Chapter 8: Shift-Share & Recentered IV

Peter Hull

Applied Econometrics II
Brown University
Spring 2024

Outline

1. SSIV Basics
2. Exogenous Shocks
3. Exogenous Shares
4. Recentered IV

What is a (Linear) SSIV?

A weighted sum of a common set of shocks, with weights reflecting heterogeneous exposure shares:

$$z_\ell = \sum_n s_{\ell n} g_n$$

- The shocks vary at a different "level," $n = 1, \ldots, N$ than the shares $\ell = 1, \ldots, L$, where we also observe an outcome y_ℓ and treatment x_ℓ

We want to use z_ℓ to estimate parameter β of the model $y_\ell = \beta x_\ell + \varepsilon_\ell$

- Could be a structural equation / potential outcomes model
- Could be misspecified, with heterogeneous treatment effects β_ℓ
- Could be a "reduced form" analysis, with $x_\ell = z_\ell$
- Could have other included controls w_ℓ

Key question: under what assumptions will this SSIV strategy work?

SSIV Examples

Instrument $z_\ell = \sum_n s_{\ell n} g_n$ for model $y_\ell = \beta x_\ell + w_\ell' \gamma + \varepsilon_\ell$

Bartik (1991); Blanchard and Katz (1992):

- β = inverse local labor supply elasticity
- x_ℓ and y_ℓ = employment and wage growth in location ℓ
- Need a labor demand shifter as an IV
- g_n = national growth of industry n
- $s_{\ell n}$ = lagged employment shares of industry n in location ℓ
- z_ℓ = predicted employment growth due to national industry trends

SSIV Examples (Cont.)

Instrument $z_\ell = \sum_n s_{\ell n} g_n$ for model $y_\ell = \beta x_\ell + w_\ell' \gamma + \varepsilon_\ell$

Autor, Dorn, and Hanson (2013, ADH) "China Shock:"

- x_ℓ = growth of Chinese import competition in location ℓ
- y_ℓ = growth of manufacturing employment, unemployment, etc.
- g_n = growth of China exports in manufacturing industry n to 8 non-U.S. countries
- $s_{\ell n}$ = 10-year lagged employment shares (over total employment)
- z_ℓ = predicted growth of Chinese import competition

SSIV Examples (Cont.)

Instrument $z_\ell = \sum_n s_{\ell n} g_n$ for model $y_\ell = \beta x_\ell + w_\ell' \gamma + \varepsilon_\ell$

"Enclave instrument," e.g. Card (2009):

- β = inverse elasticity of substitution between native and immigrant labor of some skill level (need a relative labor supply instrument)
- x_ℓ and y_ℓ = relative employment and wage in location ℓ
- g_n = national immigration growth from origin country n
- $s_{\ell n}$ = lagged shares of migrants from origin n in location ℓ
- z_ℓ = share of migrants predicted from enclaves & recent growth

What Do We Make of These?

Of course, we can always run IV with such z_ℓ ... but what does the IV estimand *identify*?

IV validity condition: $E\left[\frac{1}{L}\sum_\ell z_\ell \varepsilon_\ell\right] = 0$ for model residual ε_ℓ

- Looks a little different than normal because we're not assuming *iid* sampling, i.e. that $E\left[\frac{1}{L}\sum_\ell z_\ell \varepsilon_\ell\right] = E[z_\ell \varepsilon_\ell]$ (you'll see why soon)

What properties of shocks and shares make this condition hold?

- Is SSIV "design-based," like a charter lottery? Is it "model-based" like DiD? Or is it something new altogether?
- Note that since z_ℓ combines multiple sources of variation, it can be difficult to think about it being randomly assigned across ℓ

Outline

1. SSIV Basics✓
2. Exogenous Shocks
3. Exogenous Shares
4. Recentered IV

Identification Through Exogenous Shocks

Borusyak et al. (2022; BHJ) show how $E\left[\frac{1}{L}\sum_\ell z_\ell \varepsilon_\ell\right] = 0$ can hold from a natural experiment in the shocks g_n, allowing for endogenous shares $s_{\ell n}$

- Intuitively, need g_n to be as-good-as-randomly assigned + excludable; large N and heterogeneous exposure make this variation "useful"
- z_ℓ translates the exogenous shocks to a different "level"

Two main practical takeaways from this framework:

1. The controls w_ℓ should be carefully chosen, to isolate the desired quasi-experimental shock variation (e.g. "incomplete shares")
2. Standard errors should account for the common exposure to as-good-as-random shocks ("exposure-robust" SEs)

Shock-Level Equivalence

BHJ start with a numerical equivalence: the SSIV estimator can also be written as a (weighted) shock-level IV procedure:

$$\hat{\beta} = \frac{\sum_\ell z_\ell y_\ell^\perp}{\sum_\ell z_\ell x_\ell^\perp} = \frac{\sum_\ell \sum_n s_{\ell n} g_n y_\ell^\perp}{\sum_\ell \sum_n s_{\ell n} g_n x_\ell^\perp} = \frac{\sum_n g_n \sum_\ell s_{\ell n} y_\ell^\perp}{\sum_n g_n \sum_\ell s_{\ell n} x_\ell^\perp} = \frac{\sum_n s_n g_n \bar{y}_n^\perp}{\sum_n s_n g_n \bar{x}_n^\perp}$$

where:

- $v_\ell^\perp$ are the sample residuals from regressing v_ℓ on w_ℓ
- $s_n = \frac{1}{L}\sum_\ell s_{\ell n}$ are weights capturing the avg. "importance" of shock n
- $\bar{v}_n = \frac{\sum_\ell s_{\ell n} v_\ell}{\sum_\ell s_{\ell n}}$ are "exposure-weighted" averages of v_ℓ

I.e. $\hat{\beta}$ comes from instrumenting with g_n and weighting by s_n to estimate:

$$\bar{y}_n^\perp = \alpha + \beta \bar{x}_n^\perp + \bar{\varepsilon}_n$$

It follows that $\hat{\beta}$ is consistent if and only if this shock-level *ivreg* is...

BHJ Baseline Assumptions

A1 (Quasi-random shock assignment): $E[g_n \mid \bar{\varepsilon}, s] = \mu$ for all n

- Each shock has the same expected value, conditional on the shock-level unobsevables $\bar{\varepsilon}_n$ and average exposure s_n
- In ADH, $\bar{\varepsilon}_n$ is the average unobserved determinants of regional employment in regions most specialized in industry n
- A1 holds in ADH when the expected growth of chinese imports g_n is the same across industries with high vs. low $\bar{\varepsilon}_n$ (and s_n)

A1 ensures identification when $\sum_n s_{\ell n} = 1$ ("complete shares"):

$$E\left[\frac{1}{L}\sum_{\ell} z_\ell \varepsilon_\ell\right] = E\left[\sum_n s_n g_n \bar{\varepsilon}_n\right] = E\left[\sum_n s_n E[g_n \mid \bar{\varepsilon}, s]\bar{\varepsilon}_n\right] = \mu \underbrace{E\left[\sum_n s_n \bar{\varepsilon}_n\right]}_{=0}$$

since $E\left[\sum_n s_n \bar{\varepsilon}_n\right] = E\left[\frac{1}{L}\sum_\ell \sum_n s_{\ell n}\varepsilon_\ell\right] = E\left[\frac{1}{L}\sum_\ell \varepsilon_\ell\right] = 0$ (a normalization)

BHJ Baseline Assumptions (Cont.)

A2 (Many uncorrelated shocks): $E[\sum_n s_n^2] \to 0$ as $L \to \infty$ and $Cov(g_n, g_{n'} \mid \bar{\varepsilon}, s) = 0$ for all $n' \neq n$

- Expected Herfindahl index of average shock exposure converges to zero (implies $N \to \infty$) and shocks are cond. mutually uncorrelated
- In ADH, imposes many uncorrelated industry growth rates and sufficiently different industry specialization across locations

A2 implies a shock-level law of large numbers:

$$\frac{1}{L}\sum_{\ell} z_\ell \varepsilon_\ell = \sum_n s_n g_n \bar{\varepsilon}_n \xrightarrow{p} E\left[\sum_n s_n g_n \bar{\varepsilon}_n\right] \underbrace{= 0}_{\text{(under A1)}}$$

Note the similarity with assumptions we'd use if studying a shock-level natural experiment!

Extensions

Conditional quasi-random assignment: $E[g_n \mid \bar{\varepsilon}, q, s] = q_n'\mu$ for some observed shock-level variables q_n

- Consistency follows when $w_\ell = \sum_n s_{\ell n} q_n$ is controlled for in the IV

Weakly mutually correlated shocks: $g_n \mid (\bar{\varepsilon}, q, s)$ are clustered or otherwise mutually dependent

- Consistency follows when mutual correlation is not too strong

Estimated shocks: $g_n = \sum \ell s_{\ell n} g_{\ell n}$ proxies for an infeasible g_n^*

- Consistency may require a "leave-out" adjustment: $z_\ell = \sum_\ell s_{\ell n} \tilde{g}_{\ell n}$ for $\tilde{g}_{\ell n} = \sum_{\ell' \neq \ell} s_{\ell' n} g_{\ell' n}$ (akin to JIVE solution to many-IV bias)

Panel data: Have $(y_{\ell t}, x_{\ell t}, s_{\ell n t}, g_{nt}$ for $\ell = 1, \dots, L$ and $t = 1, \dots, T$

- Consistency can follow from either $N \to \infty$ or $T \to \infty$

Heterogeneous effects: LATE theorem logic goes through, under an appropriate first-stage monotonicity condition

Practical Consideration 1: Incomplete Shares

So far we have assumed a constant sum-of-shares: $S_\ell \equiv \sum_n s_{\ell n} = 1$ for all ℓ

- But in some settings S_ℓ varies; e.g. in ADH, S_ℓ is location ℓ's share of manufacturing employment

BHJ show that A1/A2 are not enough for validity in this case:

$$z_\ell = \sum_n s_{\ell n}(\mu + (g_n - \mu)) = \mu S_\ell + \sum_n s_{\ell n}(g_n - \mu)$$

The second term is useful variation, but the first may be correlated with ε_ℓ

- E.g. in ADH, regions w/ larger vs. smaller manufacturing employment may have different local productivity shocks

The solution: just control for sum-of-shares S_ℓ

- More generally, control for $\sum_n s_{\ell n} q_n$: e.g., in panels where q_n include time FE, interact sum-of-shares with period effects

Practical Consideration 2: Exposure Clustering

Adao, Kolesar, and Morales (2019) study a novel inference challenge when SSIV identification leverages quasi-random shocks:

- Observations with similar shares $s_{\ell 1}, \ldots, s_{\ell N}$ have correlated z_ℓ even without conventional clustering (e.g. by distance)
- When ε_ℓ is similarly clustered (e.g. when $\varepsilon_\ell = \sum_n s_{\ell n} \nu_n + \tilde{\varepsilon}_\ell$), conventional CLTs may fail to approximate the distribution of $\hat{\beta}$

AKM derive new CLTs & SEs to address this "exposure clustering"

- "Design-based," leveraging *iid*ness or clustering of shocks (not obs.)

BHJ show such SEs can be conveniently obtained by estimating SSIV at the level of identifying variation...

Exposure-Robust SEs

General shock-level regression: instrumenting with g_n and weighting by s_n:

$$\bar{y}_n^{\perp} = \alpha + \beta \bar{x}_n^{\perp} + q_n' \tau + \bar{\varepsilon}_n$$

Numerically identical to SSIV $\hat{\beta}$, when controls include $\sum_n s_{\ell n} q_n$

- Valid SEs with *ivreg [...], r* when g_n are conditionally *iid* or with *ivreg [...], cluster(c))* when shocks are clustered by some c
- Same logic applies for performing valid balance/pre-trend tests and evaluating the first-stage strength of the instrument
- See BHJ's *ssaggregate* Stata/R packages for producing $(\bar{y}_n^{\perp}, \bar{x}_n^{\perp})$

Illustration: ADH China Shock

ADH study the effects of rising Chinese import competition on US commuting zones, 1991-2000 and 2000-2007

- Treatment $x_{\ell t}$: local growth of Chinese imports in \$1,000/worker
- Main outcome $y_{\ell t}$: local change in manufacturing employment share

To address endogeneity, ADH use a SSIV $z_{\ell t} = \sum_n s_{\ell nt} g_{nt}$

- n: 397 SIC4 manufacturing industries $\times$ two periods
- g_{nt}: growth of Chinese imports in non-US economies per US worker
- $s_{\ell nt}$: lagged share of mfg. industry n in *total* employment of location ℓ

BHJ argue A1/A2 have some *ex ante* plausibility: imagine random industry productivity shocks in China affecting imports in U.S. and elsewhere

- Check balance (using exposure-robust SEs) controlling for $S_{\ell t} \times$ period
- Large effective sample (1/HHI of s_n): 58-192; shocks appear mutually uncorrelated across SIC3 sectors (but clustered within)

ADH Balance Tests

Balance variable	Coef.	SE
Production workers' share of employment, 1991	-0.011	(0.012)
Ratio of capital to value-added, 1991	-0.007	(0.019)
Log real wage (2007 USD), 1991	-0.005	(0.022)
Computer investment as share of total, 1990	0.750	(0.465)
High-tech equipment as share of total investment, 1990	0.532	(0.296)
# of industry-periods	794	

Note: clustering by SIC3 sector

Shocks do not predict industry-level observables controlling for period FE (BHJ also show balance with location-level characteristics)

SSIV Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Coefficient	-0.596	-0.489	$\boxed{-0.267}$	-0.314	-0.310	-0.290	-0.432
	(0.114)	(0.100)	$\boxed{(0.099)}$	(0.107)	(0.134)	(0.129)	(0.205)
<u>Regional controls</u>							
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
# of region-periods	1,444	1,444	1,444	1,444	1,444	1,444	1,444
# of industry-periods	796	794	794	794	794	794	794

Note: exposure-robust standard errors, clustering by SIC3 sector

Second-stage estimates are stable, once we account for incomplete shares

Outline

1. SSIV Basics✓
2. Exogenous Shocks✓
3. Exogenous Shares
4. Recentered IV

Motivation: DiD-IV as a Simple SSIV

Card (1990) leverages the Mariel Boatlift: a big migration "push" of low-skilled workers from Cuba to mostly Miami (a Cuban enclave)

- Imagine instrumenting immigrant inflows in a regression of employment growth by the lagged share of Cuban workers $s_{\ell,Cuba}$
- Want parallel trends: regions with more vs. fewer Cuban workers would have been on similar employment trends if not for the Boatlift

This can be seen as a simple shift-share setup, with binary shocks:

$$s_{\ell,Cuba} = s_{\ell,Cuba} \cdot 1 + \sum_{n \neq Cuba} s_{\ell n} \cdot 0$$

If several other migration origins n had a push shock, we could imagine pooling them together with a more traditional SSIV...

Goldsmith-Pinkham, Sorkin, and Swift (2020)

GPSS view the values of g_n as fixed, such that $z_\ell = \sum_n s_{\ell n} g_n$ is a linear combination of shares

- They then establish a (different) equivalence: SSIV $\hat{\beta}$ can be obtained from an overidentified IV procedure that uses the $s_{\ell n}$ as many IVs
- The fixed g_n just determine how the one-at-a-time share IV estimates are weighted together (recall IV mechanics!)

Sufficient identifying assumption: $E[\varepsilon_\ell \mid s_{\ell n}] = 0$ for each n (i.e. parallel trends when the ε_ℓ are in first differences

$$E\left[\sum_\ell z_\ell \varepsilon_\ell\right] = \sum_\ell \sum_n g_n E\left[E[\varepsilon_\ell \mid s_{\ell n} s_{\ell n}]\right] = 0$$

In other words, GPSS show that the SSIV estimator can be seen as pooling many Boatlift-style DiD-IVs, one for each n

Rotemberg Weights

How does SSIV combine the different DiD-IVs?

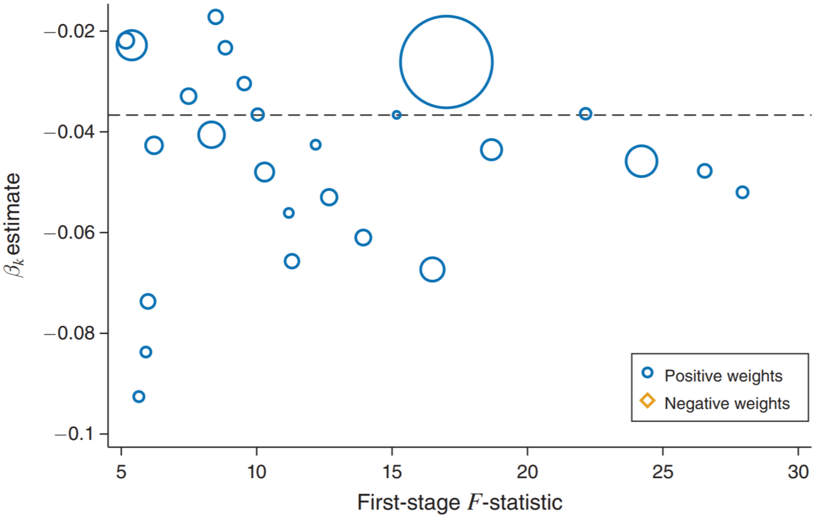
- GPSS propose "opening the black box" of overidentified IV by deriving the weights SSIV implicitly puts on each share instrument
- Call these "Rotemberg Weights" after Rotemberg (1983) – we've seen this already!

$$\hat{\beta} = \sum_n \hat{\alpha}_n \hat{\beta}_n, \text{ where } \quad \underbrace{\hat{\beta}_n = \frac{\sum_\ell s_{\ell n} y_\ell^\perp}{\sum_\ell s_{\ell n} x_\ell^\perp}}_{n\text{-specific IV estimate}} \quad \text{and} \quad \underbrace{\hat{\alpha}_n = \frac{g_n \sum_\ell s_{\ell n} x_\ell^\perp}{\sum_{n'} g_{n'} \sum_\ell s_{\ell n'} x_\ell^\perp}}_{\text{Rotemberg weight}}$$

Intuitively, SSIV puts more weight on share instruments with more extreme shocks g_n and larger first stages $\sum_\ell s_{\ell n} x_\ell^\perp$

- Weights can be negative, but are (as usual) identified

Rotemberg Weights in Card (2009)



Note: circle size proportional to Rotemberg weight; dashed line gives SSIV estimate

No negative weights, one-at-a-time share-IV estimates are all negative

Is Share Exogeneity Plausible?

Share exogeneity is not that "shares don't respond to the residual" (they can't – they're pre-determined)

- It's that all relevant unobservables are unforecastable from the shares

$E[\varepsilon_\ell \mid s_{\ell n}] = 0$ is typically violated when there are any unobserved shocks ν_n that affect ε_ℓ through the same/correlated ("generic") shares

- E.g. if $\varepsilon_\ell = \sum_n s_{\ell n} \nu_n + \tilde{\varepsilon}_\ell$, even if ν_n are uncorrelated with g_n
- Not a problem in BHJ, as long as the g_n are exogenous

GPSS show how share exogeneity can be tested, focusing on high-Rotemberg-weight-n to ease interpretation

- Balance/pre-trend tests, overid tests (under constant effects)
- Straightforward to implement; no new SEs or controls to consider
- Tests pass broadly for Card (2009), but fail badly for ADH...

Advice For Running SSIV

Determine whether or not a shock-level natural experiment is *ex ante* plausible: could you imagining a design-based analysis at the shock level?

- Shock variation may be limited or implausibly exogenous

If so, let this thinking guide your controls and standard errors

- Run analyses at the shock level to get back to familiar territory
- Check balance/pretrends and report shock-level first-stage F's

If not, think about whether you'd be comfortable with a DiD-IV using the shares as (possibly many) instruments

- Compute Rot weights to determine which shares are driving the show
- Check balance/pretrends and run overid. tests in the usual way

Remember: it could be that neither framework is appropriate!

Outline

1. SSIV Basics✓
2. Exogenous Shocks✓
3. Exogenous Shares✓
4. Recentered IV

Motivation

Many treatments/instruments have the flavor of SSIV: i.e. they're constructed from multiple sources of variation, by a known formula

- How can we leverage the fact that some, but not all, of these determinants are as-good-as-randomly assigned?

Three examples:

1. Spatial/network/GE spillover treatments: e.g. the number of neighbors selected for a randomized intervention:
   - Who got selected for the intervention & who neighbors whom
2. Regional growth of market access from transportation upgrades:
   - Location + timing of upgrades & location and size of markets
3. An individual's eligibility for a public program, e.g. Medicaid:
   - State-level policy & individual income and demographics

Borusyak and Hull (2022): Main Points

1. Non-random exposure to as-good-as-random shocks generates systematic variation in x_i, which can lead to omitted variable bias
    - Randomizing roads $\not\Rightarrow$ randomizing market access growth from them
2. Systematic variation in x_i can be removed via novel "recentering"
    - Specify many counterfactual sets of shocks
    - Compute μ_i = the average x_i across counterfactuals, by simulation — the key confounder (similar to a propensity score)
    - "Recenter" x_i by μ_i (i.e. instrument x_i with $x_i - \mu_i$) or control for μ_i
    - Conventional solutions (e.g. directly instrumenting with shocks or controlling for all features of exposure) are often infeasible
3. Recentering solution has attractive efficiency properties
    - Leverages non-random exposure, rather than discarding it
4. Same counterfactuals also yield inference tools and specification tests
    - Via randomization inference

Motivating Example 1: Market Access Effects via RCT

Theory suggests transportation upgrades affect local outcomes (e.g. land value) of regions i by increasing their market access (MA):

$$\Delta \log V_i = \beta \Delta \log MA_i + \varepsilon_i, \tag{1}$$

$$\text{where } MA_{it} = \sum_j \tau(g_t, loc_i, loc_j)^{-1} pop_j, \tag{2}$$

for road network g_t in periods $t = 1, 2$, region locations loc_j (co-determining travel cost τ), and regional population pop_j

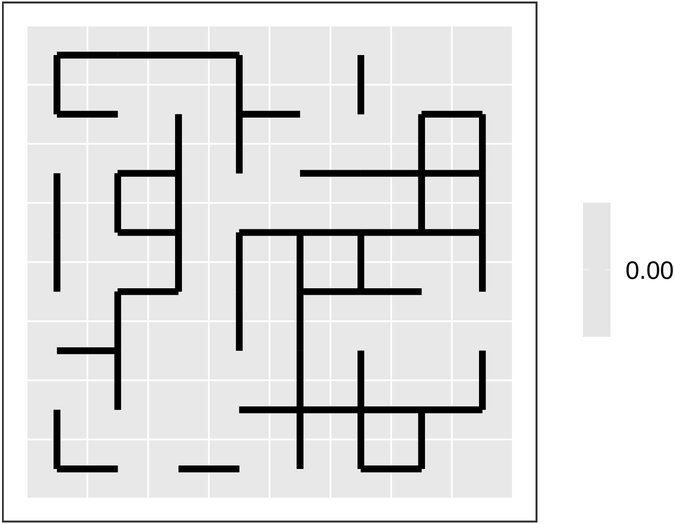
Imagine an experiment that randomly connects adjacent regions by road

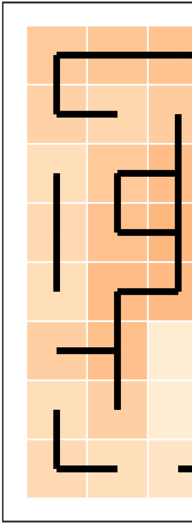
- MA only grows because of the random transportation shocks
- So can we view variation in MA growth as random and just run OLS?

Randomizing roads $\not\Rightarrow$ randomizing MA due to them!

Illustration: Market Access on a Square Island

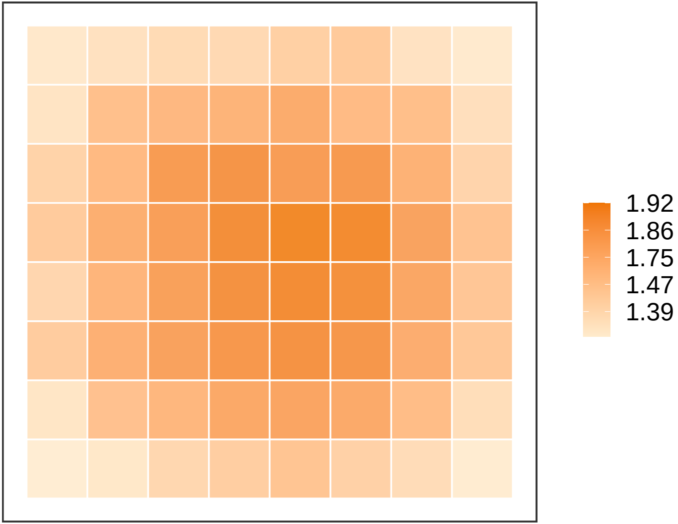
Randomly connect adjacent regions by road and compute MA growth





Expected Market Access Growth μ_i

Some regions get systematically more MA



OVB and Recentering

Systematic variation in MA growth can generate OVB

- E.g. land values fall in the periphery because of rising sea levels
- More vs less developed Chinese regions may be on different trends

Systematic variation can be removed via "recentering":

$$\begin{matrix}\text{Recentered} \\ \text{MA growth}\end{matrix} = \begin{matrix}\text{Realized} \\ \text{MA growth}\end{matrix} - \begin{matrix}\text{Expected} \\ \text{MA growth}\end{matrix}$$

Recentered MA is a valid instrument for realized MA growth

- Compares MA from actual and counterfactual shocks

Motivating Example 2: Effects of Program Eligibility

The effects of individual's eligibility x_i to a public program (e.g. Medicaid):

$$y_i = \beta x_i + \varepsilon_i$$

where x_i is determined by i's state policy g_{state_i} and demographics

- Suppose state policies g are as-good-as-random
- Yet, pre-determined demographics are endogeous $\Rightarrow$ OLS is biased

Standard "simulated instruments" solution (Currie and Gruber (1996)): use state-level variation only (a measure of policy generosity) as IV for x_i

BH approach:

- Formalize the policy experiment as "all permutations of g across states are equally likely"
- Compute $\mu_i =$ the share of states in which i would be eligible
- Leverage all variation in x_i but recenter by μ_i (or control for μ_i)
- Yields efficiency gain by better first-stage prediction, e.g. by removing i who are always or never eligible and not useful for analysis

General Setting

We have a model of $y_i = \beta x_i + \varepsilon_i$ for a fixed population $i = 1 \dots N$

- In the paper: extensions to heterogeneous effects, other controls, multiple treatments, nonlinear outcome models, panel data...

We have a candidate instrument $z_i = f_i(g, w)$, where g is a vector of shocks; w collects predetermined variables; $f_i(\cdot)$ are known mappings

- Applies to any z_i which can be constructed from observed data
- Nests reduced-form regressions: $x_i = z_i$
- Allows $g = (g_1, \dots, g_K)$ to vary at a different level than i

Assumptions:

1. Shocks are exogenous: $g \perp\!\!\!\perp \varepsilon \mid w$
2. Conditional distribution $G(g \mid w)$ is known (e.g. via randomization protocol or uniform across permutations of g)

Results

- Expected instrument, $\mu_i = \mathbb{E}[f_i(g, w) \mid w] \equiv \int f_i(g, w) dG(g \mid w)$, is the sole confounder generating OVB:

$$\mathbb{E}\left[\frac{1}{N}\sum_i z_i \varepsilon_i\right] = \mathbb{E}\left[\frac{1}{N}\sum_i \mu_i \varepsilon_i\right] \neq 0, \text{ in general}$$

- The *recentered instrument* $\tilde{z}_i = z_i - \mu_i$ is a valid instrument for x_i:

$$\mathbb{E}\left[\frac{1}{N}\sum_i \tilde{z}_i \varepsilon_i\right] = 0$$

- Regressions which control for μ_i also identify β (implicit recentering)
- **Consistency**: follows when $\tilde{z}_i$ is weakly mutually dependent across i
- **Robustness** to heterogeneous treatment effects: $\tilde{z}_i$ identifies a convex avg. of β_i under appropriate first-stage monotonicity
- **Randomization inference** provides exact confidence intervals for β (under constant effects) and falsification tests
- BH characterize the **asy. efficient** recentered IV among all $f_i(\cdot)$

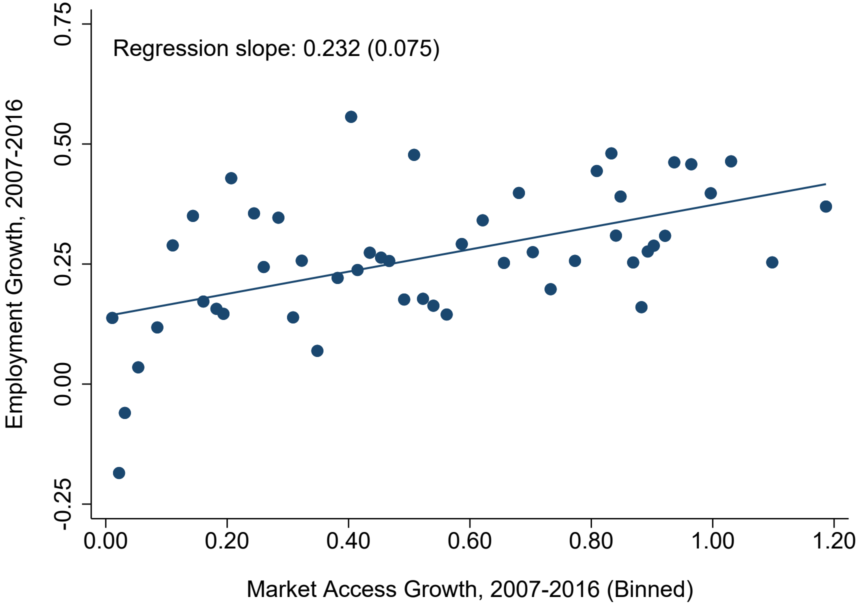
App. 1: Market Access from Chinese High-Speed Rail

BH first show how instrument recentering can address OVB when estimating the effects of market access growth

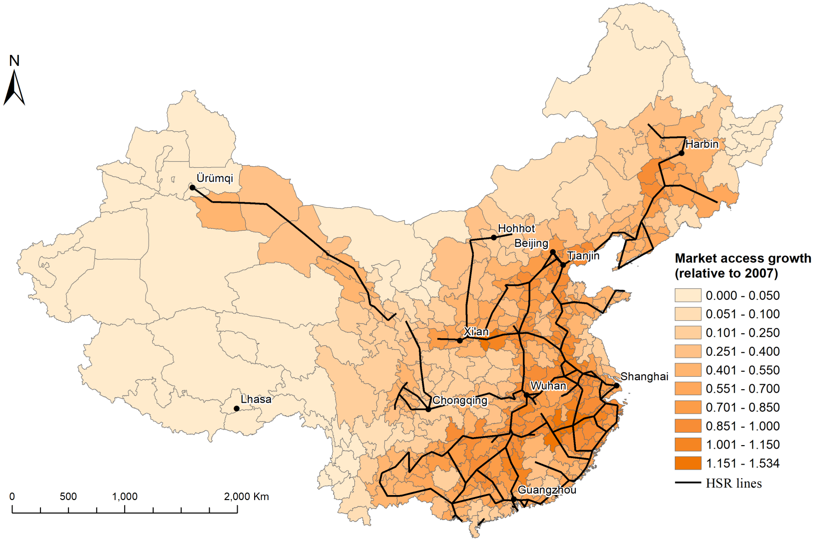
Setting: Chinese HSR; 83 lines built 2008–2016, 66 yet unbuilt

- Market access: $MA_{it} = \sum_k \exp(-0.02\tau_{ikt})\, p_{k,2000}$, where τ_{ikt} is HSR-affected travel time between prefecture capitals (Zheng and Kahn, 2013) and $p_{i,2000}$ is prefecture i's population in 2000
- Relate to employment growth in 274 prefectures, 2007-2016

Conventional OLS Regressions Suggest a Large MA Effect



High vs. Low MA Growth is Not a Convincing Contrast!



How to Find Valid Treatment-Control Contrasts?

Add controls (province FE, longitude, etc.), like Donaldson-Hornbeck 2016

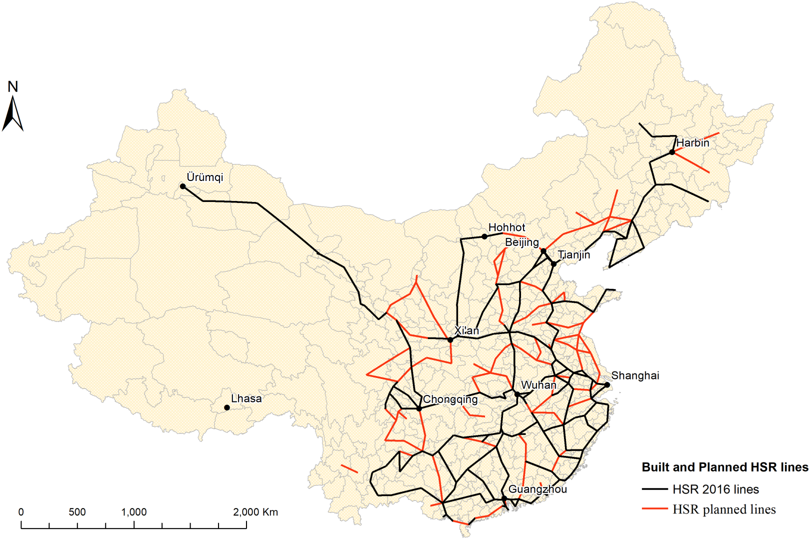
- Hard to justify *ex ante* since MA is a constructed variable
- No experimental analog

Find valid contrasts for *one* source of MA variation—a natural experiment

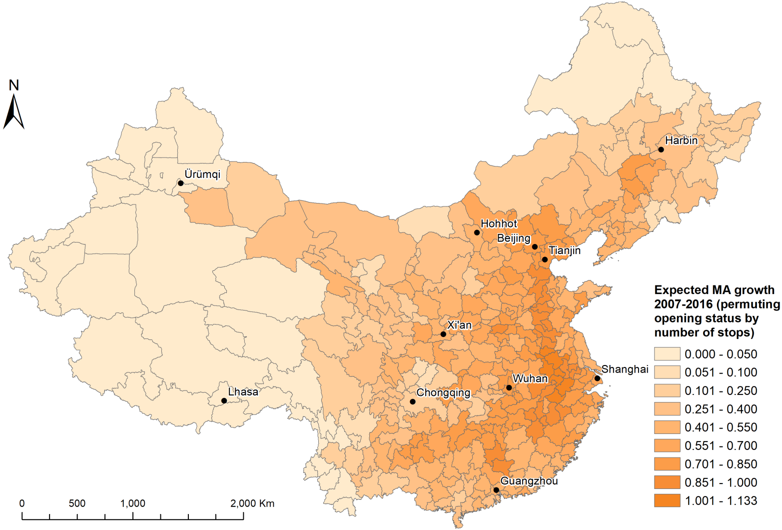
- Bartelme (2018): shocks affecting market size
- Donaldson (2018): built vs unbuilt lines
- BH's application: assume random timing of observably similar lines

Built and Planned HSR Lines

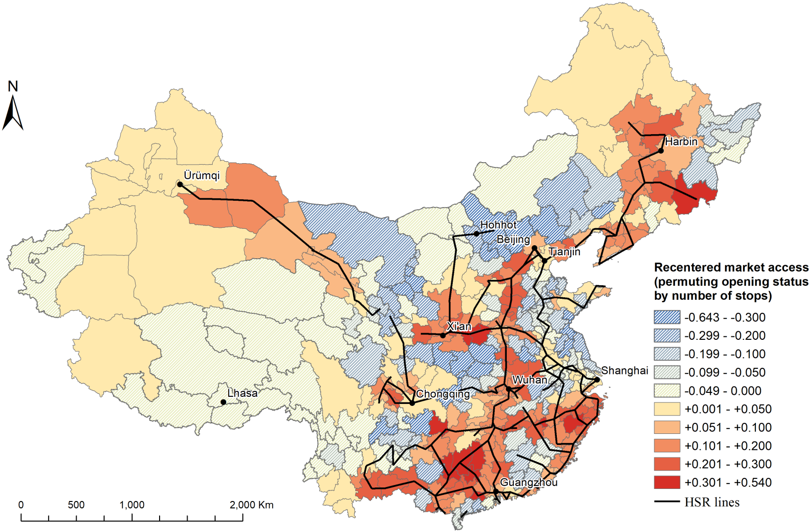
We reshuffle built & planned lines connecting the same # of regions



Expected Market Access Growth (2007–2016), μ_i



Recentered Market Access Growth (2007–2016), $\tilde{z}_i$

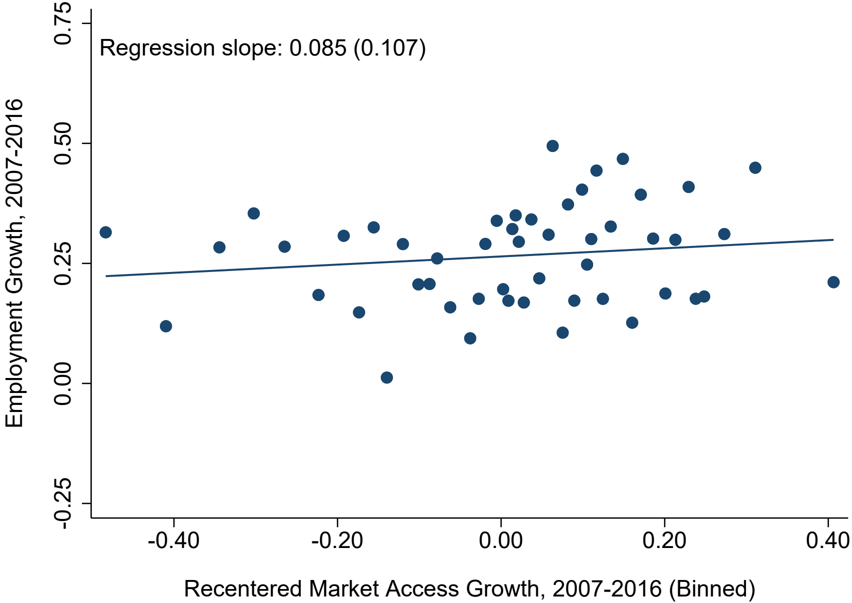


Recentered MA Balance Regressions

	Unadjusted	Recentered		
	(1)	(2)	(3)	(4)
Distance to Beijing	−0.292	0.069		0.089
	(0.063)	(0.040)		(0.045)
Latitude/100	−3.323	−0.325		−0.156
	(0.648)	(0.277)		(0.320)
Longitude/100	1.329	0.473		0.425
	(0.460)	(0.239)		(0.242)
Expected Market Access Growth			0.027	0.056
			(0.056)	(0.066)
Constant	0.536	0.014	0.014	0.014
	(0.030)	(0.018)	(0.020)	(0.018)
Joint RI p-value		0.489	0.807	0.536
R^2	0.823	0.079	0.007	0.082
Prefectures	274	274	274	274

Regressions of unadjusted and recentered market access growth on geographic features. Spatial-clustered standard errors in parentheses.

Recentered MA Doesn't Predict Employment Growth!



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

App. 2: Efficient Estimation of Medicaid Eligibility Effects

Setting: U.S. Medicaid, partially expanded in 2014 under the ACA

- 19 of 43 states with low Medicaid coverage expanded to 138% FPL
- View expansion decisions as random across states with same-party governors, but not household demographics or pre-2014 policy
- Outcomes: Medicaid takeup and private insurance crowdout

We compare two estimators, both valid under the same assumptions:

- Simulated IV: use state-level variation only; i.e. an expansion dummy
- Recentered IV: predict eligibility from expansion decisions & non-random demographics, and recenter

By using non-random variation, the recentered IV has better first-stage prediction $\implies \approx 3$ times smaller standard errors

Simulated and Recentered IV: First Stage

	(1)	(2)	(3)
Simulated IV	0.851	0.032	
	(0.113)	(0.140)	
	[0.567,1.115]	[-0.254,0.503]	
Recentered IV		0.817	0.972
		(0.171)	(0.015)
		[0.397,1.162]	[0.941,1.014]
Partial R^2	0.022	0.113	0.894
Exposed Sample	N	N	Y
States	43	43	43
Individuals	2,397,313	2,397,313	421,042

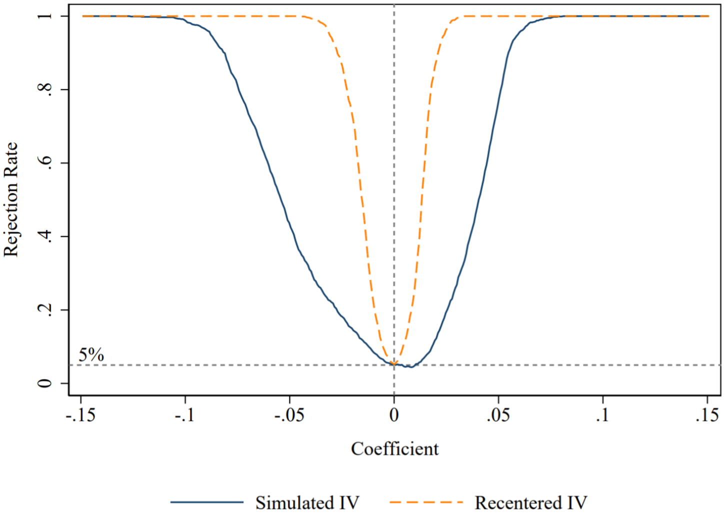
Regressions of Medicaid eligibility on the two instruments, state and year fixed effects, and an indicator for Republican governor interacted with year. State-clustered standard errors in parentheses; Wild score bootstrap 95% CI in brackets

Estimates with Simulated vs. Recentered IV

	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

Simulated and Recentered IV Power Curves



Monte Carlo simulation based on recentered IV estimates. Simulated rejection rates are from nominal 5% tests, using the wild score bootstrap

Other Settings where Recentering Is Relevant

- Network spillovers (Miguel-Kremer 2004, Carvalho et al. 2020)
- Linear shift-share IV (e.g. Autor et al. 2013, Borusyak et al. 2022)
- Nonlinear shift-share IV (e.g. Boustan et al. 2013, Berman et al. 2015, Chodorow-Reich and Wieland 2020, Derenoncourt 2021)
- IV based on centralized school assignment mechanisms (e.g. Abdulkadiroğlu et al. 2017, 2019, Angrist et al. 2020)
- Model-implied optimal IV (Adão-Arkolakis-Esposito 2021)
- Weather instruments (e.g. Gomez et al. 2007, Madestam et al. 2013)
- "Free space" instruments for mass media access (e.g. Olken 2009, Yanagizawa-Drott 2014)

Practical Takeaways

Recentering is most natural when g comes from an RCT

- Shock counterfactuals come from the experimental protocol
- Controlling for μ_i (or other baseline variables) can increase precision

Should you use recentering in observational data?

- Yes, if you can make a strong *ex ante* case for shock exogeneity and have a way to generate shock counterfactuals (similarly to SSIV)
- Build empirical (*ex post*) support with balance tests, as usual
- Multiple candidate μ_i can be included from different guesses at the shock assignment process: only one needs to be right to avoid OVB

As with SSIV, parallel trends gives an alternative path

- Harder to know what controls to include (same with any model-based identification strategy), but at least can run *ex post* pre-trend checks