PLSC 438/536: Applied Quantitative Research Design

Instructor: Prof. Shiro Kuriwaki*

Fall 2022	Yale University
Professor	Shiro Kuriwaki
Teaching Fellows	Sam Zacher (Head TF)
	Natalie Hernandez
Lecture	M/W 10:30-11:45
Review Section	W, 50 minutes

Note: This is a public version of the Fall 2022 course syllabus. It focuses on course organization and readings, while omitting details on university-specific information on dates, accommodations, and logistical notes. Future versions of the class will differ.

Course Description

Research designs are strategies to obtain empirical answers to theoretical questions. Research designs using quantitative data for social science questions are more important than ever. This class, intended for advanced students interested in social science research, trains students with best practices for designing and implementing rigorous quantitative research. We cover designs in causal inference, prediction, and missing data at a high level. This is a hands-on, application-oriented class. Exercises involve programming and statistics in addition to the social sciences (politics, economics, and policy). The final project advances a research question chosen in consultation with the instructor.

Prerequisite: Any statistics or data science course that teaches ordinary least squares regression. Past or concurrent experience with a programming language such as R is strongly recommended. Students with no prior R experience should plan on attending extra practice sessions in the first few weeks.

Assessment

The course grade will be a weighted average of the following components:

*Assistant Professor of Political Science, Yale University. https://www.shirokuriwaki.com. This syllabus borrows from class material designed by Dan Levy, Teddy Svoronos, Matt Blackwell, and Kosuke Imai. I thank them for sharing some of their course material. If you would to see any course material including problem sets, please contact me.

- Ten weekly or semi-weekly problem sets: 40%
- Stats quiz in last week of class: 5%
- Final Paper: 40%
- Participation, including pre-class reading responses for papers: 15%

Problem sets are due on Saturday the week of the content is covered in lecture as a general rule. Most problem sets will have 3 parts covering 8-10 exercises. These cover visualization, table construction, replicating statistical analyses of the papers we discuss, and discussion questions from the papers.

SCHEDULE

This course revolves around the close reading, discussion, and data analysis of the following seven academic papers and case studies:

Paper	Topics Covered
Banerjee, Abhijit, Esther Duflo, Dean Karlan, *et al.* "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." (2015) *Science*. [Dataverse]	RCT, Regression, non-compliance
Malhotra, Neil, Yotam Margalit, and Cecilia Mo, "Economic Explanations for Opposition to Immigration: Distinguishing between Prevalence and Conditional Impact." (2013). *American Journal of Political Science*. [Dataverse]	Observational regression
Miguel, Edward, Shanker Satyanath, and Ernest Sergenti (2004). "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." (2004) *Journal of Political Economy.* [Replication]	Instrumental Variables
Scheve, Kenneth and David Stasavage, "Democracy, War, and Wealth: Lessons from Two Centuries of Inheritance Taxation." (2012) *American Political Science Review*. [Replication]	Panel Data, Difference in differences, Clustered SEs
Wang, Yuhua, "Blood Is Thicker Than Water: Elite Kinship Networks and State Building in Imperial China." (2022) *American Political Science Review* [Dataverse]	Network analysis
Toeffel, Michael, Dan Levy, et al. "Improving Worker Safety in the Era of Machine Learning." (2018). HBS Case	Out of sample prediction
Cohn, Nate, "We Gave Four Good Pollsters the Same Raw Data. They Had Four Different Results." (2016), New York Times Upshot Case. [Data]	Survey weighting

We will also refer to the following textbooks:

- Our main textbook is: Kosuke Imai and Nora Webb Williams, *Quantitative Social Science: An Introduction in tidyverse*, 2022.
- We will also use parts of: Ethan Bueno de Mesquita and Anthony Fowler, *Thinking Clearly with Data: A Guide to Quantitative Reasoning and Analysis*, 2022
- As a reference for econometrics methods and implementation, we recommend *Mastering 'Metrics: The Path from Cause to Effect* or *Mostly Harmless Econometrics*, both by Joshn Angrist and Jörn-Steffen Pischke
- The main methods we will implement are also provided as 3-5 minute screencasts that students can play at their convenience: `https://vimeo.com/shirokuriwaki`

The methods we will cover is organized in three major components: Causal inference, prediction, and uncertainty, in that order. Some of the other weeks focus on using R and writing academic papers. The detailed schedule below lists topics, readings, and assignments for each class. QSS refers to the Imai text and TCD refers to Bueno de Mesquita and Fowler. Lectures are numbered by week and number.

Class	Topic	Detailed Topic and Readings	Due
0.1	**Course overview**		Before class, install R and RStudio (R screencast)
0.2	**Workflow**	RStudio Projects, tidyverse, dplyr, pipes • R for Data Science 2e, 9 "Workflow: Scripts and Projects"	
1.1	**Visualizing Data**	Prioritization of aesthetics (per Cleveland), grammar of graphics with ggplot • Rauser, "How Humans See Data" (talk)	Pset 1 (RStudio project, code style, dplyr, histograms)
2.1	**Randomized Control Trials**	Linear regression as difference in means, Estimating treatment effects with linear regression • QSS 2.1-2.4 "Causal Effects in the Counterfactual," "Randomized Controlled Trials"; 4.4 "Regression - Randomized Control Trials"	

Continued on the next page

2.2	**Discuss Banerjee et al.**	"A multifaceted program causes lasting progress for the very poor: Evidence from six countries"
		Pset 2 (analyze Banerjee *et al.* with OLS, barplots, regression tables)
3.1	**Confounding and Omitted Variable Bias**	Using regression in observational data for causation • QSS 2.5.2 "Confounding Bias", 4.3.2 "Regression with multiple predictors" • TCD 9 "Why Correlation Doesn't Imply Causation" and 10 "Controlling for Confounders"
3.2	**Units in regression**	Summarizing variables, substantive interpretation of regression coefficients, implications for changing the units of the left-hand and right-hand side
		Pset 3 (replicate Malhotra *et al.*, shapefiles, recoding values, standardizing variables)
4.1	**Discuss Malhotra et al.**	"Economic Explanations for Opposition to Immigration: Distinguishing between Prevalence and Conditional Impact."
4.2	**Non-compliance**	Never-takers, compliers, Treatment on Treated vs. Intent to treat • TCD 11 "Randomized Experiments", p. 225-231 "Noncompliance and instrumental variables"
		Pset 4 (Review of OVB in Malhotra *et al.*, for loops with `map_dfr`, TOT vs. ITT)
5.1	**Instrumental variables**	Discuss Miguel, Satyanath, Sargenti, "Economic Shocks and Civil Conflict: An Instrumental Variables Approach."
5.2	**Panel Data**	Long and wide data, reshaping with `pivot_longer` and `pivot_wider`
		Pset 5 (replicating Miguel et al. IV; line plots of trend data) (R screencast)

Continued on the next page

6.1	**Fixed Effects**	Time-invariant and unit-invariant confounding, logic of fixed effects, implementation in R with `fixest` • QSS 2.5 "Observational studies" • TCD 13 "Difference-in-Differences Designs"
6.2	**Discuss Scheve and Stasavage**	"Democracy, War, and Wealth: Lessons from Two Centuries of Inheritance Taxation." Parallel trends, Connection between TWFE and DID, unit-specific time trends
6.3	**Synthetic Control**	Pset 6 (Replicating Scheve and Stasavage TWFE, leads and lags) R screencast
7.1	**Regression for Prediction**	Prediction with new data `predict`, MSE, RMSE, and R^2 • QSS 4 "Prediction", especially 4.1-4.2
7.2	**Shrinkage**	Out of sample prediction, Overfitting, LASSO regression, tuning parameters • Mullainathan and Spiess (2017), "Machine Learning: An Applied Econometric Approach" *Journal of Economic Perspectives* Pset 7 (OSHA case, pre-case prediction) (R screencast)
8.1	**Discuss HBS Case, Worker Safety at OSHA**	"Improving Worker Safety in the Era of Machine Learning"
8.2	**Networks**	Wang, "Blood is Thicker Than Water: Elite Kinship Networks and State Building in Imperial China." • QSS 5.2 "Network Data" Pset 8 (networks with `igraph`, replicating Wang article)
9.1	**Survey Sampling**	Survey recruitment, sources of error, cross-tabulation, ratio estimator for unrepresentativeness

Continued on the next page

9.2	**Survey Weighting**	Dependence and independence, rake weighting, outcome models and weighting • "We Gave Four Good Pollsters the Same Raw Data. They Had Four Different Results."
10.1	**Probability**	Definition of probability, distributions. • QSS 6.1 "Probability", 6.2 "Conditional Probability"
10.2	**Random Variables**	Random variables as functions, Bernoulli r.v., Linearity of Expectation, Variance. • QSS 6.3 "Random Variables"
11.1	**Standard Errors**	Pset 9 (survey re-weighting, *Upshot* Florida 2016 replication) Central Limit Theorem, Normal distribution, Z-scores, Confidence intervals • QSS 7 "Uncertainty"
11.2	**Standard Errors in Regression**	SEs of Regression Coefficients, Clustered Standard Errors • QSS 7 "Uncertainty"
12.1	**In-class stats quiz**	
12.2	**Writing a research paper**	In-class exercise: Re-read • Abstract and intro of Wang (2022) • Methods and results section of Scheve and Stasavage Read / view later: • McEnery, "The Craft of Writing Effectively" (video lecture) • King, "Publication, Publication" • Fiske, "Words to the Wise on Writing Scientific Papers" Pset 10 (clustered SEs, writing structure)

Final Project

Assignment: Present original research findings, either by extending one of the papers we read in the class OR (with permission of the instructor) by writing a research paper on another topic. There are three prototypes of projects you can choose between (i.e., choose just one of these):

1. A replication of one of our course papers accompanied by an extension (e.g. with a different estimation approach or a different quantity of interest). See the end of this document for instructions specific to replication.
2. An analysis of a new dataset that asks a similar or related question as one of our course papers.
3. An analysis of an entirely different social science question using similar quantitative research designs, upon permission of the instructor. This is typically reserved for PhD students or students writing a senior essay.

Details about the paper requirements are provided in a separate handout.