

Causality, Social Sciences, and Statistics

Kosuke Imai

Department of Government and Department of Statistics
Harvard University

Beyond Curve Fitting
AAAI Spring Symposium, March 26, 2019
Stanford University

Causal Inference and Statistics

- Two methodological issues of causal inference:
 - 1 **Identification**: what can we learn if we have infinite amount of data?
↪ study design uncertainty
 - 2 **Inference**: what can we learn about identifiable quantities from a finite sample? ↪ statistical uncertainty
- Both are important, but identification precedes inference
- Statisticians have focused on inference until recently
 - Foundation of experimental analysis ↪ randomization inference
 - R. Fisher and J. Neyman
 - potential outcomes framework: $Y_i(1) - Y_i(0)$
 - Many stayed away from observational studies
 - exceptions: J. Robins, D. Rubin et al. who worked on applications
- **Causal revolution** over the last 30 years
 - influences from scientific disciplines (medical and social sciences)
 - Holland (1986) “Statistics and Causal Inference”

Scientific Applications Drive Causal Inference

- Applications lead to the use of “reasonable” methods followed by rigorous mathematical development

Regression discontinuity design
(Thistlewaite & Campbell, 1960)

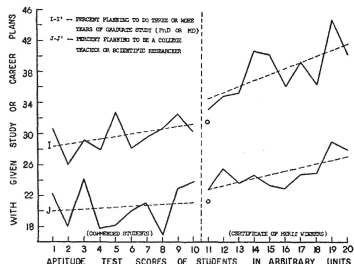


FIG. 3. Regression of study and career plans on exposure determiner.

Difference in differences design
(Card & Krueger, 1994)

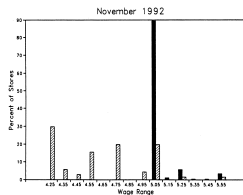
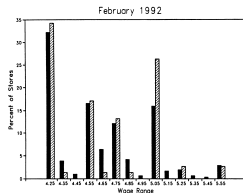


FIGURE 1. DISTRIBUTION OF STARTING WAGE RATES

- Same applies to instrumental variables, mediation analysis, sensitivity analysis, etc.

Evaluation of the Indian National Health Insurance Program

(Joint work with Anup Malani et al.)

- 150 million people worldwide face financial catastrophe due to health spending \rightsquigarrow 1/3 live in India
- In 2008, the Indian government introduced the national health insurance program (RSBY) to cover 60 million poorest families
- The government wants to expand the RSBY to 500 million Indians
- What are financial and health impacts of this expansion?
- Do beneficiaries have spillover effects on non-beneficiaries?
- Two-stage randomized design:

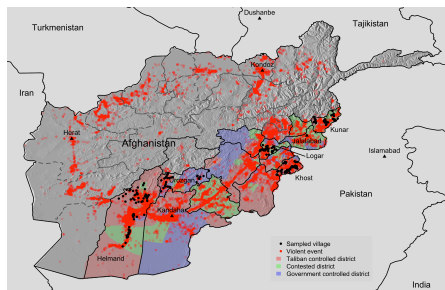
Mechanisms	Village prop.	Treatment	Control
High	50%	80%	20%
Low	50%	40%	60%

- Challenges: noncompliance, social interactions, spillover effects, general equilibrium effects

Air Strikes and Insurgency in Afghanistan

(Joint work with Jason Lyall et al.)

- War in Afghanistan (from 2001) as the longest war in US history
- War against the Taliban insurgency \rightsquigarrow winning “hearts and minds”
- Use of airstrikes \rightsquigarrow civilian casualties
- What are the impacts of airstrikes on insurgency activities and civilian attitudes?
- More than 23,000 airstrikes by USAF from 2006 to 2011
- Over 100,000 insurgency attacks against ISAF forces from 2005 to 2011
- 5 wave surveys of more than 2,000 villages
- Challenges: spillover effects across space, carryover effects over time, strategic interactions, roles of geography



Evaluation of Risk Assessment Scores in Criminal Justice System

(Joint work with Jim Greiner et al.)

- Public Safety Assessment (PSA) provides judges with numeric scores representing the risk of a defendant committing a new crime or missing a court date
- With PSA and other information about an arrestee, a judge at an initial appearance hearing decides whether to release the arrestee with or without a bail etc.
- Does the PSA help judges make better decisions? How does the PSA affect the behaviors of arrestees through judges' decisions?
- Two randomized controlled trials in Iowa
- Treatment = provision of the PSA scores to judges
 - ① Polk county: 10,400 cases (7,400 arrests) have been randomized
 - ② Linn county: 1,500 cases have been randomized
- Challenges: spillover effects across cases, heterogeneous effects across judges and arrestees, counterfactual fairness

Effects of Political Campaign Advertisement Videos

- TV advertisements are still the most popular campaign platform with more than 8 billion dollars spent in the 2018 election cycle
- Video data = image features + audio features

- We can classify the contents of ads
- We can predict which voters like which ads



- Why do certain ads appeal to some voters?
- What makes some ads more persuasive than others?
- How can we make new ads that are more effective?
- How would the opponent respond to your new ads, and how that response alter the effectiveness of your ads?

Causes and Consequences of Gerrymandering

- Redistricting after every ten years
- In many states, state legislatures control redistricting process
- Sophisticated gerrymandering attempts using voter files
- Cases currently pending in the Supreme Court

- Optimal redistricting maps
- Detecting gerrymandering \rightsquigarrow need for baseline distribution of redistricting maps
- Various simulation and optimization methods

- What are the electoral and other consequences of gerrymandering? Does gerrymandering cause polarization?
- How should we prevent gerrymandering? Does independent commission work?



Causal Challenges in Social Sciences and Statistics

- Digital and Internet revolution \rightsquigarrow Big and diverse data
 \rightsquigarrow New causal problems in social sciences
- Two methodological challenges:
 - 1 Social interactions
 - social network data; spatial and temporal data
 - spillover, carryover, and diffusion effects; strategic interactions
 - measurement challenges, importance of causal mechanisms
 - 2 Unstructured data
 - textual data; audio data; video data; maps
 - high-dimensional treatments
 - interpretability of features, prescriptive inference
- Evidence-based policy-making
 - dynamic treatment assignments and decisions
 - continuous evaluations
- Need for collaboration across disciplines and between scientists and methodologists