Causal Inference Methods in Data Science
Lecture 5: Methods for dealing with unmeasured confounding

Lin Liu

April 29, 2024

IV

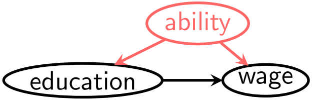
One motivating example

- Labor economists have long been interested in determining the causal effect of education on wage

- However, no randomized trials can be conducted to randomly assign people to or not to get higher education

- The only hope is to rely on observational studies

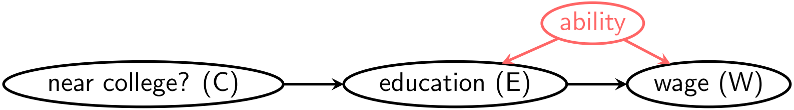
- Consider the following causal DAG:



- Obviously, if the data does not contain measurements of ability (almost impossible to measure it anyway), association between education and wage is not causation

One motivating example

- Labor economists have long been interested in determining the causal effect of education on wage
- However, no randomized trials can be conducted to randomly assign people to or not to get higher education
- The only hope is to rely on observational studies
- Instead, Card (1995) consider the following causal DAG:



- Can we identify $\tau_{E \to W}$, since $\tau_{B \to W}$ is composed of $\tau_{B \to E}$ and $\tau_{E \to W}$, and both causations, $\tau_{B \to W}$ and $\tau_{B \to E}$, are associations?

A real data analysis

let's analyze Card's data

Example 1

```r
library(ivreg)
data("SchoolingReturns", package = "ivreg")

## simple linear regression
edu_wage_ols <- lm(log(wage) ~ education + poly(experience,
    2, raw = TRUE) + ethnicity + smsa + south, data =
    SchoolingReturns)
summary(edu_wage_ols)

## IV regression
edu_wage_iv <- ivreg(log(wage) ~ education + poly(experience
    , 2, raw = TRUE) + ethnicity + smsa + south |
    nearcollege + poly(age, 2, raw = TRUE) + ethnicity +
    smsa + south, data = SchoolingReturns)
```

Another motivating example: the role of lipoprotein subfractions on heart diseases

- Lipoprotein subfractions have long been conjectured to play an essential role in the development or prevention of heart diseases

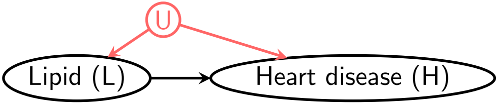
Another motivating example: the role of lipoprotein subfractions on heart diseases

- Lipoprotein subfractions have long been conjectured to play an essential role in the development or prevention of heart diseases

- But people reported conflicting data analysis, indicating mixed information on whether it is beneficial or detrimental to heart based on observational studies

Another motivating example: the role of lipoprotein subfractions on heart diseases

- Lipoprotein subfractions have long been conjectured to play an essential role in the development or prevention of heart diseases
- But people reported conflicting data analysis, indicating mixed information on whether it is beneficial or detrimental to heart based on observational studies
- The causal DAG:

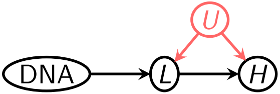


Another motivating example: the role of lipoprotein subfractions on heart diseases

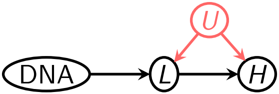
- Lipoprotein subfractions have long been conjectured to play an essential role in the development or prevention of heart diseases

- But people reported conflicting data analysis, indicating mixed information on whether it is beneficial or detrimental to heart based on observational studies

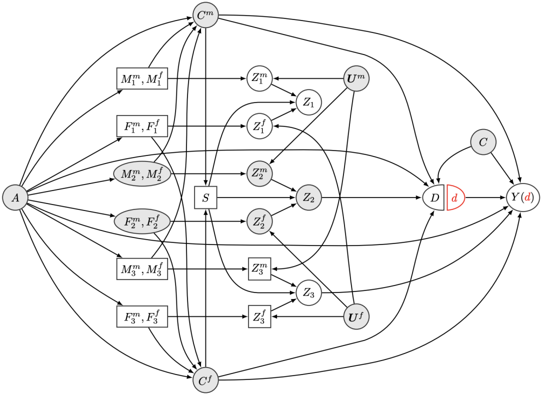
- The IV "revolution" in genetics, led by George Davey Smith from University of Bristol



the random mating process roughly renders our DNA as a random variable, not influenced by other factors (not exactly though); maybe, based on biological knowledge, the particular mutation does not biologically affect our heart ideal:

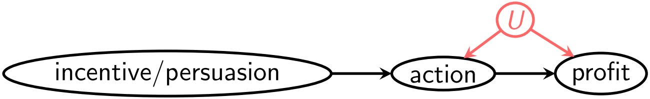


reality (REF: Almost exact Mendelian randomization)



Another motivating example: incentive or persuasion mechanism in behavior economics

- A central agent (e.g. Uber) may want to "manipulate" other agents (e.g. drivers and passengers) to increase utility
- However, the central agent cannot directly dictate what other agents do – the only thing the central agent can do is to provide incentive (e.g. money prize) or persuasion (e.g. revealing certain information of the states of the world)
- The incentive/persuasion itself may have no direct effect on the final utility
- The causal DAG



Another motivating example: effects of price on quantity

- IV was actually invented by Philip Wright (Sewell Wright's father) in 1928

- Wright wanted to study the effect of price on demand: e.g. to cut the smoking population by half, what the price of cigarette should have been?

Another motivating example: effects of price on quantity

- IV was actually invented by Philip Wright (Sewell Wright's father) in 1928

- Wright wanted to study the effect of price on demand: e.g. to cut the smoking population by half, what the price of cigarette should have been?

- The supply-demand model (from theoretical economics, possibly quite ideal):

  $$\log Q = \beta_0 + \beta_1 \log P + U$$

  U is not independent of $\log P$ (from economic theory, both determined by supply and demand curve), creating the problem of "endogeneity"

- Wright concluded to learn β_1, one needs to find some extra information to solve this "endogeneity" problem

Some other real examples

- military lottery, actually military service/war experience, psychological health (famous Vietnam war study)
- randomly giving gifts, taking covid vaccine, risk of dying from covid
- randomly giving money to students doing less well in school, actually attending school with more enthusiasm, academic achievement (famous field experiments conducted by super-star economist Roland Fryer)
- etc.

Another motivating example: RCT with non-ignorable non-compliance

- Z: randomized treatment assignment

Another motivating example: RCT with non-ignorable non-compliance

- Z: randomized treatment assignment
- A: actual treatment received

Another motivating example: RCT with non-ignorable non-compliance

- Z: randomized treatment assignment
- A: actual treatment received
- Y: outcome

Another motivating example: RCT with non-ignorable non-compliance

- Z: randomized treatment assignment
- A: actual treatment received
- Y: outcome

We are interested in estimating $\tau = \mathbb{E}[Y(a=1) - Y(a=0)]$ but experimental units might not compile with the doctor's assignment, and the non-compliance pattern might not be explained by observed data

Another motivating example: RCT with non-ignorable non-compliance

- Z: randomized treatment assignment

- A: actual treatment received

- Y: outcome

We are interested in estimating $\tau = \mathbb{E}[Y(a=1) - Y(a=0)]$ but experimental units might not compile with the doctor's assignment, and the non-compliance pattern might not be explained by observed data

This story tells us:
(1) Z causes A
(2) No unmeasured confounding between Z and $\{A, Y\}$
(3) Z causes Y only through A

Another motivating example: RCT with non-ignorable non-compliance

- Z: randomized treatment assignment
- A: actual treatment received
- Y: outcome

We are interested in estimating $\tau = \mathbb{E}[Y(a = 1) - Y(a = 0)]$ but experimental units might not compile with the doctor's assignment, and the non-compliance pattern might not be explained by observed data

This story tells us:
(1) Z causes A
(2) No unmeasured confounding between Z and $\{A, Y\}$
(3) Z causes Y only through A
Z satisfying the above three assumptions is called an "Instrumental Variable" (IV); IV is simply an IMPERFECT INTERVENTION!

One analysis strategy: intention-to-treat (ITT) analysis

- ITT analysis simply computes

$$\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]$$

One analysis strategy: intention-to-treat (ITT) analysis

- ITT analysis simply computes

$$\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]$$

- But do you think ITT analysis really answer our scientific question of interest?

One analysis strategy: intention-to-treat (ITT) analysis

- ITT analysis simply computes

$$\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]$$

- But do you think ITT analysis really answer our scientific question of interest?

- After this course, DO NOT CONFUSE ITT ANALYSIS AS IF IT IS CAUSAL!

The Instrumental Variable DAG/SWIG

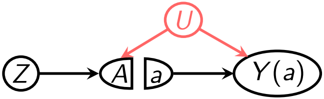
- From SWIG, one reads $Y(a) \perp\!\!\!\perp Z$ so $\mathbb{E}[Y(a)] = \mathbb{E}[Y(a)|Z]$ for all a

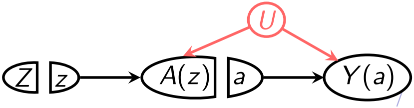
Core assumptions of IV

For simplicity we silence conditioning on the baseline confounders X

Z is an IV if

- relevance: $Z \not\perp\!\!\!\perp A$





Core assumptions of IV

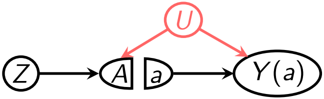
For simplicity we silence conditioning on the baseline confounders X

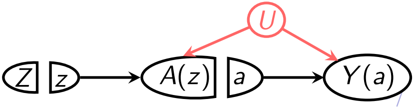
Z is an IV if

- relevance: $Z \not\perp\!\!\!\perp A$
- exogeneity (no unmeasured confounders between Z and A and between Z and Y):

$$Z \perp\!\!\!\perp (A(z), Y(z,a)) \; \forall a, z$$

[or can be relaxed to $Z \perp\!\!\!\perp Y(z,a)$]





Core assumptions of IV

For simplicity we silence conditioning on the baseline confounders X

Z is an IV if

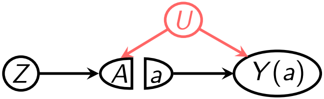
- relevance: $Z \not\perp\!\!\!\perp A$
- exogeneity (no unmeasured confounders between Z and A and between Z and Y):

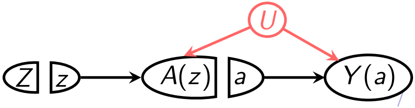
$$Z \perp\!\!\!\perp (A(z), Y(z,a)) \; \forall a, z$$

  [or can be relaxed to $Z \perp\!\!\!\perp Y(z,a)$]

- exclusion restriction (no direct effect from Z to Y):

$$Y(z,a) \equiv Y(a) \; \forall a, z$$





IV point identification: Linear SEM illustration

Let's consider the following linear SEM related to the IV DAG/SWIG: assuming U has $\mathbb{E}[U] = 0$

$$
\begin{aligned}
Y &= \tau A + \eta U + \varepsilon_Y \\
A &= \pi Z + \beta U + \varepsilon_A \\
Z &= \varepsilon_Z, \\
\varepsilon_Y &\perp\!\!\!\perp \varepsilon_A \perp\!\!\!\perp \varepsilon_Z \perp\!\!\!\perp U
\end{aligned}
$$

Then

$$
\begin{aligned}
&\mathbb{E}[A|Z] = \pi Z + \beta \mathbb{E}[U|Z] = \pi Z + \beta \mathbb{E}[U] = \pi Z \\
&\mathbb{E}[Y|Z] = \tau \mathbb{E}[A|Z] + \eta \mathbb{E}[U|Z] = \tau \pi Z + \eta \mathbb{E}[U] = \tau \pi Z = \gamma Z \\
&\text{(called "reduced-form" regression in econometrics)}
\end{aligned}
$$

so

$$\tau = \frac{\gamma}{\pi}, \text{ assuming } \pi \neq 0$$

This is the so-called 2SLS estimator of ATE under linear IV setting

IV point identification: nonparametric result

Based on the three core IV assumptions, in particular $Y(a) \perp\!\!\!\perp Z$, can we identify $\mathbb{E}[Y(a)]$ or the ACE $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$?

IV point identification: nonparametric result

Based on the three core IV assumptions, in particular $Y(a) \perp\!\!\!\perp Z$, can we identify $\mathbb{E}[Y(a)]$ or the ACE $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$?

Nobel prize paper: Imbens & Angrist 1994

IV point identification: nonparametric result

Based on the three core IV assumptions, in particular $Y(a) \perp\!\!\!\perp Z$, can we identify $\mathbb{E}[Y(a)]$ or the ACE $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$?

Nobel prize paper: Imbens & Angrist 1994

Consider binary instrument $Z \in \{0, 1\}$

IV point identification: nonparametric result

Based on the three core IV assumptions, in particular $Y(a) \perp\!\!\!\perp Z$, can we identify $\mathbb{E}[Y(a)]$ or the ACE $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$?

Nobel prize paper: Imbens & Angrist 1994

Consider binary instrument $Z \in \{0, 1\}$

Compliance table

subgroups	$A(z = 1)$	$A(z = 0)$
always taker	1	1
never taker	0	0
complier	1	0
defier	0	1

IV point identification: nonparametric result

Compliance table

subgroups	$A(z=1)$	$A(z=0)$
always taker	1	1
never taker	0	0
complier	1	0
defier	0	1

IV point identification: nonparametric result

Compliance table

subgroups	$A(z=1)$	$A(z=0)$
always taker	1	1
never taker	0	0
complier	1	0
defier	0	1

Without non-compliance, by $Y(a) \perp\!\!\!\perp Z$,

$$
\begin{aligned}
\mathbb{E}[Y(a)] &= \mathbb{E}[Y(a)|Z=a] = \mathbb{E}[Y(a)|Z=a, A(z)=a] \\
&= \mathbb{E}[Y|Z=a, A=a]
\end{aligned}
$$

IV point identification: nonparametric result

Compliance table

subgroups	$A(z=1)$	$A(z=0)$
always taker	1	1
never taker	0	0
complier	1	0
defier	0	1

Without non-compliance, by $Y(a) \perp\!\!\!\perp Z$,

$$\begin{aligned}\mathbb{E}[Y(a)] = \mathbb{E}[Y(a)|Z=a] &= \mathbb{E}[Y(a)|Z=a, A(z)=a] \\ &= \mathbb{E}[Y|Z=a, A=a]\end{aligned}$$

With non-compliance, unidentified in general

$$\begin{aligned}\mathbb{E}[Y(a)] &= \mathbb{E}[Y(a)|Z=z] \\ \Rightarrow \mathbb{E}[Y(a)] &= \mathbb{E}[Y|Z=z, A=a] \\ &+ \underbrace{P(A(z)=1-a)}_{P(A=1-a|Z=z)}\{\underbrace{\mathbb{E}[Y(a)|Z=z, A=1-a]}_{\text{unidentified}} - \mathbb{E}[Y|Z=z, A=a]\}\end{aligned}$$

IV point identification: nonparametric result

Derivation:

$$
\begin{aligned}
& \mathbb{E}[Y(a)] = \mathbb{E}[Y(a)|Z=z] \\
= & \mathbb{E}[Y(a)|Z=z, A(z)=a]P(A(z)=a|Z=z) \\
& + \mathbb{E}[Y(a)|Z=z, A(z)=1-a]P(A(z)=1-a|Z=z) \\
= & \mathbb{E}[Y|Z=z, A=a]P(A=a|Z=z) \\
& + \mathbb{E}[Y(a)|Z=z, A(z)=1-a]P(A=1-a|Z=z) \\
= & \mathbb{E}[Y|Z=z, A=a] - \mathbb{E}[Y|Z=z, A=a]P(A=1-a|Z=z) \\
& + \mathbb{E}[Y(a)|Z=z, A(z)=1-a]P(A=1-a|Z=z) \\
= & \mathbb{E}[Y|Z=z, A=a] \\
& + P(A=1-a|Z=z)\{\mathbb{E}[Y(a)|Z=z, A=1-a] - \mathbb{E}[Y|Z=z, A=a]\}
\end{aligned}
$$

by far, we have used every IV conditions but we still have a non-identifiable counterfactual quantity $\mathbb{E}[Y(a)|Z=z, A(z)=1-a]$

A more essential way of understanding non-identifiability

The following strategy is always helpful: counting free parameters by taking everything to be $\{0, 1\}$-valued

Since $A, Z \in \{0, 1\}^2$, we have only four possible values that can be calculated from the observed data $\mathbb{E}[Y|Z = 0, A = 0]$, $\mathbb{E}[Y|Z = 0, A = 1]$, $\mathbb{E}[Y|Z = 1, A = 0]$, $\mathbb{E}[Y|Z = 1, A = 1]$

IV point identification: 1st attempt

Point identification of τ needs extra modeling assumptions beyond (1) - (3)

IV point identification: 1st attempt

Point identification of τ needs extra modeling assumptions beyond (1) - (3)

One possibility is constant/homogeneous treatment effect assumption:

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

IV point identification: 1st attempt

Point identification of τ needs extra modeling assumptions beyond (1) - (3)

One possibility is constant/homogeneous treatment effect assumption:

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

Then

$$\begin{aligned}
& \mathbb{E}[Y(a)] = \mathbb{E}[Y(a)|Z = z] \\
\Rightarrow\ & \mathbb{E}[Y(a)] = \mathbb{E}[Y|Z = z, A = a] \\
& \quad + \underbrace{P(A(z) = 1 - a)}_{P(A=1-a|Z=z)}\{\mathbb{E}[Y(a)|Z = z, A = 1 - a] - \mathbb{E}[Y|Z = z, A = a]\} \\
\Rightarrow\ & \mathbb{E}[Y(1)] = \mathbb{E}[Y|Z = z, A = 1] \\
& \quad + P(A = 0|Z = z)\{\mathbb{E}[Y|Z = z, A = 0] + \tau - \mathbb{E}[Y|Z = z, A = 1]\} \\
& \mathbb{E}[Y(0)] = \mathbb{E}[Y|Z = z, A = 0] \\
& \quad + P(A = 1|Z = z)\{\mathbb{E}[Y|Z = z, A = 1] - \tau - \mathbb{E}[Y|Z = z, A = 0]\}
\end{aligned}$$

IV point identification: 1st attempt

$\forall\ z \in \{0, 1\}$:

$$\begin{aligned}\mathbb{E}[Y(1)] = \mathbb{E}[Y|Z = z, A = 1] \\ + P(A = 0|Z = z)\{\mathbb{E}[Y|Z = z, A = 0] + \tau - \mathbb{E}[Y|Z = z, A = 1]\} \\ \mathbb{E}[Y(0)] = \mathbb{E}[Y|Z = z, A = 0] \\ + P(A = 1|Z = z)\{\mathbb{E}[Y|Z = z, A = 1] - \tau - \mathbb{E}[Y|Z = z, A = 0]\}\end{aligned}$$

IV point identification: 1st attempt

$\forall\ z \in \{0, 1\}$:

$$\mathbb{E}[Y(1)] = \mathbb{E}[Y|Z = z, A = 1] + P(A = 0|Z = z)\{\mathbb{E}[Y|Z = z, A = 0] + \tau - \mathbb{E}[Y|Z = z, A = 1]\}$$

$$\mathbb{E}[Y(0)] = \mathbb{E}[Y|Z = z, A = 0] + P(A = 1|Z = z)\{\mathbb{E}[Y|Z = z, A = 1] - \tau - \mathbb{E}[Y|Z = z, A = 0]\}$$

$$\begin{aligned}\tau &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \underbrace{P(A = 0|Z = z)}_{1-\mathbb{E}[A|Z=z]}\tau + \underbrace{P(A = 1|Z = 1 - z)}_{\mathbb{E}[A|Z=1-z]}\tau \\ &\quad + \underbrace{P(A = 0|Z = z)\mathbb{E}[Y|Z = z, A = 0] + P(A = 1|Z = z)\mathbb{E}[Y|Z = z, A = 1]}_{\mathbb{E}[Y|Z=z]} \\ &\quad - \mathbb{E}[Y|Z = 1 - z]\end{aligned}$$

So: ATE τ can be computed as 2SLS (two-stage least square)

IV point identification: 1st attempt

$\forall\ z \in \{0,1\}$:

$$\mathbb{E}[Y(1)] = \mathbb{E}[Y|Z=z, A=1] + P(A=0|Z=z)\{\mathbb{E}[Y|Z=z, A=0] + \tau - \mathbb{E}[Y|Z=z, A=1]\}$$

$$\mathbb{E}[Y(0)] = \mathbb{E}[Y|Z=z, A=0] + P(A=1|Z=z)\{\mathbb{E}[Y|Z=z, A=1] - \tau - \mathbb{E}[Y|Z=z, A=0]\}$$

$$\begin{aligned}\tau &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \underbrace{P(A=0|Z=z)}_{1-\mathbb{E}[A|Z=z]}\tau + \underbrace{P(A=1|Z=1-z)}_{\mathbb{E}[A|Z=1-z]}\tau \\ &\quad + \underbrace{P(A=0|Z=z)\mathbb{E}[Y|Z=z, A=0] + P(A=1|Z=z)\mathbb{E}[Y|Z=z, A=1]}_{\mathbb{E}[Y|Z=z]} \\ &\quad - \mathbb{E}[Y|Z=1-z]\end{aligned}$$

So: ATE τ can be computed as 2SLS (two-stage least square)

$$\tau = \frac{\mathbb{E}[Y|Z=z] - \mathbb{E}[Y|Z=1-z]}{\mathbb{E}[A|Z=z] - \mathbb{E}[A|Z=1-z]} = \frac{\text{second stage LS coefficient}}{\text{first stage LS coefficient}}$$

IV point identification: 1st attempt alternative derivation

(1) Causal structural assumption gives us $Y(a) \perp\!\!\!\perp Z$, implying

$$\mathbb{E}[(Y(0) - \mathbb{E}[Y(0)])h(Z)] = 0 \ \forall h$$

IV point identification: 1st attempt alternative derivation

(1) Causal structural assumption gives us $Y(a) \perp\!\!\!\perp Z$, implying

$$\mathbb{E}[(Y(0) - \mathbb{E}[Y(0)])h(Z)] = 0 \; \forall h$$

(2) Modeling assumption gives us

$$Y(0) = Y - \tau A$$
$$\Rightarrow \mathbb{E}[Y(0)] = \mathbb{E}[Y] - \tau\mathbb{E}[A]$$

IV point identification: 1st attempt alternative derivation

(1) Causal structural assumption gives us $Y(a) \perp\!\!\!\perp Z$, implying

$$\mathbb{E}[(Y(0) - \mathbb{E}[Y(0)])h(Z)] = 0 \ \forall h$$

(2) Modeling assumption gives us

$$Y(0) = Y - \tau A$$
$$\Rightarrow \mathbb{E}[Y(0)] = \mathbb{E}[Y] - \tau \mathbb{E}[A]$$

(3) Combining (1) + (2):

$$\mathbb{E}[(Y - \tau A - \mathbb{E}[Y] + \tau \mathbb{E}[A])h(Z)] = 0$$
$$\Rightarrow \tau = \frac{\mathbb{E}[Yh(Z)] - \mathbb{E}[Y]\mathbb{E}[h(Z)]}{\mathbb{E}[Ah(Z)] - \mathbb{E}[A]\mathbb{E}[h(Z)]} = \frac{\mathrm{Cov}(Y, h(Z))}{\mathrm{Cov}(A, h(Z))}$$

Choose $h(Z) = \mathbb{E}[A|Z]$ (first stage regression), we have

$$\tau = \frac{\mathrm{Cov}(Y, \mathbb{E}[A|Z])}{\mathrm{Cov}(A, \mathbb{E}[A|Z])} \equiv \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[A|Z=1] - \mathbb{E}[A|Z=0]}$$

(two-stage least square (2SLS))

IV point identification: 2nd attempt, monotonicity assumption

Compliance table

subgroups	$A(z=1)$	$A(z=0)$
always taker	1	1
never taker	0	0
complier	1	0
defier	0	1

IV point identification: 2nd attempt, monotonicity assumption

Compliance table

subgroups	$A(z=1)$	$A(z=0)$
always taker	1	1
never taker	0	0
complier	1	0
defier	0	1

$$
\begin{aligned}
\mathbb{E}[Y(a)] = \; & \mathbb{E}[Y(a)|A(1)=1, A(0)=1]P(A(1)=1, A(0)=1) \\
& + \mathbb{E}[Y(a)|A(1)=0, A(0)=0]P(A(1)=0, A(0)=0) \\
& + \mathbb{E}[Y(a)|A(1)=1, A(0)=0]P(A(1)=1, A(0)=0) \\
& + \mathbb{E}[Y(a)|A(1)=0, A(0)=1]P(A(1)=0, A(0)=1)
\end{aligned}
$$

IV point identification: 2nd attempt, monotonicity assumption

Compliance table

subgroups	$A(z=1)$	$A(z=0)$
always taker	1	1
never taker	0	0
complier	1	0
defier	0	1

IV point identification: 2nd attempt, monotonicity assumption

Compliance table

subgroups	$A(z=1)$	$A(z=0)$
always taker	1	1
never taker	0	0
complier	1	0
defier	0	1

Instead, assume monotonicity: $A(1) \geq A(0)$ with probability 1

IV point identification: 2nd attempt, monotonicity assumption

Compliance table

subgroups	$A(z=1)$	$A(z=0)$
always taker	1	1
never taker	0	0
complier	1	0
~~defier~~	~~0~~	~~1~~

Instead, assume monotonicity: $A(1) \geq A(0)$ with probability 1

Ruling out defiers: what can we identify?

$$
\begin{aligned}
\mathbb{E}[Y(a)] = \; & \mathbb{E}[Y(a)|A(1)=1, A(0)=1]P(A(1)=1, A(0)=1) \\
& + \mathbb{E}[Y(a)|A(1)=0, A(0)=0]P(A(1)=0, A(0)=0) \\
& + \mathbb{E}[Y(a)|A(1)=1, A(0)=0]P(A(1)=1, A(0)=0) \\
& \cancel{+ \mathbb{E}[Y(a)|A(1)=0, A(0)=1]P(A(1)=0, A(0)=1)} \\
= \; & \mathbb{E}[Y(a)|A(1)>A(0)]P(A(1)>A(0)) \\
& + \mathbb{E}[Y(a)|A(1)=A(0)]P(A(1)=A(0))
\end{aligned}
$$

IV point identification: 2nd attempt, monotonicity assumption

Ruling out defiers: what can we identify?

$$\begin{aligned}\mathbb{E}[Y(a)] &= \mathbb{E}[Y(a)|A(1) > A(0)]P(A(1) > A(0)) \\ &\quad + \mathbb{E}[Y(a)|A(1) = A(0)]P(A(1) = A(0))\end{aligned}$$

IV point identification: 2nd attempt, monotonicity assumption

Ruling out defiers: what can we identify?

$$\mathbb{E}[Y(a)] = \mathbb{E}[Y(a)|A(1) > A(0)]P(A(1) > A(0)) + \mathbb{E}[Y(a)|A(1) = A(0)]P(A(1) = A(0))$$

Further: By $A(z) \perp\!\!\!\perp Z$ and $Y(a) \perp\!\!\!\perp Z$:

$$\begin{aligned}
&\mathbb{E}[Y(1)|A(1) > A(0)]P(A(1) > A(0)) \\
&= \mathbb{E}[(A(1) - A(0))Y(1)] \\
&= \mathbb{E}[A(1)Y(1)|Z = 1] - \mathbb{E}[A(0)Y(1)|Z = 0] \\
&= \mathbb{E}[AY|Z = 1] - \mathbb{E}[AY|Z = 0]
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[Y(0)|A(1) > A(0)]P(A(1) > A(0)) \\
&= \mathbb{E}[\{(1 - A(0)) - (1 - A(1))\}Y(0)] \\
&= \mathbb{E}[(1 - A)Y|Z = 0] - \mathbb{E}[(1 - A)Y|Z = 1]
\end{aligned}$$

IV point identification: 2nd attempt, monotonicity assumption

Ruling out defiers: what can we identify?

$$\mathbb{E}[Y(a)] = \mathbb{E}[Y(a)|A(1) > A(0)]P(A(1) > A(0)) + \mathbb{E}[Y(a)|A(1) = A(0)]P(A(1) = A(0))$$

Further: By $A(z) \perp\!\!\!\perp Z$ and $Y(a) \perp\!\!\!\perp Z$:

$$\begin{aligned}
&\mathbb{E}[Y(1)|A(1) > A(0)]P(A(1) > A(0)) \\
&= \mathbb{E}[(A(1) - A(0))Y(1)] \\
&= \mathbb{E}[A(1)Y(1)|Z = 1] - \mathbb{E}[A(0)Y(1)|Z = 0] \\
&= \mathbb{E}[AY|Z = 1] - \mathbb{E}[AY|Z = 0]
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[Y(0)|A(1) > A(0)]P(A(1) > A(0)) \\
&= \mathbb{E}[\{(1 - A(0)) - (1 - A(1))\}Y(0)] \\
&= \mathbb{E}[(1 - A)Y|Z = 0] - \mathbb{E}[(1 - A)Y|Z = 1]
\end{aligned}$$

IV point identification: 2nd attempt, monotonicity assumption

$$\mathbb{E}[Y(1)|A(1) > A(0)]P(A(1) > A(0)) = \mathbb{E}[AY|Z = 1] - \mathbb{E}[AY|Z = 0]$$

and

$$\mathbb{E}[Y(0)|A(1) > A(0)]P(A(1) > A(0)) = \mathbb{E}[(1 - A)Y|Z = 0] - \mathbb{E}[(1 - A)Y|Z = 1]$$

IV point identification: 2nd attempt, monotonicity assumption

$$\mathbb{E}[Y(1)|A(1) > A(0)]P(A(1) > A(0)) = \mathbb{E}[AY|Z = 1] - \mathbb{E}[AY|Z = 0]$$

and

$$\mathbb{E}[Y(0)|A(1) > A(0)]P(A(1) > A(0)) = \mathbb{E}[(1 - A)Y|Z = 0] - \mathbb{E}[(1 - A)Y|Z = 1]$$

Then take the difference:

$$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)] = \frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{P(A(1) > A(0))}$$

IV point identification: 2nd attempt, monotonicity assumption

$$\mathbb{E}[Y(1)|A(1) > A(0)]P(A(1) > A(0)) = \mathbb{E}[AY|Z = 1] - \mathbb{E}[AY|Z = 0]$$

and

$$\mathbb{E}[Y(0)|A(1) > A(0)]P(A(1) > A(0)) = \mathbb{E}[(1 - A)Y|Z = 0] - \mathbb{E}[(1 - A)Y|Z = 1]$$

Then take the difference:

$$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)] = \frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{P(A(1) > A(0))}$$

Finally,

$$\begin{aligned} P(A(1) > A(0)) &= P(A(1) = 1, A(0) = 0) \\ &= P(A(1) = 1) - P(A(1) = 1, A(0) = 1) \\ &= P(A(1) = 1) - P(A(0) = 1) \underbrace{P(A(1) = 1|A(0) = 1)}_{\equiv 1} \\ &= P(A(1) = 1) - P(A(0) = 1) \\ &= \mathbb{E}[A|Z = 1] - \mathbb{E}[A|Z = 0] \end{aligned}$$

IV point identification: 2nd attempt, monotonicity assumption

So:

$$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)] = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[A|Z=1] - \mathbb{E}[A|Z=0]}$$

IV point identification: 2nd attempt, monotonicity assumption

So:

$$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)] = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[A|Z=1] - \mathbb{E}[A|Z=0]}$$

$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)]$ is identified by 2SLS!

IV point identification: 2nd attempt, monotonicity assumption

So:

$$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)] = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[A|Z=1] - \mathbb{E}[A|Z=0]}$$

$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)]$ is identified by 2SLS!

But what is $\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)]$? The CATE conditioning on $A(1) > A(0)$

IV point identification: 2nd attempt, monotonicity assumption

So:

$$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)] = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[A|Z=1] - \mathbb{E}[A|Z=0]}$$

$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)]$ is identified by 2SLS!

But what is $\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)]$? The CATE conditioning on $A(1) > A(0)$

$A(1) > A(0)$: compliers! So 2SLS identifies CATE among compliers, termed by Imbens and Angrist as "Local Average Treatment Effect" (LATE)

IV point identification: 2nd attempt, monotonicity assumption

So:

$$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)] = \frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{\mathbb{E}[A|Z = 1] - \mathbb{E}[A|Z = 0]}$$

$\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)]$ is identified by 2SLS!

But what is $\mathbb{E}[Y(1) - Y(0)|A(1) > A(0)]$? The CATE conditioning on $A(1) > A(0)$

$A(1) > A(0)$: compliers! So 2SLS identifies CATE among compliers, termed by Imbens and Angrist as "Local Average Treatment Effect" (LATE)

This important conceptual leap together with extremely impactful applications in labor economics wins the Nobel prize

IV point identification: Can we identify ATE?

- Yes but under alternative untestable identification assumptions: Several examples (the following conditions can be relaxed to ones conditioning on all the observed covariates)

IV point identification: Can we identify ATE?

- Yes but under alternative untestable identification assumptions: Several examples (the following conditions can be relaxed to ones conditioning on all the observed covariates)

- Robins 1994: "no instrument-treatment interaction"

  (1) Three core IV assumptions

  (2) no current treatment value interaction
  $\mathbb{E}[Y(1) - Y(0)|Z = z, A = a] = \gamma^* \cdot a$

IV point identification: Can we identify ATE?

- Yes but under alternative untestable identification assumptions: Several examples (the following conditions can be relaxed to ones conditioning on all the observed covariates)
- Robins 1994: "no instrument-treatment interaction"

  (1) Three core IV assumptions

  (2) no current treatment value interaction
  $\mathbb{E}[Y(1) - Y(0)|Z = z, A = a] = \gamma^* \cdot a$
- Wang and Tchetgen Tchetgen 2018: "no unmeasured confounder-treatment interactions"

$$\mathbb{E}[Y(1) - Y(0)|U] = \mathbb{E}[Y(1) - Y(0)]$$

- In both cases, 2SLS helps identify ATE

IV point identification: Can we identify ATE?

derivation under "no instrument-treatment interaction": define mimicking counterfactual

$$\widetilde{Y}(\gamma) := Y - \gamma \cdot A$$

by SNMM, we have $\mathbb{E}[\widetilde{Y}(\gamma^*)|Z, A] = \mathbb{E}[Y(0)|Z, A]$

by exclusion restriction, we have

$$\mathbb{E}[\widetilde{Y}(\gamma^*)|Z] = \mathbb{E}[Y(0)|Z] = \mathbb{E}[Y(0)] = \mathbb{E}[\widetilde{Y}(\gamma^*)]$$

IV point identification: Can we identify ATE?

derivation under "no instrument-treatment interaction": define mimicking counterfactual

$$\widetilde{Y}(\gamma) := Y - \gamma \cdot A$$

by SNMM, we have $\mathbb{E}[\widetilde{Y}(\gamma^*)|Z, A] = \mathbb{E}[Y(0)|Z, A]$

by exclusion restriction, we have

$$\begin{aligned}
&\mathbb{E}[\widetilde{Y}(\gamma^*)|Z] = \mathbb{E}[\widetilde{Y}(\gamma^*)] \\
\Rightarrow\ &\mathbb{E}\left[(\widetilde{Y}(\gamma^*) - \mathbb{E}[\widetilde{Y}(\gamma^*)])h(Z)\right] = 0, \forall\, h \\
\Rightarrow\ &\mathbb{E}\left[(Y - \gamma^* \cdot A - \mathbb{E}[Y] + \gamma^* \mathbb{E}[A])Z\right] = 0 \\
\Rightarrow\ &\gamma^* = \frac{\mathbb{E}[(Y - \mathbb{E}[Y])Z]}{\mathbb{E}[(A - \mathbb{E}[A])Z]}
\end{aligned}$$

The issue of weak IV

- 2SLS is also called "Wald estimand" because it is a ratio

The issue of weak IV

- 2SLS is also called "Wald estimand" because it is a ratio
- Since it is a ratio, when the denominator is small, 2SLS will be gradually becoming more and more ill-defined/ill-posed
- Can we detect weak IV? Yes, this is just a (conditional) independence test between A and Z (possibly given observed covariates X)

The issue of weak IV

- 2SLS is also called “Wald estimand” because it is a ratio
- Since it is a ratio, when the denominator is small, 2SLS will be gradually becoming more and more ill-defined/ill-posed
- Can we detect weak IV? Yes, this is just a (conditional) independence test between A and Z (possibly given observed covariates X)
- Economists strongly recommend to report the first-stage F-statistic whenever using 2SLS (simply output by every regression model in R)

The issue of weak IV

- 2SLS is also called "Wald estimand" because it is a ratio
- Since it is a ratio, when the denominator is small, 2SLS will be gradually becoming more and more ill-defined/ill-posed
- Can we detect weak IV? Yes, this is just a (conditional) independence test between A and Z (possibly given observed covariates X)
- Economists strongly recommend to report the first-stage F-statistic whenever using 2SLS (simply output by every regression model in R)
- Convention: "if F-statistic is bigger than 10, one can safely use 2SLS"

Many IVs

- We have seen the following: One IV can be used to identify the causal effect of one endogenous exposure

Many IVs

- We have seen the following: One IV can be used to identify the causal effect of one endogenous exposure

- What if we have multiple, say K endogenous exposures?

Many IVs

- We have seen the following: One IV can be used to identify the causal effect of one endogenous exposure
- What if we have multiple, say K endogenous exposures?
- In general, one needs to get at least one IV per endogenous exposure – in economics, this is called the "just-identified" case
- If you have less IVs than needed, it is called the "under-identified" case
- If you have more IVs than needed, it is called the "over-identified" case

What to do with many IVs?

- To illustrate the main idea, let's again consider the linear SEM:

$$
\begin{aligned}
Y &= \tau A + \eta U + \varepsilon_Y \\
A &= \pi^\top Z + \beta U + \varepsilon_A
\end{aligned}
$$

with Z now a k-dimensional vector

What to do with many IVs?

- To illustrate the main idea, let's again consider the linear SEM:

$$Y = \tau A + \eta U + \varepsilon_Y$$
$$A = \pi^\top Z + \beta U + \varepsilon_A$$

with Z now a k-dimensional vector

- Let's write down the n-sample version of the above linear SEM

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{A}_{n\times 1}\tau + \boldsymbol{U}\eta + \boldsymbol{\varepsilon}_Y = \boldsymbol{A}\tau + \boldsymbol{\xi}$$
$$\boldsymbol{A}_{n\times 1} = \boldsymbol{Z}_{n\times k}\pi + \boldsymbol{U}\beta + \boldsymbol{\varepsilon}_A = \boldsymbol{Z}\pi + \boldsymbol{\delta}$$

What to do with many IVs?

- To illustrate the main idea, let's again consider the linear SEM:

$$Y = \tau A + \eta U + \varepsilon_Y$$
$$A = \pi^\top Z + \beta U + \varepsilon_A$$

with Z now a k-dimensional vector

- Let's write down the n-sample version of the above linear SEM

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{A}_{n\times 1}\tau + \boldsymbol{U}\eta + \boldsymbol{\varepsilon}_Y = \boldsymbol{A}\tau + \boldsymbol{\xi}$$
$$\boldsymbol{A}_{n\times 1} = \boldsymbol{Z}_{n\times k}\pi + \boldsymbol{U}\beta + \boldsymbol{\varepsilon}_A = \boldsymbol{Z}\pi + \boldsymbol{\delta}$$

- Denote $P_Z = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$, the following estimator is referred to as the 2SLS with the presence of many IVs

$$\widehat{\tau}_{\text{2SLS}} = \frac{\boldsymbol{A}^\top P_Z \boldsymbol{Y}}{\boldsymbol{A}^\top P_Z \boldsymbol{A}} = \frac{\boldsymbol{A}^\top P_Z (\boldsymbol{A}\tau + \boldsymbol{\xi})}{\boldsymbol{A}^\top P_Z \boldsymbol{A}} = \tau + \underbrace{\frac{\boldsymbol{A}^\top P_Z \boldsymbol{\xi}}{\boldsymbol{A}^\top P_Z \boldsymbol{A}}}_{\text{mean zero}}$$

Alternative popular estimator: Limited Information Maximum Likelihood (LIML)

- Denote $P_Z = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$, the following estimator is referred to as the 2SLS with the presence of many IVs

$$\widehat{\tau}_{\text{2SLS}} = \frac{\boldsymbol{A}^\top P_Z \boldsymbol{Y}}{\boldsymbol{A}^\top P_Z \boldsymbol{A}} = \tau + \frac{\boldsymbol{A}^\top P_Z \boldsymbol{\xi}}{\boldsymbol{A}^\top P_Z \boldsymbol{A}}$$

- The biased OLS:

$$\widehat{\tau}_{\text{OLS}} = \frac{\boldsymbol{A}^\top \boldsymbol{Y}}{\boldsymbol{A}^\top \boldsymbol{A}} = \tau + \frac{\boldsymbol{A}^\top \boldsymbol{\xi}}{\boldsymbol{A}^\top \boldsymbol{A}}$$

Alternative popular estimator: Limited Information Maximum Likelihood (LIML)

- Denote $P_Z = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$, the following estimator is referred to as the 2SLS with the presence of many IVs

$$\widehat{\tau}_{\text{2SLS}} = \frac{\boldsymbol{A}^\top P_Z \boldsymbol{Y}}{\boldsymbol{A}^\top P_Z \boldsymbol{A}} = \tau + \frac{\boldsymbol{A}^\top P_Z \boldsymbol{\xi}}{\boldsymbol{A}^\top P_Z \boldsymbol{A}}$$

- The biased OLS:

$$\widehat{\tau}_{\text{OLS}} = \frac{\boldsymbol{A}^\top \boldsymbol{Y}}{\boldsymbol{A}^\top \boldsymbol{A}} = \tau + \frac{\boldsymbol{A}^\top \boldsymbol{\xi}}{\boldsymbol{A}^\top \boldsymbol{A}}$$

- The LIML: let $P_Z^\perp = \boldsymbol{I} - P_Z$

$$\begin{aligned}\widehat{\tau}_{\text{LIML}} &= \frac{\boldsymbol{A}^\top (\boldsymbol{I} - \lambda P_Z^\perp)\boldsymbol{Y}}{\boldsymbol{A}^\top (\boldsymbol{I} - \lambda P_Z^\perp)\boldsymbol{A}} = \frac{\boldsymbol{A}^\top \{(1-\lambda)\boldsymbol{I} + \lambda P_Z\}\boldsymbol{Y}}{\boldsymbol{A}^\top \{(1-\lambda)\boldsymbol{I} + \lambda P_Z\}\boldsymbol{A}} \\ &= \tau + \frac{\boldsymbol{A}^\top \{(1-\lambda)\boldsymbol{I} + \lambda P_Z\}\boldsymbol{\xi}}{\boldsymbol{A}^\top \{(1-\lambda)\boldsymbol{I} + \lambda P_Z\}\boldsymbol{A}}\end{aligned}$$

LIML's λ

- The LIML: let $P_Z^{\perp} = I - P_Z$

$$\widehat{\tau}_{\text{LIML}} = \frac{\boldsymbol{A}^\top (I - \lambda P_Z^{\perp})\boldsymbol{Y}}{\boldsymbol{A}^\top (I - \lambda P_Z^{\perp})\boldsymbol{A}} = \frac{\boldsymbol{A}^\top \{(1-\lambda)I + \lambda P_Z\}\boldsymbol{Y}}{\boldsymbol{A}^\top \{(1-\lambda)I + \lambda P_Z\}\boldsymbol{A}}$$
$$= \tau + \frac{\boldsymbol{A}^\top \{(1-\lambda)I + \lambda P_Z\}\boldsymbol{\xi}}{\boldsymbol{A}^\top \{(1-\lambda)I + \lambda P_Z\}\boldsymbol{A}}$$

LIML's λ

- The LIML: let $P_Z^{\perp} = I - P_Z$

$$\widehat{\tau}_{\text{LIML}} = \frac{\boldsymbol{A}^\top (I - \lambda P_Z^{\perp}) \boldsymbol{Y}}{\boldsymbol{A}^\top (I - \lambda P_Z^{\perp}) \boldsymbol{A}} = \frac{\boldsymbol{A}^\top \{(1-\lambda) I + \lambda P_Z\} \boldsymbol{Y}}{\boldsymbol{A}^\top \{(1-\lambda) I + \lambda P_Z\} \boldsymbol{A}}$$
$$= \tau + \frac{\boldsymbol{A}^\top \{(1-\lambda) I + \lambda P_Z\} \boldsymbol{\xi}}{\boldsymbol{A}^\top \{(1-\lambda) I + \lambda P_Z\} \boldsymbol{A}}$$

- It can be "roughly" viewed as a linear combination between OLS and 2SLS

- How to set λ? LIML particularly chooses the following strategy: λ is the smallest root of the following equation

$$\det \left[(\boldsymbol{A} \ \boldsymbol{Y})_{2\times n}^\top \left\{ I - \lambda P_Z^{\perp} \right\} (\boldsymbol{A} \ \boldsymbol{Y})_{n\times 2} \right] = 0$$

The issue of many weak IVs

- When IVs are weak, it does not help to have many of them...
- Because otherwise, one could have generated so many random noises to serve as IVs to completely solve the endogeneity problem

The issue of many weak IVs

- When IVs are weak, it does not help to have many of them...

- Because otherwise, one could have generated so many random noises to serve as IVs to completely solve the endogeneity problem

- What happens when many IVs are weak? for simplicity, let's say $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{A}$ so we also have $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{\xi}$; we also have $\mathbb{E}[P_Z] \approx I$

$$\begin{aligned}
\widehat{\tau}_{\mathrm{2SLS}} &= \tau + \frac{\boldsymbol{A}^\top P_Z \boldsymbol{\xi}}{\boldsymbol{A}^\top P_Z \boldsymbol{A}} \\
&\approx \tau + \frac{\mathbb{E}[\boldsymbol{A}^\top P_Z \boldsymbol{\xi}]}{\mathbb{E}[\boldsymbol{A}^\top P_Z \boldsymbol{A}]} \\
&= \tau + \frac{\mathbb{E}[\boldsymbol{A}^\top \mathbb{E}[P_Z] \boldsymbol{\xi}]}{\mathbb{E}[\boldsymbol{A}^\top \mathbb{E}[P_Z] \boldsymbol{A}]} = \tau + \frac{\mathbb{E}[\boldsymbol{A}^\top \boldsymbol{\xi}]}{\mathbb{E}[\boldsymbol{A}^\top \boldsymbol{A}]} \\
&\approx \tau + \frac{\boldsymbol{A}^\top \boldsymbol{\xi}}{\boldsymbol{A}^\top \boldsymbol{A}} = \widehat{\tau}_{\mathrm{OLS}}
\end{aligned}$$

- People tend to view LIML as a more robust version of 2SLS under many weak IVs

Joke about IVs

- A joke among economists: it takes an economist's life-time to find a good IV
- In practice, it is difficult to find IVs for a particular social science or economic problem

Joke about IVs

- A joke among economists: it takes an economist's life-time to find a good IV
- In practice, it is difficult to find IVs for a particular social science or economic problem
- But in clinical medicine and biology, IVs seem to be much easier to find, such as non-compliance in clinical trials
- And more recently, Mendelian randomization (MR) that makes biologists both happy and sad ...

Mendelian Randomization (MR): A natural IV

- Suppose we are interested in estimating the effect of lipid level in blood (A) on heart disease (Y)

Mendelian Randomization (MR): A natural IV

- Suppose we are interested in estimating the effect of lipid level in blood (A) on heart disease (Y)

- We have a bunch of SNPs Z measured by DNA sequencing technology

Mendelian Randomization (MR): A natural IV

- Suppose we are interested in estimating the effect of lipid level in blood (A) on heart disease (Y)
- We have a bunch of SNPs Z measured by DNA sequencing technology
- Are SNPs good IV candidate?

Mendelian Randomization (MR): A natural IV

- Suppose we are interested in estimating the effect of lipid level in blood (A) on heart disease (Y)
- We have a bunch of SNPs Z measured by DNA sequencing technology
- Are SNPs good IV candidate?
- Probably yes due to Mendelian randomization so exogeneity is satisfied but might still be violated when population stratification happens: invalid IV

Mendelian Randomization (MR): A natural IV

- Suppose we are interested in estimating the effect of lipid level in blood (A) on heart disease (Y)
- We have a bunch of SNPs Z measured by DNA sequencing technology
- Are SNPs good IV candidate?
- Probably yes due to Mendelian randomization so exogeneity is satisfied but might still be violated when population stratification happens: invalid IV
- Probably not because SNPs might have multiple functions (pleiotropy) so Z might directly cause Y: invalid IV

Mendelian Randomization (MR): A natural IV

- Suppose we are interested in estimating the effect of lipid level in blood (A) on heart disease (Y)
- We have a bunch of SNPs Z measured by DNA sequencing technology
- Are SNPs good IV candidate?
- Probably yes due to Mendelian randomization so exogeneity is satisfied but might still be violated when population stratification happens: invalid IV
- Probably not because SNPs might have multiple functions (pleiotropy) so Z might directly cause Y: invalid IV
- Probably not due to GWAS study we kind of know Z is associated with A: weak IV

General strategies for dealing with weak IVs in MR

- In general, assume linear treatment effect (not necessarily completely linear model)

- Weak IV ($A - Z$ weak dependence):
  - Filter out weak IVs by hypothesis testing using F-statistic: Andrews, Stock, Sun Annual Reviews of Econometrics 2019

General strategies for dealing with weak IVs in MR

- In general, assume linear treatment effect (not necessarily completely linear model)
- Weak IV ($A - Z$ weak dependence):
  - Filter out weak IVs by hypothesis testing using F-statistic: Andrews, Stock, Sun Annual Reviews of Econometrics 2019
  - Aggregation smartly: Ye, Shao, Kang AoS 2021

General strategies for dealing with weak IVs in MR

- In general, assume linear treatment effect (not necessarily completely linear model)
- Weak IV ($A - Z$ weak dependence):
  - Filter out weak IVs by hypothesis testing using F-statistic: Andrews, Stock, Sun Annual Reviews of Econometrics 2019
  - Aggregation smartly: Ye, Shao, Kang AoS 2021
  - Alternative modeling strategy by random effect model: Zhao, Chen, Wang, Small IJE 2019

General strategies for dealing with invalid IVs in MR

- In general, assume linear treatment effect (not necessarily completely linear model)
- Invalid IV (violation of exogeneity):

General strategies for dealing with invalid IVs in MR

- In general, assume linear treatment effect (not necessarily completely linear model)
- Invalid IV (violation of exogeneity):
  - Majority rule: Kang et al. JASA 2016, less than 50% of IVs are invalid

General strategies for dealing with invalid IVs in MR

- In general, assume linear treatment effect (not necessarily completely linear model)
- Invalid IV (violation of exogeneity):
  - Majority rule: Kang et al. JASA 2016, less than 50% of IVs are invalid
  - Plurality rule: Guo et al. JRSS-B 2018, number of valid IVs $>$ the largest number of invalid IVs giving the same effect estimates

General strategies for dealing with invalid IVs in MR

- In general, assume linear treatment effect (not necessarily completely linear model)
- Invalid IV (violation of exogeneity):
  - Majority rule: Kang et al. JASA 2016, less than 50% of IVs are invalid
  - Plurality rule: Guo et al. JRSS-B 2018, number of valid IVs $>$ the largest number of invalid IVs giving the same effect estimates
  - Orthogonality between $Z \to A$ effects and $Z \to Y$ direct effects (Instrument Strength Independent of Direct Effect or InSIDE): MR-Egger (essentially a meta-analysis)

General strategies for dealing with invalid IVs in MR

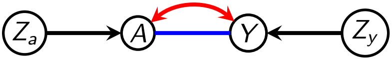
- In general, assume linear treatment effect (not necessarily completely linear model)
- Invalid IV (violation of exogeneity):
  - Majority rule: Kang et al. JASA 2016, less than 50% of IVs are invalid
  - Plurality rule: Guo et al. JRSS-B 2018, number of valid IVs $>$ the largest number of invalid IVs giving the same effect estimates
  - Orthogonality between $Z \to A$ effects and $Z \to Y$ direct effects (Instrument Strength Independent of Direct Effect or InSIDE): MR-Egger (essentially a meta-analysis)
  - GENIUS: Sun, Tchetgen Tchetgen, Walter Stat. Sci. 2020

Bidirectional IV/MR: Time ordering

- MR can also be used to orient undirected edges of ADMG with uncertainty

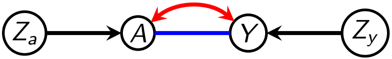
Bidirectional IV/MR: Time ordering

- MR can also be used to orient undirected edges of ADMG with uncertainty
- Simple case:



Bidirectional IV/MR: Time ordering

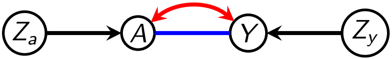
- MR can also be used to orient undirected edges of ADMG with uncertainty
- Simple case:



- Difficult to handle invalid IVs

Bidirectional IV/MR: Time ordering

- MR can also be used to orient undirected edges of ADMG with uncertainty
- Simple case:

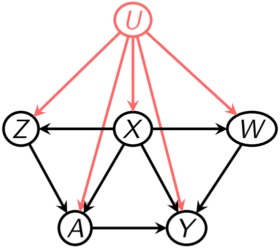


- Difficult to handle invalid IVs
- See Li and Ye, 2022 for some recent progress on testing if the effects are zero

Proximal causal inference or negative controls

Proximal causal learning (motivated from negative control in experimental biology)

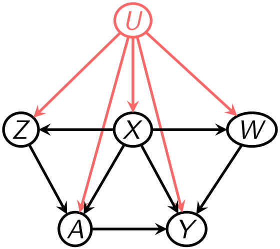
REF: Tchetgen Tchetgen, Ying, Cui, Shi, Miao. An Introduction to Proximal Causal Learning.



W: proxy of Y; Z: proxy of A

Proximal causal learning (motivated from negative control in experimental biology)

REF: Tchetgen Tchetgen, Ying, Cui, Shi, Miao. An Introduction to Proximal Causal Learning.



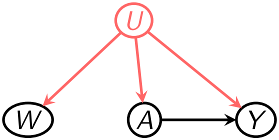
W: proxy of Y; Z: proxy of A

In the above DAG, $\tau = \mathbb{E}[Y(1) - Y(0)]$ is point identifiable without modeling assumptions, but under some extra conditions

Application of proximal causal learning

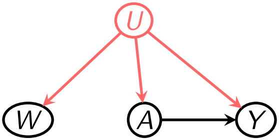
- Genomics: CRISPR-Cas9 gene-perturbation experiments – often we do not know exactly
- Environmental health:
- Proxies can also be viewed as the measurements of the true underlying biological mechanisms

Proximal causal learning comes from negative control



W: negative control outcome (NCO), not causally affected by A
Inspired from experimental biology: always compare to something that is known not to be affected by the chemical treatment
e.g. Y: death due to lung cancer, A: smoking, W: non-smoking related death (e.g. diabetes)

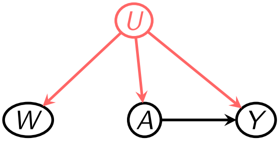
Proximal causal learning comes from negative control



W: negative control outcome (NCO), not causally affected by A
Inspired from experimental biology: always compare to something that is known not to be affected by the chemical treatment
e.g. Y: death due to lung cancer, A: smoking, W: non-smoking related death (e.g. diabetes)

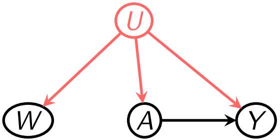
Intuition: any difference of W between $A = 1$ and $A = 0$ is due to U

Illustration via linear models



$$\mathbb{E}[Y|A,U] = \beta_{AY}A + \beta_{UY}U$$
$$\mathbb{E}[W|A,U] = \beta_{UW}U$$

Illustration via linear models

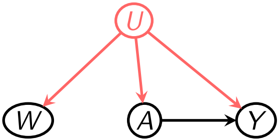


$$\mathbb{E}[Y|A, U] = \beta_{AY} A + \beta_{UY} U$$
$$\mathbb{E}[W|A, U] = \beta_{UW} U$$

The above equations imply the following linear models over observables:

$$\mathbb{E}[Y|A] = \beta_{AY} A + \beta_{UY} \mathbb{E}[U|A]$$
$$\mathbb{E}[W|A] = \beta_{UW} \mathbb{E}[U|A]$$
$$\Rightarrow \mathbb{E}[Y|A] = \beta_{AY} A + \frac{\beta_{UY}}{\beta_{UW}} \mathbb{E}[W|A]$$

Illustration via linear models



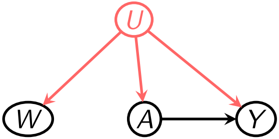
$$\mathbb{E}[Y|A,U] = \beta_{AY}A + \beta_{UY}U$$
$$\mathbb{E}[W|A,U] = \beta_{UW}U$$

The above equations imply the following linear models over observables:

$$\mathbb{E}[Y|A] = \beta_{AY}A + \beta_{UY}\mathbb{E}[U|A]$$
$$\mathbb{E}[W|A] = \beta_{UW}\mathbb{E}[U|A]$$
$$\Rightarrow \mathbb{E}[Y|A] = \beta_{AY}A + \frac{\beta_{UY}}{\beta_{UW}}\mathbb{E}[W|A]$$

When assuming $\frac{\beta_{UY}}{\beta_{UW}}$ is known, we can recover β_{AY}

Illustration via linear models

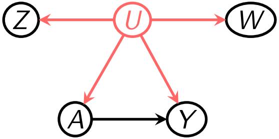


$$\mathbb{E}[Y|A,U] = \beta_{AY}A + \beta_{UY}U$$
$$\mathbb{E}[W|A,U] = \beta_{UW}U$$

So NCO is quite like IV: helpful but not enough for point identification

50% of proximal causal learning: double negative control

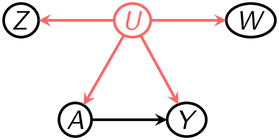
What if in addition we have a negative control treatment (NCT) Z?



Q: Is Z a valid IV?

50% of proximal causal learning: double negative control

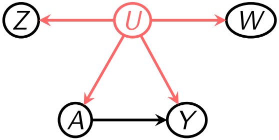
What if in addition we have a negative control treatment (NCT) Z?



Q: Is Z a valid IV?

Obviously not

50% of proximal causal learning: double negative control



$$\mathbb{E}[Y|A, Z, U] = \beta_{AY} A + \beta_{UY} U$$
$$\mathbb{E}[W|A, Z, U] = \beta_{UW} U$$
$$\mathbb{E}[U|A, Z] = \beta_{AU} A + \beta_{ZU} Z$$

Implication on observables:

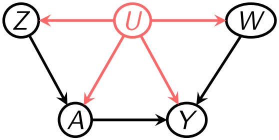
$$\mathbb{E}[Y|A, Z] = \beta_{AY} A + \beta_{UY} \mathbb{E}[U|A, Z]$$
$$\mathbb{E}[W|A, Z] = \beta_{UW} \mathbb{E}[U|A, Z]$$
$$\Rightarrow \mathbb{E}[Y|A, Z] = \beta_{AY} A + \frac{\beta_{UY}}{\beta_{UW}} \mathbb{E}[W|A, Z]$$

Non-rigorously argue yourself why we do not need to know the value of $\frac{\beta_{UY}}{\beta_{UW}}$ when $\mathbb{E}[W|A, Z]$ does depend on Z.

So it is quite important that $\mathbb{E}[U|A, Z]$ varies with Z

100% of proximal causal learning except for X

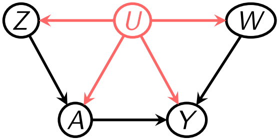
In fact, we can further relax the setting



$$\mathbb{E}[Y|A, Z, U] = \beta_{AY} A + \beta_{UY} U$$
$$\mathbb{E}[W|A, Z, U] = \beta_{UW} U$$

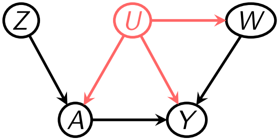
100% of proximal causal learning except for X

In fact, we can further relax the setting



$$\mathbb{E}[Y|A, Z, U] = \beta_{AY} A + \beta_{UY} U$$
$$\mathbb{E}[W|A, Z, U] = \beta_{UW} U$$

But what if



$$\mathbb{E}[Y|A, Z, U] = \beta_{AY} A + \beta_{UY} U$$
$$\mathbb{E}[W|A, Z, U] = \beta_{UW} U$$

Can you still argue $\mathbb{E}[U|A, Z]$ varies with Z?

Nonparametric identification: confounding bridge & completeness

1. Confounding bridge: there exists a function $h(a, w)$ such that

$$\mathbb{E}[Y|A, Z] = \mathbb{E}[h(A, W)|A, Z]$$

Nonparametric identification: confounding bridge & completeness

1. Confounding bridge: there exists a function $h(a, w)$ such that

$$\mathbb{E}[Y|A, Z] = \mathbb{E}[h(A, W)|A, Z]$$

2. Completeness:

$$\mathbb{E}[v(U)|Z, A] = 0 \text{ with probability } 1 \Rightarrow v(U) = 0 \text{ with probability } 1$$

Nonparametric identification: confounding bridge & completeness

1. Confounding bridge: there exists a function $h(a, w)$ such that

$$\mathbb{E}[Y|A, Z] = \mathbb{E}[h(A, W)|A, Z]$$

2. Completeness:

$$\mathbb{E}[v(U)|Z, A] = 0 \text{ with probability } 1 \Rightarrow v(U) = 0 \text{ with probability } 1$$

NOTE: Try to draw connections between these two assumptions and what we have done with linear model!

Proximal identification: Step 1

(1) "confounding bridge equation $\mathbb{E}[Y|Z, A] = \mathbb{E}[h(A, W)|Z, A]$" + "exclusion restriction: $Y \perp\!\!\!\perp Z|U, A$":

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[Y|U, A]|Z, A] = \mathbb{E}[\mathbb{E}[Y|U, Z, A]|Z, A] = \mathbb{E}[Y|Z, A] \\ \Rightarrow &\mathbb{E}[\mathbb{E}[Y|U, A]|Z, A] = \mathbb{E}[h(A, W)|Z, A] \end{aligned}$$

Proximal identification: Step 2

(1) "confounding bridge equation $\mathbb{E}[Y|Z,A] = \mathbb{E}[h(A,W)|Z,A]$" + "exclusion restriction: $Y \perp\!\!\!\perp Z|U,A$":

$$\mathbb{E}[\mathbb{E}[Y|U,A]|Z,A] = \mathbb{E}[h(A,W)|Z,A]$$

(2) "completeness: $\mathbb{E}[v(U)|Z,A] = 0 \Rightarrow v(U) = 0$" + "$W \perp\!\!\!\perp Z,A|U$"

$$\mathbb{E}[\mathbb{E}[Y|U,A]|Z,A] = \mathbb{E}[h(A,W)|Z,A] = \mathbb{E}[\mathbb{E}[h(A,W)|U,Z,A]|Z,A]$$

$$\Rightarrow \mathbb{E}[Y|U,A] = \mathbb{E}[h(A,W)|U,Z,A] = \int h(A,w)\underbrace{f(w|U,Z,A)}_{\equiv f(w|U)}\mathrm{d}w$$

Proximal identification: Step 3

(1) "confounding bridge equation $\mathbb{E}[Y|Z,A] = \mathbb{E}[h(A,W)|Z,A]$" + "exclusion restriction: $Y \perp\!\!\!\perp Z|U,A$":

$$\mathbb{E}[\mathbb{E}[Y|U,A]|Z,A] = \mathbb{E}[h(A,W)|Z,A]$$

(2) "completeness: $\mathbb{E}[v(U)|Z,A] = 0 \Rightarrow v(U) = 0$" + "$W \perp\!\!\!\perp Z, A|U$"

$$\mathbb{E}[Y|U,A] = \int h(A,w)f(w|U)\mathrm{d}w$$

(3)

$$\begin{aligned}
\mathbb{E}[Y(a)] &= \mathbb{E}[\mathbb{E}[Y(a)|U]] \\
&= \mathbb{E}[\mathbb{E}[Y|U,A=a]] \\
&= \int_u \mathbb{E}[Y|U=u,A=a]f(u)\mathrm{d}u \\
&\overset{(2)}{=} \int_u \int_w h(a,w)f(w|u)\mathrm{d}wf(u)\mathrm{d}u \\
&= \int_u \int_w h(a,w)f(w,u)\mathrm{d}w\mathrm{d}u \\
&= \int_w h(a,w)\left\{\int_u f(w,u)\mathrm{d}u\right\}\mathrm{d}w \\
&= \int_w h(a,w)f(w)\mathrm{d}w
\end{aligned}$$

Proximal identification: Complete

(1) "outcome bridge equation $\mathbb{E}[Y|Z,A] = \mathbb{E}[h(A,W)|Z,A]$" +
"exclusion restriction: $Y \perp\!\!\!\perp Z|U,A$":

$$\mathbb{E}[\mathbb{E}[Y|U,A]|Z,A] = \mathbb{E}[h(A,W)|Z,A]$$

(2) "completeness: $\mathbb{E}[v(U)|Z,A] = 0 \Rightarrow v(U) = 0$" + "$W \perp\!\!\!\perp Z,A|U$"

$$\mathbb{E}[Y|U,A] = \int h(A,w)f(w|U)\mathrm{d}w$$

(3)

$$\mathbb{E}[Y(1)] = \mathbb{E}[\mathbb{E}[Y|U,A=1]] = \int_w h(1,w)f(w)\mathrm{d}w$$

Proximal identification: IPW form

(1) "treatment bridge equation $\frac{1}{\mathbb{P}(A=1|W)} = \mathbb{E}[q(Z,A)|A=1,W]$" + "exclusion restriction: $W \perp\!\!\!\perp Z, A|U$":

$$\mathbb{E}\left[\frac{1}{\mathbb{P}(A=1|U)}|A=1,W\right] = \mathbb{E}[q(Z,A)|A=1,W]$$

(2) "completeness: $\mathbb{E}[v(U)|A,W] = 0 \Rightarrow v(U) = 0$" + "$Z \perp\!\!\!\perp Y|U$"

$$\frac{1}{\mathbb{P}(A=1|U)} = \int q(z,A)f(z|U,A=1)\mathrm{d}z$$

(3)

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\frac{AY}{\mathbb{P}(A=1|U)}\right] = \mathbb{E}\left[Aq(Z,A)Y\right]$$

Naturally, two forms give us "doubly robust" proximal ATE identification

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[Aq(Z,A)(Y - h(A,W)) + h(1,W)\right]$$

Some final words on proximal causal learning

- The origin of proximal causal learning is from the measurement error literature, in particular the work Kuroki and Pearl, 2014

Some final words on proximal causal learning

- The origin of proximal causal learning is from the measurement error literature, in particular the work Kuroki and Pearl, 2014

- If you want more intuitive explanation, see
  REF: Shi, Miao, Tchetgen Tchetgen. A Selective Review of Negative Control Methods in Epidemiology. Epidemiology 2021
  REF: Tchetgen Tchetgen, Ying, Cui, Shi, Miao. An Introduction to Proximal Causal Learning. Statistical Science 2024+

Some final words on proximal causal learning

- The origin of proximal causal learning is from the measurement error literature, in particular the work Kuroki and Pearl, 2014

- If you want more intuitive explanation, see
  REF: Shi, Miao, Tchetgen Tchetgen. A Selective Review of Negative Control Methods in Epidemiology. Epidemiology 2021
  REF: Tchetgen Tchetgen, Ying, Cui, Shi, Miao. An Introduction to Proximal Causal Learning. Statistical Science 2024+

- It is possible to use techniques from causal graphical models to design algorithms to select valid proxies from data (Kummerfield-Lim-Shi, 2022)

Other related frameworks

- Most frameworks dealing with unmeasured confounding developed in economics and statistics are related to IV or proximal causal learning (in fact, you should have realized that proxies are just generalizations of IVs)

Other related frameworks

- Most frameworks dealing with unmeasured confounding developed in economics and statistics are related to IV or proximal causal learning (in fact, you should have realized that proxies are just generalizations of IVs)
- Examples: Difference-in-Difference, Synthetic Control, Regression Discontinuity, Multiple Treatments, Bespoke IV, Data Combination ... (study on your own if interested)

Synthetic control

- We will only cover one particular method called “synthetic control” (SC), invented by econometrician Alberto Abadie and colleagues in 2003

Synthetic control

- We will only cover one particular method called "synthetic control" (SC), invented by econometrician Alberto Abadie and colleagues in 2003
- REF: the original paper published in the top 5 economics journal AER (like top 4 in math), and another case study published in JASA in 2010
- Athey & Imbens praised SC as "the most important innovation in the policy evaluation literature in the last 15 years"

Synthetic control

- We will only cover one particular method called "synthetic control" (SC), invented by econometrician Alberto Abadie and colleagues in 2003
- REF: the original paper published in the top 5 economics journal AER (like top 4 in math), and another case study published in JASA in 2010
- Athey & Imbens praised SC as "the most important innovation in the policy evaluation literature in the last 15 years"
- SC is designed to answer causal questions when we have the so-called "panel data" (longitudinal data in biostatistics)

Motivation of SC

- Suppose that one would like to study the causal effect of German reunification on GDP

Motivation of SC

- Suppose that one would like to study the causal effect of German reunification on GDP
- Data: 1960 – 2003 GDP information for Germany and 16 other countries without such a reunification
- $Y_{1,t}, t = 1, \cdots, T$: the GDPs for Germany
- $Y_{i,t}, i = 2, \cdots, N; t = 1, \cdots, T$: the GDPs for 16 other countries (untreated)
- T_0: the year of reunification, so $Y_{1,t}$ is untreated when $t \leq T_0$, but treated when $t > T_0$

The data

- Downloadable from https://doi.org/10.7910/DVN/24714
- including information on: country, year, gdp, and other time-varying covariates

SC: linear model case

- Suppose the following linear SEM for Germany:

$$Y_{1,t} = \begin{cases} \tau_t + \alpha_1^\top U_t + \varepsilon_{1,t} & t > T_0 \\ \alpha_1^\top U_t + \varepsilon_{1,t} & t \leq T_0 \end{cases}$$

where U_t is a stochastic process that changes with t

SC: linear model case

- Suppose the following linear SEM for Germany:

$$Y_{1,t} = \begin{cases} \tau_t + \alpha_1^\top U_t + \varepsilon_{1,t} & t > T_0 \\ \alpha_1^\top U_t + \varepsilon_{1,t} & t \leq T_0 \end{cases}$$

  where U_t is a stochastic process that changes with t

- Potential outcome & consistency assumption:

$$Y_{1,t} = \begin{cases} Y_{1,t}(0) = \alpha_1^\top U_t + \varepsilon_{1,t} & t \leq T_0 \\ Y_{1,t}(1) = Y_{1,t}(0) + \tau_t & t > T_0 \end{cases}$$

- ATE of the treated unit: $\mathbb{E}[Y_{1,t}(1) - Y_{1,t}(0)] = \tau_t$ for $t > T_0$

SC: linear model case

- Suppose the following linear SEM for Germany:

$$Y_{1,t} = \begin{cases} \tau_t + \alpha_1^\top U_t + \varepsilon_{1,t} & t > T_0 \\ \alpha_1^\top U_t + \varepsilon_{1,t} & t \leq T_0 \end{cases}$$

  where U_t is a stochastic process that changes with t

- Potential outcome & consistency assumption:

$$Y_{1,t} = \begin{cases} Y_{1,t}(0) = \alpha_1^\top U_t + \varepsilon_{1,t} & t \leq T_0 \\ Y_{1,t}(1) = Y_{1,t}(0) + \tau_t & t > T_0 \end{cases}$$

- ATE of the treated unit: $\mathbb{E}[Y_{1,t}(1) - Y_{1,t}(0)] = \tau_t$ for $t > T_0$

- From the single time series alone, $\tau_t, t > T_0$ is not identified

Synthetic control by other countries

- Abadie then realized that we also have data from other untreated countries – can we do something similar to matching to create a hypothetical "Germany" that was never re-unified from the data

$$Y_{i,t} = \alpha_i^\top U_t, i = 2, \cdots, N; t = 1, \cdots, T$$

Synthetic control by other countries

- Abadie then realized that we also have data from other untreated countries – can we do something similar to matching to create a hypothetical "Germany" that was never re-unified from the data

$$Y_{i,t} = \alpha_i^\top U_t, i = 2, \cdots, N; t = 1, \cdots, T$$

- Under what assumptions, can we achieve this goal?

Synthetic control by other countries

- Abadie then realized that we also have data from other untreated countries – can we do something similar to matching to create a hypothetical “Germany” that was never re-unified from the data

$$Y_{i,t} = \alpha_i^\top U_t, i = 2, \cdots, N; t = 1, \cdots, T$$

- Under what assumptions, can we achieve this goal?

- Existence of SC: there exists a set of weights $w_i, i = 2, \cdots, N$ (sum to one) such that

$$\alpha_1 = \sum_{i=2}^{N} w_i \alpha_i \tag{1}$$

Synthetic control by other countries

- Abadie then realized that we also have data from other untreated countries – can we do something similar to matching to create a hypothetical “Germany” that was never re-unified from the data

$$Y_{i,t} = \alpha_i^\top U_t, i = 2, \cdots, N; t = 1, \cdots, T$$

- Under what assumptions, can we achieve this goal?

- Existence of SC: there exists a set of weights $w_i, i = 2, \cdots, N$ (sum to one) such that

$$\alpha_1 = \sum_{i=2}^{N} w_i \alpha_i \tag{1}$$

- Under (1), we achieve identification: $t > T_0$

$$\begin{aligned}\tau_t &= \mathbb{E}[Y_{1,t}(1) - Y_{1,t}(0)] = \mathbb{E}[Y_{1,t}] - \alpha_1^\top \mathbb{E}[U_t] \\ &= \mathbb{E}[Y_{1,t}] - \sum_{i=2}^{N} w_i \alpha_i^\top \mathbb{E}[U_t] = \mathbb{E}[Y_{1,t}] - \sum_{i=2}^{N} w_i \mathbb{E}[Y_{i,t}]\end{aligned}$$

Finding the weights

- The only remaining piece is to find out the weights $\boldsymbol{w} = (w_1, \cdots, w_N)^\top$

Finding the weights

- The only remaining piece is to find out the weights $\boldsymbol{w} = (w_1, \cdots, w_N)^\top$

- A key observation under (1): for $t \leq T_0$,

$$Y_{1,t} = \sum_{i=2}^{N} w_i Y_{i,t} + \underbrace{\left(\varepsilon_{1,t} - \sum_{i=2}^{N} w_i \varepsilon_{i,t}\right)}_{\text{mean zero}}$$

Finding the weights

- The only remaining piece is to find out the weights $\boldsymbol{w} = (w_1, \cdots, w_N)^\top$

- A key observation under (1): for $t \leq T_0$,

$$Y_{1,t} = \sum_{i=2}^{N} w_i Y_{i,t} + \underbrace{\left(\varepsilon_{1,t} - \sum_{i=2}^{N} w_i \varepsilon_{i,t}\right)}_{\text{mean zero}}$$

- This observation leads to the following constrained least-square estimator of the weights:

$$\widehat{\boldsymbol{w}} = \arg \min_{0 \leq \boldsymbol{w} \leq 1, \mathbb{1}^\top \boldsymbol{w} = 1} \frac{1}{T_0} \sum_{t=1}^{T_0} \left(Y_{i,t} - \sum_{i=2}^{N} w_i Y_{i,t} \right)^2$$

Germany reunification example

- Use “Synth” package in R

Germany reunification example

- Use "Synth" package in R
- See "exercises.R"

Final comments on synthetic controls

- The theoretical justification of the constrained least square methods is tricky – the noise term is correlated with the "regressors" in the model (because $Y_{i,t}$ is determined by $\varepsilon_{i,t}$)

Final comments on synthetic controls

- The theoretical justification of the constrained least square methods is tricky – the noise term is correlated with the "regressors" in the model (because $Y_{i,t}$ is determined by $\varepsilon_{i,t}$)

- SC is also connected with matrix completion (the statistical problem that arises from the Netflix challenge)

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \cdots & \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & \cdots & \checkmark & \checkmark & \checkmark \\ \text{NA} & \checkmark & \cdots & \checkmark & \checkmark & \checkmark \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \text{NA} & \checkmark & \cdots & \checkmark & \checkmark & \checkmark \end{pmatrix}$$

for more connections, see Athey et al. '21 and Amjad, Shah, Shen '19 partial identification, nontrivial inequality constraints and a first encounter of quantum mechanics in causal inference

Partial identification instead of point identification

- So far we have largely focused on (point) identification theory except for a brief intro to sensitivity analysis

Partial identification instead of point identification

- So far we have largely focused on (point) identification theory except for a brief intro to sensitivity analysis

- Partial identification: set-valued identification

Partial identification instead of point identification

- So far we have largely focused on (point) identification theory except for a brief intro to sensitivity analysis
- Partial identification: set-valued identification
- Scenario 1: core IV conditions hold but no extra modeling/proxy assumptions

Partial identification instead of point identification

- So far we have largely focused on (point) identification theory except for a brief intro to sensitivity analysis
- Partial identification: set-valued identification
- Scenario 1: core IV conditions hold but no extra modeling/proxy assumptions
- Scenario 2: mismatched data fusion (e.g. one dataset has (X, A) but the other dataset has (A, Y))

Partial identification instead of point identification

- So far we have largely focused on (point) identification theory except for a brief intro to sensitivity analysis
- Partial identification: set-valued identification
- Scenario 1: core IV conditions hold but no extra modeling/proxy assumptions
- Scenario 2: mismatched data fusion (e.g. one dataset has (X, A) but the other dataset has (A, Y))
- Many others (e.g. very natural to consider invalid proxy)

Partial identification instead of point identification

- So far we have largely focused on (point) identification theory except for a brief intro to sensitivity analysis
- Partial identification: set-valued identification
- Scenario 1: core IV conditions hold but no extra modeling/proxy assumptions
- Scenario 2: mismatched data fusion (e.g. one dataset has (X, A) but the other dataset has (A, Y))
- Many others (e.g. very natural to consider invalid proxy)
- We will not cover the quantum mechanics part (read the materials if interested)

Trivial partial identification

- Consider binary treatment ($A \in \{0, 1\}$) and binary outcome ($Y \in \{0, 1\}$)

Trivial partial identification

- Consider binary treatment ($A \in \{0, 1\}$) and binary outcome ($Y \in \{0, 1\}$)

- $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \Pr(Y(1) = 1) - \Pr(Y(0) = 1)$
  recall observed-counterfactual by consistency:
  $Y = AY(1) + (1 - A)Y(0)$

Trivial partial identification

- Consider binary treatment ($A \in \{0, 1\}$) and binary outcome ($Y \in \{0, 1\}$)
- $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \Pr(Y(1) = 1) - \Pr(Y(0) = 1)$
  recall observed-counterfactual by consistency:
  $Y = AY(1) + (1 - A)Y(0)$
- Trivia:

$$\begin{aligned}\tau &= \Pr(Y(1) = 1, A = 1) + \Pr(Y(1) = 1, A = 0) - \Pr(Y(0) = 1, A = 1) - \Pr(Y(0) = 1, A = 0) \\ &= \Pr(Y = 1, A = 1) - \Pr(Y = 1, A = 0) + \underbrace{\Pr(Y(1) = 1, A = 0)}_{a} - \underbrace{\Pr(Y(0) = 1, A = 1)}_{b}\end{aligned}$$

$$\Rightarrow \tau \begin{cases} \geq \Pr(Y = 1, A = 1) - \Pr(Y = 1, A = 0) - \Pr(A = 1) & a = 0, b \leq \Pr(A = 1) \\ \leq \Pr(Y = 1, A = 1) - \Pr(Y = 1, A = 0) + \Pr(A = 0) & a \leq \Pr(A = 0), b = 0 \end{cases}$$

$$\Leftrightarrow \tau \begin{cases} \geq - \Pr(Y = 0, A = 1) - \Pr(Y = 1, A = 0) \\ \leq \Pr(Y = 1, A = 1) + \Pr(Y = 0, A = 0) \end{cases}$$

Conclusion:

$$- \Pr(Y = 0, A = 1) - \Pr(Y = 1, A = 0) \leq \tau \leq \Pr(Y = 1, A = 1) + \Pr(Y = 0, A = 0)$$

Trivial partial identification

- Consider binary treatment ($A \in \{0,1\}$) and binary outcome ($Y \in \{0,1\}$)
- $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \Pr(Y(1)=1) - \Pr(Y(0)=1)$
  recall observed-counterfactual by consistency:
  $Y = AY(1) + (1-A)Y(0)$
- Trivia:

$$\begin{aligned}\tau &= \Pr(Y(1)=1, A=1) + \Pr(Y(1)=1, A=0) - \Pr(Y(0)=1, A=1) - \Pr(Y(0)=1, A=0)\\ &= \Pr(Y=1, A=1) - \Pr(Y=1, A=0) + \underbrace{\Pr(Y(1)=1, A=0)}_{a} - \underbrace{\Pr(Y(0)=1, A=1)}_{b}\end{aligned}$$

$$\Rightarrow \tau \begin{cases} \geq \Pr(Y=1, A=1) - \Pr(Y=1, A=0) - \Pr(A=1) & a=0, b \leq \Pr(A=1) \\ \leq \Pr(Y=1, A=1) - \Pr(Y=1, A=0) + \Pr(A=0) & a \leq \Pr(A=0), b=0 \end{cases}$$

$$\Leftrightarrow \tau \begin{cases} \geq -\Pr(Y=0, A=1) - \Pr(Y=1, A=0) \\ \leq \Pr(Y=1, A=1) + \Pr(Y=0, A=0) \end{cases}$$

Conclusion:

$$-\Pr(Y=0, A=1) - \Pr(Y=1, A=0) \leq \tau \leq \Pr(Y=1, A=1) + \Pr(Y=0, A=0)$$

Width of this trivial bound is 1, so almost always covers 0

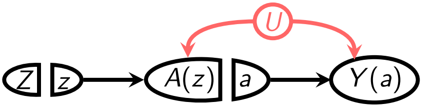
Can we do better than trivial bounds?

- Yes, using IV
  Hernan, Robins. Instruments for Causal Inference: An Epidemiologist's Dream?

Can we do better than trivial bounds?

- Yes, using IV
  Hernan, Robins. Instruments for Causal Inference: An Epidemiologist's Dream?

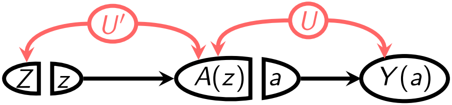
- IV SWIG (intervening both Z and A simultaneously)



Can we do better than trivial bounds?

- Yes, using IV
  Hernan, Robins. Instruments for Causal Inference: An Epidemiologist's Dream?

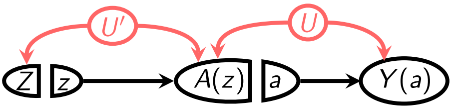
- In fact, for partial identification purpose, we can consider a more relaxed IV SWIG



Can we do better than trivial bounds?

- Yes, using IV
  Hernan, Robins. Instruments for Causal Inference: An Epidemiologist's Dream?

- In fact, for partial identification purpose, we can consider a more relaxed IV SWIG



- Even under the relaxed IV SWIG, we have
  latent-variable exclusion restriction & exogeneity

$$\Pr(Y(z=1,a)=1|U)=\Pr(Y(z=0,a)=1|U), a\in\{0,1\};$$
$$Z \perp\!\!\!\perp U; Y(z,a) \perp\!\!\!\perp Z, A(z)|U, a, z\in\{0,1\}^2$$

Narrow it down using IV: Robins-Manski bounds

Marginalizing U, "relaxed" IV core becomes marginal IV assumptions

$$Y(z,a) \perp\!\!\!\perp Z, P(Y(1,a)=1) = P(Y(0,a)=1), a, z \in \{0,1\}^2$$

Robins (1989) & Manski (1990) showed

1. When conditioning on the same z

$$\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1)|Z=z] - \mathbb{E}[Y(1)|Z=z] \Rightarrow$$

$$\tau \in \left[\begin{array}{c} \max_{z=0,1} \{-\Pr(Y=0, A=1|Z=z) - \Pr(Y=1, A=0|Z=z)\}, \\ \min_{z=0,1} \{\Pr(Y=1, A=1|Z=z) + \Pr(Y=0, A=0|Z=z)\} \end{array} \right]$$

Narrow it down using IV: Robins-Manski bounds

Marginalizing U, "relaxed" IV core becomes marginal IV assumptions

$$Y(z,a) \perp\!\!\!\perp Z, P(Y(1,a)=1) = P(Y(0,a)=1), a, z \in \{0,1\}^2$$

Robins (1989) & Manski (1990) showed

1. When conditioning on the same z

$$\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1)|Z=z] - \mathbb{E}[Y(1)|Z=z] \Rightarrow$$

$$\tau \in \left[\begin{array}{c} \max_{z=0,1} \{-\Pr(Y=0, A=1|Z=z) - \Pr(Y=1, A=0|Z=z)\}, \\ \min_{z=0,1} \{\Pr(Y=1, A=1|Z=z) + \Pr(Y=0, A=0|Z=z)\} \end{array} \right]$$

2. When conditioning on different z's: lower bound

$$\begin{aligned}
\tau &= \Pr(Y(1)=1, A=1|Z=z) + \Pr(Y(1)=1, A=0|Z=z) \\
&\quad - \Pr(Y(0)=1, A=1|Z=z') - \Pr(Y(0)=1, A=0|Z=z') \\
&= \Pr(Y=1, A=1|Z=z) - \Pr(Y=1, A=0|Z=z') \\
&\quad + \Pr(Y(1)=1, A=0|Z=z) - \Pr(Y(0)=1, A=1|Z=z') \\
&= \Pr(A=1|Z=z) - \Pr(Y=0, A=1|Z=z) - \Pr(Y=1, A=0|Z=z') \\
&\quad + \Pr(Y(1)=1, A=0|Z=z) - \Pr(Y(0)=1, A=1|Z=z') \\
&= -\Pr(Y=0, A=1|Z=z) - \Pr(Y=1, A=0|Z=z') \\
&\quad + \Pr(A=1|Z=z) + \Pr(Y(1)=1, A=0|Z=z) - \Pr(Y(0)=1, A=1|Z=z')
\end{aligned}$$

Narrow it down using IV: Robins-Manski bounds

Marginalizing U, "relaxed" IV core becomes marginal IV assumptions

$$Y(z,a) \perp\!\!\!\perp Z, P(Y(1,a)=1) = P(Y(0,a)=1), a, z \in \{0,1\}^2$$

Robins (1989) & Manski (1990) showed

1. When conditioning on the same z

$$\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1)|Z=z] - \mathbb{E}[Y(1)|Z=z] \Rightarrow$$

$$\tau \in \left[\begin{array}{c} \max_{z=0,1} \{-\Pr(Y=0, A=1|Z=z) - \Pr(Y=1, A=0|Z=z)\}, \\ \min_{z=0,1} \{\Pr(Y=1, A=1|Z=z) + \Pr(Y=0, A=0|Z=z)\} \end{array} \right]$$

2. When conditioning on different z's: lower bound

$$\begin{aligned}
\tau &= \Pr(Y(1)=1, A=1|Z=z) + \Pr(Y(1)=1, A=0|Z=z) \\
&\quad - \Pr(Y(0)=1, A=1|Z=z') - \Pr(Y(0)=1, A=0|Z=z') \\
&= \Pr(Y=1, A=1|Z=z) - \Pr(Y=1, A=0|Z=z') \\
&\quad + \Pr(Y(1)=1, A=0|Z=z) - \Pr(Y(0)=1, A=1|Z=z') \\
&= \Pr(A=1|Z=z) - \Pr(Y=0, A=1|Z=z) - \Pr(Y=1, A=0|Z=z') \\
&\quad + \Pr(Y(1)=1, A=0|Z=z) - \Pr(Y(0)=1, A=1|Z=z') \\
&= -\Pr(Y=0, A=1|Z=z) - \Pr(Y=1, A=0|Z=z') \\
&\quad + \Pr(A=1|Z=z) + \Pr(Y(1)=1, A=0|Z=z) - \Pr(Y(0)=1, A=1|Z=z') \\
&\geq -\Pr(Y=0, A=1|Z=z) - \Pr(Y=1, A=0|Z=z') \\
&\quad + \Pr(A=1|Z=z) - \Pr(A=1|Z=z')
\end{aligned}$$

Narrow it down using IV: Robins-Manski bounds

Robins (1989) & Manski (1990)

1. When conditioning on the same z

$$\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1)|Z=z] - \mathbb{E}[Y(1)|Z=z] \Rightarrow$$

$$\tau \in \left[\begin{array}{c} \max_{z=0,1}\{-\Pr(Y=0, A=1|Z=z) - \Pr(Y=1, A=0|Z=z)\}, \\ \min_{z=0,1}\{\Pr(Y=1, A=1|Z=z) + \Pr(Y=0, A=0|Z=z)\} \end{array} \right]$$

2. When conditioning on different z's: upper bound

$$\begin{aligned}
\tau &= \Pr(Y(1)=1, A=1|Z=z) + \Pr(Y(1)=1, A=0|Z=z) \\
&\quad - \Pr(Y(0)=1, A=1|Z=z') - \Pr(Y(0)=1, A=0|Z=z') \\
&= \Pr(Y=1, A=1|Z=z) - \Pr(Y=1, A=0|Z=z') \\
&\quad + \Pr(Y(1)=1, A=0|Z=z) - \Pr(Y(0)=1, A=1|Z=z') \\
&= \Pr(Y=1, A=1|Z=z) + \Pr(Y=1, A=0|Z=z') - \Pr(A=0|Z=z') \\
&\quad + \Pr(Y(1)=1, A=0|Z=z) - \Pr(Y(0)=1, A=1|Z=z') \\
&= \Pr(Y=1, A=1|Z=z) + \Pr(Y=0, A=0|Z=z') \\
&\quad - \Pr(A=0|Z=z') + \Pr(Y(1)=1, A=0|Z=z) - \Pr(Y(0)=1, A=1|Z=z')
\end{aligned}$$

Narrow it down using IV: Robins-Manski bounds

Robins (1989) & Manski (1990)

1. When conditioning on the same z

$$\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1)|Z=z] - \mathbb{E}[Y(1)|Z=z] \Rightarrow$$

$$\tau \in \begin{bmatrix} \max_{z=0,1}\{-\Pr(Y=0,A=1|Z=z) - \Pr(Y=1,A=0|Z=z)\}, \\ \min_{z=0,1}\{\Pr(Y=1,A=1|Z=z) + \Pr(Y=0,A=0|Z=z)\} \end{bmatrix}$$

2. When conditioning on different z's: upper bound

$$\begin{aligned}
\tau &= \Pr(Y(1)=1,A=1|Z=z) + \Pr(Y(1)=1,A=0|Z=z) \\
&\quad - \Pr(Y(0)=1,A=1|Z=z') - \Pr(Y(0)=1,A=0|Z=z') \\
&= \Pr(Y=1,A=1|Z=z) - \Pr(Y=1,A=0|Z=z') \\
&\quad + \Pr(Y(1)=1,A=0|Z=z) - \Pr(Y(0)=1,A=1|Z=z') \\
&= \Pr(Y=1,A=1|Z=z) + \Pr(Y=1,A=0|Z=z') - \Pr(A=0|Z=z') \\
&\quad + \Pr(Y(1)=1,A=0|Z=z) - \Pr(Y(0)=1,A=1|Z=z') \\
&= \Pr(Y=1,A=1|Z=z) + \Pr(Y=0,A=0|Z=z') \\
&\quad - \Pr(A=0|Z=z') + \Pr(Y(1)=1,A=0|Z=z) - \Pr(Y(0)=1,A=1|Z=z') \\
&\leq \Pr(Y=1,A=1|Z=z) + \Pr(Y=0,A=0|Z=z') \\
&\quad + \Pr(A=0|Z=z) - \Pr(A=0|Z=z')
\end{aligned}$$

Robins-Manski bounds

$$\tau \in \left[\max \left\{ \begin{array}{c} -\Pr(Y=0, A=1|Z=1) - \Pr(Y=1, A=0|Z=1), \\ -\Pr(Y=0, A=1|Z=0) - \Pr(Y=1, A=0|Z=0), \\ -\Pr(Y=0, A=1|Z=1) - \Pr(Y=1, A=0|Z=0) \\ +\Pr(A=1|Z=1) - \Pr(A=1|Z=0), \\ -\Pr(Y=0, A=1|Z=0) - \Pr(Y=1, A=0|Z=1) \\ +\Pr(A=1|Z=0) - \Pr(A=1|Z=1) \end{array} \right\}, \right.$$

$$\left. \min \left\{ \begin{array}{c} \Pr(Y=1, A=1|Z=1) + \Pr(Y=0, A=0|Z=1), \\ \Pr(Y=1, A=1|Z=0) + \Pr(Y=0, A=0|Z=0), \\ \Pr(Y=1, A=1|Z=1) + \Pr(Y=0, A=0|Z=0) \\ + \underbrace{\Pr(A=0|Z=1) - \Pr(A=0|Z=0)}_{\Pr(A=1|Z=0)-\Pr(A=1|Z=1)}, \\ \Pr(Y=1, A=1|Z=0) + \Pr(Y=0, A=0|Z=1) \\ + \underbrace{\Pr(A=0|Z=0) - \Pr(A=0|Z=1)}_{\Pr(A=1|Z=1)-\Pr(A=1|Z=0)} \end{array} \right\} \right]$$

Robins-Manski bounds

$$
\tau \in \left[\max \left\{ \begin{array}{c} -\Pr(Y=0, A=1|Z=1) - \Pr(Y=1, A=0|Z=1), \\ -\Pr(Y=0, A=1|Z=0) - \Pr(Y=1, A=0|Z=0), \\ -\Pr(Y=0, A=1|Z=1) - \Pr(Y=1, A=0|Z=0) \\ +\Pr(A=1|Z=1) - \Pr(A=1|Z=0), \\ -\Pr(Y=0, A=1|Z=0) - \Pr(Y=1, A=0|Z=1) \\ +\Pr(A=1|Z=0) - \Pr(A=1|Z=1) \end{array} \right\}, \min \left\{ \begin{array}{c} \Pr(Y=1, A=1|Z=1) + \Pr(Y=0, A=0|Z=1), \\ \Pr(Y=1, A=1|Z=0) + \Pr(Y=0, A=0|Z=0), \\ \Pr(Y=1, A=1|Z=1) + \Pr(Y=0, A=0|Z=0) \\ + \underbrace{\Pr(A=0|Z=1) - \Pr(A=0|Z=0)}_{\Pr(A=1|Z=0)-\Pr(A=1|Z=1)}, \\ \Pr(Y=1, A=1|Z=0) + \Pr(Y=0, A=0|Z=1) \\ + \underbrace{\Pr(A=0|Z=0) - \Pr(A=0|Z=1)}_{\Pr(A=1|Z=1)-\Pr(A=1|Z=0)} \end{array} \right\} \right]
$$

Width of the above bounds?

$$
\text{Width} \leq \underbrace{\Pr(A=0|Z=1) + \Pr(A=1|Z=0)}_{\text{sum of the probabilities of observed non-compliance}}
$$

if $\Pr(A=0|Z=1) + \Pr(A=1|Z=0) \leq \min\{1, \Pr(A=0|Z=0) + \Pr(A=1|Z=1)\}$

Can we strengthen Robins-Manski bounds?

Assume the following instead

$$Y(z=0,a=0) = Y(z=1,a=0) = Y(0)$$
$$Y(z=0,a=1) = Y(z=1,a=1) = Y(1)$$
$$Z \perp\!\!\!\perp (Y(0), Y(1))$$

Can we strengthen Robins-Manski bounds?

Assume the following instead

$$Y(z=0,a=0) = Y(z=1,a=0) = Y(0)$$
$$Y(z=0,a=1) = Y(z=1,a=1) = Y(1)$$
$$Z \perp\!\!\!\perp (Y(0),Y(1))$$

Compare with Robins-Manski's assumption

$$Y(z,a) \perp\!\!\!\perp Z, P(Y(1,a)=1) = P(Y(0,a)=1), a,z \in \{0,1\}^2$$

What are the differences?

Can we strengthen Robins-Manski bounds?

Assume the following instead

$$Y(z=0,a=0)=Y(z=1,a=0)=Y(0)$$
$$Y(z=0,a=1)=Y(z=1,a=1)=Y(1)$$
$$Z \perp\!\!\!\perp (Y(0),Y(1))$$

Compare with Robins-Manski's assumption

$$Y(z,a) \perp\!\!\!\perp Z, P(Y(1,a)=1)=P(Y(0,a)=1), a,z \in \{0,1\}^2$$

What are the differences? The new IV assumptions are cross-world and hence much stronger than the old assumptions!

Can we strengthen Robins-Manski bounds?

Assume the following instead

$$Y(z=0,a=0) = Y(z=1,a=0) = Y(0)$$
$$Y(z=0,a=1) = Y(z=1,a=1) = Y(1)$$
$$Z \perp\!\!\!\perp (Y(0), Y(1))$$

Compare with Robins-Manski's assumption

$$Y(z,a) \perp\!\!\!\perp Z, P(Y(1,a)=1) = P(Y(0,a)=1), a, z \in \{0,1\}^2$$

What are the differences? The new IV assumptions are cross-world and hence much stronger than the old assumptions!

In fact, we have

cross-world IV assumptions $\Rightarrow$ latent-variable IV assumptions $\Rightarrow$ marginal IV assumptions

Balke-Pearl bounds: tightening Robins-Manski bounds with cross-world assumption

- Recall the derivation of Robins-Manski bounds:

$$
\begin{aligned}
\tau &= \Pr(Y(1) = 1, A = 1 | Z = z) + \Pr(Y(1) = 1, A = 0 | Z = z) \\
&\quad - \Pr(Y(0) = 1, A = 1 | Z = z') - \Pr(Y(0) = 1, A = 0 | Z = z') \\
&= \Pr(Y = 1, A = 1 | Z = z) - \Pr(Y = 1, A = 0 | Z = z') \\
&\quad + \Pr(Y(1) = 1, A = 0 | Z = z) - \Pr(Y(0) = 1, A = 1 | Z = z') \\
&= \Pr(A = 1 | Z = z) - \Pr(Y = 0, A = 1 | Z = z) - \Pr(Y = 1, A = 0 | Z = z') \\
&\quad + \Pr(Y(1) = 1, A = 0 | Z = z) - \Pr(Y(0) = 1, A = 1 | Z = z') \\
&= - \Pr(Y = 0, A = 1 | Z = z) - \Pr(Y = 1, A = 0 | Z = z') \\
&\quad + \Pr(A = 1 | Z = z) + \Pr(Y(1) = 1, A = 0 | Z = z) - \Pr(Y(0) = 1, A = 1 | Z = z') \\
&\geq - \Pr(Y = 0, A = 1 | Z = z) - \Pr(Y = 1, A = 0 | Z = z') \\
&\quad + \Pr(A = 1 | Z = z) - \Pr(A = 1 | Z = z')
\end{aligned}
$$

Seemingly quite hopeless to improve!

Balke-Pearl bounds: tightening Robins-Manski bounds with cross-world assumption

- But let's do a coupling argument! Choose $z, z', z'' = 0, 1, 0$ or $1, 0, 1$

$$
\begin{aligned}
\tau &= \Pr(Y(1)=1) - \Pr(Y(0)=1) \\
&= \Pr(Y(1)=1, Y(0)=1|Z=z) + \Pr(Y(1)=1, Y(0)=0|Z=z) \\
&\quad - \Pr(Y(0)=1, Y(1)=1|Z=z') - \Pr(Y(0)=1, Y(1)=0|Z=z'') \\
&= \Pr(Y=1, Y(0)=1, A=1|Z=z) + \Pr(Y(1)=1, Y=1, A=0|Z=z) \\
&\quad + \Pr(Y=1, Y(0)=0, A=1|Z=z) + \Pr(Y(1)=1, Y=0, A=0|Z=z) \\
&\quad - \Pr(Y(0)=1, Y=1, A=1|Z=z') - \Pr(Y=1, Y(1)=1, A=0|Z=z') \\
&\quad - \Pr(Y(0)=1, Y=0, A=1|Z=z'') - \Pr(Y=1, Y(1)=0, A=0|Z=z'') \\
&\geq \Pr(Y=1, A=1|Z=z) + \underbrace{\Pr(Y(1)=1, A=0|Z=z)}_{\geq 0} \\
&\quad - \Pr(Y=1, A=1|Z=z') - \Pr(Y=1, A=0|Z=z') \\
&\quad - \Pr(Y=0, A=1|Z=z'') - \Pr(Y=1, A=0|Z=z'') \\
&\geq \Pr(Y=1, A=1|Z=z) - \Pr(Y=1|Z=z') \\
&\quad - \Pr(Y=0, A=1|Z=z'') - \Pr(Y=1, A=0|Z=z'')
\end{aligned}
$$

Balke-Pearl bounds: tightening Robins-Manski bounds with cross-world assumption

- But let's do a coupling argument! Choose $z, z', z'' = 0, 1, 0$ or $1, 0, 1$

$$\tau \geq \Pr(Y=1, A=1|Z=z) - \Pr(Y=1|Z=z') \\ - \Pr(Y=0, A=1|Z=z'') - \Pr(Y=1, A=0|Z=z'')$$

- By finessing the calculations for the blue and green terms, we also get

$$\tau \geq \Pr(Y=0, A=0|Z=z) - \Pr(Y=0|Z=z') \\ - \Pr(Y=0, A=1|Z=z'') - \Pr(Y=1, A=0|Z=z'')$$

- Upper bounds similar technique; omitted

What is driving the gap between Robins-Manski & Balke-Pearl?

- Of course, cross-world assumptions

What is driving the gap between Robins-Manski & Balke-Pearl?

- Of course, cross-world assumptions

- But is there anything more fundamental in terms of how our physical world is operating?

cross-world IV assumptions $\Rightarrow$ latent-variable IV assumptions $\Rightarrow$ marginal IV assumptions

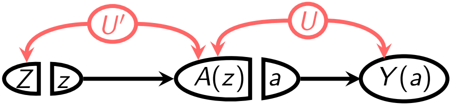
What is driving the gap between Robins-Manski & Balke-Pearl?

- Of course, cross-world assumptions

- But is there anything more fundamental in terms of how our physical world is operating?

  cross-world IV assumptions $\Rightarrow$ latent-variable IV assumptions $\Rightarrow$ marginal IV assumptions

- Let's recall the original latent variable IV SWIG

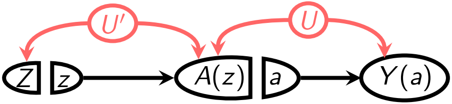


What is driving the gap between Robins-Manski & Balke-Pearl?

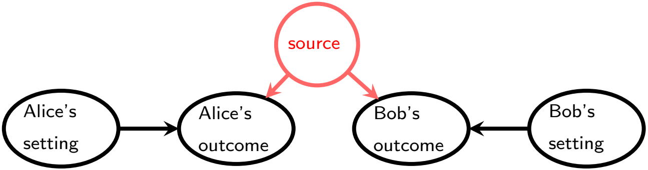
- Of course, cross-world assumptions
- But is there anything more fundamental in terms of how our physical world is operating?

  cross-world IV assumptions ⇒ latent-variable IV assumptions ⇒ marginal IV assumptions

- Let's recall the original latent variable IV SWIG

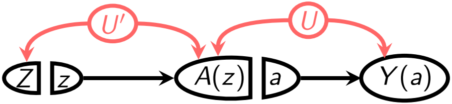


- Let's compare it with the DAG describing Bell-CHSH experiment:

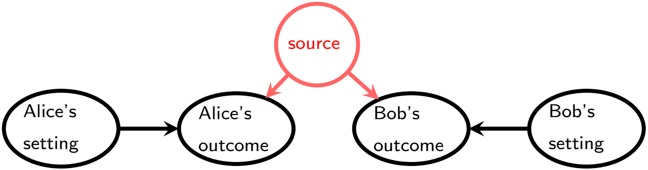


What is driving the gap between Robins-Manski & Balke-Pearl?

- Let's recall the original latent variable IV SWIG

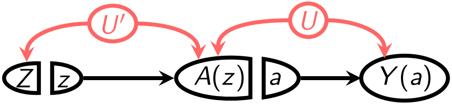


- Let's compare it with the DAG describing Bell-CHSH experiment:

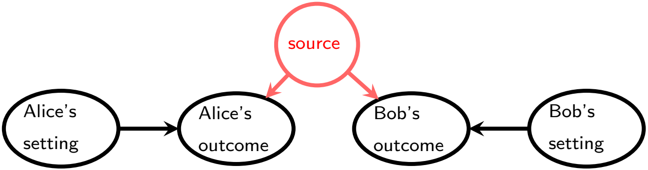


What is driving the gap between Robins-Manski & Balke-Pearl?

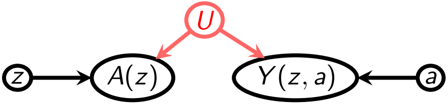
- Let's recall the original latent variable IV SWIG



- Let's compare it with the DAG describing Bell-CHSH experiment:



- Mapping the notation a bit:



Bell-CHSH experiment

One fundamental problem that quantum physicists studied back in the 1960's was if reality is local (Einstein), i.e. if all the probabilistic phenomenon observed in experiments are due to a hidden variable U or if God plays dice (Bohr)

Bell-CHSH experiment

One fundamental problem that quantum physicists studied back in the 1960's was if reality is local (Einstein), i.e. if all the probabilistic phenomenon observed in experiments are due to a hidden variable U or if God plays dice (Bohr)

Think about what statisticians usually do in practice: We always assume data are random draws from some stochastic process, e.g. regression model $Y = \beta X + \mathcal{N}(0, 1)$; but have you ever doubted why we cannot just develop data analysis methods for deterministic models? Are we statisticians fundamentally quantum? Not really

Bell-CHSH experiment

One fundamental problem that quantum physicists studied back in the 1960's was if reality is local (Einstein), i.e. if all the probabilistic phenomenon observed in experiments are due to a hidden variable U or if God plays dice (Bohr)

Think about what statisticians usually do in practice: We always assume data are random draws from some stochastic process, e.g. regression model $Y = \beta X + \mathcal{N}(0, 1)$; but have you ever doubted why we cannot just develop data analysis methods for deterministic models? Are we statisticians fundamentally quantum? Not really

Bell-CHSH experiment can be described as follows: Two particles are prepared. One particle A travels to Alice and the other Y travels to Bob, who are light years apart. Alice and Bob measure the particle spin along directions $z \in \{0, 1\}$ and $a \in \{0, 1\}$ and observe $A(z) \in \{0, 1\}$ and $Y(a) \in \{0, 1\}$

Bell-CHSH experiment

One fundamental problem that quantum physicists studied back in the 1960's was if reality is local (Einstein), i.e. if all the probabilistic phenomenon observed in experiments are due to a hidden variable U or if God plays dice (Bohr)

Think about what statisticians usually do in practice: We always assume data are random draws from some stochastic process, e.g. regression model $Y = \beta X + \mathcal{N}(0, 1)$; but have you ever doubted why we cannot just develop data analysis methods for deterministic models? Are we statisticians fundamentally quantum? Not really

Bell-CHSH experiment can be described as follows: Two particles are prepared. One particle A travels to Alice and the other Y travels to Bob, who are light years apart. Alice and Bob measure the particle spin along directions $z \in \{0, 1\}$ and $a \in \{0, 1\}$ and observe $A(z) \in \{0, 1\}$ and $Y(a) \in \{0, 1\}$

If "local realism" (i.e. existence of U) were true, then the correlation between Alice's outcome $A(z)$ and Bob's outcome $Y(a)$ must satisfy certain constraints, discovered by John **C**lauser, Michael **H**orne, Abner **S**himony, and Richard **H**olt

CHSH-like inequality

Theorem 1 (CHSH-like inequality)

Z, A, Y are all $\{0,1\}$-valued. Under latent-variable IV assumptions

$$\Pr(Y(z=1,a)=1|U) = \Pr(Y(z=0,a)=1|U), a \in \{0,1\};$$
$$Z \perp\!\!\!\perp U; Y(z,a) \perp\!\!\!\perp Z, A(z)|U, a, z \in \{0,1\}^2$$

we have

$$\begin{aligned} 0 \leq \Pr(Y(z,a)=1, A=1|Z=z) + \Pr(Y(z,1-a)=0, A=0|Z=z) \\ + \Pr(Y(1-z,a)=0, A=0|Z=1-z) \\ - \Pr(Y(1-z,1-a)=0, A=0|Z=1-z) \leq 1 \end{aligned}$$

Bell experiment showed CHSH inequality can be violated; hence Bohr were right and Einstein were wrong – reality is non-local, God does play dice, and our world is intrinsically stochastic

What does CHSH-like inequality have to do with Balke-Pearl bounds?

Theorem 2 (Theorem 5.1 of F. Richard Guo's PhD thesis)

CHSH inequality closes the gap between Balke-Pearl and Robins-Manski bounds.

Proof.
Computer assisted proof. Balke-Pearl bounds can be derived symbolically using polytope optimization algorithms. In fact, one can set up and solve the following mathematical program:

$$\max_{\cdots} \text{ or } \min_{\cdots} \Pr(Y(z=0,a=1)=1) - \Pr(Y(z=0,a=0)=1)$$

s.t. trivial inequalities for prob., consistency, marginal IV, CHSH inequality

where $\cdots$ stands for parametrized variables, including $\Pr(Y=y, A=a|Z=z)$ and $\Pr(A=a, Y(0,0)=y_{00}, Y(0,1)=y_{01}, Y(1,0)=y_{10}, Y(1,1)=y_{11}|Z=z)$. The solution to this program is in fact Balke-Pearl bounds □

Summary

cross-world IV assumptions $\Rightarrow$ latent-variable IV assumptions $\Rightarrow$ marginal IV assumptions

Summary

cross-world IV assumptions $\Rightarrow$ latent-variable IV assumptions $\Rightarrow$ marginal IV assumptions

cross-world IV assumptions $\Rightarrow$ latent-variable IV assumptions $\Rightarrow$ $\underbrace{\text{marginal IV assumptions}}_{\Rightarrow \text{Robins-Manski}}$

Summary

cross-world IV assumptions ⇒ latent-variable IV assumptions ⇒ marginal IV assumptions

cross-world IV assumptions ⇒ latent-variable IV assumptions ⇒ $\underbrace{\text{marginal IV assumptions}}_{\Rightarrow \text{Robins-Manski}}$

$\underbrace{\text{cross-world IV assumptions}}_{\Rightarrow \text{Balke-Pearl}}$ ⇒ latent-variable IV assumptions ⇒ $\underbrace{\text{marginal IV assumptions}}_{\Rightarrow \text{Robins-Manski}}$

Summary

cross-world IV assumptions ⇒ latent-variable IV assumptions ⇒ marginal IV assumptions

cross-world IV assumptions ⇒ latent-variable IV assumptions ⇒ $\underbrace{\text{marginal IV assumptions}}_{\Rightarrow \text{Robins-Manski}}$

$\underbrace{\text{cross-world IV assumptions}}_{\Rightarrow \text{Balke-Pearl}}$ ⇒ latent-variable IV assumptions ⇒ $\underbrace{\text{marginal IV assumptions}}_{\Rightarrow \text{Robins-Manski}}$

$\underbrace{\text{cross-world IV assumptions}}_{\Rightarrow \text{Balke-Pearl}}$ ⇒ latent-variable IV assumptions ⇒ $\underbrace{\text{marginal IV assumptions}}_{\Rightarrow \text{Robins-Manski}}$

⇓

CHSH inequality

Summary

cross-world IV assumptions $\Rightarrow$ latent-variable IV assumptions $\Rightarrow$ marginal IV assumptions

cross-world IV assumptions $\Rightarrow$ latent-variable IV assumptions $\Rightarrow$ $\underbrace{\text{marginal IV assumptions}}_{\Rightarrow \text{Robins-Manski}}$

$\underbrace{\text{cross-world IV assumptions}}_{\Rightarrow \text{Balke-Pearl}}$ $\Rightarrow$ latent-variable IV assumptions $\Rightarrow$ $\underbrace{\text{marginal IV assumptions}}_{\Rightarrow \text{Robins-Manski}}$

$\underbrace{\text{cross-world IV assumptions}}_{\Rightarrow \text{Balke-Pearl}}$ $\Rightarrow$ latent-variable IV assumptions $\Rightarrow$ $\underbrace{\text{marginal IV assumptions}}_{\Rightarrow \text{Robins-Manski}}$

$\Downarrow$

CHSH inequality + marginal IV assumptions

$\Downarrow$

Balke-Pearl

More references on partial identification using IVs

- REF: Balke, Pearl. Bounds on Treatment Effects from Studies with Imperfect Compliance. JASA 1997.
- REF: Swanson et al. Partial Identification of the Average Treatment Effect Using Instrumental Variables. JASA 2018.
- REF: Richardson, Robins. Analysis of the Binary Instrumental Variable Model. 2014.

Software

- R package causaloptim
- Learn how to use this package from https://sachsmc.github.io/causaloptim/articles/example-code.html
- Symbolic computation and directly giving you the formula of the bounds
- Including multiple IV bounds and outcome measurement error with proxies
- For some contrived applications in legal contexts, see Tian and Pearl, 2000 UAI

Next chapter

- Causal discovery and structure learning; some more causal graphs