Part D: Instrumental Variables

D3: Shift-Share and Other Formula Instruments

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2024

Constructed ("Formula") instruments

- So far we have considered IVs that can be viewed as-good-as-randomly assigned
- But many IVs (and RHS variables in OLS) are:
    - constructed from multiple sources of variation using a formula/algorithm
    - with some as-good-as-random sources of variation, while others are not

D3 Outline

1. Formula instruments and recentering
    - Example: parametric spillover effects
    - General setting
    - Application: Chinese high-speed railways
    - Other examples
2. Shift-share instruments
    - SSIV as leveraging a shift-level natural experiment
    - SSIV as combining diff-in-diffs

Readings: Borusyak, Hull (Ecma 2023), Borusyak, Hull, Jaravel (ReStud 2022; Econometrics Journal 2024 review)
See also Borusyak, Hull, Jaravel (2024) "A Practical Guide to Shift-Share Instruments"

Example: Miguel and Kremer (2004)

- Spillover effects of randomized deworming in Kenya
  - Y_i = educational achievement of student $i \in \mathcal{I}$
  - D_i = the number of i's neighbors (students who go to school within a certain distance from i) within $\mathcal{I}$ who have been dewormed
- Implicitly constructed from two sources of variation:
  - Who neighbors whom: $w_{ik} = \mathbf{1}\left[k \text{ is a neighbor of } i\right]$
  - Who randomly got dewormed: $g_k = \mathbf{1}\left[k \text{ was dewormed}\right]$
  - We can express as a "formula": $D_i = \sum_{k=1}^{N} w_{ik} g_k$
- In this RCT, is it correct to estimate $Y_i = \tau D_i + \varepsilon_i$ by OLS?
- What can we do to avoid OVB?

Variation 1: Nonlinear spillovers

- Now suppose $D_i =$ dummy of having at least one dewormed neighbor
  - Can we represent it as a formula in terms of the same w_{ik} and g_k? $D_i = \max_k w_{ik} g_k$
- Suppose girls were dewormed with prob. 50%, while boys with prob. 25%
- What could cause OVB here?
- What should we control for?

Variation 2: Imperfect takeup

- Suppose again $D_i =$ number of dewormed friends
  - But only offer g_k was randomized, while some kids didn't take up
- Can we represent D_i as a formula in terms of w_{ik} and new g_k?
- What would you do?

Answer key: Miguel and Kremer

- In this RCT, is it correct to estimate $Y_i = \tau D_i + \varepsilon_i$ by OLS?
  - No! Randomizing deworming $\not\Rightarrow$ randomizing the number of dewormed neighbors
  - Kids in denser areas will systematically have more dewormed neighbors (higher D_i)
  - They may also have systematically different ε_i, leaving to OVB
- How to fix this? Control for the total number of neighbors
  - With that control, randomization of g_k implies as-good-as-random D_i
- Alternatively, instrument D_i with D_i − Total # of neighbors$_i$ × Prob of deworming

Answer key: Nonlinear spillovers

- We can represent the treatment as an explicit formula, $D_i = \max_k w_{ik} g_k$
    - But it is not necessary: any algorithm $f_i(w, g)$ is already a "formula"
- Two reasons for OVB:
    - More neighbors $\Rightarrow$ higher probability of having a dewormed neighbor
    - More girls among neighbors $\Rightarrow$ same
- Sufficient to control for the probability of having a dewormed neighbor
    - Computed analytically as a function of neighbors of each sex
    - Or just by simulating the randomization protocol many times

Answer key: Incomplete takeup

- D_i is no longer fully determined by $\{w_{ik}\}$ and $\{g_k\}$
  - But we can construct an IV for D_i that is a formula: the number of neighbors invited for deworming
- Reduced-form & first-stage are not causal yet: would compare kids with more vs. fewer neighbors
  - Controlling for the total number of neighbors does the job

Outline

1. Formula instruments and recentering
   - Example: parametric spillover effects
   - General setting
   - Application: Chinese high-speed railways
   - Other examples
2. Shift-share instruments
   - SSIV as leveraging a shift-level natural experiment
   - SSIV as combining diff-in-diffs

Non-random exposure to exogenous shocks

- Borusyak and Hull (2023) setup: $Y_i = \tau D_i + \varepsilon_i$, constant linear effects
  - Extensions to heterogeneous effects, other controls, multiple treatments, 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 $D_i = Z_i$
- Assumption 1: Shock exogeneity, $g \perp\!\!\!\perp \varepsilon \mid w \quad (\text{or } g \perp\!\!\!\perp (\varepsilon, w))$
  - 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
- Assumption 2: DGP for g (conditional distribution $G(g \mid w)$) is known, e.g. the randomization protocol or uniform across some permutations of g

Formal results

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

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

- The **recentered instrument** $\tilde{Z}_i = Z_i - \mu_i$ is a valid instrument for D_i:

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

- Instrumenting D_i with Z_i and controlling for μ_i also identifies β
- Consistency: with many shocks and weak dependence of $\tilde{Z}_i$ across i
- Robustness to heterogeneous effects: identify a convex average of $\partial Y_i / \partial D_i$ under non-causal first-stage monotonicity

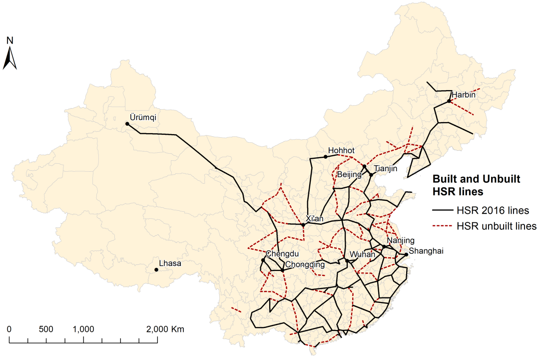
Outline

1. Formula instruments and recentering
   - Example: parametric spillover effects
   - General setting
   - Application: Chinese high-speed railways
   - Other examples
2. Shift-share instruments
   - SSIV as leveraging a shift-level natural experiment
   - SSIV as combining diff-in-diffs

Application: Effects of railway upgrades

BH study the employment effects of the Chinese high-speed railway system

- Observe 83 lines open by 2016 (also 66 planned but not yet opened lines)



First task: Specifying D_i

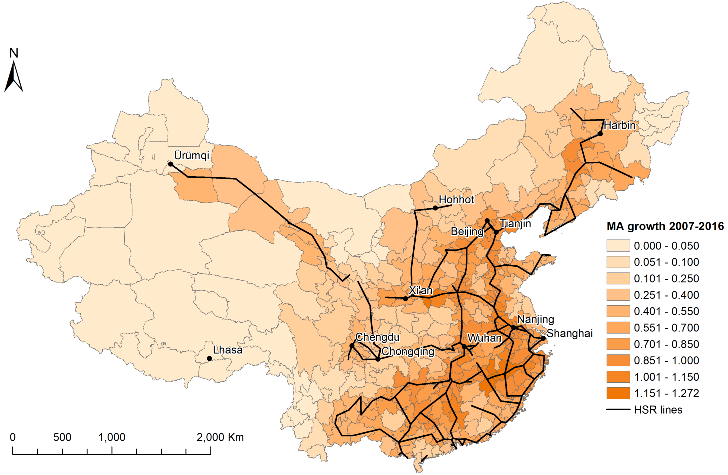
- A dummy of (at least one) line passing through city i?
  - Wouldn't capture the effects on neighboring cities
- Distance to the nearest line, computed from line locations and location of city i?
- BH choose a version suggested by models of economic geography:
  - Railways reduce travel time to big cities and thus increase regional "market access"

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

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

  - g_t is transportation network
  - loc_k is region's location on the map
  - Pop_k is time-invariant regional population [what if it changes endogenously?]
  - ε_i is effects of unobserved local shocks (e.g. productivity)

Realized 2016 network and MA growth

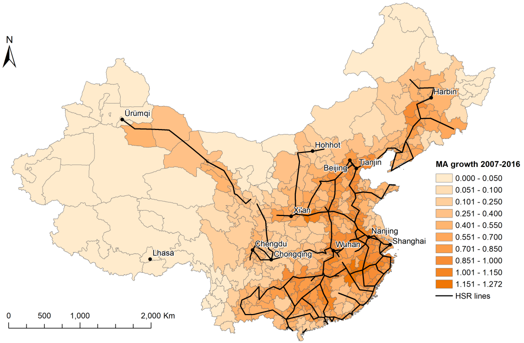


OVB problem & recentering solution

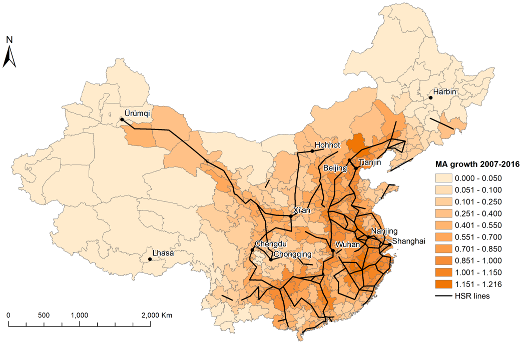
- Lines are not randomly placed: e.g. tend to connect big cities
  - And even if they were randomly placed, MA would still grow more near big cities!
  - Big cities and cities nearby can be on different trends from rest of country $\Longrightarrow$ OVB
- To recenter, need to view the HSR network as realization of some random process (a natural experiment)
  - BH assume exogenous *timing* of lines: it's random which lines get built first, within groups of lines of similar length
  - Conditioning on the set of planned lines, don't need to think about other lines that could have been
- Draw 1,000 simulated HSR networks by permuting opening statuses of lines randomly within groups
  - Recompute MA growth of each city; take the average across simulations

Actual vs. counterfactual MA growth

Actual 2016 network

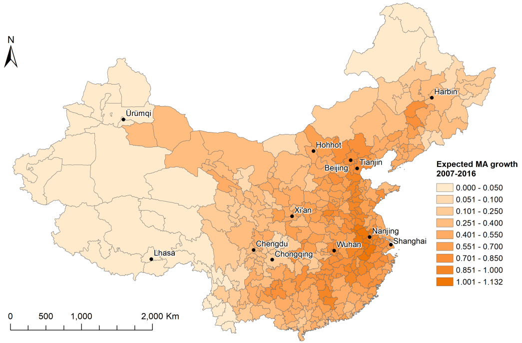


Example counterfactual 2016 network

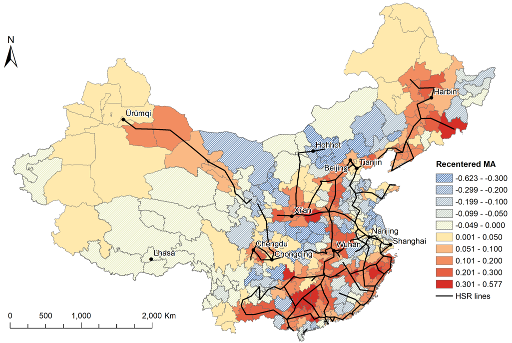


Expected and recentered MA growth

Expected MA growth



Recentered MA growth



- Treated/control group: regions with MA growth higher/lower than expected
  - Valid comparison if realized and counterfactual networks are equally likely

Testing the specification of counterfactuals

TABLE II

REGRESSIONS OF MARKET ACCESS GROWTH ON MEASURES OF ECONOMIC GEOGRAPHY.

	Unadjusted	Recentered		
	(1)	(2)	(3)	(4)
Distance to Beijing	−0.291	0.069		0.088
	(0.062)	(0.039)		(0.045)
Latitude/100	−3.324	−0.342		−0.182
	(0.646)	(0.276)		(0.319)
Longitude/100	1.321	0.485		0.440
	(0.458)	(0.237)		(0.240)
Expected Market Access Growth			0.026	0.054
			(0.056)	(0.069)
Constant	0.536	0.018	0.018	0.018
	(0.029)	(0.018)	(0.021)	(0.018)
Joint RI p-value		0.443	0.711	0.492
R^2	0.824	0.083	0.010	0.086
Prefectures	275	275	275	275

Recentered instrument should be uncorrelated with predetermined observables

Employment effects of HSR

TABLE I

EMPLOYMENT EFFECTS OF MARKET ACCESS GROWTH: UNADJUSTED AND RECENTERED ESTIMATES.

	Unadjusted OLS	Recentered IV	Controlled OLS
	(1)	(2)	(3)
Panel A: No Controls			
Market Access Growth	0.232	0.084	0.072
	(0.075)	(0.097)	(0.093)
		[−0.245, 0.337]	[−0.169, 0.337]
Expected Market Access Growth			0.317
			(0.096)

- Column 2 instruments realized MA with recentered MA
- Column 3 is OLS controlling for expected MA

Inference

- Complication: recentered MA is correlated across all regions
  - A long line affects many regions across the country
- Approach 1: restrict dependence of ε_i and thus $\tilde{Z}_i\varepsilon_i$
  - Spatially-clustered SE [used in any spatial contexts, not only recentered IVs]
- Approach 2: leverage randomness of shocks, allowing unrestricted dependence of ε_i
  - Randomization inference [for any recentered IV with constant causal effects, not only spatial]
  - (Later: Asymptotic solution when Z_i takes the shift-share form)

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: $\kappa(x) = \max\left\{1 - |x|, 0\right\}$

Spatially-clustered standard errors

Notes:

- The structure of spatial correlation within $d_{ij} < d_{max}$ is not restricted
  - E.g. doesn't have to be constant even if uniform kernel is used
- d_{max} should be small enough, such that $d_{ij} > d_{max}$ for most region pairs
  - Similar to having many clusters with non-spatially clustered SE
  - If spatial correlation is too wide-ranging, see Conley and Kelly (forth.) for other inference methods

Randomization inference leveraging shock randomness

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

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

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

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

- 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 (by test inversion) 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

Outline

1. Formula instruments and recentering
   - Example: parametric spillover effects
   - General setting
   - Application: Chinese high-speed railways
   - Other examples
2. Shift-share instruments
   - SSIV as leveraging a shift-level natural experiment
   - SSIV as combining diff-in-diffs

Example: Simulated instruments

- Currie and Gruber (1996a,b) study the effects of Medicaid eligibility on health
  - 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

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? How can we control for it?
- What is the benefit of this approach over simulated IV?
- 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 the federal poverty line 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
  - ~2x smaller standard errors

More examples

- Number of days of extreme heat measured from the local crop composition, their temperature thresholds, and local weather (Hsiao, Moscona, Sastry 2024, Zappala 2024)
- Radio signal strength, predicted from country's topography and location of transmitters (Olken 2009, Yanagizawa-Drott 2014)
- IVs from partially-randomized assignment mechanisms, e.g. for charter schools (Abdulkadiroglu et al. 2017, 2019, Narita and Yata 2023)
- Linear [and nonlinear] shift-share instruments...

Outline

1. Formula instruments and recentering
   - Example: parametric spillover effects
   - General setting
   - Application: Chinese high-speed railways
   - Other examples
2. Shift-share instruments
   - SSIV as leveraging a shift-level natural experiment
   - SSIV as combining diff-in-diffs

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 **shifts** not specific to i
  - S_{ik} are exposure **“shares”** (weights), often with $\sum_k S_{ik} = 1$ for all i

Example 0: Linear spillovers

- Cai, De Janvry, and Sadoulet (2015): spillover effects of randomized information about an insurance product
  - Y_i = person i takes up the insurance product
  - D_i = the fraction of i's friends who were given the information
  - Use OLS: $Z_i = D_i$
- Not usually understood as a shift-share design, but it is:

$$Z_i = \sum_{k=1}^{N} \frac{\mathbf{1}\left[i \text{ and } k \text{ are friends}\right]}{\# \text{ of } i\text{'s friends}} \cdot \text{Treated}_k$$

Example 1: Labor supply

- Consider the (inverse) regional labor supply equation: $\Delta w_i = \tau \Delta L_i + \varepsilon_i$
  - Δw_i = log-change in region i's average wage over some period
  - ΔL_i = log-change in i's employment
  - ε_i = labor supply shocks (e.g. migration, UI benefits)
- Need a regional labor demand shock as IV
- Regional labor demand combines demand for different industries k:
  - $\Delta L_i \approx \sum_k S_{ik} \Delta L_{ik}$
  - ΔL_{ik} = log-change of k-specific employment in i
  - S_{ik} = initial share of k in i's employment
- Consider $Z_i = \sum_k S_{ik} g_k$, replacing ΔL_{ik} with national industry shifts g_k
  - E.g. g_k = randomized subsidy
  - Bartik (1991): g_k = observed growth rate of industry employment

Example 2: Enclave instrument for migration

- Consider the (inverse) elasticity of substitution between native and immigrant workers
  - $\Delta \log \frac{\text{Immigrant wage}_i}{\text{Native wage}_i} = \tau \Delta \log \frac{\text{Immigrant employment}_i}{\text{Native employment}_i} + \varepsilon_i$
  - ε_i = change in relative labor demand in region i
- Need a relative labor supply shock as 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 & some "push shocks"
  - g_k = dummy of war in k (Llull 2017) or observed national migration rate from k (Card 2009)
  - S_{ik} = initial share of origin k in i's population

The SSIV exogeneity challenge

- How should we think about SSIV exogeneity, $\mathbb{E}\left[Z_i \varepsilon_i\right] = 0$?
  - Which properties of the shifts and/or shares can make Z_i exogenous?
  - Do we need exogeneity of both shifts and shares, or just one?
  - What is exogeneity of shifts, which don't vary across i?
  - What is exogeneity of shares, which are measured in a pre-period?
- Two narratives + sets of sufficient conditions (plausible in different applications):
  - **"Many exogenous shifts"** (Borusyak, Hull, Jaravel, 2022; BHJ): leveraging a shift-level natural experiment, translated to the observation level
  - **"Exogenous shares"** (Goldsmith-Pinkham, Sorkin, Swift, 2020; GPSS): pooling diff-in-diffs based on heterogeneous exposure shares

Outline

1. Formula instruments and recentering
   - Example: parametric spillover effects
   - General setting
   - Application: Chinese high-speed railways
   - Other examples
2. Shift-share instruments
   - SSIV as leveraging a shift-level natural experiment
   - SSIV as combining diff-in-diffs

Identification from many exogenous shifts

- Suppose the shifts are as-good-as-random: $\mathbb{E}\left[g_k \mid \text{all errors and all shares}\right] = \alpha$
- Then the expected instrument is $\mu_i = \mathbb{E}\left[\sum_k S_{ik} g_k \mid S_i, \ldots, S_{iK}\right] = \alpha \sum_k S_{ik}$
  - If the shares add up to one, $\mu_i = const \implies$ no corrections are needed
  - *Note:* No need to run 1,000 simulations, or to know the joint CDF of shifts
- Intuition: A weighted average of as-good-as-random shifts is as-good-as-random
  - If industries with high vs. low g_k are comparable, regions specializing in those industries are comparable
    - ★ True whatever the initial shares are: we don't require $\mathrm{Cov}\left[S_{ik}, \varepsilon_i\right] = 0$
  - The SSIV allows to translate random variation across k into random variation across i (under appropriate exclusion)

Consistency

- **Proposition:** $\frac{1}{N}\sum_i Z_i\varepsilon_i \to 0$ **if**

  1. Weaker notion of shift exogeneity: g_k are uncorrelated with $\bar{\varepsilon}_k = \sum_i S_{ik}\varepsilon_i / \sum_i S_{ik}$
  2. The effective number of shifts is large: $N_{eff} = 1/\sum_k s_k^2 \to \infty$ for $s_k = \frac{1}{N}\sum_i S_{ik}$
  3. Shares add up to one for each observation

- Holds because $\frac{1}{N}\sum_i Z_i\varepsilon_i = \frac{1}{N}\sum_{i,k} S_{ik}g_k\varepsilon_i = \sum_k s_k g_k \bar{\varepsilon}_k$, by changing the order of summation

"Incomplete shares" case

- What if $\sum_k S_{ik} \neq 1$, so Z_i is a weighted sum, not weighted average?
  - Then randomization of subsidies does not make Z_i as-good-as-random
- E.g. subsidies are only in manufacturing industries k
  - Regions with more manufacturing will mechanically get higher Z_i, potentially creating bias
- But recall that the expected instrument is $\mu_i = \alpha \cdot \sum_k S_{ik}$
  - Thus, controlling for $\sum_k S_{ik}$ avoids OVB

Shift-level controls

- What if shifts are as-good-as-random only after controlling for some q_k?
  - In an industry-level regression you'd control for q_k. But what about a SSIV regression?
- Expected instrument is $\mu_i = \sum_k S_{ik} \left(\alpha_0 + \alpha_1 q_k\right)$
  - To avoid OVB, it's sufficient to control for $\sum_k S_{ik}$ and $\sum_k S_{ik} q_k$
- E.g. industry subsidies are random only controlling the dummy of hi-tech industries
  - Control for the total local share of hi-tech industries
  - Then you leverage variation between regions with similar composition of hi- and low-tech industries but different exposure to subsidies within each group

Shift-level ("exposure-robust") standard errors

- Unusual clustering problem: observations with similar shares are exposed to the same shocks, both g_k and unobserved ν_k
  - Conventional clustering of SE (e.g. by state) wouldn't capture that
  - Adao, Kolesar, Morales (2019) generate placebo shifts g_k^* and find a significant reduced-form in ~50% simulations
- If each region was exposed to just one industry, we'd cluster by industry
  - But how to do it in a "fuzzy" case?
- Adao et al. derive a formula that leverages mutual independence of g_k, regardless of clustering of ε_i

Shift-level ("exposure-robust") standard errors

- BHJ propose a convenient solution:
  - SSIV estimate is exactly equal to the estimate from a *special* shift-level regression with g_k as the IV
    $$\bar{Y}_k^{\perp} = \tau \bar{D}_k^{\perp} + \gamma' q_k + \bar{\varepsilon}_k,$$
    weighted by $s_k = \frac{1}{N} \sum_i S_{ik}$, where $\bar{V}_k^{\perp} = \sum_i S_{ik} V_i^{\perp} / \sum S_{ik}$ and $V_i^{\perp}$ is residual from the regression of V_i on included i-level controls
  - In Stata and R, package *ssaggregate* helps run this special IV regression
  - Then use robust or clustered SE in the way you'd do it for any other regression leveraging these shifts

Application: "China shock" (Autor, Dorn, Hanson 2013)

- ADH study the effects of import competition on local labor market outcomes
- 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 exposure to Chinese imports:

$$D_{it} = \sum_{\text{manuf. ind. } k} \text{Initial emp share}_{ikt} \cdot \frac{\Delta \text{US imports from China}_{kt}}{\text{Initial employment}_{kt}}$$

- Control for period FE and some initial regional characteristics, including total manufacturing share of employment

ADH's shift-share approach

- OLS can be biased: imports grow in industries and regions where US productivity growth is slow
- Idea: use industry productivity growth in China as shifts
  - Construct a proxy: import growth from China in other countries (e.g. Europe)
  - Requires (1) China productivity to be exogenous and (2) productivity growth in the US and Europe to be uncorrelated
- SSIV: $Z_{it} = \sum_k S_{ikt} g_{kt}$ where
  - S_{ikt} = 10-year 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
- The shifts are the IV in the industry-level analysis in Acemoglu et al. (2016)

BHJ revisit ADH

Step 1: Picking controls

- Hypothesize that shifts g_{kt} are as-good-as-random controlling for period FEs q_t
- Regional regression should thus control for

$$\sum_k S_{ikt} q_t = \left(\sum_k S_{ikt} \right) \times q_t$$

  i.e. lagged total share of manufacturing interacted with period dummies

- Also consider the industry controls from Acemoglu et al., e.g. initial skill intensity of the industry

BHJ revisit ADH

Step 2: Balance tests to falsify conditional as-good-as-random shift assignment:

- Shifts 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

Step 3: Obtain the estimates, including the necessary controls and with SE from the equivalent industry-by-year level IV regression

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

- Controls are important: China shock g_{kt} is larger in the 2000s (post WTO entry) when overall manuf. decline is stronger for other reasons

Outline

1. Formula instruments and recentering
   - Example: parametric spillover effects
   - General setting
   - Application: Chinese high-speed railways
   - Other examples
2. Shift-share instruments
   - SSIV as leveraging a shift-level natural experiment
   - SSIV as combining diff-in-diffs

Motivating example: Mariel Boatlift

- Consider the substitution elasticity between demand for migrant and native labor
  - $D_i = \Delta$ relative immigrant/native employment, $Y_i = \Delta$ relative wage
- Many Cubans arrive in the US in 1980 ~~for exogenous reasons~~

  $\Rightarrow$ Local arrival rate of Cubans is relevant
  - But violates exogeneity if they settle in areas with growing demand for their labor
- Consider instrumenting local immigration rate with the initial share of Cubans
  - Relevance: new Cubans tend to go to places with a lot of Cubans
  - Exogeneity: parallel outcome trends in places with high vs. low initial Cuban share
    - ⋆ Violated if factors that attracted Cubans in the past are correlated with today's trends
- This is a simple shift-share IV: $Z_i = S_{i,\text{Cubans}} \cdot 1 + \sum_{k \neq \text{Cubans}} S_{ik} \cdot 0$
  - Justified by parallel trends in outcomes, not a natural experiment in shifts

Exogenous shares approach: Pooling diff-in-diffs

- This logic generalizes to having multiple shocked countries
  - Goldsmith-Pinkham, Sorkin, Swift (GPSS, 2020) develop this view
- Assume **exogenous shares**: $\mathrm{Cov}\left[\varepsilon_i, S_{ik}\right] = 0$ for every k
  - With Y_i measured in differences, this is K parallel trends assumptions
  - Strong assumption even though shares are measured in the pre-period
  - Wrong interpretation of exogeneity: *"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 one reasonable way to combine them
  - 2SLS (for small K) and LIML are other reasonable ways
  - Or just use your favorite share (e.g. of Cubans)

Rotemberg weights

- If you insist on using SSIV (and not LIML), GPSS recommend computing the importance of each share: **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 shifts 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 of two approaches to shift-share IV

- Two sets of narratives & sufficient conditions for SSIV validity
  - Pick one *ex ante*, then validate *ex post*
- **Many exogenous shifts** is appropriate when you could imagine using your shifts as IV in a shift-level analysis
- **Exogenous shares** is appropriate when you would be OK using any other combination of shares as the IV

More shift-share IV examples from the "Practical guide"

Study	Unit (i)	Outcome (y_i)	Treatment (x_i)	Level of shift variation (k)	Instrument (z_i): Share (s_{ik})	Instrument (z_i): Shift (g_k)
Hummels et al. (2014)	Worker	Wage	Imports of intermediate goods by employer	Product-by-country	$\text{Imports}_{ik}/\text{Imports}_i$	Imports from k to other countries
Nunn and Qian (2014)	Country-by-year	Conflict	Quantity of food aid (wheat) from the US	Year	Fraction of years with non-zero food aid	US wheat production in previous year
Cai et al. (2015)	Individual	Takeup of insurance	% of friends selected for an information session*	Individual	Dummy(k is friend of i)/ # of friends i has	Dummy of information session
Jaravel (2019)	Product category	Inflation and innovation	Δ Quantity demanded	Socio-demographic group	Sales of i to group k/ Total sales of i	Population change
Greenstone et al. (2020)	Region	Δ Employment	Δ Credit	Bank	Credit market share of k	Estimated credit supply shock
Aghion et al. (2022)	Firm	Δ Firm employment	Δ Firm stock of automation technologies	Technology-by-country	$\text{Imports}_{ik}/\text{Imports}_i$	Δ imports from k to other countries
Xu (2022)	Region	Δ Exports	Exposure to banking crisis*	Bank	Credit market share of k	Bankruptcy during banking crisis
Franklin et al. (2023)	Local labor market	Wage	Shift-share exposure to the intervention*	Residential neighborhood	$\text{Commuters}_{ik}/\text{Employment}_i$	Dummy of public works intervention
Mohnen (2024)	Region	Δ Young labor market outcome	Retirement rate	Age group (within 45+)	$\text{Population}_{ik}/\text{Population 45+}_i$	National retirement rate at age k

Conclusion: Formula and shift-share IV

The goal is for you to have an instinct to:

- Identify settings with formula and shift-share treatments/instruments
- 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" (or simpler controls with shift-shares)