STA4241 Final Project, Fall 2021

Your final project will consist of two parts. The first will be a write-up of your findings, and the second will be a presentation made either in class or via Zoom (TBD) at the end of the semester. 75% of your grade will be based on the written portion of your project while the remaining 25% will be based on your presentation. You are allowed to either work alone or in pairs of two, though the expectation is that you must do a more in-depth project if you work in pairs. Specifically, the write-up for individual projects must be at least 4 pages single spaced, while it must be 6 pages single spaced for groups of two. These numbers include any figures or tables that you might include in your writeup. These are not hard rules on the length of the paper. If you go slightly under or over, that's fine as long you sufficiently address the question that your project aims to answer. Every write-up will be different depending on what you decide to do and therefore you can structure your paper however you like. However, if you're looking for a standard structure to follow that applies for many of the projects, you could write a paper that includes at least four sections: An introduction that describes the problem and data set (if there is one), a methodology section that describes the types of approaches you will consider and why, a results section that concisely illustrates the important findings, and a discussion section that summarizes and interprets your findings. Again, there is no need to follow this structure, and some projects may not fit well in this structure, so feel free to do whatever you find appropriate. The presentation will be 10 minutes for individuals and 15 minutes for groups of two, and I will ask you questions about your project for about 5 minutes after you complete the presentation.

Important deadlines:

(1) November 18th: By this date you must contact me to explain what your project will be about so I can confirm that it is appropriate and provide feedback on your plan. You can do this either via email or by coming to office hours / class anytime between now and then and explaining it to me in person. You must also tell me at this point whether you will be working in a group of two and who you are working with.

(2) December 6th-8th: These are the dates that you will be giving your oral presentation. We will set up time slots to give presentations as we get closer to this date, but expect to be giving a presentation at some point during this interval. I am flexible with times, so if these dates don't work for you, we can arrange to do it slightly before or after these dates.

(3) December 8th (11:59pm): You must submit your write-up to Canvas by this date.

What does a project entail?

I am open to you doing a project on just about anything that relates to topics we have discussed in this class. This can be an applied data analysis of an interesting data set, it could be a detailed review and simulation study that further investigates the performance of some of the approaches considered in the class, it could involve you reading papers on a specific topic we covered in class and learning more about its theoretical properties and reviewing them, or it could be anything else you can think of as long as I confirm it is sufficiently related to the course material. I have provided a few ideas below for projects that you are more than welcome to use directly, but I am more than happy to see your own project ideas.

Potential project ideas:

(1) Perform a review and in-depth simulation study of all of the high-dimensional statistical approaches we have seen. This should review and describe these approaches in detail and discuss the pros/cons of each of them. Then an in-depth simulation study can be performed to highlight situations in which the various approaches perform well. This would be substantially more detailed than the simulation study from the midterm. Finally, you can interpret and discuss your findings and the implications they have for future analyses.

(2) I will provide a data set on the course website called NHANES.csv, which contains an outcome Y, environmental exposures X, and additional covariates C. Of scientific interest is whether or not any of the environmental exposures are associated with the outcome conditional on C. Additionally, of interest is whether or not any of the pairs of exposures have an interaction effect on the outcome. An additional concern is that some of these relationships might be nonlinear. For this project you will need to analyze this data set and identify whether there are any important environmental exposures and whether there appears to be interactions among the exposures.

(3) Do an in-depth analysis of any other data set that you're interested in. R has many publicly available data sets that could be useful here. For instance, the MICE data set, which we have seen in class has 83 outcomes and 145 covariates for only 60 samples. In the previous semester, many students found very interesting data sets from places such as Spotify, the stock market, and IMDB among others. This project would involve 1) Introducing and describing the data set, 2) Performing exploratory data analysis, 3) Analyzing the data using a variety of the approaches we have seen in class, and 4) Interpreting your findings. A good write-up for a project such as this would explain every decision made and justify the choice of models used for analysis.

(4) (Hard) Dig deeper into one of the approaches that we have studied in this class and discuss what you find. This will likely amount to searching the literature for papers related to the approach, choosing 2-3 of them, reading them, and summarizing your findings and what you learned beyond what we have seen in class. For instance, lasso has had countless papers written about it discussing a variety of implementation or theoretical aspects of the estimator. Additionally, there is a more advanced version of our textbook called *The elements of statistical learning*, which is written by many of the same authors and goes into more detail about the respective approaches, and this can point you in the right direction.

(5) Learn about a statistical approach that we haven't covered in the class, but is related to the ones we have seen in the class. Your write-up would then amount to being a document that is intended to teach this approach to someone who has not seen it previously. It should provide mathematical details for the approach along with examples and figures implementing the approach. Example topics that could be studied in detail would be:

- Unsupervised clustering approaches. This is covered in chapter 10 in the book, so you would have to go into more detail than simply reviewing the book chapter
- (Hard) High-dimensional regression approaches that allow for nonlinear relationships between the covariates and outcome
- (Harder) Sparse versions of PCA, LDA, or PLS
- Anything else you're interested in