Causal Inference and Missing Data

Kosuke Imai

Princeton University

POL573 Quantitative Analysis III
Fall 2016

What We Learned about Causal Inference in POL 572

- Potential outcomes framework: role of counterfactuals
- Identification and inference: assumption and credibility
- Classical randomized experiments: randomization inference
- Regression and causal effects: strong ignorability
- Cluster randomized experiments: interference, robust s.e.
- Regression discontinuity designs: sharp and fuzzy
- Instrumental variables: encouragement design, compliers
- Causal mediation analysis: causal mechanisms

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?

Identification of the Average Treatment Effect: Review

- 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.)

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

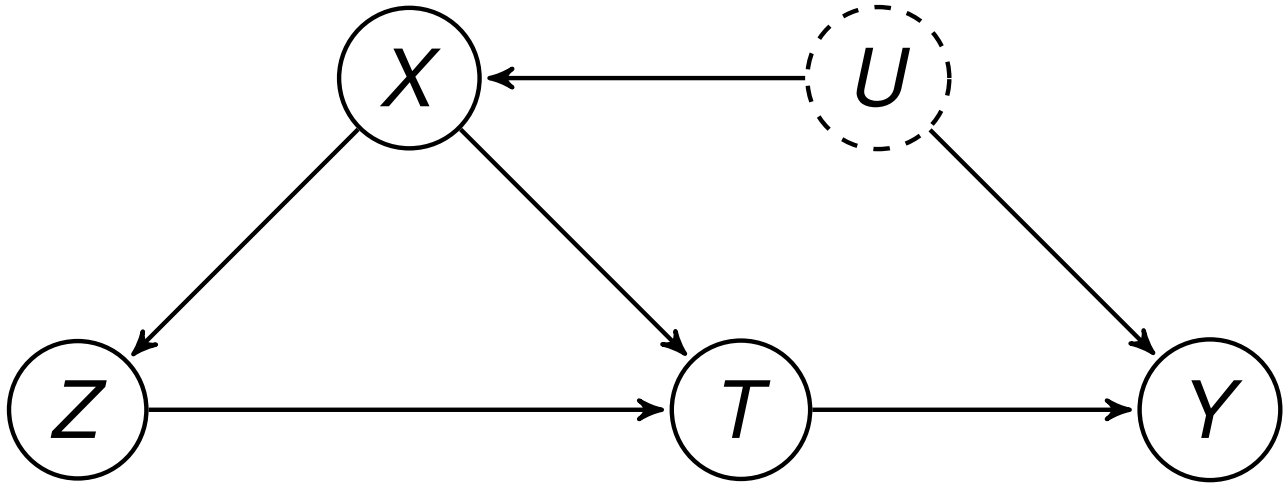
- Conditional expectation function: $\mu(t, x) = \mathbb{E}(Y_i(t) \mid T_i = t, X_i = x)$
- Regression-based Estimator:

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

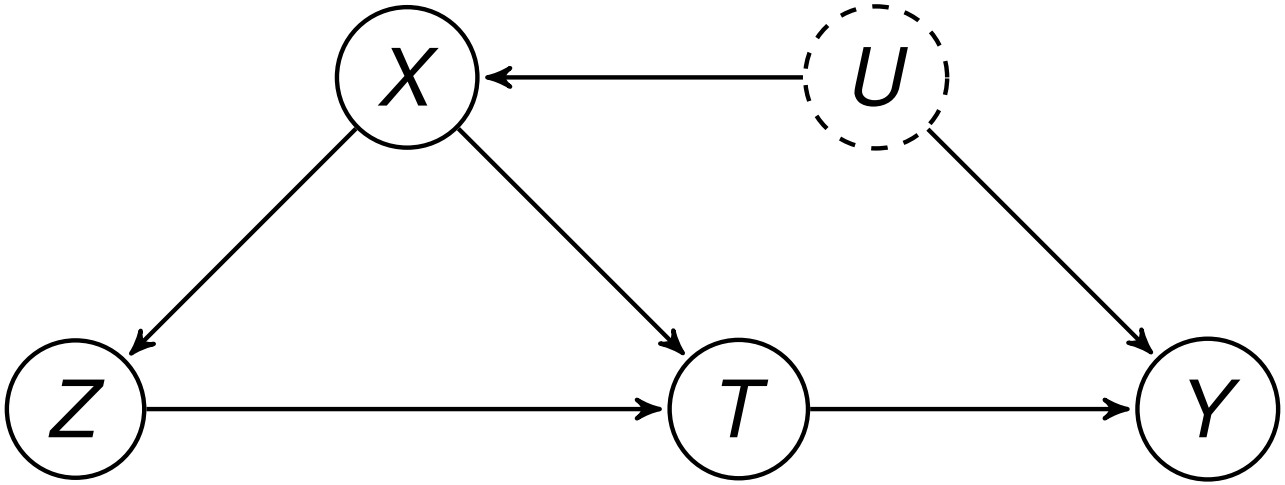
- Delta method is pain, but simulation is easy (Zelig)

Causal Directed Acyclic Graphs (Causal DAGs)

- Judea Pearl. *Causality* Cambridge UP.
- Elwert, F. (2013). Chapter 13: Graphical Causal Models in *Handbook of Causal Analysis for Social Research*
- Check out DAGitty at `http://dagitty.net/`
- Elements of DAGs:
  1. (observed and unobserved) variables: nodes or vertices
  2. (directed) arrows: *possible* causal effects
  3. Absence of variables: all common causes of any pair of variables
  4. Absence of arrows: assumed *absence* of (i.e., zero) causal effect



- Acyclic: no simultaneity, the future does not cause the past



- Parents (Children): directly causing (caused by) a node
- Ancestors (Descendents): directly or indirectly causing (caused by) a node
- Path: an acyclic sequence of adjacent nodes
  - Causal path: all arrows pointing away from T and into Y
  - Non-causal path: some arrows going against causal order
- Collider: a node on a path with two incoming arrows
  - Conditioning on a collider induces association
- Nonparametric structural equation models

D-separation

- Suppose we want to know whether a conditional independence relationship, $A \perp\!\!\!\perp B \mid C$, holds where A, B, C are sets of nodes

- Consider all possible paths from any node in A to any node in B
- Check if each path is blocked, i.e., if it includes a node such that
  1. the arrows on the path meet either head-to-tail (chain) or tail-to-tail (fork) at the node, and the node is in the set C
  2. the arrows meet head-to-head (collider) at the node, and neither the node nor any of its descendants, is in the set C.
- If all paths are blocked, then A is *d*-separated from B by C

- If *d*-separated, $A \perp\!\!\!\perp B \mid C$ holds
- If not *d*-separated (i.e., *d*-connected), $A \not\perp\!\!\!\perp B \mid C$ in at least one distribution compatible with DAG

- What variables should we condition in order to make X independent of Y? What about U and T?

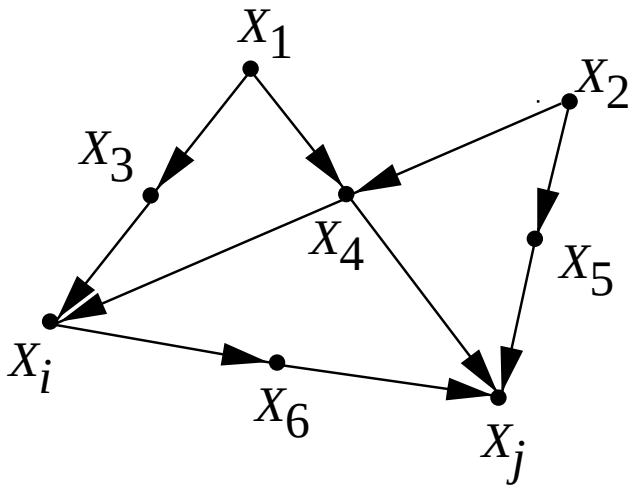
Backdoor Criterion

- Can we nonparametrically identify the average effect of T on Y given a set of variables X?

- Backdoor criterion for X:
  1. No node in X is a decendent of T
  2. X d-separates every path between T and Y that has an incoming arrow into T (backdoor path)
- Need to block all non-causal paths

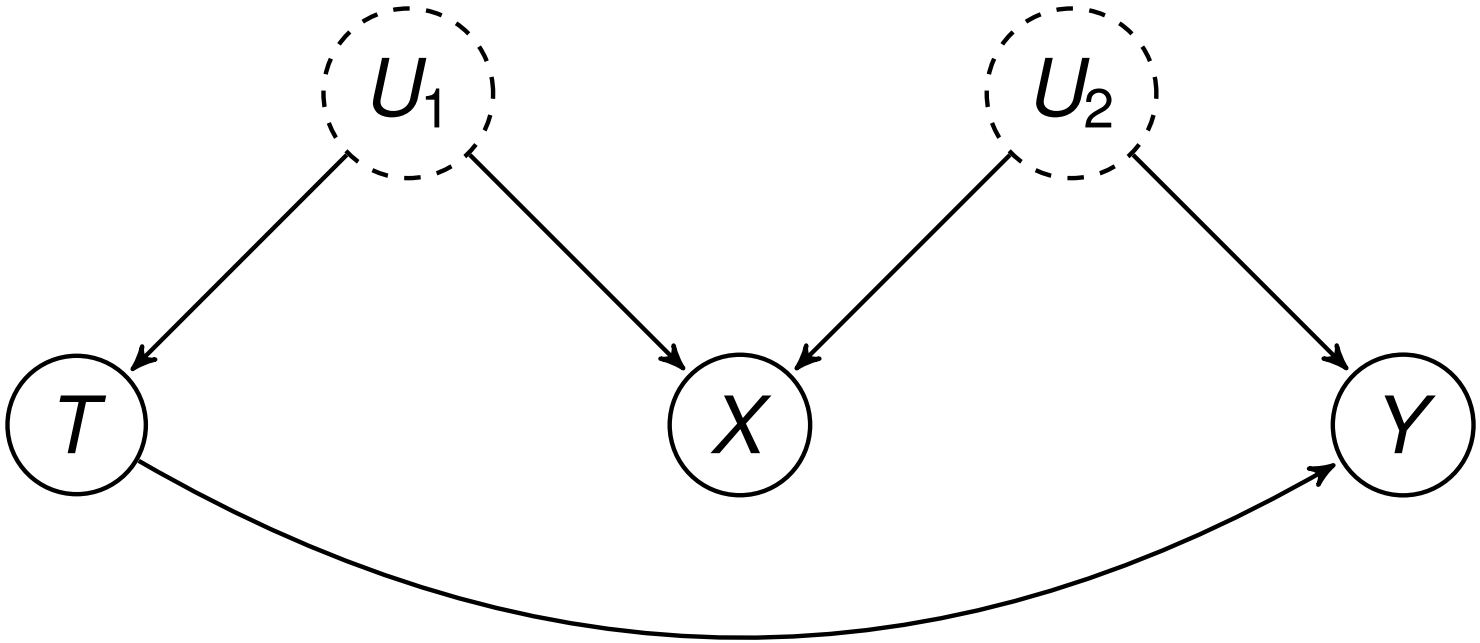
- Can we identify the causal effect of T on Y by conditioning on X? What about U?
- Can we identify the causal effect of Z on Y by conditioning on X? What about U?

An Example from Pearl



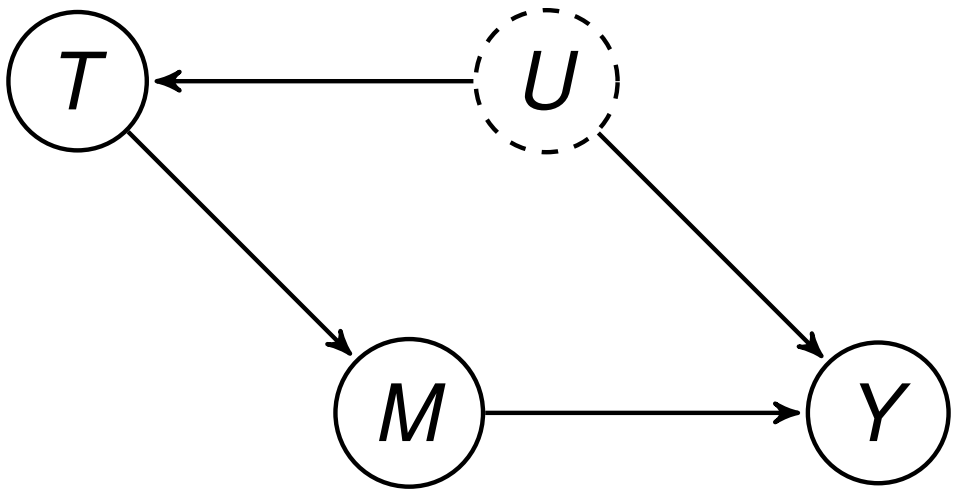
- Can we identify the average effect of X_i on X_j by conditioning on X_4 alone?
- Does conditioning on $\{X_4, X_5\}$ identify the average effect of X_i on X_j?

M-Structure



- Does conditioning on X identify the average effect of T on Y?

Frontdoor Criterion



- M can be used to identify the average causal effect of T on Y in the presence of unobserved confounder U
- Frontdoor criterion for M:
  1. M intercepts all directed paths from T to Y
  2. No backdoor path from T to M
  3. All backdoor paths from M to Y are blocked by T
- Social science applications by Glynn and Kashin

Potential Outcome vs. DAGs Controversy

- Imbens and Rubin (2015):
  Pearl's work is interesting, and many researchers find his arguments that path diagrams are a natural and convenient way to express assumptions about causal structures appealing. In our own work, perhaps influenced by the type of examples arising in social and medical sciences, we have not found this approach to aid drawing of causal inferences.

- Pearl's blog post:
  So, what is it about epidemiologists that drives them to seek the light of new tools, while economists seek comfort in partial blindness, while missing out on the causal revolution? Can economists do in their heads what epidemiologists observe in their graphs? Can they, for instance, identify the testable implications of their own assumptions? Can they decide whether the IV assumptions are satisfied in their own models of reality? Of course they can't; such decisions are intractable to the graph-less mind.

My Own View

- Potential outcomes are useful when thinking about treatment assignment mechanism $\rightsquigarrow$ experiments, quasi-experiments
- DAGs are useful when thinking about the entire causal structure $\rightsquigarrow$ complex causal relationships
- Both are better suited for causal inference than the standard regression framework

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

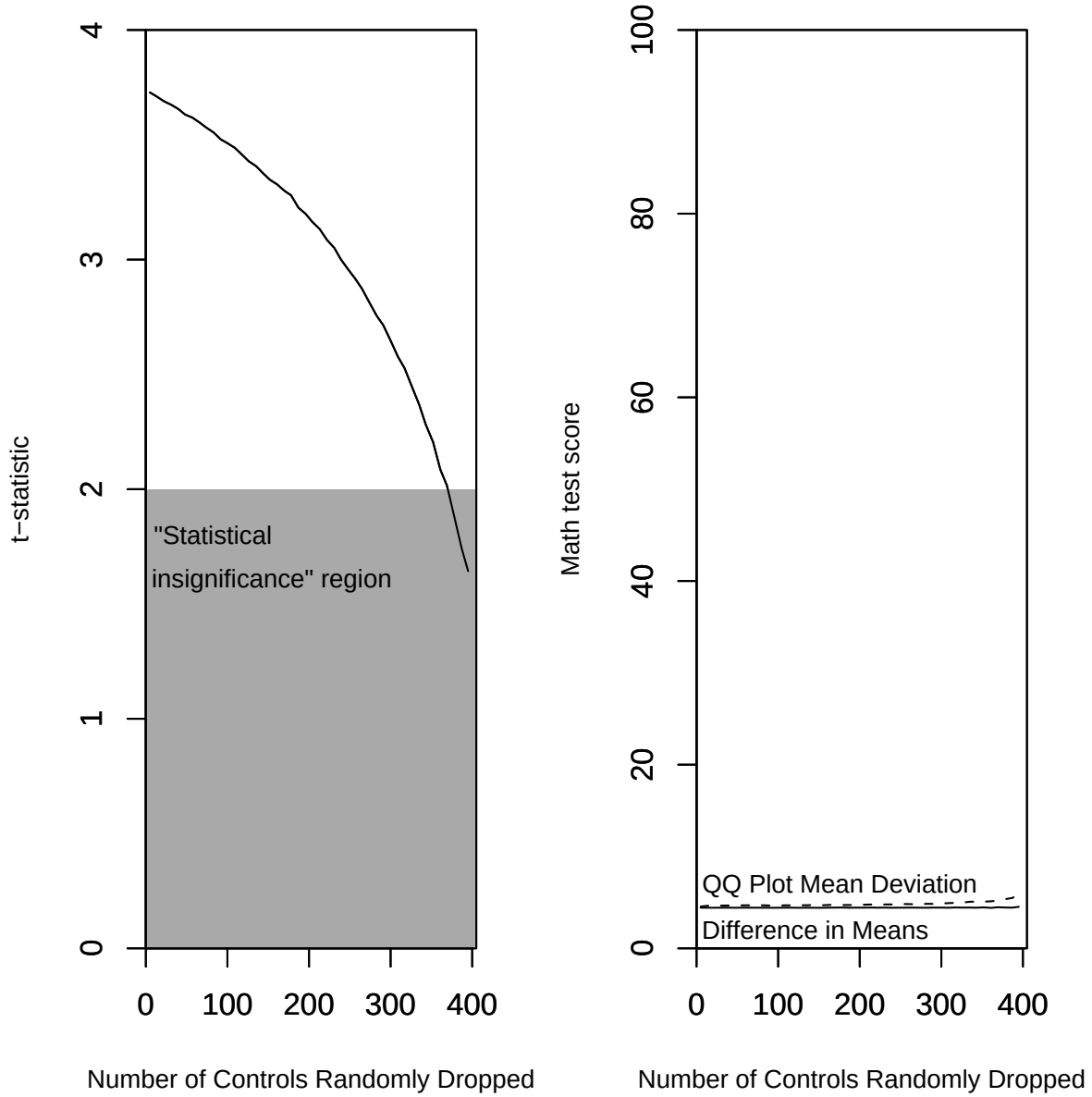
- Exact matching
- Mahalanobis distance matching: $\sqrt{(X_i - X_j)^\top \widetilde{\Sigma}^{-1}(X_i - X_j)}$
- Propensity score matching
- One-to-one, one-to-many, and subclassification
- Matching with caliper
- Which matching method to choose?
- Whatever gives you the "best" balance!
- Importance 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



Problems with Hypothesis Tests as Stopping Rules

- Balance test is a function of both balance and statistical power
- The more observations dropped, the less power the tests have
- t-test is affected by factors other than balance,

$$\frac{\sqrt{n_m}(\overline{X}_{mt} - \overline{X}_{mc})}{\sqrt{\frac{s^2_{mt}}{r_m} + \frac{s^2_{mc}}{1-r_m}}}$$

  - $\overline{X}_{mt}$ and $\overline{X}_{mc}$ are the sample means
  - s^2_{mt} and s^2_{mc} are the sample variances
  - n_m is the total number of remaining observations
  - r_m is the ratio of remaining treated units to the total number of remaining observations

Equivalence Tests for Covariate Balance

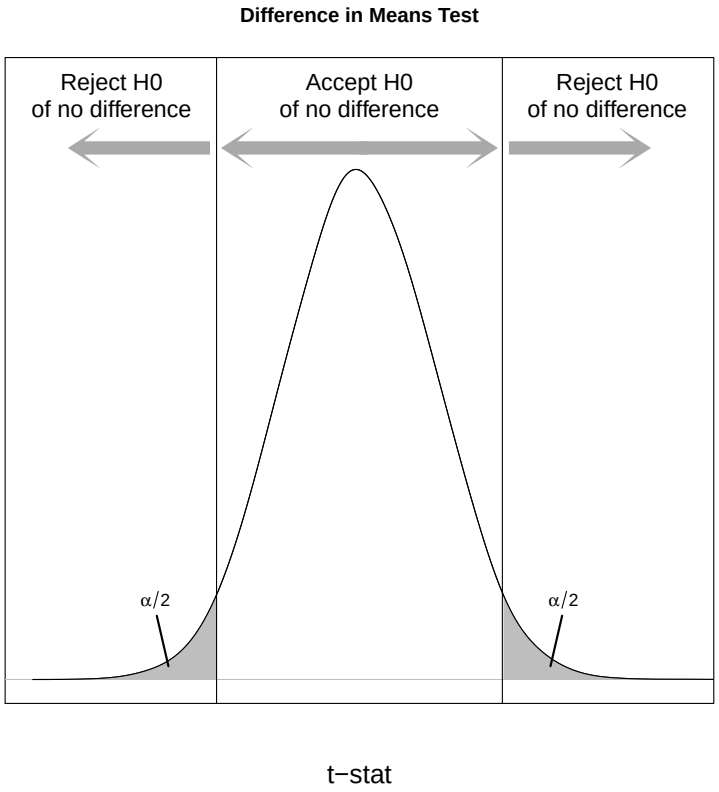
- Null hypothesis of usual significance tests: covariate is balanced
- Problem: failure to reject the null does not necessarily imply the null is correct
- Shift the burden of proof $\rightsquigarrow$ can we reject the null hypothesis that covariate is not balanced?
- Equivalence tests can be used (Hartman and Hidalgo):

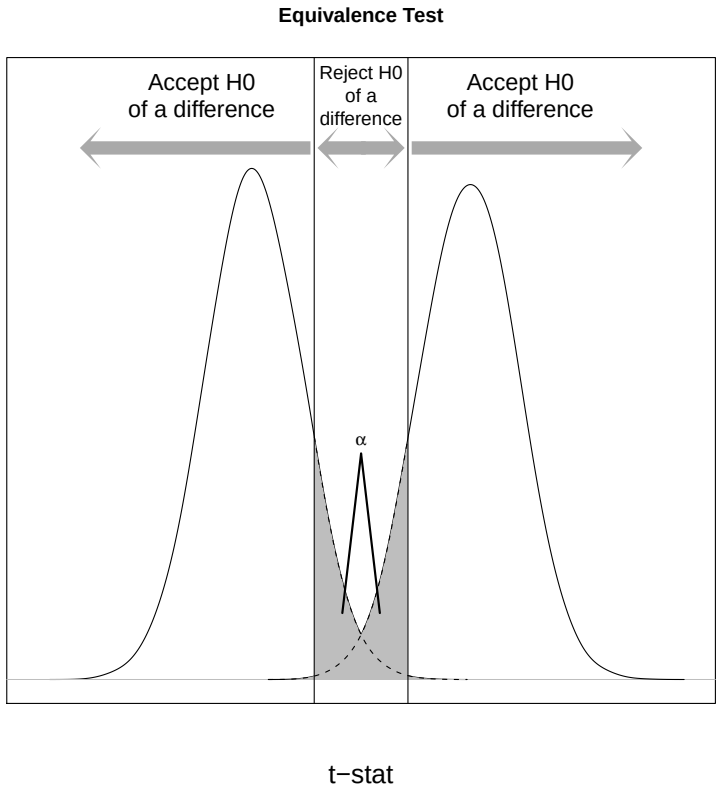
$$H_0 : |\mu_1 - \mu_0| \geq \Delta \quad \text{and} \quad H_1 : |\mu_1 - \mu_0| < \Delta$$

  for a pre-selected value of $\Delta > 0$
- Two one-sided test procedure (TOST: α level):

$$\frac{(\hat{\mu}_1 - \hat{\mu}_0) + \Delta}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}} > z_{1-\alpha} \qquad \frac{(\hat{\mu}_1 - \hat{\mu}_0) - \Delta}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}} < -z_{1-\alpha}$$

- Two groups are equivalent if and only if both are rejected
- $\alpha =$ probability of falsely concluding equivalence under the null





- Inverting the test $\rightsquigarrow$ $(1 - 2\alpha)$ level confidence interval (rather than $(1 - \alpha)$ level):

$$(\hat{\mu}_1 - \hat{\mu}_0) \pm z_{1-\alpha} \times \text{standard error}$$

- If the α level equivalence test is rejected, then this confidence interval is contained within $[-\Delta, \Delta]$

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

Statistical Uncertainty for Matching Estimators

- Three perspectives:
  1. matching (or pruning) as a preprocessing procedure (Ho et al.)
  2. matching as a weighted linear regression (Imai and Kim)
  3. matching as a sampling procedure (Abadie and Imbens)
- A general matching estimator:

$$\hat{\tau}_{match} = \frac{1}{N_1} \sum_{i=1}^{N} T_i \left(Y_i - \frac{1}{M_i} \sum_{i' \in \mathcal{M}_i} (1 - T_{i'}) Y_{i'} \right)$$

  where $\mathcal{M}_i$ is the matched set for unit i and N_1 is the number of treated units
- Can be expressed as a weighted average estimator:

$$\hat{\tau}_{match} = \frac{1}{N_1} \sum_{i=1}^{N} T_i Y_i - \frac{1}{N - N_1} \sum_{i=1}^{N} (1 - T_i) W_i Y_i$$

- Conditional variance is given by:

$$\mathbb{V}(\hat{\tau}_{match} \mid \mathbf{X}, \mathbf{T}) = \frac{1}{N_1^2} \sum_{i=1}^{N} T_i \sigma_{1i}^2 + \frac{1}{(N - N_1)^2} \sum_{i=1}^{N} (1 - T_i) W_i^2 \sigma_{0i}^2$$

- The problem: may not be able to estimate σ_{ti}^2 due to the insufficient number of observations $\rightsquigarrow$ matching, regression (see Imbens and Rubin for details)
- Unconditional variance:

$$\mathbb{V}(\hat{\tau}_{match}) = \mathbb{E}(\mathbb{V}(\tau_{match} \mid \mathbf{X}, \mathbf{T})) + \mathbb{V}(\mathbb{E}(\hat{\tau}_{match} \mid \mathbf{X}, \mathbf{T}))$$

- If we observe τ_i, then the second term is estimated by the sample variance of $\tau_i = \mathbb{E}(\hat{\tau}_{match} \mid \mathbf{X}, \mathbf{T})$. But, we only have $\hat{\tau}_i$ via matching $\rightsquigarrow$ this terms also requires the estimation of σ_{ti}^2 (see Imbens and Rubin)

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

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

- The score condition for MLE of propensity score:

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

- Optimizes a particular function of covariates $\pi'_\beta(X_i)$
- Directly optimizing the covariate balance metrics of your choice:

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

  where $\widetilde{X}_i = f(X_i)$ is any vector-valued function
- Just identified CBPS: # of parameters K = # of moment conditions L

Generalized Method of Moments (GMM)

- over-identification: K parameters and $L > K$ moment conditions
- The sample analogue of the moment condition $\mathbb{E}\{g(\theta, \text{Data}_i)\} = 0$:

$$\bar{g}(\theta, \text{Data}) \;=\; \frac{1}{N} \sum_{i=1}^{N} g(\theta, \text{Data}_i)$$

- GMM estimator (Hansen 1982):

$$\hat{\theta} \;=\; \underset{\theta \in \Theta}{\text{argmin}}\; \bar{g}(\theta, \text{Data})^\top \; \widehat{W} \; \bar{g}(\theta, \text{Data})$$

  where $\widehat{W} \xrightarrow{p} W$ with W being a positive definite matrix
- Asymptotic distribution:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, (G^\top W G)^{-1} G^\top W \Omega W G (G^\top W G)^{-1})$$

  where $G = \mathbb{E}(\frac{\partial}{\partial \theta} g(\theta, \text{Data}_i))$ and $\Omega = \mathbb{E}(g(\theta, \text{Data}_i) g(\theta, \text{Data}_i)^\top)$
- Most efficient weighting matrix: $W = \Omega^{-1}$

Application to CBPS

- Logistic regression: $\pi_\beta(X_i) = \exp(X_i^\top \beta)/\{1 + \exp(X_i^\top \beta)\}$
- CBPS estimation:

$$\hat{\beta} \quad = \quad \underset{\beta \in \Theta}{\operatorname{argmin}} \; \bar{g}_\beta(T, X)^\top \Sigma^{-1} \bar{g}_\beta(T, X)$$

where

$$\bar{g}_\beta(T, X) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \left(\frac{T_i \widetilde{X}_i}{\pi_\beta(X_i)} - \frac{(1 - T_i)\widetilde{X}_i}{1 - \pi_\beta(X_i)} \right)$$

- `CBPS: R Package for Covariate Balancing Propensity Score`
- Imai and Ratkovic (2014) *J. of Royal Stat. Soc., Series B*
- Other methods: entropy balancing (Hainmueller), Inverse probability tilting (Graham et al.), Calibration weights (Chan et al.) etc.

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	CBPS	True	logit	CBPS	True
(3) Outcome model correct							
	HT	24.25	1.09	−0.18	194.58	5.04	23.24
$n = 200$	IPW	1.70	−1.37	−0.26	9.75	3.42	4.93
	WLS	−2.29	−2.37	0.41	4.03	4.06	3.31
	DR	−0.08	−0.10	−0.10	2.67	2.58	2.58
	HT	41.14	−2.02	−0.23	238.14	2.97	10.42
$n = 1000$	IPW	4.93	−1.39	−0.02	11.44	2.01	2.21
	WLS	−2.94	−2.99	0.20	3.29	3.37	1.47
	DR	0.02	0.01	0.01	1.89	1.13	1.13
(4) Both models incorrect							
	HT	30.32	1.27	−0.38	266.30	5.20	23.86
$n = 200$	IPW	1.93	−1.26	−0.09	10.50	3.37	5.08
	WLS	−2.13	−2.20	0.55	3.87	3.91	3.29
	DR	−7.46	−2.59	0.37	50.30	4.27	3.74
	HT	101.47	−2.05	0.01	2371.18	3.02	10.53
$n = 1000$	IPW	5.16	−1.44	0.02	12.71	2.06	2.25
	WLS	−2.95	−3.01	0.19	3.30	3.40	1.47
	DR	−48.66	−3.59	0.08	1370.91	4.02	1.81

What Function of Covariates Should We Balance?

- Define $\mathbb{E}(Y_i(t) \mid X_i) = K_t(X_i)$ for $t = 0, 1$
- ATE: $\mu = \mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(K_1(X_i) - K_0(X_i))$
- Misspecified propensity score: $\pi_{\hat{\beta}}(X_i) \longrightarrow \pi_{\hat{\beta}^o}(X_i)$
- Estimator: $\hat{\mu}_{\hat{\beta}} = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{T_i Y_i}{\pi_{\hat{\beta}}(X_i)} - \frac{(1-T_i) Y_i}{1-\pi_{\hat{\beta}}(X_i)} \right\}$
- Asymptotic bias: $\mathbb{E}(\hat{\mu}_{\beta^o}) - \mu$

$$\mathbb{E}\left[\left(\frac{T_i}{\pi_{\beta^o}(X_i)} - \frac{1 - T_i}{1 - \pi_{\beta^o}(X_i)} \right) \{ \pi_{\beta^o}(X_i) K_0(X_i) + (1 - \pi_{\beta^o}(X_i)) K_1(X_i) \} \right]$$

- The improved CBPS (Fan et al. 2016):

$$\mathbb{E}\left[\left(\frac{T_i}{\pi_{\beta^o}(X_i)} - \frac{1 - T_i}{1 - \pi_{\beta^o}(X_i)} \right) K_0(X_i) + \left(\frac{T_i}{\pi_{\beta^o}(X_i)} - 1 \right) \{ K_1(X_i) - K_0(X_i) \} \right]$$

Fixed Effects Regressions in Causal Inference

- Linear fixed effects regression models are the primary workhorse for causal inference with longitudinal/panel data
- Researchers use them to adjust for unobserved time-invariant confounders (omitted variables, endogeneity, selection bias, ...):
  - "Good instruments are hard to find ..., so we'd like to have other tools to deal with unobserved confounders. This chapter considers ... strategies that use data with a time or cohort dimension to control for unobserved but fixed omitted variables"
    (Angrist & Pischke, *Mostly Harmless Econometrics*)
  - "fixed effects regression can scarcely be faulted for being the bearer of bad tidings" (Green *et al.*, *Dirty Pool*)

Linear Regression with Unit Fixed Effects

- Balanced panel data with N units and T time periods
- Y_{it}: outcome variable
- X_{it}: causal or treatment variable of interest

Assumption 1 (Linearity)

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

- $\mathbf{U}_i$: a vector of unobserved time-invariant confounders
- $\alpha_i = h(\mathbf{U}_i)$ for *any* function $h(\cdot)$
- A flexible way to adjust for unobservables
- Average contemporaneous treatment effect:

$$\beta = \mathbb{E}(Y_{it}(1) - Y_{it}(0))$$

Strict Exogeneity and Least Squares Estimator

Assumption 2 (Strict Exogeneity)

$$\epsilon_{it} \perp\!\!\!\perp \{\mathbf{X}_i, \mathbf{U}_i\}$$

- Mean independence is sufficient: $\mathbb{E}(\epsilon_{it} \mid \mathbf{X}_i, \mathbf{U}_i) = \mathbb{E}(\epsilon_{it}) = 0$
- Least squares estimator based on de-meaning:

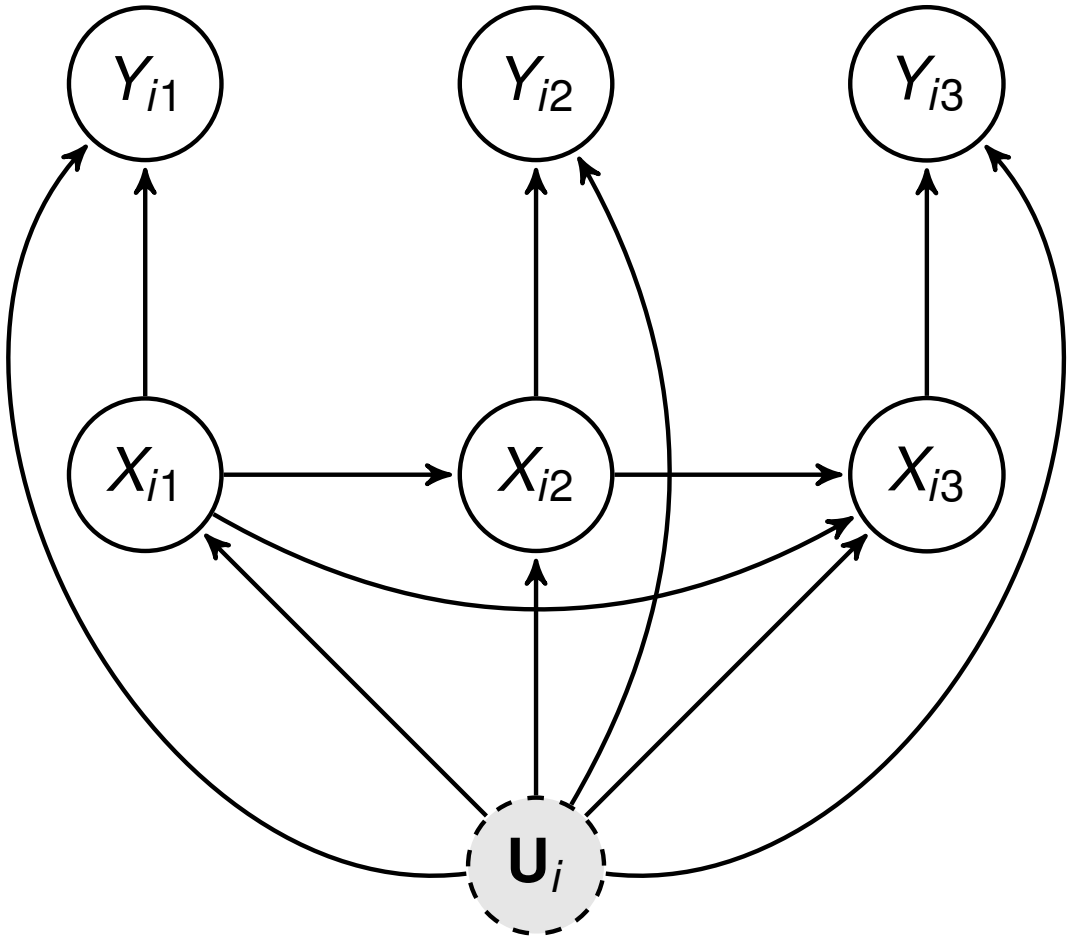
$$\hat{\beta}_{\text{FE}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \{(Y_{it} - \overline{Y}_i) - \beta(X_{it} - \overline{X}_i)\}^2$$

  where $\overline{X}_i$ and $\overline{Y}_i$ are unit-specific sample means
- ATE among those units with variation in treatment:

$$\tau = \mathbb{E}(Y_{it}(1) - Y_{it}(0) \mid C_{it} = 1)$$

  where $C_{it} = \mathbf{1}\{0 < \sum_{t=1}^{T} X_{it} < T\}$.

Causal Directed Acyclic Graph (DAG)



- arrow $=$ direct causal effect
- absence of arrows $\rightsquigarrow$ causal assumptions

Nonparametric Structural Equation Model (NPSEM)

- One-to-one correspondence with a DAG:

$$\begin{aligned} Y_{it} &= g_1(X_{it}, \mathbf{U}_i, \epsilon_{it}) \\ X_{it} &= g_2(X_{i1}, \ldots, X_{i,t-1}, \mathbf{U}_i, \eta_{it}) \end{aligned}$$

- Nonparametric generalization of linear unit fixed effects model:
  - Allows for nonlinear relationships, effect heterogeneity
  - Strict exogeneity holds
  - No arrows can be added without violating Assumptions 1 and 2

- Causal assumptions:
  1. No unobserved time-varying confounders
  2. Past outcomes do not directly affect current outcome
  3. Past outcomes do not directly affect current treatment
  4. Past treatments do not directly affect current outcome

Potential Outcomes Framework

- DAG $\rightsquigarrow$ causal structure
- Potential outcomes $\rightsquigarrow$ treatment assignment mechanism

Assumption 3 (No carryover effect)

Past treatments do not directly affect current outcome

$$Y_{it}(X_{i1}, X_{i2}, \ldots, X_{i,t-1}, X_{it}) = Y_{it}(X_{it})$$

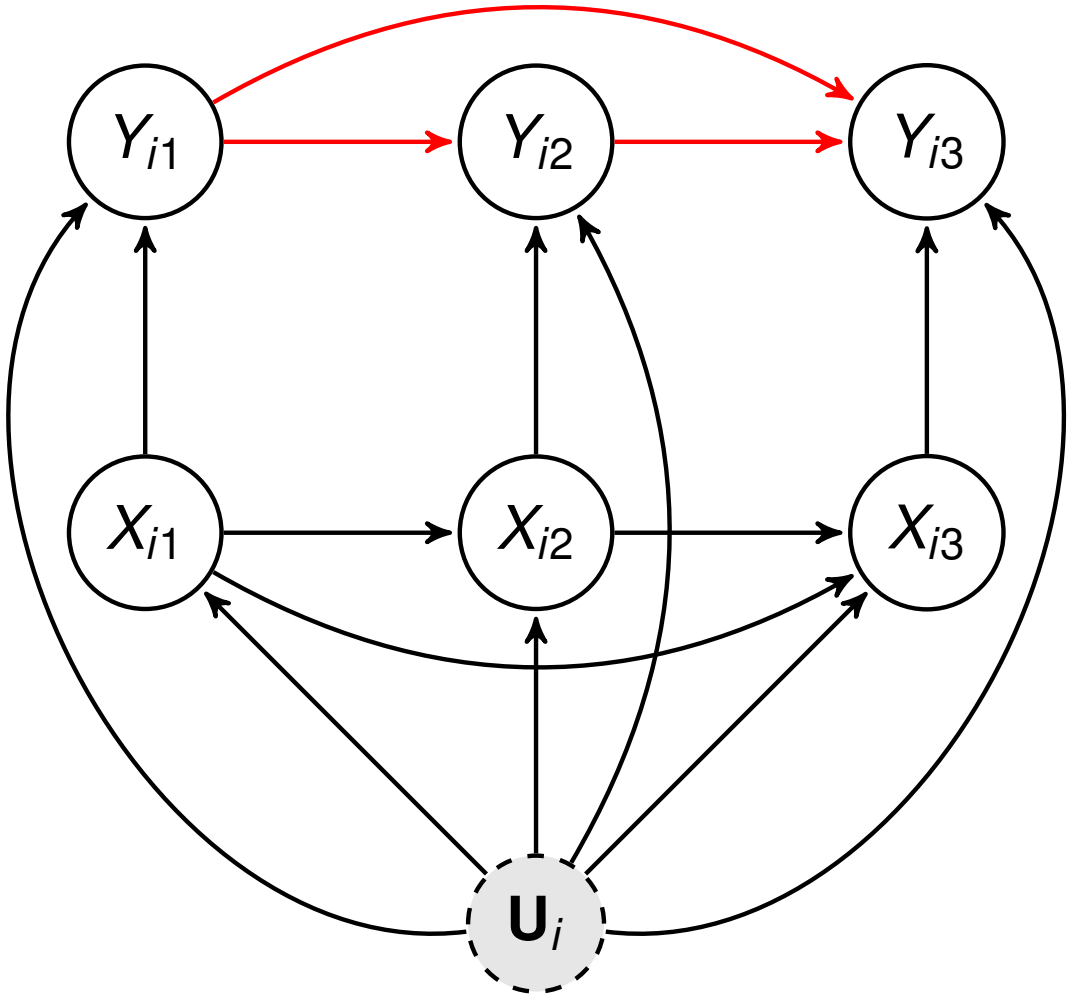
- What randomized experiment satisfies unit fixed effects model?
  1. randomize X_{i1} given $\mathbf{U}_i$
  2. randomize X_{i2} given X_{i1} and $\mathbf{U}_i$
  3. randomize X_{i3} given X_{i2}, X_{i1}, and $\mathbf{U}_i$
  4. and so on

Assumption 4 (Sequential Ignorability with Unobservables)

$$
\begin{aligned}
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} &\perp\!\!\!\perp X_{i1} \mid \mathbf{U}_i \\
&\vdots \\
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} &\perp\!\!\!\perp X_{it'} \mid X_{i1}, \ldots, X_{i,t'-1}, \mathbf{U}_i \\
&\vdots \\
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} &\perp\!\!\!\perp X_{iT} \mid X_{i1}, \ldots, X_{i,T-1}, \mathbf{U}_i
\end{aligned}
$$

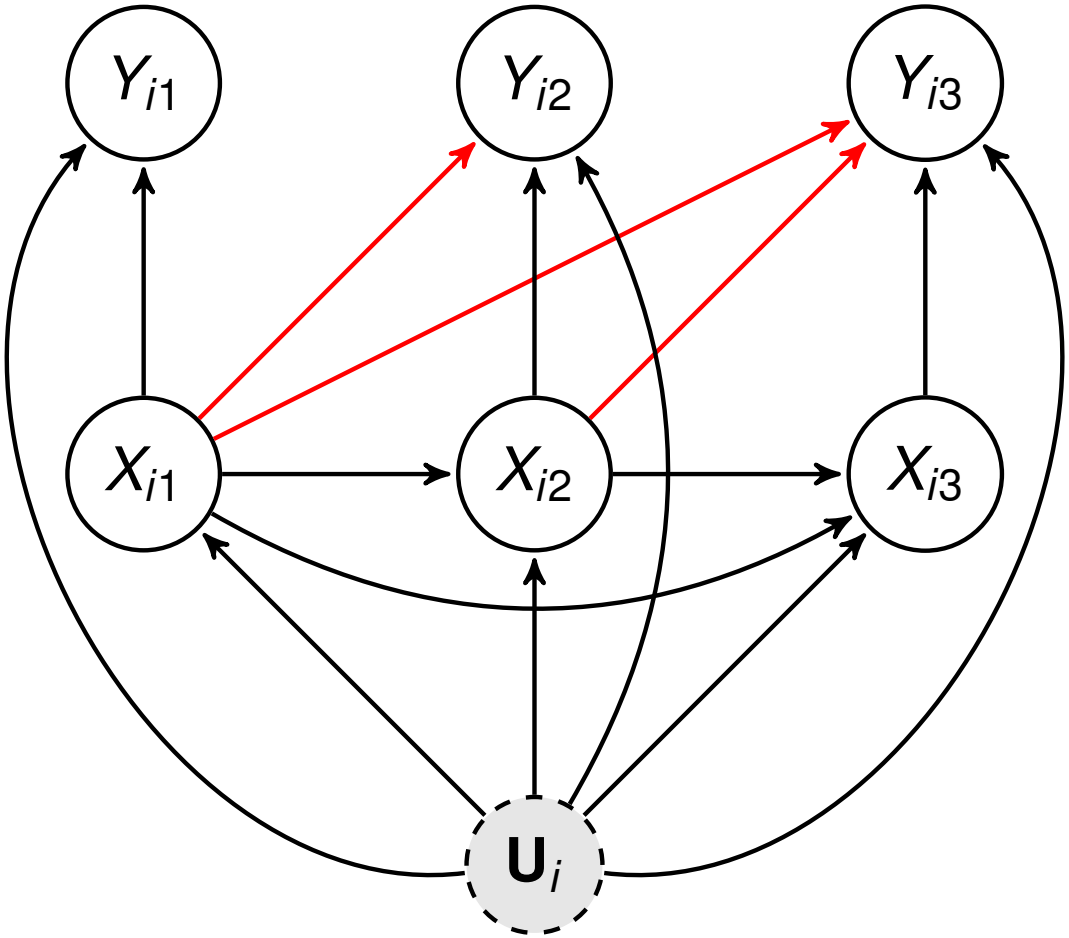
- "as-if random" assumption without conditioning on past outcomes
- Past outcomes cannot directly affect current treatment
- Says nothing about whether past outcomes can directly affect current outcome

Past Outcomes Directly Affect Current Outcome



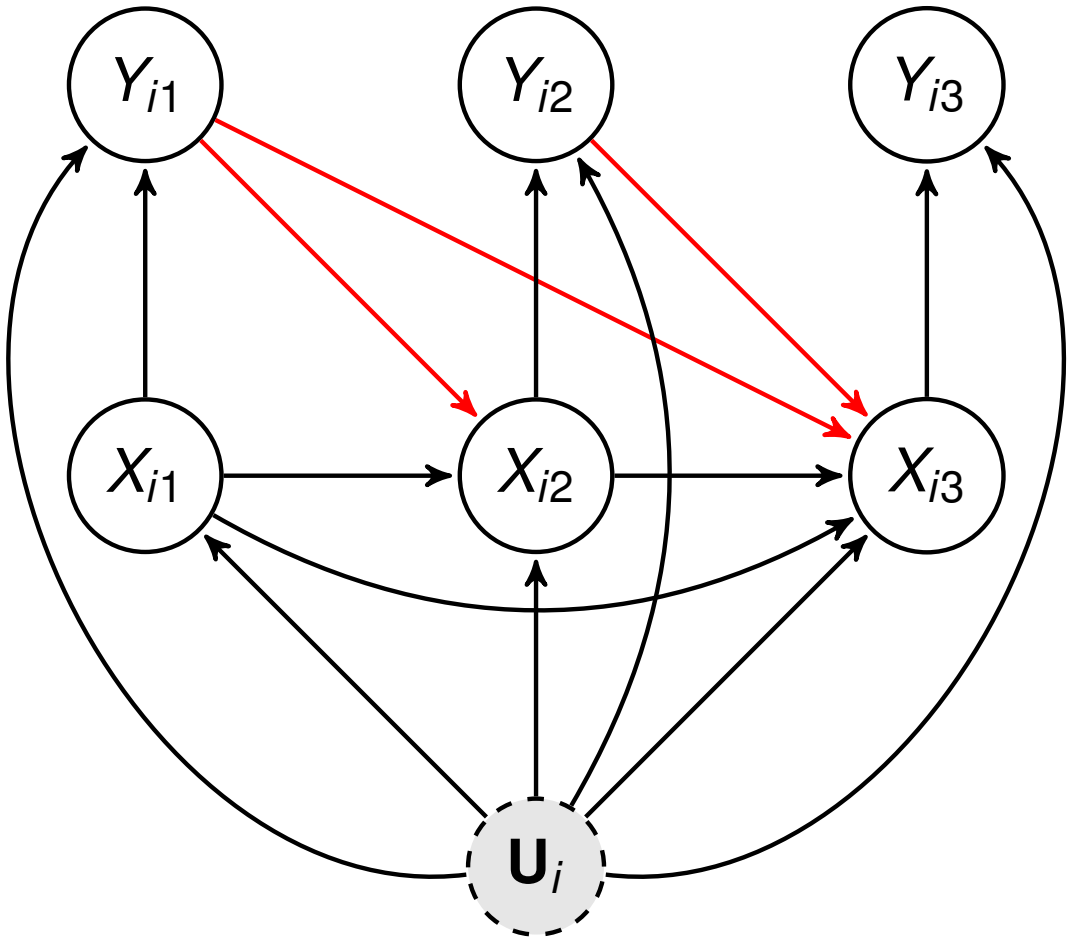
- Strict exogeneity still holds
- Past outcomes do not confound $X_{it} \longrightarrow Y_{it}$ given $\mathbf{U}_i$
- No need to adjust for past outcomes

Past Treatments Directly Affect Current Outcome



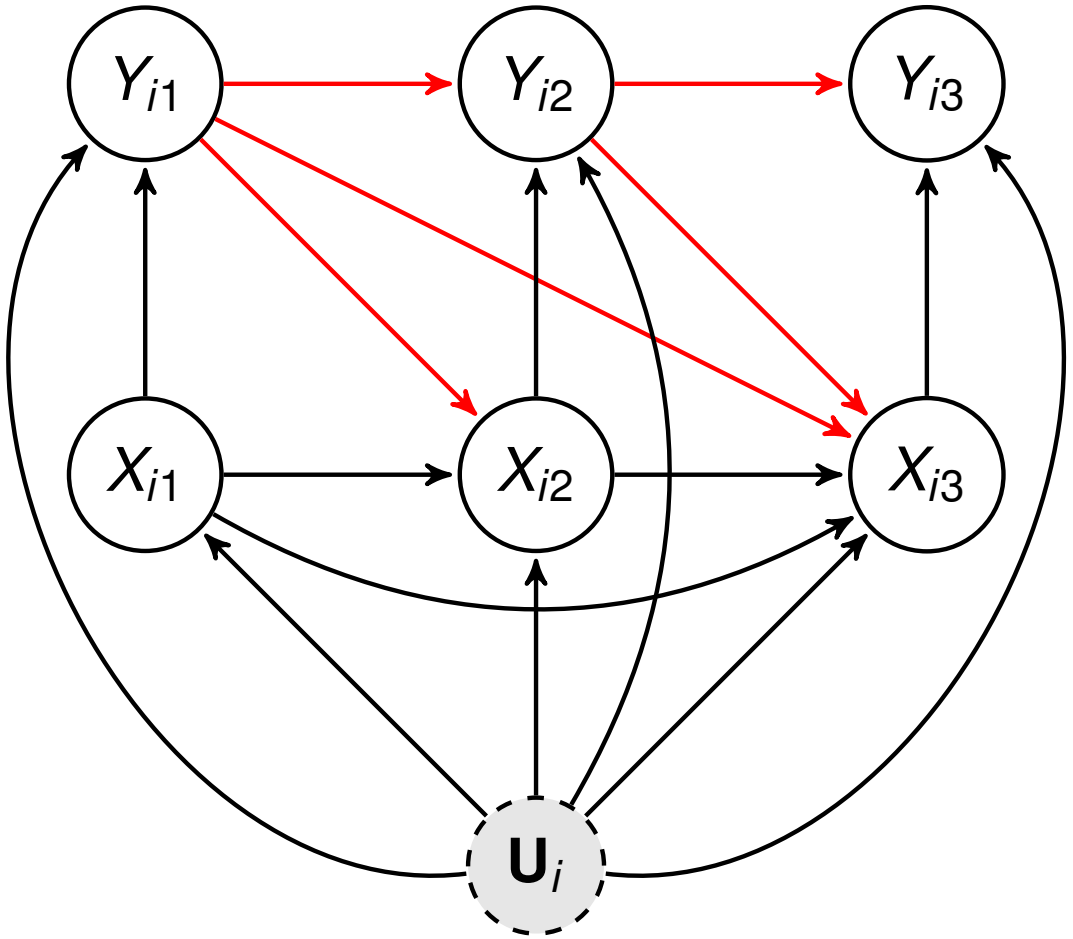
- Past treatments as confounders
- Need to adjust for past treatments
- Strict exogeneity holds given past treatments and $\mathbf{U}_i$
- Impossible to adjust for an entire treatment history and $\mathbf{U}_i$ at the same time
- Adjust for a small number of past treatments $\rightsquigarrow$ often arbitrary

Past Outcomes Directly Affect Current Treatment



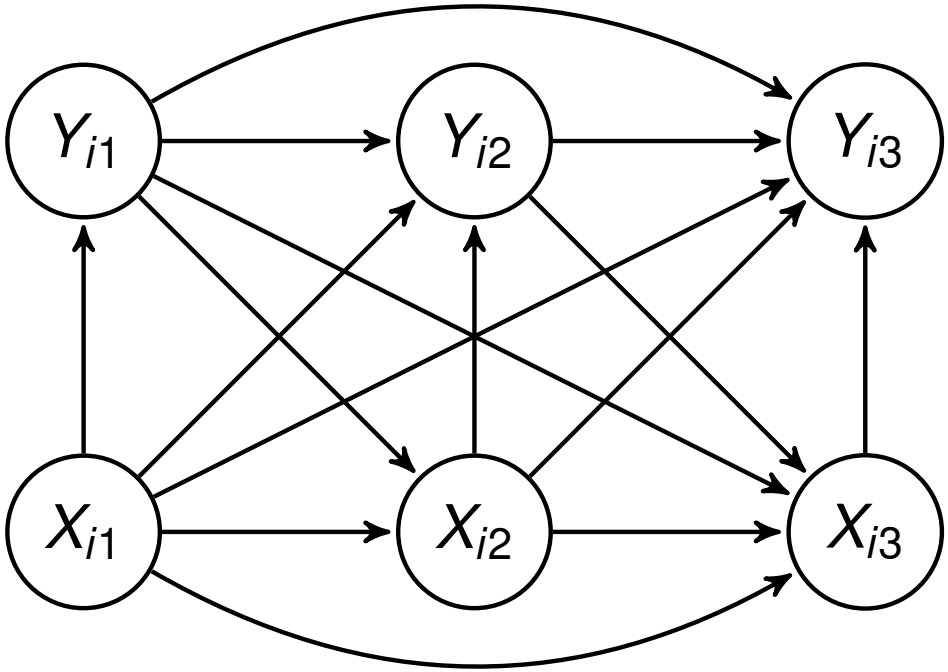
- Correlation between error term and future treatments
- Violation of strict exogeneity
- No adjustment is sufficient
- Together with the previous assumption $\rightsquigarrow$ no feedback effect over time

Instrumental Variables Approach



- Instruments: X_{i1}, X_{i2}, and Y_{i1}
- GMM: Arellano and Bond (1991)
- Exclusion restrictions
- Arbitrary choice of instruments
- Substantive justification rarely given

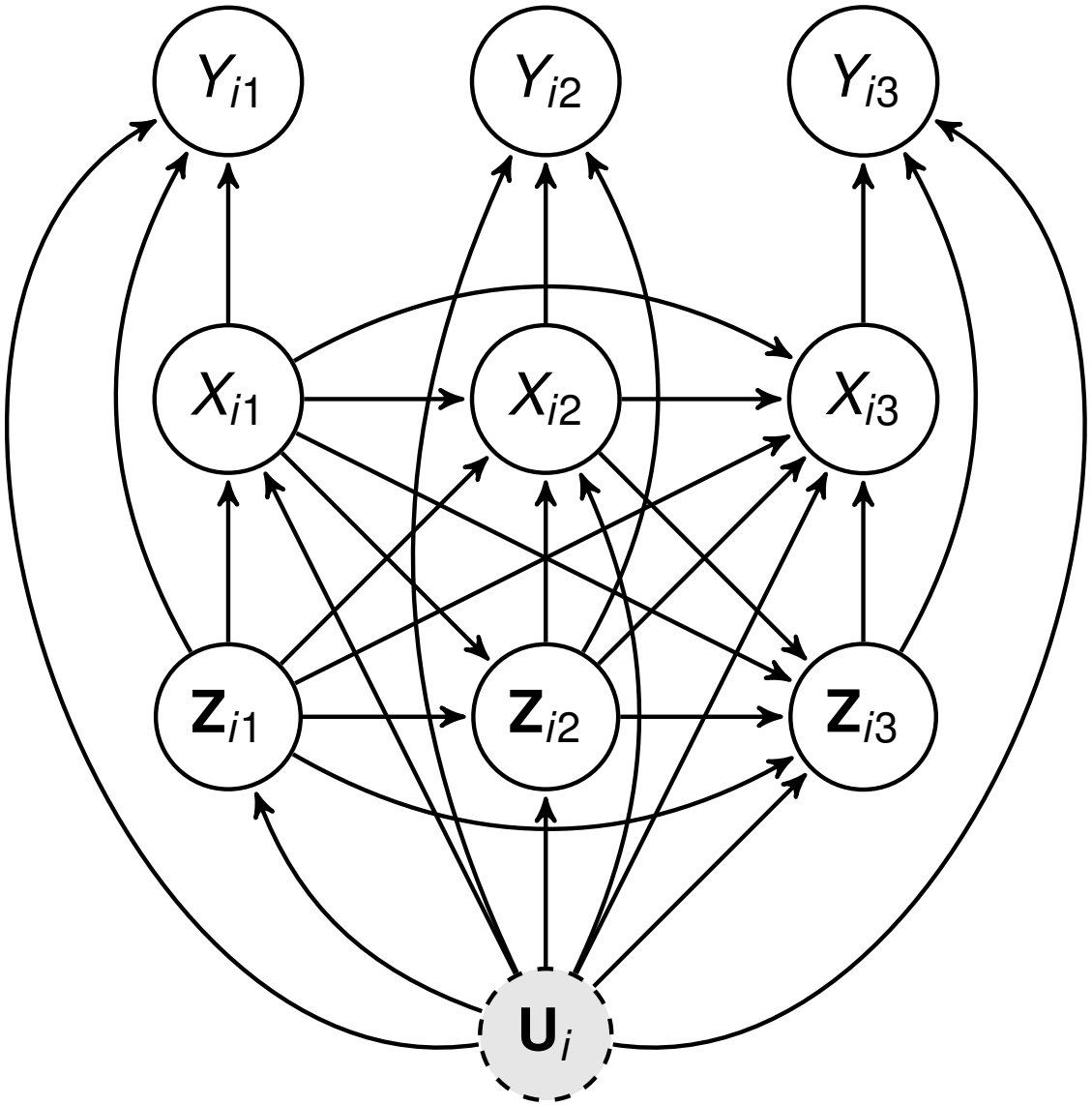
An Alternative Selection-on-Observables Approach



- Absence of unobserved time-invariant confounders $\mathbf{U}_i$
- past treatments can directly affect current outcome
- past outcomes can directly affect current treatment

- Comparison across units within the same time rather than across different time periods within the same unit
- Marginal structural models $\rightsquigarrow$ can identify the average effect of an entire treatment sequence
- Trade-off $\rightsquigarrow$ no free lunch

Adjusting for Observed Time-varying Confounders



- past treatments cannot directly affect current outcome
- past outcomes cannot directly affect current treatment
- adjusting for $\mathbf{Z}_{it}$ does not relax these assumptions
- past outcomes cannot *indirectly* affect current treatment through $\mathbf{Z}_{it}$

- Setup:
  - units: $i = 1, 2, \ldots, n$
  - time periods: $j = 1, 2, \ldots, J$
  - fixed J with $n \longrightarrow \infty$
  - time-varying binary treatments: $T_{ij} \in \{0, 1\}$
  - treatment history up to time j: $\overline{T}_{ij} = \{T_{i1}, T_{i2}, \ldots, T_{ij}\}$
  - time-varying confounders: X_{ij}
  - confounder history up to time j: $\overline{X}_{ij} = \{X_{i1}, X_{i2}, \ldots, X_{ij}\}$
  - outcome measured at time J: Y_i
  - potential outcomes: $Y_i(\bar{t}_J)$

- Assumptions:
  1. Sequential ignorability

$$Y_i(\bar{t}_J) \perp\!\!\!\perp T_{ij} \mid \overline{T}_{i,j-1} = \bar{t}_{j-1}, \overline{X}_{ij} = \bar{x}_j$$

     where $\bar{t}_J = (\bar{t}_{j-1}, t_j, \ldots, t_J)$
  2. Common support

$$0 < \Pr(T_{ij} = 1 \mid \overline{T}_{i,j-1}, \overline{X}_{ij}) < 1$$

Inverse-Probability-of-Treatment Weighting

- Weighting each observation via the inverse probability of its observed treatment sequence (Robins 1999)
- Potential weights:

$$
\begin{aligned}
w_i(\bar{t}_J, \overline{X}_{iJ}(\bar{t}_{J-1})) &= \frac{1}{P(\overline{T}_{iJ} = \bar{t}_J \mid \overline{X}_{iJ}(\bar{t}_{J-1}))} \\
&= \prod_{j=1}^{J} \frac{1}{P(T_{ij} = t_{ij} \mid \overline{T}_{i,j-1} = \bar{t}_{j-1}, \overline{X}_{ij}(\bar{t}_{j-1}))}
\end{aligned}
$$

- Stabilized potential weights:

$$
w_i^*(\bar{t}_J, \overline{X}_{iJ}(\bar{t}_{J-1})) = \frac{P(\overline{T}_{iJ} = \bar{t}_J)}{P(\overline{T}_{iJ} = \bar{t}_J \mid \overline{X}_{iJ}(\bar{t}_{J-1}))}
$$

- Observed weights: $w_i = w_i(\overline{T}_{iJ}, \overline{X}_{iJ})$ and $w_i^* = w_i^*(\overline{T}_{iJ}, \overline{X}_{iJ})$

Marginal Structural Models (MSMs)

- Consistent estimation of the marginal mean of potential outcome:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\overline{T}_{iJ} = \bar{t}_J\} w_i Y_i \xrightarrow{p} \mathbb{E}(Y_i(\bar{t}_J))$$

- In practice, researchers fit a weighted regression of Y_i on a function of $\overline{T}_{iJ}$ with regression weight w_i
- Adjusting for $\overline{X}_{iJ}$ leads to post-treatment bias
- MSMs estimate the average effect of any treatment sequence
- A pedagogical introduction with political science application: Blackwell (2013, AJPS)

- MSMs are sensitive to the misspecification of treatment assignment model (typically a series of logistic regressions)
- The effect of misspecification can propagate across time periods
- CBPS: estimate MSM weights so that covariates are balanced

A Matching Framework

- Even if these assumptions are satisfied, the the unit fixed effects estimator is inconsistent for the ATE:

$$\hat{\beta}_{\text{FE}} \xrightarrow{p} \frac{\mathbb{E}\left\{ C_i \left(\frac{\sum_{t=1}^T X_{it} Y_{it}}{\sum_{t=1}^T X_{it}} - \frac{\sum_{t=1}^T (1-X_{it}) Y_{it}}{\sum_{t=1}^T 1-X_{it}} \right) S_i^2 \right\}}{\mathbb{E}(C_i S_i^2)} \neq \tau$$

  where $S_i^2 = \sum_{t=1}^T (X_{it} - \overline{X}_i)^2/(T-1)$ is the unit-specific variance

- Key idea: comparison across time periods within the same unit
- The Within-unit matching estimator improves $\hat{\beta}_{\text{FE}}$ by relaxing the linearity assumption:

$$\hat{\tau}_{\text{match}} = \frac{1}{\sum_{i=1}^N C_i} \sum_{i=1}^N C_i \left(\frac{\sum_{t=1}^T X_{it} Y_{it}}{\sum_{t=1}^T X_{it}} - \frac{\sum_{t=1}^T (1-X_{it}) Y_{it}}{\sum_{t=1}^T (1-X_{it})} \right)$$

Constructing a General Matching Estimator

- $\mathcal{M}_{it}$: matched set for observation (i, t)
- For the within-unit matching estimator,

$$\mathcal{M}_{it}^{\text{match}} \quad = \quad \{(i', t') : i' = i, X_{i't'} = 1 - X_{it}\}$$

- A general matching estimator:

$$\hat{\tau}_{\text{match}} \quad = \quad \frac{1}{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} (\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)})$$

where $D_{it} = \mathbf{1}\{\#\mathcal{M}_{it} > 0\}$ and

$$\widehat{Y_{it}(x)} \quad = \quad \begin{cases} Y_{it} & \text{if } X_{it} = x \\ \frac{1}{\#\mathcal{M}_{it}} \sum_{(i',t') \in \mathcal{M}_{it}} Y_{i't'} & \text{if } X_{it} = 1 - x \end{cases}$$

Before-and-After Design

- No time trend for the average potential outcomes:

$$\mathbb{E}(Y_{it}(x) - Y_{i,t-1}(x) \mid X_{it} \neq X_{i,t-1}) \quad = \quad 0 \quad \text{for} \quad x = 0, 1$$

  with the quantity of interest $\mathbb{E}(Y_{it}(1) - Y_{it}(0) \mid X_{it} \neq X_{i,t-1})$

- Or just the average potential outcome under the control condition

$$\mathbb{E}(Y_{it}(0) - Y_{i,t-1}(0) \mid X_{it} = 1, X_{i,t-1} = 0) \quad = \quad 0$$

- This is a matching estimator with the following matched set:

$$\mathcal{M}_{it}^{BA} \quad = \quad \{(i', t') : i' = i, t' \in \{t-1, t+1\}, X_{i't'} = 1 - X_{it}\}$$

- It is also the first differencing estimator:

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

- "We emphasize that the model and the interpretation of β are *exactly* as in [the linear fixed effects model]. What differs is our method for estimating β" (Wooldridge; italics original).

- The identification assumptions is very different
- Slightly relaxing the assumption of no carryover effect
- But, still requires the assumption that past outcomes do not affect current treatment
- Regression toward the mean: suppose that the treatment is given when the previous outcome takes a value greater than its mean

Matching as a Weighted Unit Fixed Effects Estimator

- *Any* within-unit matching estimator can be written as a weighted unit fixed effects estimator with different regression weights
- The proposed within-matching estimator:

$$\hat{\tau}_{\text{match}} = \hat{\beta}_{\text{WFE}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} W_{it} \{ (Y_{it} - \overline{Y}_i^*) - \beta (X_{it} - \overline{X}_i^*) \}^2$$

where $\overline{X}_i^*$ and $\overline{Y}_i^*$ are unit-specific weighted averages, and

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

- We show how to construct regression weights for different matching estimators (i.e., different matched sets)
- Idea: count the number of times each observation is used for matching

- Benefits:
  - computational efficiency
  - model-based standard errors
  - robustness $\rightsquigarrow$ matching estimator is consistent even when linear unit fixed effects regression is the true model
  - specification test (White 1980) $\rightsquigarrow$ null hypothesis: linear fixed effects regression is the true model

Linear Regression with Unit and Time Fixed Effects

- Model:

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

where γ_t flexibly adjusts for a vector of unobserved unit-invariant time effects $\mathbf{V}_t$, i.e., $\gamma_t = f(\mathbf{V}_t)$

- Estimator:

$$\hat{\beta}_{\text{FE2}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \{(Y_{it} - \overline{Y}_i - \overline{Y}_t + \overline{Y}) - \beta(X_{it} - \overline{X}_i - \overline{X}_t + \overline{X})\}^2$$

where $\overline{Y}_t$ and $\overline{X}_t$ are time-specific means, and $\overline{Y}$ and $\overline{X}$ are overall means

Understanding the Two-way Fixed Effects Estimator

- β_{FE}: bias due to time effects
- β_{FEtime}: bias due to unit effects
- β_{pool}: bias due to both time and unit effects

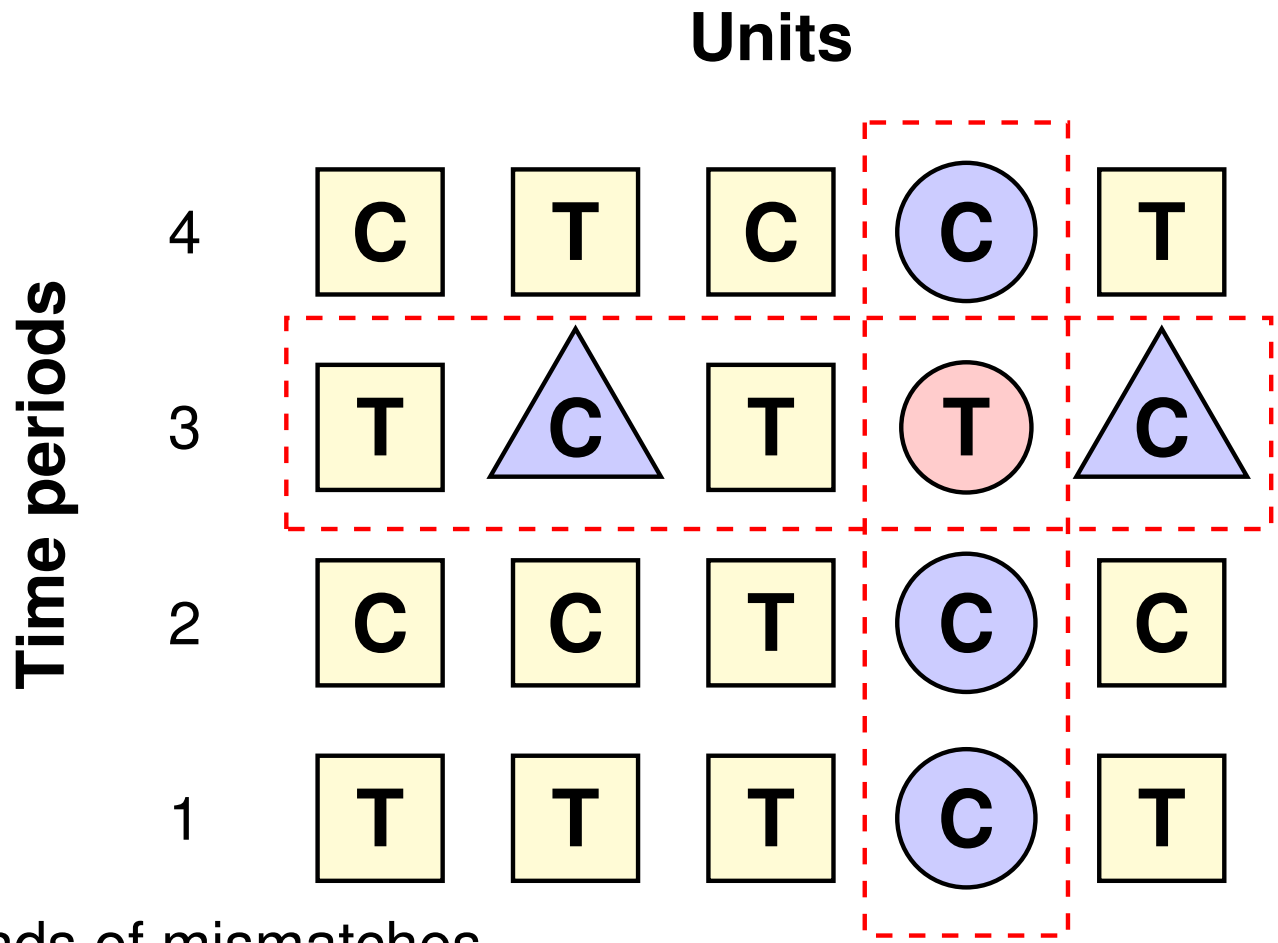
$$\hat{\beta}_{\text{FE2}} = \frac{\omega_{\text{FE}} \times \hat{\beta}_{\text{FE}} + \omega_{\text{FEtime}} \times \hat{\beta}_{\text{FEtime}} - \omega_{\text{pool}} \times \hat{\beta}_{\text{pool}}}{w_{\text{FE}} + w_{\text{FEtime}} - w_{\text{pool}}}$$

with sufficiently large N and T, the weights are given by,

$$\begin{aligned} \omega_{\text{FE}} &\approx \mathbb{E}(S_i^2) = \text{average unit-specific variance} \\ \omega_{\text{FEtime}} &\approx \mathbb{E}(S_t^2) = \text{average time-specific variance} \\ \omega_{\text{pool}} &\approx S^2 = \text{overall variance} \end{aligned}$$

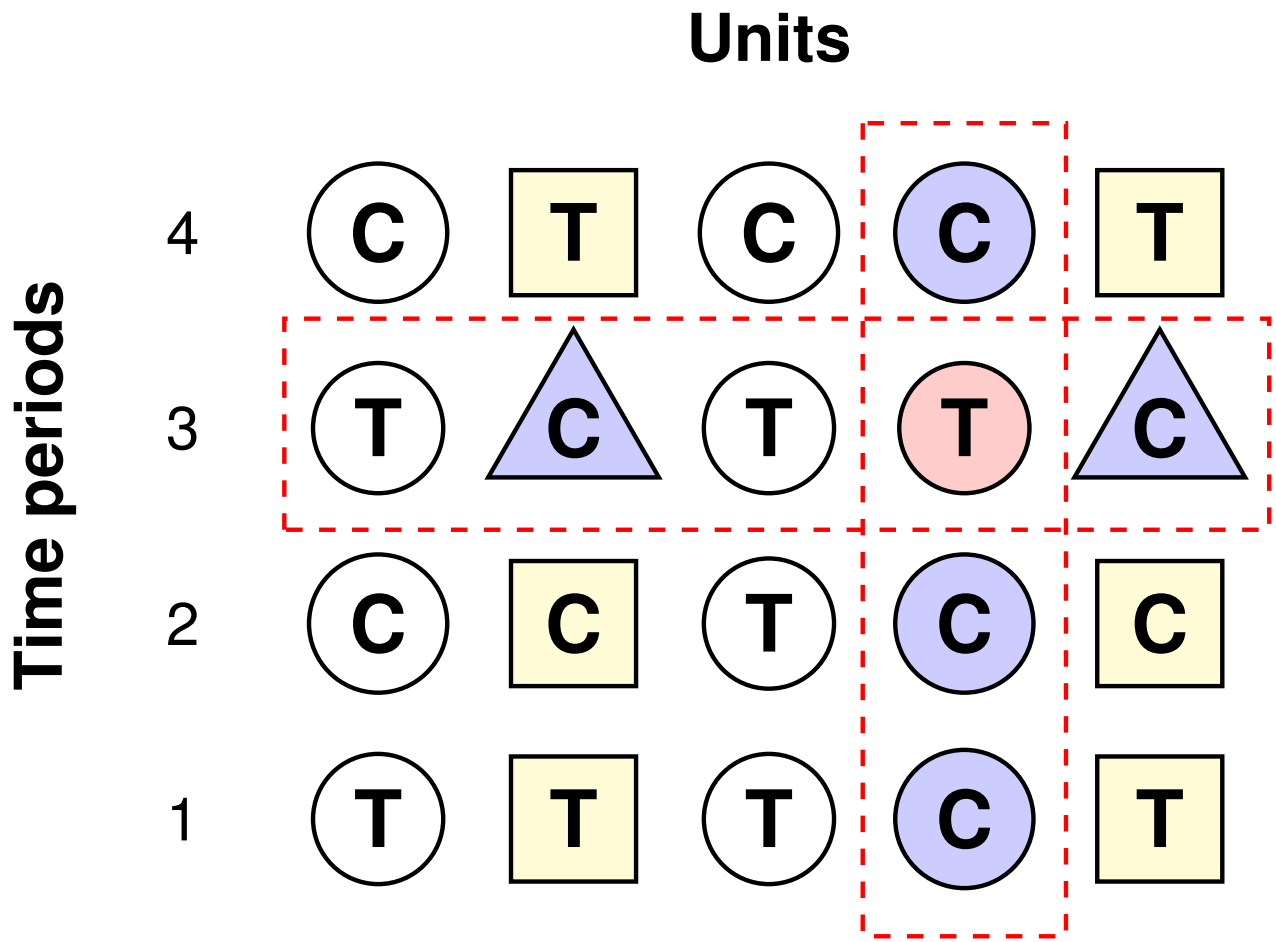
Matching and Two-way Fixed Effects Estimators

- Problem: No other unit shares the same unit and time



- Two kinds of mismatches
  1. Same treatment status
  2. Neither same unit nor same time

We Can Never Eliminate Mismatches

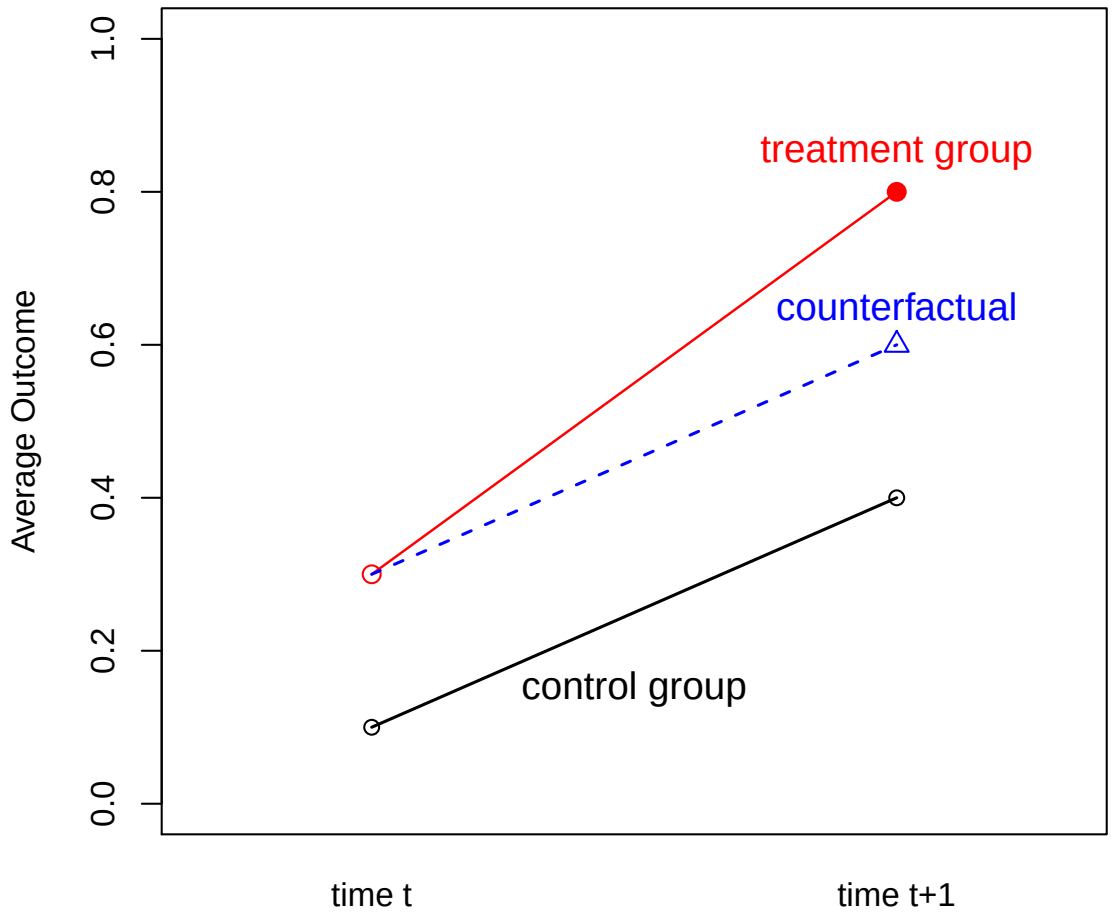


- To cancel time and unit effects, we must induce mismatches
- No weighted two-way fixed effects model eliminates mismatches

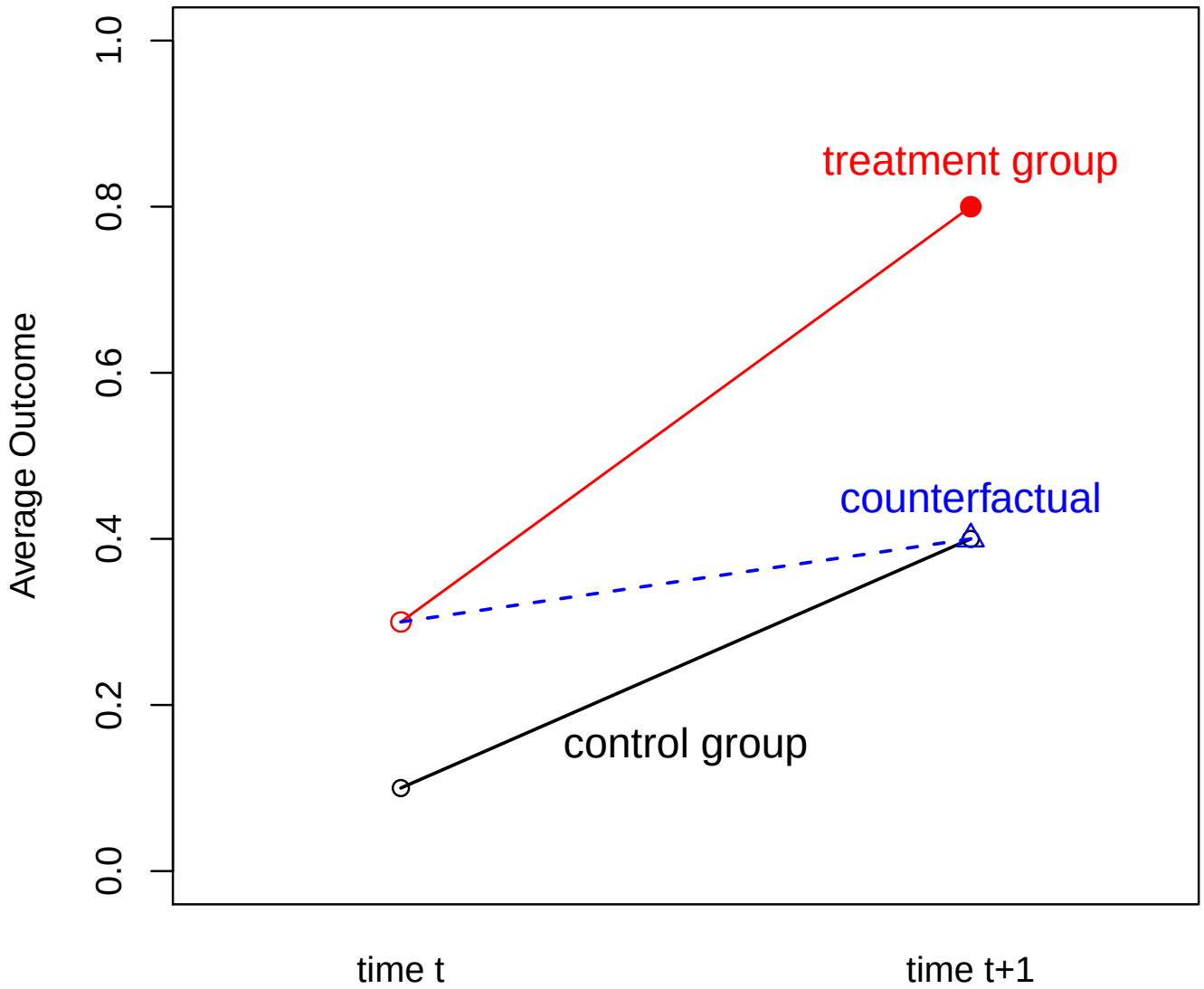
Difference-in-Differences Design

- Parallel trend assumption:

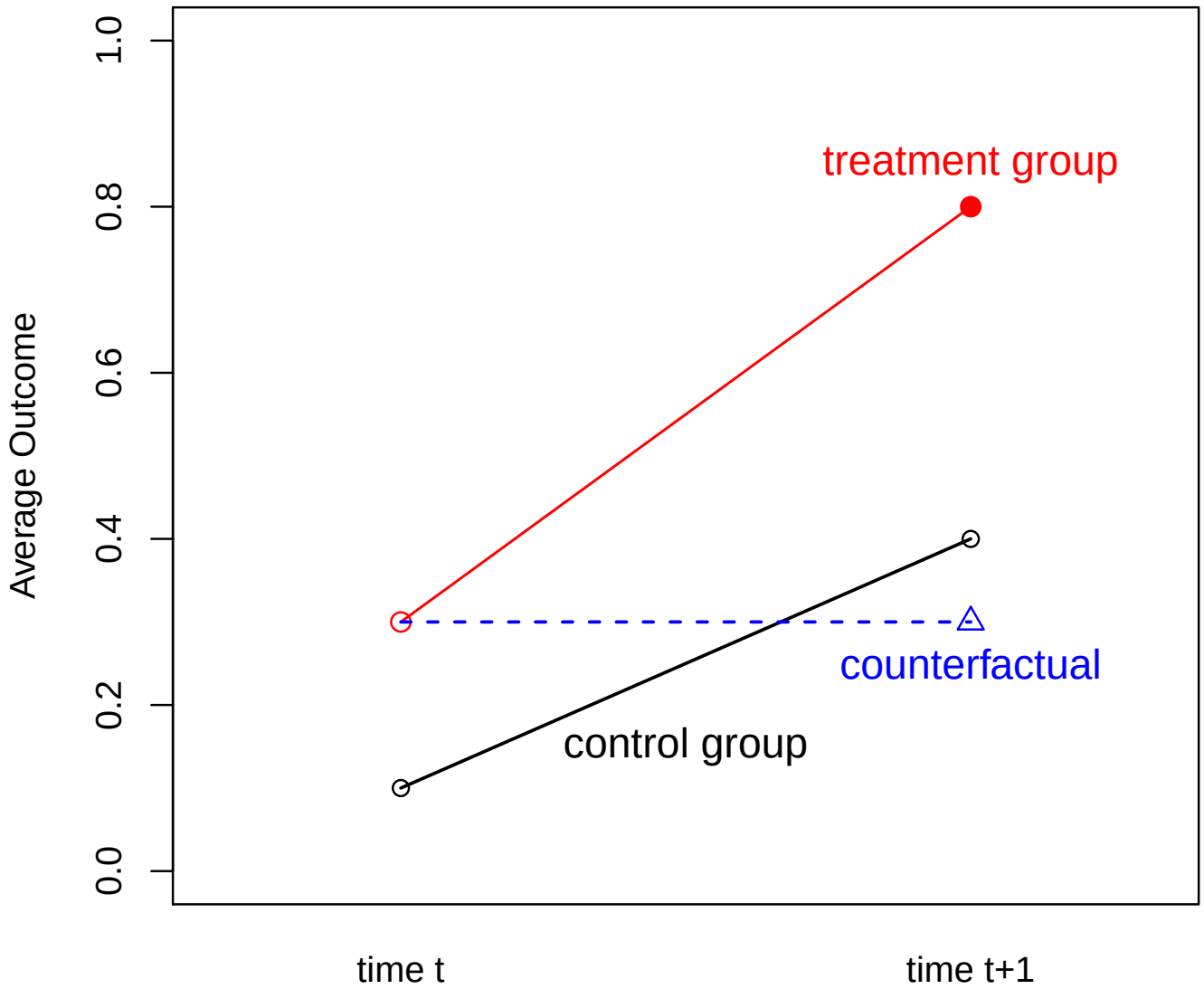
$$\begin{aligned} & \mathbb{E}(Y_{it}(0) - Y_{i,t-1}(0) \mid X_{it} = 1, X_{i,t-1} = 0) \\ = \; & \mathbb{E}(Y_{it}(0) - Y_{i,t-1}(0) \mid X_{it} = X_{i,t-1} = 0) \end{aligned}$$



Cross Section Comparison



Before-and-After Design



Formal Treatment of DiD in 2 Time-Period Case

- Two time periods: time 0 (pre-treatment) and time 1 (post-treatment)
- G_i: treatment group membership
- $Z_{ti} = tG_i$: treatment assignment indicator for $t = 0, 1$
- Potential outcomes: $Y_{0i}(0)$, $Y_{0i}(1)$, $Y_{1i}(0)$, $Y_{1i}(1)$
- Observed outcomes: $Y_{ti} = Z_{ti} Y_{ti}(1) + (1 - Z_{ti}) Y_{ti}(0)$
- Average treatment effect for the treated:

$$\tau \;=\; \mathbb{E}\{ Y_{1i}(1) - Y_{1i}(0) \mid G_i = 1\}$$

- Exogeneity assumption: $Y_{1i}(0) \perp\!\!\!\perp Z_{ti} \mid X_i, Y_{0i}$
- Difference-in-Differences assumption:

$$\mathbb{E}(Y_{1i}(0) - Y_{0i}(0) \mid G_i = 1, X_i) = \mathbb{E}(Y_{1i}(0) - Y_{0i}(0) \mid G_i = 0, X_i)$$

- DiD estimator:

$$\hat{\tau} \;=\; \{\mathbb{E}(Y_{1i} \mid G_i = 1) - \mathbb{E}(Y_{0i} \mid G_i = 1)\} - \{\mathbb{E}(Y_{1i} \mid G_i = 0) - \mathbb{E}(Y_{0i} \mid G_i = 0)\}$$

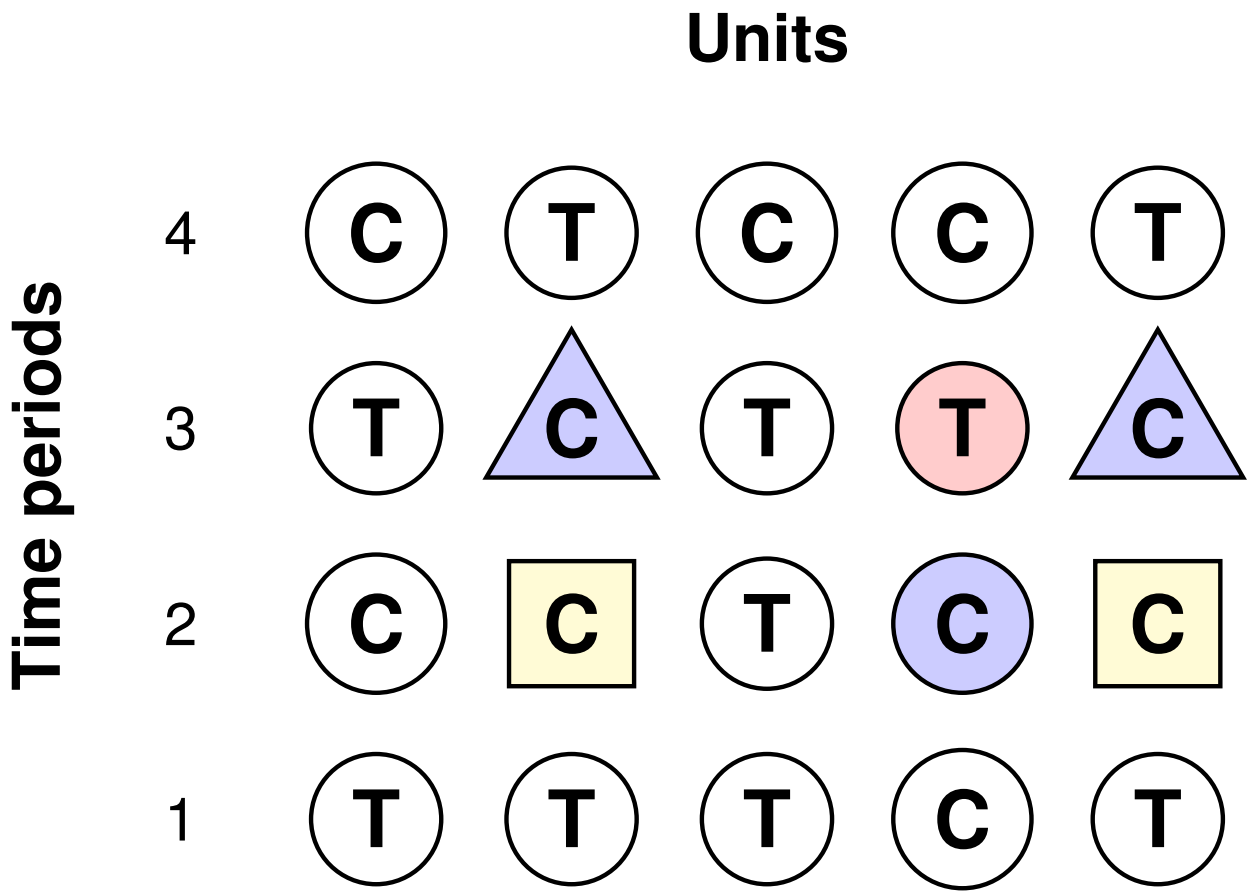
- Can be applied to repeated cross-section data

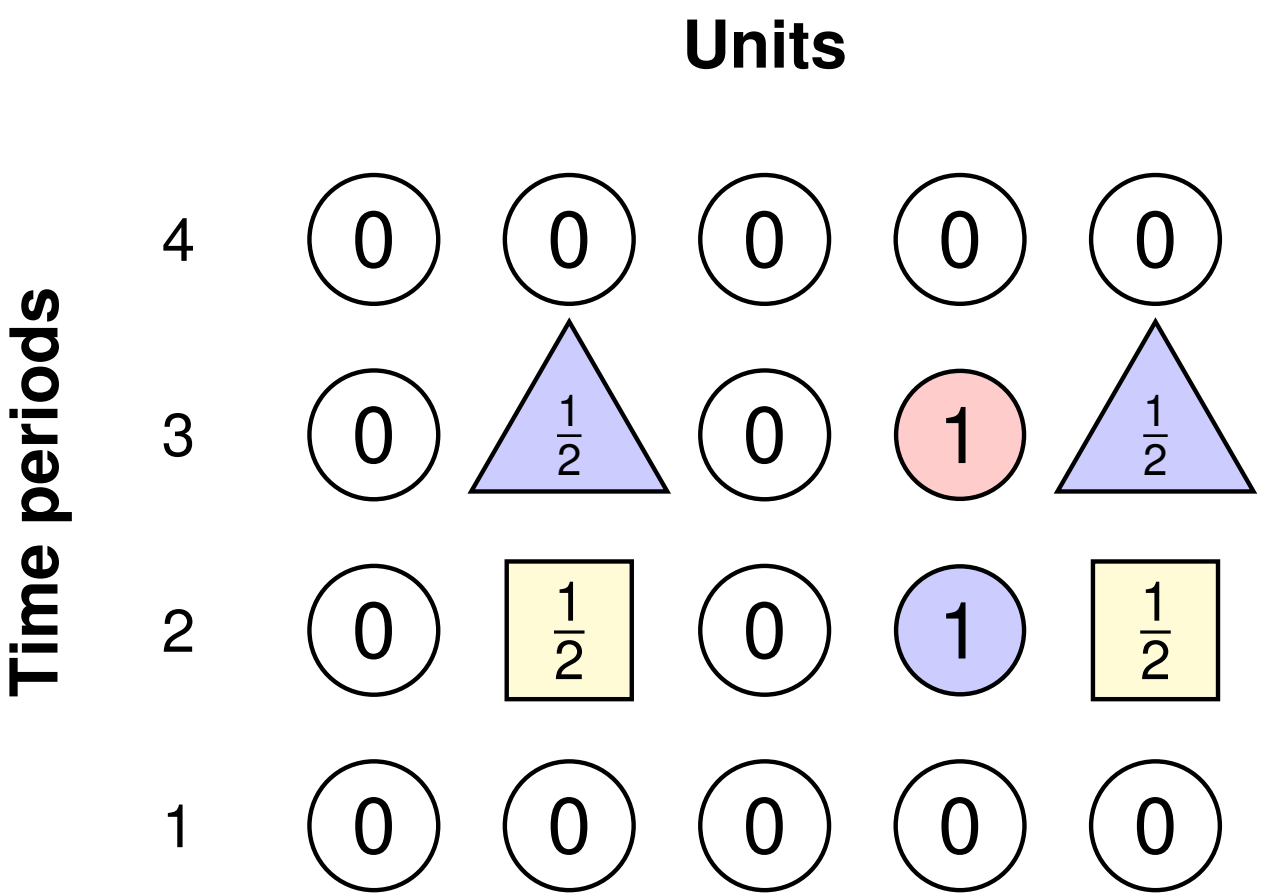
Linear Model in 2 Time-Period Case

- In this case, the standard two-way fixed effects model is suffice
- Model: $Y_{ti}(z) = \alpha_i + \beta z + \gamma t + \epsilon_{ti}$
  - $Y_{0i}(0) = \alpha_i + \epsilon_{0i}$
  - $Y_{1i}(0) = \alpha_i + \gamma + \epsilon_{1i}$
  - $Y_{1i}(1) = \alpha_i + \beta + \gamma + \epsilon_{1i}$
- Assumption: $\mathbb{E}(Y_{1i}(0) - Y_{0i}(0) \mid G_i = g) = \gamma$
- Or equivalently $\mathbb{E}(\epsilon_{1i} - \epsilon_{0i} \mid G_i = g) = 0$
- Both Z_{ti} and ϵ_{ti} can depend on α_i
- Neither stronger or weaker than the standard exogeneity assumption
- When Y_{0i} is balanced, they are equivalent

General DiD = Weighted Two-Way FE Effects

- 2×2: equivalent to linear two-way fixed effects regression
- General setting: Multiple time periods, repeated treatments





- Fast computation, standard error, specification test
- Still assumes that past outcomes don't affect current treatment
- Baseline outcome difference $\rightsquigarrow$ caused by unobserved time-invariant confounders
- It should not reflect causal effect of baseline outcome on treatment assignment

Synthetic Control Method (Abadie et al. 2010)

- One treated unit i^* receiving the treatment at time T
- Quantity of interest: $Y_{i^*T} - Y_{i^*T}(0)$
- Create a synthetic control using past outcomes
- Weighted average: $\widehat{Y_{i^*T}(0)} = \sum_{i \neq i^*} \hat{w}_i Y_{iT}$
- Estimate weights to balance past outcomes and past time-varying covariates

- A motivating autoregressive model:

$$\begin{aligned} Y_{iT}(0) &= \rho_T Y_{i,T-1}(0) + \delta_T^\top \mathbf{Z}_{iT} + \epsilon_{iT} \\ \mathbf{Z}_{iT} &= \lambda_{T-1} Y_{i,T-1}(0) + \Delta_T \mathbf{Z}_{i,T-1} + \nu_{iT} \end{aligned}$$

- Past outcomes can affect current treatment
- No unobserved time-invariant confounders

Causal Effect of ETA's Terrorism

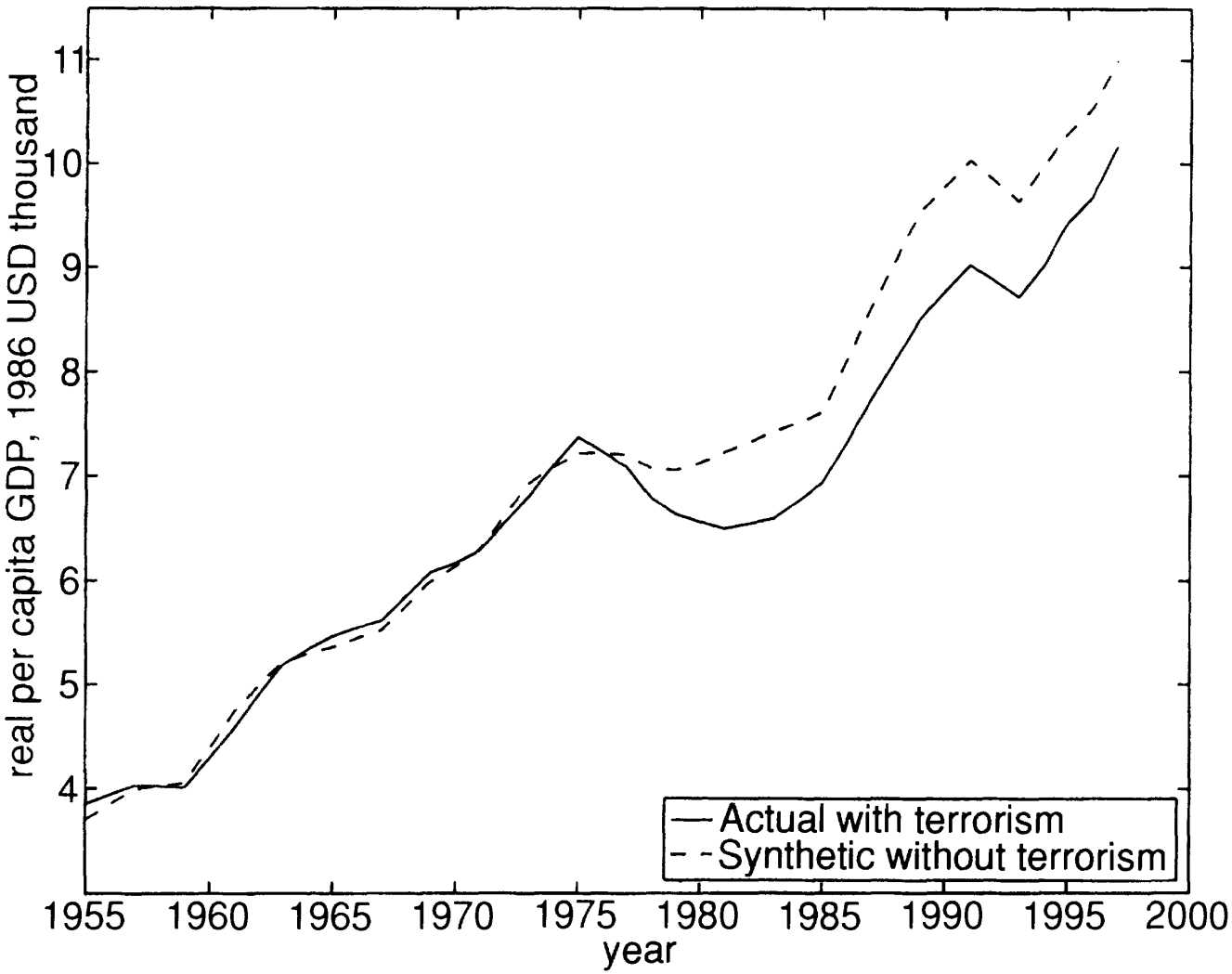


FIGURE 1. PER CAPITA GDP FOR THE BASQUE COUNTRY

Abadie and Gardeazabal (2003, AER)

- The main motivating model:

$$Y_{it}(0) \quad = \quad \gamma_t + \delta_t^\top \mathbf{Z}_{it} + \xi^\top \mathbf{U}_i + \epsilon_{it}$$

- A generalization of the linear two-way fixed effects model
- How is it possible to adjust for unobserved time-invariant confounders by adjusting for past outcomes?

- The key assumption: there exist weights such that

$$\sum_{i \neq i^*} w_i \mathbf{Z}_{it} = \mathbf{Z}_{i^* t} \text{ for all } t \leq T-1 \quad \text{and} \quad \sum_{i \neq i^*} w_i \mathbf{U}_i = \mathbf{U}_{i^*}$$

- In general, adjusting for observed confounders does not adjust for unobserved confounders

- The same tradeoff as before

Placebo Test

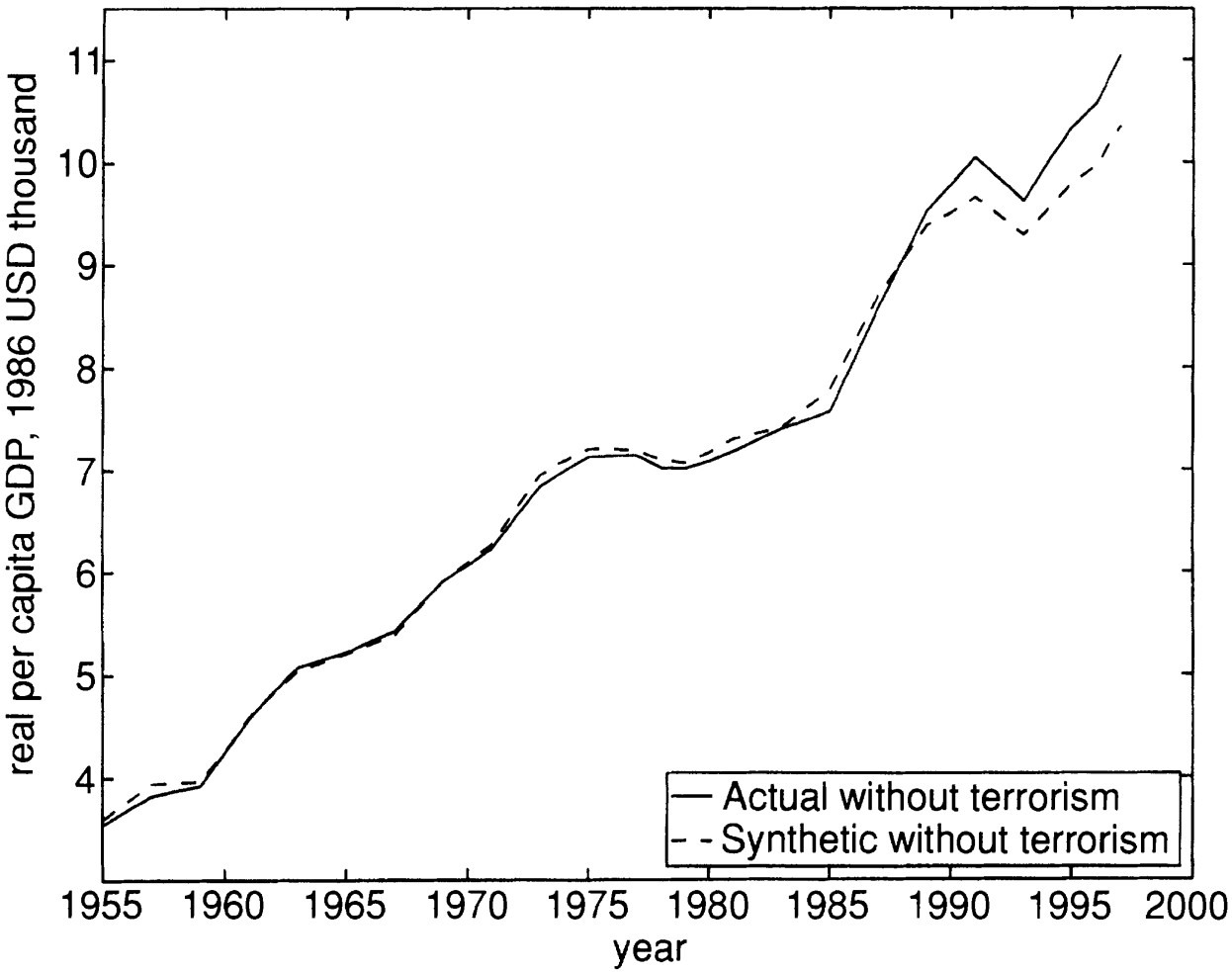


FIGURE 4. A "PLACEBO STUDY," PER CAPITA GDP FOR CATALONIA

can do this for all control units and compare them with the treated unit

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"

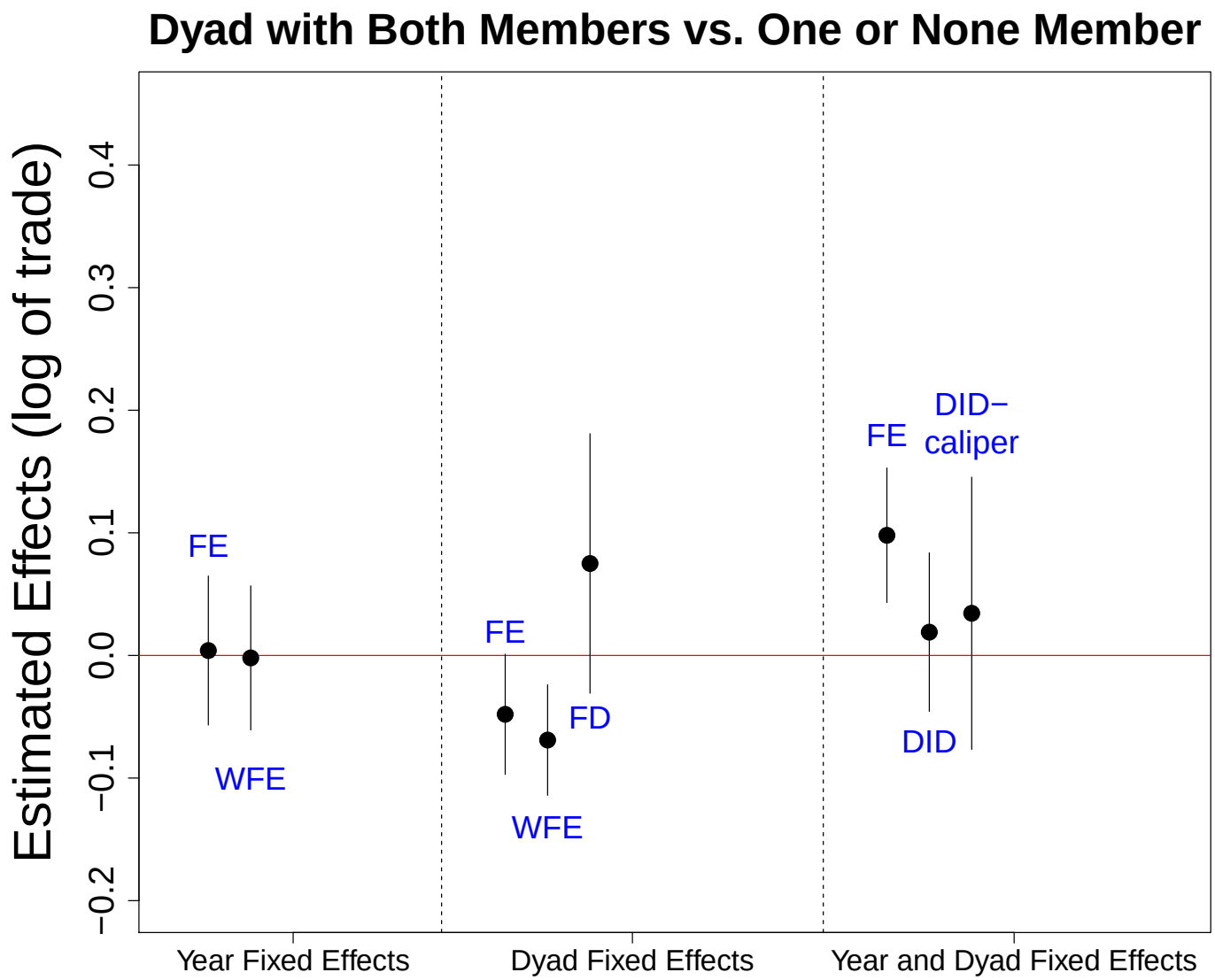
Data and Methods

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:

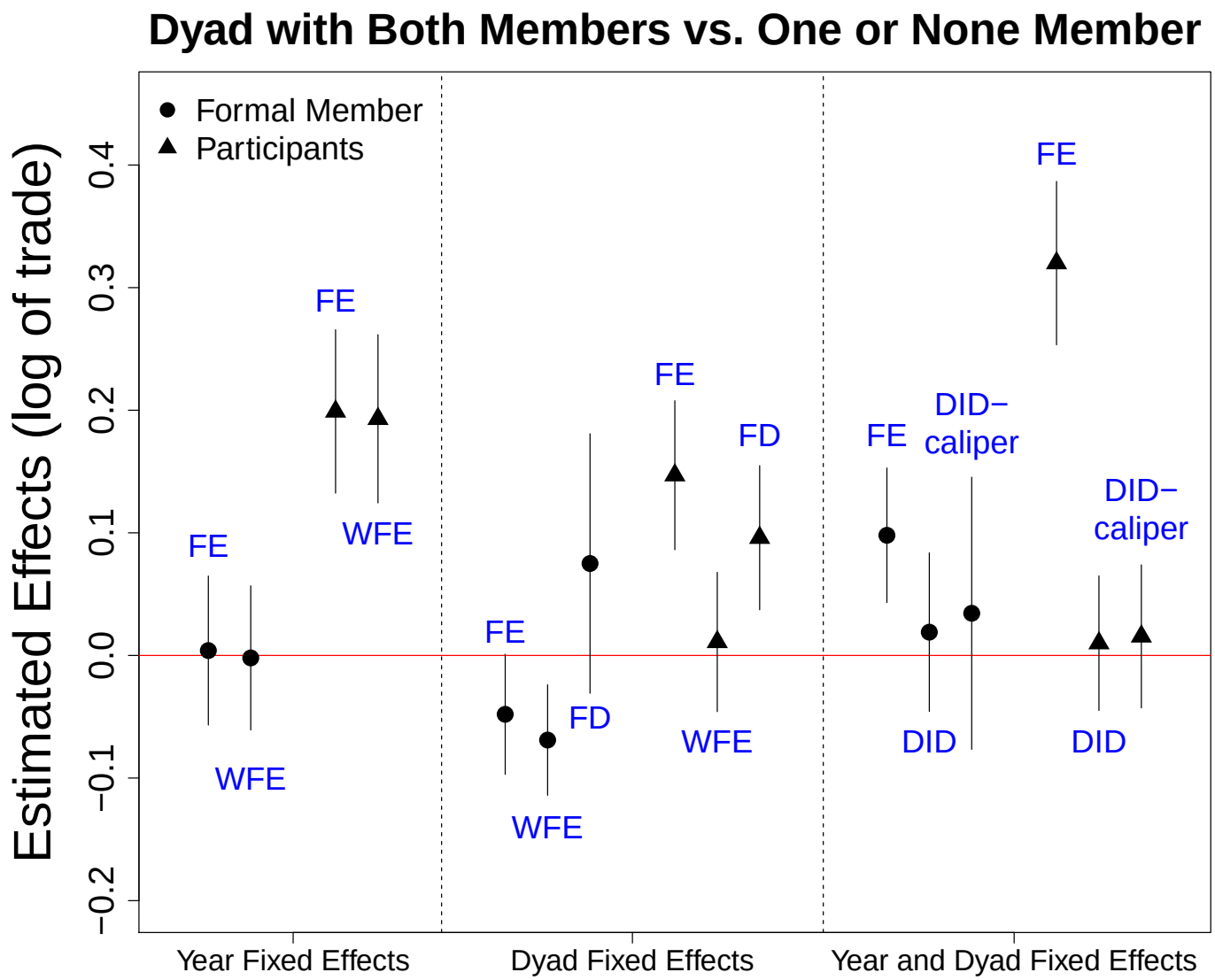
$$\ln Y_{it} = \alpha_t + \beta X_{it} + \delta^\top \mathbf{Z}_{it} + \epsilon_{it}$$

   - Y_{it}: trade volume
   - X_{it}: membership (formal/participants) Both vs. At most one
   - $\mathbf{Z}_{it}$: 15 dyad-varying covariates (e.g., log product GDP)
3. Assumptions:
   - past membership status doesn't directly affect current trade volume
   - past trade volume doesn't affect current membership status
   - Before-and-after $\rightsquigarrow$ increasing trend in trade volume
   - Difference-in-differences after conditional on past outcome?

Empirical Results: Formal Membership



Empirical Results: Participants Included



Connection to Missing Data

- Causal inference is a missing data problem
- How should we handle missing data in general?

- Observed data D_{obs}, missing data D_{mis}, "missingness" indicator R
- Missing data mechanisms (Rubin):
  1. Missing completely at random (MCAR)

$$P(R \mid D_{obs}, D_{mis}, \xi) \;=\; P(R \mid \xi)$$

  2. Missing at random (MAR)

$$P(R \mid D_{obs}, D_{mis}, \xi) \;=\; P(R \mid D_{obs}, \xi)$$

- MAR is analogous to unconfoundedness in causal inference

Likelihood Inference under Ignorability

- Ignorability
  1. MAR
  2. distinctness between the compelete data model parameters θ in $P(D \mid \theta)$ and missingness mechanism model parameters ξ

- Observed-data Likelihood:

$$\begin{aligned} P(R, D_{obs} \mid \theta, \xi) &= \int P(R, D \mid \theta, \xi) dD_{mis} \\ &= \int P(R \mid D, \xi) P(D \mid \theta) dD_{mis} \\ &= P(R \mid D_{obs}, \xi) \int P(D \mid \theta) dD_{mis} \\ &= P(R \mid D_{obs}, \xi) P(D_{obs} \mid \theta) \end{aligned}$$

you can ignore missing data under ignorability!

Handling Missing Data

- List-wise deletion: at best inefficient, most likely biased
- Weighting by modeling $P(R \mid D_{obs}, \xi)$
- Regression models $D = (Y, X)$:
  1. missing outcome $\Longrightarrow$ model $P(Y_{obs} \mid X, \theta)$ under ignorability
  2. missing predictors $\Longrightarrow$ model $P(Y \mid X_{obs}, R, \theta^*)$
- Imputation:
  1. Single imputation: mean, regression, etc.
  2. Hot-deck imputation: connection to matching
  3. Multiple imputation: modeling $P(D \mid \theta)$ and impute D_{mis}
     - EM and MCMC algorithms: `Amelia` and `mi` packages
     - M complete data sets: $D_1, D_2, \ldots, D_M$
     - Analyze each data set and combine the results: Bootstrap and (quasi/pure) Bayesian Monte Carlo simulation are easier

$$\hat{\phi} = \frac{1}{M}\sum_{i=1}^{M}\hat{\phi}_i, \quad \mathrm{Var}(\hat{\phi}) = \underbrace{\frac{1}{M}\sum_{i=1}^{M}\mathrm{Var}(\hat{\phi}_i)}_{within-imputation} + \underbrace{\left(1+\frac{1}{M}\right)}_{finite\ sample\ adjustment} \underbrace{\frac{1}{M-1}\sum_{i=1}^{M}(\hat{\phi}_i - \hat{\phi})^2}_{between-imputation}$$

Concluding Remarks

- Matching methods do:
  - make causal assumptions transparent by identifying counterfactuals
  - make regression models robust by reducing model dependence
- Matching methods cannot solve endogeneity
- Only good research design can overcome endogeneity
- Recent advances in matching methods
  - directly optimize balance
  - the same idea applied to propensity score
- Weighting methods generalize matching methods
  - Sensitive to propensity score model specification
  - Robust estimation of propensity score model
- Causal inference with panel data
  - Fixed effects regression ignores dynamics
  - Difference-in-differences and synthetic control methods
  - Marginal structural models for dynamic treatment regimes