

Programming for Data Science

Introduction to R language

Marco Beccuti

Università degli Studi di Torino
Dipartimento di Informatica

September 2021



The R Project

- Environment for statistical computing and graphics:
- Free software and Open-source;
- A simple programming language:
 - ▶ it is an open-source implementation of S language;
 - ▶ it is among the Top 10 Programming Languages in 2019 for *IEEE Spectrum Journal*.
- software and packages can be downloaded from:

www.cran.r-project.org

- Versions of R exist of Windows, MacOS, Linux and various other Unix-like OS.



Why to use R language

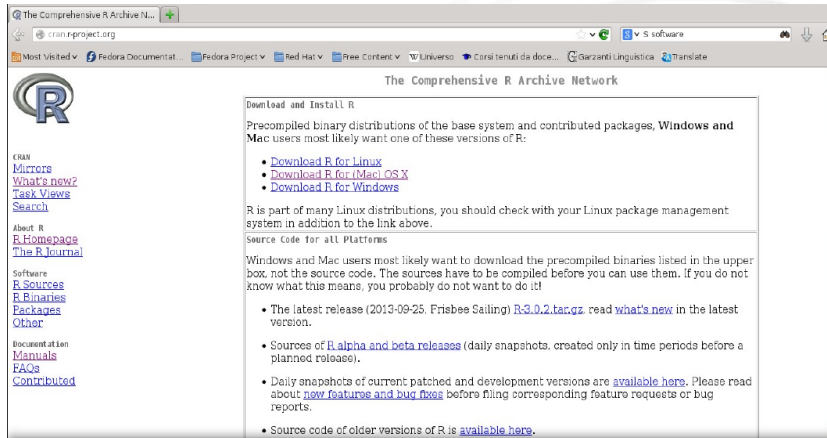
- Implement many common statistical procedures;
- Provide excellent graphics functionality;
- A convenient starting point for many data analysis projects
- Libraries (namely packages) can be automatically downloaded from:

www.cran.r-project.org

- It is standard for data mining and statistical analysis;
- Efficient data structures make programming easier.



Download and Install R language



The screenshot shows a web browser window with the address bar displaying "cran.r-project.org". The page title is "The Comprehensive R Archive Network". On the left side, there is a navigation menu with links for "CRAN", "Mirrors", "What's new?", "Task Views", "Search", "About R", "R Homepage", "The R Journal", "Software", "R Sources", "R Binaries", "Packages", "Other", "Document at Len", "Manuals", "FAQs", and "Contributed". The main content area is titled "Download and Install R" and contains the following text:

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

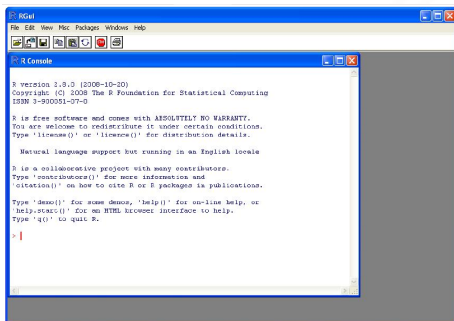
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2013-09-25, Frisbee Sailing) [R-3.0.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R.alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).

<http://cran.mirror.garr.it/mirrors/CRAN/>

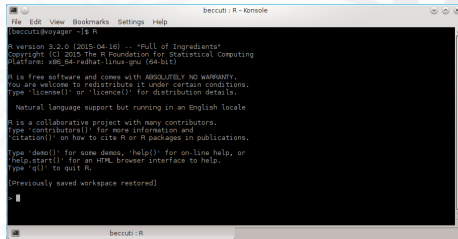
Download the appropriate version (w.r.t. your OS) and follow the instructions to install the programme.

R under GUI



A screenshot of the R GUI on a Windows operating system. The window title is 'RGui'. The menu bar includes 'File', 'Edit', 'View', 'Misc', 'Packages', 'Windows', and 'Help'. The toolbar contains icons for file operations and help. The main area is titled 'R Console' and displays the R version 3.3.0 (2006-10-20) splash screen. The text in the console reads: 'R version 3.3.0 (2006-10-20) Copyright (C) 2008 The R Foundation for Statistical Computing ISBN 3-900051-07-0 R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details. Natural language support but running in an English locale R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications. Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.' The prompt '>' is visible at the bottom left.

from Windows



A screenshot of the R console on a Linux operating system. The window title is 'beccuti: R - Konsole'. The menu bar includes 'File', 'Edit', 'View', 'Bookmarks', 'Settings', and 'Help'. The terminal shows the command 'R' being executed. The output is the R version 3.2.0 (2015-04-16) splash screen: 'R version 3.2.0 (2015-04-16) -- "Full of Ingredients" Copyright (C) 2015 The R Foundation for Statistical Computing Platform: x86_64-redhat-linux-gnu (64-bit) R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details. Natural language support but running in an English locale R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications. Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R. [Previously saved workspace restored]' The prompt 'beccuti:R' is visible at the bottom.

from Linux

R under GUI using Rstudio

RStudio allows the user to run R in a more user friendly environment.

It is open-source and available at <http://www.rstudio.com/>

The screenshot shows the RStudio interface with several panels and annotations:

- Console:** Contains R version information, license details, and a code snippet: `> m = matrix(1:20, ncol=5)`. A green text box below it states: "Console is where you can type commands and see output".
- Environment/History:** Shows the current workspace with an object 'm' of type 'int' with dimensions [1:4, 1:5]. A green text box above it lists: "1. Workspace tab shows all the active objects" and "2. History tab shows a list of commands recently used".
- Files:** Shows a file browser for the 'Home' directory with files 'ClientServer.eps' (3.9 KB) and 'ProbNetOnlySecurity.eps' (9.6 KB). A green text box below it lists: "1. Files tab shows all the files and folders in your workspace.", "2. Plots tab will show all your graphs.", "3. Packages tab will list a series of packages or add", and "4. Help tab can be used for additional info".

Starting R

R can be started:

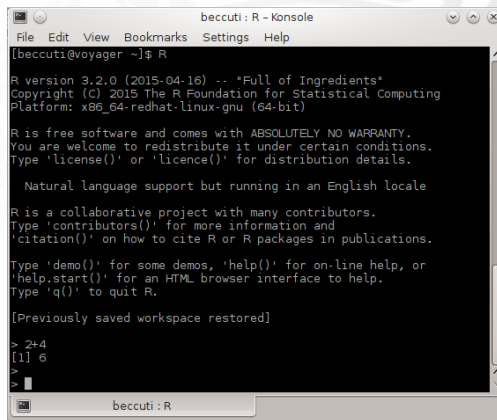
- by double-clicking on the R icon (e.g. Window);
- by double-clicking on the Rstudio icon (e.g. Window + Rstudio);
- by typing `R` in a shell (e.g. Linux).
- by typing `rstudio` in a shell (e.g. Linux + Rstudio).

How R works:

- R creates its objects in memory and saves them in a file called `.RData` (by default);
- Commands are recorded in an `.Rhistory` file, Command can be recalled using up- and down-arrow;
- Recalled commands may be edited;
- Commands may be abandoned by pressing `<Esc>`;
- To end your session type `q()` or just kill the window.
- A concept of *working directory* is introduced: each project is associated with a working folder containing each data.

Interactive R

- R defaults to an interactive mode;
- A prompt ">" is presented to users;
- Each input command is evaluated and a result returned;
- Commands
 - ▶ consist of expressions or assignments;
 - ▶ are separated by a semi-colon (;) or by a newline
 - ▶ can be grouped together using curly brackets({ and })



```
beccuti: R - Konsole
File Edit View Bookmarks Settings Help
[beccuti@voyager ~]$ R

R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

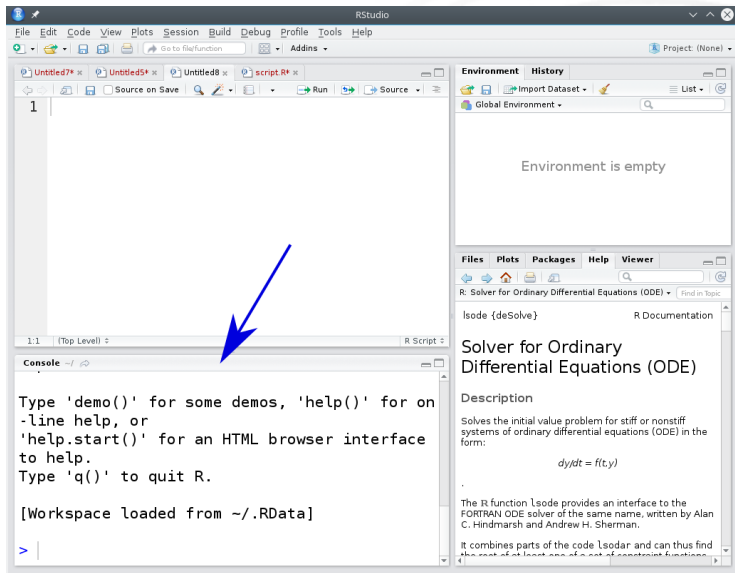
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> 2+4
[1] 6
>
>
```

RStudio prompt and script



The screenshot displays the RStudio interface with the following components:

- Script Editor:** A single line of code is visible: `1`. A blue arrow points from this line down to the console.
- Environment Pane:** Shows "Global Environment" and "Environment is empty".
- Console:** Contains the following text:

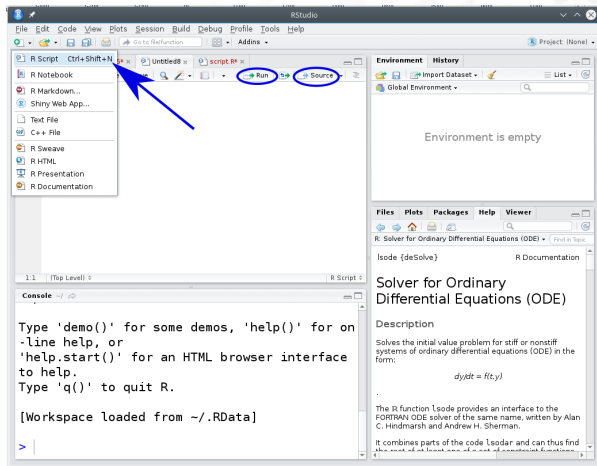
```
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
[Workspace loaded from ~/.RData]  
> |
```
- Help Pane:** Shows documentation for the `lsode` function:

```
R Documentation  
  
lsode {deSolve} Solver for Ordinary Differential Equations (ODE)  
  
Description  
Solves the initial value problem for stiff or nonstiff systems of ordinary differential equations (ODE) in the form:  

$$dy/dt = f(t,y)$$
  
  
The R function lsode provides an interface to the FORTRAN ODE solver of the same name, written by Alan C. Hindmarsh and Andrew H. Sherman.  
  
It combines parts of the code lsodar and can thus find the root of at least one of a set of constraint functions.
```

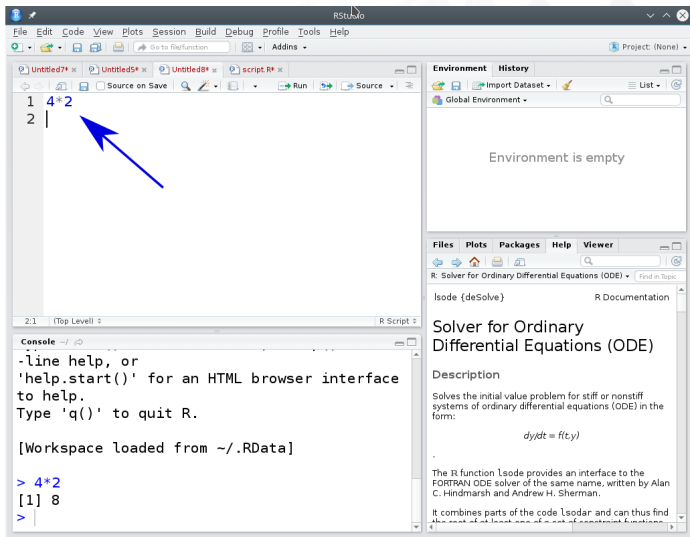
RStudio prompt and script

- R script can be used to save R commands into a file;
- Commands into R script can be executed line by line (clicking on Run) or globally (clicking on Source).



RStudio prompt and script

- Commands can be directly typed into the R script console.



The screenshot displays the RStudio environment. The main editor window shows a script with two lines of code: `1 4*2` and `2 |`. A blue arrow points to the cursor on the second line. The console at the bottom shows the output of the first line: `> 4*2` followed by `[1] 8`. The environment pane on the right indicates that the environment is empty. The bottom pane shows the help documentation for the `lsode` function, including its description and the differential equation $dy/dt = f(t,y)$.

R as a calculator

Simple Arithmetic

```
> 3 + 4  
[1]7
```

Operator precedence

```
> 2 + 3 * 5  
[1]17
```

Exponentiation

```
> 3^5  
[1]243
```

Basic mathematical functions

```
> exp(4)  
[1]54.59815  
> sqrt(4)  
[1]2
```

Predefined constant

```
> pi  
[1]3.141593  
> Inf  
[1]Inf
```


Assignments in R

It is often required to store intermediate results so that they do not need to be re-typed over and over again. To assign a value of 324 to the variable X type:

```
> X <- 324
```

or

```
> X = 324
```

Variable X can be used in next expressions:

Example

```
> X  
[1]324
```

```
> X + X  
[1]648
```

```
> sqrt(X)  
[1]18
```

```
> X = X + X; X  
[1]648
```

```
> X/4  
[1]162
```

```
> X^sqrt(X)  
[1]1.54814e + 45
```

Variable name in R

R is a case-sensitive language, hence `x` and `X` do not refer to the same variable.

Variable name:

- can be created using letters, digits and the `.` (dot) symbol;

```
> data1.address
```

```
> d1_4.f
```

- must not start with a **digit** or a `.` followed by a digit.
- some names are reserved by the system: *if, while, NULL, TRUE ...*

Variable type in R

Basic variable types are:

Numeric: integer, floating point values;

Boolean: values corresponding to **True** or **False**;

Strings: sequences of characters.

Type is determined automatically when variable is created with `<` `-` or `=` operator.

Data structures/Objects are: R provides types of different object.

Vector: a collection of elements (numbers, logical values and character strings) with same type;

Array: a generalization of a vector;

List: collections of objects of any type;
e.g. list of vectors, list of matrices, etc.

Data Frame an array in which the type of each element can be different;

Factor takes on a limited number of values;

Variable in R

- During an R session, objects are created and stored by name;
- The command `ls()` displays all currently-stored objects (workspace);
- Objects can be removed using `rm(variable_name)`;
- All the objects in the workspace are removed using `rm(list=ls())`.

Observe

At the end of each R session, you are prompted to save your workspace. If you click Yes, all objects are written to the `.RData` file. When R is re-started, it reloads the workspace from this file and the command history stored in `.Rhistry` is also reloaded.

Variable in RStudio

The screenshot shows the RStudio interface with the following components:

- Environment pane:** Shows the variable `y` with type `int [1:2, 1:5]` and values `1 2 3 4 5 6 7 8 9 10`. A red circle highlights the variable name `y`.
- Viewer pane:** Displays a table with 2 rows and 5 columns (V1-V5). A red arrow points from the variable `y` in the Environment pane to this table.
- Console pane:** Shows the R session output, including help text and the commands used to create the variable `y`.

	V1	V2	V3	V4	V5
1	1	3	5	7	9
2	2	4	6	8	10

```
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

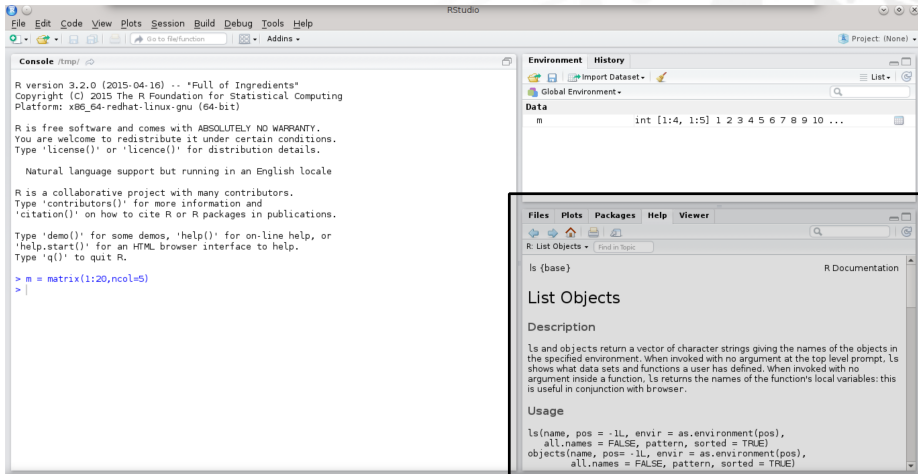
> x = c(1,2,3,4,5,6,7,8,9,10)
> y = matrix(1:10,ncol=5)
> View(y)
>
```

Getting help in R

R provides a built-in help facility.

- To get more information on any specific function, e.g. `sqrt()`, the command is:
`help(sqrt)`
or
`?sqrt`
- help on features specified by special characters must enclose in single or double quotes (e.g. `"["`) `help("[")`
- Help is also available in HTML format by running `help.start()`
- For more information use
`?help`

Getting help in Rstudio



The screenshot shows the RStudio interface. The console on the left contains the following text:

```
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> m = matrix(1:20,ncol=5)
> |
```

The Environment pane on the right shows a variable `m` of type `int` with dimensions `[1:4, 1:5]` and values `1 2 3 4 5 6 7 8 9 10 ...`.

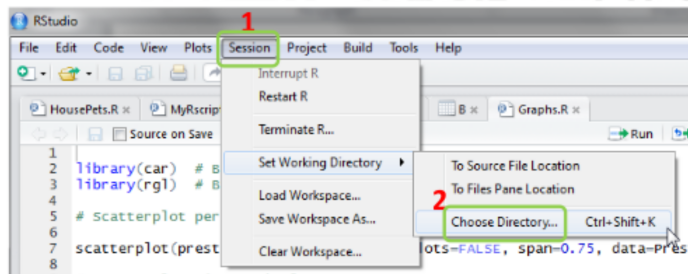
The Help pane on the right shows the documentation for the `ls` function. The title is "List Objects". The description states: "ls and objects return a vector of character strings giving the names of the objects in the specified environment. When invoked with no argument at the top level prompt, ls shows what data sets and functions a user has defined. When invoked with no argument inside a function, ls returns the names of the function's local variables: this is useful in conjunction with browser." The usage section shows the following code: `ls(name, pos = -1L, envir = as.environment(pos), all.names = FALSE, pattern, sorted = TRUE)` and `objects(name, pos = -1L, envir = as.environment(pos), all.names = FALSE, pattern, sorted = TRUE)`.

Working directory

Working directory in R:

- Working directory contains data and R scripts. It is a directory of the file-system;
- `getwd()` returns the current Working directory;
- `setwd("new_path")` sets Working directory;

Working directory in RStudio:



Packages in R

- R provides libraries of packages. Packages contain various functions and data sets for numerous purposes;
- Some packages are part of the basic installation. Others can be downloaded from CRAN:
> *install.packages("ggplot2")*
- To use functions and data sets of a package, it must be loaded into the workspace:
> *library(ggplot2)*
- To check what packages are currently loaded into the workspace:
> *search()*
- A loaded package can be removed:
> *detach("package:ggplot2")*

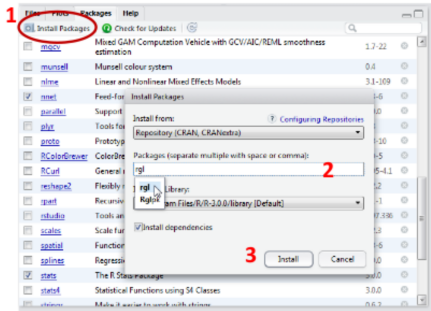
Observe:

if you terminated your session and start a new session with the saved workspace, you must load the packages again.

Packages in Rstudio

<input type="checkbox"/>	RCurl	General network (HTTP/FTP/...) client interface for R	1.95-4.1	⊗
<input type="checkbox"/>	reshape2	Flexibly reshape data: a reboot of the reshape package.	1.2.2	⊗
<input type="checkbox"/>	rpart	Recursive Partitioning	4.1-1	⊗

Before



We focus on Package tab(bottom-right)

After

<input type="checkbox"/>	RCurl	General network (HTTP/FTP/...) client interface for R	1.95-4.1	⊗
<input type="checkbox"/>	reshape2	Flexibly reshape data: a reboot of the reshape package.	1.2.2	⊗
<input type="checkbox"/>	rgl	3D visualization device system (OpenGL)	0.93.952	⊗
<input type="checkbox"/>	rpart	Recursive Partitioning	4.1-1	1.2 ⊗