Final Project – Additional Instructions

Samantha-Jo Caetano

Rough Draft Due Date (2%): Wednesday December 9, 2020 at 11:59pm ET

Peer Review Due Date (3%): Monday December 14, at 11:59pm ET

Final Report Due Date (25%): Monday December 21, 2020 at 11:59pm ET

This Final Project is to be handed in as a report.

This final project 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.

General Logistics

Here are the general expectations:

1. Everything is entirely reproducible.
2. Your paper must be written in R Markdown.
3. Your paper must have the following sections:
    1. Title, date, author, keywords, abstract, introduction, methodology (data and model), results, discussion, appendix (optional, for supporting, but not critical, material), and a reference list.

d. Your paper must be well-written, draw on relevant literature, and show your statistical skills by explaining all statistical concepts that you draw on.

e. The discussion needs to be substantial. For instance, if the paper were 10 pages long then a discussion should be at least 2.5 pages. In the discussion, the paper must include subsections on weaknesses and next steps - but these must be in proportion.

f. The report must provide a link to a GitHub repo that contains everything (apart from raw data that you git ignored if it is not yours to share). The code must be entirely reproducible, documented, and readable. The repo must be well-organised and appropriately use folders and README files.

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.

Abstract:

(~ 1 paragraph)

- Here you are provided a brief summary of the entire report.
- This is generally written as one long paragraph.

Keywords:

(~ 1-2 lines)

- Here you are providing the reader with key words.
- Usually includes anywhere from 3 to 10 key words (or "topics")
- Here is an example (based on the example text in the sections below)

Key words: Propensity Score, Causal Inference, Observational Study, Lung Cancer, Smoking

Introduction:

(~ 3-4 paragraphs)

- Here you will introduce your problem.
- First Paragraph: In this section you should start off by giving some background/context explaining the global relevance of the problem/data/analysis. For example:
    - Statistical analysis is ubiquitous to clinical research. The causal links inferred from clinical studies affect treatment options, screening techniques and help identify patient risk factors. Observational data is often more feasible, and arguably more reliable, than experimental design data. Thus, having the ability to make causal inference in this setting is key from both an economical and practical perspective. …..
- Second Paragraph: In the next paragraph you will get a little more specific about your specific problem/analysis and it should be a bit more niche. For example:
    - One popular way to make causal inference is through propensity score matching (citation). Propensity score matching was first introduced in 1983 (citation) and has become widely popular in recent years (citation). In this report, I will use propensity score matching to discern if there is a causal link between whether or not a patient smokes and whether or not a patient developed lung cancer…….
- Additional paragraphs will be of a similar nature to paragraph 2, but just introducing other ideas. For instance, maybe I would need to include a paragraph going over smoking and lung cancer (or maybe 2 paragraphs).
- Final Paragraph: In the last paragraph you will let the reader know how you will layout the rest of the report. For example:
    - Two data sets will be used to investigate how propensity score matching could be used to make inference on the causal link between smoking and lung cancer. In the Methodology section (Section 2), I describe the simulation study, the data, and the model that was used to perform the propensity score analysis. Results of the propensity score analysis are provided in the Results section (Section 3), and inferences of this data along with conclusions are presented in Conclusion section (Section 4).
- General Tip: Try to avoid first person in this section, usually using "I" or "my" will come across as though it is your own personal opinion/motivation. You want this section to come across as though it is scientifically/factually driven, usually this is a bit more compelling to a scientific reader.

Methodology:

(~ 3-4 paragraphs – length will vary depending on selected option)

- This section will vary depending on what option you selected from a, b, c, d, e in the Final Project Instructions document.

Data:

- Here you will describe the data set.
- If it is simulated you should describe how you simulated the data and why you used certain distributions and parameters. For example:
    - I simulated n survival times from the exponential distribution with mean $\beta=10$, representing the time from diagnosis until death for each of the n lung cancer patients. The exponential distribution is representative of continuous variables that are positive, making it appropriate to represent survival times. Moreover, the parameter of $\beta=10$ was selected based off previous studies which summarize characteristics regarding lung cancer patients (Include citation).
- If you did not simulate data, and instead are working with real data you should include a "Table 1" which is a Table providing baseline characteristics of the data (usually separated by treatment groups). Your text should go over key components of this table.
- You really want to use this section to set yourself up, so that when you are describing the model and results, it is evident why you selected that model, and why results would include certain variables.

Model:

- 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 will include some mathematical notation when explicitly stating the model. You should describe the notation used.
- This section really will vary depending on the option selected.

Results:

(~ 3-4 paragraphs – length will vary depending on selected option)

- Here you should relay any tables and graphs that are a result of some intended statistical analysis. There should be text describing the results of these tables and figures.
- 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:

Summary:

(~ 1 paragraph)

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

Conclusions:

(~ 2 paragraphs - length will vary depending on the results/findings)

- Here is where you explain what the results really mean, and identify any relevant findings.
- Make sure to touch on global impacts. For example,

- The propensity score analysis showed that people who smoked were 2.465 (p-value = 0.002) times more likely to develop lung cancer than people who did not smoke. Based off this result it appears as though smoking at least 1 pack of cigarettes per day for a prolonged period of time will increase one's likelihood of developing lung cancer by approximately 250%.

Weakness & Next Steps:

(~ 2 paragraphs - length will vary depending on the results/findings)

- This sub-section can be split into two, if needed
- Be careful here, especially if you are simulating data. You can never simulate every single scenario. So you will have some generalizability issues.
- 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.