Data Science for Politics — Problem Set 1

Due: Friday, February 19.

There are **two** separate pieces to this problem set.

1. Use the `problem_set_1.R` file as a template to complete your analysis. Submit this file on Gradescope to the "Problem Set 1 - Code" assignment. This piece will be autograded. You may resubmit your code on Gradescope until the due date and it will be regraded. You must use the file name `problem_set_1.R` when submitting your code to Gradescope.
2. Write up your answers to the substantive questions, including any graphs and tables you may be asked to create, and submit this file on Gradescope **as a pdf** to the "Problem Set 1 - Writeup" assignment. This piece will be graded by the professor and teaching fellows. When you upload your PDF to Gradescope, you will be asked to indicate on which page each answer is on.

Questions 13 and 14 are optional, for extra credit.

The Incumbency Advantage in the U.S. House

In this problem set, we will use a dataset on U.S. House election returns (`house.csv`) to study the incumbency advantage (and more). This dataset contains 7 variables, the first three of which uniquely identify the election by providing the state, year, and district in which the election took place. The next four columns provide the vote totals for the Democratic and Republican candidates (`vote_D` and `vote_R`), and indicate whether each candidate is an incumbent (`inc_R`, for example, takes the value 1 if the Republican candidate is an incumbent, and 0 otherwise).

1. Load the `house.csv` dataset and save it to an object named `house`. (1 point)
2. Create a new variable, called `dem_pct`, which reflects the Democratic vote percentage, defined as `100 * vote_D/(vote_D + vote_R)`.
3. Calculate the overall mean of `dem_pct` and save the value to an object named `dem_pct_mean`.
4. Calculate the mean of `dem_pct` for elections in which a Democratic incumbent is running and save the value to an object named`dem_pct_dinc`.
5. Calculate the mean of `dem_pct` for elections in which a Republican incumbent is running and save the value to an object named`dem_pct_rinc`.
6. Calculate the mean of `dem_pct` for elections in which NEITHER party has an incumbent running (an "open seat") and save the value to an object named`dem_pct_open`.
7. Compare these three quantities. What do they suggest about how incumbents perform in elections? Write a few sentences summarizing these results in your write-up file.

8. Calculate the difference between the mean Democratic vote share in elections with a Democratic incumbent and the mean Democratic vote share in open-seat races (i.e., races where neither party has an incumbent running). Save the value to an object named `dem_pct_diff`. Do you think that this quantity is "causal"? That is to say, does the difference between these two quantities reflect the causal effect of having a Democratic incumbent as opposed to having no incumbent on the Democratic vote percentage? Why or why not? (Answer this part in the write-up file.)

9. Plot a histogram of the Democratic vote share for open-seat races, and for races with a Democratic incumbent, respectively. Save your plots as image files, and include them in your write-up file. In your write-up, comment on some of the differences between the two plots. (Note: The `ggsave()` function is useful for saving ggplots to PNG or PDF.)

10. In one of the histograms, there appears to be some "lumping" at the end of the plot. What are these races? Why doesn't the other histogram seem to have this same lumping?

11. Use the `group_by` and `summarize` functions to compute the average Democratic vote share by year, for all election years. Save the results to on object called dem_pct_year, with columns `year` and `dem_pct_mean`. (Note: in `summarize`, you can name the variables that are calculated. For example, `summarize(x, new_var_name=max(some_variable))`.) Create a plot showing the change in the `dem_share_mean` variable over time. Add this plot to your write-up and comment on what you observe.

12. Calculate the average Democratic vote share for each state, and save the results to an object named `state_avg`, with columns `state` and `dem_pct_mean`. Report the averages for Massachusetts and Texas. Are the results what you would expect? Why or why not? Comment on what you find in your write-up.

13. (Extra Credit) Using any coding concepts you want, figure out which state has changed the most in absolute value, in terms of average Democratic vote share, between 1970 and 2010. Report the state, the amount it has changed, and whether it has become more Democratic or more Republican. [Note: this question is a bit more challenging than the others. You may want to use Google/StackOverflow. Answers that involve manually calculating all 50 changes will not receive credit. Include your code and your answer in your write-up.

14. (Extra Credit) In each year, calculate the average Democratic vote share for (1) Open seats, (2) seats with a Democratic incumbent, and (3) seats with a Republican incumbent. Exclude any cases where there are both Democratic and Republican incumbents. Make a plot showing comparing these three groups over time. Your plot should have clear axis labels and a legend differentiating the three groups. Customize the plot however you like to make it as clear and informative as possible. Include the plot in your write-up. Describe any trends you observe.