Advanced Instrumental Variables

Peter Hull

Northwestern Causal Inference Workshop

August 2024

Roadmap

- IV Mechanics
  - Just-Identified IV
  - Overidentification
  - Weak vs. Many-Weak Bias
- IV Interpretation
  - LATE and Generalizations
  - Characterizing Compliers
  - Diff-in-Diff and IV
- Formula Instruments
  - Shift-Share IV
  - Recentered IV

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
  - → 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

Preliminaries: Parameters, Estimands, and Estimators

Three distinct objects, not always clearly distinguished:

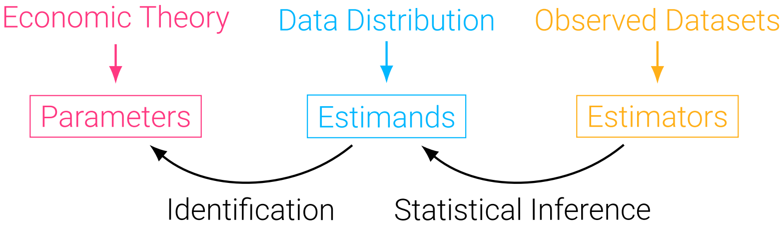
- Parameters come from economic (or other) models of the world
  - → E.g. a "structural" model of supply and demand, or a potential outcome model relating schooling to earnings
  - → They set the target for empirical analyses: what we want to know
- Estimands are functions of the population data distribution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

Preliminaries: Parameters, Estimands, and Estimators

Three distinct objects, not always clearly distinguished:

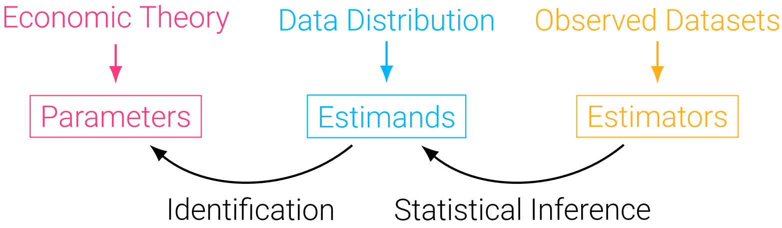
- Parameters come from economic (or other) models of the world
  - → E.g. a "structural" model of supply and demand, or a potential outcome model relating schooling to earnings
  - → They set the target for empirical analyses: what we want to know
- Estimands are functions of the population data distribution
  - → 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 Lay of the Land



Separating out the very different tasks of identification and inference can help make our lives easier

The Lay of the Land



Separating out the very different tasks of identification and inference can help make our lives easier

- Today we'll start with the mechanics of IV estimands, talk briefly about estimation, then spend some time talking about identification

An Example

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

An Example

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

- Parameter: returns to enrolling in this workshop 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 workshop enrollment and ε_i is i's outcome without the workshop

An Example

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

- Parameter: returns to enrolling in this workshop 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 workshop enrollment and ε_i is i's outcome without the workshop

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 workshops will likely boost later-life productivity

- Parameter: returns to enrolling in this workshop 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 workshop enrollment and ε_i is i's outcome without the workshop

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 (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 (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 (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 (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:

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

Population Regression and Endogeneity

In large samples ($N \to \infty$), the OLS *estimator* $\widehat{\beta}^{OLS}$ gets arbitrarily close to (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 (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 $Cov(\varepsilon_i, D_i) = 0$

Population Regression and Endogeneity

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

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

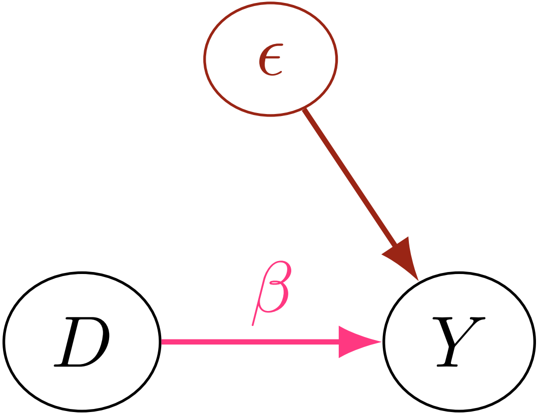
By definition, $\beta^{OLS} = \frac{Cov(Y_i, D_i)}{Var(D_i)}$. Plugging in our causal model:

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

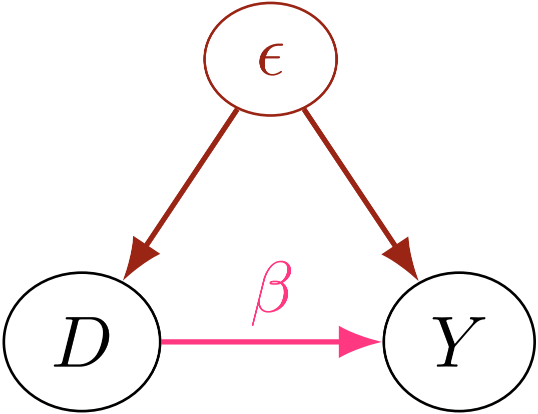
Thus, we have (regression) identification if and only if $Cov(\varepsilon_i, D_i) = 0$

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

Regression "Exogeneity"



Regression "Endogeneity"



The IV Solution

Suppose this workshop 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 workshop

The IV Solution

Suppose this workshop 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 workshop
- Randomness + no direct effects of Z_i on Y_i implies $Cov(Z_i, \varepsilon_i) = 0$

The IV Solution

Suppose this workshop 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 workshop
- 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:

$$Cov(Z_i, Y_i - \beta D_i) = 0$$

The IV Solution

Suppose this workshop 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 workshop
- 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:

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

The IV Solution

Suppose this workshop 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 workshop
- 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 workshop 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 workshop
- 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 $\widehat{\beta}^{IV} = \frac{\widehat{Cov}(Z_i, Y_i)}{\widehat{Cov}(Z_i, D_i)}$

The IV Solution

Suppose this workshop 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 workshop
- Randomness + no direct effects of Z_i on Y_i implies $Cov(Z_i, \varepsilon_i) = 0$

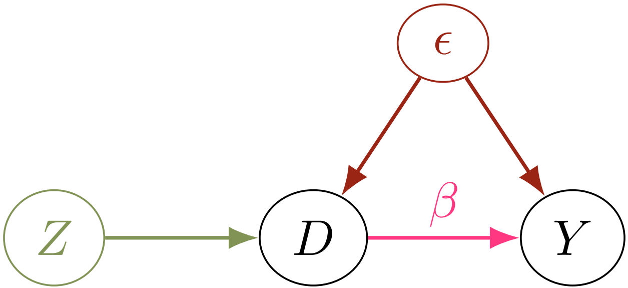
Plugging in the model for $\varepsilon_i = Y_i - \beta D_i$, we now have IV identification:

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

so long as $Cov(Z_i, D_i) \neq 0$. We can estimate this by $\widehat{\beta}^{IV} = \frac{\widehat{Cov}(Z_i, Y_i)}{\widehat{Cov}(Z_i, D_i)}$

- 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; IV feels more mysterious

- But "one weird trick" links the IV estimand back to regression:

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

Reduced Form and First Stage

We're usually pretty comfortable w/regression; IV feels more mysterious

- But "one weird trick" links the IV estimand back to regression:

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

Reduced Form and First Stage

We're usually pretty comfortable w/regression; IV feels more mysterious

- But "one weird trick" links the IV estimand back to regression:

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

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 famously used Vietnam-era draft eligibility as an instrument to estimate the earnings effects of military service

- Let $Z_i \in \{0,1\}$ be an indicator for draft eligibility, $D_i \in \{0,1\}$ be an indicator for military service, and Y_i measure later-life earnings

Angrist (1990): Draft Lottery IV

Angrist famously used Vietnam-era draft eligibility as an instrument to estimate the earnings effects of military service

- Let $Z_i \in \{0, 1\}$ be an indicator for draft eligibility, $D_i \in \{0, 1\}$ be an indicator for military service, and Y_i measure later-life earnings

Here $\beta^{IV} = \frac{Cov(Z_i, Y_i)/Var(Z_i)}{Cov(Z_i, D_i)/Var(Z_i)} = \frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$ has a special name, because Z_i is binary: the Wald estimand

Angrist (1990): Draft Lottery IV

Angrist famously used Vietnam-era draft eligibility as an instrument to estimate the earnings effects of military service

- Let $Z_i \in \{0, 1\}$ be an indicator for draft eligibility, $D_i \in \{0, 1\}$ be an indicator for military service, and Y_i measure later-life earnings

Here $\beta^{IV} = \frac{Cov(Z_i, Y_i)/Var(Z_i)}{Cov(Z_i, D_i)/Var(Z_i)} = \frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$ has a special name, because Z_i is binary: the Wald estimand

- First stage $E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]$: effect of eligibility on the *probability* of military service (b/c D_i is binary)
- Reduced form $E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]$: effect of eligibility on adult earnings (measured in 1971, 1981...)

Angrist (1990): Draft Lottery IV

Angrist famously used Vietnam-era draft eligibility as an instrument to estimate the earnings effects of military service

- Let $Z_i \in \{0,1\}$ be an indicator for draft eligibility, $D_i \in \{0,1\}$ be an indicator for military service, and Y_i measure later-life earnings

Here $\beta^{IV} = \frac{Cov(Z_i,Y_i)/Var(Z_i)}{Cov(Z_i,D_i)/Var(Z_i)} = \frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$ has a special name, because Z_i is binary: the Wald estimand

- First stage $E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]$: effect of eligibility on the *probability* of military service (b/c D_i is binary)
- Reduced form $E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]$: effect of eligibility on adult earnings (measured in 1971, 1981...)

IV interprets the latter causal effect in terms of the former

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

Earnings year	Earnings: Mean	Earnings: Eligibility Effect	Veteran Status: Mean	Veteran Status: Eligibility Effect	Wald Estimate of Veteran 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 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 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 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 $dim(Z_i) = dim(X_i)$ (“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 $dim(Z_i) = dim(X_i)$ ("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 $Cov(Z_i, X_i)$ 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 $dim(Z_i) = dim(X_i)$ ("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 $Cov(Z_i, X_i)$ is full-rank. Equivalently, $\beta^{IV} = \pi^{-1}\rho$ where:

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

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

with estimation working just as you'd think

Multiple Instruments

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

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

- We can drop any $L - J$ instruments to identify β

Multiple Instruments

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

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

- We can drop any $L - J$ instruments to identify β
- More generally, we can take any full-rank linear combination $\tilde{Z}_i^* = M\tilde{Z}_i$ for $J \times L$ matrix M such that $Cov(\tilde{Z}_i^*, X_i)$ is invertible

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

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

Multiple Instruments

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

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

- We can drop any $L - J$ instruments to identify β
- More generally, we can take any full-rank linear combination $\tilde{Z}_i^* = M\tilde{Z}_i$ for $J \times L$ matrix M 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

Two-Stage Least Squares (2SLS)

2SLS sets $M' = \pi$: the matrix of first-stage coefficients

- This makes $\tilde{Z}_i^* = \tilde{Z}_i'\pi$ the (residualized) first-stage fitted values
- Intuitively: combine IVs according to their predictiveness of X_i; we should expect this to decrease the IV standard error

Two-Stage Least Squares (2SLS)

2SLS sets $M' = \pi$: the matrix of first-stage coefficients

- This makes $\tilde{Z}_i^* = \tilde{Z}_i'\pi$ the (residualized) first-stage fitted values
- Intuitively: combine IVs according to their predictiveness of X_i; we should expect this to decrease the IV standard error

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)

Two-Stage Least Squares (2SLS)

2SLS sets $M' = \pi$: the matrix of first-stage coefficients

- This makes $\tilde{Z}_i^* = \tilde{Z}_i'\pi$ the (residualized) first-stage fitted values
- Intuitively: combine IVs according to their predictiveness of X_i; we should expect this to decrease the IV standard error

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)

Note: here we're talking about the 2SLS *estimand*, but the exact same logic holds for the 2SLS *estimator* (i.e. two stages of OLS)

- That said, you should never run 2SLS by hand. Let *ivreg2* do it!

Angrist-Krueger '91: The Power of Overidentification

	OLS		2SLS				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Years of education	.071 (.0004)	.067 (.0004)	.102 (.024)	.13 (.020)	.104 (.026)	.108 (.020)	.087 (.016)
Covariates							
9 year-of-birth dummies		✓			✓	✓	✓
50 state-of-birth dummies		✓			✓	✓	✓
Instruments							
dummy for QOB = 1			✓	✓	✓	✓	✓
dummy for QOB = 2				✓		✓	✓
dummy for QOB = 3				✓		✓	✓
QOB dummies interacted with year-of-birth dummies (30 instruments total)							✓

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the Angrist and Krueger (1991) 1980 census sample. This sample includes native-born men, born 1930–39, with positive earnings and nonallocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses. QOB denotes quarter of birth.

2SLS is a Many-Splendored Thing

There is another (I think more useful) way to understand 2SLS: as a weighted average of just-identified IVs:

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

This is the formula for a $Var(\tilde{Z}_i)$-weighted regression of reduced- form coefficients ρ on first-stage coefficients π (with no constant)

2SLS is a Many-Splendored Thing

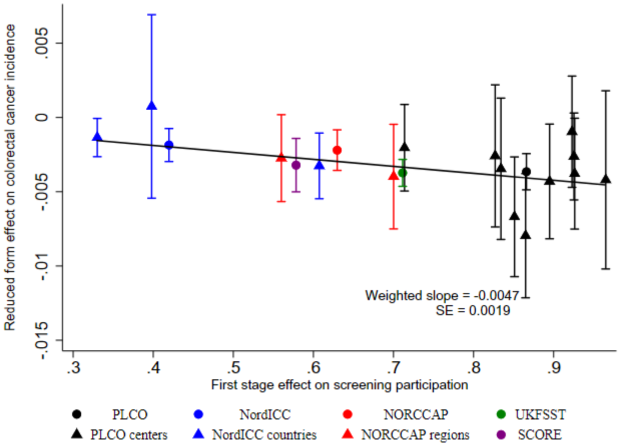
There is another (I think more useful) way to understand 2SLS: as a weighted average of just-identified IVs:

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

This is the formula for a $Var(\tilde{Z}_i)$-weighted regression of reduced- form coefficients ρ on first-stage coefficients π (with no constant)

- When $J = 1$ (one treatment), this becomes $\beta^{2SLS} = \sum_\ell \omega_\ell \beta_\ell^{IV}$ where $\omega_\ell = (\pi' Var(\tilde{Z}_i)\pi)^{-1} \pi' Var(\tilde{Z}_i)'_\ell \pi_\ell$ and $\beta_\ell^{IV} = \rho_\ell / \pi_\ell$
- Intuitively: 2SLS combines multiple “one-at-a-time” IVs β_ℓ^{IV}

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



Each dot gives a (ρ_ℓ, π_ℓ) for a trial ℓ where randomized screening offers Z_i instrument for screening participation D_i

- Slope of the weighted line-of-best fit through zero = 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

Stata will automatically give you the p-value for this test when $L > J$

- If $p > 0.05$, it means your $\hat{\beta}_\ell^{IV}$'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

Stata will automatically give you the p-value for this test when $L > J$

- If $p > 0.05$, it means your $\hat{\beta}_\ell^{IV}$'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)
- If they reject, it need not mean your instruments are invalid (b/c of treatment effect heterogeneity – more on this soon)
- Rejection doesn't tell you which IV is invalid (they all might be!)

Weak Instruments

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

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

Weak Instruments

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

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"

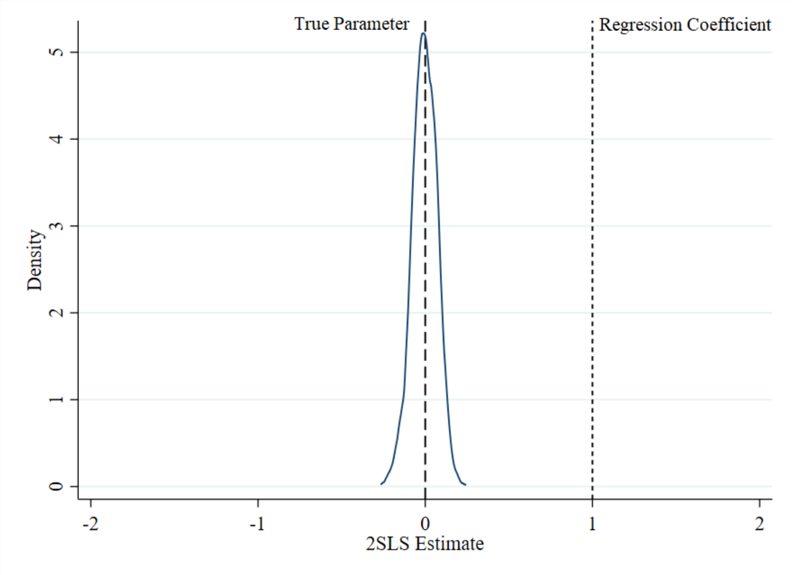
- Typically use the rule-of-thumb of $F < 10$ (Staiger and Stock 1997)
- In this case the second-stage SEs will be large and the 2SLS estimate will tend to be biased towards the corresponding OLS

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

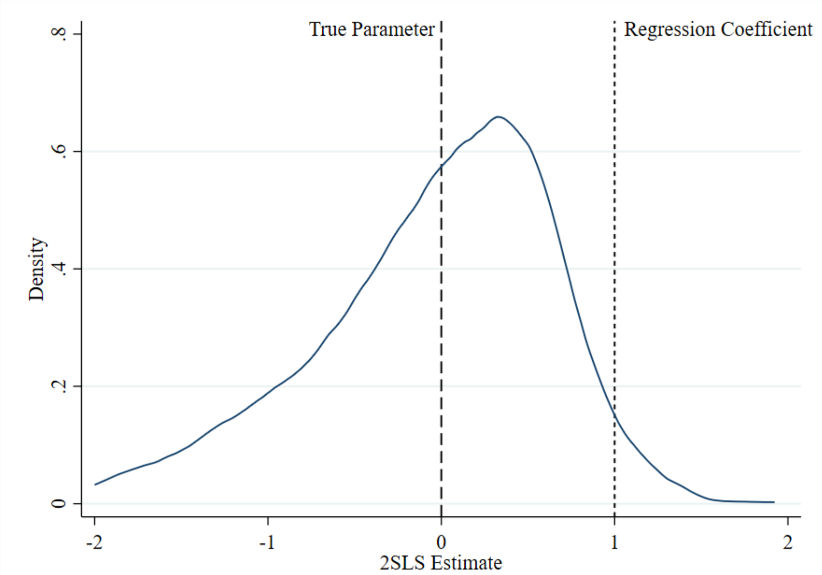
Weak Instruments: Visualized

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



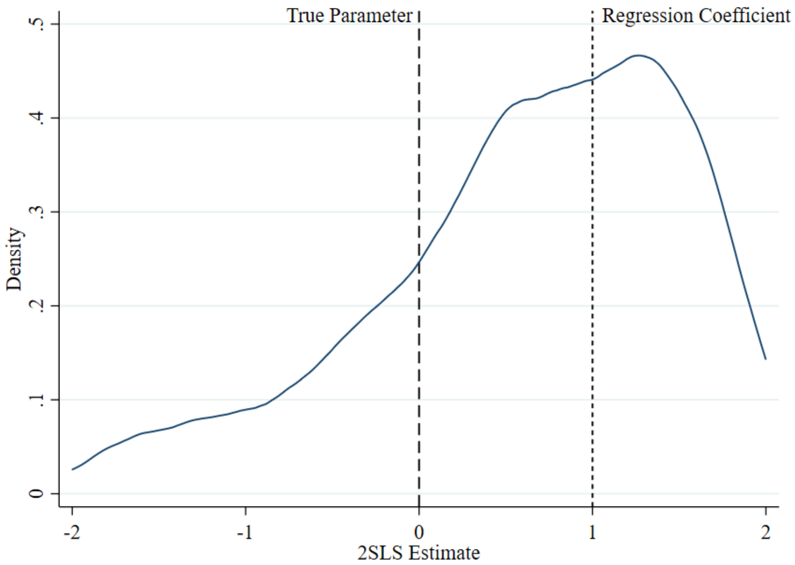
Weak Instruments: Visualized

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

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

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*

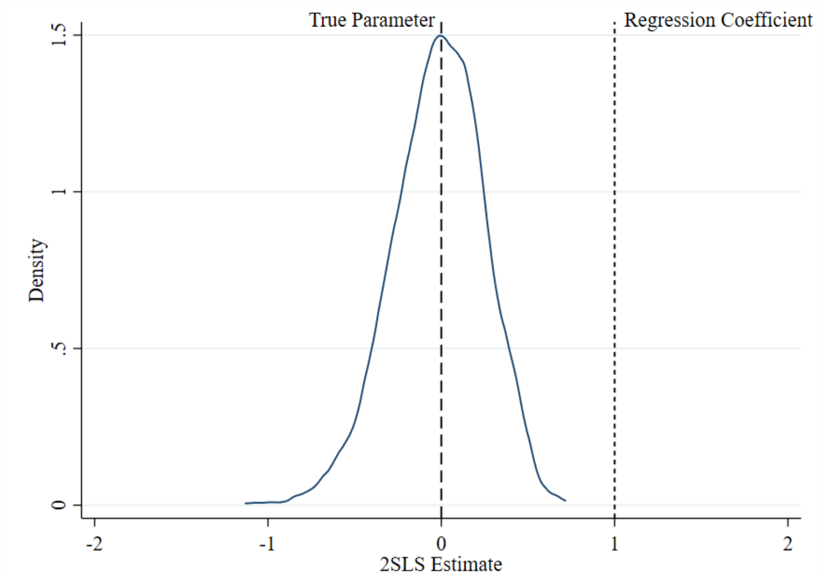
- Intuitively, a more flexible FS tends to fit D_i better $\rightarrow$ more power
- But we can have *overfitting* with lots of instruments, which essentially recreates the (endogenous) variation in D_i

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)

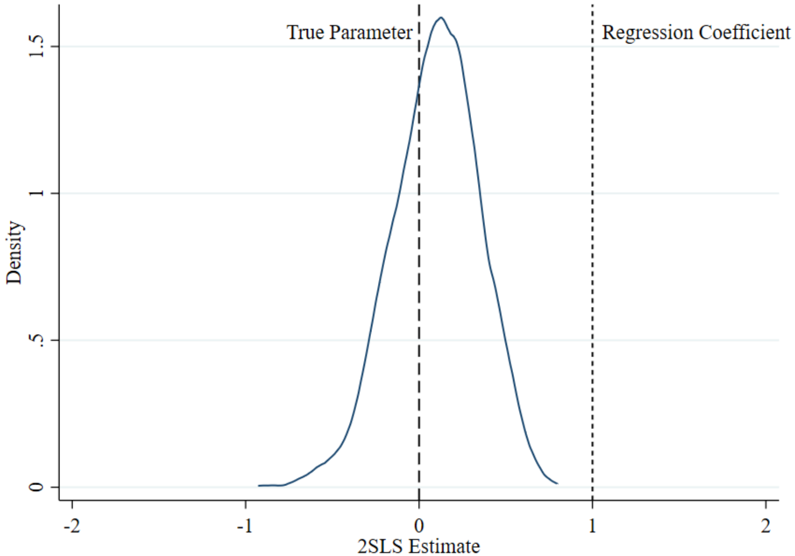
Many Instruments: Visualized

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



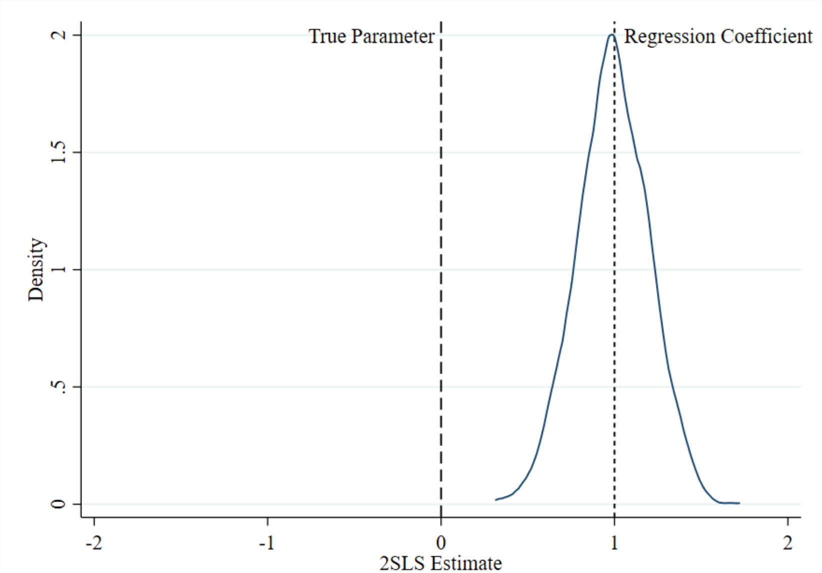
Many Instruments: Visualized

$Y_i = 0 \cdot D_i + \varepsilon_i$, $D_i = \pi Z_{i1} + \sum_{\ell>1} 0 \cdot Z_{i\ell} + \eta_i$: 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
- See Lee et al. (2020) and Keane and Neal (2022) for some discussions of additional subtleties

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
- See Lee et al. (2020) and Keane and Neal (2022) for some 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)

Roadmap

IV Mechanics

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

IV Interpretation

- LATE and Generalizations
- Characterizing Compliers
- Diff-in-Diff and IV

Formula Instruments

- Shift-Share IV
- Recentered IV

What Does IV Identify, Really?

IV was invented in the context of structural economic models, typically with a single parameter β linearly relating Y_i to X_i

- These days we understand $Y_i = \beta X_i + \varepsilon_i$ as describing a *causal* relationship, imposing a (strong) constant-effects restriction

What Does IV Identify, Really?

IV was invented in the context of structural economic models, typically with a single parameter β linearly relating Y_i to X_i

- These days we understand $Y_i = \beta X_i + \varepsilon_i$ as describing a *causal* relationship, imposing a (strong) constant-effects restriction

The Imbens-Angrist 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
- IV "exogeneity" can arise from two conceptually different assumptions of instrument *independence* and *exclusion*

Basic (Binary Treatment, Binary Instrument) Setup

Let $Y_i(0)$ and $Y_i(1)$ denote individual i's potential outcomes given a binary treatment $D_i \in \{0, 1\}$

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

Basic (Binary Treatment, Binary Instrument) Setup

Let $Y_i(0)$ and $Y_i(1)$ denote individual i's potential outcomes given a binary treatment $D_i \in \{0, 1\}$

- Observed outcomes: $Y_i = (Y_i(1) - Y_i(0))\, D_i + Y_i(0) \equiv \beta_i D_i + \varepsilon_i$

Basic (Binary Treatment, Binary Instrument) Setup

Let $Y_i(0)$ and $Y_i(1)$ denote individual i's potential outcomes given a binary treatment $D_i \in \{0, 1\}$

- Observed outcomes: $Y_i = (Y_i(1) - Y_i(0))\, D_i + Y_i(0) \equiv \beta_i D_i + \varepsilon_i$

Let $D_i(0)$ and $D_i(1)$ denote individual i's potential treatment given a binary instrument $Z_i \in \{0, 1\}$

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

Basic (Binary Treatment, Binary Instrument) Setup

Let $Y_i(0)$ and $Y_i(1)$ denote individual i's potential outcomes given a binary treatment $D_i \in \{0, 1\}$

- Observed outcomes: $Y_i = (Y_i(1) - Y_i(0))\, D_i + Y_i(0) \equiv \beta_i D_i + \varepsilon_i$

Let $D_i(0)$ and $D_i(1)$ denote individual i's potential treatment given a binary instrument $Z_i \in \{0, 1\}$

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

Under what assumptions can we causally interpret `ivreg2 Y (D=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

Imbens and Angrist (1994) Assumptions

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

   → 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

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

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, $E[D_i(1) - D_i(0)] \neq 0$

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

Local Average Treatment Effect (LATE) Identification

Imbens and Angrist showed that under these assumptions:

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

The IV estimand β^{IV} identifies a LATE: the average treatment effect $Y_i(1) - Y_i(0)$ among *compliers*: those with $1 = D_i(1) > D_i(0) = 0$

Local Average Treatment Effect (LATE) Identification

Imbens and Angrist showed that under these assumptions:

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

The IV estimand β^{IV} identifies a LATE: the average treatment effect $Y_i(1) - Y_i(0)$ among *compliers*: those with $1 = D_i(1) > D_i(0) = 0$

- Intuitively, IV can't tell us anything about the treatment effects of *never-takers* $D_i(1) = D_i(0) = 0$ or *always-takers* $D_i(1) = D_i(0) = 1$

Local Average Treatment Effect (LATE) Identification

Imbens and Angrist showed that under these assumptions:

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

The IV estimand β^{IV} identifies a LATE: the average treatment effect $Y_i(1) - Y_i(0)$ among *compliers*: those with $1 = D_i(1) > D_i(0) = 0$

- Intuitively, IV can't tell us anything about the treatment effects of *never-takers* $D_i(1) = D_i(0) = 0$ or *always-takers* $D_i(1) = D_i(0) = 1$
- Monotonicity rules out the presence of *defiers*, with $D_i(1) < D_i(0)$

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

Earnings year	Earnings: Mean	Earnings: Eligibility Effect	Veteran Status: Mean	Veteran Status: Eligibility Effect	Wald Estimate of Veteran 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.

What Does This Mean *Practically*?

Two conceptually distinct considerations: *internal* vs. *external* validity

- Context of an IV, and who the compliers likely are, may matter
- Usual “overidentification” test logic fails: two valid IVs may have different estimands

What Does This Mean *Practically*?

Two conceptually distinct considerations: *internal* vs. *external* validity

- Context of an IV, and who the compliers likely are, may matter
- Usual "overidentification" test logic fails: two valid IVs may have different estimands

In addition to as-good-as-random assignment / exclusion, we may need to worry about monotonicity when we do IV

- Sensible in many settings (e.g. draft lottery / cancer screening RCTs)
- Maybe questionable in judge IVs (see Eric's earlier slides)

What Does This Mean *Practically*?

Two conceptually distinct considerations: *internal* vs. *external* validity

- Context of an IV, and who the compliers likely are, may matter
- Usual "overidentification" test logic fails: two valid IVs may have different estimands

In addition to as-good-as-random assignment / exclusion, we may need to worry about monotonicity when we do IV

- Sensible in many settings (e.g. draft lottery / cancer screening RCTs)
- Maybe questionable in judge IVs (see Eric's earlier slides)

Note: technically, we don't need monotonicity if effects are homogenous

Generalizations and Limitations

The core logic of IA'94 extends to multivalued treatments/instruments

- IV identifies an avg. of incremental treatment effects, putting more weight on margins where the instrument shifts the treatment more
- Regressions of $\mathbf{1}[X_i \geq x]$ on Z_i identify the weights

Generalizations and Limitations

The core logic of IA'94 extends to multivalued treatments/instruments

- IV identifies an avg. of incremental treatment effects, putting more weight on margins where the instrument shifts the treatment more
- Regressions of $\mathbf{1}[X_i \geq x]$ on Z_i identify the weights

The result does *not* easily extend, however, to multiple treatments

- Contamination bias: IV coefficient on X_{ij} generally incorporates effects from other treatments X_{ik}, $k \neq j$

Generalizations and Limitations

The core logic of IA'94 extends to multivalued treatments/instruments

- IV identifies an avg. of incremental treatment effects, putting more weight on margins where the instrument shifts the treatment more
- Regressions of $\mathbf{1}[X_i \geq x]$ on Z_i identify the weights

The result does *not* easily extend, however, to multiple treatments

- Contamination bias: IV coefficient on X_{ij} generally incorporates effects from other treatments $X_{ik}, k \neq j$

Since 2SLS is a weighted average of "one-at-a-time" IVs, overidentified single-treatment specifications can have a LATE interpretation

- Need all individual IVs to have a LATE interpretation and the 2SLS weights to be convex (the latter can be checked empirically)

Generalizations and Limitations (Cont.)

We can also have a LATE interpretation in IV specifications w/controls

- E.g. if Z_i is only as-good-as-randomly assigned conditional on W_i

Generalizations and Limitations (Cont.)

We can also have a LATE interpretation in IV specifications w/controls

- E.g. if Z_i is only as-good-as-randomly assigned conditional on W_i

Key condition: the controls are flexible enough to make $E[Z_i \mid W_i]$ linear

- E.g. W_i is a set of mutually exclusive strata dummies
- IV then identifies a weighted-average of conditional-on-W_i IVs

Generalizations and Limitations (Cont.)

We can also have a LATE interpretation in IV specifications w/controls

- E.g. if Z_i is only as-good-as-randomly assigned conditional on W_i

Key condition: the controls are flexible enough to make $E[Z_i \mid W_i]$ linear

- E.g. W_i is a set of mutually exclusive strata dummies
- IV then identifies a weighted-average of conditional-on-W_i IVs

If $E[Z_i \mid W_i]$ is nonlinear, IV may be biased (even w/constant effects)

- Key practical takeaway: it's good to check sensitivity to how controls are parameterized (add interactions, higher-order polynomials, etc.)

Who Are the Compliers?

Characterizing the i that make up the IV estimand (w/$D_i(1) > D_i(0)$) is key for understanding internal vs. external validity

- Unfortunately we can't identify compliers directly: we only observe $D_i(1)$ (when $Z_i = 1$) or $D_i(0)$ (when $Z_i = 0$), not both together!

Who Are the Compliers?

Characterizing the i that make up the IV estimand (w/$D_i(1) > D_i(0)$) is key for understanding internal vs. external validity

- Unfortunately we can't identify compliers directly: we only observe $D_i(1)$ (when $Z_i = 1$) or $D_i(0)$ (when $Z_i = 0$), not both together!

As it turns out, we can still characterize compliers by their outcomes ($Y_i(0)$ and $Y_i(1)$) and by other observables X_i

- Comparing $E[X_i \mid D_i(1) > D_i(0)]$ to $E[X_i]$ can maybe shed light on how $E[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)]$ compares to $E[Y_i(1) - Y_i(0)]$

Outcomes

Computing $E[Y_i(1) \mid D_i(1) > D_i(0)]$ is surprisingly easy in the IA setup

- Define $W_i = Y_i D_i$, and note that this new outcome has potentials with respect to D_i of $W_i(1) = Y_i(1)$ and $W_i(0) = 0$

Outcomes

Computing $E[Y_i(1) \mid D_i(1) > D_i(0)]$ is surprisingly easy in the IA setup

- Define $W_i = Y_i D_i$, and note that this new outcome has potentials with respect to D_i of $W_i(1) = Y_i(1)$ and $W_i(0) = 0$
- Thus IV with W_i as the outcome identifies $E[W_i(1) - W_i(0) \mid D_i(1) > D_i(0)] = E[Y_i(1) \mid D_i(1) > D_i(0)]$

Outcomes

Computing $E[Y_i(1) \mid D_i(1) > D_i(0)]$ is surprisingly easy in the IA setup

- Define $W_i = Y_i D_i$, and note that this new outcome has potentials with respect to D_i of $W_i(1) = Y_i(1)$ and $W_i(0) = 0$
- Thus IV with W_i as the outcome identifies $E[W_i(1) - W_i(0) \mid D_i(1) > D_i(0)] = E[Y_i(1) \mid D_i(1) > D_i(0)]$

Similar logic shows that IV with $Y_i(1 - D_i)$ as the outcome and $1 - D_i$ as the treatment identifies $E[Y_i(0) \mid D_i(1) > D_i(0)]$

- So easy to do! And extends to covariates / multiple IVs...

Illustration: Angrist et al. (2013) Charter School IV

	Urban				Nonurban			
Subject	Treatment effect (1)	$E_u[Y_0 \mid D=0]$ (2)	λ_0^u (3)	λ_1^u (4)	Treatment effect (5)	$E_n[Y_0 \mid D=0]$ (6)	λ_0^n (7)	λ_1^n (8)
Panel A. Middle school								
Math	0.483***	−0.399***	0.077	0.560***	−0.177**	0.236***	0.010	−0.143***
	(0.074)	(0.011)	(0.049)	(0.054)	(0.074)	(0.007)	(0.061)	(0.042)
N	4,858				2,239			
ELA	0.188***	−0.422***	0.118**	0.306***	−0.148***	0.260***	0.102**	−0.086***
	(0.064)	(0.012)	(0.054)	(0.049)	(0.048)	(0.007)	(0.050)	(0.030)
N	4,551				2,323			

Decomposing

$$LATE = \underbrace{E[Y_i(1) \mid D_i(1) > D_i(0)] - E[Y_i(0) \mid D_i = 0]}_{\lambda_1}$$

$$- \underbrace{(E[Y_i(0) \mid D_i(1) > D_i(0)] - E[Y_i(0) \mid D_i = 0])}_{\lambda_0}$$

shows that charter compliers have typical counterfactual achievement

Covariates

For covariates X_i (not affected by D_i) we can follow a similar trick:

- Either IV'ing $X_i D_i$ on D_i or IV'ing $X_i(1 - D_i)$ on $1 - D_i$ identifies complier characteristics $E[X_i \mid D_i(1) > D_i(0)]$
- Shouldn't be very different (implicit balance test); can be averaged

Covariates

For covariates X_i (not affected by D_i) we can follow a similar trick:

- Either IV'ing $X_i D_i$ on D_i or IV'ing $X_i(1-D_i)$ on $1-D_i$ identifies complier characteristics $E[X_i \mid D_i(1) > D_i(0)]$
- Shouldn't be very different (implicit balance test); can be averaged

Can be useful to compare with always- and never-taker means:

- AT: $E[W_i \mid D_i(1) = D_i(0) = 1] = E[W_i \mid D_i = 1, Z_i = 0]$
- NT: $E[W_i \mid D_i(1) = D_i(0) = 0] = E[W_i \mid D_i = 0, Z_i = 1]$

Illustration: Angrist et al. (2023) Charter School IV

	Compliers				
	Untreated (1)	Treated (2)	Pooled (3)	Always-takers (4)	Never-takers (5)
Female	0.506	0.510	0.508	0.539	0.463
	(0.023)	(0.021)	(0.016)	(0.024)	(0.017)
Black	0.401	0.380	0.390	0.623	0.490
	(0.022)	(0.021)	(0.016)	(0.023)	(0.017)
Hispanic	0.250	0.300	0.275	0.183	0.228
	(0.02)	(0.018)	(0.013)	(0.019)	(0.014)
Asian	0.022	0.024	0.023	0.004	0.024
	(0.007)	(0.005)	(0.004)	(0.003)	(0.005)
White	0.229	0.216	0.223	0.154	0.215
	(0.018)	(0.016)	(0.012)	(0.016)	(0.014)
Special education	0.190	0.181	0.186	0.158	0.177
	(0.018)	(0.016)	(0.012)	(0.018)	(0.013)
English language learner	0.143	0.148	0.145	0.054	0.088
	(0.015)	(0.013)	(0.010)	(0.011)	(0.010)
Subsidized lunch	0.689	0.705	0.697	0.698	0.666
	(0.021)	(0.019)	(0.014)	(0.022)	(0.016)
Baseline math score	-0.274	-0.312	-0.293	-0.394	-0.301
	(0.047)	(0.041)	(0.032)	(0.045)	(0.036)
Baseline English score	-0.352	-0.349	-0.350	-0.362	-0.299
	(0.050)	(0.043)	(0.033)	(0.046)	(0.038)
Share of sample			0.546	0.197	0.257

Fancier Things

General result: $E[g(W_i, Y_i(d)) \mid D_i(1) > D_i(0)]$ for any $g(\cdot)$ and any $d \in \{0, 1\}$ is identified by β in the IV regression:

$$g(W_i, Y_i) \times \mathbf{1}[D_i = d] = \alpha + \beta \mathbf{1}[D_i = d] + \varepsilon_i$$

$$\mathbf{1}[D_i = d] = \mu + \pi Z_i + \nu_i$$

Fancier Things

General result: $E[g(W_i, Y_i(d)) \mid D_i(1) > D_i(0)]$ for any $g(\cdot)$ and any $d \in \{0, 1\}$ is identified by β in the IV regression:

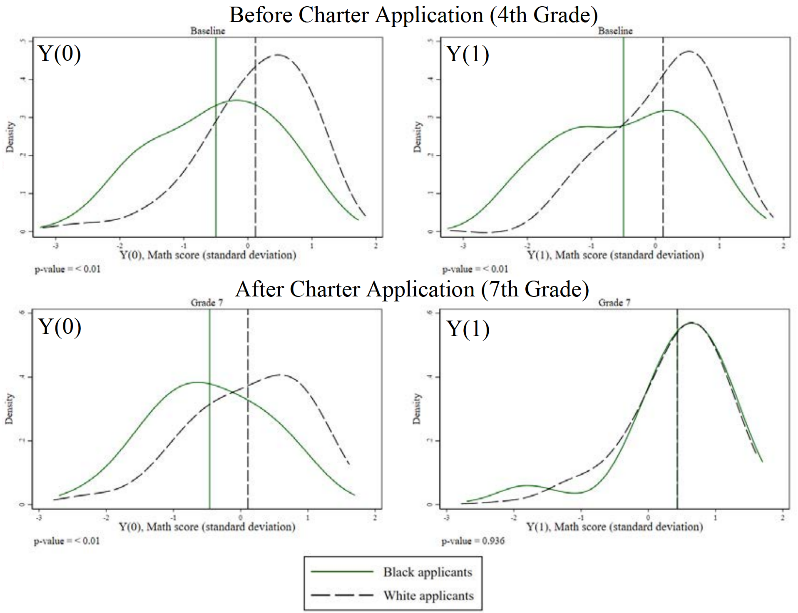
$$g(W_i, Y_i) \times \mathbf{1}[D_i = d] = \alpha + \beta \mathbf{1}[D_i = d] + \varepsilon_i$$

$$\mathbf{1}[D_i = d] = \mu + \pi Z_i + \nu_i$$

Lots of fun stuff we can do here. E.g.: distributions

- $g(W_i, Y_i) = \mathbf{1}[Y_i \leq y]$ estimates the CDF of complier potential outcomes, $F(y) = Pr(Y_i(d) < y \mid D_i(1) > D_i(0))$
- $g(W_i, Y_i) = \frac{1}{h} K(\frac{Y_i - y}{h})$ estimates the corresponding PDF, where $K(\cdot)$ is a kernel function and h is a bandwidth

Illustration: Angrist et al. (2023) Charter School IV



Outside the Basic IA Setup

The same logic applies to IV regressions with flexible controls

- E.g. *ivreg*ing $Y_i D_i$ on D_i instrumenting by Z_i and controlling for cell FE ID's a weighted avg of within-cell $E[Y_i(1) \mid D_i(1) > D_i(0)]$

Outside the Basic IA Setup

The same logic applies to IV regressions with flexible controls

- E.g. *ivreg*ing $Y_i D_i$ on D_i instrumenting by Z_i and controlling for cell FE ID's a weighted avg of within-cell $E[Y_i(1) \mid D_i(1) > D_i(0)]$
- Abadie (2003) "kappa-weighting" is an alternative approach for unweighted averages (and other things)

Outside the Basic IA Setup

The same logic applies to IV regressions with flexible controls

- E.g. *ivreg*ing $Y_i D_i$ on D_i instrumenting by Z_i and controlling for cell FE ID's a weighted avg of within-cell $E[Y_i(1) \mid D_i(1) > D_i(0)]$
- Abadie (2003) "kappa-weighting" is an alternative approach for unweighted averages (and other things)

Logic also goes through with continuous instruments

- E.g. can always mechanically decompose an IV on a binary D_i into implied "$Y(1)$" and "$Y(0)$" terms
- Unfortunately, things get trickier with non-binary treatments

What about Diff-in-Diff?

Most LATE-style discussions of IV identification are “design-based”:

i.e., the instrument is assumed to be drawn randomly, as if in an RCT

- But not all instruments are easily thought of this way (e.g. distance)

What about Diff-in-Diff?

Most LATE-style discussions of IV identification are “design-based”: i.e., the instrument is assumed to be drawn randomly, as if in an RCT

- But not all instruments are easily thought of this way (e.g. distance)

Luckily, whenever you have a reduced form & first stage that are causal and exclusion+monotonicity hold, the IV ratio has a LATE interpretation

- E.g. diff-in-diff IV, for distance measure $Z_i \in \{0,1\}$ and $t \in \{1,2\}$:

$$Y_{it} = \alpha_i + \tau_t + \beta D_{it} + \epsilon_{it}$$

$$D_{it} = \mu_i + \lambda_t + \pi Z_i \times \mathbf{1}[t=2] + \epsilon_{it}$$

  Parallel trends in Z_i make the reduced form & first stage causal

What about Diff-in-Diff?

Most LATE-style discussions of IV identification are "design-based": i.e., the instrument is assumed to be drawn randomly, as if in an RCT

- But not all instruments are easily thought of this way (e.g. distance)

Luckily, whenever you have a reduced form & first stage that are causal and exclusion+monotonicity hold, the IV ratio has a LATE interpretation

- E.g. diff-in-diff IV, for distance measure $Z_i \in \{0, 1\}$ and $t \in \{1, 2\}$:

$$Y_{it} = \alpha_i + \tau_t + \beta D_{it} + \epsilon_{it}$$

$$D_{it} = \mu_i + \lambda_t + \pi Z_i \times \mathbf{1}[t = 2] + \epsilon_{it}$$

  Parallel trends in Z_i make the reduced form & first stage causal

- Recent TWFE literature becomes relevant with fancier specifications

Example: Duflo (2001) School-Building Diff-in-Diff

TABLE 3—MEANS OF EDUCATION AND LOG(WAGE) BY COHORT AND LEVEL OF PROGRAM CELLS

	Years of education			Log(wages)		
	Level of program in region of birth			Level of program in region of birth		
	High (1)	Low (2)	Difference (3)	High (4)	Low (5)	Difference (6)
Panel A: Experiment of Interest						
Aged 2 to 6 in 1974	8.49	9.76	−1.27	6.61	6.73	−0.12
	(0.043)	(0.037)	(0.057)	(0.0078)	(0.0064)	(0.010)
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Difference	0.47	0.36	0.12	−0.26	−0.29	0.026
	(0.070)	(0.038)	(0.089)	(0.011)	(0.0096)	(0.015)
Panel B: Control Experiment						
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Aged 18 to 24 in 1974	7.70	9.12	−1.42	6.92	7.08	−0.16
	(0.059)	(0.044)	(0.072)	(0.0097)	(0.0076)	(0.012)
Difference	0.32	0.28	0.034	0.056	0.063	0.0070
	(0.080)	(0.061)	(0.098)	(0.013)	(0.010)	(0.016)

Here $Z_i =$ growing up in a region with intensive school building

Roadmap

- IV Mechanics
  - Just-Identified IV
  - Overidentification
  - Weak vs. Many-Weak Bias
- IV Interpretation
  - LATE and Generalizations
  - Characterizing Compliers
  - Diff-in-Diff and IV
- Formula Instruments
  - Shift-Share IV
  - Recentered IV

Exogenous Shocks with Non-Random Exposure

Increasingly, researchers construct instruments from multiple sources of variation — only some of which is viewed as exogenous. E.g.:

- Shift-share instruments $Z_i = \sum_k s_{ik} g_k$, which average together a set of shocks g_k with predetermined exposure shares s_{ik}
- IVs capturing spillovers across social networks/geographies
- “Simulated” instruments, capturing eligibility for a public program

Exogenous Shocks with Non-Random Exposure

Increasingly, researchers construct instruments from multiple sources of variation — only some of which is viewed as exogenous. E.g.:

- Shift-share instruments $Z_i = \sum_k s_{ik} g_k$, which average together a set of shocks g_k with predetermined exposure shares s_{ik}
- IVs capturing spillovers across social networks/geographies
- "Simulated" instruments, capturing eligibility for a public program

Generically, $Z_i = f_i(g, s)$ where $g = (g_1, \dots, g_k)$ is a set of shocks, s is some measure of shock exposure, and $f_i(\cdot)$ is a known formula

- How can we leverage exogeneity in g, allowing s to be non-random?

Example: Autor, Dorn, and Hanson (2014)

ADH study the effects of rising Chinese import competition on US commuting zones in the 1990's and 2000's

- Treatment X_i: growth of Chinese imports in CZ i
- Main outcome Y_i: change in manufacturing employment in CZ i

Example: Autor, Dorn, and Hanson (2014)

ADH study the effects of rising Chinese import competition on US commuting zones in the 1990's and 2000's

- Treatment X_i: growth of Chinese imports in CZ i
- Main outcome Y_i: change in manufacturing employment in CZ i

They use a shift-share instrument $Z_i = \sum_k s_{ik} g_k$, where:

- Shocks g_k: industry k's Chinese import growth in non-US economies
- Shares s_{ik}: pre-period employment share of industry k in CZ i

Example: Autor, Dorn, and Hanson (2014)

ADH study the effects of rising Chinese import competition on US commuting zones in the 1990's and 2000's

- Treatment X_i: growth of Chinese imports in CZ i
- Main outcome Y_i: change in manufacturing employment in CZ i

They use a shift-share instrument $Z_i = \sum_k s_{ik} g_k$, where:

- Shocks g_k: industry k's Chinese import growth in non-US economies
- Shares s_{ik}: pre-period employment share of industry k in CZ i

Idea: Z_i predicts X_i with proxies for underlying productivity shocks

- Fairly intuitive ... but what do we need for it to work?

SSIV with Exogneous Shocks

Borusyak, Hull, and Jaravel (2022) show shift-share IV can work when

- The shocks g_k are exogenous, at least conditional on some q_k
- There is enough independent variation in the shocks (i.e. large K)

SSIV with Exogneous Shocks

Borusyak, Hull, and Jaravel (2022) show shift-share IV can work when

- The shocks g_k are exogenous, at least conditional on some q_k
- There is enough independent variation in the shocks (i.e. large K)

Intuitively, $Z_i = \sum_k s_{ik} g_k$ works as a "translation device," bringing the k-level variation to bear on causal effects at the i level

- E.g. in Autor et al. (2014), an industry-level natural experiment can be used to estimate effects at the commuting zone level
- The shares s_{ik} used for this translation do not need to be exogenous!

Two Practical Considerations

Need to control for $E[Z_i \mid s, q] = E[\sum_k s_{ik} g_k \mid s, q] = \sum_k s_{ik} E[g_k \mid q]$ in order to isolate the as-good-as-random variation in shocks

- When $E[g_k \mid q] = q_k' \gamma$, this means controlling for $\sum_k s_{ik} q_k$
- E.g. in Autor et al. (2014), to isolate within-period shock variation, need to control for $\sum_k s_{ik} \times$ (period FE)

Two Practical Considerations

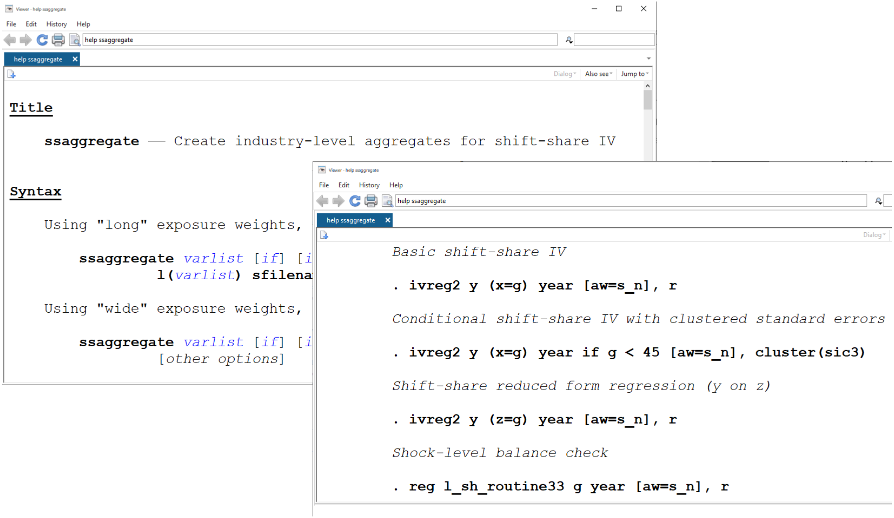
Need to control for $E[Z_i \mid s, q] = E[\sum_k s_{ik} g_k \mid s, q] = \sum_k s_{ik} E[g_k \mid q]$ in order to isolate the as-good-as-random variation in shocks

- When $E[g_k \mid q] = q_k' \gamma$, this means controlling for $\sum_k s_{ik} q_k$
- E.g. in Autor et al. (2014), to isolate within-period shock variation, need to control for $\sum_k s_{ik} \times$ (period FE)

Need to adjust SEs for a non-standard clustering: $Z_i = \sum_k s_{ik} g_k$ and $Z_j = \sum_k s_{jk} g_k$ are correlated via their common exposure to shocks

- Easy fix: use the *ssaggregate* Stata/R packages to translate SSIV estimation to the level of identifying variation (shocks)

Estimating Exogenous-Shock SSIV Regressions in Stata



To install: `ssc install ssaggregate`

Autor et al. (2014) Revisited

TABLE 4
Shift-share IV estimates of the effect of Chinese imports on manufacturing employment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Coefficient	−0.596	−0.489	−0.267	−0.314	−0.310	−0.290	−0.432
	(0.114)	(0.100)	(0.099)	(0.107)	(0.134)	(0.129)	(0.205)
Regional controls							
Autor *et al.* (2013) controls	✓	✓	✓		✓	✓	✓
Start-of-period mfg. share	✓						
Lagged mfg. share		✓	✓	✓	✓	✓	✓
Period-specific lagged mfg. share			✓	✓	✓	✓	✓
Lagged 10-sector shares					✓		✓
Local Acemoglu *et al.* (2016) controls						✓	
Lagged industry shares							✓
SSIV first stage F-stat.	185.6	166.7	123.6	272.4	64.6	63.3	27.6
No. of region-periods	1,444	1,444	1,444	1,444	1,444	1,444	1,444
No. of industry-periods	796	794	794	794	794	794	794

- Columns 3-7 include the key $\sum_k s_{ik} \times$ (period FE) control
- Standard errors / first stage F-stat.'s computed via *ssaggregate*

Beyond Shift-Share

Suppose we wanted to estimate the economic impact of new transportation upgrades (e.g. new high speed rail construction)

- Economic theory tells us to expect spillovers: easier travel increases productivity across all cities i in a country, to different extents
- E.g. *market access*: $X_i = \sum_j \tau(g, loc_i, loc_j)^{-1} pop_j$ where g captures the railroad network, loc_i gives lat/lon of cities, and pop_i is city size

Beyond Shift-Share

Suppose we wanted to estimate the economic impact of new transportation upgrades (e.g. new high speed rail construction)

- Economic theory tells us to expect spillovers: easier travel increases productivity across all cities i in a country, to different extents
- E.g. *market access*: $X_i = \sum_j \tau(g, loc_i, loc_j)^{-1} pop_j$ where g captures the railroad network, loc_i gives lat/lon of cities, and pop_i is city size

Like a more complex shift-share: $Z_i = f_i(g, s)$ for $s = (loc_i, pop_i)_{i=1}^N$

- Suppose we had a natural experiment in high-speed rail construction; can we “translate” these g shocks via the market access mapping?

Recentered IV

Borusyak and Hull (2023) show how to leverage exogenous g in $f_i(g, s)$:

1. Specify *counterfactual* exogenous shocks $g^{(1)}, \ldots, g^{(C)}$ which may well have occured (e.g. shuffle the timing of new rail construction)
2. Recompute the instrument: $Z_i^{(c)} = f_i(g^{(c)}, s)$, for $c = 1, \ldots, C$
3. Take the average ("expected instrument"): $\mu_i = \frac{1}{C} \sum_i Z_i^{(c)}$

Recentered IV

Borusyak and Hull (2023) show how to leverage exogenous g in $f_i(g, s)$:

1. Specify *counterfactual* exogenous shocks $g^{(1)}, \ldots, g^{(C)}$ which may well have occured (e.g. shuffle the timing of new rail construction)
2. Recompute the instrument: $Z_i^{(c)} = f_i(g^{(c)}, s)$, for $c = 1, \ldots, C$
3. Take the average (“expected instrument”): $\mu_i = \frac{1}{C} \sum_i Z_i^{(c)}$

Either controlling for μ_i or using the recentered $\tilde{Z}_i = Z_i - \mu_i$ avoids any bias from endogeneity/non-randomness in Z_i

- Analogous to controlling for $\sum_k s_{ik} q_k$ in shift-share IV!
- BH also discuss the analog of non-standard SSIV clustering: the solution is a bit more non-standard (randomization inference)

Illustration: High-Speed Rail in China

149 lines were built or planned (as of April 2019)

Illustration: High-Speed Rail in China

The 83 lines actually built by 2016. Suppose exact timing is random

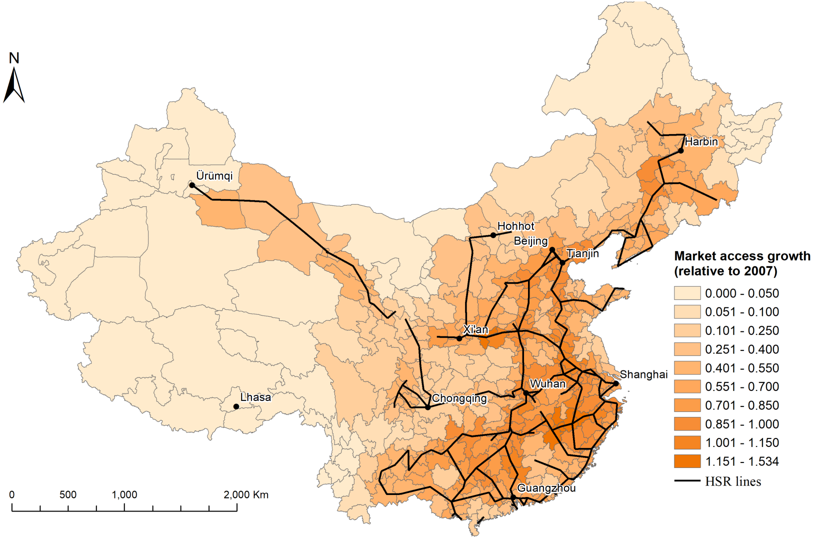


Illustration: High-Speed Rail in China

A counterfactual draw of 83 lines by 2016

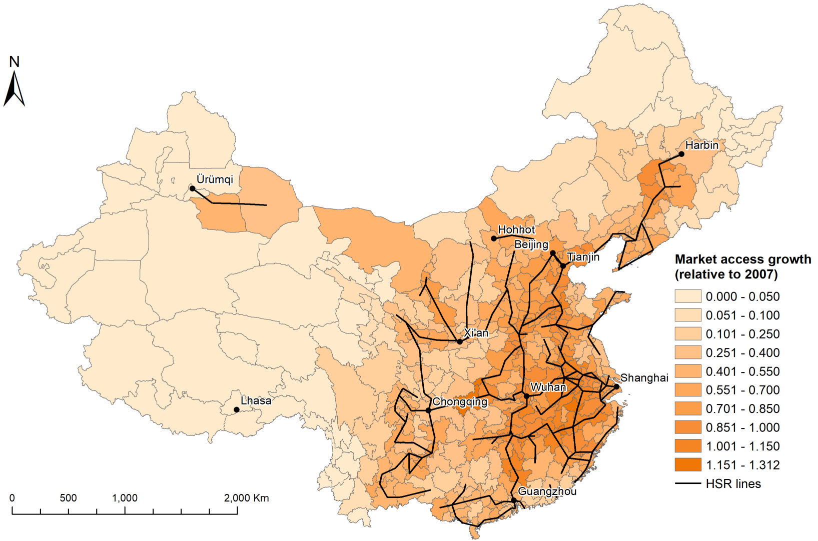


Illustration: High-Speed Rail in China

Expected MA growth, μ_i

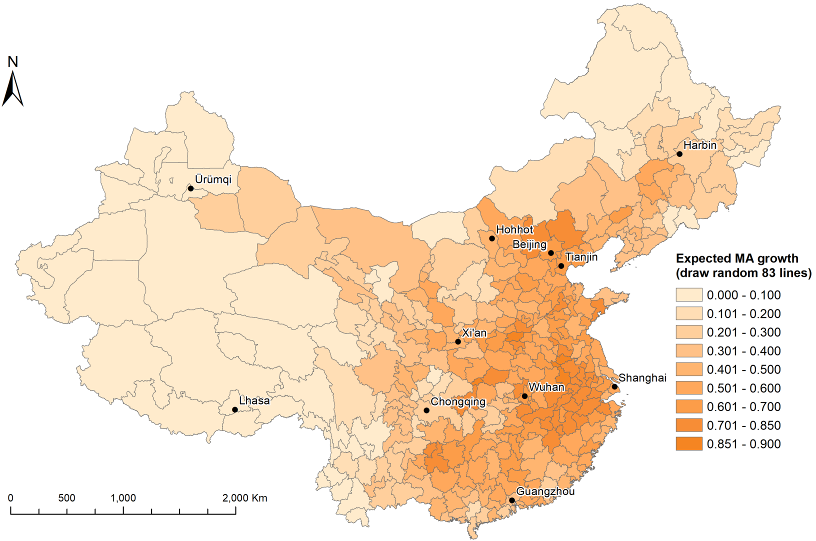
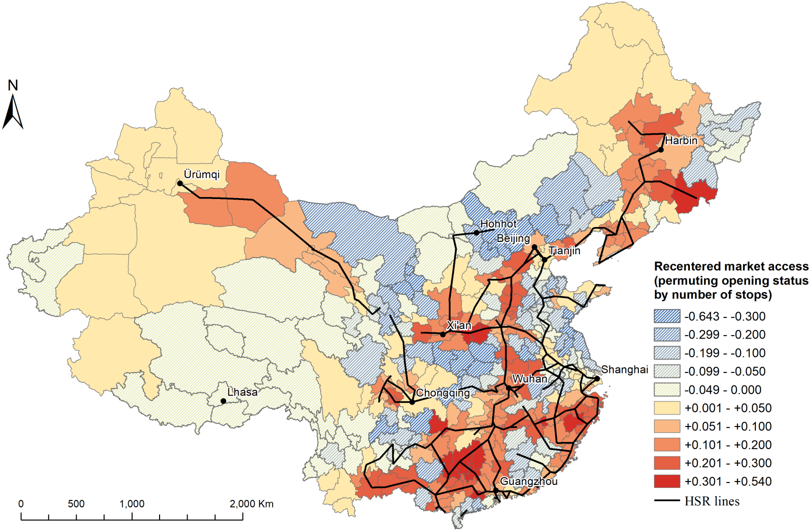


Illustration: High-Speed Rail in China

Recentered MA growth, Z_i



Adjusted Estimates of Market Access Effects

	Unadjusted OLS (1)	Recentered IV (2)	Controlled OLS (3)
Panel A. No Controls			
Market Access Growth	0.232	0.081	0.069
	(0.075)	(0.098)	(0.094)
		[-0.315, 0.328]	[-0.209, 0.331]
Expected Market Access Growth			0.318
			(0.095)
Panel B. With Geography Controls			
Market Access Growth	0.132	0.055	0.045
	(0.064)	(0.089)	(0.092)
		[-0.144, 0.278]	[-0.154, 0.281]
Expected Market Access Growth			0.213
			(0.073)
Recentered	No	Yes	Yes
Prefectures	274	274	274

Regressions of log employment growth on log market access growth in 2007–2016.

Spatial-clustered standard errors in parentheses; permutation-based 95% CI in brackets

Recentering for Power

Leveraging non-random exposure can dramatically improve precision when leveraging exogenous shocks in an IV...

- ... as long as you recenter to use this variation “safely”

Recentering for Power

Leveraging non-random exposure can dramatically improve precision when leveraging exogenous shocks in an IV...

- ... as long as you recenter to use this variation "safely"

Ex: Medicaid eligibility s a treatment combining statewide policy shocks and individual exposure (income, family structure, etc)

- In settings where policy shocks are plausibly exogenous, standard approach is to use them directly as instruments ("simulated IV")
- BH approach: use eligibility itself, but recenter: e.g. adjust for i's avg. eligibility across permutations of policies (swap MA & RI, say)

Illustration: ACA Medicaid Expansion

	Has Medicaid		Has Private Insurance		Has Employer-Sponsored Insurance	
	Simulated IV (1)	Recentered IV (2)	Simulated IV (3)	Recentered IV (4)	Simulated IV (5)	Recentered IV (6)
Panel A. Eligibility Effects						
Eligibility	0.132	0.072	−0.048	−0.023	0.009	−0.009
	(0.028)	(0.010)	(0.023)	(0.007)	(0.014)	(0.005)
	[0.080,0.216]	[0.051,0.093]	[-0.110,0.009]	[-0.040,-0.007]	[-0.034,0.052]	[-0.021,0.004]
Panel B. Enrollment Effects						
Has Medicaid			−0.361	−0.321	0.068	−0.125
			(0.165)	(0.092)	(0.111)	(0.061)
			[-0.813,0.082]	[-0.566,-0.108]	[-0.232,0.421]	[-0.263,0.070]
P-value: SIV=RIV			0.719		0.104	
Exposed Sample	N	Y	N	Y	N	Y
States	43	43	43	43	43	43
Individuals	2,397,313	421,042	2,397,313	421,042	2,397,313	421,042

1% ACS sample of non-disabled adults in 2013–14, diff-in-diff IV regressions using one of the two instruments. Controls include state and year fixed effects and an indicator for Republican governor interacted with year. State-clustered standard errors in parentheses; wild score bootstrap 95% CI in brackets

Concluding Thoughts

We’ve covered a ton of ground in a day!

- From basic IV mechanics to advanced recentering techniques

Concluding Thoughts

We've covered a ton of ground in a day!

- From basic IV mechanics to advanced recentering techniques

Some general advice on how to approch IV in the "real world":

1. Figure out what the reduced form and first stage regressions are, & what variation in the instrument(s) they leverage via the controls
2. Assess the plausibility of as-good-as-random assignment (or parallel trends) of this variation, both theoretically and empirically
3. Interrogate exclusion, which lets you interpret RF/FS, and monotonicity/complier stats if heterogeneous FX seem important

Keep Calm and `ivreg2` On!

Thank you, and good luck on your future adventures with IV!

peter_hull@brown.edu