Chapter 1: Introduction to ECON 1630

Jonathan Roth

Mathematical Econometrics I Brown University

Outline

- 1. Course Preliminaries
- 2. What is Econometrics?
- 3. Why is Econometrics Challenging?
- 4. Course Roadmap

Introducing Ourselves

Welcome to ECON 1630! I'm looking forward to teaching you all

I'm Professor Jonathan Roth

- Group OH on Zoom (link): Mon 220-250
- Individual OH: Typically Tue 3-4; sign up on my website: here, 8
 Fones Alley 014 (or Zoom by request)

Introducing Ourselves

Welcome to ECON 1630! I'm looking forward to teaching you all

I'm Professor Jonathan Roth

- Group OH on Zoom (link): Mon 220-250
- Individual OH: Typically Tue 3-4; sign up on my website: here, 8
 Fones Alley 014 (or Zoom by request)

Our Grad TAs are Moritz Poll and Max Grozovsky
Our undergrad TAs are Matt Kutam and Preetish Juneja (star students from previous semesters!)

- All TAs will hold OHs
- Grad TAs will hold weekly sessions
 - Review material and teach coding
- Times/locations for TA OHs and sections will be announced shortly on Canvas

Canvas and EdDiscussion

Course materials and communications will be posted on Canvas: https://canvas.brown.edu/courses/1100468

The Canvas page has an EdDiscussion board, which is a great place to ask (and answer) questions. The TAs will monitor the Qs.

Meeting times:

- Lectures: Mon/Wed 830-950 (S01) or 3-420pm (S02).
 Recordings will be posted online after class.
- TA Sessions: Time/location TBD
- Attendance is not required but is highly encouraged

4

Meeting times:

- Lectures: Mon/Wed 830-950 (S01) or 3-420pm (S02).
 Recordings will be posted online after class.
- TA Sessions: Time/location TBD
- Attendance is not required but is highly encouraged

Prerequisites:

- Multivariate calculus, probability/statistics, and linear algebra
- Some familiarity with reading/writing proofs and code

Meeting times:

- Lectures: Mon/Wed 830-950 (S01) or 3-420pm (S02).
 Recordings will be posted online after class.
- TA Sessions: Time/location TBD
- Attendance is not required but is highly encouraged

Prerequisites:

- Multivariate calculus, probability/statistics, and linear algebra
- Some familiarity with reading/writing proofs and code

Software:

- Default is Stata for statistical analyses (to be covered in TA sessions)
 - You're welcome to use R instead; TAs are familiar with R
- LaTeX for typing up problem sets (optional, for extra credit)
- Ask us for help if you're having any problems accessing software

Assessments:

- 6 problem sets due approximately every 2 weeks, submitted via Gradescope
- 1 midterm exam (November 3, in class)
- 1 final exam

Assessments:

- 6 problem sets due approximately every 2 weeks, submitted via Gradescope
- 1 midterm exam (November 3, in class)
- 1 final exam

Grading:

- 30% problem sets, 35% on each exam.
- I'll drop your lowest PSet grade. Use your drop wisely!
- Psets are due at 4PM on Fridays; late submissions won't be graded.
 Collaboration is OK (please list collaborators)
- The exams will be in-class, closed-book, "cheat-sheet" allowed
- 5 points extra credit if you use LaTeX for assignments. Please attach your code + output as a single PDF regardless

Al Policy

My view is that Al is a useful tool and we shouldn't ignore it. But need to use it in a way that enhances rather than impedes learning.

- Exams (70%): closed-book; no Al allowed.
- Problem sets: Al OK if it helps you learn, not if it replaces learning/effort.
 - Good: debugging R/Stata errors.
 - Bad: writing free-form answers.
- I can't police what you do on the psets, but if you don't put in effort you probably won't do well on the exams
- Include brief Al use disclosure on each pset.

Course materials:

- Main material: Lectures and lecture slides, which will be posted on Canvas
- Optional text: Stock & Watson Intro to Econometrics (4th ed)

Any questions on logistics?

Outline

- 1. Course Preliminaries ✓
- 2. What is Econometrics?
- 3. Parameters, Estimands, and Estimators
- 4. Course Roadmap

 \rightarrow The statistical toolkit that economists use to answer economic questions with data

 \rightarrow The statistical toolkit that economists use to answer economic questions with data

- Has economic inequality increased since 1960?
- How do increases in the minimum wage affect employment?
- What will the unemployment rate be next quarter?

 \rightarrow The statistical toolkit that economists use to answer economic questions with data

- Has economic inequality increased since 1960?
 - Descriptive Q: asks about how things are (or were) in reality
- How do increases in the minimum wage affect employment?
- What will the unemployment rate be next quarter?

 \rightarrow The statistical toolkit that economists use to answer economic questions with data

- Has economic inequality increased since 1960?
 - Descriptive Q: asks about how things are (or were) in reality
- How do increases in the minimum wage affect employment?
 - Causal Q: What would have happened in a counterfactual world?
- What will the unemployment rate be next quarter?

 \rightarrow The statistical toolkit that economists use to answer economic questions with data

- Has economic inequality increased since 1960?
 - Descriptive Q: asks about how things are (or were) in reality
- How do increases in the minimum wage affect employment?
 - Causal Q: What would have happened in a counterfactual world?
- What will the unemployment rate be next quarter?
 - Forecasting Q: What will happen in the future?

 \rightarrow The statistical toolkit that economists use to answer economic questions with data

What types of questions might we be interested in:

- Has economic inequality increased since 1960?
 - Descriptive Q: asks about how things are (or were) in reality
- How do increases in the minimum wage affect employment?
 - Causal Q: What would have happened in a counterfactual world?
- What will the unemployment rate be next quarter?
 - Forecasting Q: What will happen in the future?

In this course, we will focus mainly on descriptive and causal questions, with an emphasis on causal questions

Why is answering these questions hard?

- For descriptive Qs: 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 Qs: 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
 - E.g., the workers in our survey were drawn out of hat with names of all possible workers
 - If so, need to account for the fact that by chance the sample might have different characteristics from the population

Why is answering these questions hard?

- For descriptive Qs: 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
 - E.g., the workers in our survey were drawn out of hat with names of all possible workers
 - If so, need to account for the fact that by chance the sample might have different characteristics from the population
- Worst case scenario: our sample is not representative of the population that we care about
 - E.g., workers with certain characteristics were more likely to respond to the survey

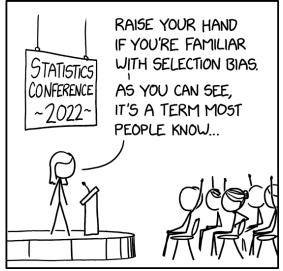


• In 1948, Chicago Tribune writes that Thomas Dewey defeats Harry Truman in the 1948 presidential election, based on survey of voters.



- 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 ≠ population → misleading results!

Selection bias referes to settings like Dewey-Truman where the sample is not drawn randomly from the population of interest



Answering causal questions is often even harder than descriptive ones.
 Why?

- Answering causal questions is often even harder than descriptive ones.
 Why?
- Causal Qs involve both a descriptive component (what are outcomes in reality?) and a *counterfactual* component (how would things have been under a different treatment?)

- Answering causal questions is often even harder than descriptive ones.
 Why?
- Causal Qs 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 Brown instead of URI?
 - Descriptive Q: how much do Brown students earn after graduation?
 - Counterfactual Q: how much would Brown students have earned if they went to URI?

- Answering causal questions is often even harder than descriptive ones.
 Why?
- Causal Qs 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 Brown instead of URI?
 - Descriptive Q: how much do Brown students earn after graduation?
 - Counterfactual Q: how much would Brown students have earned if they went to URI?
- Counterfactual Qs can't ever be answered with data alone. Need additional assumptions to learn about them!

Splitting up the problem

- When thinking about causal Qs, 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

- Sample: the data that you actually observe
 - A survey of students from Brown and URI graduates about their earnings

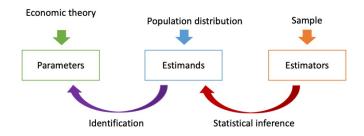
- Sample: the data that you actually observe
 - A survey of students from Brown and URI graduates about their earnings
- Estimator: a function of the data in the sample
 - Difference in earnings between Brown and URI students in survey

- Sample: the data that you actually observe
 - A survey of students from Brown and URI graduates about their earnings
- Estimator: a function of the data in the sample
 - Difference in earnings between Brown and URI students in survey
- Estimand: a function of the observable data for the population
 - Difference in earnings between all Brown and URI students

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

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

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



- Introduce potential outcomes notation
 - Super useful framework for thinking about causality!
 See the 2021 Nobel Prize writeup on Canvas!

- Introduce potential outcomes notation
 - Super useful framework for thinking about causality!
 See the 2021 Nobel Prize writeup on Canvas!
- D_i = indicator if get treatment (1 if Brown, 0 if URI)

- Introduce potential outcomes notation
 - Super useful framework for thinking about causality!
 See the 2021 Nobel Prize writeup on Canvas!
- D_i = indicator if get treatment (1 if Brown, 0 if URI)
- $Y_i(1)$ = outcome under treament = earnings at Brown
- $Y_i(0)$ = outcome under control = earnings at URI

- Introduce potential outcomes notation
 - Super useful framework for thinking about causality!
 See the 2021 Nobel Prize writeup on Canvas!
- D_i = indicator if get treatment (1 if Brown, 0 if URI)
- $Y_i(1) =$ outcome under treament = earnings at Brown
- $Y_i(0)$ = outcome under control = earnings at URI
- 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)

- Introduce potential outcomes notation
 - Super useful framework for thinking about causality!
 See the 2021 Nobel Prize writeup on Canvas!
- D_i = indicator if get treatment (1 if Brown, 0 if URI)
- $Y_i(1) =$ outcome under treament = earnings at Brown
- $Y_i(0)$ = outcome under control = earnings at URI
- 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 $Y_i = D_i Y_i(1) + (1 D_i) Y_i(0)$

• Example sample: (Y_i, D_i) for i = 1,...N. Data with earnings and where you went to school

- Example sample: (Y_i, D_i) for i = 1,...N. Data with earnings and where you went to school
- Example estimator:
 - Difference in sample mean of earnings for people who went to Brown and people who went to URI:

$$\underbrace{\frac{1}{N_1}\sum_{i:D_i=1}Y_i}_{i:D_i=1}Y_i \qquad - \underbrace{\frac{1}{N_0}\sum_{i:D_i=0}Y_i}_{i:D_i=0}$$

Avg earnings at Brown in sample Avg earnings at URI in sample

- Example sample: (Y_i, D_i) for i = 1,...N. Data with earnings and where you went to school
- Example estimator:
 - Difference in sample mean of earnings for people who went to Brown and people who went to URI:

$$\underbrace{\frac{1}{N_1}\sum_{i:D_i=1}Y_i}_{i:D_i=1}Y_i \qquad - \underbrace{\frac{1}{N_0}\sum_{i:D_i=0}Y_i}_{i:D_i=0}Y_i$$

Avg earnings at Brown in sample Avg earnings at URI in sample

- Example estimand:
 - Difference in population mean of earnings for people went to Brown and people who went to URI:

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

18

- Example sample: (Y_i, D_i) for i = 1,...N. Data with earnings and where you went to school
- Example estimator:
 - Difference in sample mean of earnings for people who went to Brown and people who went to URI:

$$\underbrace{\frac{1}{N_1} \sum_{i:D_i=1} Y_i}_{i:D_i=1} Y_i \qquad - \underbrace{\frac{1}{N_0} \sum_{i:D_i=0} Y_i}_{i:D_i=0}$$

Avg earnings at Brown in sample Avg earnings at URI in sample

- Example estimand:
 - Difference in population mean of earnings for people went to Brown and people who went to URI:

$$\underbrace{E[Y_i|D_i=1]} - \underbrace{E[Y_i|D_i=0]}$$

Avg earnings at Brown in population Avg earnings at URI in population

- Example target parameter:
 - Causal effect of Brown for Brown students:

$$E[Y_i(1)|D_i=1]$$
 - $E[Y_i(0)|D_i=1]$

 ${\sf Earnings\ at\ Brown\ for\ Brown\ students\ in\ pop} \quad {\sf Earnings\ at\ URI\ for\ Brown\ students\ in\ pop}$

- Thought experiment: suppose we had data on earnings for every Brown and URI graduate
- We can learn from the data:

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

- Thought experiment: suppose we had data on earnings for every Brown and URI graduate
- We can learn from the data:

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

• The causal effect of Brown for Brown students is

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

19

- Thought experiment: suppose we had data on earnings for every Brown and URI graduate
- We can learn from the data:

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

• The causal effect of Brown for Brown students is

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

ullet The data doesn't tell us $\underline{{\it E}[Y_i(0)|D_i=1]}$. Why not?

Earnings at URI for Brown Students

- Thought experiment: suppose we had data on earnings for every Brown and URI graduate
- We can learn from the data:

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

The causal effect of Brown for Brown students is

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

The data doesn't tell us

$$\underbrace{\textit{E}[Y_i(0)|D_i=1]} \qquad \text{. Why not?}$$

Earnings at URI for Brown Students

Because we never see Brown students going to URI!

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

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

• Why might this give us the wrong answer?

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

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

- Why might this give us the wrong answer?
- Because Brown students may be different from URI 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 Brown and URI 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 Brown and URI students is the college they went to
- Hence,

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

since we've eliminated any confounding factors

But running experiments is often hard/impossible

- Unfortunately, Brown/URI 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?"

But running experiments is often hard/impossible

- Unfortunately, Brown/URI 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.

Course Roadmap - Where we're going

- Part I (\sim 7 lectures): Review of probability/statistics. This will give us a mathematical language to talk about:
 - Statistical estimation/inference: how does the sample we observe relate to the population of interest
 - 2 Identification: how do observable features of the population relate to (causal) parameters we care about

Course Roadmap - Where we're going

- Part I (\sim 7 lectures): Review of probability/statistics. This will give us a mathematical language to talk about:
 - Statistical estimation/inference: how does the sample we observe relate to the population of interest
 - Identification: how do observable features of the population relate to (causal) parameters we care about
- Part II (\sim 9 lectures): Linear regression: We'll discuss ordinarly least squares (OLS), the workhorse model for estimation in econometrics. When does it work, and when will it fail?

Course Roadmap - Where we're going

- Part I (\sim 7 lectures): Review of probability/statistics. This will give us a mathematical language to talk about:
 - Statistical estimation/inference: how does the sample we observe relate to the population of interest
 - Identification: how do observable features of the population relate to (causal) parameters we care about
- Part II (~ 9 lectures): Linear regression: We'll discuss ordinarly least squares (OLS), the workhorse model for estimation in econometrics. When does it work, and when will it fail?
- Part III (~ 7 lectures:) Other "quasi-experimental" strategies:
 We'll discuss other strategies for "mimicking" an experiment when it's
 not available, including instrumental variables (IV) and regression
 discontinuity (RD)