Causal inference and experimental methods

Macartan Humphreys

Feb 2017

Roadmap

Roadmap

- Lecture 1: ▸ What is a cause? ▸ Potential Outcomes ▸ Estimands ▸ Endogenous subgroups

Take home ideas

- A causal claim is a claim about what did not happen.
- Random assignment to treatment is random sampling from alternative universes.
- You have to have an estimand. You should be able to describe this in terms of **potential outcomes**.
- Regression requires many more assumptions in order to lay claim to estimating average causal effect.

Lecture 1: What's a cause?

Motivation

The *intervention* based motivation for understanding causal effects:

- We want to know if a particular intervention (like aid) caused a particular outcome (like reduced corruption).
- We need to know:
  1. What happened?
  2. What would the outcome have been if there were no intervention?
- The problem
  1. . . . this is hard
  2. . . . this is impossible

  The problem in 2 is that you need to know what would have happened if things were different. You need information on a **counterfactual**

Notation

We will use:

- Y to denote an outcome (the dependent variable, the left hand side variables, the endogeneous variables)
- X to denote a cause (the independent variable, the driver, the right hand side variables, the exogeneous variables)
- $X_1, X_2, \ldots$ for particular causes

Different research projects:

- $? \rightarrow Y$: What causes Y?
- $X \rightarrow ?$: What does X do?
- $X \rightarrow Y?$: Does X cause Y?
- $? \rightarrow ??$: What's up?

Know your X and your Y.

The Potential Outcomes Framework

Potential Outcomes: Simple case

- For each unit we assume that there are two **post-treatment** outcomes: $Y_i(1)$ and $Y_i(0)$.
- eg $Y(1)$ is the outcome that **would** obtain *if* the unit received the treatment.
- The **causal effect** of Treatment (relative to Control) is:

$$\tau_i = Y_i(1) - Y_i(0)$$

- Note:
  - the causal effect is defined at the *individual level*.
  - there is no "data generating process" or functional form
  - the causal effect is defined relative to something else and so a counterfactual must be conceivable (did Germany cause the second world war?)
  - are there any substantive assumptions made here so far?

Causal claims: What is seen?

- We have talked about what's potential, now what do we *observe*?
- Say Z_i indicates whether the unit i is assigned to treatment ($Z_i = 1$) or not ($Z_i = 0$). It describes the treatment process. Then what we observe is:

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

- Say Z is a random variable, then this is a sort of data generating process. BUT the key things to note is
    - Y_i is random but the randomness comes from Z_i — the potential outcomes, $Y_i(1)$, $Y_i(0)$ are fixed
    - Compare this to a regression approach in which Y is random but the X's are fixed. eg:

$$Y \sim N(\beta X, \sigma^2) \text{ or } Y = \alpha + \beta X + \epsilon, \epsilon \sim N(0, \sigma^2)$$

Implications of the counterfactual definition

Statements about what did not happen

Inference: We define causes in terms of things that did not happen. This puts **inference** front and center.

Compare this to a **structural approach**. In a structural approach we assume a data generating process in which outcome Y is a function of X.

$$y = f_Y(x, u_Y), u_Y \sim p_Y$$

Compared to:

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

- Here knowing the effect of X on Y means knowing the functional form f_Y and knowing background conditions, u_Y.
- It does not require forming beliefs about what did not happen (the counterfactual claims are now buried inside f_Y)

Structural approach

Illustration: Say that cellphones turn on if and only if they are **intact** and have a **charged battery**.

Question: Does providing a charged battery cause cellphones to turn on?

- In the **potential outcomes** framework this requires making guesses from the data about how cellphones work with and without batteries. It is not a measure problem, but an *inference* problem.
- For the **structural approach** this figuring out whether cellphones are intact. It is masurement problem (conditional on the model).

Causal claims: Transitivity and connectedness

Now that we have a concept of causal effects available, let's answer two **questions**:

- If for a given unit A causes B and B causes C, does that mean that A causes C?

- Say A causes B — does that mean that there is a spatiotemporally continuous sequence of causal intermediates?

Causal claims: Transitivity and connectedness

Now that we have a concept of causal effects available, let's answer two **questions**:

- If for a given unit A causes B and B causes C, does that mean that A causes C?
  A boulder is flying down a mountain. You duck. This saves your life.
  So the boulder caused the ducking and the ducking caused you to survive. So: *did the boulder cause you to survive?*
- Say A causes B — does that mean that there is a spatiotemporally continuous sequence of causal intermediates?
  Person A is planning some action Y; Person B sets out to stop them; person X intervenes and prevents person B from stopping person A. In this case Person A may complete their action, producing Y, without any knowledge that B and X even exist; in particular B and X need not be anywhere close to the action. So: *did X cause Y?*

Causal claims: Contribution or attribution?

The counterfactual model is all about contribution, not attribution, except in a very conditional sense.

- Focus is on non-rival contributions
- Not: what caused Y but **what is the effect of** X?
- At most it provides a conditional account

Consider at outcome Y that might depend on two causes X_1 and X_2:

$$Y(0,0) = 0$$

$$Y(1,0) = 0$$

$$Y(0,1) = 0$$

$$Y(1,1) = 1$$

What caused Y? Which cause was most important?

Causal claims: Contribution or attribution?

The counterfactual model is all about contribution, not attribution, except in a very conditional sense.

- Focus is on non-rival contributions
- Not: what caused Y but what is the effect of X?
- At most it provides a conditional account
- This is problem for research programs that define "explanation" in terms of figuring out the things that cause Y
- Real difficulties conceptualizing what it means to say one cause is more important than another cause. What does that mean?

Causal claims: Contribution or attribution?

The counterfactual model is all about contribution, not attribution, except in a very conditional sense.

- Focus is on non-rival contributions
- Not: what caused Y but what is the effect of X?
- At most it provides a conditional account
- *Erdogan's increasing authoritarianism was the most important reason for the attempted coup*
  - More important than Turkey's history of coups?
  - What does that mean?

The difference between an *actual* cause a *counterfactual* cause

- Susie and Billy throw rocks at a bottle
- Both are great aims and, if they were the only ones throwing, would certainly hit the bottl.e
- Susie is a faster throw however and her rock hits the bottle first.
- Billy's throw is a little slower and flys past the now broken bottle.

Question: Which stone broke the bottle?

The difference between an *actual* cause a *counterfactual* cause

- Susie and Billy throw rocks at a bottle
- Both are great aims and, if they were the only ones throwing, would certainly hit the bottl.e
- Susie is a faster throw however and her rock hits the bottle first.
- Billy's throw is a little slower and flys past the now broken bottle.

Question: From a counterfactual perspective *neither* broke the bottle since the bottle would have broken in the absence of any single throw. Yet it seems obvious that Susie's throw actually broke the bottle. (To think about: what does "actually" caused actually mean?)

Causal claims: No causation without manipulation

- Some seemingly causal claims not admissible.
- To get the definition off the ground, **manipulation must be imaginable** (whether practical or not)
- This renders thinking about effects of race and gender difficult
- What does it mean to say that Aunt Pat voted for Brexit because she is old?

Causal claims: No causation without manipulation

- Some seemingly causal claims not admissible.
- To get the definition of hte ground, **manipulation must be imaginable** (whether practical or not)
- This renders thinking about effects of race and gender difficult
- **Compare**: What does it mean to say that Southern counties voted for Brexit because they have many old people?

Causal claims: Causal claims are everywhere

Which of these statements involve causal claims:

- Jack exploited Jill
- It's Jill's fault that bucket fell
- Jack is the most obstructionist member of Congress
- Melania Trump stole from Michelle Obama's speech

So:

- Policymakers and activists need causal claims

The Fundamental Problem and a Solution

Causal claims: The estimand and the rub

- The causal effect of Treatment (relative to Control) is:

$$\tau_i = Y_i(1) - Y_i(0)$$

- This is what we want to estimate
- BUT: We never can observe both $Y_i(1)$ and $Y_i(0)$!
- This is the **fundamental problem** (Holland)

Causal claims: The rub and the solution

- Now for some magic. We really want to estimate:

$$\tau_i = Y_i(1) - Y_i(0)$$

- BUT: We never can observe both $Y_i(1)$ and $Y_i(0)$
- Say we lower our sights and try to estimate an **average** treatment effect:

$$\tau = E(Y(1) - Y(0))$$

- Now make use of the fact that

$$E(Y(1) - Y(0)) = E(Y(1)) - E(Y(0))$$

- In words: *The average of differences is equal to the difference of averages*; here, the average treatment effect is equal to the difference in average outcomes in treatment and control units.
- The magic is that *while we can't hope to measure the differences; we are good at measuring averages*.

Causal claims: The rub and the solution

- So we want to estimate $E(Y(1))$ and $E(Y(0))$.
- We know that we can estimate averages of a quantity by taking the average value from a random sample of units
- To do this here we need to select a random sample of the $Y(1)$ values and a random sample of the $Y(0)$ values, in other words, we **randomly assign** subjects to treatment and control conditions.
- When we do that we can in fact estimate:

  $$E_N(Y_i(1)|Z_i = 1) - E_N(Y_i(0)|Z_i = 0)$$

  which in expectation equals:

  $$E(Y_i(1)|Z_i = 1 \text{ or } Z_i = 0) - E(Y_i(0)|Z_i = 1 \text{ or } Z_i = 0)$$

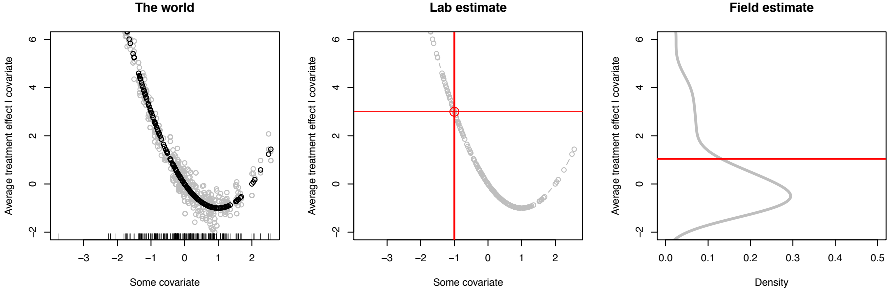
- This highlights a deep connection between **random assignment** and **random sampling**: when we do random assignment *we are in fact randomly sampling from different possible worlds*.

For some this assignment is **definitional** of experiments

The **random assignment** is critical here. In fact the assignment is, for some, **definitional** to an experiment.

- Experiments are investigations in which an intervention, in all its essential elements, is under the control of the investigator. (Cox & Reid)
- Two major types of control:
    1. control over assignment to treatment – this is at the heart of many field experiments
    2. control over the treatment itself – this is at the heart of many lab experiments
- Main focus today is on 1 and on the question: *how does control over assignment to treatment allow you to make reasonable statements about causal effects?*

Experiments



How randomization helps

This provides a **positive argument** for causal inference from randomization, rather than simply saying with randomization "everything else is controlled for"

Let's discuss:

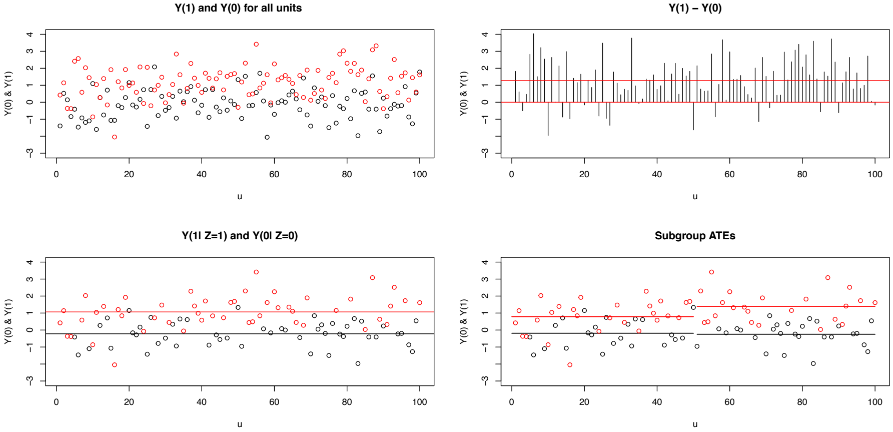
- Does the fact that an estimate is unbiased mean that it is right?
- Can a randomization "fail"?
- Where are the covariates?

Idea: random assignment is random sampling from potential worlds: to understand anything you find, you need to know the sampling weights

Potential outcomes: why randomization works

The average of the differences ≈ difference of averages

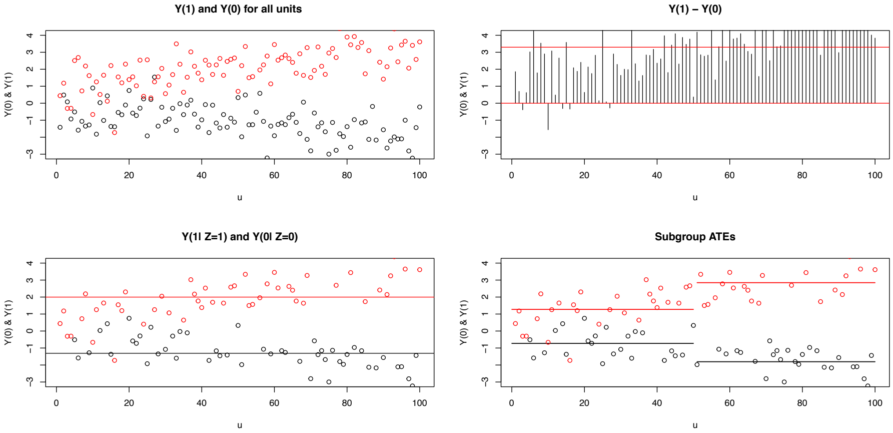
```
po.graph(N, Y0, Y1, u, Z)
```



Potential outcomes: heterogeneous effects

The average of the differences ≈ difference of averages

```
po.graph(N, Y0 - u/50, Y1+u/50, u,Z)
```



Potential outcomes: heterogeneous effects

Question: $\approx$ or $=$?

Estimands and Estimators

Estimands

- The estimand is the thing you want to estimate
- If you are estimating something you should be able to say what your estimand is
- You are responsible for your estimand. Your estimator will not tell you what your estimand is
- Just because you can calculate something does not mean that you have an estimand
- You can test a hypothesis without having an estimand

Estimands: The Average Treatment Effect

- The most common estimand is the **average treatment effect**
- This is a **summary** of individual treatment effects:

$$\tau_i = Y_i(1) - Y_i(0)$$

- For persons 1 and 2, the average is:

$$\tau = \frac{1}{2}(Y_1(1) - Y_1(0)) + \frac{1}{2}(Y_2(1) - Y_2(0))$$

- More generally:

$$\tau = \sum_i \frac{1}{n}(Y_i(1) - Y_i(0))$$

Estimands: ATE, ATT, ATC, S-, P-, C-, ITT, LATE

Say that units are randomly assigned to treatment in different strata (maybe just one); with fixed, though possibly different, shares assigned in each stratum. Then the key estimands and estimators are:

$$\begin{aligned}
\tau_{ATE} \equiv \quad & E(\tau_i) & = \quad & \sum_x \frac{w_x}{\sum_j w_j}\tau_x & \widehat{\tau}_{ATE} = \quad & \sum_x \frac{w_x}{\sum_j w_j}\widehat{\tau}_x \\
\tau_{ATT} \equiv \quad & E(\tau_i | Z_i = 1) & = \quad & \sum_x \frac{p_x w_x}{\sum_j p_j w_j}\tau_x & \widehat{\tau}_{ATT} = \quad & \sum_x \frac{p_x w_x}{\sum_j p_j w_j}\widehat{\tau}_x \\
\tau_{ATC} \equiv \quad & E(\tau_i | Z_i = 0) & = \quad & \sum_x \frac{(1-p_x)w_x}{\sum_j (1-p_j)w_j}\tau_x & \widehat{\tau}_{ATC} = \quad & \sum_x \frac{(1-p_x)w_x}{\sum_j (1-p_j)w_j}\widehat{\tau}_x
\end{aligned}$$

where x indexes strata, p_x is the share of units in each stratum that is treated, and w_x is the size of a stratum.

Here:

- ATE is Average Treatment Effect (all units)
- ATT is Average Treatment Effect on the Treated
- ATC is Average Treatment Effect on the Controls

Estimands: ATE, ATT, ATC, S-, P-, C-, ITT, LATE

In addition, each of these can be targets of interest:

- for the **population**, in which case we refer to PATE, PATT, PATC and $\widehat{PATE}, \widehat{PATT}, \widehat{PATC}$
- for a **sample**, in which case we refer to SATE, SATT, SATC, and $\widehat{SATE}, \widehat{SATT}, \widehat{SATC}$

And for different subgroups,

- given some value on a covariate, in which case we refer to CATE (conditional average treatment effect)
- for unobservable subgroups, we estimate LATE (Local Average Treatment Effect (see below).

With non-compliance we might estimate ITT —the "intention to treat" effect

Skip to ▸ Fixer or ▸ Inference 1 or ◂ Big Ideas

Exercise your potential outcomes 1

Consider the following potential outcomes table:

Unit	Y(0)	Y(1)	τ_i
1	4	3	
2	2	3	
3	1	3	
4	1	3	
5	2	3	

Questions for us: What are the unit level treatment effects? What is the average treatment effect?

Exercise your potential outcomes 2

Consider the following potential outcomes table:

In treatment?	Y(0)	Y(1)
Yes		2
No	3	
No	1	
Yes		3
Yes		3
No	2	

Questions for us: Fill in the blanks.

- Assuming a constant treatment effect of $+1$
- Assuming a constant treatment effect of -1
- Assuming an *average* treatment effect of 0

What is the actual treatment effect?

Endogeneous subgroups

Endogeneous Subgroups

Experiments often give rise to endogenous subgroups. The potential outcomes framework can make it clear why this can cause problems.

Heterogeneous Effects with Endogeneous Categories

- Problems arise in analyses of subgroups when the categories themselves are affected by treatment
- Example from our work:
    - You want to know if an intervention affects reporting on violence against women
    - You measure the share of all subjects that experienced violence that file reports
    - The problem is that which subjects experienced violence is itself a function of treatment

Heterogeneous Effects with Endogeneous Categories

It is possible that in truth no one's reporting behavior has changed, what has changed is the propensity of people with different propensities to report to experience violence:

	Violence(Treatment)		Reporting(Treatment, Violence)			
	V(0)	V(1)	R(0,1)	R(1,1)	R(0,0)	R(1,0)
Type 1 (reporter)	1	1	1	1	0	0
Type 2 (non reporter)	1	0	0	0	0	0

Expected reporting given violence in control = Pr(Type 1)

Expected reporting given violence in treatment = 100%

Question: What is the actual effect of treatment on the propensity to report violence?

Heterogeneous Effects with Endogeneous Categories

It is possible that in truth no one's reporting behavior has changed, what has changed is the propensity of people with different propensities to report to experience violence:

	Reporters		Non reporters		
	Experience Violence		Experience Violence		
	No	Yes	No	Yes	% Report
Control	25	25	25	25	$\frac{25}{25+25} = 50\%$
Treatment	25	25	50	0	$\frac{25}{25+0} = 100\%$

Heterogeneous Effects with Endogeneous Categories

This problem can arise as easily in seemingly simple field experiments. Example:

- In one study we provided constituents with information about performance of politicians
- we told politicians in advance so that they could take action
- we wanted to see whether voters punished poorly performing politicians
- what's the problem?

Heterogeneous Effects with Endogeneous Categories

Question for us:

Setting:

- Quotas for women are randomly placed in a set of constituencies in year 1. All winners in these areas are women; in other areas only some are.
- In year 2 these quotas are then lifted.

Questions Which problems face an endogenous subgroup issue?:

1. You want to estimate the likelihood that a woman will stand for reelection in treatment versus control areas in year 2.
2. You want to estimate how much incumbents are more likely to be reelected in treatment versus control areas in year 2.
3. You want to estimate how much treatment areas have more relected incumbents in elections in year 2 compared to control.

Heterogeneous Effects with Endogeneous Categories

In such cases you can:

- Examine the joint distribution of multiple outcomes
- Condition on pretreatment features only
- Engage in mediation analysis

Recap: Ten things you need to know about causal inference

1. A causal claim is a statement about what didn't happen.
2. There is a fundamental problem of causal inference.
3. You can estimate average causal effects even if you cannot observe any individual causal effects.
4. If you know that A causes B and that B causes C, this does not mean that you know that A causes C.
5. The counterfactual model is all about contribution, not attribution.
6. X can cause Y even if there is no "causal path" connecting X and Y.
7. Correlation is not causation
8. X can cause Y even if X is not a necessary condition or a sufficient condition for Y.
9. Estimating average causal effects does not require that treatment and control groups are identical.
10. There is no causation without manipulation

`http://egap.org/resources/guides/causality/`

END

Extra Slides

Missing data can create an endogeneous subgroup problem

- It is well known that missing data can undo the magic of random assignment.
- One seemingly promising approach is to match into pairs *ex ante* and drop pairs together *ex post*.
- Say potential outcomes looked like this (four units divided into two pairs):

Table 1: Full profile of potential outcomes

Pair	I	I	II	II	
Unit	1	2	3	4	Average
Y(0)	0	0	0	0	
Y(1)	-3	1	1	1	
τ	-3	1	1	1	

Missing data

- Say though that cases are likely to drop out of the sample if things go badly (eg they get a negative score or die)
- Then you might see no attrition in cases in which people that are likely to drop out if treated do not get treated.
- You might assume you have no problem (after all, no attrition).

Table 2: No missing data when the normal cases happens to be selected

Pair	I	I	II	II	
Unit	1	2	3	4	Average
Y(0)	0		0		0
Y(1)		1		1	1
$\hat{\tau}$					1

Missing data

- But in cases in which you have attrition, dropping the pair doesn't necessarily help.
- The problem is potential missingness still depends on potential outcomes
- The kicker is that the method can produce bias even if (*in fact*) there is no attrition!

Table 3: Missing data when the vulnerable cases happens to be selected

Pair	I	I	II	II	
Unit	1	2	3	4	Average
Y(0)		[0]	0		0
Y(1)	[-3]			1	1
$\hat{\tau}$					1

Missing data

[Footnote: The right way to think about this is that bias is a property of the strategy over possible realizations of data and not normally a property of the estimator conditional on the data.]

Multistage games

Multistage games can also present an endogenous group problem since collections of late stage players facing a given choice have been created by early stage players.

Multistage games

Question: Does **visibility** alter the extent to which subjects follow norms to punish antisocial behavior (and reward prosocial behavior)? Consider a trust game in which we are interested in how information on receivers affects their actions

Table 4: Return rates given investments under different conditions

		% invested	Average % returned	
		(average)	...when 10% invested	...when 50% invested
Treatment	Masked information on respondents	30% (avg)	20%	40%
	Full information on respondents	30% (avg)	0%	60%

What do we think? Does visibility make people react more to investments?

Multistage games

Imagine you could see all the potential outcomes, and they looked like \ this:

Table 5: Potential outcomes with (and without) identity protection

		Responder's return decision (given type)						Avg.
		Nice 1	Nice 2	Nice 3	Mean 4	Mean 4	Mean 6	
Offerer	Invest 10%:	60%	60%	60%	0%	0%	0%	30%
behavior	Invest 50%:	60%	60%	60%	0%	0%	0%	30%

Conclusion: Both the offer and the information condition are **completely irrelevant** for all subjects. . .

Multistage games

Unfortunately you only see a sample of the potential outcomes, and that looks like this:

Table 6: Outcomes when respondent **is visible**

		Responder's return decision (given type)						Avg.
		Nice 1	Nice 2	Nice 3	Mean 4	Mean 4	Mean 6	
Offerer	Invest 10%:				0%	0%	0%	0%
behavior	Invest 50%:	60%	60%	60%				60%

False Conclusion: When not protected, responders condition behavior *strongly* on offers (because offerers can select on type accurately)

Multistage games

Unfortunately you only see a sample of the potential outcomes, and that looks like this:

Table 7: Outcomes when respondent **is not visible**

		Responder's return decision (given type)						Avg.
		Nice	Nice	Nice	Mean	Mean	Mean	
		1	2	3	4	4	6	
Offerer	Invest 10%:			60%		0%	0%	20%
behavior	Invest 50%:	60%	60%		0%			40%

False Conclusion: When protected, responders condition behavior less strongly on offers (because offerers can select on type less accurately)

Multistage games

What to do? **Solutions?**

1. Analysis *could* focus on the effect of treatment on respondent behavior, directly.
    - This would get the correct answer but to a different question [Does information affect the share of contributions returned by subjects on average? No]
2. **Strategy method** can sometimes help address the problem, **but** that is also (a) changing the question and (b) putting demands on respondent imagination and honesty
3. First mover action could be **directly manipulated**, but unless deception is used that is also changing the question
4. First movers could be **selected** because they act in predictible ways (bordering on deception?)

Idea: Proceed with extreme caution when estimating effects beyond the first stage.

Skip to ▸ Mediation or ◂ Big Ideas