From experiments to observational data

STSCI / INFO / ILRST 3900

15 Sep 2026

Learning goals for today

At the end of class, you will be able to.

- Tie analysis of observational data to an idealized experiment
- Ask good questions which
  - involve treatments that exist (positivity assumption)
  - involve precise treatments (consistency assumption)

After class:

- Optional: Hernán, M. 2016.
  "Does water kill? A call for less casual causal inferences."
  Annals of Epidemiology 26(10):674–680.

Logistics

- First in-class quiz 9/17 (bring a writing utensil)
- Today's material will be on Quiz 2 (tentatively scheduled for Oct 8)

Linking observational data to an experiment

- Marginal exchangeability: $Y_i^{a=1}, Y_i^{a=0} \perp\!\!\!\perp A_i$ holds in conditionally experiments
- Is almost never true in observational data

Linking observational data to an experiment

- Marginal exchangeability: $Y_i^{a=1}, Y_i^{a=0} \perp\!\!\!\perp A_i$ holds in conditionally experiments
- Is almost never true in observational data

- Conditional exchangeability: $Y_i^{a=1}, Y_i^{a=0} \perp\!\!\!\perp A_i \mid L$ holds in conditionally randomized experiments
- We've typically discussed L being a single variable, but it could also be a set of variables
- But does it ever hold in observational data?

Linking observational data to an experiment

What is the effect of college degree on income at age 35

- $A_i = 1$ if four year college degree; $A_i = 0$ if no college degree
- Suppose we have information on parental income
  - $L_i = 0$: parents have high income
  - $L_i = 1$: parents have low income
- Does conditional exchangeability hold given parental income?

Linking observational data to an experiment

What is the effect of college degree on income at age 35

- $A_i = 1$ if four year college degree; $A_i = 0$ if no college degree
- Suppose we have information on parental income
  - $L_i = 0$: parents have high income
  - $L_i = 1$: parents have low income
- Does conditional exchangeability hold given parental income?
- What additional information would you gather to make conditional exchangeability plausible?

Conditional exchangeability in observational data

- Even if gathering data was possible for every covariate we want, when do we stop?

Conditional exchangeability in observational data

- Even if gathering data was possible for every covariate we want, when do we stop?
- Never 100% sure that conditional exchangeability holds
- Is it reasonable?

Conditional exchangeability in observational data

- Even if gathering data was possible for every covariate we want, when do we stop?
- Never 100% sure that conditional exchangeability holds
- Is it reasonable?
- In observational data, conditional exchangeability is an assumption we make (but can't typically verify)
- Requires expert knowledge

Conditional exchangeability in observational data

- Even if gathering data was possible for every covariate we want, when do we stop?
- Never 100% sure that conditional exchangeability holds
- Is it reasonable?
- In observational data, conditional exchangeability is an assumption we make (but can't typically verify)
- Requires expert knowledge
- Causal claims are data + outside knowledge

Formulating causal questions

Asking "good" causal questions involve

- Positivity condition: Treatments that exist
- Consistency: Treatments that are precise
- Accounts for interference

Good causal questions involve
treatments that exist

1. Treatments that exist

- Suppose we have conditional exchangeability given L
- To estimate a conditional ATE, we need to have treated and untreated individuals in each sub-group

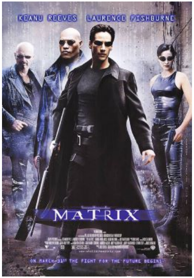
1. Treatments that exist

- Suppose we have conditional exchangeability given L
- To estimate a conditional ATE, we need to have treated and untreated individuals in each sub-group
- The assumption of **positivity** implies that for every value of ℓ:

$$Pr(A = a \mid L = \ell) > 0$$

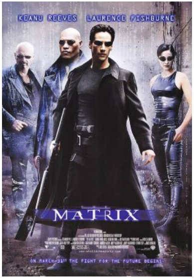
Positivity

- Suppose we are interested in the causal effect of watching the movie “The Matrix” in a theater on understanding potential outcomes



Positivity

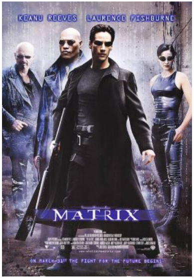
- Suppose we are interested in the causal effect of watching the movie "The Matrix" in a theater on understanding potential outcomes



- $A_i = 1$ if you saw "The Matrix" in a theater, $A_i = 0$ otherwise
- $Y_i = 1$ if you understand potential outcomes, $Y_i = 0$ otherwise
- Conditional exchangeability holds given Age ($L = \text{old}, \text{young}$)

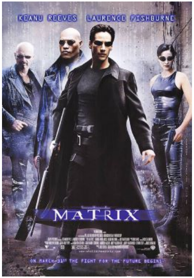
Positivity violation

- Suppose we are interested in the causal effect of watching the movie “The Matrix” in a theater on understanding potential outcomes



Positivity violation

- Suppose we are interested in the causal effect of watching the movie "The Matrix" in a theater on understanding potential outcomes



- $A_i = 1$ if you saw "The Matrix" in a theater, $A_i = 0$ otherwise
- $Y_i = 1$ if you understand potential outcomes, $Y_i = 0$ otherwise
- Conditional exchangeability holds given Age ($L = \text{old}, \text{young}$)

Positivity violation

What breaks down when positivity does not hold

Positivity violation

What breaks down when positivity does not hold

- Suppose no young people have seen "The Matrix" in theaters so that

$$Pr(\text{See "The Matrix" in theater} \mid Young) = 0.$$

- Under conditional exchangeability, standardization can get the ATE from conditional ATE
- Conditional ATE for young:

$$\underbrace{E(Y \mid A_i = 1, L = \text{young})}_{\text{Does not exist!}} - E(Y \mid A_i = 0, L = \text{young})$$

Positivity violation

What breaks down when positivity does not hold

- Suppose no young people have seen “The Matrix” in theaters so that

$$Pr(\text{See “The Matrix” in theater} \mid Young) = 0.$$

- Under conditional exchangeability, standardization can get the ATE from conditional ATE
- Conditional ATE for young:

$$\underbrace{E(Y \mid A_i = 1, L = \text{young})}_{\text{Does not exist!}} - E(Y \mid A_i = 0, L = \text{young})$$

- For IPW, we divide by propensity score (probability of treatment given covariates)
- Positivity violation causes divide by 0

Positivity violation remedy

- Question may still be interesting if positivity violation happens
- Question can't be answered from available data

Positivity violation remedy

- Question may still be interesting if positivity violation happens
- Question can't be answered from available data
- Modify question to sub-populations without positivity violation
- Example: Suppose we are interested in the causal effect of watching the movie "The Matrix" in a theater on understanding potential outcomes

Positivity violation remedy

- Question may still be interesting if positivity violation happens
- Question can't be answered from available data
- Modify question to sub-populations without positivity violation
- Example: Suppose we are interested in the causal effect of watching the movie "The Matrix" in a theater on understanding potential outcomes in the population of people older than 35

Good causal questions involve **precise treatments**

Consistency

Consistency says that

$$Y_i = Y^{a=A_i}$$

1. holds for precise treatments
2. holds with clarity about interference among units

2. Precise treatments

Imagine you are a high school counselor.

A statistician tells you

> The probability of receiving a BA in 6 years would be higher if a student initially enrolled in the State University of New York instead of a community college
>
> $$\text{P}\left(\text{BA}^{\text{Enroll in SUNY}}\right) > \text{P}\left(\text{BA}^{\text{Enroll in Community College}}\right)$$

How would you advise students?

2. Precise treatments

2. Precise treatments

6-year graduation rate	
Binghamton University, State University of New York	83%
Stony Brook, State University of New York (Established 1957)	78%
University at Buffalo, The State University of New York	74%
University at Albany, State University of New York	66%

2. Precise treatments

The treatment value
`Enroll in SUNY`
is not sufficiently precise

	6-year graduation rate
Binghamton University, State University of New York	83%
Stony Brook, State University of New York, Established 1957	78%
UB University at Buffalo, The State University of New York	74%
University at Albany, State University of New York	66%

2. Precise treatments

The treatment value

`Enroll in SUNY`

is not sufficiently precise

$$\begin{aligned} \text{BA}^{\text{Binghamton}} &\neq \text{BA}^{\text{Stony Brook}} \\ &\neq \text{BA}^{\text{Buffalo}} \\ &\neq \text{BA}^{\text{Albany}} \end{aligned}$$

	6-year graduation rate
Binghamton University, State University of New York	83%
Stony Brook, State University of New York, Established 1957	78%
University at Buffalo, The State University of New York	74%
University at Albany, State University of New York	66%

2. Precise treatments

The treatment value

`Enroll in SUNY`

is not sufficiently precise

$$
\begin{aligned}
\text{BA}^{\text{Binghamton}} &\neq \text{BA}^{\text{Stony Brook}} \\
&\neq \text{BA}^{\text{Buffalo}} \\
&\neq \text{BA}^{\text{Albany}}
\end{aligned}
$$

To advise the student,
a precise treatment
is more helpful

	6-year graduation rate
Binghamton University, State University of New York	83%
Stony Brook, State University of New York, Established 1957	78%
UB University at Buffalo, The State University of New York	74%
University at Albany, State University of New York	66%

2. Precise treatments

Consistency assumption: $Y = Y^A$

More credible when A is very precise

- it is clear how to run a hypothetical experiment
- is is clear how to inform policy

Example:

- if $a =$ SUNY, then Y^a is vague.
  To which SUNY should you send the student?
- if $a =$ Binghamton, then Y^a is clearer

A good read:
Hernán, M. 2016.
"Does water kill? A call for less casual causal inferences."
Annals of Epidemiology 26(10):674–680.

Good causal questions involve
clarity about interference

Interference

- So far: your outcome depends on **your** potential outcomes and **your** treatment

Interference

- So far: your outcome depends on **your** potential outcomes and **your** treatment
- Is this true for vaccines, ad campaigns, police interventions?

Interference

- So far: your outcome depends on **your** potential outcomes and **your** treatment
- Is this true for vaccines, ad campaigns, police interventions?
- We need to define more potential outcomes:
    - $Y^{a_{\text{you}}=1, a_{\text{your friend}}=1}$
    - $Y^{a_{\text{you}}=1, a_{\text{your friend}}=0}$
    - $Y^{a_{\text{you}}=0, a_{\text{your friend}}=1}$
    - $Y^{a_{\text{you}}=0, a_{\text{your friend}}=0}$
- Typically need to know how complex the interactions can be

Good causal questions: In math

We should study treatments that exist (positivity)

$$\mathrm{P}(A = a \mid \vec{L} = \vec{\ell}) > 0$$

with potential outcomes that are well-defined (consistency)

$$Y = Y^A$$

Well-defined potential outcomes involve precise treatments

$$\mathrm{BA}^{\text{Binghamton}} \quad \text{instead of} \quad \mathrm{BA}^{\text{SUNY}}$$

and incorporate interference when it exists

$$Y^{a_{\text{you}}, a_{\text{your friend}}} \quad \text{instead of} \quad Y^{a_{\text{you}}}$$

Learning goals for today

At the end of class, you will be able to.

- Tie analysis of observational data to an idealized experiment
- Ask good questions which
  - involve treatments that exist (positivity assumption)
  - involve precise treatments (consistency assumption)

After class:

- Optional: Hernán, M. 2016.
  "Does water kill? A call for less casual causal inferences."
  Annals of Epidemiology 26(10):674–680.