Project · Bring Your Own Data

Due Friday, April 6, 2018 (5 PM)

Project description

The assignment for this project is simple: pose an interesting question that can be addressed using regression; collect a relevant data set; and use the data, in conjuction with the tools we have learned for building multiple regression models, to answer the question you have posed. Make sure to quantify any uncertainty that arises in answering your question, and to 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.[1]

This assignment is purposely open-ended, allowing you considerable freedom to follow a path dictacted 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 modeling 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.

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 activities so far (building a predictive model for house prices; choosing the optimal price of a gallon of milk; understanding the impact of marketing displays on demand for cheese; answering questions about what factors drive variation in dengue fever; finding factors that predict variation in gas prices; and so forth). Each of these is strongly anchored in a real-world question of interest. Over the years I've seen students go to awesome lengths to collect their own data sources: driving around Austin to collect data on gas

[1] Where can one find data? Everywhere! These websites in particular have long lists of data sources: `https://github.com/caesar0301/awesome-public-datasets` 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 own experiment or survey, 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.

prices, going from one grocery store to the next to understand variation in the price of consumer goods, and so on. This kind of effort isn't necessary to do well on the project, but it is nearly always impressive when I see it. 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 STA 371H 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.[2] You can find some general advice here: `https://github.com/jgscott/learnR/blob/master/write_ups.md`.
2. The data set itself, in .csv format.
3. The R script used to analyze your data. If your analysis and plots are not 100% reproducible, you will not receive a passing grade.

[2] A reasonable length here would be 4–6 type-written pages including figures, but treat this only as a rough guideline rather than an absolute quota or limit. If you have lots of figures and tables, you might easily go over 6 pages.

The first item should be submitted as a hard copy, either in person or to my office (CBA 6.478) or mailbox (CBA 5.202). All three items should be e-mailed to the drop box at `statdropbox@gmail.com`. (Please don't send projects to my regular McCombs account or through Canvas.) The subject line of your e-mail should be: "Project: (names)," where you fill in the blank with the full names of all your group members.