Matching and Weighting Methods for Causal Inference

Kosuke Imai

Department of Politics
Center for Statistics and Machine Learning
Princeton University

May 25 – 26, 2016
Health Economic Forum
Uppsala University

Introduction

Matching and Weighting

- What is "matching"?
- Grouping observations based on their observed characteristics
  1. pairing
  2. subclassification
  3. subsetting

- What is "weighting"?
- Replicating observations based on their observed characteristics
- All types of matching are special cases with discrete weights

- What matching and weighting methods can do: flexible and robust causal modeling under selection on observables
- What they cannot do: eliminate bias due to unobserved confounding

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 $Y_i = Y_i(T_i)$

Patients	Treatment	Survival		Age	Gender
i	T_i	$Y_i(1)$	$Y_i(0)$	X_i	X_i
1	1	1	?	20	F
2	0	?	0	55	M
3	0	?	1	40	M
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	1	0	?	62	F

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

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

  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

Matching for Randomized Experiments

- Matching can be used for randomized experiments too!
- Randomization of treatment $\longrightarrow$ unbiased estimates
- Improving efficiency $\longrightarrow$ reducing variance
- Why care about efficiency? You care about your results!

- Randomized matched-pair design
- Randomized block design

- Intuition: estimation uncertainty comes from pre-treatment differences between treatment and control groups
- Mantra (Box, Hunter, and Hunter):

  "Block what you can and randomize what you cannot"

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 easier 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:

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

- Intracluster correlation coefficient ρ_t:

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

Cluster Standard Error

- Cluster robust "sandwich" variance estimator:

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

  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 non-randomly 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

Matching and Blocking for Randomized Experiments

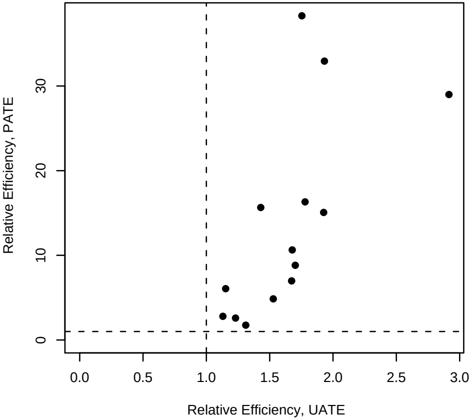
- Okay, but how should I match/block without the treatment group?
- Goal: match/block well on powerful predictors of outcome (prognostic factors)
- (Coarsened) Exact matching
- Matching based on a similarity measure:

$$\text{Mahalanobis distance} = \sqrt{(X_i - X_j)^\top \widehat{\Sigma}^{-1}(X_i - X_j)}$$

- Could combine the two

Relative Efficiency of Matched-Pair Design (MPD)

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



Cross-sectional 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?

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:

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

- Delta method is pain, but simulation is easy via Zelig

The Problem: Model Sensitivity in Causal Inference

- How most social scientists do empirical analysis:
  1. collect the data spending months or years
  2. finish recording and merging
  3. sit in front of your computer with nobody to bother you
  4. run one regression
  5. run another regression with different control variables
  6. run another regression with different functional forms
  7. run another regression with different measures
  8. run yet another regression with a subset of the data
  9. end up with 100 or 1000 *different* estimates
  10. put 5 regression results in the paper

- What's the problem?
  - "correct" specification is chosen after looking at the estimates
  - to readers of an article, it's never clear whether it represents a true test of an ex ante hypothesis or merely shows it's possible to find such results

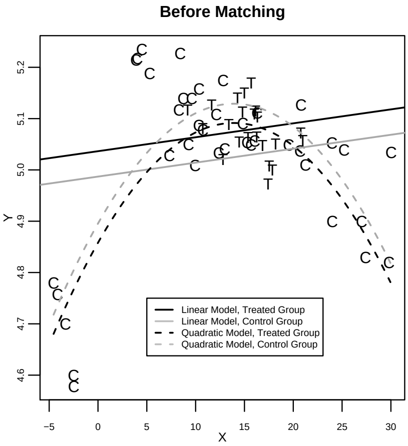
Matching as Nonparametric Preprocessing

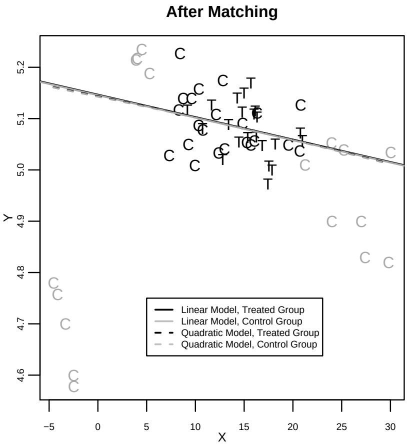
- READING: Ho *et al. Political Analysis* (2007)
- Assume exogeneity holds: matching does NOT solve endogeneity
- Need to model $\mathbb{E}(Y_i \mid T_i, X_i)$
- Parametric regression – functional-form/distributional assumptions $\Longrightarrow$ model dependence
- Non-parametric regression $\Longrightarrow$ 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 covariates
- Reduced model dependence: minimal role of statistical modeling

How Matching Reduces Model Dependence

- An artificial data set with one control variable
- Fit two regressions (with/without a quadratic term) before and after matching





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:

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

- The balancing property (no assumption):

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

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

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

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

Classical Matching Techniques

- Exact matching
- Mahalanobis distance matching: $\sqrt{(X_i - X_j)^\top \widehat{\Sigma}^{-1}(X_i - X_j)}$
- 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

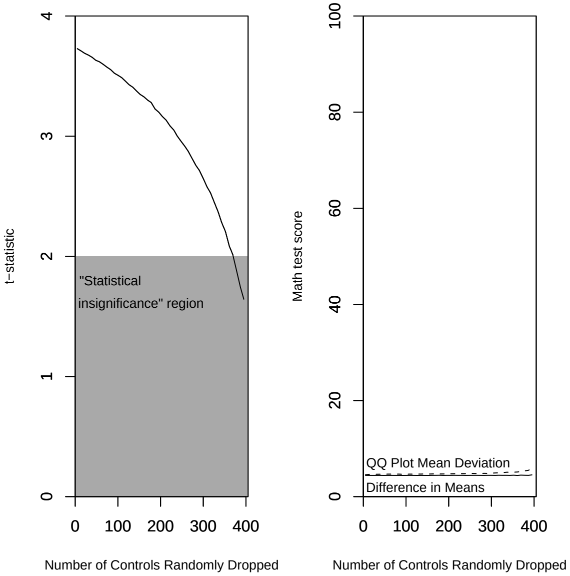
- FURTHER READING: Rubin (2006). *Matched Sampling for Causal Effects* (Cambridge UP)

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

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: minimize sum of distances
- Full matching: subclassification with variable strata size
- Genetic matching: maximize minimum p-value
- Coarsened exact matching: exact match on binned covariates
- SVM subsetting: find the largest, balanced subset for general treatment regimes
- Software: MatchIt implements various algorithms

- Another problem of matching: hard to balance in a small sample

Inverse Propensity Score Weighting

- Matching is inefficient because it throws away data
- Matching is a special case of weighting
- Weighting by inverse propensity score (Horvitz-Thompson):

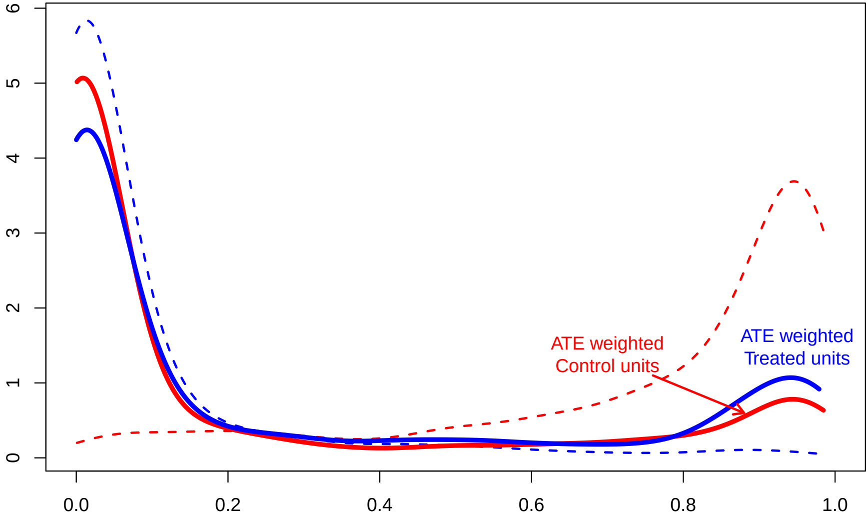
$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}\right)$$

- Unstable when some weights are extremely small
- An improved weighting scheme with normalized weights:

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

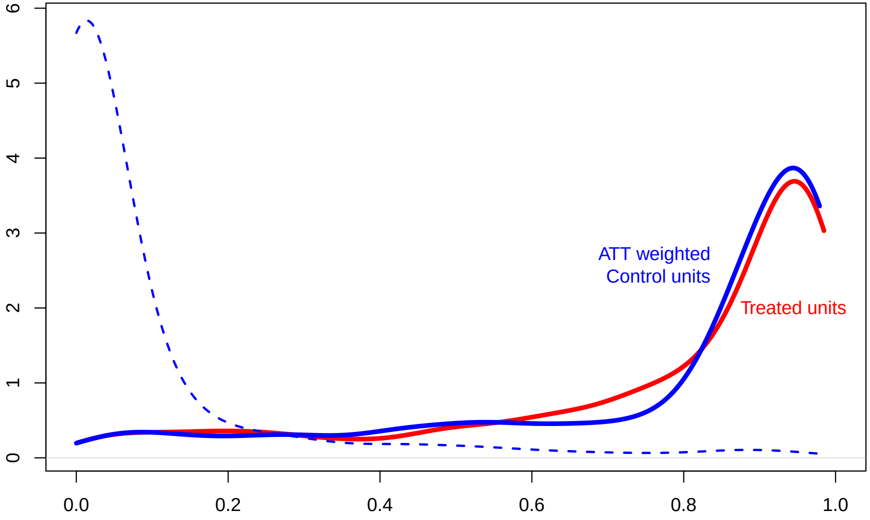
Weighting Both Groups to Balance Covariates

- 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.* :

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

- 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
- Tautology: propensity score works only when you get it right!
- In fact, estimated propensity score works even better than true propensity score when the model is correct

- 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 Great If the Model is Correct

Sample size	Estimator	Bias GLM	Bias True	RMSE GLM	RMSE 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	−0.32	−0.17	12.49	23.49
	IPW	−0.27	−0.35	3.94	4.90
	WLS	−0.07	−0.07	2.59	2.59
	DR	−0.07	−0.07	2.59	2.59
$n = 1000$	HT	0.03	0.01	4.93	10.62
	IPW	−0.02	−0.04	1.76	2.26
	WLS	−0.01	−0.01	1.14	1.14
	DR	−0.01	−0.01	1.14	1.14

Weighting Estimators Are Sensitive to Misspecification

		Bias		RMSE	
Sample size	Estimator	GLM	True	GLM	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	0.19	3.30	1.47
	DR	−48.66	0.08	1370.91	1.81

Covariate Balancing Propensity Score

- 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

- Score condition is a particular covariate balancing condition!

Estimation and Inference

- Just-identified CBPS:
  - Find the values of model parameters that satisfy covariate balancing conditions in the sample
  - Method of moments: # of parameters $=$ # of balancing conditions
- Over-identified CBPS:
  - # of parameters $<$ # of balancing conditions
  - Generalized method of moments (GMM):

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

    where

$$\bar{g}_\beta(T, X) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} \frac{T_i \pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1-T_i)\pi'_\beta(X_i)}{1-\pi_\beta(X_i)} \\ \frac{T_i \widetilde{X}_i}{\pi_\beta(X_i)} - \frac{(1-T_i)\widetilde{X}_i}{1-\pi_\beta(X_i)} \end{pmatrix}$$

    and Σ_β is the covariance of moment conditions
  - Enables misspecification test

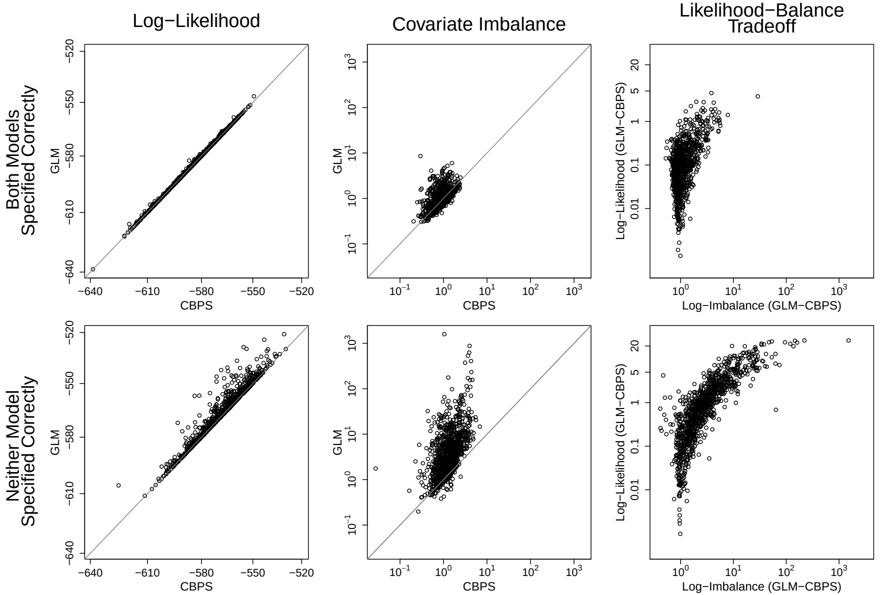
Revisiting Kang and Schafer (2007)

		Bias				RMSE			
Sample size	Estimator	GLM	CBPS1	CBPS2	True	GLM	CBPS1	CBPS2	True
(1) Both models correct									
$n = 200$	HT	0.33	2.06	−4.74	1.19	12.61	4.68	9.33	23.93
	IPW	−0.13	0.05	−1.12	−0.13	3.98	3.22	3.50	5.03
	WLS	−0.04	−0.04	−0.04	−0.04	2.58	2.58	2.58	2.58
	DR	−0.04	−0.04	−0.04	−0.04	2.58	2.58	2.58	2.58
$n = 1000$	HT	0.01	0.44	−1.59	−0.18	4.92	1.76	4.18	10.47
	IPW	0.01	0.03	−0.32	−0.05	1.75	1.44	1.60	2.22
	WLS	0.01	0.01	0.01	0.01	1.14	1.14	1.14	1.14
	DR	0.01	0.01	0.01	0.01	1.14	1.14	1.14	1.14
(2) Propensity score model correct									
$n = 200$	HT	−0.05	1.99	−4.94	−0.14	14.39	4.57	9.39	24.28
	IPW	−0.13	0.02	−1.13	−0.18	4.08	3.22	3.55	4.97
	WLS	0.04	0.04	0.04	0.04	2.51	2.51	2.51	2.51
	DR	0.04	0.04	0.04	0.04	2.51	2.51	2.51	2.51
$n = 1000$	HT	−0.02	0.44	−1.67	0.29	4.85	1.77	4.22	10.62
	IPW	0.02	0.05	−0.31	−0.03	1.75	1.45	1.61	2.27
	WLS	0.04	0.04	0.04	0.04	1.14	1.14	1.14	1.14
	DR	0.04	0.04	0.04	0.04	1.14	1.14	1.14	1.14

CBPS Makes Weighting Methods More Robust

		Bias				RMSE			
Sample size	Estimator	GLM	CBPS1	CBPS2	True	GLM	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

CBPS Sacrifices Likelihood for Better Balance



What Functions of Covariates Should We Balance?

- Bias of IPTW estimator when the propensity score is misspecified:

$$\text{bias} \quad = \quad \mathbb{E}\Bigg[\left(\frac{T_i}{\pi_{\beta^o}(X_i)} - \frac{1-T_i}{1-\pi_{\beta^o}(X_i)}\right) \times \Big\{\pi_{\beta^o}(X_i)\mathbb{E}(Y_i(0) \mid X_i) + (1-\pi_{\beta^o}(X_i))\mathbb{E}(Y_i(1) \mid X_i)\Big\}\Bigg]$$

  where β^o is the asymptotic limit of $\hat{\beta}$ under misspecification

- Balancing this weighted average leads to unbiased and efficient estimator

- Outcome model matters

Longitudinal Observational Studies

Fixed Effects Regressions in Causal Inference

- Linear fixed effects regression models are the primary workhorse for causal inference with panel data
- Researchers use them to adjust for unobserved 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*)

Questions

1. What make it possible for fixed effects regression models to adjust for unobserved confounding?

2. Are there any trade-offs when compared to the selection-on-observables approaches such as matching?

3. What are the exact causal assumptions underlying fixed effects regression models?

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

- Model:

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

- Estimator: "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

The Standard Assumption

Assumption 1 (Strict Exogeneity)

$$\mathbb{E}(\epsilon_{it} \mid \mathbf{X}_i, \alpha_i) \quad = \quad 0$$

where $\mathbf{X}_i$ *is a* $T \times 1$ *vector of treatment variables for unit* i

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

Causal Assumption I

Assumption 2 (No carryover effect)

Treatments do not directly affect future outcomes

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

- Potential outcome model:

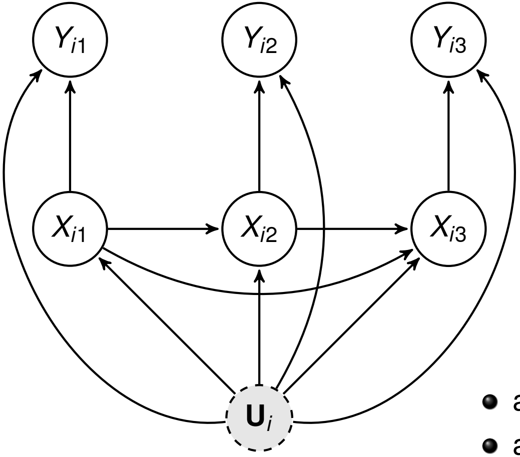
$$Y_{it}(x) \quad = \quad \alpha_i + \beta x + \epsilon_{it} \quad \text{for } x = 0, 1$$

- Average treatment effect:

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

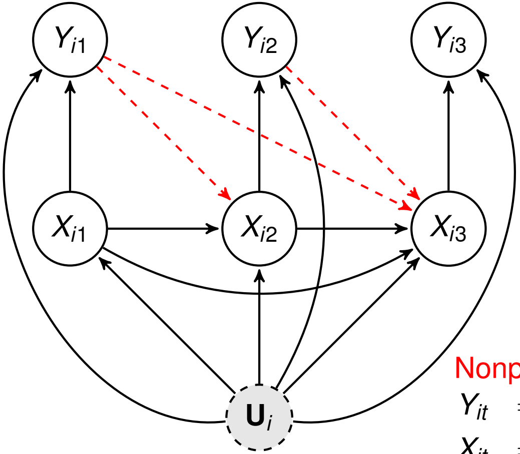
where $C_i = \mathbf{1}\{0 < \sum_{t=1}^{T} X_{it} < T\}$

Causal Directed Acyclic Graph (DAG)



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

Causal Directed Acyclic Graph (DAG)



Adding a red dashed arrow violates strict exogeneity

Nonparametric SEM (Pearl)

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

Causal Assumption II

- What randomized experiment satisfies strict exogeneity?

Assumption 3 (Sequential Ignorability with Unobservables)

$$\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} \perp\!\!\!\perp X_{i1} \mid \mathbf{U}_i$$

$$\vdots$$

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

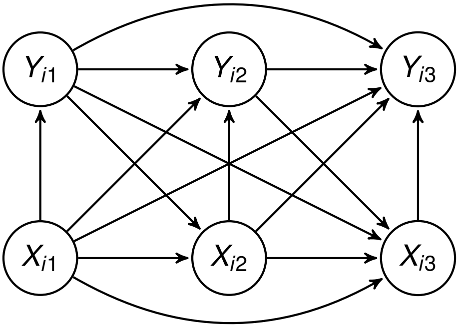
$$\vdots$$

$$\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} \perp\!\!\!\perp X_{iT} \mid X_{i1}, \ldots, X_{i,T-1}, \mathbf{U}_i$$

- The "as-if random" assumption without conditioning on the previous outcomes
- Outcomes can *directly* affect future outcomes $\rightsquigarrow$ but no need to adjust for past outcomes
- Nonparametric identification result

An Alternative Selection-on-Observables Approach

- Marginal structural models in epidemiology (Robins)
- Risk set matching (Rosenbaum)
- Trade-off: unobserved time-invariant confounders vs. direct effect of outcome on future treatment



Within-Unit Matching Estimator

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

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

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

- The Within-unit matching estimator improves $\hat{\beta}_{\text{FE}}$ by relaxing the linearity assumption:

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

Constructing a General Matching Estimator

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

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

- A general matching estimator just introduced:

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

where $D_{it} = \mathbf{1}\{\#\mathcal{M}(i,t) > 0\}$ and

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

Unit Fixed Effects Estimator as a Matching Estimator

- "de-meaning" $\rightsquigarrow$ match with all other observations within the same unit:

$$\mathcal{M}(i,t) \quad = \quad \{(i',t') : i' = i, t' \neq t\}$$

- mismatch: observations with the same treatment status

- Unit fixed effects estimator adjusts for mismatches:

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

  where K is the proportion of proper matches

- The within-unit matching estimator eliminates all mismatches

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:

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

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

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

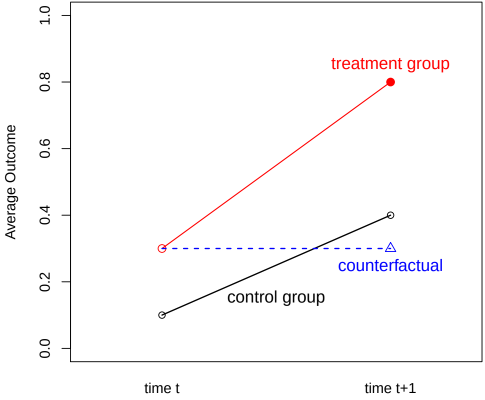
- We show how to construct regression weights for different matching estimators (i.e., different matched sets)
- Idea: count the number of times each observation is used for matching

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

Before-and-After Design

- The assumption that outcomes do not directly affect future treatments may not be credible
- Replace it with the design-based assumption:

$$\mathbb{E}(Y_{it}(x) \mid X_{it} = x') \quad = \quad \mathbb{E}(Y_{i,t-1}(x) \mid X_{i,t-1} = 1 - x')$$



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

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

- It is also the first differencing estimator:

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

- "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!

Remarks on Other Important Issues

1. Adjusting for observed time-varying confounding $\mathbf{Z}_{it}$
    - Proposes within-unit matching estimators that adjust for $\mathbf{Z}_{it}$
    - Key assumption: outcomes neither directly affect future treatments nor future time-varying confounders
2. Adjusting for past treatments
    - Impossible to adjust for all past treatments within the same unit
    - Researchers must decide the number of past treatments to adjust
3. Adjusting for past outcomes
    - No need to adjust for past outcomes if they do not directly affect future treatments
    - If they do, the strict exogeneity assumption will be violated
    - Past outcomes as instrumental variables (Arellano and Bond) $\rightsquigarrow$ often not credible

No free lunch: adjustment for unobservables comes with costs

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:

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

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

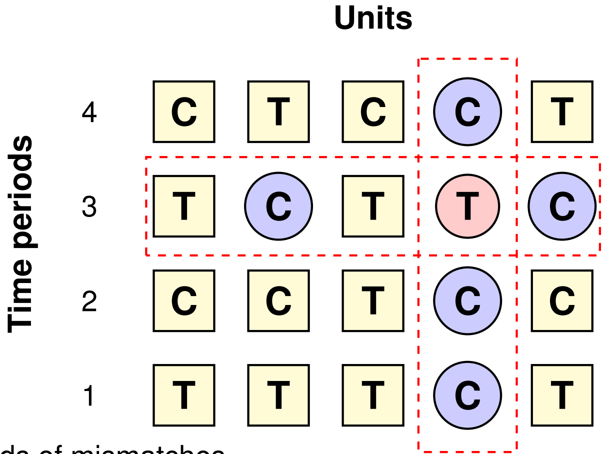
$$\hat{\beta}_{\text{FE2}} = \frac{\omega_{\text{FE}} \times \hat{\beta}_{\text{FE}} + \omega_{\text{FEtime}} \times \hat{\beta}_{\text{FEtime}} - \omega_{\text{pool}} \times \hat{\beta}_{\text{pool}}}{w_{\text{FE}} + w_{\text{FEtime}} - w_{\text{pool}}}$$

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

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

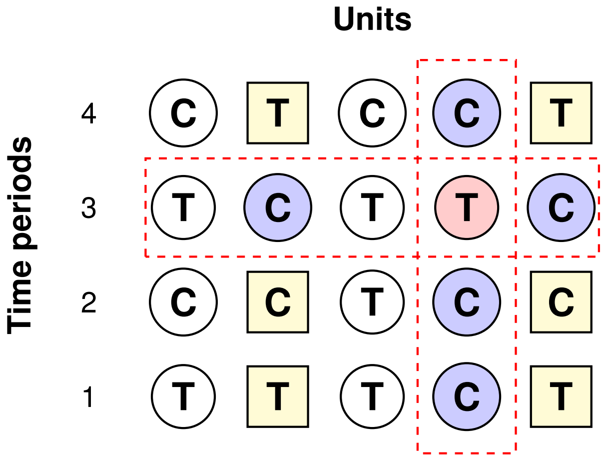
Matching and Two-way Fixed Effects Estimators

- 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

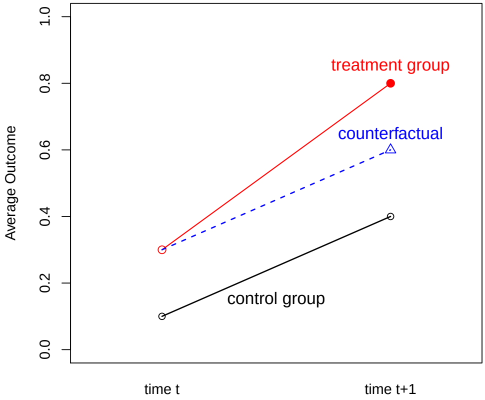


- To cancel time and unit effects, we must induce mismatches
- No weighted two-way fixed effects model eliminates mismatches

Difference-in-Differences Design

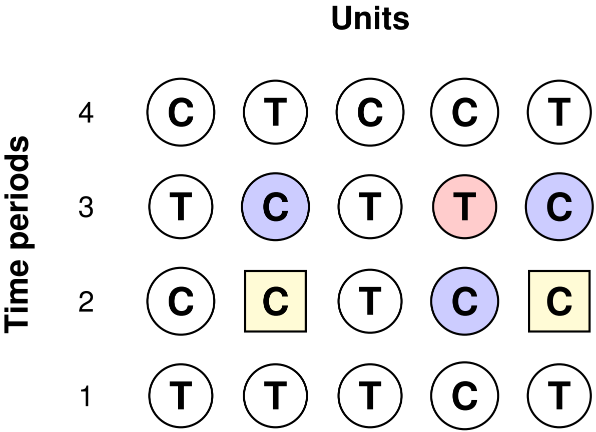
- Replace the model-based assumption with the design-based one
- Parallel trend assumption:

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

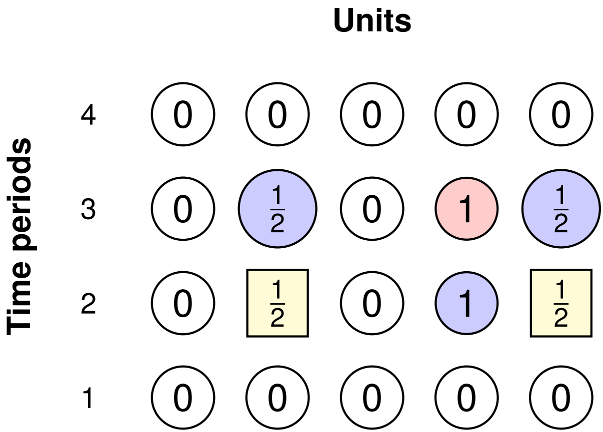


General DiD = Weighted Two-Way FE Effects

- $2 \times 2 \rightsquigarrow$ standard two-way fixed effects estimator works
- General setting: Multiple time periods, repeated treatments



- Regression weights:



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

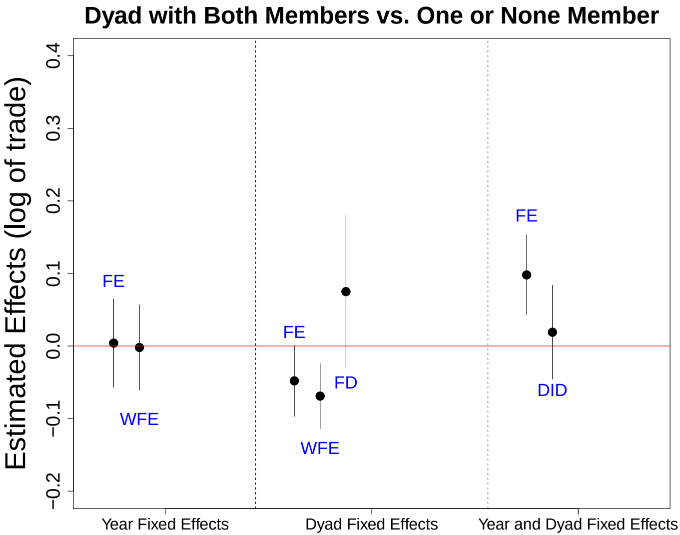
1. Data
   - Data set from Tomz et al. (2007)
   - Effect of GATT: 1948 – 1994
   - 162 countries, and 196,207 (dyad-year) observations
2. Year fixed effects model:

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

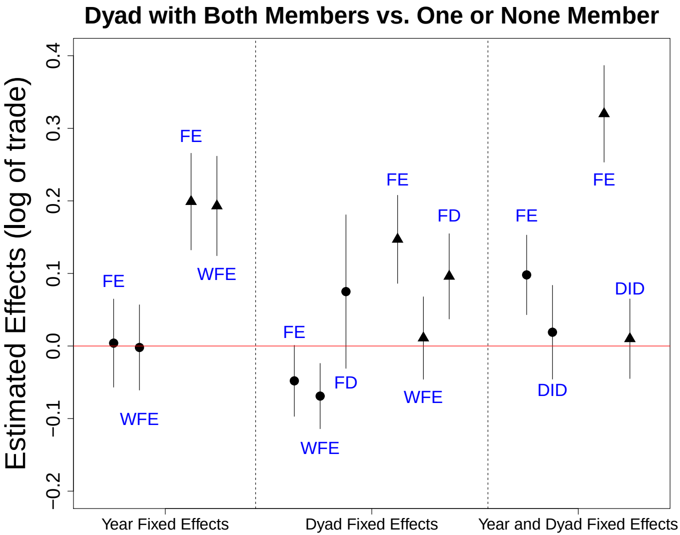
   - Y_{it}: trade volume
   - X_{it}: membership (formal/participants) Both vs. At most one
   - $\mathbf{Z}_{it}$: 15 dyad-varying covariates (e.g., log product GDP)
3. Weighted one-way fixed effects model:

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

Empirical Results: Formal Membership



Empirical Results



Synthetic Control Method

- Abadie and Gardeazabal (2003, AER); Abadie et al. (2010, JASA)
- Panel data: one treated unit, many controls
- Requirement: a long time-series of control observations before the treatment is administered at time j
  $T_{11} = 0, \ldots, T_{1,j-1} = 0, T_{1j} = 1, T_{1,j+1} = 1, \ldots, T_{1J} = 1$
- Quantity of interest: Treatment effect for the treated

$$Y_{1t}(1) - Y_{1t}(0) \;=\; Y_{1t} - Y_{1t}(0)$$

- Estimator:

$$Y_{1t}(1) - \widehat{Y_{1t}(0)} \;=\; Y_{1t} - \sum_{i=2}^{n} \hat{w}_i Y_{it}$$

  where $\hat{w}_i$ is estimated from the pre-treatment period such that

$$\hat{w} \;=\; \underset{w}{\operatorname{argmin}} \| Y_1 - \mathrm{diag}(w_i) Y_0 \|^2$$

  with $Y_1 = (Y_{11}, \ldots, Y_{1,j-1})$ and $Y_0 = (Y_{01}, \ldots, Y_{0,j-1})$
- Assumption: weights do not change over time

Causal Effect of ETA's Terrorism

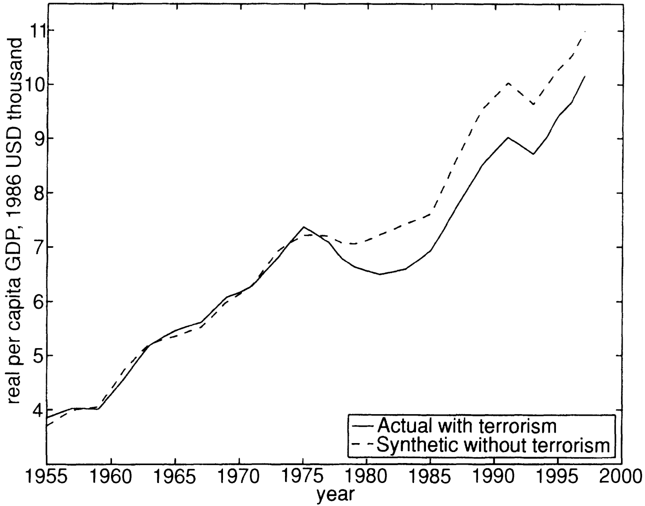


FIGURE 1. PER CAPITA GDP FOR THE BASQUE COUNTRY

Placebo Test

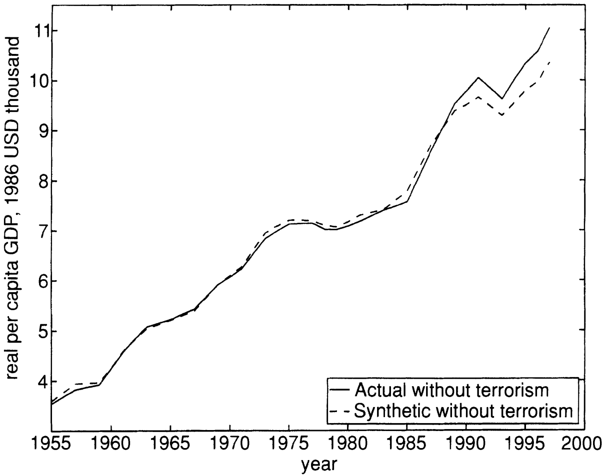


FIGURE 4. A "PLACEBO STUDY," PER CAPITA GDP FOR CATALONIA

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

Weighting with Longitudinal Data

- 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)

- Inverse-Probability-of-Treatment Weights:

$$w_i \;=\; \frac{1}{P(\overline{T}_{iJ} \mid \overline{X}_{iJ})} = \prod_{j=1}^{J} \frac{1}{P(T_{ij} \mid \overline{T}_{i,j-1}, \overline{X}_{ij})}$$

- Stabilized weights:

$$w_i^* \;=\; \frac{P(\overline{T}_{iJ})}{P(\overline{T}_{iJ} \mid \overline{X}_{iJ})}$$

Marginal Structural Models (MSMs)

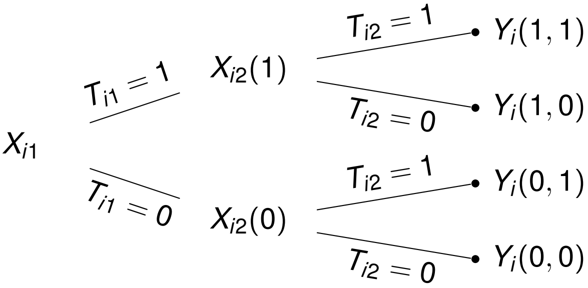
- Consistent estimation of the marginal mean of potential outcome:

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

- 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

- **Problem:** 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
- **Solution:** estimate MSM weights so that covariates are balanced

Two Time Period Case



- time 1 covariates X_{i1}: 3 equality constraints

$$\mathbb{E}(X_{i1}) \;=\; \mathbb{E}[\mathbf{1}\{T_{i1} = t_1, T_{i2} = t_2\} w_i \, X_{i1}]$$

- time 2 covariates X_{i2}: 2 equality constraints

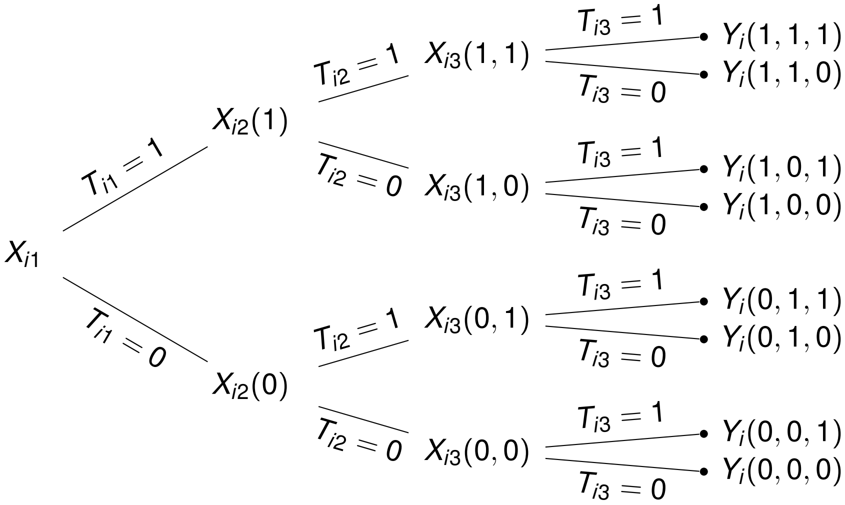
$$\mathbb{E}(X_{i2}(t_1)) \;=\; \mathbb{E}[\mathbf{1}\{T_{i1} = t_1, T_{i2} = t_2\} w_i \, X_{i2}(t_1)]$$

for $t_2 = 0, 1$

Orthogonalization of Covariate Balancing Conditions

	Treatment history: (t_1, t_2)				
Time period	(0,0)	(0,1)	(1,0)	(1,1)	Moment condition
time 1	$+$	$+$	$-$	$-$	$\mathbb{E}\left\{(-1)^{T_{i1}} w_i X_{i1}\right\} = 0$
	$+$	$-$	$+$	$-$	$\mathbb{E}\left\{(-1)^{T_{i2}} w_i X_{i1}\right\} = 0$
	$+$	$-$	$-$	$+$	$\mathbb{E}\left\{(-1)^{T_{i1}+T_{i2}} w_i X_{i1}\right\} = 0$
time 2	$+$	$-$	$+$	$-$	$\mathbb{E}\left\{(-1)^{T_{i2}} w_i X_{i2}\right\} = 0$
	$+$	$-$	$-$	$+$	$\mathbb{E}\left\{(-1)^{T_{i1}+T_{i2}} w_i X_{i2}\right\} = 0$

Extending Beyond Two Period Case



Generalization of the proposed method to J periods is in the paper

Orthogonalized Covariate Balancing Conditions

Design matrix			Treatment History Hadamard Matrix: (t_1, t_2, t_3)								Time		
			(0,0,0)	(1,0,0)	(0,1,0)	(1,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,1)			
T_{i1}	T_{i2}	T_{i3}	h_0	h_1	h_2	h_{12}	h_{13}	h_3	h_{23}	h_{123}	1	2	3
−	−	−	+	+	+	+	+	+	+	+	✗	✗	✗
+	−	−	+	−	+	−	+	−	+	−	✓	✗	✗
−	+	−	+	+	−	−	+	+	−	−	✓	✓	✗
+	+	−	+	−	−	+	+	−	−	+	✓	✓	✗
−	−	+	+	+	+	+	−	−	−	−	✓	✓	✓
+	−	+	+	−	+	−	−	+	−	+	✓	✓	✓
−	+	+	+	+	−	−	−	−	+	+	✓	✓	✓
+	+	+	+	−	−	+	−	+	+	−	✓	✓	✓

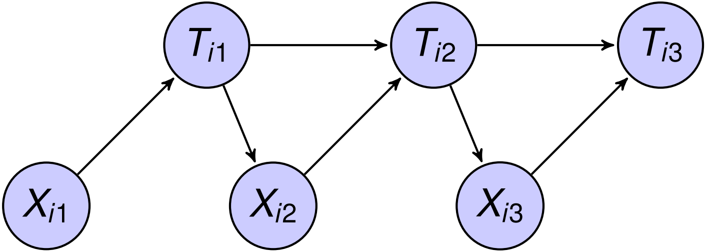
- The mod 2 discrete Fourier transform:

$$\mathbb{E}\{(-1)^{T_{i1}+T_{i3}} w_i X_{ij}\} = 0 \quad (\text{6th row})$$

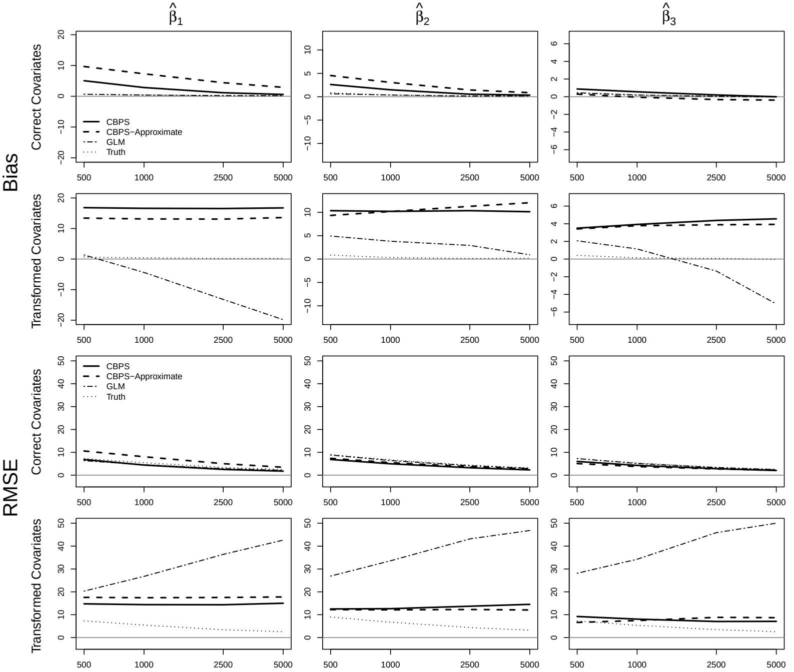
- Connection to the fractional factorial design
  - "Fractional" = past treatment history
  - "Factorial" = future potential treatments

A Simulation Study with Correct Lag Structure

- 3 time periods
- Treatment assignment process:



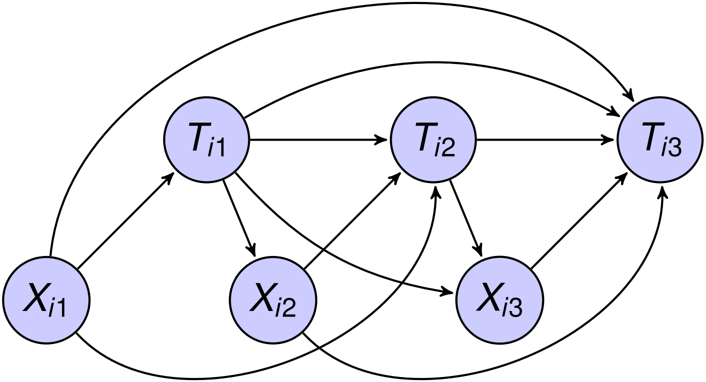
- Outcome: $Y_i = 250 - 10 \cdot \sum_{j=1}^{3} T_{ij} + \sum_{j=1}^{3} \delta^\top X_{ij} + \epsilon_i$
- Functional form misspecification by nonlinear transformation of X_{ij}



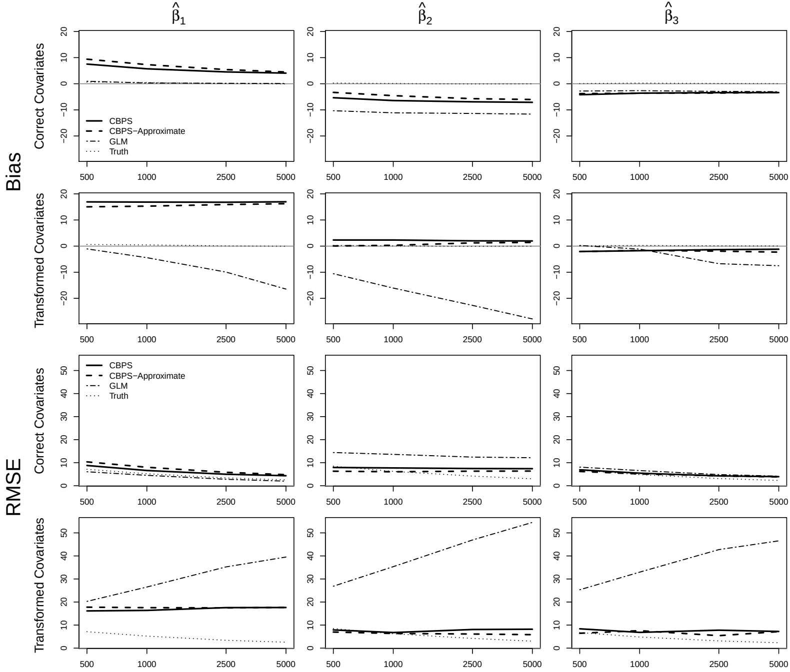
$\hat{\beta}_1$
$\hat{\beta}_2$
$\hat{\beta}_3$
Bias
RMSE
Correct Covariates
Transformed Covariates
CBPS
CBPS−Approximate
GLM
Truth
500
1000
2500
5000

A Simulation Study with Incorrect Lag Structure

- 3 time periods
- Treatment assignment process:



- The same outcome model
- Incorrect lag: only adjusts for previous lag but not all lags
- In addition, the same functional form misspecification of X_{ij}


$\hat{\beta}_1$
$\hat{\beta}_2$
$\hat{\beta}_3$
Bias
RMSE
Correct Covariates
Transformed Covariates
CBPS
CBPS−Approximate
GLM
Truth
500
1000
2500
5000

Empirical Illustration: Negative Advertisements

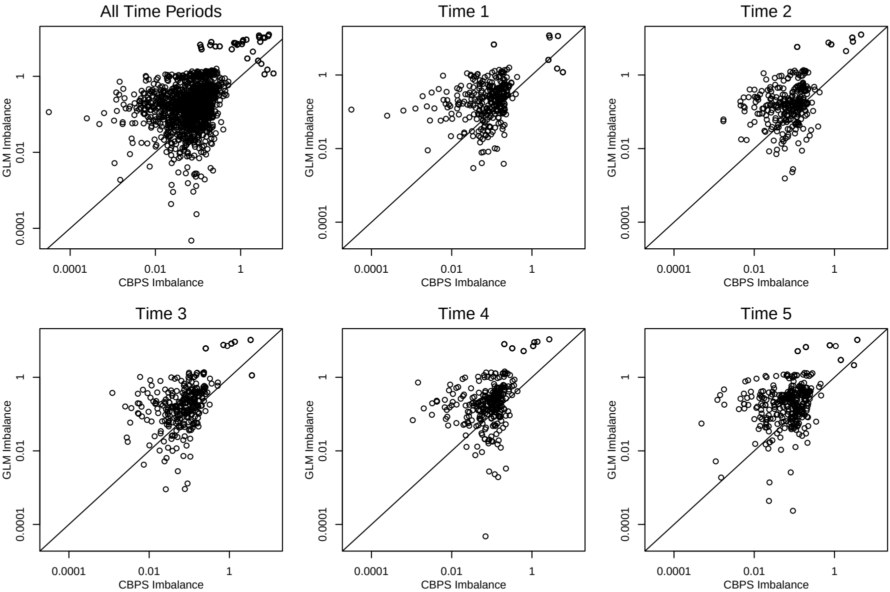
- Electoral impact of negative advertisements (Blackwell, 2013)
- For each of 114 races, 5 weeks leading up to the election

- Outcome: candidates' voteshare
- Treatment: negative ($T_{it} = 1$) or positive ($T_{it} = 0$) campaign

- Time-varying covariates: Democratic share of the polls, proportion of voters undecided, campaign length, and the lagged and twice lagged treatment variables for each week
- Time-invariant covariates: baseline Democratic voteshare, baseline proportion undecided, and indicators for election year, incumbency status, and type of office

- Original study: pooled logistic regression with a linear time trend
- We compare period-by-period GLM with CBPS

Covariate Balance



	GLM	CBPS	CBPS (approx.)	GLM	CBPS	CBPS (approx.)
(Intercept)	55.69*	57.15*	57.94*	55.41*	57.06*	57.73*
	(4.62)	(1.84)	(2.12)	(3.09)	(1.68)	(1.88)
Negative	2.97	5.82	3.15			
(time 1)	(4.55)	(5.30)	(3.76)			
Negative	3.53	2.71	5.02			
(time 2)	(9.71)	(9.26)	(8.55)			
Negative	−2.77	−3.89	−3.63			
(time 3)	(12.57)	(10.94)	(11.46)			
Negative	−8.28	−9.75	−10.39			
(time 4)	(10.29)	(7.79)	(8.79)			
Negative	−1.53	−1.95*	−2.13*			
(time 5)	(0.97)	(0.96)	(0.98)			
Negative				−1.14	−1.35*	−1.51*
(cumulative)				(0.68)	(0.39)	(0.43)
R^2	0.04	0.14	0.13	0.02	0.10	0.10
F statistics	0.95	3.39	3.32	2.84	12.29	12.23

Concluding Remarks

- Matching methods do:
  - make causal assumptions transparent by identifying counterfactuals
  - make regression models robust by reducing model dependence
- But they 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
- Other methodological challenges for causal inference: temporal and spatial dynamics, networks effects

References

- "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis*
- "Misunderstandings among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A*
- "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." *Statistical Science*
- "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society, Series B*
- "Robust Estimation of Inverse Probability Weights for Marginal Structural Models." *Journal of the American Statistical Association*
- "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Panel Data?" Working paper

All papers are available at

http://imai.princeton.edu/research

Software Implementation

- Causal inference with regression: Zelig: Everyone's Statistical Software
- Causal inference with matching: MatchIt: Nonparametric Preprocessing for Parametric Causal Inference
- Causal inference with propensity score: CBPS: Covariate Balancing Propensity Score
- Causal inference with fixed effects: wfe: Weighted Fixed Effects Regressions for Causal Inference

All software is available at
`http://imai.princeton.edu/software`