Problem set 5

Your name here

Due 11/5/2021 at 5pm

Note: *Start with the file* `ps6_2021.Rmd` *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F21/tree/main/assignments). Modify that file to include your answers. Make sure you can "knit" the file (e.g. in RStudio by clicking on the* `Knit` *button). Submit both the Rmd file and the knitted PDF via Canvas*

In this assignment we will examine data from the Cumulative CCES Common Content dataset assembled by Shiro Kuriwaki from the 2006-2020 Cooperative Congressional Election Studies. You can find the dataset and the codebook for the dataset at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/II2DB6. They are also on the course repository.

The code chunk below loads the data and creates a few variables to get you started. (You may have to change the path to get the code to run, depending on where you saved the dataset.) In subsequent code you should work with `dat`, which is created by this code chunk.

```r
cces <- readRDS("./../data/cces/cumulative_2006-2020.rds")

dat <- cces %>%
  filter(!st %in% c("IN", "KY", "TN", "NE", "KS", "SD", "ND", "ID", "HI", "AK") & year %% 2 == 0) %>%
  mutate(starthour = lubridate::hour(starttime),
         starthour = case_when(st %in% c("WA", "OR", "NV", "CA") ~ starthour - 3,
                               st %in% c("MT", "WY", "CO", "UT", "NM", "AZ") ~ starthour - 2,
                               st %in% c("OK", "TX", "MN", "IA", "MO", "AR", "LA", "WI",
                                         "IL", "MS", "AL") ~ starthour - 1),
         starthour = ifelse(starthour < 0, 24 + starthour, starthour),
         approve_pres = as.integer(approval_pres %in% c(1,2)),
         startcat = case_when(starthour >= 5 & starthour < 12 ~ "1) morning",
                              starthour >=12 & starthour < 17 ~ "2) afternoon",
                              starthour >= 17 | starthour < 1 ~ "3) evening",
                              starthour >= 1 & starthour < 5 ~ "4) late night"))
```

1) Create a histogram of `starthour`, which indicates what time (on the 24 hour clock) the respondent started the survey. Specify `binwidth = 1`.

2) Make another figure that shows the same histogram separately by year. You should see that there seems to be one pattern of survey timing for 2006, 2018, and 2020, and another for the other years. Which one is less surprising to you?

We're not sure why there are these two distinct patterns (it could be a difference in the manner of survey administration, or a coding error), but to be safe we'll focus on years with a similar pattern of response and the same president. In subsequent questions, restrict attention to 2010-2016.

3) Use `geom_smooth()` to show how the average `approve_pres` (a variable that was created above) changes over the course of the day. Interpret the result.

4) Use `group_by()` and `summarize()` to compute the average of `approve_pres` by `starthour` (another variable that was created above) and plot the result.

5) Regress `approve_pres` on `startcat` (another variable created above) and interpret the coefficients. Calculate the mean of `approve_pres` by `startcat` using `group_by()` and `summarize()` and compare this to the regression coefficients.

Looking at the results so far, one might wonder whether starting a survey in the evening causes respondents to give a higher rating to the president, while starting the survey in the late night/early morning causes respondents to give a lower rating to the president. There is a small literature suggesting that people's survey responses depend on the time of day when they take the survey.

6) In brief, how could you design a randomized experiment to evaluate this hypothesis?

7) Use `group_by()` and `summarize()` to compute the average age of respondents by `starthour` and plot the result. Interpret what you find. Does it appear that `starthour` is randomly assigned?

8) Regress `approve_pres` on age and interpret the coefficients. Does this support the idea that age might be a confounder for the relationship between `starthour` and `approve_pres`?

9) Regress `approve_pres` on age and gender and note how the coefficient on `age` differs from the previous regression. Show how to use the omitted variable bias formula to account for this difference.

10) Regress `approve_pres` on `startcat` again but now controlling for age. Using the `modelsummary` package (which will be discussed in lab), make a figure or table comparing the key coefficients (i.e. those relating to the the time of day) when you don't control for age and when you control for different polynomials of age. (Hint: You might want to specify `coef_omit = "Int|age"`, which leaves out the intercept and the age coefficients.)

11) What other variables in the dataset would you control for if you wanted to assess the causal impact of the time of day on approval of the president?

12) If there is a regression you plan to run as part of your final project, describe it here. If not, explain why not.