Design-based Identification with Formula Instruments

K. Borusyak, P. Hull and X. Jaravel

Carlos Gonzalez

Topics in Econometrics

November 4, 2024

Outline

Outline

SSIV Brush Up: Two world views

- **Example:** We are interested in finding the effect of internet usage during class-time x_i on Oxford's students grades y_i

$$y_i = \beta x_i + \varepsilon_i \tag{1}$$

- However, we can easily run into OVB and endogeneity problems...

- **A possible solution?** Students i go to different colleges k to take class, such that i spends s_{ik} share of time a week in college k. The university is quite old so colleges are prone to power outages. Let g_k be the amount of time college k is under a power outage a week, hence Wifi is not available

SSIV Brush Up: Two world views

- If rooms for different courses were randomly allocated across different colleges, then, it is plausible that $\mathbb{E}[\varepsilon_i \mid s_{ik}] = 0$ for all k. The **shares are exogenous**

- Consider the instrument $z_i = \sum_k s_{ik} \cdot g_k$. It is immediate that $\mathbb{E}[z_i \cdot \varepsilon_i] = 0$. This is the classic **share interpretation** of SSIV [Goldsmith-Pinkham et al., 2020]

- Unfortunately, shares s_{ik} are rarely exogenous. Students enroll into particular colleges according to their interests, field, etc.

SSIV Brush Up: Two world views

- Can we leverage exogeneity of shocks g_k to build valid z_i even with endogenous s_{ik}? Yes! **Shock/shift** interpretation [Borusyak et al., 2022], [Borusyak and Hull, 2023], [Borusyak et al., 2024]

- **Key takeaways from paper**
  - Review of the **shift interpretation** in SSIV
  - SSIV is in fact a very especial case of a broader set of models: **Formula models** (replace $\sum_k s_{ik} \cdot g_k$ by $f_i(s, g)$)
  - Assumptions needed for identification, consistency and inference in formula IV
  - Exogeneity of g_k is usually **not enough** for identification. We need to exploit: (i) the structure of the formula, and (ii) the **design** of the problem (the DGP/allocation rule of g_k)

UNIVERSITY OF OXFORD

Outline

Review on Notation

- Lower-level i (students), upper (shock) aggregation level k (colleges)
- Shares s_{ik}, shocks g_k, instruments $z_i = f_i(s, g)$
- Model $y_i = \beta x_i + \varepsilon_i$, $w = (s, q)$ with q some other observable covariates

Outline

Shock Exogeneity

- Two assumptions: **Shock Exogeneity** and **Known Shock Design**
- Shocks are assumed to be exogenous conditional on $w = (s, q)$, $\varepsilon \perp\!\!\!\perp g \mid w$ (power outages are independent of unobserved student heterogeneity, conditional on s_{ik}). Violated if power outages are more likely in the morning when the better students go to class.
- Additional comments
  - $x_i = z_i \implies f_i(\cdot)$ must be correct, otherwise relevance is enough
  - Why this instrument shape? Usually motivated because $x_i = f(\sum_k s_{ik}\tilde{x}_{ik})$
  - No iid requirement on x_i, y_i

Known Design: $G(g \mid w)$

- $G(g \mid w)$ is known (!!)
- Reasonable in RCT contexts where the shock generation and allocation is captured in the protocol
- Quite unreasonable everywhere else... Shock exchangability might help
  - All colleges built in the same century ($k \in C_c$) are iid
  - Similar to (local) exchangability conditions in RD or DiD
  - Still, quite unreasonable in many contexts. DGP of supply-chain shocks? One-time shocks?
- In our example, Electrical Engineering pals figured out G (instead of actually fixing the outages, a very Oxford thing to do)

UNIVERSITY OF OXFORD

Discussion on Known Design: Too much structure?

- Although IV approach, this is getting very structural...
- Instrument shape $f_i(s, g)$ is grounded on the structure of x_i. And this is about to get worse in a second
- Known design implies some deeper knowledge of the generation, allocation and propagation of shocks across observations. Very unlikely in observational designs

A Lighter Assumption

- When $f_i(s, g)$ is linear in g (as it is the case in SSIV) a joint weaker assumption suffices

- Conditionally linear shock means $\mathbb{E}[g_k \mid \varepsilon, w] = q_k' \cdot \theta$ (for some unknown θ)
  - It relaxes independence by mean independence $\mathbb{E}[g_k \mid \varepsilon, w] = \mathbb{E}[g_k \mid w]$
  - It replaces known distribution by a parametric assumption $\mathbb{E}[g_k \mid w] = q_k' \cdot \theta$
  - q_k will usually take the form of fixed effects at some aggregation level between i and k (like college century fixed effects)

Outline

Shock exogeneity is not sufficient for identification

- As in every IV setting we need exogeneity $\mathbb{E}[z_i \cdot \varepsilon_i] = 0$ and relevance $\mathbb{E}[z_i \cdot x_i]$
  - We focus on establishing exogeneity (but we will see that relevance can become an issue later on)
  - Let's see how the assumptions above yield identification
- Key insight in [Borusyak and Hull, 2023] **shock exogeneity is not enough** for formula instrument independence
  - Trivial example. Say that $\sum_k s_{ik} = S_i < 1$. Then students with higher S_i will be exposed to more shocks. S_i is likely correlated with students' exam performance

General Recipe: Re-center or control

- Further insight in [Borusyak and Hull, 2023]. Let **the expected instrument** $\mu_i = \mathbb{E}[f_i(s,g) \mid w]$, then, under conditional shock independence

$$\mathbb{E}\Big[\frac{1}{N}\sum_i z_i \cdot \varepsilon_i\Big] = \mathbb{E}\Big[\frac{1}{N}\sum_i \mu_i \cdot \varepsilon_i\Big] \tag{2}$$

- One line proof: $\mathbb{E}[z_i \cdot \varepsilon_i] = \mathbb{E}[\mathbb{E}[f_i(g,w) \cdot \varepsilon_i \mid w]] = \mathbb{E}[\mu_i \cdot \mathbb{E}[\varepsilon_i \mid w]] = \mathbb{E}[\mu_i \cdot \varepsilon_i]$

- In words: The instrument is exogenous iff the expected instrument is exogenous. Implication: **Substracting or controlling by the expected instrument** yields an exogenous instrument

- Known design is needed to recover μ_i

Implementing Re-centering and control

- Re-centering: Define $\tilde{z}_i = z_i - \mu_i$. Immediately $\mathbb{E}[\tilde{z}_i \cdot \varepsilon_i] = 0$. Note that μ_i is available because $G(\cdot)$ is known!
- Controlling (FWL): Residualize x_i and y_i on μ_i (or a vector $r(w)$ which linearly spans μ_i). Then run IV of z_i on $x_i^{\perp}$, $y_i^{\perp}$. Works because $\mathbb{E}[z_i \mid \varepsilon_i^{\perp}] = 0$
- How to obtain μ_i? Draw counterfactual g_k^j and recalculate $z_i^j = f_i(s, g^j)$. $\mu_i = \frac{1}{J} \sum z_i^j$. Under exchangability compute such average by replacing the shocks g_k by g_k^j with $j \in C_c$
- Both of these methods should be equivalent. Controlling is probably safer, because we can control by several candidates of μ_i (as long as one is correct, we should be good)

Applications

- We will now apply these conditions and general logic to different situations (SSIV being the most prominent)
- When SSIV, we can use the weaker condition. For non-linear SSIV or non-anonymous functions we need the strong assumptions

Linear SSIV: Three cases

- **Complete shares** $\sum_k s_{ik} = 1$ and **no controls** $q_k = 1$, so $\mu_i = \theta$ for all k. Just residualize x_i and y_i on a constant. Instrument independence is equivalent to independence at the (weighted) shock-level. Example: All colleges have the same expected outage time

- **Complete shares with controls**. $\mu_i = Q_i \cdot \theta$, $Q_i = \sum_k s_{ik} q_k$. Useful decomposition $z_i = Q_i\theta + \sum_k s_{ik}(g_k - q_k \cdot \theta)$ (systematic and possibly cofounding variation and idiosyncratic variation). Often times q_k take the form of fixed effects. Control by Q_i

- **Incomplete shares**. Without controls $\mu_i = S_i\theta$. Control by S_i. More generally control by Q_i where Q_i is not a weighted average of q_k provided $S_i < 1$.

Non-linear functions and non-linear SSIV

- **Mean independence is not enough.** μ_i depends on s_i in complicated ways (hence full independence is needed)

- Example. Say that the relevant variable is the log of the time a student spends on the internet, i.e. $z_i = \log(\sum_k s_{ik} \cdot g_k)$

- **Recipe still holds:** Generate μ_i from distribution/exchangability. Alternatively, linearize $f(\cdot)$ and consider the previous case

Networks and more

- Networks are a particularly interesting design. Say that electrical systems at different colleges are connected following an adjacency matrix

$$\begin{bmatrix} s_{11} & s_{12} & \dots & s_{1k} \\ \vdots & \ddots & \ddots & \vdots \\ s_{n1} & s_{n2} & \dots & s_{nk} \end{bmatrix} = \begin{bmatrix} 1 & 1/2 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & \dots & 1 \end{bmatrix}$$

- Controlling and recentering logic remains valid
- Frank's paper (non-random exposure to random shocks because of endogenous matching)

Outline

Consistency

- Main problem. When controlling/recentering *iid* is out of the window, as the expected instrument factors information from other shares and covariates.

- Conditions for consistency in SSIV case
  - **Many uncorrelated shocks** $\mathbb{E}[\sum_k (\sum_i s_{ik})^2] \to 0$ AND $Cov[g_k, g_{k'} \mid \varepsilon, w] = 0$ for all $k \neq k'$. Number of shocks (available colleges) must increase as i increases, while these shocks remaining conditionally independent (split colleges into buildings?)
  - Relevance $\frac{1}{N} \sum_i x_i \tilde{z}_i \to p \neq 0$
  - Regularity: Bounded variance of g_k and $\bar{\varepsilon}_k^2$

More on Consistency

- Overall logic: We need that an LLN goes over $\frac{1}{N}\sum_i \varepsilon_i \tilde{z}_i$ despite the mutual correlation between ε_i

- Nuance: First condition requires $\mathbb{E}[\sum_k (\sum_i s_{ik})^2] \to 0$ BUT a sufficient condition for second condition is $\mathbb{E}[\sum_i (\sum_k s_{ik})^2] > 0$. Implication: We need observations to be **effectively exposed to a small number of shocks, but different observations being exposed to different shocks!**

- Discussion: When is his reasonable and when not?

Inference

- Same problems as with consistency, *iid* is no no
- Solutions? Asymptotic approximations of variance [Adao et al., 2019]
- Transform regression into “shock level”. Robust SE in the transformed model have correct asymptotic coverage [Borusyak and Hull, 2023]
- Randomization inference (finite sample guarantees and applicable with few shocks). Simulate z_i^j under protocol or exchangability and invert to recover guarantees on $\hat{\beta}$

Outline

Conclusion and Discussion

- Unified framework for formula-based IVs under shock interpretation
- Shock exogeneity is not enough. Known design is needed. Formula structure can weaken the “design-knowledge”
- Hybrid status between reduced-form and structural econometrics. What design are you talking about in an observational study? Formula structures, known shock DGP, valid exchangability...
- Control by the expected instrument. Control by the expected instrument. Control by the expected instrument
- Consistency and inference are a bit more tricky. Many uncorrelated shocks

Bibliography I

Adao, R., Kolesár, M., and Morales, E. (2019).
Shift-share designs: Theory and inference.
The Quarterly Journal of Economics, 134(4):1949–2010.

Borusyak, K. and Hull, P. (2023).
Nonrandom exposure to exogenous shocks.
Econometrica, 91(6):2155–2185.

Borusyak, K., Hull, P., and Jaravel, X. (2022).
Quasi-experimental shift-share research designs.
The Review of economic studies, 89(1):181–213.

Borusyak, K., Hull, P., and Jaravel, X. (2024).
Design-based identification with formula instruments: A review.
The Econometrics Journal, page utae003.

Bibliography II

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020).
Bartik instruments: What, when, why, and how.
American Economic Review, 110(8):2586–2624.