Non-Random Exposure to Exogenous Shocks: Theory and Applications

Kirill Borusyak
UCL & CEPR

Peter Hull
Brown & NBER

NBER SI: Methods Section in Labor Studies, July 2021

Motivation

Many economic questions involve the causal effects of treatments x_i that are computed from multiple sources of variation by a known formula

Motivation

Many economic questions involve the causal effects of treatments x_i that are computed from multiple sources of variation by a known formula

- How can we credibly estimate the effects of such x_i when some, but not all, of its determinants are as-good-as-randomly assigned?

Motivation

Many economic questions involve the causal effects of treatments x_i that are computed from multiple sources of variation by a known formula

- How can we credibly estimate the effects of such x_i when some, but not all, of its determinants are as-good-as-randomly assigned?

1. Spatial/network/GE spillover treatments: e.g. the number of neighbors selected for a randomized intervention:

Motivation

Many economic questions involve the causal effects of treatments x_i that are computed from multiple sources of variation by a known formula

- How can we credibly estimate the effects of such x_i when some, but not all, of its determinants are as-good-as-randomly assigned?

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

Motivation

Many economic questions involve the causal effects of treatments x_i that are computed from multiple sources of variation by a known formula

- How can we credibly estimate the effects of such x_i when some, but not all, of its determinants are as-good-as-randomly assigned?

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:

Motivation

Many economic questions involve the causal effects of treatments x_i that are computed from multiple sources of variation by a known formula

- How can we credibly estimate the effects of such x_i when some, but not all, of its determinants are as-good-as-randomly assigned?

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

Motivation

Many economic questions involve the causal effects of treatments x_i that are computed from multiple sources of variation by a known formula

- How can we credibly estimate the effects of such x_i when some, but not all, of its determinants are as-good-as-randomly assigned?

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:

Motivation

Many economic questions involve the causal effects of treatments x_i that are computed from multiple sources of variation by a known formula

- How can we credibly estimate the effects of such x_i when some, but not all, of its determinants are as-good-as-randomly assigned?

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

Motivation

Many economic questions involve the causal effects of treatments x_i that are computed from multiple sources of variation by a known formula

- How can we credibly estimate the effects of such x_i when some, but not all, of its determinants are as-good-as-randomly assigned?

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

Goal: to avoid non-experimental assumptions (e.g. parallel trends)

This Paper: Contributions

1. Non-random exposure to as-good-as-random shocks generates systematic variation in x_i, which can lead to omitted variable bias

This Paper: Contributions

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

This Paper: Contributions

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"

This Paper: Contributions

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

This Paper: Contributions

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 $\mu_i =$ the average x_i across counterfactuals, by simulation — *the key confounder*

This Paper: Contributions

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*
   - Recenter x_i by μ_i (i.e. instrument x_i with $x_i - \mu_i$) or control for μ_i

This Paper: Contributions

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 $\mu_i =$ the average x_i across counterfactuals, by simulation — *the key confounder*
    - Recenter x_i by μ_i (i.e. instrument x_i with $x_i - \mu_i$) or control for μ_i
    - Alternative solutions are often infeasible/inefficient (e.g. directly instrumenting with shocks or controlling for all features of exposure)

This Paper: Contributions

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 $\mu_i =$ the average x_i across counterfactuals, by simulation — *the key confounder*
   - Recenter x_i by μ_i (i.e. instrument x_i with $x_i - \mu_i$) or control for μ_i
   - Alternative solutions are often infeasible/inefficient (e.g. directly instrumenting with shocks or controlling for all features of exposure)

3. Same counterfactuals also yield inference tools and specification tests
   - Via randomization inference

(Some) Related Literature

Methodological:

- **Propensity scores:** *Rosenbaum-Rubin 1983, Abadie 2003, Hirano-Imbens 2004*
- **Network spillovers:** *Aronow 2012, Manski 2013, Aronow-Samii 2017*
- **Linear shift-share IV:** *Borusyak et al. 2021, Adão et al. 2019*
- **Randomization inference:** *Fischer 1935, Hodges-Lehmann 1963, Rosenbaum 2002, Imbens-Rosenbaum 2005, Lehmann-Romano 2006, Athey et al. 2018*
- **Optimal instruments:** *Chamberlain 1987, 1992, Adão et al. 2021*

Applied:

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

Outline

1. Motivating examples:
   - Market access effects
   - Effects of program eligibility
2. General framework
3. Practical relevance in applications:
   - Estimate employment effects of China high-speed rail construction while addressing OVB from non-random HSR exposure
   - Efficiently estimate Medicaid eligibility effects from state-level shocks

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

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

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?

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

Start from no roads, assume equal population everywhere

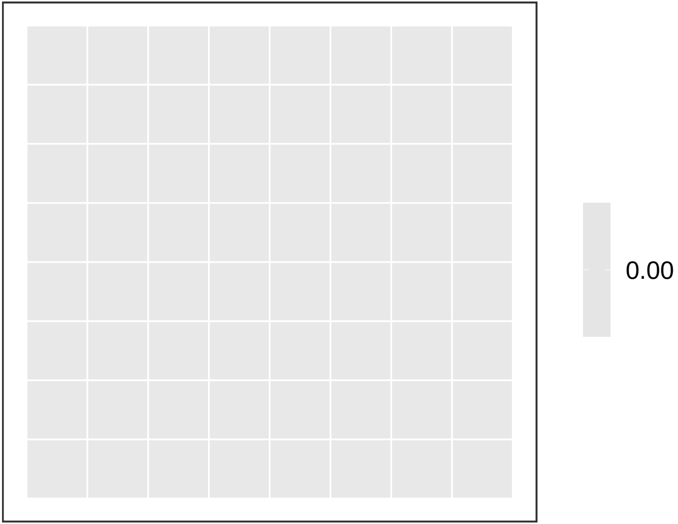


Illustration: Market Access on a Square Island

Randomly connect adjacent regions by road

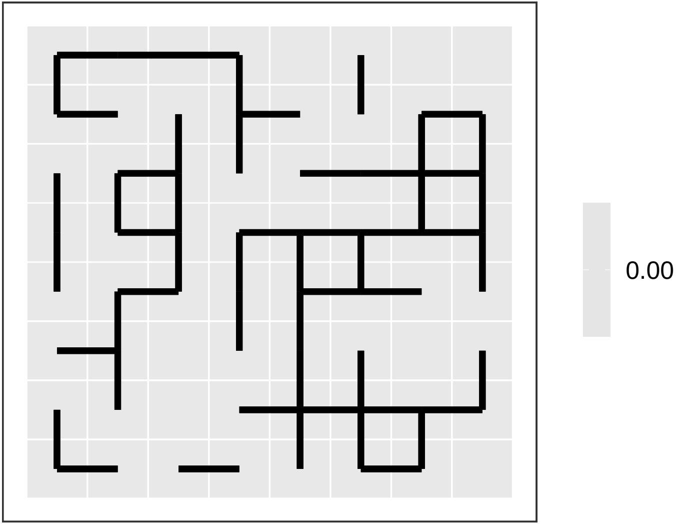


Illustration: Market Access on a Square Island

Randomly connect adjacent regions by road and compute MA growth

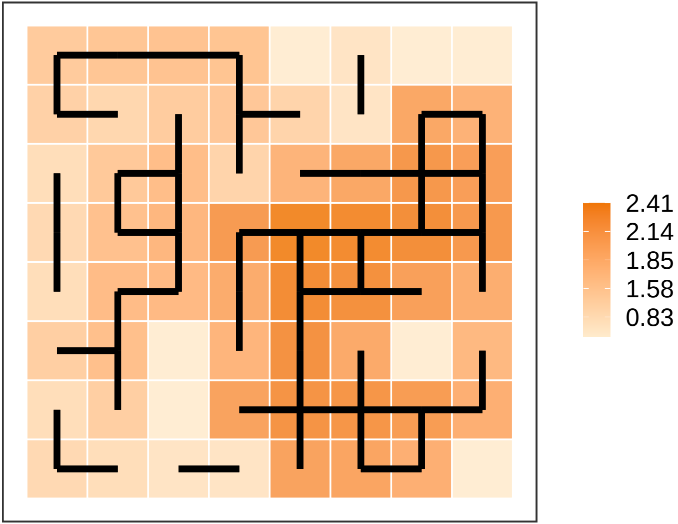


Illustration: Market Access on a Square Island

Randomly connect adjacent regions by road and compute MA growth

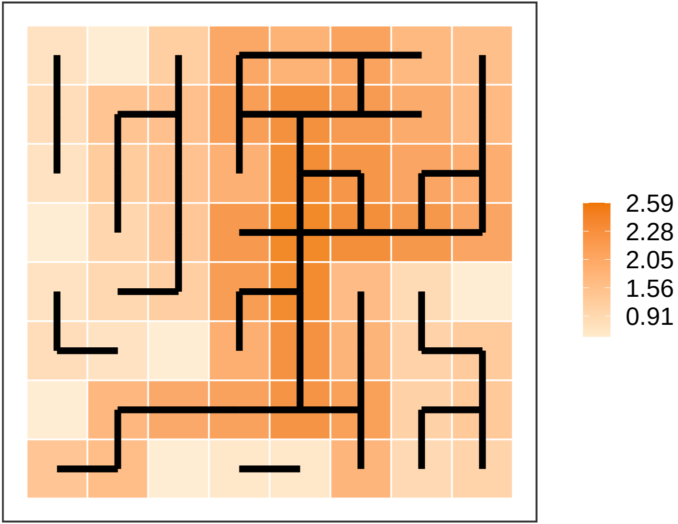
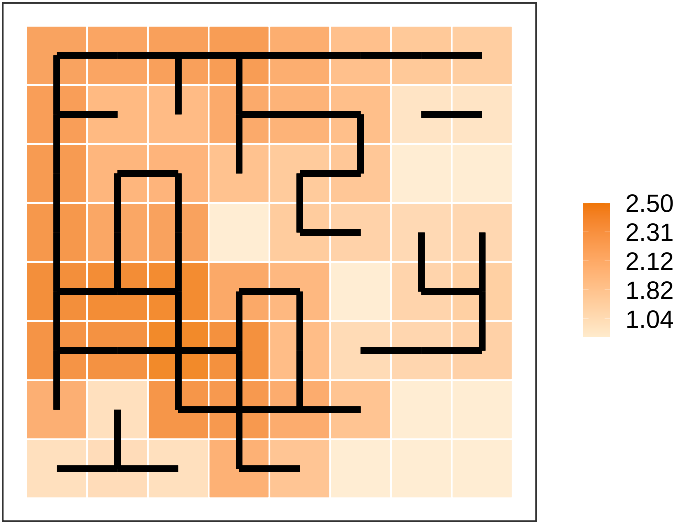


Illustration: Market Access on a Square Island

Randomly connect adjacent regions by road and compute MA growth



Expected Market Access Growth μ_i

Some regions get systematically more MA

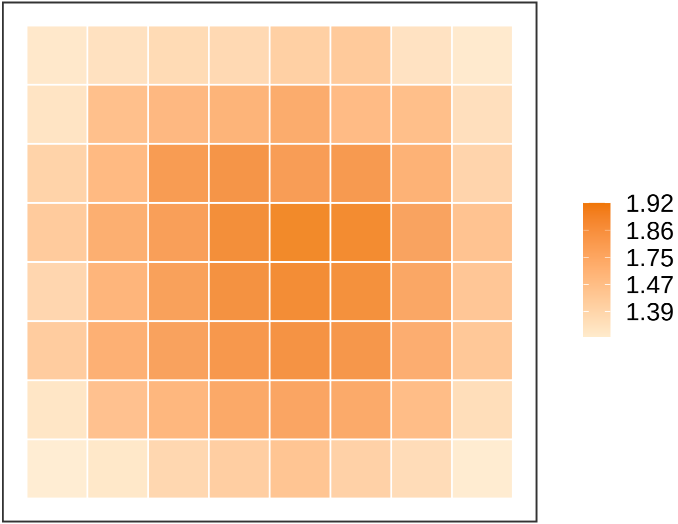


Illustration: High-Speed Rail in China

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

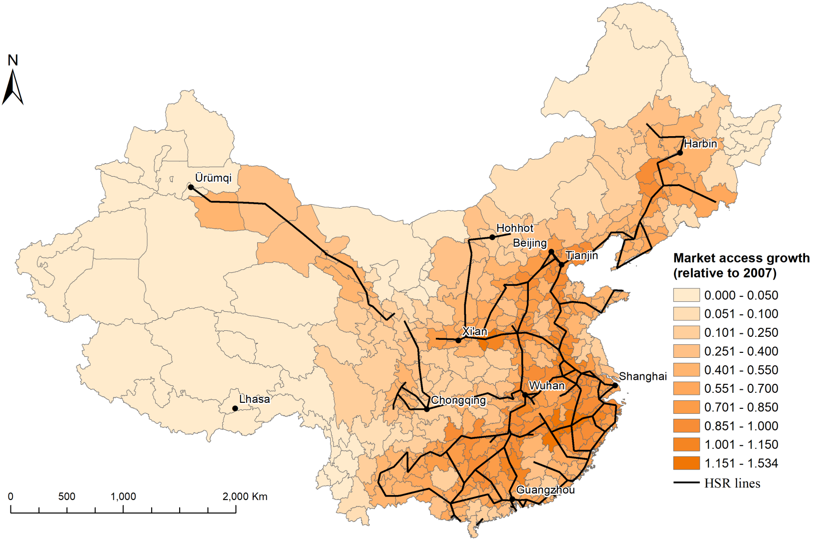


Illustration: High-Speed Rail in China

A counterfactual draw of 83 lines by 2016

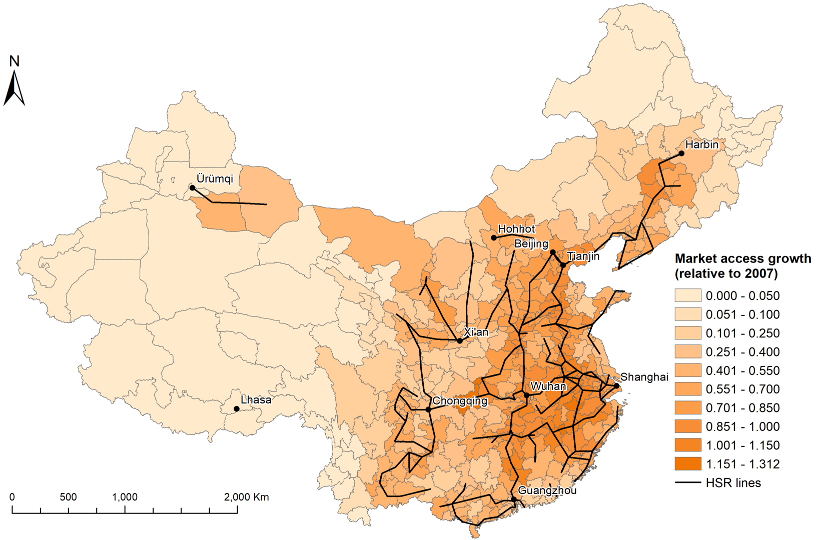
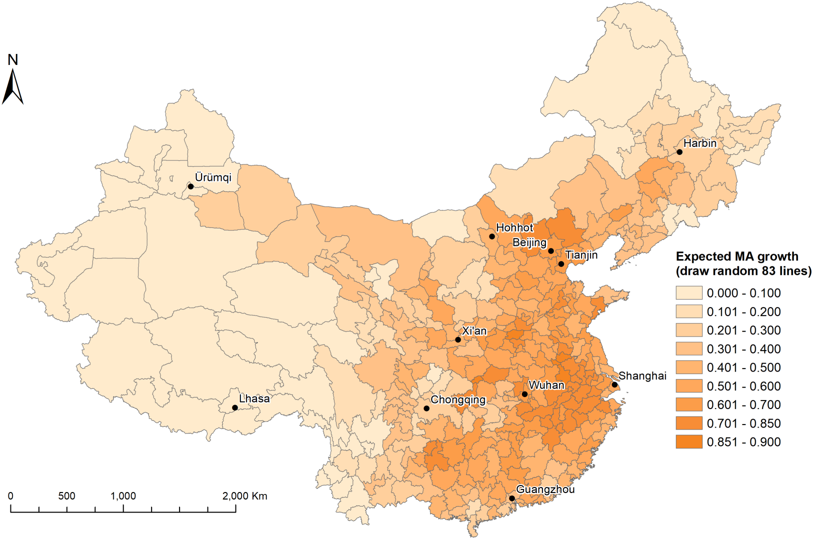


Illustration: High-Speed Rail in China

Expected MA growth, μ_i



OVB and Recentering Solution

- Systematic variation in MA growth can generate OVB

OVB and Recentering Solution

- Systematic variation in MA growth can generate OVB
  - E.g. land values fall in the periphery because of rising sea levels

OVB and Recentering Solution

- 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

OVB and Recentering Solution

- 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”:

OVB and Recentering Solution

- 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":

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

OVB and Recentering Solution

- 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":

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

  - Compares MA from actual and counterfactual shocks

OVB and Recentering Solution

- 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":

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

  - Compares MA from actual and counterfactual shocks
  - By construction, is uncorrelated with any geography-based trends in ε

OVB and Recentering Solution

- 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":

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

  - Compares MA from actual and counterfactual shocks
  - By construction, is uncorrelated with any geography-based trends in ε
  - Thus, recentered MA is a valid instrument for realized MA growth!

Avoiding Bias from Non-Random Exposure: An Algorithm

1. Measure MA from realized (exogenous) transportation shocks and preexisting geography
2. Consider many counterfactual sets of transportation shocks
   - Requires to formalize the natural experiment: what's random?
   - E.g. random timing or placement of lines
3. Recompute MA growth every time and take the average: expected MA growth, μ_i
4. Recenter realized MA growth by μ_i or add it as a control
5. Consider using counterfactual shocks for randomization inference

Motivating Example 2: Effects of Program Eligibility

The effects of individual's eligilibity 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

Motivating Example 2: Effects of Program Eligibility

The effects of individual's eligilibity 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

Motivating Example 2: Effects of Program Eligibility

The effects of individual's eligilibity 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

Motivating Example 2: Effects of Program Eligibility

The effects of individual's eligilibity 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

Motivating Example 2: Effects of Program Eligibility

The effects of individual's eligilibity 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

Our approach:

- Formalize the policy experiment as "all permutations of g across states are equally likely"

Motivating Example 2: Effects of Program Eligibility

The effects of individual's eligilibity 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

Our 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

Motivating Example 2: Effects of Program Eligibility

The effects of individual's eligilibity 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

Our approach:

- Formalize the policy experiment as "all permutations of g across states are equally likely"
- Compute μ_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)

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

Our 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 & Language

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...

General Setting & Language

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 measures predetermined "exposure"; $f_i(\cdot)$ 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

General Setting & Language

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 measures predetermined "exposure"; $f_i(\cdot)$ 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. uniform across permutations of g)

Results

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

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

Results

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

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

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

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

Results

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

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

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

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

- 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
- We characterize the **asy. efficient** recentered IV among all $f_i(\cdot)$

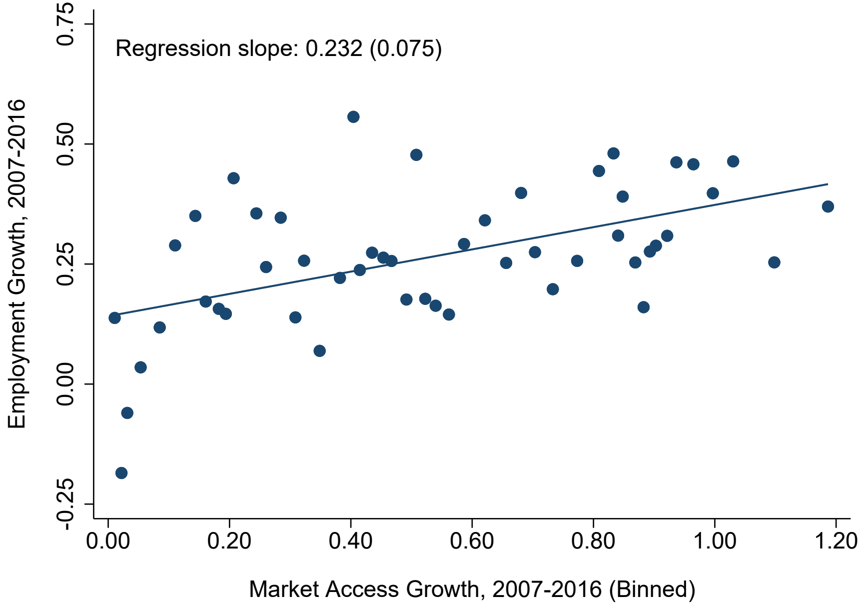
App. 1: Market Access from Chinese High-Speed Rail

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

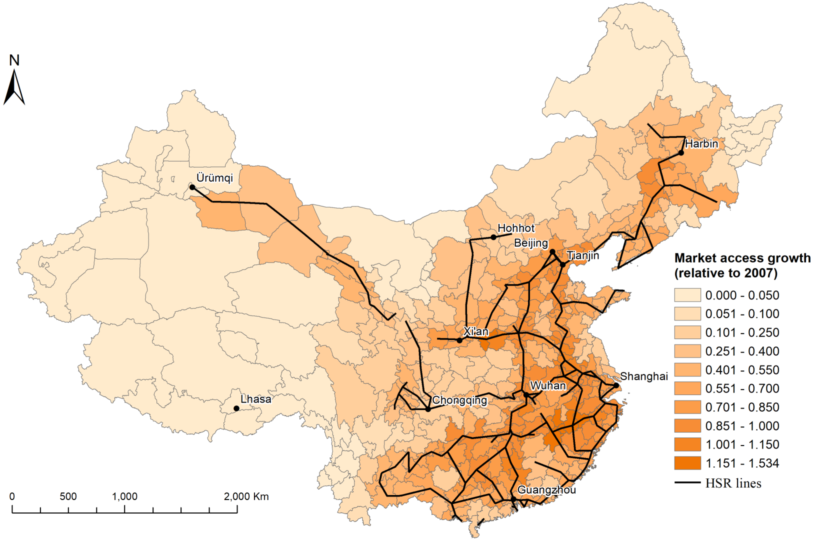
Setting: Chinese HSR; 83 lines built 2008–2016, 66 yet unbuilt

- 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

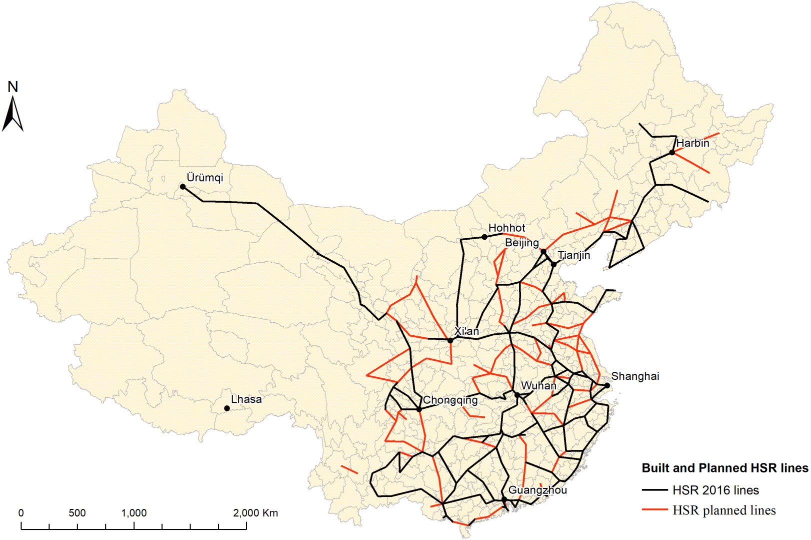


But high vs low MA growth is not the most convincing contrast!

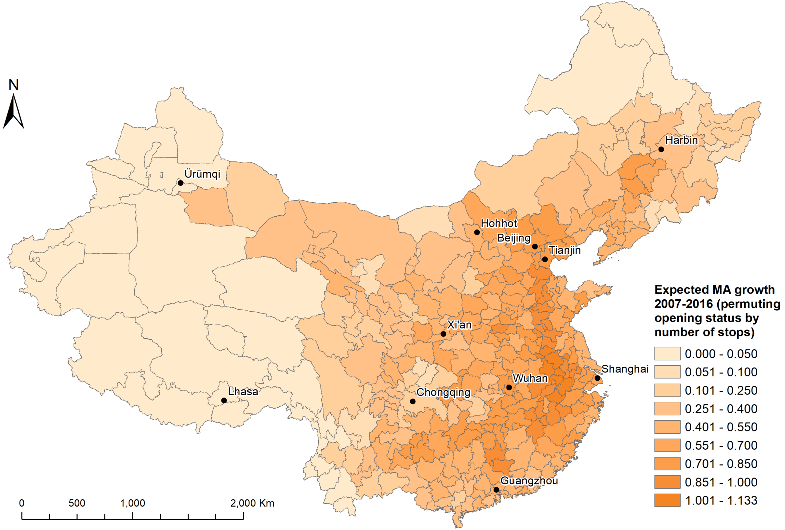


Built and Planned HSR Lines

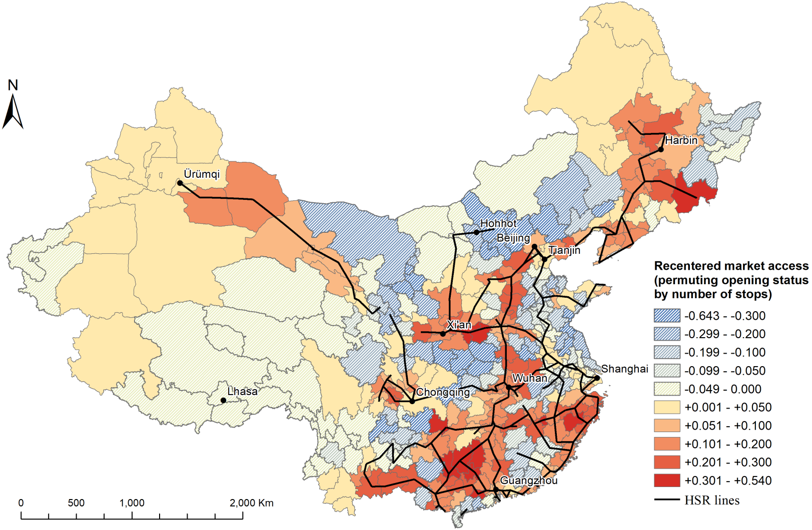
We assume random timing of built & planned lines with the same # of links ⇒ reshuffle them accordingly e.g.



Expected Market Access Growth (2007–2016), μ_i

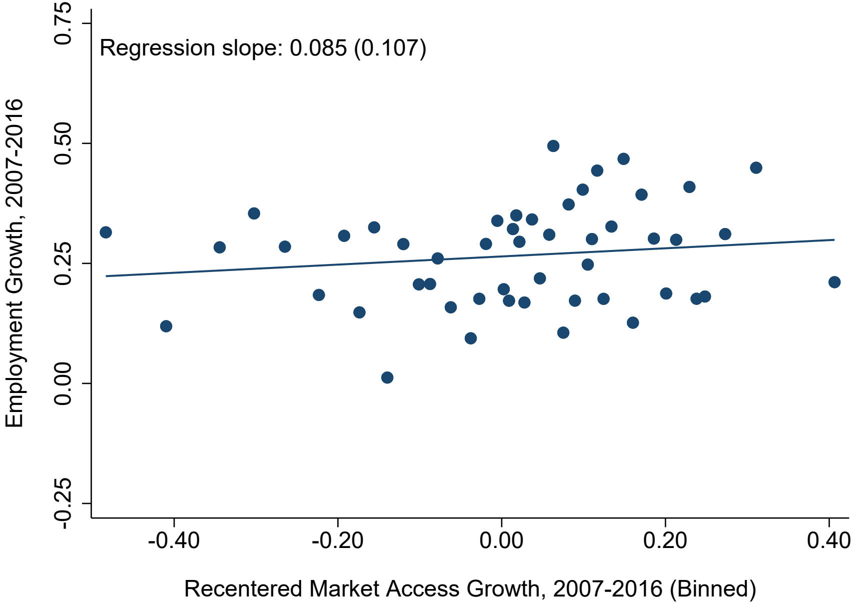


Recentered Market Access Growth (2007–2016), $\tilde{z}_i$



Specification tests pass Balance Regressions

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

Robustness LATE Weights

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

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

Compare two estimators valid under the same assumptions:

- Simulated IV: uses state-level variation only; here, simply an expansion dummy
- Our recentered IV: predict eligibility from expansion decisions & non-random demographics, and recenter
- Recentered IV has better first-stage prediction $\Rightarrow$ ≈ 3 times smaller standard errors

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. Baseline Controls						
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)
	[0.080,0.218]	[0.051,0.094]	[-0.109,0.010]	[-0.039,-0.008]	[-0.035,0.053]	[-0.021,0.004]
Panel B. With Demographics × Post						
Eligibility	0.135	0.073	−0.050	−0.024	0.003	−0.008
	(0.029)	(0.010)	(0.022)	(0.007)	(0.013)	(0.005)
	[0.082,0.223]	[0.051,0.096]	[-0.114,-0.002]	[-0.041,-0.008]	[-0.038,0.036]	[-0.020,0.005]
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. Baseline 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

First stage Pre-trends Power curve

Other Settings where Recentering Is Relevant

- Network spillovers (e.g. Miguel-Kremer 2004, Carvalho et al. 2020)
- Linear shift-share IV (e.g. Autor et al. 2013, Borusyak et al. 2021)
- 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 media access (e.g. Olken 2009, Yanagizawa-Drott 2014)

Summary

We develop a general framework for treatments and instruments computed from multiple sources of variation, only some of which are random

- Formalize the expected instrument as *the* relevant confounder
- Show that recentering by it purges OVB
- Feasible as long as researchers formalize natural experiments via counterfactual shocks

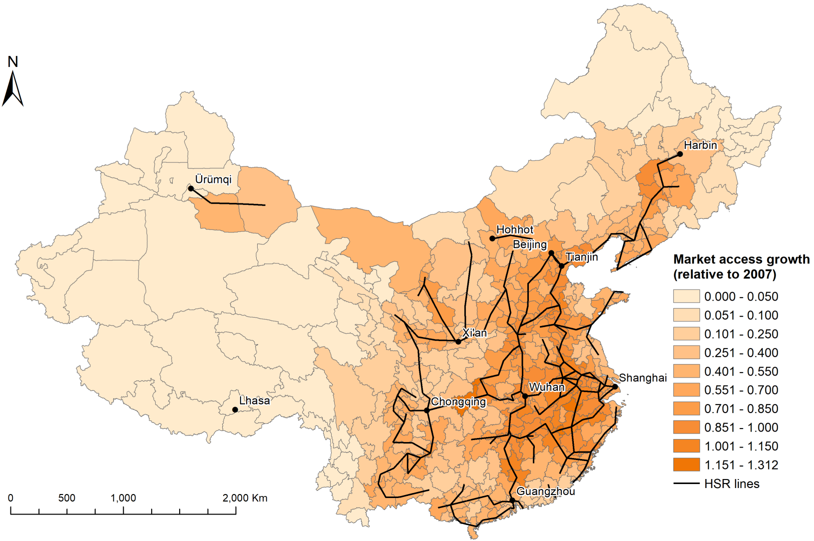
This framework empirically relevant:

- A simple recentering based on the timing of Chinese HSR construction largely "kills" OLS estimates of market access effects
- A more powerful recentered prediction of Medicaid eligibility from state-level shocks yields ≈ 3 times smaller standard errors
- Practical implications for many other common research designs

Thank You!

Appendix

Simulated HSR Map and Market Access Growth



Back

Market Access 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.

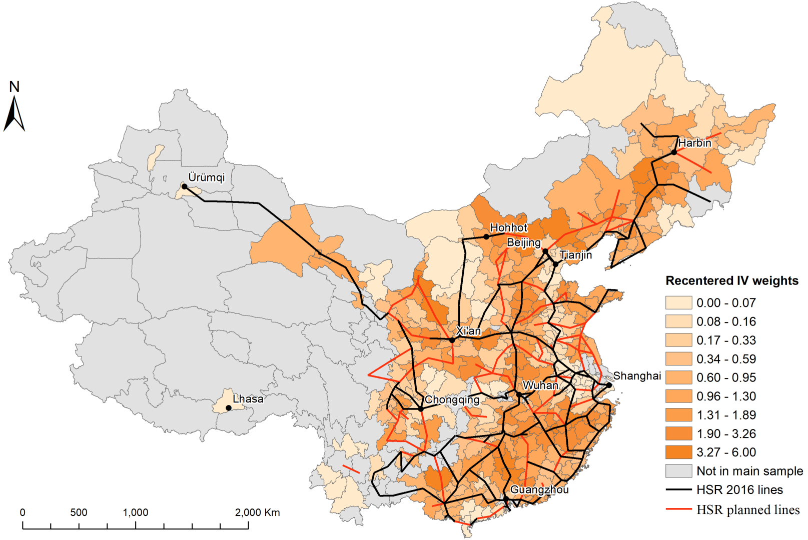
Back

Market Access Robustness Checks

	Unadjusted OLS (1)	Recentered IV (2)	Controlled OLS (3)
Panel A. Using Leave-One-Out Market Access (N=274)			
Market Access Growth	0.229	0.081	0.070
	(0.078)	(0.104)	(0.103)
		[-0.360, 0.357]	[-0.124,216]
Expected Market Access Growth			0.207
			(0.118)
Panel B. Dropping Province Capitals (N=247)			
Market Access Growth	0.215	0.068	0.060
	(0.078)	(0.104)	(0.099)
		[-0.303, 0.321]	[-0.202, 0.320]
Expected Market Access Growth			0.303
			(0.097)
Panel C. Using HSR Connectivity (N=274)			
Connectivity Growth	0.155	0.051	0.049
	(0.049)	(0.057)	(0.056)
		[-0.037, 0.149]	[-0.041, 0.145]
Expected Connectivity Growth			0.257
			(0.071)
Panel D. Adding Province Fixed Effects (N=268)			
Market Access Growth	0.108	0.099	0.097
	(0.046)	(0.070)	(0.079)
		[-0.014, 0.268]	[-0.018, 0.270]
Expected Market Access Growth			0.121
			(0.071)
Recentered	No	Yes	Yes

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

What LATE Does the Recentered IV Estimate?



Back

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 Back

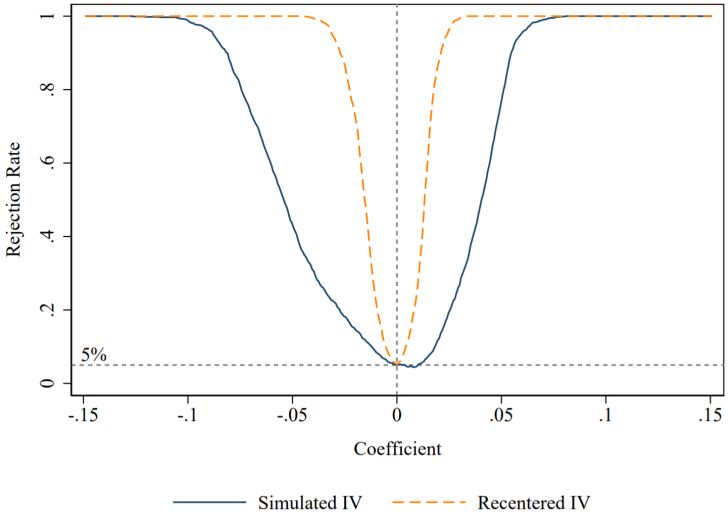
Medicaid Eligibility Pre-Trends

	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. Baseline Controls						
Eligibility	−0.022	−0.020	0.015	0.011	0.011	0.007
	(0.009)	(0.004)	(0.017)	(0.004)	(0.017)	(0.005)
	[-0.042,0.009]	[-0.028,-0.008]	[-0.021,0.071]	[0.003,0.020]	[-0.026,0.059]	[-0.005,0.020]
Panel B. With Demographics × Post						
Eligibility	−0.023	−0.020	0.019	0.014	0.016	0.011
	(0.010)	(0.004)	(0.014)	(0.004)	(0.016)	(0.005)
	[-0.040,0.012]	[-0.027,-0.009]	[-0.022,0.056]	[0.005,0.022]	[-0.029,0.049]	[-0.002,0.022]
Exposed Sample	N	Y	N	Y	N	Y
States	43	43	43	43	43	43
Individuals	2,400,142	425,112	2,400,142	425,112	2,400,142	425,112

IV regressions using one of the two instruments. Baseline 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

Back

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 Back