Coding in R Worksheet

1 Data by Hand

Suppose we have data on the weight of a mountain bike and it's price.

Brand	Weight (lb)	Price ($)
FRX Raod	37	640
TX 120	35	660
RIDER A10	30	820
FRX X60	29	790
TX 480	29	840
TX 1000	28	970

Let's input this data into two variables named brand, weight and price.
brand = c("FRX Road", "TX 120", "RIDER A10", "FRX X60", "TX 480", "TX 1000")
weight = c(37, 35, 30, 29, 29, 28)
price = c(640, 660, 820, 790, 840, 970)

We can do some basic statistics on this data

1. Find the mean and standard deviation of the prices using the command mean(price) and the command sd(price) in the command prompt.
2. Use the summary(price) command for other descriptive statistics.
3. Create a histogram of weight and price using the command hist(weight) and hist(price)
4. Create a scatterplot of the price versus weight using plot(price $\sim$ weight). Describe the relationship. Is the relationship strong? Is it linear? Is it positive or negative?

2 Data Frames

Because these three variables or objects are related, we may want to combine them into one big object called a data frame. Data frames makes a table from vectors of the same length. We invoke the command data.frame(a, b, ...,n) to combine variables a through n together. For our mountain bikes we can combine them as

mbike = data.frame(brand, weight,price)

1. Print out the data frame mbike by typing mbike at the command prompt.

2. To see only one variable you need to state the data frame followed by $ and the name of the variable like this mbike$price . Make a new variable of just the prices.

3. Make the scatter plot of the data again only using the variables from the data frame with a title like this plot(mbike$price $\sim$ mbike$weight, main = "Price vs Weight of Mountain Bikes")

4. Make the same scatter plot with red, closed circles. The command col = creates the color and the command pch = creates the marker style. You can use a number for 0 to 25 to create different styles and fill. plot(mbike$price $\sim$ mbike$weight, main = "Price vs Weight of Mountain Bikes", col = "red", pch = 16)

3 Data From a File

Many times we will import data from a file. These files can come in a variety of formats like text files (.txt) that are space delimited, tab delimited or any character that is a delimiter, comma separated values (.csv), Excel files (.xlsx or .xls), SAS files (.sas7bdat), dbase files (.dat) or any other database system.

1. Download the file lead.txt from SAKAI

2. Create a data frame titled leadpoison using the following syntax

   leadpoison = read.table(file.choose(),header =TRUE)

   A file dialog box should pop up and you can navigate to the file. Click on it and R will import the data into the data frame leadpoison.

3. Find out the first 6 data values for each variable using the command head(leadpoision)

4. Find the summary values of all numerical varaibles using the command summary(leadpoison)

5. Create a scatter plot of the variable iqf vs the variable iqv. Rather than invoke the name of the data frame followed by the dollar sign to pull the data from the data frame leadpoison, use the "data = " statement inside the plot command to call the data frame like this: plot(lqf $\sim$ lqv,data = leadpoison). Does there appear to be a linear relationship between the data?

6. Calculate the linear correlation coefficient between iqf and iqv using cor(leadpoison $ iqf, leadpoison $ iqv). What is the value? Is the relationship linear? Is it weak, moderate or strong?

7. Create a data frame with the output of a linear regression using the command fit.leadpoison = lm(iqf $\sim$ iqv, data = leadpoison). Then call summary(fit.leadpoison). Write the equation of the regression line.

4 File Path method

While the file.choose method is the easiest to open a file, it has it's drawbacks from a technical stand point. Let's open a comma separated file by pointing to the exact path

1. Download the file PUPILCOSTATLANTA.csv from SAKAI to your download folder. This is csv (comma separated variable) file of 44 all white schools in Altanta around 1938 to determine if the per pupil cost is related to school size and other factors. The variables are school name, per-pupil-cost, average daily attendance, average monthly teacher salary, percent attendance, pupil/teacher ratio.

2. Create a data frame titled cost n using the following syntax

   cost = read.table(file="C:/USER/Download/PUPILCOSTATLANTA.csv", sep = ",", header = TRUE)

   You will need to find the proper path, don't copy what I have. Note, the slashes are forward slashes and not the usual back slashes.

3. Find the name of the variables using the head(cost) command.

4. Create a scatter plot of per-pupil-cost vs pupil/teacher ratio using the names of the variables in the data set. Note, R is case sensitive and these variables are in all capital letters. This is important.

5. Find the correlation coefficient of per-pupil-cost and pupil/teacher ratio and interpret.

6. Create the linear regression of per-pupil-cost vs pupil/teacher ratio and write the equation.

7. Create the scatter plot again using the same code you did in part 4. Underneath this code type abline(fit.cost)

   where fit.cost is the name of the linear regression. The command abline creates a line with intercept a and a slope b. You can also type abline(157.21, -2.90)

5 Other files

If the text file is delimited by other characters than a space, you can use the following syntax

a = read.table("C:/file",header = TRUE, sep = " ")

Comma Delimited

a = read.table("C:/file",header = TRUE, sep = ", ")

Tab Delimited

a = read.table(”C:/file”,header = TRUE, sep = ” \t ”)

Comma Separted Files (.csv) a = read.csv(”C:/file”,header = TRUE)