# Data Analysis Using R: Chapter01

罗智超 *(ROKIA.ORG)*

## Contents

### 通过本章你将学会

- 配置你的工作环境
- 安装常用的 R 包
- 掌握 RSTUDIO 的基本功能
- 掌握 git 代码管理功能

## 安装 R

- 安装 R 环境

– www.r-project.org

- 安装 R 应用包

– 选择 CRAN 镜像（为厦大而骄傲，居然为中国高校中为数不多的几个镜像之一）

– github

要想在 CRAN 上面发布一个包难度类似发表一篇一类核心的文章，因此，Hadley 开发了一个 devtools 包），这样任何人都可以将自己开发的包上传到 github 上面，供别人下载安装

```
# 安装 rticles 包
install.packages("devtools")

library(devtools)
devtools::install_github("rstudio/rticles")
```

– R-forge

```
install.packages("quantmod", repos = "http://R-Forge.R-project.org")
install.packages("TTR", repos = "http://R-Forge.R-project.org")
install.packages("FinancialInstrument", repos = "http://R-Forge.R-project.org")
install.packages("blotter", repos = "http://R-Forge.R-project.org")
install.packages("quantstrat", repos = "http://R-Forge.R-project.org")

install.packages("PerformanceAnalytics", dependencies=TRUE)
install.packages("xts", dependencies=TRUE)
```

– Bioconductor.org

## 安装 RSTUDIO

- 操作演示

## 安装 GIT

- 操作演示

### Test Github config

## 配置 RSTUDIO+GIT+GITHUB

- 详见配置指南

( 1 ) 注册 GIT 账号

( 2 ) 创建一个 repository：DataAnalysis

( 3 ) 下载 git、rstudio 并安装

( 4 ) 在 rstudio-tools-global options-Git/Svn 里面设置 git.exe 的路径，重启 rstudio

( 5 ) 打开 Git 终端配置用户账户信息 (注意区分大小写)

下面三句分别配置用户名、邮件地址以及创建公钥

git config –global user.name “zhichaoluo”

git config –global user.email ”zhichao.luo@gmail.com”

ssh-keygen -t rsa -C zhichao.luo@gmail.com

( 6 ) 登陆 github.com，在 Personal settings-SSH keys-Add SHH key，将 ( 6 ) 第三句中创建的 key 的内容 copy 进去。

( 7 ) 在 rstudio 中新建 project from version control-Git 配置第 ( 2 ) 步中创建的 repository 的地址

git@github.com:zhichaoluo/DataAnalysis.git

有两种传输协议模式 https 和 SSH 模式，由于我们在第 ( 5 )( 6 ) 步骤创建了 SHH key 所以，我们可以选择这个模式。如果选择 https 模式，每次提交更新都要提示输入用户名密码，非常麻烦。

有关于 Git 的详细介绍可以参考下文廖雪峰

## 安装 CTEX

- 操作演示
- ctex.org
- mactex

## RSTUDIO 功能介绍

- 参数配置 ( 全局、项目 )
- 新建项目
- 文艺编程 (case_Reproducible Report.rmd)
- RMARKDOWN
- 操作演示

## RSTUODIO+GITHUB 代码管理

- 操作演示

### 关于 R 工作空间管理的一些基础函数

```
# 注意：R 是区分大小写，R 里面的目录要用反斜杠/或者\\
getwd()
setwd("D:\\RPROJECT\\DataAnalysis\\data")
ls()
rm()
options(digits=3)
save.image("filename")
```

### 常用的 R 包

```
google+top 100 r packages
dplyr
ggplot2
lubridate
stringr
reshape2
RColorBrewer
zoo
xts
scales
car
knitr
rmarkdown
devtools
rticles

RODBC
RJDBC
RSQLite
sessionInfo()
```

### 如何查看 R 包源代码

- 简单的函数（非类函数），直接在 R 里面输入函数名就可以查看源代码，注意函数名后面不要加 () 在命令行输入：help 和 help() 的结果不一样，前者显示 help 函数的源代码，后者显示 help() 的帮助文档
- 对于类函数，直接输入函数名不能显示出源代码，例如：

summary

function (object, ...) UseMethod("summary")

这时候需要用到 methods() 函数，用法 methods(FunctionName) 如下：

> methods(summary) [1] summary.aov summary.aovlist summary.aspell *[4] summary.connection summary.data.frame summary.Date [7] summary.default summary.ecdf* summary.factor ……

Non-visible functions are asterisked 加星号标注不能直接输入函数名来看代码，因为它不在默认命名空间中。但是可以通过 getAnywhere() getS3method() 来查看。

找到这个类函数里面你所关注的函数，输入函数名，回车，就可以查看代码了，如：

> summary.data.frame

对于非类函数使用 methods 会报出错误：

> methods("sample") [1] sample.int Warning message: In methods("sample") : function 'sample' appears not to be generic

对于具体的函数，要搞懂它，可能看这些信息还不够，需要下载 *.tar.gz，查看里面的源代码。这时候 linux 下的 find 命令就非常有用，具体可以问问谷哥和度娘。

- 直接上 CRAN 下载源代码包。对于加星号标注的是不可见的方法流程如下：

(1) 登入 R 主页 http://www.r-project.org/ ，点击 Download 下的 CRAN

(2) 选择一个镜像

(3) 里面的 Source Code for all Platforms 就可以下载各种源码了，下面以下载程序包源码包为例，点 packages

(4) 选择 sorted 的方式，推荐 by name

(5) 找到你感兴趣的包, 比如 abind, 点进去就可以看见 Package source 这一项, 用 tar.gz 封装的, download 就可以了，解压后就能看见源码了。一般源码都在 R 目录里面。

## 本周“大牛”

- K. Pearson 1879 年毕业于剑桥大学数学系；曾参与激进的政治活动。出版几本文学作品，并且作了三年的律师实习。1884 年进入伦敦大学学院 (University College, London)，教授数学与力学，从此待在该校一直到 1933 年。
- K. Pearson 最重要的学术成就，是为现代统计学打下基础。自从达尔文演化论问世后，关于演化的本质争论不断，在这方面他深受 Galton ( 达尔文表哥,「 优生学 」一词的发明者 ) 与 Weldon 影响。Weldon 1893 年提出「 所谓变异，遗传与天择事实上只是『 算术 』」的想法。这促使 K. Pearson 在 1893-1912 年间写出 18 篇〈在演化论上的数学贡献〉的文章，而这门「 算术 」也就是今日的统计。许多熟悉的统计名词如标准差，成分分析，卡方检定都是他提出的。
- K. Pearson、Galton 与 Weldon 为了推广统计在生物上的应用，于 1901 年创立统计的元老期刊《Biometrika》，由 K. Pearson 主编至死，但是 K. Pearson 的主观强，经常对他本人认为有「 争议 」的文章，删改或退稿，并因此与英国本世纪最有才华的统计学家 Fisher 结下梁子。
- 1906 年 Weldon 死后，K. Pearson 不再注意生物问题，而专心致志于将统计发展成一门精确的科学。