Background

The 2000 U.S. Presidential election was eventually settled when George Bush officially beat Al Gore by 537 votes, out of nearly 6 million cast, in the state of Florida. This gave Bush the 25 Electoral College votes he needed to become President under the U.S. Constitution. During the month that it took to resolve this election, many allegations of voting irregularities were made. One of the strongest arguments concerned Palm Beach County, the second largest of Florida’s 67 counties, with over 400,000 votes cast for a total of ten candidates. Here, it was alleged, the so-called “butterfly ballot” design had misled many voters into voting for a different candidate from the one they intended. Specifically, it was claimed that many voters who had intended to vote for Al Gore in fact voted for Pat Buchanan, the candidate of the Reform Party. The initial returns gave Buchanan 3,407 votes in Palm Beach, compared with about 1,300 that he might have expected based on his overall vote in the state. The excess of about 2,000 votes, had they indeed been credited to Gore, would have given Gore the state of Florida and hence the White House.

Within a few days of the election, numerous papers appeared on various websites, containing statistical analyses of the election results, and focusing specifically on the Buchanan vote in Palm Beach County using various political or demographic covariates to predict the number of votes he should have received. Most such analyses in fact put the predicted number of Buchanan votes at considerably less than 1,300. For example, Palm Beach overall is a Democratic county—Al Gore gained 62% of the total vote there—so one would have expected the right wing Buchanan to gain a smaller percentage of votes than he did over the whole state. Comparisons like this have generally been taken as strengthening the argument that Gore should have won, but they also raise questions about the correct application of regression methodology to the results of an election.

The purpose of this assignment is for you to examine this hypothesis in detail, by carrying out a regression analysis of the number of Buchanan votes across all 67 Florida counties, using both the votes for other candidates and demographic variables, obtained from the Census, as covariates. Your focus will be on the use of graphics as an aid to assist in and simplify the analysis.
1. The two data files `CountyVotingData.txt` and `CountyDemographicData.txt`.

The first data set consists of the provisional election returns (i.e., counts) from the Florida Division of Elections for all 10 candidates, and the second contains demographic data compiled from the U.S. Census Bureau.

We have prepared a merged data file which also includes variables for Buchanan and Bush percentage votes.

Download the data file `CountyData.csv` and read it into R. The data set consists of the provisional election returns (i.e., counts) from the Florida Division of Elections for all 10 candidates, and the demographic data compiled from the U.S. Census Bureau. The demographic data include the following variables:

(a) Pop: county population in 1997,

(b) Whi: percentage of whites in 1996,

(c) Bla: percentage of blacks in 1996,

(d) Hisp: percentage of Hispanics in 1996 (note that the percentages of whites, blacks and Hispanics sometimes add up to more than 100, because Hispanics include other races),

(e) $\geq 65$: percent of the population aged 65 and over (actually calculated from the 1996 population aged 65 and over, divided by the 1997 total population),

(f) HS: percentage of the population graduating from high school (1990 census)

(g) Coll: percentage of the population graduating from college (1990 census),

(h) Inc: Mean personal income (1994).

2. As an initial analysis of the data, plot the Buchanan percentage vote against 12 covariates (including the percentage votes for the other candidates). In each case, mark Palm Beach County with an X.

3. Add a loess trend line to each of the 12 plots. What do you conclude about the relevance of these variables for a regression analysis?

4. Carry out a multiple regression using the proportion of voters who supported Buchanan in each county as the dependent variable (the one to be predicted and the demographic variables plus the percentage voting for Bush as the independent variables. Why are we not including the percentage voting for Gore?

5. Compute the studentized residuals and plot them against the fitted values as well as in a q-q plot. Do you have any reason for concern? If so, why?

6. Redo the regression analysis dropping those predictors which seem to contribute in only a limited fashion to the predictive ability of the model, and reexamine the residual plots.
7. If we thought of the votes for Buchanan as being binomially distributed by county, then we would expect the variance to vary by the relative size of the Buchanan vote. One way to stabilize the variance in such circumstances, at least approximately, might be a square-root transformation. Repeat the regression analysis predicting the square root of the proportion of voters who supported Buchanan as the response variable.

8. Compute the studentized residuals for this new analysis and plot the residuals against the fitted values and in a q-q plot. Has the transformation helped with any problems?

9. Redo the regression analysis dropping those predictors which seem to contribute in only a limited fashion to the predictive ability of the model, and reexamine the residual plots. Does the regression model now look reasonable? Explain. [Note: Once you transform the Y variable the “significan” predictors may change.]

10. In class we looked at a simple scatterplot with the Buchanan results by county. In addition to Palm Beach, there were a couple of additional potential outlying observations. Check to see what has happened to the predictions for these counties. Are they any longer a concern in your analyses?

11. The boxcox() command in R allows you to find an “optimal” transformation for this dataset. The boxcox() function is in the MASS library, so you will need to type

   library(MASS)

   before issuing the boxcox command. Once you have loaded the MASS library, a help file is accessible via ?boxcox. Recall from class that the boxcox function is similar to that of lm().

   What is the “optimal” value of the Box-Cox parameter \( \lambda \)? How well does your approach with square-roots work as an approximation. [Be careful in reading your output here since the Box-Cox transformation does a rescaling of \( Y \) that may make coefficients hard to compare.]

12. Your goal is to predict the Palm Beach vote. Thus you need to refit the regression model omitting the values for Palm Beach itself. Carry out the revised analysis, and reexamine the residual plots.

13. Using the new model based on omitting Palm Beach, compute a predicted value for Palm Beach as well as a 95% prediction interval. Translate your prediction interval back into one for the Buchanan vote total.

14. Based on your analyses, explain why Buchanan’s Palm Beach vote can or cannot be explained away as normal statistical variation. What does your analysis have to say about whether Gore should have won the 2000 election?

   Be sure to include your graphs and an edited version of the regression analyses output as part of the solutions. Throughout, if you find that additional residual and/or other plots are helpful, by all means include them as part of your solutions.