Final assignment

Due December 5, 2023, at 9pm

(Your name here)

Note: Start with the file `final_assignment_2023.qmd` (available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F23/tree/main/assignments). Modify that file to include your answers. Make sure you can "render" the file (e.g. in RStudio by clicking on the `Render` button). Submit both the qmd file and the PDF via Canvas.

Further note: *Collaboration is not permitted on this assignment.* You may ask clarifying questions on the course stackoverflow (please do not share code). You may not share code or discuss answers with anyone else otherwise.

For this assignment, you will analyze a dataset of your own choice. Choose a dataset that interests you, but that also

- has a respectable number of observations (rows), e.g. at least 40
- allows you to carry out all the tasks in the list below, which means you need at least 3 variables

Ideas for finding datasets:

- Many journals require authors to share replication data, often at the Harvard Dataverse (links below), so that if you find a paper that interests you that was published in one of those journals in recent years, you should be able to access the data
  - Quarterly Journal of Economics
  - American Journal of Political Science
  - American Political Science Review
- Other datasets you might find interesting
  - Penn World Tables
  - Varieties of Democracy – see also `R` package
  - Comparative study of electoral systems (CSES)

  - Cooperative election study (formerly CCES)
  - American National Election Study (ANES)
  - British Household Panel Survey
  - British Election Study (BES)
  - Roper iPoll US public opinion data
  - World Bank open data sets

- The University of Chicago Library has lists of resources for political science, economics, and other subject areas (some more useful than others)

You may find a dataset that is not available in formats we have used so far (`.RData`, `.RDS`, `.csv`). (Check what formats are available: e.g. on Harvard Dataverse there is often an `.RData` version or a `.csv` version, though you may need to click on "Access data" to see it.) There are many tools for reading data formatted for other statistical software into `R`. The `haven` package reads SPSS, Stata, and SAS files into `R`; the `readxl` package reads Excel spreadsheets into `R`; the `read.table()` command is helpful for reading in different kinds of text files (`.txt`, `.tab`). If you google "read X file in R" where X is e.g. SPSS or SAS or Stata you'll find some help.

Your task is to use your chosen dataset to display your understanding of concepts from the class and your ability to execute skills we have worked on.

1. Data description

What do the rows of the dataset represent (people, municipalities, a country-year, etc)? What are some of the important variables you will be using, and (if not obvious) how were they measured? Is this a sample from a population (if so what is that population?), or is this the whole population? (If your dataset is not a sample from a population, you will assume that is is for the purpose of inference below, as is standard in social science research.)

(Your answer here.)

2. Potential research questions

What research questions might this data be useful for answering?

(Your answer here.)

3. Dataset loading, variable creation and other data manipulation

The tasks below will probably require you to create new variables and exclude missing observations; you may also want to rename some variables for your convenience. Put your data loading and data manipulation code here so we can see the skills you have accumulated in this area.

```
# your code here
```

4. Inference for a mean

Estimate the mean of 3 important variables in the dataset. For each one, report a 95% confidence interval in a table. For one of the variables, state a null hypothesis that could be considered interesting to test and report a p-value for that null hypothesis. Explain in words what your p-value means. In all cases explain what assumptions you are making.

(Your code and data here – you get the idea.)

5. Difference between means

Define two means that could be interesting to compare. (It could be two variables in the dataset, or a single variable for two groups of observations defined by another variable, e.g. average growth rate of democracies and non-democracies). Estimate the difference in the two means, report a confidence interval for that difference, and report a p-value. (Do not use regression for this question.) Explain in words what your p-value means. Explain what assumptions you are making.

6. Scatterplot

Choose a dependent variable (Y) and an independent variable (X). Present a scatterplot with Y on the vertical axis and X on the horizontal axis. Both X or Y should be numerical variables, not categorical variables such as religion or race. It is helpful if either X or Y is a continuous variable (i.e. one that takes on many values) rather than a discrete variable; if not, you may *want to use `jitter()` to avoid plotting many points on top of each other. Comment on any apparent relationship between X and Y and why it might arise.*

7. Regression with one predictor

Regress Y on X to estimate the BLP of Y as a function of X. Interpret the slope coefficient. In a table, report a 95% confidence interval and p-value for the slope coefficient based on three approaches:

- *classical standard errors as computed by `lm()`*
- *robust (Huber-White) standard errors as computed by `estimatr::lm_robust()`*
- *bootstrap standard errors (naive bootstrap) State any assumptions you are making.*

8. Confidence interval for a prediction

Based on the model above, what is your prediction for Y at the highest value of X in your dataset? Report a 95% confidence interval for that prediction, stating and justifying any assumptions. (Hint: you'll need the variance rule and/or the bootstrap.)

9. Regression with more than one predictor

Again regress Y on X, but now include additional predictors other than X. Explain why you might want to include these additional predictors in the model. Fit at least two models with different sets of predictors or transformations of predictors. Report the results in a single regression table (using `modelsummary::modelsummary()` or a similar approach). Compare the coefficient on X across models.

10. Interactions

Again regress Y on X, but this time include an interaction between X and a numerical predictor W. Tell us what W is and explain why the relationship between X and Y could plausibly depend on W. Report a confidence interval and p-value for the interaction term, stating any assumptions you use. Explain in words what the estimated interaction term means.