Lecture 13
Matching methods

Outline

- Outcome regression V.S. Matching

- Find matched sets

    - Matching metrics and algorithms

    - Check covariate balancing

- Estimate ATT after matching

    - Bias adjustment

- Suggested reading: Imbens and Rubin book Chapter 15 & 18, Peng's book Chapter 15

Causal estimand

- If we treat the units as sampled from a population
  - Population average treatment effect: $\text{PATE} = \text{ATE} = \mathbb{E}(Y_i(1) - Y_i(0))$
  - Average treatment effect for the treated: $\text{PATT} = \text{ATT} = \mathbb{E}(Y_i(1) - Y_i(0) \mid W_i = 1)$
  - Average treatment effect for the control: $\text{ATC} = \mathbb{E}(Y_i(1) - Y_i(0) \mid W_i = 0)$

$$\text{ATE} = P(W_i = 1) \times \text{ATT} + P(W_i = 0) \times \text{ATC}$$

- In randomized experiments, ATE is equivalent to ATT, because treatment and control groups are comparable in expectation

- In observational studies, we can be interested in ATT
  - Many dataset can have a modest number of treated units, but a relatively large pool of possible controls
  - Treated units are more well defined
  - Control units may include units that never have a chance to receive treatment

Outcome regression estimator

- The outcome regression estimator is the same as in conditional randomized experiment

- Under unconfoundedness assumption

$$\tau = \mathbb{E}\left(\mathbb{E}\left(Y_i^{\mathrm{obs}}\middle| \boldsymbol{X}_i, W_i = 1\right) - \mathbb{E}\left(Y_i^{\mathrm{obs}}\middle| \boldsymbol{X}_i, W_i = 0\right)\right)$$

model assumptions on the potential outcomes

- Define the conditional expectations

$$\mu_w(\boldsymbol{x}) = \mathbb{E}\left(Y_i^{\mathrm{obs}}\middle| \boldsymbol{X}_i = \boldsymbol{x}, W_i = w\right) = \mathbb{E}(Y_i(w)| \boldsymbol{X}_i = \boldsymbol{x})$$

- We can estimate the conditional expectations via a regression model and obtain $\hat{\mu}_w(\boldsymbol{x})$
  - Regress Y_i^{obs} on $\boldsymbol{X}_i$ on the treated units and control units separately

- Estimator for the ATE: implement unobserved potential outcome by regression estimates

$$\hat{\tau}_{\mathrm{reg}} = \frac{1}{N}\left\{\sum_{i=1}^{N} W_i\left(Y_i^{\mathrm{obs}} - \hat{\mu}_0(\boldsymbol{X}_i)\right) + (1 - W_i)\left(\hat{\mu}_1(\boldsymbol{X}_i) - Y_i^{\mathrm{obs}}\right)\right\}$$

Regression estimator V.S. Matching

Scatter Plot of Potential and Observed Outcomes

- Estimator for the ATT from regression

$$\hat{\tau}_{\text{reg}} = \frac{1}{N_t} \sum_{i=1}^{N} W_i \left(Y_i^{\text{obs}} - \hat{\mu}_0(\boldsymbol{X}_i) \right)$$

  - Model-based imputation of unobserved potential outcomes
  - Drawbacks:
    - biased imputation if model is wrong
    - If the imbalance of the covariates between the two groups is large, the model-based results heavily relies on extrapolation in the region with little overlap, which is sensitive to the model specification assumption
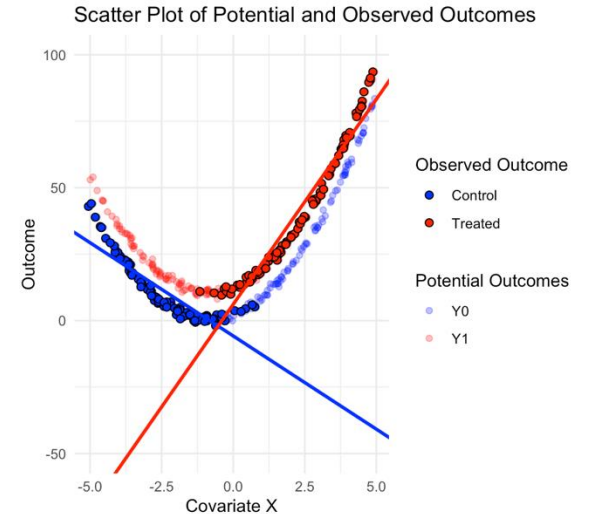
- Matching: nonparametric imputation
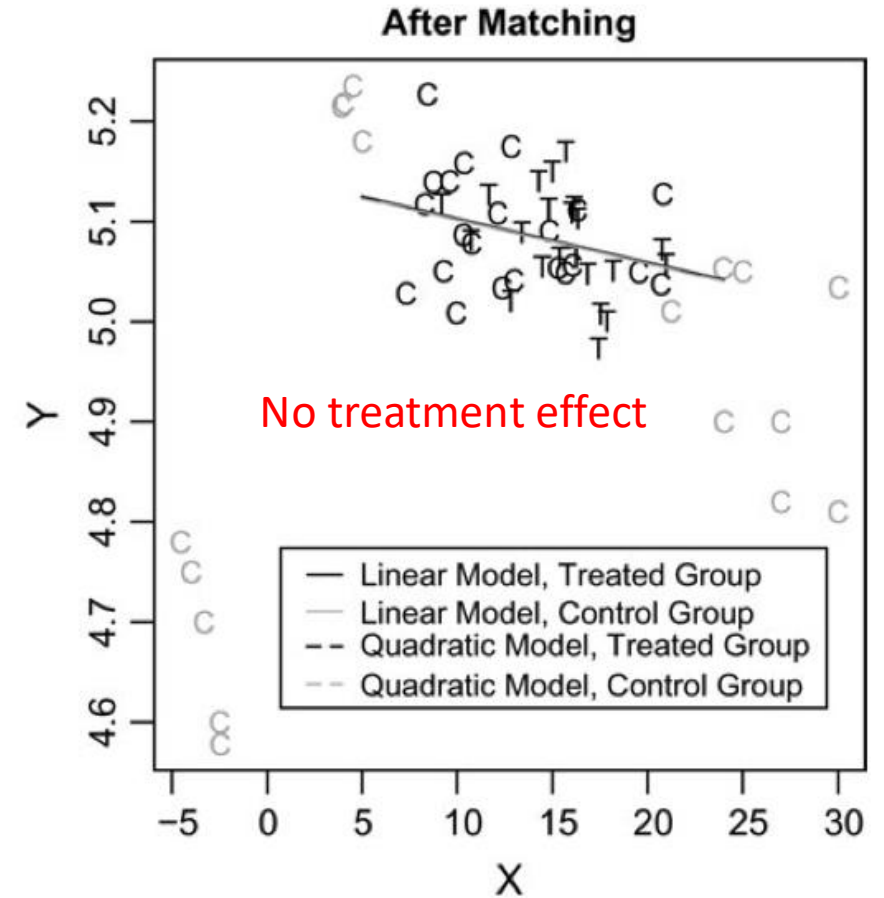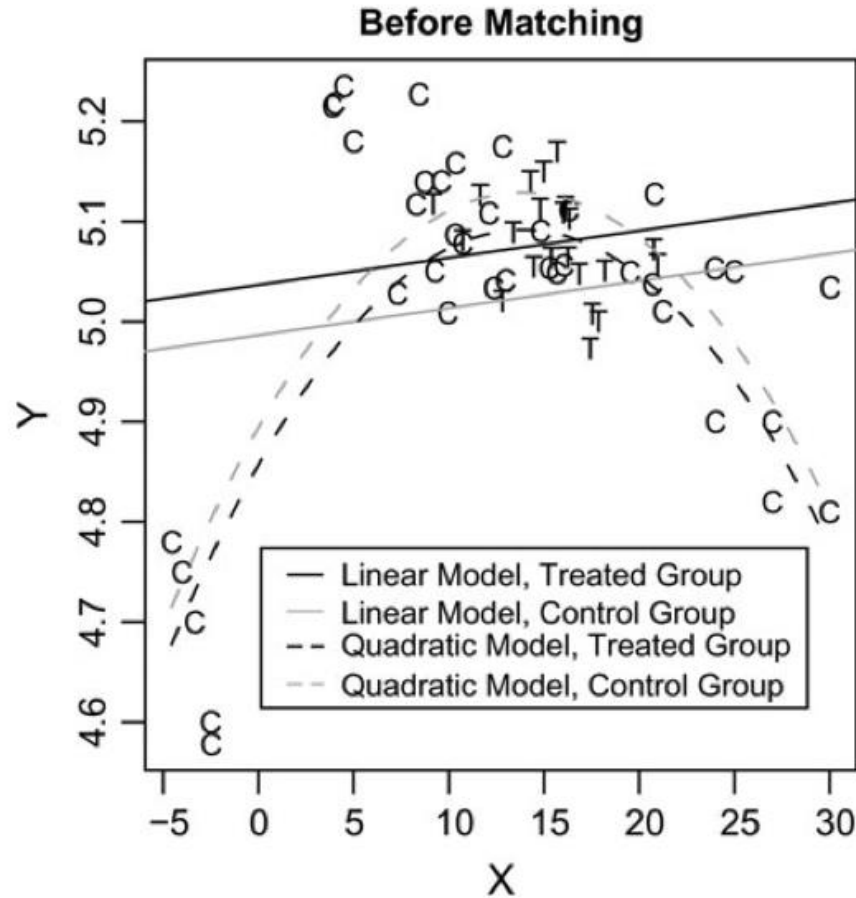
$$\hat{\tau}_{\text{match}} = \frac{1}{N_t} \sum_{i=1}^{N} W_i \left(Y_i^{\text{obs}} - \frac{1}{|\mathcal{M}_i^c|} \sum_{i' \in \mathcal{M}_i^c} Y_{i'}^{\text{obs}} \right)$$

  - $\mathcal{M}_i^c$: matched set of controls for treated unit i

A simulation data example

- Linear regression: positive treatment effect
- Quadratic regression: negative treatment effect
- Both are wrong!!

Before Matching

Linear Model, Treated Group
Linear Model, Control Group
Quadratic Model, Treated Group
Quadratic Model, Control Group

After Matching

No treatment effect

Linear Model, Treated Group
Linear Model, Control Group
Quadratic Model, Treated Group
Quadratic Model, Control Group

- At the two extreme tails of X, there are no treatment units at all

How to find matched sets?

- Matching with replacement v.s. matching without replacement
  - Whether we restrict each control to match with at most one treated unit or not
  - Matching without replacement: harder matching algorithm but easier statistical inference

- Exact match: perfect covariate balance X_i for the matched control(s) are the same as the treated unit
  - Infeasible when covariate is continuous / many covariates

- Coarsened exact matching (Lacus et al. 2011 Political Anal.)
  - discretize covariates so that you can perform exact match

- Matching based on a distance
  - Define a distance measure for any two units: $D(X_i, X_j)$
  - Aim to make units within matched sets as close as possible

Matching based on a distance

- Mahalanobis metric matching

$$D(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)^\top \widehat{\mathbb{V}(\mathbf{X})}^{-1}(\mathbf{X}_i - \mathbf{X}_j)}$$

$\widehat{\mathbb{V}(\boldsymbol{X})} = \frac{N_t \widehat{\Sigma}_t + N_c \widehat{\Sigma}_c}{N_t + N_c}$, $\widehat{\Sigma}_t$ and $\widehat{\Sigma}_c$ are sample covariance matrices for the treated and control

- Propensity score matching

$$D(\boldsymbol{X}_i, \boldsymbol{X}_j) = \left| \ln\left(\frac{\hat{e}(\boldsymbol{X}_i)}{1 - \hat{e}(\boldsymbol{X}_i)} \right) - \ln\left(\frac{\hat{e}(\boldsymbol{X}_j)}{1 - \hat{e}(\boldsymbol{X}_j)} \right) \right|$$

- Hybrid matching methods
  - Ensure exact matching in some key covariates: sex
  - First stratify units by key covariates, match within each strata using distance-based matching

Matching based on a distance

Nearest-neighbor (NN) matching:

- Define $\mathcal{M}_i^c$ as the set of indices of M closest control units

$$\mathcal{M}_i^c = \left\{ j : W_j = 0, \sum_{l|W_l=0} 1_{\{D(X_i, X_j) \leq D(X_i, X_l)\}} \leq M \right\}$$

- Matching with replacement

Greedy algorithm

- Define an order of the treated units
- Match M control units with the shortest distance, set them aside, and repeat
- match most difficult units first: order treated units in a descending order of $\hat{e}(X_i)$

Optimal matching

- $D : N_t \times N_c$ bipartite matrix of pairwise distance or a cost matrix
- Select $N_t M$ elements of D such that there is only M elements in each row and at most one element in each column and the sum of pairwise distances is minimized

Optimal matching

- $D\colon N_t \times N_c$ matrix of pairwise distance or a cost matrix
- Select $N_t M$ elements of D such that there is only M element in each row and at most one element in each column and the sum of pairwise distances is minimized

- Linear Sum Assignment Problem (LSAP)
  - Binary $N_t \times N_c$ matching matrix: S with $S_{ij} \in \{0,1\}$
  - Optimization problem

$$\min_S \sum_{i=1}^{N_t} \sum_{j=1}^{N_c} S_{ij} D_{ij} \quad \text{subject to} \sum_{i=1}^{N_t} S_{ij} \leq 1, \; \sum_{j=1}^{N_c} S_{ij} = M$$

  - can apply the Hungarian algorithm

A simple illustrative example

- Consider 7 units

- Matching based on the linearized estimated propensity score

$$\hat{l}(\boldsymbol{X}_i) = \ln\left(\frac{\hat{e}(\boldsymbol{X}_i)}{1 - \hat{e}(\boldsymbol{X}_i)}\right)$$

- Treated unit 1 matched with control unit 5
- Treated unit 2 matched with control unit 3

- NN, greedy algorithm and optimal matching result in the same matched sets here

Unit	W_i	$\hat{e}(X_i)$	$\hat{\ell}(X_i)$
1	1	0.577	0.310
2	1	0.032	−3.398
3	0	0.136	−1.846
4	0	0.003	−5.913
5	0	0.310	−0.798
6	0	0.000	−9.424
7	0	0.262	−1.033

Further restrictions on the matched sets

- Rejecting matches of poor quality
  - For some units, even the closest match may not be close enough
  - Drop treated units if it's hard to find a good match. E.x., drop i if
  $$D\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) > d_{\max} = 0.1$$
    - Often eliminate only treated units with propensity score very close to 1

- How to determine M?
  - $M = 1$
  - Matching with Caliper: controls that are outside of some distance (caliper) of a treated unit are not allowed to be matched with the treated units.
    - Keep all controls j satisfying $D\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) \leq d_{\mathrm{cal}}$
    - Can use greedy algorithm
    - Optimal matching: define $D_{ij} = \infty$ if $D_{ij} > d_{\mathrm{cal}}$
  - M increases with sample size
  - Smaller M, smaller bias but larger variance; larger M, larger bias but smaller variance

Check covariate balancing after matching

- Statistics we can use to assess the balancing of a particular covariate
  - **Standardized mean difference** (also called the normalized difference, not the t-statistics)

$$\Delta_{ct} = \frac{\frac{1}{N_t}\sum_{i=1}^{N} W_i \left(X_{ik} - \frac{1}{|\mathcal{M}_i^c|}\sum_{i' \in \mathcal{M}_i^c} X_{i'k} \right)}{\sqrt{s_t^2}}$$

  May compare Δ_{ct} with 0.1
    - Before matching, we may calculate the denominator of Standardized mean difference as $\sqrt{(s_t^2 + s_c^2)/2}$
  - **Log ratio of the sample variances** $\Gamma_{ct} = \ln(s_t) - \ln(s_c)$
  - Comparing the distribution function in the treated group and control group
    - Empirical cdf: $\hat{F}_c(x) = \frac{1}{N_c}\sum_{i:W_i=0} \mathbf{1}_{X_i \leq x}$, and $\hat{F}_t(x) = \frac{1}{N_t}\sum_{i:W_i=1} \mathbf{1}_{X_i \leq x}$

    - Proportion of treated units outside of the 2.5% and 97.5% quantiles of the control distribution

$$\hat{\pi}_t^{0.05} = \left(1 - \left(\hat{F}_t \left(\hat{F}_c^{-1}(0.975) \right) \right) \right) + \hat{F}_t \left(\hat{F}_c^{-1}(0.025) \right) \right)$$

Love plot
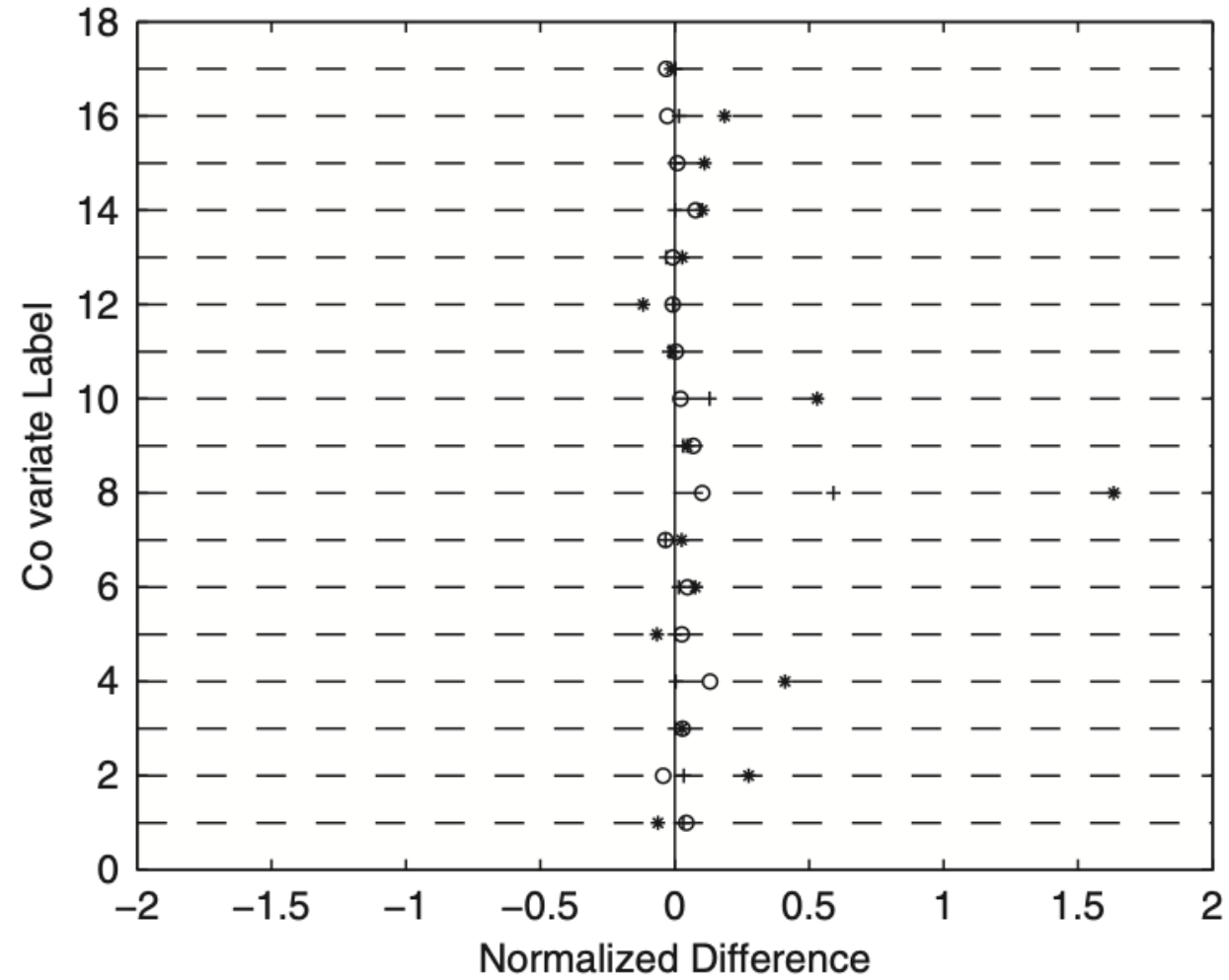
Figure 15.2. Covariate balance before (*) and after (+) lps and after Mahalanobis (o) matching, for the Reinisch barbiturate data

How to estimate ATT after matching

- Unless exact matching, under unconfoundedness, the probability of assignment to the treatment is only approximated the same within each matched set

- In practice, one may **ignore** the potential bias, and analyze the datasets as from a pairwise randomized experiment

$$\hat{\tau}_i^{\text{match}} = Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}}, \qquad \hat{\tau}_t^{\text{match}} = \frac{1}{N_t} \sum_{i:W_i=1} \hat{\tau}_i^{\text{match}}$$

$$\hat{\mathbb{V}}\left(\hat{\tau}_t^{\text{match}}\right) = \frac{1}{N_t(N_t - 1)} \sum_{i:W_i=1} \left(Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}} - \hat{\tau}_t^{\text{match}}\right)^2$$

- Another approach is to apply outcome regression on the matched dataset
  - Treat matching as a pre-processing step to improve covariate balancing in the dataset
  - Reduce bias in matching
  - Or we can use regression to only adjust for the potential biases (see later)

The minimum wage data

- An influential study by Card and Krueger (1995)
- The goal is to evaluate the effect of raising the state minimum wage in Ney Jersey in 1993

- They collected data on employment at fast-food restaurants in Ney Jersey (treated group) and in neighboring state of Pennsylvania (control group)

- Each unit is a restaurant
- Pre-treatment covariates: initial number of employees, starting wage, average time until first raise, identity of the chain
- Outcome: number of employees after the raise in the minimum wage

The minimum wage data

Table 18.1. *The Card-Krueger New Jersey and Pennsylvania Minimum Wage Data*

	(N = 347)		(N_c = 68) (controls)		(N_t = 279) (treated)		Nor	Log Ratio
	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Dif	of STD
initial empl	17.84	(9.62)	20.17	(11.96)	17.27	(8.89)	−0.28	−0.30
burger king	0.42	(0.49)	0.43	(0.50)	0.42	(0.49)	−0.02	−0.01
kfc	0.19	(0.40)	0.13	(0.34)	0.21	(0.41)	0.20	0.17
roys	0.25	(0.43)	0.25	(0.44)	0.25	(0.43)	0.00	−0.00
wendys	0.14	(0.35)	0.19	(0.40)	0.13	(0.33)	−0.18	−0.18
initial wage	4.61	(0.34)	4.62	(0.35)	4.60	(0.34)	−0.05	−0.02
time until raise	17.96	(11.01)	19.05	(13.46)	17.69	(10.34)	−0.11	−0.26
pscore	0.80	(0.05)	0.79	(0.06)	0.81	(0.04)	0.28	−0.35
final empl	17.37	(8.39)	17.54	(7.73)	17.32	(8.55)		

The minimum wage data

Estimated propensity score model:
Higher initial employment, lower propensity score

$$\hat{l}(X_i) = 1.93 - 0.03 \times \text{ initial empl}$$

Table 18.2. *Estimated Parameters of Propensity Score for the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data*

Variable	Est	(s. e.)	t-Stat
Intercept	1.93	(0.14)	14.05
Linear terms			
initial empl	−0.03	(0.01)	−2.17

The minimum wage data on 20 units

Unit i	State W_i	chain X_{i1}	initial empl X_{i2}	final empl Y_i^{obs}
1	NJ	BK	22.5	40.0
2	NJ	KFC	14.0	12.5
3	NJ	BK	37.5	20.0
4	NJ	KFC	9.0	3.5
5	NJ	KFC	8.0	5.5
6	PA	BK	10.5	15.0
7	PA	KFC	13.8	17.0
8	PA	KFC	8.5	10.5
9	PA	BK	25.5	18.5
10	PA	BK	17.0	12.5
11	PA	BK	20.0	19.5
12	PA	BK	13.5	21.0
13	PA	BK	19.0	11.0
14	PA	BK	12.0	17.0
15	PA	BK	32.5	22.5
16	PA	BK	16.0	20.0
17	PA	KFC	11.0	14.0
18	PA	KFC	4.5	6.5
19	PA	BK	12.5	31.5
20	PA	BK	8.0	8.0

- Matching order:
  if we rank based on $\hat{e}(X_i)$: 5, 4, 2, 1, 3
- Matching metric:
  - Only based on $\hat{l}(X_i)$: 20, 8, 7, 11, 15

  - If we want exact match on the chain brand
    5 <-> 8, 4 <->17, 2 <->7, 1<-> 11, 3 <-> 15

  - If we want to match on Mahalanobis distance, can code the restaurant brand by 0/1 indicators, then 5 <-> 20, 4 <-> 8

The minimum wage data on 20 units

i	m_i^c	Y_i^{obs}	$Y_{m_i^c}^{obs}$	$\hat{\tau}_i^{match}$	i	m_i^c	Y_i^{obs}	$Y_{m_i^c}^{obs}$	$\hat{\tau}_i^{match}$
1	11	40.0	19.5	20.5	1	11	40.0	19.5	20.5
2	7	12.5	17	−4.5	2	7	12.5	17.0	−4.5
3	15	20.0	22.5	−2.5	3	15	20.0	22.5	−2.5
4	8	3.5	10.5	−7	4	17	3.5	14	-10.5
5	20	5.5	8.0	−2.5	5	8	5.5	10.5	-5

$\hat{\tau}_t^{match}$: +0.8 $\hat{\tau}_t^{match}$: −0.4

$\hat{\mathbb{V}}\left(\hat{\tau}_t^{match}\right)$: 5.0^2 5.4^2

The bias of matching estimators (1-1 matching)

- Individual treatment effect is estimated with a bias due to matching discrepancy

$$\mathbb{E}_{\text{sp}}\left[\hat{\tau}_i^{\text{match}}\middle| W_i = 1, X_i, X_{m_i^c}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1) - Y_{m_i^c}(0)\middle| X_i, X_{m_i^c}\right] = \mu_{\text{t}}(X_i) - \mu_{\text{c}}(X_{m_i^c})$$

$$= \tau(X_i) + (\mu_{\text{c}}(X_i) - \mu_{\text{c}}(X_{m_i^c})).$$

We refer to the last term of this expression,

$$B_i = \mu_{\text{c}}(X_i) - \mu_{\text{c}}(X_{m_i^c}),$$

as the *unit-level bias* of the matching estimator.

- If we can have estimates of B_i, then we can potentially correct for the biases
- We can obtain the estimates of B_i by outcome regression: only need an estimate $\hat{\mu}_0(X_i)$

$$\hat{\tau}_i^{\text{match}} = Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}} + \hat{B}_i$$

Three types of regression

- Regression on the differences

$$Y_i^{obs} - Y_{m_i^c}^{obs} = \tau + \left(X_i - X_{m_i^c}\right)\beta_d + v_i = \tau + D_i\beta_d + v_i$$

$$\hat{B}_i = \left(X_i - X_{m_i^c}\right)\hat{\beta}_d.$$

- Regression only on the matched control

$$Y_{m_i^c} = \alpha_c + X_{m_i^c}\beta_c + v_{ci}$$

$$\hat{B}_i = (X_i - X_{m_i^c})\hat{\beta}_c$$

- Regression on both the treated and the matched controls (pooled sample)

$$\tilde{Y}_i = \alpha_p + \tau_p \cdot \tilde{W}_i + \tilde{X}_i\beta_p + v_i$$

$$\hat{B}_i = (X_i - X_{m_i^c})\hat{\beta}_p$$

- These methods differ in their robustness to model assumptions and efficiency

Results on the 20 units

	Difference Regression (Approach #1)	Control Regression (Approach #2)	Pooled Regression (Approach #3)
Regression coefficients			
Intercept	−1.30	4.21	12.01
Treatment indicator	−	−	1.63
Restaurant chain	−1.20	2.65	−7.32
Initial employment	1.43	0.62	0.39

- Different regression methods differ a lot because small sample size
- In real data, they are typically similar