Assignment 2 Instructions

STA304 - Fall 2025

Samantha-Jo Caetano

Instructions

Please read all instructions carefully.

This can be a group assignment. You are expected to work on this either independently or in a group of up to 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, but only share your writing and code with your groupmates. 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 are allowed to use Generative Artificial Intelligence to support your work, but it is recommended that you perform your own proofreading and editing after using any Generative AI. Please read through the "Generative AI" policy on the course syllabus and in the instructions of this assignment to ensure that your usage is inline with the requirements of this assessment.

There is a starter Qmd file (called Assignment2-startercode.qmd) available for you to use to start your code. We suggest you read the entire assignment before starting.

Submission Due: Thursday November 13th at 11:59pm ET

Your submission will consist of three components:

1. .qmd file (submitted as a Group)
2. .pdf file (submitted as a Group)
3. Assignment 2 - Group Work Survey (completed as an individual - even if you worked alone)

Group Work Submission

Your complete .qmd file AND the resulting pdf (i.e., the one you 'Render to PDF' from your .qmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/404574/assignments/1625363) by 11:59PM ET, on November 13th.

Please note that only one group member needs to submit the .qmd and .pdf files onto Quercus in ONE submission. We will be directly marking the LATEST submission of the .pdf (submitted on/before the due date/time). All group members will receive the same grade. We will only be accepting submissions through this Quercus page (i.e., we *not* be accepting email submissions). Please consult the course syllabus for other inquiries.

There are three attempts to submit this assignment, to account for the possibility of an error in your first attempt/submission. If you submit prior to the November 13 11:59pm ET deadline, then we will grade the latest submission that came in prior to November 13 11:59pm ET.

There is a one week grace period available for this assignment, if your group chooses to use the grace period then do NOT submit any documentation until after November 13, 2025. Note: if you use the grace period we will grade the *latest* submission, so please ensure that you are including BOTH the pdf and Rmd/qmd in your upload/submission.

Assignment grading

This assignment is to be a report. The page limit is 10 pages (this does not include the Generative AI Statement, Ethics Statement, the Bibliography nor any Appendices).

In this report you will perform a data analysis and describe your insights/findings. Thus, the assignment requires coding, analysis and written communication. We recommend you spellcheck and proofread your written work.

We will be directly marking the pdf files, so 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 qmd (or Rmd) and pdf files for this assignment to receive full marks in terms of reproduciblity. **If you do NOT submit both the pdf AND qmd in your submission you will receive a 20% grade deduction.**

This assignment will be graded based off the rubric available on the Assignment Quercus page (link: https://q.utoronto.ca/courses/404574/assignments/1625363) - the rubric will be available at least one week in advance of the due date. TAs will look over each section and select the appropriate grade for that section based off a brief overview (one-time read over) of that section. Your assignment should be well understood to the average university level student after reading it once.

We 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 size of the class and limited time for marking).

Assignment 2: Report

Poststratification used to predict election results

Objective

To predict the overall popular vote of the next Canadian federal election (tentatively 2029) using a regression model with post-stratification. Since Canada has a multi-party system, it is recommended for this post-stratification to be done for multiple political parties (e.g., Liberal, Conservative, NDP, etc.).

The model choice is up to you. With that being said, the model should still be appropriate (e.g., logistic regression for binary outcome).

Description:

In this assignment you will create an "Introduction", "Data", "Model" (or "Methods"), "Results" and a "Discussion" section of a report, based on a post-stratification analyses. It is recommended that you use the Canadian Census as the "census" data, and data from the CES2021 package as "survey" data.

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 on the sample data. Note: any model is acceptable, but some justification (either practical or statistical) should be given.
3. Load in the census data (Canadian Census data).
4. Calculate $\hat{y}^{PS}$.

Canadian Census - Census Data

The Canadian Census is collected by Statistics Canada every 5 years. The last collection unit was in 2021. Please note that this data is only for University of Toronto users, this data is *not* to be shared publicly. The data is available on Quercus only, you will need to download it from quercus and add it to the files that you are working in (either locally or on Jupyterhub). Do not post the census data on any public websites. A .csv file of the data and a codebook are provided only through Quercus.

Canadian Election Study - Survey Data

I have included the data and code book in starter code file. There is some documentation here http://www.ces-eec.ca/2021-canadian-election-study/ that you may find useful.

Report Components

1 Introduction (2-4 paragraphs)

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.
- Introduce the importance of the analysis.
- Get the reader interested/excited about analysis.
- Provide some background/context explaining the overall 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 the research question.
- Introduce any hypotheses (hypotheses should be decided on prior to performing your analysis and should have some mild justification).
- Inline referencing.

2 Data (3-5 paragraphs)

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 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.
- 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, numbered and labelled plot/graph/figure (No more than 4 plots).
- Text explaining/highlighting each table or figure.
- In line referencing if needed.
- Reference the programming language/software used to complete this section.

3 Methods (2-4 paragraphs)

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 each methodology you are using. 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).
- Give some justification for why this model was selected.
- Here you will also give an explanantion of the poststratification process. I.e., explaining $\hat{y}^{PS}$.
- This should include a description of what poststratification 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 implement in the Results, based on the sample (and census) data. Here you should briefly recall the variables that you used 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 "province" 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, poststratification by province, 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) (E.g. The mean is `` `r mean(x)` ``.).

4 Results (1-3 paragraphs)

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:

- A well-formatted, numbered, and labelled table showcasing the predicted proportion of the popular vote for the political parties included in your analysis.
- The results of the analyses 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)` ``.).

5 Discussion (3-6 paragraphs)

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

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

- A brief recap of the hypotheses and methods.
- 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.

6 Generative AI or Workflow Statement (2-4 paragraphs)

If you have used generative AI tools (e.g., ChatGPT, other writing assistants) to help write your report, please include a brief reflection on how you used these tools. This should include: what specific tasks you used the AI for, and how you ensured that the final report was your own work and aligned with the assignment's requirements. Please note: The use of AI tools should supplement, not replace, your own critical thinking and analysis. Ensure that you cite and properly attribute any content generated by AI.

If you did not use generative AI tools on this assessment, please include a brief statement outlining your workflow for completing this assignment. This statement should include timelines and a general description of any resources you used.

7 Ethics Statement (1-3 paragraphs)

Explain how you ensured that your analysis is reproducible (e.g., documenting code, using proper statistical methods).

Since the CES 2019 data is publicly available, describe whether or not this the work completed in your report needs Research Ethics Board approval for the report to be made publicly available. Be sure to specifically discuss the privacy of human participants in this study.

8 Bibliography

Provide at least 5 external academic citations, including:

- The 2021 Canadian Federal Election Study.
- References related to R coding used in your analysis.
- Any generative AI tools used for writing or analysis.

9 Appendix (Optional)

Any additional notes/derivations that are supplementary to the report can be added in an appendix. This section will not be directly graded, but may be included for completion-sake.