

Stat 256: Causality

Instrumental Variables

Chad Hazlett

UCLA

chazlett@ucla.edu

Spring 2026

Where are we? Where are we going?

- We know now what to do when we can argue for backdoor/CI.
- And we know we can use do-calc to find any other adjustment that, given a non-parametric DAG **and no other assumptions**, can ID an effect (including front-door, napkin, and any others if one exists).
- Next we consider other options, all of which require some assumption not represented non-parametrically by an SCM/DAG.
- First up: **Instrumental variables (IV)**: “what if you can find an **exogenous source of variation** that drives the treatment?”

The intuition: a randomized encouragement

Suppose we can't randomize *exercise* (because we can't force people to exercise, and we shouldn't forbid them either), but we *can* randomize a *free voucher* for a 12-week program.

Half the sample gets a voucher; we observe who actually enrolls, then measure blood pressure (BP).

Good news: we have an experiment with respect to the voucher.

- $\mathbb{E}[\text{BP} \mid \text{voucher}] - \mathbb{E}[\text{BP} \mid \text{no voucher}]$ is identified. This is the **intent-to-treat (ITT)** or **reduced form** effect of the voucher.
- **But it isn't the effect of the program.** How do we get back to that?

Backing it out: four kinds of people

Imagine four groups defined by how they respond to the encouragement:

- *Never-takers*: don't enroll either way. Voucher has no effect on their enrollment $\Rightarrow$ no effect on their BP.
- *Always-takers*: enroll either way. Same: voucher doesn't change anything for them.
- *Compliers*: enroll *iff* they get the voucher. For them, voucher means program, no voucher means no program: effect of voucher (ITT) is effect of program.
- *Defiers*: enroll *iff* they don't get the voucher. We assume away.

Punchline: the ITT over everyone will be an average over these and so will be a *watered-down* version of the program effect. Specifically,

Backing it out: the algebra

Averaging the ITT over the four types:

$$\begin{aligned}
 \text{ITT} &= \underbrace{(\text{ITT} \mid \text{never-taker})}_{=0} \cdot \Pr(\text{never-taker}) \\
 &+ \underbrace{(\text{ITT} \mid \text{always-taker})}_{=0} \cdot \Pr(\text{always-taker}) \\
 &+ (\text{ITT} \mid \text{complier}) \cdot \Pr(\text{complier}) \\
 &+ \underbrace{(\text{ITT} \mid \text{defier})}_{=0, \text{ by assumption}} \cdot \Pr(\text{defier})
 \end{aligned}$$

$$\Rightarrow \text{ITT} \mid \text{complier} = \frac{\text{ITT}}{\Pr(\text{complier})}$$

So we can recover an ITT-for-compliers.

And for compliers, the ITT is the ATE.

From ITT-on-compliers to LATE

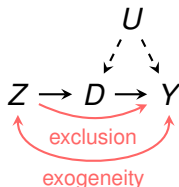
One more thing you need to know: $\Pr(\text{complier})$ is the “first stage” effect:

$\Pr(\text{complier}) = \mathbb{E}[D_1 - D_0] = \mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0]$ (by ignorability of Z), which equals the proportion of compliers (if no defiers).

Putting it together: We call the ATE among the compliers the “local average treatment effect” (LATE):

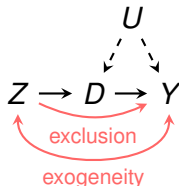
$$\text{LATE} = \frac{\text{ITT}}{\Pr(\text{complier})} = \frac{\mathbb{E}[Y | Z = 1] - \mathbb{E}[Y | Z = 0]}{\mathbb{E}[D | Z = 1] - \mathbb{E}[D | Z = 0]}$$

Basic IV setup with DAGs



- Z is the instrument, D is the treatment, and U is the unmeasured confounder...would typically ruin your day.
- Absence of $Z - Y$ arc, however directed, is known by two assumptions:
 - no common causes of instrument and the outcome (**exogeneity**)
 - no direct or indirect effect of the instrument on the outcome except through treatment (**exclusion**)
- First-stage relationship: Z affects D (**relevance**)
- (Actually Z - D confounding is workable, but will complicate the interpretation we will use.)

Basic IV setup with DAGs



The IV-DAG irony:

- DAGs are great to convey the IV idea; even anti-DAGs probably draw one in the air when they talk about IV.
- But, DAGs+do-calc won't tell you that IV identifies a causal effect:
 - Identification depends on a “monotonicity” assumption (Z only affects D in one direction)

IVs in empirical work

1. Uses of IV with randomization:

- Killer application: RCTs with non-compliance: ITT is unbiased, but IV is the undiluted effect of treatment (in the compliers)
- Randomized encouragement designs are also great way to get an RCT from a non-randomizable treatment.

IVs in empirical work

2. Then there are observational studies that attempt to “find an instrument”. More challenging. To give you a flavor:

- Angrist (1990): Draft lottery as an IV for military service (income as outcome)
- Acemoglu et al (2001): settler mortality as an IV for institutional quality (GDP/capita as outcome)
- Levitt (1997): being an election year as IV for police force size (crime as outcome)
- Kern & Hainmueller (2009): having West German TV reception in East Berlin as an instrument for West German TV watching (outcome is support for the East German regime)
- Acharya, Blackwell, Sen (2015): cotton suitability as IV for proportion of population in slavery in 1860 (outcome is white attitudes today)

IV then and now

- IV was originally framed as a system of structural equations with constant treatment effects, estimated by 2SLS. Useful intuition, but ties identification to a strong constancy assumption.
- We'll work mostly in the modern *potential outcomes* language — LATE, compliers, monotonicity — which makes assumptions and target estimand transparent under heterogeneity.
- The SEM/2SLS treatment is in the appendix for reference (and because you'll see it everywhere in applied work).

POM for Instrumental Variables

Definition (Instrument)

Z_i : Binary instrument for *unit* i .

$$Z_i = \begin{cases} 1 & \text{if unit } i \text{ "encouraged" to receive treatment} \\ 0 & \text{if unit } i \text{ "encouraged" to receive control} \end{cases}$$

Definition (Potential Treatments)

D_{zi} indicates *potential* treatment status given $Z_i = z$

Assumption (Consistency of potential treatments)

Observed treatments are realized as

$$D_i = Z_i \cdot D_{1i} + (1 - Z_i) \cdot D_{0i} \text{ so } D_i = \begin{cases} D_{1i} & \text{if } Z_i = 1 \\ D_{0i} & \text{if } Z_i = 0 \end{cases}$$

Potential Outcome Model for Instrumental Variables

Following Angrist, Imbens, and Rubin (1996), we can define:

Definition

- *Compliers*: $D_{1i} > D_{0i}$ ($D_{0i} = 0$ and $D_{1i} = 1$).
- *Always-takers*: $D_{1i} = D_{0i} = 1$.
- *Never-takers*: $D_{1i} = D_{0i} = 0$.
- *Defiers*: $D_{1i} < D_{0i}$ ($D_{0i} = 1$ and $D_{1i} = 0$).

Only (D_{0i} or D_{1i}) is observed, not both, so we cannot identify which group an individual belongs to.

- Though, further assumptions often help, e.g. no defiers, and sometimes no always-takers.

Potential Outcome Model for Instrumental Variables

Definition (Potential Outcomes)

Given the binary instrument $Z_i \in \{1, 0\}$ and the binary treatment $D_i \in \{1, 0\}$ every unit now has four potential outcomes $Y_i(D, Z)$:

$$Y_i(D_i = 1, Z_i = 1); Y_i(D_i = 1, Z_i = 0); Y_i(D_i = 0, Z_i = 1); Y_i(D_i = 0, Z_i = 0)$$

- Under exclusion restriction, we will have
$$Y_i(d, z = 0) = Y_i(d, z = 1) = Y_i(d)$$

Potential Outcome Model for Instrumental Variables

Estimand (LATE)

$$\alpha_{LATE} = \mathbb{E}[Y_{1i} - Y_{0i} | D_{1i} > D_{0i}]$$

is the Local Average Treatment Effect among compliers.

NB: this estimand varies with the particular instrument Z .

Special case: one-sided noncompliance s.t. $D_{0i} = 0$

- 1 Z is randomized, so compliers in the encouraged arm are representative of all compliers.
- 2 With no always-takers, *the encouraged compliers are just the treated.*
- 3 So the effect among the treated equals the effect among the compliers,
 $\alpha_{ATT} = \alpha_{LATE}$.

Identification of the LATE under IV

Identification Assumption

- 1 *Ignorability of the Instrument:* $(Y_{0i}, Y_{1i}, D_{0i}, D_{1i}) \perp\!\!\!\perp Z_i$
- 2 *First Stage:* $P(D_{1i} = 1) \neq P(D_{0i} = 1)$
- 3 *Monotonicity:* $D_{1i} \geq D_{0i}$ for all i (no defiers)

Identification Result

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0 | D_1 > D_0] &= \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]} \\ &= \frac{\text{Intent to Treat Effect of } Z \text{ on } Y}{\text{First Stage Effect of } Z \text{ on } D} \\ &= \frac{\text{Intent to Treat Effect}}{\text{Proportion of Compliers}}\end{aligned}$$

Identification of the LATE under IV

Identification Assumption

- 1 *Ignorability of the Instrument:* $(Y_{0i}, Y_{1i}, D_{0i}, D_{1i}) \perp\!\!\!\perp Z_i$
- 2 *First Stage:* $P(D_{1i} = 1) \neq P(D_{0i} = 1)$
- 3 *Monotonicity:* $D_{1i} \geq D_{0i}$ for all i (no defiers)

Identification Result

- *Effect of Z on Y is identified (ITT)*
- *Effect of Z on D is identified and is the $Pr(\text{complier})$*
- *All of the effect of Z on Y is through compliers; none is through always-takers, none is through never-takers, and defiers are ruled out, so*

$$ITT = ATE(\text{compliers})Pr(\text{complier}) + 0 + 0$$

$$ATE(\text{compliers}) = \frac{ITT}{Pr(\text{compliers})}$$

Comments on Identification Assumptions

Ignorability of the Instrument:

$$\{Y_{0i}, Y_{1i}, D_{0i}, D_{1i}\} \perp\!\!\!\perp Z_i$$

- The $D_z \perp\!\!\!\perp Z_i$ part implies Z_i allows identification of first-stage. (This can be omitted if we aren't worried about heterogeneous effects).
- The other part, $Y_d \perp\!\!\!\perp Z_i$ encodes both “exogeneity” and “exclusion restriction”.
How so?
- Random assignment *not* sufficient for exclusion. Why not?

First Stage/ Relevance

- Implies that Z induces variation in D
- Fully testable by regressing D on Z . (Though strength is important too).

Monotonicity: $D_1 \geq D_0$

- Rules out defiers. Not testable but sometimes qualitatively easy to argue.

Instrumental Variable: Estimators

Estimator (IV Estimator)

Standard implementation is the IV estimator:

$$\hat{\beta}_{IV} = (\mathbf{Z}^\top \mathbf{D})^{-1} \mathbf{Z}^\top \mathbf{Y}$$

- 1 Typically estimate using `ivreg()`
- 2 If ID assumptions only hold after conditioning on X , you can add covariates:
 - As before, $\mathbf{X} = [1; X \ \mathbf{D}]$ and $\mathbf{Z} = [1 \ \mathbf{X} \ \mathbf{Z}]$, estimate with $(\mathbf{Z}^\top \mathbf{X})^{-1} \mathbf{Z}^\top \mathbf{Y}$
- 3 However, with heterogenous effects, β_{IV} estimated in this way does not necessarily have a clear causal interpretation!
 - complier group changes depending on strata of X .
 - There are numerous tools for “characterizing the compliers”
 - κ -weighting: weights s.t. any model you run gets you a result as if only estimated for compliers (Abadie, 2003): Local average response function (LARF).

IV identification in DAGs

DAGs were useful in talking through IV, but unsatisfying that they do not identify IV.

- $\exists$ graphical criteria for showing valid IVs, but they don't show why they work.
- Annotating the DAG, plus rules to work with such DAGs could be productive.

One proposal¹: Sub-population DAGs for compliers and non-compliers (rule-out defiers),

- for compliers, the SCM says $D = f_D(Z) = Z$.
- for non-compliers, $D = f_D(U_D)$.

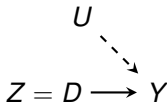


Figure: Among compliers

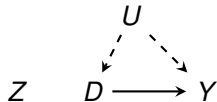


Figure: Among noncompliers

¹This is not unlike the proposal of Steiner et al 2017.

IV identification in DAGs

From the **whole-population DAG** (with Z randomized) we identify:

- $\Pr(\text{comp}) = \mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]$ (first stage)
- $\text{ITT} = \mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]$ (effect of Z on Y)

The two **sub-population DAGs** each do one job:

- *Compliers* ($Z = D$): the effect of Z on Y is the effect of D on Y , so

$$\text{ITT} \mid \text{comp} = \text{ATE} \mid \text{comp}.$$
- *Non-compliers* (no $Z \rightarrow D$, no $Z \rightarrow Y$): Z is disconnected from Y , so

$$\text{ITT} \mid \text{non-comp} = 0.$$

By total expectation:

$$\begin{aligned} \text{ITT} &= (\text{ITT} \mid \text{comp}) \cdot \Pr(\text{comp}) + \underbrace{(\text{ITT} \mid \text{non-comp}) \cdot \Pr(\text{non-comp})}_{=0} \\ &= (\text{ATE} \mid \text{comp}) \cdot \Pr(\text{comp}). \end{aligned}$$

Therefore

$$\text{LATE} = \text{ATE} \mid \text{comp} = \frac{\text{ITT}}{\Pr(\text{comp})}.$$

Summary

- IV was our first strategy outside of do-calculus
- It is very widely used in empirical practice.
- We will talk about sensitivity analysis for IV
- But first we will look at other non-do-calc-identifiable opportunities, i.e. those that flow from imposing other “non-NP” type assumptions.

OVB with instrumental variables

We can apply the same OVB machinery to IV. Walks through, applied to the Card (1993) returns-to-education example.

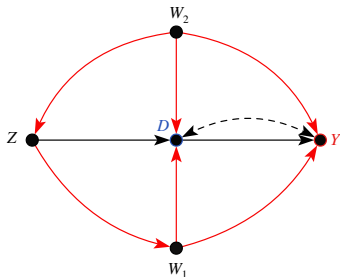
Running example: returns to education (Card 1993)

- Estimating returns to education has long been a challenge.
- OLS with $N = 3010$, covariates $X = \{\text{race, experience, regional}\}$ suggests 7.5% higher earnings per year of education.
- Confounding may be severe: ability, motivation, family/social.
- Card uses *proximity to college* as instrument.

These assumptions may fail!

The IV estimate can be biased — worse than OLS — if *Proximity* is not a valid instrument due to uncontrolled:

- “side-effects” (e.g. high-wage jobs near colleges)
- “common-cause confounders” (e.g. unmeasured SES)



How “bad” would such violations need to be to change the conclusion?

IV regression: three estimators

1. Indirect least squares:

Reduced form: $\text{Earnings} = \hat{\lambda}_{\text{res}} \text{Proximity} + \mathbf{X}\hat{\beta}_{\text{res}} + \hat{\epsilon}_{y,\text{res}}$

First stage: $\text{Education} = \hat{\theta}_{\text{res}} \text{Proximity} + \mathbf{X}\hat{\psi}_{\text{res}} + \hat{\epsilon}_{d,\text{res}}$

2. Two-stage least squares:

2nd stage: $\text{Earnings} = \hat{\tau}_{2\text{SLS},\text{res}} \widehat{\text{Education}} + \mathbf{X}\hat{\beta}_{2\text{SLS},\text{res}} + \hat{\epsilon}_{2\text{SLS},\text{res}}$

Same answer:

$$\hat{\tau}_{2\text{SLS}} = \hat{\tau}_{\text{ILS}} := \frac{\hat{\lambda}_{\text{res}}}{\hat{\theta}_{\text{res}}} = \frac{0.042}{0.319} \approx 0.132.$$

IV regression: Anderson–Rubin

3. Anderson–Rubin (Fieller):

$$\text{Earnings} - \tau_0 \text{Education} = \hat{\phi}_{\tau_0, \text{res}} \text{Proximity} + \mathbf{X} \hat{\beta}_{\tau_0, \text{res}} + \hat{\epsilon}_{\tau_0, \text{res}}$$

- τ_0 that makes $\hat{\phi}_{\tau_0, \text{res}} = 0$ is the point estimate
- $\{\tau_0 : \text{do not reject } \hat{\phi}_{\tau_0, \text{res}} = 0\}$ is the CI
- Correct coverage regardless of instrument strength

All give 13.2% higher earnings per year of education.

How would these change had we included W ?

Start with the reduced form

We wanted to run

$$Y = \hat{\lambda} Z + \mathbf{X}\hat{\beta} + \hat{\gamma} W + \hat{\epsilon}_y.$$

But W was not observed, so we ran

$$Y = \hat{\lambda}_{\text{res}} Z + \mathbf{X}\hat{\beta}_{\text{res}} + \hat{\epsilon}_{y,\text{res}}.$$

How would including W change inference about λ ?

Same OVb analysis, different variable names

- $R^2_{Y \sim W|Z, \mathbf{X}}$: W 's share of outcome variance
- $R^2_{Z \sim W|\mathbf{X}}$: W 's share of *instrument* variance

$$\begin{aligned}
 |\widehat{\text{bias}}| &= \sqrt{\frac{R^2_{Y \sim W|Z, \mathbf{X}} R^2_{Z \sim W|\mathbf{X}}}{1 - R^2_{Z \sim W|\mathbf{X}}}} \text{ df} \cdot \widehat{\text{se}}(\hat{\lambda}_{\text{res}}) \\
 &= \text{BF} \sqrt{\text{df}} \cdot \widehat{\text{se}}(\hat{\lambda}_{\text{res}})
 \end{aligned}$$

$$\begin{aligned}
 \widehat{\text{se}}(\hat{\lambda}) &= \sqrt{\frac{1 - R^2_{Y \sim W|Z, \mathbf{X}}}{1 - R^2_{Z \sim W|\mathbf{X}}}} \frac{\text{df}}{\text{df} - 1} \cdot \widehat{\text{se}}(\hat{\lambda}_{\text{res}}) \\
 &= \text{SeF} \sqrt{\text{df} / (\text{df} - 1)} \cdot \widehat{\text{se}}(\hat{\lambda}_{\text{res}})
 \end{aligned}$$

An embarrassingly simple sensitivity analysis

Change the critical t rather than the point estimate.

Familiar $1 - \alpha$ CI:

$$\hat{\lambda}_{\text{res}} \pm t_{\alpha, \text{df}}^* \cdot \widehat{\text{se}}(\hat{\lambda}_{\text{res}}).$$

Adjust for confounding (limited by R^2):

$$\hat{\lambda}_{\text{res}} \pm t_{\alpha, \text{df} - 1, R^2}^\dagger \cdot \widehat{\text{se}}(\hat{\lambda}_{\text{res}}),$$

where

$$t_{\alpha, \text{df} - 1, R^2}^\dagger := \text{SeF} \sqrt{\text{df} / (\text{df} - 1)} \cdot t_{\alpha, \text{df} - 1}^* + \text{BF} \sqrt{\text{df}}.$$

Routine reporting: (extreme) robustness values

What would it take to reduce the coefficient until we no longer reject (with confidence $1 - \alpha$) that it is $100q\%$ as large?

Using the RF coefficient:

- $RV_{q,\alpha}(\lambda)$: minimal strength of association with both Z and Y needed.
- $XR_{q,\alpha}(\lambda)$: minimal strength with Z alone, even if W explains *all* residual outcome variance.

Sensitivity of RF and FS

What can we learn about IV from sensitivity of RF and FS? Recall $\tau = \lambda/\theta$.

Sensitivity of reduced form: for $H_0 : \tau = 0$, sensitivity of RF *is exactly* sensitivity of IV.

“If it ain’t in the reduced form, it ain’t there.” (Angrist & Pischke)

Sensitivity of first stage: If the sign of the FS isn’t guaranteed, you cannot rule out unbounded estimates / CIs.
AR is consistent with this logic.

Sensitivity with the AR regression

$$Y_{\tau_0} = \hat{\phi}_{\tau_0, \text{res}} Z + \mathbf{X} \hat{\beta}_{\tau_0, \text{res}} + \hat{\epsilon}_{\tau_0, \text{res}}$$

The AR confidence interval:

$$\text{CI}_{1-\alpha}(\tau) = \left\{ \tau_0 : t_{\hat{\phi}_{\text{res}, \tau_0}}^2 \leq (t_{\alpha, \text{df}-1}^*)^2 \right\}.$$

Replace $t_{\alpha, \text{df}-1}^*$ with $t_{\alpha, \text{df}-1, \mathbf{R}^2}^{\dagger \max}$:

$$\text{CI}_{1-\alpha, \mathbf{R}^2}^{\max}(\tau) = \left\{ \tau_0 : t_{\hat{\phi}_{\text{res}, \tau_0}}^2 \leq (t_{\alpha, \text{df}-1, \mathbf{R}^2}^{\dagger \max})^2 \right\}.$$

Analytical solution via quadratic inequality. Yields $\text{RV}_{q, \alpha}(\tau)$, $\text{XRV}_{q, \alpha}(\tau)$ and $\text{RV}_{\geq q, \alpha}(\tau)$, $\text{XRV}_{\geq q, \alpha}(\tau)$.

Card example: sensitivity of RF and FS

1. Reduced form: $H : \tau = 0$.

Outcome: *Earnings* (log)

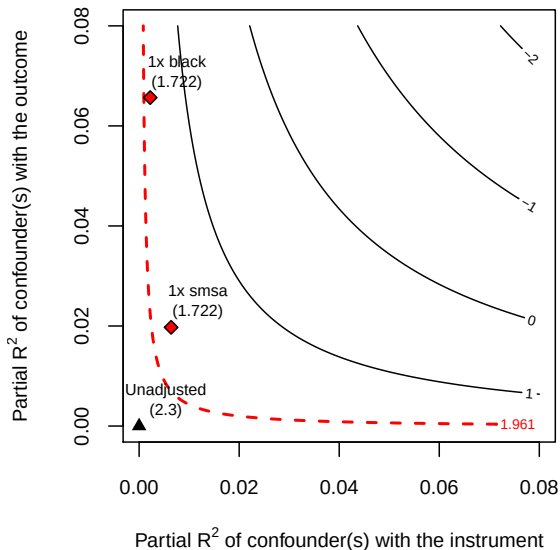
Instrument	Est.	SE	t	$R^2_{Y \sim Z X}$	$XRV_{q^*, \alpha}$	$RV_{q^*, \alpha}$
<i>Proximity</i>	0.042	0.018	2.33	0.18%	0.05%	0.67%
<i>Bound</i> ($1 \times s\text{msa}$): $R^2_{Y \sim W Z, X} = 2\%$, $R^2_{Z \sim W X} = 0.6\%$, $t^\dagger = 2.55$ df = 2994, $q^* = 1$, $\alpha = 0.05$						

2. First stage: stability.

Outcome: *Education* (years)

Instrument	Est.	SE	t	$R^2_{D \sim Z X}$	$XRV_{q^*, \alpha}$	$RV_{q^*, \alpha}$
<i>Proximity</i>	0.32	0.088	3.64	0.44%	0.31%	3.02%
<i>Bound</i> ($1 \times s\text{msa}$): $R^2_{D \sim W Z, X} = 0.5\%$, $R^2_{Z \sim W X} = 0.6\%$, $t^\dagger = 2.26$						

Card example: reduced form contours



Card example: sensitivity of the IV (AR)

Outcome: *Earnings* (log)

Treatment	Est.	LL	UL	t	$XRV_{\geq q^*, \alpha}$	$RV_{\geq q^*, \alpha}$
<i>Education</i>	0.132	0.025	0.285	2.33	0.05%	0.67%

Bound ($1 \times \text{msa}$): $R^2_{Y_0 \sim W|Z, \mathbf{X}} = 2\%$, $R^2_{W \sim Z|\mathbf{X}} = 0.6\%$, $t^\dagger = 2.55$

df = 2994, $q^* = 1$, $\alpha = 0.05$

Table: Minimal sensitivity reporting of IV (AR)

- Identical to RF results because the null is $\tau = 0$.
- More interesting under other null hypotheses.

Card example: CI-limit contours

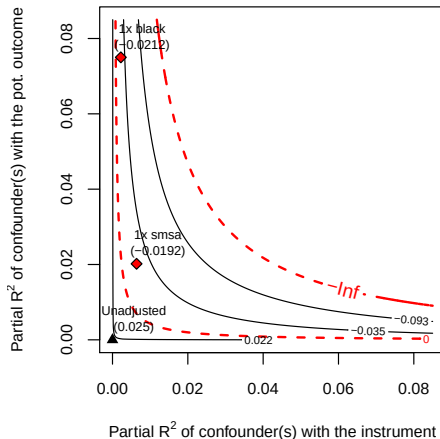


Figure: Sensitivity contours: lower 95% limit

IV sensitivity: conclusions

- OLS sensitivity to confounding can be as simple as adjusting your critical value.
- For null of zero in IV, sensitivity of the reduced form *is all you need*.
- General sensitivity for IV is also easy: in the AR framework, explore how postulated confounding moves point estimates, CI limits, RVs, and bounds.

IV with constant effects

- Write down a causal model for Y_i with constant effects and an unmeasured confounder, U_i :

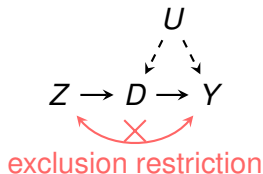
$$Y_i(d, u) = \alpha + \tau d + \gamma u + \eta_i$$

- Under consistency,

$$Y_i = \alpha + \tau D_i + \gamma U_i + \eta_i$$

- Without U , we have error term $\gamma U_i + \eta_i$, and $\text{Cov}(\gamma U_i + \eta_i, D_i) \neq 0$.

The role of the instrument



- If we have a true instrument Z_i , then

$$\text{Cov}(\gamma U_i + \eta_i, Z_i) = 0$$

- Any association of Y and Z is due to D . Specifically,

$$\begin{aligned} \text{Cov}(Y_i, Z_i) &= \text{Cov}(\alpha + \tau D_i + \gamma U_i + \eta_i, Z_i) \\ &= \text{Cov}(\alpha, Z_i) + \text{Cov}(\tau D_i, Z_i) + \text{Cov}(\gamma U_i + \eta_i, Z_i) \\ &= 0 + \tau \text{Cov}(D_i, Z_i) + 0 \\ &\implies \tau = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)} \end{aligned}$$

IV estimator with constant effects

We can see this as a ratio of two linear regression estimates: treatment effect:

$$\tau = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)} = \frac{\text{Cov}(Y_i, Z_i)/\mathbb{V}[Z_i]}{\text{Cov}(D_i, Z_i)/\mathbb{V}[Z_i]}$$

So many words and names:

$$\begin{aligned} \text{Wald} = 2\text{SLS} = \text{ILS} &= \frac{\text{effect of } Z \text{ on } Y}{\text{effect of } Z \text{ on } D} \\ &= \frac{\text{Reduced form effect (aka intent-to-treat)}}{\text{First stage effect}} \end{aligned}$$

With binary instrument, this is merely two DIMs:

$$\tau = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]}$$

More generically,

$$\tau_{IV} = (Z^\top X)^{-1}(Z^\top Y)$$

where X contains D and any covariates.

Assumptions

Let's take stock of the assumptions we have used:

- 1 Exclusion and exogeneity of instrument
- 2 Relevance
- 3 Constant treatment effects

Weak instruments

$$\hat{\tau}_{IV} = \frac{\widehat{\text{Cov}}(Y_i, Z_i)}{\widehat{\text{Cov}}(D_i, Z_i)}$$

$$\xrightarrow{p} \tau + \frac{\text{Cov}(Z_i, U_i)}{\text{Cov}(Z_i, D_i)}$$

- If $\text{Cov}(Z_i, D_i)$ is small, then even very small violations of exogeneity $\text{Cov}(Z_i, U_i) \neq 0$ can lead to severe bias.
- Can easily be much worse than bias of OLS!
- Standard to communicate the strength of the first-stage. $F > 10$ is usual rule, but its just a rule.
- Better: Use Anderson-Rubin approach, which constructs CIs that have correct coverage even with weak instruments

What about covariates?

- No covariates up until now. What if we have a set of covariates X_i that we are also conditioning on?
- Let's start with linear models for both the outcome and the treatment:

$$Y_i = X_i^\top \beta + \tau D_i + \varepsilon_i$$

$$D_i = X_i^\top \alpha + \gamma Z_i + \nu_i$$

- Now, we assume that X_i are **exogenous** along with Z_i :

$$Z \perp \nu; \quad Z \perp \varepsilon$$

$$X \perp \nu; \quad X_i \perp \varepsilon$$

- ...but D_i is (still) **endogenous**: $\mathbb{E}[D_i \varepsilon_i] \neq 0$

Getting the reduced form

- We can plug the treatment equation into the outcome equation:

$$\begin{aligned}
 Y_i &= X_i^\top \beta + \tau[X_i^\top \alpha + \gamma Z_i + \nu_i] + \varepsilon_i \\
 &= X_i^\top \beta + \tau[X_i^\top \alpha + \gamma Z_i] + [\tau \nu_i + \varepsilon_i] \\
 &= X_i^\top \beta + \tau[X_i^\top \alpha + \gamma Z_i] + \varepsilon_i^* \\
 &= X_i^\top \beta + \tau \mathbb{E}[D_i | X_i, Z_i] + \varepsilon_i^*
 \end{aligned}$$

- Because Z_i and X_i are uncorrelated with ν_i and ε_i , then $\mathbb{E}[D_i | X_i, Z_i]$ is also independent of ε_i^* .
- Thus, the population regression coefficient of a Y_i on $[X_i^\top \alpha + \gamma Z_i]$ is the average treatment effect, τ .

Two-stage least squares (2SLS)

Heuristic procedure:

- 1 Run regression of treatment on covariates and instrument
 - 2 Construct fitted values of treatment
 - 3 Run regression of outcome on covariates and fitted values
- This is not how we actually estimate 2SLS because the standard errors from second stage are wrong.
 - You can instead find that the 2SLS estimator is simply $(\mathbf{Z}^\top \mathbf{X})^{-1}(\mathbf{Z}^\top \mathbf{Y})$, where $\mathbf{Z} = [X, Z]$, and $\mathbf{X} = [X, D]$. Can compute SEs from there.