Stat 21 Final Project

Swarthmore College

Due: Dec 14th

Contents

This assignment is due on to be submitted on Gradescope on **December 14**. I also ask that you please submit your final projet to the `final-project` channel on Slack so that you have the chance to share and discuss your work with your classmates. On both Slack and Gradescope, please submit your project as a single **PDF** document.

There are three broad options for you to choose for your final project. Each of these three options has their own statistical merit *and* each option is customizable so that you can spend time researching something that is important to you. My hope is that this project is more enjoyable than laborious. For each of the final project options I've provided a guide on what to include in your report. You may type your final report in Word or OpenOffice or another editor (or for option 2 you may want to use R Markdown to type your report).

To make this project go as smoothly as possible for everyone, please make sure you have okayed your final project idea with me before Monday, Dec 7th. (This form will be posted on Moodle on Nov 30th.)

Option 1

Present a critique of a published paper that presents either an ANOVA model or a MLR model. The objectives for this option are for you to gain experience critically reading and analyzing scientific articles, apply what you have learned in class to understand a published statistical analysis, and to practice communicating statistical information.

One way you can find articles that are of interest to you, is by going to https://www.jstor.org/ and in the search bar enter "linear regression" (for example) in quotation marks. Then refine your search on the left hand side to filter only by "Journals" and then filter again by the subject area of your choice. This is an opportunity to further your knowledge in any area while simultaneously practicing statistical reasoning.

What should the project include?

1. A explanation of the problem or purpose presented in the paper. What problems or questions did the researcher set out to investigate?
2. Background or literature review. How did the researcher build their statistical model? How was the data collected and why were these variables chosen?
3. Methodology. What model did the researcher use and why? What software did they use? What kind of statistical summaries and graphical techniques did they use?
4. Results and conclusions. What conclusions did the researchers draw from their analysis? How do you interpret the results of their model?

5. Discussion and critique. What did you learn from this analysis? Were there any weaknesses to the methods used? Do all the necessary assumptions seem reasonable? Do you agree with the conclusion or do you think there is a better way to address the research question?

6. References. You must cite your sources throughout your report and list them all at the end of your document. (APA or MLA format is fine.)

Option 2

Find a data set and build your own statistical model: ANOVA, SLR, or MLR. The objectives for this option are for you to develop an appreciation for the difficulties of sharing and interpreting raw data, to practice building a statistical model from scratch, and to practice communicating statistical information.

If you do this option I ask that you limit your analysis to estimation and data exploration (rather than inference). It is incredibly difficult, time consuming, and costly to obtain a random sample of data to use for inferential conclusions. This is really only verifiable with experiments but experimental data is often kept private. It can be really fun however to find a data set on a topic that interests you so I want to keep this option open to you. My advice to you is this: keep the model simple (you don't need to have a bunch of predictors for example) and supplement your estimated model with many plots!

Alternatively, if you find census data, that is, if you find data for an entire population, you can design a randomization procedure to collect a representative sample from this census and then draw inferential conclusions about the population. If you do this however, you should discuss the relative merits of your inferential conclusions about the population versus just studying the population directly.

Here are some useful links for finding publicly available data sets

- https://dataverse.harvard.edu/dataverse/harvard/
- https://www.census.gov/data/academy/data-gems/2018/api.html
- https://opendataphilly.org/

What should the project include?

1. A explanation of the problem or purpose. What problems or questions did you set out to investigate? What are the key issues raised? How were the data collected?

2. Background or literature review. What sources or background readings did you consult? What literature exists on your topic?

3. Methodology. What did you do, and how did you do it? What statistical and graphical techniques did you use?

4. Results and conclusions, the summary and presentation of your data analyses. What did you find out? This might include tables, graphs, or verbal summaries.

5. Discussion and critique. What did you learn about the problem or question you set out to investigate? What were weaknesses and strengths of your analysis? If you had more time or resources, how could your project be improved?

6. References. You must cite your sources throughout your report and list them all at the end of your document. (APA or MLA format is fine.)

Option 3

Write a report reflecting on the content of a presentation from the #BlackinStats virtual conference. (You should have attended at least two different talks so that you have a choice on which talk to present.) The objectives for this option are for you to build an understanding and appreciation for the work of Black statisticians and data scientists, to familiarize yourself with the work of the presenters, and to reflect on the content of the presentations in the context of what you know about the fields of statistics and data science at large.

You may find the Google Scholar search engine and LinkedIn to be useful with this option.

What should the project include?

For your analysis of the presentation/work you must include:

1. A survey of the presenter's work. Find other published or presented work by the speaker. Read the abstracts of their most recent publications. How would you describe their work? What does their research or professional work entail? What kind of problems do they solve and what kind of statistical methods do they use?

2. A summary of the talk. What was the purpose of the talk? What was the main message? How did the topic compare to what we learned in class this semester? What kind of statistical methods did the talk discuss?

3. Discussion and reflection. What did you learn from this presentation? Did the talk inspire any new questions or ideas? Was there anything in the talk that you did not understand? Anything you wish we could have covered in class?

4. References. You must cite your sources throughout your report and list them all at the end of your document. (APA or MLA format is fine.)