Project · Bring Your Own Data

Due Monday, May 11, 2020 (10 AM)

Project description

The assignment for this project is simple: pose an interesting question that can be addressed using statistical learning; collect or find a relevant data set; and use the data, in conjunction with the tools we have learned in class, to answer the question you have posed. If relevant, quantify any uncertainty that arises in answering your question. You should also address any shortcomings in the answer provided by your data and analysis. You will be evaluated both on the technical correctness (50%) and the overall intellectual quality (50%) of your approach and write-up. See the footnote at right for lists of possible data sources; also note that you will find lots of publicly available data sets on Covid-19 these days, some of which would potentially make an excellent project.[1]

This assignment is purposely open-ended, allowing you considerable freedom to follow a path dictated by your own intellectual curiosity. Strive to write something that a statistically literate person of wide-ranging interests (for example, a future employer) would find engaging and impressive.

Advice

The best projects I have seen over the years have been *problem-driven* rather than *process-driven*. A problem-driven project is one where you start with an interesting question, together with a data set capable of addressing that question, and let the statistical approach be guided by that question. A process-driven project, on the other hand, is one where you find some data set, without necessarily having a strong idea of the question(s) you'd like to use that data set to answer. You then fit some models on that data set, focusing on the process of what you did, as opposed to the substantive question you'd like to answer. Most of the least inspiring projects I have seen over the years fall into this mold; they often involve downloading a data set from Kaggle and simply "kicking the tires" on that data somewhat aimlessly.

My strong advice here is to undertake a project that is problem-driven. For good examples of problem-driven questions, you should recall some of those we've used for in-class problems and homework so far (building a predictive model for house prices; demand curves for beer; energy demand; and so forth). Each of these is strongly anchored

[1] These websites in particular have long lists of data sources: `https://github.com/caesar0301/awesome-public-datasets`, `https://www.google.com/publicdata/directory`, and `http://stats-for-change.github.io/data.html`. If you want to branch out even further, here's a short list of other sources you might consider: major newspapers, the U.S. census, the Federal Reserve, academic journals, the Economist, Twitter, the World Bank, ESPN.com or other sports sites, Craigslist, Amazon prices, EBay, the Bureau of Labor Statistics, Facebook, the World Economic Forum, the OECD Factbook, the CIA World Factbook, the Securities and Exchange Commission, Yahoo finance, Google Public Data Explorer, your own vital signs, your favorite blogs, your other classes, and your friends. If you know how to write a program that will scrape a website, your options are almost limitless here, but even if you don't, there are lots and lots of structured data sets out there waiting to be downloaded and analyzed.

in a real-world question of interest. Over the years I've seen students go to some really cool lengths to collect their own data sources, often scraping web sources. The reason these groups tend to do well on the project isn't because they've gone to such an effort; it's because they had a very clear question of interest, and they collected a data set that was appropriate for answering that question.

Many, many students in the past have told me that the project has been the centerpiece of their class experience—the opportunity to let their curiosity take over, and to create something that they can talk about in future job interviews as showing clear evidence that they've learned some valuable data-science skills. I encourage you to make the most of the opportunity.

To turn in

You should turn in the following three items. As with the homework assignments, you may work in groups of 4 people or fewer, or you may turn in your own project. If you work in a group, only one set of these items (bearing all of your names) needs to be turned in.

1. A written project report that describes your question, your data sources, your methodological approach, and your conclusions. This should be prepared in RMarkdown format and compiled to a PDF document that you post on GitHub. Send me both the link and the PDF itself.
2. The data set itself, in .csv format.
3. A link to the raw .Rmd file used to analyze your data and prepare your report. If your analysis and plots are not 100% reproducible, you will not receive a passing grade.

All three items should be e-mailed to the e-mail drop box at `statdropbox@gmail.com`. (Please don't send projects to my regular McCombs account or through Canvas: Gmail has a higher limit for size of attachments, and I want to have these projects all in one place.) The subject line of your e-mail should be: "SDS 323 Project: (names)," where you fill in the blank with the full names of all your group members.

Organizing your report

A reasonable length here would be 5–10 type-written pages including figures and tables, but treat this only as a rough guideline rather than

an absolute quota or limit. There is a hard max of 15 pages including tables and figures: I won't grade anything beyond that. Please organize your report into the following sections.

Abstract: summarize your question, your methods, your results, and your main conclusions in a few hundred words or less.

Introduction: Introduce the question you're trying to answer at a reasonable level of detail. Give background and motivation for why it's important.

Methods: Describe your data set and the methods you will use to analyze it.

Results: Tables, figures, and text that illustrate your findings. Keep the focus on the numbers here. You will interpret your results in the next section.

Conclusion: Interpret what you found. What are the main lessons we should take away from your report?

Appendix: optional. Any details (like extra figures, etc) that didn't fit in well with the main report, but that you think are important.

The main body of your report (excluding the appendix) should probably have no more than about 4-6 figures and tables. Use those to focus on the most important results. Put the rest in the appendix if you need more. Number all figures and tables and refer to them by number (Figure 2, Table 3, etc) in the text where appropriate. Do not include a figure or table in the report or the appendix if you don't discuss it in the text.