HW7: Chapter 7, Sections 8.1 to 8.3

Your Name Here

The code below just loads some packages and makes it so that enough digits are printed that you won't get confused by rounding errors.

```
library(dplyr) # functions like summarize
library(ggplot2) # for making plots
library(readr)

options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

Crowdedness and GDP

Danielle Vasilescu and Howard Wainer (*Chance*, 2005) used data from the United Nations Center for Human Settlements to investigate aspects of living conditions for several countries. Among the variables they looked at were the country's per capita gross domestic product (GDP, in dollars) and Crowdedness, defined as the average number of persons per room living in homes there. Suppose we want to estimate the relationship between these variables, using GDP as the explanatory variable and Crowdedness as the response.

The following code reads the data in:

```
crowdedness <- read.csv("http://www.evanlray.com/data/sdm4/Crowdedness.csv")
```

(a) Create an appropriate plot of the data.

(b) Find a transformation of the data so that the simple linear regression model conditions are as well satisfied as possible. You do not need to show all of the steps in your process; you can just keep your final selected transformation. (It's also fine if you want to keep all of the steps you took for your records.) For your final selected transformation, please create 3 plots: (1) a scatter plot with the transformed variables, (2) a scatter plot of the residuals vs. the transformed explanatory variable, and (3) a histogram or density plot of the residuals. No need to discuss these plots in this part.

(c) Discuss all of the linear regression model conditions based on your transformed variables. For each condition, you should write a sentence or two describing whether or not the condition is satisfied and why. If your conclusion is based on the plots you made for part (b), please clearly indicate which plot or plots you are looking at and describe a specific characteristic of that plot that your conclusion is based on.

(d) What are the interpretations of the estimated intercept and slope? Please interpret the coefficient estimates in context on the scale of the *transformed* data.

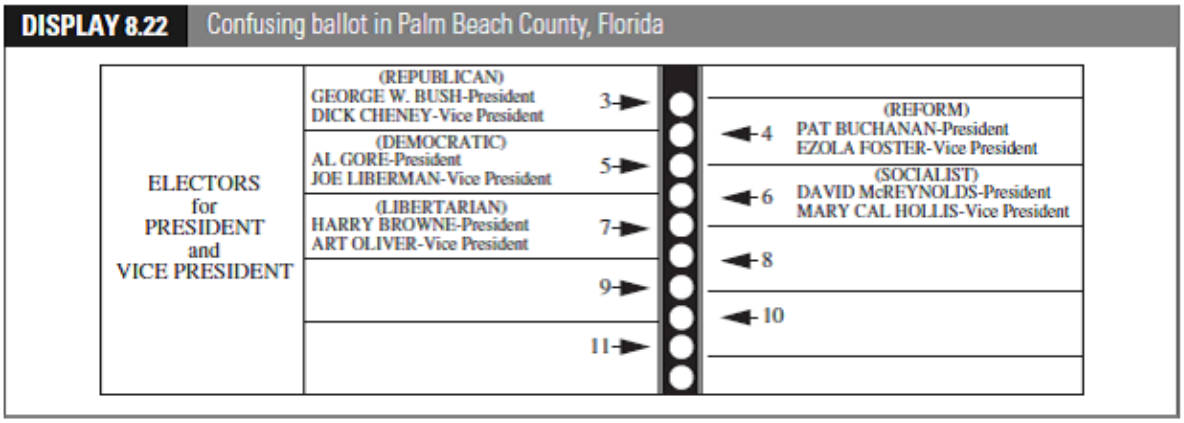
(e) Find a set of three Bonferroni-adjusted confidence intervals with familywise confidence level of 95% for the median crowdedness in the "population" for countries with a GDP of $5000, $25000, and $45000. Interpret your intervals in context. You can use the predict function to generate the confidence intervals on the transformed scale, but you will have to then transform back to the original data scale.

The Dramatic US Presidential Election of 2000 (Sleuth3 Exercise 8.25)

Quote from the book:

> The US 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 outcome in the state of Florida would determine the next president. ... When the roughly 6 million Florida votes had been counted, Bush was shown to be leading by only 1,738, and the narrow margin triggered an automatic recount. The recount, completed in the evening of November 9, showed Bush's lead to be less than 400.
>
> Meanwhile, angry Democratic voters in Palm Beach County complained that a confusing "butterly" lay-out ballot caused them to accidentally vote for the Reform party candidate Pat Buchanan instead of Gore. The ballot, as illustrated in Display 8.22 [included in this repository, or knit the document to view], listed presidential candidates on both a left-hand and a right-hand page. Voters were to register their vote by punching the circle corresponding to their choice, from the column of circles between the pages. It was suspected that since Bush's name was listed first on the left-hand page, Bush voters likely selected the first circle. Since Gore's name was listed second on the left-hand side, many voters – who already knew who they wished to vote for – did not bother examining the right-hand side and consequently selected the second circle in the column: the one actually corresponding to Buchanan. Two pieces of evidence supported this claim: Buchanan had an unusually high percentage of teh vote in that county, and an unusually large number of ballots (19,000) were discarded because voters had marked two circles (possibly by inadvertently voting for Buchanan and then trying to correct the mistake by then voting for Gore [though we don't have data to check this theory]).
>
> [We have] a data set containing the numbers of votes for Cuchanan and Bush in all 67 counties in Florida. What evidence is there in teh scatterplot of Display 8.24 that Buchanan received more votes than expected in Palm Beach County? Analyze the data without Palm Beach County results to [fit a model for] predicting Buchanan votes from Bush votes. Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach from this result – assuming the relationship is the same in this county as in the others. If it is assumed that Buchanan's actual count contains a number of votes intended for Gore, what can be said about the likely size of this number from the prediction interval? (Consider transformation.)

Here is a picture of the ballot (will show up in the knitted pdf):



The following code reads in the data:

```
votes <- read.csv("http://www.evanlray.com/data/sleuth3/ex0825_2000_election.csv")
```

Please conduct the analysis outlined in the book's description. You will need to use the `filter` function to create a copy of the data set without the observation for Palm Beach County. When you have found a suitable transformation of the data, please create a scatter plot with the transformed variables, a scatter plot of the residuals vs. the transformed explanatory variable, and a histogram or density plot of the residuals. Discuss whether the linear model conditions are satisfied and any limitations of this analysis (does this analysis prove that the high rate of voting for Buchanan in Palm Beach County was caused by the ballot? Why or why not?).