# Introduction

Alexander Torgovitsky

ECON 31720: Applied Microeconometrics
University of Chicago, Fall 2020

# Logistics



## Schedule

- Lectures Tuesdays and Thursday 2:40–4:00
- Francesco's discussion section Wednesdays 4:10–5:00 (for now)
- My office hours are Fridays 4:10–5:00
- Francesco's office hours TBD

## Remote learning stuff

- Canvas is the main organizing tool, everything should be there
- Zoom lectures trying to recreate in-person experience
- In-person there is usually a lot of discussion

$\rightarrow$ So please ask questions — just start talking and interrupt me

- Any suggestions—let me know—this is new territory for me too

## Applied (vs Theoretical) Econometrics

- Methods that are (or can be) currently implemented in many applications
- This is not a value statement — most of my research is theoretical

→ But it takes months (years) of computation/tweaking to apply

- That's not what most applied work looks like in economics

→ If you work with data, you *must* think about common applied work

- Can you design an automobile if you don't know how to drive?

## *Micro*econometrics

- As opposed to *macro*econometrics or financial econometrics
- Also not a value statement — it is most empirical work in economics
- Greater focus on identification, causal inference
- Less focus on sampling issues (e.g. nonstationary time series)

### This course is for 2nd year PhD students in economics/Booth

- For those who want to do empirical work and/or econometric theory
- Many others have taken it (Political Science, Harris, Statistics, CS, . . . )
- I assume familiarity with (much of) the 1st year Econ PhD sequence
- I take a *laissez faire* approach to enrollment — up to you

### This course is *difficult*

- "Applied" econometrics $\neq$ easy econometrics

$\rightarrow$ I've never quite understood why many people assume that

- It does mean we won't use abstract mathematics
- It does *not* mean that we won't be mathematically rigorous
- Even more difficult are the non-technical skills of argumentation

$\rightarrow$ e.g. "Does this assumption make sense in this application?"

## Slides and supplemental notes

- These slides will be the main source
- I will also use the “blackboard” occasionally
- Lengthier derivations contained in supplemental notes to each slide deck

## Papers

- The supplemental notes contains a *curated* list of important papers

$\rightarrow$ *curated* $\Rightarrow$ short — I chose them carefully, so read them!

## Books?

- Unfortunately there are no textbooks that are sufficiently up-to-date
- Wooldridge (2010) is pretty good, but a bit outdated now
- I am not a fan of “Mostly Harmless” by Angrist and Pischke (2009)

$\rightarrow$ Nevertheless, I strongly encourage you to read it while taking this course

## Organization

- 4 problem sets, due roughly every 2–3 weeks throughout the course
- Due dates in syllabus, plan ahead — late problem sets not accepted
- Mixture of derivation (“theory”) and programming problems
- Everything except code must be in LaTeX

## Evaluation

- We grade on aesthetics too — take pride in your work — polish

→ Proofs, code, discussion should be clean, clear and concise (CCC)

- Comment your code, make your graphs and tables beautiful

## The problems are usually very hard, but …

- I design them to teach you something, not to torture you
- If you are doing a lot of algebra, you are probably doing it wrong
- If you are writing thousands of lines of code, probably also wrong

## Problem set guidelines

- You may not use high-level statistical commands
- → e.g. `lm` in R — instead we want you to write your own `lm`
- Important part of learning; now you will *really* understand the methods
- I realize everything is open source — ok to look, but don't copy
- Non-statistical routines (optimization, matrix inversion, etc.) are ok

## Languages

- I recommend R, Julia, or Python
- Other serious programming languages are ok, just ask me first
- R should be your choice if you want to do empirical work
- → Stata still popular in economics, but nowhere else—very error-prone
- Matlab is popular; I encourage replacing it with Julia (open source, *fast*)
- Learn something new — now is the time to build human capital

### What you will (hopefully) learn

- How to formally and precisely justify common empirical methods
- How the methods are commonly presented and discussed
- How the methods work on a deep level
- What the attendant assumptions mean, and common sources of violation
- A taste of some “frontier” techniques — the market rewards novelty
- To be a better programmer — get used to it, it’s your life now

### What you will not learn

- How to come up with a good empirical question/project
- How to find the appropriate data

$\rightarrow$ I wish I could teach you this, but that’s not my expertise

- There are many (*many*) empirical researchers at U Chicago who can

## Description

- Much of statistics is concerned with **descriptive tasks**
- These tasks concern "what is" questions
  1. What is the median income in the United States?
  2. What is the median income for 40 year old males in the US?
  3. What is the coefficient on age in a regression of income on A,B, and C?
- Focus on how to use data to formulate and answer "what is" questions

## Causality

- In contrast, this course is about **causality** and **causal inference**
- Causal inference focuses on using data to answer "what *if*" questions
- Requires thinking about a model of the "if" under consideration
- It's strictly harder than description

→ Because "what if" is always going to depend on "what is"

- But most interesting questions in economics are "what if" …

### Counterfactual questions

- What would happen if a job training program were expanded? [Labor]
- What would happen to prices/welfare if two firms merged? [IO]
- What would different monetary policy do to real output? [Macro]
- What effect would this medication have on heart disease? [Biostatistics]
- What will happen to global temps if emissions decrease? [Climatology]

### Causal inference

- Notice that these are all "what if" questions

$\Rightarrow$ Answering a counterfactual requires thinking about causality

- Theory alone might (*might*) allow us to provide a signed answer

$\rightarrow$ Even when it does, it will rarely tell us the *magnitude*

- Empiricism is needed, thus the role of causal inference in economics

**The focus of this course is on causal inference for empirical micro**

### Academics are often pretty good at it

- Most economists are pretty careful about it these days
→ Maybe not all of our friends who study macroeconomics
- Empirical political scientists, sociologists, epidemiologists all take care

### But the general public just doesn't *get* it

- Causal inference is one of the game-changing, mind-altering insights
→ Much like comparative advantage, returns to scale, equilibrium
- Many of the points may seem obvious ex-post; they are not ex-ante
- How many times have you heard someone say something like this?
→ "The economy is (was) booming because of Trump (Obama)"
- Really? How do you know the counterfactual?
- Journalists are unfortunately some of the worst offenders ... LOL

# Causal Reasoning (or Lack Thereof) in the *NY Times*



The New York Times

**The Morning**

September 17, 2020

By David Leonhardt

**Good morning. The C.D.C. makes the case for masks. The West Coast gets a rainy reprieve. And U.S. higher education faces troubles bigger than the virus.**

## The college money crisis

The coronavirus has caused severe budget problems for American higher education. But many colleges' financial troubles are much larger than the virus. They have been building for years and stem, above all, from a breakdown in this country's hodgepodge system of paying for higher education.

Given the importance of higher education — for scientific research, entrepreneurship and ultimately American living standards — I want to use today's newsletter to talk about this breakdown.

The New York Times

**The Morning**

September 17, 2020

(important person)

By David Leonhardt

**Good morning. The C.D.C. makes the case for masks. The West Coast gets a rainy reprieve. And U.S. higher education faces troubles bigger than the virus.**

## The college money crisis

The coronavirus has caused severe budget problems for American higher education. But many colleges' financial troubles are much larger than the virus. They have been building for years and stem, above all, from a breakdown in this country's hodgepodge system of paying for higher education.

Given the importance of higher education — for scientific research, entrepreneurship and ultimately American living standards — I want to use today's newsletter to talk about this breakdown.

# Causal Reasoning (or Lack Thereof) in the *NY Times*



The New York Times

**The Morning**

September 17, 2020

By David Leonhardt

(well-educated)

**Good morning. The C.D.C. makes the masks. The West Coast gets a rainy re U.S. higher education faces troubles l the virus.**

## The college money crisis

The coronavirus has caused severe budget problem higher education. But many colleges' financial troul larger than the virus. They have been building for y above all, from a breakdown in this country's hodge paying for higher education.

Given the importance of higher education — for scientific research, entrepreneurship and ultimately American living standards — I want to use today's newsletter to talk about this breakdown.

**David Leonhardt**

Leonhardt in 2012

| | |
|---|---|
| **Born** | January 1, 1973 (age 47)[1]<br>New York City, New York, U.S. |
| **Nationality** | American |
| **Education** | B.S., Applied Mathematics (1994) |
| **Alma mater** | Yale University |
| **Occupation** | Journalist, columnist |
| **Employer** | *The New York Times* |
| **Known for** | Pulitzer Prize for Commentary, 2011; Washington bureau chief, *The New York Times* (2011–2013) |

These budget cuts have left most colleges struggling for resources, even as elite colleges, both private and public, can raise substantial revenue from tuition and alumni donations. Not surprisingly, inequality in higher education has grown. Many poor and middle-class students who excel in high school attend colleges with inadequate resources and low graduation rates — and end up with student debt but no degree.

And research repeatedly shows that college matters: Graduates are more likely than nongraduates to be employed, to earn good salaries, to be happy and to live long lives.

causal statement

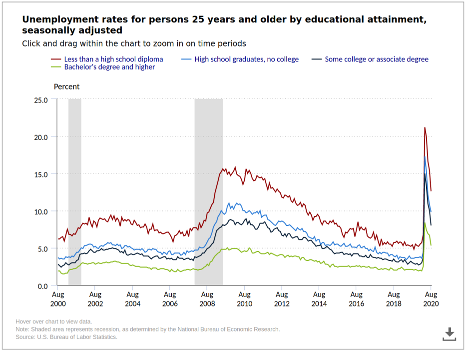
- Enormous economics literature on the causal effects of education
- The difficulty is the *non-random* assignment of education

$\rightarrow$ Those who attend college may be different than those who do not

- Clever strategies have been used to circumvent this
- Surely that must be what Leonhardt is linking to . . .

These budget cuts have left most colleges struggling for resources, even as elite colleges, both private and public, can raise substantial revenue from tuition and alumni donations. Not surprisingly, inequality in higher education has grown. Many poor and middle-class students who excel in high school attend colleges with inadequate resources and low graduation rates — and end up with student debt but no degree.

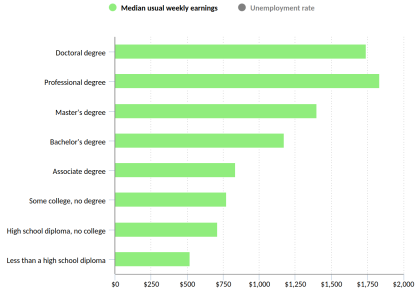
And research repeatedly shows that college matters: Graduates are more likely than nongraduates to be employed, to earn good salaries, to be happy and to live long lives.



These budget cuts have left most colleges struggling for resources, even as elite colleges, both private and public, can raise substantial revenue from tuition and alumni donations. Not surprisingly, inequality in higher education has grown. Many poor and middle-class students who excel in high school attend colleges with inadequate resources and low graduation rates — and end up with student debt but no degree.

And research repeatedly shows that college matters: Graduates are more likely than nongraduates to be employed, to earn good salaries, to be happy and to live long lives.

**Median weekly earnings and unemployment rate by educational attainment, 2017**




Click legend items to change data display. Hover over chart to view data.
Note: Data are for persons age 25 and over. Earnings are for full-time wage and salary workers.
Source: U.S. Bureau of Labor Statistics.

View Chart Data

These budget cuts have left most colleges struggling for resources, even as elite colleges, both private and public, can raise substantial revenue from tuition and alumni donations. Not surprisingly, inequality in higher education has grown. Many poor and middle-class students who excel in high school attend colleges with inadequate resources and low graduation rates — and end up with student debt but no degree.

And research repeatedly shows that college matters: Graduates are more likely than nongraduates to be employed, to earn good salaries, to be happy and to live long lives.

(serious researchers)

## Priceless: The Nonpecuniary Benefits of Schooling

Philip Oreopoulos

Kjell G. Salvanes

but did he actually read it?

JOURNAL OF ECONOMIC PERSPECTIVES
VOL. 25, NO. 1, WINTER 2011
(pp. 159-84)

Download Full Text PDF (Complimentary)

These budget cuts have left most colleges struggling for resources, even as elite colleges, both private and public, can raise substantial revenue from tuition and alumni donations. Not surprisingly, inequality in higher education has grown. Many poor and middle-class students who excel in high school attend colleges with inadequate resources and low graduation rates — and end up with student debt but no degree.

And research repeatedly shows that college matters: Graduates are more likely than nongraduates to be employed, to earn good salaries, to be happy and to live long lives.

(pg. 2)

Research on nonpecuniary effects of schooling is at an exciting and potentially productive stage. An accumulation of evidence suggests many ways, in and out of labor markets, that nonpecuniary effects of schooling might be quantitatively important. However, this suggestive evidence is plagued by two difficulties in drawing causal inferences. One difficulty, which is endemic in the literature on effects of schooling, is that a higher amount of schooling may be correlated with a wide array of other factors, like persistence, family background, perhaps even genetics. A persuasive argument about identifying the causal effects must find a way to disentangle the effect of schooling alone. A second difficulty, which is specific to the study of nonpecuniary effects, is that more schooling generates more income, and higher income will affect people's lives as well. Thus, in thinking about nonpecuniary effects of schooling, it's necessary to separate the effects taken alone from the effects of the higher incomes brought about by schooling.

These budget cuts have left most colleges struggling for resources, even as elite colleges, both private and public, can raise substantial revenue from tuition and alumni donations. Not surprisingly, inequality in higher education has grown. Many poor and middle-class students who excel in high school attend colleges with inadequate resources and low graduation rates — and end up with student debt but no degree.

And research repeatedly shows that college matters: Graduates are more likely than nongraduates to be employed, to earn good salaries, to be happy and to live long lives.

causal statement

# How Working-Class Life Is Killing Americans, in Charts

By David Leonhardt and Stuart A. Thompson March 6, 2020

When the economists Anne Case and Angus Deaton first published their research on "deaths of despair" five years ago, they focused on middle-aged whites. So many white working-class Americans in their 40s and 50s were dying of suicide, alcoholism and drug abuse that the overall mortality rate for the age group was no longer falling – a rare and shocking pattern in a modern society.

But as Case and Deaton continued digging into the data, it became clear that the grim trends didn't apply only to middle-aged whites. Up and down the age spectrum, deaths of despair have been surging for people without a four-year college degree:

These budget cuts have left most colleges struggling for resources, even as elite colleges, both private and public, can raise substantial revenue from tuition and alumni donations. Not surprisingly, inequality in higher education has grown. Many poor and middle-class students who excel in high school attend colleges with inadequate resources and low graduation rates — and end up with student debt but no degree.

And research repeatedly shows that college matters: Graduates are more likely than nongraduates to be employed, to earn good salaries, to be happy and to live long lives.

- Enormous economics literature on the causal effects of education
- The difficulty is the *non-random* assignment of education

→ Those who attend college may be different than those who do not

- Clever strategies have been used to circumvent this
- Surely that must be what Leonhardt is linking to . . . nope!

**Point: This stuff matters, and not "just" for research**

# Outline

### The Neyman-Fisher-Roy-Quandt-Rubin ("Potential outcomes") model

- Took off in biostatistics in 1970s, now widely used in economics
- Want to know how $D$ affects $Y$, so imagine "potential" $Y(d)$
- How does $Y(d)$ relate to a model of economic behavior?

### Latent variable models (for lack of a better name)

- Classic econometrics textbook: $Y = g(D, U)$

example:
$Y = \alpha + \beta D + u$

→ What is $U$? How do we interpret it?

- Allows for a broad range of assumptions

need to write a paragraph

### Directed acylcic graphs and the "do calculus"

- Associated with Judea Pearl — popular in computer science

titan / jerk

- The least popular of the three in economics (by far)

$$\begin{tikzcd} A \arrow[r] & B \\ C \arrow[ur] & \end{tikzcd}$$

Can't do parametric assumptions

→ Too general in certain ways, not enough in others

easier to get computers to follow this

### Variables

- $\mathcal{D}$ is a mutually exclusive and exhaustive set of states (“treatments”)
- e.g. training/no training $\mathcal{D} = \{0, 1\}$, prices $\mathcal{D} = [0, +\infty)$, etc.
- For each $d \in \mathcal{D}$ there is a **potential outcome** $Y(d)$ — a random variable
- $Y(d)$ is what $Y$ *would have* been if state were $d$ — imagination

### Relationship between actual and potential

- We observe the actual outcome $Y$ and actual state, $D \in \mathcal{D}$
- These are related to the potential outcomes via $Y(d, t) = \sum_t \mathbb{1}\{T = t\} Y(d)$

$$Y = \sum_{d \in \mathcal{D}} Y(d)\mathbb{1}[D = d] = Y(D)$$

$$\mathcal{D} = \{0, 1\}: \qquad Y = DY(1) + (1 - D)Y(0) = Y(D)$$

- $Y = Y(d)$ is observed for $d = D$, but not for $d \neq D$

# Latent Variable Notation



## Latent variable model

- Many empirical models in economics look like a special case of:

ex

$$Y = g(D, U), \quad = \alpha + \beta D + u \; \ldots$$

Suppose $E[u] = E[Du] = 0$
$\Rightarrow$ projection

- $Y$ and $D$ are outcome and treatment just as in potential outcomes
- $U$ is an unobservable (latent variable), $g$ is a function
- $U$ is everything that affects $Y$ besides $D$ (“all causes” interpretation)

↳ stronger Ass than example

## Latent variables generate potential outcomes

- This model also allows us to “imagine” outcomes if $d$ were different:

do(d) in Pearl

$$Y(d) \equiv g(d, U) \quad \text{for every } d \in \mathcal{D}$$

- So, a latent variable model *implies* a potential outcomes model. “Everything Else Interpretation”
- Latent variable models are thus strictly more general

Potential Outcome $\Rightarrow$ Latent Variable

### Benefits of potential outcomes

- Generality is not always good — simplicity and parsimony are virtues
- Many find potential outcomes easier to understand, conceptualize
- Widely used across fields (political science, epidemiology, etc.)

### Benefits of latent variables

- Enables clarity of what is embedded in the potential outcomes
- Generality enables additional restrictions, complex counterfactuals

$\rightarrow$ We will see examples of this later

### This course will use both notation, sometimes simultaneously

- The best of both worlds
- Potential outcomes as baseline, latent variable when complexity needed
- Some people have strong views, field-specific norms (e.g. labor vs. IO)
- Remember, *it's just notation* — use the above to translate

### Famous example: female labor supply (e.g. Heckman, 1974)

- How to interpret wage regressions among married women?

$\rightarrow$ But many women don't work (1960s–70s) — what determines this?

- Probably something to do with their potential wages $\Rightarrow$ selection bias

### "Missing data" problem — one-sided version of binary treatment

- $D \in \{0, 1\}$ is whether the woman is working or not
- If $D = 1$, observe wages $Y = Y(1)$ — but if $D = 0$ observe nothing
- Compare to $Y = DY(1) + (1 - D)Y(0)$ — just there's no $Y(0)$ here

$Y(0) = N/A$ (missing)

### One-sided selection is not in vogue, but just as relevant

- Testing for COVID-19 — who is getting tested and why?
- Election polls — who is responding and why? (2% response rates!!!)
- In both cases $\mathbb{E}[Y]$ is (arguably) what we want, but we get $\mathbb{E}[Y|D = 1]$

# Outline

## Models (assumptions)

- Allow us to interpret the data, formalize counterfactual questions
- So, the model you use depends on the question you want to answer

$\rightarrow$ The model should be rich enough to contemplate these counterfactuals

- As we will see, there will be other considerations too

$\rightarrow$ Are the assumptions (model) strong enough? Too strong?

## Target parameters

- How do we communicate what the model says about a counterfactual?
- Many ways to summarize/evaluate the outcome of an “if” question

$\rightarrow$ Who would be impacted under the counterfactual and how?

- These summaries are low(er)–dimensional features of the full model
- We will call them **target parameters** — what we want to know

### Setup

- $D \in \{0, 1\}$ indicates participation in a job training program
- $Y$ is earnings at some later date, potential outcomes $Y(0)$ and $Y(1)$

### Possible counterfactuals

1. What if the program were shut down?
2. What if the program were expanded?
3. What if random people were compelled to take or not take the program?

### Potential target parameters

1. Average treatment on the treated (ATT): $\mathbb{E}[Y(1) - Y(0)|D = 1]$
2. Average treatment on the untreated (ATU): $\mathbb{E}[Y(1) - Y(0)|D = 0]$
3. Average treatment effect (ATE): $\mathbb{E}[Y(1) - Y(0)]$

Or quantiles of the above, or cost-adjusted versions, etc. ...

## The question

- What can we conclude about the target parameter based on …?

1. The distribution of observable data — for example $(Y, D)$
2. Our maintained assumptions (the model)

- This is the *crucial* question of **identification**

## Preview of formality

- We will formalize identification mathematically in a future lecture
- Briefly, it is a *property* of the model/target parameter given data
- Different from how the phrase is (wrongly) used colloquially

$\rightarrow$ Often conflated with judgments of "credibility"

- We want to separate subjective from objective — will return to this

### Things all empiricists can agree on

- We have a question (target parameter)
- We have a model (assumptions)
- We have some data (we are empiricists)
- We want to use the model and data to answer the question

### Identification as logical consistency

- Identification is about which answers are *logically consistent* with this
- If target parameter is not **point identified** (definition discussed later)

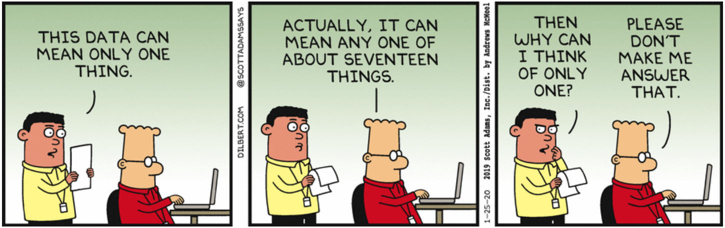
$\Rightarrow$ At least two target parameter values consistent with model and data

- No basis (illogical) to choose one value and exclude the other

### Nothing controversial here

- The problem is that formal identification analysis can be difficult
- So, researchers often use models where they don't know the properties

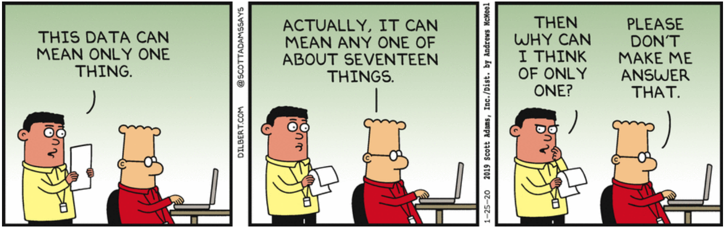
# Identification is a Natural Concept




"I think my model is point-identified"

# Identification is a Natural Concept



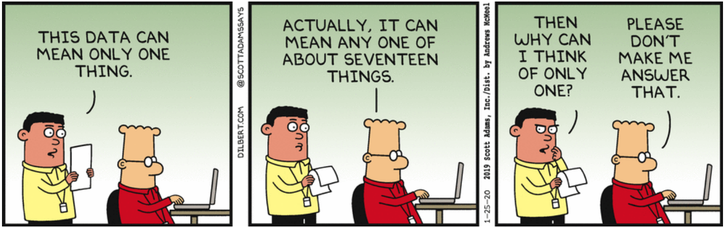



"Actually, it's partially identified"

"I think my model is point-identified"

# Identification is a Natural Concept






"Actually, it's partially identified"

"macroeconomists. . . "

"I think my model is point-identified"

You didn't do the identification analysis

## Population vs sample

- In practice, we only observe a finite **sample**, e.g. $\{(Y_i, D_i)\}_{i=1}^{n}$
- From this we know the **sample distribution**
- However, we don't know the **population distribution** of $(Y, D)$
- **Statistical inference** is using the sample to learn about the population

## Identification is logically prior to statistical inference

- It is useful to separate identification from statistical inference:

$$\text{sample} \quad \underbrace{\rightarrow}_{\text{statistical inference}} \quad \text{population} \quad \underbrace{\rightarrow}_{\text{identification}} \quad \text{unobserved parameters}$$

- The second arrow is logically the first thing to consider
- Can't recover a parameter when we know the population distribution?

$\rightarrow$ Then you also couldn't recover it with the sample distribution!

- Informally, identification is sometimes seen as having "infinite data"

### Job training example

- Suppose the counterfactual is ending the program
- Suppose target parameter is the ATT:

$$\text{ATT} \equiv \mathbb{E}[Y(1) - Y(0)|D = 1] = \underbrace{\mathbb{E}[Y|D = 1]}_{\text{fnc. of pop. dist.}} - \underbrace{\mathbb{E}[Y(0)|D = 1]}_{\text{fnc. of unobs.}}$$

### Identification vs statistical inference

- The first term is a function of the population distribution
- Using the sample to understand this from data is the domain of statistics
- The question of identification is about the second term
- What can we say about $\mathbb{E}[Y(0)|D = 1]$ under different assumptions?
- Must answer this question *before* we can construct an estimate of ATT

$\Rightarrow$ Identification is prior to statistical inference

## The Ideal: Forward engineering

1. Start with a question
2. Obtain some data that could be used to address the question
3. Develop one or more models that allow one to answer the question
4. Formulate specific target parameters in these models
5. Analyze their identification
6. Based on this analysis, construct and analyze estimators
7. Interpret results
8. Update knowledge/questions/data/models, then repeat

## This course

- We will embrace The Ideal when discussing current practice
- We will discuss how/when common practice deviates from The Ideal
- And how one might adjust common practice to work towards The Ideal

### Reverse engineering

- Currently fashionable to start with a common estimator

$\rightarrow$ For example ordinary least squares, or two-stage least squares

- Then attempt to interpret this estimator with regards to the counterfactual
- Even a methodological literature about formalizing these interpretations

### Reverse engineering is backward

- Why would we start with the *estimator*?
- Question $\rightarrow$ model $\rightarrow$ target parameter $\rightarrow$ estimator
- Reverse engineering is starting with the last part and working backwards

$\rightarrow$ Quite the opposite logic of forward engineering

- So, why do people do it?

$\rightarrow$ Common estimators are in common statistical packages (e.g. Stata)

- It's been embraced in influential circles — we will fight this trend

### Approximate definition?

- Structural: Fully specified model of an economic agent’s decisions
- Reduced form: A methodology that can be implemented in Stata
- “You know it when you see it”
- Terrible, terrible names — let’s not use them

### Debates about methodology

- One of the exciting aspects about empirical economics
- Structural (S) vs. reduced form (RF) is the big one
- RF fans say S models are not credible, make too many assumptions
- S fans say RF models have vague connection to economic questions
- The truth lies somewhere in between; we will embrace the best of both
- True believers of both approaches often fall short of The Ideal

# Outline

### Summary

1. Counterfactual questions are usually what we are interested in
2. Answering a counterfactual question requires a causal model
3. Two main models in economics: potential outcomes, latent variables
4. Regardless of the model, will have one or more target parameters
5. Identification is a crucial concept
6. Ideal approach is forward engineering from question to estimation results
7. Common practice to deviate from this through reverse engineering

### Next steps

- We will start talking about methods
- We will often talk about reverse engineering — important to know
- But we will also point towards forward engineering when possible
- We won't cover traditional "structural" models — a bit too specific