Reading and saving data in Rstudio

Marianne Huebner, Michigan State University

2015-05-01

## Reading data into Rstudio from the menu

The upper right window in Rstudio has a link "Import Dataset" under the Environment tab. If your data file is a textfile, it will ask you to click on the dataset in your folder. If you dataset is from the web, it will ask you to copy the URL.

In either case a window opens showing the input file and there are several choices: *Heading yes, no (if the columns have names)* Separator: Whitespace, Comma, Semicolon, Tab  
*Decimal: Period, Comma (different countries use different decimal marks)* Quote: Double quote, Single quote, None \*na.strings: NA (sometimes missing data are indicated with NA or with text or with numbers)

For example, the first two lines in the input file for cereal data (<http://stt.msu.edu/Academics/ClassPages/uploads/FS14/200-1/Ch08_Cereals.csv>) are

name, calories, sugar  
100%\_Bran, 70, 6

This indicates that there is a heading (variable names: name, calories, sugar) The separator between columns is a Comma. The default decimal mark is Period. There are no quotes or space. There are no missing data.

Another example is:

"Waist" "Weight" "BodyFat"  
"1" 32 175 6  
"2" 36 181 21

Here the separator is Whitespace, the quote is Double quote, and there is a heading. The number "1" and "2" indicate row numbers.

## Reading internet data into R

Suppose you have a larger data list in a file. You can read it from the folder where you save the Session. On the STT webpage there are some data saved that you can access with a URL. They can be read into R directly with the URL as follows. The R command "head" will show you the top few lines, the command "tail" what show you the last few lines

body=read.table("http://stt.msu.edu/Academics/ClassPages/uploads/FS14/200-1/Ch08\_Body\_fatv2.txt", header=T, sep="")  
  
head(body)

## Waist Weight BodyFat  
## 1 32 175 6  
## 2 36 181 21  
## 3 38 200 15  
## 4 33 159 6  
## 5 39 196 22  
## 6 40 192 31

tail(body)

## Waist Weight BodyFat  
## 15 44 246 38  
## 16 33 160 10  
## 17 41 215 27  
## 18 34 159 12  
## 19 34 146 10  
## 20 44 219 28

In the line that reads in the data the "header=TRUE" or "header=T" says that there are variable names in the data file. We need to say how the columns are separated, in this case with a space, sep="" or with a tab sep="".

Have a look at the upper right window under the "Environment" tab. The dataset "body" is shown. If you click on "body" it will give you a quick overview what is in the data set. In this case there are three variables, all integers (=discrete numerical data) and the first few observations of each variable.

The data set "body" has three variables (stored in columns) and 20 observations each. You can access the first column, Waist measurements, as follows

# body[row number, column number]  
# if the row number is left empty, all rows are shown, anmely all 20 Waist measurments  
body[,1]

## [1] 32 36 38 33 39 40 41 35 38 38 33 40 36 32 44 33 41 34 34 44

# or using the variable name  
body$Waist

## [1] 32 36 38 33 39 40 41 35 38 38 33 40 36 32 44 33 41 34 34 44

To see all measurements of the 4th individual:

body[4,]

## Waist Weight BodyFat  
## 4 33 159 6

## Saving data to your folder

You can save the data to your session folder in three ways:

save(body, file="mydata.Rdata")  
write.csv(body, file="mydata.csv")  
write.table(body, file="mydata.txt")

Check your folder. If you click on "mydata.Rdata", this will open R window and load the data immediately (but not Rstudio). To open these data in Rstudio, click on File > Open File> mydata.Rdata. The other two files are opened with Excel or a text editor, respectively.

## Reading data from text files into R

Suppose you have a dataset in a file (in your Session folder!).

# Read the bodyfat data that you just saved.  
# If there are variable names, add "header=TRUE"  
# Indicate the type of separator (commas "'", space "", tab "\t" etc)  
body<- read.table("mydata.txt", header=TRUE, sep="")

Sometimes variable names are complicated, such as "Waist (in)". In this case Rstudio thinks the first line has more columns while the next line has only three columns. You can open the text file, simplify the variable name so that it is one word, and save the data with a new filename. If you make changes to any dataset, leave the original and save the changes in another file.

The structure in R called *data frame* is a collection of all variables. To extract values for the variable "Weight" we use the name of the dataset and the name of the Variable after a $ sign. Let's see the first three weight measurements from this dataset:

body$Weight[1:3]

## [1] 175 181 200

## Reading Excel files into R

To read an excel file instead of a textfile, you can use "read.csv", if it is a ".csv" file" or type the command install.packages("xlsx") and then do, for example, the following.

library(xlsx)  
bodyfat<-read.xlsx("Bodyfat.xlsx", sheetName = "Sheet1")