PLSC30500, Fall 2023

Part 0. Introduction

Andy Eggers

Welcome!

This course

- Instructor: Andy Eggers
- Teaching assistant: Moksha Sharma
- part of a sequence:
  - Intro to Quant Soc Sci **(this course)** (fall)
  - Causal Inference (winter)
  - Linear Models (spring)

Our objectives

- give a strong foundation for further study
- give a taste of what is fun about quantitative social science
  - mathematical rigor and clarity
  - thinking about estimation, uncertainty, causality

Broad plan

Five modules:

- Probability (1.1, 1.2, 2.1, 2.2)
- Summarizing distributions (3.1, 3.2, 4.1, 4.2)
- Estimation (5.1, 5.2, 6.1)
- Inference (6.2, 7.1, 7.2)
- Regression (8.1, 8.2, 9.1, 9.2)

Expectations about background

Useful (not required) to have exposure to

- math (semi-recently)
- probability & statistics
- econometrics/regression modeling
- programming

If you have don't have much exposure to X, you may have to work harder on X. If you have lots of exposure to all of the above, we believe you can still learn something.

Expectations for the course

- Read the syllabus (link from Canvas page and Github)
- Prepare for class: attempt the main reading (Aronow & Miller); ask for easier readings if necessary
- If you are stuck on reading/assignments:
  1. Use google first, or e.g. ChatGPT
  2. Ask your question on our private StackOverflow (https://stackoverflowteams.com/c/uchicagopolmeth)
  3. Or if you're brave, ask on the real StackOverflow (https://stackoverflow.com/) if it's about R or CrossValidated (https://stats.stackexchange.com/) if it's about stats.
- If you are confused in class, ask a question

Please also *answer* questions on our private StackOverflow.
If you email me a question, I am likely to tell you to put it on our StackOverflow.

Labs

Taught by Moksha, Fridays, Cobb 301.

- Lab 1: 12:30-1:20
- Lab 2: 1:30-2:20

Assessment

- 40% problem sets (8 in all)
- 10% class participation
- 20% in-class midterm on October 19
- 30% final take-home exam due December 5

Websites

All slides and assignments will be distributed via the course Github:
https://github.com/UChicago-pol-methods/IntroQSS-F23
Download files one by one, or `git clone` and frequently update.
Homework submission via Canvas page.

Technical setup

By lab on Friday (ideally sooner), make sure you do this:

1. install R from https://cran.rstudio.com/
2. Install RStudio from https://www.rstudio.com/products/rstudio/download/
3. In RStudio install `tidyverse` and `tinytex`

If you can "knit" the first homework (`ps1_2023_probability.qmd`) into a PDF, you are all set.

Motivation

What most applied social scientists "know" about statistics

Most social scientists "know" a few things about

- Linear regression (OLS) and two other estimation techniques (logit, probit)
- **Statistical inference** (standard errors, p-values, null hypothesis)

That's it.

What most applied social scientists "know" about linear regression (OLS)

- We use **regression** (ordinary least squares, OLS) to measure relationships between a **dependent variable** (DV, left-hand-side (LHS) variable, outcome variable) Y and **independent variables** (right-hand-side (RHS) variables, covariates, predictors) X_1, X_2, X_3, etc: e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$
- We call the estimated "effect" of each variable a regression **coefficient**
- For regression to work, you need a lot of assumptions, e.g. relationships have to be linear, the error term (or the dependent variable) needs to be normally distributed
- A regression coefficient for X_1 measures how much Y is predicted to change with a **one-unit increase** in X_1, holding fixed X_2, X_3, etc
- Sometimes this coefficient is a good estimate of the **(causal) effect** of X_1 on Y, i.e. what would happen if you changed X_1
- You can use an **interaction term** to get a coefficient that measures how the "effect" of X_1 depends on the value of X_2:
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 + \ldots$

A regression table

WHY DO CONGRESSWOMEN OUTPERFORM CONGRESSMEN?

TABLE 2 Evidence of Sex-Based Selection

	District Ideology	District & Member Ideology	Widows
	(1)	(2)	(3)
Female	0.12	0.106	
	(0.048)**	(0.046)**	
Female * Constituent Ideology	0.584	0.656	
	(0.205)***	(0.193)***	
Member Ideology		−0.426	
		(0.098)***	
Female Nonwidow			0.138
			(0.068)**
Widows			−0.104
			(0.119)
Constant	18.817	18.764	13.814
	(2.388)***	(2.409)***	(1.633)***
Observations	7404	7404	9067
R-squared	0.89	0.89	0.66
Fixed Effects	District & year	District & year	State & year
F-test: Widows = Nonwidows			p = 0.054*

What most applied social scientists "know" about other estimation techniques

- When the dependent variable is **binary** (i.e. only 0 or 1), you shouldn't use OLS
- Instead you should use **logit** or **probit**
- Logit and probit coefficients are hard to understand

What most applied social scientists "know" about statistical inference

- Statistical software gives you a **standard error** for each regression coefficient. A bigger standard error means we are more uncertain what that coefficient really is.
- The **null hypothesis** is usually the claim that there is no relationship. We do **hypothesis testing** to see if we can reject the null hypothesis.
- If the **p-value** on your coefficient is below .05, your coefficient is **statistically significant** and you can **reject the null hypothesis**. This means the coefficient probably isn't zero because the relationship is unlikely to have occurred by chance. If you get a p-value above .05, you didn't find anything and your analysis didn't work.

What we want you to know

You need to know what is above – at least, the correct parts! (e.g. reading regression table, interpreting interaction terms)
But also, we want you

- to avoid the misconceptions (the stuff in orange and red)
- to understand common approaches to uncertainty (e.g. standard errors) and hypothesis testing (e.g. p-values): the logic behind these approaches and their limits.
- to see what coin flips and urns have to do with social science.