STAT 21 SPRING 2019 HOMEWORK 5 DUE TUES APR 9; NO LATE HW

1. RETIRE STATISTICAL SIGNIFICANCE (8 POINTS)

Read the article, “Scientists rise up against statistical significance” at https://www.nature.com/articles/d41586-019-00857-9.

(a) The article claims, “…researchers have been warned that a statistically non-significant result does not ‘prove’ the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome).” Explain why failing to reject the null hypothesis does not prove that there is no effect. (What does it mean instead?)

(b) In the graphic “Beware false conclusions”, results are shown from two studies: one that found “significant” results, and another that found “non-significant” results. The article claims that it is “ludicrous” to say that the second study found “no association.” Briefly explain why this is the case.

(c) Regarding the same two studies in part (b), the article claims that it is “absurd” to say that the two studies are in conflict, even though one was “significant” and the other was “not significant”. Briefly explain why this is the case.

(d) In the section titled “Quit categorizing”, the article claims that, “Statistically significant estimates are biased… Consequently, any discussion that focuses on estimates chosen for their significance will be biased.” Briefly explain why this is the case.

2. HOME PRICES (15 POINTS)

Download the dataset on home prices from Moodle. These data give the selling prices (in thousands of dollars) of randomly sampled homes in Albuquerque, New Mexico, together with various features of the home: square footage, the age of the home, the number of features possessed out of a list of 11 (e.g., dishwasher, refrigerator, washer and dryer, etc.), whether it is located in the Northeast section of the city, whether it was custom-built, whether it is on a corner lot, and the amount of real estate taxes assessed. Realtors use data such as these to determine the proper selling price for a home.

(a) Read in the dataset using the following commands:

```
# read in dataset
setwd("~/fiil in your pathname")
data <- read.csv("homeprices.csv")

# define variables
price <- data[,"price"]
sqft <- data[,"sqft"]
NE <- data[,"NE"]
```

The size of a home, measured in square feet, should be an important determinant of the home’s selling price. Homes in the Northeast section of the city may also be more expensive, and the effect of square footage may differ in the Northeast section — in other words, square footage and being in the Northeast may have an interaction.

First, make a scatterplot of price vs square footage (sqft). Color the points in the Northeast section (NE = 1) one color and the other points another color. To do this, add the option col=NE+1 to your plot command (you can vary the +1 to some other value to change the colors). To fit regression lines separately for homes in and out of the Northeast, see the R commands handout for the billionaires in class 18 (available on Moodle). If sqft and NE had an interaction, what would be true about the regression lines? Do you see evidence for such an interaction? **(Hand in the scatterplot.)**

(b) Now fit a regression model for price based on sqft and NE, and include an interaction term. Do you see evidence for a statistically significant interaction? **(Hand in the regression table output.)**

(c) Are there outliers or influential points? Calculate the leverages and Cook's distances, and indicate which point(s) may be outliers or influential points in your scatterplot from (a).

(d) Repeat steps (a) and (b) with any outliers or influential points deleted. (To delete a row from the dataset, use data <- data[-k,], where k is the row number you want to delete. Make sure to re-define your variables after deleting the row.) For the purposes of this assignment, delete exactly one point. How do your answers change? **(Hand in the new scatterplot and regression output.)**

3. THE 113TH SENATE (18 POINTS)

This question explores data on votes from the 113th Senate, which was in session during the 2013 and 2014 calendar years. The dataset senate.csv gives data on each senator, the state he or she represents, his or her party affiliation (if any), and their voting record for 657 votes that occurred during this period. For the voting data, 1 indicates Yea, 9 indicates Nay, 5 indicates abstention, and 0 indicates that the person was not a senator at the time of that vote. Your assignment is to carry out a principal components analysis and plot the data on the first two principal components.

In this dataset, the raw data are the results of the 657 votes. These data are actually categorical, whereas principal components should strictly speaking be used only with quantitative data, but we will still be able to discover structure in the data nonetheless.

(a) Read in the dataset using the following commands:

```
# read in dataset
setwd("~/directoryname")
data <- read.csv("senate.csv")
dim(data)     # should be 105 rows, 660 columns

# define variables
name <- data[,"name"]
state <- data[,"state"]
party <- data[,"party"]
votes <- data[,4:660]
```

(b) Run a principal components analysis (PCA) using the following commands. The center=T and scale=T options standardize the data before running PCA. The first line runs the PCA and stores the results in pca. The second line saves the PC values for each senator in a variable called PCscores. The last two lines plot the first two PC's in a scatterplot. **Hand in the scatterplot.**

```
pca <- prcomp(votes, center=T, scale=T)
PCscores <- pca$x
plot(PCscores[,1], PCscores[,2], type="n")                # set up plot with no points
text(PCscores[,1], PCscores[,2], label=name, cex=.4)    # plot the names
```

(c) Let's color each senator red or blue by party affiliation. Use these commands to create a vector of colors:

```
partycolor <- rep(NA, nrow(data))
partycolor[party=="D"] <- "blue"
partycolor[party=="R"] <- "red"
```

Then redo the commands from step (b), but in the last line use the option col=partycolor to add color to the scatterplot. **What does the first PC represent?**

(d) There are a number of outliers, such as John Kerry. **Why are these points outliers?** A possibly helpful web page:

https://en.wikipedia.org/wiki/113th_United_States_Congress

(e) Remove all the points that are outliers for the reason that Kerry is an outlier. You should remove 8 rows. Be careful to remove the correct rows; each row removed changes the remaining row numbers, so double-check the row numbers at each step. After removing the rows, be sure to redefine the variables from part (a) and the colors from part (c). When you have done so, try running the PCA again using the commands in part (b). You will get an error. **What is causing the error?** You may find the command sort(apply(data, 2, sd), decreasing=T) to be useful; this calculates the SD of each column of the dataset.

(f) To fix the problem, remove the problematic votes (you should remove 15 columns). Again, be careful to remove the correct columns. Be sure to re-define your variables from part (a) and the colors from part (c) when you are done. Re-run the PCA and make a scatterplot of the first two PC's. **Hand in the scatterplot.**

(g) Use the command summary(pca) to print a list of the variances associated with each PC and the cumulative proportion of variance represented. **How much variance is accounted for by the first PC? By the second? How many PC's do you need to account for 90% of the variance of the dataset?**

(h) In your scatterplot, what does the first PC represent? Can you determine what the second PC represents? (I am not sure myself what the answer is.) **What outliers remain, and what makes them outliers?**

(i) Are there any independent senators? Who are they, and do they resemble Democrats or Republicans more closely?

(j) According to the first two PC's, which party is more "coherent" in the sense that their PC values are more similar? Can you tell what accounts for the differences in the less coherent party?