Sargan Lecture 2
Weak Identification with Many Instruments

Anna Mikusheva

MIT

June, 2024

Introduction

- We consider IV models with *many weak* instruments
  - Estimation with many instruments
  - How to determine that instruments are weak?
  - Weak identification robust inferences
  - Open questions

Introduction

- **Example 1**: Angrist and Krueger (1991)

$$wage_i = \beta \; education_i + controls + e_i,$$

- Instrument is quarter of birth
- First stage is heterogeneous: law depends on state and birth cohort
- Instruments used: QOB ($\times$ state dummy) ($\times$ year dummy)
  - year of birth (30)
  - year and state of birth (180)
  - year and state of birth, and their interactions (1530)
- Staiger and Stock (1997)- IV may be weak
- Hansen et al. (2008)- instruments are many

Introduction

- IV regression often uses interactions between instruments and covariates. Why?
  - Extract more information - exclusion restriction is conditional
  - Search for optimal instrument
  - TSLS has LATE (causal) interpretation only if IV is fully saturated- Blandhol et al (2022)

Introduction

- **Example 2**: 'Judges design'
- Bhuller, Dahl, Loken and Mogstad (JPE, 2020): "Incarceration, Recidivism, and Employment"

$$\textit{recidivism}_i = \beta \;\; \textit{incarceration}_i + \textit{controls} + e_i,$$

- Instruments: "judge stringency" = the average incarceration rate in other cases a judge has handled
- This is a form of JIVE with instrument-dummies for judge assignment
- Sample size is roughly proportional to the number of judges
- Other known examples: Mendelian randomization as instruments, name-based estimators of inter-generational mobility

Setup

- Linear IV model with one endogenous variable:

$$\begin{cases} Y_i = \beta X_i + (\delta' W_i) + e_i \\ X_i = \pi' Z_i + (\gamma' W_i) + v_i \end{cases}$$

  where $Z_i \in \mathbb{R}^K$ s.t. $\mathbb{E}[e_i|Z_i, W_i] = \mathbb{E}[v_i|Z_i, W_i] = 0$
- Data is i.i.d., $i = 1, ..., N$
- Many instruments: $K \to \infty$ as $N \to \infty$ (up to $K = \lambda N$)
- Weak instruments: π is small in some sense
- For most results errors are heteroskedastic
- What we assume away- heterogeneous treatment effects

Outline

1. Estimation
2. Weak Identification: detection
3. Weak IV robust inferences
4. Adding covariates
5. Conclusions and Open questions

Overview

Setup

- Assume away covariates (we will add them in the last section)

$$\begin{cases} Y_i = \beta X_i + e_i \\ X_i = \pi' Z_i + v_i \end{cases}$$

  where $Z_i \in \mathbb{R}^K$ s.t. $\mathbb{E}[e_i|Z_i] = \mathbb{E}[v_i|Z_i] = 0$
- Data is i.i.d., $i = 1, ..., N$
- For most results errors are heteroskedastic

TSLS

Most commonly known estimator is Two-Stage Least Squares (TSLS)

- First stage: regress X_i on Z_i via OLS and find best linear predictor

$$\widehat{X}_i = \widehat{\pi} Z_i$$

- Second stage: regress Y_i on $\widehat{X}_i$ (exogenous part of X_i) via OLS

TSLS

Another interpretation of Two-Stage Least Squares (TSLS)

- First stage- finding the optimal instrument = best predictor

$$\widehat{X}_i = \widehat{\pi} Z_i$$

- Second stage: just identified IV regression of Y_i on X_i using $\widehat{X}_i$ as the instruments
- Optimal instrument under homoskedasticity: $\mathbb{E}[X_i|Z_i]$ (Chamberlain, 1987)
- Concentration parameter $\frac{\pi' Z' Z \pi}{\sigma_v^2}$ plays as effective sample size (Stock and Yogo, 2005)

Estimation with Many IV

- First stage: $X_i = \pi' Z_i + v_i$
- If many regressors in the first stage, they might 'overfit' the noise
- Estimated optimal instrument is endogenous $\mathbb{E}[\widehat{X}_i e_i] \neq 0$
- For homoscedastic TSLS: $\widehat{X}_i = \hat{\pi}' Z_i = \pi' Z_i + v' Z (Z'Z)^{-1} Z_i$

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \widehat{X}_i e_i\right] = \frac{K}{N}\sigma_{ev}$$

- Endogeneity is growing in K, leads to bias
- Bias of the IV estimator increases with the number of moment conditions/instruments (Bekker, 1994, Newey and Smith, 2004)

Estimation with Many IV

Suggestions on how to remove endogeneity:

- Sample splitting (Angrist and Krueger, 1995):
  - split sample to halves
  - select/estimate optimal instrument on one half
  - estimate β on the other half
- Jackknife (Angrist et al., 1999)
  - estimate optimal instrument for observation i on sample excluding i
  - use estimated optimal instrument

Estimation with Many IV

$$\widehat{\beta}_{TSLS} = \frac{X'P_Z Y}{X'P_Z X} = \frac{\sum_{i,j} X_i P_{ij} Y_j}{\sum_{i,j} X_i P_{ij} X_j}$$

- $\widehat{\beta}_{TSLS} - \beta = \frac{X'P_Z e}{X'P_Z X}$, where $P_Z = Z(Z'Z)^{-1}Z'$
- Bias comes from $\mathbb{E}[X'P_Z e] = \mathbb{E}[v'P_Z e] = \sum_i P_{ii}\mathbb{E}[v_i e_i]$ the diagonal of the projection matrix, $trace(P_Z) = K$
- Idea: remove the diagonal

Estimation with Many IV

$$\widehat{\beta}_{TSLS} = \frac{X' P_Z Y}{X' P_Z X} = \frac{\sum_{i,j} X_i P_{ij} Y_j}{\sum_{i,j} X_i P_{ij} X_j}$$

- Idea: remove the diagonal

$$\widehat{\beta}_{JIV} = \frac{\sum_{i \neq j} X_i P_{ij} Y_j}{\sum_{i \neq j} X_i P_{ij} X_j}$$

- It is very close to jackknife (numerical differences are tiny)
- Diagonal removal done to many estimators: JIVE-LIML and JIVE-Fuller (Hausman et al., 2012), JIVE-ridge (Hansen et al, 2014)

Estimation with Many IV

$$\begin{cases} Y_i = \beta X_i + e_i, \\ X_i = \pi' Z_i + v_i, \end{cases}$$

- TSLS is consistent when $\frac{\pi' Z' Z \pi}{K} \to \infty$ (Chao and Swanson, 2005)
- When $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$, JIVE, JIVE-Fuller and JIVE-LIML are consistent (Hausman et al, 2012)
- When $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$, JIVE, JIVE-Fuller and JIVE-LIML are asymptotically gaussian
    - Wald confidence sets and t-statistics can be used
    - Estimation of standard errors is non-trivial (Hausman et al, 2012)

Estimation with Many IV: Summary

- Many instruments can be hurtful if they do not extract additional information from the first stage
- Over-fitting creates a bias
- One should avoid using TSLS with many instruments
- Jack-knifing or diagonal removal is very fruitful idea

Other ideas in the literature

- Use Machine Learning for instrument selection on first stage
    - Information Criteria (Donald and Newey, 2001)
    - LASSO (Belloni et al, 2012)
    - Ridge (Carrasco, 2012)
    - Random forest, neural nets, etc.
- Pluses: If data satisfy assumptions of ML algorithm consistency $\Rightarrow$ asymptotic efficiency
- Minuses: we do not know what happens when ML is not consistent
- Angrist and Frandsen (2022): biases of ML first stage comparable to TSLS without gain in efficiency
- If you want to use ML on the first stage- DO SAMPLE-SPLITTING!

Overview

What is Weak Identification?

- If $\frac{\pi' Z' Z\pi}{\sqrt{K}} \to \infty$, then JIVE or JIVE-LIML are consistent and asymptotically gaussian
- What if there are better estimators (work well for weaker cases)?
  - Negative statement: in the best possible scenario – only π and β are unknown, if $\frac{\pi' Z' Z\pi}{\sqrt{K}} \asymp const$, there exists no asymptotically consistent robust test (Mikusheva and Sun, 2022)
- How to know in practice if $\frac{\pi' Z' Z\pi}{\sqrt{K}}$ is large enough to trust Wald confidence sets?

Weak Identification: detection

- Mikusheva and Sun (2022): pre-test for weak identification
- Our pre-test is based on the empirical measure:

$$\widetilde{F} = \frac{1}{\sqrt{K}\sqrt{\widehat{\Upsilon}}} \sum_{i=1}^{N} \sum_{j \neq i} P_{ij} X_i X_j,$$

  here $\widehat{\Upsilon}$ is an estimate of uncertainty in the first stage
- If $\widetilde{F} > 4.14$, then the JIVE- Wald test has less than 10 % size distortion
- Suggestion: if $\widetilde{F}$ is low, one should use “robust” tests
- Stata package implementing pre-test and robust tests: `manyweakiv` (beta version)

Re-visiting Angrist and Krueger (1991)

- Research question: returns to education. Y_i is the log weekly wage, X_i is education
- Instruments: quarter of birth. Justification is related to compulsory education laws:
  - 180 instruments: 30 quarter and year of birth interactions (QOB-YOB) and 150 quarter and state of birth interactions (QOB-POB)
  - 1530 instruments: full interactions among QOB-YOB-POB
- The sample contains 329,509 men born 1930-39 from the 1980 census
- This paper sparked the weak IV literature. It is a running example for multiple papers

Re-visiting Angrist and Krueger (1991)

	FF	$\widetilde{F}$	JIVE-Wald	Robust AR	Robust LM
180 instruments	2.4	13.4	[0.066,0.132]	[0.008,0.201]	[0.067,0.135]
1530 instruments	1.3	6.2	[0.024,0.121]	[-0.047, 0.202]	[0.022,0.127]

Table: Angrist and Krueger (1991) Pre-test Results

Notes: Results on pre-tests for weak identification and confidence sets for IV specification underlying Table VII Column (6) of Angrist and Krueger (1991). The confidence sets are constructed via analytical test inversion.

Overview

Weak IV-Robust Tests: Refresher, Fixed K

- $Y_i = \beta X_i + e_i$, Z_i-instrument ($\mathbb{E}[e_i|Z_i] = 0$)
- $H_0 : \beta = \beta_0$. Define $e(\beta_0) = Y - \beta_0 X$
- AR (Anderson-Rubin) statistics:

$$e(\beta_0)' Z \Sigma^{-1} Z' e(\beta_0) \sim \chi^2_K$$

  Σ is a covariance matrix of $e'Z$ or a good estimate of it
- Size is robust to weak IV

What Changes with $K \to \infty$?

- Homoskedastic AR statistics for fixed K:

$$\frac{1}{\sigma^2} e(\beta_0)' Z(Z'Z)^{-1} Z' e(\beta_0) \sim \chi^2_K$$

- χ^2_K is a diverging distribution for large K
- $e(\beta_0)' P_Z e(\beta_0)$ has a non-zero mean $\mathbb{E} e' P_Z e = \sum_{i=1}^{N} P_{ii} \mathbb{E} e_i^2$
- Idea: remove the diagonal $\sum_{i \neq j} e_i(\beta_0) P_{ij} e_j(\beta_0)$
- Use CLT for quadratic forms (U-statistics)

AR test with many instruments

- The infeasible leave-one-out AR is

$$AR_0(\beta_0) = \frac{1}{\sqrt{K\Phi_0}} \sum_{i \neq j} e_i(\beta_0) P_{ij} e_j(\beta_0),$$

  for $\Phi_0 = \frac{2}{K} \sum_{i \neq j} P_{ij}^2 \sigma_i^2 \sigma_j^2$
- Under $H_0 : \beta = \beta_0$ we have $AR_0(\beta_0) \Rightarrow N(0, 1)$
- Need $K \to \infty$ for asymptotic distribution
- Rejects for large values of AR
- Mikusheva and Sun(2023) for estimate the variance

Weak IV-Robust Tests: LM

- Problem: AR is not efficient if identification is strong
- AR uses all instruments "equally"
- LM intends to test a "powerful" combination of instruments $e'Z\pi$,
- Idealistic LM is based on the linear combination $e'(\beta_0)Z\widehat{\pi} = e'(\beta_0)P_Z X$
- Leave-one-out gives us $LM^{1/2} \propto \sum_{i \neq j} e_i(\beta_0)P_{ij}X_j$

Robust LM

- The infeasible leave-one-out LM is

$$LM^{1/2}(\beta_0) = \frac{1}{\sqrt{K\Psi}} \sum_{i \neq j} e_i(\beta_0) P_{ij} X_j,$$

- Under $H_0 : \beta = \beta_0$ we have $LM^{1/2}(\beta_0) \Rightarrow N(0,1)$ as $N, K \to \infty$
- Reject when $\left| LM^{1/2}(\beta_0) \right|$ is large (two-sided test)
- Mikusheva and Sun (2024) suggest how to estimate variance

Re-visiting Angrist and Krueger (1991)

	FF	$\widetilde{F}$	JIVE-Wald	Robust AR	Robust LM
180 instruments	2.4	13.4	[0.066,0.132]	[0.008,0.201]	[0.067,0.135]
1530 instruments	1.3	6.2	[0.024,0.121]	[-0.047, 0.202]	[0.022,0.127]

Table: Angrist and Krueger (1991) Pre-test Results

Notes: Results on pre-tests for weak identification and confidence sets for IV specification underlying Table VII Column (6) of Angrist and Krueger (1991). The confidence sets are constructed via analytical test inversion.

Power Trade-off

- Under the alternative $\beta = \beta_0 + \Delta$, we have :

$$LM^{1/2} \Rightarrow \Delta \frac{\mu^2}{\sqrt{K\Psi}} + \mathcal{N}(0,1),$$

$$AR \Rightarrow \Delta^2 \frac{\mu^2}{\sqrt{K\Phi}} + \mathcal{N}(0,1)$$

- $\mu^2 \approx \pi' Z' Z \pi$
- When $\frac{\mu^2}{\sqrt{K}} \to \infty$, AR and LM are asymptotically consistent for fixed alternatives β

Power Trade-off

- Under the alternative $\beta = \beta_0 + \Delta$, we have :

$$LM^{1/2} \Rightarrow \Delta \frac{\mu^2}{\sqrt{K\Psi}} + \mathcal{N}(0,1),$$

$$AR \Rightarrow \Delta^2 \frac{\mu^2}{\sqrt{K\Phi}} + \mathcal{N}(0,1)$$

- When $\frac{\mu^2}{\sqrt{K}} \to \infty$ but $\frac{\mu^2}{K} \to 0$ local alternatives are:
  - for AR $\{\Delta : \frac{\Delta^2\mu^2}{\sqrt{K}} \leq C\}$ i.e. $|\Delta| \propto \sqrt{\frac{\sqrt{K}}{\mu^2}}$
  - for LM $\{\Delta : \frac{|\Delta|\mu^2}{\sqrt{K}} \leq C\}$ i.e. $|\Delta| \propto \frac{\sqrt{K}}{\mu^2}$
  - AR has slower speed of detection

Conditional Switch Test: CLR

- We may think about combining three statistics optimally

$$\begin{pmatrix} AR(\beta_0) - \Delta^2 \frac{\mu^2}{\sqrt{K\Phi}} \\ LM^{1/2}(\beta_0) - \Delta \frac{\mu^2}{\sqrt{K\Psi}} \\ \widetilde{F} - \frac{\mu^2}{\sqrt{K\Upsilon}} \end{pmatrix} \Rightarrow \mathcal{N}\left(\mathbf{0}, \Sigma\right).$$

- AR and LM are for testing β_0 and $\widetilde{F}$ for assessing the strength of identification
- Lim, Wang and Zhang (2022) - suggests an optimal combination test
- Ayyar, Matsushita and Otsu (2022) - suggestions on how to build CLR test

Overview

Adding covariates: what is the problem?

- Linear IV model with one endogenous variable:

$$\begin{cases} Y_i = \beta X_i + \delta' W_i + e_i, \\ X_i = \pi' Z_i + \gamma' W_i + v_i, \end{cases}$$

- TSLS: regress Y_i on $\widehat{X}_i = \widehat{\pi}' Z_i + \widehat{\gamma}' W_i$ and on W_i.
- Equivalent to partialling out W from Y, X and Z and running TSLS without covariates
- Let $M_W = I - W(W'W)^{-1}W'$ be partialling out operator
- $Y^{\perp} = M_W Y$, $X^{\perp} = M_W X$, $Z^{\perp} = M_W Z$, $P^{\perp} = P_{Z^{\perp}}$

$$\widehat{\beta}_{TSLS} = \frac{(X^{\perp})' P^{\perp} Y^{\perp}}{(X^{\perp})' P^{\perp} X^{\perp}}$$

Adding covariates: what is the problem?

- When there are no covariates (W_i) the bias was removed by removing a diagonal
- Could we do a similar thing: partial out covariates and remove the diagonal from P_Z?
- Would the following estimator work?

$$\widehat{\beta} = \frac{\sum_{i \neq j} X_i^{\perp} P_{ij}^{\perp} Y_j^{\perp}}{\sum_{i \neq j} X_i^{\perp} P_{ij}^{\perp} Y_j^{\perp}} = \frac{(X^{\perp})' P_{JIV}^{\perp} Y^{\perp}}{(X^{\perp})' P_{JIV}^{\perp} X^{\perp}}$$

- No. This is the same as $\widehat{\beta} = \frac{X' M_W P_{JIV}^{\perp} M_W Y}{X' M_W P_{JIV}^{\perp} M_W X}$
- Matrix $M_W P_{JIV}^{\perp} M_W$ has a non-trivial diagonal and produces bias in the estimator

Adding covariates: what is the problem?

$$\widehat{\beta}_{TSLS} = \frac{X' M_W P^{\perp} M_W Y}{X' M_W P^{\perp} M_W X}$$

- What if we do this in opposite order:

$$\widehat{\beta} = \frac{\sum_{i \neq j} X_i (M_W P^{\perp} M_W)_{ij} Y_j}{\sum_{i \neq j} X_i (M_W P^{\perp} M_W)_{ij} X_j}$$

- It does not work either

$$\sum_{j \neq i} (M_W P^{\perp} M_W)_{ij} W_j \neq 0$$

it loses partialling out property

Adding covariates: estimation

- Solution proposed in Chao, Swanson and Woutersen (2023): find $\theta_1, ..., \theta_n$ and diagonal matrix D_θ:

$$M_W(P^\perp - D_\theta)M_W \text{ has zero diagonal}$$

  this problem is linear and solvable for well-balanced designs
- Suggested estimator

$$\widehat{\beta} = \frac{X' M_W(P^\perp - D_\theta)M_W Y}{X' M_W(P^\perp - D_\theta)M_W X}$$

- Chao, Swanson and Woutersen (2023) has proof of consistency and asymptotic gaussianity under some assumptions

Adding covariates: robust inference

$$\begin{cases} Y_i = \beta X_i + \delta' W_i + e_i, \\ X_i = \pi' Z_i + \gamma' W_i + v_i, \end{cases}$$

- We can create a weak IV robust test for $H_0 : \beta = \beta_0$ using this idea

$$AR(\beta_0) = \frac{1}{\sqrt{K\Phi}}(Y - \beta_0 X)' M_W (P^{\perp} - D_{\theta}) M_W (Y - \beta_0 X)$$

- Under the null $AR(\beta_0) \Rightarrow N(0, 1)$, reject when $AR(\beta_0)$ is large

Overview

Conclusions and Open Questions

- Many instruments come with costs - one needs to find an optimal way to combine them
- Uncertainty about the first stage produces biases of TSLS
- Jackknifing or deleting diagonal is productive idea for both estimation and inference
- The knife-edge case for consistency happens when $\frac{\pi' Z' Z \pi}{\sqrt{K}} \asymp const$
- There is a pre-test for weak identification robust to heteroscedasticity when $K \to \infty$, which depends on the estimator one uses with it
- Robust tests (AR and LM) use the leave-one-out quadratic forms
- Adding many covariates is non-trivial

(Relatively simple) open questions

- Open question: there is a pre-test for whether one can trust JIVE-Wald confidence set/ t-test. JIVE-LIML is more efficient (Hausman et al, 2012), but there is no pre-test for it
- Open question: there is no pre-test that accommodates many covariates either
- Open question: unclear what to do with inferences when there are multiple endogenous variables (sub-vector inference)

(Hard) open questions

- Open question: many instruments framework accommodates well heterogeneous first stage, what to do about heterogeneous structural equation (non-parametric IV)
- Open question: How to use ML on the first stage? Sample splitting?
- Open question: Many instruments in Time Series- do not even know how to approach...