CWRU DSCI351-351M-451: HW1

Prof.:Roger French, TA:JiQi Liu, Student:Anish Mitra

August 30, 2017, Due Tuesday September 4th, before class

Contents

1.1 So by now I believe everyone has

1.1.1 Logged into your ODS VDI, or your VUVlab VDI.

```
- If not send the help@case.edu helpdesk an email,
  - directed to CSE-IT, saying you are in DSCI351/351M/451
  - and should have access to the ODS VDI,
  - that you use Citrix REceiver to connect to.
```

1.1.2 Your H: drive is big enough

- so that you can Git clone your personal fork of the Prof repo,
- from Bitbucket down to `H:\Git\` folder.

1.1.3 You'll notice in HW1

- That there is an initial read Leek's structure of a data analysis.
- This all students should do.

1.1.4 Ask Questions in CWRU-DSCI Slack Channel for DSCI351-351M-451

- This is the easier way to ask and answer questions
- You can use @JiQi Liu and @Roger French
  - To direct a question to us
  - But anyone can answer the questions

1.2 351, 351M and 451 students

- will both do the last part of the homework,
  - where you are doing some R coding,
  - inside the R code blocks of the Rmd file
  - (between the "‘*r* and the "‘ that closes the R code block in the Rmd file.

1.3 And 451 students

- will start writing about what they are considering for their Semester Project.
  - Read about the 451 Semester Project in 1-assignments>SemProj-451>1808-451-SemProj-Overview.pdf
- Your SemProj will have 3 in-class report outs on progress, and a final full report.
- It should ideally be related to your thesis research,
  - and be a data analysis project that will help advance your research.

- We will be defining and refining what you will do your Semester Project on,
  - in the next few weeks.

1.4 Here are answers to a few questions we usually get about HW1

A. "I am having trouble converting the Rmd file containing homework 1 into a PDF. I was able to save it as a .txt file–is it okay if I submit that instead of a PDF?"

Once you have made a *.Rmd file, you compile it to make the pdf, by hitting the Knit button at the top of the Rstudio text editor, or you can click on the Knit button to choose Knit to PDF from the choices. You can also use the keyboard shortcut Cntrl+Shift+K. (You can find lots of keyboard shortcuts help with Alt+Shift+K).

And if you open the homework 1 Rmd file named "1808-351-351M-451-hw1-NAME.Rmd" and change NAME to your own name. Then you can immediately compile that Rmd file to make the pdf. This way you'll know that its not some error in what you have added to the file's text. Compiling to pdf, uses the LaTeX publishing distribution on your VDI. So if you are trying this on your own personal computer, it won't work, since you probably don't have a LaTeX distribution, such as MikTeX (for windows), MacTeX (for Macs), or TexLive for Linux, installed, so can't produce a LaTeX pdf output.

B. "I was also wondering where we are supposed to submit our homework assignments. Are we supposed to upload them onto BitBucket?"

You will upload your *.Rmd file (so we can see your coding style and commenting), and your compiled Pdf file to our Canvas Assignment page in canvas.case.edu for the DSCI351-351M-451 class.

1.5 Read the LEEK handout

- in ./readings/1503LeekDataAnalyticStyle-outline.txt

- This is located in ./class/Leek/Leek-ADataAnalysisStructureAndOrganizing.pdf
- and look at Leek's book in ./readings/Texts/Leek-DataAnalyticStyle.pdf

1.5.1 On organizing a data analysis

- Jeff Leek is a biostatistician
- At Johns Hopkins School of Public Health

1.5.2 Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

1.6 DSCI 451 HW1 Assignment

- Make an separate Rmarkdown file (*.Rmd) discussing your Semester Project
- Which you compile to produce a pdf report
- About a data analysis question of interest to you
  - Use figures, sections, math equations and hyperlinks
  - Author, License, Version number, Changelog
- Submit the .Rmd file and the pdf on Canvas

1.6.1 For DSCI451 Grad Students

- This Rmd will be an initial proposal
  - Of your semester data science project
- Explain background of questin
- Experiments or approaches
- Data types and characterisitics.
- Use figures, sections, math equations and hyperlinks

1.7 DSCI 351/351M/451

- Enter all work into this rmd file
- Make sure it compiles and submit both the rmd and pdf
- Rmd documents run code in block segments
  - defined and closed by "‘ with an option to provide parameters

```
print('hello world')
```

```
## [1] "hello world"
```

- Answer all questions in below in the provided code blocks

1.7.1 Basic R operations

- Show an example of addition, subtraction, multiplication, division, and an exponential below

```
#Example of addition
print("Addition")
```

```
## [1] "Addition"
```

```
x <- 1 + 1
x
```

```
## [1] 2
```

```
#Example of subtraction
print("Subtraction")
```

```
## [1] "Subtraction"
```

```
y <- 2 - 2
y
```

```
## [1] 0
```

```
#Example of multiplication
print("Multiplication")
```

```
## [1] "Multiplication"
```

```
z <- 2 * 2
z
```

```
## [1] 4
```

```
#Example of division
print("Division")
```

```
## [1] "Division"
```

```
a <- 2 / 2
a
```

```
## [1] 1
```

```
#Example of an exponential
print("Exponential")
```

```
## [1] "Exponential"
```

```
b = 2^3
b
```

```
## [1] 8
```

1.7.2 Working with data frames

- Data frames are an important data format in R
- Example data can be loaded from base R
- Run the code below to load the iris dataset into your environment

- This data set will be used for the later problems

```
data("iris")
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

- Give the class of each of the columns in the iris data set
- Explain in code comments what is a factor and how it differs from a character

```
x <- lapply(iris, class)
x
```

```
## $Sepal.Length
## [1] "numeric"
##
## $Sepal.Width
## [1] "numeric"
##
## $Petal.Length
## [1] "numeric"
##
## $Petal.Width
## [1] "numeric"
##
## $Species
## [1] "factor"
```

```
#The difference is that a factor is a categorical variable whereas a factor #variable is not.This basic
```

- Use the table() function to determine how many species there are
  - and how many observation each one has (Species column in the data frame)

```
class(iris)
```

```
## [1] "data.frame"
```

```
table(iris[5])
```

```
##
##     setosa versicolor  virginica
##         50         50         50
```

- Use the subset() function create a new data frame of only versicolor flower data

```
versicolor <- iris[5]
versicolordataframe <- subset.data.frame(iris, versicolor == "versicolor")
versicolordataframe
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 51           7.0         3.2          4.7         1.4 versicolor
## 52           6.4         3.2          4.5         1.5 versicolor
## 53           6.9         3.1          4.9         1.5 versicolor
## 54           5.5         2.3          4.0         1.3 versicolor
```

```
## 55          6.5         2.8          4.6         1.5 versicolor
## 56          5.7         2.8          4.5         1.3 versicolor
## 57          6.3         3.3          4.7         1.6 versicolor
## 58          4.9         2.4          3.3         1.0 versicolor
## 59          6.6         2.9          4.6         1.3 versicolor
## 60          5.2         2.7          3.9         1.4 versicolor
## 61          5.0         2.0          3.5         1.0 versicolor
## 62          5.9         3.0          4.2         1.5 versicolor
## 63          6.0         2.2          4.0         1.0 versicolor
## 64          6.1         2.9          4.7         1.4 versicolor
## 65          5.6         2.9          3.6         1.3 versicolor
## 66          6.7         3.1          4.4         1.4 versicolor
## 67          5.6         3.0          4.5         1.5 versicolor
## 68          5.8         2.7          4.1         1.0 versicolor
## 69          6.2         2.2          4.5         1.5 versicolor
## 70          5.6         2.5          3.9         1.1 versicolor
## 71          5.9         3.2          4.8         1.8 versicolor
## 72          6.1         2.8          4.0         1.3 versicolor
## 73          6.3         2.5          4.9         1.5 versicolor
## 74          6.1         2.8          4.7         1.2 versicolor
## 75          6.4         2.9          4.3         1.3 versicolor
## 76          6.6         3.0          4.4         1.4 versicolor
## 77          6.8         2.8          4.8         1.4 versicolor
## 78          6.7         3.0          5.0         1.7 versicolor
## 79          6.0         2.9          4.5         1.5 versicolor
## 80          5.7         2.6          3.5         1.0 versicolor
## 81          5.5         2.4          3.8         1.1 versicolor
## 82          5.5         2.4          3.7         1.0 versicolor
## 83          5.8         2.7          3.9         1.2 versicolor
## 84          6.0         2.7          5.1         1.6 versicolor
## 85          5.4         3.0          4.5         1.5 versicolor
## 86          6.0         3.4          4.5         1.6 versicolor
## 87          6.7         3.1          4.7         1.5 versicolor
## 88          6.3         2.3          4.4         1.3 versicolor
## 89          5.6         3.0          4.1         1.3 versicolor
## 90          5.5         2.5          4.0         1.3 versicolor
## 91          5.5         2.6          4.4         1.2 versicolor
## 92          6.1         3.0          4.6         1.4 versicolor
## 93          5.8         2.6          4.0         1.2 versicolor
## 94          5.0         2.3          3.3         1.0 versicolor
## 95          5.6         2.7          4.2         1.3 versicolor
## 96          5.7         3.0          4.2         1.2 versicolor
## 97          5.7         2.9          4.2         1.3 versicolor
## 98          6.2         2.9          4.3         1.3 versicolor
## 99          5.1         2.5          3.0         1.1 versicolor
## 100         5.7         2.8          4.1         1.3 versicolor
```

- Give the mean and median of each of the numeric columns for the versicolor data frame
- Why might the mean and median of the entire iris dataset be misleading?

```
#Mean and median for sepal length
print("Mean");lapply(versicolordataframe[1], mean); print("Median");lapply(versicolordataframe[1], media
```

```
## [1] "Mean"
```

```
## $Sepal.Length
```

```
## [1] 5.936
```

```
## [1] "Median"
```

```
## $Sepal.Length
## [1] 5.9
```

```
#Mean and median for sepal width
print("Mean");lapply(versicolordataframe[2], mean); print("Median");lapply(versicolordataframe[2], media
```

```
## [1] "Mean"
```

```
## $Sepal.Width
## [1] 2.77
```

```
## [1] "Median"
```

```
## $Sepal.Width
## [1] 2.8
```

```
#Mean and median for petal length
print("Mean");lapply(versicolordataframe[3], mean); print("Median");lapply(versicolordataframe[3], media
```

```
## [1] "Mean"
```

```
## $Petal.Length
## [1] 4.26
```

```
## [1] "Median"
```

```
## $Petal.Length
## [1] 4.35
```

```
#Mean and median for petal width
print("Mean");lapply(versicolordataframe[4], mean); print("Median");lapply(versicolordataframe[4], media
```

```
## [1] "Mean"
```

```
## $Petal.Width
## [1] 1.326
```

```
## [1] "Median"
```

```
## $Petal.Width
## [1] 1.3
```

```
print("The mean and median might be misleading since we don't know the standard deviation and therefore
```

```
## [1] "The mean and median might be misleading since we don't know the standard deviation and therefor
```

```
print("The dataset for iris also has three different species so this implies that the mean and median m
```

```
## [1] "The dataset for iris also has three different species so this implies that the mean and median
```

1.7.3 Modeling and plotting

- Use the lm() and plot() functions to build a simple linear model
  - of versicolor petal length as a function of petal width
- What are the dependant and independant variables in this case?
- Add the model to the plot with abline()

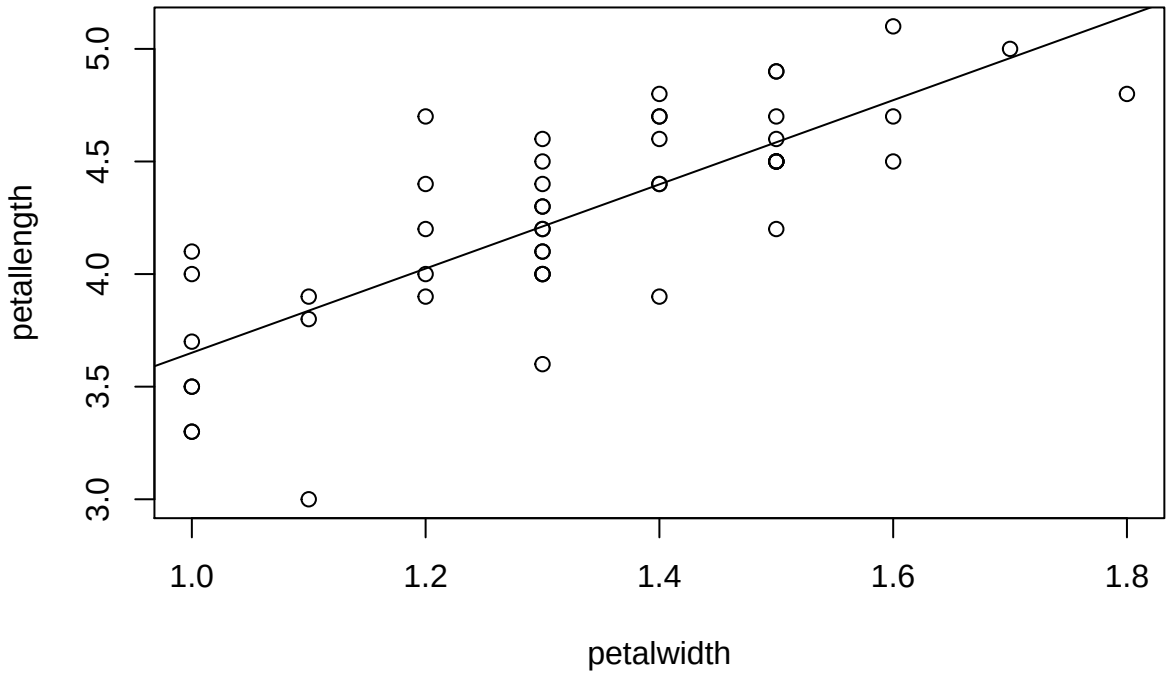
```
print("The dependent variable is the versicolor petal length and the dependent variable is the petal wi
```

```
## [1] "The dependent variable is the versicolor petal length and the dependent variable is the petal wi
```

```
petallength <- versicolordataframe$Petal.Length
petalwidth <- versicolordataframe$Petal.Width
graphofversicolor <- plot(petalwidth, petallength)
graphofversicolor
```

```
## NULL
```

```
fitofversicolor <- lm(petallength ~ petalwidth, graphofversicolor)
abline(fitofversicolor)
```



- Print the summary of this model

```
summary(fitofversicolor)
```

```
##
## Call:
## lm(formula = petallength ~ petalwidth, data = graphofversicolor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8375 -0.1441 -0.0114  0.1984  0.6755
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7813     0.2838   6.276 9.48e-08 ***
## petalwidth    1.8693     0.2117   8.828 1.27e-11 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2931 on 48 degrees of freedom
## Multiple R-squared:  0.6188, Adjusted R-squared:  0.6109
## F-statistic: 77.93 on 1 and 48 DF,  p-value: 1.272e-11
```

1.8 Links

- http://www.r-project.org
- http://rmarkdown.rstudio.com/

<!– # Keep a complete change log history at bottom of file. # Complete Change Log History # v0.00.00 - 1405-07 - Nick Wheeler made the blank script ##########