Final Project

Your final "exam" is a project which you can complete in pairs or by yourself. It is due on Friday June 12. There will be a question and answer session on Friday May 22, 1-2pm (focusing on data sets for submission) and again on Friday May 29 at 4-5pm (focusing on analysis issues).

If you wish to work with a partner but do not have one, please fill out this form: https://forms.gle/QaqQXBB4fstCe6BH8 Be aware that this means I will share your name and email with at least one other student. There are no guarantees. You should fill out the form NO LATER THAN Friday, May 15. You'll find a link to this form on CCLE under Final Project.

Summary

Find a data set and show us your understanding of multiple regression analysis.

Milestones

1. Due May 26 (Tuesday). You and your partner should find a data set online, or collect your own data, that has these characteristics:

a) At least 50 independent observations
b) A numerical response variable
c) At least 5 predictor variables, at least 3 of which are numerical.

You will submit: a 2 paragraph description that explains (a) Who collected the data, (b) where the data are stored, (c) why the data were collected (including any questions they were collected to answer), (d) how the data were collected and (e) what your primary analysis question will be. This will be 20% of the grade. You should also include a link to a data dictionary/codebook if such a thing exists. A data dictionary is a document or webpage that explains what the variables and observations mean and the method used to collect the data.

If you submit by May 20, I will give you feedback on whether the data are suitable.

Sources to consider: data.gov, any open-data portals (LA City and Santa Monica and San Francisco, for example, have excellent portals), sports websites, data scraping, wikipedia (this may need some data scraping), entertainment sites, data from your personal devices, data from a study you do at home.

Data you should NOT use: data from this course or textbook, data from other textbooks, data from educational websites (ask us if you have any doubts, or make a pitch if you think an exception should be made.)

I expect everyone to get the full 20% on this part, but encourage you to make sure you are submitting a valid and productive data set before the deadline to make sure.

In the past, problematic data sets have included

datasets with too many categorical predictors or a categorical response variable

datasets with a binary or fixed-value response variable. (For example, the response is a person's answer to a Likert-scale type question in which they're asked to rate something from 1-5, or the response is a variable that counts something and the highest value is small)

datasets with correlated values, such as time-series

"nested" data. for example, your response variable is student test scores, but you have multiple students in multiple schools. This means responses within a school are likely to be correlated with each other.

2. Due June 12 (80%)

A report of your analysis. This should include

a) An abstract paragraph that explains the primary question you were hoping to answer and your answer, as well as a brief description to the data.

b) A link to the data

c) A discussion of your findings. The analysis should be clear and legible and should clearly state your final model, a discussion of the strengths and weaknesses of the model with respect to model validity and "pesky" points.

Your analysis should also include a brief explanation of how you achieved the final result, but you don't have to show intermediate models. For example, if you tried a number of transformations of variables, no need to show us all of those models. It's enough to tell us what you tried and show us the final "best" model. It is understood that the "best" may not be great, but your job is to tell us how it is an improvement on the more basic model.

Things to consider in your discussion

1) Why is your question interesting to you? Convince us of its importance or relevance or entertainment value.

2) What analysis preparation did you undertake? This may mean selecting a subset of data, recoding variables, imputing missing values, fixing errors, reformatting the data, etc.

3) What transforms of the variables did you consider and why?

4) How did you select which variables belong in the model and which do not?

5) Are there any influential or leverage points that might affect your conclusions?

6) Are any of the predictor values surprising? Interpret them (or some of them) to help us understand. (Maybe you were surprised by what was dropped from, or what remained in, the model? Or by the sign (positive or negative)?

7) Provide graphics that help you explain your answer.

There are no length guidelines, but I do not feel the need for this to be very long. (3-5 pages).

Please use this outline. If you think another would be better, that's ok, but please run it by me first.

I. Abstract (200 or so words).

II. Description of the questions to be addressed and overview of the data. (This may be a condensed version of Milestone 1.)

III. Results. Your answers to the questions, along with a brief justification. This will likely include providing your final model. A tabular form (showing the slope estimates and their p-values) is usually sufficient for this.

IV. Discussion. This is where you will explain the model-fitting process, and strengths and weaknesses, and other things. (See "Things to consider")

V. Conclusion. A wrap-up, and maybe what additional data you'd like to see, or what you'd do if you had more time.