Week 1: Introduction

Econ 672

Samuel Rowe, PhD

Outline of Week

- Overview of the Course
- What is Causal Inference
- Correlation and Causation
- Importance of Theory
- Question about Questions
- Steps to Get Started
- Programming Topics

OVERVIEW OF THE COURSE

Overview of Course

- What is program analysis and evaluation?
  - How can we establish that a treatment or program causes an outcome instead of being associated with an outcome?
- What is causal inference?
  - How can we infer causal relationships?
- The course will be divided into 2 halves
  - The first half will be theoretical foundations
  - The latter half will be developing a toolbox

Overview of the Course

- Theoretical Foundations
  - Potential Outcomes and Experimental Design
  - Directed Acyclic Graphs
  - Regression Review
- Developing a Toolbox
  - Subclassification and Matching
  - Instrumental Variables

Overview of the Course

- Developing a Toolbox (cont.)
  - Regression Discontinuity Design
  - Panel Data
  - Difference-in-Differences
  - Synthetic Control Method
  - Additional Topics
    - Advances in Diff-in-Diff
    - Comparative Interrupted Time Series

Overview of Course

- Randomized Control Trials
  - Gold standard for assessing causal inference
  - What happens when we don't have an RCT?
- Four Overarching Questions for Each Identification Strategy
  - What are the strengths?
  - What are the weaknesses?
  - What are the assumptions?
  - Are the assumptions testable?

WHAT IS CAUSAL INFERENCE?

What is Causal Inference?

- Causal inference is the study of cause and effect
  - Causal inference builds off of work from Rubin and Pearl
  - Rubin's causal model (Rubin, 1974)
    - Builds a basis of potential outcomes
- Pearl (2010)
  - "its aim is to infer probabilities under conditions that are changing, for example, changes induced by treatments or external interventions"

What is Causal Inference?

- Pearl (2010)
  - An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables
    - Examples of association concepts: Correlation and regression
  - A causal concept is any relationship that cannot be defined from the distribution alone
    - We need to go one step further to establish causation
    - Examples of causal concepts: Randomized, influence, effect

CORRELATION AND CAUSATION

Correlation and Causation

- “Correlation is not causation” is a common statement
- Correlations in observational data might not reflect causal relationship at all
  - Observational data are data not generated through randomized trials
  - Think survey or administrative data
- Correlations may be spurious in observational data

Correlation and Causation

- There may be a causal relationship in observational data without any observed correlation
- We will use an example of a sailboat on a windy day

Correlation and Causation

- On a windy day in a boat a sailor maneuvers a rudder to counter the wind blowing
- The boat appears to be going in a straight line while the rudder in the water seems erratic
- It might seem that the rudder is broken
- If the sailor randomized the rudder, then the boat would zigzag across the lake
- The sailor's movement is endogenous in response to unobserved wind
  - Cancels out relationship between rudder and boat direction
- There is a causal relationship between the rudder and the boat direction but the unobserved wind

Correlation and Causation

- Correlations fail at revealing causal effects
  - Human beings do not make decisions randomly
  - Human sort into preferable or optimal decision
    - Like microeconomics shows us
  - Human will select or sort into treatment if it is optimal for them without randomization
    - Those correlations will not reveal causation
- We need theory, literature, and prior knowledge

IMPORTANCE OF THEORY

Importance of Theory

- Causal inference seems atheoretical
  - It is not; we need theory
  - Identification strategies are tools
  - Lead with a question, go to theory, set up a hypothesis, test with toolbox
- Without prior knowledge, causal estimates are less believable
  - Theory, literature, or deep institutional knowledge provides corroborate causal estimates
- Economic theory helps support results
  - We need to be suspicious of correlations as causation
  - Economic agents make decision they thought to be in their best interest – sort into treatment/program

Importance of Theory

- We need data in order to test a hypothesis built from theory or prior knowledge
- There are two types of data
  - Data generated through randomized trials
  - Data not generated through randomized trials
- Data from Randomized Trials
  - Less common in economic fields, but growing
  - Harder to collect due to financial or moral reasons
  - Generated in a prospective manner
  - Researcher is an active participant

Importance of Theory

- Data not from Randomized Trials: Observational Data
  - Collected in a retrospective manner
  - Researcher is a passive actor in data generation process
- Survey data
  - E.g.: Public-use microdata from Current Population Survey (CPS) or American Community Survey (ACS)
- Administrative data
  - Data generated for administrative processes
  - Data not originally intended for public consumption
  - E.g.: Unemployment Insurance, compliance data from federal agencies, marketing data from firms, etc.

Importance of Theory

- Observational Data might have biases within them
  - Who is selected or not selection into the data
  - Is the administrative data a census of all the population (UI) or a subset (FLSA Wage Violations)?
- It's easy to assume that correlation is causation, but we need theory and knowledge to estimate causal relationship

QUESTION ABOUT QUESTIONS

Question about Questions

- Angrist and Pischke (2009) provides a great questions in their opening chapter
- There are four questions of interest
  - 1) What is the causal relationship of interest?
  - 2) What is the experiment that could ideally be used to capture causal effect of interest?
  - 3) What is your identification strategy?
  - 4) What is your mode of statistical inference?

Causal Relationship of Interest

- We will focus on research questions of cause and effect
  - There are other types of program evaluation, but for this course will focus on “impact evaluations” or causal inference
- How do we assess the causal relationship?
- What is the counterfactual?
  - What does theory tell us about the unobserved world if an alternative choice had been made?

Causal Relationship of Interest

- What does theory tell us causal effects for causal research questions?
  - What does the price elasticity of demand tell us about taxing inelastic or elastic goods to raise revenue?
  - What does microeconomic theory tell us about price controls and patents or natural monopolies?
- What does theory tell us about
  - “RTW” laws and unionization?
  - carbon tax instead of regulation for reducing carbon emissions?
  - What does theory tell us about zoning laws and housing prices?
  - paid family and medical leave and parental labor force participation
  - Investing in public transportation instead of highways?

Causal Relationship of Interest

- We will learn about and utilize DAGs
  - Directed acyclic graphs are a useful tool
  - Graphically shows theory
  - Graphically shows research design
  - Graphically show (or not show) your assumptions
- Another tool to setting up a causal relationship of interest

Ideal Experiment

- Fundamentally Answerable Question
  - Your research question should be able to be answered by an experiment (randomized control trial − RCT)
  - It's good think about good research questions
- Fundamentally Unanswerable Question (FUQ)
  - No experiment will be able to answer the question of interest if it is FUQ

Ideal Experiment

- It good to think about an ideal experiment
- Think about a no-constraint experiment to answer the research question of interest
- Example
  - We want to assess a training program for returning ex-offenders in the justice system
  - It might not morally be responsible to deny services for a sensitive population in an RCT
  - But, it good to think about an ideal experiment

Identification Strategy

- You will likely not have assess to an RCT or data that have been randomized into treatment
- You will likely be working with observational data
  - Data not generated by randomized trials
  - Observational data may come from survey data (e.g.: CPS, ACS) or administrative data
  - Understanding the data-generation process will be an important step will discuss later

Identification Strategy

- Working with Observational Data
  - What is your strategy to identify a causal effect from observational data or your identification strategy
  - There may be bias within the data-generation process for observational data
  - Individuals make decisions to optimize their well-being
  - Selection or sorting into treatment causes bias
- We will be discussing several identification strategies in the toolbox section
  - What are their strengths?
  - What are their weaknesses?
  - What are the assumptions?
  - Are the assumptions testable?

Statistical Inference

- How will you test your hypothesis?
  - Most focus on p-values
- How are your errors distributed?
  - Are they clustered? Homogenous?
- Cunningham (2021) provides a nice review of randomized inference
  - Different from distribution of coefficients
  - What happens when we randomized our treatment instead of randomizing a sampling of observations like bootstrapping?

Statistical Inference

- Angrist and Pischke (2009) provide a great haiku
  - "T-stats looks too good;"
  - "Try clustered standard errors -"
  - "Significance gone"

STEPS TO GET STARTED

Steps to Get Started

- Empirical Analysis is a cornerstone of scientific methodologies
- We will concisely review some of the topics we have discussed into steps to get started
- First Step
  - Formulate a research question
  - It needs to be an answerable question

Steps to Get Started

- Second Step
  - Develop or set up an economic model
  - This should describe the relationship of interest mathematically
- Third Step
  - The model will allow us to set up a testable hypotheses or falsifiable predictions
  - We can test these hypotheses with data

Steps to Get Started

- Fourth Step
  - Set up an econometric model or identification strategy
  - What is the functional form of the relationship between two variables?
    - Linear? Polynomial? Interaction? Logarithmic?
  - How do we deal with observed and unobserved variables and confounders?

Steps to Get Started

- Comparative Statics
  - Theoretical descriptions of the causal effects contained within the model
  - These are based upon ceteris paribus or all else constant (similar to a partial derivative)
  - What happens with a change in the treatment or program of interest holding everything else constant
  - Covariate balance (observed and unobserved) will be a frequent topic of discussion with all else constant
  - Without ceteris paribus, the estimate of the causal relationship would be confounded or biased

Steps to Get Started

- Example: Estimating Price Elasticity of Demand
- Select a Question
  - We want to know what is the price elasticity of demand for a product due to tax reasons
  - Problem is we don't observe demand and supply curves − just price and quantity at equilibrium
- Select a economic model
  - We can use the price elasticity of demand

$$\varepsilon = \frac{\partial \log Q}{\partial \log P}$$

Steps to Get Started

- Set up an econometric model

$$\log Q_D = \alpha + \delta \log P + \lambda X + u$$

- We are interested in estimating delta
- The functional form is logarithmic
- We need:
  - Lots of data that provides variation
  - We need price to be independent of u
  - We need price to be exogenous
- We run into an endogenous issue with Q and P

PROGRAMMING TOPICS

Programming

- We will switch over to Stata
- Knowledge of theoretical foundations and identification strategies is more important
  - However, it is good to know programming
  - Marketable skill
- Please use github.com
  - I try to update all of the do files to make them easily accessible for everyone
  - We can use ELMS, but GitHub is easier
  - www.github.com/rowesamuel/ECON672

Programming

- Use of Statistical Packages (e.g.: Stata, R)
  - Transparent
  - Replication and Reproducible
  - Easily shared
    - (e.g.: GitHub)
  - Efficient
    - Mistakes can be fixed more easily
- Access to the tools you need
  - Can be used for data management
  - Can be used for data analysis
  - Can be used for program evaluation and analysis

Programming

- Don’t use Excel!
  - But I like Excel, it’s easy
  - Excel is easy to use upfront but you pay in the long-run
  - Not transparent, might not be sequential, very hard to reproduce
  - Broken tabs, broken links, no order of sequence
  - A use of a “do file” or programming scripts reduces or eliminates these excel issues

Programming

- Utilize “ETL”
  - Extract -> Transform -> Load
- Extract
  - Take data from their raw form
  - We will be utilizing GitHub (or ELMS depending upon the size of the data)
- Transform
  - Data management and transformation of data
  - Any mistake can be easily fixed in scripts, such as do files
- Load
  - Most of “loading” will be csv output or esttab or figures and graphs

Programming Topics

- Organizing your files
- Macros
- Looping
- Tempfiles
- By, Sort, and Egen
  - Powerful combination
- Weights
- Esttab – Outputting results
  - Try not to copy and paste (doesn't work with ETL)