Applied Micro-Econometrics,Fall 2023

Lecture 5: Matching

Zhaopeng Qu

Nanjing University Business School

11/7/2023

Review the Last Lecture

OLS and Controls

- The main identification strategy of OLS regression is **Control**, ie. putting **covariates** into the regression as control variables.

- The main identifying assumption of an OLS regression is

  - **Conditional Independence Assumption(CIA)**: which means that if we can "balance" covariates X then we can take the treatment D as randomized, thus

$$(Y_1, Y_0) \perp\!\!\!\perp D | X$$

  - Then ATE or ATT can be obtained to estimate the CEF

$$\delta = E[Y_{1i} - Y_{0i} \mid X_i]$$

- Essentially the strategy compares treatment and control subjects who have **the same observable characteristics**, which is often called **Selection on observables**.

Internal v.s. External Validity

- There are five primary threats to the internal validity of a multiple regression study:

    1. Omitted variables
    2. Functional form misspecification
    3. Errors in variables (measurement error in the regressors)
    4. Sample selection
    5. Simultaneous causality

- the data structure may violate the 2th OLS regression assumption, thus random sampling.

    1. Times series(including Panel)
    2. Cluster data
    3. Spatial data

    - Last but not least, the magnitude of $\hat{\beta}$ matters.

Matching: Introduction

Introduction

- In observational studies, we cannot obtain the causal effect directly because the counterfactural outcome of the treated group is unknown(in other words we cannot find a proper control group).

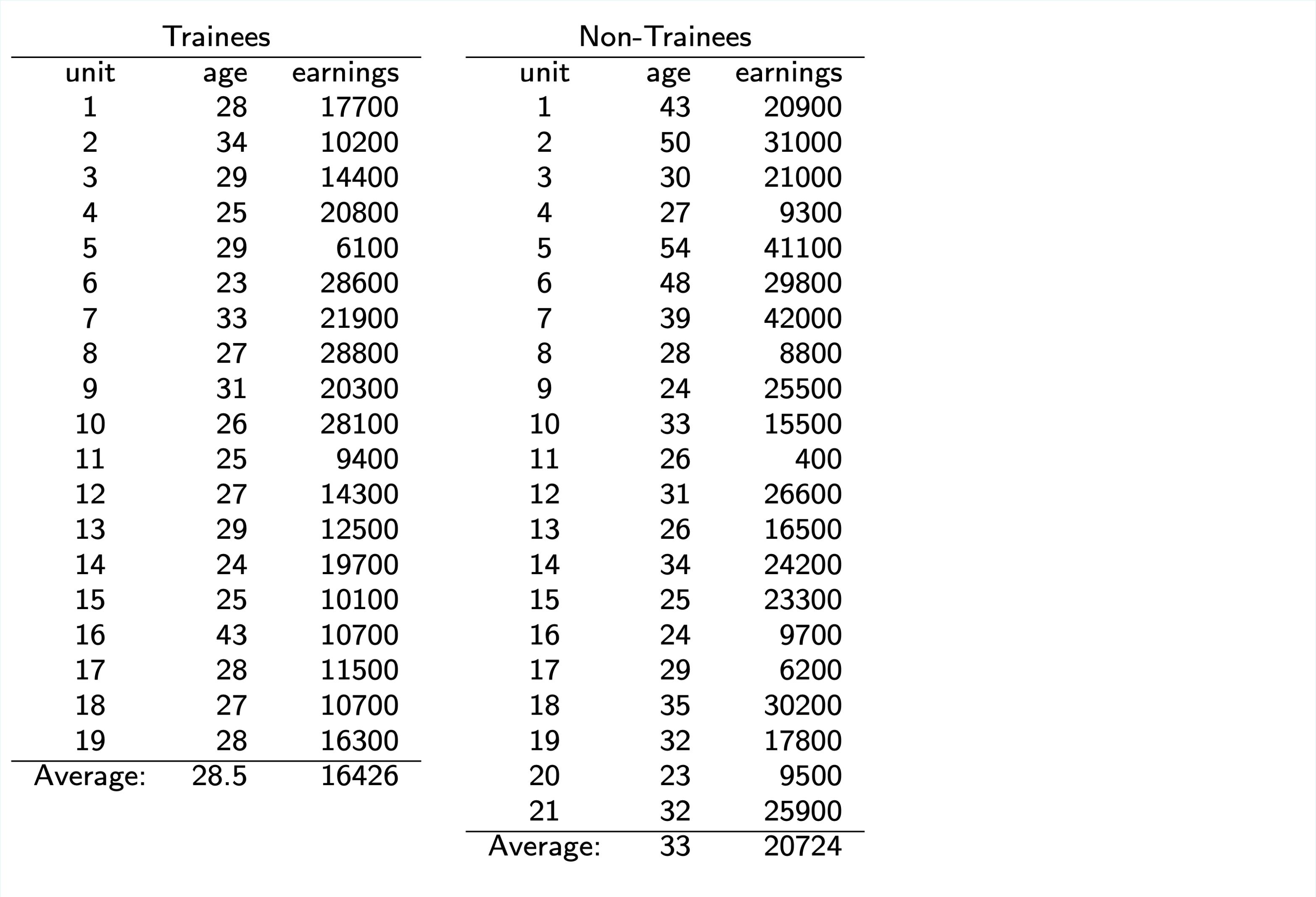
- The idea of matching method is quite simple.
  - **What if we can construct a reasonable control group by selecting some(or all) samples in untreated group in some way**

- Then we can obtain the treatment effect easily by making a difference

$$\delta_i = Y_{i1} - Y_{i0}^c$$

  - Y_i^c is the corresponding counterfactual outcomes by matching(selecting) the sample from untreated group.

A Trainning Example



Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

A Trainning Example

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample

unit	age	earnings
8	28	8800
Average:		

A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
Average:		

A Trainning Example

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample

unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
Average:		

A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
Average:		

A Trainning Example

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample

unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
Average:		

A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
Average:		

A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
Average:		

A Trainning Example

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample

unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
Average:		

A Trainning Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:	33	20724			

A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
Average:		

A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
Average:		

A Trainning Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:	33	20724			

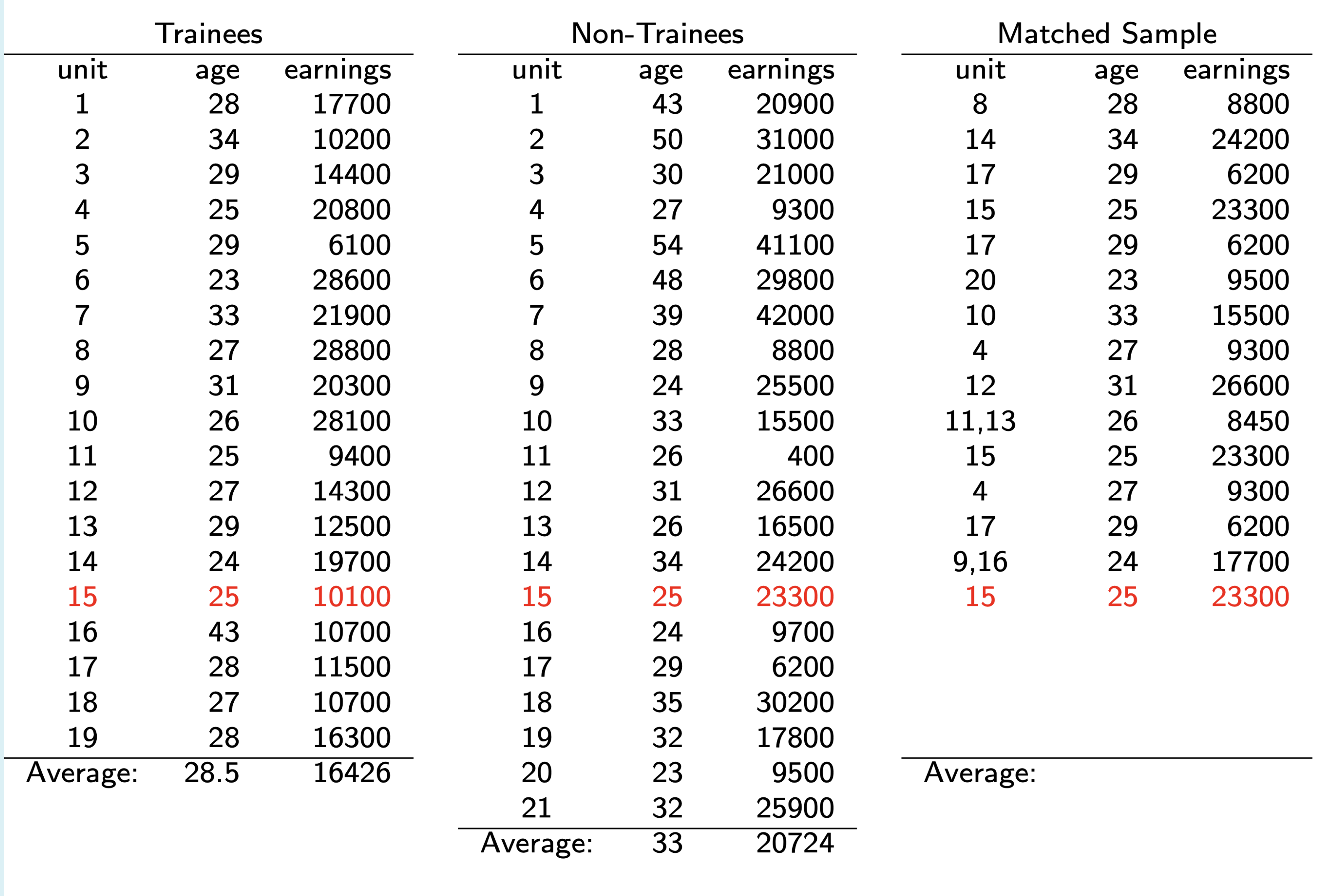
A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
Average:		

A Trainning Example



Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:	33	20724			

A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
15	25	23300
1	43	20900
Average:		

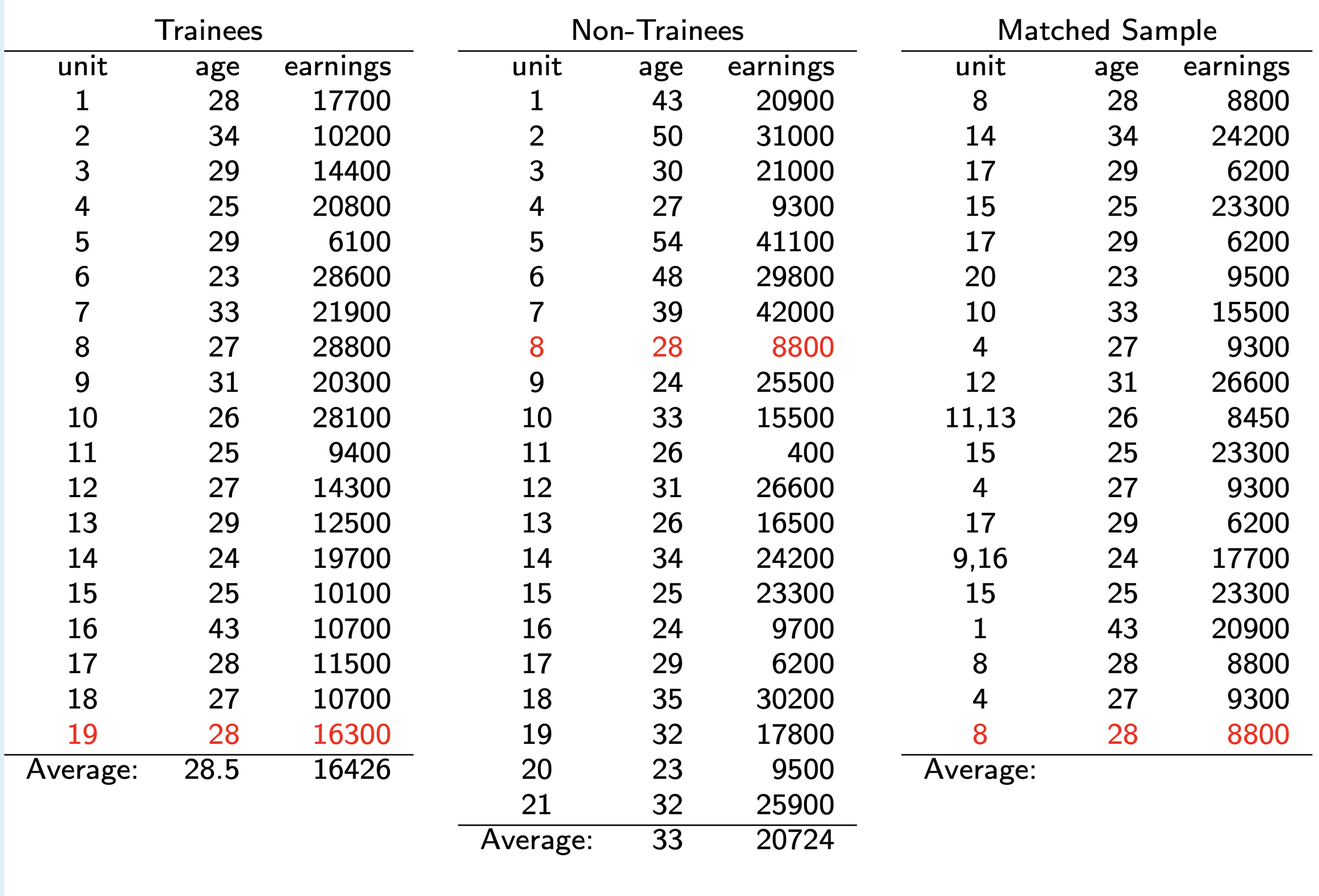
A Trainning Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

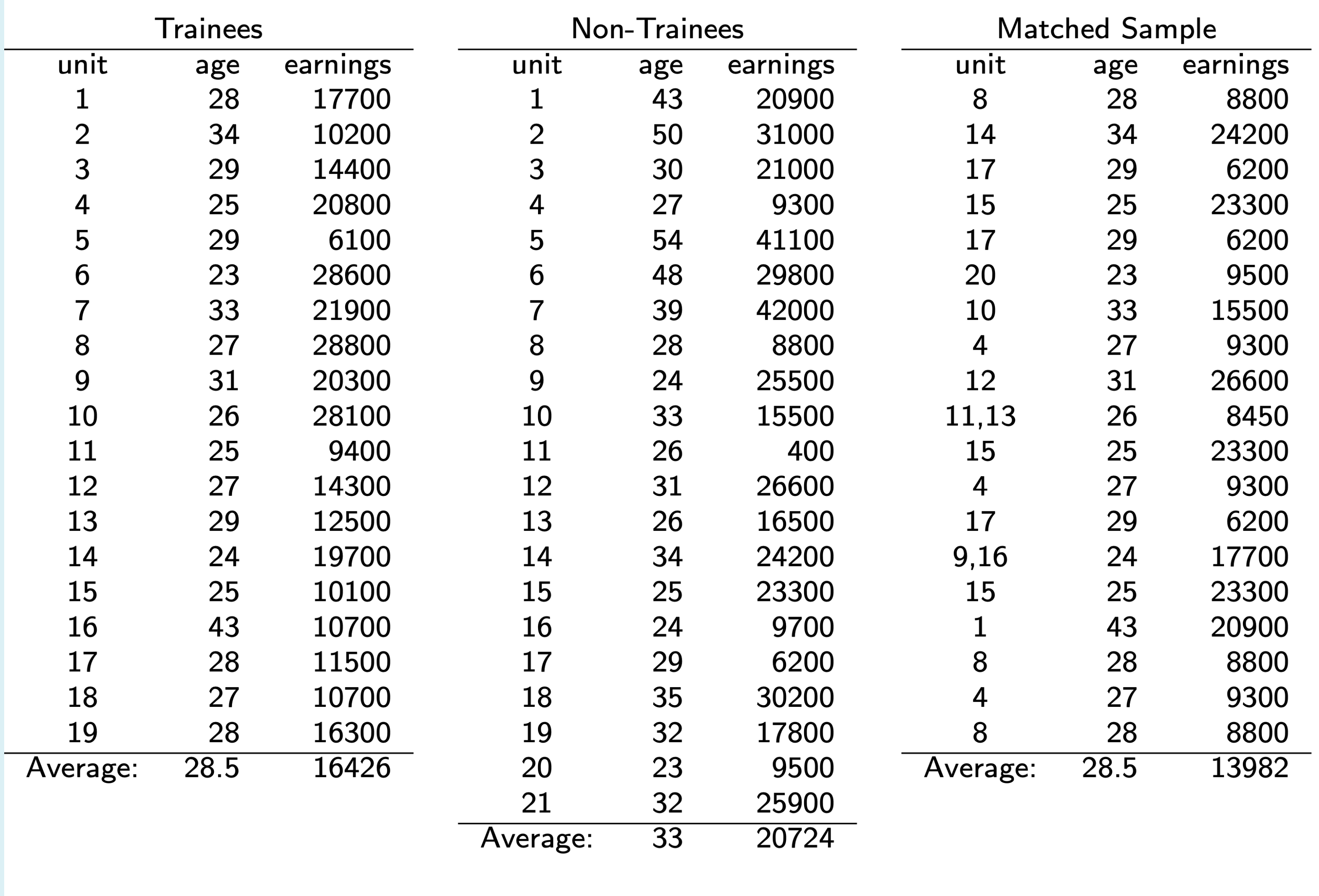
Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
15	25	23300
1	43	20900
8	28	8800
Average:		

A Trainning Example



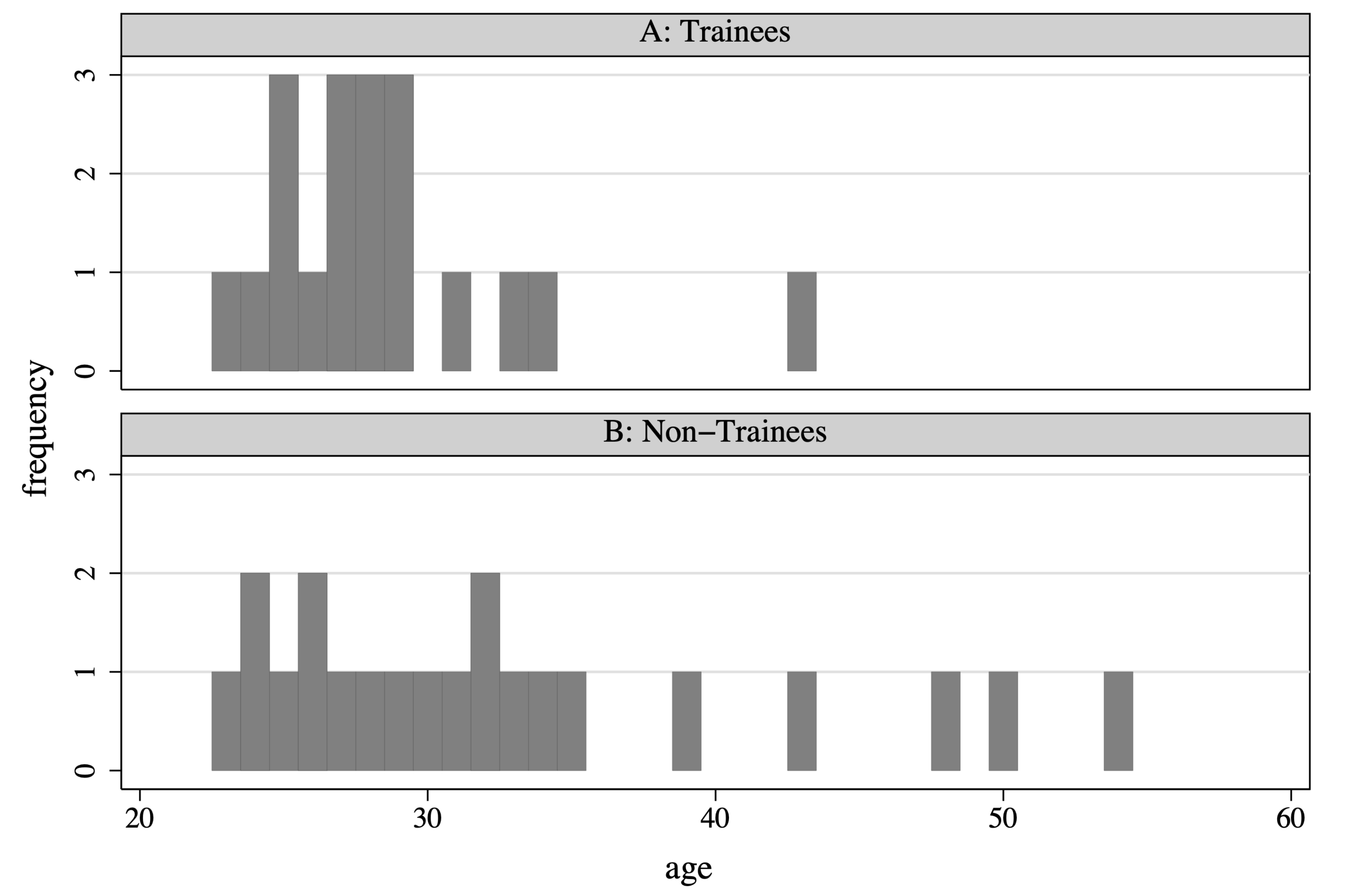
Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800	8	28	8800
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:	33	20724			

A Trainning Example

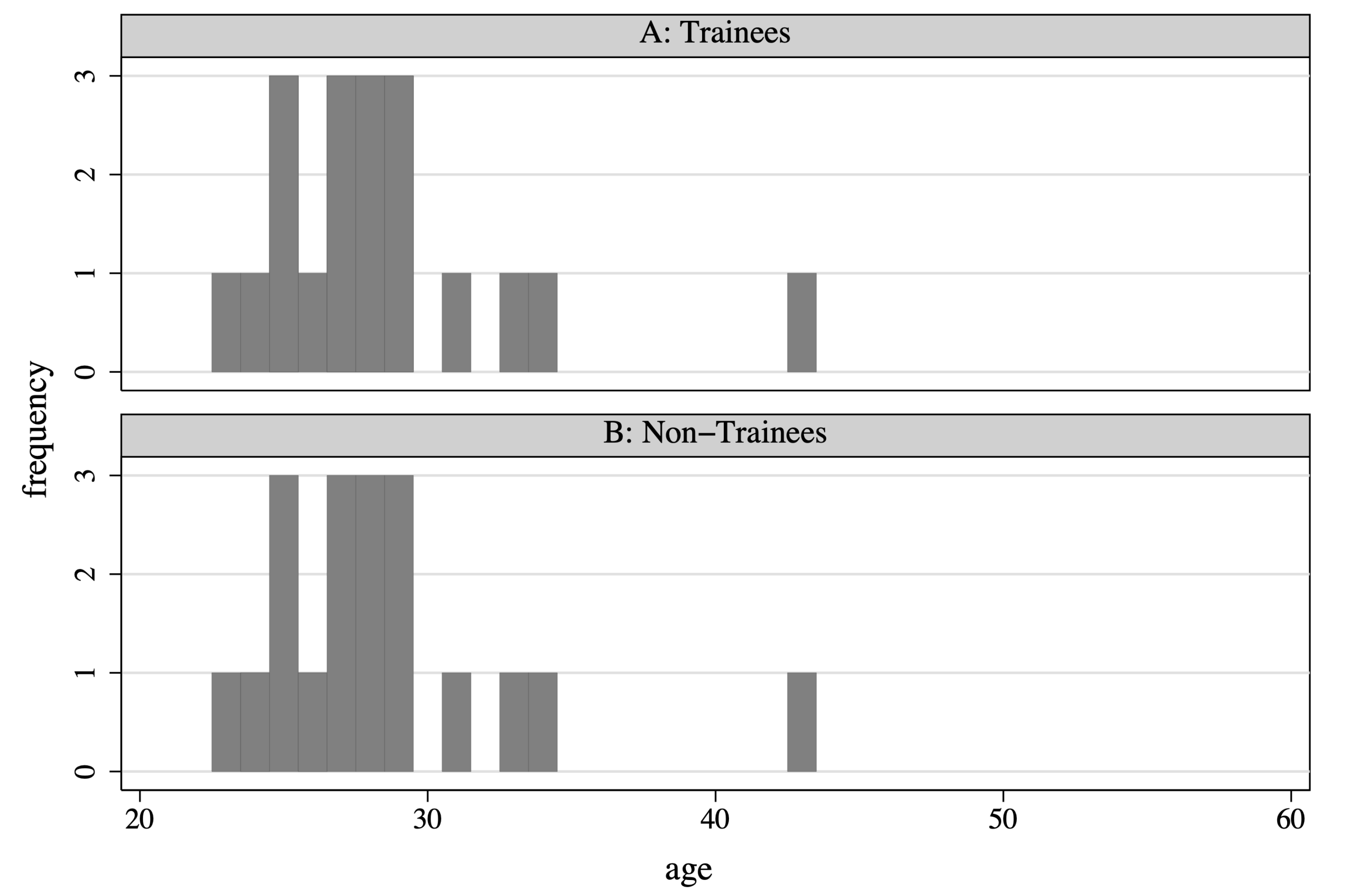


Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800	8	28	8800
Average:	28.5	16426	20	23	9500	Average:	28.5	13982
			21	32	25900			
			Average:	33	20724			

A Trainning Example: before matching



A Trainning Example: after matching

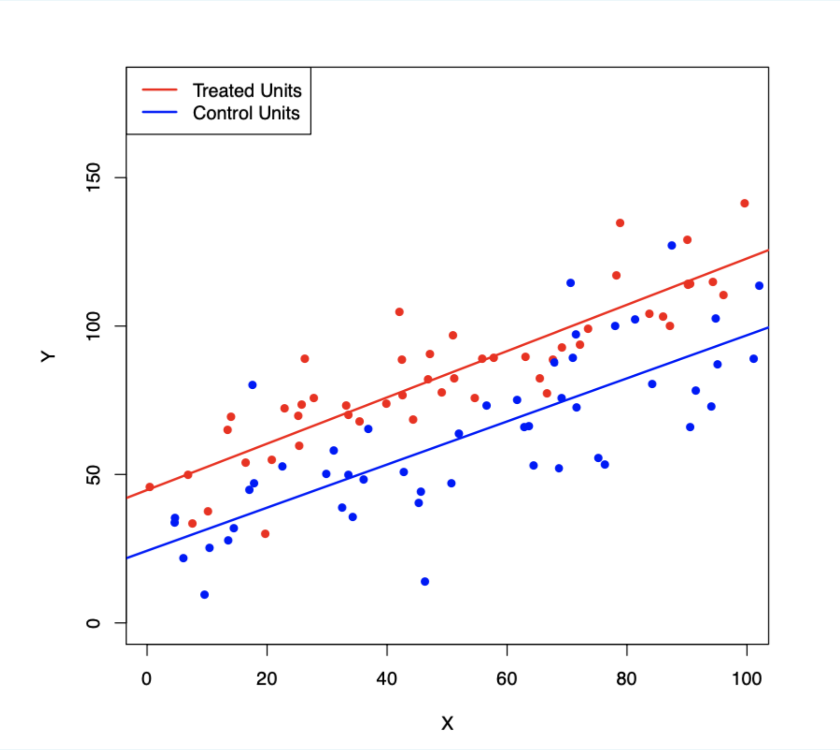


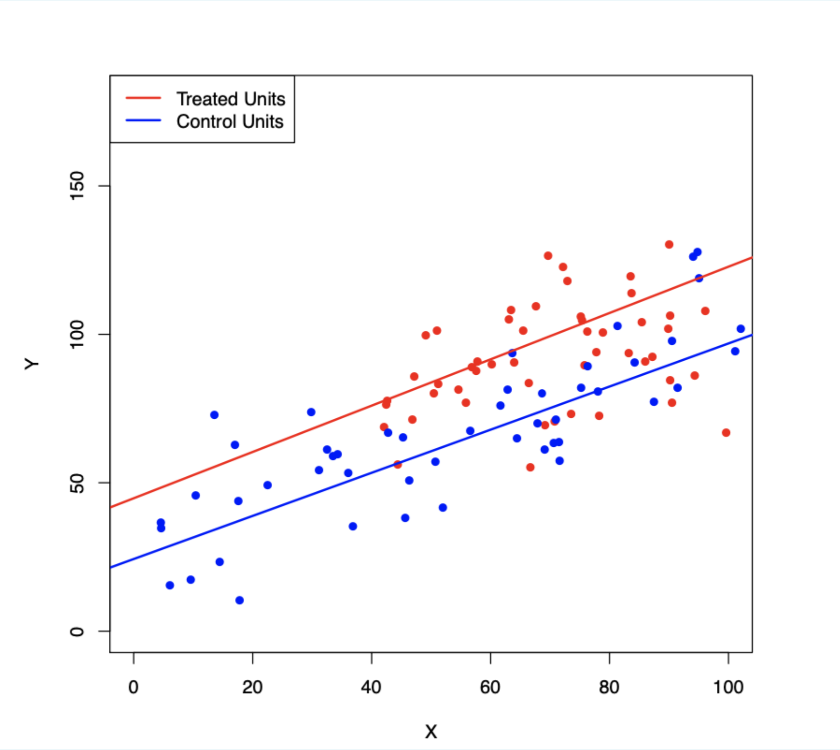
Two Assumptions

- Two assumptions: one old and one new.

1. **Conditional independence:** $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$

2. **Overlap:** $0 < \Pr(D_i = 1 \mid X_i) < 1$





Matching Estimators: Exact matching is hard

- The training case is an example of **Exact** matching which means that only units with identical covariate values are used to construct the control group.

- But what if we have multiple covariates using to match, thus $X = (X_1, X_2, \ldots X_k)'$?

  - In this case, it is **impossible** to find proper units with identical values in all covariates $X_1, X_2, \ldots X_k$.

- Two complementary solutions running in parallel

  1. **lower the accuracy of the comparison**.

  - From *find a unit in the untreated group with the **same** covariate values* to *find a unit in the untreated group with **similar** covariate values*.

  1. **Directly reduce dimensionality** by converting multiple variables into a single numerical value.

- Actually matching methods develop on both directions.

Matching: Theory and Application

Formally

- Construct a counterfactual for each individual with $\mathrm{D}_i = 1$.
- Based on **CIA**,the counterfactual for i should only use individuals that match X_i.
- Let there be N_T treated individuals and N_C control individuals.
- There is a weight to adjust Y_j: $w_i(j)\ (i = 1, \ldots, N_T;\ j = 1, \ldots, N_C)$
- Assume $\sum_j w_i(j) = 1$. Our estimate for the counterfactual of treated i is

$$\widehat{\mathrm{Y}_{0i}} = \sum_{j \in (D=0)} w_i(j)\mathrm{Y}_j$$

Matching: Theory and Application

More formally

- If our estimated counterfactual for treated individual i is

$$\widehat{\mathrm{Y}_{0i}} = \sum_j w_i(j)\mathrm{Y}_j$$

- then our estimated treatment effect (for individual i) is

$$\hat{\tau}_i = \mathrm{Y}_{1i} - \widehat{\mathrm{Y}_{0i}} = \mathrm{Y}_{1i} - \sum_j w_i(j)\mathrm{Y}_j$$

$\therefore$ a generic matching estimator for the treatment effect on the treated is

$$\hat{\tau}_M = \frac{1}{N_T} \sum_{i \in (\mathrm{D}=1)} \left(\mathrm{Y}_{1i} - \widehat{\mathrm{Y}_{0i}}\right) = \frac{1}{N_T} \sum_{i \in (\mathrm{D}=1)} \left(\mathrm{Y}_{1i} - \sum_{j \in (D=0)} w_i(j)\mathrm{Y}_j\right)$$

Matching: Theory and Application

Weight for it†

- **Question** How to obtain these weights?

- **Answer** Many options, but need to choose carefully / responsibly.

E.g., if $w_i(j) = \frac{1}{N_C}$ for all (i, j), then we're back to a difference in means.

- **Right Answer** Choose weights $w_i(j)$ that indicate ***how close*** $\mathbf{X}_j$ is to $\mathbf{X}_i$.

Proximity

- Our weights $w_i(j)$ should be a measure of ***how close*** X_j is to X_i.
- If X is **discrete**, then we can consider equality, *i.e.*, $w_i(j) = \mathbb{I}(\mathrm{X}_i = \mathrm{X}_j)$, scaling as necessary to get $\sum_j w_i(j) = 1$.

Proximity

- Our weights $w_i(j)$ should be a measure of ***how close*** X_j is to X_i.
- If X is **continuous**, then we need *proximity* rather than *equality*.

Nearest-neighbor matching chooses the closest control observation using the **Euclidean** distance between X_i and X_j, *i.e.*,

$$\|(X_i - X_j)\| = \sqrt{(X_i - X_j)'(X_i - X_j)} = \sqrt{\Sigma_{n=1}^{k}(X_{ni} - X_{nj})^2}$$

- $\hat{\tau}_i = \mathrm{Y}_{1i} - \mathrm{Y}^i_{0j}$, where Y^i_{0j} is i's nearest neighbor in the control group.
- **Estimator:** $\hat{\tau}_M = \frac{1}{N_T}\sum_i \hat{\tau}_i$
- Produces causal estimates if CIA is valid *and* we have sufficient overlap.

Proximity

- Our weights $w_i(j)$ should be a measure of ***how close*** X_j is to X_i.
- If X is **continuous**, then we need *proximity* rather than *equality*.
- The Euclidean distance is not invariant to changes in the **scale of the X's**. A more commonly used distance is the **normalized Euclidean distance**

$$\|(X_i - X_j)\| = \sqrt{(X_i - X_j)' V_X^{-1} (X_i - X_j)}$$

- where V_X^{-1} is the symmetric and positive semidefinite variance matrix of X of X.
- No scale problem but still no correlations between Xs.

Proximity

Our weights $w_i(j)$ should be a measure of ***how close*** X_j is to X_i.

If X is **continuous**, then we need *proximity* rather than *equality*.

Nearest-neighbor matching with Mahalanobis distance chooses the single closest control using Mahalanobis distance between X_i and X_j, *i.e.*,

$$\|(X_i - X_j)\| = \sqrt{(X_i - X_j)' \Sigma_X^{-1} (X_i - X_j)}$$

where Σ_X^{-1} is the covariance matrix of X.

- **Estimator:** $\hat{\tau}_M = \frac{1}{N_T} \sum_i \hat{\tau}_i$ where $\left(\hat{\tau}_i = \mathrm{Y}_{1i} - \mathrm{Y}_{0j}^{i}\right)$
- Produces causal estimates if CIA is valid *and* we have sufficient overlap.
- No scale problem and taking correlations between Xs into account.

More neighbors?

- Why limit ourselves to a **single** or **some** "best" match?
- If we're going to let a function/algorithm choose the *nearest* match, can't we also let the function/algorithm choose *how many* matches?
- Furthermore, if $N_C \gg N_T$, it we're throwing away *a lot* of information.
- Don't forget we assume that $\sum_j w_i(j) = 1$, then we could use the property of c.d.f to transform the weight in a distribution.

More neighbors!

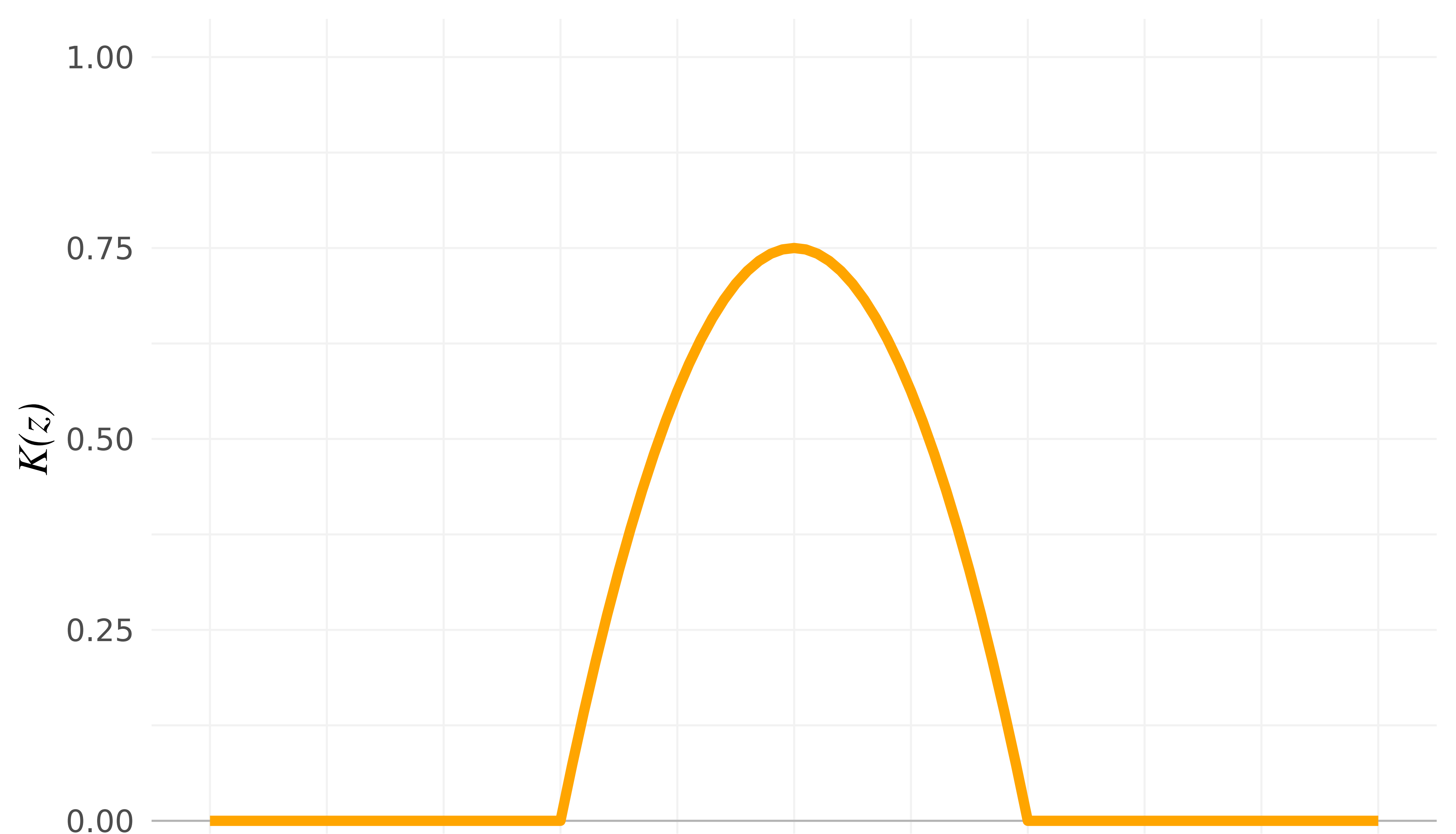
- Kernel matching gives positive weight to all control observations within some **bandwidth** h, with higher weight for closer matches determined by some **kernel function** $K(\cdot)$,

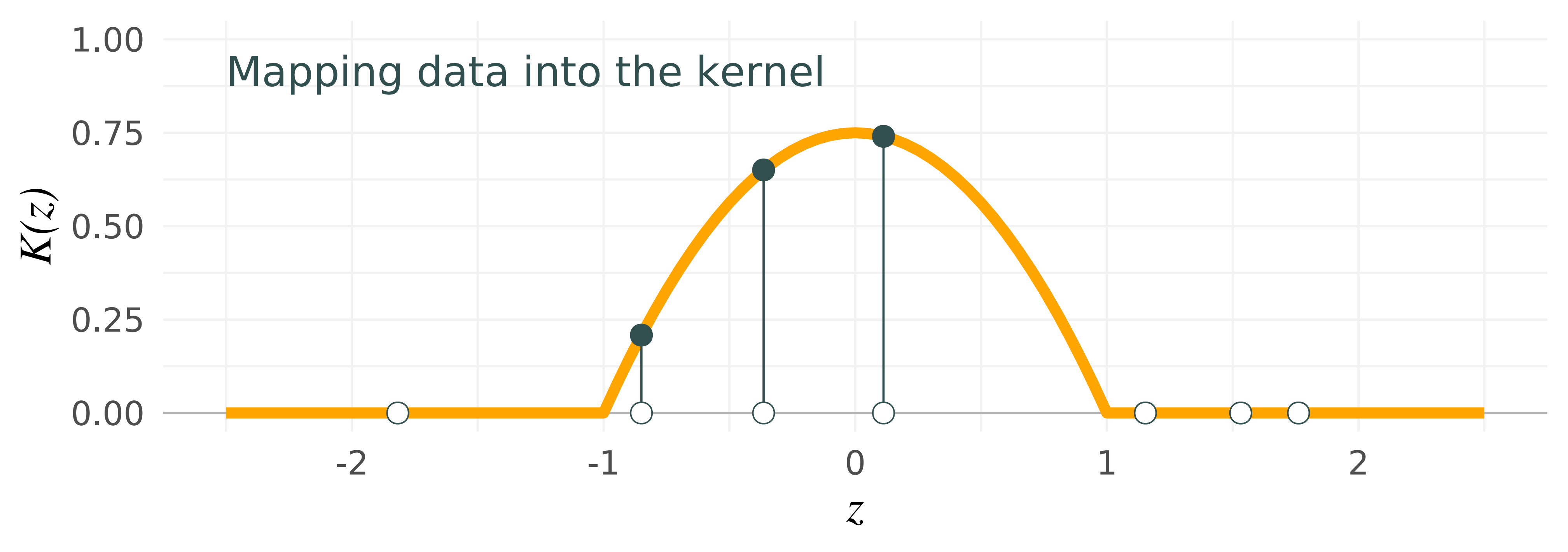
$$w_i(j) = \frac{K\left(\frac{\mathrm{X}_j - \mathrm{X}_i}{h}\right)}{\sum_{j \in (D=0)} K\left(\frac{\mathrm{X}_j - \mathrm{X}_i}{h}\right)}$$

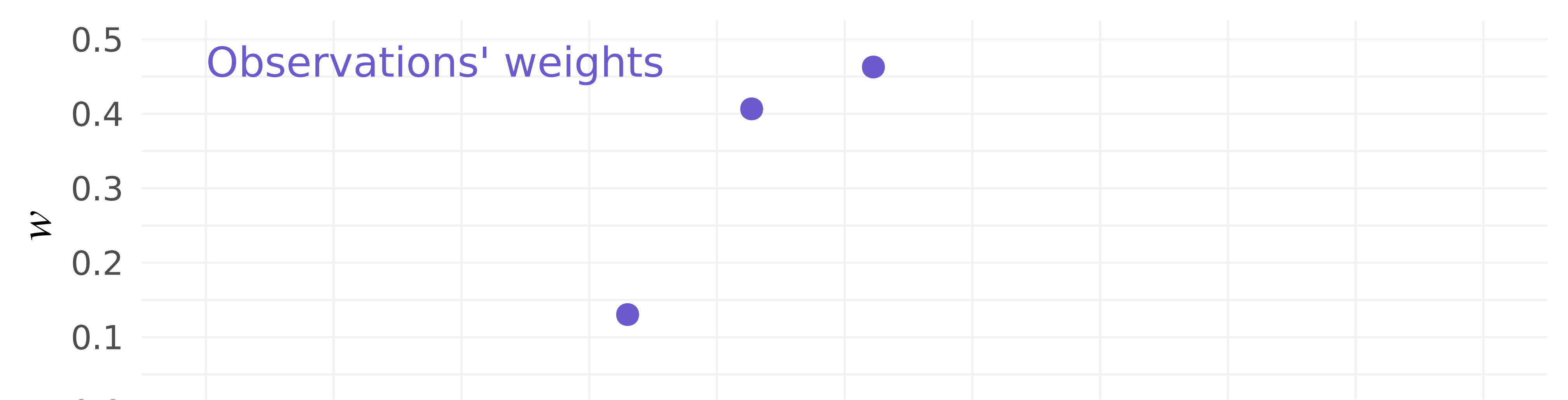
Example The *Epanechnikov kernel* is defined as

$$K(z) = \frac{3}{4}(1 - z^2) \times \mathbb{I}(|z| < 1)$$

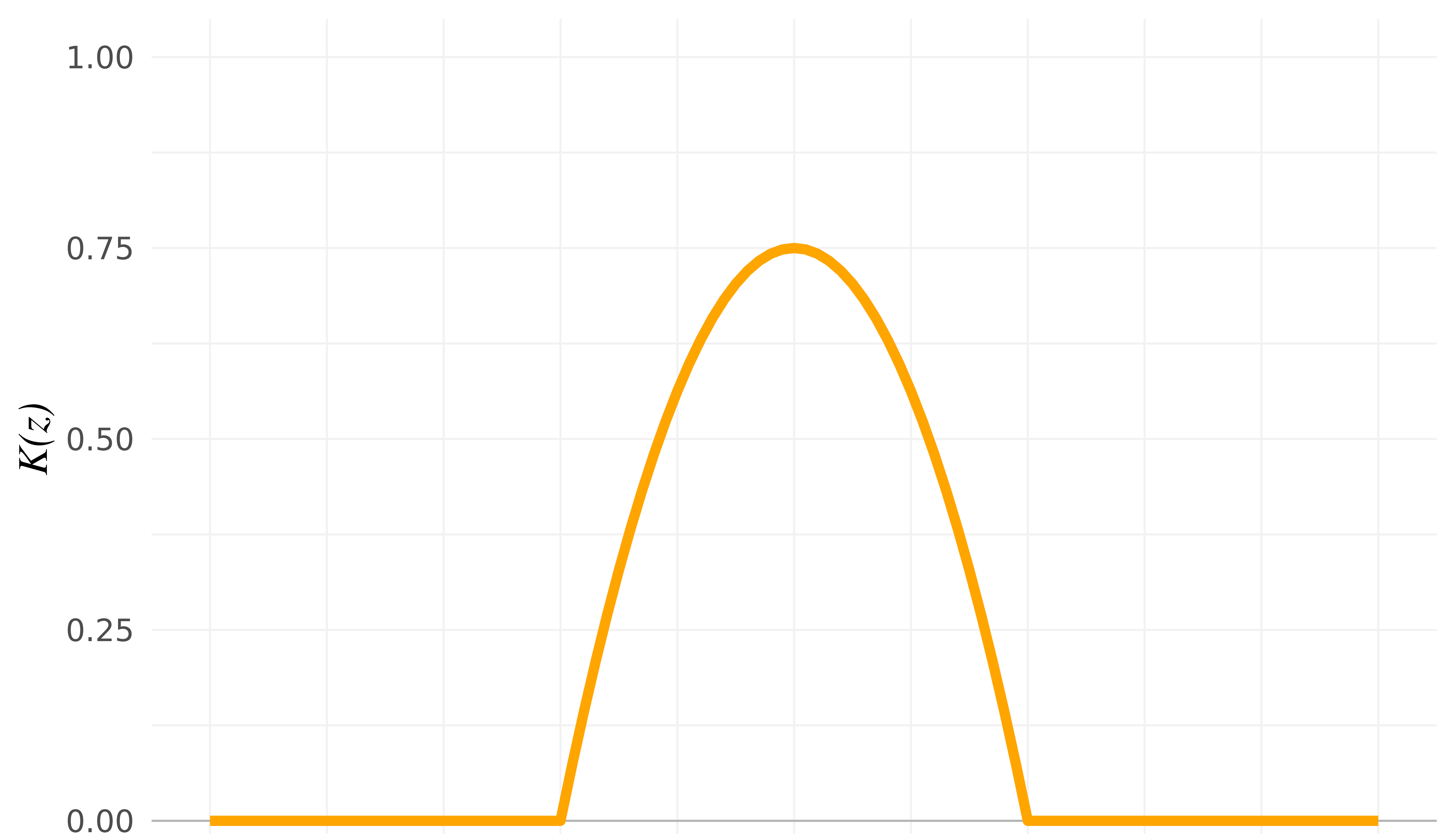
The Epanechnikov kernel $K(z) = \frac{3}{4}\left(1 - z^2\right) \times \mathbb{I}(|z| < 1)$




Mapping data into the kernel
1.00
0.75
0.50
0.25
0.00
$K(z)$
-2
-1
0
1
2
z

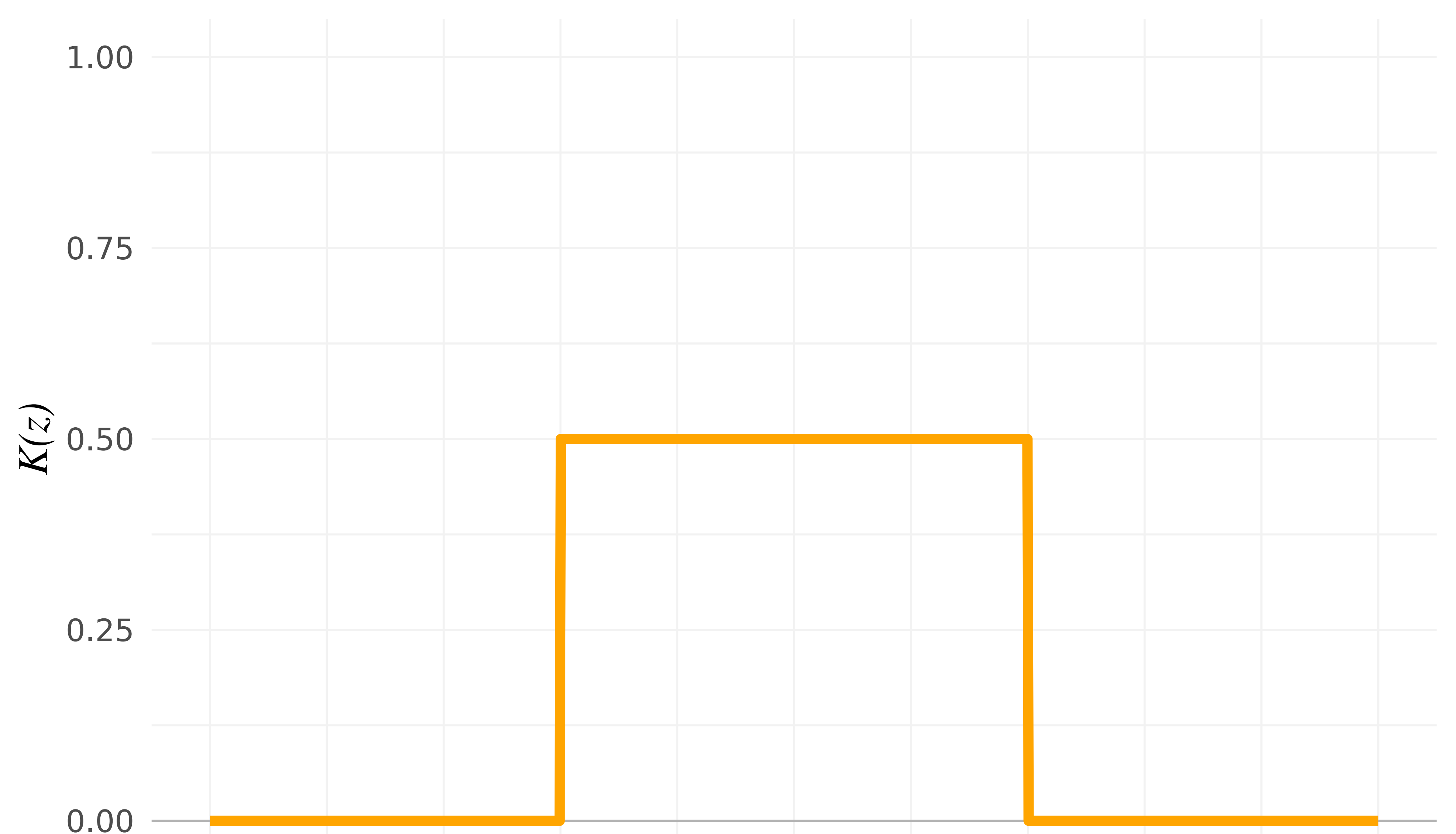

Observations' weights
0.5
0.4
0.3
0.2
0.1
w

The Epanechnikov kernel $K(z) = \frac{3}{4}\left(1 - z^2\right) \times \mathbb{I}(|z| < 1)$

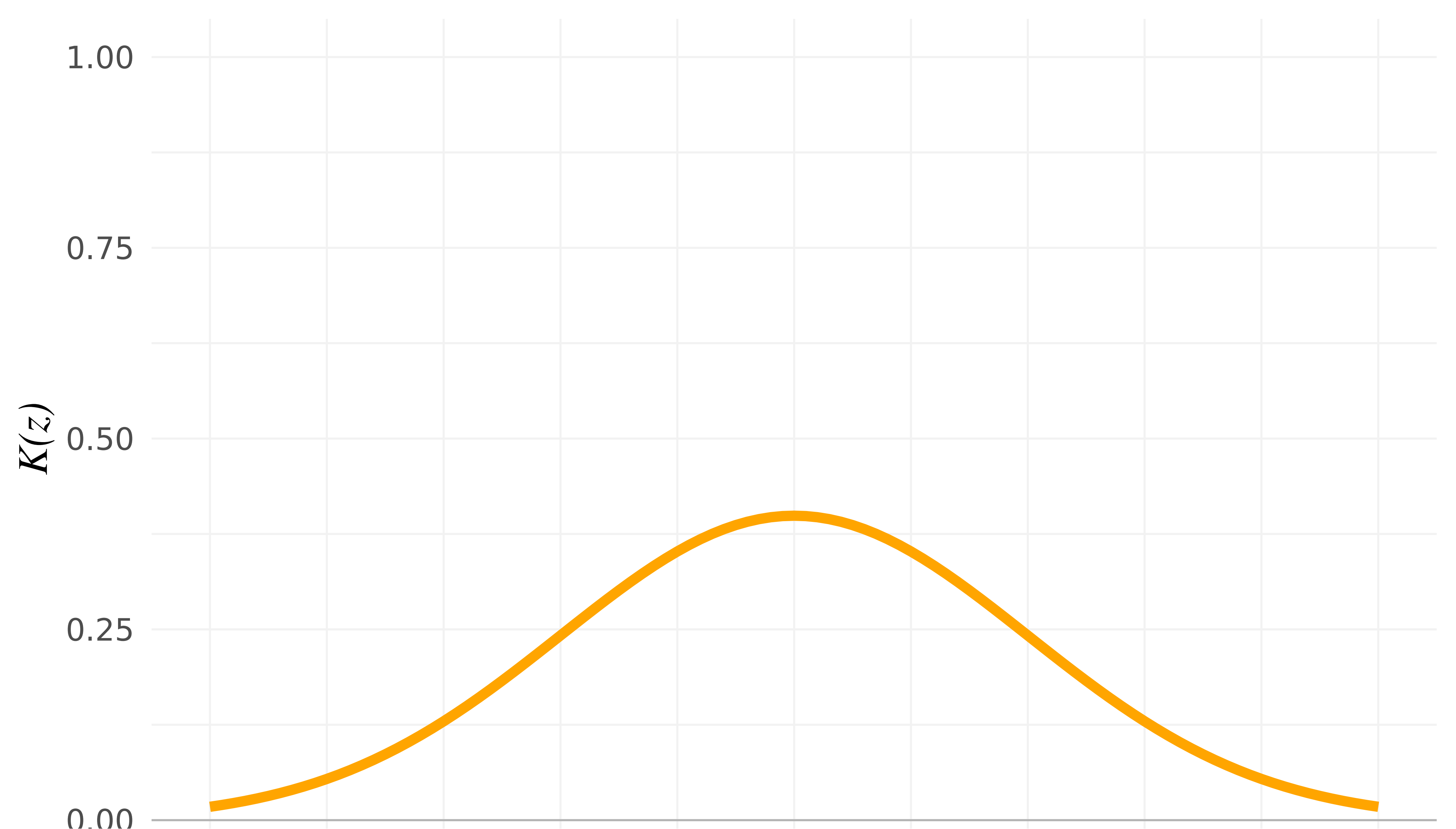


The Triangle kernel $K(z) = (1 - |z|) \times \mathbb{I}(|z| < 1)$

The Uniform kernel $K(z) = \frac{1}{2} \times \mathbb{I}(|z| < 1)$



The Gaussian kernel $K(z) = (2\pi)^{-1/2} \exp\left(-z^2/2\right)$



Propensity-score methods

Propensity-score methods

The magic

- It turns out that if $(\mathrm{Y}_{0i}, \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|\mathrm{X}_i$, then we actually **only** need to match/condition on $p(\mathrm{X}_i) = E[\mathrm{D}_i|\mathrm{X}_i]$.

- $p(\mathrm{X}_i)$ is the **propensity score**,the probability of treatment given X_i.

- **Propensity-score theorem** If $(\mathrm{Y}_{0i}, \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|\mathrm{X}_i$, then $(\mathrm{Y}_{0i}, \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|p(\mathrm{X}_i)$.

- This theorem extends CIA assumption from a multiple dimensions to a one-dimensional score, avoiding the curse of dimensionality.

Propensity-score methods

Theorem If $(\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i | \mathrm{X}_i$, then $(\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i | p(\mathrm{X}_i)$.

Proof

$$\Pr\left[\mathrm{D}_i = 1 \middle| \mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)\right]$$

$$= E\left[\mathrm{D}_i \middle| \mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)\right]$$

$$= E\left[\ E\left(\mathrm{D}_i \middle| \mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i),\ \mathrm{X}_i\right) \middle| \mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)\right]$$

$$= E\left[\ E\left(\mathrm{D}_i \middle| \mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ \mathrm{X}_i\right) \middle| \mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)\right]$$

Propensity-score methods

Theorem If $(\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|\mathrm{X}_i$, then $(\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|p(\mathrm{X}_i)$.

Proof

$$\Pr\left[\mathrm{D}_i = 1\middle|\mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)\right] = \cdots = E\left[E\left(\mathrm{D}_i\middle|\mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ \mathrm{X}_i\right)\middle|\mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)\right]$$

$$= E\left[E\left(\mathrm{D}_i\middle|\mathrm{X}_i\right)\middle|\mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)\right]$$

$$= E\left[p(\mathrm{X}_i)\middle|\mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)\right]$$

$$= p(\mathrm{X}_i)$$

$$\therefore\ (\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|\mathrm{X}_i \implies (\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|p(\mathrm{X}_i) \quad ✔$$

Propensity-score methods

Intuition

Question

- X_i carries way more information than $p(\mathrm{X}_i)$, so how can we still get conditional independence of treatment by only conditioning on $p(\mathrm{X}_i)$?

Answer 1 Conditional independence of treatment isn't about *extracting all of the information* possible from X_i. We actually *only care about creating a situation* in which $\mathrm{D}_i|$something is independent of $(\mathrm{Y}_{0i},\ \mathrm{Y}_{1i})$.

Answer 2 Back to our main concern: **selection bias**. People select into treatment. If X says two people were equally likely to be treated, and if X_i explains all of selection (CIA), then there cannot be selection between these two people.

Propensity-score methods

Estimation

- **Question**:where do propensity scores come from?

- there are a lot of ways to estimate it.

1. Flexible (*i.e.*, interactions) logit specification
2. Kernel regression (remember kernel functions?)
3. Many others—machine learning, series-logit estimator, *etc.*

- The most common way is to use logit regression.

Propensity-score methods

Estimation

From *MHE* (p. 83)

Question

> A big question here is how to best model and estimate $p(X_i)$...

Answer

> The answer to this is inherently application-specific. A growing empirical literature suggests that a logit model for the propensity score with a few polynomial terms in continuous covariates works well in practice...

Propensity-score methods

Major requirements

- still have two **major** requirements for any of these methods to work.

1. Is the **conditional-independence assumption** true?

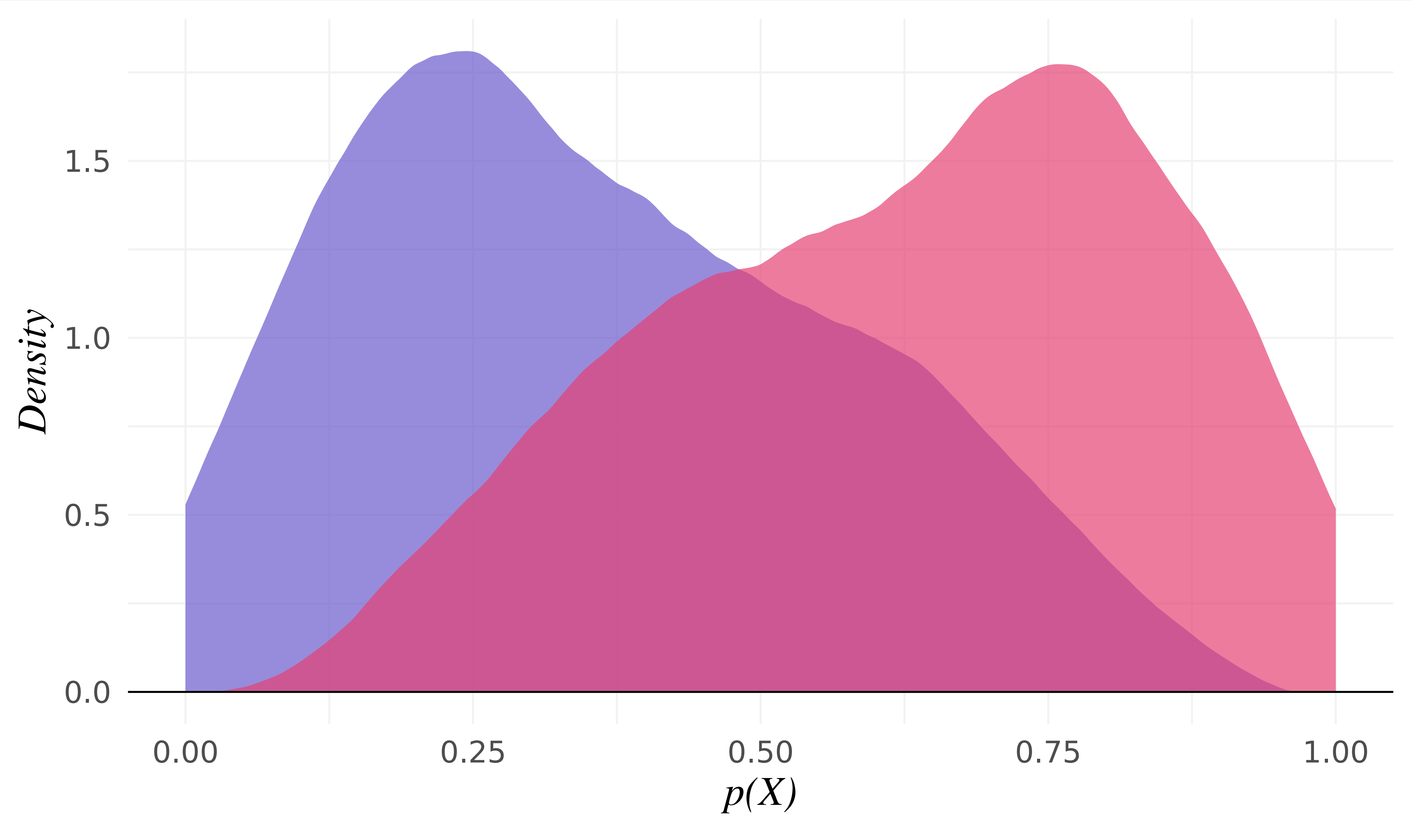
2. Do we have **overlap** between treatment and control units.

We can look for evidence of (**2**) in the data—particularly if we're using propensity-score methods.†

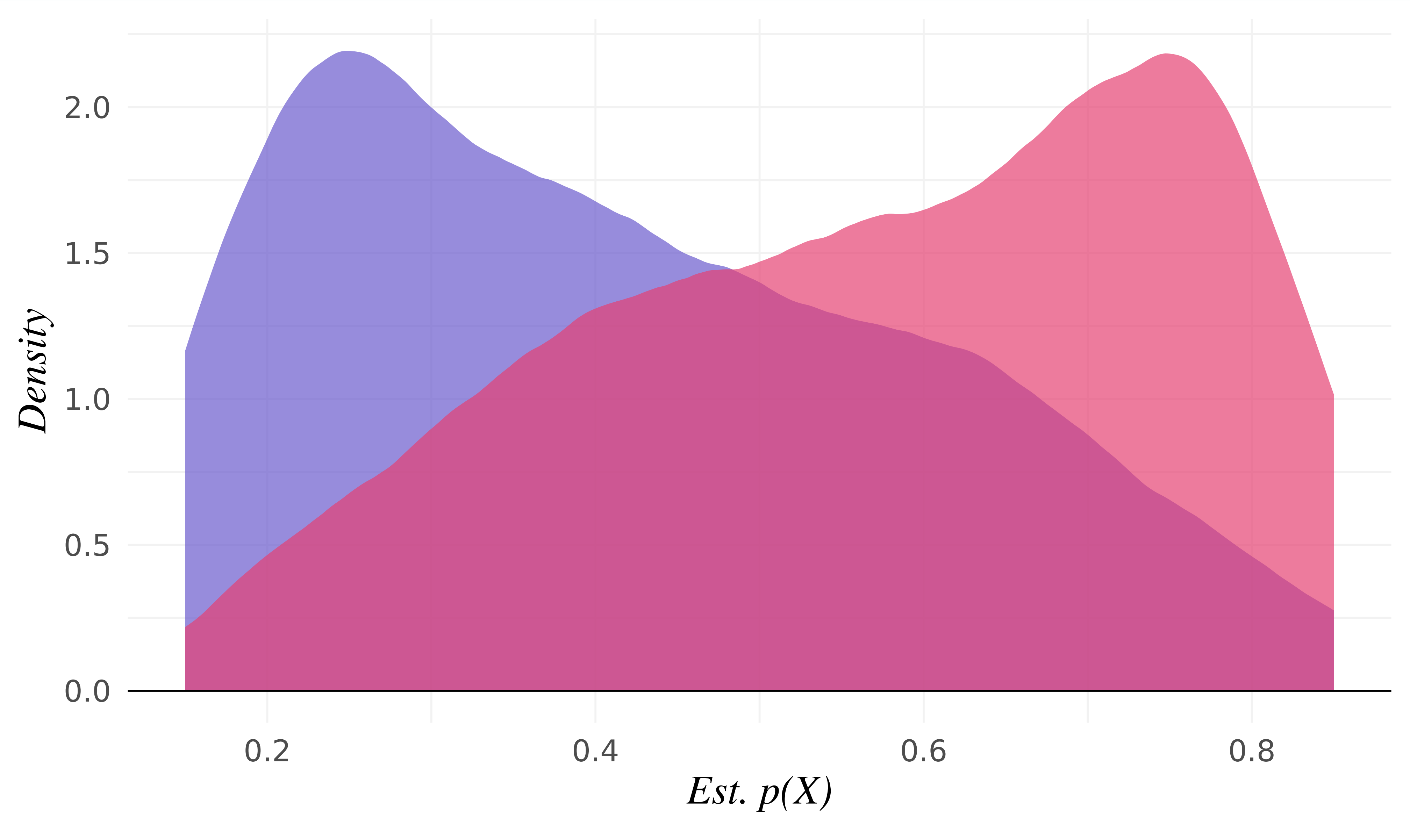
How? Plot the distributions of $p(\mathrm{X}_i)$ for **T** and **C**.

† Checking for overlap in X-space, can be tough as the dimensions of X expand.

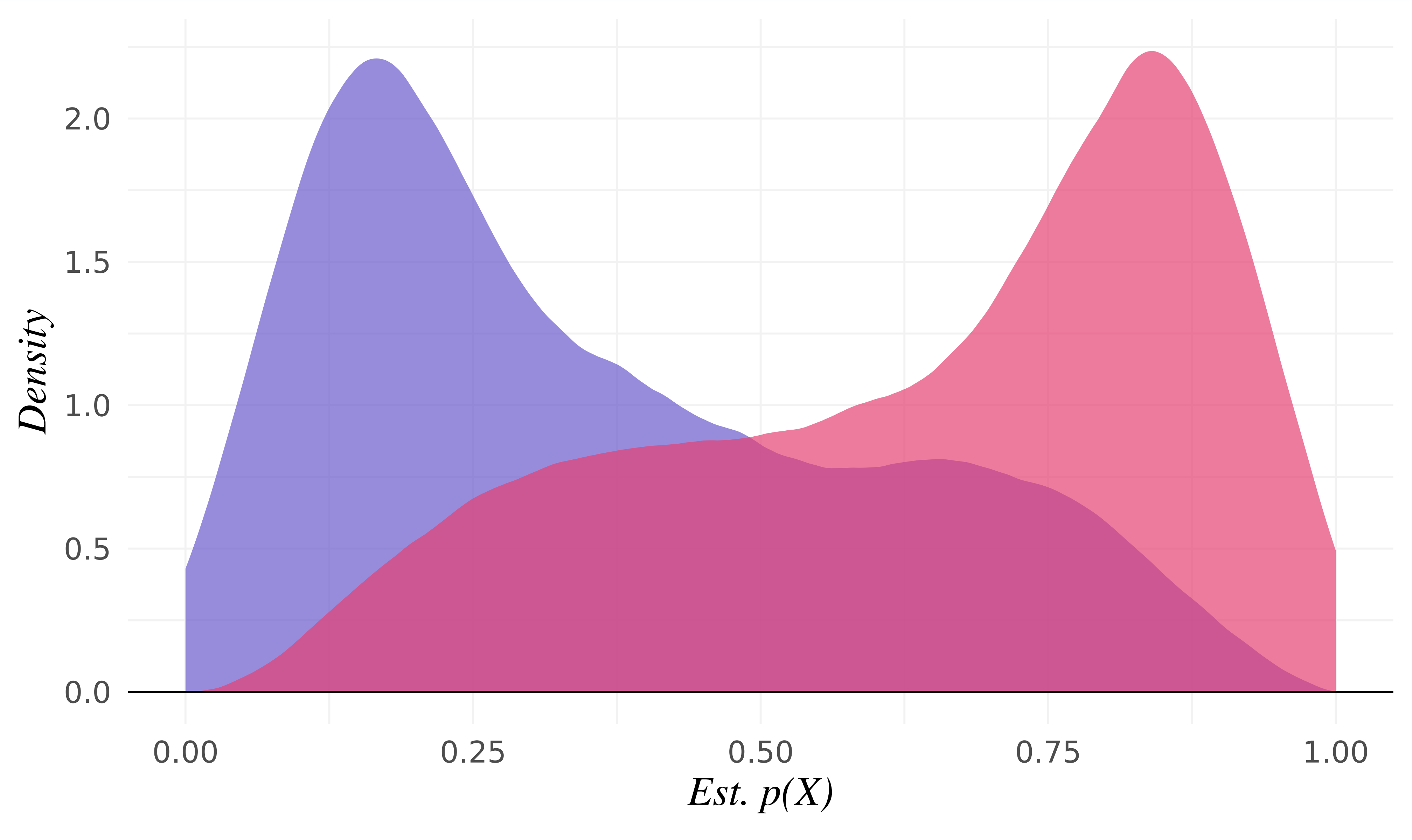
Missing overlap in $p(X_i)$



Authentic (enforced) overlap in $p(X_i)$



Logit-based $\hat{p}(\mathrm{X}_i)$ hiding some of the missing overlap in $p(\mathrm{X}_i)$



Matching in practice

Choosing Variables

- **Question**: Which variables among many available ones should be used to match treatment and control units?

- **Answer**: All variables that you think are likely to be **confounders**.(*Recall ”good and bad controls story”*)

    - all variables that determine both treatment uptake and the outcome.
    - Pre-treatment covariates are the best.
    - Post-treatment variables,especially the outcomes should not be used.

- Similar to OLS regression analysis, different results of including different variable sets can be considered as the **sensitivity analysis**.

Matching in practice

With or Without Replacement

- Matching with replacement means that control units can be used as a match for **more than once**.
    - each control unit is “placed back” into the controls after being used once.
- **Two advantages**:
    - treatment and control units after matching will be better balanced.
    - the order in which we match the units does not matter, in turn the matching algorithm is reduced in complexity.
- Nonetheless, it is very common to match with replacement.

Matching in practice

1:1 Matching v.s 1:m

- 1:1 matching: each treated unit can be matched to only one control.
- 1:m matching: each one can be matched to more than one control.
- **Benefit**: This can be useful in large samples where there are more control units than treated units, because the inclusion of more units will increase the precision of our estimates.
- **Cost**: often the second, third and fourth matches may be poorer than the first match, meaning that we may end up including control units that are not very similar to the treatment

Matching in practice

Assessing Balance

- As in RCTs,after carrying out matching we should first carry out balance tests to compare the treatment and control units.
- If matching was successful, then by definition they should be very similar to each other in terms of their covariates.
- Balance tests are particularly useful in matching because they might be able to help us choose between different distance metrics or matching with vs. without replacement.

Matching in practice

In a Summary

- Choosing the **"best"** matching method highly depends on the unique characteristics of the dataset as well as the goals of the analysis.
  - Similar to the logic of Machine learning
- Therefore, sensitivity analysis is very crucial to Matching Method.

Matching in practice

Matching v.s Regression

- Both matching and regression rely on CIA (selection on observables). Most biases we could suffer in regression, such as OVB, measurement error, and simultaneous causality, will not be avoided even if we use matching.

- Why we still need matching?

    - Due to its non-parametric characteristics, matching does not impose any restrictions on empirical specification or estimate specific parameters of the CEF function.

- Regression does not account for the common support issue.

--

- Using matching alone is **less common** in economics, more frequently combined with other methods like DID and SCM.