MAT8406 Project Overview

Spring 2019

The goal of the project is to produce an essay that addresses a realistic empirical question by conducting a thorough regression analysis. Students are encouraged to work in groups of **about 3 people**. You may investigate any empirical question you choose: you will find the data, decide on the analyses to perform, and draw all the conclusions. I can help you, or course, but I will deliberately avoid explicit guidance. The project is meant to be an open-ended exercise.

The project report is due on Monday of the final week (May 6).

Description:

The most important thing is to demonstrate conceptual mastery of the course material and its implementation. Your essay should clearly state your empirical question and write down your regression model(s) as we did in class, e.g. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$. You should provide a motivation for which variables you include in your regressions and clear definitions of any constructed variables like indicator variables for categories. You should carefully address diagnostic issues, model selection and inference, as indicated in the model building procedures in the right diagram.

There is no formal requirement on the length or format of the essay. The goal is to write something that is clear, readable, and thorough; however you feel you can best accomplish those goals is fine. The suggested length is 10 pages including tables and graphs (no need to include a ton of computer code/output).

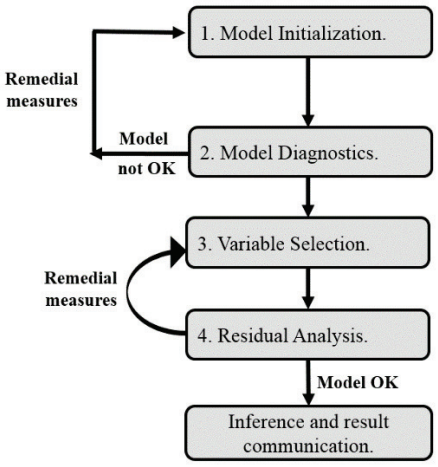

Diagram: Strategy for regression analysis.

Proposal:

The project proposal due in week 8 should include the following five things:

1. a description of your essay's empirical question,
2. why it is important,
3. the data sources you will use and how you will get access to them, and
4. the methods you will use, and
5. preliminary results you have obtained.

Submit the proposal to Blackboard by Mar 9. Make sure the name of each group member is in the proposal.

I will provide feedback on each proposal, discussing your preliminary results, details of your approach, additional strategies in addressing the empirical question, and any problems you bring up. In-person meetings can of course be scheduled. The more you have attempted in your proposal, the better the feedback you will get. Of course, feel free to talk to me before the proposal is due about your idea. You can always get informal feedback.

Data:
Feel free to find and use the data that is most intriguing to you. Here are some sources you might find useful (only a few ideas):

- Wharton Research Data Services (WRDS): http://wrds.wharton.upenn.edu/
- City of Philadelphia data: https://www.opendataphilly.org/
- IPUMS: U.S. census data: http://www.ipums.org
- Prediction and data mining competitions (all sorts of application areas): http://www. kaggle.com/
- Compustat: firm level data for publicly traded firms

Some data sets are interesting, but are not high quality in one way or another. That's fine. If your data set is limited in one way, think about exploring/expanding your project in a different direction. Are there different ways of using those variables? Different outcomes you could predict? Different ways to evaluate model quality? Interactions that are interesting? Diagnostics and transformations that are useful? If you have only 5 variables, then you'll want to explore these issues carefully. If you have 5000, you are going to be more worried about variable selection methods. Different techniques for different projects and none are a priori better or worse.

Project Grading:
People often ask how the project and the proposal are graded. The proposal is not formally graded, beyond you turning it in on time. It's a way for you to get feedback to make your project better. Only under extreme circumstances will I "reject" a proposal (this has happened only once). As for the project itself, the goal of the project is to demonstrate that you can do a thoughtful, thorough job of investing a real-world empirical question using the material from class. Doing so is an "A" project. So you do not have to: *(i)* use every single technique from class; some will apply to your project and some won't; *(ii)* come up with an earth-shattering question or result; *(iii)* have the best data.

The project is deliberately open ended, and to reflect that spirit, so is the grading. I will not post a "sample A+" project, or anything of that sort. There's no specific structure or content required, so different projects look very different.

Timeline:

Saturday, 03/09: Project proposal due
Thursday, 05/02: Project presentation in class
Monday, 05/06: Project final report due