Design and Analysis of Two-Stage Randomized Experiments

Kosuke Imai

Department of Government and Department of Statistics
Harvard University

IDSS Distinguished Seminar Series

Massachusetts Institute of Technology

May 7, 2019

Joint work with Zhichao Jiang and Anup Malani

Methodological Motivation

- Causal inference revolution over the last three decades
- The first half of this revolution $\rightsquigarrow$ no interference between units

- In social sciences, interference is the rule rather than the exception
- How should we account for spillover effects?

- Experimental design solution:

  two-stage randomized experiments (Hudgens and Halloran, 2008)

Empirical Motivation: Indian Health Insurance Experiment

- 150 million people worldwide face financial catastrophe due to health spending $\rightsquigarrow$ 1/3 live in India
- In 2008, Indian government introduced the national health insurance program (RSBY) to cover about 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?

- We conduct an RCT to evaluate the impact of expanding RSBY in the State of Karnakata

Study Design

- Sample: 10,879 households in 435 villages
- Experimental conditions:
  - A Opportunity to enroll in RSBY essentially for free
  - B No intervention
- Time line:
  1. September 2013 – February 2014: Baseline survey
  2. April – May 2015: Enrollment
  3. September 2016 – January 2017: Endline survey
- Two stage randomization:

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

Causal Inference and Interference between Units

1. Causal inference without interference between units
   - Potential outcomes: $Y_i(1)$ and $Y_i(0)$
   - Observed outcome: $Y_i = Y_i(D_i)$
   - Causal effect: $Y_i(1) - Y_i(0)$
2. Causal inference with interference between units
   - Potential outcomes: $Y_i(d_1, d_2, \ldots, d_N)$
   - Observed outcome: $Y_i = Y_i(D_1, D_2, \ldots, D_N)$
   - Causal effects:
     - Direct effect $= Y_i(D_i = 1, \mathbf{D}_{-i} = \mathbf{d}) - Y_i(D_i = 0, \mathbf{D}_{-i} = \mathbf{d})$
     - Spillover effect $= Y_i(D_i = d, \mathbf{D}_{-i} = \mathbf{d}) - Y_i(D_i = d, \mathbf{D}_{-i} = \mathbf{d}')$

Fundamental problem of causal inference
$\rightsquigarrow$ only one potential outcome is observed

What Happens if We Ignore Interference?

- Completely randomized experiment
  - Total of N units with N_1 treated units
  - $\Pr(D_i = 1) = N_1/N$ for all i
- Difference-in-means estimator is unbiased for the average direct effect:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{\mathbf{d}_{-i}}\Big\{Y_i(D_i=1,\mathbf{D}_{-i}=\mathbf{d}_{-i})\underbrace{\mathbb{P}(\mathbf{D}_{-i}=\mathbf{d}_{-i}\mid D_i=1)}_{1/\binom{N-1}{N_1-1}} - $$

$$-Y_i(D_i=0,\mathbf{D}_{-i}=\mathbf{d}_{-i})\underbrace{\mathbb{P}(\mathbf{D}_{-i}=\mathbf{d}_{-i}\mid D_i=0)}_{1/\binom{N-1}{N_1}}\Big\}$$

- Bernoulli randomization (or large sample) simplifies the expression

$$\frac{1}{N2^{N-1}}\sum_{i=1}^{N}\sum_{\mathbf{d}_{-i}}\left\{Y_i(D_i=1,\mathbf{D}_{-i}=\mathbf{d}_{-i}) - Y_i(D_i=0,\mathbf{D}_{-i}=\mathbf{d}_{-i})\right\}$$

- Cannot estimate spillover effects

What about Cluster Randomized Experiment?

- Setup:
  - Total of J clusters with J_1 treated clusters
  - Total of N units: n_j units in cluster j
  - Complete randomization of treatment across clusters
  - All units are treated in a treated cluster
  - No unit is treated in a control cluster
- Partial interference assumption:
  - No interference across clusters
  - Interference within a cluster is allowed
- Difference-in-means estimator is unbiased for the average total effect:

$$\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\{Y_{ij}(D_{1j}=1, D_{2j}=1, \ldots, D_{n_j j}=1) \\ - Y_{ij}(D_{1j}=0, D_{2j}=0, \ldots, D_{n_j j}=0)\}$$

- Cannot estimate spillover effects

Two-stage Randomized Experiments

- Individuals (households): $i = 1, 2, \ldots, N$
- Blocks (villages): $j = 1, 2, \ldots, J$
- Size of block j: n_j where $N = \sum_{j=1}^{J} n_j$

- Binary treatment assignment mechanism: $A_j \in \{0, 1\}$
- Binary encouragement to receive treatment: $Z_{ij} \in \{0, 1\}$
- Binary treatment indicator: $D_{ij} \in \{0, 1\}$
- Observed outcome: Y_{ij}

- Partial interference assumption: No interference across blocks
  - Potential treatment and outcome: $D_{ij}(\mathbf{z}_j)$ and $Y_{ij}(\mathbf{z}_j)$
  - Observed treatment and outcome: $D_{ij} = D_{ij}(\mathbf{Z}_j)$ and $Y_{ij} = Y_{ij}(\mathbf{Z}_j)$
- Number of potential values reduced from 2^N to 2^{n_j}

Intention-to-Treat Analysis: Causal Quantities of Interest

- Average outcome under the treatment $Z_{ij} = z$ and the assignment mechanism $A_j = a$:

$$\overline{Y}_{ij}(z, a) \;=\; \sum_{\mathbf{z}_{-i,j}} Y_{ij}(Z_{ij} = z, \mathbf{Z}_{-i,j} = \mathbf{z}_{-i,j}) \mathbb{P}_a(\mathbf{Z}_{-i,j} = \mathbf{z}_{-i,j} \mid Z_{ij} = z)$$

- Average direct effect of encouragement on outcome:

$$\text{ADE}^Y(a) \;=\; \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left\{ \overline{Y}_{ij}(1, a) - \overline{Y}_{ij}(0, a) \right\}$$

- Average spillover effect of encouragement on outcome:

$$\text{ASE}^Y(z) \;=\; \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left\{ \overline{Y}_{ij}(z, 1) - \overline{Y}_{ij}(z, 0) \right\}$$

- Horvitz-Thompson estimator for unbiased estimation

Effect Decomposition

- Average total effect of encouragement on outcome:

$$\text{ATE}^Y = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left\{\overline{Y}_{ij}(1,1) - \overline{Y}_{ij}(0,0)\right\}$$

- Total effect = Direct effect + Spillover effect:

$$\text{ATE}^Y = \text{ADE}^Y(1) + \text{ASE}^Y(0) = \text{ADE}^Y(0) + \text{ASE}^Y(1)$$

- In a two-stage RCT, we have an unbiased estimator,

$$\mathbb{E}\left[\frac{\sum_{j=1}^{J}\mathbf{1}\{A_j = a\}\frac{n_j}{N}\frac{\sum_{i=1}^{n_j} Y_{ij}\mathbf{1}\{Z_{ij}=z\}}{\sum_{i=1}^{n_j}\mathbf{1}\{Z_{ij}=z\}}}{\frac{1}{J}\sum_{j=1}^{J}\mathbf{1}\{A_j = a\}}\right] = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\overline{Y}_{ij}(z,a)$$

- Halloran and Struchiner (1995), Sobel (2006), Hudgens and Halloran (2008)

When Should We Use Two-stage Randomization?

- Do we care about spillover effects?
  - Yes $\rightsquigarrow$ two-stage randomization
  - No
    - Interested in direct effects $\rightsquigarrow$ individual randomization
    - Interested in total effects $\rightsquigarrow$ cluster randomization

- Do we lose statistical power if there is no spillover effect?
  - variance of the difference-in-means estimator:

$$\frac{1-\rho}{J^2}\left\{\mathbb{V}(Y_{ij}(1))\sum_{a=0}^{1}\frac{J_a}{np_a}+\mathbb{V}(Y_{ij}(0))\sum_{a=0}^{1}\frac{J_a}{n(1-p_a)}\right\}-\frac{1-\rho}{Jn}\mathbb{V}(Y_{ij}(1)-Y_{ij}(0))$$

  - large intracluster correlation coefficient $\rightsquigarrow$ more efficient
  - large variation in p_a $\rightsquigarrow$ less efficient
  - trade-off between detection of spillover effects and statistical efficiency

Complier Average Direct Effect

- Goal: Estimate the treatment effect rather than the ITT effect
- Use randomized encouragement as an instrument
  1. Monotonicity: $D_{ij}(Z_{ij} = 1) \geq D_{ij}(Z_{ij} = 0)$
  2. Exclusion restriction: $Y_{ij}(z_{ij}, d_{ij}) = Y_{ij}(z'_{ij}, d_{ij})$ for any z_{ij} and z'_{ij}
- Generalization to the case with spillover effects
  1. Monotonicity: $D_{ij}(1, \mathbf{z}_{-i,j}) \geq D_{ij}(0, \mathbf{z}_{-i,j})$ for any $\mathbf{z}_{-i,j}$
  2. Exclusion restriction: $Y_{ij}(\mathbf{z}_j, \mathbf{d}_j) = Y_{ij}(\mathbf{z}'_j, \mathbf{d}_j)$ for any $\mathbf{z}_j$ and $\mathbf{z}'_j$
- Compliers: $C_{ij}(\mathbf{z}_{-i,j}) \; = \; \mathbf{1}\{D_{ij}(1, \mathbf{z}_{-i,j}) = 1, D_{ij}(0, \mathbf{z}_{-i,j}) = 0\}$
- Complier average direct effect of encouragement (CADE(z, a)):

$$\frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \{Y_{ij}(1, \mathbf{z}_{-i,j}) - Y_{ij}(0, \mathbf{z}_{-i,j})\} C_{ij}(\mathbf{z}_{-i,j}) \mathbb{P}_a(\mathbf{Z}_{-i,j} = \mathbf{z}_{-i,j} \mid Z_{ij} = z)}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} C_{ij}(\mathbf{z}_{-i,j}) \mathbb{P}_a(\mathbf{Z}_{-i,j} = \mathbf{z}_{-i,j} \mid Z_{ij} = z)}$$

- We propose a consistent estimator of the CADE

Key Identification Assumption

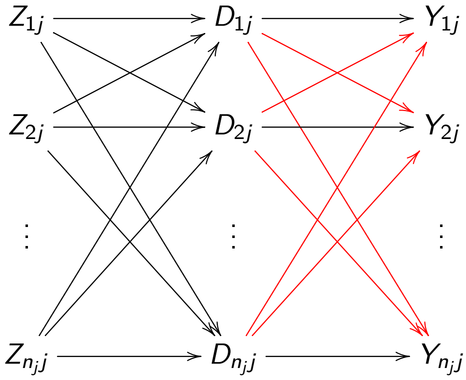
- Two causal mechanisms:
  - Z_{ij} affects Y_{ij} through D_{ij}
  - Z_{ij} affects Y_{ij} through $\mathbf{D}_{-i,j}$
- Idea: if Z_{ij} does not affect D_{ij}, it should not affect Y_{ij} through $\mathbf{D}_{-i,j}$

Assumption (Restricted Interference for Noncompliers)

If a unit has $D_{ij}(1, \mathbf{z}_{-i,j}) = D_{ij}(0, \mathbf{z}_{-i,j}) = d$ *for any given* $\mathbf{z}_{-i,j}$*, it must also satisfy* $Y_{ij}(d, \mathbf{D}_{-i,j}(Z_{ij} = 1, \mathbf{z}_{-i,j})) = Y_{ij}(d, \mathbf{D}_{-i,j}(Z_{ij} = 0, \mathbf{z}_{-i,j}))$

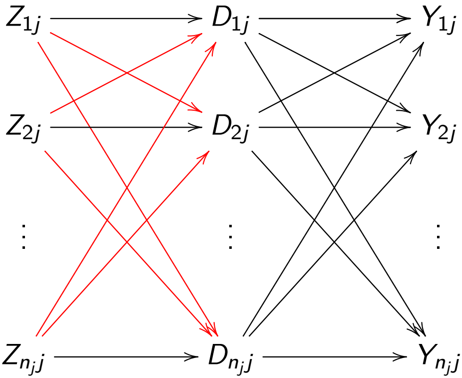
Scenario I: No Spillover Effect of the Treatment Receipt on the Outcome

$$Y_{ij}(d_{ij}, \mathbf{d}_{-i,j}) \ = \ Y_{ij}(d_{ij}, \mathbf{d}'_{-i,j})$$



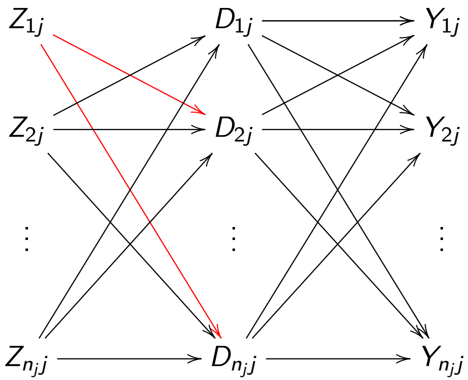
Scenario II: No Spillover Effect of the Treatment Assignment on the Treatment Receipt

$D_{ij}(z_{ij}, \mathbf{z}_{-i,j}) \ = \ D_{ij}(z_{ij}, \mathbf{z}'_{-i,j})$ (Kang and Imbens, 2016)



Scenario III: Limited Spillover Effect of the Treatment Assignment on the Treatment Receipt

$$\text{If } D_{ij}(1, \mathbf{z}_{-i,j}) = D_{ij}(0, \mathbf{z}_{-i,j}) \text{ for any given } \mathbf{z}_{-i,j},$$
$$\text{then } D_{i'j}(1, \mathbf{z}_{-i,j}) = D_{i'j}(0, \mathbf{z}_{-i,j}) \text{ for all } i' \neq i$$



Identification and Consistent Estimation

1. Identification: monotonicity, exclusion restriction, restricted interference for noncompliers

$$\lim_{n_j \to \infty} \text{CADE}(z, a) \;=\; \lim_{n_j \to \infty} \frac{\text{ADE}^Y(a)}{\text{ADE}^D(a)}$$

2. Consistent estimation: additional restriction on interference (e.g., Savje et al.)

$$\frac{\widehat{\text{ADE}}^Y(a)}{\widehat{\text{ADE}}^D(a)} \xrightarrow{p} \lim_{n_j \to \infty, J \to \infty} \text{CADE}(z, a)$$

Randomization Inference

- Variance is difficult to characterize

Assumption (Stratified Interference (Hudgens and Halloran. 2008))

$Y_{ij}(z_{ij}, \mathbf{z}_{-i,j}) = Y_{ij}(z_{ij}, \mathbf{z}'_{-i,j})$ and $D_{ij}(z_{ij}, \mathbf{z}_{-i,j}) = D_{ij}(z_{ij}, \mathbf{z}'_{-i,j})$ *if* $\sum_{i'=1}^{n_j} z_{ij} = \sum_{i=1}^{n_j} z'_{ij}$

- Under stratified interference, our estimand simplifies to,

$$\begin{aligned} & \text{CADE}(a) \\ = \quad & \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\{Y_{ij}(1,a) - Y_{ij}(0,a)\}\mathbf{1}\{D_{ij}(1,a)=1, D_{ij}(0,a)=0\}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{1}\{D_{ij}(1,a)=1, D_{ij}(0,a)=0\}} \end{aligned}$$

- Compliers: $C_{ij} = \mathbf{1}\{D_{ij}(1,a)=1, D_{ij}(0,a)=0\}$
- Consistent estimation possible without additional restriction
- We propose an approximate asymptotic variance estimator

Connection to the Two-stage Least Squares Estimator

- The model:

$$\begin{aligned} Y_{ij} &= \sum_{a=0}^{1} \alpha_a \mathbf{1}\{A_j = a\} + \sum_{a=0}^{1} \underbrace{\beta_a}_{\text{CADE}} D_{ij} \mathbf{1}\{A_j = a\} + \epsilon_{ij} \\ D_{ij} &= \sum_{a=0}^{1} \gamma_a \mathbf{1}\{A_j = a\} + \sum_{a=0}^{1} \delta_a Z_{ij} \mathbf{1}\{A_j = a\} + \eta_{ij} \end{aligned}$$

- Weighted two-stage least squares estimator:

$$w_{ij} = \frac{1}{\Pr(A_j) \Pr(Z_{ij} \mid A_j)}$$

- Transforming the outcome and treatment: multiplying them by $n_j J/N$
- Randomization-based variance is equal to the weighted average of cluster-robust HC2 $\left(1 - \frac{J_a}{J}\right)$ and individual-robust HC2 variances $\left(\frac{J_a}{J}\right)$

Results: Indian Health Insurance Experiment

- A household is more likely to enroll in RSBY if a large number of households are given the opportunity

Average Spillover Effects	Treatment	Control
Individual-weighted	0.086 (s.e. $=$ 0.053)	0.045 (s.e. $=$ 0.028)
Block-weighted	0.044 (s.e. $=$ 0.018)	0.031 (s.e. $=$ 0.021)

- Households will have greater hospitalization expenditure if few households are given the opportunity

Complier Average Direct Effects	High	Low
Individual-weighted	-1649 (s.e. $=$ 1061)	1984 (s.e. $=$ 1215)
Block-weighted	-485 (s.e. $=$ 1258)	3752 (s.e. $=$ 1652)

Concluding Remarks

- In social science research,
  1. people interact with each other $\rightsquigarrow$ interference
  2. people don't follow instructions $\rightsquigarrow$ noncompliance
- Two-stage randomized controlled trials:
  1. randomize assignment mechanisms across clusters
  2. randomize treatment assignment within each cluster
- Spillover effects as causal quantities of interest
- Our contributions:
  1. Identification condition for complier average direct effects
  2. Consistent estimator for CADE and its variance
  3. Connections to regression and instrumental variables
  4. Application to the India health insurance experiment
  5. Implementation as part of R package experiment

Send comments and suggestions to Imai@Harvard.Edu
Other research at `https://imai.fas.harvard.edu`