Regression Project Assignment

This note provides detailed information about the capstone project assignment. The capstone project provides you the opportunity to analyze a problem of your choice, using the analytical tools that are taught in D&D.

Objectives

The immediate objective of the capstone project is to provide practical experience at data analysis and regression modeling. The statistical models developed in D&D are used for decision-making in a wide variety of settings, and the capstone project assignment is designed to bridge the gap between classroom learning and real-world application. A second objective is to provide teamwork experience and project management experience addressing quantitative problems.

The capstone project involves four basic steps: (1) Identify, as a group, a question that can be answered with regression analysis. (2) Obtain suitable data that can be used to analyze the problem. This may involve data from your former employer, a former client, prior academic experience, or publicly available data. (3) Perform the analysis using techniques discussed in class. (4) Communicate and document your results in a clear, polished report.

Forming Groups

The assignment is to be completed in groups of **at most 4** students, from your same D&D section. You can form a group among your classmates based on broad interests. Alternatively, you can request to be assigned to a group by filling up this form as soon as possible: `https://forms.gle/EDbWg3KgcWXnYN1s5`

We encourage you to spend a few moments before then looking through the example projects that we have posted on Canvas. From these projects you can quickly gain a general understanding of capstone projects and what types of data you use.

Choosing a Topic

Once you have formed a group, you should decide on a topic. There are two main sources for project topics:

1. **Publicly available datasets** If you select this option, you should start by looking through the list of publicly available datasats linked on Canvas. will then need to do the following:

2. **Source your own project topic:** ideally motivated by a company you have worked for in the past or see yourself working for in the future. You can also see Canvas for some examples of publicly available data sets. need to do the following: If you are unsure about your project, please consult with us (the CAs or me) either via e-mail or in person. Again, look at the sample projects for examples.

Logistics

By the end of the day on **Thursday October 16**: Submit a very brief description of your idea via `https://forms.gle/EDbWg3KgcWXnYN1s5`, detailing your project topic and where you plan to find data. In the form, list the names of all members of your group.

On **Sunday November 9**, as part of a lab assignment, a **project proposal** is due. The purpose of this proposal is for me to provide you with some advance feedback on your project. The project proposal should be a **one page** document with the following information: the names of the project members, the project title, a concise statement of what questions are to be addressed, and a summary of the data to be used and its source. Be sure to describe what information is contained in a typical data point from the data to be used - i.e., is it one customer, one person, one firm, one country, one year, etc. Most importantly, you must make it clear what your main left hand side (Y) variable is — the data you wish to explain. You should also describe the most important right hand side (X) variables in your data.

Your final report is due **Friday December 4**. The report itself should be (at most) 10 pages in length (not including tables and figures – see below). The first page must be a one-page Executive Summary of the problem, principal findings, and recommendations. The body of the report should include a detailed statement of the problem and questions that are addressed; the data and their sources; the analysis conducted and regression results; and your principal findings, recommendations and conclusions. In your analysis, we expect you to carefully justify the regression model(s), to present alternative regressions specifications (if appropriate), and to verify that the standard regression assumptions are satisfied (as best as possible) for your model. You are encouraged to present informative plots

and descriptive statistics that contribute to the analysis and conclusions. You are also encouraged to think about what are the inferences possible from your analysis, what the limits of your analysis are, and what recommendations (if any), you would give to your client based on your analysis.

In addition to the ten pages of the main report, you may include a technical appendix containing supplementary data tables and graphs. If your data can be printed onto several pages or less, the data set should be included in this appendix, clearly labeled for the reader. The final report should be a polished document of the sort presentable to a client after a professional consulting study. Unlike some consulting reports, however, your report will be read by someone with an advanced knowledge of statistics. Reports should be submitted via Canvas. Please be sure that the names of all the students in your group are on the report.

Finally, on the last day of class (Friday December 4), each group will make a short (few minutes with max three slides) presentation in class, explaining their project and their main findings.

Grading Criteria

Each project group should hand in only one proposal and one capstone project report. All project team members will receive the same grade on the project. Besides the usual administrative criteria (e.g., the project meets the deadline), we will use the following criteria:

1. Originality of the problem.
2. Usefulness and practicality of the analysis.
3. Quality and correctness of the analysis.
4. Quality of the exposition, writing, and presentation.

Because this is a D&D project, item 3 receives more weight than the other three criteria do individually. We take into account, on a more subjective basis, unusual effort or findings that are valuable to a real client. We also give credit for entrepreneurial, clever and difficult data collection efforts.

Strategies and Recommendations

A useful technique is to write the analysis as though it was a consulting report. Here are project titles for some of the better projects that have been done in the recent past:

- Analysis of Teacher Retention Factors for *Teach for America*
- Feeding the Hungry in San Francisco: An Analysis of the Factors Influencing the Demand for Free Meals at St. Anthony Dining Room
- Determinants of Patients Leaving Without Treatment and Wait Times in Emergency Rooms
- Analyzing Data to Accurately Predict SAT Score Improvement Guarantees
- Math & Marijuana: Factors affecting math performance

Each group must select a topic and obtain the data. Part of your grade depends on the originality and relevance of your topic and the data. Regressing stock prices on a market index is not very original, though it is perhaps relevant to finance. On the other hand, forecasting the sales of a software product in different stores is more original and requires unique data. In the past, students have used data from employers, former clients, surveys, government agencies, library references, and local area businesses.

In general, the best and most rewarding capstone projects have been those where the students worked to address a specific problem faced by a specific company, non-profit institution, or government agency. You are encouraged, indeed expected, to use whatever industry or worldly knowledge your group possess to identify and select your topic.

Obtaining the data: some considerations

There are several options that you can follow to obtain the data:

Own firm/past experience: You may be able to use data from your previous employer. Students tend to find this option very rewarding, as they can use their specific knowledge to make the analysis more insightful. It can also speed up the coordination with the firm to obtain the data.

GSB Library: Stanford University is subscribed to a number of important business databases. If there is a particular topic that you would like to explore, you can sign up for an orientation session with a librarian to get some help. The GSB library also provides a guide with many data links:

http://libguides.stanford.edu/az.php

Websites: There is an increasing number of websites that offer public data sets for competitions. There are also researchers that make their data available. We keep a current list on Canvas. Here are some examples:

- `www.kaggle.com` and `challengedata.ens.fr` (Data competition websites)
- http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataSets.htm (Research data on several topics)
- Government data from the Census, the US BLS, the US DOT, and many others, including many international data sources.

Regarding the quality of data, here are some hints that may help:

- *Topic/Variables:* It is essential that you understand what your variables are, to test meaningful relationships in the data. If you have additional industry or institutional knowledge, it is even better. Even though this may seem obvious, some competition data websites provide codified data where you *cannot* tell what each variable represents!
- *Size:* It is useful if you can have more than a hundred observations. You also want to keep the size manageable: if you have a very large data set (gigabytes), you should probably focus on a particular subset of it.
- *Categorical vs Numerical variables:* It is very useful if your data have at least one numerical explanatory variable (e.g. price, quantity), versus categorical variables (e.g. color, day of the week).