BIOS: 4120 Lab 5

February 18-19, 2020

Objectives

In today's lab we will:

1. Review & expand what we've covered in R
2. Learn how to construct a Scatter plot
3. Explore linear regression and correlation functions

R Practice and Review:

Set a variable named 'diarrhea' and read in the data using the read.delim function.

```
diarrhea<- read.delim("http://myweb.uiowa.edu/pbreheny/data/diarrhea.txt")
```

Split a Data Frame by Factors

Referring to the goal of the study, we want to know whether Bismuth salicylate was an effective treatment for diarrheal disease in children. Therefore, we want to compare stool volumes between the control and treatment group. How can we find the mean for each of the groups?

You can use the function 'mean', subset the data using brackets [], and specify which group you would like the mean for. As you can see below, you have to write two lines of code.

```
#Average stool volume for control group

mean(diarrhea$Stool[diarrhea$Group=="Control"])

#Average stool volume for treatment group

mean(diarrhea$Stool[diarrhea$Group=="Treatment"])
```

Many times in R, you can get the same desired result using different approaches. We can obtain the same results as above in only one line of code. For example, as practiced in lab last week, you can use the 'by' function to specify that the mean of stool volumes should be given for the control and treatment group, respectively.

```
#Average stool volume for Control and Treatment Group

by(diarrhea$Stool, diarrhea$Group, mean)
```

Note: The third argument asks what function (FUN) you are interested in. You can replace 'mean' with other commands such as summary, min, and max.

Here are a couple more questions to practice and solidify what we have learned so far in R.

1. How would you read in the Tips dataset from the course website?
2. How do you access just the Tip column? Can you find the mean?
3. How do you access just the Tip column but only for the smokers? What's the mean tip from smokers?
4. How do you access just the Tip column but only when the total bill is less than 15? What is the mean for this group?

(For checking purposes, the answers regarding the means are presented below)

```
## [1] 2.998279
## [1] 3.00871
## [1] 2.05025
```

Manipulating Columns

You can perform operations on whole columns of data at a time. For instance:

```
tips$Pctgs <- tips$Tip/tips$TotBill
summary(tips$Pctgs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03564 0.12913 0.15477 0.16080 0.19148 0.71034
```

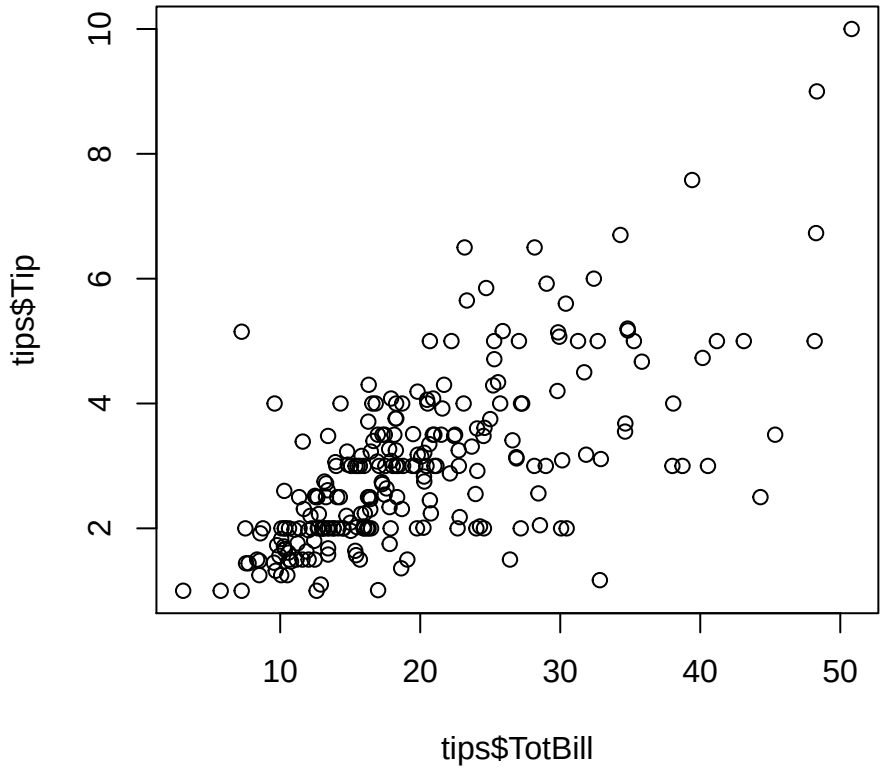
```
tips$Pctgs2 <- tips$Pctgs*100
summary(tips$Pctgs2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.564  12.913  15.477  16.080  19.148  71.034
```

Scatter Plots

As discussed in lecture, when we are interested in visualizing the connection between two continuous variables we can use a scatter plot which is done using the following function:

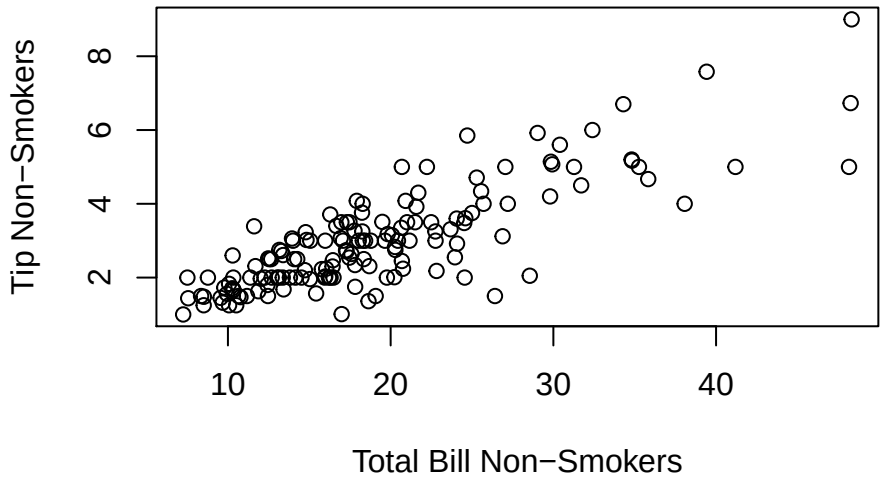
```
plot(tips$TotBill, tips$Tip)
```



Looking at this plot we can see a positive association between the variables and that there is a good amount of variation. Also , you can see horizonatal lines are beginning to form. *What's causing these lines? Why would the scatter plot not be used to compare other variables in this dataset?*

Maybe we're only curious about the total bill compared to the tip for individuals who don't smoke. Here's how we could plot that:

```
plot(tips$TotBill[tips$Smoker=="No"], tips$Tip[tips$Smoker=="No"],
     xlab = 'Total Bill Non-Smokers', ylab = 'Tip Non-Smokers')
```



Notice that we were able to make reading the plot easier by changing the x and y axis labels using the parameters xlab and ylab.

Constructing Models

In R, the function we will use is called 'lm', which stands for "linear model". Regression is useful because it can be generalized to all kinds of settings through the notion of a model, as alluded to by its name in R.

To create a model in R, the code looks like this:

```
mod <- lm(Tip~TotBill,data=tips)
(mod)
```

```
##
## Call:
## lm(formula = Tip ~ TotBill, data = tips)
##
## Coefficients:
## (Intercept)      TotBill
##      0.9203       0.1050
```

Note: The general function should look like lm(y variable ~ x variable, data = the dataset).

Printing out the model itself gives you the intercept and slope. This specific output tells us that for every additional dollar that a meal costs, the waiter can expect to get 10.5 cents more on his tip.

We could have also calculated this using the output from cor() function. By itself, the cor() function calculates the correlation coefficient between two vectors (shown below using the tip and and total bill amounts). When

we multiply that output by the ratio of the standard deviations of those two variables, we end up with the slope of the model.

```
cor(tips$Tip, tips$TotBill)
```

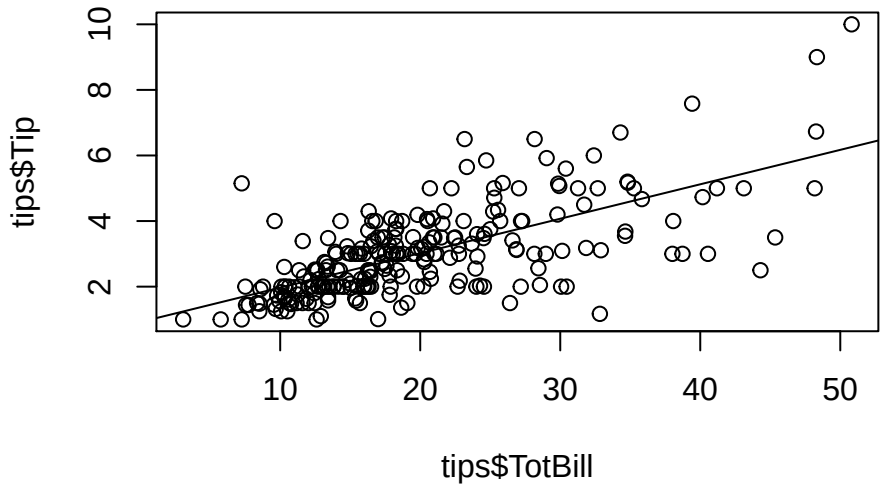
```
## [1] 0.6757341
```

```
cor(tips$Tip,tips$TotBill)*sd(tips$Tip)/sd(tips$TotBill)
```

```
## [1] 0.1050245
```

If you are interested in adding this regression line to a plot of the data, you can use the abline() funciton, and just put the name of your model inside the parentheses.

```
plot(tips$TotBill, tips$Tip)
abline(mod)
```



Example:

A. Make a scatter plot of tip amount vs total bill for only individuals who smoke.

B. Now add a regression line to the plot.

Now let's shift back to looking at the entire dataset.

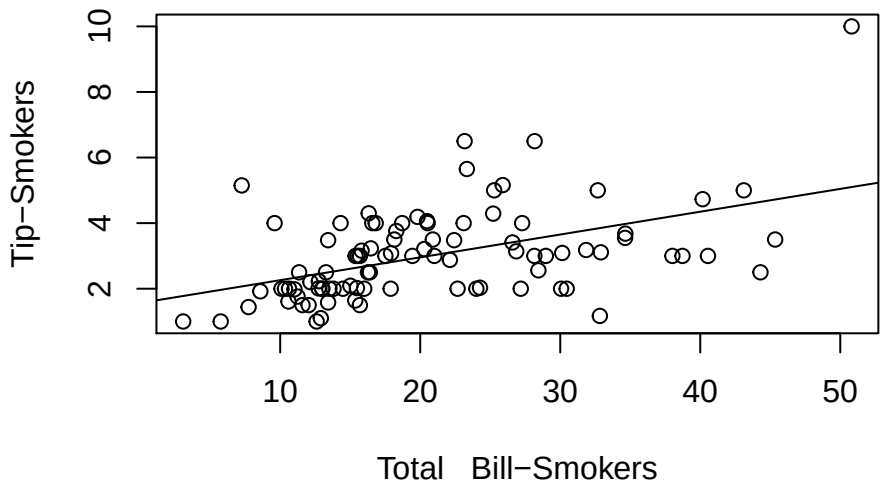
C. Suppose a table is 1 standard deviation above average in terms of total bill. How many dollars above average in terms of tip would you expect it to be?

D. Suppose a table is $2 above average tip. How many dollars above the average total bill would you expect it to be?

E. Suppose a table is $10 above the average total bill. What would we expect the tip to be? Find a solution to this with and without using the "mod" model we created above.

```r
# Part A
plot(tips$TotBill[tips$Smoker=="Yes"], tips$Tip[tips$Smoker=="Yes"],
     xlab = 'Total   Bill-Smokers', ylab = 'Tip-Smokers')

# Part B
smokermodel <- lm(Tip[tips$Smoker=="Yes"]~TotBill[tips$Smoker=="Yes"],data=tips)
abline(smokermodel)
```




```r
# Part C
cor(tips$TotBill,tips$Tip)*sd(tips$Tip)
```

```
## [1] 0.9349715
```

```r
# Part D
Zx <- 2/sd(tips$Tip)
Zy <- Zx * cor(tips$Tip,tips$TotBill)
Zy * sd(tips$TotBill)
```

```
## [1] 8.695428
```

```r
# Part E
Zx <- 10/sd(tips$TotBill)
Zy <- Zx * cor(tips$TotBill,tips$Tip)
(y <- mean(tips$Tip) + Zy * sd(tips$Tip))
```

```
## [1] 4.048524
```

```r
mod$coefficients[1] + mod$coefficients[2]*(mean(tips$TotBill)+10)
```

```
## (Intercept)
##    4.048524
```