

# Matching Methods

Prof. Tzu-Ting Yang  
楊子霆

Institute of Economics, Academia Sinica  
中央研究院經濟研究所

September 10, 2024

# Observational Studies

- Randomized Controlled Trial (RCT) is called the “gold standard” for causal inference
  - In a RCT, researcher can assign treatments randomly to the individuals
  - Therefore, **treatment status is unrelated to any observed and unobserved confounders**
    - Treatment and control group should be similar in all characteristics
    - Thus, we can use the observed outcomes of control group to approximate the counterfactual outcomes of treatment group

## Observational Studies

- But implementing a randomized experiment in social science is very expensive and sometimes has ethical issues
  - In social science, many empirical studies use **non-experimental data**
    - It means researchers **can NOT assign treatment**
  - We call this type of empirical researches as **observational studies**

# Observational Studies

- We want to **design** observational studies that **approximate experiments**:
  - “The planner of an observational study should always ask himself: How would the study be conducted if it were possible to do it by controlled experimentation” (Cochran 1965)
- How to do that?
  - Need to directly control for the observed variables
  - Use indirect methods to adjust for unobserved variables
  - Make “other thing equal” in observed and unobserved variables

## Matching Methods: Main Idea

# Main Idea of Matching

重要假設

- Assume all confounding factors are **observable** to researchers
- Matching is a way to eliminate selection bias
  - By constructing a control group with the same observable characteristics as the treatment group
- This can be accomplished by **matching** treated and untreated units with the same observable characteristics.

# Main Idea of Matching

- **Example:**

- We want to estimate the causal effect of job training program on worker's earnings
- Suppose **age** is the only confounding factors that affect both earnings and job training decision

# Matching: A Numerical Example

Trainees			Non-Trainees		
unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900
2	34	10200	2	50	31000
3	29	14400	3	30	21000
4	25	20800	4	27	9300
5	29	6100	5	54	41100
6	23	28600	6	48	29800
7	33	21900	7	39	42000
8	27	28800	8	28	8800
9	31	20300	9	24	25500
10	26	28100	10	33	15500
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
			20	23	9500
			21	32	25900
Avg:		28.5	Avg:		33
		16426			20724

# Matching: A Numerical Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	✓28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

同齡匹配

# Matching: A Numerical 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			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
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			
		20	23	9500				
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
10	26	28100	10	33	15500			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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	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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
19	28	16300	19	32	17800			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical 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			
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical Example

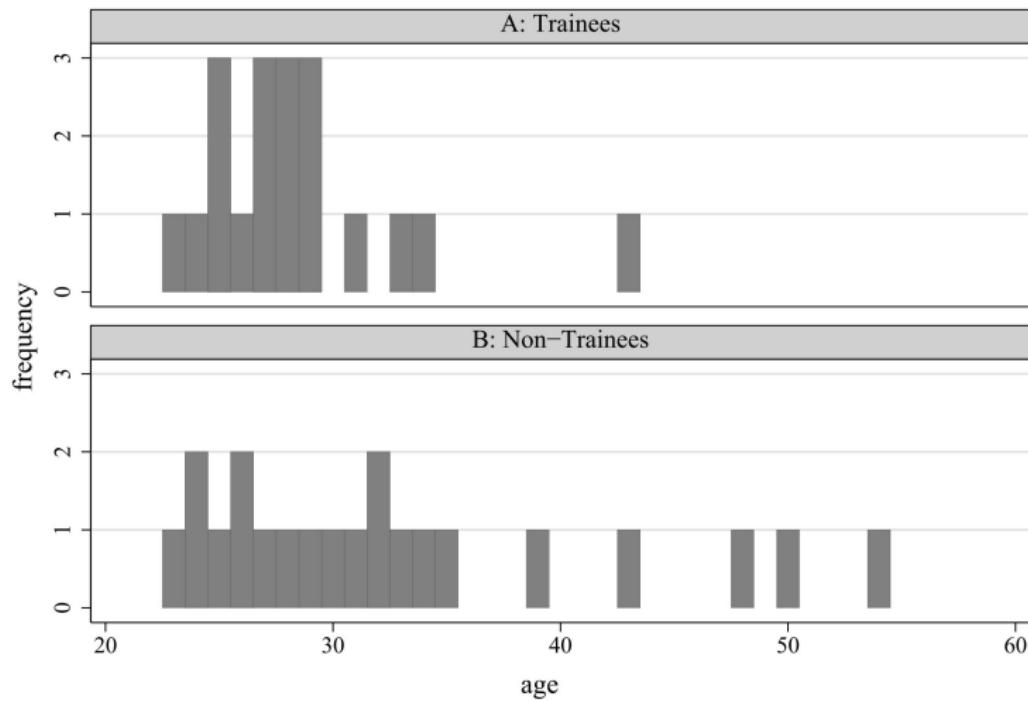
Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	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
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:		

# Matching: A Numerical Example

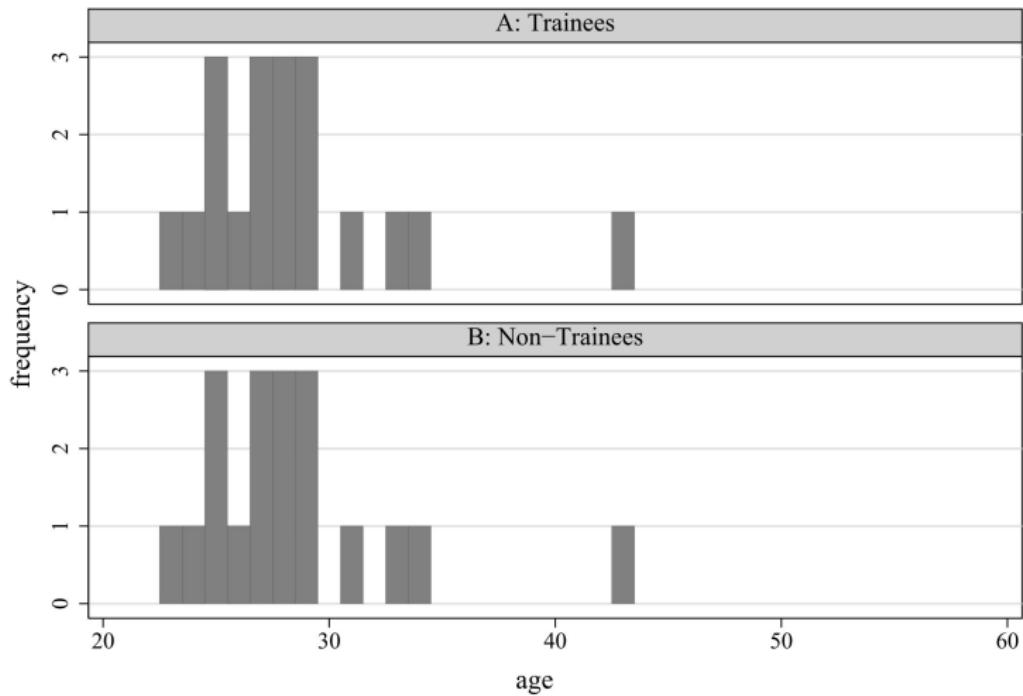
全部配對完

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
			20	23	9500			
			21	32	25900			
Avg:	28.5	16426	Avg:	33	20724	Avg:	28.5	13982

# Age Distribution: Before Matching



# Age Distribution: After Matching



# Treatment Effect Estimates

Difference in average earnings between trainees and non-trainees:

- Before matching:

$$16426 - 20724 = \underline{-4298}$$

- After matching:

$$16426 - 13982 = \underline{2444}$$

加入職業訓練有效果

# Matching Methods: Potential Outcomes Framework

# Conditional Independence Assumption (CIA)

## Conditional Independence Assumption

$$(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | X_i$$

- $X_i$  are observable characteristics (covariates) with value of  $k$
- CIA asserts that conditional on observable characteristics  $X_i$ , potential outcomes are independent of treatment assigned
  - This assumption is also called **selection on observable**

# Conditional Independence Assumption (CIA)

## Conditional Independence Assumption

$$(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | X_i$$

- This implies  $X_i$  are key factors that determine the value of potential outcome for each individual
  - $Y_i^0 = g(X_i)$
  - $Y_i^1 = g(X_i)$
- Thus, people with the same value of covariates  $X_i$ , they should have similar value of potential outcomes
  - $E[Y_i^0 | X_i = k, D_i = 1] = E[Y_i^0 | X_i = k, D_i = 0]$
  - $E[Y_i^1 | X_i = k, D_i = 1] = E[Y_i^1 | X_i = k, D_i = 0]$

對照組

對照組

# Conditional Independence Assumption (CIA)

- Job training example:
  - $D_i$ : join job training
  - $Y_i^1$ : potential earnings for joining job training
  - $Y_i^0$ : potential earnings for not joining job training
  - $X_i$ : age, education...etc.

# Conditional Independence Assumption (CIA)

- Suppose only **age** can affect an individual's potential earnings
  - CIA suggests for those with the same age, their training decision are unrelated to the potential earnings
- That is, once controlling age, treatment and control group should be comparable (apple-to-apple comparison)
  - $E[Y_i^0 | X_i = 40, D_i = 1] = E[Y_i^0 | X_i = 40, D_i = 0]$
  - $E[Y_i^1 | X_i = 40, D_i = 1] = E[Y_i^1 | X_i = 40, D_i = 0]$

# Common Support Assumption

## Common Support Assumption

$$0 < \Pr(D_i = 1|X_i) < 1$$

- For each value of covariates  $X_i$ , there is a positive probability of being both treated and untreated
- In other words, it is NOT possible to perfectly predict one's treatment status by using specific value of  $X_i$ 
  - For example, this exclude:

需要有  
配對

- All individuals with age 40 are in treatment group:  
 $\Pr(D_i = 1|X_i = 40) = 1$
- All individuals with age 40 are in control group:  
 $\Pr(D_i = 1|X_i = 40) = 0$
- It ensures that there is sufficient overlap in the characteristics of treated and untreated units to find adequate matched sample

# Identification Results for Matching

*same  $X_i$*

$$\begin{aligned} & \underbrace{\mathbb{E}[Y_i | X_i, D_i = 1] - \mathbb{E}[Y_i | X_i, D_i = 0]}_{\text{ODO at given } X_i} \\ &= \mathbb{E}[Y_i^1 | X_i, D_i = 1] - \mathbb{E}[Y_i^0 | X_i, D_i = 0] \\ &= \mathbb{E}[Y_i^1 | X_i, D_i = 1] - \mathbb{E}[Y_i^0 | X_i, D_i = 1] \\ &\quad + \mathbb{E}[Y_i^0 | X_i, D_i = 1] - \mathbb{E}[Y_i^0 | X_i, D_i = 0] \\ &= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 | X_i, D_i = 1]}_{\text{CATT}} \\ &\quad + \underbrace{\mathbb{E}[Y_i^0 | X_i, D_i = 1] - \mathbb{E}[Y_i^0 | X_i, D_i = 0]}_{\text{Selection Bias}} \\ &= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 | X_i, D_i = 1]}_{\text{CATT}} + \underbrace{0}_{\text{Selection Bias}} \end{aligned}$$

- Remember CIA ensures  $\mathbb{E}[Y_i^0 | X_i, D_i = 1] = \mathbb{E}[Y_i^0 | X_i, D_i = 0]$

# Identification Results for Matching

$$\underbrace{\mathbb{E}[Y_i|X_i, D_i = 1] - \mathbb{E}[Y_i|X_i, D_i = 0]}_{\text{ODO at given } X_i} = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0|X_i, D_i = 1]}_{\text{CATT}} + \underbrace{0}_{\text{Selection Bias}}$$
$$= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0|X_i, D_i = 0]}_{\text{CATU}}$$
$$= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0|X_i]}_{\text{CATE}}$$

# Identification Results for Matching

- Under CIA, matching estimator represent **conditional average treatment effect (CATE)**
  - Causal effect for the people with specific value of  $X_i$ 
    - Causal effect of job training on annual earnings for people with age 40
    - Causal effect of job training on annual earnings for treated people with age 40
- How to obtain ATT, ATU, and ATE?
  - Take average of CATT, CATU, or CATE over all subgroups (all possible X-values)

# Review: The Law of Iterated Expectations (LIE)

## The Law of Iterated Expectations (LIE)

$$E[Y_i] = E[E[Y_i|X_i]]$$

- Intuitively, there are two ways to compute average outcome  $Y$  (e.g. earnings)

1  $E[Y_i] = \sum_{Y_i} Y_i \times Pr(Y_i = y)$

- Each value of outcome times its probability

2  $E[E[Y_i|X_i]] = \sum_{X_i} E[Y_i|X_i] \times Pr(X_i = x) = E[Y_i]$

- Average outcome of subgroup times the share of subgroup

# Review: The Law of Iterated Expectations (LIE)

## Example

- Suppose we want to compute average income in Taiwan
  - 1 We know 40% of people in Taiwan earn 1 million NTD per year, 40% earn 2 million NTD, and 20% earn 3 million NTD

$$\begin{aligned} E[Y_i] &= \sum_{Y_i} Y_i \times Pr(Y_i = y) \\ &= 1 \times 0.4 + 2 \times 0.4 + 3 \times 0.2 \\ &= 1.8 \end{aligned}$$

# Review: The Law of Iterated Expectations (LIE)

## Example

- 2 Suppose we know the average annual income for male is 2 million and average annual income for female is 1.6 million
  - The share of male to total population is 50%

$$\begin{aligned} E[E[Y_i|X_i]] &= \sum_{X_i} E[Y_i|X_i] \times Pr(X_i = x) \\ &= 2 \times 0.5 + 1.6 \times 0.5 \\ &= 1.8 \\ &= E[Y_i] \end{aligned}$$

# Identification Results for Matching: ATT

- Using a matching method, we can also identify CATT

$$\underbrace{E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0]}_{\text{ODO at given } X_i} = \underbrace{E[Y_i^1 - Y_i^0|X_i, D_i = 1]}_{\text{CATT}}$$

- Applying LIE, we can identify ATT by averaging all of the X-specific effects in treatment group (CATT):

$$\underbrace{E[E[Y_i^1 - Y_i^0|X_i, D_i = 1|D_i = 1]]}_{\text{CATT}} = \underbrace{E[Y_i^1 - Y_i^0]|D_i = 1}_{\text{ATT}}$$

# Identification Results for Matching: ATT

## A Numerical 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
			20	23	9500			
			21	32	25900			
Avg: 28.5 16426			Avg: 33 20724			Avg: 28.5 13982		

# Identification Results for Matching: ATT

## A Numerical 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
			20	23	9500			
			21	32	25900			
Avg: 28.5 16426			Avg: 33 20724			Avg: 28.5 13982		

# Identification Results for Matching: ATT

## A Numerical Example

- Using matching methods, we get CATT for all possible ages

$$\begin{aligned} \mathbb{E}[Y_i^1 - Y_i^0 | X_i = 28, D_i = 1] &= \frac{(17700 - 8800) + (11500 - 8800) + (16300 - 8800)}{3} \\ &= 6366.67 \end{aligned}$$

⋮

$$\begin{aligned} \mathbb{E}[Y_i^1 - Y_i^0 | X_i = 34, D_i = 1] &= \frac{10200 - 24200}{1} \\ &= 14,000 \end{aligned}$$

# Identification Results for Matching: ATT

## A Numerical Example

- Applying LIE, we can identify ATT by averaging CATT over all possible ages:

$$\begin{aligned} & \underbrace{\mathbb{E}[\mathbb{E}[Y_i^1 - Y_i^0 | X_i, D_i = 1]]}_{\text{CATT}} \\ &= \mathbb{E}[Y_i^1 - Y_i^0 | X_i = 28, D_i = 1] \times \frac{3}{19} + \dots \\ &+ \mathbb{E}[Y_i^1 - Y_i^0 | X_i = 34, D_i = 1] \times \frac{1}{19} \\ &= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 | D_i = 1]}_{\text{ATT}} \quad \text{加權平均} \end{aligned}$$

# Identification Results for Matching: ATU

- Note that using a matching method, we can also identify CATU

$$\underbrace{E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0]}_{\text{ODO at given } X_i} = \underbrace{E[Y_i^1 - Y_i^0|X_i, D_i = 0]}_{\text{CATU}}$$

- Applying LIE, we can identify ATU by averaging all of the  $X$ -specific effects in treatment group (CATU):

$$\underbrace{E[E[Y_i^1 - Y_i^0|X_i, D_i = 0|D_i = 0]]}_{\text{CATU}} = \underbrace{E[Y_i^1 - Y_i^0]|D_i = 0}_{\text{ATU}}$$

# Identification Results for Matching: ATE

- Using a matching method, we can identify CATE

$$\underbrace{E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0]}_{\text{ODO at given } X_i} = \underbrace{E[Y_i^1 - Y_i^0|X_i]}_{\text{CATE}}$$

- Applying LIE, we can identify ATE by averaging all of the  $X$ -specific effects (CATE):

$$\underbrace{E[E[Y_i^1 - Y_i^0|X_i]]}_{\text{CATE}} = \underbrace{E[Y_i^1 - Y_i^0]}_{\text{ATE}}$$

## Matching Methods: Estimation

# Matching Estimator

- Again, we usually only have sample
  - Part of population data
- Suppose our sample is  $N$  individuals
- Treatment is job training and outcome is earning
  - $N_1$  individuals choose to join job training: treatment group
  - $N_0$  individuals choose not join it ( $N_0 = N - N_1$ ): control group

# Matching Estimator

## Estimation for ATT

- Suppose we want to estimate ATT
  - Average treatment effect for treatment group
- In that case, a matching estimator of  $\alpha_{ATT}$  can be constructed as:

$$\hat{\alpha}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

代替  $Y_i$   
(看不到)

- We want to match **treated** individual  $i$ 's outcome  $Y_i$ 
  - We impute  $Y_i^0$  using untreated units  $Y_{j(i)}$  in control group
  - $Y_{j(i)}$ : the outcome of an untreated observation  $j$  such that  $X_{j(i)}$  is the same or closest value to  $X_i$  among the untreated observations.

# Matching Estimator

## Estimation for ATT

- We can also use the average:

$$\hat{\alpha}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left\{ Y_i - \left( \frac{1}{M} \sum_{m=1}^M Y_{jm(i)} \right) \right\}$$

- Works well when we can find good matches for each treated unit, so  $M$  is usually small (typically,  $M = 1$  or  $M = 2$ )
- Perfect matches are often not available

# Matching Estimator

## Estimation for ATU

- Suppose we want to estimate ATU
  - Average treatment effect for control group
- In that case, a matching estimator of  $\alpha_{ATU}$  can be constructed as:

代替  $\gamma$  (看到)

$$\hat{\alpha}_{ATU} = \frac{1}{N_0} \sum_{D_i=0} (Y_{j(i)} - Y_i)$$

- We want to match **untreated** individual  $i$ 's outcome  $Y_i$ 
  - We impute  $Y_i^1$  using treated units  $Y_{j(i)}$  in treatment group
  - $Y_{j(i)}$ : the outcome of an treated observation  $j$  such that  $X_{j(i)}$  is the same or closest value to  $X_i$  among the treated observations.

# Matching

## Estimation for ATE

- We can also use matching to estimate ATE
  - Average treatment effect for both treatment and control groups
- In that case, we match in both directions:
  - If observation  $i$  is treated, we impute  $Y_i^0$  using untreated units  $Y_{j(i)}$  in control group
  - If observation  $i$  is untreated, we impute  $Y_i^1$  using treated units  $Y_{j(i)}$  in treatment group
- The matching estimator for ATE is:

做兩次

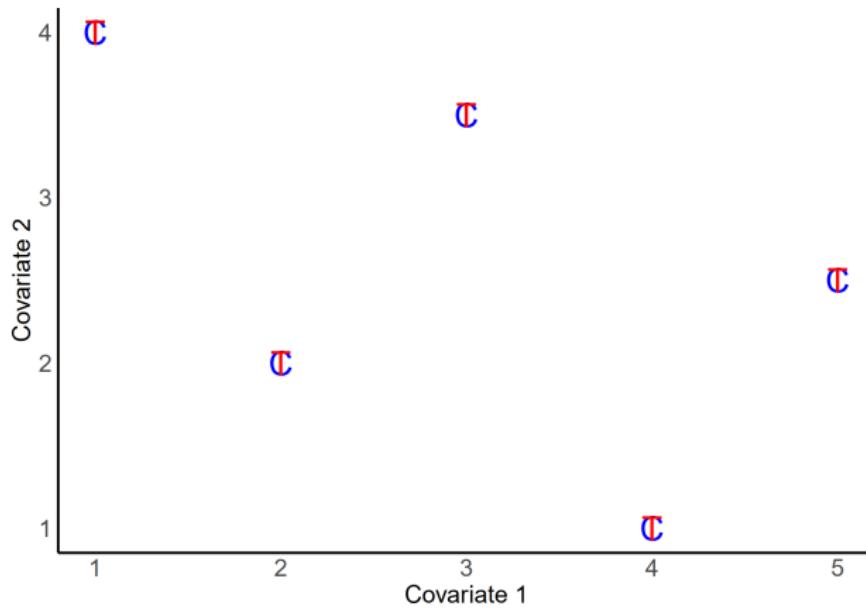
$$\hat{\alpha}_{ATE} = \frac{1}{N} \left\{ \underbrace{\sum_{D_i=1} (Y_i - Y_{j(i)})}_{\text{做兩次}} + \underbrace{\sum_{D_i=0} (Y_{j(i)} - Y_i)}_{\text{做兩次}} \right\}$$

# Exact Matching

- Match each treated unit to a control unit that has exactly the same covariate values  
*配對樣本需完全一樣*
- This is called **exact matching** and can be thought of as the gold standard for matching

# Exact Matching

## A Numerical Example



Source: Ben Elsner's slides

# Exact Matching and the “Curse of Dimensionality”

- Matching becomes unfeasible with many covariates
- This is also true even if we divided each of covariates into coarse categories (subclassification)

# Exact Matching and the “Curse of Dimensionality”

- Assume we have  $k$  covariates and divided each of them into 3 coarse categories
  - age could be “young”, “middle age” or “old”
  - income could be “low”, “medium” or “high”
- The number of subclassification cells is  $3^k$ .
  - For  $k = 10$ , we obtain  $3^{10} = 59049$
- Many cells may contain only treated or untreated observations
  - We may not be able to construct matched sample
  - Violate common support assumption

# Approximate Matching

- In most cases, we just match similar units
- This is called **approximate matching**
- There are three main methods for approximate matching:
  - 1 **Distance Matching:** minimize distance in covariates  $X$
  - 2 **Coarsened Exact Matching:** match within coarsened subgroups
  - 3 **Propensity Score Matching:** match on likelihood of being treated

# Distance Matching

## Measure Closeness

- We usually use more than one characteristics to construct a matched sample
- When the vector of matching covariates has more than one variables ( $k > 1$ )

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{pmatrix}$$

- We need to define a **distance metric** to measure “closeness” to construct a matched sample

# Distance Matching

Measure Closeness

- The usual **Euclidean distance** is:

$$\begin{aligned} \|X_i - X_j\| &= \sqrt{(X_i - X_j)'(X_i - X_j)} \\ &= \sqrt{\sum_{k=1}^K (X_{ki} - X_{kj})^2}. \end{aligned}$$

- Sum up the differences between treatment group and control group over  $k$  characteristics
  - **Drawback:** The Euclidean distance is NOT invariant to changes in the scale of the  $X$ 's
  - For this reason, we often use alternative distances that are invariant to changes in scale

# Distance Matching

Measure Closeness

- A commonly used distance is the **normalized Euclidean distance**

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

where

$$\hat{V} = \begin{pmatrix} \hat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\sigma}_K^2 \end{pmatrix}.$$

- $\hat{\sigma}_k^2$  is the variance of variable  $k$

# Distance Matching

Measure Closeness

- Notice that, the normalized Euclidean distance is equal to:

$$\|X_i - X_j\| = \sqrt{\sum_{k=1}^K \frac{(X_{ki} - X_{kj})^2}{\hat{\sigma}_k^2}}.$$

- ⇒ Changes in the scale of  $X_{ki}$  affect also  $\hat{\sigma}_k$ , and the normalized Euclidean distance does not change

# Distance Matching

## Criteria for a Good Match

- k-Nearest-neighbor Matching

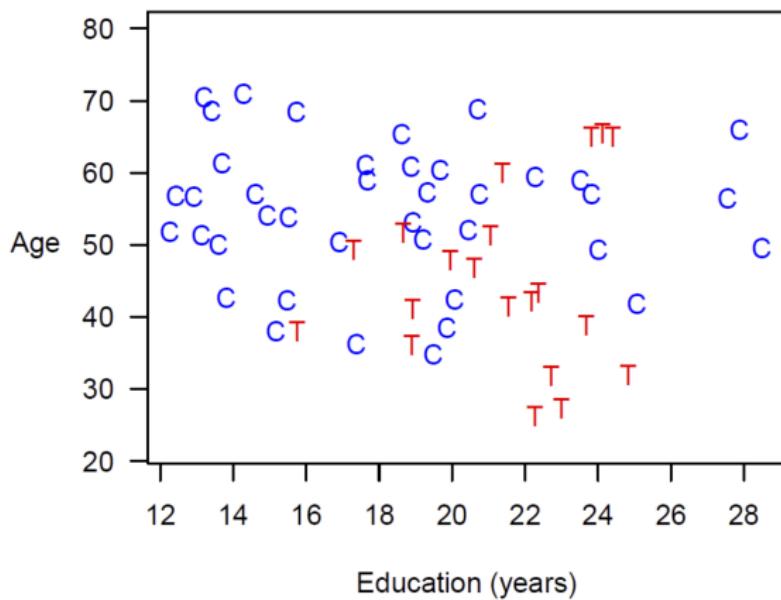
- Match with the nearest neighbor or the  $k$  nearest neighbors in terms of normalized Euclidean distance
- Drop the unmatched units

- Radius Matching

- Match with all control units within a certain radius of the treated unit

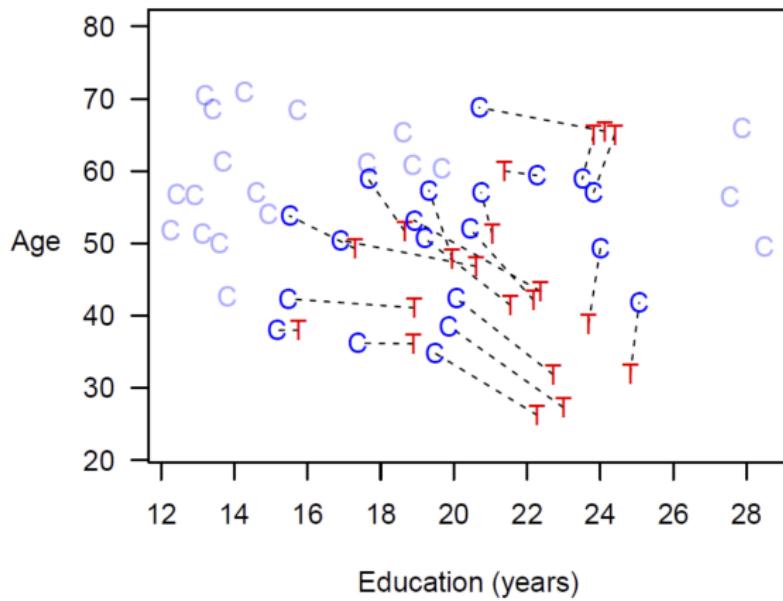
# Distance Matching

k-Nearest-neighbor Matching: K=1



# Distance Matching

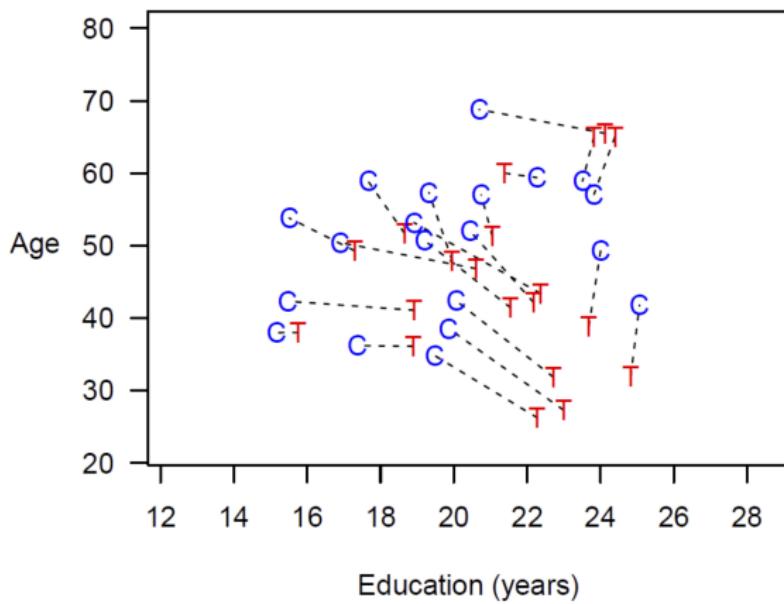
k-Nearest-neighbor Matching: K=1



Source: Ben Elsner's slides

# Distance Matching

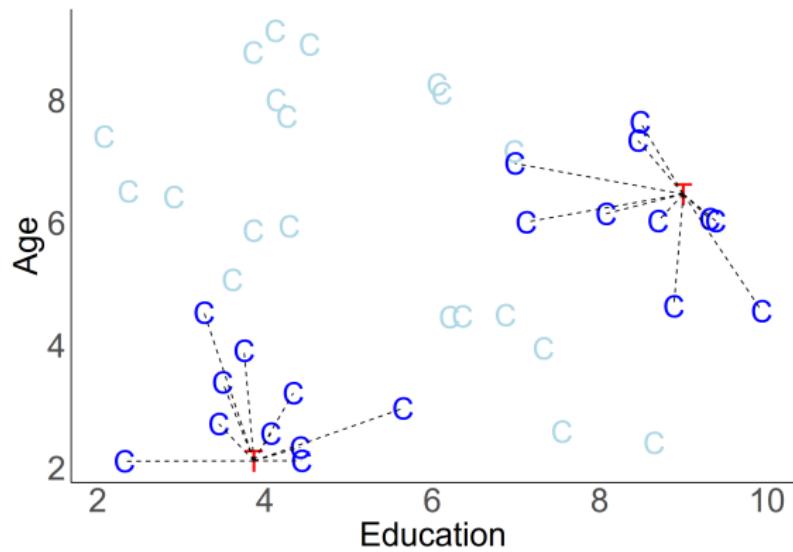
k-Nearest-neighbor Matching: K=1



Source: Ben Elsner's slides

# Distance Matching

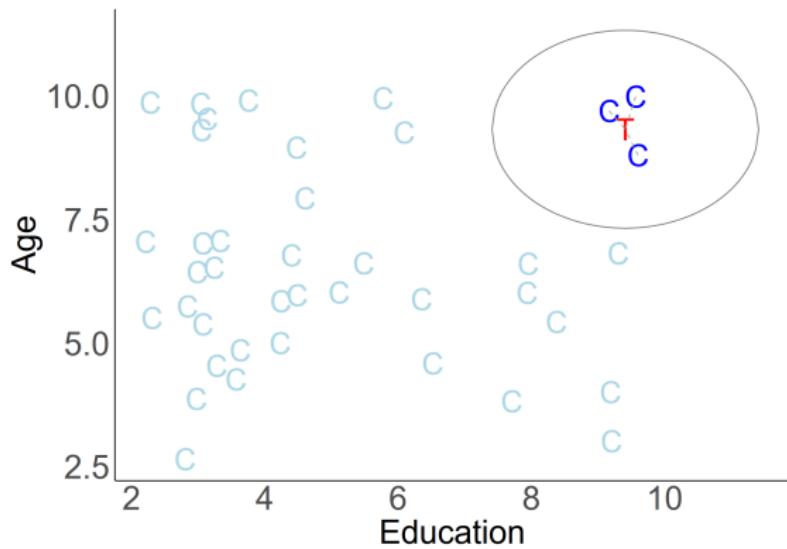
k-Nearest-neighbor Matching: K=10



Source: Ben Elsner's slides

# Distance Matching

## Radius Matching



Source: Ben Elsner's slides

# Matching and the “Curse of Dimensionality”

- Matching discrepancies  $\|X_i - X_{j(i)}\|$  tend to increase with  $k$ , the dimension of  $X$
- It is difficult to find good matches in large dimensions: you need many observations if  $k$  is large

## Propensity Score Matching: Main Idea

# Propensity Score Matching

h

- Instead of matching over ~~10~~ dimensions, the method of **propensity score matching (PSM)** allows the matching problem to be reduced to a single dimension
  - The **propensity score** is defined as the treatment probability conditional on a set of observed variables  $X_i$ :
$$p(X_i) = E[D_i|X_i] = \Pr(D_i = 1|X_i)$$
  - Intuitively, propensity score  $p(X_i)$  summarized all information of a set of covariates  $X_i$  into a single value
  - Then, we can just control (match)  $p(X_i)$  to eliminate selection bias

# Propensity Score Matching

- Rosenbaum and Rubin (1983) proved that CIA (selection on observables) implies:

$$(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | p(X_i)$$

- Conditioning on the propensity score  $p(X_i)$  is enough to make treatment status be independent of the potential outcomes
- Substantial dimension reduction in the matching variables!

# Propensity Score Matching

## Propensity Score Theorem

Suppose the CIA holds, such that  $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | X_i$ . Then  
 $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | p(X_i)$

- If potential outcomes are independent of treatment status conditional on a set of covariates  $X_i$
- Then, potential outcomes are independent of treatment status  $D_i$  conditional on the propensity score  $p(X_i)$

# Propensity Score Matching

- Goal of Proof:

- Assume that  $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | X_i$ . Then:
  - $\Rightarrow Pr(D_i = 1 | Y_i^1, Y_i^0, p(X_i)) = p(X_i) = Pr(D_i = 1 | p(X_i))$
  - $\Rightarrow (Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | p(X_i)$

# Propensity Score Matching

Proof: Assume that  $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | X_i$ . Then:

$$\begin{aligned} Pr(D_i = 1 | Y_i^1, Y_i^0, p(X_i)) &= E[D_i | Y_i^1, Y_i^0, p(X_i)] \\ &= E[E[D_i | Y_i^1, Y_i^0, p(X_i), X_i] | Y_i^1, Y_i^0, p(X_i)] \\ &= E[E[D_i | Y_i^1, Y_i^0, X_i] | Y_i^1, Y_i^0, p(X_i)] \\ &= E[E[D_i | X_i] | Y_i^1, Y_i^0, p(X_i)] \\ &= E[p(X_i) | Y_i^1, Y_i^0, p(X_i)] \\ &= p(X_i) \end{aligned}$$

# Propensity Score Matching

Using a similar argument, we obtain

$$\begin{aligned} \Pr(D_i = 1 | p(X_i)) &= E[D_i | p(X_i)] \\ &= E[E[D_i | p(X_i), X_i] | p(X_i)] \\ &= E[E[D_i | X_i] | p(X_i)] \\ &= E[p(X_i) | p(X_i)] \\ &= p(X_i) \end{aligned}$$

$$\begin{aligned} \Rightarrow \Pr(D_i = 1 | Y_i^1, Y_i^0, p(X_i)) &= p(X_i) = \Pr(D_i = 1 | p(X_i)) \\ \Rightarrow (Y_i^1, Y_i^0) &\perp\!\!\!\perp D_i | p(X_i) \end{aligned}$$

# Propensity Score Matching

- From CIA, to get causal effect, we need only control for covariates that affect the probability of treatment
- The propensity score theorem says something more:
  - The only covariate you really need to control for is the probability of treatment itself  $p(X_i) = \Pr(D_i = 1 | X_i)$**

利用機率  
來配對

# Identification Results or PSM

- Similar to identification results of matching estimator, the only difference is that we control  $p(X_i)$  instead of a set of covariates  $X_i$
- Remember CIA and propensity score theorem ensures

$$E[Y_i^0 | p(X_i), D_i = 1] = E[Y_i^0 | p(X_i), D_i = 0]$$

(看不見)

代替

(看得見)

# Identification Results or PSM

$$\begin{aligned} & \underbrace{\mathbb{E}[Y_i | p(X_i), D_i = 1] - \mathbb{E}[Y_i | p(X_i), D_i = 0]}_{\text{ODO at given } p(X_i)} \\ &= \mathbb{E}[Y_i^1 | p(X_i), D_i = 1] - \mathbb{E}[Y_i^0 | p(X_i), D_i = 0] \\ &= \mathbb{E}[Y_i^1 | p(X_i), D_i = 1] - \mathbb{E}[Y_i^0 | p(X_i), D_i = 1] \\ &\quad + \mathbb{E}[Y_i^0 | p(X_i), D_i = 1] - \mathbb{E}[Y_i^0 | p(X_i), D_i = 0] \\ &= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 | p(X_i), D_i = 1]}_{\text{CATT}} \\ &\quad + \underbrace{\mathbb{E}[Y_i^0 | p(X_i), D_i = 1] - \mathbb{E}[Y_i^0 | p(X_i), D_i = 0]}_{\text{Selection Bias}} \\ &= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 | p(X_i), D_i = 1]}_{\text{CATT}} + \underbrace{0}_{\text{Selection Bias}} \\ &= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 | p(X_i), D_i = 0]}_{\text{CATU}} = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 | p(X_i)]}_{\text{CATE}} \end{aligned}$$

# Identification Results or PSM

ATT

- Note that using a matching method, we can also identify CATT

$$E \left[ \underbrace{E[Y_i | p(X_i), D_i = 1] - E[Y_i | p(X_i), D_i = 0]}_{\text{ODO at given } p(X_i)} \right] = \underbrace{E[Y_i^1 - Y_i^0 | p(X_i), D_i = 1]}_{\text{CATT}}$$

- Applying LIE, we can identify ATT by averaging all of the  $p(X)$ -specific effects for treatment group (CATT):

$$E \left[ \underbrace{E[Y_i^1 - Y_i^0 | \cancel{p(X_i)}, D_i = 1 | D_i = 1]}_{\text{CATT}} \right] = \underbrace{E[Y_i^1 - Y_i^0 | D_i = 1]}_{\text{ATT}}$$

# Identification Results or PSM

ATU

- Note that using a matching method, we can also identify CATU

$$\underbrace{E[Y_i|p(X_i), D_i = 1] - E[Y_i|p(X_i), D_i = 0]}_{\text{ODO at given } p(X_i)} = \underbrace{E[Y_i^1 - Y_i^0|p(X_i), D_i = 1]}_{\text{CATU}}$$

- Applying LIE, we can identify ATU by averaging all of the  $p(X)$ -specific effects for treatment group (CATU):

$$\underbrace{E[E[Y_i^1 - Y_i^0|p(X_i), D_i = 0|D_i = 0]]}_{\text{CATU}} = \underbrace{E[Y_i^1 - Y_i^0|D_i = 0]}_{\text{ATU}}$$

# Identification Results or PSM

ATE

- Using a matching method, we can identify CATE

$$\underbrace{\mathbb{E}[Y_i|p(X_i), D_i = 1] - \mathbb{E}[Y_i|p(X_i), D_i = 0]}_{\text{ODO at given } p(X_i)} = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0|p(X_i)]}_{\text{CATE}}$$

- Applying LIE, we can identify ATE by averaging all of the  $p(X)$ -specific effects (CATE):

$$\mathbb{E}[\underbrace{\mathbb{E}[Y_i^1 - Y_i^0|p(X_i)]}_{\text{CATE}}] = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0]}_{\text{ATE}}$$

## Propensity Score Matching: Estimation

# Propensity Score Matching

## Estimation

- There are two ways to estimate causal effect of treatment using PSM

### 1 Nearest Neighbor:

- By matching each treated observation to the untreated observation with the same or similar values of the propensity score

### 2 Weighting Approach *skip*

- Skip the cumbersome matching procedure and re-weight sample

# Propensity Score Matching

Estimation: Nearest Neighbor

- There are two steps to estimate causal effect of treatment using PSM with nearest neighbor
  - 1 Estimate the propensity score:  $\hat{p}(X) = \hat{Pr}(D_i = 1|X_i)$  using logit or porbit regression

$$D_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \dots + \beta_k X_i^k + \epsilon_i$$

- 2 By matching each treated observation to the observation (control group) with the same or similar values of the propensity score  $\hat{Pr}(D_i = 1|X_i)$

# Propensity Score Matching

## A Numerical Example

Trainees			Non-Trainees			Matched Sample		
unit	pro-score	earnings	unit	pro-score	earnings	unit	pro-score	earnings
1	0.28	17700	1	0.43	20900			
2	0.34	10200	2	0.50	31000			
3	0.29	14400	3	0.30	21000			
4	0.25	20800	4	0.27	9300			
5	0.29	6100	5	0.54	41100			
7	0.33	21900	7	0.39	42000			
8	0.27	28800	8	0.28	8800			
9	0.31	20300	9	0.24	25500			
10	0.26	28100	10	0.33	15500			
11	0.25	9400	11	0.26	400			
12	0.27	14300	12	0.31	26600			
13	0.29	12500	13	0.26	16500			
14	0.24	19700	14	0.34	24200			
15	0.25	10100	15	0.25	23300			
16	0.43	10700	16	0.24	9700			
17	0.28	11500	17	0.29	6200			
18	0.27	10700	18	0.35	30200			
19	0.28	16300	19	0.32	17800			
			20	23	9500			
			21	32	25900			
Avg:		16426	Avg:		20724	Avg:		

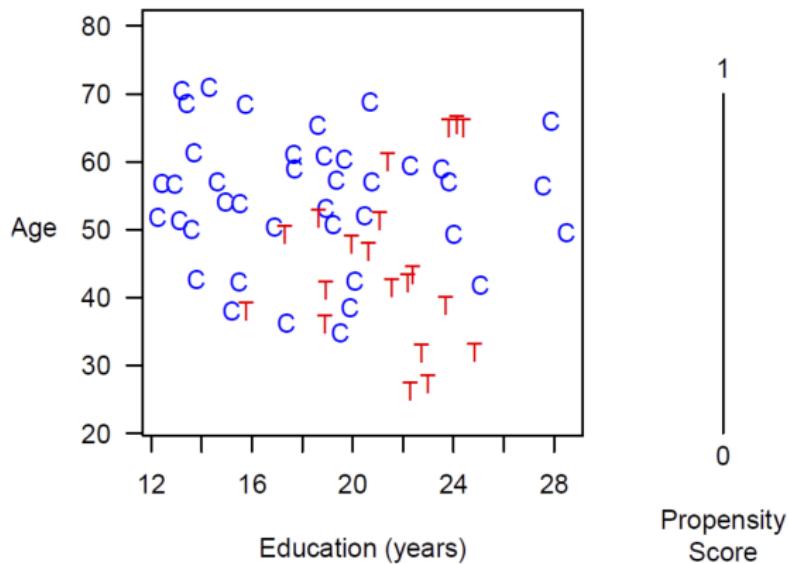
# Propensity Score Matching

## A Numerical Example

Trainees			Non-Trainees			Matched Sample		
unit	pro-score	earnings	unit	pro-score	earnings	unit	pro-score	earnings
1	0.28	17700	1	0.43	20900	8	0.28	8800
2	0.34	10200	2	0.50	31000	14	0.34	24200
3	0.29	14400	3	0.30	21000	17	0.29	6200
4	0.25	20800	4	0.27	9300	15	0.25	23300
5	0.29	6100	5	0.54	41100	17	0.29	6200
6	0.23	28600	6	0.48	29800	20	0.23	9500
7	0.33	21900	7	0.39	42000	10	0.33	15500
8	0.27	28800	8	0.28	8800	4	0.27	9300
9	0.31	20300	9	0.24	25500	12	0.31	26600
10	0.26	28100	10	0.33	15500	11,13	0.26	8450
11	0.25	9400	11	0.26	400	15	0.25	23300
12	0.27	14300	12	0.31	26600	4	0.27	9300
13	0.29	12500	13	0.26	16500	17	0.29	6200
14	0.24	19700	14	0.34	24200	9,16	0.24	17700
15	0.25	10100	15	0.25	23300	15	0.25	23300
16	0.43	10700	16	0.24	9700	1	0.43	20900
17	0.28	11500	17	0.29	6200	8	0.28	8800
18	0.27	10700	18	0.35	30200	4	0.27	9300
19	0.28	16300	19	0.32	17800	8	0.28	8800
			20	23	9500			
			21	32	25900			
Avg:	16426		Avg:	20724		Avg:	13982	

# Propensity Score Matching

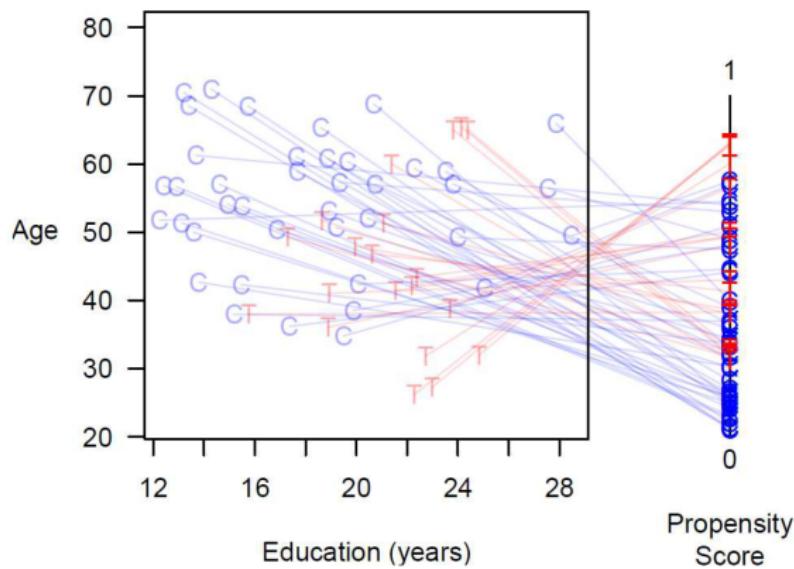
## A Numerical Example



Source: Ben Elsner's slides

# Propensity Score Matching

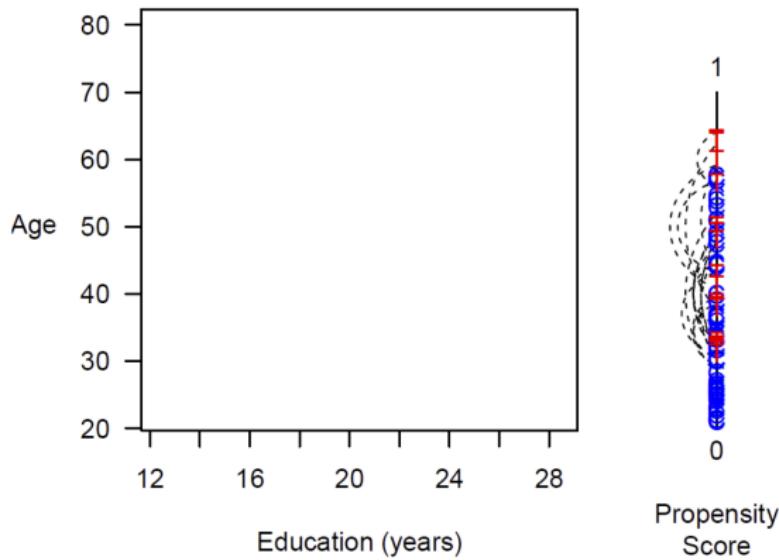
## A Numerical Example



Source: Ben Elsner's slides

# Propensity Score Matching

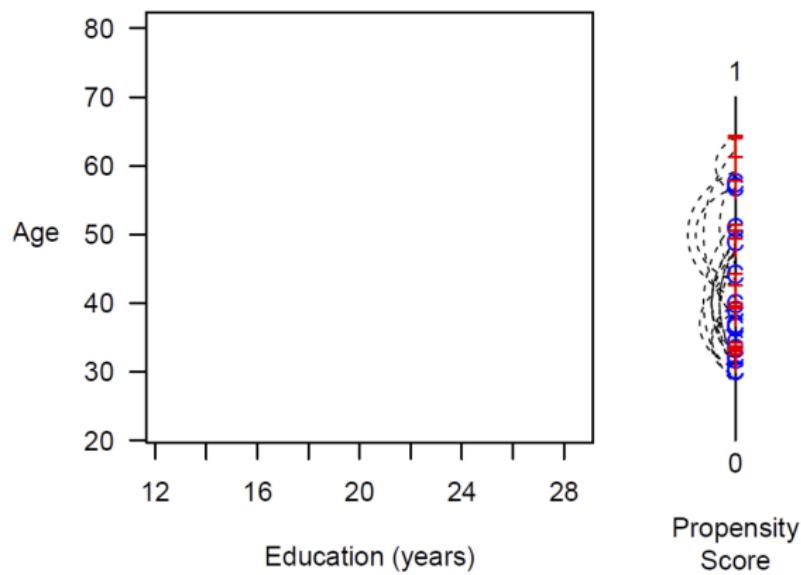
## A Numerical Example



Source: Ben Elsner's slides

# Propensity Score Matching

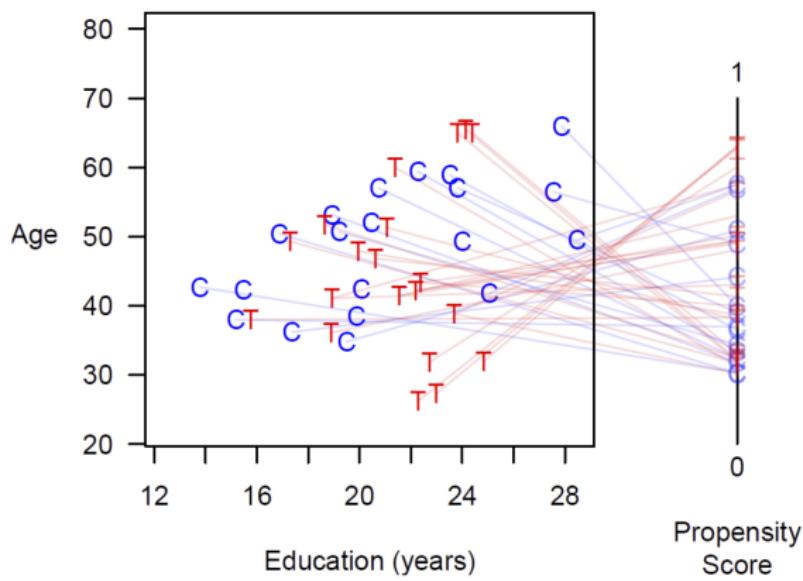
## A Numerical Example



Source: Ben Elsner's slides

## Propensity Score Matching

## A Numerical Example



*Source:* Ben Elsner's slides

# Propensity Score Matching

## Statistical Inference

- A valid method to calculate standard errors has not been known until very recently (see, Abadie and Imbens, 2016)
- Abadie, Alberto, and Guido W. Imbens. "**Matching on the Estimated Propensity Score.**" *Econometrica* 84.2 (2016): 781-807.
  - Need to take into account that propensity scores are estimated
  - The adjustment for ATE is always negative: smaller standard errors
  - The adjustment for ATT can be positive or negative: smaller or larger standard errors

## Propensity Score Matching – STATA Example

# STATA Example

Dehejia et al. (1999)

Rajeev H. Dehejia; Sadek Wahba (1999) “**Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs**” Journal of the American Statistical Association

- The authors wants to examine the effect of job training on workers' earnings
- We use this example to go through the procedure of implementing PSM

# STATA Example

Dehejia et al. (1999)

- See **matching.do**
- Use `lalonde.dta`
- Install the following ado files:
  - `psmatch2.ado`

# STATA Example

## Step 1: Test Differences in Outcomes in Pre-matching Data

```
1 ttest re78, by(treat)  
2 reg re78 treat,r
```

- Test differences in outcome for treatment group and control group

## STATA Example

#### Step 1: Test Differences in Outcomes in Pre-matching Data

```
. ** Step 1: Test Differences in Outcomes in Pre-matching Data  
. ttest re78, by(treat)
```

### Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	429	6984.17	352.1654	7294.162	6291.981	7676.359
1	185	6349.144	578.4229	7867.402	5207.95	7490.338
combined	614	6792.834	301.4942	7470.731	6200.748	7384.921
diff		635.0262	657.1374		-655.4917	1925.544

diff = mean(θ) - mean(1) t = 0.9664  
H<sub>0</sub>: diff = 0 degrees of freedom = 612

Ha: diff < 0  
Pr(T < t) = 0.8329

Ha: diff != 0  
Pr(|T| > |t|) = 0.3342

Ha:  $\text{diff} > 0$   
 $\Pr(T > t) = 0.1671$

# STATA Example

## Step 2: Test Differences in Covariates in Pre-matching Data

```
1 ttest age, by(treat)
2 ttest educ, by(treat)
3 reg age treat,r
4 reg educ treat,r
```

- Test differences in sample characteristics for treatment group and control group

# STATA Example

## Step 2: Test Differences in Covariates in Pre-matching Data

```
. ttest age, by(treat)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	429	28.0303	.5207845	10.78665	27.00669 29.05392
1	185	25.81622	.5260475	7.155019	24.77836 26.85408
combined	614	27.36319	.3987723	9.881187	26.58007 28.14632
diff		2.214087	.8652112		.5149437 3.91323

diff = mean(0) - mean(1) t = 2.5590  
Ho: diff = 0 degrees of freedom = 612

Ha: diff < 0 Pr(T < t) = 0.9946 Ha: diff != 0 Pr(|T| > |t|) = 0.0107 Ha: diff > 0 Pr(T > t) = 0.0054

# STATA Example

## Step 3: PSM Estimation – teffects psmatch

Syntax:

```
1 teffects psmatch (outcome) (treatment covariates,  
logit), nn(#) ate
```

- **nn(#)**: specify number of matches per observation; default is nn(1)
  - The number of variables generated may be more than nn(#) because of tied distances
- **logit**: use logit to predict propensity score (the default)
- **ate**: estimate average treatment effect in population (the default)
- **atet**: estimate average treatment effect on the treated

# STATA Example

## Step 3: PSM Estimation – teffects psmatch

Example:

```
1 teffects psmatch (re78) (treat age educ black  
hispan nodegree married re74 re75, logit), nn  
(1) atet  
2 teffects psmatch (re78) (treat age educ black  
hispan nodegree married re74 re75, logit), nn  
(1) ate
```

- Outcome: re78 (earnings in 1978)
- Treatment: treat (get job training or not)

# STATA Example

## Step 3: PSM Estimation – teffects psmatch

```
. teffects psmatch (re78) (treat age educ black hispan nodegree married re74 re75, logit), nn(1) ate
```

Treatment-effects estimation  
Estimator : propensity-score matching  
Number of obs = 614  
Outcome model : matching  
Matches: requested = 1  
Treatment model: logit  
min = 1  
max = 4

re78	AI Robust				
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
<b>ATE</b> treat (1 vs 0)	-304.6074	1076.527	-0.28	0.777	-2414.562 1805.347

# STATA Example

## Step 3: PSM Estimation – teffects psmatch

```
. teffects psmatch (re78) (treat age educ black hispan nodegree married re74 re75, logit), nn(1) atet
```

```
Treatment-effects estimation          Number of obs      =       614
Estimator    : propensity-score matching   Matches: requested =        1
Outcome model : matching                 min =        1
Treatment model: logit                  max =        4
```

re78	AI Robust				
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
ATET treat (1 vs 0)	1968.8	1126.321	1.75	0.080	-238.7493 4176.349

# STATA Example

## Step 3: PSM Estimation – teffects psmatch

# week 4

Understanding the matching process:

```
1 teffects psmatch (re78) (treat age educ black  
hispan nodegree married re74 re75), nn(1)  
atet gen(matchnum)
```

- **gen(matchnum)**: specifies that the observation numbers of the nearest neighbors be stored in the new variables matchnum1, matchnum2, ....
- This option is required if you wish to perform postestimation based on the matching results

# STATA Example

## Step 3: PSM Estimation – teffects psmatch

Understanding the matching process:

```
1 predict ps1, ps
2 predict y0 y1, po
3 predict te
```

- **predict ps1, ps:** predict propensity score (i.e. probability of getting treatment)
- **predict y0 y1, po:** generate the potential outcome with or without treatment
- **predict te:** get treatment effect for each observation

# STATA Example

## Step 3: PSM Estimation – teffects psmatch

ps1	y0	y1	te
.3612301	14421.13	9930.046	-4491.084
.7753658	1525.014	3595.894	2070.88
.3217561	2158.959	24909.45	22750.49
.2236759	701.9201	7506.146	6804.226
.2983612	14344.29	289.7899	-14054.5
.3009301	8900.347	4056.494	-4843.853

# STATA Example

## Step 3: PSM Estimation – teffects psmatch



對照 實驗

	id	matchnum1	treat	re78	ps1	y0	y1	te
1	1	254	1	9930.046	.3612301	14421.13	9930.046	-4491.084
2	254		0	14421.13	.3614458	14421.13	9930.046	-4491.084

機率相近

# STATA Example

## Step 3: PSM Estimation – psmatch2

Syntax:

```
1 psmatch2 treatment covariates, out(outcome) n(#)
    logit ate
```

- **n(#)**: specify number of matches per observation; default is nn(1)
  - The number of variables generated may be more than n(#) because of tied distances
- **out(var)**: specify an outcome variable
- **ate**: display ATT, ATU, ATE

# STATA Example

## Step 3: PSM Estimation – psmatch2

Example:

```
1 psmatch2 treat age educ black hispan nodegree  
    married re74 re75, out(re78) logit n(1) ate
```

- The PSM estimate is similar to the one using teffects

# STATA Example

Compare teffects psmatch and psmatch2



The **teffects psmatch** command has one very important advantage over **psmatch2**

- **teffects psmatch** takes into account the fact that propensity scores are estimated rather than known when calculating standard errors.
- **teffects psmatch** calculates standard errors based on this paper:
  - Abadie, Alberto, and Guido W. Imbens. "**Matching on the Estimated Propensity Score.**" *Econometrica* 84.2 (2016): 781-807.

# STATA Example

Compare teffects psmatch and psmatch2

不會重複配對 (影響不大)

- But **psmatch2** can allow matching without replacement, which is quite useful.

# STATA Example

## Step 3: PSM Estimation – psmatch2

Example:

```
1 psmatch2 treat age educ black hispan nodegree  
    married re74 re75, out(re78) logit n(1)  
    noreplace
```

- **noreplace**: STATA will perform PSM without replacement so that each untreated observation can be used only once.

# STATA Example

## Step 4: Post Matching Analysis – teffects psmatch

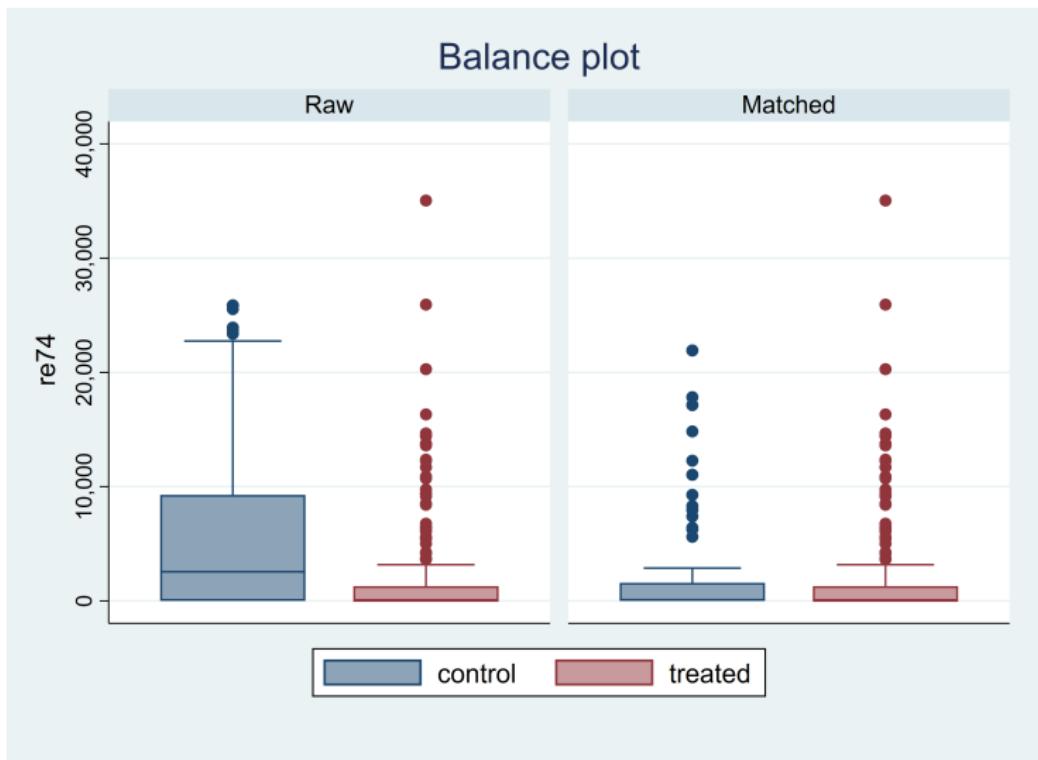
Example:

```
1 tebalance box re74  
2 tebalance density educ  
3 tebalance density
```

- **tebalance box**: Produces box plots that are used to check for balance in matched samples after **teffects**
- **tebalance density**: Produces density plots that are used to check for covariate balance after estimation by a **teffects**
- If you do not specify variable, it will plot the density of propensity score

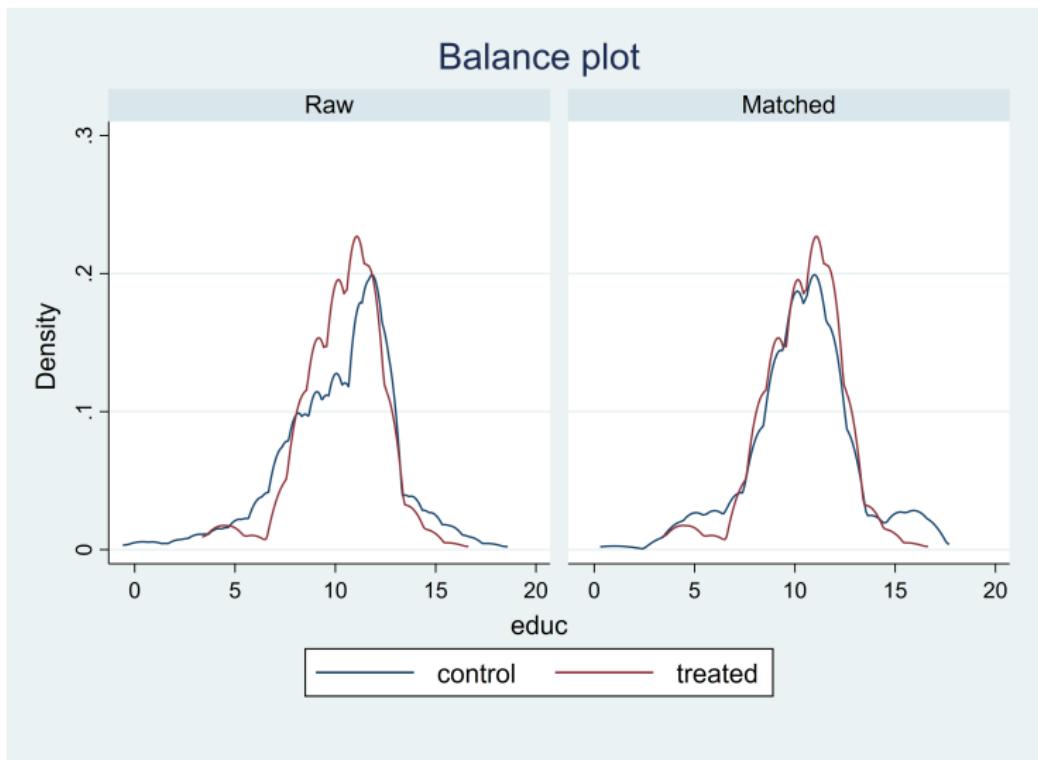
# STATA Example

## Step 4: Post Matching Analysis – teffects psmatch



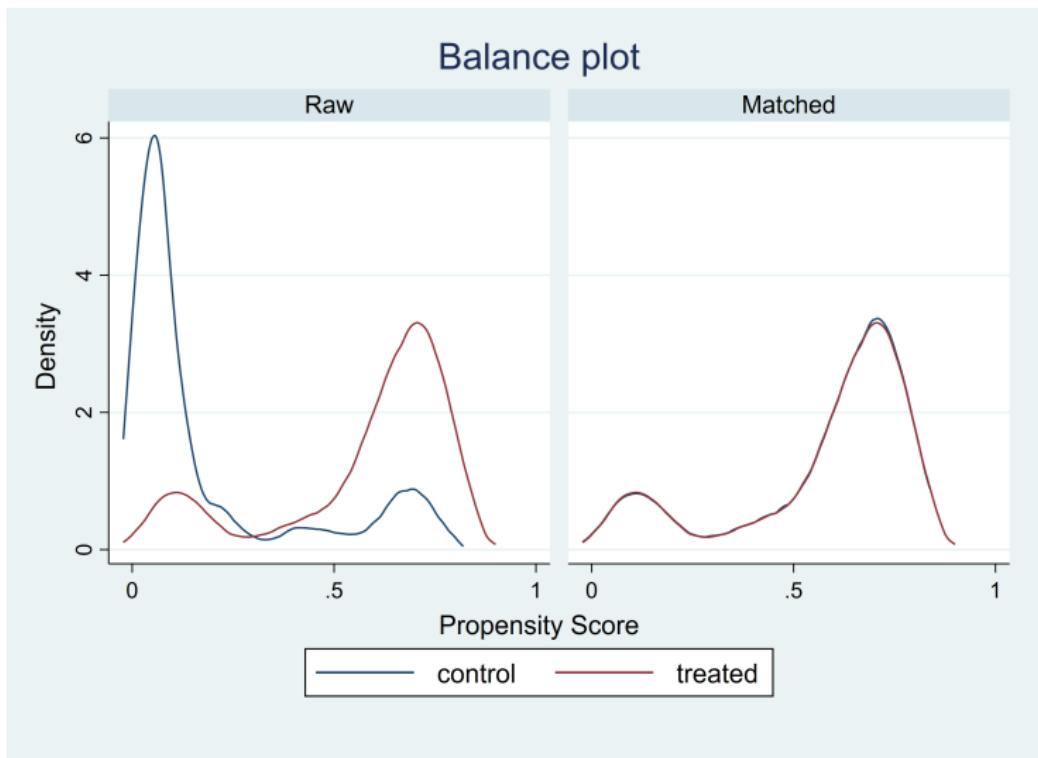
# STATA Example

## Step 4: Post Matching Analysis – teffects psmatch



# STATA Example

## Step 4: Post Matching Analysis – teffects psmatch



# STATA Example

## Step 4: Post Matching Analysis – psmatch2

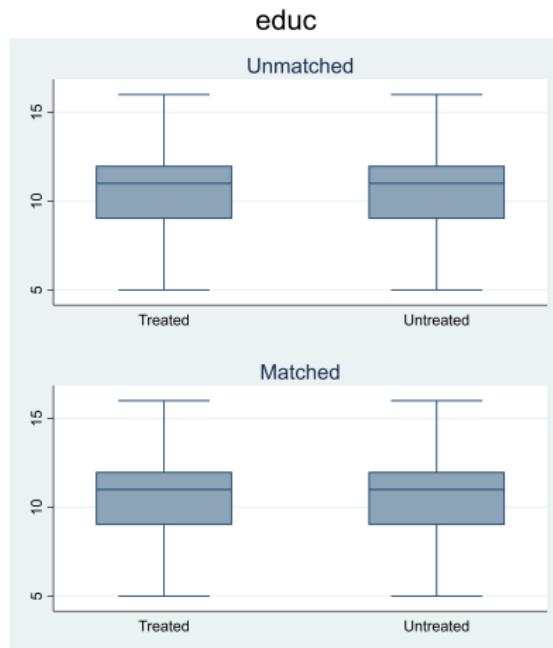
Example:

```
1 pstest age educ black hispan nodegree married  
      re74 re75, both  
2 pstest educ, box both  
3 pstest _pscore, density both
```

- command **pstest**: calculates and optionally graphs several measures of the extent of balancing of the variables between two groups.
- option **both**: compares the extent of balancing between the two samples before and after having performed matching.
- option **box**: draw box plot to compare two groups
- option **density**: draw density plot to compare two groups

# STATA Example

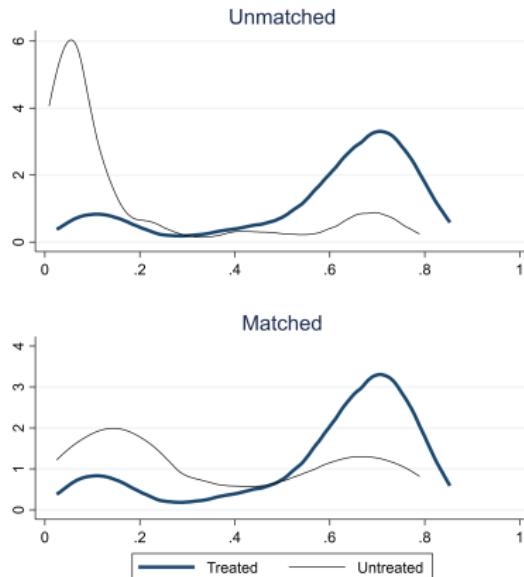
## Step 4: Post Matching Analysis – psmatch2



# STATA Example

## Step 4: Post Matching Analysis – psmatch2

psmatch2: Propensity Score



## Propensity Score Matching – R Example

# R Example

Dehejia et al. (1999)

- See **matching.R**
- Use `lalonde.dta`
- Install the following package:
  - `matchit`

# R Example

## Step 1: Test Differences in Outcomes in Pre-matching Data

```
1 summary(lm(re78~treat,data=lalonde))
```

- Test differences in outcome for treatment group and control group

# R Example

## Step 2: Test Differences in Covariates in Pre-matching Data

```
1 summary(lm(age~treat,data=lalonde))
```

- Test differences in sample characteristics for treatment group and control group

# R Example

## Step 3: PSM Estimation – matchit

Example:

```
1 m.out1 <- matchit(treat ~ age + educ + black + hispan  
+ nodegree + married + re74 + re75, data =  
lalonde, method = "nearest", distance = "logit",  
replace=FALSE)  
2 m.data1 <- match.data(m.out1,distance ="pscore")  
3 summary(lm(re78~treat, data = m.data1))
```

- **matchit:** perform PSM
- **match.data:** get matched sample
- **replace=FALSE:** matching without replacement

# R Example

## Step 3: PSM Estimation – matchit

Example:

```
1 m.out1 <- matchit(treat ~ age + educ + black + hispan  
+ nodegree + married + re74 + re75, data =  
lalonde, method = "nearest", distance = "logit",  
replace=TRUE)  
2 m.data1 <- match.data(m.out1,distance ="pscore") #  
create ps matched data set from previous output  
3 summary(lm(re78~treat, data = m.data1,weights=weights  
))^I^I^I
```

- **replace=TRUE**: matching with replacement

# Propensity Score Matching

## Drawback

- PSM is hugely popular method to estimate treatment effects even if it relies on **less convincing assumption:**
  - Selection on observables (CIA)

# Suggested Readings

- Chapter 2, Mastering Metrics: The Path from Cause to Effect
- Chapter 3, Mostly Harmless Econometrics
- Chapter 5, Causal Inference: The Mixtape