Part D: Instrumental Variables

D3: Shift-Share and Other Formula Instruments

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2023

Constructed ("Formula") instruments

- So far we have considered IVs that can be viewed as-good-as-randomly assigned
- But some IVs are more complex:
  - constructed from multiple sources of variation
  - some as-good-as-random, some are not
- Why/when would one do that?
- What are the assumptions for IV validity?
- How to do this correctly?

D3 Outline

1. Structure and examples of shift-share IVs
2. SSIV as leveraging a shock-level natural experiment
3. SSIV as combining diff-in-diffs
4. Formula instruments and recentering

Readings: Borusyak, Hull, Jaravel (2022), Borusyak and Hull (forthcoming)

Shift-share IV

- General structure of **shift-share IV** (SSIV):

$$Z_i = \sum_{k=1}^{K} S_{ik} g_k$$

  - $g_1, \ldots, g_K$ are a set of common **shocks** (or **shifts**) not specific to i
  - S_{ik} are exposure **shares**, often with $\sum_k S_{ik} = 1$ for all i

SSIV examples: Bartik instrument

- Consider estimating the inverse elasticity of regional labor supply: $Y_i = \tau D_i + \varepsilon_i$
  - $Y_i =$ log-change in region i's average wage
  - $D_i =$ log-change in i's employment over some period
  - $\varepsilon_i =$ labor supply shocks (e.g. migration, UI benefits)
- Need a region labor demand shock as IV
- Labor demand comes from different industries k: $D_i \approx \sum_k S_{ik} D_{ik}$
  - $S_{ik} =$ initial share of k in i's employment
  - $D_{ik} =$ log-change of k-specific employment in i
- Build a SSIV $Z_i = \sum_k S_{ik} g_k$ based on some shocks g_k that do not vary by region:
  - Observed growth rates of industry employment $\Rightarrow$ **Bartik (1991) instrument**
  - Or specific industry labor demand shifts, e.g. change in import tariffs
  - $Z_i =$ prediction for D_i using some shocks and initial exposure shares

SSIV examples: Enclave instrument

- "Enclave instrument" for migration (e.g. Card 2009):
    - τ = inverse elasticity of substitution between native and immigrant labor
    - Y_i = change in relative immigrant/native wage
    - D_i = change in relative immigrant/native employment
    - Need relative labor supply shock as an IV
    - New immigrants from country k tend to go where there are historic enclaves of k's immigrants
    - Z_i = migration intensity prediction from historic enclaves & national inflows or "push shocks"
    - S_{ik} = initial share of origin k in i's population
    - g_k = observed national migration growth from k (Card 2009) or dummy of war in k (Llull 2017)

SSIV examples: Spillovers

- Miguel and Kremer (2004): spillover effects of randomized deworming in Kenya
  - Y_i = educational achievement of student i
  - D_i = the number of i's neighbors (students who go to school within a certain distance from i) who have been dewormed
  - Use OLS: $Z_i = D_i$. Not usually understood as a shift-share design, but it is
  - $S_{ik} = ?$
  - $g_k = ?$

SSIV examples: China shock

- Autor, Dorn, Hanson (ADH, 2013): effect of import competition with China on regional labor markets in the US
    - Y_i = growth of manufacturing employment rate, unemployment rate, etc.
    - D_i = growth of import competition in region i (imports per US worker)
    - Endogeneity: D_i is affected by low productivity & demand in i
    - $Z_i = \sum_k S_{ik} g_k$ = predicted growth of import competition
    - S_{ik} = 10-year lagged share of manufacturing industry k in i's total employment
    - g_k = growth of industry import competition with China in 8 other countries (e.g. Australia)

Identification approaches

- Relevance makes sense: $Z_i = \sum_k S_{ik} g_k$ predicts $D_i = \sum_k S_{ik} D_{ik}$ if g_k predicts D_{ik}
- But how should think about exogeneity, $\mathbb{E}\left[\frac{1}{N}\sum_i Z_i \varepsilon_i\right] = 0$?
  - *Note:* notation for non-random samples
- Two narratives + sets of sufficient conditions:
  - "**Exogenous shocks**" (Borusyak, Hull, Jaravel; BHJ, 2022):
    shock-level natural experiment, translated to the observation level
  - "**Exogenous shares**" (Goldsmith-Pinkham, Sorkin, Swift, 2020):
    combining diff-in-diffs in heterogeneous exposure

Outline

Simple shock-level regressions

- Acemoglu, Autor, Dorn, Hanson, Price (2016) study the effect import competition with China on employment across $K \approx 400$ *industries*
  - IV with g_k: import competition with China in 8 other countries
  - A natural experiment: China growth is as-if random across industries

$$\mathbb{E}\left[g_k \mid \varepsilon_k, \text{pre-trends}, \text{industry characteristics}, \ldots\right] = \theta \qquad \text{for all } k$$

  - With covariates q_k, e.g. dummies of 10 broad sectors (electronics, food, etc.):

$$\mathbb{E}\left[g_k \mid \varepsilon_k, q_k, \text{pre-trends}, \text{industry characteristics}, \ldots\right] = \theta' q_k \qquad \text{for all } k$$

  (where q_k always includes an intercept)

- Why aggregate to regional level?

Spillovers

- ADH and Acemoglu et al. (2016) estimate different economic parameters
- Import competition in k can reallocate workers to other industries $\Rightarrow$ SUTVA violation
- If regional economies are isolated islands, SUTVA holds in ADH
    - Otherwise can incorporate spillovers across regions, e.g. via trade or migration (Adao, Arkolakis, Esposito 2022)
- Some outcomes are not well-defined at the industry level, e.g. unemployment

Translating to observation level

- Does the shock-level natural experiment imply exogeneity of SSIV?
- Yes, but $Q_i = \sum_k S_{ik} q_k$ must be controlled for
- $q_k = 1$, $S_i \equiv \sum_k S_{ik} = 1$ (**complete shares**): $Q_i = 1$
  - Weighted average of as-good-as-random shocks is as-good-as-random
  - Even if (lagged) shares are endogenous: $\mathrm{Cov}\left[S_{ik}, \varepsilon_i\right] \neq 0$
    - E.g. if $\varepsilon_i = \sum_k S_{ik}\nu_k + \tilde{\varepsilon}_i$ where ν_k are some other industry shocks (say, automation)

Translating to observation level (2)

- $q_k = 1$, $S_i \neq 1$ (**incomplete shares**): $Q_i = \sum_k S_{ik} q_k = \sum_k S_{ik} = S_i$
  - Must control for the sum of exposure shares
  - In ADH, k are manufacturing industries (no China competition for services)
  - $S_i =$ initial manufacturing share in region i
  - Z_i is mechanically correlated with S_i
  - Is that a problem? Import competition does grow more in manuf.-heavy regions
  - But $\mathrm{Cov}\left[S_i, \varepsilon_i\right] \neq 0$ via any reason for overall manuf. decline, other than China
  - Correct specification $\neq$ no OVB!
- $q_k =$ dummies for broad sectors, $S_i = 1$
  - Control for $Q_i =$ initial employment shares in each broad sector
  - To translate within-sector variation in g_k to the regional level

BHJ equivalence result

BHJ (Prop. 1): Consider SSIV estimator $\hat{\tau}$ from

$$Y_i = \tau D_i + \gamma' X_i + \varepsilon_i$$

instrumenting D_i by $Z_i = \sum_k S_{ik} g_k$ and controlling for X_i that include $Q_i = \sum_k S_{ik} q_k$.

This $\hat{\tau}$ can be obtained from a shock-level IV regression

$$\bar{y}_k^{\perp} = \tau \bar{d}_k^{\perp} + \theta' q_k + \bar{\varepsilon}_k,$$

- instrumenting $\bar{d}_k^{\perp}$ by g_k
- weighted by $s_k = \frac{1}{N} \sum_i S_{ik}$ capturing the average importance of shock k
- where $\bar{v}_k = \sum_i S_{ik} V_i / \sum_i S_{ik}$ are exposure-weighted averages of V_i
  - e.g. $\bar{\varepsilon}_k$ is average residual of observations i with a high exposure to k
- and $V_i^{\perp}$ are residuals from regressing V_i on X_i (in the sample)

BHJ equivalence result: Proof

- Proof by exchanging the order of summation:

$$\hat{\tau} = \frac{\sum_i Z_i Y_i^{\perp}}{\sum_i Z_i D_i^{\perp}} = \frac{\sum_{i,k} S_{ik} g_k Y_i^{\perp}}{\sum_{i,k} S_{ik} g_k D_i^{\perp}} = \frac{\sum_k g_k \sum_i S_{ik} Y_i^{\perp}}{\sum_k g_k \sum_i S_{ik} D_i^{\perp}}$$

$$= \frac{\sum_k g_k \sum_i S_{ik} Y_i^{\perp}}{\sum_k g_k \sum_i S_{ik} D_i^{\perp}} = \frac{\sum_k s_k g_k \bar{y}_k^{\perp}}{\sum_k s_k g_k \bar{d}_k^{\perp}} = \frac{\sum_k s_k (g_k - \hat{\theta}' q_k) \bar{y}_k^{\perp}}{\sum_k s_k (g_k - \hat{\theta}' q_k) \bar{d}_k^{\perp}}$$

where the last equality holds because, when X_i includes Q_i,

$$\sum_k s_k q_k \bar{v}_k^{\perp} = \frac{1}{N} \sum_{i,k} S_{ik} q_k V_i^{\perp} = \frac{1}{N} \sum_i Q_i V_i^{\perp} = 0$$

SSIV consistency

- Since one can view SSIV as using g_k as the IV, as-good-as-random assignment of g_k implies consistency of $\hat{\tau}$
  - Specifically, g_k should not correlate with $\bar{\varepsilon}_k$ (controlling for q_k):
  - In ADH, $\bar{\varepsilon}_k$ is unobserved determinants of regional employment, averaged among regions with a high employment share of k
- **BHJ (Prop. 4):** $\hat{\tau}$ is consistent for the (constant-effect) τ if
  1. $\mathbb{E}\left[g_k \mid \bar{\varepsilon}, \boldsymbol{q}, \boldsymbol{S}\right] = q_k'\theta$ for some θ (conditionally as-good-as-random shocks)
  2. $\mathbb{E}\left[\sum_k s_k^2\right] \to 0$ (many shocks with dispersed average exposure)
  3. $\mathrm{Cov}\left[g_k, g_{k'} \mid \bar{\varepsilon}, \boldsymbol{q}, \boldsymbol{S}\right] = 0$ for $k \neq k'$ (uncorrelated shocks)
  4. $\frac{1}{N}\sum_i D_i Z_i \xrightarrow{p} \pi \neq 0$ (relevance)
     - Typical i should have concentrated shock exposure (but to different shocks across i)
- If you can use g_k as IV (**exogenous shocks**), you can use it in SSIV across i, too!

Exposure-robust inference

- Complication: observations with similar shares are exposed to the same shocks, both g_k and unobserved ν_k
  - Conventional clustering of SE wouldn't capture that (e.g. by state or Conley spatial clustering)
- Adao, Kolesar, Morales (2019) derive corrected formula
  - Leverages independence of g_k, regardless of correlations in ε_i
- BHJ show SE from the shock-level equivalent regression are valid
  - Convenient solution, directly extends to autocorrelation, spatial clustering, etc.
  - In Stata and R, package *ssaggregate* does the conversion

Extensions

- "Estimated shocks"
  - Things are more complicated when g_k is an equilibrium object (e.g. national employment growth rate by industry or migration inflow by origin country)
- Panel data
  - In panels, exogenous shock variation can come from the cross-section *or* the time series (Nakamura and Steinsson 2014, Nunn and Qian 2014)
- Heterogeneous effects
  - LATE logic goes through, even if Z_i is misspecified (but D_i is specified correctly)

Application: ADH

- Region i = commuting zone ($N = 722$)
- Industry k = SIC4 manufacturing industry ($K = 397$)
- Two periods t: 1991–2000 and 2000–2007
- Y_{it} = local change in manufacturing employment rate
  - D_{it} = local growth of Chinese imports in \$1,000/worker
- X_{it} include period FE and (non-lagged) total manufacturing share
- $Z_{it} = \sum_k S_{ikt} g_{kt}$ where:
  - S_{ikt} = lagged share of k in total employment of i; $\sum_k S_{ikt}$ = lagged total share of manufacturing in employment
  - g_{kt} = growth of Chinese imports in eight non-US countries in \$1,000/US worker
- If q_{kt} = period FE, Q_{it} = ?

Application: ADH

- Region i = commuting zone ($N = 722$)
- Industry k = SIC4 manufacturing industry ($K = 397$)
- Two periods t: 1991–2000 and 2000–2007
- Y_{it} = local change in manufacturing employment rate
    - D_{it} = local growth of Chinese imports in \$1,000/worker
- X_{it} include period FE and (non-lagged) total manufacturing share
- $Z_{it} = \sum_k S_{ikt} g_{kt}$ where:
    - S_{ikt} = lagged share of k in total employment of i; $\sum_k S_{ikt}$ = lagged total share of manufacturing in employment
    - g_{kt} = growth of Chinese imports in eight non-US countries in \$1,000/US worker
- If q_{kt} = period FE, $Q_{it} = \sum_k S_{ikt} q_{kt}$ = period FE $\times$ lagged total manuf. share

BHJ revisit ADH

Balance tests to verify conditional as-good-as-random shock assignment:

- Shocks are uncorrelated with industry observables, controlling for period FE
- SSIV is uncorrelated with regional observables, controlling for period FE $\times$ lagged total manuf. share

Balance variable	Coef.	SE
Panel A: Industry-level balance		
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)
No. of industry-periods	794	
Panel B: Regional balance		
Start-of-period % of college-educated population	0.915	(1.196)
Start-of-period % of foreign-born population	2.920	(0.952)
Start-of-period % of employment among women	−0.159	(0.521)
Start-of-period % of employment in routine occupations	−0.302	(0.272)
Start-of-period average offshorability index of occupations	0.087	(0.075)
Manufacturing employment growth, 1970s	0.543	(0.227)
Manufacturing employment growth, 1980s	0.055	(0.187)
No. of region-periods	1,444	

BHJ revisit ADH

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 (0.114)	−0.489 (0.100)	−0.267 (0.099)	−0.314 (0.107)	−0.310 (0.134)	−0.290 (0.129)	−0.432 (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

- Adding Q_{it} changes the estimate: China shock g_{kt} is larger in the 2000s (post WTO entry) when overall manuf. decline is stronger for other reasons

Outline

1. Structure and examples of shift-share IVs
2. SSIV as leveraging a shock-level natural experiment
3. SSIV as combining diff-in-diffs
4. Formula instruments and recentering

Brazilian trade liberalization

- Dix-Carneiro and Kovak (2016) study the long-run labor market effects of Brazilian trade liberalization in early 1990s, by OLS in a cross-section of regions:

$$Y_{i,\text{post}} - Y_{i,\text{pre}} = \tau D_i + \text{controls} + \varepsilon_i, \qquad D_i = Z_i = \sum_k S_{ik} g_k$$

  - $K = 20$ tradable industries (agriculture + 19 manuf. industries)
  - $S_{ik} \approx$ pre-period employment shares relative to total tradable employment
  - g_k is change in tariffs (> 0 in agriculture, < 0 in all manuf. industries)
- Narrative for OLS validity?
  - *"Along with regional differences in industry mix, the cross-industry variation in tariff cuts provides the identifying variation"*
  - Tariff cuts are driven by heterogeneity in initial levels from 1957
- Could the exogenous shocks approach be used?

Not a natural experiment in shocks

- Only 20 industries
- Agriculture is ≈40% of employment (Herfindahl $\sum_k s_k^2$ is large)
- Because of how tariffs changed, D_i has a 99% correlation with $S_{i,\text{agriculture}}$
  - Essentially a DiD with continuous treatment intensity $S_{i,\text{agriculture}}$
  - Should be justified by parallel trends, not a natural experiment in shocks
- Goldsmith-Pinkham, Sorkin, Swift (GPSS, 2020) develop this view

Exogenous shares approach

- Assume **exogenous shares**: $\mathrm{Cov}[\varepsilon_i, S_{ik}] = 0$ for every k
  - With Y_i measured in differences, this is PTA (K times)
  - Strong assumption even though shares are measured in the pre-period
  - Wrong: *"shares are not affected by ε_i"* (they can't be)
  - Correct: *"all unobservables are uncorrelated with everything about local shares"*
  - Rules out any unobserved ν_k shocks that affect regions based on S_{ik}
- Then we have K valid IVs: $S_{i1}, \ldots, S_{iK}$
  - SSIV $Z_i = \sum_k S_{ik} g_k$ is just a reasonable way to combine them
  - 2SLS (for small K) and LIML are other reasonable ways
  - Or just using your favorite share (e.g. of agriculture)
  - GPSS prove a numerical equivalence: SSIV estimator is GMM with $S_{i1}, \ldots, S_{iK}$ as IVs and a weight matrix that depends on g_k

Rotemberg weights

- If you insist on using SSIV (and not LIML), GPSS recommend computing **Rotemberg weights** $\hat{\alpha}_k$:
  - $\hat{\tau} = \sum_k \hat{\alpha}_k \hat{\tau}_k$ for $\hat{\tau}_k$ that uses S_{ik} as IV one at a time
  - $\hat{\alpha}_k$ are higher for k with more extreme shocks and larger first stages
  - $\hat{\alpha}_k$ add up to one but need not be positive
- Then scrutinize validity of the share IVs with highest Rotemberg weights

Summary

- Two sets of narratives & formal conditions for SSIV validity
  - Pick one *ex ante*, then validate *ex post*
- Exogenous shocks is appropriate when you could imagine using your shocks as IVs in some shock-level analysis
  - Check balance at the shock level
  - Include share-aggregated controls (especially with incomplete shares)
  - Use exposure-robust inference
- Exogenous shares is appropriate when you would be OK using any other combination of shares as the IV
  - Scrutinize share IVs with high Rotemberg weights
  - Report LIML (or just switch to it). Run overidentification test (w/ usual caveats)
- Pre-trend & balance tests no SSIV at the observation level are useful in both cases

Outline

Formula treatments and instruments

- SSIVs are only a special case of treatments and instruments constructed from multiple sources of variation
- Let's develop an instinct to:
  - Identify settings that are in this class
  - Ask which determinants are as-good-as-random and which are non-random
  - Understand what it means to call your shocks as-good-as-random, by thinking of counterfactuals shocks
  - Recognize that OVB is possible even with as-good-as-random shocks
  - Know how to fix OVB, via "recentering"
  - Have no fear of designs with "Non-Random Exposure to Exogenous Shocks" (following Borusyak and Hull (forthcoming) and related work)

Example 1: Miguel and Kremer (2004)

- $D_i = Z_i =$ the number of kid i's dewormed neighbors
- Implicitly constructed from two sources of variation: who neighbors whom S_{ik} and who gets dewormed g_k
- g_k are as-good-as-random, S_{ik} are non-random (potentially correlated with errors)
- g_k were randomized according to some randomization protocol: say, stratified by gender
    - We can rerun the protocol many times and see which sets of kids could as likely have been dewormed instead
- OVB is still possible: Z_i is mechanically correlated with the numbers of male neighbors and female neighbors
- OVB is fixed by controlling for this number of neighbors of each gender

Example 2: Nonlinear spillovers

- Now suppose $D_i = Z_i =$ dummy of having at least one dewormed neighbor

- Constructed from the same S_{ik} and g_k but nonlinear in $(g_1, \ldots, g_K)$:

$$Z_i = \max_k S_{ik} g_k$$

- What could cause OVB here?

- How to fix it?

Example 3: Effects of transportation

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

$$\Delta Y_i = \tau \Delta \log MA_i + \varepsilon_i$$

$$\text{where } MA_{it} = \sum_{k=1}^{N} \text{TravelTime}(\text{loc}_i, \text{loc}_k, g_t)^{-1} \text{Pop}_k, \qquad t = 0, 1$$

    - g_t is transportation network
    - loc_k is region's location on the map
    - Pop_k is regional population (assume time-invariant)
    - ε_i is effects of unobserved local shocks (e.g. amenities or productivity)
- Consider best-case scenario of "exogenous transportation shocks"
    - At $t = 0$ no transportation; at $t = 1$ roads are built in a RCT
    - Randomizing the network $\not\Longrightarrow$ as-good-as-random $\Delta \log MA_i$

Illustration: Market access on a square island

Start from no roads, assume $\text{Pop}_i = 1$ everywhere $\Longrightarrow \log MA_{i0} = \log MA_{i1} = 0$

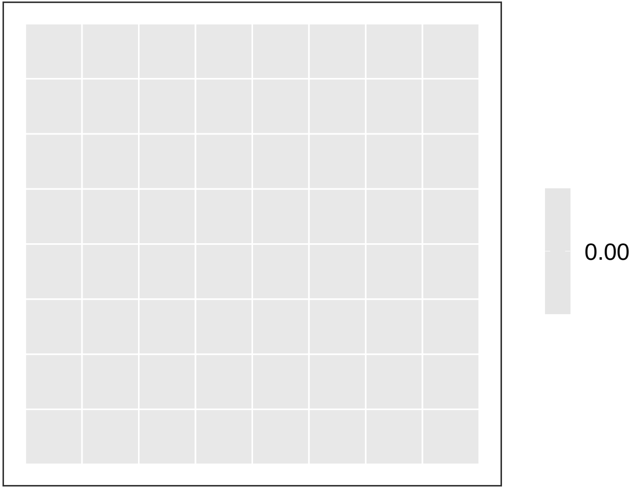


Illustration: Market access on a square island

Randomly connect adjacent regions by road

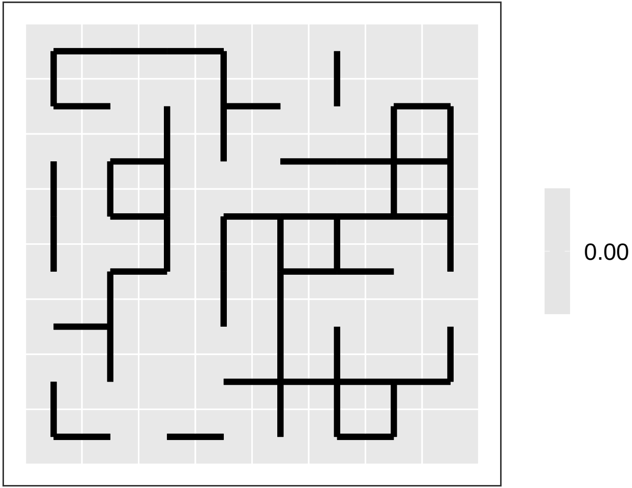
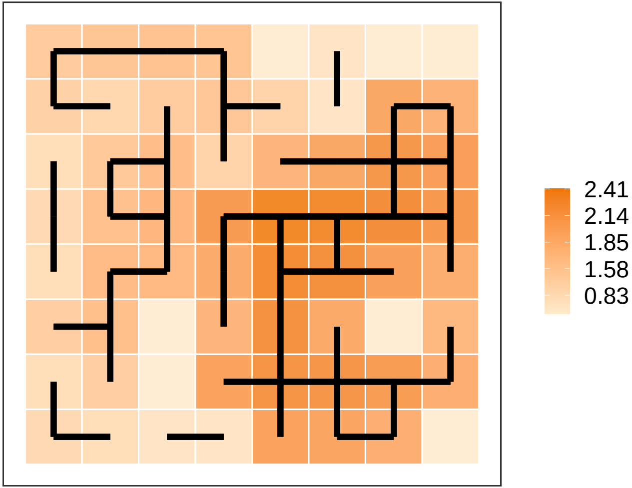


Illustration: Market access on a square island

Get variation in $\Delta \log MA_i$. Is it as-good-as-random?



OVB problem

- No! Market access growth is systematically higher in the center
  - Central regions have higher "propensity" to be near random lines
  - And could have systematically different ε_i, leading to OVB
- Can we measure *i*'s propensity to get MA growth from random lines? Yes!
  - Simulate random counterfactual networks $g^{(s)}$ for many $s = 1, \ldots, S$, holding $w = (\text{loc}_k, \text{Pop}_k)_{k=1}^K$ fixed;
  - Compute $\Delta \log MA_i \left(g^{(s)}; w\right)$ by the formula;
  - Average across simulations to get **expected MA growth**

$$\mu_i(w) = \mathbb{E}\left[\Delta \log MA_i \left(g^{(s)}; w\right) \mid w\right] \approx \frac{1}{S} \sum_s \Delta \log MA_i \left(g^{(s)}; w\right)$$

Illustration: Market access on a square island

$\Delta \log MA_i$ in a random **counterfactual** network draw

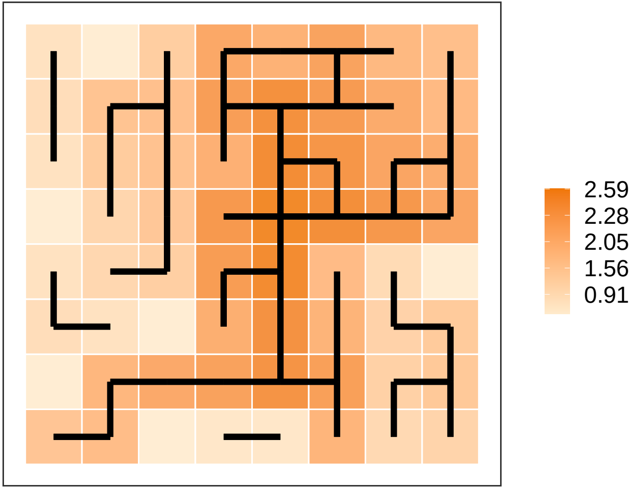


Illustration: Market access on a square island

Yet another counterfactual network draw

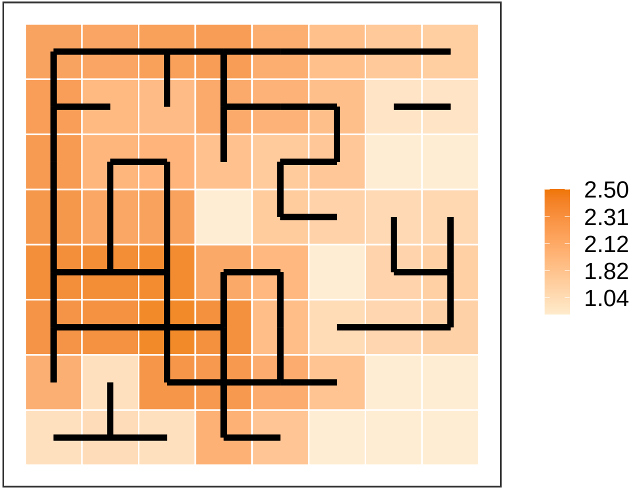
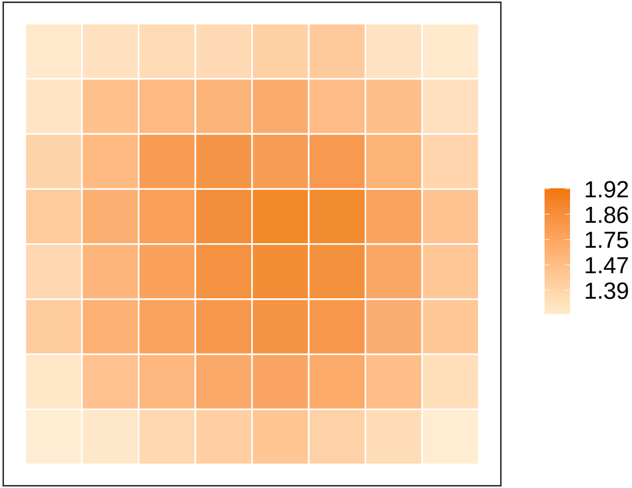


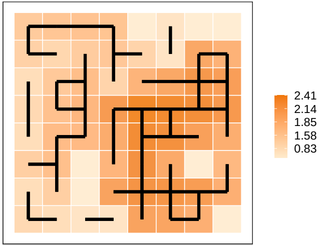
Illustration: Market access on a square island

Average across 1,000 draws: expected MA growth μ_i

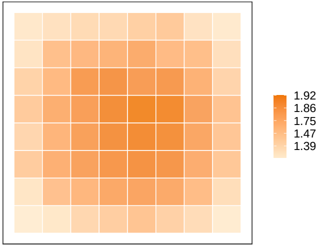


How to use μ_i?

Actual $\Delta \log MA_i(g; w)$

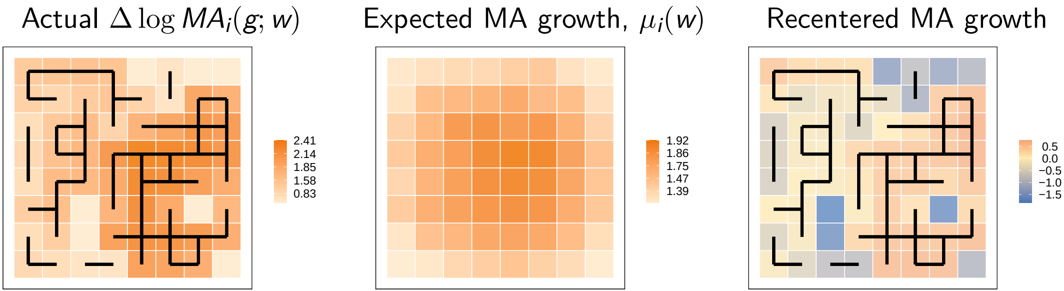


Expected MA growth, $\mu_i(w)$



- How to avoid OVB? Can regress ΔY_i on $\Delta \log MA_i$ by OLS, controlling for μ_i
- Or instrument $\Delta \log MA_i$ by **recentered** MA growth, $\tilde{Z}_i = \Delta \log MA_i - \mu_i$

How to use μ_i?



- How to avoid OVB? Can regress ΔY_i on $\Delta \log MA_i$ by OLS, controlling for μ_i
- Or instrument $\Delta \log MA_i$ by **recentered** MA growth, $\tilde{Z}_i = \Delta \log MA_i - \mu_i$

How recentering & controlling corrections work

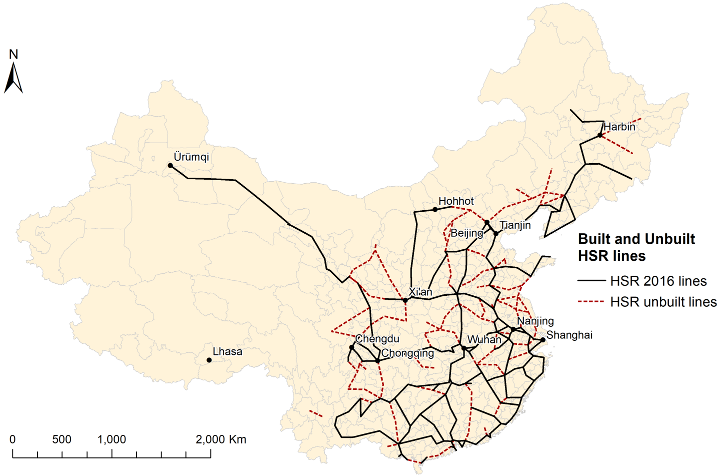
- Recentering reduced-form: ΔY_i on $\tilde{Z}_i = \Delta \log MA_i - \mu_i$
  - Treated group: regions that got more MA growth than expected because certain connections got built and not others
  - Control group: regions with less MA growth than expected
  - Valid if realized and counterfactual networks are equally likely
- First stage: should be close to 1. Why?
- Controlling approach: OLS of ΔY_i on $\Delta \log MA_i$ controlling for μ_i
  - Same using recentered IV + controlling for μ_i
  - Can help efficiency by removing some variation from ε_i — like any other predetermined control (e.g. coordinates or initial MA level)

Recentering in practice

- What if shocks don't come from an RCT?
- Researcher claiming a natural experiment should specify shock counterfactuals they have in mind
  - Defines a natural experiment, as opposed to a quasi-experiment (as in diff-in-diffs)
- BH study the effects of Chinese high-speed railways (HSR) on employment growth
  - Observe 149 *planned* HSR lines: 83 open by 2016 and 66 don't
  - Assume *timing* of opening is random within groups of similar lines
  - Generate counterfactual networks by reshuffling opening status of planned lines within groups

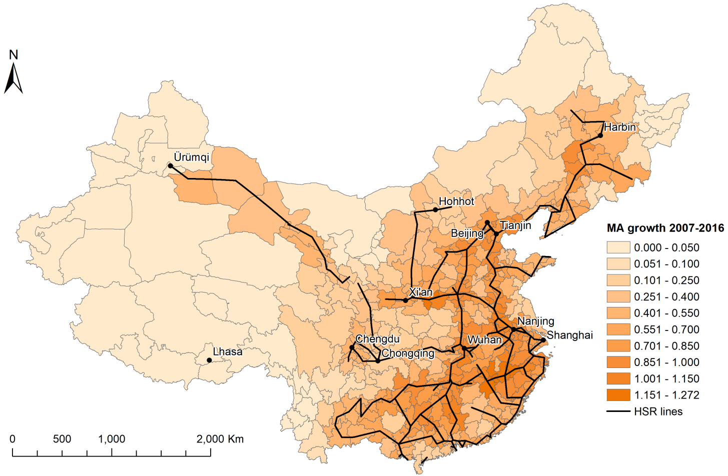
HSR application

Planned HSR lines



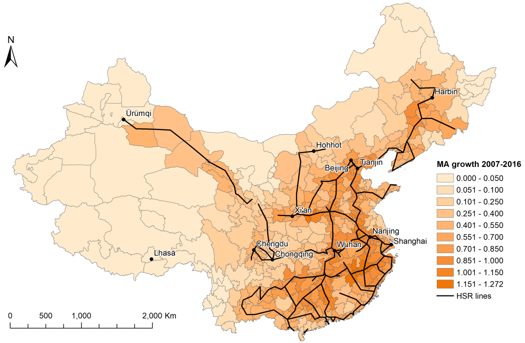
HSR application

Actual network and MA growth

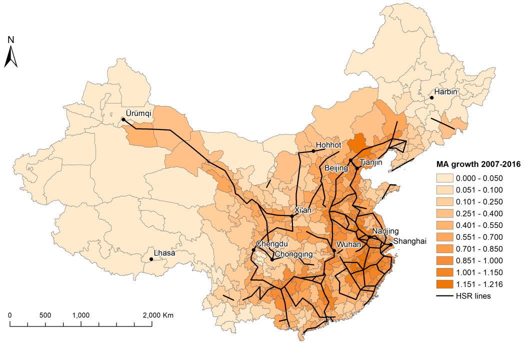


HSR application

Actual 2016 network

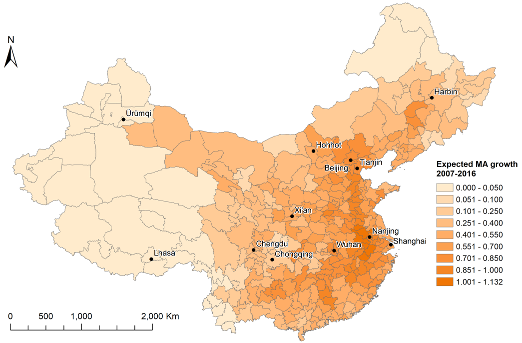


Example counterfactual 2016 network

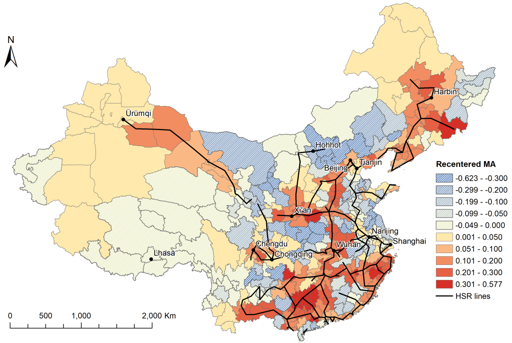


HSR application

Expected MA growth



Recentered MA growth



BH's formal framework

- Outcome equation $Y_i = \tau D_i + \varepsilon_i$ (for a fixed sample)
  - Extensions to heterogeneous effects, other controls, multiple treatments, nonlinear outcome models, panel data
- Consider a candidate instrument $Z_i = f_i(g; w)$, where $g = (g_1, \ldots, g_K)$ are shocks, w collects predetermined variables, f_i are known formulas
  - Nests reduced-form regressions when $D_i = Z_i$
- Assume shock exogeneity: $g \perp\!\!\!\perp \varepsilon \mid w$
  - Exclusion: shocks g don't causally affect Y_i other than through D_i
  - Independence: g is assigned independently of potential outcomes, conditionally on w
- Assume conditional distribution $G(g \mid w)$ is known (e.g. via randomization protocol or uniform across some permutations of g)

Identification

- These assumptions imply that the sole confounder generating OVB is the **expected instrument** $\mu_i = \mathbb{E}\left[f_i(g; w) \mid w\right] \equiv \int f_i(g; w) dG(g \mid w)$:

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

- Thus, **recentered instrument** $\tilde{Z}_i = Z_i - \mu_i$ satisfies $\mathbb{E}\left[\frac{1}{N}\sum_i \tilde{Z}_i\varepsilon_i\right] = 0$
- If $\tilde{Z}_i$ is relevant, recentered IV estimator is consistent as long as $\tilde{Z}_i$ are weakly mutually dependent, regardless of mutual correlation in ε_i
- What about inference?
  - Conventional inference restricts dependence of $\tilde{Z}_i\varepsilon_i$
  - Randomization inference leverages shock counterfactuals

Spatially-clustered standard errors

- Conley spatially-clustered standard errors are based on

$$\widehat{Var}\left(\sum_i \tilde{Z}_i \varepsilon_i\right) = \sum_{i,j:\ d(i,j)<d_{max}} \kappa\left(\frac{d(i,j)}{d_{max}}\right) \cdot \tilde{Z}_i \hat{\varepsilon}_i \hat{\varepsilon}_j \tilde{Z}_j'$$

  - $d(i,j)$ is geographic distance
  - d_{max} is the distance cutoff such that $\mathrm{Cov}\left[\tilde{Z}_i \varepsilon_i, \tilde{Z}_j \varepsilon_j\right] = 0$ if $d(i,j) > d_{max}$
  - $\kappa(\cdot)$ is a kernel function:
    - ★ Uniform kernel: $\kappa(x) = \mathbf{1}\left[|x| \leq 1\right]$
    - ★ Bartlett kernel: $\kappa(x) = \max\left\{1 - |x|, 0\right\}$

Randomization inference

- To test the **sharp null** $\tau = b$ (assuming constant effects), compute statistic

$$T(g) = \frac{1}{N}\sum\nolimits_i (Y_i - bD_i)\,(f_i(g;w) - \mu_i(w))$$

- For many simulated counterfactual shocks $g^{(s)}$, compute

$$T(g^{(s)}) = \frac{1}{N}\sum\nolimits_i (Y_i - bD_i)\left(f_i(g^{(s)};w) - \mu_i(w)\right)$$

- Check that $T(g)$ is not in the tails of the distribution of $T(g^{(s)})$
  - If $\tau = b$ holds, no reason for ε_i to correlate with more $f_i(g,w)$ than $f_i(g^{(s)},w)$
  - But if $\tau \neq b$, $T(g^{(s)})$ are centered around 0 while $T(g)$ is not
- Tests and confidence intervals are valid in finite samples, with no assumptions on ε
- This statistic is natural but any statistic $T(g; Y - bD, w)$ would work, too

Almost done

✓ We tried to develop an instinct to:

- Identify settings that are in formula instruments class
- Ask which determinants are as-good-as-random and which are non-random
- Understand what it means to call your shocks as-good-as-random, by thinking of counterfactuals shocks
- Recognize that OVB is possible even with as-good-as-random shocks
- Know how to fix OVB, via "recentering"

$\rightarrow$ Final task: have no fear of designs with non-random exposure to exogenous shocks

Example 4: Simulated instruments

- Curie and Gruber (1996a,b) study the effects of Medicaid eligibility on health outcomes
- OLS is surely biased because richer households are less likely to be eligible
- Assume variation in eligibility policy across states is exogenous
  - But policy is a complicated object: set of eligibility rules
  - Construct a scalar measure of policy generosity as IV
  - "**Simulated instrument**": % of population nationally that would be eligible under policy of *i*'s state

Example 4: Simulated instruments

- What do you think of the simulated instrument: Exogeneity? Relevance?
- How can we recast household i's Medicaid eligibility D_i as a formula treatment?
- What is a household's expected eligibility?
- What does recentering / controlling for it mean here?
- What if $D_i =$ Medicaid takeup, rather than eligibility?

Application to Obamacare

- Borusyak and Hull (2021) estimate crowding-out effects of Medicaid takeup (D_i) on private health insurance (Y_i)
- Leverage eligibility expansions to 146% of FPL under the Affordable Care Act
  - 11 of 13 states with Democratic governor, 8 of 30 states with Republican governor
  - View expansion decisions as random across states with same-party governors, but not household demographics or pre-2014 policy
- Compare two IVs:
  - Simulated IV: expansion dummy (controlling for governor's party)
  - Recentered IV: predict eligibility from expansion decisions & non-random demographics, and recenter
- By not fearing non-random exposure, recentered IV has much better first-stage
  - $\sim$2x smaller standard errors