**Data Analysis Project**

BIOST311: Regression Methods in the Health Sciences

Nina Galanter

Spring 2023

# Overview

The course project is a data analysis, consisting of six parts: (1) a project proposal; (2) a statistical analysis plan; (3) a draft report; (4) a peer review; (5) a slide presentation, and (6) a final report. I want you to implement components of the analysis as you learn new concepts in both class and discussion section. My goal in having multiple parts to the project, due at different points throughout the quarter, is to provide feedback (and have your classmates provide feedback) multiple times so that your final report is as polished as possible. With that in mind, please do the best job you can on each component—the better the document you hand in, the better comments we can give you!

There are multiple valid methods to choose from for this project. You may choose to analyze the data differently than I would. You should spend your time choosing a method that you can justify, implement, and interpret correctly.

Each of the following components of the project has its own guidelines, each listed in separate documents.

The project proposal (due April 14 at 11:59 pm)

This component of the project addresses two questions:

* What dataset are you using?
* What scientific questions are you answering?

I encourage you to find a dataset that interests you—since being excited about the data will make you more excited about analyzing it—but I have a few criteria that each dataset must meet.

1. The dataset must have more than five variables

2. The dataset must have an outcome variable that is continuous, binary, or time-to-event

3. You must be able to pose at least two interesting scientific questions (judged by me) that you can answer using the data

Some of you may already be involved in research—If you can get permission to use these data, I highly encourage it!

If you don't have a research project, a couple of options for finding datasets are

* [University of California at Irvine Machine Learning Repository](https://archive.ics.uci.edu/ml/index.php)
* [FiveThirtyEight](https://data.fivethirtyeight.com/)
* [Data and Story Library](https://dasl.datadescription.com/)
* [JSE Data Archive](https://jse.amstat.org/jse_data_archive.htm)
* [Kaggle competition datasets](https://www.kaggle.com/datasets)
* [Larry Winner’s Miscellaneous Datasets](https://users.stat.ufl.edu/~winner/datasets.html)
* [List of Data Resources](https://stat2labs.sites.grinnell.edu/DataResources.html)

I also have a handful of datasets that you can use if you're unable to find one that interests you. Please ask if this is the case, and I will provide a dataset that you (hopefully) find interesting!

I will provide you with an R Markdown template for this component of the project.

The statistical analysis plan (due May 5 at 11:59 pm)

In a statistical analysis plan (SAP), you write down explicitly all the statistical analyses (e.g., descriptive statistics, plots and tables, and inferential analyses) that you will conduct using the data. See the guidelines for the SAP for more information about the required elements.

An additional component of the statistical analysis plan is to run one preliminary analysis on your data, using one of the techniques we have learned in class or discussion section.

I will provide you with an R Markdown template for this component of the project.

The first draft (due May 19 at 11:59 pm)

The final report is due on June 8, but I think you should get started early, and I want you to have a high-quality report by the end of the quarter. With that in mind, I’m asking that you hand in a first draft with time for us to read your drafts and hand back comments, and for your classmates to complete the peer review.

I will provide guidelines on the components of the report in a separate document but will not provide you with a template file.

Peer review(due May 26 at 11:59 pm)

You will be paired with one of your classmates to complete a peer review of the first draft. You will read your partner’s draft and provide comments and critiques regarding the statistical analysis, writing, and overall effectiveness.

See the peer review document for more information. I will provide you with an R Markdown template for this component of the project.

Presentation (due May 31 or June 2 at noon)

You will give a 7 minute (without questions) slide presentation on your project on either May 31 or June 2. Presentation order will be determined later in the quarter. A final copy of your slides is due at noon on the day you will be presenting. I will provide you with an optional powerpoint template for this component of the project, but you may use any organization you think will be effective and any program (powerpoint, R markdown, beamer, etc) to create your slides.

You will also think of 2 questions to ask about your peer review partner’s presentation.

The final report (due June 8 at 11:59 pm)

You will write up a full account of your statistical analysis in this report, including: your scientific questions; your statistical hypotheses; your statistical methods, including for descriptive statistics, figures and tables, and inferential analyses; your results; and your statistical and scientific conclusions. All of this will be written in a style suitable to be included in a manuscript.

By turning in a draft earlier in the quarter, I hope that you will have enough time to incorporate feedback before handing in the final version.