STA304 - Fall 2021

Assignment 2 Instructions

Samantha-Jo Caetano

Instructions

This is an individual assignment. You are expected to work on this independently. 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 your peers. It is expected that all code and written work should be written by yourself (unless they are taken from the materials provided in this course or are from a credible source which you have cited). Please note, this assignment is fairly open, so the context of most of the work completed here should not match your peers.

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

Submission Due: Friday October 22nd 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/236142/assignments/723611) by 11:59PM ET, on Friday, October 22, 2021. Late assignments are not accepted. Please consult the course syllabus for other inquiries. **If you do NOT submit the pdf in your LATEST submission you will receive a grade of 0.**

Please note, that this assignment is fairly writing intensive, the university offers many writing resources/supports. Please check out our Course Resources page (here) for resources regarding writing (as well as RMarkdown, R coding, tidyverse, LaTeX, mental health, etc.).

Assignment grading

This assignment will consist of one report-style data analysis. This assignment will has aspects focussed on coding, data analysis and communication/writing of the goals, data, analysis (methods+results) and findings of a statistical analysis. We recommend you spellcheck and proofread your written work. The deliverable will be a report-style pdf (and the Rmd that rendered the pdf). 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 LATEST submission you will receive a grade of 0.**

If you do not knit from Rmd directly to pdf then the document is not reproducible and/or will not be well-formatted, making you ineligible to receive full points in these sections of the Assignment rubric.

This assignment will be graded based off the rubric available on the Assignment Quercus page (link: https://q.utoronto.ca/courses/236142/assignments/723611). 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).

Report

Description:

In this project you will write a report on a regression analysis of observational data, available in the Open Toronto Data Portal. Your main methodology will include one of the following:

- Frequentist Linear Regression (i.e., using `lm()`)
- Frequentist Logistic Regression (i.e., using `glm(..., family="binomial")`)
- Bayesian Linear Regression (i.e., using `brm()`)
- Bayesian Logistic Regression (i.e., using `brm(..., family = bernoulli())`)

There is one resource to finding data:

1. The Toronto Open Data Portal https://open.toronto.ca/. Data from the portal can be downloaded via the opendatatoronto R package, some info available here: https://sharlagelfand.github.io/opendatatoronto/. Additionally, there is a tutorial on how to access data from this portal here: https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-toronto-open-data-portal

Minimum Requirements:

- Sample size must be at least 30.
- Data must have at least 3 variables (see next two points to clarify specific requirements of these variables).
- The dependent variable must be appropriate based on the type of regression you perform (i.e., linear regression should have a numerical/quantitative dependent variable; and logistic regression should have a binary dependent variable).
- Your regression must have at least two independent variables: one must be categorical and one must be numerical.
- Your submission should load in the data directly into R, to ensure the Rmd is reproducible (Please note, there is still an option to load in a csv onto the assignment submission page, but this is left as a last resort. If you do this (i.e., load in the data locally instead of through the R package) you will not be eligible for full points in the reproducibility section of the rubric).

You will find some data, form an interesting question and answer the question through your analysis. Your question should be stated clearly so that the reader can quickly identify it in the introduction (and repeat it, maybe more formally, in the methods section).

The report will consist of 6 sections: Introduction, Data, Methods, Results, Conclusions, and Bibliography.

There should be no evidence that this is a class assignment. The pdf should be a well-formatted report-style document. I should be able to take a screenshot of the pdf and paste it into a newspaper/blog. There should be no raw code or unformatted/code-type output in the pdf. All output, tables, figures, etc. should be nicely formatted. Plots and tables need to be labelled and numbered.

Introduction

The goal of the Introduction section is to introduce the overall "research question" 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 (both statistical and non-statistical) and prep the reader for the following sections.
- Introduce research question. **It might even be a good idea to bold it.**
- Introduce hypotheses (this does not mean H_0, this means in text what do you think/"hypothesize" will happen).
- There should be support (i.e., inline citations) in this section. When writing statements to draw interest, make sure your writing is objective and/or supported.

Data

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

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

- A detailed description of the data collection process (there are resources in the portal to get to know the data - I suggest picking a data set that you can learn about the collection process as much as possible).
- A summary of the cleaning process. Again, one (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, labelled and numbered table.
- At least 1 aesthetically pleasing plot/graph/figure (No more than 4 plots - if you have more than 4 then create an Appendix section at the end of the report). You should probably use a scatterplot at some point.
- Text explaining/highlighting each table or figure.
- Some text (and perhaps graphical summaries) of the variables you will perform the regression on. This should help prep the reader in understanding why the regression 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 **Method** section should include the following:

- A complete explanation of what each methodology you are using entails. So an explanation of the linear or logistic regression.
- The mathematical model (this should have parameters - not estimators/estimates). (E.g., The simple linear regression model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, with some small changes being acceptable, but the idea is you should have Greek letters, not numbers, in this section).
- An explanation of the model for a general science reader (i.e., not a statistician).

- A description of why the model is appropriate (based off assumptions, variable types and practical rationale).
- Explain/justify any assumptions.
- An explanation of the parameters of interest (i.e., What does β represent? How should the reader interpret the numbers presented in the next section?).
- You will likely use some model selection technique (or a combination), you should explain your model selection process/criteria in this section. Please note, there are lots of (correct) ways to select your model, some include: **p-value/significance cut-off** (i.e., strict forward/backward selection), **coefficient of determination** (i.e.,R^2 or $R^2_{adjusted}$), **practical rationale** (e.g., I am interested in if age correlates to bedtime, so age must stay in the model), **likelihood ratio tests** (LRTs), **Akaike Information Criteria** (AIC), **Bayesian Information Criteria** (BIC), etc. This is not an exhaustive list. It is expected that the first 3 (i.e., p-values, R^2, and practical rationale) have been covered in this class or a pre-requisite course to some degree. If you use any other model selection technique (e.g., you use LRT, AIC, BIC, etc.) which is not covered in this course or your pre-requisite course then you must explain the process and include a citation. In all cases, you should be explaining the process, and it's recommended to still include a citation, regardless.
- 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:

- A well formatted and labelled table containing the numerical output of the model (i.e., the estimates (and p-values/standard errors) from the linear/logistic regression model).
- An explanation/interpretation of the results.
- You may want to comment on the implementation of the model selection process. (For example, in the Methods if you used p-values for selection, then explain that the p-value of was and it was retained in the model based off it being below some prespecified threshold, as outlined in the Methods section.). You don't need to spend a lot of text on this, but if it flows well in your report/writing then I would just include it as it will tie in nicely with the Methods section.
- Some commentary on whether or not the results seem reasonable (based off the original plots).
- It is recommended to include a plot with the overlaid regression line atop, but this is not required.
- 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) ` ``.).
- Reference any packages and any software used to complete this section.

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 research question/goal, 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 well-thought out commentary on any drawbacks/limitations/weaknesses. (I.e., if you were to hand off this assignment to a friend/colleague what would you tell them to work on/look out for?)
- Recommendations for Next Steps for future analyses/reports. (I.e., again, if you were to hand off this assignment to a friend/colleague what would you tell them to work on/look out for?)

Bibliography

Your **Bibliography** section should include a list of well-formatted citations. No specific reference style is required, but you must use some formal reference style in order to ensure that the references are accessible.

General Notes:

- As a reminder, here is a tutorial on grabbing Toronto Open Data https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-toronto-open-data-portal.
- All tables/figures should be well labelled, numbered and clean (i.e., should not have coded names like `x_1`).
- Everything in the report should be written in full sentences/paragraphs.
- There should be no raw R code or code-like output in the pdf. Any output should be nicely formatted and legible to a general science student (who may not be able to read/understand code).
- 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. Your bibliography should have well over 5 citations, beyond the starter code (this is not a strict lower-bound cutoff, but if you are looking up anything you should cite it, and I imagine you will look on at least 5 different sites/texts/books/papers to support the data, research question, methods and develop your code).
- 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, you are selecting this topic/data and the reader/marker may not be familiar with it. 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.