PMAP 8131 Applied Research Methods

Matching Methods

Matteo Zullo

Georgia State University & Georgia Institute of Technology

July 6, 2022

Outline

Table of Contents

Introduction

- Exact Matching (EM)
- Propensity Score Matching (PSM)
  - Nearest Neighbor Matching (NNM)
  - Optimal Matching (OM)
  - Full Matching (FM)
- Inverse Probability Weighting (IPW)
- Genetic Matching (GM)

Introduction

Assess covariate balance: Standardized Mean Differences

$$SMD = \frac{|\overline{X}_{treat} - \overline{X}_{control}|}{(\sqrt{S^2_{treated} - S^2_{control}})/2}$$

- where:
  - X_{treat} and $X_{control}$ are mean characteristics
  - S^2_{treat} and $S^2_{control}$ are variances of characteristics

Introduction

- Example: Age
  - $Age_{treat} \sim (35, 3)$
  - $Age_{control} \sim (38, 4)$

$$SMD = \frac{|35 - 38|}{\sqrt{(9 + 16)/2}} = 0.24$$

Table of Contents

Exact Matching

Exact Matching (EM)

- Treated and controls are the exact same

Coarsened Exact Matching (EM)

- Treated and controls are roughly the same
  - Binning covariate values

Exact Matching

id	treat	age	earnings
001	0	20	50,000
002	0	21	47,000
003	0	54	110,000
004	1	24	52,000
005	1	29	60,000
006	1	57	125,000

Table: Coarsened Exact Matching via binning

Exact Matching

id	treat	age	age_binned	earnings
001	0	20	19-29	50,000
002	0	21	19-29	47,000
003	0	54	49-59	110,000
004	1	24	19-29	52,000
005	1	29	19-29	60,000
006	1	57	49-59	125,000

Table: Coarsened Exact Matching via binning

Table of Contents

Propensity Score Matching

1. Estimate logistic regression model
   - Dependent variable: Binary treatment indicator
   - Independent variable: Covariates predicting treatment

$$P(T_i = 1) = \frac{1}{1 + e^{-\beta X}}$$

2. For each unit, treated and untreated, save predicted probability of receiving treatment (i.e., propensity score)
3. Match treated and controls based on propensity scores

Propensity Score Matching

Predictive equation (Brookhart et al., 2006)

- Variables to include
  - Affect both X and Y
  - Affect Y, but not X
- Variables not to include
  - Affect X, but not Y
  - Affected by X and affect Y

Propensity Score Matching

Nearest Neighbor Matching (NNM)

- Match each treated unit with $(1 : k)$ control units within maximum specified distance of d propensity scores
  - $k =$ neighbors
  - $d =$ caliper (usually between 0.1-0.25)
- STATA's *psmatch2*

Propensity Score Matching

Optimal Matching (OM)

- Minimizes **average distance** between treated and control units given maximum treat-to-control ratio

Propensity Score Matching

Full matching (FM)

- Minimizes **average weighted distance** between treated and control units across different strata
  - Strata are defined from original sample via exact matching, coarsened exact matching, or PSM
  - Matching happens within strata

Propensity Score Matching

NNM vs OM vs FM: 1-to-1

- Units
  - T: $0.80, 0.72$
  - C: $0.88, 0.73, 0.68$
- NNM
  - $T = (0.80, 0.72), C = (0.73, 0.68)$
  - avg. distance $= (0.07 + 0.04) = 0.055$
- FM
  - $T = (0.80, 0.72), C = (0.88, 0.73)$
  - avg. distance $= (0.08 + 0.01) = 0.045$

Propensity Score Matching

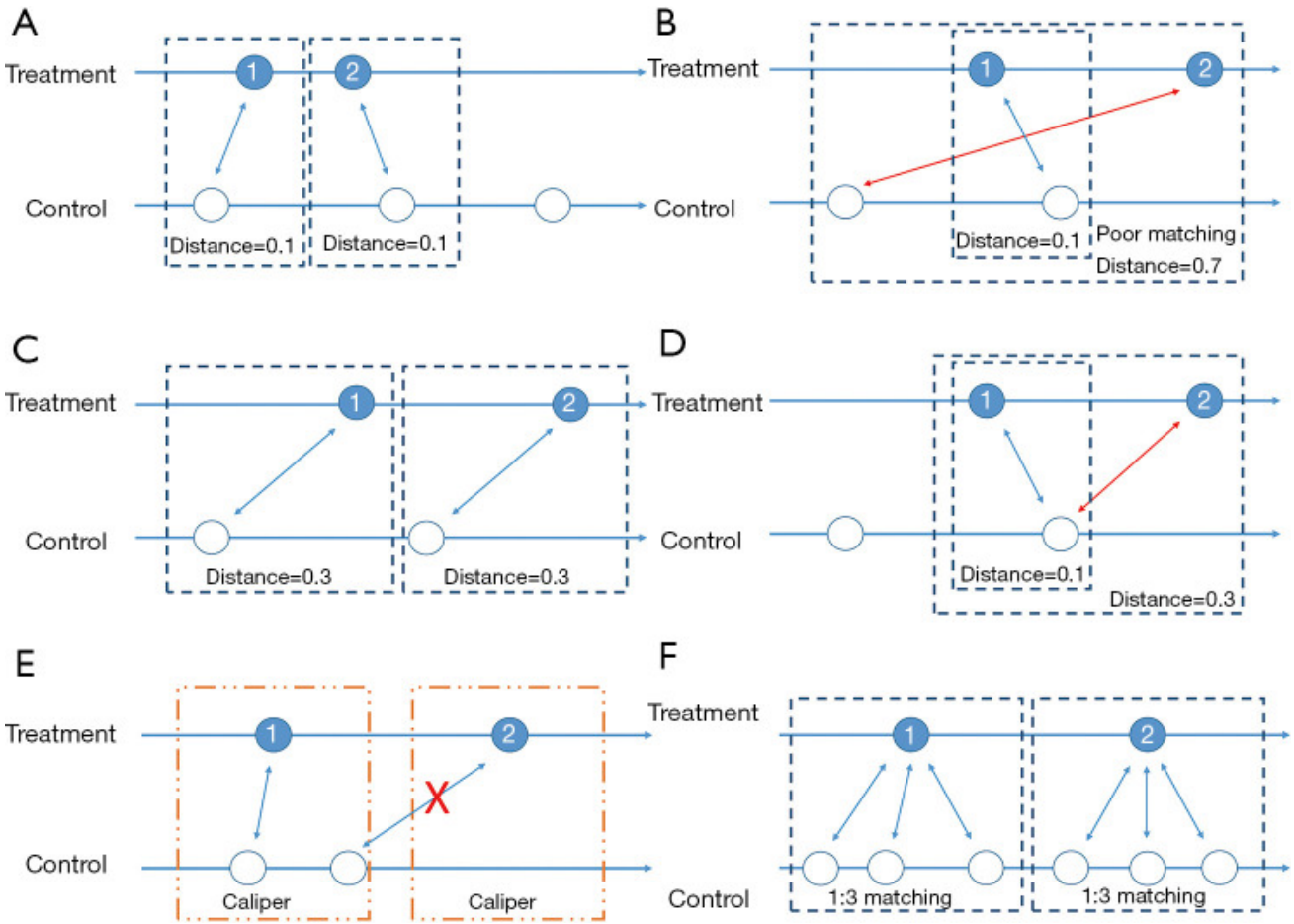

Figure: NNM, OM, FM (Zhao et al., 2021)

Propensity Score Matching

- The PSM paradox: "When you do better, you do worse"
  - "When you do better"
    - When propensity scores are ≈ 0.5
  - "You do worse"
    - PS-based selection discards closer matches on covariates

Propensity Score Matching

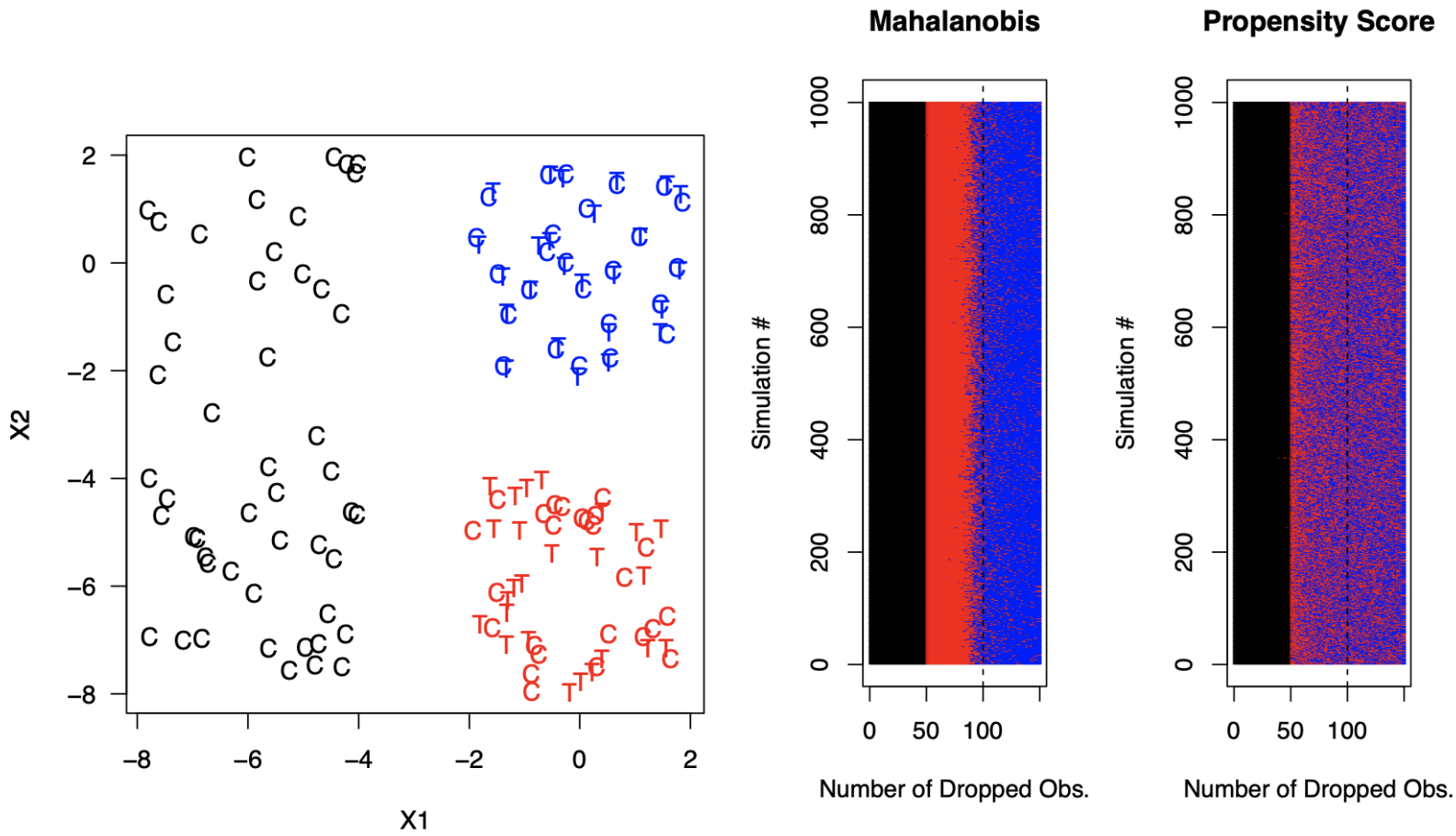

Figure: PSM vs covariate matching (King & Nielsen, 2019)

Propensity Score Matching

- The PSM paradox: "When you do better, you do worse"

id	treat	age_binned	ps
001	0	19-29	0.5
002	0	19-29	0.5
003	0	49-59	0.5
004	1	19-29	0.5
005	1	19-29	0.5
006	1	49-59	0.5

  - Final samples: $\{001, 002, 004, 005\}$, $\{001, 002, 005, 006\}$
  - PS selection: Matched sample 1 has better *age_binned* balance but propensity scores ignore that

Inverse Probability Weighting

1. Estimate propensity scores
2. Calculate inverse probability weights

$$ipw_i = \begin{cases} \frac{1}{P_i}, & \text{if } T_i = 1 \\ \frac{1}{1-P_i}, & \text{if } T_i = 0 \end{cases}$$

3. Calculate weighted outcome difference

$$Y_1^1 - Y_0^1 = \sum_{i=1}^{T} ipw_i Y_i - \sum_{i=1}^{N-T} ipw_i Y_i$$

Inverse Probability Weighting

IPW vs PSM

- Pro: Preserving full sample size (no units discarded!)
- Con: Keeping units outside common support

Inverse Probability Weighting

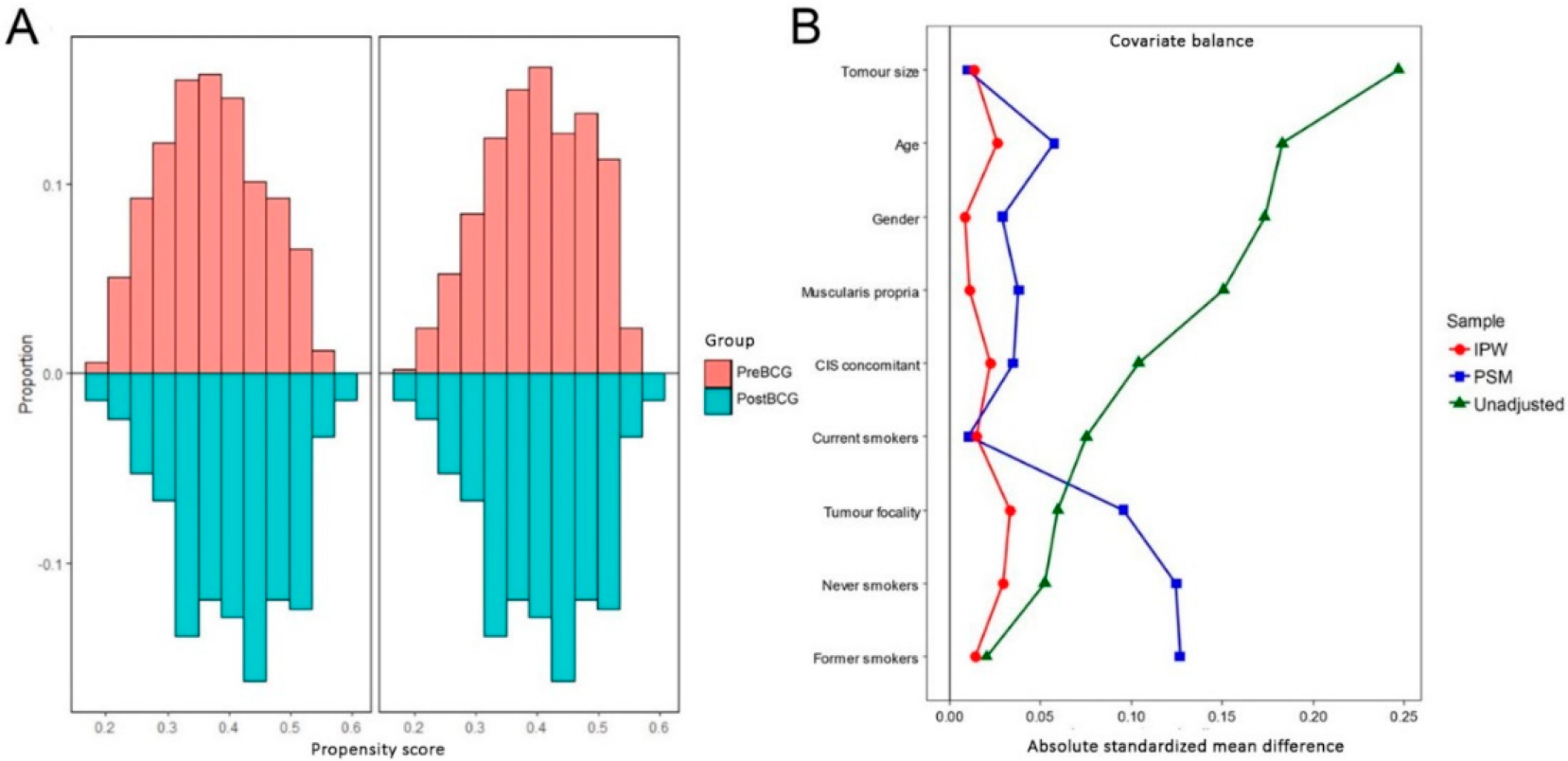

Figure: Covariate balance with PSM and IPW (Krajewski et al., 2020)

Table of Contents

Genetic Matching

1. Initialize vector of weights W on covariates as 1
2. Estimate propensity scores for each unit (optional)
3. Match using Weighted Mahalanobis Distance (WMD)
4. Randomly (i.e., "genetically") modify W deploying K new sets of weights $W_1, \ldots, W_K$ (i.e., "generations")
5. Choose the vector W_k which minimizes SMD

Genetic Matching

1. Initialize vector of weights W on covariates as 1
2. Estimate propensity scores for each unit (optional)
3. Match using Weighted Mahalanobis Distance (WMD)
   - For each treatment, calculate WMD with any control
   - Sequentially match treatments and controls
   - Calculate resulting covariate balance using SMD
4. Randomly (i.e., "genetically") modify W deploying K new sets of weights $W_1, \ldots, W_K$ (i.e., "generations")
   - For each set of weights W_k, repeat step 3
   - Calculate resulting covariate balance using SMD
5. Choose the vector W_k which minimizes SMD
   - STOP if final balance improves on initial, ELSE restart

Genetic Matching

Weighted Mahalanobis Distance

- WMD between treatment i and control j

$$WMD_{ij}(X_i, X_j) = \sqrt{(X_i - X_j)'(\hat{\Sigma}^{-\frac{1}{2}})' W (\hat{\Sigma}^{-\frac{1}{2}})(X_i - X_j)}$$

  - Note: Might include propensity scores in covariates!

- WMD matching
  - Starting with 1st treatment, match to closest control
  - Match 2nd treatment to closest of $N - T - 1$ controls
  - etc.

Genetic Matching

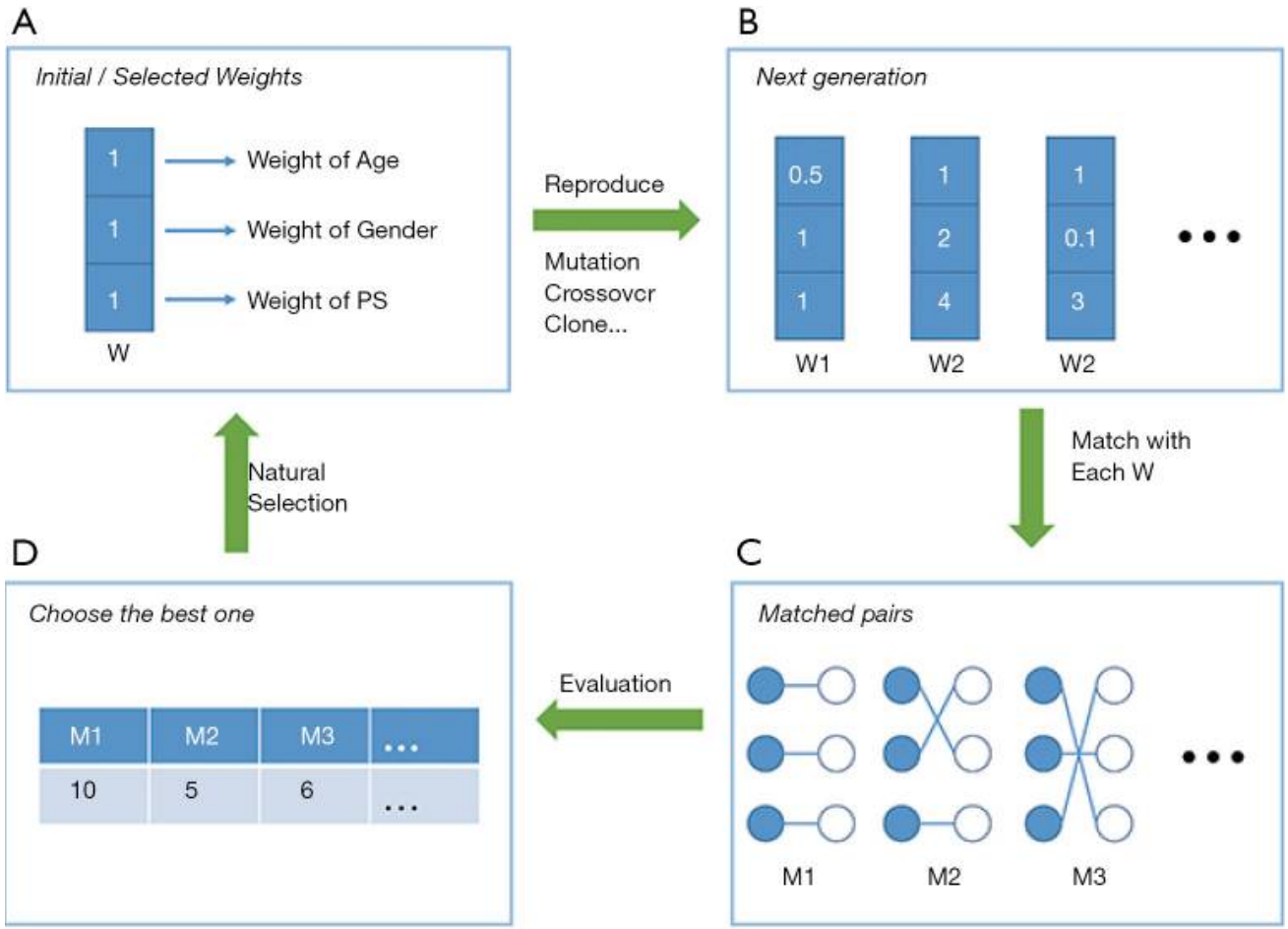

Figure: GM (Zhao et al., 2021)

Genetic Matching

- WMD matching: Covariates $X = [age, ps]$

id	treat	age	ps	WMD_1	WMD_2	WMD_3
001	0	20	0.41	–	–	–
002	0	21	0.42	–	–	–
003	0	54	0.62	–	–	–
004	1	24	0.43	0.22	0.16	2.50
005	1	29	0.46	0.60	0.59	2.02
006	1	57	0.64	2.62	2.55	0.74

Table: Genetic Matching on propensity score and age

Genetic Matching

GM vs PSM vs IPW

- Pros: Matching on covariate space
- Cons: Computationally expensive