Data Analysis Using R: Chapter04

罗智超 (ROKIA.ORG)　　1814347@qq.com

1　通过本章你将学会

- 数据导入

  –文本文件

  –EXCEL 文件

  –其他统计软件

  –批量导入数据

  –数据库

- 数据输出
- 网络爬虫

2　从剪切板读取

data<-read.delim(“clipboard”)

3　从键盘读入

- scan(), readline(), print(), and cat()

4 链接方法 file(), url() 等

```
uci <- "http://archive.ics.uci.edu/ml/machine-learning-databases/"
uci <- paste(uci,"echocardiogram/echocardiogram.data",sep="")
ecc <- read.csv(uci)
```

5 文件及目录相关函数

- file.info() 获取文件信息
- list.dirs()、dir()、list.files()、file.info(“.”): 返回目录里面的文件信息
- file.exists(): 判断是否存在某文件
- dir.create(“newfolder”) 创建目录
- dir.create(path=“a1/b2/c3”,recursive = TRUE) 创建多级目录
- file.rename(“tmp”, “tmp2”) 目录重命名
- unlink(“tmp2”, recursive = TRUE) 删除目录
- file.create(“A.txt”) 创建一个空文件
- file.append(“A.txt”, rep(“B.txt”, 10)) 合并文件
- readLines(“A.txt”) 查看文件内容
- getwd() 获取当前工作目录
- setwd() 设定当前工作目录

6 读入文本文件

- read.table
- read.csv
- read.delim

- read.fwf
- read.table 详细说明

7 读入固定宽度文件

```
mydata<-read.fwf("data.txt",widths=c(1,4,3))
```

8 redLines(),scan()

大部分情况下，用 *read.table* 函数可以将文本文件读入 *R*，但有时也有无法使用的时候，如文件中的观察可能是多行的，这时就要使用 *readLines()* 可以用 readLines 交互式的输入数据 *scan() 可以读入更复杂的文件格式

9 读入数据练习

- 将世界城市列表导入到 R

10 读入 EXCLE 文件

- 远离 EXCEL!!!

```
# Support 64bit system
install.packages("XLConnect")
library("XLConnect")
df = readWorksheetFromFile("data.xls",
                           sheet=1, header=TRUE)
```

```
library(xlsx)#very slow
df<-read.xlsx("excelfile.xlsx",
              sheetIndex=1,header=TRUE,
              colIndex=,rowIndex=)
```

```
install.packages("RODBC")
library(RODBC)
channel <- odbcConnectExcel("myfile.xls")
mydataframe <- sqlFetch(channel, "mysheet")
odbcClose(channel)
```

11 通过 ODBC 访问

```
# Only support 32-bit system.
library(RODBC)
myconn <-odbcConnect("mydsn", uid="user", pwd="password")
crimedat <- sqlFetch(myconn, Crime)
pundat <- sqlQuery(myconn, "select * from Punishment")
close(myconn)
```

12 访问 ORACLE

- RJDBC 配置说明

13 读入比较大的数据文件

- Use data.table library fread()
- 使用 read.table 时明确 colClasses 和 nrows，设置 comment.char=“”

14 读取其他统计软件数据

- library foreign
- library haven **New**
  - 支持 SAS SPSS Stata

15 访问 ORACLE

- Using RORACLE package

16 访问 Sqlite

- Using RSQLite package

```
library("RSQLite")
drv <- dbDriver("SQLite")
con <- dbConnect(drv, dbname = "d:/mydb.s3db")
db_u<-dbGetQuery(con, "select * from table1" )
dbDisconnect()
```

17 批量读入外部文件

- 方法一：保存成独立文件
- 方法二：保存成 list

```
# 方法一

fileName <- dir("D:/tempdata/csv")
scode<-substr(fileName,1,6)
N=length(fileName)

for(i in 1:N){
assign(paste("s",scode[i], sep=""), read.csv(fileName[i],header=TRUE))
}
```

```
# 方法二

fileName <- dir("D:/tempdata/csv")
cls <- c("character","character","character","numeric","numeric","numeric","numeric")
stocklist<- lapply(fileName,function(x) read.csv(x,header=TRUE,colClasses=cls,stringsAs
allstcok<- do.call(rbind,stocklist)
```

18 输出数据集

```
write.table(dataframe, file = "output.csv",
            sep = ",", col.names = NA,append=TRUE)
save(df,file="data/df.RData")
load("data/df.RData")
```

19 网络爬虫

- 天气数据爬虫程序
- 爬取Wikipedia article traffic statistics数据

20 本周“大牛”

- Hadley Wickham 是 RStudio 的首席科学家以及 Rice University 统计系的助理教授。他是著名图形可视化软件包 ggplot2 的开发者，以及其他许多被广泛使用的软件包的作者，代表作品如 plyr、reshape2 等。
- 统计之都对他的采访