Data Analysis and Econometrics (Econ 132) Fall 2017:

Visualization, Causation, and Prediction

(As of 2017/10/5 with latest changes colored red; This syllabus will evolve along the way)

Class meets in WTS A60

Tuesday and Thursday 1:00-2:15 pm

TA section Thursday 7:00-8:00 pm

Instructor:

Yusuke Narita (yusuke.narita@yale.edu)

Room 38, 37 Hillhouse Avenue

Assistant: Brooke Williams (brooke.williams@yale.edu)

Office hours: Tuesday 3-4 pm or by appointment

TA:

Soumitra Shukla (soumitra.shukla@yale.edu)

Room 20, 37 Hillhouse Avenue

Office hours: Wednesday 4-5 pm or by appointment (place to be announced each week)

Overview

This course introduces ways to harness data to answer questions of social and intellectual interests. It combines tools used in a variety of fields (economics, statistics, machine learning, business) for a variety of goals (description/visualization, causal evaluation, and prediction with and without policy changes). This course aims for students to understand and critically discuss:

1) Econometric, statistical, and machine learning methods

2) Computational/algorithmic implementations of methods in programming languages (like Matlab, Python, R, and/or Stata)
3) Key applications (from a variety of domains like brain and cognitive sciences, education, development, health, industrial organization, labor, linguistics, marketing, politics, public finance, trade; see the bold questions in the following Roadmap)

Students will understand not only how methods work in theory but will also replicate and extend empirical applications by working with data by themselves.

Students will finish the course equipped with a workman's familiarity with the tools of data science, facility with data handling and statistical programming, and—hopefully—a good understanding of what questions you want to answer and how best to do it. That's a lot of ground to cover, so plan your time accordingly and be prepared to spend many hours every week.

Prerequisites

Students should be familiar with basic concepts in every one of probability, statistics, algebra, and calculus. I strongly suggest students to take econ 131 before this 132 course. Students not already proficient in a statistical programming language like Python, R, Stata, or Matlab will be required to learn and use a language in the first few weeks of the semester. You can find a list of code and data resources at the end of this syllabus.

Requirements and Grades

Five problem sets (mix of mathematical, programming, and empirical exercises): 40%

Midterm (closed-book): 30%

Final (closed-book): 30%

You are also asked to do a few hours of reading and "watching" (videos, podcasts, etc.) distributed every week. Their content will be asked for in problem sets and exams.

Key Dates (subject to changes)

August 31: First class meets

September 5: Problem set 1 handed out

September 14: Problem set 1 due, Problem set 2 handed out

September 28: Problem set 2 due, Problem set 3 handed out

October 12: Problem set 3 due

October 26: Midterm

October 31: Problem set 4 handed out

November 14: Problem set 4 due, Problem set 5 handed out

November 28: Problem set 5 due

December 17 (2pm): Final

Roadmap (subject to changes)

References with a star * are mandatory. Others without * are optional.

1. Introduction
   1.1. Reference: "Math is Music; Statistics is Literature (Or, Why are There No Six-year-old Novelists?)" (by De Veaux, College, and Velleman) *Amstat News*

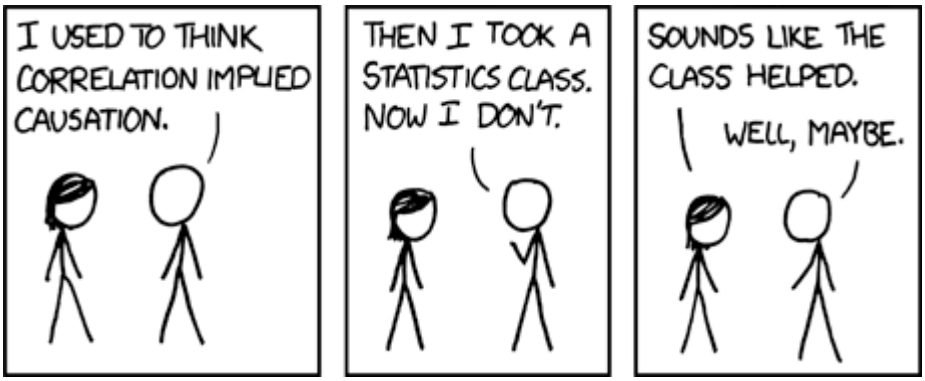
2. Description/Visualization of Data (About 3 lectures)
   2.1. *Reference: "An Economist's Guide to Visualizing Data" (by Schwabish) *Journal of Economic Perspective*
   2.2. Unstructured Data
      2.2.1. Application: Sounds and Videos --- **How Do We Learn to Speak?**
         2.2.1.1. Reference: *Reality Mining: Using Big Data to Engineer a Better World* (by Eagle and Green) MIT Press
         2.2.1.2. Reference: "Predicting the Birth of a Spoken Word" (by Roy, Frank, DeCamp, Miller, and Roy) *Proceedings of the National Academy of Sciences*
      2.2.2. Application: Brain --- **How Are Neurons Connected?**
         2.2.2.1. Reference: *Connectome: How the Brain's Wiring Makes Us Who We Are* (by Seung) Mariner Books
   2.3. Structured Data
      2.3.1. Conditional Expectation Function
      2.3.2. *Reference: "Introduction" and "Conditional Prediction" (by Manski) in *Identification for Prediction and Decision*, Harvard University Press
         2.3.2.1. Application: **Where is the American Dream?**
            2.3.2.1.1. Reference: "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States" (by Chetty, Hendren, Kline, and Saez) *Quarterly Journal of Economics*
      2.3.3. Linear Regression and Generalized Linear Models (e.g. Logit regression)
         2.3.3.1. Application: **Is Police Use of Force Racially Biased?**
            2.3.3.1.1. Reference: "An Empirical Analysis of Racial Differences in Police Use of Force" (by Fryer). Working Paper.

"We are now inside a Big Data model made with our new 7-D printer. The extra dimensions include time, money, number of Twitter followers, and total awesomeness."

3. Evaluation of Causality (About 15 lectures)

   Main reference for this segment: *Mastering 'Metrics: The Path from Cause to Effect* (by Angrist and Pischke) Princeton University Press

   3.1. Motivation: Correlation ≠ Causation



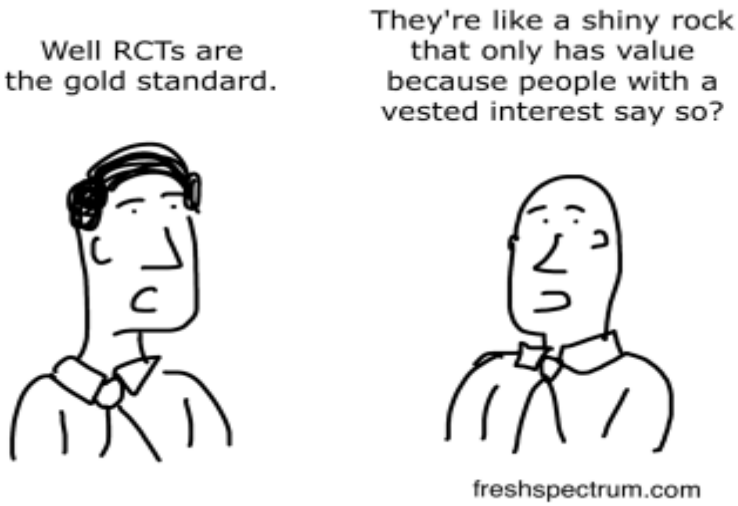
   3.2. Method: Causal effects & omitted variable bias (aka selection, endogeneity)

   3.3. Method: Randomized experiment & large-sample inference

      3.3.1. *Reference: Angrist and Pischke, chapter 1

      3.3.2. Application: **Are Online Ads Effective?**

         3.3.2.1. Reference: "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment" (by Blake, Nosko, Tadelis) *Econometrica*



3.4. Method: Natural experiment 1 – Selection on observables

3.4.1. *Reference: Angrist and Pischke, chapter 2

3.4.2. Application: **Are Expensive Private Colleges Worthwhile?**

3.4.2.1. Reference: "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables," (by Dale and Krueger) *Quarterly Journal of Economics*

3.5. Method: Natural Experiment 2 – Propensity Score

3.5.1. Application: **Are Charter Schools Better than Public Schools?**

3.5.1.1. Reference: "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation" (by Abdulkadiroglu, Angrist, Narita, and Pathak) *Econometrica*.

3.6. Method: Natural Experiment 3 – Instrumental Variables

3.6.1.1. *Reference: Angrist and Pischke. Chapter 3

3.6.1.2. Application: **Why Are There Rich and Poor Countries?**

3.6.1.2.1. Reference: "The Colonial Origins of Comparative Development: An Empirical Investigation" (by Acemoglu, Johnson, and Robinson) *American Economic Review*.

3.7. Method: Natural experiment 4 – Regression Discontinuity Design

3.7.1. *Reference: Angrist and Pischke. Chapter 4.

3.7.2. Sharp Regression Discontinuity as Compromised Selection-on-observables

3.7.3. Application: **Who Benefits from Electoral Voting?**

3.7.3.1. Reference: “Voting Technology, Political Responsiveness, and Infant Health: Evidence from Brazil” (by Fujiwara) *Econometrica*

3.7.4. Fuzzy Regression Discontinuity Design as Compromised IV

3.7.5. Application: **Does Class Size Matter?**

3.7.5.1. Reference: “Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement,” (by Angrist and Lavy) *Quarterly Journal of Economics*

3.7.6. Threats to Regression Discontinuity Designs

3.7.6.1. Reference: “Class-Size Caps, Sorting, and the Regression Discontinuity Design (by Urquiola and Verhoogen) *American Economic Review*

3.8. Method: Grouped/Panel Data Approach 1 – Fixed Effects

3.8.1. Application: **Is Schooling Worthwhile?**

3.8.1.1. Reference: “"Estimates of the Economic Return to Schooling from a New Sample of Twins" (by Ashenfelter and Krueger) *American Economic Review*

3.9. Method: Grouped/Panel Data Approach 2 – Difference-in-differences

3.9.1. *Reference: Angrist and Pischke. Chapter 5.

3.9.2. Application: **What Made Donald Trump the President?**

3.9.2.1. Reference: “The China Shock: Learning from Labor-Market Adjustment to Large Changes in Trade” (by Autor, Dorn, and Hanson) *Annual Review of Economics*

3.9.2.2. Reference: “A Note on the Effect of Rising Trade Exposure on the 2016 Presidential Election” (by Autor, Dorn, Hanson, and Majlesi) Working Paper.

3.9.2.3. Reference: “Social Media and Fake News in the 2016 Election” (by Allcott and Gentzkow) Working Paper.

3.10. Method: Grouped/Panel Data Approach 3 – Event Study

3.10.1. Application: **What Effects Did Trump’s Win Have on the Economy?**

3.10.1.1. Reference: "What Do Financial Markets Think of the 2016 Election?" (by Wolfers and Zitzewitz) Working Paper

3.11. Advanced Method: Exploiting Tiny Data via Exact (Permutation) Inference

3.11.1. *Reference: “Fisher’s Randomization Inference” (by Blackwell) http://www.mattblackwell.org/files/teaching/s05-fisher.pdf

3.11.2. Application: **Do Management Consultants Help?**

3.11.2.1. Reference: “Does Management Matter: Evidence from India” (by Bloom, Eifert, McKenzie, Mahajan, and Roberts) *Quarterly Journal of Economics*

3.12. Advanced Method: Ethics in Causal Inference & Experimental Design

3.12.1. Application: **How to Cure Opioid Addiction?**

3.12.1.1. Reference: “Algorithms for the Multi-Armed Bandit Problem” (by Kuleshov and Precup) *Journal of Machine Learning Research*

4. Prediction (a.k.a. Machine Learning, About 5 lectures)

Main reference for this segment: *Elements of Statistical Learning* (by Hastie, Tibshirani, and Friedman) https://statweb.stanford.edu/~tibs/ElemStatLearn/

4.1. Motivation: Causation in the past ≠ Good prediction for the future

4.1.1.1. *Reference: “To Explain or to Predict” (by Shmueli), *Statistical Science*

4.1.2. Application: **Do Classmates Matter?**

4.1.2.1. Reference: “From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation” (by Carrell, Sacerdote, and West), *Econometrica*

4.2. Method: Predictive Regression with Regularization

4.2.1. *Reference: "High-Dimensional Methods and Inference on Structural and Treatment Effects" (Belloni, Chernozhukov, and Hansen) *Journal of Economic Perspectives*

4.2.2. Application: **How to Sort Out Annoying Spam Emails?**

4.2.3. Application: **What Made People Watch Movies?**

4.2.3.1. Reference: “Something to Talk About: Social Spillovers in Movie Consumption” (by Gilchrist and Sands) *Journal of Political Economy*

4.3. Method: Neural Networks

4.3.1. *Reference: “Neural Networks” chapter 13 of *Elements of Statistical Learning* (by Hastie, Tibshirani, and Friedman) https://statweb.stanford.edu/~tibs/ElemStatLearn/

4.3.2. Application: **Where is a Cat?**

4.3.2.1. Reference: “ImageNet Classification with Deep Convolutional Neural Networks” (by Krizhevsky, Sutskever, and Hinton) *Advances in Neural Information Processing Systems 25 (NIPS 2012)*

4.3.3. Application: **How to Predict Poverty?**

4.3.3.1. Reference: "Combining Satellite Imagery and Machine Learning to Predict Poverty" (by Jean, Burke Xie, Davis, Lobell, and Ermon) *Science*

4.4. Method: Trees and Forests

4.4.1. *Reference: "Tree-Based Methods" chapter 8 of *An Introduction to Statistical Learning* (by James, Witten, Hastie, and Tibshirani)

4.4.2. Application: **How to Better Sort Out Annoying Spam Emails?**

5. Harder Predictions with Policy Changes (a.k.a. Counterfactual, About 2 lectures)

Main reference for this segment: *Identification for Prediction and Decision* (by Manski) Harvard University Press

5.1. Motivation: ML prediction ≠ Every prediction

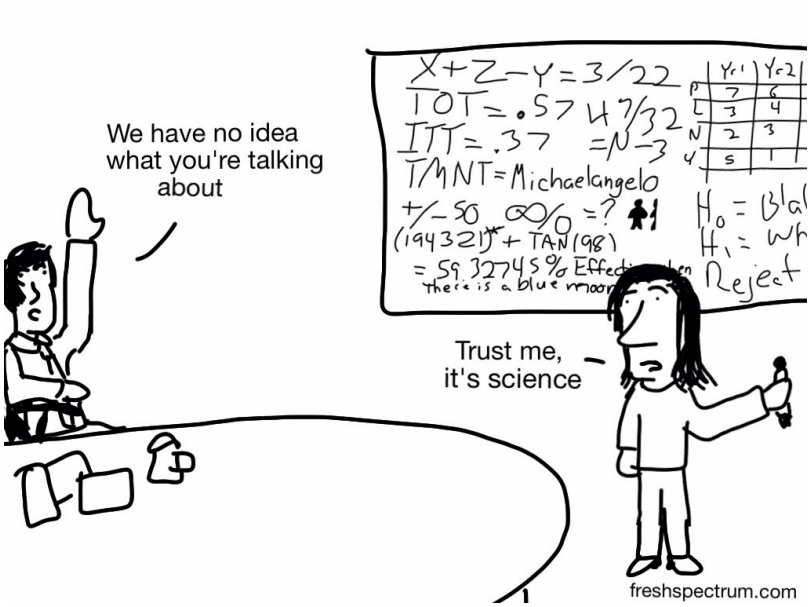
5.2. Method: Structural modelling of human behavior

5.2.1. Reference: "Econometrics" (by Varian) chapter 12 of *Microeconomic Analysis* (3rd Edition), W. W. Norton

5.3. Method: Random utility demand models

5.3.1. *Reference: Reference: "Revealed Preference Analysis" and "Studying Human Decision Processes" (by Manski) chapters 13 and 15 of *Identification for Prediction and Decision*, Harvard University Press

5.3.2. Application: **Who'd use a new train line?**



Fun Background Reading

The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century (by Salsburg) Great history of statistics

Freakonomics: A Rogue Economist Explores the Hidden Side of Everything (by Levitt and Dubner) Examples of idea-driven data analyses in academia

Moneyball: The Art of Winning an Unfair Game (by Lewis) Examples of data-driven data analysis in the sports industry

Further References (Use them only as "dictionaries")

Introductory Econometrics: A Modern Approach (by Wooldridge) South-Western College Pub

Introduction to Econometrics (by Stock and Watson) Pearson

Elements of Statistical Learning (by Hastie, Tibshirani, and Friedman) https://statweb.stanford.edu/~tibs/ElemStatLearn/

Computer Age Statistical Inference: Algorithms, Evidence and Data Science (by Efron and Hastie) http://web.stanford.edu/~hastie/CASI

Resources on General Coding and Data Management

Learn Enough to Be Dangerous series (by Hartl):

1) Command Line (https://www.learnenough.com/command-line-tutorial)
2) Version Control (https://www.learnenough.com/git-tutorial)
3) Text Editor (https://www.learnenough.com/text-editor-tutorial)

"Code and Data for the Social Sciences: A Practitioner's Guide" (by Gentzkow and Shapiro, http://web.stanford.edu/~gentzkow/research/CodeAndData.xhtml)

Resources on Programming Languages

Any of the following four languages is appropriate for this course. Python and R are for free, but Matlab and Stata are not. Please speak to me or the TA if you would like advice on which to use. CodeAcademy, DataCamp, Coursera, EdX, and Lynda.com have good resources on some or all of these languages.

Matlab (based on the matrix data structure, widely used for numerical methods in engineering and macroeconomics)

- *Matlab: An Introduction with Applications* (by Gilat) Wiley
- Online tutorial *Learn Matlab Basics* (https://www.mathworks.com/support/learn-with-matlab-tutorials.html?requestedDomain=www.mathworks.com)

Python (full-fledged programming language, used for data analytics too)

- *Quantitative Economics* (by Sargent and Stachurski) http://lectures.quantecon.org/py/index.html
- *Introduction to Python for Econometrics, Statistics and Data Analysis* (by Sheppard) https://www.kevinsheppard.com/images/0/09/Python_introduction.pdf
- Online tutorial *Learn Python the Hard Way* (by Shaw) https://learnpythonthehardway.org/

R (designed from the ground-up for data analysis, especially popular in statistics and machine learning)

- "Econometrics in R" (by Farnsworth, https://ocw.mit.edu/courses/economics/14-381-statistical-method-in-economics-fall-2013/study-materials/MIT14_381F13_EcnomtrisInR.pdf)
- Online tutorial *Swirl* (http://swirlstats.com)
- Online tutorial *R Programming* (by Peng, Leek, and Caffo, https://www.coursera.org/learn/r-programming)
- Online tutorial *Introducing R* (by Rodriguez, http://data.princeton.edu/R)
- See also *R Inferno* (http://www.burns-stat.com/documents/books/the-r-inferno/) to understand common problems in R

Stata (designed from the ground-up for data analysis, popular (only) in economics and other social sciences)

- *Microeconometrics using Stata* (by Cameron and Trivedi) Stata Press
- *Statistics with Stata* (by Hamilton) Duxbury Press
- Online tutorial *Stata Tutorial* (by Rodriguez, http://data.princeton.edu/stata/)

Coding Tutors @ Yale

Saul Downie (saul.downie@yale.edu)
Blue Dog Cafe in HGS
Monday: 7:30 - 9:30 pm, Wednesdays: 7:30 - 9:30 pm, Friday: 3:00 - 4:00 pm

Daniela Aizencang (daniela.aizencang@yale.edu)
Center for Teaching and Learning in Sterling
Tuesday: 9:30 - 12:00 pm, Thursday: 9:30 - 12:00 pm

StatLab Statistical Consulting (http://csssi.yale.edu/data-and-gis/csssi-statisical-consulting/csssi-statistical-and-gis-consultants)

Yale Center for Research Computing Bootcamp - Writing Efficient R Code (http://research.computing.yale.edu/training/hpc-bootcamp-writing-efficient-r-code)

Yale Center for Research Computing Bootcamp - Scripting with Python (http://research.computing.yale.edu/hpc-bootcamp-scripting-python)