Causal Inference with Formula Instruments

Peter Hull

Brown University
November 2024

Based on Borusyak, Hull, & Jaravel '22, Borusyak & Hull '23, and Borusyak, Hull, & Jaravel '24

Motivation

We often estimate causal effects or structural parameters using IVs that combine multiple sources of variation via a known formula

- We often think that some, but not all, of these sources are exogenous

Motivation

We often estimate causal effects or structural parameters using IVs that combine multiple sources of variation via a known formula

- We often think that some, but not all, of these sources are exogenous

Ex. 1: $z_i = s_i g_i$, where g_i is drawn in an RCT and s_i is a predetermined variable expected to predict the first stage (Coussens and Spiess, 2021)

Motivation

We often estimate causal effects or structural parameters using IVs that combine multiple sources of variation via a known formula

- We often think that some, but not all, of these sources are exogenous

Ex. 1: $z_i = s_i g_i$, where g_i is drawn in an RCT and s_i is a predetermined variable expected to predict the first stage (Coussens and Spiess, 2021)

Ex. 2: Shift-share $z_i = \sum_k s_{ik} g_k$, where g_k is a shift varying at a different "level" (e.g. industries) and s_{ik} are local (e.g. regional) exposure shares

Motivation

We often estimate causal effects or structural parameters using IVs that combine multiple sources of variation via a known formula

- We often think that some, but not all, of these sources are exogenous

Ex. 1: $z_i = s_i g_i$, where g_i is drawn in an RCT and s_i is a predetermined variable expected to predict the first stage (Coussens and Spiess, 2021)

Ex. 2: Shift-share $z_i = \sum_k s_{ik} g_k$, where g_k is a shift varying at a different "level" (e.g. industries) and s_{ik} are local (e.g. regional) exposure shares

Ex. 3: Formulas capturing spatial/network/GE spillovers of exogenous shocks to other units with heterogeneous/non-random exposure

- E.g. a z_i counting the # of neighbors selected for an intervention

Motivation

We often estimate causal effects or structural parameters using IVs that combine multiple sources of variation via a known formula

- We often think that some, but not all, of these sources are exogenous

Ex. 1: $z_i = s_i g_i$, where g_i is drawn in an RCT and s_i is a predetermined variable expected to predict the first stage (Coussens and Spiess, 2021)

Ex. 2: Shift-share $z_i = \sum_k s_{ik} g_k$, where g_k is a shift varying at a different "level" (e.g. industries) and s_{ik} are local (e.g. regional) exposure shares

Ex. 3: Formulas capturing spatial/network/GE spillovers of exogenous shocks to other units with heterogeneous/non-random exposure

- E.g. a z_i counting the # of neighbors selected for an intervention

How can we just leverage the exogenous shocks for identification?

Identification via Expected Instrument Adjustment

The Problem: non-random exposure to exogenous shocks generally makes such formulas invalid instruments

- Randomizing shocks $\not\Rightarrow$ randomized formulas based on them

Identification via Expected Instrument Adjustment

The Problem: non-random exposure to exogenous shocks generally makes such formulas invalid instruments

- Randomizing shocks $\not\Rightarrow$ randomized formulas based on them

A Solution: use knowledge of the "design" of shocks to derive the *expected instrument* μ_i: the average z_i across repeated draws of shocks

Identification via Expected Instrument Adjustment

The Problem: non-random exposure to exogenous shocks generally makes such formulas invalid instruments

- Randomizing shocks $\not\Rightarrow$ randomized formulas based on them

A Solution: use knowledge of the "design" of shocks to derive the *expected instrument* μ_i: the average z_i across repeated draws of shocks

- E.g. redraw g_i from an RCT protocol or permute shocks which could well have been randomly exchanged, recompute+average the formula

Identification via Expected Instrument Adjustment

The Problem: non-random exposure to exogenous shocks generally makes such formulas invalid instruments

- Randomizing shocks $\not\Rightarrow$ randomized formulas based on them

A Solution: use knowledge of the "design" of shocks to derive the *expected instrument* μ_i: the average z_i across repeated draws of shocks

- E.g. redraw g_i from an RCT protocol or permute shocks which could well have been randomly exchanged, recompute+average the formula
- This μ_i is the sole confounder of z_i, akin to a propensity score

Identification via Expected Instrument Adjustment

The Problem: non-random exposure to exogenous shocks generally makes such formulas invalid instruments

- Randomizing shocks $\not\Rightarrow$ randomized formulas based on them

A Solution: use knowledge of the "design" of shocks to derive the *expected instrument* μ_i: the average z_i across repeated draws of shocks

- E.g. redraw g_i from an RCT protocol or permute shocks which could well have been randomly exchanged, recompute+average the formula
- This μ_i is the sole confounder of z_i, akin to a propensity score

Controlling for μ_i or using the *recentered IV* $\tilde{z}_i = z_i - \mu_i$ avoids bias

- Or controlling for w_i that are known to linearly span μ_i

(Some) Empirical Settings Where This May be Relevant

- **Network spillovers:** *Miguel and Kremer 2004, Gerber and Green 2012, Acemoglu et al. 2015, Jaravel et al. 2018, Carvalho et al. 2020*
- **Effects of transportation:** *Baum-Snow 2007, Donaldson and Hornbeck 2016, Lin 2017, Donaldson 2018, Ahlfeldt and Feddersen 2018, Bartelme 2018*
- **Simulated instruments:** *Currie and Gruber 1996a,b, Cullen and Gruber 2000, East and Kuka 2015, Cohodes et al. 2016, Frean et al. 2017*
- **Shift-share/Bartik IV:** *Autor et al. 2013, Adão et al. 2021, Kovak 2013*
- **Nonlinear shift-share IV:** *Boustan et al. 2013, Berman et al. 2015, Basso and Peri 2015, Chodorow-Reich and Wieland 2020, Derenoncourt 2021*
- **IVs from assignment mechanisms:** *Abdulkadiroglu et al. 2017, 2019*
- **Weather IVs:** *Gomez et al. 2007, Madestam et al. 2013*
- **IVs for mass media access:** *Olken 2009, Yanagizawa-Drott 2014*

Example: Market Access Effects in an RCT

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

$$\Delta \log V_i = \beta \Delta \log MA_i + \varepsilon_i,$$
$$\text{where } MA_{it} = \sum_j \tau(g_t, loc_i, loc_j)^{-1} pop_j,$$

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

Example: Market Access Effects in an RCT

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

$$\Delta \log V_i = \beta \Delta \log MA_i + \varepsilon_i,$$
$$\text{where } MA_{it} = \sum_j \tau(g_t, loc_i, loc_j)^{-1} pop_j,$$

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

Example: Market Access Effects in an RCT

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

$$\Delta \log V_i = \beta \Delta \log MA_i + \varepsilon_i,$$
$$\text{where } MA_{it} = \sum_j \tau(g_t, loc_i, loc_j)^{-1} pop_j,$$

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?

Example: Market Access Effects in an RCT

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

$$\Delta \log V_i = \beta \Delta \log MA_i + \varepsilon_i,$$
$$\text{where } MA_{it} = \sum_j \tau(g_t, loc_i, loc_j)^{-1} pop_j,$$

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?

No, because of the non-random components of the formula

Illustration: Market Access on a Square Island

Start from no roads, assume equal population everywhere

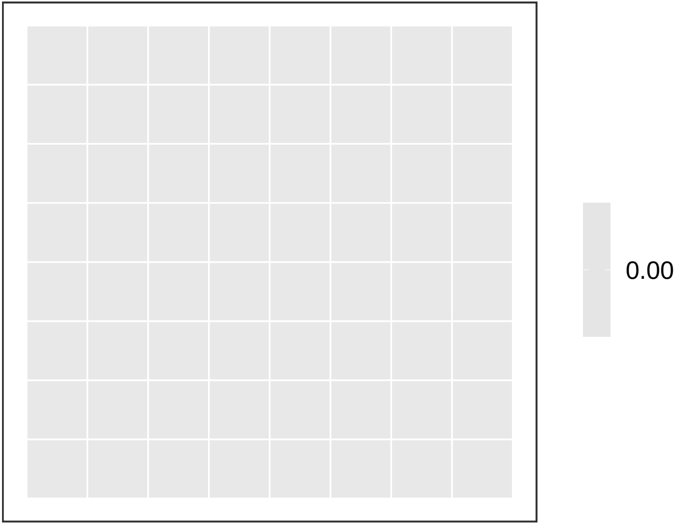


Illustration: Market Access on a Square Island

Randomly connect adjacent regions by road

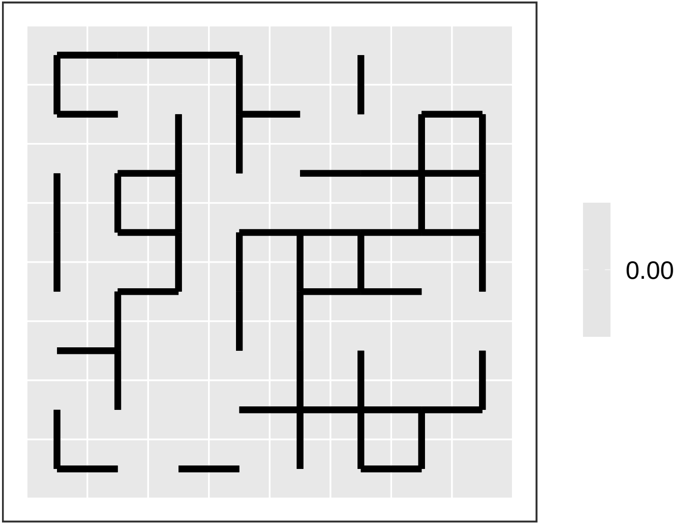


Illustration: Market Access on a Square Island

Randomly connect adjacent regions by road and compute MA growth

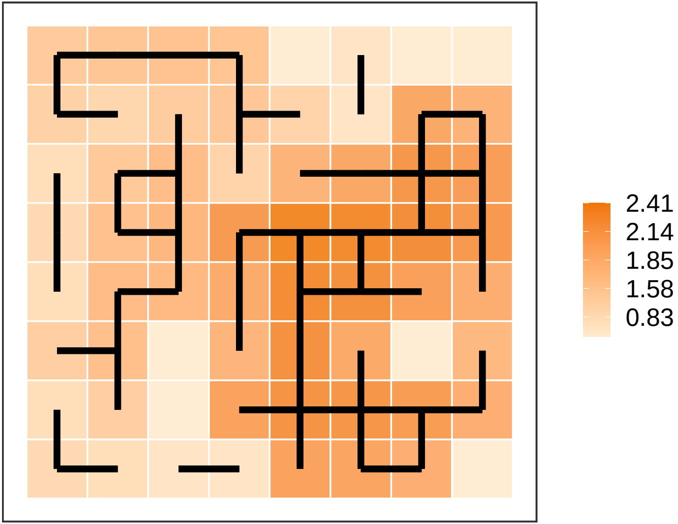


Illustration: Market Access on a Square Island

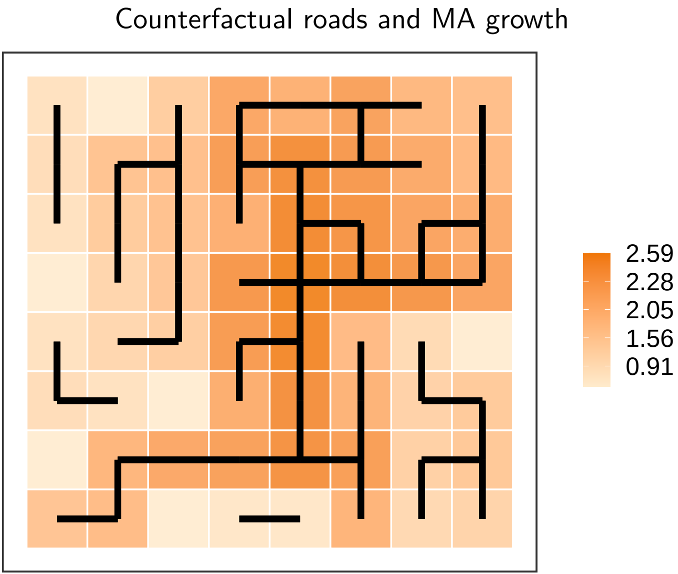
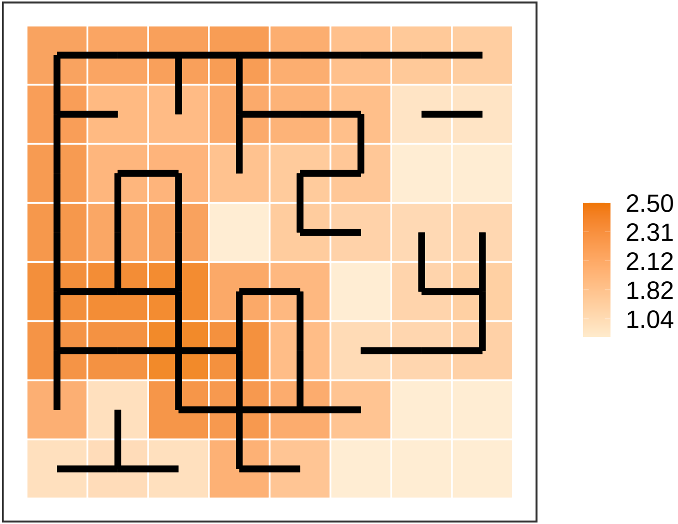


Illustration: Market Access on a Square Island

Counterfactual roads and MA growth



Expected Market Access Growth μ_i

Some regions get systematically more MA

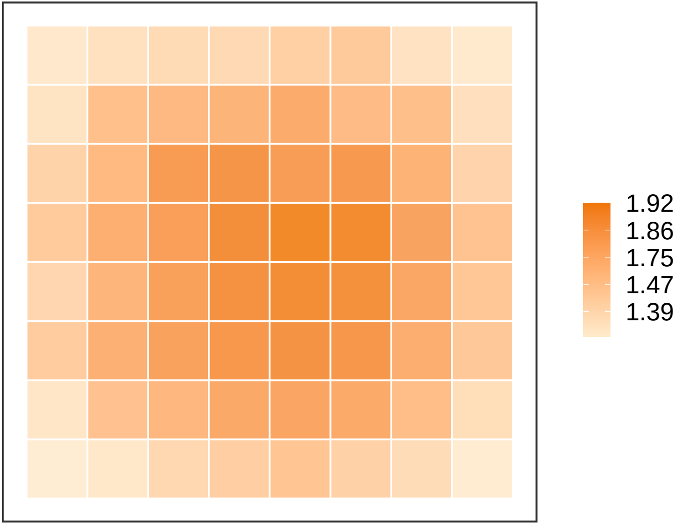


Illustration: High-Speed Rail in China

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

Illustration: High-Speed Rail in China

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

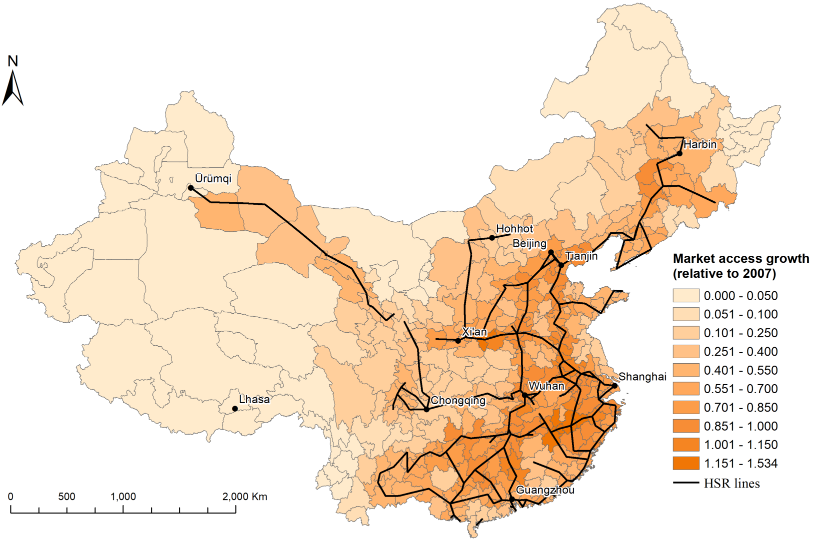


Illustration: High-Speed Rail in China

A counterfactual draw of 83 lines by 2016

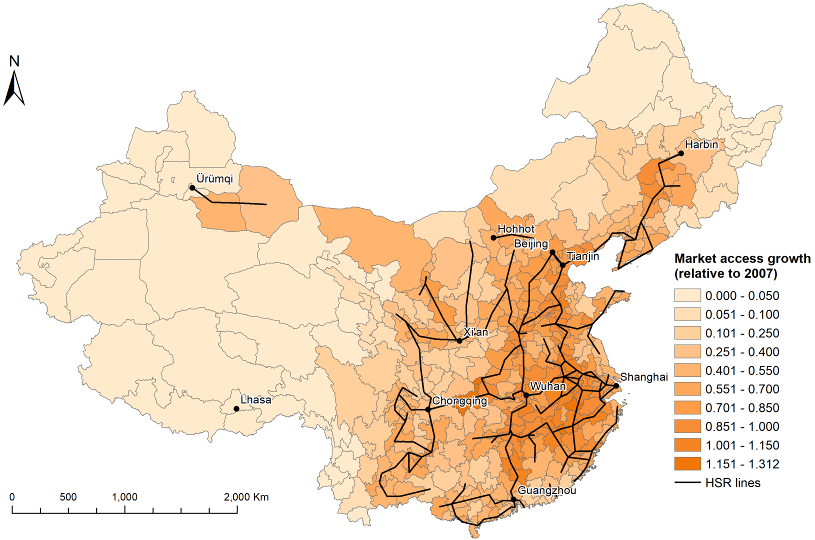
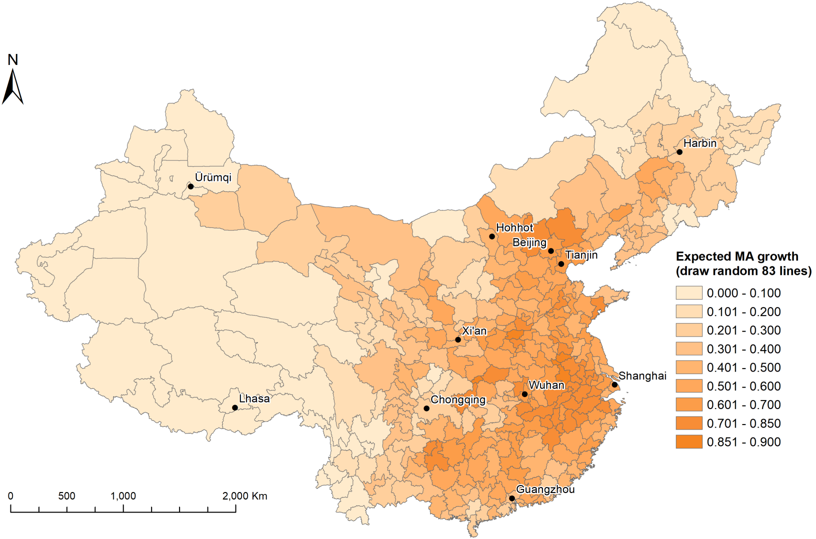


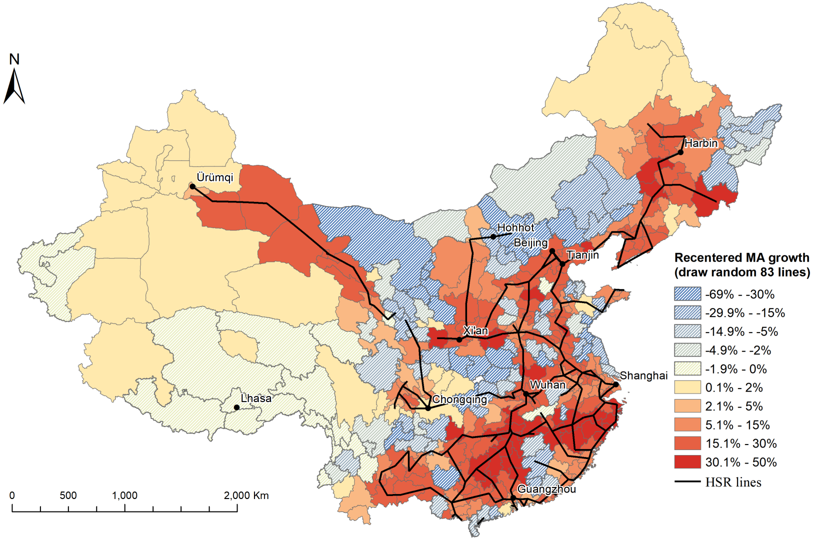
Illustration: High-Speed Rail in China

Expected MA growth, μ_i



Recentered MA growth

Recentered MA growth, $\Delta \log MA_i - \mu_i$



General Setting

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

- Extensions: heterogeneous effects, other controls, multiple treatments, panel data...

General Setting

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

- Extensions: heterogeneous effects, other controls, multiple treatments, panel data...

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

- 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

General Setting

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

- Extensions: heterogeneous effects, other controls, multiple treatments, panel data...

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

- 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: $g \perp\!\!\!\perp \varepsilon \mid w$, for $w = (s, q)$
2. Conditional distribution $G(g \mid w)$ is known (e.g. via randomization protocol or uniform across permutations of g)

Results

The 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 bias:

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

Results

The 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 bias:

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

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

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

Results

The 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 bias:

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

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)

Results

The 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 bias:

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

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**: with many shocks and $\tilde{z}_i$ weakly dependent across i
- **Robustness** to heterogeneous treatment effects: $\tilde{z}_i$ identifies a convex avg. of $\partial y_i/\partial x_i$ under appropriate first-stage monotonicity
- **Randomization inference** provides exact confidence intervals for β (under constant effects) and falsification tests

Special Case: Linear Formulas

When $z_i = \sum_k s_{ik} g_k$ is linear in the shocks (i.e. shift-share IV), we need only specify the conditional shock *mean*:

$$\mu_i = E[z_i \mid w] = \sum_k s_{ik} E[g_k \mid w]$$

and in fact we can work with a weaker shock exogeneity assumption (see Borusyak, Hull, and Jaravel 2022)

Special Case: Linear Formulas

When $z_i = \sum_k s_{ik} g_k$ is linear in the shocks (i.e. shift-share IV), we need only specify the conditional shock *mean*:

$$\mu_i = E[z_i \mid w] = \sum_k s_{ik} E[g_k \mid w]$$

and in fact we can work with a weaker shock exogeneity assumption (see Borusyak, Hull, and Jaravel 2022)

If we assume $E[g_k \mid w] = q_k' \gamma$ for some shock-level controls q_k, this tells us it's enough to control for $c_i = \sum_k s_{ik} q_k$

- Special case: $E[g_k \mid w] = \mu_g$ (i.e. unconditionally exogenous shocks), so c_i is the "sum of shares" $\sum_k s_{ik}$ (often $= 1$ in practice)

Contrast: Outcome-Based Models

Rather than focusing on the design of exogenous shocks, we could model the unobserved error ε_i's dependence on w:

$$E[\varepsilon_i \mid g, w] = q_i{}'\gamma, \quad \text{for } q_i \in w$$

E.g., if y_i and x_i are in first differences, $E[\varepsilon_i \mid g, w] = \gamma$ is "parallel trends"

Contrast: Outcome-Based Models

Rather than focusing on the design of exogenous shocks, we could model the unobserved error ε_i's dependence on w:

$$E[\varepsilon_i \mid g, w] = q_i{}'\gamma, \quad \text{for } q_i \in w$$

E.g., if y_i and x_i are in first differences, $E[\varepsilon_i \mid g, w] = \gamma$ is "parallel trends"

- Goldsmith-Pinkham et al. (2020) formalize this approach for shift-share IV: $E[\varepsilon_i \mid g, w] = E[\varepsilon_i \mid s_{ik}] = 0$ for all shares s_{ik}

Contrast: Outcome-Based Models

Rather than focusing on the design of exogenous shocks, we could model the unobserved error ε_i's dependence on w:

$$E[\varepsilon_i \mid g, w] = q_i{}'\gamma, \quad \text{for } q_i \in w$$

E.g., if y_i and x_i are in first differences, $E[\varepsilon_i \mid g, w] = \gamma$ is "parallel trends"

- Goldsmith-Pinkham et al. (2020) formalize this approach for shift-share IV: $E[\varepsilon_i \mid g, w] = E[\varepsilon_i \mid s_{ik}] = 0$ for all shares s_{ik}

A very strong assumption, which makes *any* formula of (g, w) a valid IV

- E.g. individual shares in shift-share IV, or any transformation of them

Contrast: Outcome-Based Models

Rather than focusing on the design of exogenous shocks, we could model the unobserved error ε_i's dependence on w:

$$E[\varepsilon_i \mid g, w] = q_i{}'\gamma, \quad \text{for } q_i \in w$$

E.g., if y_i and x_i are in first differences, $E[\varepsilon_i \mid g, w] = \gamma$ is "parallel trends"

- Goldsmith-Pinkham et al. (2020) formalize this approach for shift-share IV: $E[\varepsilon_i \mid g, w] = E[\varepsilon_i \mid s_{ik}] = 0$ for all shares s_{ik}

A very strong assumption, which makes *any* formula of (g, w) a valid IV

- E.g. individual shares in shift-share IV, or any transformation of them

In practice, researchers may have stronger priors on how observed shocks are assigned than the right model for ε_i

- Does parallel trends hold in logs vs. levels? (Roth and Sant'Anna '23)
- What are the right features of w to include in q_i?

Application 1: The China Shock (Autor et al. 2013)

ADH study the effects of rising Chinese import competition on US commuting zones over two periods: 1991-2000 and 2000-2007

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

Application 1: The China Shock (Autor et al. 2013)

ADH study the effects of rising Chinese import competition on US commuting zones over two periods: 1991-2000 and 2000-2007

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

To address endogeneity, they use a shift-share IV $z_{it} = \sum_n s_{int} g_{nt}$

- n: 397 SIC4 manufacturing industries $\times$ two periods
- g_{nt}: growth of Chinese imports in non-US economies per US worker
- s_{int}: lagged share of manufacturing industry n in *total* employment of location i; hence $\sum_n s_{int}$ is i's manufacturing employment share

Application 1: The China Shock (Autor et al. 2013)

ADH study the effects of rising Chinese import competition on US commuting zones over two periods: 1991-2000 and 2000-2007

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

To address endogeneity, they use a shift-share IV $z_{it} = \sum_n s_{int} g_{nt}$

- n: 397 SIC4 manufacturing industries $\times$ two periods
- g_{nt}: growth of Chinese imports in non-US economies per US worker
- s_{int}: lagged share of manufacturing industry n in *total* employment of location i; hence $\sum_n s_{int}$ is i's manufacturing employment share

Design-based justification: random industry productivity shocks in China, jointly affecting imports in the U.S. and elsewhere, proxied by g_{nt}

- If g_{nt} is as-good-as-randomly assigned within (but not across) periods, the expected instrument is $\mu_{it} = \sum_n s_{int} \times Post_t$

ADH Balance Tests

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	

- Panel A regresses industry characteristics on the g_{nt} shocks, controlling for period FE
- Panel B regresses location characteristics on the z_{it} instrument, controlling for manufacturing employment share $\times$ period FE

ADH Estimates

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

Note: columns 3-7 control for mfg. employment share $\times$ period FE

Application 2: Chinese HSR (Borusyak and Hull, 2023)

Let's return to the motivating market access application

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

Application 2: Chinese HSR (Borusyak and Hull, 2023)

Let's return to the motivating market access application

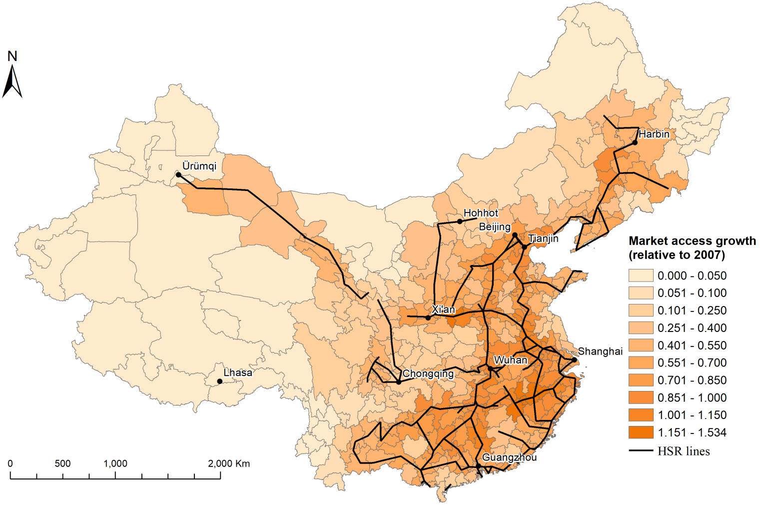
Setting: Chinese HSR; 83 lines built 2008–2016, 66 yet unbuilt

- 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

Design: which planned lines opened by some date is as-good-as-random, conditional on line observables (e.g. line length/complexity)

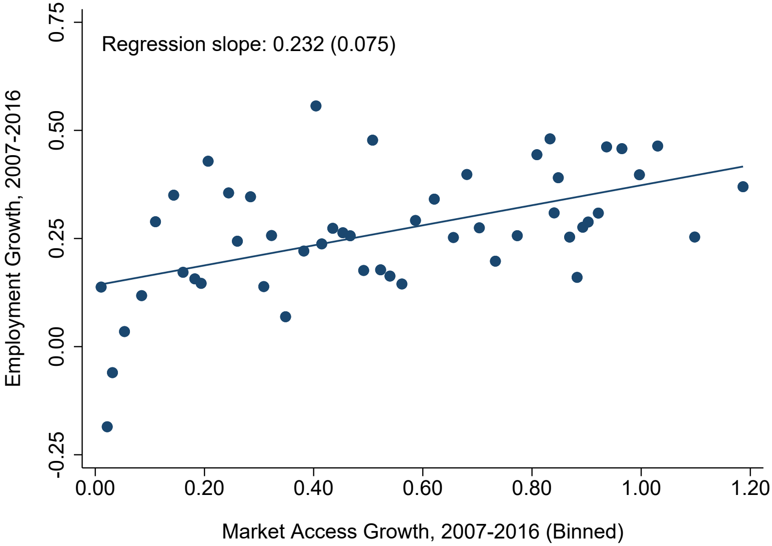
- Expected market access growth given by permuting line openings among observably similar lines

HSR Lines and Market Access



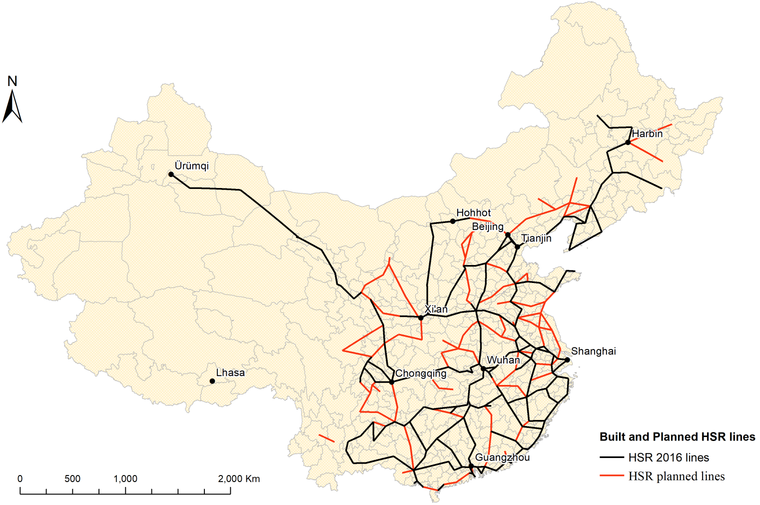
Naive OLS compares dark (“treatment”) vs light (“control”) regions

Naive OLS Suggests a Big Market Access Effect...



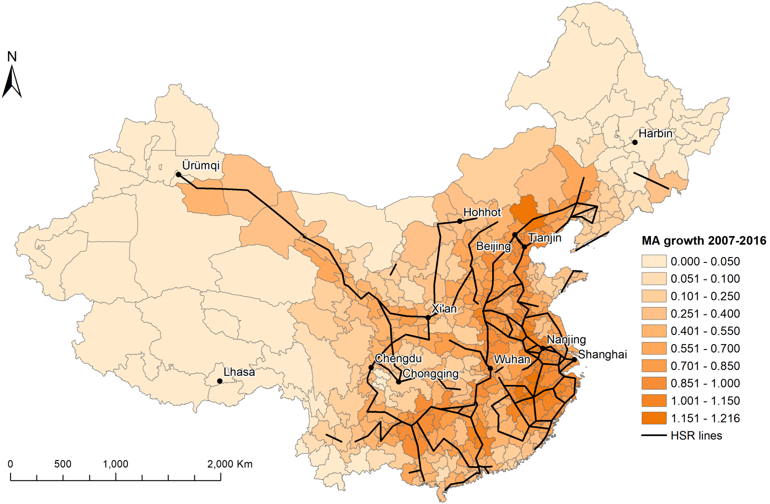
... but we probably shouldn't believe it

HSR Lines and Counterfactuals



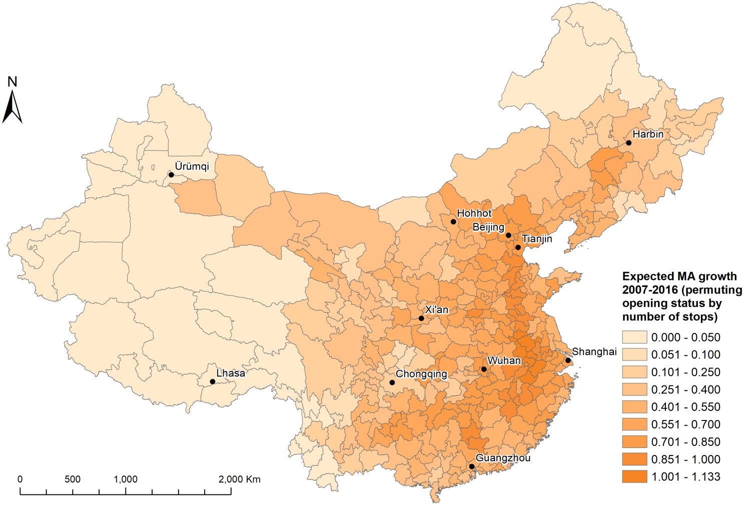
Counterfactuals permute which lines opened by 2016, conditional on length

An Example Counterfactual HSR Network



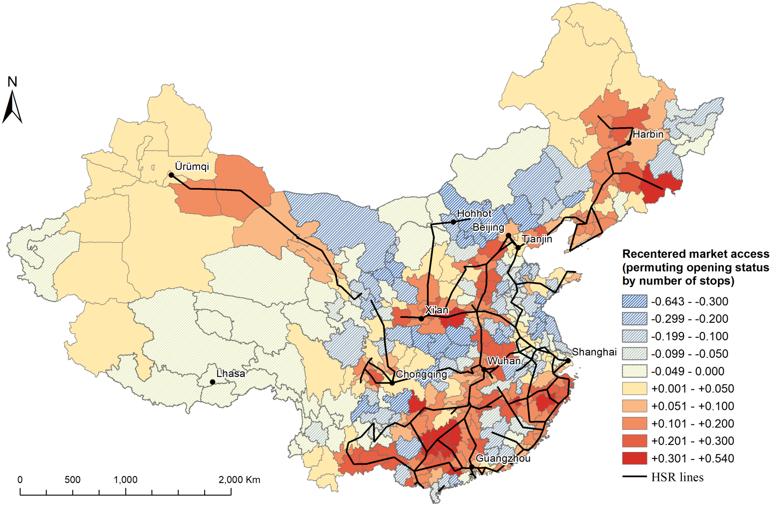
Seems ok...

Expected Market Access Across Counterfactuals



Darker regions see more MA growth regardless of which lines are built first

Recentered Market Access



Recentered IV compares region that saw more MA growth than expected (in red) to those that saw less MA growth than expected (in blue)

Balance Tests

	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

Recentered MA growth can't be reliably predicted from geography

No Market Access Effect with Recentering/Controlling

	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

Summary

A "design-based" approach to formula IVs sheds new light on longstanding identification strategies in economics (e.g. shift-share IV)...

- ... while also suggesting novel strategies leveraging more complex instrument constructions (e.g. market access models)
- We're still learning of new empirical settings using formula IVs!

Summary

A "design-based" approach to formula IVs sheds new light on longstanding identification strategies in economics (e.g. shift-share IV)...

- ... while also suggesting novel strategies leveraging more complex instrument constructions (e.g. market access models)
- We're still learning of new empirical settings using formula IVs!

Some open questions:

- Asymptotic inference (non-standard clustering; e.g. Adao et al. '19)
- Estimated / machine learnt shock assignment processes?
- Using recentered IV for structural models (stay tuned...)

Summary

A "design-based" approach to formula IVs sheds new light on longstanding identification strategies in economics (e.g. shift-share IV)...

- ... while also suggesting novel strategies leveraging more complex instrument constructions (e.g. market access models)
- We're still learning of new empirical settings using formula IVs!

Some open questions:

- Asymptotic inference (non-standard clustering; e.g. Adao et al. '19)
- Estimated / machine learnt shock assignment processes?
- Using recentered IV for structural models (stay tuned...)

Thank you!