Matching and Weighting Methods for Causal Inference

Kosuke Imai

Princeton University

Methods Workshop, Duke University

References to Relevant Papers

- "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* (2007)
- "Misunderstandings among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* (2008)
- "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." *Statistical Science* (2009)
- "Covariate Balancing Propensity Score." Working paper
- "On the Use of Linear Fixed Effects Regression Models for Causal Inference." 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`

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

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:

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

  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

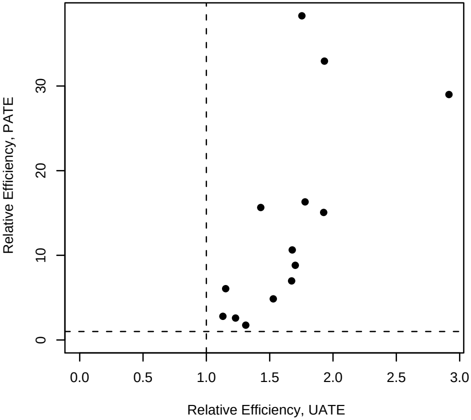
- Okay, but how should I match/block without the treatment group?
- Goal: match/block well on powerful predictors of outcome (prognostic factors)
- (Coarsened) Exact matching
- Matching based on a similarity measure:

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

- 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!



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

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

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 (no assumption):

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

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

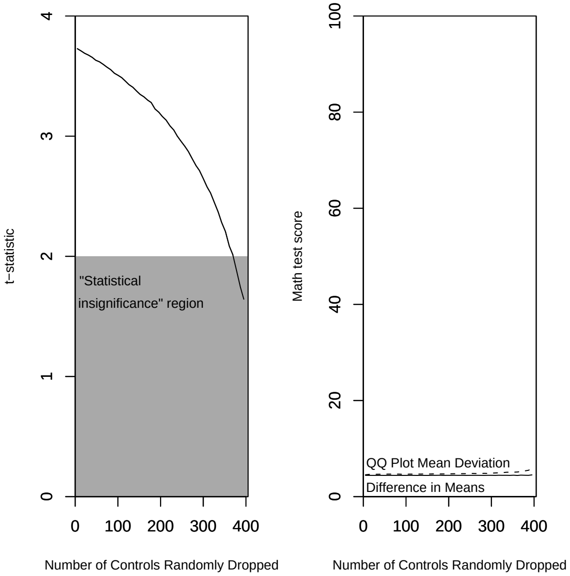
- FURTHER READING: Rubin (2006). *Matched Sampling for Causal Effects* (Cambridge UP)

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,

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

  - $\overline{X}_{mt}$ and $\overline{X}_{mc}$ are the sample means
  - s_{mt}^2 and s_{mc}^2 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

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):

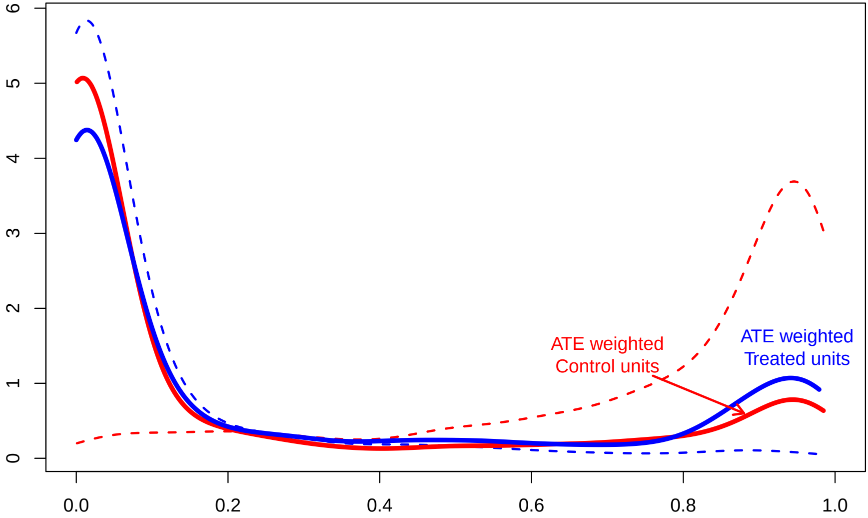
$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}\right)$$

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

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

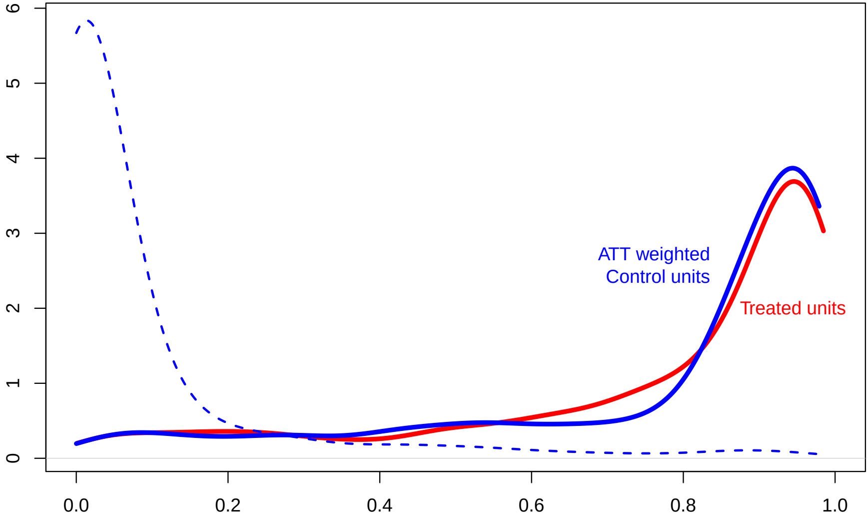
Weighting Both Groups to Balance Covariates

- 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.* :

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

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

Marginal Structural Models for Longitudinal Data

- Units $i = 1, \ldots, N$ and time $j = 1, \ldots, J$
- Eventual outcome Y_i measured at time J
- Treatment and covariate history: $\overline{T}_{ij}$ and $\overline{X}_{ij}$
- Quantity of interest: (marginal) ATE $= \mathbb{E}\{Y_i(\bar{t})\}$
- Sequential ignorability assumption:

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

- Inverse-probability-of-treatment weight:

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

- Stabilized weight: multiply w_i by $P(\overline{T}_{iJ})$
- Analysis: *weighted* regression of Y_i on $\overline{T}_{iJ}$
- FURTHER READINGS: Robins *et al.* (2000), Blackwell (2013)

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):

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

where

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

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

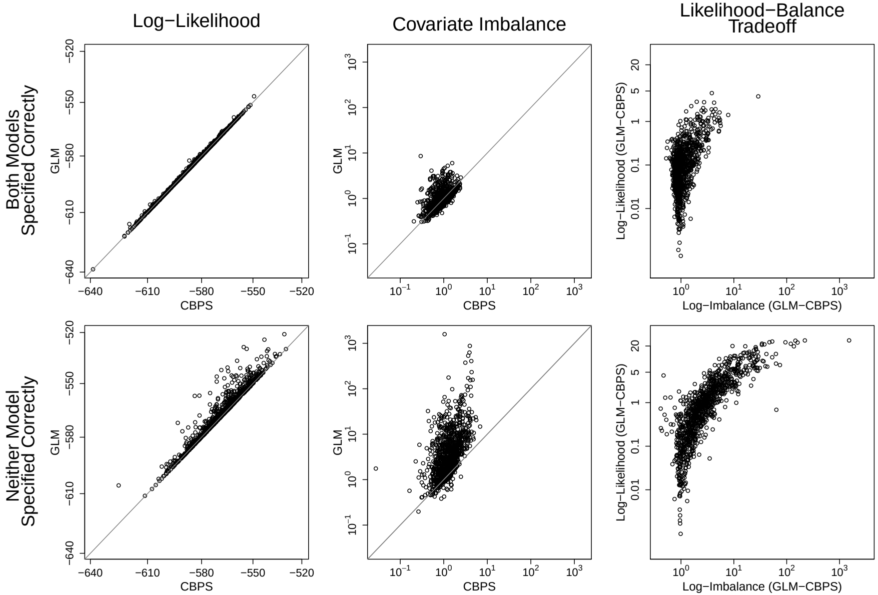
Revisiting Kang and Schafer (2007)

		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



A Close Look at Fixed Effects Regression

- Fixed effects models are a primary workhorse for causal inference
- Used for stratified experimental and observational data
- Also used to adjust for unobservables in observational studies:
  - "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*)
- Common claim: Fixed effects models are superior to matching estimators because the latter can only adjust for observables
- **Question:** What are the exact causal assumptions underlying fixed effects regression models?

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

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

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

Matching Representation of Simple Regression

- Cross-section simple linear regression model:

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

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

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

where

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

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

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

One-Way Fixed Effects Regression

- Simple (one-way) FE model:

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

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

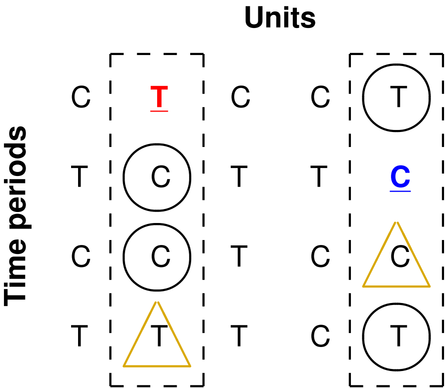
- $\hat{\beta}_{FE}$ may be biased for the ATE even if X_{it} is exogenous within each unit
- It converges to the weighted average of conditional ATEs:

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

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

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

Mismatches in One-Way Fixed Effects Model



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

Matching Representation of Fixed Effects Regression

Proposition 1

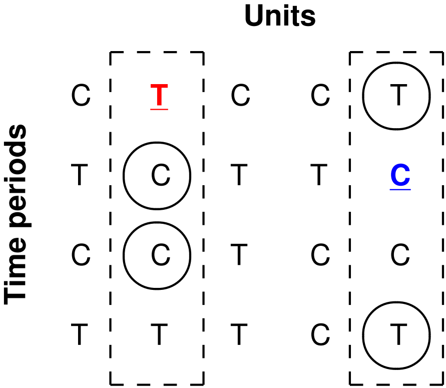
$$\hat{\beta}^{FE} = \frac{1}{K}\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right)\right\},$$

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

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

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

Unadjusted Matching Estimator



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

Unadjusted Matching = **Weighted** FE Estimator

Proposition 2

The unadjusted matching estimator

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

where

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

is equivalent to the weighted fixed effects model

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

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

Equal Weights

Treatment

C	T	C	C	T
T	C	T	T	C
C	C	T	C	C
T	T	T	C	T

Weights

0	1	0	0	$\frac{1}{2}$
0	$\frac{1}{2}$	0	0	1
0	$\frac{1}{2}$	0	0	0
0	0	0	0	$\frac{1}{2}$

Different Weights

Treatment					Weights				
C	T	C	C	T	0	1	0	0	$\frac{3}{4}$
T	C	T	T	C	0	$\frac{2}{3}$	0	0	1
C	C	T	C	C	0	$\frac{1}{3}$	0	0	0
T	T	T	C	T	0	0	0	0	$\frac{1}{4}$

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

First Difference = Matching = Weighted One-Way FE

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

Treatment

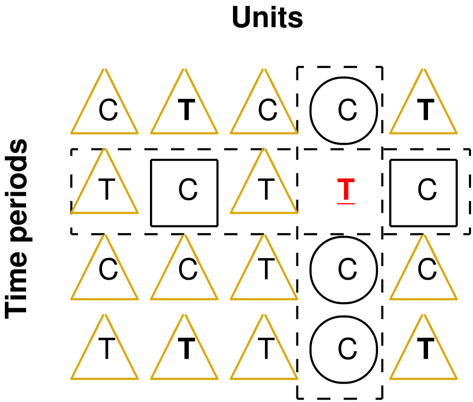
C	T	C	C	T
T	C	T	T	C
C	C	T	C	C
T	T	T	C	T

Weights

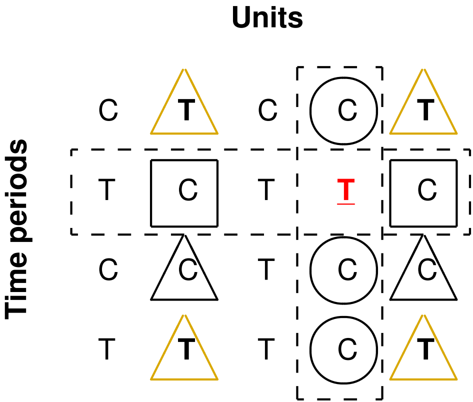
0	1	0	0	0
0	1	0	0	0
0	0	0	0	0
0	0	0	0	0

Mismatches in Two-Way FE Model

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

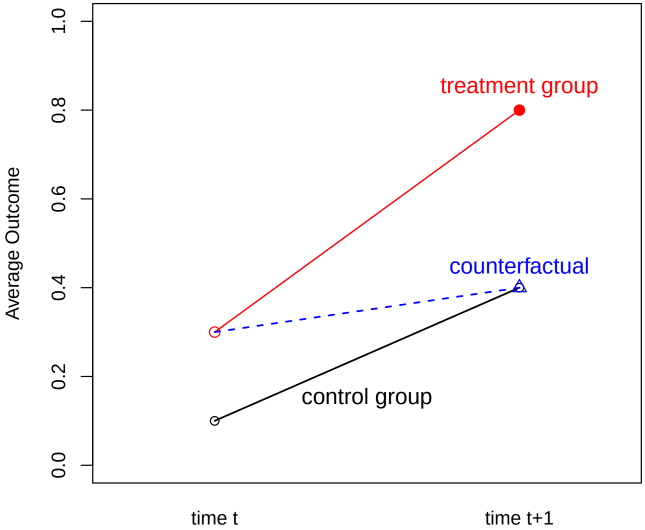


- **Triangles**: Two kinds of mismatches
  - Same treatment status
  - Neither same unit nor same time

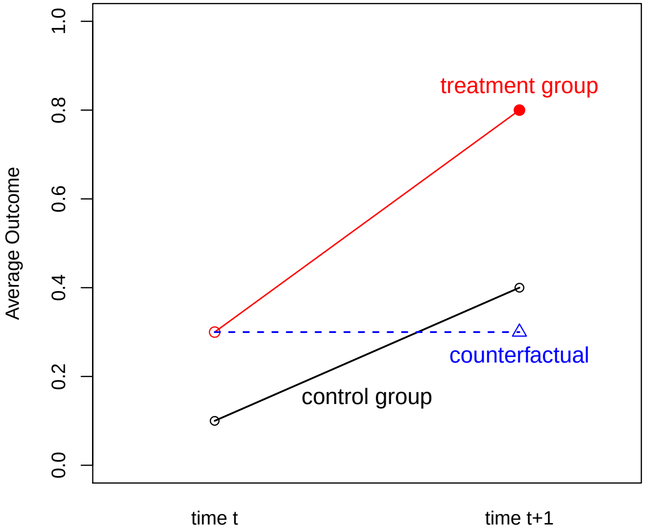


- Some mismatches can be eliminated
- You can NEVER eliminate them all

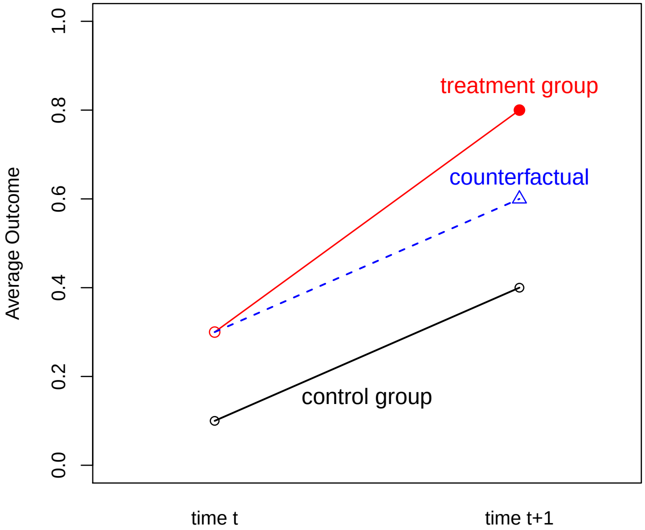
Cross Section Analysis = Weighted **Time** FE Model



First Difference = Weighted **Unit** FE Model



What about Difference-in-Differences (DiD)?



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

- 2×2: standard two-way fixed effects
- General setting: Multiple time periods, repeated treatments

Treatment

C	**T**	C	C	**T**
T	$\boxed{\text{C}}$	T	<u>**T**</u>	$\boxed{\text{C}}$
C	C	T	C	C
T	**T**	T	C	**T**

Weights

0	0	0	0	0
0	$\boxed{\frac{1}{2}}$	0	<u>**1**</u>	$\boxed{\frac{1}{2}}$
0	$\frac{1}{2}$	0	1	$\frac{1}{2}$
0	0	0	0	0

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

Effects of GATT Membership on International Trade

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

Data and Methods

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

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

  - X_{it}: *Formal* membership/*Participant* (1) Both vs. One, (2) One vs. None, (3) Both vs. One/None
  - Z_{it}: 15 dyad-varying covariates (e.g., log product GDP)
- Year fixed effects: standard, weighted, and first difference
- Two-way fixed effects: standard and difference-in-differences

Empirical Results

Comparison	Membership	Year Fixed Effects: Standard	Year Fixed Effects: Weighted	Dyad Fixed Effects: Standard	Dyad Fixed Effects: Weighted	Dyad Fixed Effects: First Diff.	Year and Dyad Fixed Effects: Standard	Year and Dyad Fixed Effects: Diff.-in-Diff.
Both vs. Mix	Formal	0.004	−0.002	−0.048	−0.069	0.075	0.098	0.019
	(N=196,207)	(0.031)	(0.030)	(0.025)	(0.023)	(0.054)	(0.028)	(0.033)
	White's p-value		1.000		0.064	0.000		0.058
	Participants	0.199	0.193	0.147	0.011	0.096	0.320	0.010
	(N=196,207)	(0.034)	(0.035)	(0.031)	(0.029)	(0.030)	(0.034)	(0.028)
	White's p-value		0.998		0.000	0.102		0.000
Both vs. One	Formal	−0.006	−0.005	−0.034	−0.061	0.076	0.105	0.016
	(N=175,814)	(0.031)	(0.031)	(0.025)	(0.023)	(0.055)	(0.028)	(0.033)
	White's p-value		1.000		0.031	0.000		0.034
	Participants	0.180	0.174	0.161	0.020	0.099	0.332	0.009
	(N=187,651)	(0.035)	(0.036)	(0.031)	(0.029)	(0.030)	(0.034)	(0.029)
	White's p-value		0.999		0.000	0.086		0.000
One vs. None	Formal	0.007	0.046	−0.011	−0.094	0.031	0.082	−0.020
	(N=109,702)	(0.053)	(0.056)	(0.041)	(0.041)	(0.067)	(0.043)	(0.378)
	White's p-value		0.276		0.058	0.000		0.789
	Participants	0.163	0.171	0.181	−0.034	0.053	0.244	0.007
	(N=70,298)	(0.072)	(0.079)	(0.062)	(0.058)	(0.063)	(0.066)	(0.085)
	White's p-value		0.046		0.004	0.000		0.026
covariates		dyad-varying covariates		year-varying covariates			no covariate	

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
- Next methodological challenges for causal inference: temporal and spatial dynamics, networks effects