Final Project InstructionsStatistics 506

Overview

Your final project will take the form of a short (~2 page) report on data analyses you design to answer a substantive question of your choosing. You have two options from which to choose for posing your substantive question.

1. Pose a research question that can be answered using publicly available data.
2. Pose a question about how a statistical method performs in an atypical situation. Answer this question using a Monte Carlo study.

This is an **individual project**.

List of topics

If you are taking option #1, you may choose from the following data-sets:

1. The Medicare provider utilization and payment data, https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners.
    - This data set tracks information on procedures and services that physicians or other health care providers perform which are sent to Medicare for insurance coverage.
    - Your question should involve region (state and/or zip code) and pull in at least one variable from another data source (e.g. income tax statistics can be found at https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2020-zip-code-data-soi).
    - An example of an appropriate research question using this data: Are there characteristics of a zip code which predict higher Medicare reimbursement amongst Radiation Oncologists? Here is a paper which addresses this question.
2. The Commercial Buildings Energy Consumption Survey (CBECS), https://www.eia.gov/consumption/commercial/data/2018/index.php?view=microdata.

- This data set contains information on a sample of all commercial buildings in the US and their energy characteristics, consumption, and expenditures.
- This is a complex survey design, so you will need to account for that in the analysis. The user guide can provide more information about this.
- An example of an appropriate research questions using this data: Does the window-to-wall ratio (portion of an exterior wall which is windows) affect the energy consumption of a commercial building, for several different measures of energy consumption? Here is a paper which addresses this question.

3. Choose an alternate dataset.

    - If you choose an alternate dataset, it must have some level of complexity to it. Examples of this complexity could be:
        - Extreme size.
        - A complex survey design.
        - Multiple data files that will need merging.
        - An abnormal amount of data-processing.
        - Needing to pull in other data to obtain certain variables (e.g. census data for zip codes).

If you are taking option #2, you may choose from the following statistical topics.

1. What is the impact on collinearity on multiple linear regression? Choose a non-trivial case of collinearity to question and explore alongside the basic bivariate situation (e.g. multivariate collinearity with bivariate independence).
2. Poisson regression is occasionally used in place of logistic regression. Address when (if any) these approaches produce similar results, and choose a secondary non-trivial question to explore (e.g. does the presence of omitted variables affect one model more than the other.)
3. Pose your own statistical question. Good examples of this are typically assumption violations of standard statistical techniques (e.g. what would happen if the true errors had a very non-normal distribution?).

In all these methodological topics, your simulation should be comprehensive. E.g. for the collinearity question, it is not sufficient to set the correlation to a single value, run a MC, and call it good. Instead, you should look at things such as the impact of the collinearity as correlation changes, or whether the strength of the errors plays a roll, or whether non-collinear variables are affected.

Any of the software we discussed in class (R, SAS or Stata) can be used. If you want to use something else, please check with me prior to submitting your abstract. Do not use Python as that is the focus of the other class.

Project proposal

The project proposal is due November 14th at the start of class. Any late submission incurs a 2% penalty on the final project grade.

The proposal should be a statement of your research question and a brief (1/2 page) description of your analysis plan. This will not be graded, but will need to be approved by me. If necessary, we may iterate on the proposal. After you submit it, I will either approve it, or send it back to you with comments for amending. If I send it back, please submit an updated version is a timely fashion - the longer it takes to finish your proposal, the less time you have for your final project.

If two students somehow ask the exact same question, I will send it back to both of you to work our between yourselves - I'd encourage you to brainstorm together to find a second question and pick who will work on which question.

Final Project

The final project is due December 8th at noon. **The final project will not be accepted late.**

The final project should be a report (submitted as a PDF). It should function as a small research paper - a short introduction to the data or the statistical methodology, a clear description of the research question, a summary of your approach and results (if doing data analysis) or your simulations and results (if examinining a statistical methodology), and a conclusion.

The report should be about 2 pages (no more than 3). If you have graphs, figures, or tables to include, you may include up to two in the document; any remaining graphs, figures, or tables should go in an appendix. It is unlikely that much if any actual code should feature in the report.

Note that unlike a problem set, you can (and probably should) exclude from the report some of the details. For example, if doing data analysis, you do not need to go into details about how you cleaned the data; it is sufficient to quickly summarize any major changes you needed to make.

Code

Your report should not include the code, unless there are special circumstances where you need to demonstrate a *small* bit of code to make a point. Instead, your analysis code should be placed in a GitHub repository. Provide a link to the repository in the final report.