Use of Matching Methods for Causal Inference in Experimental and Observational Studies

Kosuke Imai

Department of Politics
Princeton University

April 13, 2009

This Talk Draws on the Following Papers:

- Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. (2007). “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” *Political Analysis*, Vol. 15, No.3 (Summer), pp. 199–236.
- Imai, Kosuke, Gary King, and Elizabeth A. Stuart. (2008). “Misunderstandings among Experimentalists and Observationalists about Causal Inference.” *Journal of the Royal Statistical Society, Series A*, Vol. 171, No. 2 (April), pp. 481–502.
- Imai, Kosuke, Gary King, and Clayton Nall. “The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation.” (with discussions) *Statistical Science*, Forthcoming.

Overview

- What is matching?
- Nonparametric covariate adjustment method

- Use of Matching in Experimental Studies
  - Goal: efficiency gain
  - Covariate adjustment *prior to* treatment assignment
  - matched-pair design

- Use of Matching in Observational Studies
  - Goal: bias reduction
  - Ignorability (unconfounding, selection on observables)
  - Matching as nonparametric preprocessing

Covariate Adjustments in Experiments

- Randomization of treatment guarantees unbiasedness
- Adjusting for covariates may lead to efficiency gain
- Dangers of post-randomization covariate adjustment
  - Bias due to statistical methods
  - Bias due to post-hoc analysis

- Make adjustments *before* the randomization of treatment
- Employ design-based inference rather than model-based
- Difference-in-means rather than regression

Randomized-block Design

- Form a group of units based on the pre-treatment covariates so that the observations within each block are similar
- Complete randomization of the treatment within each block
- Inference based on the weighted average of within-block estimates

- Blocking can never hurt; unbiased and no less efficient
- Efficiency gain is greater if across-block heterogeneity is greater
- Difference in asymptotic variance:

$$\mathrm{Var}(\overline{Y(1)}_x + \overline{Y(0)}_x)$$

where $\overline{Y(t)}_x$ is the within-block mean of $Y_i(t)$

An Example: A Randomized Survey Experiment

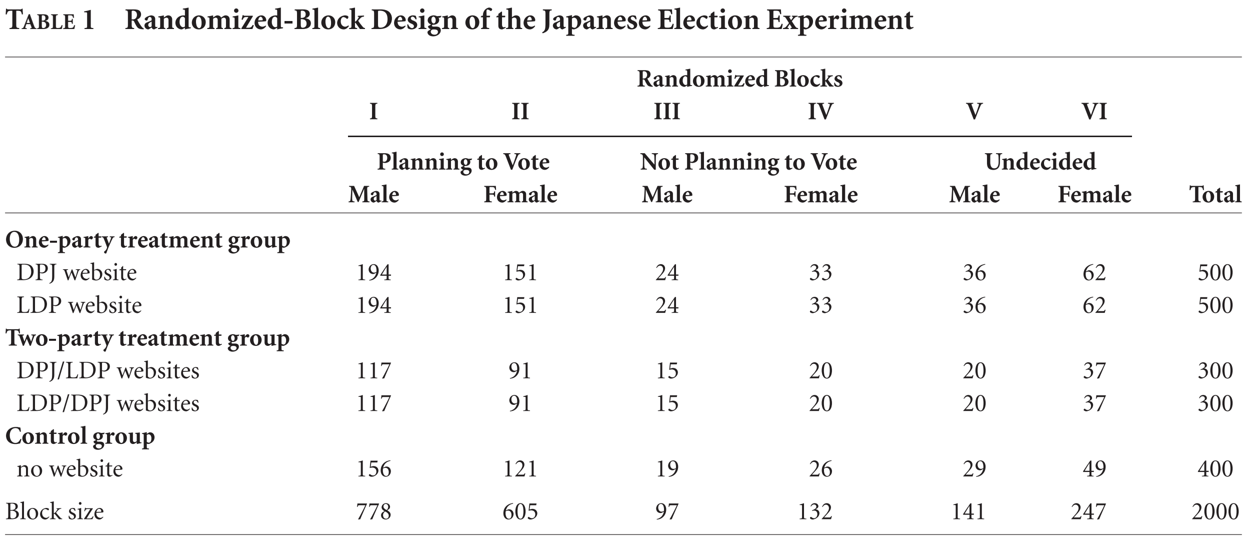


TABLE 1 **Randomized-Block Design of the Japanese Election Experiment**

	Randomized Blocks						
	I	**II**	**III**	**IV**	**V**	**VI**	
	Planning to Vote		**Not Planning to Vote**		**Undecided**		
	Male	**Female**	**Male**	**Female**	**Male**	**Female**	**Total**
One-party treatment group							
DPJ website	194	151	24	33	36	62	500
LDP website	194	151	24	33	36	62	500
Two-party treatment group							
DPJ/LDP websites	117	91	15	20	20	37	300
LDP/DPJ websites	117	91	15	20	20	37	300
Control group							
no website	156	121	19	26	29	49	400
Block size	778	605	97	132	141	247	2000

Horiuchi, Imai, and Taniguchi (2007, *AJPS*)

Matched-Pair Design (MPD)

- Blocking where the size of all blocks is 2
- Create pairs of units based on the pre-treatment covariates so that the units within a pair are similar to each other
- Randomly assign the treatment within each matched-pair
- Inference based on the average of within-pair differences

- Difference in variances:

$$\frac{1}{n/2}\mathrm{Cov}(Y_{ij}(1), Y_{i'j}(0))$$

- Greater within-pair similarity leads to greater efficiency
- Multivariate blocking/matching methods

Under-usage of Matching in Experiments

- Most applied experimental research conducts simple randomization of the treatment
- Among all experiments published in *APSR*, *AJPS*, and *JOP* (since 1995) and those listed in Time-sharing Experiments for the Social Sciences, only one uses matching methods!
- Two key analytical results:
  1. Randomized-block design **always** yields more efficient estimates.
  2. Matched-pair design **usually** yields more efficient estimates.

An Example: Cluster-Randomized Experiments

- Unit of randomization = clusters of individuals
- Unit of interest = individuals
- CREs among political science field experiments: 68% (out of 28)
- Public health & medicine: CREs have "risen exponentially since 1997" (Campbell, 2004)
- Education (classrooms – students)
- Psychology (groups – individuals)
- Sociology (neighborhoods – households), etc.

Design and Analysis of CREs

- Cluster randomization → loss of efficiency & specialized methods
- Prop. of polisci CREs which completely ignore the design: $\approx$ 50%
- Prop. of polisci CREs which use *design-based* analysis: 0%
- Prop. of polisci CREs which make more assumptions than necessary: 100%

- Matched-Pair Design (MPD) to improve efficiency:
  1. Pair clusters based on the similarity of background characteristics
  2. Within each pair, randomly assign one cluster to the treatment group and the other to the control group

- Use of MPDs in CREs:
  - Prop. of public health CREs: $\approx$ 50% (Varnell *et al.*, 2004)
  - Prop. of polisci CREs: 0%

Methodological Recommendations Against MPDs

- "Analytical limitations" of MPDs (Klar and Donner, 1997):
- Echoed by other researchers (exception Thompson, 1998)
- Echoed by other researchers and clinical standard organizations (e.g., Medical Research Council in UK)

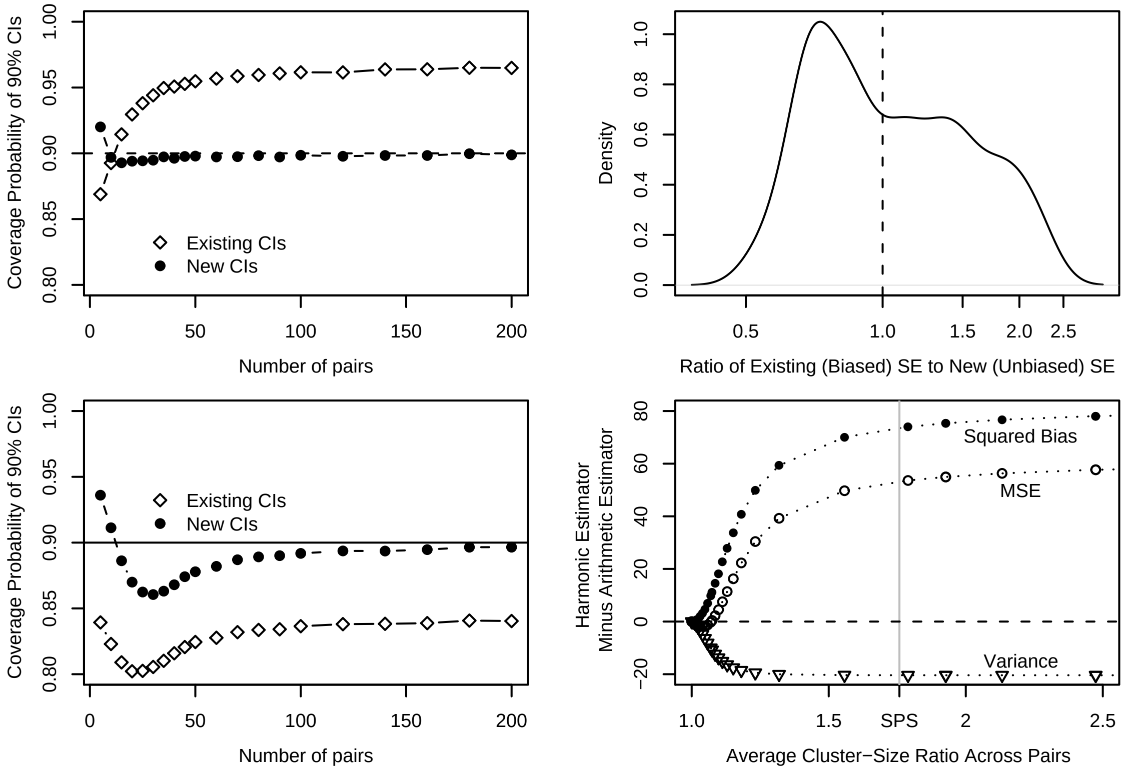
- In 10 or fewer pairs, MPDs can lose power (Martin *et al.* 1993)
- No formal definition of causal effects to be estimated
- No formal evaluation of the existing estimators for MPDs

- It's okay to use MPDs!
  1. overcome analytical limitations
  2. develop a new design-based estimator
  3. conduct design-based inference

Seguro Popular de Salud (SPS)

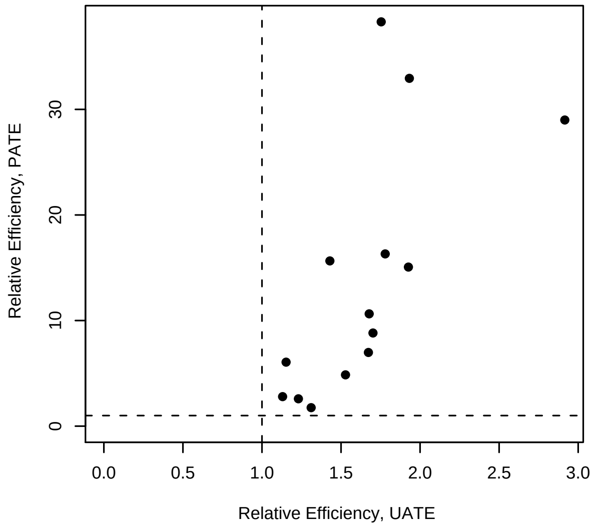
- Evaluation of the Mexican universal health insurance program
- Aim: "provide social protection in health to the 50 million uninsured Mexicans"
- A key goal: reduce out-of-pocket health expenditures
- Sounds obvious but not easy to achieve in developing countries
- Individuals must affiliate in order to receive SPS services
- $12,824$ "health clusters"
- 100 clusters nonrandomly chosen for randomized evaluation
- Pairing 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

Design-based vs. Model-based Inference



Relative Efficiency of MPD

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



Fundamental Challenges of Observational Studies

- Researchers cannot randomly assign treatments
- Association is NOT causation
- How can we "design" an observational study close to an randomized experiment?
- "A good way to do econometrics is to look for good natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for us." (D. McFadden)
- Advantages: large *n*, representative sample, double-blinded, etc.
- Trade-off between internal and external validity

Ignorability Assumption

- Conditional on observed pre-treatment covariates X_i, the treatment is "randomized"

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \mid X_i = x \quad \text{for all } x$$

- Also called unconfoundedness, no omitted variable, selection on observables
- The assumption of overlap

$$0 < \Pr(T_i = 1 \mid X_i = x) < 1 \quad \text{for all } x$$

More on Ignorability

- Strong assumption
- Unobserved confounders that causally affect both treatment and outcome
- Not testable from the observed data
- Conditioning on too much can hurt
- But in practice it's hard to justify ignoring observed covariate imbalance

- How can we make the assumption credible?
- What information was relevant when making decisions?

Identification of the Average Treatment Effect

- Under exogeneity,

$$\mathbb{E}(Y_i(1) - Y_i(0)) \quad = \quad \mathbb{E}\left\{\mathbb{E}(Y_i \mid T_i = 1, X_i) - \mathbb{E}(Y_i \mid T_i = 0, X_i)\right\}$$

- X_i can be high-dimensional and difficult to model
- ATT (ATE for the treated):

$$\mathbb{E}(Y_i(1) - Y_i(0) \mid T_i = 1) \; = \; \mathbb{E}(Y_i \mid T_i = 1) - \mathbb{E}\left\{\mathbb{E}(Y_i \mid T_i = 0, X_i)\right\}$$

- Need to model $\mathbb{E}(Y_i \mid T_i, X_i)$ or $\mathbb{E}(Y_i \mid T_i = 0, X_i)$
- Non-parametric regression – curse of dimensionality
- Parametric regression – functional-form/distributional assumptions

Matching as Nonparametric Preprocessing

- Assume ignorability holds
- Preprocess the data so that treatment and control groups are similar to each other in terms of the observed pre-treatment covariates
- Goal of matching: achieve balance

$$\widetilde{F}(X \mid T = 1) \quad = \quad \widetilde{F}(X \mid T = 0)$$

  where $\widetilde{F}(\cdot)$ is the *empirical* distribution

- Exact matching: impossible in most cases
- Maximize balance via matching
- Parametric adjustment for remaining imbalance
- Minimal role of statistical models; reduced model dependence

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 under exogeneity:

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

- Ignorability given the propensity score:

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

- Dimension reduction
- But, true propensity score is unknown: propensity score tautology
- Possible to extend it to non-binary treatment

Methods to Improve Covariate Balance

- Matching: Each treated unit is paired with a similar control unit based on the pre-treatment covariates.
- Subclassification: Treated and control units are grouped to form subclasses based on the pre-treatment covariates so that within each subclass treated units are similar to control units.
- Weighting: Weight each observation within the treated or control groups by the inverse of the probability of being in that group.

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

- The goal of all three methods is to improve balance

Double Robustness Property

- Why care about propensity score?
- Propensity score model specification can be difficult when X_i is high-dimensional
- Doubly-robust estimator:

$$\begin{aligned}\hat{\tau}_{DR} \quad \equiv \quad & \left\{\frac{1}{n}\sum_{i=1}^{n}\hat{m}_1(X_i) + \frac{1}{n}\sum_{i=1}^{n}\frac{T_i(Y_i - \hat{m}_1(X_i))}{\hat{\pi}(X_i)}\right\} \\ & - \left\{\frac{1}{n}\sum_{i=1}^{n}\hat{m}_0(X_i) + \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)(Y_i - \hat{m}_0(X_i))}{1-\hat{\pi}(X_i)}\right\},\end{aligned}$$

  where $m_t(X_i) = \mathbb{E}(Y_i(t) \mid X_i)$ for $t = 0, 1$.
- Consistent if either the propensity score model or the outcome model is correct
- Efficient if both models are correct

Common Matching Methods

- 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 matching
- Caliper matching
- Subclassification on propensity score
- Optimal/Genetic matching
- Matching with and without replacement

- Which matching method to choose?
- Whatever gives you the “best” balance!

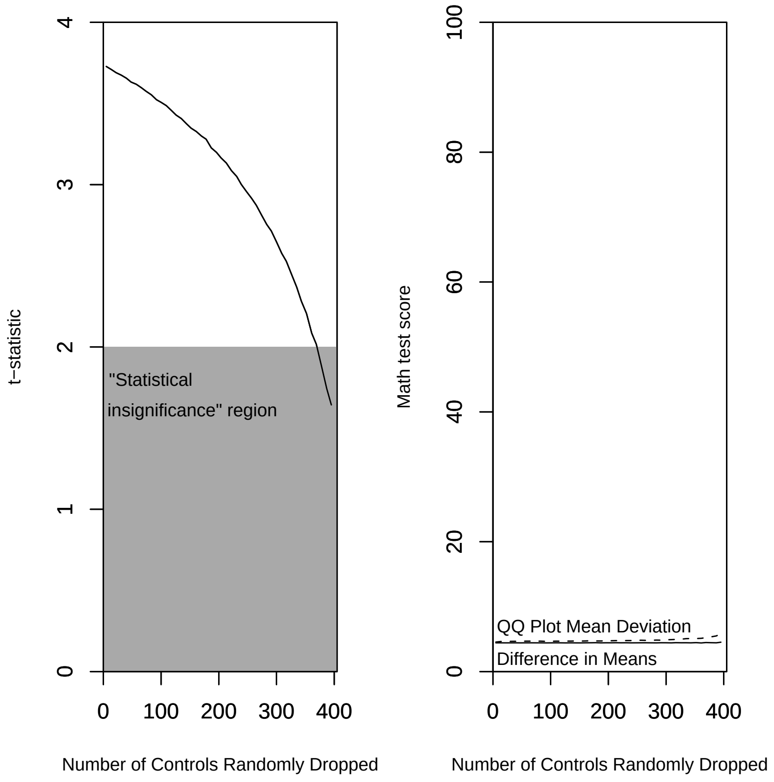
How to Check Balance

- 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

- School Dropout Demonstration Assistance Program.
- Treatment: school “restructuring” programs.
- Outcome: dropout rates.
- We look at the baseline math test score.
- “Silly” matching algorithm: randomly selects control units to discard.



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

- Balance is a characteristic of sample rather than population
- Even in experiments, (pre-randomization) matching is preferred

Balance Test Fallacy in Experiments

- Hypothesis tests should be used to examine the process of randomization itself but not to look for “significant imbalance”
- Imbalance is a sample concept not a population one, and cannot be eliminated or reduced by randomization
- Only matched-pair or randomized-block designs can eliminate or reduce imbalance

Recent Developments of Matching Methods

- The main problem of matching: balance checking
- Propensity score tautology
- Skip balance checking altogether
- Specify desired degree of balance before matching

- Simple implementation: exact restrictions on key confounders
- Fine matching
- Coarsened exact matching
- Synthetic matching

Concluding Recommendations

- For experimenters:
  1. Unbiasedness should not be your goal.
  2. Use matching methods to improve efficiency.
  3. “block what you can and randomize what you cannot.”
     - Randomized-block design is always more efficient.
     - Matched-pair design is often more efficient.

- For observationalists:
  1. Balance should be assessed by comparing the *sample* covariate differences between the treatment and control groups.
  2. Do not use hypothesis tests to assess balance.
  3. No critical threshold – observed imbalance should be minimized.

Software Implementation

- MatchIt: An R package available at CRAN
- Three commands: `matchit()`, `summary()`, `plot()`
- A number of matching methods are available

- After preprocessing, analyze the data with Zelig, an R package available at CRAN
- Three commands: `zelig()`, `setx()`, `sim()`
- A number of models are available