Lecture 1: Introduction

Mitch Downey[1]

January 21, 2024

[1] IIES, Stockholm University. Slides heavily based on those developed by Markus Jäntti based on the textbook by Bruce Hansen.

What is econometrics? Conceptual answer

Experience has shown that each of these three view-points, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And it is this unification that constitutes econometrics. (Ragnar Frisch, as quoted in Hansen)

What is econometrics?

- Conceptual answer: Ragnar Frisch (quoted in Hansen)
    - Combines “statistics, economic theory, and mathematics” for “a real understanding of the quantificative relations in modern economic life”
    - “It is the unification of all three that is powerful”
- Relative to mathematics and statistics, we think more about human behavior
- Mathematics thinks about equilibria and fixed points of functions
    - $p, q \in \mathbb{R}^2, p = S(q), q = D(p)$
    - p^* is a fixed point if $p^* = S\big(D(p^*)\big)$
- We build specific functions to capture features of human behavior
    - Decision theory: What should be the axiomatic foundations of decisions?
        - Daniel McFadden brings this to econometrics
    - General equilibrium theory: How do these decisions map to aggregate outcomes?
        - Arrow-Debreu: Under what conditions will an axiomatic supply function and axiomatic demand function generate equilibrium?

What is econometrics?

- Conceptual answer: Ragnar Frisch (quoted in Hansen)
    - Combines “statistics, economic theory, and mathematics” for “a real understanding of the quantificative relations in modern economic life”
    - “It is the unification of all three that is powerful”
- Relative to mathematics and statistics, we think more about human behavior
- Mathematics thinks about equilibria and fixed points of functions
- We build specific functions to capture features of human behavior
- Statistics thinks about distributions and random variables
    - You would learn 1000 times as much about Weibull distributions in a statistics class
    - Common statistics papers: Let $Y = f(X)$...
      Last 2 pages: We use data on education and earnings
    - No focus on *realism* of assumptions, only *generality* of assumptions
        - The worst of the economic theory papers are like this too
    - Last page: “Of course, in an application, domain experts would need to assess whether...”
        - This is us!
- These are cultural definitions as much as conceptual ones

What is econometrics? Generative answer

- Instead of defining it by its *differences* or absence, let's define it constructively
- What is economics?
  - Constrained optimization
  - Choices are solutions to objective functions, subject to constraints (decision rules)

What is econometrics? Generative answer

- Instead of defining it by its *differences* or absence, let's define it constructively
- What is economics?
  - Constrained optimization
  - Choices are solutions to objective functions, subject to constraints (decision rules)
- Econometrics version 1: Under what conditions can we invert decision process to infer objectives and constraints from observed choices?
  - This tradition is often called "**structural econometrics**"
  - Strongly tied to theory
  - Write probabilistic model, take is seriously, derive likelihood function, estimate with either method of moments, maximum likelihood, or Bayesian estimation
  - Main criticisms:
    - Hard to interpret what is driving results
    - Hard to know how results are affected if some assumptions are wrong

What is econometrics? Generative answer

- Instead of defining it by its *differences* or absence, let's define it constructively
- What is economics?
  - Constrained optimization
  - Choices are solutions to objective functions, subject to constraints (decision rules)
- Econometrics version 2: If decisions are endogenous (optimal choice given the conditions), how can we separate effects of choices from effects of conditions?
  - This tradition is often called "**reduced form causal inference**"
  - Sometimes seen as distinct from "real" (theoretically motivated) economics
  - Actually a deep ideological connection to theory
  - Question: What is effect of fertilizer on agricultural output?
    - Agricultural scientist: Collect good fertilizer and output data. Compare.
    - Economist: But everyone optimizes. Two farmers making different choices about fertilizer must have done so for some reason.
  - Question: What is effect of better school on later life outcomes?
    - Sociology: Come up with good measures of schools and later life outcomes. Compare.
    - Economist: But everyone optimizes. There's a reason one kid went to a better school.
  - Old economics: Those fields do their thing, we do ours
  - Modern economics: Integrates the traditions of those fields (original data collection, better and more careful measurement) into our causal inference tradition

What is econometrics? Generative answer

- What is economics?
  - Constrained optimization
  - Choices are solutions to objective functions, subject to constraints (decision rules)
- Econometrics version 1: Under what conditions can we invert decision process to infer objectives and constraints from observed choices?
  - This tradition is often called "**structural econometrics**"
- Econometrics version 2: If decisions are endogenous (optimal choice given the conditions), how can we separate effects of choices from effects of conditions?
  - This tradition is often called "**reduced form causal inference**"
- Most content of this course is common to both versions
  - "Identification strategy": What variation in the data determines the parameter estimate, and how should that affect my interpretation of the parameter?
- Econometrics II focuses on reduced form causal inference
- Mehran Ebrahimian (SSE) teaches an applied Industrial Organization elective on structural econometrics
- The most successful modern work blends these

The probability approach to econometrics

- Economic theories are often deterministic
- Econometrics is probabilistic
- Are there **random variables**?
  - Formal definition next lecture
- If not, no role for econometrics
  - No **inference**, no **endogeneity**
  - Still questions about **identification**, **invertability**, and **falsifiability**
  - Mainly micro theory: Revealed preference theory[2]
- If so, ***where does the randomness come from?***
  - This question long ignored
  - Recently gaining more attention as important
  - We will try to focus on this, you try to pay attention
  - Key distinction: **Design-based** vs. **model-based** randomness

[2] Chambers, Christopher P., and Federico Echenique. Revealed preference theory. Vol. 56. *Cambridge University Press*, 2016. Chambers, Christopher P., Federico Echenique, and Eran Shmaya. "The axiomatic structure of empirical content." *American Economic Review* 104.8 (2014): 2303-2319.

The probability approach to econometrics

- Economic theories are often deterministic
- Econometrics is probabilistic
- Where does the randomness come from?
- Everyone's goal: Reduce and/or weaken assumptions
  - Structural: Go from **fully parametric** (randomness comes from a specific distributional family) to **semi-parametric** or **non-parametric** (no assumptions on the specific distributions)
  - Reduced form: Fewer assumptions about incentives, information sets, and constraints
- These have methodological implications
  - **Maximum likelihood estimation** mostly requires fully parametric model
  - **Generalized method of moments** can solve less parametric models
  - Unspecified models (reduced form econometrics) can only be formulated as **potential outcomes** with **regression analysis** used to estimate **conditional expectations**
- But assumptions have power
  - Key tradeoff: Weaker assumptions make it more difficult to learn things from data
  - This should be in the background of every lecture

What is this course?

- This course is not an applied econometrics course
  - Econometrics II is, and it’s very good
- This is a theory course. Why?
  - Econometrics II will build on this foundation
  - We will build that foundation slowly & carefully
  - It is good for you to work through math in abstract contexts
- This course will include plenty of pointers to applied work to help applied-minded students stay interested and motivated
- **This course is supposed to be difficult.**

Goal 1: Applied micro students can read these papers by the end

- Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. "Sampling-based versus design-based uncertainty in regression analysis." *Econometrica* 88, no. 1 (2020): 265-296.
    - Where does randomness come from? Why does that matter for inference?
- Borusyak, Kirill, Peter Hull, and Xavier Jaravel. "Quasi-experimental shift-share research designs." *The Review of Economic Studies* 89.1 (2022): 181-213.
    - Where can identifying variation come from in shift-share designs and why does it matter for inference?
- Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. "Bartik instruments: What, when, why, and how." *American Economic Review* 110.8 (2020): 2586-2624.
    - Where can identifying variation come from in shift-share designs and why does it matter for inference?
- Roth, Jonathan, and Pedro HC Sant'Anna. "Efficient estimation for staggered rollout designs." *Journal of Political Economy: Microeconomics* 1.4 (2023): 669-709.
    - What variation drives identification in panel models? Should this affect inference?
- Shen, Dennis, Peng Ding, Jasjeet Sekhon, and Bin Yu. "Same Root Different Leaves: Time Series and Cross-Sectional Methods in Panel Data." *Econometrica* 91, no. 6 (2023): 2125-2154.
    - What variation drives identification in panel models? Should this affect inference?
    - Are modern state-of-the-art fancy methods a dramatic shift in econometric analysis, or a marginal refinement?

Goal 2: Macro/finance students can read more broadly

- Stockholm macro/finance is top notch
- Most of it captures a very specific style of work
  - State of the art, and at the forefront of that style
- Even outside of that, people understand their advisor's work
- Other work is substantively relevant but stylistically different
- **Time and time again**: I have seen macro/finance students ignore that work
  - Sometimes macro labor students ignore structural labor
  - Sometimes HANK students ignore DSGE models
  - Sometimes macro students ignore quantitative urban/trade models since these rely much more on statistical assumptions
- **They are penalized for this**
- I believe they ignore this work because they do not and cannot understand it
- You ignore lots of related work (same questions) if you only understand the way your advisor/community approaches problems
- **Reading work doesn't mean you like it**
  - Perfectly fine: The approach of my advisor/community is the correct approach to tackling the problems I care about, on philosophical and scientific grounds

Goal 2: Macro/finance students can read more broadly

- Stockholm macro/finance is top notch
- Most of it captures a very specific style of work
    - State of the art, and at the forefront of that style
- Even outside of that, people understand their advisor's work
- Other work is substantively relevant but stylistically different
- **Time and time again**: I have seen macro/finance students ignore that work
    - Sometimes macro labor students ignore structural labor
    - Sometimes HANK students ignore DSGE models
    - Sometimes macro students ignore quantitative urban/trade models since these rely much more on statistical assumptions
- **They are penalized for this**
- I believe they ignore this work because they do not and cannot understand it
- You ignore lots of related work (same questions) if you only understand the way your advisor/community approaches problems
- **Reading work doesn't mean you like it**
    - Perfectly fine: The approach of my advisor/community is the correct approach to tackling the problems I care about, on philosophical and scientific grounds
    - Dangerous: But I actually can't read or understand other approaches to the same problems I study...
- My goal: Give you the foundations to read that work.

Notation and terms

- *data*, *dataset*, or *sample*: multiple measurements of a set of variables
  - Examples: GDP per capita, unemployment rate, interest rates; earnings, educational attainment, age
- *Unit of observation*: Type of entity for which/whom measurement is taken
  - Country or quarter; individual or person
- *Observation*: distinct instances of the measurement
- In almost all instances, the variables we are looking at are *random variables* (an unknown, a function from the sample space), not the *realized value* (a fixed number).

Notation and terms

- *Parameters*: Mostly Greek letters
  - Mostly lower case are scalars, upper case are matrices
  - Sometimes upper case refers to the set of parameters: $\theta \in \Theta$
- *Estimator* of a parameter θ: denoted by a “hat”, “bar” (mostly a mean) or “tilde” (often an alternative to the hat estimator)
  - $\widehat{\theta}, \bar{\theta}, \tilde{\theta}$
- Central to understanding inference: the difference between a *parameter* – an unknown constant–, an *estimator* of those parameters – a function of sample data/random variables and therefore a *random variable*–, and an *estimate* – a realized value of an estimator)
- The variance of an estimator $\widehat{\theta}$ is often denoted as $\boldsymbol{V}_{\widehat{\theta}}$ and is the variance matrix of the quantity $\sqrt{n}(\widehat{\theta} - \theta)$: $\boldsymbol{V}_{\widehat{\theta}} = \text{Var}[\sqrt{n}(\widehat{\theta} - \theta)]$.

Data and data structures

- Typical data structures:
  - cross-sectional: index i, $(y_i, \boldsymbol{X}_i', \boldsymbol{Z}_i')$
  - (discrete) time series : index t, $(y_t, \boldsymbol{X}_t', \boldsymbol{Z}_t')$
  - panel (cross-sections of time series): index i, t: $(y_{it}, \boldsymbol{X}_{it}', \boldsymbol{Z}_{it}')$
- We sometimes assume that data are composed of mutually *independent* observations.
- Independence is a strong assumption and is defined in terms of the joint probability being the *product* of the marginal probabilities of the observations.
- This assumption is often not appropriate. We will discuss...
  - ... weaker assumptions (e.g., mean independence)
  - ... conditions in which we can do without it (sometimes by substituting combinations of other assumptions)
  - ... practical circumstances in which it is likely violated

Course organization

- The way this course is organized
    - Part 1: Estimation (me). **What is a "good" guess for some parameter's value?**
    - Part 2: Inference (Markus). **How plausible is it that the true parameter is some other value or set of values?**
- Alternative distinctions a course could be based around
    - (Note: These are distinctions I want you to pay attention to)
    - Reduced form methods vs. Structural methods
    - Causal inference vs. Descriptive methods
    - Regressions vs. Everything else
    - (Note: These three distinctions are similar and related but not identical)
    - Different estimators and estimation philosophies
      (next lecture will be organized around this)
    - Design-based vs. Model-based inference (i.e., what's random?)

Statistical and econometric program packages

- Statistical or econometric software package is a central requirement of econometric modelling.
    - Reason 1: Goal of econometrics is data analysis
    - Reason 2: Simulations are often easier than proofs and as valuable

Statistical and econometric program packages

- Statistical or econometric software package is a central requirement of econometric modelling.
    - Reason 1: Goal of econometrics is data analysis
    - Reason 2: Simulations are often easier than proofs and as valuable
- There are many possible packages to choose from and no single package is unambiguously best.
- Nearly all economists use one of five programs:
    - Stata (I use)
        - Still the most common in applied work (perhaps changing)
        - Best for replication code and data and inter-operability with coauthors
        - Better data management tools for standard datasets
        - Lots of standard, useful, mainstream tools are built in
        - Stata Corp. adds modern tools to each new version (getting much worse)
        - Users (incl. econometricians) code their own new tools and estimators (getting worse)

Statistical and econometric program packages

- There are many possible packages to choose from and no single package is unambiguously best.
- Nearly all economists use one of five programs:
  - Stata (I use)
  - R (this course)
    - Will plausibly eclipse Stata in near future
    - Better for non-standard stuff (more general; object oriented), clumsier and less efficient for a lot of standard stuff on standard-structured datasets
    - Better for programmers: Increasingly the choice of applied econometricians
    - Open source

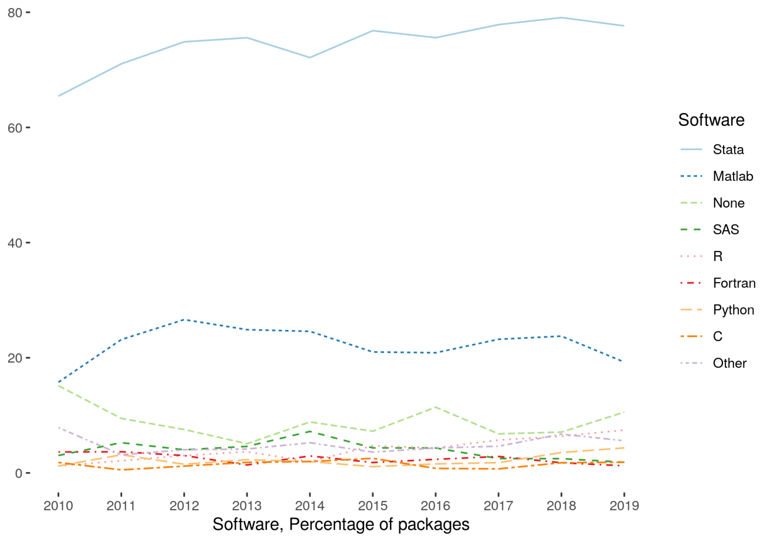

Figure: Software norms (from the AEA Replication Archive)

Statistical and econometric program packages

- There are many possible packages to choose from and no single package is unambiguously best.
- Nearly all economists use one of five programs:
    - Stata (I use)
    - R (this course)
    - Matlab ("similar" to R)
        - Main package used by macroeconomists (data, methods)
        - Best for solving dynamic programming models
        - Can do statistics, but only clunkily

Statistical and econometric program packages

- There are many possible packages to choose from and no single package is unambiguously best.
- Nearly all economists use one of five programs:
  - Stata (I use)
  - R (this course)
  - Matlab (“similar” to R)
  - Python
    - Best for “data science” tasks (text analysis, web scraping, but also machine learning)
    - (Note: David Stromberg teaches excellent elective on this stuff that all applied micro students should take [and some macro too!])
    - General programming language: Can do statistics, but not what it’s designed for
    - Lots of programmers put new “data science” tools in Python or R
    - Rarely are new statistics/econometrics tools coded in Python

Statistical and econometric program packages

- There are many possible packages to choose from and no single package is unambiguously best.
- Nearly all economists use one of five programs:
  - Stata (I use)
  - R (this course)
  - Matlab ("similar" to R)
  - Python: Best for "data science" tasks
  - Julia
    - General programming language
    - No idea what it's for, but some macroeconomists use it
    - The type of irrational "I love it so much I want to wear a t-shirt about it" borderline-cultish love that makes me skeptical...

Statistical and econometric program packages

- All demonstrations and examples in this class will be given in **R**.
- The **R** homepage https://www.r-project.org is useful:
    - the documentation sections lists many newbie documents
    - the program, documentation and central extension packages available for download
- Many guides to specialized areas, including econometrics, exist.

A note on work practices

- Two things to be mindful of when starting your career
  - You will revisit code you write today for many years
    - Writing a paper takes 1-3 years
    - Publication takes 1-4 years
    - Referees will request work that requires you to edit and change code from the earliest phase of the project
  - Replication is extremely important and many journals require you prepare detailed, complete, easy-to-follow replication code
- Some people are using github for version control
- Best advice is to read two things, ***EARLY*** before starting your first independent project, and take them to heart:
  - Gentzkow & Shapiro: Code and Data for the Social Sciences (2014): https://web.stanford.edu/ gentzkow/research/CodeAndData.pdf "spend less time wrestling with code, and more time on the research problems that got you interested in the first place."
  - AEA Replication guidelines: https://aeadataeditor.github.io/aea-de-guidance/preparing-for-data-deposit.html
- Best resource: Older students