STAT 230A Final Project[*]

Spring 2023

The goal of the final project is to apply what you've learned in this course to conduct a statistical analysis. It should be an in-depth regression analysis of a question that interests you. This question may come from one of your other courses, your research interests, your future career interests, etc.

You may work either individually or in pairs to complete the project. It consists of three primary delivarables:

- Project proposal (7% of course grade): due Friday, April 7 at 10:00 PM.
- Final project report (30% of course grade): due Thursday, May 12 at 10:00 PM.
- Self-evaluation (1% of course grade): due Friday, May 13 at 10:00 PM.

1 Data

It is best to start with the question of interest and find relevant data second. As you're looking for data, keep in mind that your regression analysis must be done in R. Once you find a data set, you should make sure you are able to load it into R, especially if it is in a format we haven't used in class before. If you're having trouble loading your data set into R, ask for help as soon as possible, so you can make any necessary adjustments before the project proposal is due.

In order for you to have the greatest chance of success with this project it is important that you choose a manageable dataset. This means that the data should be readily accessible and sufficiently large to permit interesting use of non-trivial techniques from the course. In general datasets must have at least 100 observations and at least 10 variables (exceptions can be made but you must speak with me first).

Please do not reuse datasets used in examples/homework/labs in class.

2 Data resources

- Duke University provides a guide to finding datasets for projects like this, with links to several possible data sources.
- Google Dataset Search allows you to explore 25 million datasets from various sources using search filters.
- The fivethirtyeight package in R contains 128 datasets on topics including politics, sports analytics, and popular culture.
- TidyTuesday publishes a new dataset every week, as well as a list of links to data portals from news organizations.

[*] This document is adapted from materials created originally by Prof. Elizabeth Purdom at UC Berkeley and Prof. Maria Tackett at Duke University.

- The UC Irvine Machine Learning Repository provides 557 datasets and a searchable interface.
- Kaggle maintains a large collection of datasets published by users.

If you identify other interesting data sources you are strongly encouraged to share them with your classmates on Ed Discussion.

3 Detailed timeline and deliverables

- **Friday, April 7: project proposals due 10:00 PM.** The proposal is a draft of the introduction section of your project as well as a regression analysis plan and evidence that you can load the needed data into R. The proposal should be no more than 3 pages (excluding appendices). **Please submit only one document per group to Gradescope, tagging all group members**.

  You will receive written feedback on this proposal from the professor or the GSI. If your proposal is not sufficiently clear for the course staff to properly evaluate it, you may be asked to resubmit a revised proposal. After proposals have been submitted, all major changes to your project topic or reconfiguration of project partnerships must be cleared in advance with the professor.

- **Thursday May 12: final project due 10:00 PM**. The purpose of the report is for you to demonstrate your ability to Kask meaningful questions and answer them with the results from regression analysis, and your proficiency in using R and in interpreting and presenting results. Focus on methods that help you answer your research questions. You do not have to apply every statistical procedure we learned. Also pay attention to your presentation. Neatness, coherency, and clarity will count.

  At a minimum, your report should have the following sections: an introduction, a section describing your final regression model, a discussion section outlining conclusions drawn from the model and limitations of the analysis, a conclusion summarizing the main findings in answer to the research question, an additional work section describing other models and diagnostics you tried, and a references section listing the sources you used. The first four sections should be no more than 3 pages in total, and the “additional work” section should be no more than 5.

  At the end of your report, you should include a code appendix containing all the R code used to conduct your analysis. If you used R Markdown to write your report, you may either append the entire .Rmd source, or extract the code chunks and present them separately. The code appendix will not count toward the page limit. **Please submit only one document per group to Gradescope, tagging all group members**.

- **Friday May 13: self-evaluation due 10:00 PM.** Please submit a short reflection on the experience of working on the project, including the strengths of your approach and anything you would have handled differently in retrospect. Please also comment on the differences (in context, content, etc.) you see between your project and the kinds of data analyses you are likely to encounter in your work after completing your program at Berkeley.

  For projects completed in pairs, each partner must write and submit the reflection individually, and must also include a description of how the work was distributed throughout the project. You should be specific about how you decided to organize the work and what each group member contributed.

  These reports do not need to be formal, but should be 1-2 paragraphs. They will be graded for completion only.

4 Assessment

The proposal and final report will be graded holistically. Below we provide the rubrics used to do this grading.In addition, if the end-of-project self-evaluations show highly disproportionate distribution of effort between project partners, grades may be adjusted to reflect this.

Proposal rubric:

- **Research question and data** (40% of grade): Is a research question chosen and clearly defined? Is the question challenging, interesting, and specific? Is a specific dataset identified that is relevant for this question, and can the group load it successfully in R in a usable form?
- **Choice of analysis** (40% of grade): Is the model and analysis chosen relevant to the research question? Does it make effective use of the data in hand? Is a clear plan given for making modeling choices and assessing assumptions, with attention to the broader scientific context?
- **Clarity and concision** (20% of grade): Is the report organized logically with a clear structure? Are individual steps in the analysis explained clearly? Is writing concise and grammatically correct? Are length and formatting requirements observed?

Final report rubric:

- **Choice of analysis** (33% of grade): Is an appropriate research question chosen and clearly defined? Is the model and analysis chosen relevant to the research question? Does it make effective use of the data in hand? Are modeling choices and assumptions properly justified with reference either to the data itself (e.g. through diagnostics or exploratory data analysis) or to the broader scientific context?
- **Interpretation and evaluation** (33% of grade): Is model output interpreted correctly? Are conclusions justified by specific results from the analysis? Are weaknesses of the approach and alternate explanations or methodologies considered and discussed?
- **Plots, figures, and code** (17% of grade): Do plots and figures convey important information that contributes to the central arguments of the report? Are plots and figures labeled appropriately and referred to effectively in the text? Are computations performed in R correct and well-documented?
- **Clarity and concision** (17% of grade): Is the report organized logically with a clear structure? Are individual steps in the analysis explained clearly? Do conclusions drawn from the data analysis respond meaningfully to the research question? Is writing concise and grammatically correct? Are length and formatting requirements observed?