Assignment 2 Instructions

STA304 - Winter 2025

Samantha-Jo Caetano

Instructions

Please read all instructions carefully.

This is 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 completion of the work, but it is recommended that you perform your own proofreading and editing following the usage of 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 March 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/374323/assignments/1481116) by 11:59PM ET, on March 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 on 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. If you do NOT submit the pdf or do NOT submit the rmd/qmd in your submission you will receive a 20% deduction.

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 March 13 11:59pm ET deadline, then we will grade the latest submission that came in prior to March 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 March 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 in total (this includes the Generative AI Statement, Ethics Statement, but does not include the Bibliography or 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/374323/assignments/1481116) - 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

Stratified Sampling Survey Analysis

Objective

You will write a report that mimics an academic paper using survey data from the 2019 Canadian Federal Election Study (CES) phone OR web data. You are to assume that the data was collected via stratified random sampling, where the stratification variable is either **province, education level, or gender** (you can choose one of these to assume as the stratification variable). The starter code has cleaned up versions of the data, but more documentation and information regarding accessing the entire raw files is available here in the cesR documentation.

Based on this dataset you choose (i.e., phone or web survey), you will:

1. **Select one political party of interest** (Liberal, Conservative, New Democratic Party, etc.).
2. **Calculate the proportion of votes** that the selected party is expected to receive.
3. **Compute a 95% confidence interval** for the proportion of votes that the selected party is expected to win.
4. **Create a logistic regression model** predicting the log odds of voting for the selected party of interest, including (at minimum) the stratification variable and one other numeric variable (i.e., age or some other numeric variable you choose from the CES study).

Deliverables:

You will produce a report (pdf) that is completely digestible to a university level student (who understands what a confidence interval generally is and knows a bit about regression models, but may not be familiar with mathematical theory or R code/output). The restrictions are as follows:

- Maximum page length is 10 pages (for the pdf) (not including the bibliography or any appendices)
- Standard margin sizes (i.e., 2.5cm)
- Standard font sizes (i.e., 12 pt font)
- Any plots, tables, and output are neatly presented and organized.
- No visible code in the pdf.

Note: Any text beyond 10 pages (except the Bibliography & Appendices) will not be read by the grader.

Assignment Components:

1 Introduction (2-5 paragraphs)

In this section you will briefly describe your report. Explain the importance of the subsequent analysis and prepare the reader for what they will read in the subsequent sections. Provide an overview of the research question. Briefly describe the 2019 Canadian Federal Election Study and its relevance. State the purpose of the report.

2 Data (2-5 paragraphs)

Briefly introduce the data and key variables of interest. If you do any general data cleaning or data processing you should describe it (in a reproducible manner) here. Identify the stratification variable used. Include at least one plot displaying the distribution of the stratification variable. If you do any data cleaning or data processing to the you should describe it (in a reproducible manner) in this section.

3 Methods Section (2-5 paragraphs)

Include the formula for calculating the confidence interval for proportions (do not include specific numbers yet) and provide a description of its components. Present the logistic regression model, specifying the independent variables and expected interpretation of coefficients (parameters, not estimates), and describe the model. In this section you are preparing the reader for how to interpret the numbers displayed in the next section (Results).

4 Results (3-6 paragraphs)

Present a **table** showing the estimated proportion of votes for the selected party along with the 95% confidence interval, and include text describing this table and the key takeaways.

Provide a **table or formula** of the estimated logistic regression model, and include text describing this table/formula and the key takeaways. Interpret the estimates from the logistic regression model. Specifically, commenting on how the predictor variables relate to the outcome variable.

5 Discussion (3-5 paragraphs)

Restate the objective. Summarize key findings.

Discuss limitations of the analysis (e.g., potential biases, missing variables, survey errors).

Provide recommendations for future research or improvements.

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 2019 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.