Statistics and Causal Inference

Kosuke Imai

Princeton University

February 2014
Academia Sinica, Taipei

Overview of the Workshop

A quick tour of modern causal inference methods

1. **Randomized Experiments**
   - Classical randomized experiments
   - Cluster randomized experiments
   - Instrumental variables
2. **Observational Studies**
   - Regression discontinuity design
   - Matching and weighting
   - Fixed effects and difference-in-differences
3. **Causal Mechanisms**
   - Direct and indirect effects
   - Causal mediation analysis

Introduction

What is Causal Inference?

- Comparison between factual and counterfactual for each unit
- Incumbency effect:
  What would have been the election outcome if a candidate were not an incumbent?
- Resource curse thesis:
  What would have been the GDP growth rate without oil?
- Democratic peace theory:
  Would the two countries have escalated crisis in the same situation if they were both autocratic?

Defining Causal Effects

- Units: $i = 1, \ldots, n$
- "Treatment": $T_i = 1$ if treated, $T_i = 0$ otherwise
- *Observed* outcome: Y_i
- Pre-treatment covariates: X_i
- Potential outcomes: $Y_i(1)$ and $Y_i(0)$ where $Y_i = Y_i(T_i)$

Voters	Contact	Turnout		Age	Party ID
i	T_i	$Y_i(1)$	$Y_i(0)$	X_i	X_i
1	1	1	?	20	D
2	0	?	0	55	R
3	0	?	1	40	R
⋮	⋮	⋮	⋮	⋮	⋮
n	1	0	?	62	D

- Causal effect: $Y_i(1) - Y_i(0)$

The Key Assumptions

- The notation implies three assumptions:
  1. No simultaneity (different from endogeneity)
  2. No interference between units: $Y_i(T_1, T_2, \ldots, T_n) = Y_i(T_i)$
  3. Same version of the treatment
- Stable Unit Treatment Value Assumption (SUTVA)
- Potential violations:
  1. feedback effects
  2. spill-over effects, carry-over effects
  3. different treatment administration

- Potential outcome is thought to be "fixed": data cannot distinguish fixed and random potential outcomes
- Potential outcomes across units have a distribution
- Observed outcome is random because the treatment is random

- Multi-valued treatment: more potential outcomes for each unit

Average Treatment Effects

- Sample Average Treatment Effect (SATE):

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$$

- Population Average Treatment Effect (PATE):

$$\mathbb{E}(Y_i(1) - Y_i(0))$$

- Population Average Treatment Effect for the Treated (PATT):

$$\mathbb{E}(Y_i(1) - Y_i(0) \mid T_i = 1)$$

- Treatment effect heterogeneity: Zero ATE doesn't mean zero effect for everyone! $\Longrightarrow$ Conditional ATE
- Other quantities: Quantile treatment effects etc.

Randomized Experiments

Classical Randomized Experiments

- Units: $i = 1, \ldots, n$
- May constitute a simple random sample from a population
- Treatment: $T_i \in \{0, 1\}$
- Outcome: $Y_i = Y_i(T_i)$
- Complete randomization of the treatment assignment
- Exactly n_1 units receive the treatment
- $n_0 = n - n_1$ units are assigned to the control group
- Assumption: for all $i = 1, \ldots, n$, $\sum_{i=1}^{n} T_i = n_1$ and

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i, \quad \Pr(T_i = 1) = \frac{n_1}{n}$$

- Estimand = SATE or PATE
- Estimator = Difference-in-means:

$$\hat{\tau} \equiv \frac{1}{n_1} \sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) Y_i$$

Estimation of Average Treatment Effects

- Key idea (Neyman 1923): Randomness comes from treatment assignment (plus sampling for PATE) alone
- Design-based (randomization-based) rather than model-based
- Statistical properties of $\hat{\tau}$ based on design features

- Define $\mathcal{O} \equiv \{Y_i(0), Y_i(1)\}_{i=1}^n$
- Unbiasedness (over repeated treatment assignments):

$$\begin{aligned}
\mathbb{E}(\hat{\tau} \mid \mathcal{O}) &= \frac{1}{n_1}\sum_{i=1}^{n}\mathbb{E}(T_i \mid \mathcal{O})\,Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\{1-\mathbb{E}(T_i \mid \mathcal{O})\}\,Y_i(0) \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) \;=\; \text{SATE}
\end{aligned}$$

- Over repeated sampling: $\mathbb{E}(\hat{\tau}) = \mathbb{E}(\mathbb{E}(\hat{\tau} \mid \mathcal{O})) = \mathbb{E}(\text{SATE}) = \text{PATE}$

Relationship with Regression

- The model: $Y_i = \alpha + \beta T_i + \epsilon_i$ where $\mathbb{E}(\epsilon_i) = 0$
- Equivalence: least squares estimate $\hat{\beta}$ =Difference in means
- Potential outcomes representation:

$$Y_i(T_i) = \alpha + \beta T_i + \epsilon_i$$

- Constant additive unit causal effect: $Y_i(1) - Y_i(0) = \beta$ for all i
- $\alpha = \mathbb{E}(Y_i(0))$
- A more general representation:

$$Y_i(T_i) = \alpha + \beta T_i + \epsilon_i(T_i) \quad \text{where} \quad \mathbb{E}(\epsilon_i(t)) = 0$$

- $Y_i(1) - Y_i(0) = \beta + \epsilon_i(1) - \epsilon_i(0)$
- $\beta = \mathbb{E}(Y_i(1) - Y_i(0))$
- $\alpha = \mathbb{E}(Y_i(0))$ as before

Bias of Model-Based Variance

- The design-based perspective: use Neyman's exact variance
- What is the bias of the model-based variance estimator?
- Finite sample bias:

$$\begin{aligned} \text{Bias} &= \mathbb{E}\left(\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(T_i - \overline{T}_n)^2}\right) - \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right) \\ &= \frac{(n_1 - n_0)(n-1)}{n_1 n_0 (n-2)}(\sigma_1^2 - \sigma_0^2) \end{aligned}$$

- Bias is zero when $n_1 = n_0$ or $\sigma_1^2 = \sigma_0^2$
- In general, bias can be negative or positive and does not asymptotically vanish

Robust Standard Error

- Suppose $\mathbb{V}(\epsilon_i \mid T) = \sigma^2(T_i) \neq \sigma^2$
- Heteroskedasticity consistent robust variance estimator:

$$\mathbb{V}(\widehat{(\hat{\alpha}, \hat{\beta})} \mid T) = \left(\sum_{i=1}^{n} x_i x_i^\top \right)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_i^2 x_i x_i^\top \right) \left(\sum_{i=1}^{n} x_i x_i^\top \right)^{-1}$$

  where in this case $x_i = (1, T_i)$ is a column vector of length 2
- Model-based justification: asymptotically valid in the presence of heteroskedastic errors
- Design-based evaluation:

$$\text{Finite Sample Bias} = -\left(\frac{\sigma_1^2}{n_1^2} + \frac{\sigma_0^2}{n_0^2} \right)$$

- Bias vanishes asymptotically

Cluster Randomized Experiments

- Units: $i = 1, 2, \ldots, n_j$
- Clusters of units: $j = 1, 2, \ldots, m$
- Treatment at cluster level: $T_j \in \{0, 1\}$
- Outcome: $Y_{ij} = Y_{ij}(T_j)$
- Random assignment: $(Y_{ij}(1), Y_{ij}(0)) \perp\!\!\!\perp T_j$
- Estimands at unit level:

$$\begin{aligned} \text{SATE} &\equiv \frac{1}{\sum_{j=1}^{m} n_j} \sum_{j=1}^{m} \sum_{i=1}^{n_j} (Y_{ij}(1) - Y_{ij}(0)) \\ \text{PATE} &\equiv \mathbb{E}(Y_{ij}(1) - Y_{ij}(0)) \end{aligned}$$

- Random sampling of clusters and units

Merits and Limitations of CREs

- Interference between units within a cluster is allowed
- Assumption: No interference between units of different clusters
- Often easy to implement: Mexican health insurance experiment

- Opportunity to estimate the spill-over effects
- D. W. Nickerson. Spill-over effect of get-out-the-vote canvassing within household (*APSR*, 2008)

- Limitations:
  1. A large number of possible treatment assignments
  2. Loss of statistical power

Design-Based Inference

- For simplicity, assume equal cluster size, i.e., $n_j = n$ for all j
- The difference-in-means estimator:

$$\hat{\tau} \equiv \frac{1}{m_1}\sum_{j=1}^{m} T_j \overline{Y}_j - \frac{1}{m_0}\sum_{j=1}^{m}(1-T_j)\overline{Y}_j$$

  where $\overline{Y}_j \equiv \sum_{i=1}^{n_j} Y_{ij}/n_j$
- Easy to show $\mathbb{E}(\hat{\tau} \mid \mathcal{O}) = \text{SATE}$ and thus $\mathbb{E}(\hat{\tau}) = \text{PATE}$
- Exact population variance:

$$\mathbb{V}(\hat{\tau}) = \frac{\mathbb{V}(\overline{Y_j(1)})}{m_1} + \frac{\mathbb{V}(\overline{Y_j(0)})}{m_0}$$

- Intracluster correlation coefficient ρ_t:

$$\mathbb{V}(\overline{Y_j(t)}) = \frac{\sigma_t^2}{n}\{1 + (n-1)\rho_t\} \leq \sigma_t^2$$

Cluster Standard Error

- Cluster robust variance estimator:

$$\widehat{\mathbb{V}((\hat{\alpha}, \hat{\beta}) \mid T)} = \left(\sum_{j=1}^{m} X_j^\top X_j \right)^{-1} \left(\sum_{j=1}^{m} X_j^\top \hat{\epsilon}_j \hat{\epsilon}_j^\top X_j \right) \left(\sum_{j=1}^{m} X_j^\top X_j \right)^{-1}$$

  where in this case $X_j = [1\, T_j]$ is an $n_j \times 2$ matrix and $\hat{\epsilon}_j = (\hat{\epsilon}_{1j}, \ldots, \hat{\epsilon}_{n_j j})$ is a column vector of length n_j

- Design-based evaluation (assume $n_j = n$ for all j):

$$\text{Finite Sample Bias} = -\left(\frac{\mathbb{V}(\overline{Y_j(1)})}{m_1^2} + \frac{\mathbb{V}(\overline{Y_j(0)})}{m_0^2} \right)$$

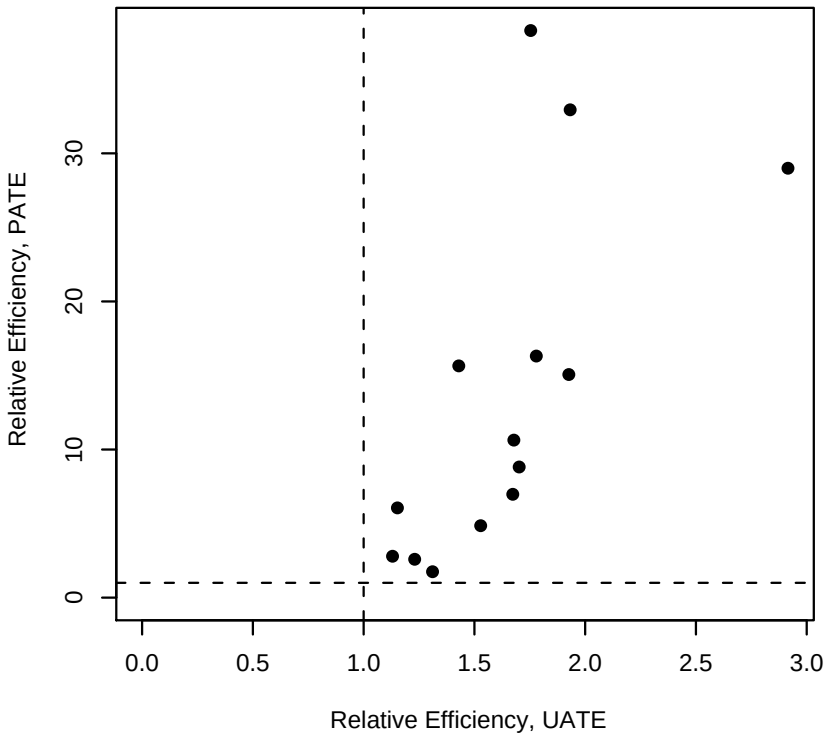
- Bias vanishes asymptotically as $m \to \infty$ with n fixed
- **Implication:** cluster standard errors by the unit of treatment assignment

Example: Seguro Popular de Salud (SPS)

- Evaluation of the Mexican universal health insurance program
- Aim: "provide social protection in health to the 50 million uninsured Mexicans"
- A key goal: reduce out-of-pocket health expenditures
- Sounds obvious but not easy to achieve in developing countries
- Individuals must affiliate in order to receive SPS services
- 100 health clusters nonrandomly chosen for evaluation
- Matched-pair design: based on population, socio-demographics, poverty, education, health infrastructure etc.
- "Treatment clusters": encouragement for people to affiliate
- Data: aggregate characteristics, surveys of $32,000$ individuals

Relative Efficiency of Matched-Pair Design (MPD)

- Compare with completely-randomized design
- Greater (positive) correlation within pair $\rightarrow$ greater efficiency
- UATE: MPD is between 1.1 and 2.9 times more efficient
- PATE: MPD is between 1.8 and 38.3 times more efficient!



Partial Compliance in Randomized Experiments

- Unable to force all experimental subjects to take the (randomly) assigned treatment/control
- Intention-to-Treat (ITT) effect $\neq$ treatment effect
- Selection bias: self-selection into the treatment/control groups

- Political information bias: effects of campaign on voting behavior
- Ability bias: effects of education on wages
- Healthy-user bias: effects of exercises on blood pressure

- Encouragement design: randomize the encouragement to receive the treatment rather than the receipt of the treatment itself

Potential Outcomes Notation

- Randomized encouragement: $Z_i \in \{0, 1\}$
- Potential treatment variables: $(T_i(1), T_i(0))$
  1. $T_i(z) = 1$: would receive the treatment if $Z_i = z$
  2. $T_i(z) = 0$: would not receive the treatment if $Z_i = z$
- Observed treatment receipt indicator: $T_i = T_i(Z_i)$
- Observed and potential outcomes: $Y_i = Y_i(Z_i, T_i(Z_i))$
- Can be written as $Y_i = Y_i(Z_i)$

- No interference assumption for $T_i(Z_i)$ and $Y_i(Z_i, T_i)$
- Randomization of encouragement:

$$(Y_i(1), Y_i(0), T_i(1), T_i(0)) \perp\!\!\!\perp Z_i$$

- But $(Y_i(1), Y_i(0)) \not\perp\!\!\!\perp T_i \mid Z_i = z$, i.e., selection bias

Principal Stratification Framework

- Imbens and Angrist (1994, *Econometrica*); Angrist, Imbens, and Rubin (1996, *JASA*)
- Four principal strata (latent types):
  - compliers $(T_i(1), T_i(0)) = (1, 0)$,
  - non-compliers $\begin{cases} always-takers & (T_i(1), T_i(0)) = (1, 1), \\ never-takers & (T_i(1), T_i(0)) = (0, 0), \\ defiers & (T_i(1), T_i(0)) = (0, 1) \end{cases}$

- Observed and principal strata:

	$Z_i = 1$	$Z_i = 0$
$T_i = 1$	Complier/Always-taker	Defier/Always-taker
$T_i = 0$	Defier/Never-taker	Complier/Never-taker

Instrumental Variables and Causality

- Randomized encouragement as an instrument for the treatment
- Two additional assumptions
  1. Monotonicity: No defiers

$$T_i(1) \geq T_i(0) \quad \text{for all } i.$$

  2. Exclusion restriction: Instrument (encouragement) affects outcome only through treatment

$$Y_i(1,t) \;=\; Y_i(0,t) \quad \text{for } t = 0,1$$

  Zero ITT effect for always-takers and never-takers
- ITT effect decomposition:

$$\begin{aligned} \text{ITT} &= \text{ITT}_c \times \Pr(\text{compliers}) + \text{ITT}_a \times \Pr(\text{always} - \text{takers}) \\ &\quad + \text{ITT}_n \times \Pr(\text{never} - \text{takers}) \\ &= \text{ITT}_c \Pr(\text{compliers}) \end{aligned}$$

IV Estimand and Interpretation

- IV estimand:

$$\begin{aligned} \text{ITT}_c &= \frac{\text{ITT}}{\Pr(\text{compliers})} \\ &= \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(T_i \mid Z_i = 1) - \mathbb{E}(T_i \mid Z_i = 0)} \\ &= \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(T_i, Z_i)} \end{aligned}$$

- ITT_c = Complier Average Treatment Effect (CATE)
- Local Average Treatment Effect (LATE)
- CATE $\neq$ ATE unless ATE for noncompliers equals CATE
- Different encouragement (instrument) yields different compliers
- Debate among Deaton, Heckman, and Imbens in *J. of Econ. Lit.*

Violation of IV Assumptions

- Violation of exclusion restriction:

$$\text{Large sample bias} \quad = \quad \text{ITT}_{\text{noncomplier}} \, \frac{\Pr(\text{noncomplier})}{\Pr(\text{complier})}$$

- Weak instruemnts (encouragement)
- Direct effects of encouragement; failure of randomization, alternative causal paths

- Violation of monotonicity:

$$\text{Large sample bias} \quad = \quad \frac{\{\text{CATE} + \text{ITT}_{\text{defier}}\} \Pr(\text{defier})}{\Pr(\text{complier}) - \Pr(\text{defier})}$$

- Proportion of defiers
- Heterogeneity of causal effects

An Example: Testing Habitual Voting

- Gerber *et al.* (2003) *AJPS*
- Randomized encouragement to vote in an election
- Treatment: turnout in the election
- Outcome: turnout in the next election

- Monotonicity: Being contacted by a canvasser would *never* discourage anyone from voting
- Exclusion restriction: being contacted by a canvasser in this election has no effect on turnout in the next election other than through turnout in this election
- CATE: Habitual voting for those who would vote if and only if they are contacted by a canvasser in this election

Concluding Remarks

- Even randomized experiments often require sophisticated statistical methods
- Deviation from the protocol:
  1. Spill-over, carry-over effects
  2. Noncompliance
  3. Missing data, measurement error
- Beyond the average treatment effect:
  1. Treatment effect heterogeneity
  2. Causal mechanisms
- Getting more out of randomized experiments:
  1. Generalizing experimental results
  2. Deriving individualized treatment rules
  3. Studying dynamic treatment regimes

Observational Studies

Challenges of Observational Studies

- Randomized experiments vs. Observational studies
- Tradeoff between internal and external validity
  - **Endogeneity**: selection bias
  - Generalizability: sample selection, Hawthorne effects, realism
- Statistical methods cannot replace good research design
- "Designing" observational studies
  - Natural experiments (haphazard treatment assignment)
  - Examples: birthdays, weather, close elections, arbitrary administrative rules and boundaries
- "Replicating" randomized experiments
- Key Questions:
  1. Where are the counterfactuals coming from?
  2. Is it a credible comparison?

Coping with Endogeneity in Observational Studies

- Selection bias in observational studies
- Two common research design strategies:
  1. Find a plausibly exogenous treatment
  2. Find a plausibly exogenous instrument
- A valid instrument satisfies the following conditions
  1. Exogenously assigned – no confounding
  2. It monotonically affects treatment
  3. It affects outcome only through treatment – no direct effect
- Challenge: plausibly exogenous instruments with no direct effect tends to be weak
- Another strategy: regression discontinuity design

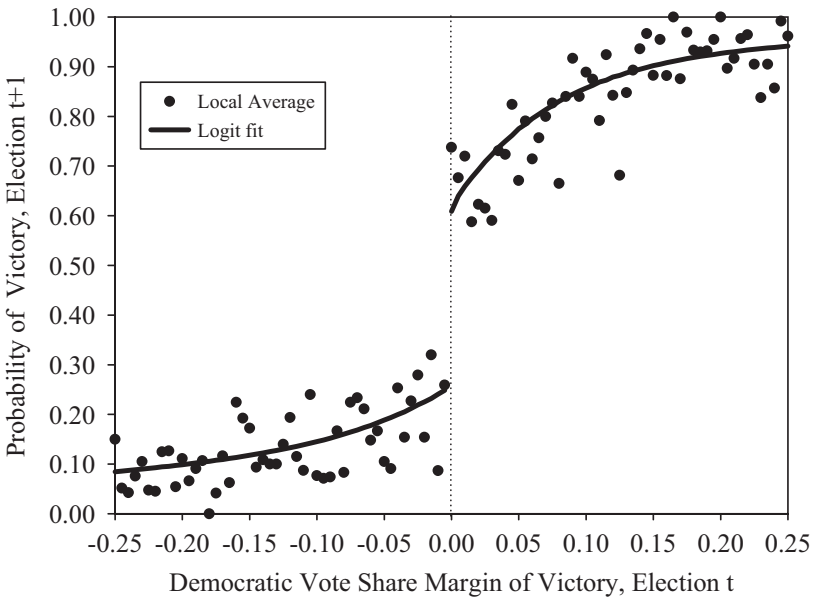
Regression Discontinuity Design

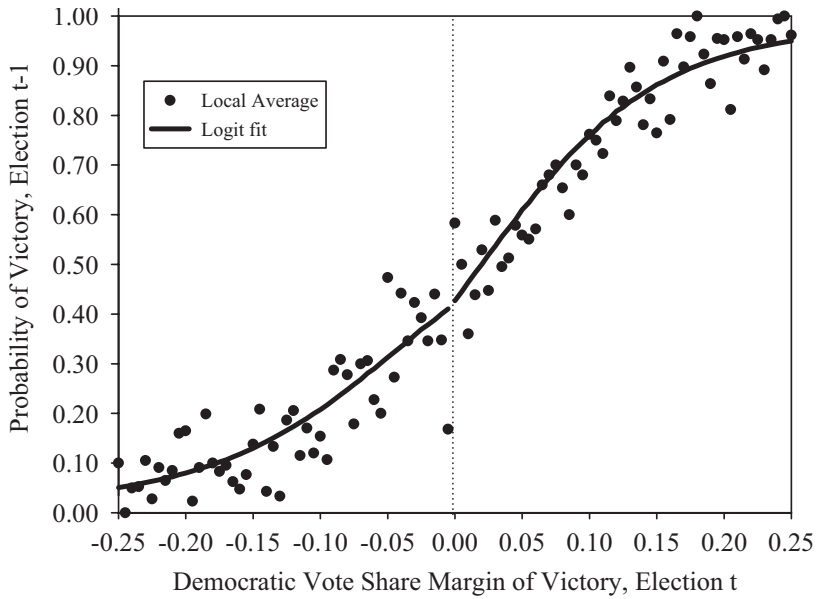
- Idea: Find an arbitrary cutpoint c which determines the treatment assignment such that $T_i = \mathbf{1}\{X_i \geq c\}$
- Assumption: $\mathbb{E}(Y_i(t) \mid X_i = x)$ is continuous in x
- Estimand: $\mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = c)$
- Regression modeling:

$$\begin{aligned}
\mathbb{E}(Y_i(1) \mid X_i = c) &= \lim_{x \downarrow c} \mathbb{E}(Y_i(1) \mid X_i = x) = \lim_{x \downarrow c} \mathbb{E}(Y_i \mid X_i = x) \\
\mathbb{E}(Y_i(0) \mid X_i = c) &= \lim_{x \uparrow c} \mathbb{E}(Y_i(0) \mid X_i = x) = \lim_{x \uparrow c} \mathbb{E}(Y_i \mid X_i = x)
\end{aligned}$$

- Advantage: internal validity
- Disadvantage: external validity
- Make sure nothing else is going on at $X_i = c$

Close Elections as RD Design (Lee)





- Placebo test for natural experiments
- What is a good placebo?
  1. expected not to have any effect
  2. closely related to outcome of interest

Fuzzy Regression Discontinuity Design

- Sharp regression discontinuity design: $T_i = \mathbf{1}\{X_i \geq c\}$
- What happens if we have noncompliance?
- Forcing variable as an instrument: $Z_i = \mathbf{1}\{X_i \geq c\}$
- Potential outcomes: $T_i(z)$ and $Y_i(z, t)$

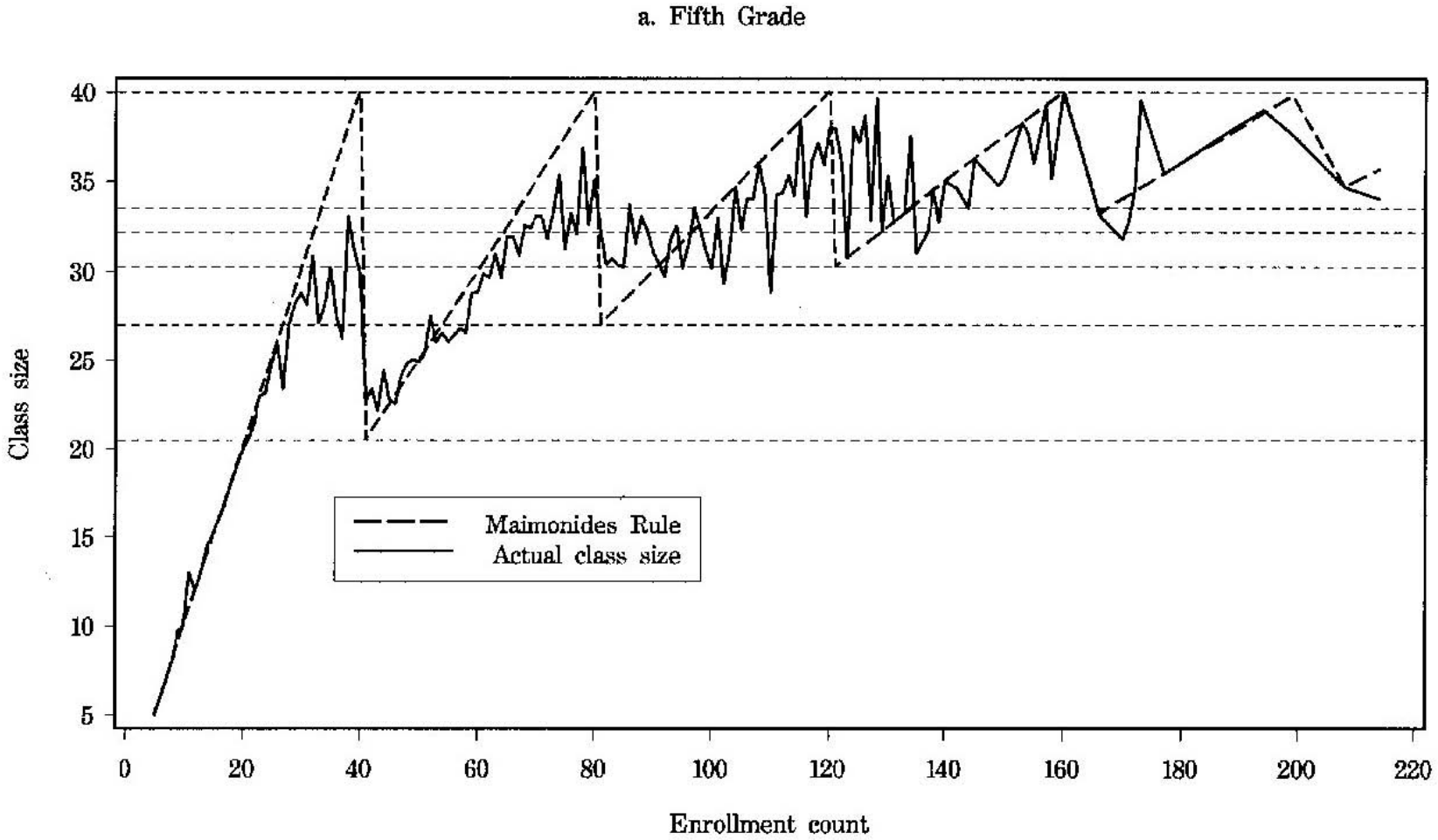
- Monotonicity: $T_i(1) \geq T_i(0)$
- Exclusion restriction: $Y_i(0, t) = Y_i(1, t)$
- $\mathbb{E}(T_i(z) \mid X_i = x)$ and $\mathbb{E}(Y_i(z, T_i(z)) \mid X_i = x)$ are continuous in x
- Estimand: $\mathbb{E}(Y_i(1, T_i(1)) - Y_i(0, T_i(0)) \mid \textit{Complier}, X_i = c)$
- Estimator:

$$\frac{\lim_{x \downarrow c} \mathbb{E}(Y_i \mid X_i = x) - \lim_{x \uparrow c} \mathbb{E}(Y_i \mid X_i = x)}{\lim_{x \downarrow c} \mathbb{E}(T_i \mid X_i = x) - \lim_{x \uparrow c} \mathbb{E}(T_i \mid X_i = x)}$$

- Disadvantage: external validity

An Example: Class Size Effect (Angrist and Lavy)

- Effect of class-size on student test scores
- Maimonides' Rule: Maximum class size $= 40$



Identification of the Average Treatment Effect

- Assumption 1: Overlap (i.e., no extrapolation)

$$0 < \Pr(T_i = 1 \mid X_i = x) < 1 \text{ for any } x \in \mathcal{X}$$

- Assumption 2: Ignorability (exogeneity, unconfoundedness, no omitted variable, selection on observables, etc.)

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid X_i = x \text{ for any } x \in \mathcal{X}$$

- Conditional expectation function: $\mu(t, x) = \mathbb{E}(Y_i(t) \mid T_i = t, X_i = x)$
- Regression-based Estimator:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)\}$$

- Standard error: delta method is pain, but simulation is easy (Zelig)

Matching as Nonparametric Preprocessing

- READING: Ho *et al.* *Political Analysis* (2007)
- Assume exogeneity holds: matching does NOT solve endogeneity
- Need to model $\mathbb{E}(Y_i \mid T_i, X_i)$
- Parametric regression – functional-form/distributional assumptions $\Longrightarrow$ model dependence
- Non-parametric regression $\Longrightarrow$ curse of dimensionality
- Preprocess the data so that treatment and control groups are similar to each other w.r.t. the observed pre-treatment covariates

- Goal of matching: achieve balance = independence between T and X
- "Replicate" randomized treatment w.r.t. observed covaraites
- Reduced model dependence: minimal role of statistical modeling

Sensitivity Analysis

- Consider a simple pair-matching of treated and control units
- Assumption: treatment assignment is "random"
- Difference-in-means estimator
- Question: How large a departure from the key (untestable) assumption must occur for the conclusions to no longer hold?
- Rosenbaum's sensitivity analysis: for any pair j,

$$\frac{1}{\Gamma} \leq \frac{\Pr(T_{1j} = 1)/\Pr(T_{1j} = 0)}{\Pr(T_{2j} = 1)/\Pr(T_{2j} = 0)} \leq \Gamma$$

- Under ignorability, $\Gamma = 1$ for all j
- How do the results change as you increase Γ?
- Limitations of sensitivity analysis
- FURTHER READING: P. Rosenbaum. *Observational Studies*.

The Role of Propensity Score

- The probability of receiving the treatment:

$$\pi(X_i) \equiv \Pr(T_i = 1 \mid X_i)$$

- The balancing property:

$$T_i \perp\!\!\!\perp X_i \mid \pi(X_i)$$

- Exogeneity given the propensity score (under exogeneity given covariates):

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \mid \pi(X_i)$$

- Dimension reduction
- But, true propensity score is unknown: propensity score tautology (more later)

Classical Matching Techniques

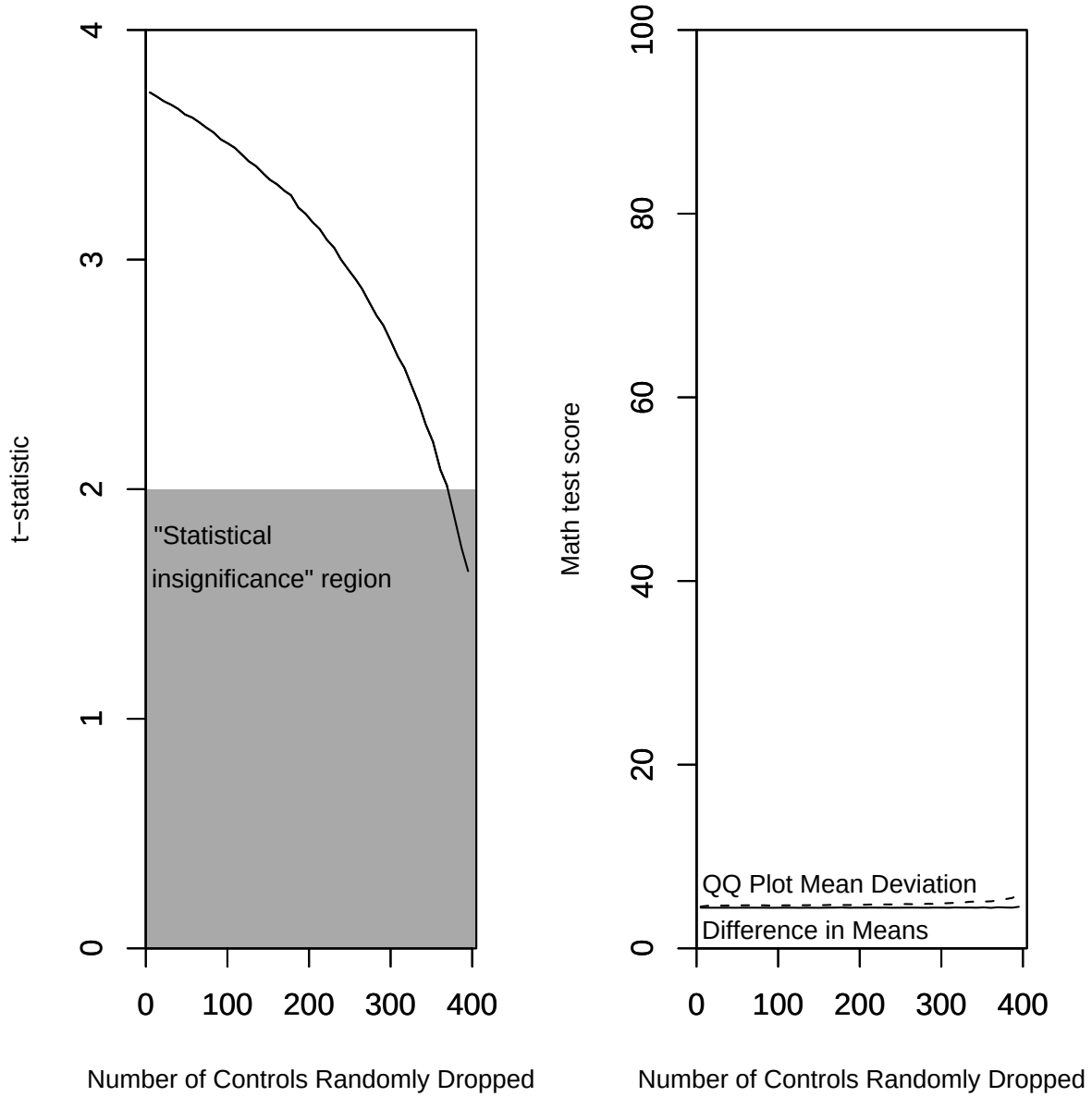
- Exact matching
- Mahalanobis distance matching: $\sqrt{(X_i - X_j)^\top \widetilde{\Sigma}^{-1}(X_i - X_j)}$
- Propensity score matching
- One-to-one, one-to-many, and subclassification
- Matching with caliper
- Which matching method to choose?
- Whatever gives you the "best" balance!
- Importance of substantive knowledge: propensity score matching with exact matching on key confounders
- FURTHER READING: Rubin (2006). *Matched Sampling for Causal Effects* (Cambridge UP)

How to Check Balance

- Success of matching method depends on the resulting balance
- How should one assess the balance of matched data?
- Ideally, compare the joint distribution of all covariates for the matched treatment and control groups
- In practice, this is impossible when X is high-dimensional
- Check various lower-dimensional summaries; (standardized) mean difference, variance ratio, empirical CDF, etc.

- Frequent use of balance test
  - t test for difference in means for each variable of X
  - other test statistics; e.g., χ^2, F, Kolmogorov-Smirnov tests
  - statistically insignificant test statistics as a justification for the adequacy of the chosen matching method and/or a stopping rule for maximizing balance

An Illustration of Balance Test Fallacy



Recent Advances in Matching Methods

- The main problem of matching: balance checking
- Skip balance checking all together
- Specify a balance metric and optimize it

- Optimal matching (Rosenbaum, Hansen): minimize sum of distances
- Genetic matching (Diamond and Sekhon): maximize minimum p-value
- Coarsened exact matching (King et al.): exact match on binned covariates
- SVM subsetting (Ratkovic): find the largest, balanced subset for general treatment regimes

Inverse Propensity Score Weighting

- Matching is inefficient because it throws away data
- Weighting by inverse propensity score

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}\right)$$

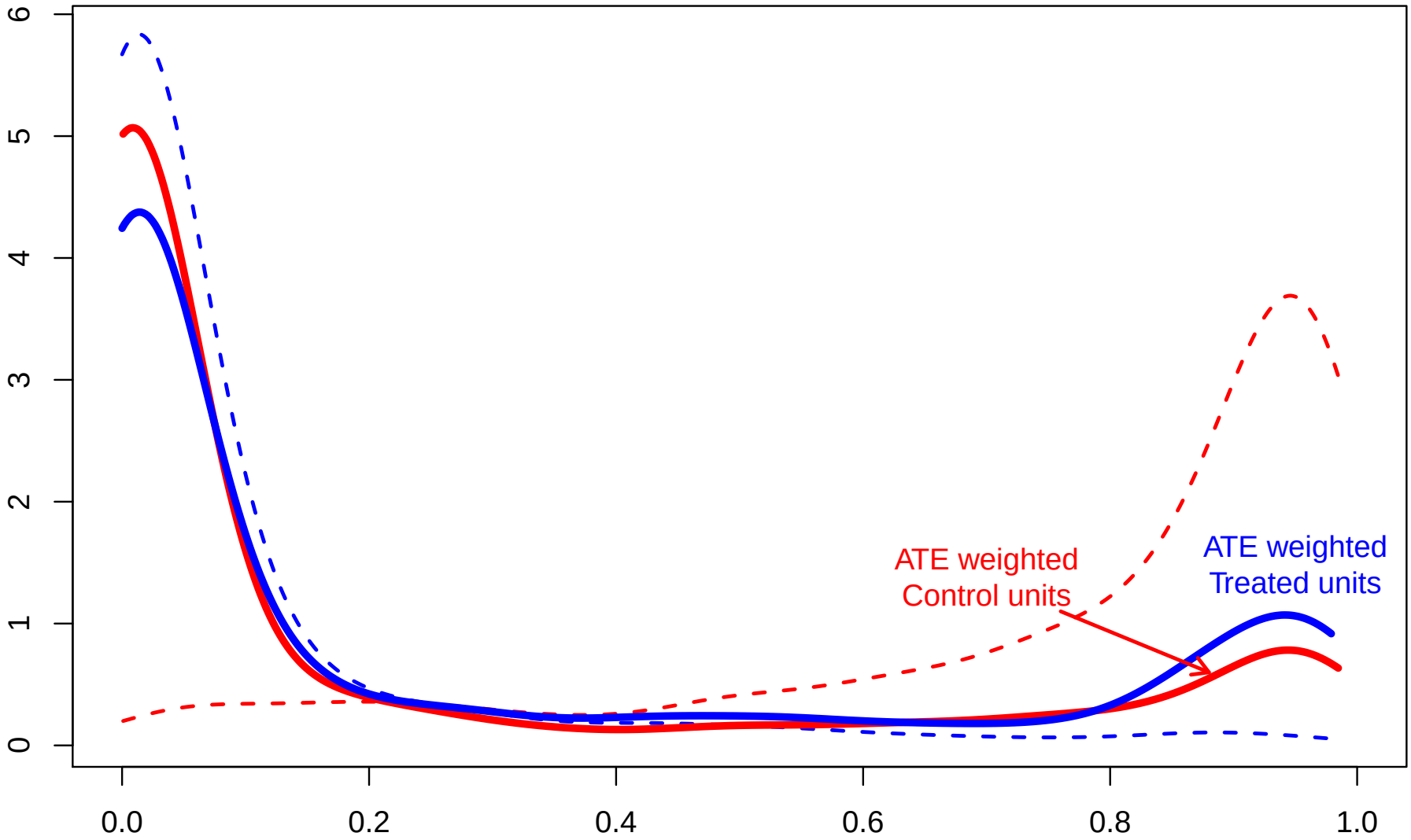
- An improved weighting scheme:

$$\frac{\sum_{i=1}^{n}\{T_i Y_i/\hat{\pi}(X_i)\}}{\sum_{i=1}^{n}\{T_i/\hat{\pi}(X_i)\}} - \frac{\sum_{i=1}^{n}\{(1-T_i)Y_i/(1-\hat{\pi}(X_i))\}}{\sum_{i=1}^{n}\{(1-T_i)/(1-\hat{\pi}(X_i))\}}$$

- Unstable when some weights are extremely small

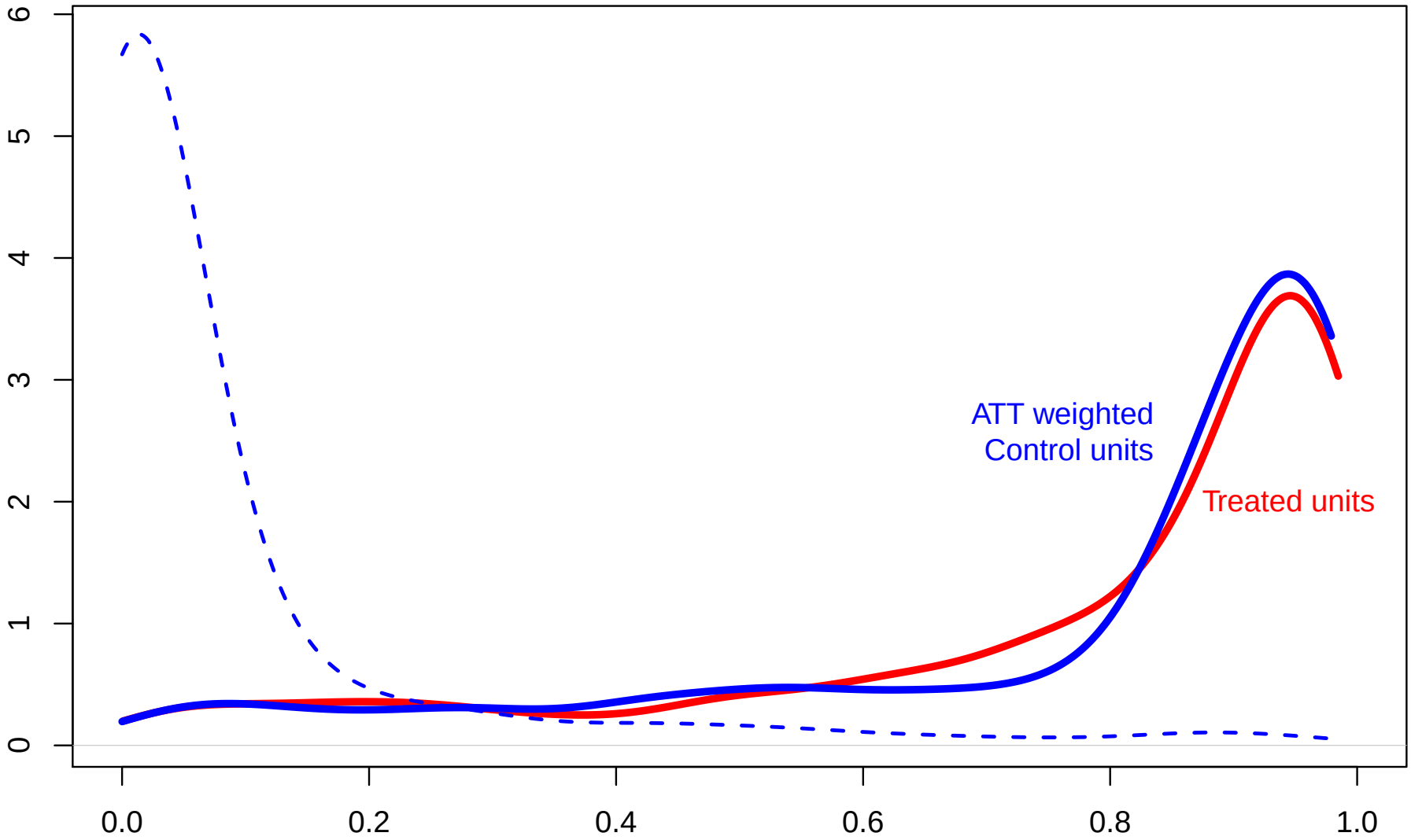
Weighting Both Groups to Balance Covariates

- Balancing condition: $\mathbb{E}\left\{\frac{T_i X_i}{\pi(X_i)} - \frac{(1-T_i)X_i}{1-\pi(X_i)}\right\} = 0$



Weighting Control Group to Balance Covariates

- Balancing condition: $\mathbb{E}\left\{T_i X_i - \frac{\pi(X_i)(1-T_i)X_i}{1-\pi(X_i)}\right\} = 0$



Efficient Doubly-Robust Estimators

- The estimator by Robins *et al*. :

$$\begin{aligned}\hat{\tau}_{DR} \equiv & \left\{ \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(1, X_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{T_i(Y_i - \hat{\mu}(1, X_i))}{\hat{\pi}(X_i)} \right\} \\ & - \left\{ \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(0, X_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - T_i)(Y_i - \hat{\mu}(0, X_i))}{1 - \hat{\pi}(X_i)} \right\}\end{aligned}$$

- Consistent if either the propensity score model or the outcome model is correct
- (Semiparametrically) Efficient
- FURTHER READING: Lunceford and Davidian (2004, *Stat. in Med.*)

Propensity Score Tautology

- Propensity score is unknown
- Dimension reduction is purely theoretical: must model T_i given X_i
- Diagnostics: covariate balance checking
- In practice, adhoc specification searches are conducted
- Model misspecification is always possible

- Theory (Rubin *et al.*): ellipsoidal covariate distributions
  $\Longrightarrow$ equal percent bias reduction
- Skewed covariates are common in applied settings

- Propensity score methods can be sensitive to misspecification

Kang and Schafer (2007, *Statistical Science*)

- Simulation study: the deteriorating performance of propensity score weighting methods when the model is misspecified
- Setup:
  - 4 covariates X_i^*: all are *i.i.d.* standard normal
  - Outcome model: linear model
  - Propensity score model: logistic model with linear predictors
  - Misspecification induced by measurement error:
    - $X_{i1} = \exp(X_{i1}^*/2)$
    - $X_{i2} = X_{i2}^*/(1 + \exp(X_{1i}^*) + 10)$
    - $X_{i3} = (X_{i1}^* X_{i3}^*/25 + 0.6)^3$
    - $X_{i4} = (X_{i1}^* + X_{i4}^* + 20)^2$
- Weighting estimators to be evaluated:
  1. Horvitz-Thompson
  2. Inverse-probability weighting with normalized weights
  3. Weighted least squares regression
  4. Doubly-robust least squares regression

Weighting Estimators Do Fine If the Model is Correct

		Bias		RMSE	
Sample size	Estimator	logit	True	logit	True
(1) Both models correct					
$n = 200$	HT	0.33	1.19	12.61	23.93
	IPW	−0.13	−0.13	3.98	5.03
	WLS	−0.04	−0.04	2.58	2.58
	DR	−0.04	−0.04	2.58	2.58
$n = 1000$	HT	0.01	−0.18	4.92	10.47
	IPW	0.01	−0.05	1.75	2.22
	WLS	0.01	0.01	1.14	1.14
	DR	0.01	0.01	1.14	1.14
(2) Propensity score model correct					
$n = 200$	HT	−0.05	−0.14	14.39	24.28
	IPW	−0.13	−0.18	4.08	4.97
	WLS	0.04	0.04	2.51	2.51
	DR	0.04	0.04	2.51	2.51
$n = 1000$	HT	−0.02	0.29	4.85	10.62
	IPW	0.02	−0.03	1.75	2.27
	WLS	0.04	0.04	1.14	1.14
	DR	0.04	0.04	1.14	1.14

Weighting Estimators are Sensitive to Misspecification

		Bias		RMSE	
Sample size	Estimator	logit	True	logit	True
(3) Outcome model correct					
$n = 200$	HT	24.25	−0.18	194.58	23.24
	IPW	1.70	−0.26	9.75	4.93
	WLS	−2.29	0.41	4.03	3.31
	DR	−0.08	−0.10	2.67	2.58
$n = 1000$	HT	41.14	−0.23	238.14	10.42
	IPW	4.93	−0.02	11.44	2.21
	WLS	−2.94	0.20	3.29	1.47
	DR	0.02	0.01	1.89	1.13
(4) Both models incorrect					
$n = 200$	HT	30.32	−0.38	266.30	23.86
	IPW	1.93	−0.09	10.50	5.08
	WLS	−2.13	0.55	3.87	3.29
	DR	−7.46	0.37	50.30	3.74
$n = 1000$	HT	101.47	0.01	2371.18	10.53
	IPW	5.16	0.02	12.71	2.25
	WLS	−2.95	0.37	3.30	1.47
	DR	−48.66	0.08	1370.91	1.81

Covariate Balancing Propensity Score

- IMAI AND RATKOVIC (2014; JRSSB)
- Recall the dual characteristics of propensity score
  1. Conditional probability of treatment assignment
  2. Covariate balancing score

- Implied moment conditions:
  1. Score equation:

$$\mathbb{E}\left\{\frac{T_i \pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1-T_i)\pi'_\beta(X_i)}{1-\pi_\beta(X_i)}\right\} = 0$$

  2. Balancing condition:

$$\mathbb{E}\left\{\frac{T_i \widetilde{X}_i}{\pi_\beta(X_i)} - \frac{(1-T_i)\widetilde{X}_i}{1-\pi_\beta(X_i)}\right\} = 0$$

  where $\widetilde{X}_i = f(X_i)$ is any vector-valued function
- Use generalized method of moments for estimation
- Estimate β to minimize imbalance

Revisiting Kang and Schafer (2007)

		Bias			RMSE		
Sample size	Estimator	logit	CBPS	True	logit	CBPS	True
(1) Both models correct							
$n = 200$	HT	−0.01	0.73	0.68	13.07	4.04	23.72
	IPW	−0.09	−0.09	−0.11	4.01	3.23	4.90
	WLS	0.03	0.03	0.03	2.57	2.57	2.57
	DR	0.03	0.03	0.03	2.57	2.57	2.57
$n = 1000$	HT	−0.03	0.15	0.29	4.86	1.80	10.52
	IPW	−0.02	−0.03	−0.01	1.73	1.45	2.25
	WLS	−0.00	−0.00	−0.00	1.14	1.14	1.14
	DR	−0.00	−0.00	−0.00	1.14	1.14	1.14
(2) Propensity score model correct							
$n = 200$	HT	−0.32	0.55	−0.17	12.49	4.06	23.49
	IPW	−0.27	−0.26	−0.35	3.94	3.27	4.90
	WLS	−0.07	−0.07	−0.07	2.59	2.59	2.59
	DR	−0.07	−0.07	−0.07	2.59	2.59	2.59
$n = 1000$	HT	0.03	0.15	0.01	4.93	1.79	10.62
	IPW	−0.02	−0.03	−0.04	1.76	1.46	2.26
	WLS	−0.01	−0.01	−0.01	1.14	1.14	1.14
	DR	−0.01	−0.01	−0.01	1.14	1.14	1.14

CBPS Makes Weighting Methods Work Better

		Bias				RMSE			
	Estimator	logit	CBPS1	CBPS2	True	logit	CBPS1	CBPS2	True
(3) Outcome model correct									
$n = 200$	HT	24.25	1.09	−5.42	−0.18	194.58	5.04	10.71	23.24
	IPW	1.70	−1.37	−2.84	−0.26	9.75	3.42	4.74	4.93
	WLS	−2.29	−2.37	−2.19	0.41	4.03	4.06	3.96	3.31
	DR	−0.08	−0.10	−0.10	−0.10	2.67	2.58	2.58	2.58
$n = 1000$	HT	41.14	−2.02	2.08	−0.23	238.14	2.97	6.65	10.42
	IPW	4.93	−1.39	−0.82	−0.02	11.44	2.01	2.26	2.21
	WLS	−2.94	−2.99	−2.95	0.20	3.29	3.37	3.33	1.47
	DR	0.02	0.01	0.01	0.01	1.89	1.13	1.13	1.13
(4) Both models incorrect									
$n = 200$	HT	30.32	1.27	−5.31	−0.38	266.30	5.20	10.62	23.86
	IPW	1.93	−1.26	−2.77	−0.09	10.50	3.37	4.67	5.08
	WLS	−2.13	−2.20	−2.04	0.55	3.87	3.91	3.81	3.29
	DR	−7.46	−2.59	−2.13	0.37	50.30	4.27	3.99	3.74
$n = 1000$	HT	101.47	−2.05	1.90	0.01	2371.18	3.02	6.75	10.53
	IPW	5.16	−1.44	−0.92	0.02	12.71	2.06	2.39	2.25
	WLS	−2.95	−3.01	−2.98	0.19	3.30	3.40	3.36	1.47
	DR	−48.66	−3.59	−3.79	0.08	1370.91	4.02	4.25	1.81

Matching Representation of Difference-in-Means

Units	**1**	**2**	**3**	**4**	**5**
Treatment status	**T**	**T**	**C**	**C**	**T**
Outcome	Y_1	Y_2	Y_3	Y_4	Y_5

- Estimating the Average Treatment Effect (ATE) via matching:

$$
\begin{aligned}
Y_1 &- \frac{1}{2}(Y_3 + Y_4) \\
Y_2 &- \frac{1}{2}(Y_3 + Y_4) \\
\frac{1}{3}(Y_1 + Y_2 + Y_5) &- Y_3 \\
\frac{1}{3}(Y_1 + Y_2 + Y_5) &- Y_4 \\
Y_5 &- \frac{1}{2}(Y_3 + Y_4)
\end{aligned}
$$

Matching Representation of Simple Regression

- Simple linear regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- Binary treatment: $X_i \in \{0, 1\}$
- Equivalent matching estimator:

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y_i(1)} - \widehat{Y_i(0)} \right)$$

where

$$\widehat{Y_i(1)} = \begin{cases} Y_i & \text{if } X_i = 1 \\ \frac{1}{\sum_{i'=1}^{N} X_{i'}} \sum_{i'=1}^{N} X_{i'} Y_{i'} & \text{if } X_i = 0 \end{cases}$$

$$\widehat{Y_i(0)} = \begin{cases} \frac{1}{\sum_{i'=1}^{N} (1 - X_{i'})} \sum_{i'=1}^{N} (1 - X_{i'}) Y_{i'} & \text{if } X_i = 1 \\ Y_i & \text{if } X_i = 0 \end{cases}$$

- Treated units matched with the average of non-treated units

One-Way Fixed Effects Regression

- Simple (one-way) FE model:

$$Y_{it} = \alpha_i + \beta X_{it} + \epsilon_{it}$$

- Commonly used by applied researchers:
  - Stratified randomized experiments (Duflo *et al.* 2007)
  - Stratification and matching in observational studies
  - Panel data, both experimental and observational

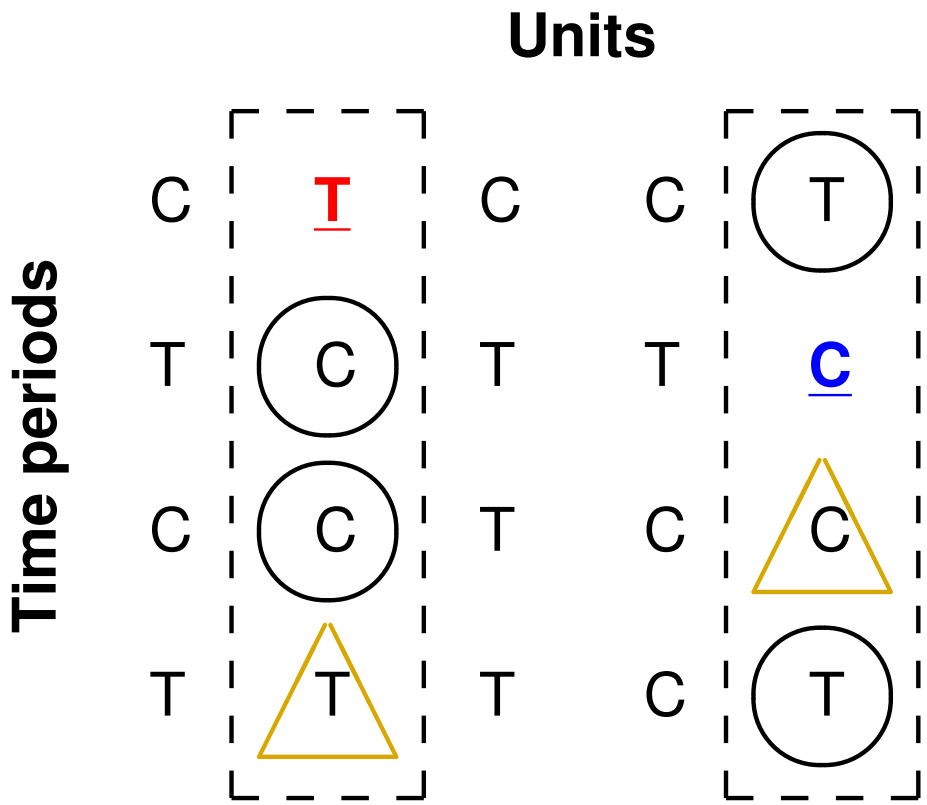
- $\hat{\beta}_{FE}$ may be biased for the ATE even if X_{it} is exogenous within each unit
- It converges to the weighted average of conditional ATEs:

$$\hat{\beta}_{FE} \xrightarrow{p} \frac{\mathbb{E}\{\text{ATE}_i \, \sigma_i^2\}}{\mathbb{E}(\sigma_i^2)}$$

  where $\sigma_i^2 = \sum_{t=1}^{T} (X_{it} - \overline{X}_i)^2 / T$

- How are counterfactual outcomes estimated under the FE model?
- Unit fixed effects $\Longrightarrow$ within-unit comparison

Mismatches in One-Way Fixed Effects Model



- T: treated observations
- C: control observations
- **Circles**: Proper matches
- **Triangles**: "Mismatches" $\Longrightarrow$ attenuation bias

Matching Representation of Fixed Effects Regression

Proposition 1

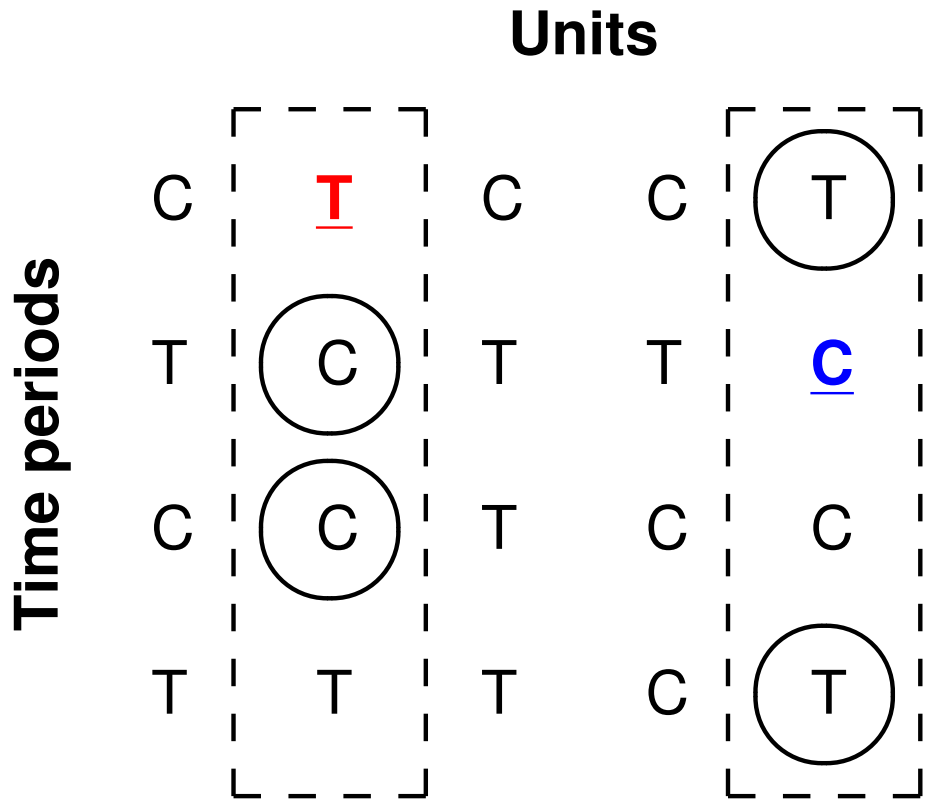
$$\hat{\beta}^{FE} = \frac{1}{K}\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right)\right\},$$

$$\widehat{Y_{it}(x)} = \begin{cases} Y_{it} & \text{if } X_{it} = x \\ \frac{1}{T-1}\sum_{t' \neq t} Y_{it'} & \text{if } X_{it} = 1 - x \end{cases} \quad \text{for} \quad x = 0, 1$$

$$K = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it} \cdot \frac{1}{T-1}\sum_{t' \neq t}(1 - X_{it'}) + (1 - X_{it}) \cdot \frac{1}{T-1}\sum_{t' \neq t} X_{it'}\right\}.$$

- K: average proportion of proper matches across all observations
- More mismatches $\Longrightarrow$ larger adjustment
- Adjustment is required except very special cases
- "Fixes" attenuation bias but this adjustment is not sufficient
- Fixed effects estimator is a special case of matching estimators

Unadjusted Matching Estimator



- Consistent if the treatment is exogenous within each unit
- Only equal to fixed effects estimator if heterogeneity in either treatment assignment or treatment effect is non-existent

Unadjusted Matching = **Weighted** FE Estimator

Proposition 2

The unadjusted matching estimator

$$\hat{\beta}^M = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right)$$

where

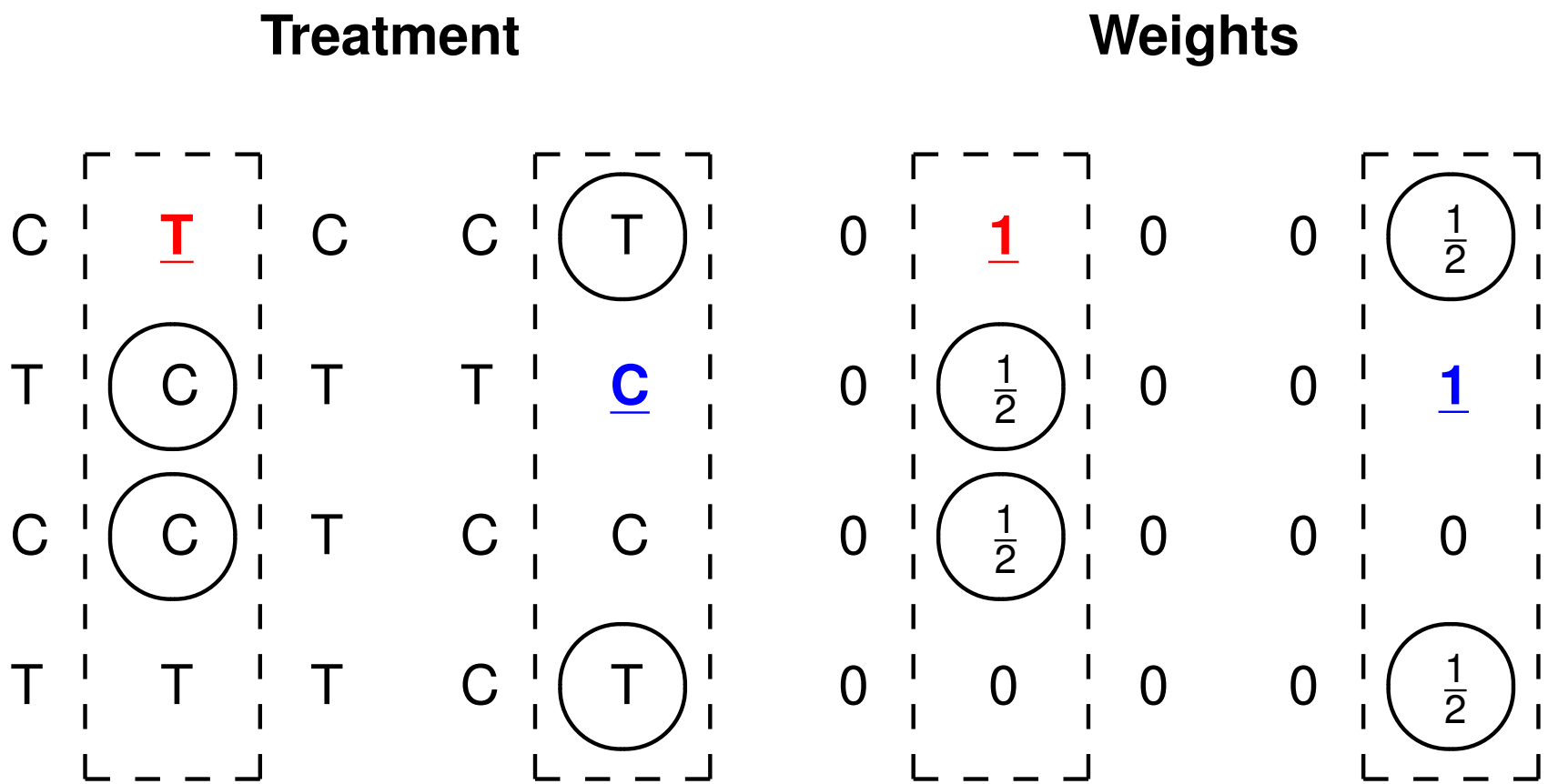
$$\widehat{Y_{it}(1)} = \begin{cases} Y_{it} & \text{if } X_{it}=1 \\ \frac{\sum_{t'=1}^{T} X_{it'} Y_{it'}}{\sum_{t'=1}^{T} X_{it'}} & \text{if } X_{it}=0 \end{cases} \quad \text{and} \quad \widehat{Y_{it}(0)} = \begin{cases} \frac{\sum_{t'=1}^{T}(1-X_{it'})Y_{it'}}{\sum_{t'=1}^{T}(1-X_{it'})} & \text{if } X_{it}=1 \\ Y_{it} & \text{if } X_{it}=0 \end{cases}$$

is equivalent to the weighted fixed effects model

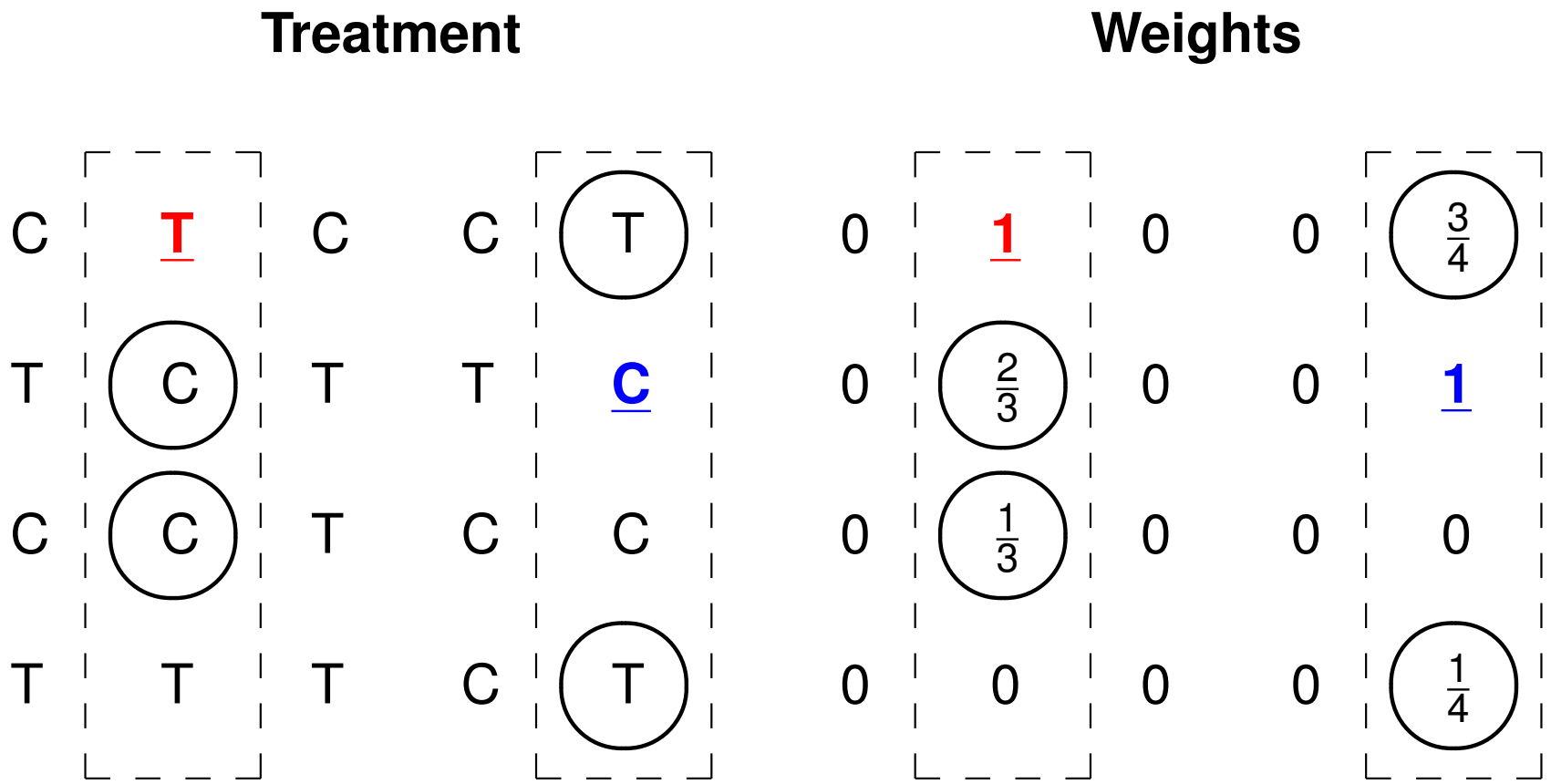
$$(\hat{\alpha}^M, \hat{\beta}^M) = \underset{(\alpha,\beta)}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(Y_{it} - \alpha_i - \beta X_{it})^2$$

$$W_{it} \equiv \begin{cases} \frac{T}{\sum_{t'=1}^{T} X_{it'}} & \text{if} \quad X_{it}=1, \\ \frac{T}{\sum_{t'=1}^{T}(1-X_{it'})} & \text{if} \quad X_{it}=0. \end{cases}$$

Equal Weights



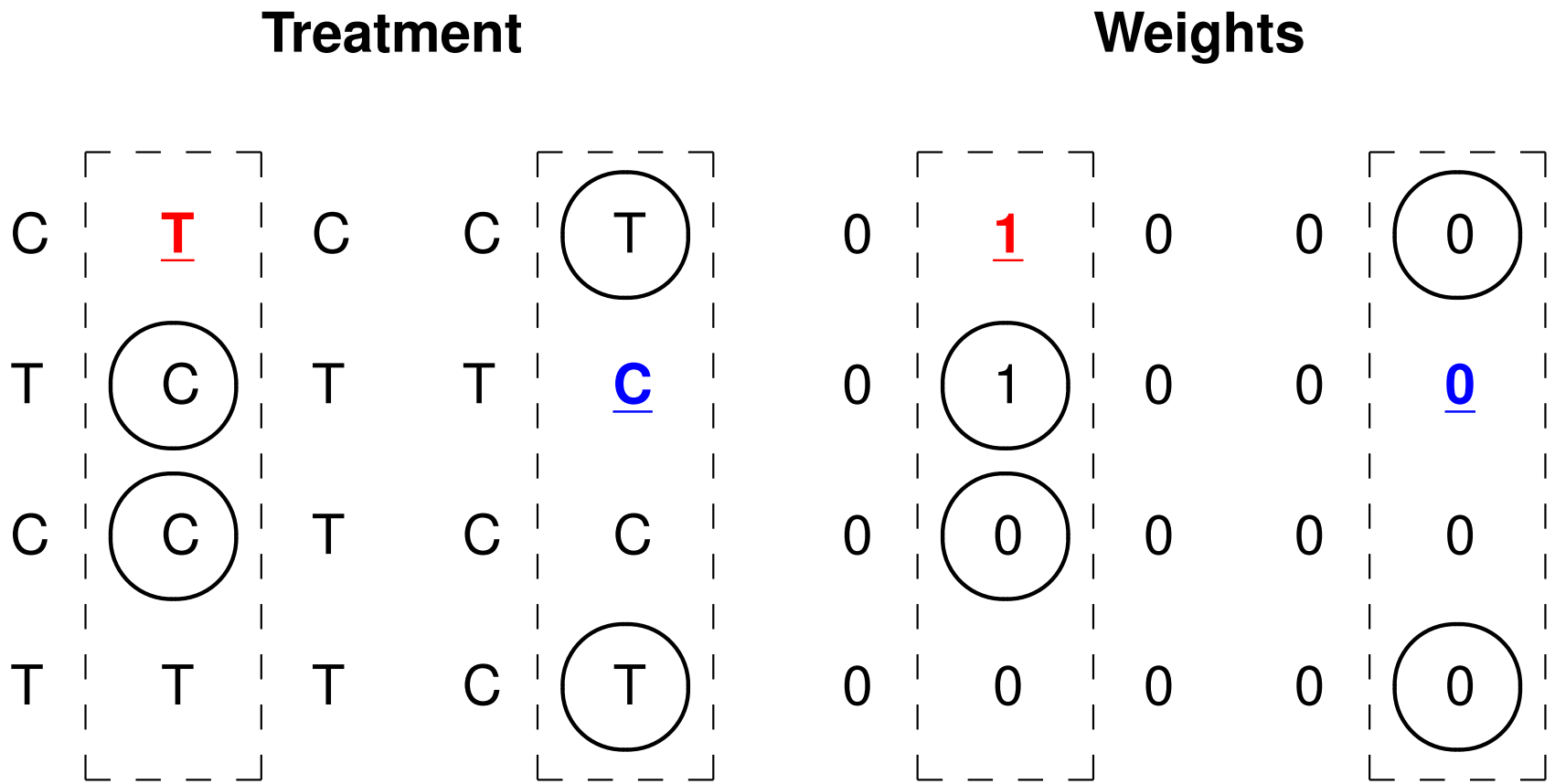
Different Weights



- Any within-unit matching estimator leads to weighted fixed effects regression with particular weights
- Can derive regression weights given *any* matching estimator for various quantities (ATE, ATT, etc.)

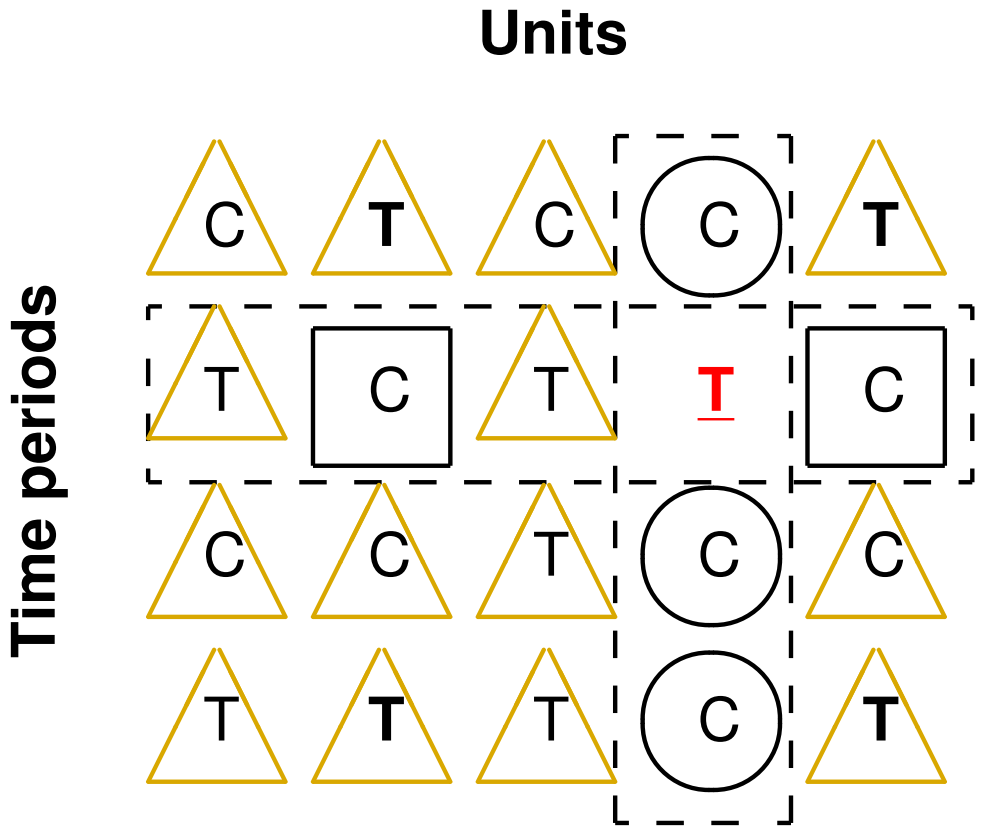
First Difference = Matching = Weighted One-Way FE

- $\Delta Y_{it} = \beta \Delta X_{it} + \epsilon_{it}$ where $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$, $\Delta X_{it} = X_{it} - X_{i,t-1}$



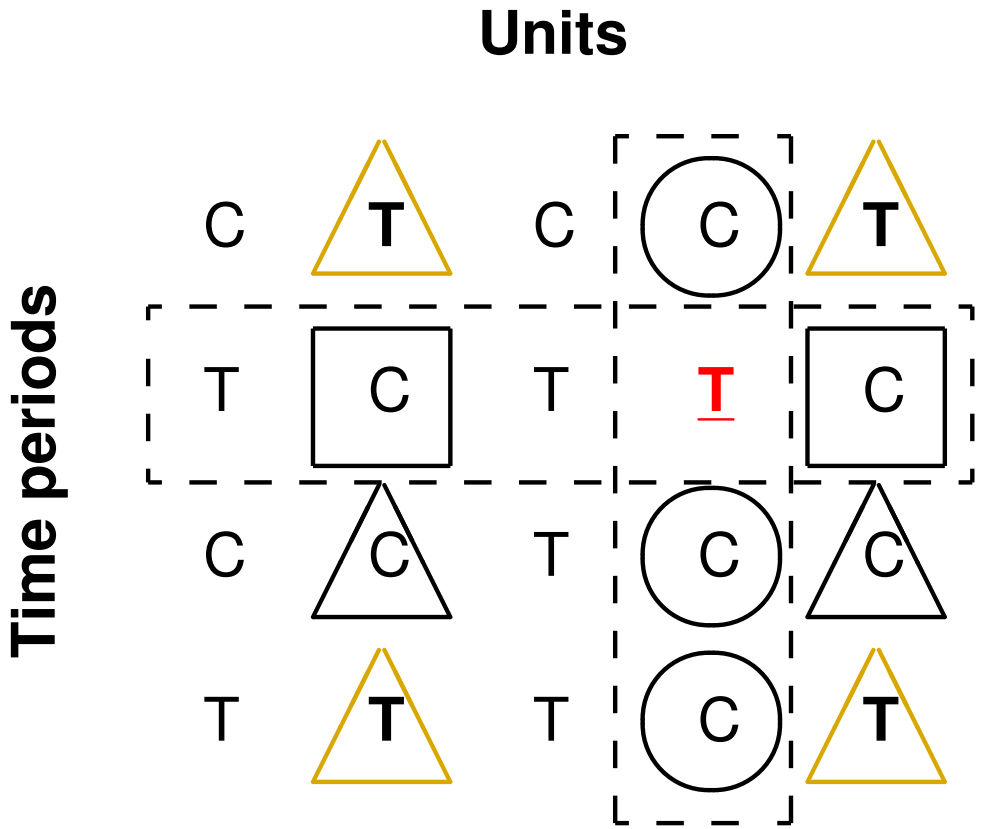
Mismatches in Two-Way FE Model

$$Y_{it} = \alpha_i + \gamma_t + \beta X_{it} + \epsilon_{it}$$



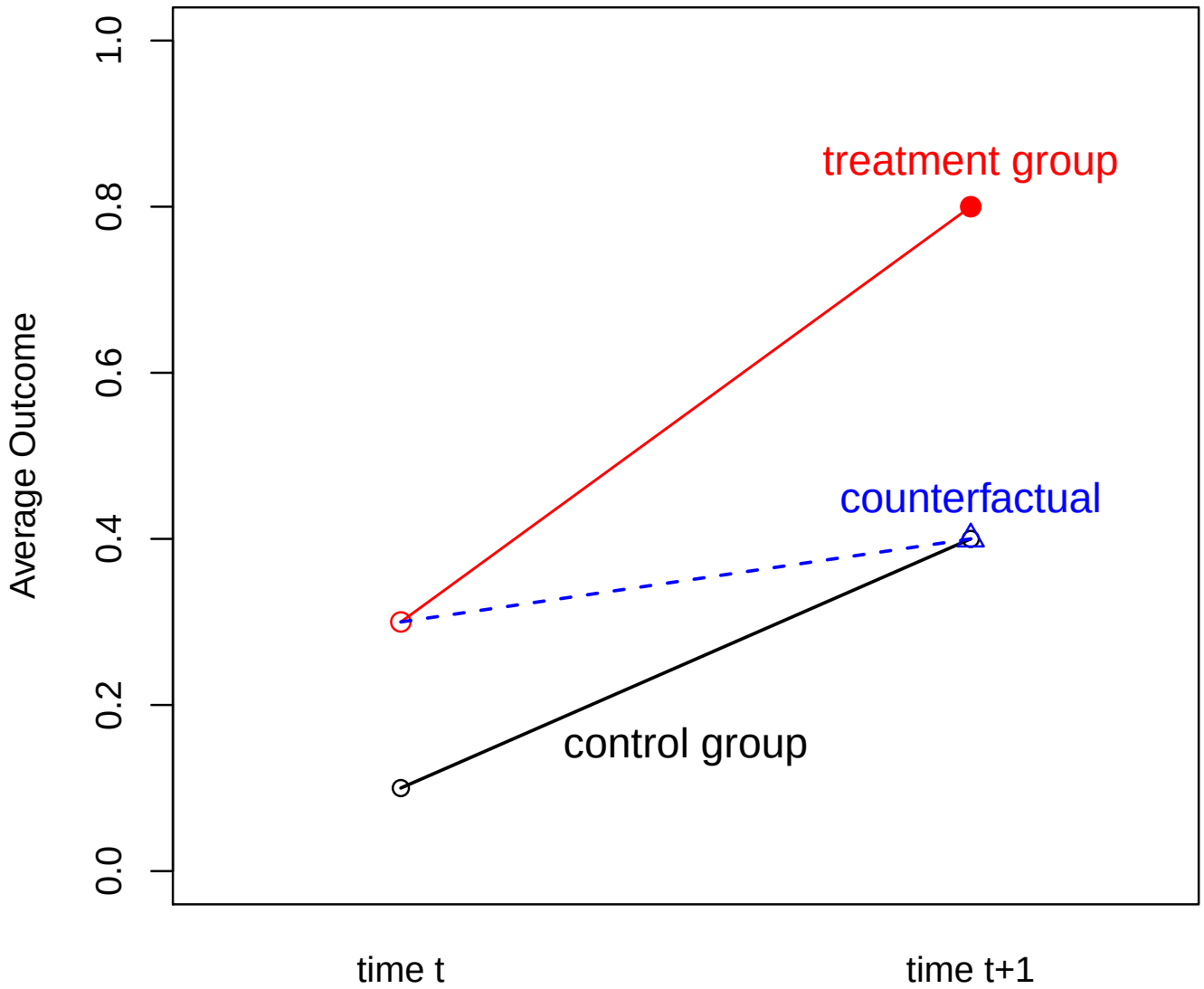
- **Triangles**: Two kinds of mismatches
  - Same treatment status
  - Neither same unit nor same time

Mismatches in Weighted Two-Way FE Model

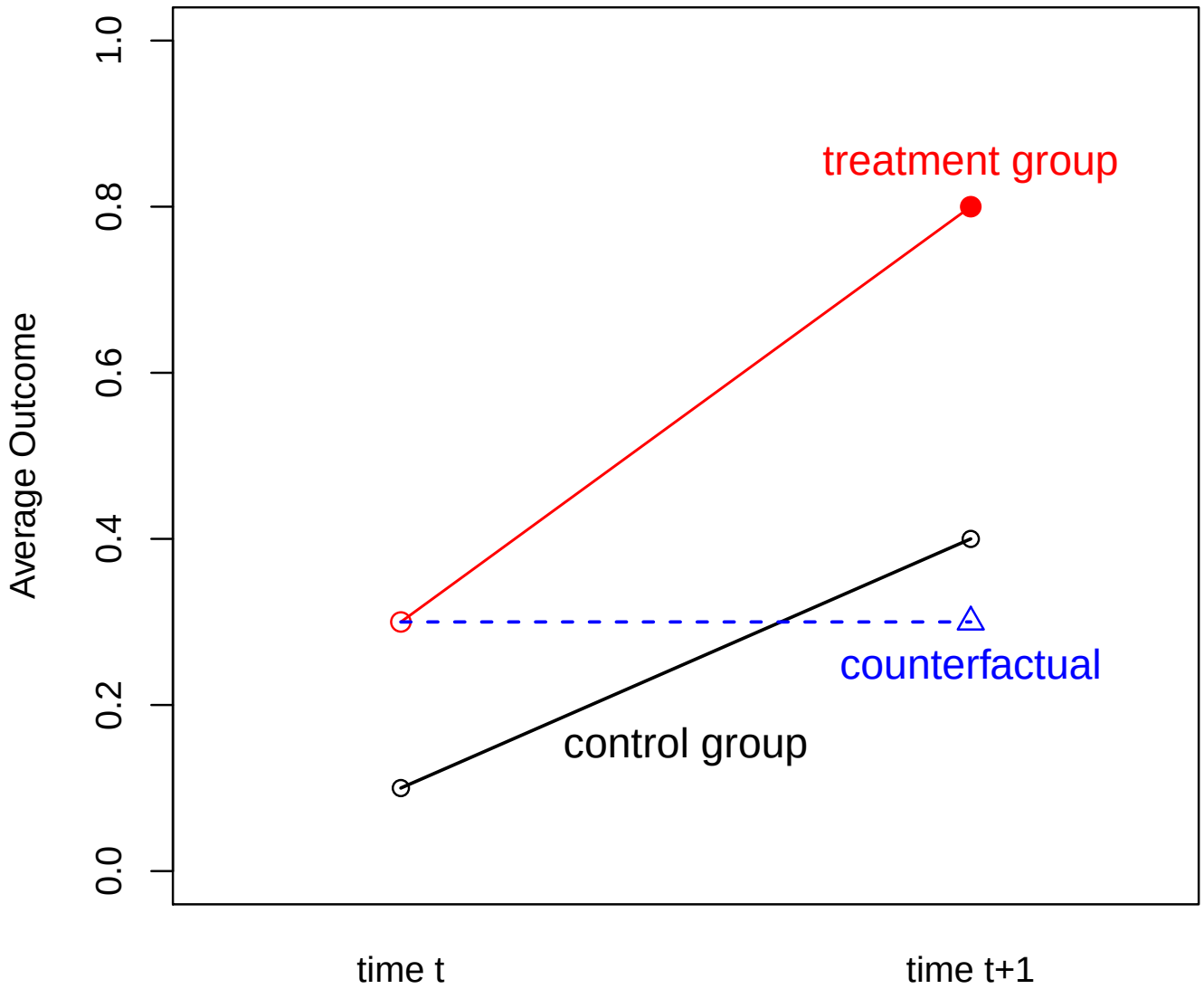


- Some mismatches can be eliminated
- You can NEVER eliminate them all

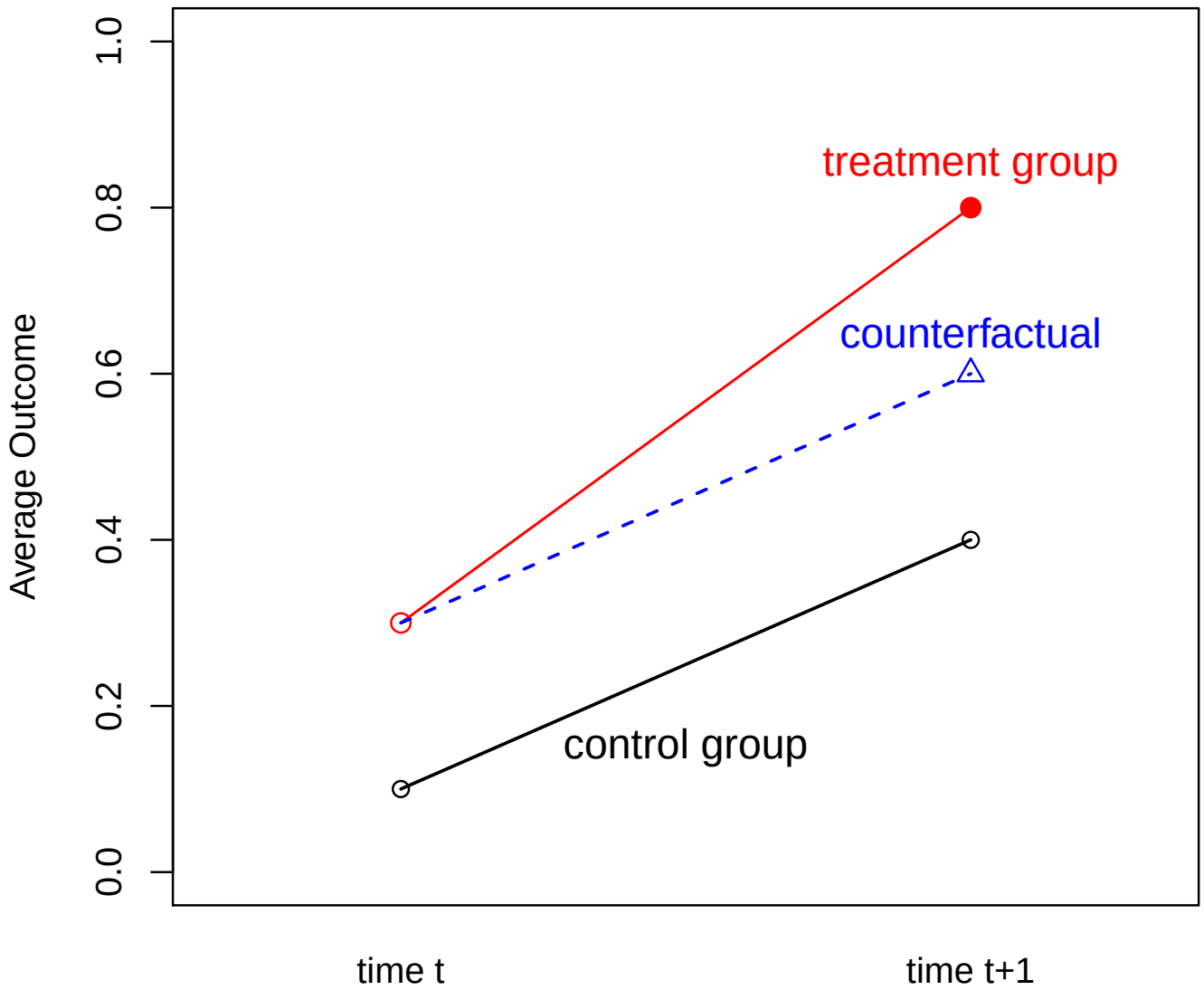
Cross Section Analysis = Weighted **Time** FE Model



First Difference = Weighted **Unit** FE Model



What about Difference-in-Differences (DiD)?



Two-Way Fixed Effects = DiD in 2 Time-Period Case

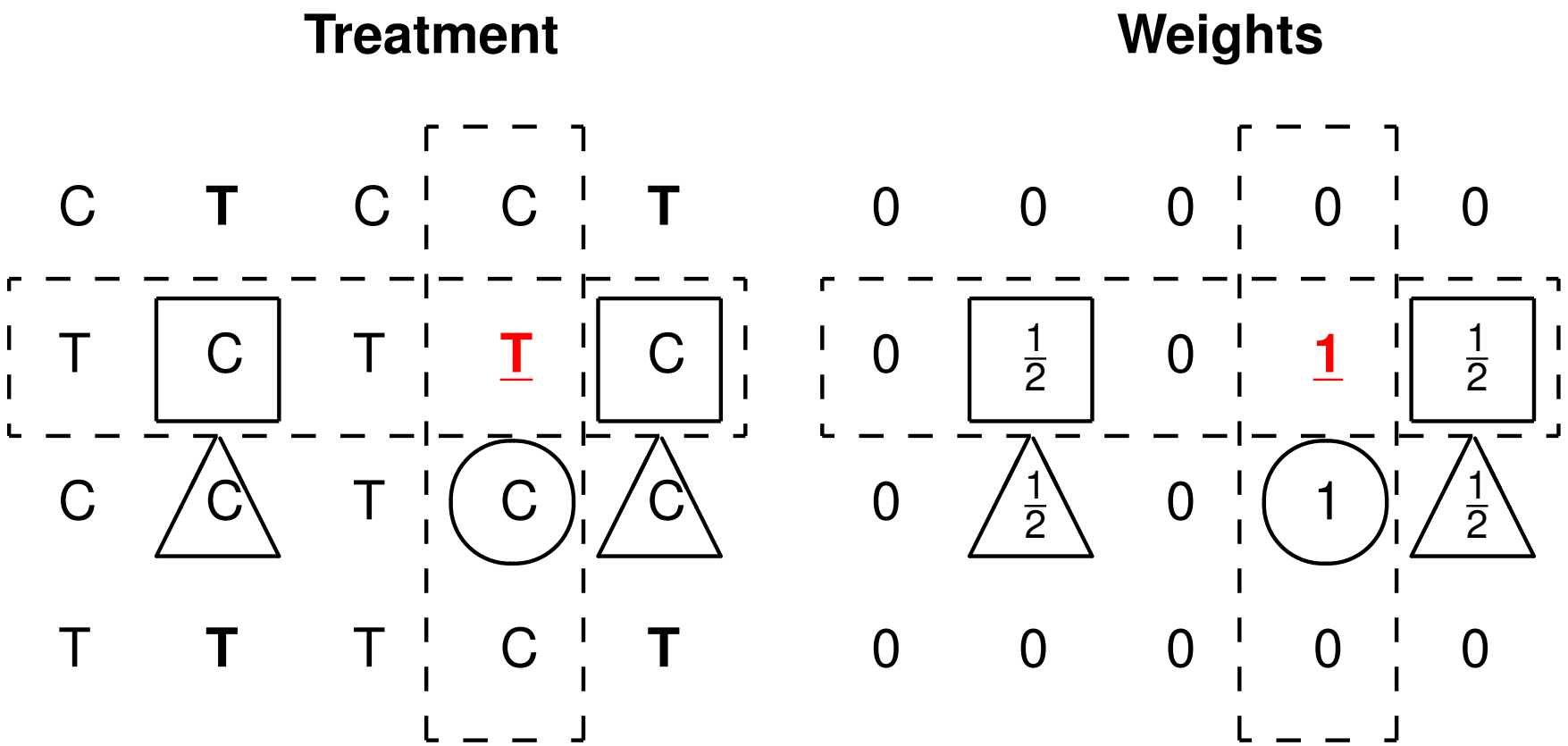
- Two-way fixed effects model:

$$Y_{ti}(z) = \alpha_i + \beta z + \gamma t + \epsilon_{it}$$

  where $z = 0, 1$ is the treatment status and t is the time
- The model implies:
  - $Y_{i0}(0) = \alpha_i + \epsilon_{i0}$
  - $Y_{i1}(0) = \alpha_i + \gamma + \epsilon_{i1}$
  - $Y_{i1}(1) = \alpha_i + \beta + \gamma + \epsilon_{i1}$
- Assumption: $\mathbb{E}(Y_{i1}(0) - Y_{i0}(0) \mid Z_{i1} = z) = \gamma$
- Or equivalently $\mathbb{E}(\epsilon_{i1} - \epsilon_{i0} \mid Z_{i1} = z) = 0$
- Both Z_{it} and ϵ_{it} can depend on α_i
- Neither stronger or weaker than the standard exogeneity assumption
- When $Y_{i0} = Y_{i0}(0)$ is balanced, they are equivalent

General DiD = Weighted Two-Way (Unit and Time) FE

- General setting: Multiple time periods, repeated treatments
- Standard two-way fixed effects $\neq$ DiD



- Weights can be negative $\Longrightarrow$ the method of moments estimator
- Fast computation is available

Effects of GATT Membership on International Trade

1. Controversy
   - Rose (2004): No effect of GATT membership on trade
   - Tomz et al. (2007): Significant effect with non-member participants

2. The central role of fixed effects models:
   - Rose (2004): one-way (year) fixed effects for dyadic data
   - Tomz *et al.* (2007): two-way (year and dyad) fixed effects
   - Rose (2005): "I follow the profession in placing most confidence in the fixed effects estimators; I have no clear ranking between country-specific and country pair-specific effects."
   - Tomz *et al.* (2007): "We, too, prefer FE estimates over OLS on both theoretical and statistical ground"

Data and Methods

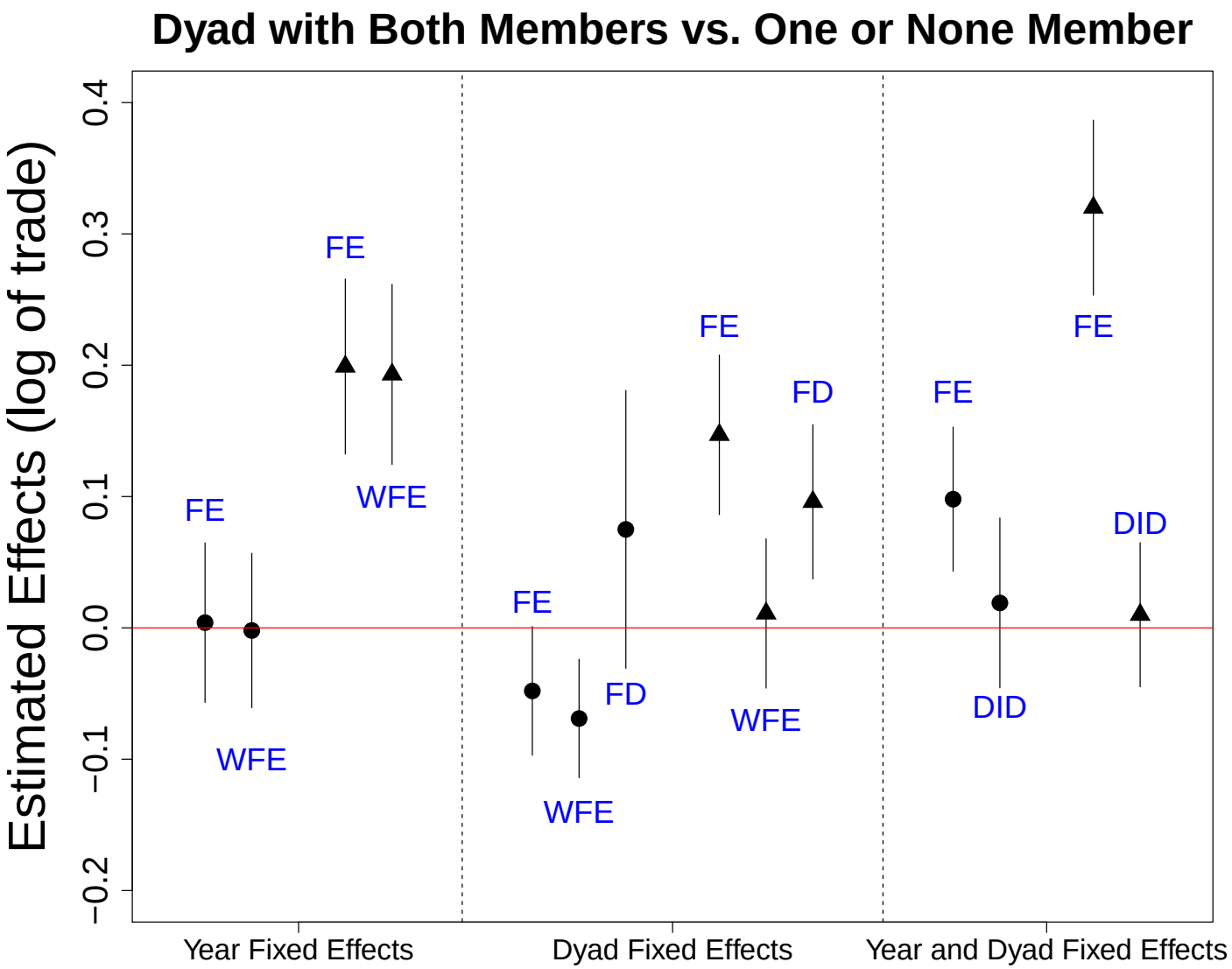
1. Data
   - Data set from Tomz et al. (2007)
   - Effect of GATT: 1948 – 1994
   - 162 countries, and 196,207 (dyad-year) observations
2. Year fixed effects model:

$$\ln Y_{it} = \alpha_t + \beta X_{it} + \delta^\top Z_{it} + \epsilon_{it}$$

   - X_{it}: membership (formal/participants) Both vs. One/None
   - Z_{it}: 15 dyad-varying covariates (e.g., log product GDP)
3. Weighted one-way fixed effects model:

$$\underset{(\alpha,\beta,\delta)}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} (\ln Y_{it} - \alpha_t - \beta X_{it} - \delta^\top Z_{it})^2$$

Empirical Results



Concluding Remarks

- Matching and weighting methods do:
  - make causal assumptions transparent by identifying counterfactuals
  - make regression models robust by reducing model dependence
- Matching and weighting methods cannot solve endogeneity
- Only good research design can overcome endogeneity
- Recent advances in matching and weighting methods: directly optimize balance
- Next methodological challenges: panel data
  - Fixed effects regression assumes no carry-over effect
  - They do not model dynamic treatment regimes

Causal Mechanisms

Identifying Causal Mechanisms

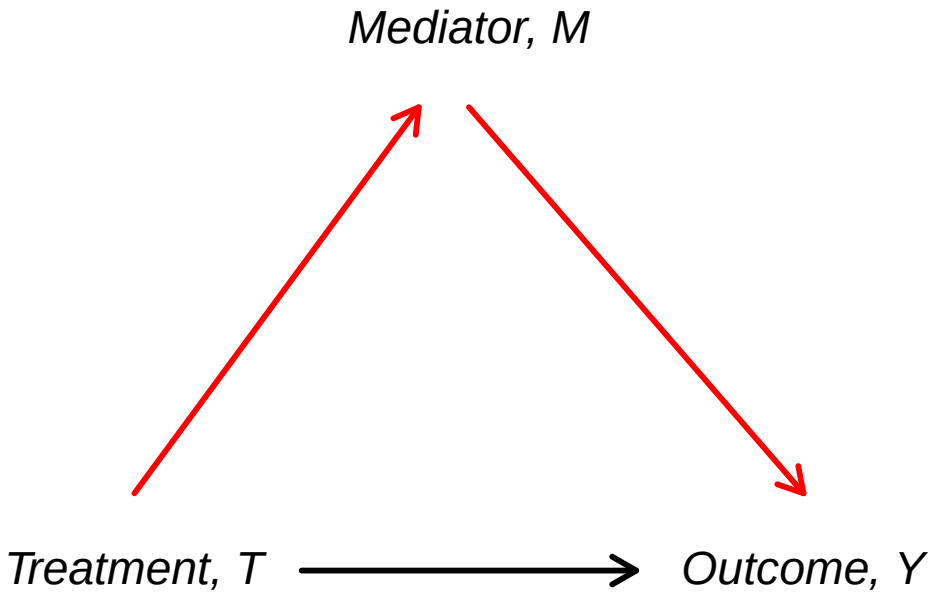
- Randomized experiments as **gold standard** for causal inference
- But, experiments are a **black box**
- Can only tell *whether* the treatment causally affects the outcome
- Not *how* and *why* the treatment affects the outcome
- Qualitative research uses process tracing

- Question: How can we learn about causal mechanisms from experimental and observational studies?

- IMAI, KEELE, TINGLEY, AND YAMAMOTO (2011) UNPACKING THE BLACK BOX OF CAUSALITY. *American Political Science Review*

Causal Mediation Analysis

- Graphical representation



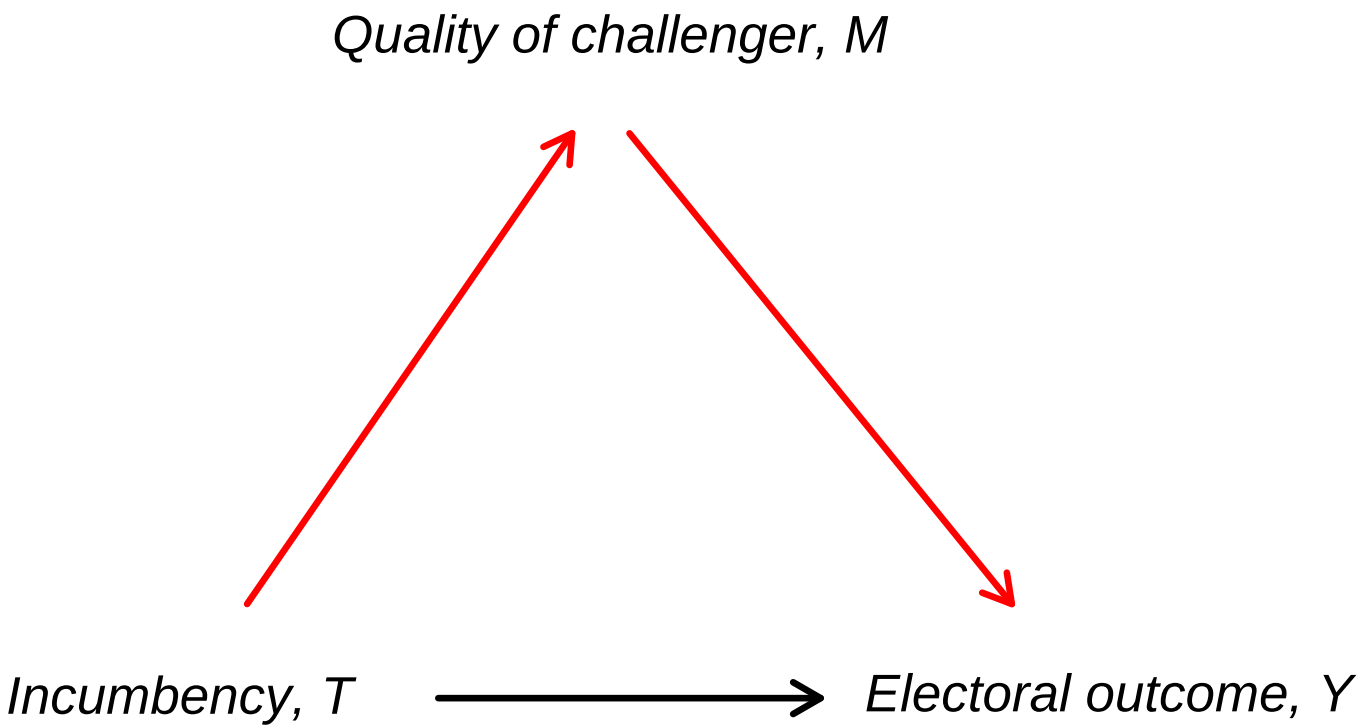
- Goal is to decompose total effect into direct and indirect effects
- Alternative approach: decompose the treatment into different components
- Causal mediation analysis as quantitative process tracing

Decomposition of Incumbency Advantage

- Incumbency effects: one of the most studied topics in American politics
- Consensus emerged in 1980s: incumbency advantage is positive and growing in magnitude

- New direction in 1990s: Where does incumbency advantage come from?
- Scare-off/quality effect (Cox and Katz): the ability of incumbents to deter high-quality challengers from entering the race
- Alternative causal mechanisms: name recognition, campaign spending, personal vote, television, etc.

Causal Mediation Analysis in Cox and Katz



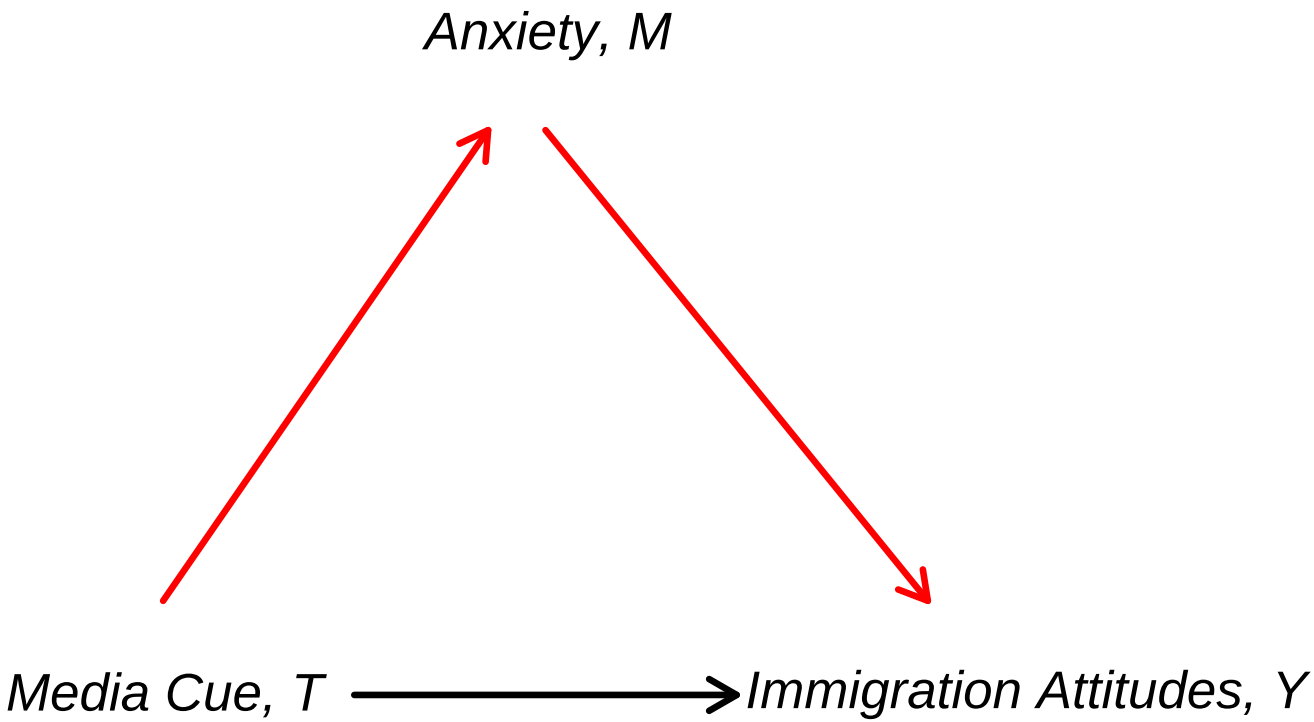
- How much of incumbency advantage can be explained by scare-off/quality effect?
- How large is the mediation effect relative to the total effect?

Psychological Study of Media Effects

- Large literature on how media influences public opinion
- A media framing experiment of Brader *et al.*:
  1. (White) Subjects read a mock news story about immigration:
     - Treatment: Hispanic immigrant in the story
     - Control: European immigrant in the story
  2. Measure attitudinal and behavioral outcome variables:
     - Opinions about increasing or decrease immigration
     - Contact legislator about the issue
     - Send anti-immigration message to legislator
- Why is group-based media framing effective?: role of emotion
- Hypothesis: Hispanic immigrant increases anxiety, leading to greater opposition to immigration

- The primary goal is to examine how, not whether, media framing shapes public opinion

Causal Mediation Analysis in Brader *et al.*



- Does the media framing shape public opinion by making people anxious?
- An alternative causal mechanism: change in beliefs
- Can we identify mediation effects from randomized experiments?

The Standard Estimation Method

- Linear models for mediator and outcome:

$$
\begin{aligned}
Y_i &= \alpha_1 + \beta_1 T_i + \xi_1^\top X_i + \epsilon_{1i} \\
M_i &= \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{2i} \\
Y_i &= \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{3i}
\end{aligned}
$$

  where X_i is a set of pre-treatment or control variables
  1. Total effect (ATE) is β_1
  2. Direct effect is β_3
  3. Indirect or mediation effect is $\beta_2\gamma$
  4. Effect decomposition: $\beta_1 = \beta_3 + \beta_2\gamma$.

- Some motivating questions:
  1. What should we do when we have interaction or nonlinear terms?
  2. What about other models such as logit?
  3. In general, under what conditions can we interpret β_1 and $\beta_2\gamma$ as causal effects?
  4. What do we really mean by causal mediation effect anyway?

Potential Outcomes Framework of Causal Inference

- Observed data:
  - Binary treatment: $T_i \in \{0, 1\}$
  - Mediator: $M_i \in \mathcal{M}$
  - Outcome: $Y_i \in \mathcal{Y}$
  - Observed pre-treatment covariates: $X_i \in \mathcal{X}$
- Potential outcomes model (Neyman, Rubin):
  - Potential mediators: $M_i(t)$ where $M_i = M_i(T_i)$
  - Potential outcomes: $Y_i(t, m)$ where $Y_i = Y_i(T_i, M_i(T_i))$
- Total causal effect:

$$\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0))$$

- **Fundamental problem of causal inference**: only one potential outcome can be observed for each i

Back to the Examples

- $M_i(1)$:
  1. Quality of her challenger if politician i is an incumbent
  2. Level of anxiety individual i would report if he reads the story with Hispanic immigrant

- $Y_i(1, M_i(1))$:
  1. Election outcome that would result if politician i is an incumbent and faces a challenger whose quality is $M_i(1)$
  2. Immigration attitude individual i would report if he reads the story with Hispanic immigrant and reports the anxiety level $M_i(1)$

- $M_i(0)$ and $Y_i(0, M_i(0))$ are the converse

Causal Mediation Effects

- Causal mediation (Indirect) effects:

$$\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0))$$

- Causal effect of the change in M_i on Y_i that would be induced by treatment
- Change the mediator from $M_i(0)$ to $M_i(1)$ while holding the treatment constant at t
- Represents the mechanism through M_i
- Zero treatment effect on mediator $\Longrightarrow$ Zero mediation effect

- Examples:
  1. Part of incumbency advantage that is due to the difference in challenger quality induced by incumbency status
  2. Difference in immigration attitudes that is due to the change in anxiety induced by the treatment news story

Total Effect = Indirect Effect + Direct Effect

- Direct effects:

$$\zeta_i(t) \equiv Y_i(1, M_i(t)) - Y_i(0, M_i(t))$$

- Causal effect of T_i on Y_i, holding mediator constant at its potential value that would realize when $T_i = t$
- Change the treatment from 0 to 1 while holding the mediator constant at $M_i(t)$
- Represents all mechanisms other than through M_i

- Total effect = mediation (indirect) effect + direct effect:

$$\tau_i = \delta_i(t) + \zeta_i(1-t) = \frac{1}{2}\{(\delta_i(0) + \zeta_i(0)) + (\delta_i(1) + \zeta_i(1))\}$$

Mechanisms, Manipulations, and Interactions

Mechanisms

- Indirect effects: $\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0))$
- Counterfactuals about treatment-induced mediator values

Manipulations

- Controlled direct effects: $\xi_i(t, m, m') \equiv Y_i(t, m) - Y_i(t, m')$
- Causal effect of directly manipulating the mediator under $T_i = t$

Interactions

- Interaction effects: $\xi(1, m, m') - \xi(0, m, m')$
- The extent to which controlled direct effects vary by the treatment

What Does the Observed Data Tell Us?

- Recall the Brader *et al.* experimental design:
  1. randomize T_i
  2. measure M_i and then Y_i

- Among observations with $T_i = t$, we observe $Y_i(t, M_i(t))$ but not $Y_i(t, M_i(1-t))$ unless $M_i(t) = M_i(1-t)$
- But we want to estimate

$$\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0))$$

- For $t = 1$, we observe $Y_i(1, M_i(1))$ but not $Y_i(1, M_i(0))$
- Similarly, for $t = 0$, we observe $Y_i(0, M_i(0))$ but not $Y_i(0, M_i(1))$

- We have the identification problem $\Longrightarrow$ Need assumptions or better research designs

Counterfactuals in the Examples

1. Incumbency advantage:
   - An incumbent ($T_i = 1$) faces a challenger with quality $M_i(1)$
   - We observe the electoral outcome $Y_i = Y_i(1, M_i(1))$
   - We also want $Y_i(1, M_i(0))$ where $M_i(0)$ is the quality of challenger this incumbent politician would face if she is not an incumbent

2. Media framing effects:
   - A subject viewed the news story with Hispanic immigrant ($T_i = 1$)
   - For this person, $Y_i(1, M_i(1))$ is the observed immigration opinion
   - $Y_i(1, M_i(0))$ is his immigration opinion in the counterfactual world where he still views the story with Hispanic immigrant but his anxiety is at the same level as if he viewed the control news story

In both cases, we can't observe $Y_i(1, M_i(0))$ because $M_i(0)$ is not realized when $T_i = 1$

Sequential Ignorability Assumption

- Proposed identification assumption: Sequential Ignorability (SI)

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x, \qquad (1)$$

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, X_i = x \qquad (2)$$

- In words,
  1. T_i is (as-if) randomized conditional on $X_i = x$
  2. $M_i(t)$ is (as-if) randomized conditional on $X_i = x$ and $T_i = t$

- Important limitations:
  1. In a standard experiment, (1) holds but (2) may not
  2. X_i needs to include all confounders
  3. X_i must be pre-treatment confounders $\Longrightarrow$ post-treatment confounder is not allowed
  4. Randomizing M_i via manipulation is not the same as assuming $M_i(t)$ is as-if randomized

Sequential Ignorability in the Standard Experiment

Back to Brader *et al.*:

- Treatment is randomized $\Longrightarrow$ (1) is satisfied
- But (2) may not hold:
  1. Pre-treatment confounder or X_i: state of residence
     those who live in AZ tend to have higher levels of perceived harm and be opposed to immigration
  2. Post-treatment confounder: alternative mechanism
     beliefs about the likely negative impact of immigration makes people anxious

- Pre-treatment confounders $\Longrightarrow$ measure and adjust for them
- Post-treatment confounders $\Longrightarrow$ adjusting is not sufficient

Nonparametric Identification

Under SI, both ACME and average direct effects are nonparametrically identified (can be consistently estimated without modeling assumption)

- ACME $\bar{\delta}(t)$

$$\int \int \mathbb{E}(Y_i \mid M_i, T_i = t, X_i) \{dP(M_i \mid T_i = 1, X_i) - dP(M_i \mid T_i = 0, X_i)\} dP(X_i)$$

- Average direct effects $\bar{\zeta}(t)$

$$\int \int \{\mathbb{E}(Y_i \mid M_i, T_i = 1, X_i) - \mathbb{E}(Y_i \mid M_i, T_i = 0, X_i)\} dP(M_i \mid T_i = t, X_i) dP(X_i)$$

Implies the general mediation formula under any statistical model

Traditional Estimation Methods: LSEM

- Linear structural equation model (LSEM):

$$\begin{aligned} M_i &= \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{i2}, \\ Y_i &= \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{i3}. \end{aligned}$$

- Fit two least squares regressions separately
- Use product of coefficients ($\hat{\beta}_2\hat{\gamma}$) to estimate ACME
- Use asymptotic variance to test significance (Sobel test)

- Under SI and the no-interaction assumption ($\bar{\delta}(1) \neq \bar{\delta}(0)$), $\hat{\beta}_2\hat{\gamma}$ consistently estimates ACME
- Can be extended to LSEM with interaction terms

- Problem: Only valid for the simplest LSEM

Popular Baron-Kenny Procedure

- The procedure:
  1. Regress Y on T and show a significant relationship
  2. Regress M on T and show a significant relationship
  3. Regress Y on M and T, and show a significant relationship between Y and M

- The problems:
  1. First step can lead to false negatives especially if indirect and direct effects in opposite directions
  2. The procedure only anticipates simplest linear models
  3. Don't do star-gazing. Report quantities of interest

Proposed General Estimation Algorithm

1. Model outcome and mediator
   - Outcome model: $p(Y_i \mid T_i, M_i, X_i)$
   - Mediator model: $p(M_i \mid T_i, X_i)$
   - These models can be of any form (linear or nonlinear, semi- or nonparametric, with or without interactions)
2. Predict mediator for both treatment values $(M_i(1), M_i(0))$
3. Predict outcome by first setting $T_i = 1$ and $M_i = M_i(0)$, and then $T_i = 1$ and $M_i = M_i(1)$
4. Compute the average difference between two outcomes to obtain a consistent estimate of ACME
5. Monte-Carlo or bootstrap to estimate uncertainty

Example: Binary Mediator and Outcome

- Two logistic regression models:

$$\begin{aligned} \Pr(M_i = 1 \mid T_i, X_i) &= \text{logit}^{-1}(\alpha_2 + \beta_2 T_i + \xi_2^\top X_i) \\ \Pr(Y_i = 1 \mid T_i, M_i, X_i) &= \text{logit}^{-1}(\alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i) \end{aligned}$$

- Can't multiply β_2 by γ
- Difference of coefficients $\beta_1 - \beta_3$ doesn't work either

$$\Pr(Y_i = 1 \mid T_i, X_i) = \text{logit}^{-1}(\alpha_1 + \beta_1 T_i + \xi_1^\top X_i)$$

- Can use our algorithm (example: $\mathbb{E}\{Y_i(1, M_i(0))\}$)
  1. Predict $M_i(0)$ given $T_i = 0$ using the first model
  2. Compute $\Pr(Y_i(1, M_i(0)) = 1 \mid T_i = 1, M_i = \widehat{M}_i(0), X_i)$ using the second model

Sensitivity Analysis

- Standard experiments require sequential ignorability to identify mechanisms
- The sequential ignorability assumption is often too strong
- Need to assess the robustness of findings via sensitivity analysis
- Question: How large a departure from the key assumption must occur for the conclusions to no longer hold?
- Parametric sensitivity analysis by assuming

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x$$

but not

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, X_i = x$$

- Possible existence of unobserved *pre-treatment* confounder

Parametric Sensitivity Analysis

- Sensitivity parameter: $\rho \equiv \text{Corr}(\epsilon_{i2}, \epsilon_{i3})$
- Sequential ignorability implies $\rho = 0$
- Set ρ to different values and see how ACME changes

- **Result:**

$$\bar{\delta}(0) = \bar{\delta}(1) = \frac{\beta_2 \sigma_1}{\sigma_2} \left\{ \tilde{\rho} - \rho \sqrt{(1 - \tilde{\rho}^2)/(1 - \rho^2)} \right\},$$

where $\sigma_j^2 \equiv \text{var}(\epsilon_{ij})$ for $j = 1, 2$ and $\tilde{\rho} \equiv \text{Corr}(\epsilon_{i1}, \epsilon_{i2})$.

- When do my results go away completely?
- $\bar{\delta}(t) = 0$ if and only if $\rho = \tilde{\rho}$
- Easy to estimate from the regression of Y_i on T_i:

$$Y_i = \alpha_1 + \beta_1 T_i + \epsilon_{i1}$$

Interpreting Sensitivity Analysis with R squares

- Interpreting ρ: how small is too small?
- An unobserved (pre-treatment) confounder formulation:

$$\epsilon_{i2} = \lambda_2 U_i + \epsilon'_{i2} \quad \text{and} \quad \epsilon_{i3} = \lambda_3 U_i + \epsilon'_{i3}$$

- How much does U_i have to explain for our results to go away?
- Sensitivity parameters: R squares
  1. Proportion of previously unexplained variance explained by U_i

$$R^{2*}_M \equiv 1 - \frac{\text{var}(\epsilon'_{i2})}{\text{var}(\epsilon_{i2})} \quad \text{and} \quad R^{2*}_Y \equiv 1 - \frac{\text{var}(\epsilon'_{i3})}{\text{var}(\epsilon_{i3})}$$

  2. Proportion of original variance explained by U_i

$$\widetilde{R}^2_M \equiv \frac{\text{var}(\epsilon_{i2}) - \text{var}(\epsilon'_{i2})}{\text{var}(M_i)} \quad \text{and} \quad \widetilde{R}^2_Y \equiv \frac{\text{var}(\epsilon_{i3}) - \text{var}(\epsilon'_{i3})}{\text{var}(Y_i)}$$

- Then reparameterize ρ using (R^{2*}_M, R^{2*}_Y) (or $(\widetilde{R}^2_M, \widetilde{R}^2_Y)$):

$$\rho \quad = \quad \text{sgn}(\lambda_2\lambda_3) R^*_M R^*_Y \quad = \quad \frac{\text{sgn}(\lambda_2\lambda_3)\widetilde{R}_M \widetilde{R}_Y}{\sqrt{(1 - R^2_M)(1 - R^2_Y)}},$$

  where R^2_M and R^2_Y are from the original mediator and outcome models

- $\text{sgn}(\lambda_2\lambda_3)$ indicates the direction of the effects of U_i on Y_i and M_i

- Set (R^{2*}_M, R^{2*}_Y) (or $(\widetilde{R}^2_M, \widetilde{R}^2_Y)$) to different values and see how mediation effects change

Reanalysis: Estimates under Sequential Ignorability

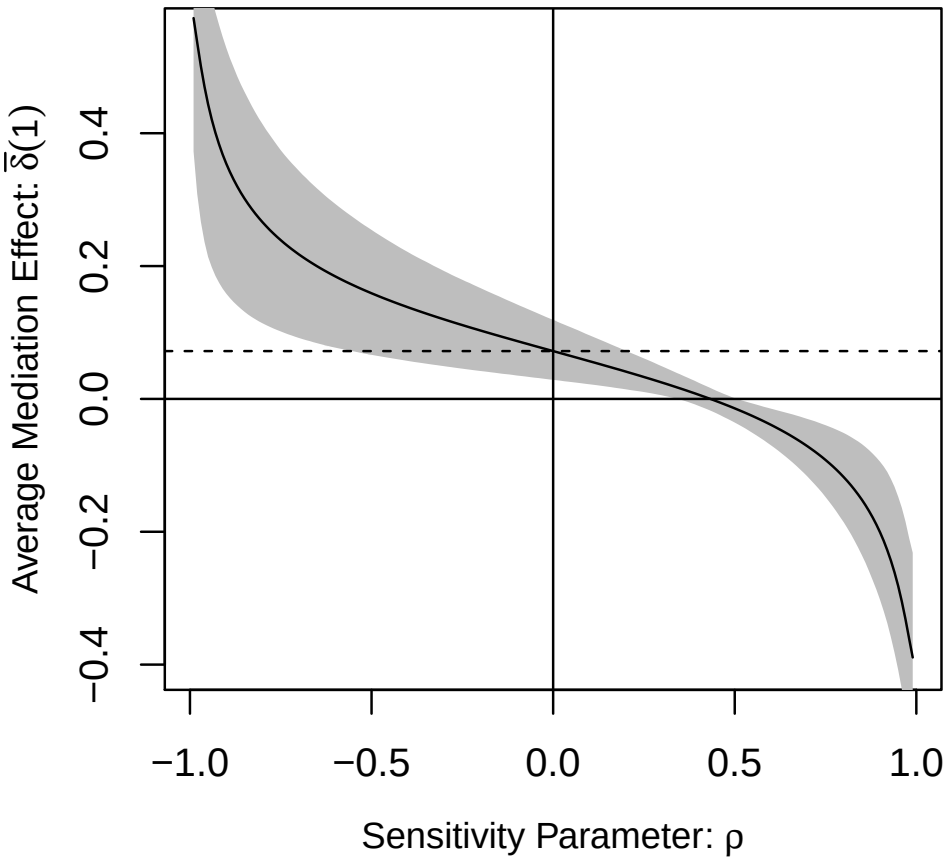
- Original method: Product of coefficients with the Sobel test

  — Valid only when both models are linear w/o T–M interaction (which they are not)

- Our method: Calculate ACME using our general algorithm

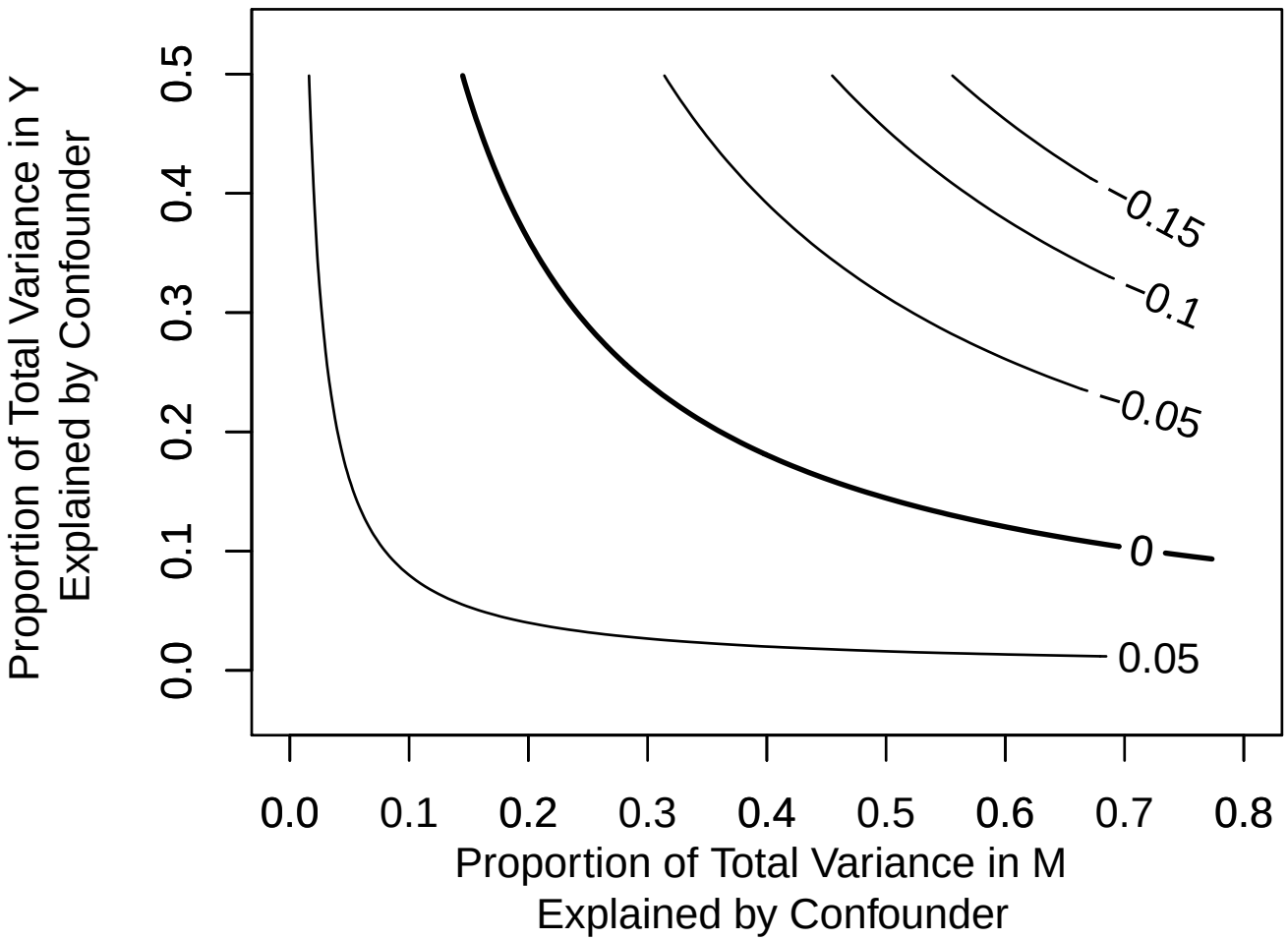
Outcome variables	Product of Coefficients	Average Causal Mediation Effect (δ)
Decrease Immigration	.347	.105
$\bar{\delta}(1)$	[0.146, 0.548]	[0.048, 0.170]
Support English Only Laws	.204	.074
$\bar{\delta}(1)$	[0.069, 0.339]	[0.027, 0.132]
Request Anti-Immigration Information	.277	.029
$\bar{\delta}(1)$	[0.084, 0.469]	[0.007, 0.063]
Send Anti-Immigration Message	.276	.086
$\bar{\delta}(1)$	[0.102, 0.450]	[0.035, 0.144]

Reanalysis: Sensitivity Analysis w.r.t. ρ



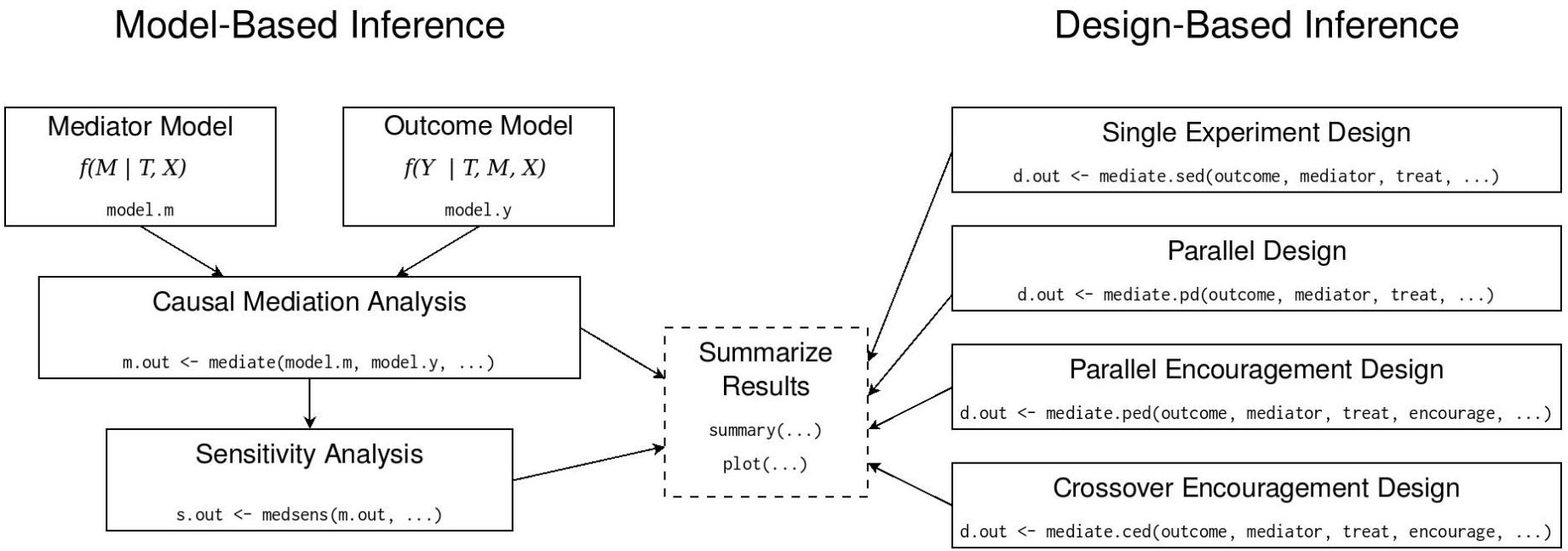
- ACME $>$ 0 as long as the error correlation is less than 0.39 (0.30 with 95% CI)

Reanalysis: Sensitivity Analysis w.r.t. $\tilde{R}^2_M$ and $\tilde{R}^2_Y$



- An unobserved confounder can account for up to 26.5% of the variation in both Y_i and M_i before ACME becomes zero

Open-Source Software "Mediation"



Implementation Examples

1. Fit models for the mediator and outcome variable and store these models

```
> m <- lm(Mediator ~ Treat + X)
> y <- lm(Y ~ Treat + Mediator + X)
```

2. Mediation analysis: Feed model objects into the `mediate()` function. Call a summary of results

```
> m.out<-mediate(m, y, treat = "Treat",
                 mediator = "Mediator")
> summary(m.out)
```

3. Sensitivity analysis: Feed the output into the `medsens()` function. Summarize and plot

```
> s.out <- medsens(m.out)
> summary(s.out)
> plot(s.out, "rho")
> plot(s.out, "R2")
```

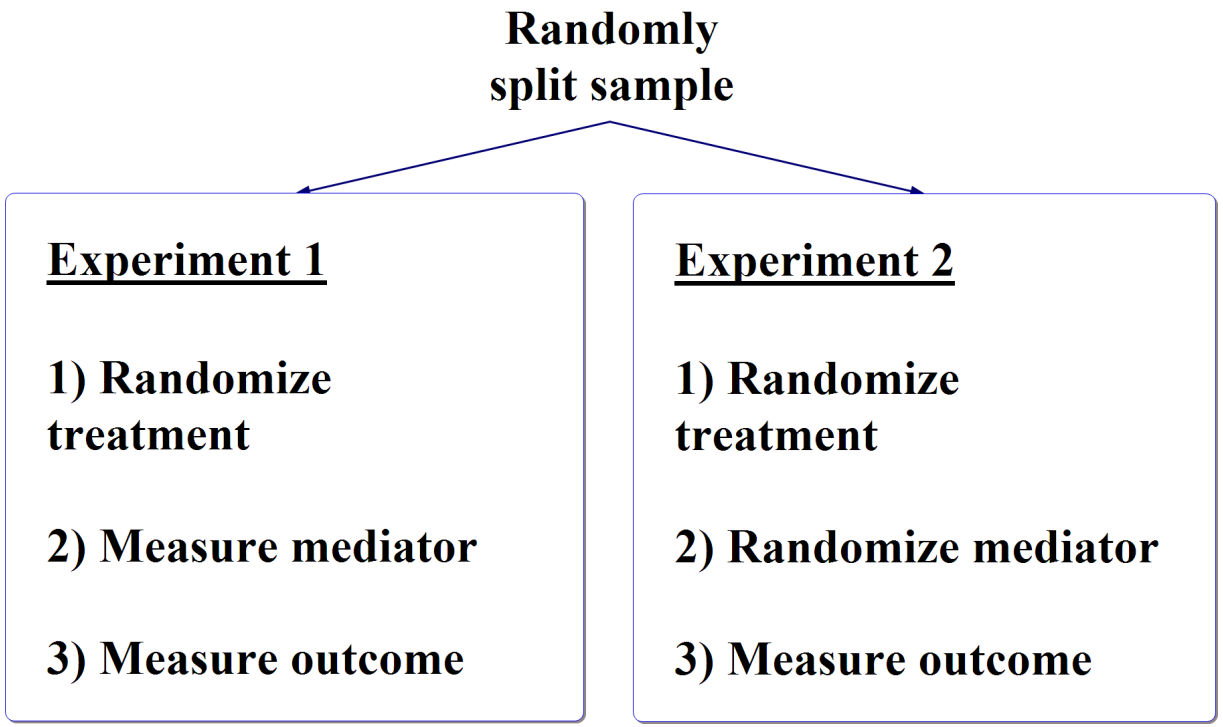
Beyond Sequential Ignorability

- Without sequential ignorability, standard experimental design lacks identification power
- Even the sign of ACME is not identified

- Need to develop alternative experimental designs for more credible inference
- Possible when the mediator can be directly or indirectly manipulated
- All proposed designs preserve the ability to estimate the ACME under the SI assumption
- Trade-off: statistical power

- These experimental designs can then be extended to natural experiments in observational studies

Parallel Design



- Must assume no direct effect of manipulation on outcome
- More informative than standard single experiment
- If we assume no T–M interaction, ACME is point identified

Why Do We Need No-Interaction Assumption?

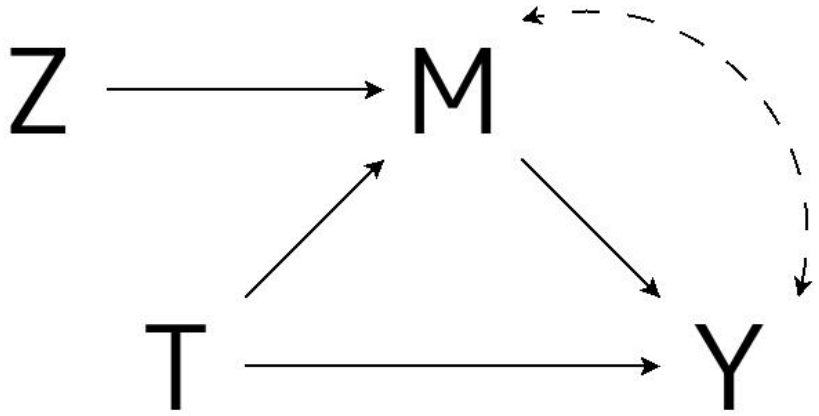
- Numerical Example:

Prop.	$M_i(1)$	$M_i(0)$	$Y_i(t,1)$	$Y_i(t,0)$	$\delta_i(t)$
0.3	1	0	0	1	-1
0.3	0	0	1	0	0
0.1	0	1	0	1	1
0.3	1	1	1	0	0

- $\mathbb{E}(M_i(1) - M_i(0)) = \mathbb{E}(Y_i(t,1) - Y_i(t,0)) = 0.2$, but $\bar{\delta}(t) = -0.2$
- The Problem: Causal effect heterogeneity
  - T increases M only *on average*
  - M increases Y only *on average*
  - $T - M$ interaction: Many of those who have a positive effect of T on M have a negative effect of M on Y (first row)
- A solution: sensitivity analysis (see Imai and Yamamoto, 2013)
- Pitfall of "mechanism experiments" or "causal chain approach"

Encouragement Design

- Direct manipulation of mediator is difficult in most situations
- Use an instrumental variable approach:



- Advantage: allows for unobserved confounder between M and Y
- Key Assumptions:
  1. Z is randomized or as-if random
  2. No direct effect of Z on Y (a.k.a. exclusion restriction)

Example: Social Norm Experiment on Property Taxes

- Lucia Del Carpio. "Are Neighbors Cheating?"
- Treatment: informing average rate of compliance
- Outcome: compliance rate obtained from administrative records
- Large positive effect on compliance rate $\approx$ 20 percentage points
- Mediators:
  1. social norm (not measured; direct effect)
  2. M_1: beliefs about compliance (measured)
  3. M_2: beliefs about enforcement (measured)
- Instruments:
  1. Z_1: informing average rate of enforcement
  2. Z_2: payment-reminder
- Assumptions:
  1. Z_1 affects Y only through M_1 and M_2
  2. Z_2 affects Y only through M_1
- Results:
  - Average direct effect is estimated to be large
  - The author interprets this effect as the effect of social norm

Crossover Design

- Recall ACME can be identified if we observe $Y_i(t', M_i(t))$
- Get $M_i(t)$, then switch T_i to t' while holding $M_i = M_i(t)$

- Crossover design:
  1. Round 1: Conduct a standard experiment
  2. Round 2: Change the treatment to the opposite status but fix the mediator to the value observed in the first round
- Very powerful – identifies mediation effects for each subject
- Must assume no carryover effect: Round 1 must not affect Round 2
- Can be made plausible by design

Example: Labor Market Discrimination

EXAMPLE Bertrand & Mullainathan (2004, AER)

- Treatment: Black vs. White names on CVs
- Mediator: Perceived qualifications of applicants
- Outcome: Callback from employers

- Quantity of interest: Direct effects of (perceived) race
- Would Jamal get a callback if his name were Greg but his qualifications stayed the same?

- Round 1: Send Jamal's actual CV and record the outcome
- Round 2: Send his CV as Greg and record the outcome

- Assumption: their different names do not change the perceived qualifications of applicants
- Under this assumption, the direct effect can be interpreted as blunt racial discrimination

Designing Observational Studies

- Key difference between experimental and observational studies: treatment assignment
- Sequential ignorability:
  1. Ignorability of treatment given covariates
  2. Ignorability of mediator given treatment and covariates
- Both (1) and (2) are suspect in observational studies

- Statistical control: matching, propensity scores, etc.
- Search for quasi-randomized treatments: "natural" experiments

- How can we design observational studies?
- Experiments can serve as templates for observational studies

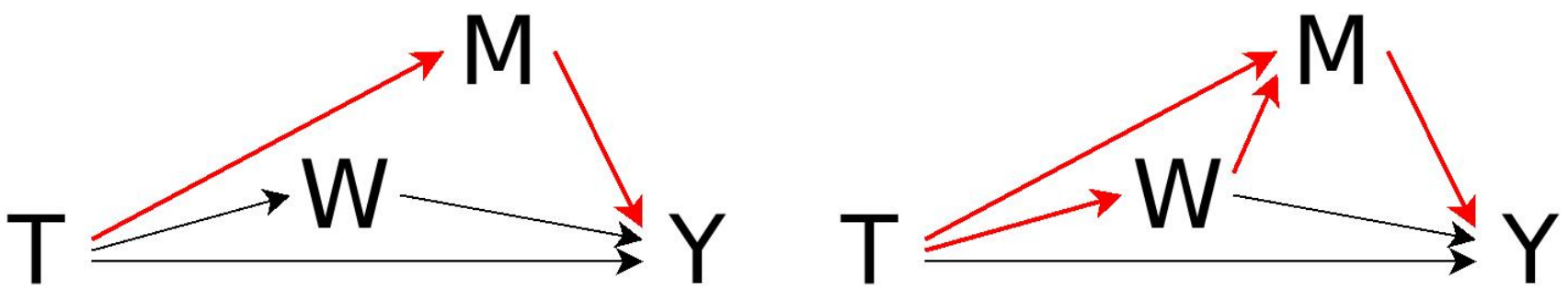
Cross-over Design in Observational Studies

EXAMPLE Back to incumbency advantage

- Use of cross-over design (Levitt and Wolfram)
  1. 1st Round: two non-incumbents in an open seat
  2. 2nd Round: same candidates with one being an incumbent
- Assume challenger quality (mediator) stays the same
- Estimation of direct effect is possible

- Redistricting as natural experiments (Ansolabehere et al.)
  1. 1st Round: incumbent in the old part of the district
  2. 2nd Round: incumbent in the new part of the district
- Challenger quality is the same but treatment is different
- Estimation of direct effect is possible

Multiple Mediators



- Quantity of interest = The average indirect effect with respect to M
- W represents the alternative observed mediators
- Left: Assumes independence between the two mechanisms
- Right: Allows M to be affected by the other mediators W
- Applied work often assumes the independence of mechanisms
- Under this independence assumption, one can apply the same analysis as in the single mediator case
- For causally dependent mediators, we must deal with the heterogeneity in the $T \times M$ interaction as done under the parallel design $\Longrightarrow$ sensitivity analysis

Concluding Remarks

- Even in a randomized experiment, a strong assumption is needed to identify causal mechanisms
- However, progress can be made toward this fundamental goal of scientific research with modern statistical tools
- A general, flexible estimation method is available once we assume sequential ignorability
- Sequential ignorability can be probed via sensitivity analysis
- More credible inferences are possible using clever experimental designs
- Insights from new experimental designs can be directly applied when designing observational studies
- Multiple mediators require additional care when they are causally dependent