Canonical Research Designs V: Bartik and Simulated Instruments

Paul Goldsmith-Pinkham

April 7, 2022

Roadmap for Today

- In some cases, the source of exogeneous variation (either in an IV setting, or just OLS) is straightforward
  - There is a single policy or source of variation
- However, in other settings, there are more complicated sources of variation exploited to identify effects. Today we'll focus on two:
  - Bartik (shift-share) instruments: three recent papers on commonly used identification approach
  - Simulated instruments: reframe an older literature in a new light using Borusyak and Hull (2022) paper
- Key historical feature of these approaches is that they had an “intuitive” feature of identification, but formal properties were not established for several decades
  - Analagous to staggered DinD lit!

Bartik instruments are used everywhere

Bartik

Labor supply + immigration

Trade

Finance, innovation, others...

- Thread that links all Bartik applications:
  - local markets composed of many "categories"
  - need for identification
- Approach has been used since the early 90:
  - sometimes called "shift-share" or "industry mix" instruments

Examples of Bartik instruments in many subfields

Immigration: Altonji and Card (1991), Card (2001)

Bank Lending: Amiti and Weinstein (2018), Greenstone, Mas and Nguyen (2015)

Market Size + Demography: Acemoglu and Linn (2004), Jaravel (2018)

Labor Supply Elasticity: Blanchard and Katz (1992), **Bartik** (1991)

Fiscal Multipliers: Nakamura and Steinsson (2014)

Trade + Labor: Autor, Dorn, and Hanson (2013), Autor, Dorn and Hanson (2018), etc.

Foreign Aid: Nunn and Qian (2014)

Portfolio Allocation: Calvet, Campbell, and Sodini (2009)

Trade + Prices: Piveteau and Smagghue (2017), de Roux et al. (2017)

Automation: Acemoglu and Restrepo (2017)

Many paths lead to Bartik

- Diverse literature leads to many motivations and justifications for Bartik approach
- Two distinct approaches in the literature:
  1. Applied micro statistical approach: interested in a reduced form causal relationship; need an instrument that is uncorrelated with error term; make argument that Bartik instrument is defensible
  2. Structural approach: interested in particular parameters from model; assumptions of model motivate certain estimating equations
- So what is the Bartik approach anyway?

Motivation: local labor market approaches + reduced form

Consider a local labor market regression like the following:

$$y_l = \beta_0 + \beta x_l + \epsilon_l$$

- $\mathbb{E}[x_l \epsilon_l] \neq 0 \Rightarrow$ need an instrument to estimate β
- E.g. Autor, Dorn and Hanson (2013) setting:
  - l: location (commuting zone)
  - y_l: manufacturing employment *growth*
  - x_l: import exposure to China *growth*
  - β: effect of rise of China on manufacturing employment
  - an instrument for location-level exposure to trade with China

The Bartik instrument

Accounting identity #1:

$$x_l = \sum_{k=1}^{K} z_{lk} g_{lk}$$

- z_{lk}: location-industry shares (Z_l)
- g_{lk}: location-industry growth (in imports) rates (G_l)

Accounting identity #2:

$$\underbrace{g_{lk}}_{\text{location-industry}} = \underbrace{g_k}_{\text{industry}} + \underbrace{\tilde{g}_{lk}}_{\substack{\text{idiosyncratic} \\ \text{location- industry}}}$$

Infeasible Bartik:

$$B_l = \sum_{k=1}^{K} z_{lk} g_k$$

This gives us a simple 2SLS structure

$$
\begin{aligned}
y_l &= \beta_0 + \beta x_l + \epsilon_l \\
x_l &= \pi_0 + \pi_1 B_l + u_l \\
B_l &= \sum_{k=1}^{K} z_{lk} g_k \\
g_{lk} &= g_k + \tilde{g}_{lk}
\end{aligned}
$$

China shock: e.g., Autor, Dorn and Hanson (2013)

- z_{lk}: location (l) industry (k) composition
- g_{lk}: location (l) industry (k) growth in imports from China
- g_k: industry *k* growth of imports (from China)

Other instruments have this structure

$$
\begin{aligned}
y_l &= \beta_0 + \beta x_l + \epsilon_l \\
x_l &= \pi_0 + \pi_1 B_l + u_l \\
B_l &= \sum_{k=1}^{K} z_{lk} g_k \\
g_{lk} &= g_k + \tilde{g}_{lk}
\end{aligned}
$$

Immigrant enclave: e.g., Altonji and Card (1991)

- z_{lk}: share of people from foreign k living in l (in a base period)
- g_{lk}: growth in number of people from k to l
- g_k: growth in people from k nationally

Other instruments have this structure

$$
\begin{aligned}
y_l &= \beta_0 + \beta x_l + \epsilon_l \\
x_l &= \pi_0 + \pi_1 B_l + u_l \\
B_l &= \sum_{k=1}^{K} z_{lk} g_k \\
g_{lk} &= g_k + \tilde{g}_{lk}
\end{aligned}
$$

Bank-lending relationships: e.g., Greenstone, Mas and Nguyen (2015)

- z_{lk}: location (l) share of loan origination from bank k
- g_{lk}: loan growth in location l by bank k
- g_k: part of loan growth due to bank supply shock

Other instruments have this structure

$$y_l = \beta_0 + \beta x_l + \epsilon_l$$
$$x_l = \pi_0 + \pi_1 B_l + u_l$$
$$B_l = \sum_{k=1}^{K} z_{lk} g_k$$
$$g_{lk} = g_k + \tilde{g}_{lk}$$

Market size and demography: e.g., Acemoglu and Linn (2004)

- z_{lk}: spending share on drug l from age group k
- g_{lk}: growth in spending of group k on drug l
- g_k: growth in spending of group k (due to population aging)

What's necessary for consistency?

$$y_l = \beta_0 + \beta x_l + \epsilon_l$$
$$x_l = \pi_0 + \pi_1 B_l + u_l$$
$$B_l = \sum_{k=1}^{K} z_{lk} g_k$$
$$g_{lk} = g_k + \tilde{g}_{lk}$$

- We need B_l to be a valid instrument
- Requires two conditions with constant effects:
  1. Relevance: $\pi_1 \neq 0$, e.g. $Cov(B_l, x_l) \neq 0$
  2. Exclusion: $E(B_l \epsilon_l) = 0$
- Key flaw in this literature until recently: economic + statistical content of exclusion has been vague and sometimes confused

Key thing to remember from today

- Assuming independence or exogeneity on the basis of a model does not necessarily make it true
    - E.g. Hausman instruments in IO models – model may assume that exclusion restriction is satisfied, but not necessarily true in reality
- Assuming that two things are independent because they don't seem "related" doesn't make it true
    - Bartik literature many times argues that national nature of shocks "decouples" the instrument from local market conditions. However, it still exploits local characteristics. Need to make very specific arguments to validate claim (will come to this).
- When evaluating an identification strategy, you should be able to describe counterfactual claims using the measure. This is typically not concrete in Bartik – try to make it concrete! What is exactly changing in China? Why is it random?

More general econometric set-up

$$
\begin{aligned}
y_{lt} &= \mathbf{D}_{lt}\beta_0 + x_{lt}\beta + \epsilon_{lt}, \\
x_{lt} &= \mathbf{D}_{lt}\tau + B_{lt}\gamma + \eta_{lt} \\
\mathbf{D}_{lt} &= \text{controls, f.e.} \\
g_{lkt} &= g_{kt} + \tilde{g}_{lkt} \\
B_{lt} &= \sum_{k=1}^{K} z_{lk0} g_{kt}, \\
&\{\{x_{lt}, \mathbf{D}_{lt}, \epsilon_{lt}\}_{t=1}^{T}\}_{l=1}^{L}, \text{ iid, } L \to \infty
\end{aligned}
$$

Assumptions for IV in terms of B_{lt}:

- Exogeneity: $\mathbb{E}\left[B_{lt}\epsilon_{lt} | \mathbf{D}_{lt}\right] = 0$
- Relevance: $\text{Cov}\left[B_{lt}, x_{lt} | \mathbf{D}_{lt}\right] \neq 0$

Question:

- What do these statements about B_{lt} imply about z_{lk0} and g_{kt}?

Recent Literature on this topic

- Three papers addressed this question, and can be split into two grouping
- The division between papers can be split based on focus on z_{lk0} vs. g_{kt}
  1. Goldsmith-Pinkham, Sorkin and Swift (2020) focus on z_{lk0} and make an analogy to difference-in-differences
  2. Adao, Kolesar and Morales (2019) and Borusyak, Hull and Jaravel (2020) focus on g_{kt}, and make a strong connection to the design based approach (e.g. these are as-if random shocks)
- Key problem, historically, in this literature, was the lack of a coherent defense of the identifying variation
  - These papers provide a way of doing this! But you have to pick one approach

Understanding the identifying assumption in GPSS: Three special cases

1. One time period, two industries
2. T time periods, two industries
3. One time period, K industries

Special case #1: One time period, two industries

- $z_{l2} = 1 - z_{l1}$
- Bartik:

$$\begin{aligned} B_l &= z_{l1} g_1 + z_{l2} g_2 = z_{l1} g_1 + (1 - z_{l1}) g_2 \\ &= g_2 + (g_1 - g_2) z_{l1} \end{aligned}$$

First-stage:

$$\begin{aligned} x_l &= \gamma_0 + \gamma B_l + \eta_l \\ x_l &= \underbrace{\gamma_0 + \gamma g_2}_{\text{constant}} + \underbrace{\gamma(g_1 - g_2)}_{\text{coefficient}} z_{l1} + \eta_l \end{aligned}$$

The instrument is z_{l1}, while g_k affects relevance

▸ Why OLS is biased

Special case #2: T time periods, two industries

Panel Bartik:

$$B_{lt} = z_{l10} g_{1t} + z_{l20} g_{2t} = g_{2t} + \underbrace{\Delta_{gt}}_{g_{1t}-g_{2t}} z_{l10}$$

First stage:

$$x_{lt} = \tau_l + \tau_t + \gamma B_{lt} + \eta_{lt}$$
$$x_{lt} = \tau_l + \underbrace{(\tau_t + \gamma g_{2t})}_{\tilde{\tau}_t} + \underbrace{\gamma \Delta_{gt}}_{\tilde{\gamma}_t} z_{l10} + \eta_{lt}$$

- Industry shares times time period is the instrument
- (Updated industry shares: similar)

Special case #2: T time periods, two industries

- Analogy to continuous difference-in-differences
  - Δ_{gt} is size of policy
  - z_{l10} is exposure to policy
- Sometimes a "pre-period" before policy: test for parallel pre-trends
  - E.g., in ADH, what happens from 1970 to 1990?

Special case #3: One time period, K industries

- G: $K \times 1$ vector of g_k
- Z: $L \times K$, matrix of Z_l
- $Y^\perp$, $X^\perp$, $B = (ZG)$: $L \times 1$, vectors of $y_l^\perp$, $x_l^\perp$ and B_l
- Ω: $K \times K$

$$\hat{\beta}_{Bartik} = \frac{B'Y^\perp}{B'X^\perp}$$
$$\hat{\beta}_{GMM} = \frac{(X^{\perp\prime}Z)\Omega(Z'Y^\perp)}{(X^{\perp\prime}Z)\Omega(Z'X^\perp)}$$

If $\Omega = (GG')$, then $\hat{\beta}_{Bartik} = \hat{\beta}_{GMM}$

Full general result with T time periods and K industries

Two estimators are numerically identical:

- TSLS with Bartik instrument
- GMM with industry shares $\times$ time period as instruments and a particular weight matrix

$$\hat{\beta}_{Bartik} = \frac{\mathbf{B}'\widetilde{\mathbf{Y}}^{\perp}}{\mathbf{B}'\widetilde{\mathbf{X}}^{\perp}}$$

$$\hat{\beta}_{GMM} = \frac{(\mathbf{X}^{\perp\prime}\widetilde{\mathbf{Z}})\Omega(\widetilde{\mathbf{Z}}'\mathbf{Y}^{\perp})}{(\mathbf{X}^{\perp\prime}\widetilde{\mathbf{Z}})\Omega(\widetilde{\mathbf{Z}}'\mathbf{X}^{\perp})}$$

$\Omega = (\mathbf{GG}')$, and $\widetilde{\mathbf{Z}}$ is an $LT \times KT$ stacked vector of Z_0 interacted with time fixed effects and $\mathbf{G}$ is a $KT \times 1$ stacked vector of growth rates g_{kt}.

When is the estimator consistent for the estimand of interest?

What is the identification condition?

$$\hat{\beta}_{Bartik} = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T} \sum_{k=1}^{K} z_{lkt} g_{kt} y_{lt}^{\perp}}{\sum_{l=1}^{L} \sum_{t=1}^{T} \sum_{k=1}^{K} z_{lkt} g_{kt} x_{lt}^{\perp}}$$

Two ideas:

- "Shares" : talk about properties of z_{lkt}
    - Conditional exogeneity
    - *model based* – diff-in-diff style approach
- "Shocks" (Borusyak, Hull and Jaravel (2018)): talk about properties of g_{kt}
    - Random, and a large number (equivalent industry-level regression)
    - *design-based* (in spirit) – IV strategy

When are these views plausible? What do they mean?

Shares

Conditional exogeneity:

- Typically: exogenous to changes in error term, not levels of outcome
- Standard in diff-in-diff (exclusion): in a period, exposure to an industry matters for outcome only through x

Shocks

- Large number of industries (shares are misspecified, need it to average out)
- Random shocks across industries – need the shocks to be conditionally random

How do we choose?

- The shocks approach is more design-based (which can be appealing), but requires an argument why shocks are randomly assigned
- The shares approach is model-based, so suffers from same issues as dind, but may more naturally work in your setting.

Is ADH about shocks or shares?

Shocks:

- Explains why $g_{kt}^{high-income}$ rather than g_{kt}^{US} (hard to rationalize under shares)
- Natural in a trade model: why would imports from China rise (in a trade model)? Independent industry-specific shocks

Shares:

- Explains why z_{lkt-1} rather than z_{lkt} (hard to rationalize under shocks)
- Explains why it is important for identification to study local labor markets (as opposed to parameter of interest where we want to think about spillovers)

Bottom line: a little hard to tell what exactly ADH are assuming; ADH approach does not appear to satisfy testable assumptions under GPSS, but do appear to under BHJ.

Decomposing Bartik (GPSS 2020)

(Special case of Rotemberg (1983), proposition 1)

$$\hat{\beta}_{Bartik} = \sum_k \hat{\alpha}_k \hat{\beta}_k, \quad \sum_k \hat{\alpha}_k = 1$$

IV estimate using only the k^{th} instrument:

$$\hat{\beta}_k = (Z_k' X)^{-1} Z_k' Y$$

"Rotemberg" weight:

$$\hat{\alpha}_k = \frac{g_k Z_k' X}{\sum_{k=1}^{K} g_k Z_k' X}$$

▸ Proof

Interpretation: sensitivity to misspecification elasticity

Conley, Hansen and Rossi (2012); Andrews, Gentzkow and Shapiro (2017)
Local misspecification: $\epsilon_{lt} = L^{-1/2} V_{lt} + \tilde{\epsilon}_{lt}, \quad Cov(V_{lt}, Z_{lt}) \neq 0,$

- $\sqrt{L}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \tilde{\beta}, \qquad \mathbb{E}\left[\tilde{\beta}\right] =$ bias (misspecification) of Bartik instrument
- $\sqrt{L}\left(\hat{\beta}_k - \beta_0\right) \xrightarrow{d} \tilde{\beta}_k, \qquad \mathbb{E}\left[\tilde{\beta}_k\right] =$ bias (misspecification) of kth instrument

Suppose $\beta_0 \neq 0$. Percentage bias:

$$\frac{\mathbb{E}[\tilde{\beta}]}{\beta_0} = \sum_k \alpha_k \frac{\mathbb{E}[\tilde{\beta}_k]}{\beta_0}$$

Industry with high α_k:

- an industry where it matters whether it is misspecified (endogenous)
  - because it is “important” in the estimate

Top five industries (out of 397)

	$\hat{\alpha}_k$	$g_k^{\text{high-income}}$	$\hat{\beta}_k$
Games and Toys	0.182	174.841	-0.151
Electronic Computers	0.182	85.017	-0.620
Household Audio and Video	0.130	118.879	0.287
Computer Equipment	0.076	28.110	-0.315
Telephone Apparatus	0.058	37.454	-0.305
	0.628/1.379		-0.230

*The **main source of variation in exposure is** within-manufacturing specialization in industries subject to different degrees of import competition...there is differentiation according to **local labor market reliance on labor-intensive industries**...By 2007, China accounted for over 40 percent of US imports in four four-digit SIC industries (**luggage, rubber and plastic footwear, games and toys, and die-cut paperboard**) and over 30 percent in 28 other industries, including **apparel, textiles, furniture, leather goods, electrical appliances, and jewelry**.*

— Autor, Dorn and Hanson (2013) , pg. 2123

Three tests of the identifying condition (under GPSS (2020)

1. Confounds (or correlates)
2. Pre-trends
3. Alternative estimators and overidentification

- There are *also* tests for BHJ – similar to assuming strict ignorability, you can test for balance on observables (like the confounds above) of industries and locations

Alternative estimators and overidentification tests

Basic insight in GPSS: many instruments

- Estimators (maximum likelihood): LIML, Hausman, Newey, Woutersen, Chao and Swanson (2012) HFUL (heteroskedasticity-Fuller (1977))
- Estimators (two-step): TSLS (problematic), Bartik TSLS, MBTSLS (Anatolyev (2013), and Kolesar et al (2015))

Interpretation:

- Gap between maximum likelihood and two-step estimators is evidence of misspecification

Also, overidentification tests, which provides evidence of misspecification

Test #3: Alternative estimators and overidentification

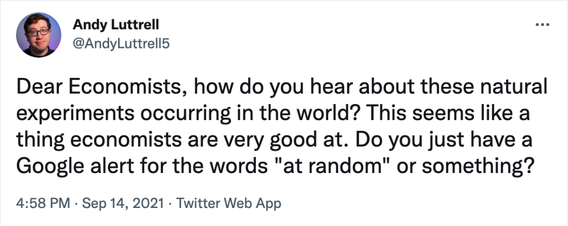
	Δ Emp	Over ID Test
OLS	-0.17	
	(0.04)	
TSLS (Bartik)	-0.62	
	(0.11)	
TSLS	-0.22	872.69
	(0.06)	[0.00]
MBTSLS	-0.33	
	(0.05)	
LIML	-2.07	1348.50
	(3.52)	[0.00]
HFUL	-1.13	1141.08
	(0.04)	[0.00]
Year and Census Division FE	Yes	
Controls	Yes	
Observations	1,444	

Switching gears: Economists have a nose for randomness

- Paraphrasing a Yale prof:

  > *Economists are really good at doing almost the right thing in empirical work.*
  >
  > *-Anonymous Yale Professor*

- Economists are clever at finding things that look convincingly “random”
  - Sometimes, it is easy to know how to use this randomness
  - **This paper is about when it is hard**


Andy Luttrell
@AndyLuttrell5

Dear Economists, how do you hear about these natural experiments occurring in the world? This seems like a thing economists are very good at. Do you just have a Google alert for the words "at random" or something?

4:58 PM · Sep 14, 2021 · Twitter Web App

Borusyak and Hull (2022) on exploiting randomness in IV

- Two key parts to this paper:
  1. Highlighting how seemingly complicated research designs can be framed as generalized propensity scores
  2. How complicated research designs can suffer from *interference*

- There are many interesting results that spiral out from these two insights, but these are the key kernels (third piece is thinking about uncertainty using randomization inference, but deeply tied to other pieces))

- Will first start with showing how complicated research designs $\rightarrow$ propensity scores

Non-Random Exposure to Exogenous Shocks: Theory and Applications

Kirill Borusyak
UCL and CEPR

Peter Hull
UChicago and NBER*

January 2021

Abstract

We develop new tools for estimating the causal effects of treatments or instruments that combine multiple sources of variation according to a known formula. Examples include treatments capturing spillovers in social and transportation networks, simulated instruments for policy eligibility, and shift-share instruments. We show how exogenous shocks to some, but not all, determinants of such variables can be leveraged while avoiding omitted variables bias. Our solution involves specifying counterfactual shocks that may as well have been realized and adjusting for a summary measure of non-randomness in shock exposure: the average treatment (or instrument) across such counterfactuals. We further show how to use shock counterfactuals for valid finite-sample inference, and characterize the valid instruments that are asymptotically efficient. We apply this framework to address bias when estimating employment effects of market access growth from Chinese high-speed rail construction, and to boost power when estimating coverage effects of expanded Medicaid eligibility.

Research designs from simple to complex

- Consider the trivial research design, following an RCT that randomly assigns $x_i \in \{0, 1\}$, and we want to estimate the effect of x_i on y_i:
$$y_i = \alpha + x_i\beta + \epsilon_i$$
- The research design is effectively a coin flip: $E(x_i) = p$, and each x_i is independent for each i
  - β is identified thanks to this coin flip design
- This is true even when we have covariates, w_i that stratify the experiment. We just need to control for w_i correctly: $E(x_i|w_i) = p(w_i)$ and we can estimate the ATE directly
- Effectively, the (potentially) endogenous w_i affects treatment, but if we condition correctly, we can still identify a causal effect

Research designs from simple to complex- Medicaid eligibility

- Now imagine the eligibility rules for Medicaid were being randomly assigned
  - Drawn from a bag just like marbles, completely randomly

- We can now estimate the effect of Medicaid eligibility on things like child mortality
  - Issue: eligibility is also a function of many *endogenous* features
- We consider a known function, f_i, and eligibility rules, g_i, such that $x_i = f(g_i, w_i)$ maps the w_i characteristics using the randomly drawn eligilibility rules
  - Much like w_i strata case, but more complex b/c can be high-dimensional / non-linear

Simulated instruments as a way to get a handle on this

- The challenge is that g_i is a complicated variable – it is a set of rules of that potentially complicated and hard to map to an "instrument" or "treatment"

- You don't want to just use x_i because it contains endogenous w_i

- Currie and Gruber (1996) solution: construct a variable $z_i = \sum_j f(g_i, w_j)$ which takes w from a random population (outside the state) and uses it to construct a "predicted" x
  - Intuitively, hold fixed the g_i and average over some distribution of w_j

Saving Babies: The Efficacy and Cost of Recent Changes in the Medicaid Eligibility of Pregnant Women

Janet Currie

University of California, Los Angeles and National Bureau of Economic Research

Jonathan Gruber

Massachusetts Institute of Technology and National Bureau of Economic Research

A key question for health care reform in the United States is whether expanded health insurance eligibility will lead to improvements in health outcomes. We address this question in the context of the dramatic changes in Medicaid eligibility for pregnant women that took place between 1979 and 1992. We build a detailed simulation model of each state's Medicaid policy during this era and use this model to estimate (1) the effect of changes in the rules on the fraction of women eligible for Medicaid coverage in the event of pregnancy and (2) the effect of Medicaid eligibility changes on birth

Simulated instruments as a way to get a handle on this

- The challenge is that g_i is a complicated variable – it is a set of rules of that potentially complicated and hard to map to an "instrument" or "treatment"

- You don't want to just use x_i because it contains endogenous w_i

- Currie and Gruber (1996) solution: construct a variable $z_i = \sum_j f(g_i, w_j)$ which takes w from a random population (outside the state) and uses it to construct a "predicted" x
  - Intuitively, hold fixed the g_i and average over some distribution of w_j

To the extent that relevant state- and year-specific characteristics are not captured by state and year dummies (i.e., they are not constant within a state or across states within a year), the coefficient on the fraction eligible will be biased by omitted variables. Suppose, for example, that a state recession is associated with both increases in eligibility and a higher incidence of low birth weight. Then this source of variation in eligibility could induce a spurious positive correlation between Medicaid eligibility and low birth weight.

In order to overcome this potential problem, we instrument the actual fraction eligible with a measure of the generosity of Medicaid in a state and year that depends only on the state's eligibility rules. To create our instrument, which we label the "simulated fraction eligible," we first take a sample of 3,000 women from the CPS in each year. We then calculate the fraction of this sample of women who would be eligible for Medicaid in each state. By using the same group of women in each state simulation, we obtain an estimate of the fraction eligible that depends only on the legislative environment and is independent of other characteristics of states. This measure can be thought of as a convenient parameterization of legislative differences affecting women in different states and years: the generosity of state Medicaid policy can be naturally summarized in terms of the effect it would have on a given, nationally representative, population. Furthermore, we reduce the sampling variability in our estimates that derives from having relatively small cells for some states in the CPS.[9]

This paper's approach vs. simulated instrument

- This is not the most efficient way to exploit this variation

- Remember our propensity score example: if we could just condition directly on w_i, then we would not worry about endogeneity
  - The solution, then, is to construct a propensity score and condition on that!

- Intuitively, "the eligibility rules for Medicaid were being randomly assigned"
  - In other words, we assert a counterfactual distribution over the policy rules $Pr(g)$
  - This allows us to construct the propensity score for a given individual

$$p(w_i) = Pr(x_i|w_i) = \sum_g f_i(g, w_i) Pr(g)$$

- With pscore in hand, estimation is straightforward, and known to be semiparametrically efficient!

Recentering vs. Controlling?

- The paper makes a big point about the recentering concept
    - I'm not sure why one would do this vs. just controlling
- This is particularly true in this setting (rather than with interference)
- My suggestion when reading the paper: just think of this as conditioning on propensity scores
    - Can revisit if you are worried about interference

The return on propensity scores in an empirical example

- Medicaid Empirical example in this paper: ACA medicaid expansion

- ACA expanded Medicaid in only some states thanks to NFIB v. Seblius allowing choice by states

- Interested in understanding compositional shifts in health care across states
  - Use ACS micro data and consider structural equation

$$y_{it} = x_{it}\beta + \alpha_{s(i)} + \alpha_t + \epsilon_{it}$$

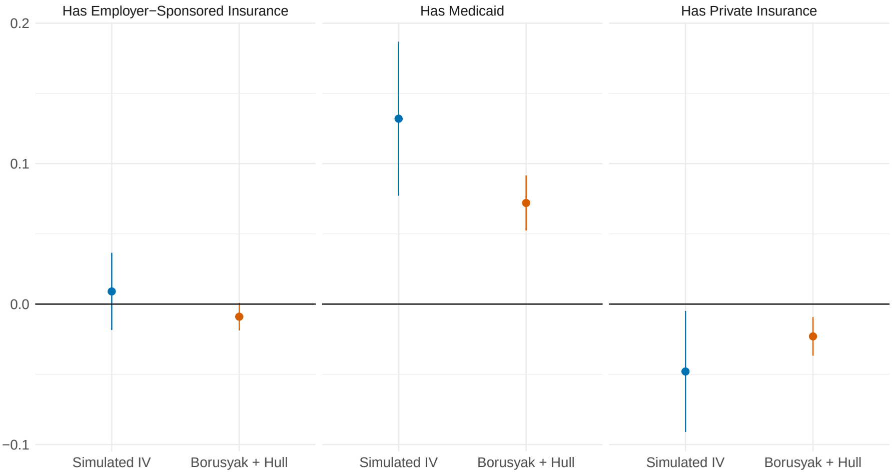
  - y_{it} are different health insurance take-up; x_{it} is Medicaid eligibility for individual i

- "Simulated" IV: dummy for whether state expanded Medicaid z_{sim}

- Borusyak and Hull IV: construct a person-level indicator for whether a person is eligible under their state's law z_{bh}
  - Also identify the $p(w)$ that they are eligible on average across others states' laws
  - They recenter ($\tilde{z}_{bh} = z_{bh} - p(w)$) but you could just control...

The return on propensity scores in an empirical example

- **Much** more precise
- Makes sense!
- Seems like we should use it...



Discussion: When beautiful theory meets data roadblocks

- Many times, the data necessary to calculate eligibility and the outcomes of interest are not in the same dataset. This is especially true in simulated IV settings (I know from experience)
  - Appendix D.2 talks about these issues but it's a little vague
- Clarify ideas: Cohodes et al. (2016) consider effects of Medicaid in the 1980s on long-term education outcomes
  - Parental income (the relevant w_i for eligbility during childhood) is not known in the same dataset for y_i.
  - x_i is average eligibility for types born in state s at age a of race r in period t.
  - They just want the variation that comes from state laws, not the demographics
  - Construct Sim. IV that takes the average share of individuals in national population that would be eligible under states' laws
- Hopefully you're getting it now! Proposal from paper:
  - Construct $Pr(x_i|\tilde{w})$ that randomizes over *states* and control (or recenter)
- Note: this type of within-state demographic info is actually used in Mahoney (2015) for bankruptcy simulated instruments!

Discussion: Heterogeneity in the underlying data

- Something harder to take from the paper
  - How to consider aggregation issues

- State-level variation, but maybe some individuals experienced Medicaid expansion, while other experienced contraction
  - Less likely in Medicaid case, but possible in other settings

- E.g. monotonicity violation when there is heterogeneous treatment
  - Paper discusses some points on this in Appendix, but would be useful to be a bit more concrete in some examples
  - Something I tried to work on with Aronow and Sorkin but ran into serious data issues!

Second kernel of the paper: interference

- Medicaid example is simple to think about, and clarifies idea that:
  1. Can convert high-dimensional variation into simple treatment effects
  2. Can be more *efficient* (e.g. smaller s.e.)

- However, you can take this much further.

- Consider the design of a railroad. Imagine the world in which a railroad designer randomly threw darts on a map to decide where to construct train lines
  - Similar to the analogy of "drawing" the Medicaid eligibility rules
  - But now, how do we think about the "random" piece interacting with different places?
  - Let's start with something simple first

Interference in network settings

- Consider a setting where the researcher want to measure the impact of a randomized experiment on a network
  - In other words, for a given person i, and observed network W, we randomly treat some subset of individuals on the network.
  - We want to know what the effect of having more treated individuals connected to you x_i is on y_i
- Insight from paper: since the position in *network* affects probability of being connected to individuals, some individuals will inherently get more exposure!
  - Analogous to the friendship paradox
- Need to construct an analogous propensity score for the network setting, and control for that
  - Since we have a true RCT, this is not too hard!
  - (But we do have to make decisions about what the spillover is)
  - Aronow and Samii (2017) made serious progress on the network context

Interference in spatial settings

- Things can be more confusing than an RCT, but this same insight applies
  - Even with random shocks (darts on a board), some locations / people attract more treatment than others
  - Consider the application from the paper

- Estimate the impact of market access growth (*MA*) on land values growth (*V*) in China
  - *MA* is influenced by transportation networks, and measures aggregated access to other populations

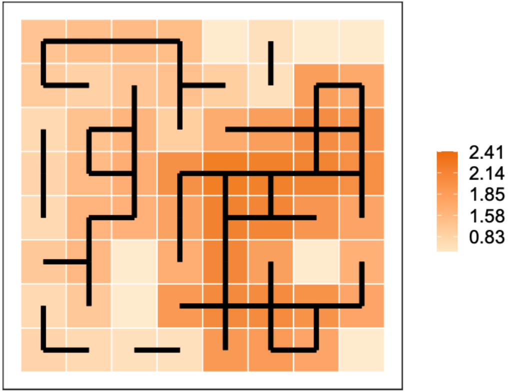
$$MA_{it} = \sum_j \tau(\mathbf{g}_t, loc_i, loc_j)^{-1} pop_j$$

- Want to estimate the effect of MA_{it} using "random" variation in network changes!
  - Can we just run the OLS? No!

Stylized Example of Market Access on a Square Island

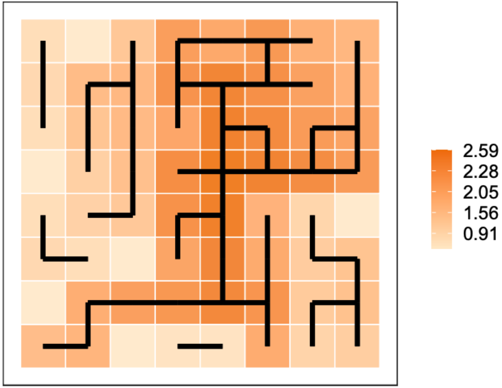
- Take a square with square villages and randomly assign roads

- How does market access change?



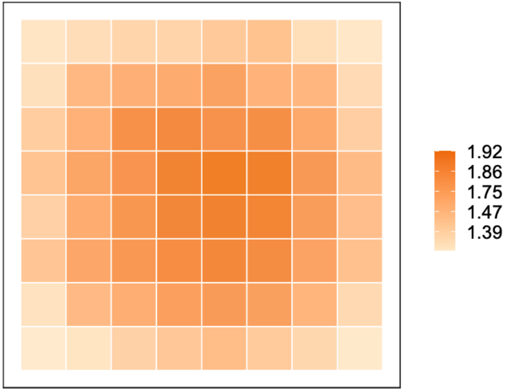
Stylized Example of Market Access on a Square Island

- Take a square with square villages and randomly assign roads
- How does market access change?
- If we rerandomize, does it look different?



Stylized Example of Market Access on a Square Island

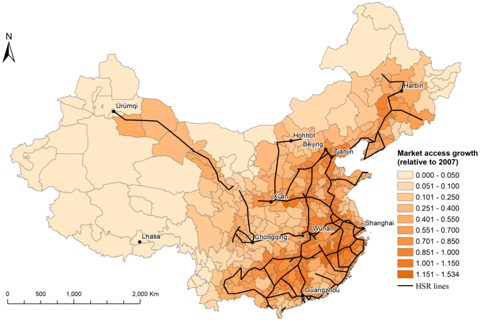
- Take a square with square villages and randomly assign roads
- How does market access change?
- If we rerandomize, does it look different?
- As with the network, some places get more market access than others on average!
- Need to account for this propensity difference



China: defining the counterfactual distribution

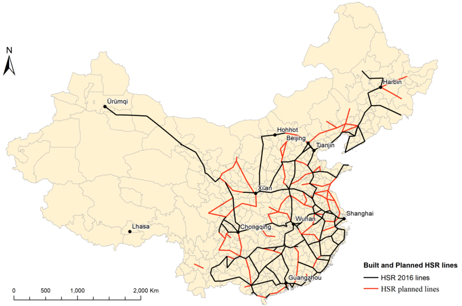
- In the stylized example, lines are laid randomly, making it easy to define the propensity scores
  - What about in China?

- What is the plausible counterfactual?



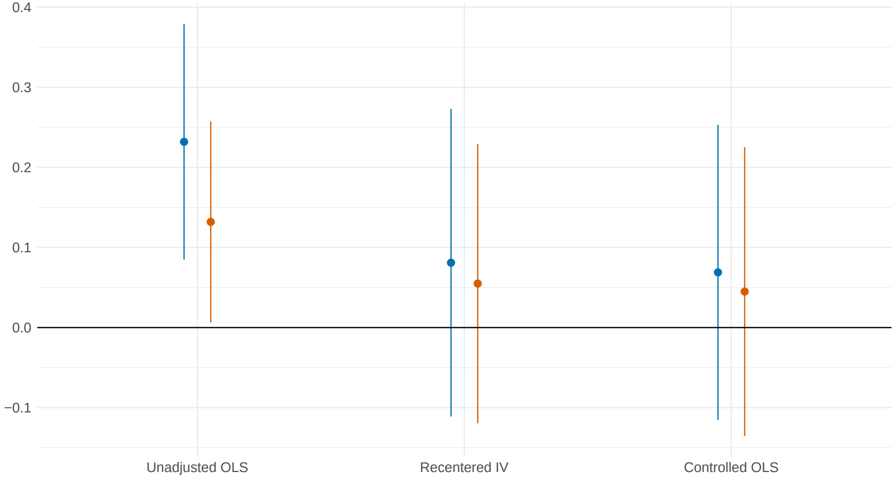
China: defining the counterfactual distribution

- In the stylized example, lines are laid randomly, making it easy to define the propensity scores
  - What about in China?
- What is the plausible counterfactual?
- Paper proposes an idea, and analagous to other examples
  - Use *planned* lines are randomized between unbuilt but planned, and built lines
  - Calculate distribution of propensity score by constructing MA_i under each counterfactual scenario



China railroads: the punchline

- There was substantial bias from using OLS!
- Makes sense – geography is king...
- No effect in randomized setting



Defining the counterfactual distribution

- If one takes issue with the counterfactuals, that is reasonable (but of course, challenging to prove one way or the other)
- Key issue: this paper is just making **text** what was already **subtext**
  - There was always an assumption about some counterfactual comparison in these designs!
- The issue is that many of these paper do not understand how to describe the randomization aspect of their research design
  - Consequentially, they cannot describe the "as-if random" component coherently
  - If a researcher has an alternative proposal, they should try that and see what estimates are available!
- Also suggests that reserchers can show a "range" of estimates under different scenarios

Key takeaways from paper

- Provide a toolbox for contexts when economists have found good “as-if” random variation (and can describe the counterfactual distribution)
- Show that in cases where treatment is not influenced by others’ treatment status, approach maps very tightly with traditional propensity methods, and can be much more efficient
- In spatial and network cases where treatment spillovers exist, show how to adjust for bias arising from units location on network or graph (or relevant characteristic)

Caveats

- We focused on Currie and Gruber (1996a, b) cases of simulated instruments, but there are other cases of “simulated instruments”
  - Gruber and Saez is an individual-level instrument so the monotonicity point doesn’t apply
  - Gruber and Saez is all about characteristics (income) responding endogenously
- These are tax elasticities and not clear that exclusion restriction holds
- This tax literature is strongly tied to functional form or additional assumptions (see Blomquist, Newey, Kumar and Liang (2021))