Basic R Programming

Author: Chiachun Chiu (jiajyun.ciou@gmail.com (mailto:jiajyun.ciou@gmail.com))

2015年03月16日

About R 

> R (http://www.r-project.org/about.html) is a **language** and **environment** for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R. …… R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

How/where to obtain and install R?

If you use linux as your default os, you can install R from the package repositories of each distribution directly. Alternatively, you can download R binary-version or source code from **CRAN** if you use M$ windows or Mac OS.

Ubuntu users

- Add **deb http://my.favorite.cran.mirror/bin/linux/ubuntu (http://my.favorite.cran.mirror/bin/linux/ubuntu) [version name of ubuntu]/** into sources.list
- sudo apt-key adv –keyserver keyserver.ubuntu.com –recv-keys E084DAB9
- sudo apt-get update
- sudo apt-get install r-base

Redhat/Fedora/CentOS users

- Add **EPEL** repository
- sudo yum install R

Other OS users - CRAN (mirror-sites in Taiwan)

- NTU (http://cran.csie.ntu.edu.tw/)
- YZU (http://ftp.yzu.edu.tw/CRAN/)

Using RStudio as your default R-programming IDE

About RStudio (http://www.rstudio.com/)

> RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

Install the most suitable version of RStudio for your needs.

- Desktop version: Access RStudio locally.
- Server version: Access via a web browse.

Other choices

- REVOLUTION ANALYTICS : Revolution R (http://www.revolutionanalytics.com/revolution-r-enterprise)
- : Vim-R-plugin (http://www.lepem.ufc.br/jaa/vim-r-plugin.html)
- Any other text editors: gedit, emacs(+ESS), eclipse and etc.

Resources of learning R

Books & web sites

- The R Book (http://as.wiley.com/WileyCDA/WileyTitle/productCd-0470973927.html)
- Introduction to Probability and Statistics Using R (http://ipsur.org/)
- Introduction to Scientific Programming and Simulation Using R (http://www.amazon.com/Introduction-Scientific-Programming-Simulation-Chapman/dp/1420068725)
- The Art of R Programming (http://www.amazon.com/The-Art-Programming-Statistical-Software/dp/1593273843)
- R Tutorial (http://www.cyclismo.org/tutorial/R/)
- Programming in R (http://manuals.bioinformatics.ucr.edu/home/programming-in-r)
- Advanced R (http://adv-r.had.co.nz/)

Forums & communities

- R-bloggers (http://www.r-bloggers.com/)
- Taiwan R User Group (http://www.meetup.com/Taiwan-R)
- Stack Overflow (http://stackoverflow.com/questions/tagged/r)
- Ptt R_Language (https://www.ptt.cc/bbs/R_Language/index.html)

The most important step when begining to learn R is using help()

help() & help.search()

```
help(help)
help.search("standard deviation")
```

? & ??

```
?mean
??hypergeometric
```

Package installation & PATH setting

Installing packages

```
# Download Pkgs from CRAN repository & install
install.packages('rmarkdown',                  # Package name
                 repo="http://cran.csie.ntu.edu.tw", # The url of CRAN-repository
                 destdir="~/Download",         # The directory where downloaded pkgs are stored
                 lib=.libPaths()[1])           # The directory where to install pkgs

# Install Pkgs from downloaded source code
install.packages('~/Download/rmarkdown_0.5.1.tar.gz',
                 repos=NULL,
                 type="source",
                 lib=.libPaths()[1])
```

Setting PATH

```
.libPaths(new)
```

Some Pkgs shoud be downloaded/installed from R-forge (https://r-forge.r-project.org/)

Set `install.packages(Pkg, repo='http://R-Forge.R-project.org')`

Using the package installed

```
library(Pkg)
require(Pkg) # Avoid to use this!
```

What is the difference between require() and library() (http://stackoverflow.com/questions/5595512/what-is-the-difference-between-require-and-library)

Bioconductor

About Bioconductor (http://www.bioconductor.org/)

> Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development.

Install Pkgs from Bioconductor

```
source("http://bioconductor.jp/biocLite.R") # Mirror site in Japan
biocLite("Package Name")
```

Basic operation

```
5+5
5-3
5*3
5/3
5^3
10%%3

# Variable declaration
x <- 5 # '<-' is assign operator in R, which is equivalent to '='
y <- function(i) mean(i)
```

Data and object types

Data types

- numeric: c(1:3, 5 ,7)
- character: c("1","2","3"); LETTERS[1:3]
- logical: TRUE; FALSE
- complex: 1, b, 3

Object types

- **vector**: the data types of all elements in a vector must be consistent!

```
x <- 1:5
y <- c(6,7,8,9,10)
z <- x - y
print(z)
```

```
## [1] -5 -5 -5 -5 -5
```

```
# Vectorized code performs better!
a <- 1:100000
system.time(mean(a))
```

```
##    user  system elapsed
##       0       0       0
```

```
total <- 0
system.time(for (i in a) {total <- total + i; total/100000})
```

```
##    user  system elapsed
##   0.040   0.000   0.041
```

- **matrix**

```
x <- matrix(rnorm(100), nr=20, nc=5)
print(x)
```

```
##              [,1]       [,2]       [,3]       [,4]       [,5]
##  [1,]  1.061073327  0.16918686 -0.06332983  0.18502634  1.53782180
##  [2,] -2.198584485 -0.11298950  1.83166296  1.29242442 -0.99674780
##  [3,]  0.091173055  1.83391152  1.01952557  1.02375958  0.63436552
##  [4,] -2.764016056  2.32801138  0.16697182 -0.08757874 -1.02939810
##  [5,]  0.907844361  0.17523699  0.62210775  0.58176899  0.51919079
##  [6,]  0.464909500  0.44614795  0.53591792 -0.56231754  0.32303444
##  [7,] -0.154265058 -1.71913498  0.08241890 -0.71482833 -0.05796294
##  [8,]  0.051215002 -0.22018453  0.96643633 -0.94732819 -0.57563312
##  [9,] -0.625286782 -0.21714787  0.32574418 -0.36198554  0.99808041
## [10,] -0.804825830  0.87572601 -1.41002646 -1.76296089  0.15584392
## [11,]  2.115877990  1.29374224 -0.30746104 -0.77154569  0.07066943
## [12,] -0.482892174  0.16815861  0.53816792  1.04005936  0.57007232
## [13,]  0.174797263  0.37595993  0.30428119 -0.16849691  0.40437195
## [14,] -0.975788940 -0.07958521  0.37545193  1.08159152 -0.46566451
## [15,] -0.408961092  0.13661077 -0.79918839  0.81350170  0.93155066
## [16,]  1.760093697 -1.30313228 -0.61877764 -0.35185296 -1.07378676
## [17,] -1.393555267 -2.23416186 -1.87091754 -1.78409466 -1.11475631
## [18,] -1.492962866 -0.35010233  3.11650703  0.06084068 -0.35560597
## [19,]  0.146860899 -0.95841843 -0.11137265  0.15591210 -1.72724113
## [20,] -0.002815384 -0.23550138 -0.27074177  0.18945211  0.49373900
```

```
x[1,3]
x[2:4,]
x[,3:5]
x %*% t(x)

# A matrix is a vector with subscripts!
x[1:3]
x[1:3,1]
```

- **array**

```
y <- array(rnorm(64), c(8,4,2))
print(y) # An array is also a vector with subscripts!
```

```
## , , 1
##
##            [,1]       [,2]       [,3]       [,4]
## [1,]  0.29626565  0.85907326  0.37145797  0.3759709
## [2,]  0.03271301  1.56081671 -0.40933852  0.4879662
## [3,]  0.09431310 -0.63121001 -0.65800667 -1.3624521
## [4,]  0.33525617  1.82403830  0.11957709 -0.6410010
## [5,]  0.08114325  0.52315298 -1.59316418 -1.7534226
## [6,]  1.53482058 -0.07411969  0.46223350  2.4801945
## [7,]  1.52694017  0.41640368 -0.86909372 -0.2344898
## [8,] -0.57808513  1.30839382  0.04567981  1.6877330
##
## , , 2
##
##            [,1]       [,2]       [,3]       [,4]
## [1,]  1.1647846  0.22678385  0.21999112 -0.4926574
## [2,] -1.3677179 -0.30990515 -0.11823946 -0.3748530
## [3,]  1.7830684  0.95224588 -2.58860639  2.0753362
## [4,] -0.1109423  0.01243485 -0.53975661  1.0348551
## [5,] -1.2792004  0.20633487  0.48396615  0.1805941
## [6,]  0.2225891 -0.56688274 -0.03399781  0.4548411
## [7,]  0.5700414  0.59499381  0.40185669 -0.8397162
## [8,]  1.2124715 -1.03291098  0.53448620  0.1941677
```

- **list**: the data types of elements in a list could be complex

```
x<-list(1:5, c("a","b","c"), matrix(rnorm(10),nr=5,nc=2))
print(x)
```

```
## [[1]]
## [1] 1 2 3 4 5
##
## [[2]]
## [1] "a" "b" "c"
##
## [[3]]
##            [,1]       [,2]
## [1,] -0.36323571  2.28438880
## [2,]  0.52022800 -0.02840625
## [3,]  1.04497913 -0.66628924
## [4,]  0.04955235  0.33631166
## [5,]  2.22026407 -1.25465550
```

```
x$mylist <- x
print(x)
```

```
## [[1]]
## [1] 1 2 3 4 5
##
## [[2]]
## [1] "a" "b" "c"
##
## [[3]]
##            [,1]       [,2]
## [1,] -0.36323571  2.28438880
## [2,]  0.52022800 -0.02840625
## [3,]  1.04497913 -0.66628924
## [4,]  0.04955235  0.33631166
## [5,]  2.22026407 -1.25465550
##
## $mylist
## $mylist[[1]]
## [1] 1 2 3 4 5
##
## $mylist[[2]]
## [1] "a" "b" "c"
##
## $mylist[[3]]
##            [,1]       [,2]
## [1,] -0.36323571  2.28438880
## [2,]  0.52022800 -0.02840625
## [3,]  1.04497913 -0.66628924
## [4,]  0.04955235  0.33631166
## [5,]  2.22026407 -1.25465550
```

- **data frame**: a data frame is collection of multiple lists with the same length

```
df<-data.frame(num=1:10,
        char=LETTERS[1:10],
        logic=sample(c(TRUE,FALSE), 10, replace=TRUE))

df
```

```
##    num char logic
## 1    1    A  TRUE
## 2    2    B FALSE
## 3    3    C FALSE
## 4    4    D  TRUE
## 5    5    E  TRUE
## 6    6    F FALSE
## 7    7    G FALSE
## 8    8    H FALSE
## 9    9    I  TRUE
## 10  10    J  TRUE
```

```
df$char
```

```
##  [1] A B C D E F G H I J
## Levels: A B C D E F G H I J
```

```
df$logic[5:7]
```

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

- **factor**: An R **factor** might be viewed simply as **a vector with a bit more information added** (though, as seen below, it's different from this internally). That extra information consists of a record of the distinct values in that vector, called **levels**.

```
x <- c(5, 12, 32, 12)
xf <- factor(x)
print(xf)
```

```
## [1] 5  12 32 12
## Levels: 5 12 32
```

So…. a factor looks like a vector, right?

```
str(xf) # Here str stands for structure. This function shows the internal structure of any R object.
```

```
##  Factor w/ 3 levels "5","12","32": 1 2 3 2
```

```
unclass(xf)
```

```
## [1] 1 2 3 2
## attr(,"levels")
## [1] "5"  "12" "32"
```

```
length(xf)
```

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

What??? What are you talking about?

```
x <- c(5, 12, 13, 12)
xff <- factor(x, levels=c(5, 12, 13, 88))
xff
```

```
## [1] 5  12 13 12
## Levels: 5 12 13 88
```

```
xff[2] <- 88
xff
```

```
## [1] 5  88 13 12
## Levels: 5 12 13 88
```

```
xff[2] <- 28 # You cannot sneak in an "illegal" level
```

```
## Warning in `[<-.factor`(`*tmp*`, 2, value = 28): invalid factor level, NA
## generated
```

- **table**: Another common way to store information is in a table.

```
# One way table
a <- factor(c("A","A","B","A","B","B","C","A","C"))
a
```

```
## [1] A A B A B B C A C
## Levels: A B C
```

```
a.table <- table(a)
a.table
```

```
## a
## A B C
## 4 3 2
```

```
attributes(a.table)
```

```
## $dim
## [1] 3
##
## $dimnames
## $dimnames$a
## [1] "A" "B" "C"
##
##
## $class
## [1] "table"
```

```
# Two way table
a <- c("Sometimes","Sometimes","Never","Always","Always","Sometimes","Sometimes","Never")
b <- c("Maybe","Maybe","Yes","Maybe","Maybe","No","Yes","No")
twoway.table <- table(a,b)
twoway.table
```

```
##            b
## a          Maybe No Yes
##   Always       2  0   0
##   Never        0  1   1
##   Sometimes    2  1   1
```

```
# An example
sexsmoke<-matrix(c(70,120,65,140),ncol=2,byrow=TRUE)
rownames(sexsmoke)<-c("male","female")
colnames(sexsmoke)<-c("smoke","nosmoke")
sexsmoke <- as.table(sexsmoke)
sexsmoke
```

```
##        smoke nosmoke
## male      70     120
## female    65     140
```

Control structures

Conditional excutions

- **equal**: ==
- **not equal**: !=
- **greater/less than**: >, <
- **greater/less than or equal**: >=, <=

Logical operators

- **and**: &, &&
- **or**: |, ||
- **not**: !

if-else statements

```
if (cond1==TRUE) {cmd1} else {cmd2}
```

```
# Example
if (1 == 0) {
   print(1)
} else {
   print(2)
}
```

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

ifelse statements (ternary operator in R)

```
ifelse(test, true_value, false_value)
```

```
x <- 1:10
ifelse(x<5|x>8, x, 0)
```

```
##  [1]  1  2  3  4  0  0  0  0  9 10
```

switch-case statements

```
AA <- 'foo'
switch(AA,
       foo = {print('AA is foo')},
       bar = {print('AA is bar')},
       {print('Default')}
)
```

```
## [1] "AA is foo"
```

Loops

For loop

```
for (var in vector) {
   statement
}
```

```
# Example
mydf <- iris
head(mydf)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
myve <- NULL
for (i in 1:nrow(mydf)) {
   myve <- c(myve, mean(as.numeric(mydf[i, 1:3])))
}
myve
```

```
##   [1] 3.333333 3.100000 3.066667 3.066667 3.333333 3.666667 3.133333
##   [8] 3.300000 2.900000 3.166667 3.533333 3.266667 3.066667 2.800000
##  [15] 3.666667 3.866667 3.533333 3.333333 3.733333 3.466667 3.500000
##  [22] 3.433333 3.066667 3.366667 3.366667 3.200000 3.333333 3.400000
##  [29] 3.333333 3.166667 3.166667 3.433333 3.600000 3.700000 3.166667
##  [36] 3.133333 3.433333 3.300000 2.900000 3.333333 3.266667 2.700000
##  [43] 2.966667 3.366667 3.600000 3.066667 3.500000 3.066667 3.500000
##  [50] 3.233333 4.966667 4.700000 4.966667 3.933333 4.633333 4.333333
##  [57] 4.766667 3.533333 4.700000 3.933333 3.500000 4.366667 4.066667
##  [64] 4.566667 4.033333 4.733333 4.366667 4.200000 4.300000 4.000000
##  [71] 4.633333 4.300000 4.566667 4.533333 4.533333 4.666667 4.800000
##  [78] 4.900000 4.466667 3.933333 3.900000 3.866667 4.133333 4.600000
##  [85] 4.300000 4.633333 4.833333 4.333333 4.233333 4.000000 4.166667
##  [92] 4.566667 4.133333 3.533333 4.166667 4.300000 4.266667 4.466667
##  [99] 3.533333 4.200000 5.200000 4.533333 5.333333 4.933333 5.100000
## [106] 5.733333 3.966667 5.500000 5.000000 5.633333 4.933333 4.800000
## [113] 5.100000 4.400000 4.566667 4.966667 5.000000 6.066667 5.733333
## [120] 4.400000 5.266667 4.433333 5.733333 4.633333 5.233333 5.466667
## [127] 4.600000 4.666667 4.933333 5.333333 5.433333 6.033333 4.933333
## [134] 4.733333 4.766667 5.600000 5.100000 5.000000 4.600000 5.133333
## [141] 5.133333 5.033333 4.533333 5.300000 5.233333 4.966667 4.600000
## [148] 4.900000 5.000000 4.666667
```

while loop

```
while (condition) statements
```

```
# Example
z <- 0
while (z < 5) {
   z <- z + 2
   print(z)
}
```

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

apply loop

For matrix/array

```
apply(X, MARGIN, FUN, ARGS)

# Examples
apply(iris[,1:3], 1, mean)

x <- 1:10

apply(as.matrix(x), 1, function(i) {
   if (i < 5)
      i - 1
   else
      i/i
})
```

For vector/list

```
lapply(X, FUN)
sapply(X, FUN)
```

```
# Examples
mylist <- as.list(iris[1:3, 1:3])
mylist
```

```
## $Sepal.Length
## [1] 5.1 4.9 4.7
##
## $Sepal.Width
## [1] 3.5 3.0 3.2
##
## $Petal.Length
## [1] 1.4 1.4 1.3
```

```
lapply(mylist, sum) # Compute sum of each list component and return result as list
```

```
## $Sepal.Length
## [1] 14.7
##
## $Sepal.Width
## [1] 9.7
##
## $Petal.Length
## [1] 4.1
```

```
sapply(mylist, sum) # Compute sum of each list component and return result as vector
```

```
## Sepal.Length  Sepal.Width Petal.Length
##         14.7          9.7          4.1
```

More apply functions

- **tapply**
- **mapply**

Advanced R programming

Garbage collection

- **rm()**
- **gc()**

```
x <- as.matrix(read.table("test.csv", sep="\t")) # x is a 4500000 x 220 matrix
y <- apply(x, 1, mean)
rm(list=c("x","y"))
gc()
```

Use data.table to speed up acquirement of data

See Introduction to the data.table package in R (http://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.pdf)

> Fast aggregation of large data (e.g. 100GB in RAM), fast ordered joins, fast add/modify/delete of columns by group using no copies at all, list columns and a fast file reader (fread). Offers a natural and flexible syntax, for faster development. - from CRAN (http://cran.r-project.org/web/packages/data.table/index.html)

```
library(data.table)
grpsize <- ceiling(1e7/26^2)
DF <- data.frame(
   x=rep(LETTERS, each=26*grpsize),
   y=rep(letters, each=grpsize),
   v=runif(grpsize*26^2),
   stringsAsFactors=FALSE)
system.time(ans1 <- DF[DF$x=="R" & DF$y=="h",])
```

```
##    user  system elapsed
##   1.423   0.018   1.441
```

```
DT <- as.data.table(DF)
setkey(DT, x, y)
system.time(ans2 <- DT[list("R","h")])
```

```
##    user  system elapsed
##   0.004   0.000   0.004
```

Version control

> Version control is a system that records changes to a file or set of files over time so that you can recall specific versions later. For the examples in this book you will use software source code as the files being version controlled, though in reality you can do this with nearly any type of file on a computer. –from git (http://git-scm.com/book/en/v2/Getting-Started-About-Version-Control)

Installing git

- **ubuntu users**: sudo apt-get install git
- **RedHat/Fedora/CentOS users**: sudo yum install git
- **Mac users**: http://git-scm.com/download/mac (http://git-scm.com/download/mac)
- **Windows users**: http://git-scm.com/download/win (http://git-scm.com/download/win)

Git setup

```
git config --global user.name "YOUR NAME"
git config --global user.email you.email@address.org
git config --global core.ui true
git config --global core.editor vim

# For windows users
git config --global core.quotepath off
```

Git basics

```
## Initializing a repository in an existing directory
# Go to the project's directory and type
git init

# Add files you want to track
git add LICENSE
git add READ.md
got commit -m 'First commit. Add LICENSE & READ.md'

# Add new files
git add R.Rmd
git add helloworld.r
git commit -m 'Second commit. Add R.Rmd, helloworld.r'
git remote add origin
git push -u origin master

# Recover your codes to the last commit
git checkout -- filename
git reset --hard


## Cloning an existing repository
git clone https://github.com/godkin1211/Rcourses.git
git pull https://github.com/godkin1211/Rcourses.git
```

References

- 猴子都能懂的git入門 (http://backlogtool.com/git-guide/cn/intro/intro1_1.html)
- ProGit (http://git-scm.com/book/en/v2)
- Git Reference (http://gitref.org/creating/)
- ProGit (http://git-scm.com/book/en/v2)