STAT225 **Project Instructions** May 14, 2020, 2pm

Overview

For your end of semester project, your group will select a dataset and carry out a thorough data analysis. There will be two components to the data analysis: a presentation (pre-recorded) and a report, both due on May 14th at 2pm, a final exam period you will be required to attend on Zoom. You will also need to submit a proposal through Gradescope and dataset (details below) by Wednesday next week (4/22) at 10pm ET.

As a group, you will need to find a dataset together. This dataset should meet the following conditions:

- Has one quantitative response variable
- Has 5+ quantitative predictor variables
- Preferably, at least one pair of X-Y relationships has a non-linear relationship (use GGally::ggpairs() to explore a dataset rapidly)
- Measurements should not be aggregated over large, dissimilarly-sized groups (for example, no data at the country or state level), but they may be aggregated over similarly sized groups (for example, number of games played by players or teams).
- For every variable you might include in a regression model, you will want 15 observations per predictor. So, you will want at least 75 observations.
- Your dataset should not be large - no more than 500 total observations. Take a random sample of 500 observations if it is larger.

I recommend looking at the following sites if you do not have a dataset readily available:

- R package openintro https://www.openintro.org/stat/extras.php
- Datasets available in R: https://vincentarelbundock.github.io/Rdatasets/datasets.html
- Professor Horton's IS5 dataset repository: https://nhorton.people.amherst.edu/is5/data/

You are welcome to find a dataset from a different site, or a dataset you've used in a prior class, but not a concurrent STEM course!

Methodology

For the analysis, you will pick at least five of the nonparametric techniques (and OLS MLR) we have or will have learned in this course to apply to your dataset. Every procedure should involve your response variable, Y, in some way.

You will need to perform one of the following tests (or calculate the corresponding CI):

- Binomial Test
- Fisher's Sign Test and
- Wilcoxon Signed Rank Test
- Wilcoxon Rank Sum Test
- Permutation/Randomization Test
- Kendall's τ or Spearman's ρ Test

And at least one of each of the following:

- Kolmogorov Smirnov Test (with empirical CDF[s] plotted)
- Kernel Density Estimation (not smoothing related)
- OLS multiple linear regression (for comparison purposes)
- JHM multiple regression
- Generalized additive model (which will require you attempting a variety of smoothers)

The six methods you use should tell a cohesive story about your data and should relate to each other in some way. Models should be formally stated, i.e., $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \varepsilon, \varepsilon \sim N(0, \sigma_\varepsilon)$ or $\hat{Y} = \hat{\beta}_0 + \hat{f}_1(x_1) + \hat{f}_2(x_2) + \varepsilon, \varepsilon \sim$ with center at 0, and relevant tables should be printed using `kableExtra`, see for example https://bookdown.org/yihui/rmarkdown-cookbook/kableextra.html and the cheat sheet on Moodle.

Proposal

Your proposal should contain the following content (please use this structure):

1. GROUP: Provide your group indicator "GroupX"
2. MEMBERS: List the members of your group
3. TITLE: Your title
4. PURPOSE: Describe the general topic/phenomenon you want to study, as well some focused questions that you hope to answer and specific hypotheses that you intend to assess.
5. DATA: Describe the data that you plan to use, with specifications of where it can be found (URL) and a short description. You may want to combine data from multiple sources into one file. Also consider how reliable the data or expected data source is? What are its shortcomings? Identify any potential biases that you might find in it. You should include your dataset in a proposal .RMD you post and pin in your Slack channel (where you've uploaded it to a personal website (like GitHub) and are reading the file directly from the source, your website, or the R package that contains the data), OR send me your dataset in the Slack channel if you've pre-processed it but don't know how to upload it (in this case, it should be a in a flat file format – .csv, .xlsx, etc, and under 5 MB given the dataset restrictions stated above. I will only upload it to my website - nothing more.)).
6. POPULATION: Specify what the observational units are (i.e., the rows of the data frame), describe the larger population/phenomenon to which you'll try to generalize, and (if appropriate) estimate roughly how many such individuals there are in the population.
7. RESPONSE VARIABLE(S): What is the response variable? What are its units? Estimate the range of possible values that it may take on.
8. EXPLANATORY VARIABLES: Describe the variables that you'll examine for each observational unit (i.e., the columns of the data frame which correspond to a particular row). Carefully define each variable and describe how each was measured. For categorical variables, list the possible categories; for quantitative variables, specify the units of measurement. You may want to add more variables later on, but you should have at least three or four explanatory variables (predictors).
9. EXPLORATORY ANALYSIS: use `GGally::ggpairs()` as many times as needed (no more than 6-8 variables per `ggpairs` call). There should be no more than four `ggpairs` calls – whittle your candidate variables down to 12 if needed.

Report

The final report will be turned in via Gradescope. This should be a formal data analysis report with full write ups for all methods (including hypotheses, significance levels, assumption checking, and formal conclusions where appropriate – no need to check assumptions for KDEs, JHM multiple regression, and GAMs). You should also include a discussion of the most troublesome area of your project–was it data manipulation? Was there a particularly tricky variable that you had to deal with? Were there modeling decisions you made that you weren't sure about? I want to know what your biggest takeaways are! This report can be as short as 3-6 pages – only the relevant details are needed. Plots should only be included if they are part of the discussion. You should include smoothed scatterplots, for example, and it would be best if any lines you are adding to the same scatterplot are all plotted on the *same* instance of the scatterplot (there are obvious exceptions to this – use your best judgment). I expect your report to be largely complete by 5/5 - 5/7 to give you and your group members time for your other three courses that may have exams during Finals week.

Presentation

You will present your analysis to the class in a pre-recorded Zoom session with your group. Each member can pre-record their part of the slides (please use R Markdown to produce slides with ioslides, beamer, slidy, or PowerPoint for example – see https://bookdown.org/yihui/rmarkdown/presentations.html), perhaps with a voiceover of a shared screen on Zoom. The final slides must be submitted on Gradescope in .PDF form no later than 5/13 at 10pm. You are welcome to produce the presentation with an .HTML/.PPT version of the slides, as they allow bullets to appear, disappear, etc. There will be a short Q&A session at the end of your presentation. I recommend you finish a preliminary version of the presentation before 5/8, so that I can give you feedback on how to improve the presentation (and possibly your report). I also recommend that you upload your videos to YouTube by 5/13 at 10pm to A. give them time to upload and B. give me time to organize a playlist before the day of the presentations. Upload them as unlisted and share the link, or upload them as private and share them with `rmcshane@amherst.edu`.

Group Participation

Your group members should all be contributing to this project. If you feel as if you are shouldering an unfair burden, please reach out to me and let me know. There will be a group assessment at the end of the project, to be completed on 5/14 after the presentations are complete. I will be monitoring your group Slack channels to get a general impression of contributions as well. I can also steer your group in the right direction more quickly if I have access to your thinking.