Part C: Instrumental Variables

C3: Judge IV Design

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2025

C3 Outline

Readings:

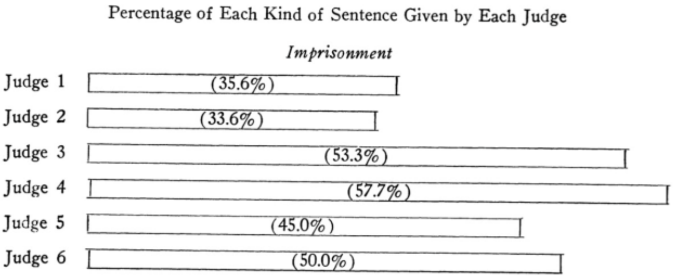
- Scott Cunnigham's "Mixtape" textbook, Ch. 7.8.2
- Goldsmith-Pinkham, Hull, Kolesar (in progress)

Setting

- There are many situations where:
  1. the treatment (usually binary) is decided by one of K "judges" (examiners, caseworkers, ...)
  2. judge's decision is discretionary
  3. judges are assigned to cases randomly (perhaps within strata, e.g. location-period)
- Examples of treatments:
  - Incarceration *(Kling 2006, Mueller-Smith 2015)*, bail *(Arnold, Dobbie, Yang 2018)*
  - Patent granting *(Sampat and Williams 2019)*
  - Credit ratings *(Rieber and Schechinger 2019)*
  - Psychotherapy treatment *(Blæhr and Søgaard 2021)*
  - Several types of job training programs for the unemployed *(Humlum, Munch, Rasmussen 2024)*

Idea

- Judges typically vary by leniency



(Gaudet, Harris and John 1933, reproduced from Scott Cunningham's "Mixtape")

- Can use this heterogeneity to instrument for treatment
  - Leniency is unobserved $\implies$ what should we do?
  - Under which assumptions is the answer causal?

Estimated leniency?

- Notation: judge assignment $Q_i = k \in \{1, \ldots, K\}$; $Z_{ki} = \mathbf{1}\left[Q_i = k\right]$
- Consider binary treatment. Popular naïve idea:
  - Measure judge leniency as % of lenient decisions: $\hat{L}_k = \frac{\sum_i Z_{ki} D_i}{\sum_i Z_{ki}}$
  - Then instrument D_i with $\hat{L}_{Q_i}$
  - If assignment is random only within strata, control for strata FE
- Problem: $\hat{L}_k$ is noisy if there are many judges (and not so many cases per judge)
  - $\hat{L}_{Q_i}$ is influenced by D_i, which correlates with $\varepsilon_i \implies$ bias
  - Conventional inference doesn't take into account estimation noise

Correct IV framework

- Note that $\hat{L}_k =$ OLS estimates of L_k from a first-stage

$$D_i = \sum_k L_k Z_{ik} + u_i$$

$\Rightarrow$ Using fitted values $\hat{L}_{Q_i}$ as IV $\Longleftrightarrow$ 2SLS with $Z_1, \ldots, Z_K$ instruments

- Problem of noisy $\hat{L}_k$ is the standard many weak IV problem
  - Without covariates, JIVE is natural: uses leave-out leniency $\hat{L}_i = \frac{\sum_{j\neq i} \mathbf{1}[Q_j = Q_i] D_j}{\sum_{j \neq i} \mathbf{1}[Q_j = Q_i]}$
  - With (many) covariates (e.g., strata FEs with few cases per strata), better to use UJIVE ("unbiased JIVE"; Kolesar (2013)): a version of JIVE that is consistent with many controls

Underlying assumptions

With judge assignment dummies $Z_{i1}, \ldots, Z_{iK}$ as IVs, what about:

- Independence?
- Exclusion?
- Monotonicity?

Underlying assumptions

With judge assignment dummies $Z_{i1}, \dots, Z_{iK}$ as IVs, what about:

- Independence?
  - Guaranteed by random assignment, as long as strata FEs are controlled for
- Exclusion?
  - Does the judge directly make only one decision D_i?
  - Can the judge indirectly affect others treatments, e.g. by affecting who will be making those decisions?
- Joint monotonicity: very strong (as usual with multiple IVs)
  - A judge who is more lenient on average should be weakly more lenient on everyone
  - Violated if judges put different weights on different characteristics of the case

Tests for monotonicity and exclusion

1. Reject joint monotonicity if the ranking of judges by leniency varies by subgroup of cases based on observables (see Dobbie, Goldin, and Yang (2018))
2. Frandsen, Lefgren, and Leslie (2023):
    - Under all LATE assumptions, comparing any two judges gives causal effects for some complier population
    - Causal effects cannot be too large: e.g. bounded by the range of possible outcomes
    - Reject exclusion or monotonicity if there is a pair of judges with similar $\mathbb{E}\left[D_i \mid Q_i\right]$ but very different $\mathbb{E}\left[Y_i \mid Q_i\right]$ (like Sarsons (2015) and Kitagawa (2015))

- Should we panic if monotonicity doesn't hold?
    - Not with homogeneous effects (or in De Chaisemartin (2017) "Tolerating defiance")
    - Frandsen et al. (2023): 2SLS identifies a convex average of causal effects under "average monotonicity": for all i, $D_i(k)$ is positively correlated with L_k across k

References I

De Chaisemartin, C. (2017): "Tolerating defiance? Local average treatment effects without monotonicity: Tolerating defiance?" *Quantitative Economics*, 8, 367–396.

Dobbie, W., J. Goldin, and C. S. Yang (2018): "The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 108, 201–240.

Frandsen, B., L. Lefgren, and E. Leslie (2023): "Judging Judge Fixed Effects," *American Economic Review*, 113, 253–277.

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063.

Kolesar, M. (2013): "Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity," 1–45.

Sarsons, H. (2015): "Rainfall and Conflict: A Cautionary Tale," *Journal of Development Economics*, 115, 62–72.