Lab12

Jonathan Stiefel

11/3/2020

Learning Objectives

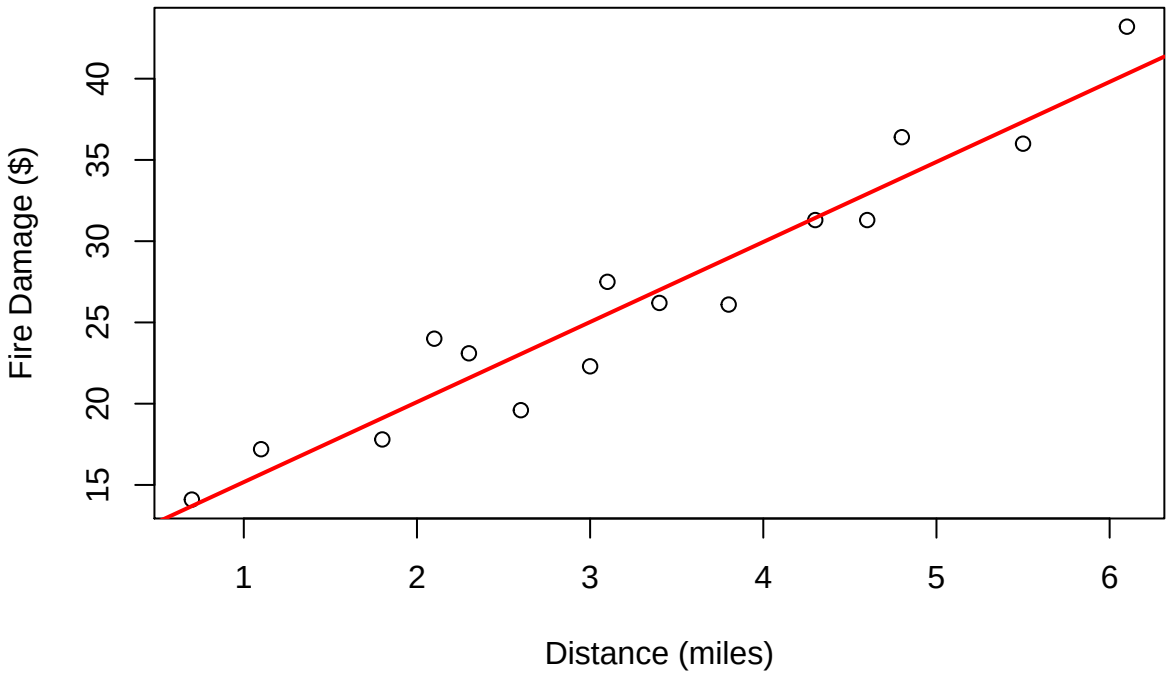
- Use scatterplots and the correlation value to interpret the association between two variables.
- Interpret a linear regression model.
- Use residual plots to evaluate if a linear model is appropriate.
- Compute confidence intervals and hypothesis tests to make inferences about the slope β for a linear regression model.
- Merge two data frames in R (Refer to Lab 11 for instructions)

Problems

Problem 1 Let's go back to the fire data from Weekly Assignment (Lab) 11. Remember that a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the residence and the nearest fire station. This study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage y (in dollars) and the distance x (in miles) between the fire and the nearest fire station are recorded for each fire. The results are tabulated in the csv file named "fire_damage.csv." Last time you fit a regression line to the data. To interpret this data you need to do this again, and print out the summary table of the linear regression:

```
fire <- read.csv("fire_damage.csv") #reads in file
Damage <- fire$Damage #creates Damage variable
Distance <- fire$Distance #creates Distance variable

plot(Damage~Distance, ylab = "Fire Damage ($)",xlab = "Distance (miles)") #creates scatter plot of Dam
reg <- lm(Damage~Distance)
abline(reg, col = "red", lwd = 2)
```




```
summary(reg)
```

```
##
## Call:
## lm(formula = Damage ~ Distance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4573 -1.4750 -0.1308  1.7555  3.4055
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2507     1.4171   7.234 6.61e-06 ***
## Distance      4.9256     0.3919  12.570 1.20e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.311 on 13 degrees of freedom
## Multiple R-squared:  0.924,  Adjusted R-squared:  0.9181
## F-statistic:   158 on 1 and 13 DF,  p-value: 1.196e-08
```

(a) Since you determined that a linear model is appropriate for this data in Lab 11, you can now check the utility of the hypothesized model, that is whether x really contributes information for the prediction of y using the straight line model. First test the null hypothesis that the slope, $\beta = 0$, i.e., that there is no linear relationship between fire damage and the distance from the nearest first station, against the the alternative that distance and fire damage are positively linearly related.

since the p value is small we can reject the null hypothesis that beta = 0 and assume that there is a linear relationship between distance and fire damage.

```
pt(12.57,13)
```

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

(b) To gain additional information about the relationship between the distance and fire damage variables, construct and interpret a 95% confidence interval for the slope.

```
qt(.05,13)
```

```
## [1] -1.770933
```

```
mean(reg)
```

```
## Warning in mean.default(reg): argument is not numeric or logical: returning NA
```

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

(c) Why is the linear model or linear regression model also referred to as a least squares line? The least squares line, aka linear regression model, is the smallest sum of the square of errors.

Problem 2: Gapminder Data Let's go back to the gapminder data we analyzed in Labs 4-6, using descriptive statistics. We looked at the distributions of the variables per capita income (expressed as per-capita gross domestic product (gdp) in units of dollars per year), life expectancy, and per capita CO2 emissions (tonnes per year). It would be to better understand the relationships of these variables to each other. To do so we can create scatter plots, calculate correlation coefficients, and–when appropriate–conduct linear regression. Data for the year 2007 was provided by the gapminder website in two separate csv files: (1) "Gap_CO2.csv", and (2) Gap_Other.csv, which were uploaded to your RStudio Cloud project space.

(a) To start, read in both csv files, assign a name to each dataframe, and display the names of the variables in both data frames.

```
GDP <- read.csv("Gap_Other.csv")
CO2 <- read.csv("Gap_CO2.csv")

names(GDP)
```

```
## [1] "X"         "Country"   "continent" "lifeExp"   "pop"       "gdpPercap"
```

```
names(CO2)
```

```
## [1] "Country" "CO2"
```

(b) Merge the files based on a common variable (or column header). You can refer to the start of Lab 11, to see how to merge two dataframes. I showed you how to do this with the baseball data.

```
merge <-merge(GDP, CO2, by="Country")
```

(c) List the variables included in the dataframe you created to check that the file merged properly. Also make sure you know what each variable is and if you don't ask us, refer back to your previous labs, or search online.

```
names(merge)
```

```
## [1] "Country"   "X"         "continent" "lifeExp"   "pop"       "gdpPercap"
## [7] "CO2"
```

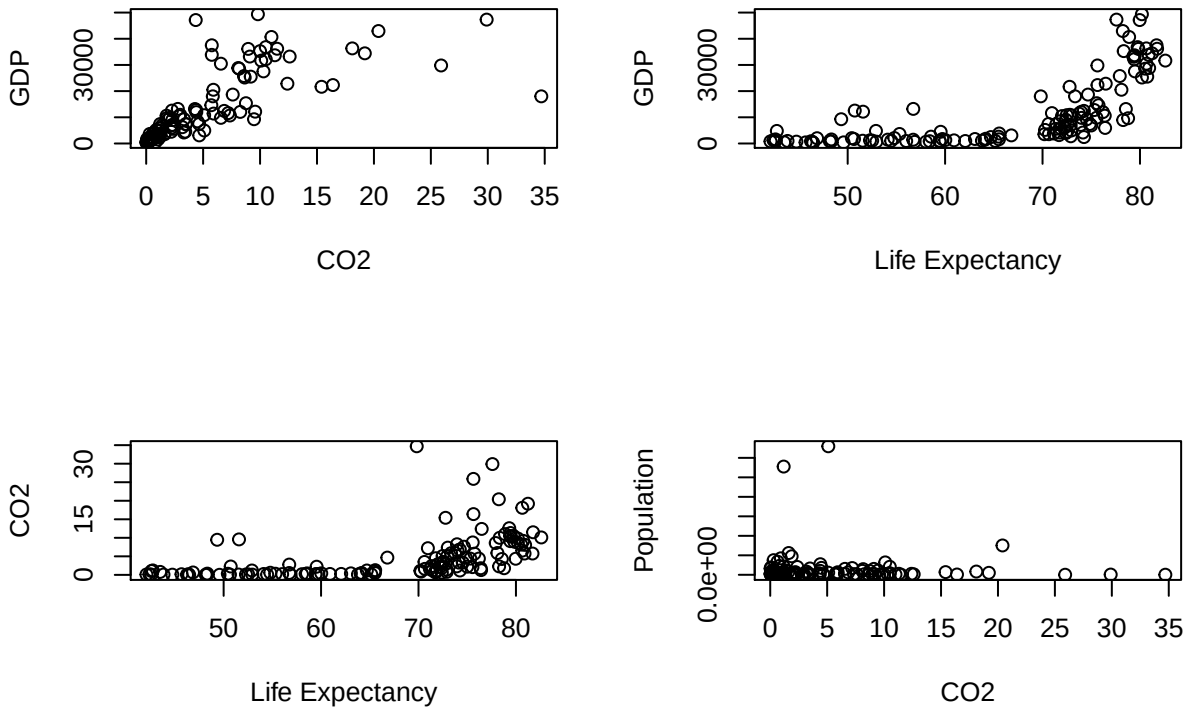
(d) Create a 2x2 panel of four scatter plots for (1) CO2 vs gdpPercap, (2) lifeExp vs gdpPercap, (3) lifeExp vs CO2, and (4) CO2 vs. pop. Use descriptive labels for all plots.

```
par(mfrow = c(2 ,2)) #creates 2 by 2 plottings area
plot(merge$CO2,merge$gdpPercap, xlab = "CO2", ylab = "GDP")

plot(merge$lifeExp,merge$gdpPercap, xlab = "Life Expectancy", ylab = "GDP")

plot(merge$lifeExp,merge$CO2, xlab = "Life Expectancy", ylab = "CO2")

plot(merge$CO2,merge$pop, xlab = "CO2", ylab = "Population")
```



(e) Describe the data. How do per capita GDP and per capita CO2 emissions affect life expectancy? Why is there no relationship between per capita CO2 emissions and population? What is the relationship between GDP and CO2 emissions? Do these relationships make sense to you?

Population and CO2 emission have no relationship because the CO2 data is per capita, while the population data is not. The plot of GDP vs CO2 shows this, that CO2 is really dependent of how much money a nation has, which is related to industry and manufacturing scale. The other two also makes sense because in general a country with a high life expectancy has a good medical care system, which means that the country has money so they have high GDP and CO2 emissions.

(f) Explain why the CO2 vs gdpPercap data set is the most appropriate among the four datasets for a linear regression model. Calculate the correlation and discuss.

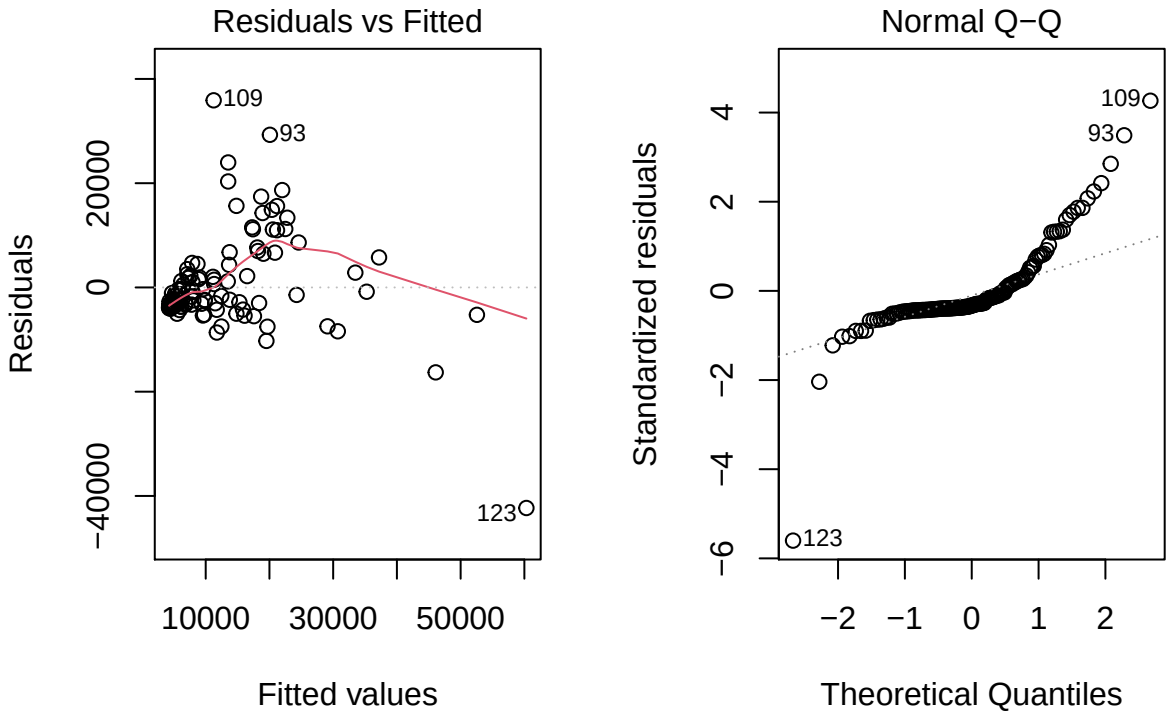
```
vGDP <- merge$gdpPercap
vCO2 <- merge$CO2
cor(vGDP,vCO2)
```

```
## [1] 0.7546162
```

This data set most closely depicts a linear trend.

(g) Create a linear model for the CO2 vs gdpPercap data set.

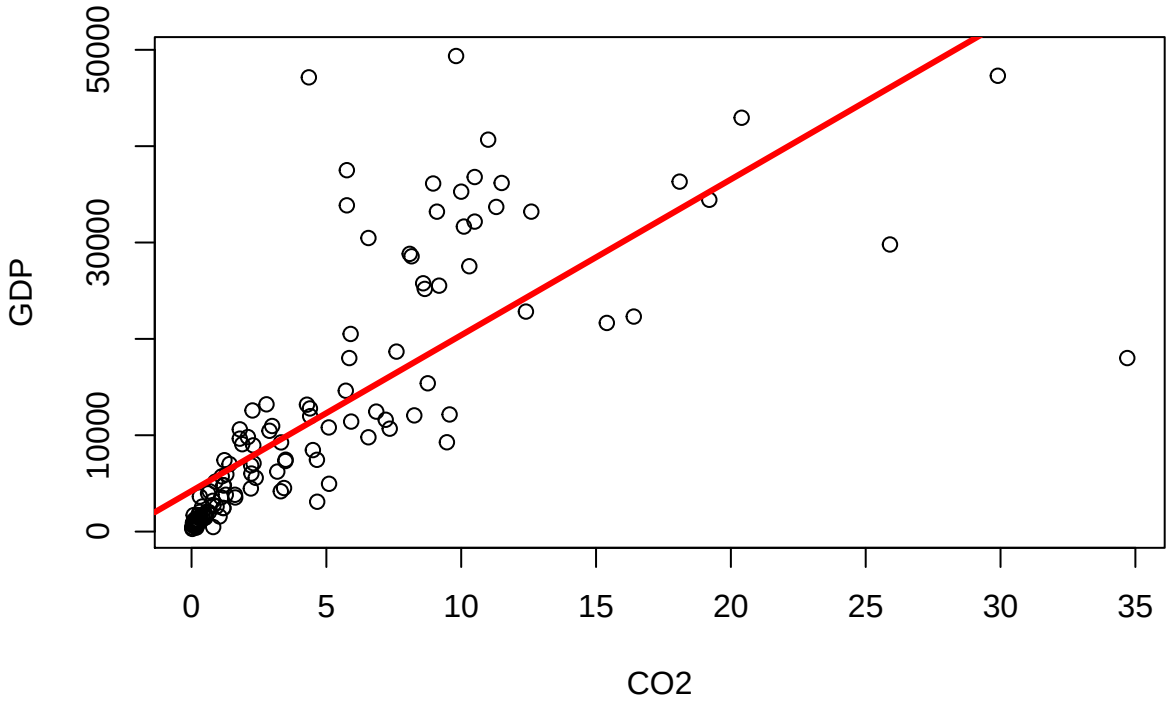
```
vGDP <- merge$gdpPercap
vCO2 <- merge$CO2
par(mfrow = c(1 ,2))
model <- lm(vGDP~vCO2)
plot(model,which = 1)
plot(model,which = 2)
```



(h) Create a new scatter plot for "CO2 vs GDP". Use the `abline( )` function to draw the regression line onto the plot in any color you want. I would use `lwd = 3` to make a thick enough line so it shows up nicely.

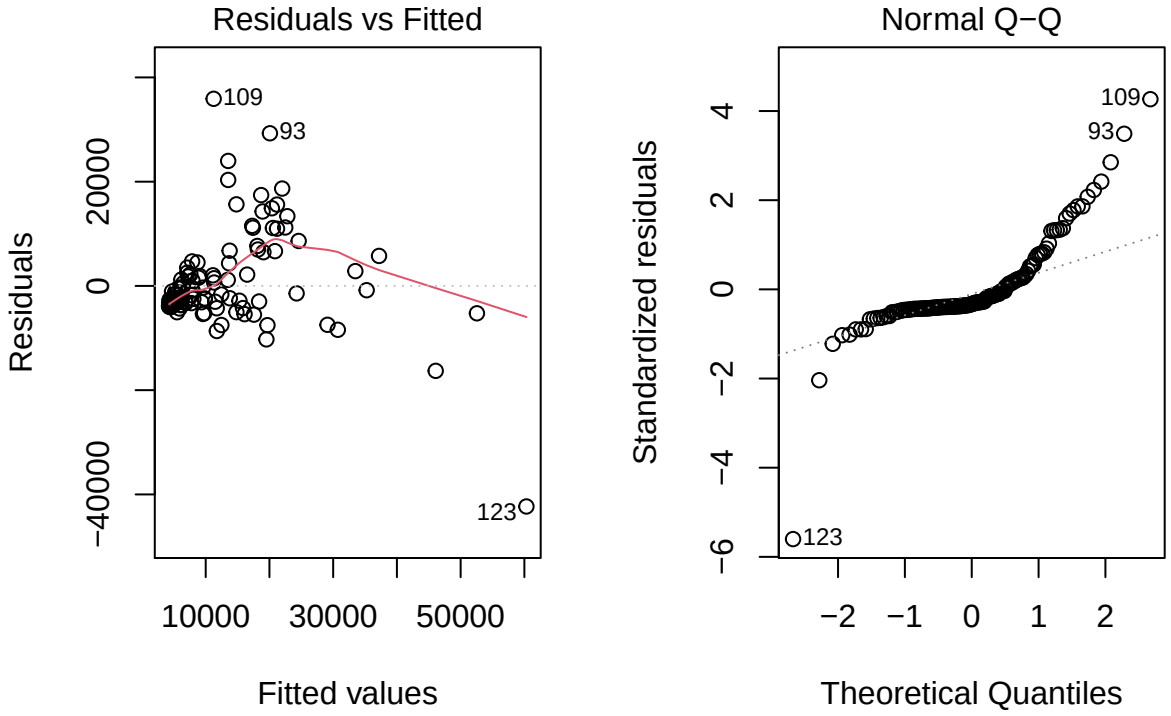
```
CO2 <- merge$CO2
GDP <- merge$gdpPercap

plot(merge$CO2,merge$gdpPercap, xlab = "CO2", ylab = "GDP")
model <- lm(GDP~CO2)
abline(model, col="red",lwd=3)
```



(i) Does the linear model seem appropriate? Justify your answer by creating a 1 by 2 panel and plotting (1) the residuals versus the fitted values, and (2) a normal Q-Q plot, as shown at the beginning of lab (code is in the lab notes). Discuss.

```
par(mfrow = c(1 ,2))
plot(model,which = 1)
plot(model,which = 2)
```

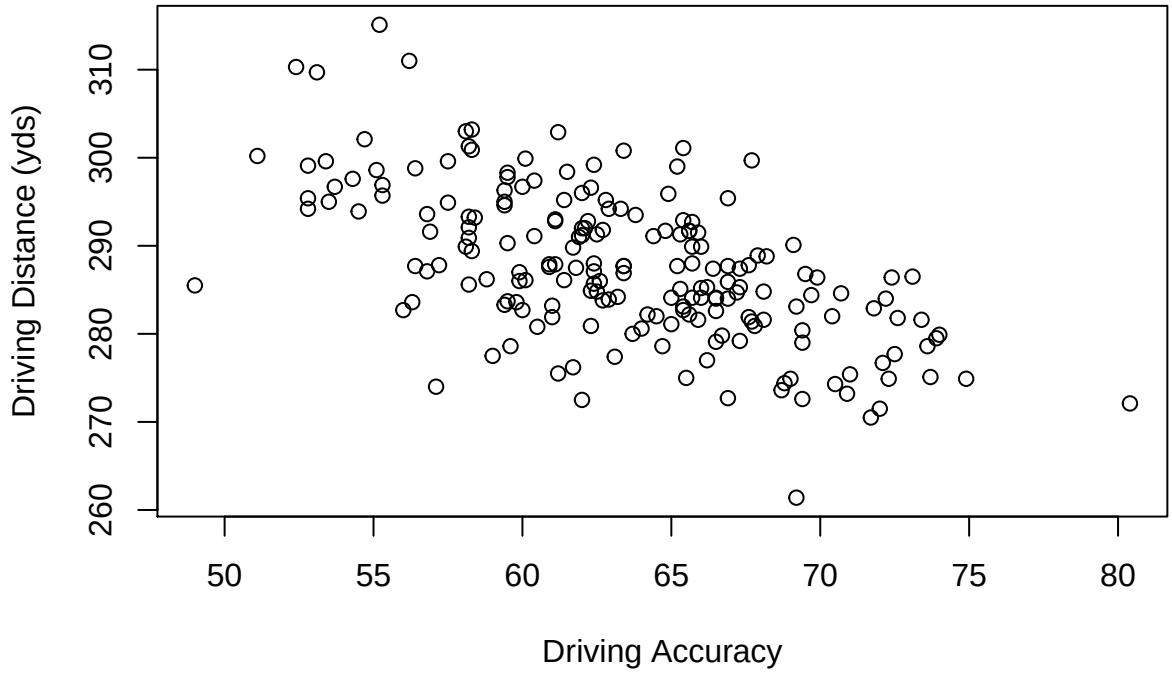


Problem 3 Golf For this question we are going to look at data from the 2008 season of Professional Golf. We will be trying to predict a golfer's Fairway Accuracy (as a percentage) from their Average Driving Distance (in yards).

(a) Read in the data set and create a scatter plot of Fairway Accuracy versus Average Driving Distance. Be sure to put them on the appropriate axes, and label the axes. Compute the correlation and describe the relationship.

The graph shows a correlation between the two variables so we can say that they are related.

```
pga <- read.csv("pga.csv") #reads in pga file
DD <- pga$Ave_Driving_Distance #creates driving distance variable
DA <- pga$Fairway_Accuracy #creates driving accuracy variable
L <- lm(DD~DA) #creates linear regression
plot(DA,DD,xlab = "Driving Accuracy",ylab = "Driving Distance (yds)") #plots
```



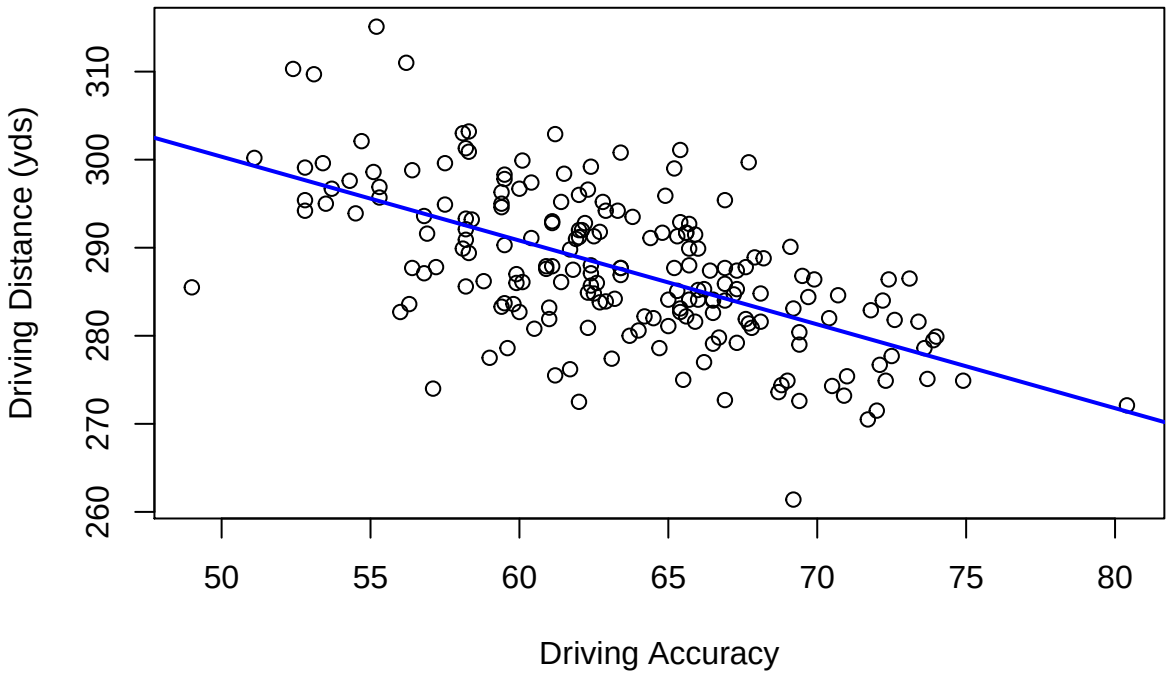

```
summary(L)$coefficents[2,2]
```

```
## NULL
```

(b) Fit a linear regression model and interpret the slope.

The slope indicates that in general an increase in distance means a decrease in accuracy.

```
plot(DA,DD,xlab = "Driving Accuracy",ylab = "Driving Distance (yds)") #plots
abline(L, lwd = "2",col="blue")
```



(c) What is the average accuracy for a player that has an average driving distance of 288 yards? There was actually a player with this average driving distance and they had a fairway accuracy of 62.4%. What is their residual value? mu = 174.9-0.388*288 = 63.2% res = 62.4-63.2=-0.8%* **

(d) How much would the accuracy change on average if a player increases their average driving distance by 20 yards? -0.388*20=-7.76%* ***

(e) Does the linear model seem appropriate? Justify your results. The linear model seems appropriate because the residual is small, which means that the line matches well with the data.

(f) One of the assumptions of the residuals is that they have constant standard deviation σ. What is the estimate of σ for this model? se = sqrt(4.349/195) = 0.14 ****

(g) Create and interpret a 95% confidence interval for the slope of the regression line. Would you conclude that the slope is significant? Justify your answer.

```
qt(0.95,df=195)
```

```
## [1] 1.652705
```

-0.388+1.653*0.0363= -0.328 -0.388-1.653*0.0363= -0.448 [-0.448,-0.328]

The slope is significant because the slope is not within the CI.

Problem 4: Exporting Plots Since R creates nice plots, you might want to export a plot to include in a report. There are a multiple options for getting the plot out of R. Three are listed here

1. The easiest way is to right click on the plot pane, and select "save image as" –> png, and save the image on your computer.
2. If you paste the plotting command in the console (e.g., `plot(y~x)`), the plot will be created in the "plots" pane on the bottom right section of the RStudio window. There is a button there labeled "Export", and if you click that, you are given the option to save the plot as an Image file, as a PDF, or

copy it to the Clipboard. If you choose Clipboard, another window pops up that allows you to modify the dimensions of the plot if you wish, and then copy it to the clipboard. If you click the "Copy Plot" button, then you can open up a document (e.g., a report in MS Word) and choose to "Paste" the plot right in.

3. A third option for saving your plot, if you don't like pointing and clicking, you can add this line of code after your plot commands: dev.print(pdf, "figure_name.pdf"). R will save the plot in the project space on the bottom right section of the RStudio window (sample place as the csv files). You can then select the the file, and click on "More"–>"Export to download the figure to your computer.

Use one of these methods to export one figure you created today. Copy the figure into an MS Word file and submit to Moodle separately from the R Markdown file.