Introductions and Overview

Dr. Adam Soliman

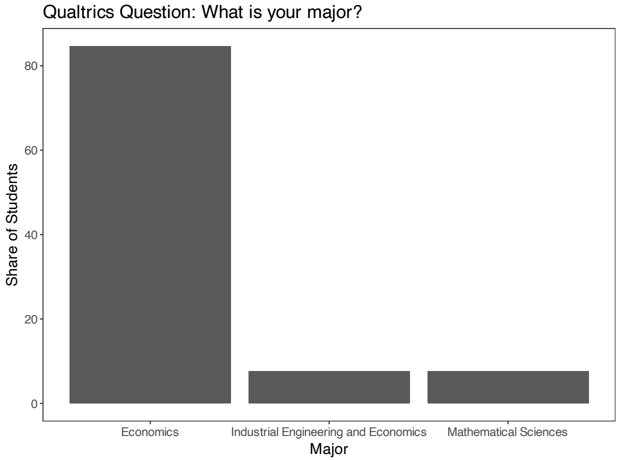
Introduction to Econometrics (ECON 4050)
Clemson University
Spring 2025

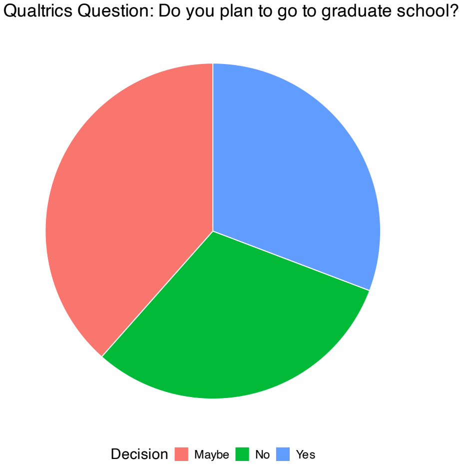
Outline

A little bit about your econometrics professor

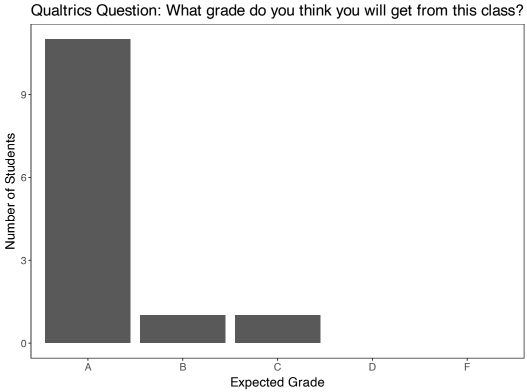
- I have been in school for a while...from Boston University (undergrad) to Michigan State (masters) to Duke (PhD) to London School of Economics (post-doc)
- My first job was teaching math and economics in Dubai
- I study topics in the economics of crime & some research questions of interest include
  1. What are the impacts of cracking down on rogue doctors during the opioid epidemic on street drug prices, overdose mortality, and other doctors behavior?
  2. Do police respond to changes in punishment severity?
  3. What happens to neighborhood crime when investors buy many properties?

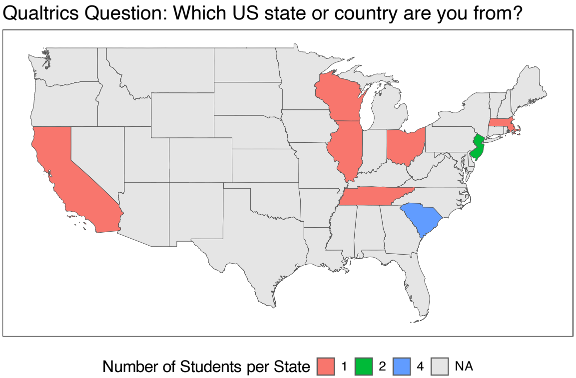
A little bit about your classmates (Major and Grad School)





A little bit about your classmates (Grade and Location)





What is econometrics?

The statistical toolkit (techniques and methods) used to answer economic questions with data

What is econometrics?

The statistical toolkit (techniques and methods) used to answer economic <u>questions</u> with <u>data</u>

What types of questions might we be interested in:

- Has economic inequality increased since 1960?
- Does raising the minimum wage reduce employment for low-skilled workers?
- What will the unemployment rate be next quarter?

What is econometrics?

The statistical toolkit (techniques and methods) used to answer economic <u>questions</u> with <u>data</u>

What types of questions might we be interested in:

- Has economic inequality increased since 1960?
  - **Descriptive Question:** asks about how things are (or were) in reality
- Does raising the minimum wage reduce employment for low-skilled workers?
- What will the unemployment rate be next quarter?

What is econometrics?

The statistical toolkit (techniques and methods) used to answer economic <u>questions</u> with <u>data</u>

What types of questions might we be interested in:

- Has economic inequality increased since 1960?
  - **Descriptive Question:** asks about how things are (or were) in reality
- Does raising the minimum wage reduce employment for low-skilled workers?
  - **Causal Question:** What would have happened in a counterfactual world?
- What will the unemployment rate be next quarter?

What is econometrics?

The statistical toolkit (techniques and methods) used to answer economic <u>questions</u> with <u>data</u>

What types of questions might we be interested in:

- Has economic inequality increased since 1960?
  - **Descriptive Question:** asks about how things are (or were) in reality
- Does raising the minimum wage reduce employment for low-skilled workers?
  - **Causal Question:** What would have happened in a counterfactual world?
- What will the unemployment rate be next quarter?
  - **Forecasting Question:** What will happen in the future?

Causal questions will be our focus in this course

More causal questions:

1. Does immigration *lead to* lower wages and/or higher unemployment for locals?
2. Does getting a college degree *afford* higher wages?
3. Do higher public debt levels *lead to* lower economic growth?
4. Does the neighborhood you grew up in have an *impact* on your life outcomes?

Causal questions will be our focus in this course

More causal questions:

1. Does immigration *lead to* lower wages and/or higher unemployment for locals?
2. Does getting a college degree *afford* higher wages?
3. Do higher public debt levels *lead to* lower economic growth?
4. Does the neighborhood you grew up in have an *impact* on your life outcomes?

- Note that many other factors could have caused each of these outcomes
- Often, we'll want to focus on the causal impact of just one of these factors (immigration, minimum wage, education, etc.)
- Econometrics is about spelling out *conditions* under which we can *claim to measure causal relationships*
- We will encounter most basic of those conditions, and talk about some potential pitfalls

Expectations

- Ask questions: to me, your TA, each other
- Make mistakes, but always try
- You will be learning a powerful set of skills that apply well outside of this course, so enjoy where you can and try to think of where else they may be useful

Logistics

1) Schedule and Location:
   - Lectures: 4050-001 TR from 11:00AM to 12:15PM in Brackett Hall 111
   - Labs: 4051-001 T 5:30PM to 8:30PM or ECON 4051-002 Tuesday 5:30PM-8:30PM in Powers Hall 112
   - My office hours are on Zoom, please sign up in advance for a slot here
   - Our Teaching Assistant (TA) is Haoran Li and will hold office hours (TBD)
2) Content:
   - The main course materials are currently posted on the Github course website. Other communication will be via Canvas or Email.
3) Software:
   - R for statistical analyses (to be covered in Lab sessions and throughout the course)

Logistics

4) Assessments:
   - Assignments (Course and Lab) 20%
   - Coding Midterm (Take Home) 25%
   - Theory Exam (In Class) 20%
   - Final project and associated presentation: 30% and 5%, respectively
5) Prerequisites:
   - ECON2110 and 2120 Principle of Microeconomics and Macroeconomics
   - MATH1080 Calculus One
   - STAT3090/MATH3090 Introduction to Statistics

Resources

- **Econometrics**
  - SciencesPo Online Book
  - Causal Inference: The Mixtape by Cunningham
  - Ben Lambert's youtube channel
- **Metrics and 'R'**
  - ModernDive
  - Introduction to Econometrics with R
  - Awesome R Learning Resources
  - R for Data Science

Outline

Why is answering these questions hard?

- For descriptive questions: we only observe data for a **sample** of individuals, not for the full **population**
  - Example: we want to know how the distribution of income in the US has changed, but we only observe income for a survey of workers

Why is answering these questions hard?

- For descriptive questions: we only observe data for a **sample** of individuals, not for the full **population**
  - Example: we want to know how the distribution of income in the US has changed, but we only observe income for a survey of workers
- Best case scenario: Our sample is **randomly** selected from the population
  - For example, workers in the survey were drawn out of hat with names of all possible workers

Why is answering these questions hard?

- For descriptive questions: we only observe data for a **sample** of individuals, not for the full **population**
  - Example: we want to know how the distribution of income in the US has changed, but we only observe income for a survey of workers
- Best case scenario: Our sample is **randomly** selected from the population
  - For example, workers in the survey were drawn out of hat with names of all possible workers
- Worst case scenario: Our sample is *not representative* of the population that we care about
  - For example, workers with certain characteristics were more likely to respond to the survey

An example from before we were born

- In 1948, Chicago Tribune writes that Thomas Dewey defeats Harry Truman in the 1948 presidential election, based on survey of voters.

An example from before we were born $\Longrightarrow$ error known as selection bias

- In 1948, Chicago Tribune writes that Thomas Dewey defeats Harry Truman in the 1948 presidential election, based on survey of voters.
- But their survey was conducted by phone. In 1948, only rich people had phones: sample $\neq$ population $\Longrightarrow$ misleading results!

Why is answering these questions hard? (Part II)

- Answering causal questions is often *even harder* than descriptive ones because they involve <u>both</u> a descriptive component (what are outcomes in reality?) and a *counterfactual* component (how would things have been under a different treatment?)

Why is answering these questions hard? (Part II)

- Answering causal questions is often *even harder* than descriptive ones because they involve <u>both</u> a descriptive component (what are outcomes in reality?) and a *counterfactual* component (how would things have been under a different treatment?)
- Example: what is the causal effect on your earnings of going to Clemson instead of USC?
  - Descriptive Question: how much do Clemson students earn after graduation?
  - Counterfactual Question: how much would Clemson students have earned if they went to USC?
- Counterfactual questions can't ever be answered with data alone. Need additional assumptions to learn about them!

Splitting up the problem

- When thinking about causal questions, it's often easier to split the problem in two
- **Identification:** what could we learn about the parameters we care about (causal effects) if we had the observable data for the entire population
  - Need to make assumptions about how observed outcomes relate to outcomes that would have been realized under different treatments
- **Statistics**: what can we learn about the full population that we care about from the finite sample that we have?
  - Need to understand the process by which our data is generated from the full population

Framework for thinking about these steps

- **Sample:** the data that you actually observe
  - A survey of students from Clemson and USC graduates about their earnings
- **Estimator:** a function of the data in the sample
  - Difference in earnings between Clemson and USC students in survey
- **Estimand:** a function of the observable data for the *population*
  - Difference in earnings between all Clemson and USC students
- **Target (aka structural) parameter:** what we actually care about
  - Causal effect on earnings of going to Clemson relative to USC

Framework for thinking about these steps

- **Sample:** the data that you actually observe
  - A survey of students from Clemson and USC graduates about their earnings
- **Estimator:** a function of the data in the sample
  - Difference in earnings between Clemson and USC students in survey
- **Estimand:** a function of the observable data for the *population*
  - Difference in earnings between all Clemson and USC students
- **Target (aka structural) parameter:** what we actually care about
  - Causal effect on earnings of going to Clemson relative to USC
- The process of learning about the *estimand* from the estimator constructed with your *sample* is called **statistical estimation/inference**
- The process of learning about the *parameter* from the *estimand* is called **identification**

Let's add some math...by introducing potential outcomes notation

- D_i = indicator if get treatment (1 if Clemson, 0 if USC)
- $Y_i(1)$ = outcome under treatment = earnings at Clemson
- $Y_i(0)$ = outcome under control = earnings at USC

- Observed outcome Y_i is $Y_i(1)$ if $D_i = 1$ and $Y_i(0)$ if $D_i = 0$. (Y_i is your actual earnings)
- We can write the observed outcome as $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$

What is the causal effect on earnings of going to Clemson instead of USC?

- **Sample**: (Y_i, D_i) for $i = 1, ...N$. Data with earnings and where you went to school

What is the causal effect on earnings of going to Clemson instead of USC?

- **Sample**: (Y_i, D_i) for $i = 1, ...N$. Data with earnings and where you went to school
- **Estimator**: Difference in sample mean of earnings for people who went to Clemson and people who went to USC:

$$\underbrace{\frac{1}{N_1}\sum_{i:D_i=1} Y_i}_{\text{Avg earnings at Clemson in sample}} - \underbrace{\frac{1}{N_0}\sum_{i:D_i=0} Y_i}_{\text{Avg earnings at USC in sample}}$$

What is the causal effect on earnings of going to Clemson instead of USC?

- **Sample**: (Y_i, D_i) for $i = 1, \ldots N$. Data with earnings and where you went to school
- **Estimator**: Difference in sample mean of earnings for people who went to Clemson and people who went to USC:

$$\underbrace{\frac{1}{N_1} \sum_{i:D_i=1} Y_i}_{\text{Avg earnings at Clemson in sample}} \quad - \quad \underbrace{\frac{1}{N_0} \sum_{i:D_i=0} Y_i}_{\text{Avg earnings at USC in sample}}$$

- **Estimand**: Difference in population mean of earnings for people went to Clemson and people who went to USC:

$$\underbrace{E[Y_i | D_i = 1]}_{\text{Avg earnings at Clemson in population}} \quad - \quad \underbrace{E[Y_i | D_i = 0]}_{\text{Avg earnings at USC in population}}$$

What is the causal effect on earnings of going to Clemson instead of USC?

- **Sample**: (Y_i, D_i) for $i = 1, ...N$. Data with earnings and where you went to school
- **Estimator**: Difference in sample mean of earnings for people who went to Clemson and people who went to USC:

$$\underbrace{\frac{1}{N_1}\sum_{i:D_i=1} Y_i}_{\text{Avg earnings at Clemson in sample}} - \underbrace{\frac{1}{N_0}\sum_{i:D_i=0} Y_i}_{\text{Avg earnings at USC in sample}}$$

- **Estimand**: Difference in population mean of earnings for people went to Clemson and people who went to USC:

$$\underbrace{E[Y_i|D_i = 1]}_{\text{Avg earnings at Clemson in population}} - \underbrace{E[Y_i|D_i = 0]}_{\text{Avg earnings at USC in population}}$$

- **Target parameter**: Causal effect of Clemson for Clemson students:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Clemson for Clemson students in pop}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at USC for Clemson students in pop}}$$

Why is causal identification hard?

- Thought experiment: suppose we had data on earnings for *every* Clemson and USC graduate
- We can learn from the data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Clemson for Clemson Students}} \quad \text{and} \quad \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at USC for USC students}}$$

Why is causal identification hard?

- Thought experiment: suppose we had data on earnings for *every* Clemson and USC graduate

- We can learn from the data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Clemson for Clemson Students}} \quad \text{and} \quad \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at USC for USC students}}$$

- The causal effect of Clemson for Clemson students is

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Clemson for Clemson Students}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at USC for Clemson Students}}$$

Why is causal identification hard?

- Thought experiment: suppose we had data on earnings for *every* Clemson and USC graduate

- We can learn from the data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Clemson for Clemson Students}} \quad \text{and} \quad \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at USC for USC students}}$$

- The causal effect of Clemson for Clemson students is

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Clemson for Clemson Students}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at USC for Clemson Students}}$$

- The data doesn't tell us $\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at USC for Clemson Students}}$. Why not?

- One idea to solve this problem would be to assume that:

$$\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at USC for Clemson Students}} = \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at USC for USC Students}}$$

- Why might this give us the wrong answer?

- One idea to solve this problem would be to assume that:

$$\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at USC for Clemson Students}} = \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at USC for USC Students}}$$

- Why might this give us the wrong answer?
- Because Clemson students may be different from USC students in other ways that would affect their earnings (regardless of where they went to college)
  - Academic ability, family background, career goals, etc.
- These differences are referred to as *omitted variables* or *confounding factors*

What about experiments?

- The gold standard for learning about causal effects is a randomized controlled trial (RCT), aka experiment
- Suppose that the Clemson and USC administration randomized who got into which college (assume these are the only 2 colleges for simplicity)
- Since college is randomly assigned, the only thing that differs between Clemson and USC students is the college they went to
- Hence,

$$\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at USC for Clemson Students}} = \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at USC for USC Students}}$$

since we've eliminated any confounding factors

But running experiments is often hard/impossible

- Unfortunately, Clemson/USC have not let us randomize who gets into which college
  - At least not yet! If you could convince them to do this, it'd make for a cool senior thesis!
- Likewise, it is difficult to convince states to randomize their minimum wages, or other policies
- In some cases, randomization is not just difficult but would be immoral
  - "What is the causal effect of spousal death on labor supply?"
- In this course, we'll discuss tools economists try to use when running experiments is not possible

Outline

Course Roadmap – Where we're going

- **Topics**
  1. Simple Linear Regression
  2. Introduction to Causality
  3. Multiple Linear Regression
  4. Linear Regression Extensions
  5. Sampling
  6. Confidence Intervals & Hypothesis Testing
  7. Regression Inference
  8. ChatGPT and coding
  9. Regression Discontinuity
  10. Difference-in-Differences
  11. Panel Data

Course Roadmap – Where we're going

- **Topics**
  1. Simple Linear Regression
  2. Introduction to Causality
  3. Multiple Linear Regression
  4. Linear Regression Extensions
  5. Sampling
  6. Confidence Intervals & Hypothesis Testing
  7. Regression Inference
  8. ChatGPT and coding
  9. Regression Discontinuity
  10. Difference-in-Differences
  11. Panel Data

- **Important Dates**
  - February 6 (No Class): One Page Proposal for Final Project Due
  - March 4 (No Class): Take Home Coding Midterm
  - Spring Break: March 18, March 20, and March 25 (No Classes)
  - April 10: In Class Theory Midterm
  - Final Project Presentations: April 15, April 17, April 22, April 24
  - Final Project: Due by April 30 at 3PM