Data Entry Using R and RStudio

JuJu

Data may be entered into R in a number of ways. Three commonly used methods will be discussed.

## Manual Entry

Perhaps the easiest way to enter small datasets is to enter each variable individually and then combine them into a **data frame**. Using the data from BPS7e problem 29.48, this might look like:

sex = c(rep("Female",12),rep("Male",7))  
 mass = c(36.1, 54.6, 48.5, 42.0, 50.6, 42.0, 40.3, 33.1, 42.4,  
 34.5, 51.1, 41.2, 51.9, 46.9, 62, 62.9, 47.4, 48.7, 51.9)  
 rate = c(995, 1425, 1396, 1418, 1502, 1256, 1189, 913, 1124, 1052,   
 1347, 1204, 1867, 1439, 1792, 1666, 1362, 1614, 1460)  
 gender = c(rep(1,12),rep(2,7))  
 bps7.29.48 = data.frame(sex, mass, rate, gender)

We can now check to see if the data frame has been created by entering:

ls()

## [1] "bps7.29.48" "gender" "mass" "rate" "sex"

Note that the listing also shows the individual variables that were used to create the data frame. These can be deleted by using ***rm()***.

rm("sex", "mass", "rate", "gender")  
 ls()

## [1] "bps7.29.48"

The attributes of the data frame and some summary statistics can be computed using the ***attributes*** and ***summary*** functions.

attributes(bps7.29.48)

## $names  
## [1] "sex" "mass" "rate" "gender"  
##   
## $class  
## [1] "data.frame"  
##   
## $row.names  
## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

summary(bps7.29.48)

## sex mass rate gender   
## Length:19 Min. :33.10 Min. : 913 Min. :1.000   
## Class :character 1st Qu.:41.60 1st Qu.:1196 1st Qu.:1.000   
## Mode :character Median :47.40 Median :1396 Median :1.000   
## Mean :46.74 Mean :1370 Mean :1.368   
## 3rd Qu.:51.50 3rd Qu.:1481 3rd Qu.:2.000   
## Max. :62.90 Max. :1867 Max. :2.000

Notice that while ***sex*** was treated as a categorical variable, ***gender*** was treated as if it was cardinal. R is smart in that it recognizes the difference between cardinal and categorical (which it calls ``factor’’) variables. To make ***gender*** a factor variable we can enter

bps7.29.48$gender = factor(bps7.29.48$gender,levels=c(1,2),labels=c("F","M"))

Using ***summary*** we can see that ***gender*** is treated as a factor, or categorical, variable.

summary(bps7.29.48)

## sex mass rate gender  
## Length:19 Min. :33.10 Min. : 913 F:12   
## Class :character 1st Qu.:41.60 1st Qu.:1196 M: 7   
## Mode :character Median :47.40 Median :1396   
## Mean :46.74 Mean :1370   
## 3rd Qu.:51.50 3rd Qu.:1481   
## Max. :62.90 Max. :1867

## Using RStudio

RStudio will read comma and tab delimited text files. The data for BPS7e can be downloaded from <http://bcs.whfreeman.com/bps7e>. Save the PC-Text.ZIP file to your drive. Then, extract to files. Within the folder that is created will find a number of “Chapter” folders. Each of these has a number of files corresponding to examples (eg—–.txt files) and exercises (ex——.txt files).

We can use the GUI interface to import the file. Select **Tools -> Import Dataset -> From Local File** and navigate to the ***ex29-48METAB2.txt*** file. Double click on the file or enter its name in the **File Name** area and click **Open**.

In the new window, change the **Name** to something that is fairly easy to type. In this case we can use **bps7.29.48** to represent the 48th exercise from the 29th chapter of BPS7e. Be sure that **Heading** is set to **Yes**. Note that the file is tab delimited and not comma separated. The rest of the default values are probably okay. Click on **Import**.

Rstudio will now submit ***view(bps7.29.48)*** to the console so that you can check to see if the file was properly imported.

## Reading Comma Separated Value (CSV) Files

R has a utility for reading comma separated value (CSV) ASCII files. These files can reside on the host machine or on a server. If the files are in standard CSV format, either of:

# To make the next line work you will have to change the path  
 HtWt = read.csv("htwt.csv")  
 summary(HtWt)

## Height Weight Group   
## Min. :51.0 Min. : 82.0 Min. :1.00   
## 1st Qu.:56.0 1st Qu.:108.2 1st Qu.:1.00   
## Median :59.5 Median :123.5 Median :2.00   
## Mean :62.1 Mean :139.6 Mean :1.55   
## 3rd Qu.:68.0 3rd Qu.:166.8 3rd Qu.:2.00   
## Max. :79.0 Max. :228.0 Max. :2.00

# This reads the file from the given URL  
 htwt = read.csv("http://facweb1.redlands.edu/fac/jim\_bentley/downloads/math111/htwt.csv")  
 summary(htwt)

## Height Weight Group   
## Min. :51.0 Min. : 82.0 Min. :1.00   
## 1st Qu.:56.0 1st Qu.:108.2 1st Qu.:1.00   
## Median :59.5 Median :123.5 Median :2.00   
## Mean :62.1 Mean :139.6 Mean :1.55   
## 3rd Qu.:68.0 3rd Qu.:166.8 3rd Qu.:2.00   
## Max. :79.0 Max. :228.0 Max. :2.00

will create a data frame that contains the ***htwt*** data. Note the use of forward slashes instead of backslashes.

The group variable will be imported as a numeric. To help R function efficiently, it will need to be converted to a factor variable using the method from above.

## Saving and Loading Data Frames

Regardless of how they were created, data frames may be saved in R as part of the R workspace. The workspace contains all of the variables, data frames, and functions that you have defined. A workspace is a snapshot of your work to the point of the save.

In RStudio, to save a workspace click on **Session -> Save Workspace As**. Navigate to the folder in which you wish to save the file and provide a descriptive file name. Now click on **Save**. Your workspace is now safely tucked away on your drive. This file can later be **Load**ed or you can open it by double clicking on the file.

History files store the commands that you used during your R session. These can be saved and loaded in a manner similar to that of workspaces. These files are are text files and can be edited using Wordpad or something similar. RStudio hides these behind a tab in the upper right window.