STA304 - Summer 2021

Assignment 2 Instructions

Samantha-Jo Caetano

Instructions

This is a group assignment. You are expected to work on this in a group of up to for 4. You are expected to work exclusively with your group-mates and not other groups. You are more than welcome to discuss ideas, code, concepts, etc. regarding this assignment with your class mates. Please do not share your code or your written text with peers outside of your group. It is expected that all code and written work should be written by members of your group (unless they are taken from the materials provided in this course or are from a credible source which you have cited). You have the option to work in a group of smaller than 4, but please note that we do not recommend this, as the workload of this assignment is for groups of size 4. Please note, this assignment is fairly open, so the context of most of the work completed here should not match that of other groups.

There is a starter Rmd file (called Assignment2.Rmd) available for you to use to start your code.

Submission Due: Friday May 28th at 11:59pm ET

Your complete .Rmd file that you create for this assignment AND the resulting pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/220076/assignments/613555) by 11:59PM ET, on Friday, May 28th. **If you do NOT submit the pdf in your LATEST submission you will receive a grade of 0.**

Please note that only one group member needs to submit the .Rmd and .pdf files onto Quercus in ONE submission. We will be directly marking on the LATEST submission of the .pdf (submitted on/before the due date/time). All group members will receive the same grade. *If your LATEST submission does not contain a .pdf AND .Rmd then you will receive a 0 on this Assignment.*

Late assignments are NOT accepted. Please consult the course syllabus for other inquiries.

Assignment grading

There is one part to this assignment and it is to produce a report on a data analysis. The report focusses on theory/methodology, data analysis and communication/writing. We recommend you spellcheck and proofread your written work. We will be directly marking the pdf files, thus please ensure that your final submission looks as you want it to look before submitting it.

As mentioned above, this assignment will be marked based on the output in the pdf submission. You must submit both the Rmd and pdf files for this assignment to receive full marks in terms of reproducible. **If you do NOT submit the pdf in your LATEST submission you will receive a grade of 0.**

This assignment will be graded based off the rubric available on the Assignment Quercus page (link: https://q.utoronto.ca/courses/220076/assignments/613555). Please note that only one group member needs to submit the .Rmd and .pdf files onto Quercus in ONE submission. All group members will receive the same

grade. TAs will look over each section and select the appropriate grade for that section based off a coarse overview (one-time read over) of that section. Your assignment should be well understood to the average university level student after reading it once. I would suggest you make sure your document looks clean, aesthetically pleasing, and has been proofread. You will be able to see the rubric grade for each section. There may be some comments/feedback provided (by the TAs) if the same issue seems to be arising in multiple sections, but you will likely receive no comments/feedback (due to the scaling of the class and marking).

Group Work

You are expected to work on this in a group of up to 4. Your group-mates can consist of any students currently enrolled in the class and you can choose how many other members you'd like to work with and who those members are. Please note, that due to the fast-paced nature of this summer course we will not be manually adjusting groups. Two days prior to the assignment deadline we will be locking in the groups and we will NOT be making any changes to groups beyond that point. Additionally, our teaching team will not be working through group dynamics. Please procure your group carefully, and bring questions/comments to our attention as early as possible.

Report

Objective

To predict the overall popular vote of the next Canadian federal election (tentatively 2023) using a regression model with post-stratification.

Please note that there is NO requirement on the type of model you use. You can use a standard model (i.e., simple or multiple regression), a multilevel model or a Bayesian model (standard or multilevel). The model choice is up to you. With that being said, the model should still be appropriate (e.g., logistic regression for binary outcome, or if you assume a prior distribution you should justify the prior in some way).

Description:

In this assignment you will create an "Introduction", "Data", "Model" (or "Methods"), "Reults" and a "Conclusions" section of a report, based on a post-stratification analyses. It is recommended that you use the General Social Survey (GSS) as the "census" data, and data from the CES2019 package as "survey" data.

The idea is, as a small team (of size 1-4) you will work through the following steps:

1. Load in the sample/survey data (CES 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 (GSS data).
4. Calculate $\hat{y}^{PS}$.

General Social Survey (GSS) - Census Data

You will need to grab the GSS data from the CHASS website (I cannot post it for copyright/privacy reasons). Instructions for how to access and load in this data are available in the first 30 lines of the gss_cleaning.R code. Additionally, the gss_cleaning.R document has code that I used to clean the data. You do NOT need to describe the cleaning in this R script in your report, you only need to describe any additional cleaning that YOUR GROUP had done.

CES - Survey Data

Here is a resource for grabbing the CES2019 data: https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-canadian-election-study. There is some code available in the Assignment2.Rmd where I go through selecting and grabbing the CES2019phone data set.

Additionally, Paul and Rohan have some more documentation here https://hodgettsp.github.io/cesR/ that you may find useful.

Report Components

Introduction

The goal of the Introduction section is to introduce the overall "problem" to the reader.

Your **Introduction** section should include the following:

- Describe the data and the problem in 2-3 clear sentences.
- Should introduce the importance of the analysis.
- Get the reader interested/excited about analysis.
- Provide some background/context explaining the global relevance of the problem/data/analysis.
- Introduce terminology and prep the reader for the following sections. For example, here you should explain different political terms if they are niche.
- Introduce research question.
- Introduce any hypotheses (hypotheses should be decided on prior to performing your analysis and should have some mild justification).

Data

The goal of the Data section is to introduce the reader to the data set, showcase some meaningful aspects of the data, and get them thinking about potential hypotheses/findings.

Your **Data** section should include the following:

- A description of the data collection process.
- A summary of the cleaning process (if you cleaned the data). Someone (who is NOT necessarily familiar with Tidyverse functions) should be able to read this section and reproduce your cleaning process based off reading your description.

- A description of the important variables.
- Some appropriate numerical summaries (at minimum center and spread, but something else may be more appropriate). If there are a lot, please put them in a well formatted and labelled/numbered table.
- At least 1 aesthetically pleasing plot/graph/figure (No more than 4 plots).
- Text explaining/highlighting each table or figure.
- Some text (and perhaps graphical summaries) of the variables you will use in your model. This should help prep the reader in understanding why the subsequent analysis is important/interesting and whether it is appropriate.
- In line referencing/text if needed.
- Reference the programming language/software used to complete this section.

Methods

The goal of the Methods section is to introduce the reader to the statistical methods that you will be using to analyze the data.

Your **Methods** section should include the following:

- A complete explanation of what each methodology you are using entails. So a thorough explanation of the regression model and a thorough explanation of poststratification.
- Here you will describe the chosen model (e.g., if you decide to perform linear regression you must write out the mathematical model, with symbols (not numbers) and describe the parameters and variables included) and give some justification for why this model was selected.
- Be sure to explicitly state the model. You should describe the notation used (i.e., β_0 is the intercept term which represents..., etc.).
- Here you will also give an explanantion of the poststratification process. I.e., explaining

$$\hat{y}^{PS}$$

.
- This should include a description of what post-stratification is (in non-statistical language) and a description on why it is useful.

- As part of the poststratification technique you should also describe the cell/bin splits that you will disply/implement in the Results, based on the sample data. Here you should briefly recall the variables that you are using to create the cells (again, the full description of these should be in the Data section). 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).
- Explain any/all assumptions.
- An explanation of the parameters of interest.
- An explanation of the method for a general science reader (i.e., not a statistician).
- A description of why the method is appropriate (based off assumptions, variable types and practical rationale).
- If you want to include some additional analysis (e.g., standard error, post-stratification by state, etc.) then you should describe your methodology here. Additionally, if you do this be sure to include any citations/references that may be needed by the reader.
- In line referencing
- In line R code (if needed).

Results

The goal of the Results section is to present the results of the statistical analyses to the reader.

Your **Results** section should include the following:

- The results of the methodologies included in the report.
- An explanation/interpretation of the results.
- Some commentary on whether or not the results seem reasonable.
- Text explaining/highlighting each table or figure.
- In line referencing.
- In line R code to produce output in text (E.g. The mean is `` `r mean(x)` ``.).

Conclusions

The goal of the Conclusions section is to present the story of your analysis to the reader.

Your **Conclusions** section should include the following:

- A brief recap of the hypotheses, methods, and results.
- State (or re-iterate) your key results.
- State any reasonable conclusions drawn from the results.
- An explanation/interpretation of the results.
- Some commentary on any drawbacks/limitations.
- Recommendations for Next Steps for future analyses/reports.

Bibliography

A well formatted bibliography, including references in a well formatted list. These should have been referred to in the text above.

General Notes:

- All tables/figures should be well labelled and clean.
- Everything should be written in full sentences/paragraphs.

- There should be no evidence that this is a class assignment, I should be able to take a copy of this report and paste it into a newspaper/blog without needing to implement any edits.
- There should be no raw code in the pdf. All output should be nicely formatted/presentable.
- You will also need a reference/bibliography section. You should reference the data, any outside code/documentation and any ideas/concepts that are taken outside of the course.
- Note, we are not marking grammar, but we are looking for clarity. If you need help with writing there are resources posted on the Course Info>Resources page of Quercus. It is important that you communicate in a clear and professional manner. I.e., no slang or emojis should appear.
- Be specific. Remember, the reader/marker may not be familiar with the topic or specifically what your team/grooup did. A good principle is to assume that your audience is not aware of the subject matter.
- Remember to end each section with a concluding sentence. This means reiterating the key points from your writing.