

STA304 - Fall 2023

Assignment 2 Instructions

Emily Somerset & 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 for 4. You have the option to work in a group of smaller than 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 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 after you use 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 matches with the requirements of this assessment.

Please note, this assignment is fairly open, so the context of most of the work completed here should not match other groups.

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

Submission Due: Thursday Nov 16th at 11:59pm ET

Your submission will consist of three components:

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

Group Work Submission

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/course/s/317099/assignments/1172216>) by 11:59PM ET, on November 16th.

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 submission does not contain a .pdf AND .Rmd then you will receive a 0 on this Assignment.*

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 16, 2023. Please note there is an Ouriginal

“Draft” Submission page for your group to check your Original Score and to familiarize yourself with the submission process.

Please only submit your final work to the actual submission page. There will not be multiple attempts awarded for errors in submitting on the wrong page or incomplete submissions. So please be careful and mindful of what you are submitting and use the draft page to practice the submission process.

Individual Submission

Individual Item to Submit

In order to keep a record of group contributions and as a support to report an misconduct of group members there is a survey (link: <https://q.utoronto.ca/courses/317099/quizzes/353284>) . Please complete this survey between Thursday November 16th 12:01am ET and Friday November 24 at 11:59pm E.T. Completion of this survey will be worth 1% of this assignment. The survey should not take more than 5 minutes to complete. So NO LATE submissions will be accepted (i.e., no submissions of the survey beyond November 24 at 11:59pm E.T.). There is no time limit on the survey, but you must submit it before November 24 at 11:59pm ET. You have one attempt/submission of this survey.

This survey must be completed by ALL STUDENTS (even if you are working on Assignment 2 as an individual).

Assignment grading

For this assignment, you will produce a report on a data analysis. The report focuses on theory/methodology, data analysis and communication/writing. We recommend you spell check 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 reproducibility. **If you do NOT submit the pdf in your 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/317099/assignments/1172216>). Please note that only one group member needs to submit. 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 quick 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 size of the class and number of available TA hours).

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 course we will not be manually adjusting groups. We will be locking in the groups on November 16, 2023 and we will NOT be making any changes to groups beyond that point.

All group members will receive the same grade on Assignment 2.

Additionally, our teaching team will not be working through group dynamics (e.g., tardiness, work allocation, etc.) since there is the option to work alone. With that being said, we will not tolerate any forms of harassment, bullying, name-calling, etc. If there is any reports of this, our teaching team will investigate this

and students' grade(s) may be affected. 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 2025) 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 state and justify the prior in some way).

Description:

In this assignment you will create an “Introduction”, “Data”, “Model” (or “Methods”), “Results” 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 2021 Canadian Election Study (CES) as the “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 type of 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, LRTs, AIC, BIC, literature review, data availability etc.)
3. Load in the census data (GSS data).
4. Calculate \hat{y}^{PS} .

General Social Survey (GSS) - Census Data

The GSS data can be accessed by loading the `gss_clean.csv` file. Instructions for how to load in this data are available in the `Assignment2-starter_code.Rmd`. Additionally, the `gss_cleaning.R` document has code that I used to clean the data. You do NOT need to describe the cleaning included in this R script in your report, you only need to describe any additional cleaning that YOUR GROUP had done.

Canadian Election Study (CES) - Survey Data

The 2021 CES survey data can be accessed by loading the `ces2021.RData` file. Instructions for how to load in this data are available in the `Assignment2-starter_code.Rmd`. The 2021 Canadian Election Study Codebook (.pdf) can be used to match survey questions to responses in the CES data. There is some code available in the `Assignment2-starter_code.Rmd` where I go through identifying and interpreting columns in the CES data.

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.
- 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.

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 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.
- If there is missing data in the variables you will use in your model, include some appropriate numerical summaries (at minimum the proportion of missingness in each variable).
- At least 1 aesthetically-pleasing plot/graph/figure (No more than 4 plots).
- Text explaining/highlighting each table or figure.
- Inline referencing 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 each methodology you are using. So a thorough explanation of the regression model and a thorough explanation of poststratification.
- A complete explanation of how you decide to deal with missing data (imputation, deletion) should be given so that your analysis is reproducible.
- Give some justification for how missing data (if any) was dealt with.
- 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 explanation 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 display/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 “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.
- Inline referencing
- Inline 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.
- Inline referencing.
- Inline 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.

Appendix

The goal of the Appendix is to include any secondary information to the reader.

Your **Appendix** section should include the following: - Your Generative Artificial Intelligence (AI) Statement (see below for more information). - Any additional plots or calculations that are not of primary necessity to the report, but should be included for completion-sake.

Your **Generative AI Statement** should include a detailed description of any generative artificial intelligence tools used along with a reference(s)/citation(s). Please include a list of prompts, and explain in detail how the result of the prompts were used in this assignment (which sections of the assignment - if any - were constructed via the usage of generative AI and how). It is expected that this section will likely be about 100-200 words, but maybe longer depending on your usage of different tools.

If you did not use any generative AI then please still include this section and explain that you did not use any generative AI tools. If this is the case, you do not need to include any references/citations to generative AI tools.

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/group 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.
- You are more than welcome to perform a prediction of a different election (e.g., predict the 2024 U.S.A. election or the outcome of the next British Columbia provincial election) in lieu of the next Canadian federal election, just be sure to still perform a regression and poststratification (i.e., create a model on sample data and poststratify on some census data).
- If you end up using other data (i.e., not CES 2019 for sample or not GSS 2016 for census) to perform the task please include your csv files in order for us to assess reproducibility.