    5    5 -0.6946 -2.0000 -1.3681     71
## 110 928    99 16.9 23.0 22.6   6    4    7  1.5000  0.8682  0.8682     77
## 111 929    83 16.9 19.8 22.1   6    5    3 -4.0000 -3.7588 -4.0000     99
## 112 930    70 15.7 18.6 20.7   7    7    7  0.0000 -1.0419 -4.0000     83
##      wind rain
## 107  West Rain
## 108 South  Dry
## 109 South Rain
```

```
## 110 South  Dry
## 111  West Rain
## 112 South  Dry
```

```r
ozone=subset(ozone,select=-c(obs))
colnames(ozone)
```

```
##  [1] "maxO3"  "T9"     "T12"    "T15"    "Ne9"    "Ne12"   "Ne15"
##  [8] "Vx9"    "Vx12"   "Vx15"   "maxO3v" "wind"   "rain"
```

```r
Tmean=(ozone$T9+ozone$T12+ozone$T15)/3
Tmean
```

```
##   [1] 17.50000 17.70000 17.46667 19.46667 19.43333 18.60000 15.76667
##   [8] 17.10000 19.03333 21.03333 18.93333 18.43333 19.10000 17.76667
##  [15] 16.46667 18.83333 24.16667 20.66667 23.60000 26.86667 24.43333
##  [22] 31.13333 24.20000 22.46667 21.76667 23.96667 22.20000 22.13333
##  [29] 27.53333 29.53333 18.70000 19.00000 23.16667 19.46667 21.06667
##  [36] 22.13333 20.86667 20.86667 16.53333 14.66667 18.16667 18.50000
##  [43] 19.73333 18.73333 15.53333 18.36667 19.30000 20.96667 21.36667
##  [50] 21.73333 21.86667 25.26667 26.96667 26.26667 26.56667 28.56667
##  [57] 29.20000 24.86667 28.66667 22.40000 20.60000 19.86667 18.36667
##  [64] 21.63333 20.63333 20.60000 18.66667 19.30000 21.63333 24.26667
##  [71] 21.76667 28.33333 23.70000 19.60000 21.03333 24.13333 25.70000
##  [78] 27.06667 25.93333 31.33333 23.53333 24.56667 24.06667 18.43333
##  [85] 18.26667 19.03333 19.80000 21.53333 17.10000 18.20000 18.13333
##  [92] 18.53333 20.06667 18.83333 19.46667 17.86667 17.23333 17.43333
##  [99] 16.56667 14.50000 13.96667 15.66667 16.96667 16.96667 18.40000
## [106] 16.83333 16.23333 16.26667 19.70000 20.83333 19.60000 18.33333
```

```r
colnames(ozone)
```

```
##  [1] "maxO3"  "T9"     "T12"    "T15"    "Ne9"    "Ne12"   "Ne15"
##  [8] "Vx9"    "Vx12"   "Vx15"   "maxO3v" "wind"   "rain"
```

```r
ozone$Tmax=max(ozone$T9,ozone$T12,ozone$T15)
colnames(ozone)
```

```
##  [1] "maxO3"  "T9"     "T12"    "T15"    "Ne9"    "Ne12"   "Ne15"
##  [8] "Vx9"    "Vx12"   "Vx15"   "maxO3v" "wind"   "rain"   "Tmax"
```

```r
ozone=data.frame(ozone,Tmean)
colnames(ozone)
```

```
##  [1] "maxO3"  "T9"     "T12"    "T15"    "Ne9"    "Ne12"   "Ne15"
##  [8] "Vx9"    "Vx12"   "Vx15"   "maxO3v" "wind"   "rain"   "Tmax"
## [15] "Tmean"
```

4 Mathematical operations

4.1 Random numbers simulation

```r
sample(1:10,5)
```

```
## [1] 7 3 8 9 1
```

```r
sample(1:10,5) # another time
```

```
## [1] 7 9 2 8 6
```

```r
sample(1:10,10,replace=TRUE)
```

```
##  [1] 1 8 2 9 6 7 2 3 9 8
```

```r
sample(c("yes","not"),10,replace=TRUE,prob=c(0.7,0.3))
```

```
##  [1] "yes" "yes" "yes" "not" "not" "yes" "yes" "yes" "yes" "yes"
```

```r
rbinom(10,size=20,prob=.5) # Binomial distribution
```

```
##  [1] 16 11 11 12 11  9  9  9 10  9
```

```r
rnorm(10) # Gaussian distribution
```

```
##  [1]  0.09453009 -0.10211704 -1.78599242  0.25571374  1.43583628
##  [6]  0.98103957 -0.19170978 -0.31481358 -1.92940820 -0.12774358
```

```r
runif(10,min=0,max=1) # Uniform distribution
```

```
##  [1] 0.07396396 0.51405142 0.38085818 0.30674079 0.09780119 0.90172092
##  [7] 0.08966613 0.55866674 0.21945320 0.95840149
```

4.2 Calculations

```r
sqrt(5)*1:10
```

```
##  [1]  2.236068  4.472136  6.708204  8.944272 11.180340 13.416408 15.652476
##  [8] 17.888544 20.124612 22.360680
```

```r
sqrt(1:10)
```

```
##  [1] 1.000000 1.414214 1.732051 2.000000 2.236068 2.449490 2.645751
##  [8] 2.828427 3.000000 3.162278
```

```r
exp(1:10)
```

```
##  [1]     2.718282     7.389056    20.085537    54.598150   148.413159
##  [6]   403.428793  1096.633158  2980.957987  8103.083928 22026.465795
```

```r
log(1:10)
```

```
##  [1] 0.0000000 0.6931472 1.0986123 1.3862944 1.6094379 1.7917595 1.9459101
##  [8] 2.0794415 2.1972246 2.3025851
```

```r
(1:10)^2
```

```
##  [1]   1   4   9  16  25  36  49  64  81 100
```

```r
sin(1:10)+cos(1:10)
```

```
##  [1]  1.3817733  0.4931506 -0.8488725 -1.4104461 -0.6752621  0.6807548
##  [7]  1.4108889  0.8438582 -0.4990118 -1.3830926
```

```r
round(c(1.4345,1.926),digits=0)
```

```
## [1] 1 2
```

```r
round(c(1.4345,1.926),digits=1)
```

```
## [1] 1.4 1.9
```

```r
round(c(1.4345,1.926),digits=2)
```

```
## [1] 1.43 1.93
```

```r
trunc(c(1.4345,1.926))
```

```
## [1] 1 1
```

4.3 Functions

```r
testfonc=function(x)
  {
  a=x^2
  b=x-3
  return(list(a=a,b=b))
}
resfonc=testfonc(3)
resfonc$a
```

```
## [1] 9
```

```r
resfonc$b
```

```
## [1] 0
```

5 Descriptive statistics

5.1 Summaries

```r
summary(ozone)
```

```
##      maxO3              T9              T12             T15
##  Min.   : 42.00   Min.   :11.30   Min.   :14.00   Min.   :14.90
##  1st Qu.: 70.75   1st Qu.:16.20   1st Qu.:18.60   1st Qu.:19.27
##  Median : 81.50   Median :17.80   Median :20.55   Median :22.05
##  Mean   : 90.30   Mean   :18.36   Mean   :21.53   Mean   :22.63
##  3rd Qu.:106.00   3rd Qu.:19.93   3rd Qu.:23.55   3rd Qu.:25.40
##  Max.   :166.00   Max.   :27.00   Max.   :33.50   Max.   :35.50
##       Ne9             Ne12            Ne15           Vx9
##  Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :-7.8785
##  1st Qu.:3.000   1st Qu.:4.000   1st Qu.:3.00   1st Qu.:-3.2765
##  Median :6.000   Median :5.000   Median :5.00   Median :-0.8660
##  Mean   :4.929   Mean   :5.018   Mean   :4.83   Mean   :-1.2143
##  3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:7.00   3rd Qu.: 0.6946
##  Max.   :8.000   Max.   :8.000   Max.   :8.00   Max.   : 5.1962
##       Vx12             Vx15            maxO3v          wind      rain
##  Min.   :-7.878   Min.   :-9.000   Min.   : 42.00   East :10   Dry :69
##  1st Qu.:-3.565   1st Qu.:-3.939   1st Qu.: 71.00   North:31   Rain:43
##  Median :-1.879   Median :-1.550   Median : 82.50   South:21
```

```
##  Mean   :-1.611   Mean   :-1.691   Mean   : 90.57   West :50
##  3rd Qu.: 0.000   3rd Qu.: 0.000   3rd Qu.:106.00
##  Max.   : 6.578   Max.   : 5.000   Max.   :166.00
##       Tmax          Tmean
##  Min.   :35.5   Min.   :13.97
##  1st Qu.:35.5   1st Qu.:18.36
##  Median :35.5   Median :19.77
##  Mean   :35.5   Mean   :20.84
##  3rd Qu.:35.5   3rd Qu.:23.26
##  Max.   :35.5   Max.   :31.33
```

```r
mean(ozone$T12)
```

```
## [1] 21.52679
```

```r
median(ozone$T12)
```

```
## [1] 20.55
```

```r
var(ozone$T12)
```

```
## [1] 16.34036
```

```r
sd(ozone$T12) # Standart deviation
```

```
## [1] 4.042321
```

```r
cov(ozone$T12,ozone$maxO3) # Covariance
```

```
## [1] 89.36026
```

```r
cor(ozone$T12,ozone$maxO3) # Correlation
```

```
## [1] 0.7842623
```

```r
quantile(ozone$T12)
```

```
##    0%   25%   50%   75%  100%
## 14.00 18.60 20.55 23.55 33.50
```

```r
quantile(ozone$T12,prob=seq(0,1,by=0.1))
```

```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
## 14.00 17.60 18.20 18.96 19.80 20.55 21.78 22.87 24.56 27.69 33.50
```

5.2 Contingency table

```r
table(ozone$wind,ozone$rain)
```

```
##
##         Dry Rain
##   East    8    2
##   North  21   10
##   South  16    5
##   West   24   26
```