STAT-S 470/670 In-class Data Analysis Quiz 09/29/14

Complete the following two data analysis problems, and submit a file of your saved plots and R code electronically to Canvas. In your code file, provide the file names of your save plots and brief answer under code to each question.

1. Example in lecture: Use productivity measures to grade 15 employees who work on an assembly line. Each employee was tested once, then again a month later.

Employee #	Test 1	Test 2
1	50	58
2	35	46
3	15	40
4	64	76
5	53	62
6	18	39
7	40	57
8	24	41
9	16	31
10	67	75
11	46	62
12	64	64
13	32	54
14	71	65
15	16	51

(a) Use boxplots with notches to compare distributions of two sets of test scores. Based upon the notches in the two box plots is there a significant difference in the median values between test 1 and test 2?

(b) Find median and F-pseudosigma for each set of test scores.

(c) Use the estimates found in (b) as mean and standard deviation of normal distributions to generate two new sets of test scores, each of 500 data.

(d) Use notched boxplots to compare distributions of the two new datasets.

(e) Use Letter-value boxplots to compare distributions of two new datasets.

2. The second problem will be concerned with the pressure data set in R.

```
data(pressure)
?pressure
attach(pressure)
y=pressure
x=temperature
```

This data set concerns the relationship between the vapor pressure of mercury as a function of temperature. There are two columns in the data set: the temperature in degrees Celcius (C) and the corresponding vapor pressure of mercury measured in millimeters of Mercury (mmHg). For this problem you are interested in modelling the pressure as some function of temperature $P = f(T)$, so pressure will be the "y" variable and temperature will be the "x" variable. To bring this problem into the 21st century, I am going to begin by converting the units of measurement to SI units. This will also avoid some problems with log transforms if necessary (notice the minimum value of "temperature" is 0)...

```
y=y * 0.1333 # pressure is now in kiloPascals
x=x + 273.15  # Temp in Kelvin instead of Celcius
```

(a) Create a scatterplot of pressure versus temperature.

(b) If we want to transform the "y" variable in this plot for **straightness** then construct an appropriate plot for this data and fit a line to this which will tell us what power p transform we need to use for the y variable. (Hint: You may round the slope down to the nearest 1/2 integer).

(c) It is clear that the data is not straight and the above transformation won't work!! When you get desperate enough to find an answer, you'll eventually start reading the literature. Fortunately, I found from Industrial & Engineering Chemistry Research that suggested to me that a reciprocal transform of the temperature might be worth a try.... So now let's try plotting the data to the reciprocal of temperature using plot($y \sim I(1/x)$). A word about what the term $I(1/x)$

means in R - this means take the formula "as is". So mathematically, $I(1/x)$ will treat the explanatory variable as "$1/x$" both when we plot the data and when we fit a linear model to it.

(d) Fit a linear model to y using $1/x$ as the response variable using lm($y \sim I(1/x)$) and then based upon your estimated slope construct an appropriate plot for straightness just like we did in part b except use the variable $1/x$ as the new "x" variable. Based upon the plot you constructed in part d and the slope of the line which best fits this straightness plot, what power p should we choose for the y variable?

(e) We can use the Box-Cox transformation in R, by typing in boxcox($y \sim I(1/x)$). This recommends using $p = 0$, so the log transformation is recommended. Does the plot produced by plot(log(y) $\sim I(1/x)$) seem linear to you? If so, fit this model using R and report out the equation for the final model.