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

Kosuke Imai

Department of Politics
Princeton University

April 27, 2007

This Talk Draws on the Following Papers:

- Imai, Kosuke, Gary King, and Elizabeth A. Stuart. "Misunderstandings among Experimentalists and Observationalists: Balance Test Fallacies in Causal Inference."
- 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*, Forthcoming.
- Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. (2007). "Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment." *American Journal of Political Science*, Vol. 51, No. 3 (July), pp. 670-689.
- Imai, Kosuke. "Randomization-based Analysis of Randomized Experiments under the Matched-Pair Design: Variance Estimation and Efficiency Considerations."

Matching Methods in Experimental Studies

1. Matched-Pair Design:
   1. Create pairs of observations based on the pre-treatment covariates so that the two observations within each matched-pair are similar.
   2. Randomly assign the treatment within each matched-pair (either simple randomization or complete randomization across pairs).
   3. Inference is based on the average of within-pair estimates.

2. Randomized-block Design:
   1. Form a group or block of (more than two) observations based on the pre-treatment covariates so that the observations within each block are similar.
   2. Complete randomization of the treatment within each block.
   3. Inference is based on the weighted average of within-block estimates.

- Matching and blocking can be based on a single covariate or a function of multiple covariates (e.g., Mahalanobis distance).

An Example of Randomized-block Design

- Japanese election experiment: randomly selected voters are encouraged to look at the party manifestos on the official party websites during Japan's 2004 Upper House election.

	Randomized blocks						
	I	II	III	IV	V	VI	
	Planning to vote		Not planning to vote		Undecided		
	M	F	M	F	M	F	Total
Treatment groups							
DPJ website	194	151	24	33	36	62	500
LDP website	194	151	24	33	36	62	500
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

Matching Methods in Observational Studies

1. Matching:
   - Each treated unit is paired with a similar control unit based on the pre-treatment covariates.
2. 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.
3. Weighting:
   - Weight each observation within the treated or control groups by the inverse of the probability of being in that group.

- Weighting is based on the propensity score (i.e., the probability of receiving the treatment).
- Matching and subclassification are based on the propensity score and other measures of similarity.

Notations and Quantities of Interest

- Notations
  - Population size: N.
  - Sample size: n.
  - Sample selection: I_i.
  - Binary treatment assignment: T_i.
  - Potential outcomes: $Y_i(1)$ and $Y_i(0)$.
  - Observed outcome: $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ for $I_i = 1$.
  - Observed and unobserved pre-treatment covariates: X_i and U_i.
- Quantities of Interest
  1. Unit treatment effect: $\text{TE}_i \equiv Y_i(1) - Y_i(0)$.
  2. Sample average treatment effect: $\text{SATE} \equiv \frac{1}{n} \sum_{i \in \{I_i = 1\}} \text{TE}_i$.
  3. Population average treatment effect: $\text{PATE} \equiv \frac{1}{N} \sum_{i=1}^{N} \text{TE}_i$.
- Super-population
  - N units are sampled from a population of infinite size.
  - Super-population average treatment effect: $\text{SPATE} \equiv E[Y_i(1) - Y_i(0)]$.

A Decomposition of Causal Effect Estimation Error

- A simple estimator of the PATE, SATE, or SPATE:

$$D \equiv \left(\frac{1}{n/2} \sum_{i \in \{I_i=1, T_i=1\}} Y_i \right) - \left(\frac{1}{n/2} \sum_{i \in \{I_i=1, T_i=0\}} Y_i \right).$$

- Estimation error for the PATE: $\Delta \equiv \text{PATE} - D$.
- Consider an additive model: $Y_i(t) = g_t(X_i) + h_t(U_i)$ for $t = 0, 1$.
- New decomposition:

$$\Delta \;=\; \Delta_S + \Delta_T \;=\; (\Delta_{S_X} + \Delta_{S_U}) + (\Delta_{T_X} + \Delta_{T_U}).$$

  1. Δ_S: sample selection error due to observables (Δ_{S_X}) and unobservables (Δ_{S_U}).
  2. Δ_T: treatment imbalance due to observables (Δ_{T_X}) and unobservables (Δ_{T_U}).
- Generalizable to SPATE, various average treatment effects on the treated, and other settings.

Sample Selection Error

- Definition:

$$\Delta_S \equiv \text{PATE} - \text{SATE} = \frac{N-n}{N}(\text{NATE} - \text{SATE}),$$

  where NATE is the *nonsample average treatment effect*, i.e., $\text{NATE} \equiv \sum_{i \in \{I_i=0\}} \text{TE}_i/(N-n)$.

- When does Δ_S equal 0?
  1. $N = n$: The sample is a census of the population.
  2. SATE = NATE: The average treatment effect is the same between the sample and nonsample.
  3. Give up PATE and focus on SATE.

Decomposition of Sample Selection Error

- Sample selection error due to observables:

$$\Delta_{S_X} = \frac{N-n}{N} \int [g_1(X) - g_0(X)] \, d[\widetilde{F}(X \mid I=0) - \widetilde{F}(X \mid I=1)],$$

  where $\widetilde{F}(\cdot)$ represents the empirical distribution function.

- When does Δ_{S_X} equal 0?
  1. $\widetilde{F}(X \mid I=0) = \widetilde{F}(X \mid I=1)$: identical distributions of X_i.
  2. $g_1(X_i) - g_0(X_i) = \alpha$: constant treatment effect over X_i.

- The same argument applies to the sample selection error due to unobservables:

$$\Delta_{S_U} = \frac{N-n}{N} \int [h_1(U) - h_0(U)] \, d[\widetilde{F}(U \mid I=0) - \widetilde{F}(U \mid I=1)].$$

Treatment Imbalance Error

- Decomposition of estimation error due to treatment imbalance:

$$\Delta_{T_X} = \int \frac{g_1(X) + g_0(X)}{2} \, d[\widetilde{F}(X \mid T=0, I=1) - \widetilde{F}(X \mid T=1, I=1)],$$
$$\Delta_{T_U} = \int \frac{h_1(U) + h_0(U)}{2} \, d[\widetilde{F}(U \mid T=0, I=1) - \widetilde{F}(U \mid T=1, I=1)].$$

- Δ_{T_X} and Δ_{T_U} vanish if the treatment and control groups are **balanced** (i.e., have identical empirical distributions):

$$\widetilde{F}(X \mid T=1, I=1) = \widetilde{F}(X \mid T=0, I=1),$$
$$\widetilde{F}(U \mid T=1, I=1) = \widetilde{F}(U \mid T=0, I=1).$$

- The first condition is verifiable from the observed data but the second is not.

The Role of Matching Methods in Causal Inference

- The decomposition formalizes the role of matching methods in reducing the estimation error.
- In both experimental and observational studies, matching methods try to reduce Δ_{T_X} given by:

$$\int \frac{g_1(X) + g_0(X)}{2} d[\widetilde{F}(X \mid T = 0, I = 1) - \widetilde{F}(X \mid T = 1, I = 1)],$$

  which represents the estimation error due to the treatment imbalance in observables.
- They do so by trying to achieve:

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

- In practice, some imbalance remains and needs to be adjusted by randomization (in experimental studies) and model adjustments (observational studies – matching as "nonparametric preprocessing").

Consequences of Design Choices on Estimation Error

Design Choice	Δ_{S_X}	Δ_{S_U}	Δ_{T_X}	Δ_{T_U}
Random sampling	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$		
Focus on SATE rather than PATE	$= 0$	$= 0$		
Weighting for nonrandom sampling	$= 0$	$=?$		
Large sample size	$\to ?$	$\to ?$	$\to ?$	$\to ?$
Random treatment assignment			$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$
Complete blocking			$= 0$	$=?$
Exact matching			$= 0$	$=?$
Assumption				
No selection bias	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$		
Ignorability				$\overset{\text{avg}}{=} 0$
No omitted variables				$= 0$

What is the Best Design? None!

	Δ_{S_X}	Δ_{S_U}	Δ_{T_X}	Δ_{T_U}
Ideal Experiment	$\to 0$	$\to 0$	$= 0$	$\to 0$
Randomized Clinical Trials (Limited or no blocking)	$\neq 0$	$\neq 0$	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$
Randomized Clinical Trials (Complete blocking)	$\neq 0$	$\neq 0$	$= 0$	$\overset{\text{avg}}{=} 0$
Field Experiment (Limited or no blocking)	$\neq 0$	$\neq 0$	$\to 0$	$\to 0$
Survey Experiment (Limited or no blocking)	$\to 0$	$\to 0$	$\to 0$	$\to 0$
Observational Study (Representative data set, Well-matched)	≈ 0	≈ 0	≈ 0	$\neq 0$
Observational Study (Unrepresentative data set, Well-matched)	$\neq 0$	$\neq 0$	≈ 0	$\neq 0$

Fallacies in Experimental Studies

1. 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.

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

Proof that Randomized-block Design is Always Better

1. Variance under the complete-randomized design:

$$\text{Var}_c(D) = \frac{1}{n/2}\left\{\text{Var}(Y_i(1)) + \text{Var}(Y_i(0))\right\}.$$

2. Variance under the randomized-block design:

$$\text{Var}_b(D) = \frac{1}{n/2}\sum_{x\in\mathcal{X}} w_x \left\{(\text{Var}_x(Y_i(1)) + \text{Var}_x(Y_i(0)))\right\},$$

   where $w_x = n_x/n$ is the known population weight for the block formed by units with $X_i = x$.

3. Then, since $\text{Var}(Y(t)) \geq E[\text{Var}_x(Y_i(t))] = \sum_{x\in\mathcal{X}} w_x \text{Var}_x(Y_i(t))$, we have $\text{Var}_c(D) \geq \text{Var}_b(D)$.

Proof that Matched-Pair Design is Usually Better

- Variance under the Matched-Pair design:

$$\text{Var}_m(D) = \frac{1}{n}\{\text{Var}(Y_{1j}(1) - Y_{2j}(0)) + \text{Var}(Y_{2j}(1) - Y_{1j}(0))\}$$

- Then,

$$\begin{aligned}&\text{Var}_c(D) - \text{Var}_m(D)\\ =\ &\frac{n-1}{n(2n-1)}\text{Cov}(Y_{1j}(1) + Y_{1j}(0), Y_{2j}(1) + Y_{2j}(0))\end{aligned}$$

- Thus, unless matching induces negative correlation, the matched-pair design yields more efficient estimates.

Relative Efficiency of the Matched-Pair Design

- One can empirically evaluate the efficiency of the matched-pair design relative to the complete-randomized design.
- Question: Are match-makers dumb?

- Under the matched-pair design, I derive the bounds on the variance one would obtain if the complete-randomized design were employed.
- The bounds are based on:

$$\begin{aligned}\frac{2(n-1)}{2n-1}\mathrm{Var}(Y_{ij}(t)) &\leq \mathrm{Var}(Y_i(t)) \\ &\leq \frac{2(n-1)}{2n-1}\mathrm{Var}(Y_{ij}(t)) + \frac{2}{2n-1}E\{Y_{ij}(t)^2\}\end{aligned}$$

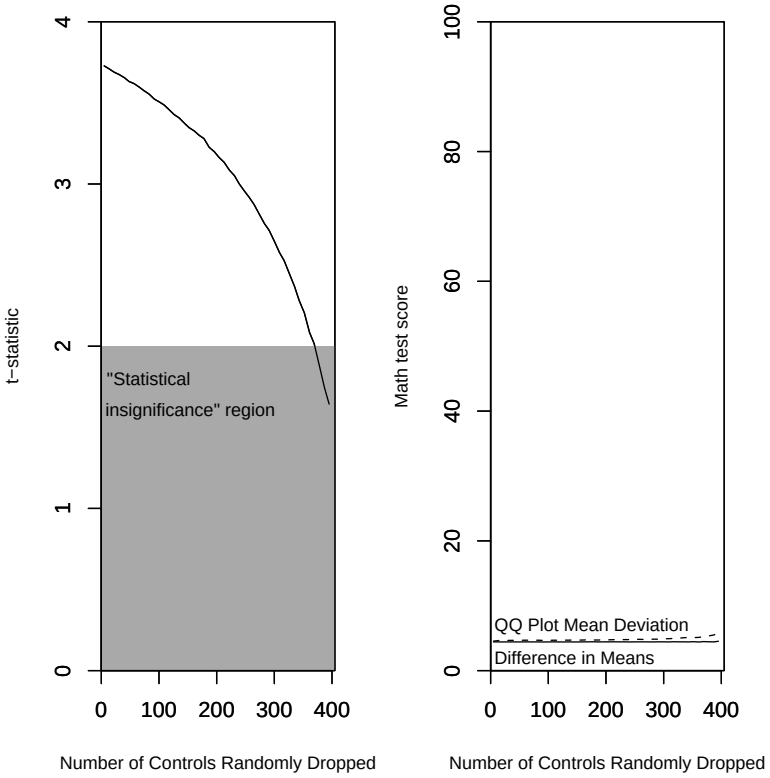
  for $t = 0, 1$.
- The bounds can be estimated without bias.

Assessing the Balance in Matching Methods

- The success of any matching method depends on the resulting covariate balance.
- How should one assess the balance of matched data?
  - Ideally, one would like to compare the joint distribution of all covariates for the matched treatment and control groups.
  - In practice, this is impossible when X is high-dimensional.
- Standard practice: use of balance test
  - t test for difference in means for each variable of X.
  - other test statistics such as χ^2, F, Kolmogorov-Smirnov tests are also used.
  - statistically insignificant test statistics are used as a justification for the adequacy of the chosen matching method and a stopping rule for maximizing balance.
  - the practice is widespread across disciplines (economics, education, management science, medicine, public health, psychology, and statistics).

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

1. Balance test is a function of both balance and statistical power: the more observations dropped, the less power the tests have.
2. 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.
3. Even a small imbalance can greatly affect the estimates.
   - Linear regression: $E(Y \mid T, X) = \theta + T\beta + X\gamma$.
   - Bias: $E(\hat{\beta} - \beta \mid T, X) = G\gamma$ where $\hat{\beta}$ is the difference in means estimator and G contains vector of coefficients from regression of each of the covariates in X on a constant and T.

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.

- For everyone:
  1. There is no best design.
  2. Minimize each component of the estimation error via design and analysis: sample selection and treatment imbalance.