Evaluating the Validity and Robustness of Instrumental-Variable Analyses

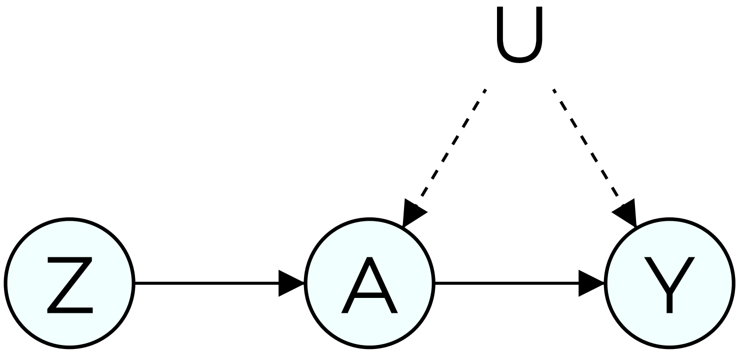
18 July 2024

Kai Cooper **Penn**
Guilherme Duarte **Penn**
Luke Keele **Penn**
Dean Knox **Penn**
Kennedy Mattes **Harvard**
Jonathan Mummolo **Princeton**

The Problem

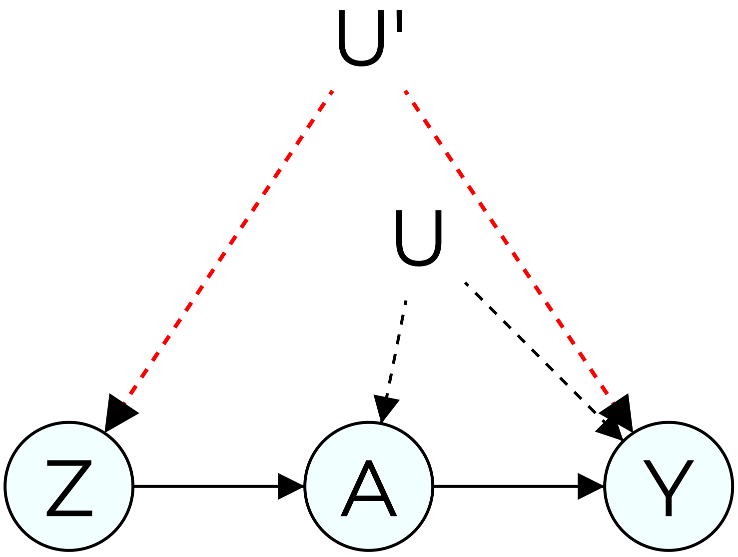
- Is voting habit forming? (Davenport et al., 2010)
  - Z: instrument (encouragement to vote at $t = 1$)
  - A: treatment (voting at $t = 1$)
  - Y: outcome (voting at $t = 2$)
  - U: unobserved confounder (e.g. political interest)

- We want to investigate the effect of A on Y:



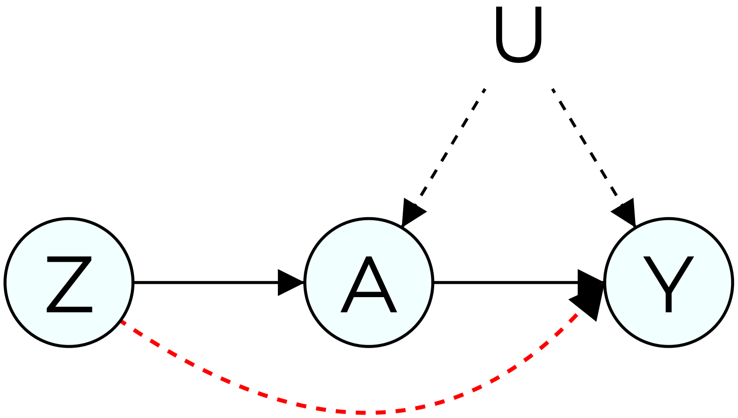
IV Theory Review

1. **Exogeneity**: $(Y(a), A(z)) \perp\!\!\!\perp Z$:
    - Violation: Confounding between Z and Y



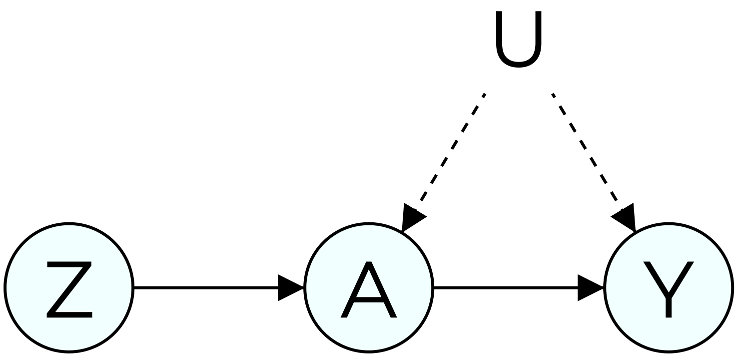
IV Theory Review

1. **Exogeneity**: $(Y(a), A(z)) \perp\!\!\!\perp Z$:
    - Violation: Confounding between Z and Y
2. **Exclusion restriction**: $Y(a, z) = Y(a)$
    - Violation: Direct arrow from Z to Y



IV Theory Review

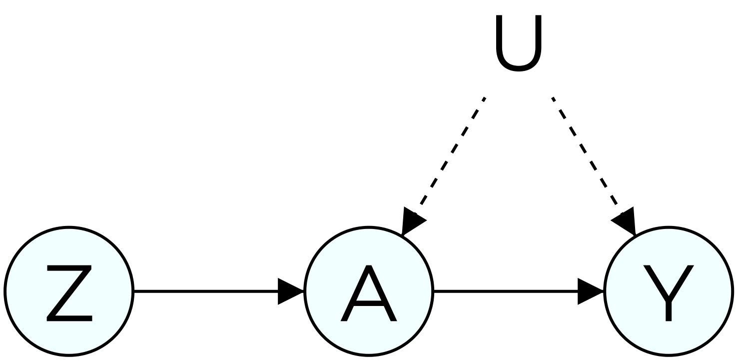
1. **Exogeneity**: $(Y(a), A(z)) \perp\!\!\!\perp Z$:
   - Violation: Confounding between Z and Y
2. **Exclusion restriction**: $Y(a, z) = Y(a)$
   - Violation: Direct arrow from Z to Y
3. **No defiers**: $A(Z = 1) \geq A(Z = 0)$
   - Violation: Units challenge their assignment



IV Theory Review

1. **Exogeneity**: $(Y(a), A(z)) \perp\!\!\!\perp Z$:
   - Violation: Confounding between Z and Y
2. **Exclusion restriction**: $Y(a, z) = Y(a)$
   - Violation: Direct arrow from Z to Y
3. **No defiers**: $A(Z = 1) \geq A(Z = 0)$
   - Violation: Units challenge their assignment

- LATE = $E[Y(a_1) - Y(a_0)|\text{compliers}]$ is identifiable if one assumes 1, 2, and 3 (Imbens & Angrist, '94)



Are these assumptions testable?

“fundamentally untestable, and its validity has to be argued in the context of a particular application”
(Imbens & Angrist,'94' on monotonicity/IV assumptions)

Are these assumptions testable?

"fundamentally untestable, and its validity has to be argued in the context of a particular application"
(Imbens & Angrist,'94' on monotonicity/IV assumptions)

- Those assumptions indeed have observable implications (Pearl, '95; Balke & Pearl, '97)
    - These are useful for falsification tests

Testing/Sensitivity Framework

- **Framework**:

Testing/Sensitivity Framework

- **Framework**:
    1. Define assumptions, collect available data, state an estimand

Testing/Sensitivity Framework

- **Framework**:
  1. Define assumptions, collect available data, state an estimand
  2. Test if data contradicts assumptions

Testing/Sensitivity Framework

- **Framework**:
  1. Define assumptions, collect available data, state an estimand
  2. Test if data contradicts assumptions
  3. Derive sharp bounds for the estimand

Testing/Sensitivity Framework

- **Framework**:
    1. Define assumptions, collect available data, state an estimand
    2. Test if data contradicts assumptions
    3. Derive sharp bounds for the estimand
    4. Sensitivity: check how violations affect results

Testing/Sensitivity Framework

- **Framework**:
    1. Define assumptions, collect available data, state an estimand
    2. Test if data contradicts assumptions
    3. Derive sharp bounds for the estimand
    4. Sensitivity: check how violations affect results

- Based on automated partial identification (Duarte et al., '23; Duarte, '24)
    - When a quantity is not identified, we still get sharp bounds

Testing/Sensitivity Framework

- **Framework**:
  1. Define assumptions, collect available data, state an estimand
  2. Test if data contradicts assumptions
  3. Derive sharp bounds for the estimand
  4. Sensitivity: check how violations affect results

- Based on automated partial identification (Duarte et al., '23; Duarte, '24)
  - When a quantity is not identified, we still get sharp bounds
  - We can evaluate bounds when assumptions are relaxed

Evaluating IV Assumptions

Testing

$$P(y_1, a_1|z_1) - P(y_1, a_1|z_0) \geq 0$$
$$P(y_0, a_1|z_1) - P(y_0, a_1|z_0) \geq 0$$
$$P(y_0, a_0|z_0) - P(y_0, a_0|z_1) \geq 0$$
$$P(y_1, a_0|z_0) - P(y_1, a_0|z_1) \geq 0$$
$$P(a_1|z_1) - P(a_1|z_0) \geq 0$$

Evaluating IV Assumptions

Testing

$$
\begin{gathered}
P(y_1, a_1|z_1) - P(y_1, a_1|z_0) \geq 0 \\
P(y_0, a_1|z_1) - P(y_0, a_1|z_0) \geq 0 \\
P(y_0, a_0|z_0) - P(y_0, a_0|z_1) \geq 0 \\
P(y_1, a_0|z_0) - P(y_1, a_0|z_1) \geq 0 \\
P(a_1|z_1) - P(a_1|z_0) \geq 0
\end{gathered}
$$

Sensitivity

$$\theta \in [0, 0.2],$$

$$\psi \in [0, 0.01]$$

Evaluating IV Assumptions

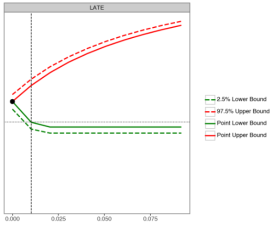
Testing

$$P(y_1, a_1|z_1) - P(y_1, a_1|z_0) \geq 0$$
$$P(y_0, a_1|z_1) - P(y_0, a_1|z_0) \geq 0$$
$$P(y_0, a_0|z_0) - P(y_0, a_0|z_1) \geq 0$$
$$P(y_1, a_0|z_0) - P(y_1, a_0|z_1) \geq 0$$
$$P(a_1|z_1) - P(a_1|z_0) \geq 0$$

Sensitivity

$$\theta \in [0, 0.2],$$
$$\psi \in [0, 0.01]$$

Applications



Testing Assumptions

Observable Implications

- In 1995, Pearl derived *Instrumental Inequalities* (Pearl, '95):

$$\max_a \sum_y [\max_z P(y, a|z)] \leq 1, \text{ under assumptions } 1, 2$$

  - If inequalities are violated, then IV assumptions are invalid

Observable Implications

- In 1995, Pearl derived *Instrumental Inequalities* (Pearl, '95):

$$\max_a \sum_y [\max_z P(y, a|z)] \leq 1, \text{ under assumptions } 1, 2$$

  - If inequalities are violated, then IV assumptions are invalid
- In practice, falsification tests can detect large violations

Observable Implications

- In 1995, Pearl derived *Instrumental Inequalities* (Pearl, '95):

$$\max_a \sum_y [\max_z P(y, a|z)] \leq 1, \text{ under assumptions } 1, 2$$

  - If inequalities are violated, then IV assumptions are invalid
- In practice, falsification tests can detect large violations
- Smaller violations may go undetected

Observable Implications

- In 1995, Pearl derived *Instrumental Inequalities* (Pearl, '95):

$$\max_a \sum_y [\max_z P(y, a|z)] \leq 1, \ \text{under assumptions } 1, 2$$

    - If inequalities are violated, then IV assumptions are invalid
- In practice, falsification tests can detect large violations
- Smaller violations may go undetected
- Note: if exogeneity is satisfied (e.g. by random assignment), this is a test of the exclusion restriction

Generalized IV Falsification Test

- Kedagni and Mourifie (2020) proved that if a model satisfies exclusion restriction and exogeneity, then:

$$\max_z P(y_1, a|z) + \max_z P(y_1, a'|z) \leq 1$$

$$\max_z P(y_1, a|z) - \min_z P(y_1|z) - \min_z P(y_1, a|z) + P(y_0, a'|z) \leq 0$$

$$\max_z P(y_0, a|z) - \min_z P(y_0|z) - \min_z P(Y_0, a|z) - P(y_1, a'|z) \leq 0$$

$$\min_z P(y_0|z) + \min_z P(y_1|z) + \min_z P(y_0, a|z) + P(y_1, a'|z) \min_z P(y_1, a|z) + P(y_0, a'|z) \geq 1$$

Generalized IV Falsification Test

- Kedagni and Mourifie (2020) proved that if a model satisfies exclusion restriction and exogeneity, then:

$$\max_z P(y_1, a|z) + \max_z P(y_1, a'|z) \leq 1$$

$$\max_z P(y_1, a|z) - \min_z P(y_1|z) - \min_z P(y_1, a|z) + P(y_0, a'|z) \leq 0$$

$$\max_z P(y_0, a|z) - \min_z P(y_0|z) - \min_z P(Y_0, a|z) - P(y_1, a'|z) \leq 0$$

$$\min_z P(y_0|z) + \min_z P(y_1|z) + \min_z P(y_0, a|z) + P(y_1, a'|z) \min_z P(y_1, a|z) + P(y_0, a'|z) \geq 1$$

- There are no other observable implications (sharpness)

Monotonicity Falsification Test

- If we also assume monotonicity (Balke & Pearl, '97'):

$$P(y_1, a_1|z_1) - P(y_1, a_1|z_0) \geq 0$$
$$P(y_0, a_1|z_1) - P(y_0, a_1|z_0) \geq 0$$
$$P(y_0, a_0|z_0) - P(y_0, a_0|z_1) \geq 0$$
$$P(y_1, a_0|z_0) - P(y_1, a_0|z_1) \geq 0$$
$$P(a_1|z_1) - P(a_1|z_0) \geq 0$$

Monotonicity Falsification Test

- If we also assume monotonicity (Balke & Pearl, '97'):

$$P(y_1, a_1|z_1) - P(y_1, a_1|z_0) \geq 0$$
$$P(y_0, a_1|z_1) - P(y_0, a_1|z_0) \geq 0$$
$$P(y_0, a_0|z_0) - P(y_0, a_0|z_1) \geq 0$$
$$P(y_1, a_0|z_0) - P(y_1, a_0|z_1) \geq 0$$
$$\mathbf{P(a_1|z_1) - P(a_1|z_0) \geq 0}$$

- The ATE of Z on A being positive is a weak test

Monotonicity Falsification Test

- If we also assume monotonicity (Balke & Pearl, '97'):

$$P(y_1, a_1|z_1) - P(y_1, a_1|z_0) \geq 0$$
$$P(y_0, a_1|z_1) - P(y_0, a_1|z_0) \geq 0$$
$$P(y_0, a_0|z_0) - P(y_0, a_0|z_1) \geq 0$$
$$P(y_1, a_0|z_0) - P(y_1, a_0|z_1) \geq 0$$
$$\mathbf{P(a_1|z_1) - P(a_1|z_0) \geq 0}$$

- The ATE of Z on A being positive is a weak test
- The test is sharp (Kitagawa, 2015)

Sensitivity Analysis

Sensitivity Analysis

- How much can these assumptions be violated before the data is uninformative?

Sensitivity Analysis

- How much can these assumptions be violated before the data is uninformative?
- Let θ be the proportion of defiers and ψ, of E.R. violators
  - How do bounds change in response to their values?

Sensitivity Analysis

- How much can these assumptions be violated before the data is uninformative?
- Let θ be the proportion of defiers and ψ, of E.R. violators
  - How do bounds change in response to their values?
- *Sensitivity function*: bounds as function of violations / data
  - E.g. what are LATE bounds given θ or ψ and $P(Y, A, Z)$?

Sensitivity Analysis

- How much can these assumptions be violated before the data is uninformative?
- Let θ be the proportion of defiers and ψ, of E.R. violators
  - How do bounds change in response to their values?
- *Sensitivity function*: bounds as function of violations / data
  - E.g. what are LATE bounds given θ or ψ and $P(Y, A, Z)$?
- How can we derive sensitivity functions?
  - Use *Autobounds* (Duarte et al., '23) to get numerical approximations

Sensitivity Analysis

- How much can these assumptions be violated before the data is uninformative?
- Let θ be the proportion of defiers and ψ, of E.R. violators
  - How do bounds change in response to their values?
- *Sensitivity function*: bounds as function of violations / data
  - E.g. what are LATE bounds given θ or ψ and $P(Y, A, Z)$?
- How can we derive sensitivity functions?
  - Use *Autobounds* (Duarte et al., '23) to get numerical approximations
  - Use Autobounds-Ext (Duarte, '24) to derive closed-form solutions

Sensitivity Analysis

- How much can these assumptions be violated before the data is uninformative?
- Let θ be the proportion of defiers and ψ, of E.R. violators
  - How do bounds change in response to their values?
- *Sensitivity function*: bounds as function of violations / data
  - E.g. what are LATE bounds given θ or ψ and $P(Y, A, Z)$?
- How can we derive sensitivity functions?
  - Use *Autobounds* (Duarte et al., '23) to get numerical approximations
  - Use Autobounds-Ext (Duarte, '24) to derive closed-form solutions
  - Both are based on the principles of automated partial id.

Sensitivity Analysis

- How much can these assumptions be violated before the data is uninformative?
- Let θ be the proportion of defiers and ψ, of E.R. violators
  - How do bounds change in response to their values?
- *Sensitivity function*: bounds as function of violations / data
  - E.g. what are LATE bounds given θ or ψ and $P(Y, A, Z)$?
- How can we derive sensitivity functions?
  - Use *Autobounds* (Duarte et al., '23) to get numerical approximations
  - Use Autobounds-Ext (Duarte, '24) to derive closed-form solutions
  - Both are based on the principles of automated partial id.
  - One states a causal question, introduces data and assumptions, and gets sharp bounds on the estimand

Sensitivity Analysis

- Sensitivity functions depend on *exact* violations:
  - θ, ψ take precise values

Sensitivity Analysis

- Sensitivity functions depend on *exact* violations:
  - θ, ψ take precise values
- But we want to understand how bounds change across a range of violations:
  - E.g. defiers are *at most* 0.2 of units, restrict $\theta \in [0, 0.2]$
  - E.g. E.R. violation units are *at most* 0.01, restrict $\psi \in [0, 0.01]$

Sensitivity Analysis

- Sensitivity functions depend on *exact* violations:
  - θ, ψ take precise values
- But we want to understand how bounds change across a range of violations:
  - E.g. defiers are *at most* 0.2 of units, restrict $\theta \in [0, 0.2]$
  - E.g. E.R. violation units are *at most* 0.01, restrict $\psi \in [0, 0.01]$
- **Sensitivity analysis**: optimize bounds over possible θ and ψ using *Autobounds*:

Empirical Applications

Simulation

- We simulate a scenario with $N = 10^6$ units:
  - 10% of defiers and 31.5% of units violating exclusion restriction
  - 4.6% of units violate both assumptions at the same time

Simulation

- We simulate a scenario with $N = 10^6$ units:
  - 10% of defiers and 31.5% of units violating exclusion restriction
  - 4.6% of units violate both assumptions at the same time

- Test the inequalities against the data
  - Detection of no defiers and E.R. violations (p-value < 0.01)
  - $E[A(z_1) - A(z_0)] = 0.1$: naive ATE_A test fails to detect them

Simulation

- We simulate a scenario with $N = 10^6$ units:
  - 10% of defiers and 31.5% of units violating exclusion restriction
  - 4.6% of units violate both assumptions at the same time
- Test the inequalities against the data
  - Detection of no defiers and E.R. violations (p-value < 0.01)
  - $E[A(z_1) - A(z_0)] = 0.1$: naive ATE_A test fails to detect them
- Sensitivity Analysis:
  - Proportion of violating units is at least 0.03
  - LATE can be signed in (0.03, 0.07)
  - LATE can't be signed above 0.07

Reanalysis: Davenport et al. (2010)

- Testing habit forming with turnout encouragement
  - instrument Z: encouragement to vote
  - treatment A: voting in 2006 Michigan elections
  - outcome Y: voting in subsequent elections

Reanalysis: Davenport et al. (2010)

- Testing habit forming with turnout encouragement
  - instrument Z: encouragement to vote
  - treatment A: voting in 2006 Michigan elections
  - outcome Y: voting in subsequent elections
- Are there defiers? Is the exclusion restriction violated?

Reanalysis: Davenport et al. (2010)

- Testing habit forming with turnout encouragement
  - instrument Z: encouragement to vote
  - treatment A: voting in 2006 Michigan elections
  - outcome Y: voting in subsequent elections

- Are there defiers? Is the exclusion restriction violated?

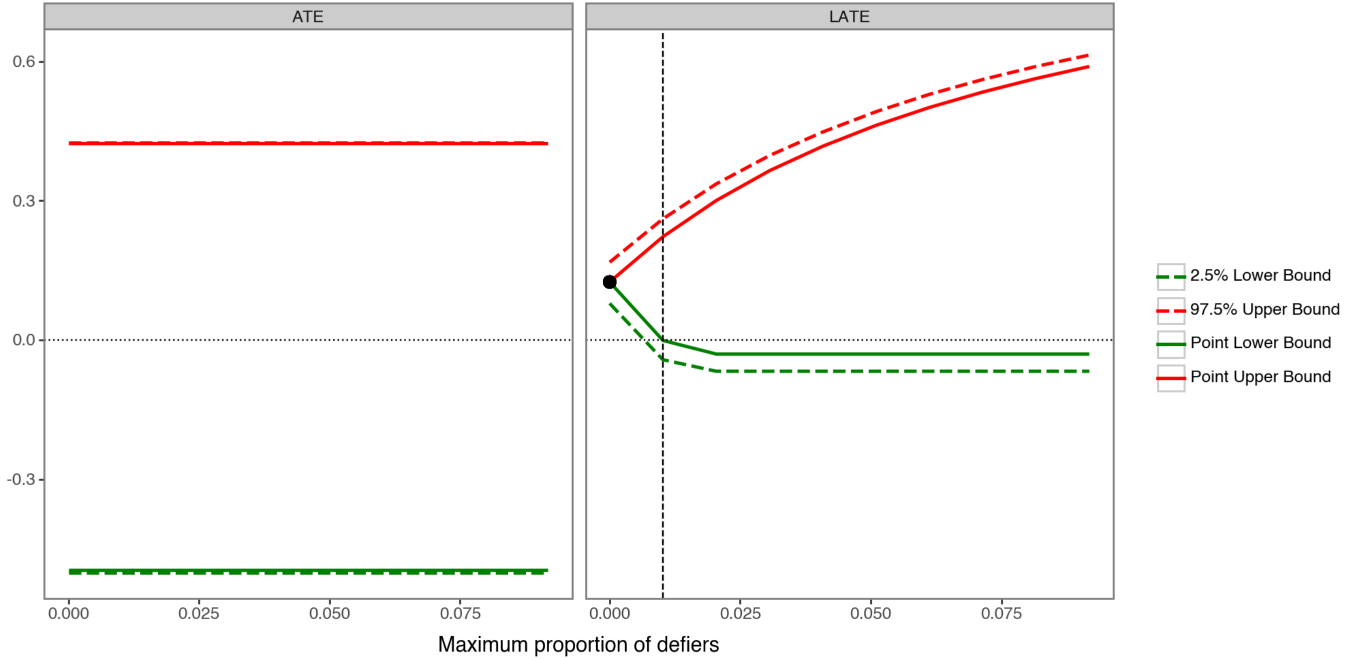
- Testing results:
  - Tests: cannot reject violations
  - ATE: $[-0.494, 0.423]$, with 95% CI of $[-0.5, 0.425]$
  - LATE: 0.124, with 95% CI of $[0.08, 0.17]$

Reanalysis: Davenport et al. (2010)

- Testing habit forming with turnout encouragement
  - instrument Z: encouragement to vote
  - treatment A: voting in 2006 Michigan elections
  - outcome Y: voting in subsequent elections
- Are there defiers? Is the exclusion restriction violated?
- Testing results:
  - Tests: cannot reject violations
  - ATE: $[-0.494, 0.423]$, with 95% CI of $[-0.5, 0.425]$
  - LATE: 0.124, with 95% CI of $[0.08, 0.17]$
- Just because we did not detect violations, it does not mean they are not there, so we proceed with sensitivity analysis

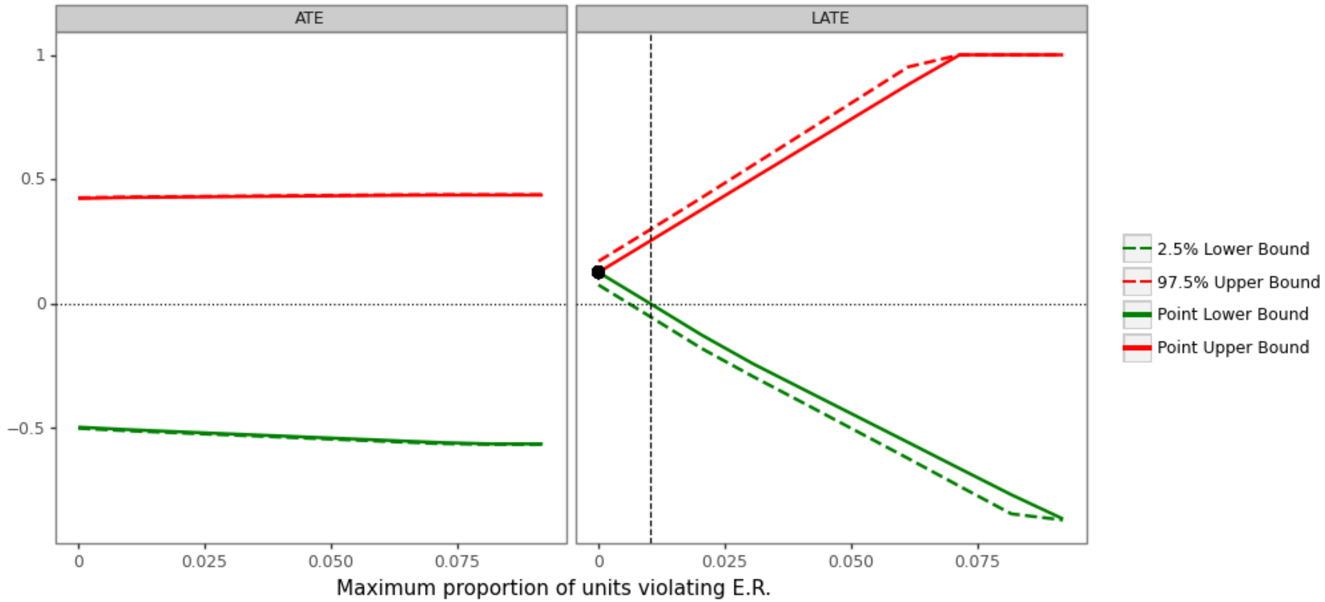
Reanalysis: Davenport et al. (2010)

- Sensitivity Analysis:
  - How robust are those results to violations of no defiers?
  - ATE is not much affected
  - LATE is positive if the proportion of defiers is $< 1\%$
  - Violations cause nonlinear impact on the LATE



Reanalysis: Davenport et al. (2010)

- Sensitivity Analysis:
  - How robust are those results to violations of E.R.?
  - ATE is not much affected
  - LATE is positive if the proportion of violations is is $< 1\%$
  - Violations cause linear impact on the LATE



More Complex Scenarios

Judge IV Design and Issues

- Use judge random assignment as natural experiment:
  - E.g. estimate the effect of pre-trial detention on conviction
  - Z: judge random assignment
  - A: pre-trial detention
  - Y: conviction

Judge IV Design and Issues

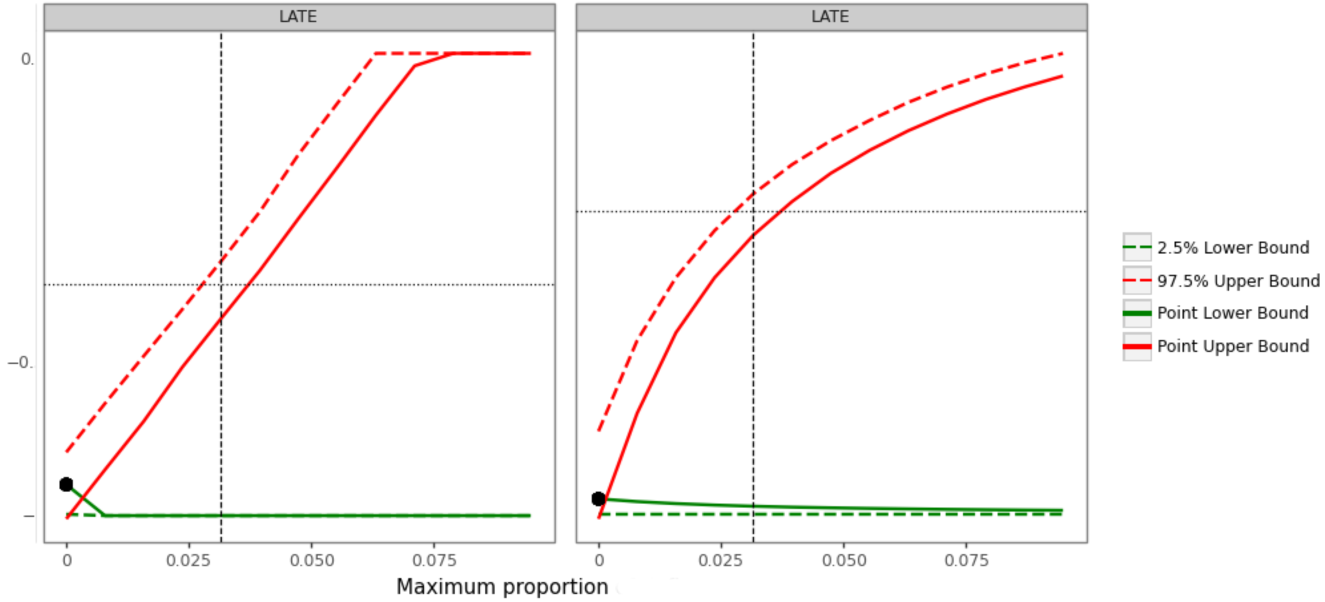
- Use judge random assignment as natural experiment:
    - E.g. estimate the effect of pre-trial detention on conviction
    - Z: judge random assignment
    - A: pre-trial detention
    - Y: conviction
- Complications:
    - Instrument is many valued
        - Who is a defier?
    - Exclusion restriction violated
        - Trial judge can read the case notes of arraignment judge

Judge IV Design and Issues

- Use judge random assignment as natural experiment:
  - E.g. estimate the effect of pre-trial detention on conviction
  - Z: judge random assignment
  - A: pre-trial detention
  - Y: conviction
- Complications:
  - Instrument is many valued
    - Who is a defier?
  - Exclusion restriction violated
    - Trial judge can read the case notes of arraignment judge
- Reanalysis of Stevenson (2018): positive effect of 0.13
- Our paper: derive results for many-valued instrument
- Today: compare two judges at a time (more severe to more lenient)

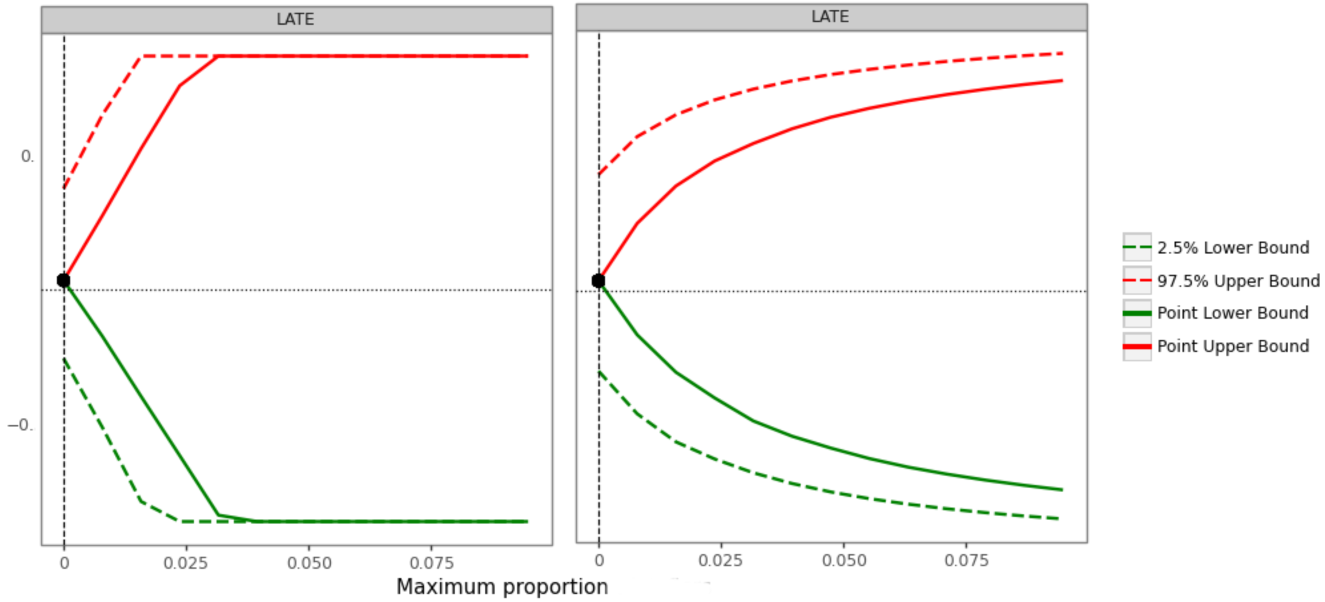
Reanalysis of Stevenson (2018)

- Comparison between the most extreme judges (z_0 / z_n):
  - "No defiers" assumption is rejected (p-value < 0.01)
  - The bounds cross in the region where θ is 0.0015, with LATE equal to -0.84
  - High negative LATE contradicts the main result of the paper, suggesting small violations cause high bias to the LATE estimate, even when evaluated at the minimum θ



Reanalysis of Stevenson (2018)

- Comparison between somewhat extreme judges (z_2 / z_n):
  - "No defiers" is not rejected for this case
  - No pre-existent violation, so LATE is identifiable at 0.044
  - LATE is unsigned if we allow for small θ deviation (close to 0)



Conclusions

- IV assumptions often characterized as untestable
- We *can* empirically evaluate key assumptions
  - Falsify monotonicity/exclusion restriction
  - Sensitivity analysis for defiers and E.R. violations
- Show we can reject assumptions in practice
- In applications, IV results extremely sensitive to minor violations
- Extensions:
  - Characterize robustness in the IV literature
  - Use framework for other models, e.g. factorial experiments

Guilherme Duarte

gjduarte@upenn.edu