

# Introductions and Overview

Dr. Adam Soliman

Introduction to Econometrics (ECON 4050)  
Clemson University  
Fall 2024

# Outline

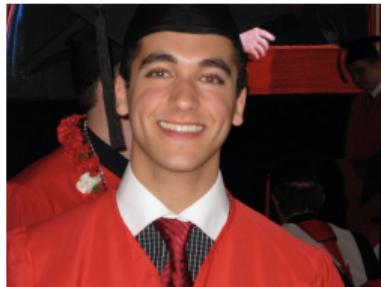
1 Course Preliminaries

2 Why is Econometrics Challenging?

3 Course Roadmap

## A little bit about your econometrics professor

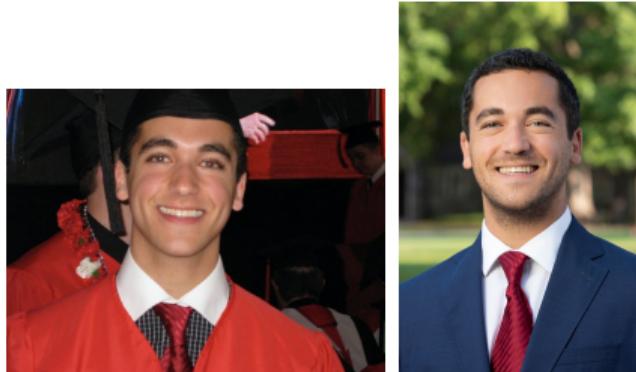
- I have been in school for a while...from Boston University (undergrad) to Michigan State (masters) to Duke (PhD) to London School of Economics (post-doc)



- My first job was teaching math and economics in Dubai

## 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
  - What are the impacts of cracking down on rogue doctors during the opioid epidemic on street drug prices, overdose mortality, and other doctors behavior?
  - Do police respond to changes in punishment severity?
  - What happens to neighborhood crime when investors buy many properties?

## What is econometrics?

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

## What is econometrics?

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

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) that to answer economic questions with data

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) that to answer economic questions with data

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) that to answer economic questions with data

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

## A little bit about your classmates

Insert some info from survey

## Expectations

- Ask questions: to me, your TAs, 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: Tues/Thurs from PM to PM in Powers Hall
- ▶ Lab: Thursday from PM to PM in Powers Hall
- ▶ My office hours are Wednesdays from 3PM to 5PM on Zoom
- ▶ Our Teaching Assistants (TAs) are , and
  - ★ All TAs will hold OHs. Times/locations for TA OHs will be announced shortly

## 2) Content and communication:

- ▶ We will **exclusively** use Slack for our interactions. Please do not contact me by email unless there's a legitimate reason to
  - ★ I will be checking Slack sparingly so please help one another in the meantime
- ▶ The main course materials are currently posted on [the Github course website](#). Other communication will be via Slack and Canvas.

# Logistics

## 3) Software:

- ▶ R for statistical analyses (to be covered in Lab sessions and throughout the course)

## 4) Assessments:

- ▶ In-class quizzes/attendance: 10%
- ▶ Lab work: 25%
- ▶ 1 midterm exam: 25%
- ▶ Final project and associated presentation: 30% and 10%, 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

1 Course Preliminaries

2 Why is Econometrics Challenging?

3 Course Roadmap

## 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  $\Rightarrow$  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  $\Rightarrow$  misleading results!

## Why is answering these questions hard? (Part II)

- Answering causal questions is often *even harder* than descriptive ones because they involve both 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 both 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  $i = 1, \dots, N$ . 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  $i = 1, \dots, N$ . 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  $i = 1, \dots, N$ . 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:

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

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

- **Sample:**  $(Y_i, D_i)$  for  $i = 1, \dots, N$ . 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:

$$\underbrace{E[Y_i(1)|D_i = 1]}$$

Earnings at Clemson for Clemson Students

and

$$\underbrace{E[Y_i(0)|D_i = 0]}$$

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

1 Course Preliminaries

2 Why is Econometrics Challenging?

3 Course Roadmap

# 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 Regression Discontinuity
- ⑨ Instrumental Variables
- 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 Regression Discontinuity
- ⑨ Instrumental Variables
- ⑩ Difference-in-Differences
- ⑪ Panel Data

- Important Dates

- ▶ October 3rd: No Class
- ▶ October 15th: No Class (Fall Break)
- ▶ October 17th: Midterm Exam
- ▶ November 5th: No Class (Election Day)
- ▶ November 28th: No Class (Thanksgiving Break)
- ▶ Final Project Presentations (Tentatively): November 21st, November 26th, December 3rd, and December 5th

**And now...**

## A brief introduction to R