STA238 - Winter 2021

Final Project Instructions

Samantha-Jo Caetano

Rough Draft Due: Monday April 5 at 11:59pm ET

Peer Review Due: Friday April 9 at 11:59pm ET

Final Submission Due: Friday April 16th at 11:59pm ET

Instructions

This is an individual project. You are expected to work on this independently. You are more than welcome to discuss ideas, code, concepts, etc. regarding this assessment 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 project is fairly open, so the context of most of the work completed here should not match your peers.

Rough Draft Due: Monday April 5 at 11:59pm ET

Your submission will consist of one component:

1. .pdf file

Submission of Rough Draft

As an individual, via Quercus, submit a PDF of your rough draft on Quercus (link: https://q.utoronto.ca/courses/204754/assignments/581741) by 11:59pm ET on Monday, April 5, 2021. (Peer reviews start on Tuesday April 6 at 12:01am ET.) This submission is only a pdf, it is recommended you generate it from an Rmd, but it is not a requirement for this part of the project.

At a minimum this Rough Draft must include:

- your title (this must be descriptive)
- At least 4 sentences introducing your Data.
- At least 2 sentences stating the research question and relaying to the reader why it is interesting/useful (related to the data described above).
- At least 1 sentence describing the how you will use at least 2 of the methodologies in this class to answer your research question.
- Minimum 100 words (but it is recommended that you include more).
- All text is in full sentences.

You will be awarded 2% for completion of the total 30% for the Final Project. It is recommended that you also include the (partially completed) reference section here, but this is optional. You You do not need to include your name in the pdf if you prefer to stay anonymous to your peers. The point of this is to get feedback on your work (and to make sure you have at least started thinking about this by April 2nd) so you are more than welcome to include other sections that you wish to get feedback on. If you do NOT submit a Rough Draft/Proposal, then you will not be able to participate in the peer review process, and thus this 5% of your final grade (2% proposal + 3% peer review) will go onto the Final Report.

There is no required format for the Rough Draft, other than what is stated above. Thus, you can just submit a pdf with a title and a paragraph (with at least 100 words addressing the points above) and you will receive the completion points for the Rough Draft. The requirements are quite minimal for this Rough Draft, we just want you to start thinking about this project as early as possible.

Further recommendations:

- do NOT include your name in the pdf (if you want to remain anonymous to your peers).
- Include more than 100 words to help yourself get started.
- Include inline citations.
- Include other sections that are partially completed/written up.
- Include graphs/figures and tables.
- Include bibliography.
- Include anything that you think might be helpful to receive feedback on! :)
- The pdf for the Rough Draft does not need to be generated from an Rmd file. So if you are running low on time, then just quickly write something in word, convert it to pdf and submit that.

Disclaimer: There will be no extensions granted for this Rough Draft submission since the following submission is dependent on this date.

Peer Review Due: Friday April 9 at 11:59pm ET

Your submission will consist of three components:

1. comments/feedback on first peer
2. comments/feedback on second peer
3. comments/feedback on third peer

Submission of Peer Reviews

As an individual, on April 6, you will randomly be assigned three rough drafts (submitted by other students) to provide feedback on. You have until April 9, 2021 11:59pm ET to provide feedback to your peers. If you provide feedback to one peer you will receive 1%, if you provide feedback to two peers you will receive 2% if you provide feedback to three (or more) peers you will receive the full 3%.

Providing feedback is obviously subjective, so we have established a set of minimum requirements (for each peer review - thus you'll need to do this three times to receive the full 3%):

- Your feedback must include at least 5 comments (meaningful/useful bullet points).
- One comment that states something you really liked about the Rough Draft.
- One comment on the readability of the writing (I.e. address any edits, grammar, typos, etc.)
- One comment on how interesting/compelling the writing and potential analysis is.
- One comment/suggestion a foreseeable analysis, weakness, next step, data, reference, etc. (Just give them something useful to work off of.)
- One follow-up question you have for the writer.

Disclaimer: There will be no extensions granted for this submission since the following submission is dependent on this date.

Disclaimer: Please remember that you are providing feedback here. All comments should be professional and kind. It is challenging to receive criticism, and arguably more challenging to provide criticism, and even more challenging to give criticism strictly through text. Please remember that your goal here is to help your peers advance their writing/analysis. Any feedback that is inappropriate will receive a 0 on this section.

Final Submission Due: Friday April 16th at 11:59pm ET

Your submission will consist of two components:

1. .Rmd file
2. .pdf file

Submission

As an indiviual you will submit one .Rmd file that you created for this **project 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/204754/assignments/581740) by 11:59PM ET, on Friday, April 16th.

We will be directly marking on the LATEST submission of the .pdf (submitted on/before the due date/time). *If your LATEST submission does not contain a .pdf AND an .Rmd then you will receive a 0 on this Project.* There will be a short grace period instilled (~1 extra hour) to account for technically difficulties. Anything submitted after the grace period will not be accepted. **Late projects are NOT accepted.** Furthermore, all submissions must be made onto quercus. **Email submissions are NOT accepted.** Please consult the course syllabus for other inquiries regarding extensions.

Project grading

There are three parts to this project. You must complete all three parts to be considered for the full 30%. For instance, if you do NOT submit a Final Report you will not receive the completion points from the rough draft and peer review process.

As mentioned above, this project will be marked based on the output in the pdf submission. You must submit both the Rmd and pdf files for this project to receive full marks in terms of reproducibility. Furthermore, this is an individual project. You are expected to work individually. The workload level is higher than that of an assignment, since this is a project. Thus, it is recommended that you start early.

This project will be graded based off the rubric available on the Assignment Quercus page (link: https://q.utoronto.ca/courses/204754/assignments/581740). TAs will look over each section (on the submitted pdf) and select the appropriate grade for that section based off a coarse overview (one-time read over) of that section (of the pdf). Your project should be well understood to the average university level student after reading it once. I would suggest you make sure your (pdf) document looks clean, aesthetically pleasing, and has been proofread. Since this is a final project, the process to review your grade on this assessment is handled by the Department of Statistical Sciences. Thus, you will need to apply through a process (TBA at a later date) to see the graded rubric and potentially inquire about a regrading. 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).

Description:

In this project you will write a report on a data analysis in which your main methodology will comprise of a collection of techniques taught in STA238 Winter 2021. The methodology must include the following:

- at least one simple linear regression;
- at least one confidence interval (either through a bootstrap or the Z/t approach);
- at least one maximum likelihood estimator derivation (I would recommend putting the mathematics in the Appendix);
- at least one hypothesis test of the mean;
- at least one goodness of fit test;
- at least one Bayesian credible interval. (Put derivations of the posterior into the Appendix).

Please keep in mind that this analysis is for our course. Thus the analysis should be to **answer a question about an underlying random process we have data from**. 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 repeated maybe more formally as a hypothesis test in the methods section).

The report will consist of 8 sections: Abstract, Introduction, Data, Methods, Results, Conclusions, Bibliography and Appendix.

There should be no evidence that this is a class project, I should be able to take a screenshot of this and paste it into a newspaper/blog. There should be no raw code. All output, tables, figures, etc. should be nicely formatted.

This will allow you to look at some interesting aspects of the data. Please find some open source data through any R package that has not been used on a previous assignment in this course. Some examples of R packages with data that we have used in this course are `dplyr`, `nycflights13`, etc. Here is a list of R packages available: https://cran.r-project.org/web/packages/available_packages_by_name.html. Additionally, if you prefer to use some other data available through a website (e.g., kaggle, github, etc.) that is also an option so long as the data is open, free and ethically viable for you to analyze. If you are unsure about whether your data is appropriate please visit one of our office hours and we will be happy to discuss.

Based off the above criteria, the following three packages are OFF LIMITS. You CAN NOT use data from any of the following sources: The Toronto Open Data Portal, survey data from the 2019 Canadian Election Study, Statistics Canada Portal.

Make sure that the data is appropriate for your methods. Pick something that is interesting to investigate and has variables appropriate for the methodology you are going to perform. Again, the analysis should be to **answer a question about an underlying random process we have data from**. Please visit office hours, post on Piazza or email our teaching team at sta238@utoronto.ca for clarification on appropriate data.

If you use data from the Statistics Canada Portal, CES, or Open Toronto Data Portal you will receive a 0 on this project. (I.e., do NOT use data from open Toronto data on this project; do NOT use data from CES on this project; do NOT use data from Stats Canada data on this project)

The material and text on this project should be different from that of your previous assignments in this course. Thus, you should NOT directly copy your previous assignment work. We highly encourage you use feedback from previous assignments to amend/proofread/update your Final Project. If your work is a direct copy of a previous submission or is a direct copy of another person's submission this is considered an academic offense.

Abstract

The goal of the abstract is to provide the reader with a summary of the report.

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

- One or two sentences describing the introduction.
- One or two sentences describing the data.
- One or two sentences describing the methods.
- One or two sentences describing the results.
- One or two sentences describing the conclusions.

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.
- Introduce research question.
- Introduce hypotheses.

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).
- 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 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 perform the bootstrap on (don't do the bootstrap here - just prep the reader for what is coming in later sections). This should help prep the reader in understanding why the CI 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
- Explain any assumptions.
- An explanation of the parameters of interest (i.e., mean and variance/percentile/etc).
- Any rigorous mathematical computations (i.e., the MLE derivation or the posterior derivation) should go into the Appendix.

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 biblopgraphy, 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 supplementary, non-primary information.

Your appendix should include:

- the MLE derivation.
- the Bayesian posterior distribution derivation.

General Notes:

- A standard report would normally not include this much variation in methodology. It is asked here since we want you to display your understanding of the course material.
- You are allowed to change your data, methods, analyses, etc. from what was in your draft. We recommend using the feedback from the draft to edit your final report, but it is not mandatory that you stick to the original proposal.
- Again, this analysis is for our course. The analysis should be to **answer a question about an underlying random process we have data from**. Hence the question cannot be "what is the mean of the dataset I collected".

- Your question should be stated clearly so that the reader can quickly identify it in the introduction (and repeated maybe more formally as a hypothesis test, or some other methodology, in the methods section).
- It is expected that you include at minimum the required methodology in your report, but you can include more (i.e., maybe you want to look at more variables).
- All tables/figures should be well labelled and clean.
- Everything in this project should be written in full sentences/paragraphs.
- There should be no evidence that is a class project, I should be able to take a screenshot of this section and paste it into a newspaper/blog.
- There should be no raw code. Any output should be nicely formatted.
- You will also need a reference 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.
- Use full sentences.
- Be specific. Remember, you are selecting this 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 with a conclusion. This means reiterating the key points from your writing.