What Does IV Identifies: An Overview

Yagan Hazard

May 8, 2025— Advanced IV Methods

Credits

These slides were first constructed by Peter Hull.

Outline

Preliminaries

IV Mechanics

- Just-Identified IV
- Overidentification
- Weak vs. Many-Weak Bias

IV Interpretation

- LATE Fundamentals
- Generalizations and Limitations

Preliminaries: Parameters, Estimands, and Estimators

Three distinct objects, not always clearly distinguished:

Preliminaries: Parameters, Estimands, and Estimators

Three distinct objects, not always clearly distinguished:

- Parameters come from economic (or other) models of the world
  - $\rightarrow$ E.g. a "structural" model of supply and demand, or a potential outcome model relating schooling to earnings
  - $\rightarrow$ They set the target for empirical analyses: what we want to know

Preliminaries: Parameters, Estimands, and Estimators

Three distinct objects, not always clearly distinguished:

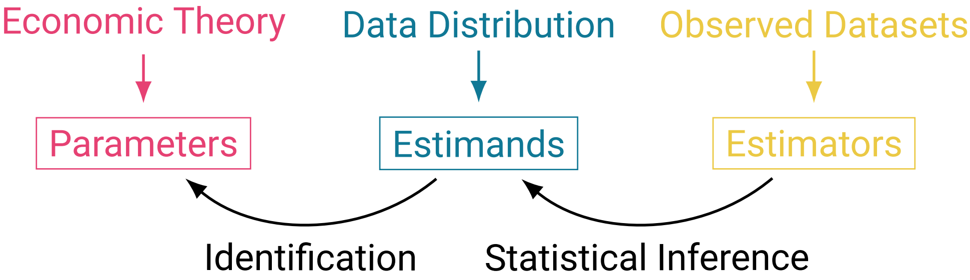
- Parameters come from economic (or other) models of the world
  - $\rightarrow$ E.g. a "structural" model of supply and demand, or a potential outcome model relating schooling to earnings
  - $\rightarrow$ They set the target for empirical analyses: what we want to know
- Estimands are functions of the population data distribution
  - $\rightarrow$ E.g. a difference in means or ratio of population regression coef's
  - $\rightarrow$ We make assumptions to link parameters & estimands (identification)

Preliminaries: Parameters, Estimands, and Estimators

Three distinct objects, not always clearly distinguished:

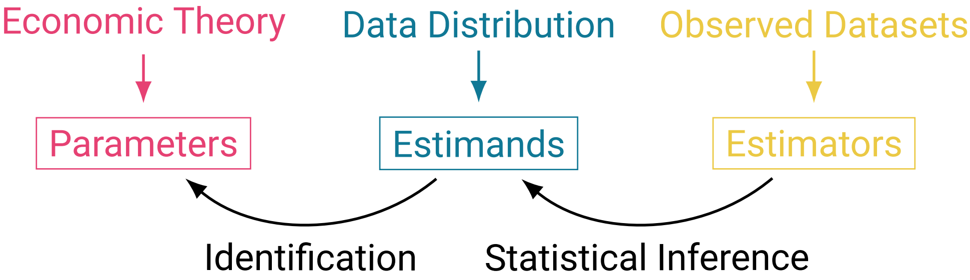
- Parameters come from economic (or other) models of the world
  - → E.g. a "structural" model of supply and demand, or a potential outcome model relating schooling to earnings
  - → They set the target for empirical analyses: what we want to know
- Estimands are functions of the population data distribution
  - → E.g. a difference in means or ratio of population regression coef's
  - → We make assumptions to link parameters & estimands (identification)
- Estimators are functions of observed data (i.e. the "sample")
  - → E.g. a difference in sample means or ratio of OLS coefficients
  - → Since data are random, so are estimators. Each has a distribution
  - → We use knowledge of estimator distributions to learn about estimands (inference) and thus identified parameters

The Key Concepts in Statistical Inference



Separating out the very $\neq$ tasks of identification and inference helps tremendously!

The Key Concepts in Statistical Inference



Separating out the very $\neq$ tasks of identification and inference helps tremendously!

Structure for today:

- recap of how IV estimands are structured [+ brief mention of estimation] ,
- then focus on what they *identify*.

An Example

Human capital theory (e.g. Becker, 1957) tells us that taking econometrics classes will likely boost later-life productivity

An Example

Human capital theory (e.g. Becker, 1957) tells us that taking econometrics classes will likely boost later-life productivity

- Parameter: returns to enrolling in this class β, measured in some outcome Y_i (e.g. lifetime earnings or # of top publications)
- Simple causal model: $Y_i = \beta D_i + \varepsilon_i$, where $D_i \in \{0, 1\}$ indicates class enrollment and ε_i is i's outcome without the class

An Example

Human capital theory (e.g. Becker, 1957) tells us that taking econometrics classes will likely boost later-life productivity

- Parameter: returns to enrolling in this class β, measured in some outcome Y_i (e.g. lifetime earnings or # of top publications)
- Simple causal model: $Y_i = \beta D_i + \varepsilon_i$, where $D_i \in \{0, 1\}$ indicates class enrollment and ε_i is i's outcome without the class

We observe Y_i and D_i for some sample of individuals $i = 1, \ldots, N$

- We fire up Stata and `reg Y D, r`.

An Example

Human capital theory (e.g. Becker, 1957) tells us that taking econometrics classes will likely boost later-life productivity

- Parameter: returns to enrolling in this class β, measured in some outcome Y_i (e.g. lifetime earnings or # of top publications)
- Simple causal model: $Y_i = \beta D_i + \varepsilon_i$, where $D_i \in \{0, 1\}$ indicates class enrollment and ε_i is i's outcome without the class

We observe Y_i and D_i for some sample of individuals $i = 1, \dots, N$

- We fire up Stata and `reg Y D, r`. How do we interpret the results?

Population Regression and Endogeneity

In large samples ($N \to \infty$), the OLS *estimator* $\widehat{\beta}^{OLS}$ gets arbitrarily close to [i.e., consistently estimates] the regression *estimand* β^{OLS}

Population Regression and Endogeneity

In large samples ($N \to \infty$), the OLS *estimator* $\widehat{\beta}^{OLS}$ gets arbitrarily close to [i.e., consistently estimates] the regression *estimand* β^{OLS}

- The *identification* question is thus whether or not $\beta^{OLS} = \beta$

Population Regression and Endogeneity

In large samples ($N \to \infty$), the OLS *estimator* $\widehat{\beta}^{OLS}$ gets arbitrarily close to [i.e., consistently estimates] the regression *estimand* β^{OLS}

- The *identification* question is thus whether or not $\beta^{OLS} = \beta$

By definition, $\beta^{OLS} = \frac{Cov(Y_i, D_i)}{Var(D_i)}$.

Population Regression and Endogeneity

In large samples ($N \to \infty$), the OLS *estimator* $\widehat{\beta}^{OLS}$ gets arbitrarily close to [i.e., consistently estimates] the regression *estimand* β^{OLS}

- The *identification* question is thus whether or not $\beta^{OLS} = \beta$

By definition, $\beta^{OLS} = \frac{Cov(Y_i, D_i)}{Var(D_i)}$. Plugging in our causal model:

$$\begin{aligned} \beta^{OLS} &= \frac{Cov(\beta D_i + \varepsilon_i, D_i)}{Var(D_i)} \\ &= \beta + \frac{Cov(\varepsilon_i, D_i)}{Var(D_i)} \end{aligned}$$

Thus, we have (regression) identification if and only if $Cov(\varepsilon_i, D_i) = 0$.

Otherwise, selection bias: people with certain potential outcomes ε_i are more/less likely to take this class, such that $Cov(\varepsilon_i, D_i) \neq 0$

Population Regression and Endogeneity

In large samples ($N \to \infty$), the OLS *estimator* $\widehat{\beta}^{OLS}$ gets arbitrarily close to [i.e., consistently estimates] the regression *estimand* β^{OLS}

- The *identification* question is thus whether or not $\beta^{OLS} = \beta$

By definition, $\beta^{OLS} = \frac{Cov(Y_i, D_i)}{Var(D_i)}$. Plugging in our causal model:

$$\begin{aligned}\beta^{OLS} &= \frac{Cov(\beta D_i + \varepsilon_i, D_i)}{Var(D_i)} \\ &= \beta + \frac{Cov(\varepsilon_i, D_i)}{Var(D_i)}\end{aligned}$$

Population Regression and Endogeneity

In large samples ($N \to \infty$), the OLS *estimator* $\widehat{\beta}^{OLS}$ gets arbitrarily close to [i.e., consistently estimates] the regression *estimand* β^{OLS}

- The *identification* question is thus whether or not $\beta^{OLS} = \beta$

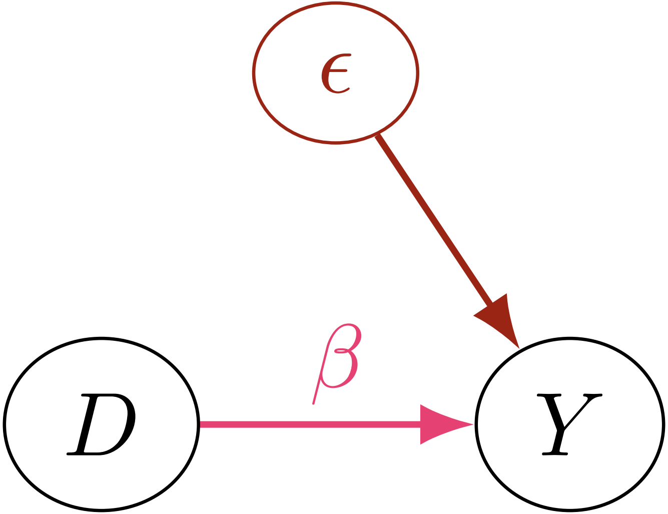
By definition, $\beta^{OLS} = \frac{Cov(Y_i, D_i)}{Var(D_i)}$. Plugging in our causal model:

$$\begin{aligned} \beta^{OLS} &= \frac{Cov(\beta D_i + \varepsilon_i, D_i)}{Var(D_i)} \\ &= \beta + \frac{Cov(\varepsilon_i, D_i)}{Var(D_i)} \end{aligned}$$

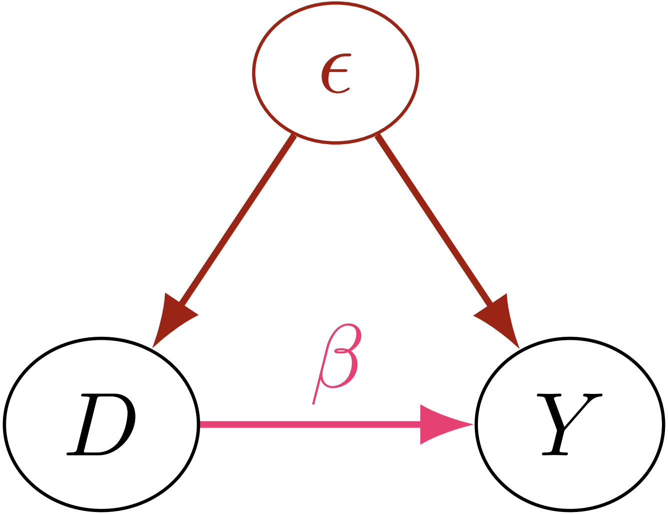
Thus, we have (regression) identification if and only if $Cov(\varepsilon_i, D_i) = 0$.

Otherwise, selection bias: people with certain potential outcomes ε_i are more/less likely to take this class, such that $Cov(\varepsilon_i, D_i) \neq 0$

Regression "Exogeneity"



Regression "Endogeneity"



Outline

The IV Solution

Suppose this class was “oversubscribed” (i.e., more people wanted to attend than we had space for), so seats were determined by lottery.

- $Z_i \in \{0, 1\}$ indicates randomized admission to the class

The IV Solution

Suppose this class was “oversubscribed” (i.e., more people wanted to attend than we had space for), so seats were determined by lottery.

- $Z_i \in \{0, 1\}$ indicates randomized admission to the class
- Randomness + no direct effects of Z_i on Y_i implies $Cov(Z_i, \varepsilon_i) = 0$.

The IV Solution

Suppose this class was “oversubscribed” (i.e., more people wanted to attend than we had space for), so seats were determined by lottery.

- $Z_i \in \{0, 1\}$ indicates randomized admission to the class
- Randomness + no direct effects of Z_i on Y_i implies $Cov(Z_i, \varepsilon_i) = 0$.

Plugging in the model for $\varepsilon_i = Y_i - \beta D_i$, we now have IV identification:

$$Cov(Z_i, Y_i - \beta D_i) = 0 \implies \beta = \frac{Cov(Z_i, Y_i)}{Cov(Z_i, D_i)} \equiv \beta^{IV},$$

so long as $Cov(Z_i, D_i) \neq 0$.

The IV Solution

Suppose this class was "oversubscribed" (i.e., more people wanted to attend than we had space for), so seats were determined by lottery.

- $Z_i \in \{0, 1\}$ indicates randomized admission to the class
- Randomness + no direct effects of Z_i on Y_i implies $Cov(Z_i, \varepsilon_i) = 0$.

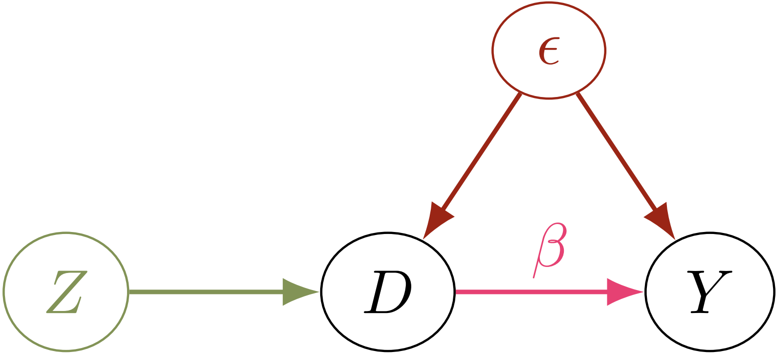
Plugging in the model for $\varepsilon_i = Y_i - \beta D_i$, we now have IV identification:

$$Cov(Z_i, Y_i - \beta D_i) = 0 \implies \beta = \frac{Cov(Z_i, Y_i)}{Cov(Z_i, D_i)} \equiv \beta^{IV},$$

so long as $Cov(Z_i, D_i) \neq 0$.We can estimate this by $\hat{\beta}^{IV} = \frac{\widehat{Cov}(Z_i,Y_i)}{\widehat{Cov}(Z_i,D_i)} = (\boldsymbol{Z'D})^{-1}\boldsymbol{Z'Y}$

[Or, in Stata, `ivreg2 Y (D=Z), r`]

The IV Solution



Note: no arrow connecting ε and Z (“as-good-as-random assignment”), and no arrow from Z to Y directly (“exclusion”). We’ll come back to both.

Reduced Form and First Stage

We're usually pretty comfortable w/regression; how does it connect to IV?

$$\beta^{IV} = \frac{Cov(Z_i, Y_i)}{Cov(Z_i, D_i)} = \frac{Cov(Z_i, Y_i)/Var(Z_i)}{Cov(Z_i, D_i)/Var(Z_i)} \equiv \frac{\rho}{\pi}$$

where ρ and π come from two population regressions:

$$Y_i = \kappa + \rho Z_i + \nu_i \quad \text{The "reduced form"}$$

$$D_i = \mu + \pi Z_i + \eta_i \quad \text{The "first stage"}$$

Angrist (1990): Draft Lottery IV

Angrist (1990) [Vietnam draft lottery as IV to estimate earnings effects of military service]

- $Z_i \in \{0, 1\}$ an indicator for draft eligibility,
- $D_i \in \{0, 1\}$ an indicator for military service,
- Y_i measures later-life earnings.

Angrist (1990): Draft Lottery IV

Angrist (1990) [Vietnam draft lottery as IV to estimate earnings effects of military service]

- $Z_i \in \{0,1\}$ an indicator for draft eligibility,
- $D_i \in \{0,1\}$ an indicator for military service,
- Y_i measures later-life earnings.

$$\beta^{IV} = \frac{Cov(Z_i, Y_i)/Var(Z_i)}{Cov(Z_i, D_i)/Var(Z_i)} = \frac{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]}{E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]}$$ as Z_i is binary

$\rightarrow$ the famous Wald estimand

Regression on binary var.

Angrist (1990): Draft Lottery IV

Angrist (1990) [Vietnam draft lottery as IV to estimate earnings effects of military service]

- $Z_i \in \{0,1\}$ an indicator for draft eligibility,
- $D_i \in \{0,1\}$ an indicator for military service,
- Y_i measures later-life earnings.

$$\beta^{IV} = \frac{Cov(Z_i, Y_i)/Var(Z_i)}{Cov(Z_i, D_i)/Var(Z_i)} = \frac{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]}{E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]}$$ as Z_i is binary

$\rightarrow$ the famous Wald estimand

`Regression on binary var.`

- $E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]$: effect of eligibility on the *probability* of military service (because D_i is binary)
- $E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]$: effect of eligibility on adult earnings [in 1971, 1981...]

IV interprets the latter causal effect [reduced form] in terms of the former [first stage] .

IV Estimates of the Effects of Military Service on the Earnings of White Men born in 1950

Earnings year	Earnings		Veteran Status		Wald Estimate of Veteran Effect
	Mean	Eligibility Effect	Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8 (210.5)	.267	.159 (.040)	-2,741 (1,324)
1971	3,338	-325.9 (46.6)			-2050 (293)
1969	2,299	-2.0 (34.5)			

Note: Adapted from Table 5 in Angrist and Krueger (1999) and author tabulations. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 individuals in the sample.

Adding Controls

We might only think our Z_i is exogenous controlling (linearly) for some vector W_i

- Just add controls to the reduced form and first stage! $\beta^{IV} = \frac{\rho}{\pi}$ for

$$Y_i = \kappa + \rho Z_i + W_i'\phi + \nu_i$$

$$D_i = \mu + \pi Z_i + W_i'\psi + \eta_i$$

  with the estimator $\hat{\beta}^{IV}$ defined analogously

Adding Controls

We might only think our Z_i is exogenous controlling (linearly) for some vector W_i

- Just add controls to the reduced form and first stage! $\beta^{IV} = \frac{\rho}{\pi}$ for

$$Y_i = \kappa + \rho Z_i + W_i'\phi + \nu_i$$

$$D_i = \mu + \pi Z_i + W_i'\psi + \eta_i$$

  with the estimator $\hat{\beta}^{IV}$ defined analogously

- The Frisch-Waugh-Lovell theorem tells us the effective instrument is now $\tilde{Z}_i$: the residuals from regressing Z_i on W_i in the population

Adding Controls

We might only think our Z_i is exogenous controlling (linearly) for some vector W_i

- Just add controls to the reduced form and first stage! $\beta^{IV} = \frac{\rho}{\pi}$ for

$$Y_i = \kappa + \rho Z_i + W_i' \phi + \nu_i$$

$$D_i = \mu + \pi Z_i + W_i' \psi + \eta_i$$

  with the estimator $\hat{\beta}^{IV}$ defined analogously

- The Frisch-Waugh-Lovell theorem tells us the effective instrument is now $\tilde{Z}_i$: the residuals from regressing Z_i on W_i in the population
- E.g. if W_i is a vector of dummies for randomization strata in an RCT, then $\tilde{Z}_i$ captures the within-strata variation in Z_i

Multiple Treatments

We might be interested in a multi-dimensional model: $Y_i = X_i'\beta + \varepsilon_i$

- Instrument vector Z_i, with $L = \dim(Z_i) = \dim(X_i) = J$ ("just-identified")
- Population residuals $\tilde{Z}_i$, given some vector of controls W_i

Multiple Treatments

We might be interested in a multi-dimensional model: $Y_i = X_i'\beta + \varepsilon_i$

- Instrument vector Z_i, with $L = \dim(Z_i) = \dim(X_i) = J$ ("just-identified")
- Population residuals $\tilde{Z}_i$, given some vector of controls W_i

Suppose $Cov(\tilde{Z}_i, \varepsilon_i) = 0$. Then, just as before, we have identification:

$$Cov(\tilde{Z}_i, Y_i - X_i'\beta) = 0 \implies \beta = Cov(\tilde{Z}_i, X_i)^{-1} Cov(\tilde{Z}_i, Y_i) \equiv \beta^{IV},$$

so long as $Cov(\tilde{Z}_i, X_i)$ [a $L \times L$ matrix] is full-rank.

Multiple Treatments

We might be interested in a multi-dimensional model: $Y_i = X_i'\beta + \varepsilon_i$

- Instrument vector Z_i, with $L = \dim(Z_i) = \dim(X_i) = J$ ("just-identified")
- Population residuals $\tilde{Z}_i$, given some vector of controls W_i

Suppose $Cov(\tilde{Z}_i, \varepsilon_i) = 0$. Then, just as before, we have identification:

$$Cov(\tilde{Z}_i, Y_i - X_i'\beta) = 0 \implies \beta = Cov(\tilde{Z}_i, X_i)^{-1} Cov(\tilde{Z}_i, Y_i) \equiv \beta^{IV},$$

so long as $Cov(\tilde{Z}_i, X_i)$ [a $L \times L$ matrix] is full-rank. Equivalently, $\beta^{IV} = \pi^{-1}\rho$ where:

$$Y_i = Z_i'\rho + W_i'\phi + \nu_i$$

$$X_i = \pi Z_i + W_i'\psi_i + \eta_i,$$

[Estimation as in simpler case: take residuals from J 1^{st}-stages and plug them into 2^{nd} stage.]

Multiple Instruments

What happens when $\dim(Z_i) = L > J = \dim(X_i)$? Overidentification:

$$Cov(\tilde{Z}_i, Y_i - X_i'\beta) = 0 \implies \underbrace{Cov(\tilde{Z}_i, Y_i)}_{L\times 1} = \underbrace{Cov(\tilde{Z}_i, X_i)}_{L\times J} \underbrace{\beta}_{J\times 1}$$

so we can drop any $L - J$ instruments and still identify β.

More generally, we can take any full-column rank linear combination $\tilde{Z}_i^* = M\tilde{Z}_i$ [with M a $J \times L$ full-column rank matrix] such that $Cov(\tilde{Z}_i^*, X_i)$ is invertible

$$\underbrace{M \cdot Cov(\tilde{Z}_i, Y_i)}_{Cov(\tilde{Z}_i^*, Y_i)} = \underbrace{M \cdot Cov(\tilde{Z}_i, X_i)}_{Cov(\tilde{Z}_i^*, X_i)} \beta$$

$$\implies \beta = \left(M \cdot Cov(\tilde{Z}_i, X_i)\right)^{-1} M \cdot Cov(\tilde{Z}_i, Y_i)$$

This defines a *class* of IV estmands/estimators, indexed by M

Multiple Instruments

$$\beta = \left(M \cdot Cov(\tilde{Z}_i, X_i)\right)^{-1} M \cdot Cov(\tilde{Z}_i, Y_i)$$

This defines a *class* of IV estmands/estimators, indexed by the $J \times L$ matrix M.

Note that in the just-identified [i.e., $L = J$] case, the choice of M is irrelevant.

I.e., there is only one IV estimand, $\beta^{IV} = Cov(\tilde{Z}_i, X_i)^{-1} Cov(\tilde{Z}_i, Y_i)$.

Why?

Multiple Instruments

$$\beta = \left(M \cdot Cov(\tilde{Z}_i, X_i)\right)^{-1} M \cdot Cov(\tilde{Z}_i, Y_i)$$

This defines a *class* of IV estmands/estimators, indexed by the $J \times L$ matrix M.

Note that in the just-identified [i.e., $L = J$] case, the choice of M is irrelevant.

I.e., there is only one IV estimand, $\beta^{IV} = Cov(\tilde{Z}_i, X_i)^{-1} Cov(\tilde{Z}_i, Y_i)$.

Why? Because in this case, M is a full-column rank $L \times L$ matrix $\Rightarrow$ M is invertible, and therefore for any invertible M,

$$\begin{aligned}\left(M \cdot Cov(\tilde{Z}_i, X_i)\right)^{-1} M \cdot Cov(\tilde{Z}_i, Y_i) &= Cov(\tilde{Z}_i, X_i)^{-1} \cdot M^{-1} \cdot M \cdot Cov(\tilde{Z}_i, Y_i) \\ &= Cov(\tilde{Z}_i, X_i)^{-1} \cdot Cov(\tilde{Z}_i, Y_i).\end{aligned}$$

Two-Stage Least Squares (2SLS)

2SLS sets $M = Cov(\tilde{Z}_i, X_i)'Var(\tilde{Z}_i)^{-1} = \pi$: the $J \times L$ matrix of first-stage coefficients

- this makes $\tilde{Z}_i^* = \pi\tilde{Z}_i$ the (residualized) first-stage fitted values
- intuitively: it combines IVs according to their predictiveness of X_i.

  [We should expect this to decrease the IV standard error by increasing fitted values variation.]

$$\beta^{2SLS} = (\underbrace{\text{Cov}\left(\tilde{Z}_i, X_i\right)' \text{Var}\left(\tilde{Z}_i\right)^{-1}}_{M} \text{Cov}\left(\tilde{Z}_i, X_i\right))^{-1} \underbrace{\text{Cov}\left(\tilde{Z}_i, X_i\right)' \text{Var}\left(\tilde{Z}_i\right)^{-1}}_{M} \text{Cov}\left(\tilde{Z}_i, Y_i\right)$$

Since the first-stage from regressing X_i on this $\tilde{Z}_i^*$ is one (by construction), β^{2SLS} can be obtained in two *stages*:

1. Regress X_i on Z_i and W_i (first stage)
2. Regress Y_i on first-stage fitted values and W_i (second stage)

Are we talking about an estimand or an estimator so far?

Two-Stage Least Squares (2SLS)

2SLS sets $M = Cov(\tilde{Z}_i, X_i)' Var(\tilde{Z}_i)^{-1} = \pi$: the $J \times L$ matrix of first-stage coefficients

- this makes $\tilde{Z}_i^* = \pi \tilde{Z}_i$ the (residualized) first-stage fitted values
- intuitively: it combines IVs according to their predictiveness of X_i.

  [We should expect this to decrease the IV standard error by increasing fitted values variation.]

$$\beta^{2SLS} = (\underbrace{\mathrm{Cov}\left(\tilde{Z}_i, X_i\right)' \mathrm{Var}\left(\tilde{Z}_i\right)^{-1}}_{M} \mathrm{Cov}\left(\tilde{Z}_i, X_i\right))^{-1} \underbrace{\mathrm{Cov}\left(\tilde{Z}_i, X_i\right)' \mathrm{Var}\left(\tilde{Z}_i\right)^{-1}}_{M} \mathrm{Cov}\left(\tilde{Z}_i, Y_i\right)$$

Since the first-stage from regressing X_i on this $\tilde{Z}_i^*$ is one (by construction), β^{2SLS} can be obtained in two *stages*:

1. Regress X_i on Z_i and W_i (first stage)
2. Regress Y_i on first-stage fitted values and W_i (second stage)

Are we talking about an estimand or an estimator so far? $\rightarrow$ the 2SLS estimand!

Two-Stage Least Squares (2SLS)

Are we talking about an estimand or an estimator so far?

Two-Stage Least Squares (2SLS)

Are we talking about an estimand or an estimator so far? $\rightarrow$ the 2SLS estimand!

Yet the exact same logic holds for the 2SLS estimator [i.e., two stages of OLS] as it simply takes the logic to the sample. But do not 2SLS by hand (let softwares do it)! Why? Manual 2SLS Mistakes

$$\hat{\beta}^{2SLS} = [\widehat{\text{Cov}}\left(\tilde{Z}_i, X_i\right)' \widehat{\text{Var}}\left(\tilde{Z}_i\right)^{-1} \widehat{\text{Cov}}\left(\tilde{Z}_i, X_i\right)]^{-1} \widehat{\text{Cov}}\left(\tilde{Z}_i, X_i\right)' \widehat{\text{Var}}\left(\tilde{Z}_i\right)^{-1} \widehat{\text{Cov}}\left(\tilde{Z}_i, Y_i\right)$$

$$= \left(X' P_{\tilde{Z}} X\right)^{-1} X' P_{\tilde{Z}} Y$$ where X and Y stack obs. of X_i' and Y_i

$P_{\tilde{Z}}$ is the sample projection matrix on (sample) residualized $\tilde{Z}_i$. Specifically,

$P_{\tilde{Z}} = \hat{\tilde{Z}}\left(\hat{\tilde{Z}}'\hat{\tilde{Z}}\right)^{-1}\hat{\tilde{Z}}'$ where $\hat{\tilde{Z}}$ stacks residuals from sample proj. of Z_i on controls.

Since $P_{\tilde{Z}}$ is an idempotent and symmetric matrix, we can rewrite

$$\hat{\beta}^{2SLS} = \left(X' P_{\tilde{Z}} X\right)^{-1} X' P_{\tilde{Z}} Y = \left(X' P_{\tilde{Z}} P_{\tilde{Z}} X\right)^{-1} X' P_{\tilde{Z}} Y = \left(X' P_{\tilde{Z}}' P_{\tilde{Z}} X\right)^{-1} X' P_{\tilde{Z}}' Y$$
$$= \left(\left(P_{\tilde{Z}} X\right)' P_{\tilde{Z}} X\right)^{-1} \left(P_{\tilde{Z}} X\right)' Y$$

i.e., formula for OLS reg. of Y_i on $\hat{X}_i = 1^{\text{st}}$-stage fitted values (partialling out controls): 2.S.L.S.!

2SLS Is a Many-Splendored Thing

Another really useful way to understand 2SLS with multiple instruments [and possibly multiple treatments] is as a weighted average of just-identified IVs:

$$\beta^{2SLS}$$

$$= (\underbrace{\mathrm{Cov}\left(\tilde{Z}_i, X_i\right)' \mathrm{Var}\left(\tilde{Z}_i\right)^{-1}}_{\pi} \mathrm{Cov}\left(\tilde{Z}_i, X_i\right))^{-1} \underbrace{\mathrm{Cov}\left(\tilde{Z}_i, X_i\right)' \mathrm{Var}\left(\tilde{Z}_i\right)^{-1}}_{\pi} \mathrm{Cov}\left(\tilde{Z}_i, Y_i\right)$$

$$= \left(\pi \,\mathrm{Var}(\tilde{Z}_i)\, \underbrace{\mathrm{Var}(\tilde{Z}_i)^{-1} Cov(\tilde{Z}_i, X_i')}_{\pi'}\right)^{-1} \pi \,\mathrm{Var}(\tilde{Z}_i)\, \underbrace{\mathrm{Var}(\tilde{Z}_i)^{-1} Cov(\tilde{Z}_i, Y_i)}_{\rho} \text{ as } \mathrm{Var}(\tilde{Z}_i)\,\mathrm{Var}(\tilde{Z}_i)^{-1} = I$$

$$= (\pi \,\mathrm{Var}(\tilde{Z}_i)\pi')^{-1} \pi \,\mathrm{Var}(\tilde{Z}_i)\rho,$$

i.e., formula for a $Var(\tilde{Z}_i)$-weighted reg. of reduced- form coefficients ρ [$L \times 1$ vector] on the (transposed) matrix of 1^{st}-stage coefficients π' [π is $J \times L$, π' is $L \times J$.] [w/o constant] .

2SLS Is a Many-Splendored Thing

$$\underbrace{\beta^{2SLS}}_{J\times 1} = (\underbrace{\pi}_{J\times L}\underbrace{Var(\tilde{Z}_i)}_{L\times L}\pi')^{-1}\pi Var(\tilde{Z}_i)\underbrace{\rho}_{L\times 1}$$

When $J = 1$ [one treatment] and still $L > 1$ [multiple instruments] , this becomes

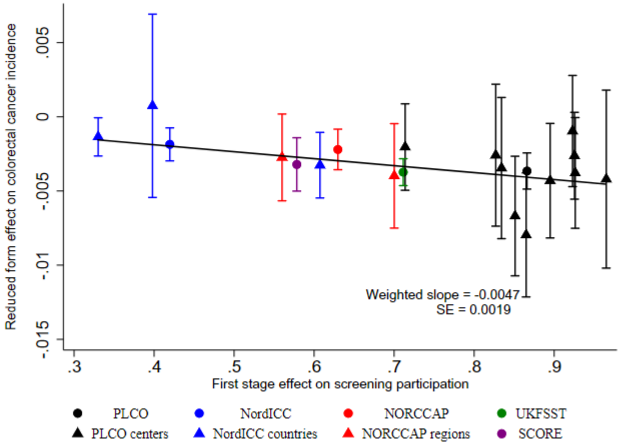
$$\beta^{2SLS} = \sum_{\ell} \omega_\ell \beta_\ell^{IV}$$

where $\omega_\ell = (\pi Var(\tilde{Z}_i)\pi')^{-1}\underbrace{\pi}_{1\times L}\underbrace{Var(\tilde{Z}_i)_{\cdot\ell}}_{L\times 1}\pi_\ell$ and $\beta_\ell^{IV} = \rho_\ell/\pi_\ell$.

[$Var(\tilde{Z}_i)_{\cdot\ell}$ denotes the ℓ^{th} column of $Var(\tilde{Z}_i)$.]

So 2SLS with one endogenous treatment and multiple instruments combines multiple "one-at-a-time" just-identified IVs β_ℓ^{IV}.

Angrist-Hull '23: "Visual IV" for Cancer Screening Trials



Each dot = a (ρ_ℓ, π_ℓ) for a trial ℓ where randomized screening offers Z_i instrument for screening participation D_i. Slope of weighted line-of-best fit through 0 = 2SLS estimate.

Overidentification Tests

Under the constant-effects causal model of $Y_i = X_i'\beta + \varepsilon_i$, overidentification gives a way to test instrument validity

- All just-identified IVs should coincide: i.e. $\beta_\ell^{IV} = \beta$ for all ℓ
- Graphically: the R^2 from visual IV plots should $= 1$ in large samples

Overidentification Tests

Under the constant-effects causal model of $Y_i = X_i'\beta + \varepsilon_i$, overidentification gives a way to test instrument validity

- All just-identified IVs should coincide: i.e. $\beta_\ell^{IV} = \beta$ for all ℓ
- Graphically: the R^2 from visual IV plots should $= 1$ in large samples

Softwares like Stata automatically compute the p-value for this test when $L > J$

- If $p > 0.05$, it means the $\hat{\beta}_\ell^{IV}$'s are all pretty similar to each other

Overidentification Tests

Under the constant-effects causal model of $Y_i = X_i'\beta + \varepsilon_i$, overidentification gives a way to test instrument validity

- All just-identified IVs should coincide: i.e. $\beta_\ell^{IV} = \beta$ for all ℓ
- Graphically: the R^2 from visual IV plots should $= 1$ in large samples

Softwares like Stata automatically compute the p-value for this test when $L > J$

- If $p > 0.05$, it means the $\hat{\beta}_\ell^{IV}$'s are all pretty similar to each other

Don't place too much stock in overidentification tests, however:

- They tend to have low power (b/c individual $\hat{\beta}_\ell^{IV}$ tend to be noisy)
- Rejection doesn't tell us which IV is invalid (they all might be!)
- If they reject, it need not mean the instruments are invalid

  [because of treatment effect heterogeneity $\rightarrow$ we're getting there!]

Weak Instruments `skip`

When running just-identified IV, people worry about instrument "strength"

- Specifically the first stage F-statistic, which tests if $\pi = 0$

Weak Instruments `skip`

When running just-identified IV, people worry about instrument "strength"

- Specifically the first stage F-statistic, which tests if $\pi = 0$

If π is small relative to its standard error, we say the IV is "weak"

- Typically use the rule-of-thumb of $F < 10$ (Staiger and Stock 1997)
- In this case the second-stage SEs will be large and the 2SLS estimate will tend to be biased towards the corresponding OLS

Weak Instruments `skip`

When running just-identified IV, people worry about instrument "strength"

- Specifically the first stage F-statistic, which tests if $\pi = 0$

If π is small relative to its standard error, we say the IV is "weak"

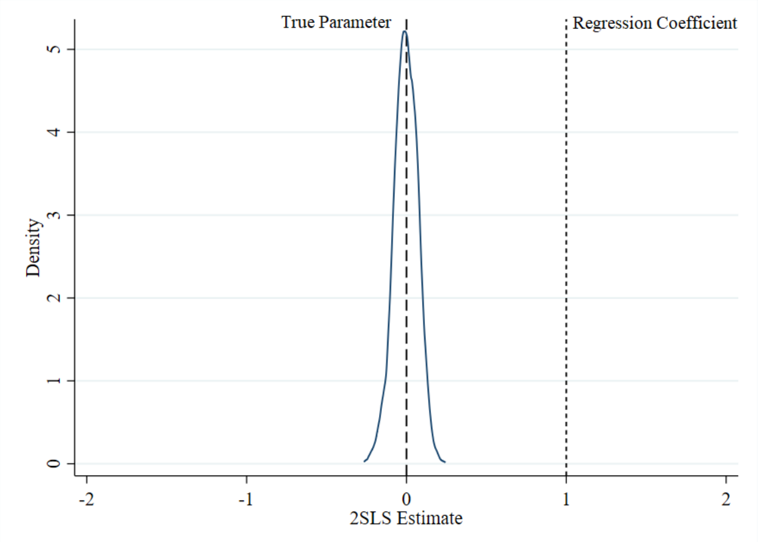
- Typically use the rule-of-thumb of $F < 10$ (Staiger and Stock 1997)
- In this case the second-stage SEs will be large and the 2SLS estimate will tend to be biased towards the corresponding OLS

Much has been made of this over the years, but Angrist and Kolesár (2022) show that we shouldn't worry too much

- The SE increase tends to be large enough to "cover up" the bias, so you're unlikely to reject the null of $\beta = 0$ spuriously

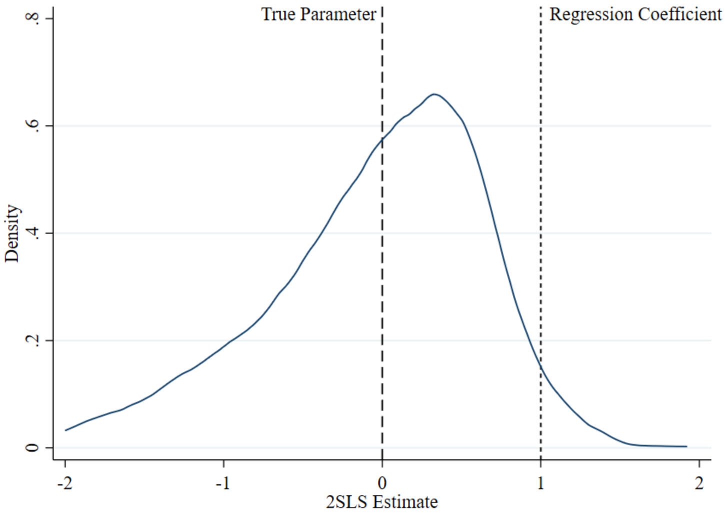
Weak Instruments: Visualized

Monte Carlo: $Y_i = 0 \cdot D_i + \varepsilon_i$, $D_i = \pi Z_i + \eta_i$: $\pi = Var(\varepsilon_i) = Var(\eta_i) = 1$



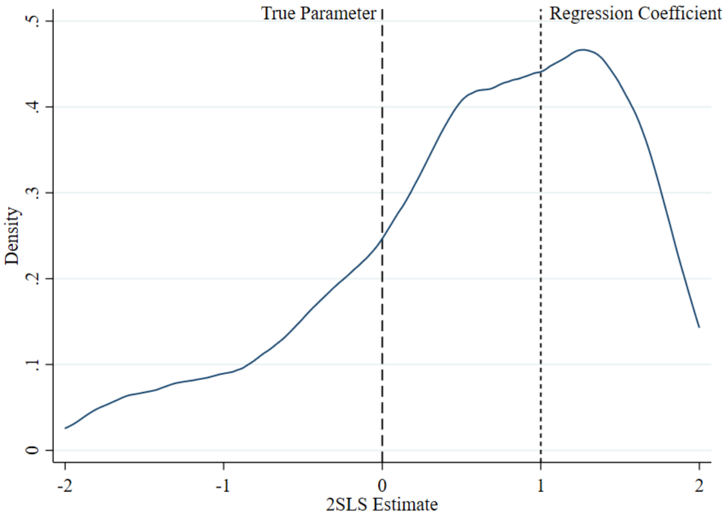
Weak Instruments: Visualized

Monte Carlo: $Y_i = 0 \cdot D_i + \varepsilon_i$, $D_i = \pi Z_i + \eta_i$: $\pi = 0.1$ (Weaker)



Weak Instruments: Visualized

Monte Carlo: $Y_i = 0 \cdot D_i + \varepsilon_i$, $D_i = \pi Z_i + \eta_i$: $\pi = 0.01$ (Very Weak)



Many IVs

A thornier problem is many-weak bias, when overidentified

- This also tends to manifest in low first-stage F's, and also causes 2SLS to be biased towards OLS

Unlike when just-id., however, with many weak IVs the SE's go *down*

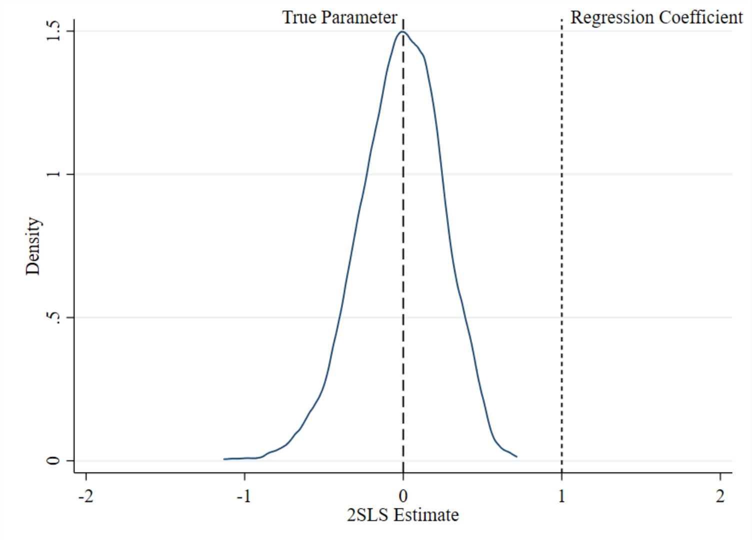
- Intuitively, a more flexible FS tends to fit D_i better $\rightarrow$ more power
- But we can have *overfitting* with lots of instruments, which essentially recreates the (endogenous) variation in D_i

This became a high-profile problem with Angrist-Krueger '91, where the QOB instrument was interacted with many state/year FEs

- These days folks don't make this mistake ... but many-IV bias can be lurking in other settings with constructed instruments (e.g. judge IV)

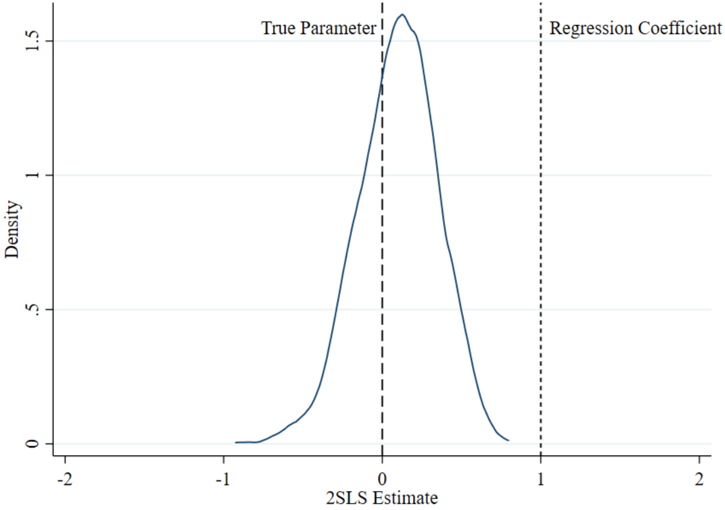
Many Instruments: Visualized

$Y_i = 0 \cdot D_i + \varepsilon_i$, $D_i = \pi Z_{i1} + \sum_{\ell>1} 0 \cdot Z_{i\ell} + \eta_i$: IV w/ Z_{i1} only



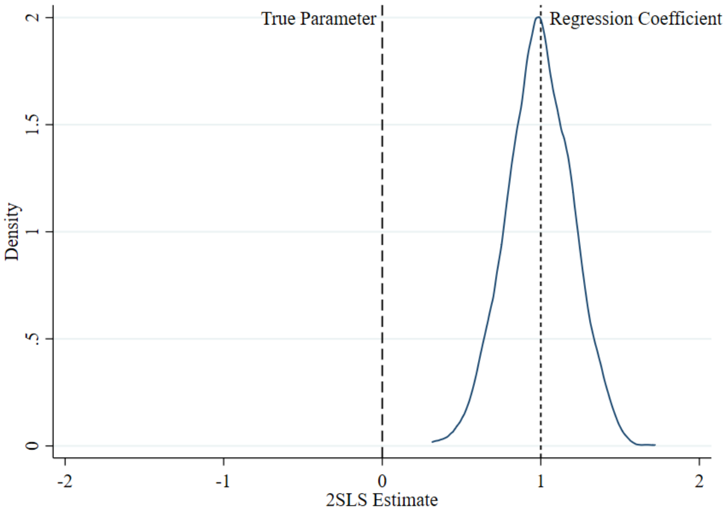
Many Instruments: Visualized

$Y_i = 0 \cdot D_i + \varepsilon_i$, $D_i = \pi Z_{i1} + \sum_{\ell>1} 0 \cdot Z_{i\ell} + \eta_i$: IV w/ $Z_{i1}, \ldots, Z_{i10}$



Many Instruments: Visualized

$Y_i = 0 \cdot D_i + \varepsilon_i$, $D_i = \pi Z_{i1} + \sum_{\ell>1} 0 \cdot Z_{i\ell} + \eta_i$: IV w/ $Z_{i1}, \ldots, Z_{i100}$



What to Do?

Aim for few instruments, and check your F's after every *ivreg*

- State of the art: Montiel Olea and Pflueger '15; `weakivtest` in Stata
- Staiger-Stock rule-of-thumb ($F > 10$) still seems widely held
- Cf. Lee et al. (2020), Keane and Neal (2022) for discussions of additional subtleties

If your F is small, some things to consider:

- Is there a better functional form for your instrument?
- Do interactions with covariates help? (note: beware many-weak!)
- Does changing the covariate set help? (note: beware invalidity!)
- Check results w/a more robust approach (e.g. Anderson-Rubin, JIVE)

Outline

Preliminaries

IV Mechanics

- Just-Identified IV
- Overidentification
- Weak vs. Many-Weak Bias

IV Interpretation

- LATE Fundamentals
- Generalizations and Limitations

What Does IV Identify, Really?

IV was invented for structural economic models (SEMs), typically w/ a single parameter β linearly relating Y_i to X_i. *Yet* modern view of $Y_i = \beta X_i + \varepsilon_i$ is that it describes a causal relationship and imposes a (strong) linear-and-constant-effects restriction.

The Imbens-Angrist 1994 LATE result revolutionized our understanding of IV estimands, and clarified some subtle points around IV identification:

- β^{IV} often identifies a convex average of heterogeneous effects under first-stage monotonicity: Z_i only affects X_i in one direction...
  though there are notable exceptions! [E.g., multiple treatments]
- IV "exogeneity" $[Cov(Z_i, \varepsilon_i) = 0]$ can arise from two conceptually different assumptions of instrument independence and exclusion.

Basic (Binary Treatment, Binary Instrument) Setup

$Y_i(0), Y_i(1)$: denote indiv. i's potential outcomes given a binary treatment $D_i \in \{0, 1\}$.

$\rightarrow$ Observed outcomes: $Y_i = Y_i(0)(1 - D_i) + Y_i(1)D_i = (Y_i(1) - Y_i(0))\, D_i + Y_i(0)$

$D_i(0), D_i(1)$: denote indiv. i's potential treatments given a binary instrument $Z_i \in \{0, 1\}$.

$\rightarrow$ Observed treatment: $D_i = (D_i(1) - D_i(0))\, Z_i + D_i(0)$

Under what assumptions can we causally interpret the regression of Y on D instrumented by Z?

Imbens and Angrist (1994) Assumptions

1. *As-good-as-random assignment*: $Z_i \perp (Y_i(0), Y_i(1), D_i(0), D_i(1))$

   $\rightarrow$ Consider the Angrist draft lottery, or Angrist-Krueger's QoB IV

2. *Exclusion*: Z_i only affects Y_i through its effect on D_i

   $\rightarrow$ Implicit in the potential outcome notation: $Y_i(d)$ is not indexed by Z_i

3. *Relevance*: Z_i is correlated with D_i

   $\rightarrow$ Equivalently, given Assumption 1, $E[D_i(1) - D_i(0)] \neq 0$

4. *Monotonicity*: $D_i(1) \geq D_i(0)$ for all i (i.e., almost-surely)

   $\rightarrow$ The instrument can only shift the treatment in one direction

   I.e., there are only "Always-takers" $D_i(1) = D_i(0) = 1$, "Never-takers" $D_i(1) = D_i(0) = 0$, and "Compliers" $1 = D_i(1) > D_i(0) = 0$ but no "Defiers $0 = D_i(1) < D_i(0) = 1$.

Local Average Treatment Effect (LATE) Identification

Imbens and Angrist (1994) showed that under these assumptions:

$$\beta^{IV} = E[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)]$$

Proof: By independence, $E\left[D_i \mid Z_i = 1\right] - E\left[D_i \mid Z_i = 0\right] = E\left[D_i(1) - D_i(0)\right]$

Similarly,

$$\begin{aligned}
E\left[Y_i \mid Z_i = 1\right] - E\left[Y_i \mid Z_i = 0\right] =& E\left[Y_i(0) + \left(Y_i(1) - Y_i(0)\right) D_i(1) \mid Z_i = 1\right] \\
&- E\left[Y_i(0) + \left(Y_i(1) - Y_i(0)\right) D_i(0) \mid Z_i = 0\right] \\
=& E\left[\left(Y_i(1) - Y_i(0)\right)\left(D_i(1) - D_i(0)\right)\right]
\end{aligned}$$

By monotonicity, $D_i(1) - D_i(0) \in \{0, 1\}$. Thus:

$$\frac{E\left[\left(Y_i(1) - Y_i(0)\right)\left(D_i(1) - D_i(0)\right)\right]}{E\left[D_i(1) - D_i(0)\right]} = E\left[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)\right]$$

Implications

- Potential outcomes notation makes clear independence vs. exclusion
  - $\rightarrow$ A lottery'd Z_i can ensure independence, but exclusion can still fail.

- First-stage monotonicity becomes important under heterogeneous effects
  - $\rightarrow$ Otherwise, β^{IV} identifies a non-convex average: proof

$$\frac{E\left[(Y_i(1)-Y_i(0))\,(D_i(1)-D_i(0))\right]}{E\left[D_i(1)-D_i(0)\right]} = \frac{c}{c-d} E\left[Y_i(1)-Y_i(0) \mid D_i(1) > D_i(0)\right] + \frac{-d}{c-d} \underbrace{E\left[Y_i(1)-Y_i(0) \mid D_i(1) < D_i(0)\right]}_{\text{Avg. effect for defiers}}$$

  - $\rightarrow$ Monotonicity: clearly sensible in some settings [e.g., draft lottery], but can be questionable in others [e.g., judge IVs].

- The LATE result formalizes a key limitation of overidentification tests
  - $\rightarrow$ Two valid IVs can identify different LATEs [*internal* vs. *external* validity].

Multivalued Ordered Treatments ("Variable Treatment Intensity")

Angrist and Imbens (1995) show that when $D_i \in \{0, \ldots, \bar{d}\}$ and a generalization of the LATE assumptions hold, IV identifies an "Average Causal Response" (ACR)

$$\frac{E\left[Y_i \mid Z_i=1\right]-E\left[Y_i \mid Z_i=0\right]}{E\left[D_i \mid Z_i=1\right]-E\left[D_i \mid Z_i=0\right]}=\sum_{d=1}^{\bar{d}} \omega_d E\left[Y_i(d)-Y_i(d-1) \mid D_i(1) \geq d>D_i(0)\right]$$

with $\omega_d=\frac{\operatorname{Pr}\left(D_i(1) \geq d>D_i(0)\right)}{\sum_{d'=1}^{\bar{d}} \operatorname{Pr}\left(D_i(1) \geq d'>D_i(0)\right)}$ convex weights [under monotonicity] .

Averages unit causal responses $E\left[Y_i(d)-Y_i(d-1) \mid D_i(1) \geq d>D_i(0)\right]$

$\rightarrow$ More weight on margins d with more first-stage "action"/"bite"

$\rightarrow$ Note: "Compliers" with $D_i(1)>D_i(0)$ can be double-counted at different margins

$\rightarrow$ Weights are identified by difference in treatment CDF when $Z_i=1$ vs. when $Z_i=0$.

Multivalued Ordered Treatments

Proof

Assumptions:

- Independence and Exclusion: $(Y_i(0), \ldots, Y_i(\bar{d}), D_i(0), D_i(1)) \perp\!\!\!\perp Z_i$
- First-Stage: $E(D_i(1) - D_i(0)) \neq 0$
- Monotonicity: $D_i(1) - D_i(0) \geq 0, \forall i$ (sign wlog).

Let $I(A)$ be the indicator function for the event A. Define the following indicators: $\lambda_{Zd} = I\left(D(Z) \geq d\right)$ for $Z = 0, 1$ and $d = 0, 1, 2, \ldots, \bar{d} + 1$. Note that $\lambda_{Z0} = 1$ and $\lambda_{Z\bar{d}+1} = 0$ for all Z. In terms of the λ_{Zd}, Y can be written as [subscript i dropped to reduce notational burden]

$$\begin{aligned} Y &= Z \cdot Y(D(1)) + (1 - Z) \cdot Y(D(0)) \\ &= \left\{ Z \cdot \sum_{d=0}^{\bar{d}} Y(d) \left(\lambda_{1d} - \lambda_{1d+1}\right) \right\} + \left\{ (1 - Z) \cdot \sum_{d=0}^{\bar{d}} Y(d) \left(\lambda_{0d} - \lambda_{0d+1}\right) \right\}. \end{aligned}$$

as $\lambda_{Zd} - \lambda_{Zd+1} = I\left(D(Z) \geq d\right) - I\left(D(Z) \geq d+1\right) = I\left(d \leq D(Z) < d+1\right)$
and $Y(D(Z)) = \sum_{d=0}^{\bar{d}} Y(d) I\left(d \leq D(Z) < d+1\right)$.

Multivalued Ordered Treatments

Proof

Under independence, $E[Y \mid Z=1] - E[Y \mid Z=0] = E[Y(D(1)) - Y(D(0))]$ which equals

$$E\left\{\sum_{d=0}^{\bar{d}} Y_j \cdot [\lambda_{1d} - \lambda_{1d+1} - \lambda_{0d} + \lambda_{0d+1}]\right\} = E\left\{\sum_{d=1}^{\bar{d}} [(Y_d - Y_{d-1}) \cdot (\lambda_{1d} - \lambda_{0d})] + Y_0 \cdot \underbrace{(\lambda_{10} - \lambda_{00})}_{=0}\right\}$$

$$= E\left\{\sum_{d=1}^{\bar{d}} (Y_d - Y_{d-1}) \cdot (\lambda_{1d} - \lambda_{0d})\right\}$$

because $\lambda_{Z0} = 1$ for $Z = 0, 1$. Note that $\lambda_{1d} \geq \lambda_{0d}$ under monotonicity and that λ_{1d} and λ_{0d} equal 0 or 1 . Therefore, $\lambda_{1d} - \lambda_{0d}$ equals 0 or 1 , and we can write the previous expression as

$$\sum_{d=1}^{\bar{d}} E\left[Y_d - Y_{d-1} \mid \lambda_{1d} - \lambda_{0d} = 1\right] \cdot \Pr\left(\lambda_{1d} - \lambda_{0d} = 1\right)$$

$$= \sum_{d=1}^{\bar{d}} E\left[Y_d - Y_{d-1} \mid D(1) \geq d > D(0)\right] \cdot \Pr\left(D(1) \geq d > D(0)\right)$$

Multivalued Ordered Treatments

Proof

Similarly, for the denominator, $D = Z \cdot D(1) + (1 - Z) \cdot D(0)$ and, because d plays the role played by $Y(d)$ in previous derivations for the numerator,

$$
\begin{aligned}
& E[D \mid Z = 1] - E[D \mid Z = 0] \\
& \quad = E\left\{\sum_{d=0}^{\bar{d}} d \cdot (\lambda_{1d} - \lambda_{1d+1} - \lambda_{0d} + \lambda_{0d+1})\right\} \\
& \quad = E\left\{\sum_{d=1}^{\bar{d}} \underbrace{(d - (d-1))}_{=1} \cdot (\lambda_{1d} - \lambda_{0d})\right\} = \sum_{d=1}^{\bar{d}} \Pr\left(D(1) \geq d > D(0)\right)
\end{aligned}
$$

QED.

Multivalued Ordered Treatments

Weights identification

Notice that each weight $\omega_d = \dfrac{\Pr\left(D_i(1) \geq d > D_i(0)\right)}{\sum_{d'=1}^{\bar{d}} \Pr\left(D_i(1) \geq d' > D_i(0)\right)}$ are identified too.

The denominator is identified (as derived in the previous slide) by $E[D \mid Z = 1] - E[D \mid Z = 0]$.

As for the numerator of ω_d, assuming that D_i is continuously distributed with no mass point,

$$\begin{aligned} P[D_i(1) \geq d > D_i(0)] &= P[D_i(1) \geq d] - P[D_i(0) \geq d] \\ &= P[D_i \geq d \mid Z_i = 1] - P[D_i \geq d \mid Z_i = 0] \quad \text{under independence} \\ &= E[I(D_i \geq d) \mid Z_i = 1] - E[I(D_i \geq d) \mid Z_i = 0] \end{aligned}$$

which corresponds to the coefficient on Z_i in a regression of $I(D_i \geq d)$ on $(1, Z_i)$.

Notice this also equals the gap between the CDF of D_i when $Z_i = 0$ vs. $Z_i = 1$ [guaranteed to be positive under monotonicity]. Indeed,

$$P[D_i(1) \geq d] - P[D_i(0) \geq d] = \underbrace{P[D_i(0) < d]}_{F_{D_i(0)}(d)} - \underbrace{P[D_i(1) < d]}_{F_{D_i(1)}(d)}$$

Continuous Treatments

Angrist, Graddy and Imbens (2000) extend this to a continuous treatment.

- Potential outcome function $q_i(p)$: demand for fish in market i at hypothetical price p
- Treatment variable p_i: price of fish in market i
- Observed outcome is $q_i(p_i)$
- $\frac{\partial}{\partial p} q_i(p) = q_i'(p)$: slope of the demand curve [price elasticity of demand if q_i and p_i are in logs] , a.k.a. a i-specific (marginal) treatment effect of p on q_i [as $q_i'(p) = \frac{q_i(p) - q_i(p-h)}{h}$]
- Instrument: bad weather indicator $stormy_i \in \{0, 1\}$ [= exogenous supply shock].

Assumptions:

- Independence and Exclusion: $(q_i(p), p_i(z)) \perp\!\!\!\perp stormy_i, \forall (p, z) \in Supp(p_i) \times \{0, 1\}$
- First-Stage: $E(p_i(1) - p_i(0)) \neq 0$
- Monotonicity: $p_i(1) - p_i(0) \geq 0, \forall i$ (sign wlog).

Continuous Treatments

Then the 2SLS/Wald estimand identifies

$$\frac{E[q_i \mid stormy_i = 1] - E[q_i \mid stormy_i = 0]}{E[p_i \mid stormy_i = 1] - E[p_i \mid stormy_i = 0]}$$
$$= \frac{\int E[q_i'(t) \mid p_i(1) \geq t > p_i(0)] \cdot P[p_i(1) \geq t > p_i(0)]dt}{\int P[p_i(1) \geq t > p_i(0)]dt}$$

a.k.a. the continuous average causal response formula.

Continuous Treatments

Proof

Let us use $Z_i = stormy_i$ to lighten the notation. Let us also assume that p_i has a lower bound, and (wlog) that this lower bound is 0. Then by the fundamental theorem of calculus,

$$q_i = Z_i q_i(p_i(1)) + (1 - Z_i) q_i(p_i(1)) = Z_i \left[q_i(0) + \int_0^{p_i(1)} q_i'(t)dt \right] + (1 - Z_i) \left[q_i(0) + \int_0^{p_i(0)} q_i'(t)dt \right].$$

By independence + monotonicity [implying $p_i(1) > p_i(0)$], $E[q_i \mid Z_i = 1] - E[q_i \mid Z_i = 0]$ equals

$$E[q_i(p_i(1)) - q_i(p_i(0))] = E\left[q_i(0) + \int_0^{p_i(1)} q_i'(t)dt \right] - E\left[q_i(0) + \int_0^{p_i(0)} q_i'(t)dt \right].$$

Under monotonicity, $p_i(1) > p_i(0)$ hence [assuming all necessary regularity conditions allowing the use of Fubini theorem to interchange the integration and expectation operators]

$$E[q_i \mid Z_i = 1] - E[q_i \mid Z_i = 0] = E\left[\int_{p_i(0)}^{p_i(1)} q_i'(t)dt \right] = E\left[\int_0^{\infty} q_i'(t) I(p_i(1) \geq t > p_i(0))dt \right]$$

$$= \int_0^{\infty} E\left[q_i'(t) \mid p_i(1) \geq t > p_i(0) \right] P[p_i(1) \geq t > p_i(0)]dt$$

Continuous Treatments

Proof

Similarly, for the denominator, $p_i = Z_i p_i(1) + (1 - Z_i)p_i(0)$, and because p plays the role of $q_i(p)$ in previous derivations for the numerator [and $\frac{\partial}{\partial p} = 1$]

$$E[p_i \mid Z_i = 1] - E[p_i \mid Z_i = 0] = E\left[\int_{p_i(0)}^{p_i(1)} \underbrace{p_i'(t)}_{=1}\, dt\right] = E\left[\int_0^{\infty} I(p_i(1) \geq t > p_i(0))dt\right]$$
$$= \int_0^{\infty} P[p_i(1) \geq t > p_i(0)]dt$$

QED.

Continuous Treatments

Building intuition

Two special cases help building intuition about the continuous ACR identification result.

1. Linear causal response function. $q_i(p) = \alpha_{0i} + \alpha_{1i}p$ for random coefficients $(\alpha_{0i}, \alpha_{1i})$.

$$\frac{E[q_i \mid Z_i = 1] - E[q_i \mid Z_i = 0]}{E[p_i \mid Z_i = 1] - E[p_i \mid Z_i = 0]} = \frac{E[q_i(p_i(1)) - q_i(p_i(0))]}{E[p_i(1) - p_i(0)]} \text{ by independence}$$
$$= \frac{E[\alpha_{1i}\,(p_i(1) - p_i(0))]}{E[p_i(1) - p_i(0)]},$$

a weighted average (across markets) of the effect of price on demand with weights proportional to the price change induced by the weather instrument in market i.

Did we use monotonicity at this point?

Continuous Treatments

Building intuition

Two special cases help building intuition about the continuous ACR identification result.

1. Linear causal response function. $q_i(p) = \alpha_{0i} + \alpha_{1i}p$ for random coefficients $(\alpha_{0i}, \alpha_{1i})$.

$$\frac{E[q_i \mid Z_i = 1] - E[q_i \mid Z_i = 0]}{E[p_i \mid Z_i = 1] - E[p_i \mid Z_i = 0]} = \frac{E[q_i(p_i(1)) - q_i(p_i(0))]}{E[p_i(1) - p_i(0)]} \text{ by independence}$$
$$= \frac{E[\alpha_{1i}\,(p_i(1) - p_i(0))]}{E[p_i(1) - p_i(0)]},$$

a weighted average (across markets) of the effect of price on demand with weights proportional to the price change induced by the weather instrument in market i.

Did we use monotonicity at this point? No. But if we want to ensure that the weights $\frac{p_i(1)-p_i(0)}{E[p_i(1)-p_i(0)]}$ are non-negative, we need the monotonicity assumption.

Continuous Treatments

Building intuition

2. Non-linear but homogeneous demand function. $q_i(p) = Q(p) + \eta_i$ where $Q(.)$ is a non-stochastic function and η_i an additive random error [independent from everything] . In this case, $q_i'(p) = Q'(p)$ every day or in every market, and

$$\frac{E[q_i \mid Z_i = 1] - E[q_i \mid Z_i = 0]}{E[p_i \mid Z_i = 1] - E[p_i \mid Z_i = 0]} = \frac{\int (Q'(t) + \overbrace{E[\eta_i \mid p_i(1) \geq t > p_i(0)]}^{=0}) P[p_i(1) \geq t > p_i(0)] dt}{\int P[p_i(1) \geq t > p_i(0)] dt}$$

$$= \int Q'(t)\omega(t)dt, \quad \text{where } \omega(t) \equiv \frac{P[p_i(1) \geq t > p_i(0)]}{\int P[p_i(1) \geq r > p_i(0)] dr}$$

a weighted average *along* the length of the causal response function [assumed here common to all markets] , placing more weight on derivatives at prices where the instrument shifts CDF of prices most sharply. [Indeed, if $(p_i(1), p_i(0))$ are continuously distributed w/o point mass, $P[p_i(1) \geq t > p_i(0)] = F_{p_i(0)}(t) - F_{p_i(1)}(t) \geq 0$ under monotonicity.]

References I

Regression coefficient on a constant and binary variable

Recall that in a univariate regression of Y on $(1, Z)$, the coef. on Z is given by $\dfrac{Cov(Y,Z)}{Var(Z)}$ [while the coef. on the constant is given by $E(Y) - \dfrac{Cov(Y,Z)}{Var(Z)}E(X)$].

Then one can observe that

$$
\begin{aligned}
Cov(Y,Z) &= E(YZ) - E(Y)E(Z) \\
&= E(Y \mid Z=1)P(Z=1) \\
&\quad - \left[E(Y \mid Z=1)P(Z=1) + E(Y \mid Z=0)(1-P(Z=1))\right] P(Z=1) \\
&= \underbrace{P(Z=1)(1-P(Z=1))}_{=Var(Z)} \left[E(Y \mid Z=1) - E(Y \mid Z=0)\right]
\end{aligned}
$$

back

Avoid Manual 2SLS

Although easy, you should never literally run 2SLS in two stages. Cf. Mostly Harmless Econometrics (MHE) 4.6 for details, but here is summary/refresher:

1. Point estimates will be right, but s.e. generally won't be
2. Risk of omitting some controls from second stage in first stage (or vice versa)
3. Risk of "Forbidden regressions": e.g. regressing Y_i on probit/logit fits for X_i
4. Risk of regressing on $\hat{X}_i$ and $\hat{X}_i^2$, instead of instrumenting X_i^2 directly [e.g., by adding Z_i^2 as extra instrument]

back

Proof: LATE Identification Without Monotonicity

$$
\begin{aligned}
&E\left[(Y_i(1)-Y_i(0))\left(D_i(1)-D_i(0)\right)\right] \\
&= \; E\left[(Y_i(1)-Y_i(0)) \mid D_i(1)>D_i(0)\right] \cdot P[D_i(1)>D_i(0)] \\
&\quad - E\left[(Y_i(1)-Y_i(0)) \mid D_i(1)<D_i(0)\right] \cdot P[D_i(1)<D_i(0)]
\end{aligned}
$$

and $E\left[(D_i(1)-D_i(0))\right]=1 \cdot P(D_i(1)>D_i(0))-1 \cdot P(D_i(1)<D_i(0))$.

Using the notation $P(D_i(1)>D_i(0)) \equiv c$ and $P(D_i(1)<D_i(0)) \equiv d$ and taking the ratio of these two quantities yields the result.

back

Covariance Property

Claim: Let X be a random variable and $f, g : \mathbb{R} \to \mathbb{R}$ be increasing functions. Then $\mathrm{Cov}(f(X), g(X)) \geq 0$.

Proof: Since f and g are increasing, then $(f(x) - f(y))(g(x) - g(y)) \geq 0$ for all $x, y \in \mathbb{R}$.
Assume X, Y are independent and identically distributed. By the above observation and the monotonicity of expectations, we get $E[(f(X) - f(Y))(g(X) - g(Y))] \geq 0$.
Expanding, we get $E[f(X)g(X)] - E[f(X)g(Y)] - E[f(Y)g(X)] + E[f(Y)g(Y)] \geq 0$.
Now due to independence, the LHS becomes

$$E[f(X)g(X)] - E[f(X)]E[g(Y)] - E[f(Y)]E[g(X)] + E[f(Y)g(Y)]$$

Then due to identically distributed, the LHS further becomes $2E[f(X)g(X)] - 2E[f(X)]E[g(X)]$.
So together we have $2\,\mathrm{Cov}(f(X), g(X)) \geq 0$. **QED**.

back