Problem Set 3

Samantha-Jo Caetano

Due: Monday November 2, 2020 at 11:59pm ET

Logistics

- Individual-level survey data:
    - Request access to the Democracy Fund + UCLA Nationscape 'Full Data Set': https://www.voterstudygroup.org/publication/nationscape-data-set. This could take a day or two. Please start early.
    - Given the expense of collecting this data, and the privilege of having access to it, if you don't properly cite this dataset then you will get zero for this problem set.
    - Once you have access then pick a survey of interest. We will use "ns20200102.dta" in the example.
    - Use the example R code to get started preparing this dataset, and then go on cleaning and preparing it based on what you need.
    - Make graphs and tables about the survey data and write beautiful sentences and paragraphs explaining everything.

- Post-stratification data:
    - We will use the American Community Surveys (ACS).
    - Please create an account with IPUMS: https://usa.ipums.org/usa/index.shtml
    - You want the 2018 5-year ACS. Then you need to select some variables. This will depend on what you want to model and the survey data, but some options include: REGION, STATEFIP, AGE, SEX, MARST, RACE, HISPAN, BPL, CITIZEN, EDUC, LABFORCE, INCTOT. Have a look around and see what you are interested in.
    - Download the relevant post-stratification data (it's probably easiest to change the data format to CSV). Again, this can take some time. Please start this early.
    - This will be a large file. Don't attempt to push it to GitHub (use the .gitignore file - see here: https://carpentries-incubator.github.io/git-Rstudio-course/02-ignore/index.html).
    - Given the expense of collecting this data, and the privilege of having access to it, if you don't properly cite this dataset then you will get zero for this problem set.
    - Clean and prepare post-stratification dataset.
    - Use the ACS data to create the post-stratification dataset. Remember that you need cell counts for the sub-populations in your model. See examples in the readings.

- Modelling.
    - You will want to explain vote intention based on a variety of explanatory variables. Construct the vote intention variable so that it is binary (either 'supports Trump' or 'supports Biden').

- You are welcome to use `lm()` but you would need to explain the nuances of this decision in the model section (Hint: start here: https://statmodeling.stat.columbia.edu/2020/01/10/linear-or-logistic-regression-with-binary-outcomes/).
- That said, you should probably use logistic regression if it is all possible for you. If you don't know where to start then look at (in increasing levels of complexity) `glm()`, `lme4::glmer()`, or `brms::brm()`. There are examples of each in the readings. Note: whether you use a Bayesian model or a Frequentist model is up to you.
- Think very deeply about model fit, diagnostics, and other similar things that you need in order to convince someone that your model is appropriate.
- You have flexibility of the model that you use, (and hence the cells that you'll need to create next). In general, the more cells the better, but you may want fewer cells for simplicity in the writing process and to ensure a decent sample in each cell.
- Apply your trained model to the post-stratification dataset to make the best estimate of the election result that you can. The specifics will depend on your modelling approach but will likely involve `predict()`, `add_predicted_draws()`, and similar. See the examples in the readings. We are primarily interested in the distribution of your forecast of the overall popular vote, and how the explanatory variables affect this. But great submissions would go beyond that.
- Create beautiful graphs and tables of your model and results.
- Create wonderful paragraphs talking about and explaining everything.

- Write up:
  - Using R Markdown, please write a report about your analysis and compile it into a PDF.
  - The paper must be well-written, draw on relevant literature, and show your statistical skills by explaining all statistical concepts that you draw on.
  - The paper must have the following sections:
    - title, name/s, and date,
    - model,
    - results,
    - discussion (weaknesses & next steps), and
    - references.
  - The report may use appendices for supporting, but not critical, material. (This is optional).
  - This report should be roughly 2 pages long. If you use an advanced model or wish to do some subsequent analysis 3-4 pages will be accepted.
  - The discussion needs to be substantial and well thought out. In the discussion, in addition to subsections about the existing model and results and similar, the paper must include subsections on weaknesses and next steps - but these must be in proportion.