Homework 3: due Tuesday, 11 October

1. **Florida Vote Count.** In 2000, the US Supreme Court decided the presidential election after reviewing recounts of the ballots from some parts of Florida. A major issue was the count in Palm Beach county, where many voters claimed that the ballot design was confusing and had led to mistaken votes for Buchanan (Reform Party) that were intended for Gore (Democratic Party). We will compare the actual votes for Buchanan in Palm Beach with the predicted number of votes for Buchanan based on the other counties in Florida. Data taken from Agresti and Presnell (2002) *Statistical Science,* 17, 436-440. (Available through U of M Library E-Journals.)

The data are given in florida2000.xls with these columns:

County: name of Florida County
Palm Beach: an indicator variable equal to 1 for Palm Beach county, 0 for all others
Population: number of county residents
Black percent: percent of Black county residents
Hispanic percent: percent of Hispanic county residents
White percent: percent of White county residents
percent 65: percent of county residents aged 65 and over
percent college: percent of population who had attended college
Income: county average annual income in $1000s
Gore: number of county votes for Al Gore
Bush: number of county votes for George W. Bush
Buchanan: number of county votes for Pat Buchanan
Nader: number of county votes for Ralph Nader
Total votes: number of county votes

The aim in this problem is to estimate the vote for Buchanan in Palm Beach from these county characteristics, **without using the data from Palm Beach to fit or check the model.**

a. Using all the data, plot the vote for Bush (horizontal axis) against the vote for Buchanan (vertical axis). Where is Palm Beach in the plot? What does this suggest about the vote for Buchanan in Palm Beach?

b. Make a prediction data set in which Buchanan's vote in Palm Beach is set to missing.

   Compute *log(Buchanan)* and square root of Buchanan (look up the SQRT function). Rescale votes for the other presidential candidates as percents of total votes. Finally, check whether any variables need log transformation. List any variables that you log transform.

c. Fit *log(Buchanan)* on all the predictors *excluding* percent votes for Gore, county name, Palm Beach indicator, and Buchanan—this is *Model A.* Present the SAS table of regression coefficients, standard errors, t statistics, and p-values.

   Make a studentized residual plot for *Model A.* Do you see any problems?

d. Reduce *Model A* by dropping nonsignificant predictors, and call this *Model B.* Use your own judgement in this model reduction. Check your reduced model with a residual plot, and report the SAS table of regression coefficients, standard errors, t statistics, and p-values.

e. Some analysts claimed the square root transformation of Buchanan's vote was better than log. Fit the square root of Buchanan's vote to your predictors in *Model A* and call this *Model C.* Make a studentized residual plot—do you see any problems?

f. Predict the number of votes for Buchanan in Palm Beach county from models *A, B, C* and calculate a 95% confidence interval for each prediction using the standard error for a new observation, not the standard error for a predicted mean. Make a table to display these results and back-transform so that the units in your table are *votes*. Are there any differences between the predictions?

g. Buchanan got 3407 votes in Palm Beach county. Estimate the number of Gore votes lost by mistakes to Buchanan using each of your Buchanan vote estimates *A–C*, and add these results to your table. Explain how you made your estimates.

2. **Grades.** The spreadsheet Grades.xls contains 4 homework scores, midterm score, and final exam scores. The first column is student number, and the observation `student = 0` gives the total possible points for each homework and test.

a. Find the percent score on each homework. For each student, drop the lowest percent homework score and average the rest. If a student missed two homeworks, drop one and give the second a score of zero. Call this result the "homework average."

b. Compute the weighted course average, using these weights: 60% for the homework average, 20% midterm, and 20% final exam. For students 5001–5006, print the homework average and weighted course average.

c. Draw a histogram of the course averages.