Problem Set 3

Samantha-Jo Caetano

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

This Problem Set consists of one question, which is to be submitted as a 2 page report.

This problem set should be completed in an R markdown file and should be knit to a pdf document. Your submission will have 3 parts: (i) Output/Final Copy of Report; (ii) R markdown code, .Rmd file; (iii) link to a Github repository of your code (this will include your .R scripts for cleaning the code).

Please have all three files available for submission at the due date - if you do not submit your personal .Rmd files then you will receive a 0 on this Problem Set.

Your Objective

To predict the overall popular vote of the 2020 American federal election using a regression model with post-stratification.

Note: There are videos posted, as well as the **ProblemSet3-Additional-Instructions.pdf** providing information on how to access the data and run an analysis.

What you will do

- You will use the post-stratification technique described in the MRP videos, slides and paper provided in Week 6.
- The idea is, as a small team (of size 1-4) you will work through the following steps:
  1. Load in the sample data
  2. Build a model (any model is acceptable) on the sample data

     Note: any model is acceptable, but some justification (either practical or statistical) should be given. (Some options: meaningful variables, p-values, AIC, BIC, etc.)
  3. Load in the census data
  4. Calculate $\hat{y}^{PS}$
- The report will consist of the following sections (more details in "Report Details" section below):
  - Title & Authors
  - Model
  - Results
  - Discussion
  - References
- You are all working with the same data and same overall problem, so you do not need to include an introduction or data description.

Submission Process

- As a team, via Quercus, submit a PDF of your paper. Again, in your paper you must have a link to the associated GitHub repo in an appendix.
- Via Quercus you will need to submit the following three files:
  - pdf of your final report.
  - your group .Rmd file.
  - a link to a Github repository with your materials.
- Everyone in the team receives the same mark.
- There should be no evidence that this is a class assignment.
- Due to the American election, late assignments will not be accepted. Reports received on or before Monday November 2, 2020 at 11:59pm ET will be accepted and considered for full marks. Any submissions after the grace period will receive a mark of 0 on this problem set.

Report Details

Below are notes about what should be in each main section, I have included sub-sections that are optional for you to include:

Title & Authors:

- Include an aptly named title for your report.
- Include authors names and the date.

Model:

Model Specifics:
(1-2 paragraphs)

- Here you will describe the chosen model (e.g., if you decide to perform linear regression you must write out the model and describe the parameters and variables included) and give some justification for why this model was selected.
- This should be no more than two paragraphs long (ideally it will be only one paragraph) and will include some mathematical notation when explicitly stating the model. You should describe the notation used (i.e., β_0 is the intercept which represents…., etc.)

Post-Stratification:
(1 paragraph)

- Here you will describe the post-stratification technique.
- This should include a sentence or two about what post-stratification is (in non-statistical language) and a sentence or two on why it is useful.
- As part of the stratification technique you should also describe the cell/bin splits in the sample data. Here you should describe the variables that you are using to create the cells. You can briefly justify the choice to include or exclude certain variables when creating the cells/bins. (For example, choosing "state" because it is likely to influence voter outcome because of…., or not including "eye colour" because it is not available in the census data).

Additional Information (change this title to something more apt):
(Optional – 1-2 paragraphs)

- If you want to include some additional analysis (e.g., standard error, post-stratification by state, etc.) then you should describe your methodology here.
- Note, this section is optional and you will not lose marks by not including this section.

Results:

(1 paragraph)

- Here you should relay the value calculated for $\hat{y}^{PS}$ and briefly describe what the number represents.
- Any “additional analysis” results should also be described here.
- Be concise in this section. Simply relay the facts (in a digestible way).
- Note about describing the results:
    - Do not just write: We calculated $$\hat{y}^{PS}$$ to be 0.529.
    - Do write: We estimate that the proportion of voters in favour of voting for <Party Name> to be 0.529. This is based off our post-stratification analysis of the proportion of voters in favour of <Party Name> modelled by a <type of model> model, which accounted for <list of variables in the model>.
    - As a minimum, you pretty much only need to include the statement above (in your own words and filled in accordingly) in this section, but you likely will have some other results to relay.

Discussion:

<u>Summary</u>:

(1 paragraph)

- Summarize what was done earlier
- The idea is to tie everything together.

<u>Conclusions</u>:

(1 paragraph)

- Here is where you say who you think will win the primary vote and why.
- Elaborate on the results and make your prediction.
    - Example: Based off the estimated proportion of voters in favour of voting for <Party Name> being 0.529, we predict that <Party Name> will win the election.

<u>Weakness & Next Steps</u>:

(1-2 paragraphs)

- This sub-section can be split into two, if needed
- Addressing weaknesses of the analysis.
- Addressing future steps of the analysis.
    - Hint: a good future step might be to compare with the actual election results and do a post-hoc analysis (or at least a survey) of how to better improve estimation in future elections.

References:

- Include a bibliography that includes all ideas that were employed that were not your own.
- Do your best to be as thorough as possible here. It is only right to give credit where credit is due.
- Referencing should be consistent, organized and well formatted.
- This can go beyond the recommended page limit.