Advanced Data Analysis

Statistics 5291 — Spring 2020

Course Project

Find a data set and conduct a statistical analysis, summarizing your findings in (1) a 5-minute oral presentation, and (2) a written report of not more than 5-7 pages.

Data and analysis

Your data set can be taken from a published or on-line source, but your analysis should be original and entirely your own work. There is no expectation that you employ highly sophisticated methods; rather, you will be evaluated on the correctness of your interpretations, and on how useful your conclusions would be to a non-statistician. You oral presentation and your written report should be directed to someone who is not an expert in statistics, but is highly interested in the subject matter. Can any surprising conclusions can be reached based on the data? What interesting questions can't be addressed because of how the data were collected, or can't be answered conclusively? That said, it is a course project, so the data analysis shouldn't be *too* simplistic. A multiple regression analysis would be okay, a simple regression or one-way ANOVA would not. You are not limited to methods studied in this course.

Written report

It is suggested that you organize your write-up in three parts. In the first part you should describe the source of the data set, and identify the scientific questions of interest. Describe how the data were collected. The second part should consist of a summary of findings and conclusions. The first two parts should be no longer than a page or so each. In the third part you can include a detailed description of your analyses. Begin with descriptive statistics and graphical summaries of your variables, and move on to statistical modeling and inference. Report and interpret p-values and confidence intervals as appropriate. Be explicit about the assumptions that underlie your methods, and the impact on your conclusions if those assumptions are violated. You might consider a fourth section of your report, to address any limitations or shortcomings of the study, and/or suggest additional questions for future research.

Oral presentation

Each student is required to give a 5-minute talk to the class. We will schedule four sessions, two each on May 6 and May 7, at the times stated below. The format will be modeled after that of the "Speed Presentations" at the *Joint Statistical Meetings*. **You are expected to remain in attendance for the entire session; any student who gives their talk then leaves will receive zero credit on the project, and will fail the course.** Support your classmates by being a courteous and attentive audience. You do not have to attend the other three sessions, just the one where your talk is scheduled.

- Wed May 6, 9:00am–11:30am
- Wed May 6, 12:30pm–3:00pm
- Thu May 7, 9:00am–11:30am
- Thu May 7, 12:30pm–3:00pm

Email the instructor (rcn2112@columbia.edu) to indicate your preferred time slot(s). For the spring 2020 edition of this course we will conduct these sessions via Zoom meetings.

Deadline

Your written report is due by 10:00am on Monday, May 11.

Preliminary feedback

You are strongly encouraged to discuss your project idea(s) with the instructor, the earlier the better, so you can get redirected before heading too far down the wrong path.