

STAT444/STAT844/CM764: Final Project

Spring 2024

Overview

The final project in this class is to write a small applied statistics paper. You will choose a data set, analyze it using methods from this course and/or from the course textbook, and write it up in the style of the journal *The Annals of Applied Statistics*.

Each group will use the same dataset. Groups will prepare and submit one project proposal per group, and will present together. Individuals will each write their own report.

It is expected that group members' reports have very similar analyses, and may fit the same models and make conclusions based on the same summaries. All writing and analysis will be done individually: reports **should not share text**, and all model fitting, numeric summaries, and plots should be **created individually**, even if all group members end up fitting the same model and producing plots of the same quantities.

Formal Requirements

1. **Dataset.** Find a dataset with a continuous response and at least two continuous covariates. Here are some suggested resources for finding datasets:
 - *Elements of Statistical Learning* datasets: .
 - *Generalized Additive Models* datasets: 'R' package 'gamair' , available on 'CRAN'.
 - *Statistical Models* by A. C. Davison datasets: 'R' package 'SMPracticals', available on 'CRAN'.

Don't overthink the choice of dataset: it needs to be just interesting enough for you to write a paper on, but not so challenging that you can't make progress and stick to the page limit. A dataset from a textbook is totally fine. I prefer a simple analysis done well to an overly complicated analysis done poorly.

2. **Proposal:** each group will prepare a project proposal. The purpose of this document is to sketch out what you're planning on doing. This forces you to start the project, and gives me a chance to correct your path if I think what you're proposing is too easy to get full marks or too hard to complete. Requirements:
 - Maximum 2 pages in the AoAS LaTeX template (see below),

- Cite your **dataset** and demonstrate that it meets the requirements above: show a brief exploratory data analysis that details the response variable of interest and any interesting covariates.
- Describe the underlying research question that you will address in your analysis. This should come from the source where you got your data: why were these data collected and what did the original scientists want to learn? You should go and read the original cited source of the data, not just their description in the book, and provide a brief summary in your own words.
- Give a brief **plan** of how you will analyze the data to answer the research question. Do not actually *do* any analysis (that would just be doing the project!), but describe what kind(s) of exploratory data analyses and model(s) you will try. Show that you have started to think about this.

The purpose of the proposal stage is to demonstrate that you have a feasible project.

3. **Paper:** each *individual* will submit a final paper containing their full written analysis and conclusions. Requirements:

- Maximum length of 6 pages in the AoAS LaTeX template (see below). Inclusive of figures, tables, references, and appendices. This is **hard** but realistic of academic writing and industry communication: you have to be concise.
- **Structure:** your report should have the following sections:
 - *Introduction:* describe the research question you will address and how you will address it.
 - *Data:* describe your data source. How were the data collected, how are they structured? Perform a brief exploratory analysis that articulates how your data relate to the research question, and foreshadows what you expect to find in your modelling.
 - *Methods:* describe the model(s) you will fit in **full technical detail**. Be concise but thorough, using clear and complete mathematical notation. The only **requirement** is that you must use at least one model or method that we learned in class.
 - *Results:* Give the results of fitting your model(s) to your data. Explain the relevance of your analysis to the research question. Check model assumptions and explain why they appear satisfied, or if not satisfied how this might affect the interpretation of the results.
 - *Conclusions:* make a final conclusion about the original research question, and substantiate your conclusion using your results. Discuss limitations of the analysis, and what future work you might do.

4. **Presentation:** each *group* will give a 5 minute presentation about their project, covering the elements described under **paper** above. Maximum 3 slides in LaTeX Beamer format– no fancy stuff like PowerPoint or Keynote. There will be a 2 minute question period following each group’s presentation. Time limit will be strictly enforced: you will be cut off at 5 minutes.

5. **Peer Assessment:** you will provide feedback to your peers using the project rubric, with comments added, during their presentations. Requirements:

- Score each presentation using the same project rubric that we use to grade your projects.
- Provide brief, constructive comments: name one thing you thought they did well, and one thing that they can improve.

You must attend both presentation days and submit feedback for each presentation to get full marks for your presentation and peer assessments.

Due Dates

Here are the due dates for each component:

Component	Due Date	Notes
Proposal	June 13, 11:59PM	Submit on Crowdmark, one per group
Paper	August 2, 11:59PM	Submit on Crowdmark, one per person.
Presentations	July 23/25	In class
Peer Assessment	July 23/25	During presentations, in class

I cannot accept late papers: because this assessment is due during the final exam period, I am not permitted to accept late papers.

Formatting

Your proposal and paper **must** be formatted in the LaTeX template for the journal *The Annals of Applied Statistics*. This is the same template used for all your assignments so far, so you should be familiar by now. Link: <https://vtex-soft.github.io/texsupport.ims-aoas/> Assessments not in the proper format will be heavily penalized according to the rubric. For full marks, you must adhere to the AoAS style guide, as though you were to submit this to that journal: <https://imstat.org/journals-and-publications/annals-of-applied-statistics/annals-of-applied-statistics-manuscript-submission/>. The style guide is more suggestive than prescriptive, which makes this harder than it could be, so you'll have to put some work in to this.

STAT844

Grad students will complete the project as described, with the following modifications:

- Students will complete the project **individually**.
- Your models/methods must include at least one advanced topic that goes a bit beyond what is covered in class.

These modifications are meant to encourage you to take on a more interesting project. You are encouraged to (a) discuss your project with me directly, early on in the class, and (b) use data and questions from your research, if at all possible. I am comfortable encouraging students to try something more risky and interesting, as long as we discuss it first.