Homework 1 - Linear Regression Review

NAME: Your Name

NETID: Your NetID

DUE DATE: Friday, February 12, 2016 by noon

Homework 1 Instructions

1. For each problem:
    a) Answer all questions
    b) Insert code chunks directly under any problems that require you to use R and type in code as needed. In particular, make sure code chunks are included for any requested plots.
    c) Answer any questions related to the problem in the .Rmd document directly under the question
    d) Note: Occasionally when you insert a code chunk it may not go where you intend it to. If this happens, you can cut and paste it into the correct spot. Make sure the code chunk is aligned to the left margin of this document. Often it may be easier to just store a code chunk on your clip board and paste it in when you need one.
    e) You may need to knit your document occasionally to answer questions related to R output.
2. Submit your homework as a pdf document compiled from R markdown. Your file should be named *LastFirst*-HW1.pdf.

Question 1 (10 points)

Make up a data set that you will find to be a relevant source of examples. The data set can be anything, but is intended to be helpful for you to translate statistical model and tests into real world statements. Suggested sources: - a data set that you already have, or something made up to be close to it - a data set you expect to collect in future research - data that you would like to collect if only it were possible - data (possibly collected from somewhere else) associated with your field of study - data about some hobby or area of personal interest to you

You do not need to make up values for your data, but you should answer

(2 points)

a) How many experimental units (subjects, observations etc) are in the data?
b) What measurements are being taken?

In particular, you should record (but not be able to control) at least one of each of the following measures

(6 points)

a) Continuous variables that could take any value (may need to be positive)
b) Binary variables (yes/no, true/false, 1/0 etc)
c) Count variables (0, 1, 2, ...)

(2 points)

Additionally, you should have some variables that you have controlled. These could be due to selection (specified numbers of men and women, specified numbers of dogs in certain weight categories) or because of active treatments you have applied (applying certain pesticides to some plants and not to others, for example).

Question 2 (Adapted from Ramsey and Schafer Ex 8.25)

The U.S. presidential election of November 7, 2000 was one of the closest in history. As returns were counted on election night it became clear that the election would be determined by the outcome in the state of Florida. At one point in the evening the state was projected to be carried by the Democratic nominee, Al Gore but that projection was retracted a few hours later. Early in the morning of November 8, the networks projected that George W. Bush had carried the state and with it the presidency. This projection, too, was later retracted. It ultimately took weeks of legal arguments about voting irregularities to decide a winner.

One of the controversies at the time (although it did not play a significant role in the legal decisions) was the "butterfly ballot" of Palm Beach County. A number of Democratic voters complained that this ballot was confusing: voters were required to fill in a circle going down the middle of the ballot with the candidate's names on either side. Filling in the second circle counted as a vote for the Reform Party candidate, Pat Buchanan. However, since Al Gore's name appeared second on the left hand side of the ballot, it was speculated that some voters who intended to vote for Gore may have mistakenly filled in the second circle instead of the third.

Two pieces of evidence were advanced in support of this. The first was a large number of ballots discarded because both the second and third circles were filled in. The second was that Buchanan had an unusually high percentage of the vote in that county.

The file Election2000.csv contains the number of votes for George Bush and for Pat Buchanan in the 2000 election in each of the 67 Florida counties. The aim of this exercise is to determine if Palm Beach was anomalous among them.

1. (10 points) Plot the data along with a line of best fit. Do any patterns in the data emerge?
2. (10 points) What is the influence of Palm Beach county on this fit as measured by cooks disstance? Does it appear unusual? (The **cooks.distance()** function in R will provide cook's distance for a fitted model.)
3. (10 points) Re-do the analysis without Palm Beach. Examine the residuals, does anything concern you about them? (Remember that **x[-j]** will remove the j'th element of **x**.)
4. (10 points) Try transforming the data (without Palm Beach) a number of ways. What model do you feel best fits the linear regression assumptions? Express the formula for the model.
5. (10 points) Is the model significant at the 0.05 level? How does the model imply the number of votes for Buchanan will change as the number of votes for Bush increases?
6. (10 points) Plot the studentized residuals and produce a QQ plot of the residuals for your final model. Do the normal assumptions appear to have been met? In the MASS library, the **studres()** function will provide studentized residuals.
7. (10 points) Produce a prediction interval for the number of votes for Pat Buchanan in Palm Beach county. Does the observed number of votes for Buchanan fall within this interval? You can obtain prediction intervals from the **predict()** function by specifying **interval="prediction"**.

8. (10 points) *Carefully* interpret your statistical findings. On what assumptions do they rely? What does this analysis tell you about what Gore's vote would have been without the butterfly ballot?