MY457/MY557: Causal Inference for Observational and Experimental Studies

Week 7: Instrumental Variables 1

Daniel de Kadt

Department of Methodology
LSE

Winter Term 2025

Course Outline

- **Week 1:** The potential outcomes framework
- **Week 2:** Randomized experiments
- **Week 3:** Selection on observables I
- **Week 4:** Selection on observables II
- **Week 5:** Selection on observables III
- Week 6: Reading week
- **Week 7:** Instrumental variables I
- **Week 8:** Instrumental variables II
- **Week 9:** Regression discontinuity
- **Week 10:** Difference-in-differences I
- **Week 11:** Difference-in-differences II

Example: Segregation, Inequality, and Poverty

Does residential segregation lead to racialised economic outcomes?

Ananat (2011) studies this relationship at the city-level in the USA, focused on two outcomes:

1. Black poverty rates
2. Black-white income inequality

But this is a very hard question to study. Why?

Hard to imagine that there are not many confounders:

- Residential segregation has numerous causes
- Some of those causes must surely cause racialised economic outcomes
- These problems become especially acute over long time periods
- Selection problems via selective migration

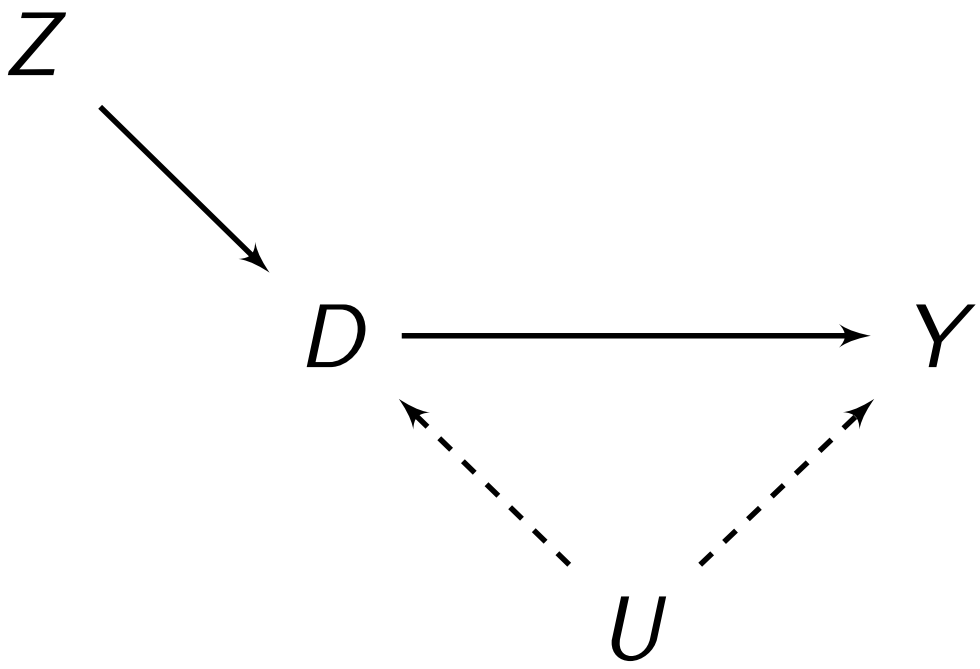
Example: Segregation, Inequality, and Poverty

The design problem in the author's own words:

> To test for these or other patterns of outcomes requires empirical variation approaching a randomized experiment. Ideally, one would conduct the following test using two initially identical cities with small open economies:
>
> 1. At time zero, one city would be assigned perfect residential segregation, the other perfect residential integration.
>
> 2. Each city would be randomly assigned black residents from the initial black skill distribution and white residents from the initial white distribution.
>
> 3. Then, the relationship between segregation and the income distribution of the offspring generation would be measured. This is the individual-treatment effect of segregation.
>
> 4. Finally, residents would be allowed to move, and aggregate demand for cities (rent, migration) by race and skill would be measured to determine tastes for segregation and its consequences. This is the selection effect of segregation.

Enter instrumental variables (IV)...

Instrumental Variables: Graphical Intuition



Idea: Find some variable Z that induces 'as-if random' variation in D.

Study only that variation in D, and how to is related to Y.

Ignore (partial out) variation in D attributable to U.

Example: Segregation, Inequality, and Poverty

Ananat (2011) proposes the railroad division index (RDI):

1. Digitize 19th century city maps
2. From each city centre, draw a 4km-radius circle
3. Measure how dispersed the city's area is in terms of neighborhoods

RDI should affect post-Great Migration segregation

FIGURE 1. THE NATURAL EXPERIMENT—2 EXAMPLES

Example: 'First Stage' and Falsification

TABLE 1—TESTING RDI AS AN INSTRUMENT

	First stage	Falsification checks					
		1910 city characteristics					
Outcome:	1990 dissimilarity index[a] (1)	Physical area (square miles/ 1,000)[a] (2)	Pop. (1,000s)[b] (3)	Ethnic dissimilarity index[a] (4)	Ethnic isolation index[a] (5)	Percent black[b] (6)	Street-cars per cap. (1,000s) (1915)[a] (7)
RDI	**0.357** **(0.088)**	−3.993 (11.986)	0.666 (1.36)	0.076 (0.185)	0.027 (0.070)	−0.0006 (0.0100)	−0.132 (0.183)
Track length per square kilometer	18.514 (10.731)	−574.401 (553.669)	75.553 (135)	15.343 (53.249)	−12.439 (17.288)	**9.236** **(0.650)**	3.361 (20.507)
Mean of dependent variable	0.568	14.626	1,527	0.311	0.055	1.442 percent	179
N	121	58	121	49	49	121	13

Focus on column 1: This is the 'first stage', how RDI affects segregation

Note also columns 2-7: Essentially balance checks. (SOO anyone?)

Example: IV Results

TABLE 2—THE EFFECTS OF SEGREGATION ON POVERTY AND INEQUALITY AMONG BLACKS AND WHITES

	OLS: Effect of 1990 dissimilarity index		Main results: 2SLS RDI as instrument for 1990 dissimilarity		Falsification: Reduced form effect of RDI among cities far from the south	
Outcome:	Whites (1)	Blacks (2)	Whites (3)	Blacks (4)	Whites (5)	Blacks (6)
Within-race poverty and inequality						
Gini index	**−0.079**	**0.459**	**−0.334**	**0.875**	−0.110	0.167
	(0.037)	**(0.093)**	**(0.099)**	**(0.409)**	(0.066)	(0.424)
Poverty rate	**−0.073**	**0.182**	**−0.196**	**0.258**	−0.036	−0.136
	(0.019)	**(0.045)**	**(0.065)**	**(0.108)**	(0.035)	(0.094)

Focus on columns 3 and 4: These are the IV estimates (estimated using two-stage least squares – 2SLS – more later)

If assumptions satisfied, these give the estimated effect of the variation in segregation induced by RDI on the outcome of interest. This is a very specific interpretation!

Instrumental Variables: Back to Basics

The motivating example is a case of 'classical' instrumental variables in an observational study.

We are going to learn IV from the 'modern' perspective, which subsumes the classical perspective.

You may have encountered classical IV before – try to set aside some of what you know!

To do this, we will begin by studying IV in experimental settings with just a binary treatment and a binary instrument.

Next week we will cover some extensions of IV, including continuous treatments.

Noncompliance in Randomised Experiments

Let's begin by returning to randomised experiments (it's safe there!).

Randomised experiments can have compliance problems: Despite randomisation, units may control whether they are actually treated.

Canonical example: Non-compliance in JTPA Experiment

	Not Enrolled in Training	Enrolled in Training	Total
Assigned to Control	3,663	54	3,717
Assigned to Training	2,683	4,804	7,487
Total	6,346	4,858	11,204

This is yet another selection problem, our age-old concern!

Implication: Even in a randomised experiment, we may not be able to naïvely compare groups...

Instrumental Variables: Setup

Assume an encouragement: $Z_i \in \{0, 1\}$

We now define treatment potential outcomes under Z: $D_{zi} \in \{D_{1i}, D_{0i}\}$

1. $D_{zi} = 1$: would receive the treatment if $Z_i = z$
2. $D_{zi} = 0$: would not receive the treatment if $Z_i = z$

   e.g., $D_{1i} = 1$ encouraged to take treatment and takes treatment

   Note: encouragement $\neq$ treatment

   Instead: treatment $= f$(encouragement)

We can also define our outcome potential outcomes: $Y_{(Z_i, D_{Z_i i})i}$

What is observed in a given trial?

- Observed treatment indicator: $D_i = D_{Z_i i}$ for $Z_i = z$
- Observed outcome of Y_i: $Y_i = Y_{(Z_i, D_{Z_i i})i}$ for $Z_i = z$
- Thus observed outcome of Y_i can also be written as $Y_i = Y_{Z_i i}$

Compliance Types

Given our setup, we can define four compliance types:

- Unit i is a complier if: $D_{1i} = 1$ and $D_{0i} = 0$
- and a non-complier if $\begin{cases} \text{Always-takers:} & D_{1i} = D_{0i} = 1 \\ \text{Never-takers:} & D_{1i} = D_{0i} = 0 \\ \text{Defiers:} & D_{1i} = 0 \text{ and } D_{0i} = 1 \end{cases}$

Or, written as principal strata:

This basically sais that we have 4 types of compliance types but we can only identify 2 categories - compliers or defiers

Treatment	Encouragement: $Z_i = 1$	Encouragement: $Z_i = 0$
$D_i = 1$	Complier/Always-taker	Defier/Always-taker
$D_i = 0$	Defier/Never-taker	Complier/Never-taker

Causal Estimand: The ITT

Intention-to-Treat:

$$\tau_{ITT} = \frac{1}{N}\sum_{i=1}^{N}(Y_{(1,D_{1i})i} - Y_{(0,D_{0i})i})$$

or equivalently

$$\tau_{ITT} = \mathbb{E}[Y_{(1,D_{1i})i} - Y_{(0,D_{0i})i}]$$

Read: Effect of encouragement on outcome (regardless of treatment status)

Note: If there is non-compliance, self-selection into the treatment/control groups may mean $\tau_{ITT} \neq \tau_{ATE}$

In experiments we call this an encouragement design, with randomised Z such that $\{Y_{zd}\} \perp\!\!\!\perp Z$. In such settings, our identification result is:

$$\tau_{ITT} = \mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]$$

IV: Assumptions

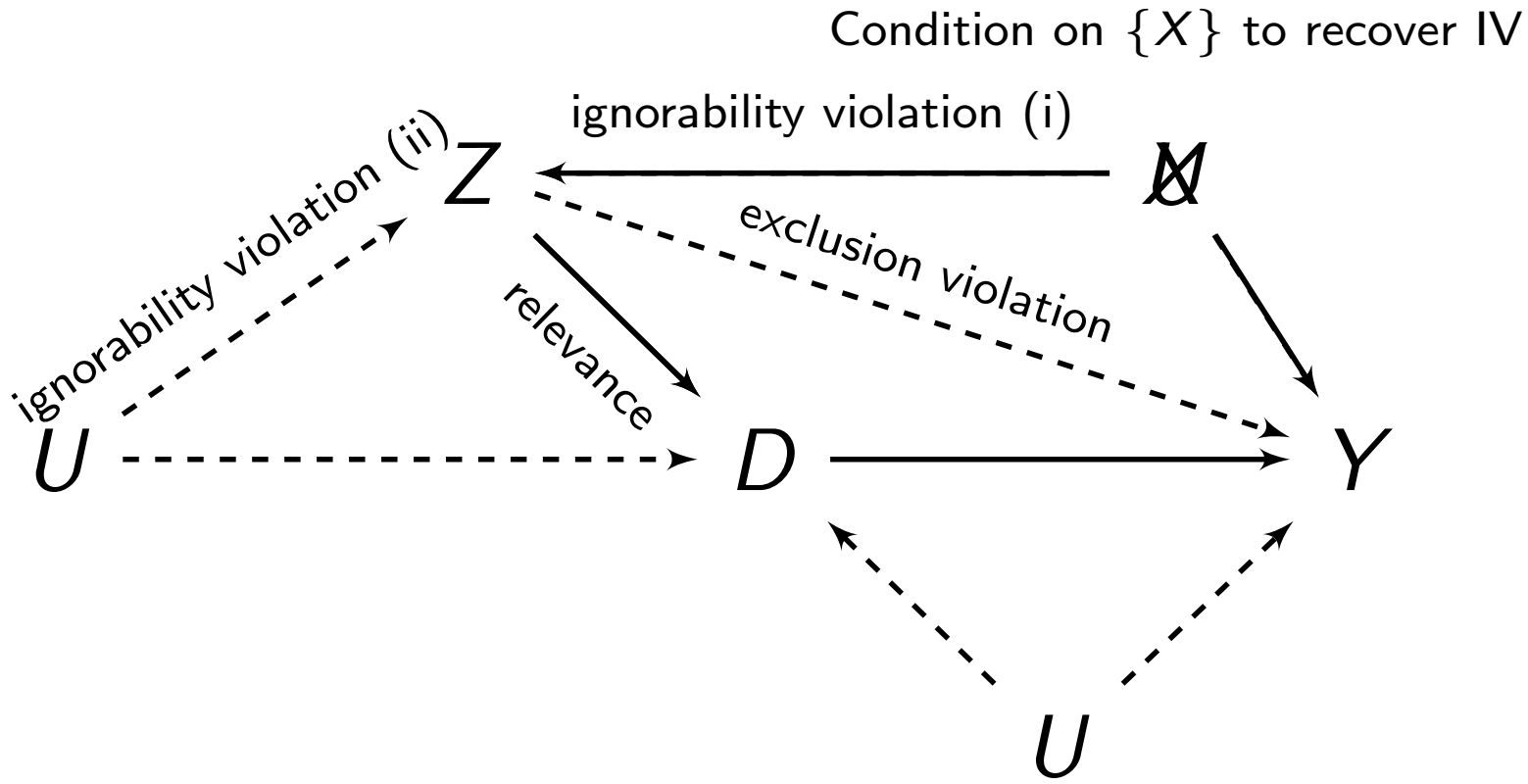
The ITT only allows us to say something about the effect of Z on Y, but what about the effect of D?

Idea: Perhaps we can (under some assumptions) express the effect of D on Y in terms of the ITT.

Five assumptions give us just such an identification result:

1. SUTVA
2. Relevance of the instrument: $0 < P(Z = 1) < 1$ and $P(D_1 = 1) \neq P(D_0 = 1)$
3. Ignorability or exogeneity of the instrument: $\{Y_{zd}, D_z\} \perp\!\!\!\perp Z$
   (i) $\rightsquigarrow$ $\{Y_{zd}\} \perp\!\!\!\perp Z$ (sufficient for ITT)
   (ii) $\rightsquigarrow$ $\{D_z\} \perp\!\!\!\perp Z$
4. Exclusion restriction: $Y_{1,d} = Y_{0,d}$ for $d = 0, 1$.
5. Monotonicity: $D_1 \geq D_0$ ('no defiers')

IV: Relevance, Ignorability, and Exclusion



Identification Proof: Decomposing τ_{ITT}

τ_{ITT} can be decomposed into a combination of subgroup ITTs:

$$\begin{aligned}\tau_{ITT} &= \tau^{c}_{ITT} \times \Pr(\text{compliers}) + \tau^{a}_{ITT} \times \Pr(\text{always-takers}) \\ &\quad + \tau^{n}_{ITT} \times \Pr(\text{never-takers}) + \tau^{d}_{ITT} \times \Pr(\text{defiers})\end{aligned}$$

where

$$\begin{aligned}\tau^{c}_{ITT} &= \mathbb{E}[Y_{1i,D_{1i}} - Y_{0i,D_{0i}} \mid D_{1i} = 1, D_{0i} = 0], \\ \tau^{a}_{ITT} &= \mathbb{E}[Y_{1i,D_{1i}} - Y_{0i,D_{0i}} \mid D_{1i} = D_{0i} = 1], \text{ etc.}\end{aligned}$$

Under monotonicity and exclusion restriction, this simplifies as:

$$\begin{aligned}\tau_{ITT} &= \tau^{c}_{ITT} \times \Pr(\text{compliers}) + \tau^{a}_{ITT} \times \Pr(\text{always-takers}) \\ &\quad + \tau^{n}_{ITT} \times \Pr(\text{never-takers}) + 0 \quad [\because \text{monotonicity}] \\ &= \tau^{c}_{ITT} \times \Pr(\text{compliers}) + 0 \times \Pr(\text{always-takers}) \\ &\quad + 0 \times \Pr(\text{never-takers}) \quad [\because \text{exclusion restriction}] \\ &= \tau^{c}_{ITT} \times \Pr(\text{compliers})\end{aligned}$$

Identification Result

Therefore, τ^c_{ITT} can be nonparametrically identified:

$$\begin{aligned}\tau^c_{ITT} &= \frac{\tau_{ITT}}{\Pr(\text{compliers})} \\ &= \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(D_i \mid Z_i = 1) - \mathbb{E}(D_i \mid Z_i = 0)}\end{aligned}$$

τ^c_{ITT} is the Local Average Treatment Effect (LATE) for compliers:

$$\tau^c_{ITT} = \tau_{LATE} = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_{1i} = 1, D_{0i} = 0]$$

LATE has a clear causal meaning, but interpretation is often tricky:

- How do we generalise from compliers to everyone else?
- We can never identify who exactly the compliers actually are
- Different encouragements (instruments) may yield different compliers

IV: Plug-in Estimator

Recall the LATE identification result:

$$\tau_{LATE} = \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(D_i \mid Z_i = 1) - \mathbb{E}(D_i \mid Z_i = 0)} = \frac{\mathrm{Cov}(Y_i, Z_i)}{\mathrm{Cov}(D_i, Z_i)}$$

A plug-in estimator is called the Wald estimator:

$$\widehat{\tau_{LATE}} = \frac{\frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i) Y_i}{\frac{1}{n_1}\sum_{i=1}^{n} Z_i D_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i) D_i} = \frac{\widehat{\mathrm{Cov}}(Y_i, Z_i)}{\widehat{\mathrm{Cov}}(D_i, Z_i)}$$

where $n_1 = \sum_{i=1}^{n} D_i$ and $n_0 = n - n_1$

- The Wald estimator is consistent, but not unbiased in finite samples
- The small sample bias may be considerable when the instrument is weak (i.e. when $\widehat{\mathrm{Cov}}(D_i, Z_i) \simeq 0$, more later)

IV: Two Stage Least Squares Estimator

$\widehat{\tau_{LATE}}$ can also be estimated via two-stage least squares (2SLS), a traditional regression-based instrumental variables estimator. Note that the same small sample bias concerns apply!

Consider two regression functions that generate our potential outcomes:

1. $D_z = \mu + \rho Z + \eta$ (first stage)
2. $Y_{zd} = \gamma + \alpha D + \varepsilon$ (second stage)

2SLS estimator runs OLS twice to estimate these stages:

- Stage 1: Regress D on Z and obtain fitted values ($\hat{D}$'s)
- Stage 2: Regress Y on $\hat{D}$

Note: As always, we assert homogeneous treatment effects! Becomes an issue when controlling for X.

Can be implemented in `R` with using `lm` (but your SEs will need to be corrected) or with `AER::ivreg`.

Example: 'First Stage' in Ananat (2011)

TABLE 1—TESTING RDI AS AN INSTRUMENT

	First stage	Falsification checks					
		1910 city characteristics					
Outcome:	1990 dissimilarity index[a] (1)	Physical area (square miles/ 1,000)[a] (2)	Pop. (1,000s)[b] (3)	Ethnic dissimilarity index[a] (4)	Ethnic isolation index[a] (5)	Percent black[b] (6)	Street-cars per cap. (1,000s) (1915)[a] (7)
RDI	**0.357**	−3.993	0.666	0.076	0.027	−0.0006	−0.132
	(0.088)	(11.986)	(1.36)	(0.185)	(0.070)	(0.0100)	(0.183)
Track length per square kilometer	18.514	−574.401	75.553	15.343	−12.439	**9.236**	3.361
	(10.731)	(553.669)	(135)	(53.249)	(17.288)	**(0.650)**	(20.507)
Mean of dependent variable	0.568	14.626	1,527	0.311	0.055	1.442 percent	179
N	121	58	121	49	49	121	13

Example: 'Second Stage' in Ananat (2011)

TABLE 2—THE EFFECTS OF SEGREGATION ON POVERTY AND INEQUALITY AMONG BLACKS AND WHITES

	OLS: Effect of 1990 dissimilarity index		Main results: 2SLS RDI as instrument for 1990 dissimilarity		Falsification: Reduced form effect of RDI among cities far from the south	
Outcome:	Whites (1)	Blacks (2)	Whites (3)	Blacks (4)	Whites (5)	Blacks (6)
Within-race poverty and inequality						
Gini index	**−0.079**	**0.459**	**−0.334**	**0.875**	−0.110	0.167
	(0.037)	**(0.093)**	**(0.099)**	**(0.409)**	(0.066)	(0.424)
Poverty rate	**−0.073**	**0.182**	**−0.196**	**0.258**	−0.036	−0.136
	(0.019)	**(0.045)**	**(0.065)**	**(0.108)**	(0.035)	(0.094)

Ignorability Violations

Researchers often under-appreciate that the causal interpretation of IV hinges on the ignorability of Z.

When is that more plausible than the ignorability of D? Do we risk returning to SOO world?

Consider, e.g. the canonical paper by Acemoglu et al (2001) which has $19,000$ citations and a Nobel prize:

- Study effect of institutions on economic outcomes
- Use settler mortality rates to instrument for institutional types
- But surely disease environment is not ignorable?
- Is this actually any better than a naïve SOO analysis?

Falsification tests can help:

- Balance tests (*a la* selection on observables)
- Placebo tests (all types)

Exclusion Violations

More attention has typically been paid to exclusion violations.

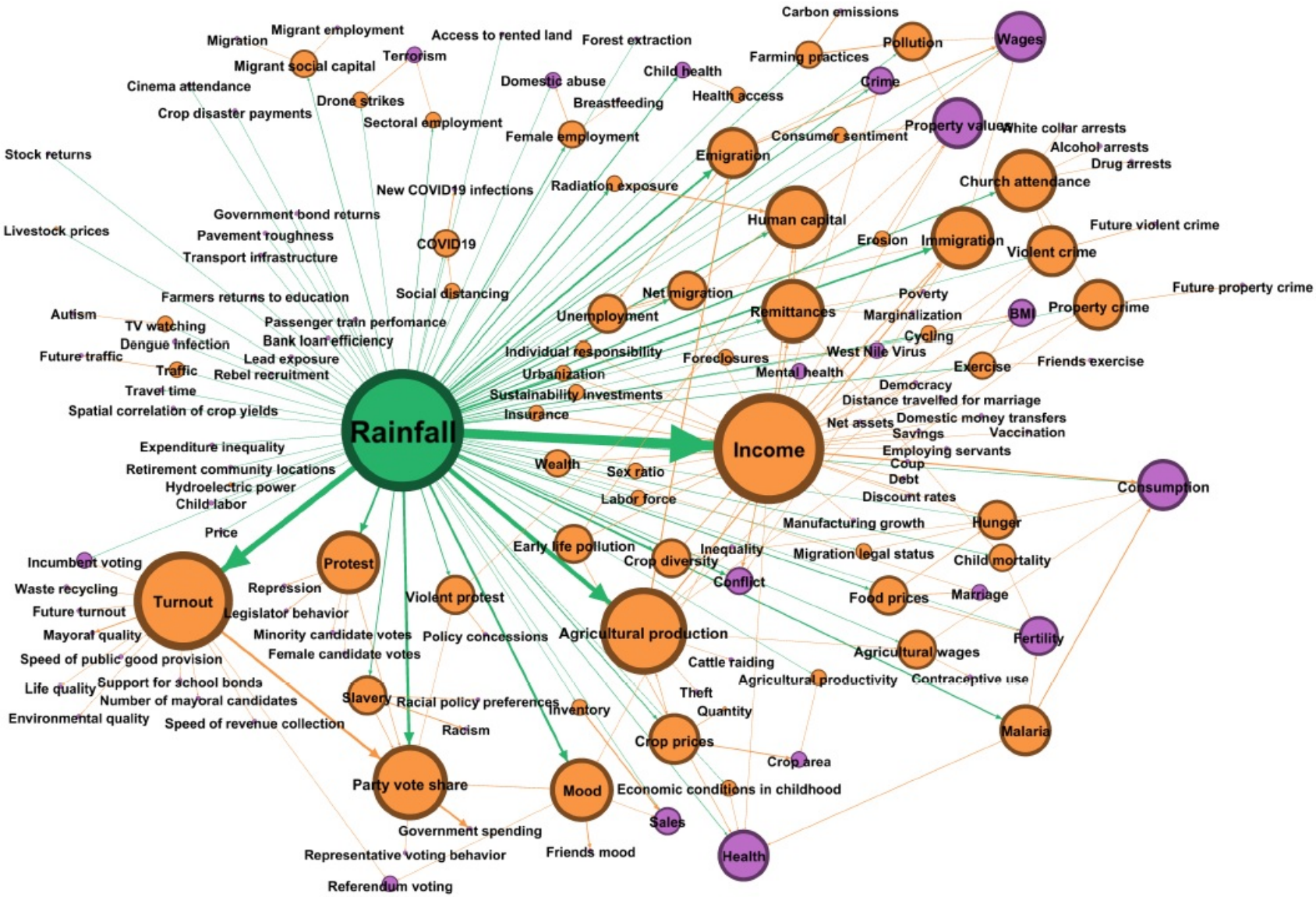
Violations of the exclusion restriction are typically unobservable – it is akin to speculation about mechanisms in a causal graph

Again, falsification tests can help:

- Placebo outcome tests on alternative Y'
- Placebo population tests

One common problem is that people often want to use the same instrument multiple times...

Example: Rainfall as an Instrument (Mellon, 2024)



Exclusion Violations: A Bayesian Approach

Intuitively, note that the size of the exclusion restriction problem is roughly proportional to the ratio of the LATE and the exclusion violation.

That is, if the LATE is large and the exclusion violation very small, we can perhaps ignore the problem.

There are some Bayesian solutions, e.g. the 'plausibly exogenous' framework (Conley et al. 2012):

- Place a prior on the exclusion restriction violation
- Estimate the IV given that prior

Weak IV

Weak instruments – those that only weakly affect D – have different asymptotic properties to non-weak instruments

Question: When is an instrument 'relevant enough'?

Traditionally, researchers focused on the first stage F-statistic (greater than 10 was considered good)

Lots of ongoing debate, see Stock & Yogo (2005), Lee et al. (2022), Angrist & Kolesár (2023)

But at a fundamental level, what exactly are we doing here? If the instrument has only a very weak influence on treatment, what variation in D are we really studying in the first place?

Better LATE than Nothing?

Short of further assumptions, τ_{LATE} is not generally equal to τ_{ATE} or τ_{ATT}.

Consider, however, one-sided non-compliance:

- $D_{0i} = 0$ (where $Z_i = 0$)
- $D_{1i} \in \{0, 1\}$ (where $Z_i = 1$)

In this setting, $\tau_{LATE} = \tau_{ATT}$. Why?

- We now have no always takers: $D_{0i} = 0 \ \forall i$
- Recall that $\tau^c_{ITT} = \tau_{LATE} = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_{1i} = 1, D_{0i} = 0]$
- Now, $\mathbb{E}[Y_{1i} - Y_{0i} \mid D_{1i} = 1, D_{0i} = 0] = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_{1i} = 1]$ (why?)
  $D_{0i} = 0$ is true for all, so conditioning adds no information.
- And $\mathbb{E}[Y_{1i} - Y_{0i} \mid D_{1i} = 1] = \mathbb{E}[Y_{1i} - Y_{0i} \mid Z_i = 1, D_i = 1]$ (why?)
  Given $Z_i = 0$ for control units and $D_{0i} = 0 \ \forall i$, if $D_i = 1$ then $Z_i = 1$
- So: $\mathbb{E}[Y_{1i} - Y_{0i} \mid Z_i = 1, D_i = 1] = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1] = \tau_{ATT}$

Questions of external validity still remain, however. (See the Deaton and Imbens exchange.)

Characterising Compliers

We can't observe compliers, but may be able to characterize compliers in terms of some covariates X

Marbach & Hangartner (2020) offer simple and intuitive method:

1. Observe $f(X)$ (e.g. mean) for always-takers (treated in the non-encouraged group)
2. Observe $f(X)$ for never-takers (control in the encouraged group)
3. Subtract off the weighted $f(X)$ and you are left with the $f(X)$ for compliers.

Aronow & Carnegie (2013) suggest we can go even further:

1. Estimate $P_{C_i} = \Pr(D_{1i} > D_{0i})$, the compliance score
2. Use inverse compliance score weighting to move from LATE to ATE

   (But only if our estimation of P_{C_i} works well!)