Columbia Business School

B8114: Applied Regression Analysis

Syllabus – Fall 2022 B-Term DRAFT

Professor Paul Glasserman
212-854-4102
pg20@columbia.edu

Overview

The goal of this course is to provide students with practical experience in building and analyzing regression models to address business problems.

The course picks up where the core course in Managerial Statistics left off. We will begin with a brief review of regression analysis as covered in the core and then move on to new topics, including model selection, interaction effects, nonlinear effects, classification problems, and forecasting.

All material will be covered through examples, exercises, and cases. In addition, students will work in groups on a final project of their choosing. The goal of the project is to address a specific business problem through statistical analysis.

Coursework and Grading

Grades: 50% Homework, 35% Final project, 15% Class participation

Homework

We will have four homework assignments. The first homework must be done individually (type C assignment). Subsequent homeworks may be done in groups of up to three students (type A).

We will have two sets of short review questions to be answered on Canvas. These must be done individually.

Class participation

Students should be present, prepared, participating. Class will be more engaging for everyone if we maintain an active discussion. Class participation is as much about asking good questions as

it is about knowing the answers. **As part of your class participation score, you will be asked to evaluate the final projects of other groups.**

Final project

Final projects should be done in groups. The target group size will depend on the final enrollment in the course. Given the limited time available, final projects should not be overly ambitious. A successful project poses an interesting question and runs one or more regressions to address the question. The scope of a project is about the same as a homework assignment or lecture topic. The main difference is that you're posing the question, not me.

Finding a good combination of an interesting question and relevant data takes time, so you should be working on this from the beginning. There are at least three broad categories of data sources to consider:

- Data you have through your work
- Data available through public sources, such as governments, NGOs, foundations, academic journals, news organizations
- Data designed for statistical exercises, such as Kaggle and the UCI repository

I have listed these sources in order of interest (most interesting at top) and ease of access (easiest at bottom). Given the vast amounts of interesting data available in the second category (and hopefully the first), the third category should be a last resort.

We will have project presentations in class. The presentation deck and supporting data analysis are the deliverables for the projects.

Possible Data Sources

Here are a few examples of data sources to explore. You should be able to find many others based on your interests.

General repositories:

https://dataverse.harvard.edu/

https://archive.ics.uci.edu/ml/index.php

https://www.openml.org/

Economic data:

https://fred.stlouisfed.org/

https://tracktherecovery.org/

http://www.thebillionpricesproject.com/datasets/

Government and NGO data:

https://opendata.cityofnewyork.us/

https://ourworldindata.org/

https://databank.worldbank.org/reports.aspx?source=world-development-indicators#

http://ghdx.healthdata.org/gbd-2019

Politics, sports, surveys:

https://www.propublica.org/datastore/

https://fivethirtyeight.com/

https://www.basketball-reference.com/

https://www.pewresearch.org/download-datasets/

https://today.yougov.com/

https://worldmanagementsurvey.org/survey-data/download-data/

Software

You are free to use whatever software you like. I have designed the course around Minitab on Windows. My slides will show you how to use Minitab, and I have also made some short videos to illustrate basic tasks. Minitab strikes a nice balance between the ease of use of Excel and the power of a programming language.

Business school students registered for this course will be provided with a license for Minitab. In the meantime, everyone can download a free 30-day license.

For students who have learned some Python and would like to build on that knowledge, I have written scripts to illustrate how some of the results we discuss in class using Minitab could have been obtained in Python (using the StatsModels library). These are purely FYI.

The homework assignments are designed to be done in Minitab. You are free to do them in Python if you prefer, but only if you are sufficiently proficient to overcome any obstacles you may encounter. I will do my best to answer any Minitab questions, but I won't be able to provide support for other software options.

Class Schedule

1. Review of Basic Regression: Explaining Wine Quality: A review of material from the core, including interpreting coefficients, confidence intervals, p-values, R-square, dummy variables.

2. Which Variables Matter? Variable Selection and Effect Sizes: How do you decide which variables to keep in a model? What's the difference between statistical significance and practical importance?

3. Predicting the 2020 Presidential Vote (and Predictive R-Square). To what extent do economic conditions determine the vote? With limited data, how do we balance the need to consider special circumstances against the dangers of overfitting?

4. Capturing Nonlinearities and Interaction Effects. How do weather and work affect bike-share usage? Do student-teacher ratios affect test scores? Have temperatures hit an inflection point? These questions lead us to extend the scope of linear regression through variable transformations.

5. Born to Run Regressions: Who Likes What on Spotify. Sometimes, the answer is in the residuals.

6. What's Wrong With My Regression? Understanding Assumptions. Theoretical conditions for regression are like driving rules – you need to understand them to know when you can safely break them.

7. What Makes Lawyers (Un)happy? Regression Goes Logistic. Using logistic regression to understand differences in survey responses.

8. From Yelp Reviews to Restaurant Closures: Classification. Using logistic regression to predict restaurant closures from online reviews.

9. CART: Classification and Regression Trees. Translating data to decisions.

10. Forecasting Walmart's Sales. Decomposing time-series data into trend and seasonality.

11. Autoregressive Models. Serial correlation is everywhere. How to measure it, how to address it in forecasting.

12. Project Presentations