Shift-Share IVs (Shortened from Goldsmith-Pinkman and Peter Hull's Lecture Notes online)

April 15, 2025

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:

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

Bartik Instrument:

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

Other instruments have this structure

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

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

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

More general econometric set-up

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

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

- 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. Borusyak, Hull and Jaravel (2022) 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

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

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

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

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

$\Omega = (\mathbf{GG}')$, 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?

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 diff-in-diff, but may more naturally work in your setting.

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:

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

Interpretation: sensitivity to misspecification

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

- $\sqrt{L}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \tilde{\beta}, \qquad \mathbb{E}\left[\tilde{\beta}\right] =$
  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 over-identification

- 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

Borusyak, Hull, and Javarel, 2022

$$\hat{\beta}_{Bartik} = \frac{\sum_{l=1}^{L}\sum_{k=1}^{K} z_{lk} g_k y_l^{\perp}}{\sum_{l=1}^{L}\sum_{k=1}^{K} z_{lk} g_k x_l^{\perp}} = \frac{\sum_{k=1}^{K} g_k \sum_{l=1}^{L} z_{lk} y_l^{\perp}}{\sum_{k=1}^{K} g_k \sum_{l=1}^{L} z_{lk} x_l^{\perp}} \equiv \frac{\sum_k z_k g_k \bar{y}_k^{\perp}}{\sum_k z_k g_k \bar{x}_k^{\perp}}$$

where we have

- $\bar{y}_k^{\perp} = \frac{\sum_l z_{lk} y_l^{\perp}}{\sum_l z_{lk}}, \bar{x}_k^{\perp} = \frac{\sum_l z_{lk} x_l^{\perp}}{\sum_l z_{lk}}$ are the "exposure-weighted" average of local variables.
- $z_k = \frac{1}{L}\sum_l z_{lk}$ are weights capturing the importance of industry/categ. k.

This casts the estimator as an IV regression instrumented by g_k (weighted by z_k) to estimate

$$\bar{y}_k^{\perp} = \alpha + \beta \bar{x}_k^{\perp} + \bar{\epsilon}_k \quad \bar{\epsilon}_k = \frac{\sum_l z_{lk}\epsilon_l}{\sum_l z_{lk}}$$

Note that this is now a regression at the industry/category level, so it naturally relies on a different set of assumptions relative to GPSS.

Borusyak, Hull, and Javarel, Assumption 1

- We first need quasi-random shock assignment: $E[g_k|\bar{\epsilon}_k, z_k] = \mu$ for all k.
- Consider ADH: this needs the expected growth of Chinese imports for each sector g_k is the same for high vs low $\bar{\epsilon}_k$ (average unobserved determinants of regional employment growth specializing in sector k).
- We can show that

$$E[\sum_k z_k g_k \bar{\epsilon}_k] = E[\sum_k z_k E[g_k|\bar{\epsilon}_k, z_k]\bar{\epsilon}_k] = \mu E[\sum_k z_k \bar{\epsilon}_k] = 0$$

- The final piece of the equation is zero since $\sum_k z_{lk} = 1$

$$E[\sum_k z_k \bar{\epsilon}_k] = E[\frac{1}{L}\sum_l \sum_k z_{lk}\epsilon_l] = E[\frac{1}{L}\sum_l \epsilon_l] = 0$$

Borusyak, Hull, and Javarel, Assumption 2

- Given the identification argument, we next need an appropriate law of large numbers (at the shock level).
- The key assumption is that assumption there are many uncorrelated shocks such that
  - $E[\sum_k z_k^2] \to 0$ as $L \to \infty$. In other words, we will need z_k vanishing small when the sample size grows. (Number of shocks grows with the sample).
  - $Cov(g_k, g_{k'}|\bar{\epsilon}, z) = 0, \forall k' \neq k$. The shocks are mutually uncorrelated.
- We can then show that the sample analog

$$\sum_k z_k g_k \bar{\epsilon_k} \to_p \underbrace{E[\sum_k z_k g_k \bar{\epsilon}_k] = 0}_{\text{Assumption 1}}$$