Lecture 9: Matching

Introduction to Econometrics,Spring 2026

Zhaopeng Qu
Nanjing University Business School
May 14 2026

Review the Last Two Lectures

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. **Censored, Trancated and Selection Samples**
  5. **Simultaneous causality**

- The data structure may violate the 2th OLS regression assumption, thus **random sampling**.
  - adjusted the s.e. by **clustering** or other methods.
- Last but not least, the economic **magnitude** of $\hat{\beta}$ matters.
  - economic significance is as important as statistical significance.

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/adjust/control for the covariates X to make 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**.
  - Besides the regression, we can also use **matching** to achieve this goal.

Matching: Introduction

Introduction

- In **observational** studies, as opposed to RCTs, we cannot directly determine the causal effect because the **counterfactual** outcome of the treated group is unknown.
- In other words, we **cannot find a suitable control group** to compare with the treated 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?**
  - A **reasonable control group** should be **similar** to the treated group in terms of the covariates X before the treatment.
  - A similar but not equivalent question is find a suitable treat group by selecting some samples in treated group.
- For simplicity, we focus on the former question, ie. **constructing a control group**, which is more common in practice.

Introduction

- Suppose Y_{i1} and Y_{i0} are the outcomes of the treated and untreated group, respectively.
- And we can use some or all samples from untreated group to construct the **counterfactual outcomes** of the treated group Y_{i1}^c
- Then **the average treatment effect(ATE or ATT)** easily by making the difference

$$\delta_{ATE} = E[Y_{i1} - Y_{i1}^c]$$
$$\delta_{ATT} = E[Y_{i1} - Y_{i1}^c | D = 1]$$

- **Question**: How can we use samples from the untreated group to get the counterfactual outcomes of the treated group, Y_{1i}^c?
- **Answer**: select **the untreated samples** that are **similar to the treated ones** in terms of **the covariates** X_i
- **Assumption**: If CIA holds, thus $(Y_1, Y_0) \perp\!\!\!\perp D|X$, then the treatment status can be seen as randomized given the covariates X_i.

Example: Training Program Evaluation

- **Question**: What is the causal effect of a training program on the wage of workers?
- A simple OLS regression model can be written as

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

  - **The treatment variable** D_i is the treatment status, $D_i = 1$ if the worker has received the training program, $D_i = 0$ otherwise.
  - **The outcome** is the log-wage of workers Y_i, and the covariates X_i form a vector including variables such as age, education, experience, etc.

- Then the key coefficient $\hat{\beta}_1$ is the **difference in mean values log-wage** between the treated group and untreaded group

$$\hat{\beta}_1 = E[Y_{1i}|X_i] - E[Y_{0i}|X_i]$$

- If all the OLS assumptions hold, then the estimated coefficient $\hat{\beta}_1$ can be interpreted as **the causal effect**.

Unmatched Samples by trainning status

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

- The average wage gap between the treated group and the untreated group is

$$\delta = E[Y_{1i} - Y_{0i}] = 16426 - 20724 = -4298$$

- It appears that joining the training program will **reduce** the wages of workers by 4298.
- **Question**: Can you find the bias of this estimation?

OLS Regression for the Training Program

- **Answer**: Yes, we may suffer the **OVB**.The common way to solve this problem is to add some covariates into the regression model.

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

- where X_i is the covariate of the workers, like age, education, experience, etc.
- The main identifying assumption of the OLS regression here is
  - **Conditional Independence Assumption(CIA)**: which means that if we can "balance" covariates X then we can take the treatment D as randomized.
- However, we may still suffer the **misspecification** of the model under the **CIA**, which can also make estimates β_1 biased.
- Therefore, we may use a method which can balance the covariates X like OLS but **without the assumption of function forms**:
  - **Matching**

A Trainning Example: matching samples

- Assume that the covariates X is the **age of the workers**, and to see **how the matching method works**.
- We pick **the untreated samples** that are **similar to the treated samples** in terms of the age of the workers.

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: matching samples

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: matching samples

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: matching samples

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: matching samples

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: matching samples

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: matching samples

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: matching samples

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: matching samples

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: matching samples

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: matching samples

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
Average:		

A Trainning Example: matching samples

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: matching samples

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: matching samples

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
Average:		

A Trainning Example: matching samples

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: matching samples

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
Average:		

A Trainning Example: matching samples

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: matching samples

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: matching samples

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
4	27	9300
8	28	8800
Average:		

An Illustrated Example: matched samples

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
4	27	9300
8	28	8800
Average:	28.5	13982

- The average wage gap between the treated group and the **matched untreated group** is

$$\delta = E[Y_{1i} - Y_{0i}] = 16426 - 13982 = 2444$$

- Now, joining the training program will **increase** the wage of workers by 2444 .

Two Assumptions: One Old and One New

- We still rely on the **Conditional independence Assumption(CIA)**, which is akin to running an OLS regression.

$$(\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i | \mathrm{X}_i$$

  - More specifically, we assume that the potential incomes for the workers are **independent of th e training status** given the age of the workers.

- It means that if **CIA** are not satisfied, then both the OLS and the matching estimator will be biased.

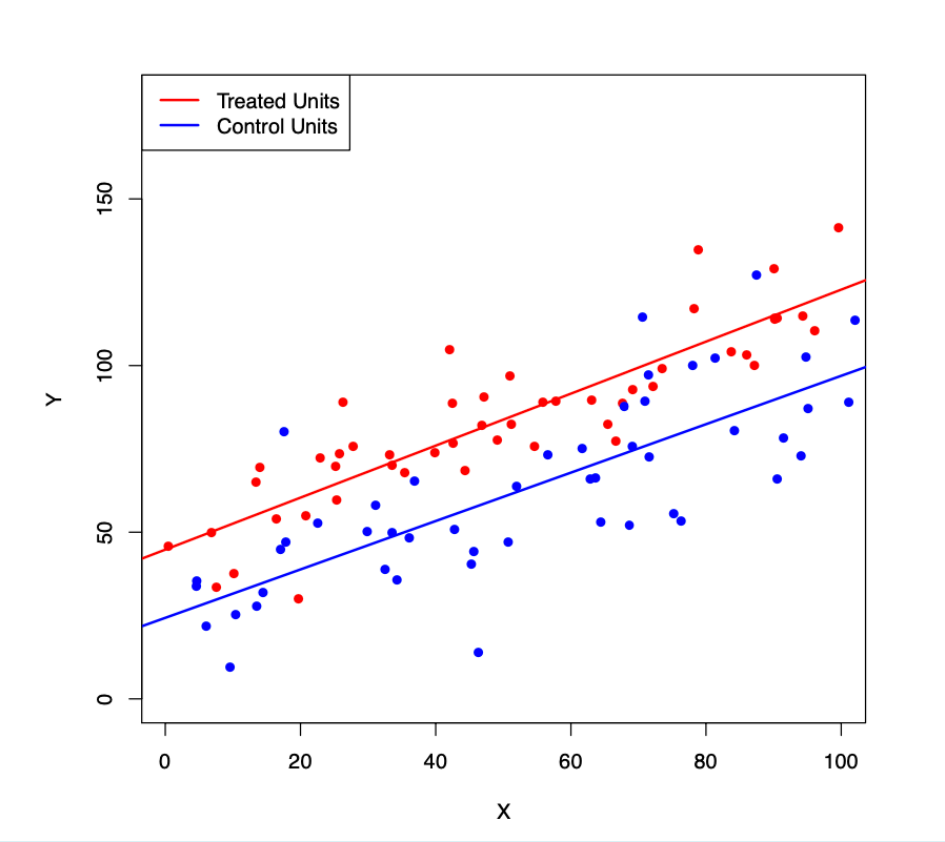
  - Matching is ***not a silver bullet*** for **OVB** in OLS.

- Besides, do you notice that **there are some untreated samples that are not matched with any treated samples**?

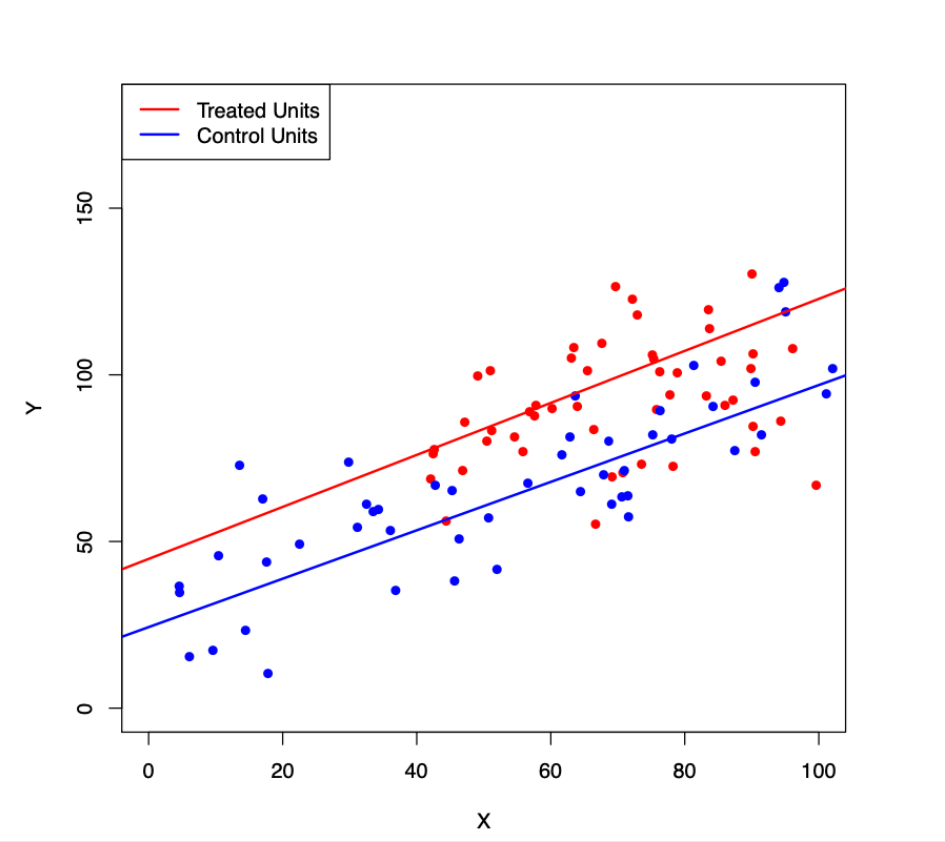
  - It means that these samples are ***not used*** in the estimation of the average treatment effect at all.

  - This is due to the **Overlap Assumption**, a new assumption in the matching method that was not discussed in the OLS regression.

The Overlap Assumption in OLS



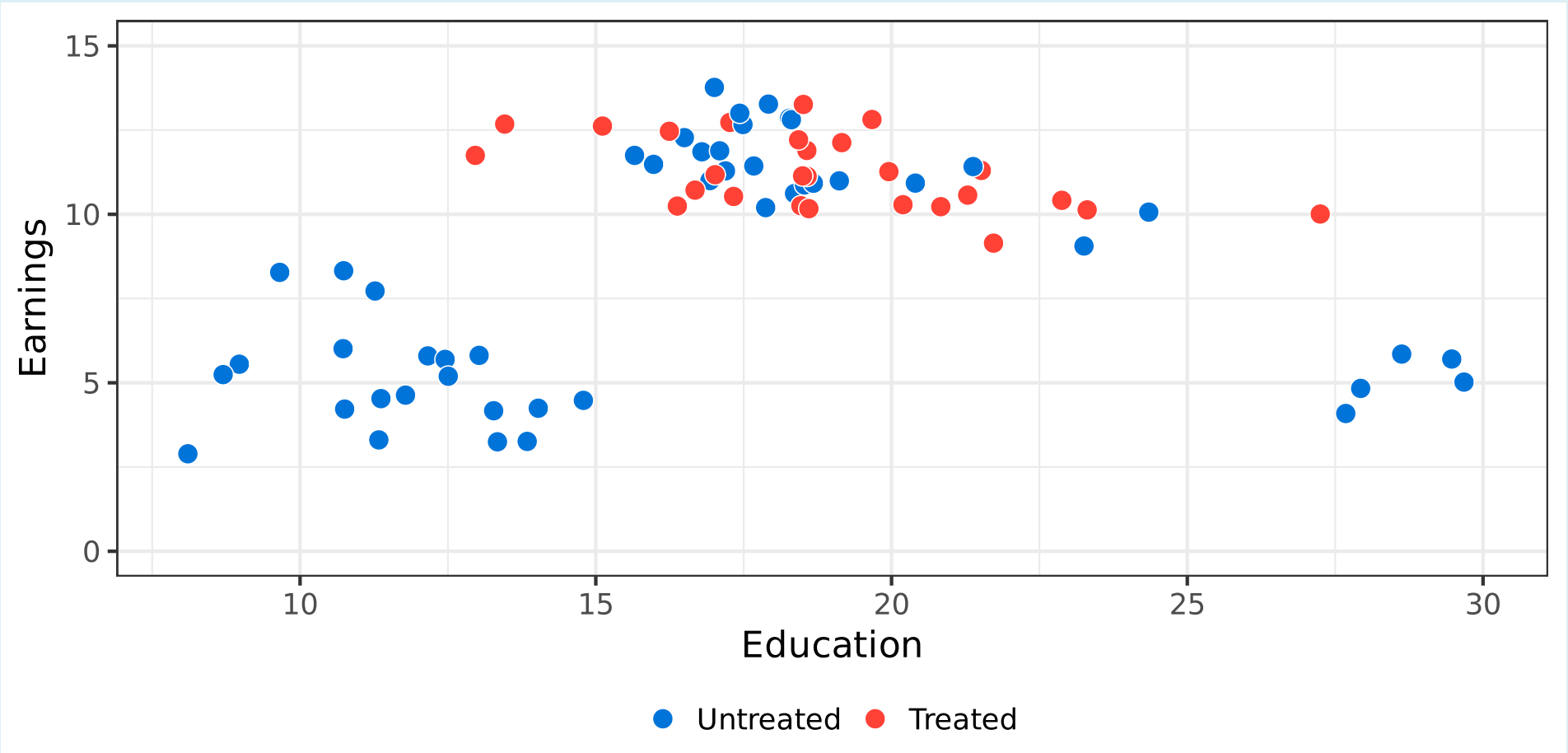
The overlap assumption is satisfied



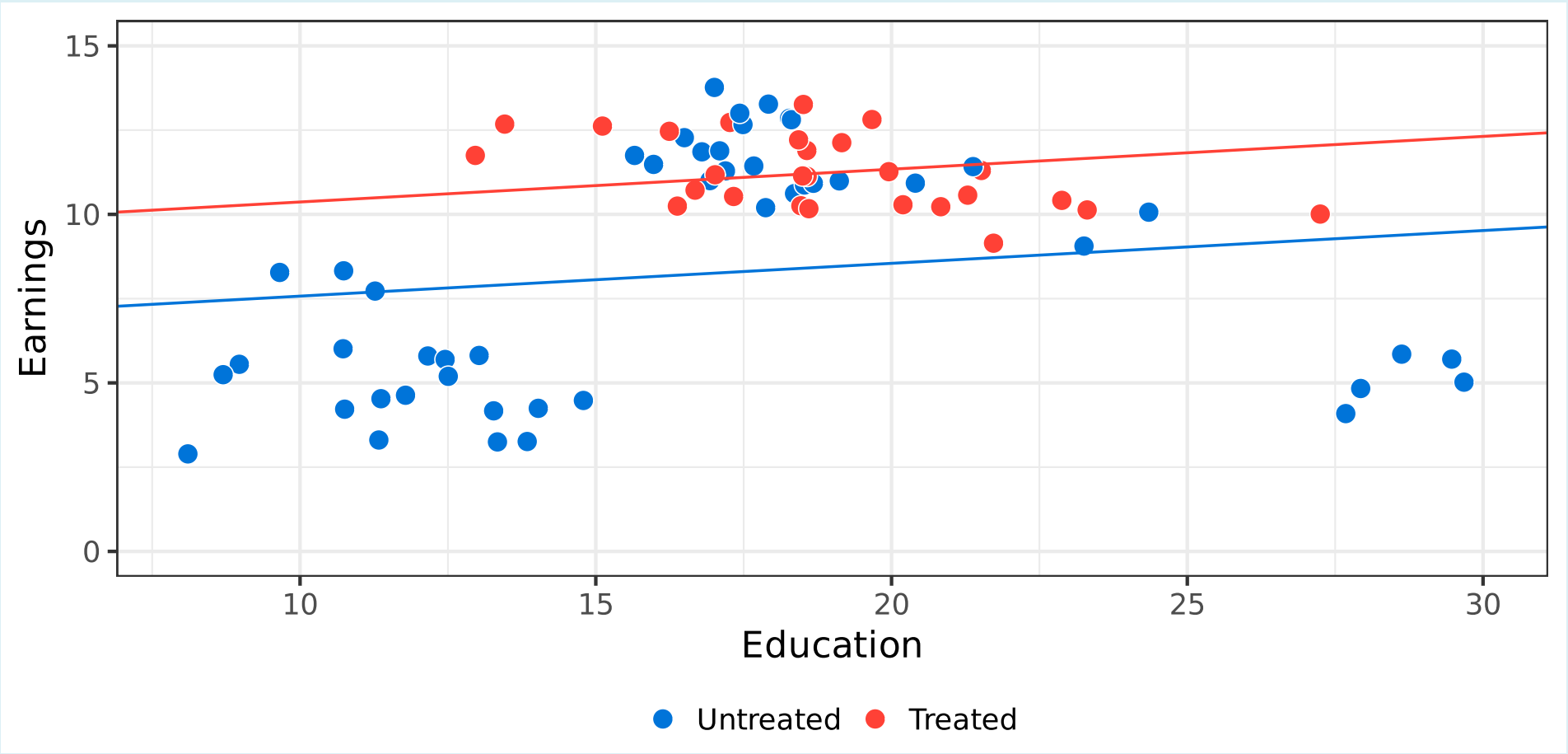
The overlap assumption is violated

- In the OLS regression, the overlap assumption is not **explicitly required**, which may lead to biased estimates.

OLS vs Matching: Overlap Assumption



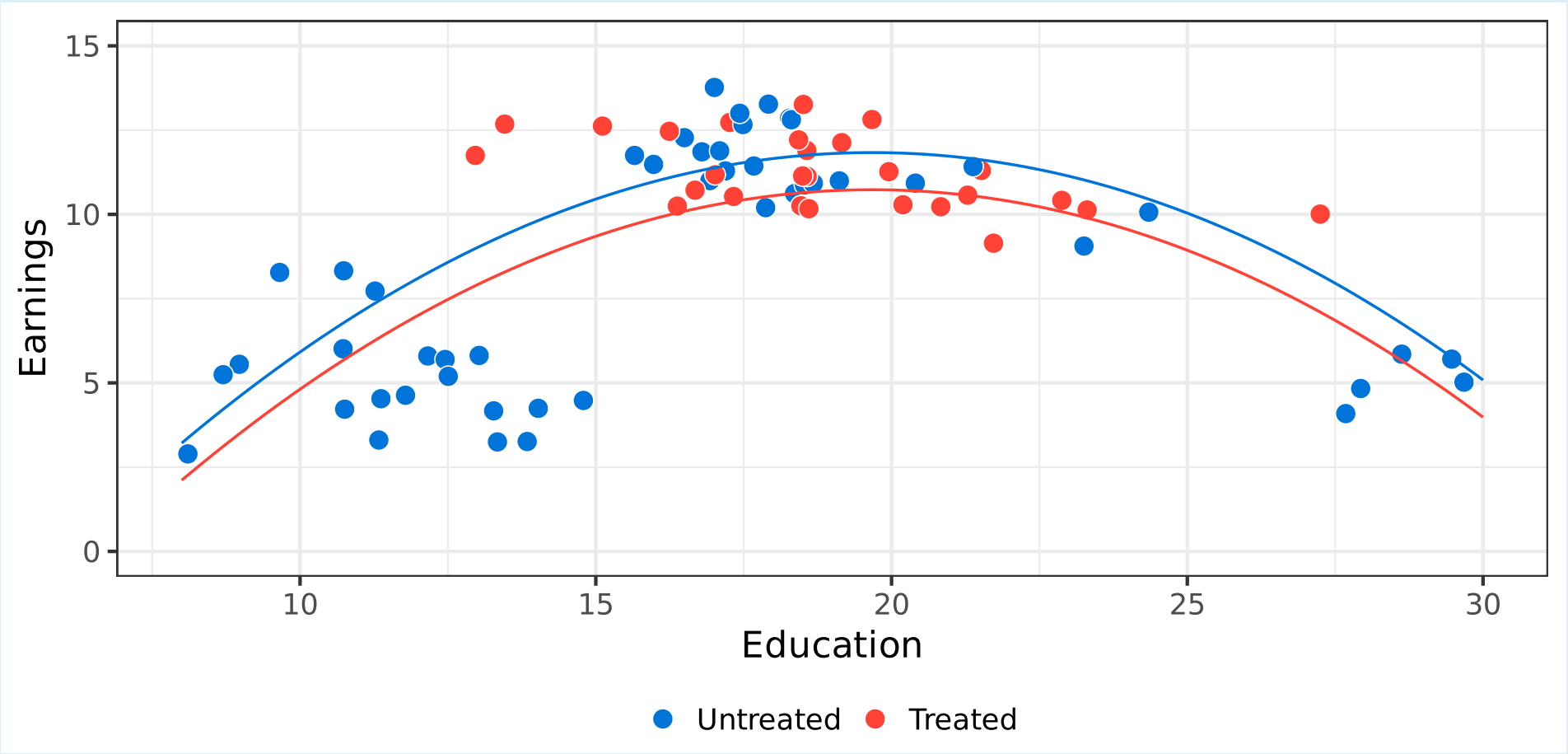
OLS vs Matching: Overlap Assumption



$$\text{Earnings} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Treatment} + u$$

- Basic take-away: *treated group* is **higher** than the *untreated group* in the OLS regression.

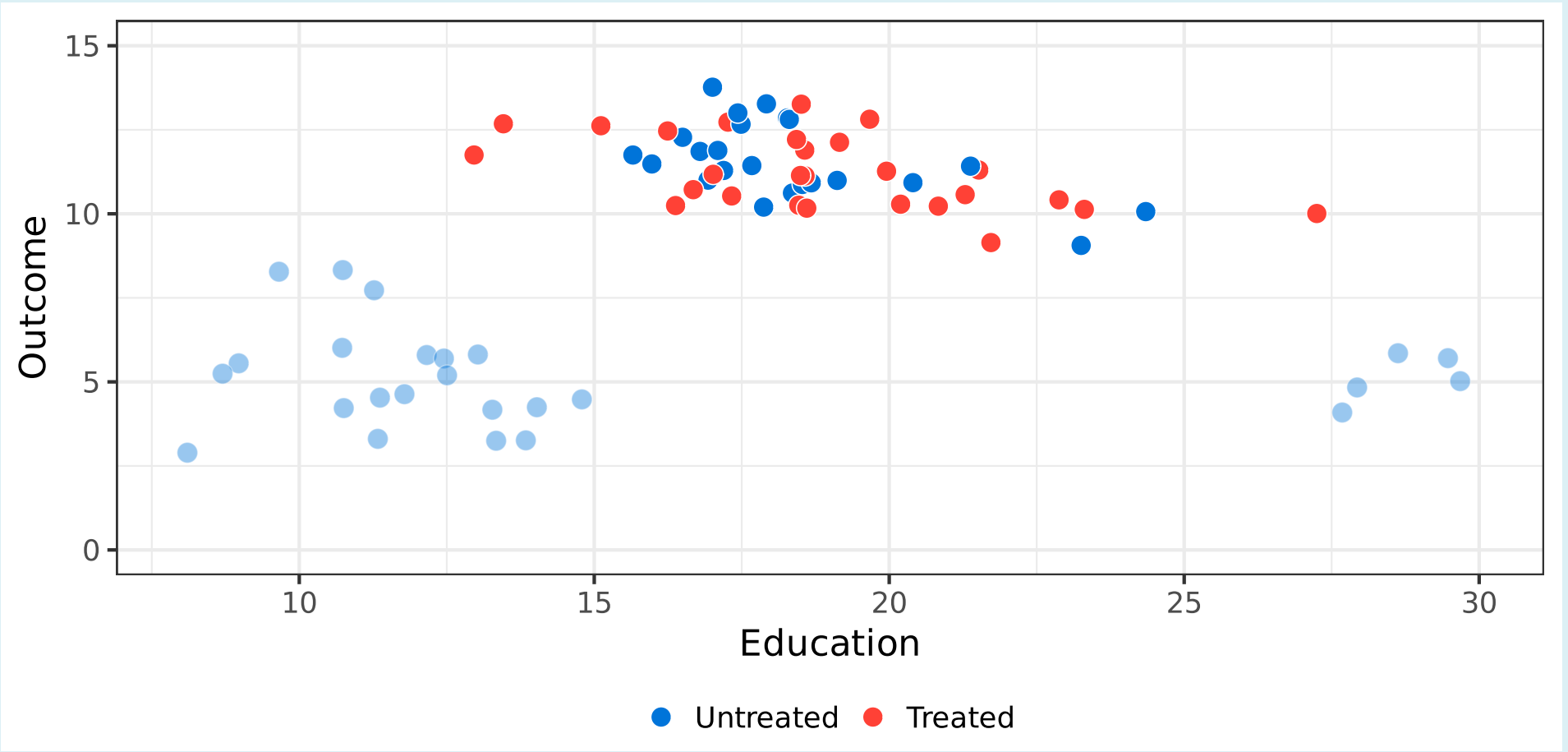
Overlap Assumption in Nonlinearity



$$\text{Earnings} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Education}^2 + \beta_3 \text{Treatment} + u$$

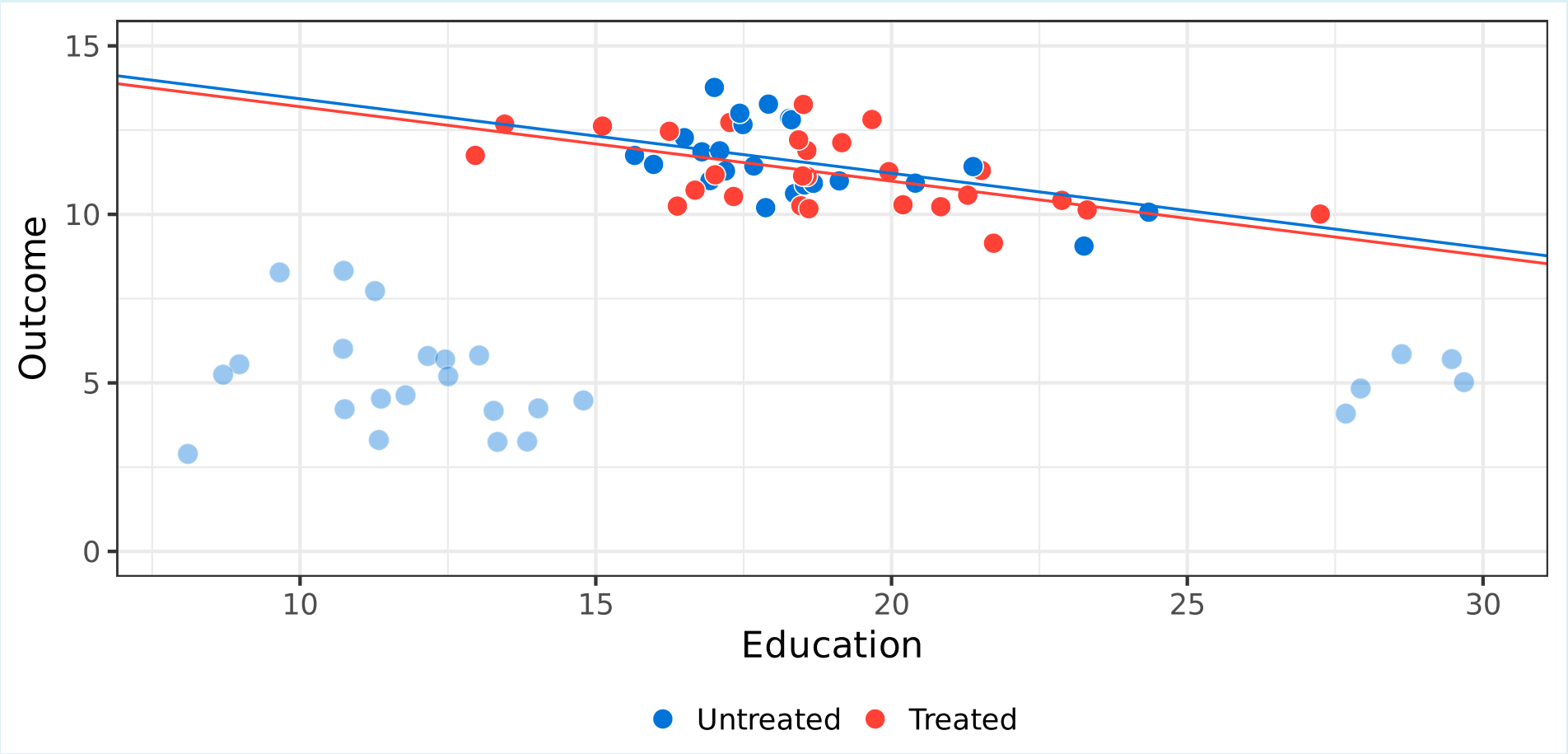
- Basic take-away: *untreated group* is **higher** than the *treated group* in the OLS regression.

Overlap Assumption: Common Support



- The **common support region** is the region where *both the treated and untreated groups have data.*

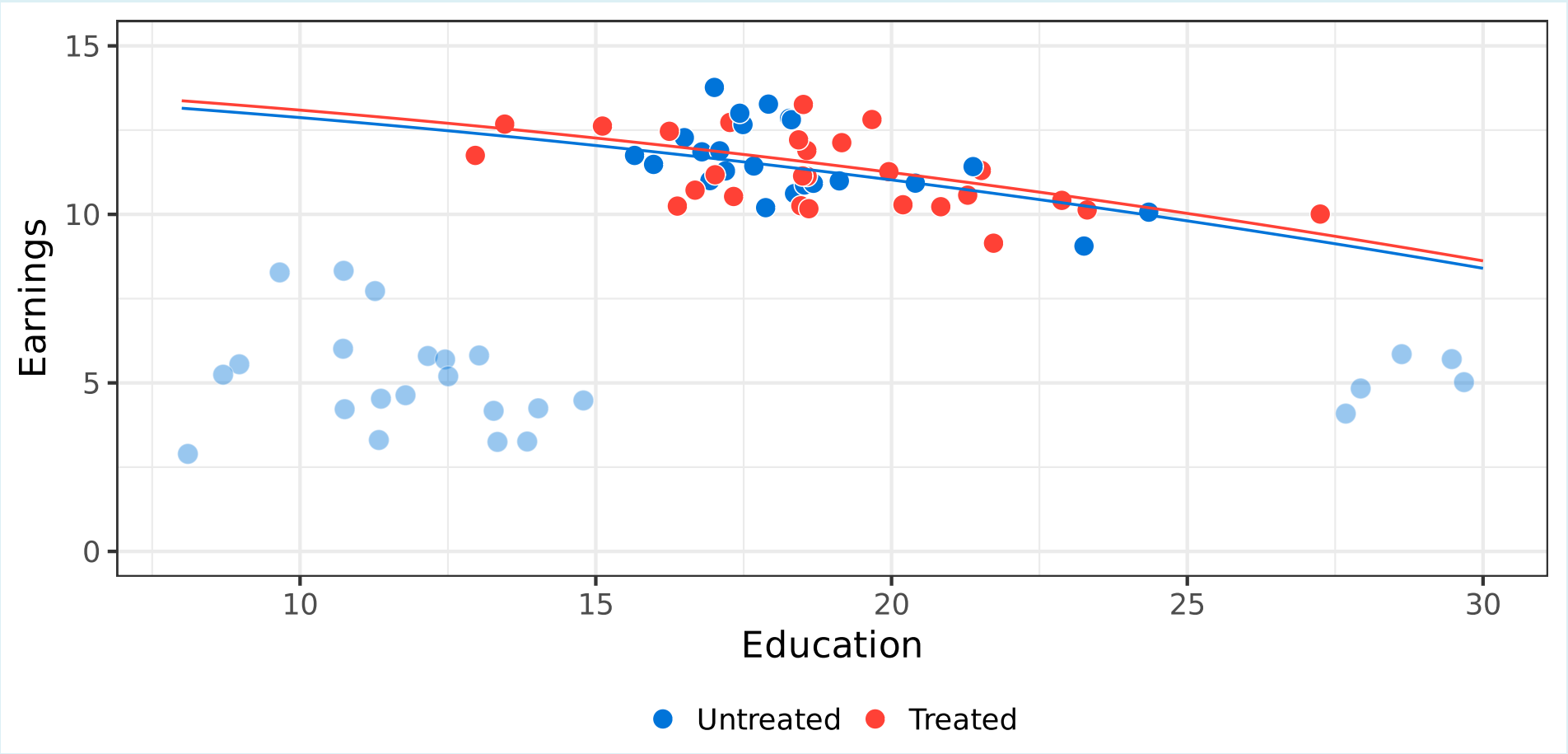
Common Support: After Trimming Data



$$\text{Earnings} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Treatment} + u$$

- Basic take-away: *treated group* is **very close** to the *untreated group* in the OLS regression.

Common Support: After Trimming Data



$$\text{Earnings} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Education}^2 + \beta_3 \text{Treatment}$$

- Basic take-away: *treated group* is still **very close** to the *untreated group* in the OLS regression.

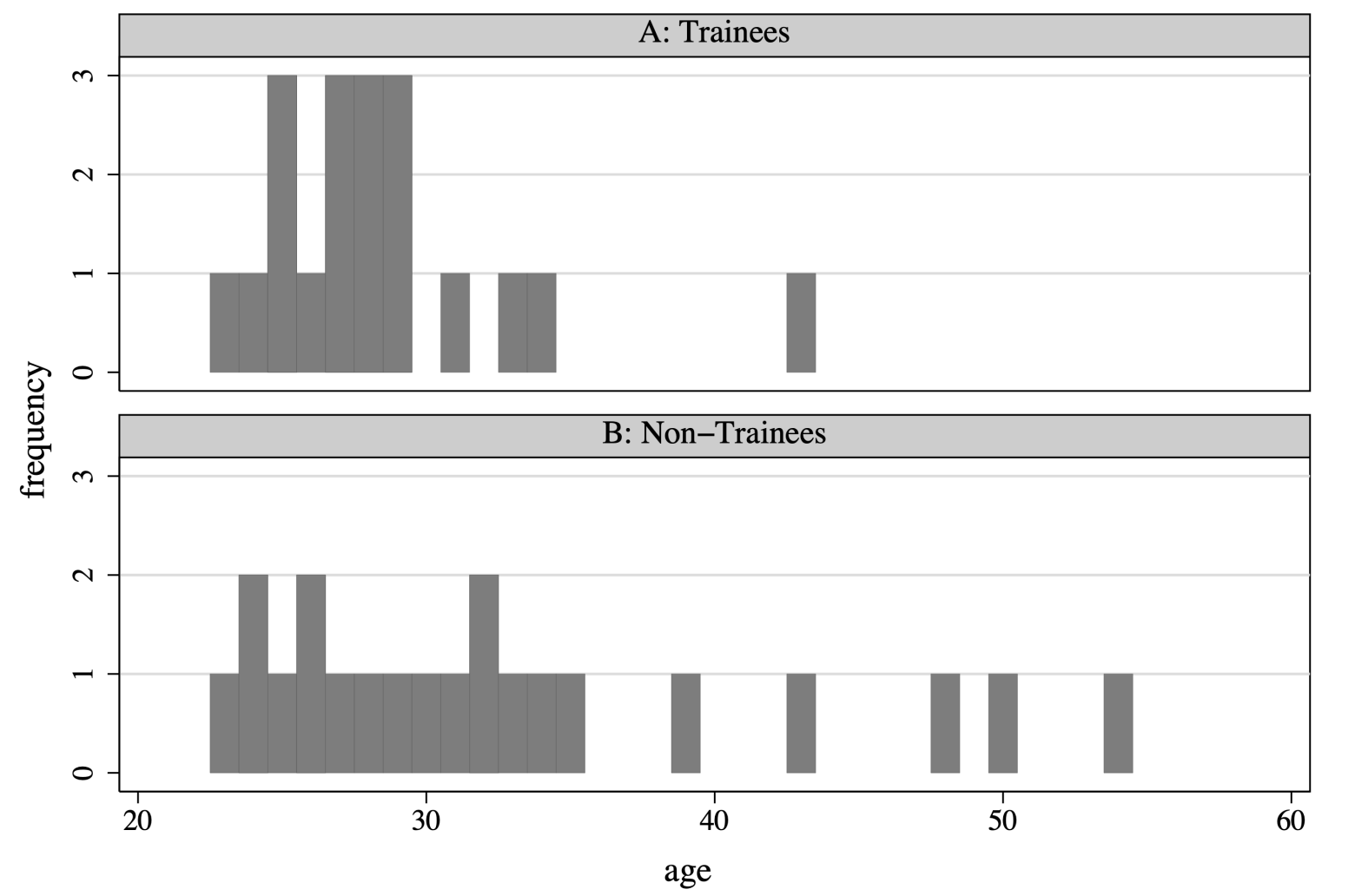
Two Assumptions: One Old and One New

- The **Overlap Assumption** is to ensure that we can find a matched untreated sample for each treated sample.Mathematically, it is expressed as:

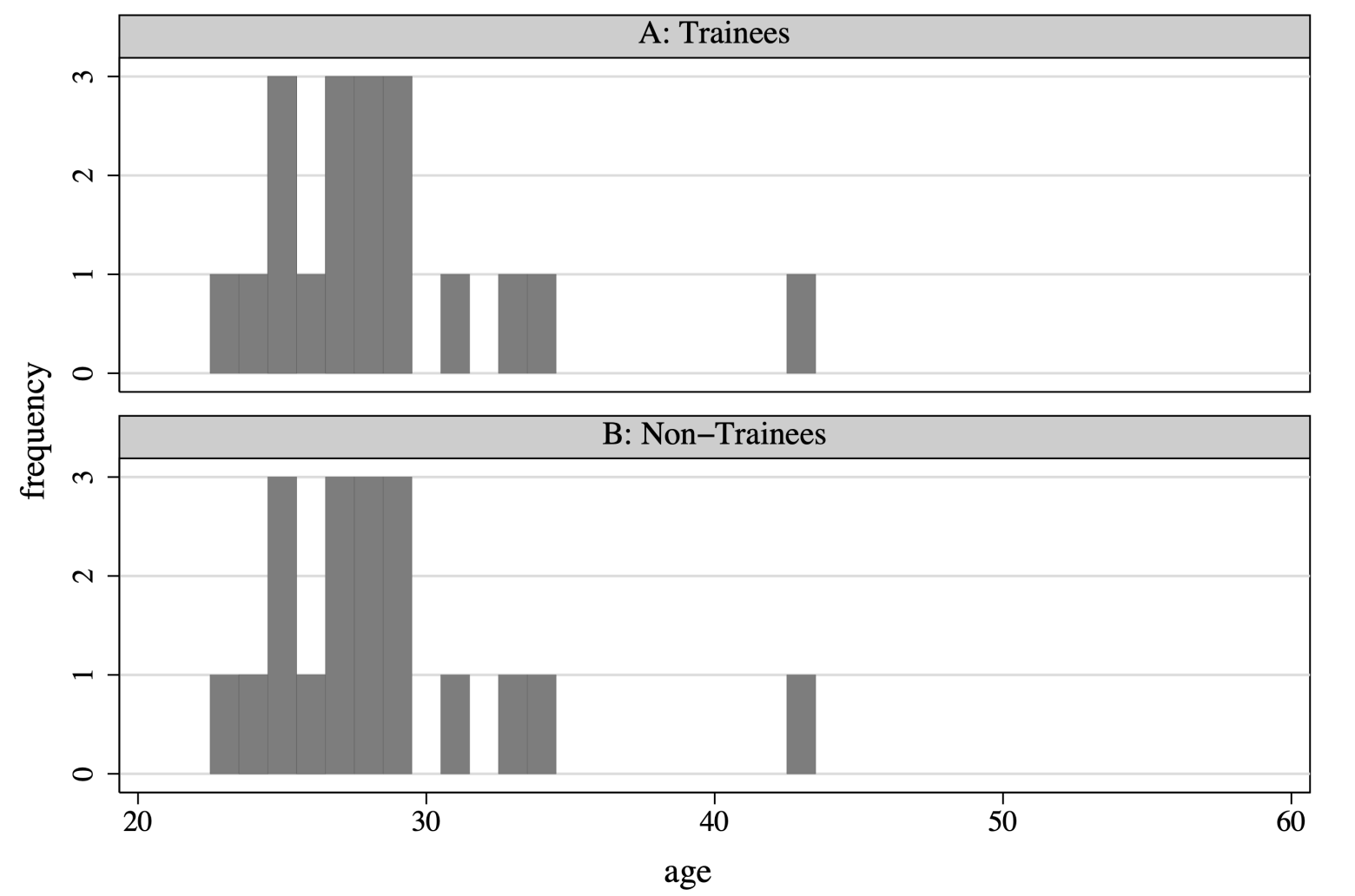
$$0 < \Pr(\mathrm{D}_i = 1 \mid \mathrm{X}_i) < 1$$

- This implies that the **likelihood of receiving treatment** is neither 0 nor 1 for any given covariates.
  - If the **probability of receiving treatment** is 0 for some X_i, then no samples with these characteristics receive treatment, **making them unavailable for matching**.
  - If the **probability of receiving treatment** is 1 for some X_i, then all samples with these characteristics receive treatment, **making it impossible to find untreated matches**.
  - Including either case in our comparison would **bias the average treatment effect estimation**.
- It suggests that we **change the samples explicitly** based on the covariates to ensure that the **overlap assumption** is satisfied.

A Trainning Example: before matching



A Trainning Example: after matching



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 $X = (X_1, X_2, \dots X_k)'$?
  - In this case, it is **impossible** to find proper units with identical values in all covariates $X_1, X_2, \dots X_k$.
- Two complementary solutions are running in parallel, representing the directions in which the matching method is developing.

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*.

2. **Directly reduce dimensions**.
   - Converting multiple variables into *a single numerical value*, then use the numerical value to match the samples.

Matching Estimator

Introduction

- The matching estimator can be divided into three steps: **Matching**, **Estimation** and **Inference**.
- **Matching**: Find a control group for each treated individual based on the covariates.
    1. define how to measure the similarity between the treated and untreated samples.
    2. choose the criteria to match the samples.
    3. evaluate the quality of the matching.
- **Estimation**: Estimate the average treatment effect(making a difference) using the matched samples.
- **Inference**: Test the statistical significance of the treatment effect(ATT or ATE) using the matched samples.
- We will mainly focus on the **Matching**, the **Estimation** and **Inference** will be skipped.

Reweight as Counterfactuals

- **Basic settings**: all notations are the same as before,like Y_{1i}, Y_{0i}, D_i, and X_i.
  - the **sample size** here is the only one need to noted : N_T treated individuals and N_C control individuals.
- The **counterfactual for treated individual** i that what we need is Y_{1i}^C, then **the question is how to construct it by matching**?
- **Answer**: select **the untreated samples** that are **similar to the treated ones** in terms of **the covariates** X_i.
- Suppose we can find a set of untreated samples $j = 1, 2, \ldots, N_C$ that are similar to the treated sample i ,then the counterfactual for treated individual i is

$$Y_{1i}^C = Y_{0j}$$

Reweight as Counterfactuals

- In a more general sense,the counterfactual for treated individual i can be constructed as a reweighting of the untreated samples.

$$Y_{1i}^{C} = \sum_{j=1}^{N_C} w_i(j) Y_{0j}$$

- where $w_i(j)$ is a **weight** of untreated individual j for treated individual i, and normally $\sum_j w_i(j) = 1$
- In the former case, the weights

$$w_i(j) = \begin{cases} 1 & \text{if} X_i = X_j \\ 0 & \text{otherwise} \end{cases}$$

- Then the counterfactual for treated individual i is

$$Y_{1i}^{C} = \sum_{j=1}^{N_C} 0 * Y_{01} + \ldots + 0 * Y_{0(j-1)} + 1 * Y_{0j} + 0 * Y_{0(j+1)} + \ldots + 0 * Y_{0N_C} = Y_{0j}$$

Matching Estimator

- Then individual treatment effect, δ_i, is

$$\delta_i = \mathrm{Y}_{1i} - \mathrm{Y}_{1i}^{C} = \mathrm{Y}_{1i} - \sum_j w_i(j)\mathrm{Y}_{0j}$$

- A matching estimator for the average treatment effect on the treated(ATT) is

$$\hat{\delta}_M = \frac{1}{N_T} \sum_{i \in (\mathrm{D}=1 \cap \mathrm{C})} \delta_i = \frac{1}{N_T} \sum_{i \in (\mathrm{D}=1 \cap \mathrm{C})} (\mathrm{Y}_{1i} - \mathrm{Y}_{1i}^{c}) = \frac{1}{N_T} \sum_{i \in (\mathrm{D}=1)} \left(\mathrm{Y}_{1i} - \sum_{j \in (D=0 \cap \mathrm{C})} w_i(j)\mathrm{Y}_{0j} \right)$$

- Where C is the **common support region** of the treated and untreated individuals.
- And $i \in (\mathrm{D} = 1 \cap \mathrm{C})$ means that i is a treated individual and i is in the **common support region**.
- And $j = 1, 2, \ldots, N^C$ and $i = 1, 2, \ldots, N^T$.

Weight to Matching

- **Question**: **How to obtain these weights**,**thus** $w_i(j)$?
- **Answer**: *It is easy and hard at the same time.*
- E.g. if $w_i(j) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$
- In this case, the weights are **equal** for all the untreated samples.

$$\begin{aligned} \hat{\delta}_M &= \frac{1}{N_T} \sum_{i \in (\mathrm{D}=1 \cap \mathrm{C})} \left(\mathrm{Y}_{1i} - \sum_{j \in (D=0 \cap \mathrm{C})} w_i(j) \mathrm{Y}_{0j} \right) \\ &= \frac{1}{N_T} \sum_{i \in (\mathrm{D}=1 \cap \mathrm{C}) \& j=i} \left(\mathrm{Y}_{1i} - \mathrm{Y}_{0j} \right) \end{aligned}$$

- Then we're back to a **difference in means**, except now it's based on the N_T matched samples.

Weight to Matching

- **More Reasonable Weights**: The weights $w_i(j)$ should be **related with covariates** X_i in treated group and X_j in untreated group.
  - The idea: **more similar the covariates are, more weight the untreated sample should have**.

Proximity: When X is Discrete

- If X is **discrete**,then we can use the **equality** of X to construct the weights.Thus

$$w_i(j) = \mathbb{I}(X_i = X_j)$$

  - Where $\mathbb{I}(\cdot)$ is an **indicator function**,

$$\mathbb{I}(X) = \begin{cases} 1 & \text{if } X_i = X_j \\ 0 & \text{otherwise} \end{cases}$$

- This is the **Exact Matching** what we did in the trainning case.

Proximity: When X is Continuous

- If X is **continuous**, then we may not find a unit with the same covariate values.Then we may need **proximity** rather than **equality**.
- Then the weight $w_i(j)$ can be ***a measure of how close*** X_j of untreated group is to X_i of the treated group.

$$|X_i - X_j|$$

  - If the gap(distance) is **small**, then the weight is **large**, and vice versa.
- **Question**: What do "**small**" and "**large**" mean in the previous sentence?
  - It depends on the **distance metric**.
- If we just pick **the smallest one** as we did in the trainning case, then we have the **Nearest Neighbor Matching**.

Math Review: Distance between two vectors

- If X_i and X_j are both single-dimensional variables, then the distance between them is the difference between them,

$$|X_i - X_j|$$

- What if X_i and X_j are both multi-dimensional variables,thus k-dimensional vectors as follows

$$X_i = (X_{i1}, X_{i2}, \ldots, X_{ik}) \quad \text{and} \quad X_j = (X_{j1}, X_{j2}, \ldots, X_{jk})$$

- **Question**: how to measure the distance between two vectors?

- **Answer**: The **Euclidean distance** can be as the measure of the distance between X_i and X_j,

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

Proximity: When X is a Vector

- The Euclidean distance is not invariant to changes in the **scale** of X. 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, thus

$$V_X^{-1} = \begin{bmatrix} \hat{\sigma}_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \hat{\sigma}_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \hat{\sigma}_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \hat{\sigma}_k^2 \end{bmatrix}$$

- $\hat{\sigma}_k^2$ is the variance of the k-th variable.
- **No scale problem** but still *no correlations* between Xs.

Proximity: When X is a Vector

- **Mahalanobis distance** between X_i and X_j is defined as

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

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

$$\Sigma_X^{-1} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \hat{\sigma}_{13} & \cdots & \hat{\sigma}_{1k} \\ \hat{\sigma}_{11} & \hat{\sigma}_{22} & \hat{\sigma}_{23} & \cdots & \hat{\sigma}_{2k} \\ \hat{\sigma}_{11} & \hat{\sigma}_{32} & \hat{\sigma}_{33} & \cdots & \hat{\sigma}_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{k1} & \hat{\sigma}_{k2} & \hat{\sigma}_{k3} & \cdots & \hat{\sigma}_{kk} \end{bmatrix}$$

- $\hat{\sigma}_{jk}$ is the covariance between the j-th and k-th variables.
- **No scale problem** and **taking correlations between Xs into account**.

Many Matching Methods

- There many methods to choose the matchers and weights. Here are some of them:
- **Exact Matching**:
  - the weight $w_i(j)$ is 1 if j is the exact match of i and 0 otherwise.
  - **problem**: The exact match may not be found.
- **Nearest Neighbor Matching(NNM)**:
  - The smallest distance between the treated and untreated group.
  - the weight $w_i(j)$ is 1 if j is the nearest neighbor of i and 0 otherwise.
  - **problem**: The nearest neighbor may not be a good match.
- **Caliper Matching**: Samples within a certain range are matched.
  - the weight $w_i(j)$ is 1 if j is in the range and 0 otherwise.
  - **problem**: How to choose the range?

Many Matching Methods

- **Radius Matching**: all the samples within a certain range are matched.
  - the weight $w_i(j)$ is 1 if j is in the range and 0 otherwise.
  - **problem**: How to choose the **radius**?
- **Subclassification** : Divide the treated and untreated group into subclasses based on the covariates and then match within each subclass.
  - the weight $w_i(j)$ is 1 if j is in the same subclass as i and 0 otherwise.
  - **problem**: How to choose the subclasses?
- **Kernel Matching**: The weight is based on the **kernel function**, which is an estimated density function of the covariates.
  - all the samples in the untreated group are used to estimate the counterfactual outcome.
  - the weight is based on the **specific kernel function**

The curse of dimensionality

- As the dimension of $\mathbf{X}$ expands (i.e., matching on more variables), whatever matching method we use, it *becomes increasingly difficult* to find a suitable or closely matched control for each treated sample, even if we have a large sample size.
- Need alternative ways to **shrink** the dimensions of $\mathbf{X}$.
- **Propensity scores**
- It turns out that if CIA is satisfied,then we actually **only** need to match/conditional on the **propensity score** $p(x)$, instead of the entire $\mathbf{X}_i$.

Propensity-Score Methods

The Magic of Propensity Scores

- The **propensity score** is defined as *the **probability of treatment** given X_i, thus

$$p(\mathrm{X}_i) = p(D_i = 1|X_i) = E[\mathrm{D}_i|\mathrm{X}_i]$$

- Recall the **CIA** assumption:

$$(\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|\mathrm{X}_i$$

- Formally the **Propensity Score Theorem** is saying that
  - If $(\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|\mathrm{X}_i$, then $(\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|p(\mathrm{X}_i)$.
  - **Explanation**: If we control/adjust/balance the **propensity score** instead of the raw covariates, then the treatment is as good as random.
- This theorem extends CIA assumption from **multiple dimensions** to **a one-dimensional score**, avoiding the curse of dimensionality.

Propensity-Score Theorem

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

To prove this theorem, we will show

$$\Pr(\mathrm{D}_i = 1 \mid \mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)) = p(\mathrm{X}_i)$$

i.e., D_i is independent of $(\mathrm{Y}_{0i},\ \mathrm{Y}_{1i})$ after conditioning on $p(\mathrm{X}_i)$.

Propensity-Score Theorem

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

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

$$= E\left[\mathrm{D}_i \middle| \mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)\right] \quad \because \ D_i \text{ is a dummy variable}$$

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

$$= E\left[\ E\left(\mathrm{D}_i \middle| \mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ \mathrm{X}_i\right) \middle| \mathrm{Y}_{0i},\ \mathrm{Y}_{1i},\ p(\mathrm{X}_i)\right] \quad \because \ \text{if X is known, then p(x) is known. But it's not vice versa.}$$

Propensity-Score Theorem

Theorem If $(Y_{0i},\ Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i},\ Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

$$\Pr\Big[D_i = 1 \Big| Y_{0i},\ Y_{1i},\ p(X_i)\Big] = \cdots = E\Big[\ E\Big(D_i \Big| Y_{0i},\ Y_{1i},\ X_i\Big) \Big| Y_{0i},\ Y_{1i},\ p(X_i)\Big]$$

$$= E\Big[\ E\Big(D_i \Big| X_i\Big) \Big| Y_{0i},\ Y_{1i},\ p(X_i)\Big] \ \because \text{CIA}$$

$$= E\Big[p(X_i) \Big| Y_{0i},\ Y_{1i},\ p(X_i)\Big] \ \because \text{P-Scores' definition}$$

$$= p(X_i) \neq f(Y_{0i},\ Y_{1i})$$

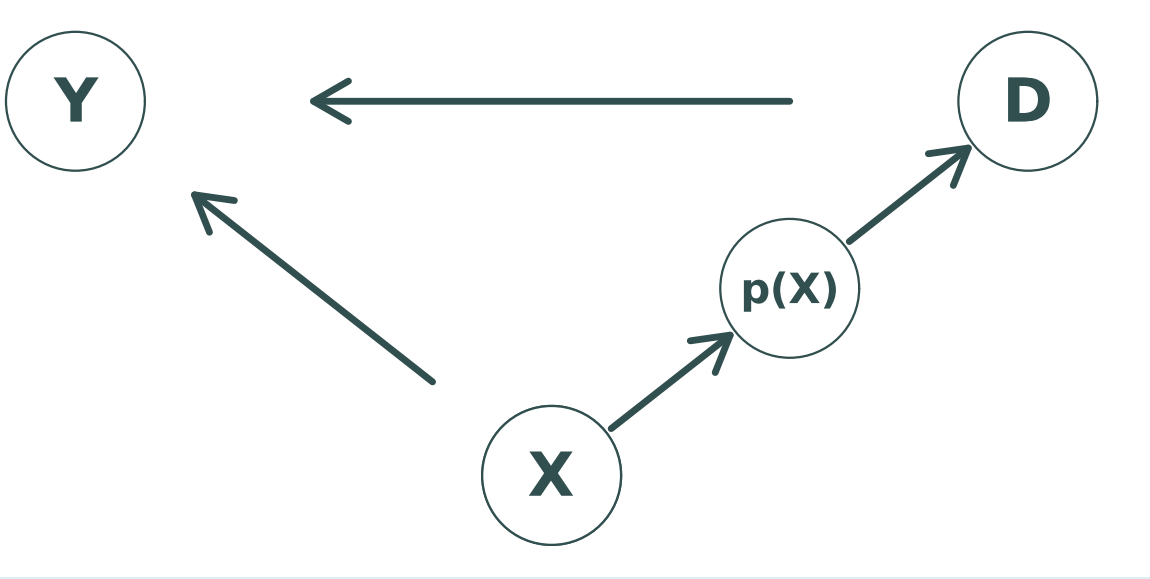
$$\therefore\ (Y_{0i},\ Y_{1i}) \perp\!\!\!\perp D_i | X_i \implies (Y_{0i},\ Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$$

Propensity-Score Matching

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



Propensity-Score Matching

Estimation: Binary Dependent Regression

- **Question**:How to obtain the propensity scores $p(\mathrm{X}_i)$?
- Recall the definition of propensity score, **does it sound familiar?**

$$p(\mathrm{X}_i) = \Pr(\mathrm{D}_i = 1 \mid \mathrm{X}_i)$$

  - Yes,it is the **binary dependent** regression model that the independent variables are the covariates X_i.
- We learned that it can be estimated by three models as follows:

  1. **LPM**
  2. **Logit**
  3. **Probit**

- Of course there are another ways to estimate it like **machine learning methods**, but the most common way is to use **logit** regression.

Propensity-Score Matching

Estimation: Logit Regression

- The **logit model of the propensity score** is given by

$$p(\mathrm{X}_i) = E(D = 1|X_i) = \frac{1}{1 + e^{(-\mathbf{X}_i\beta)}}$$

- Where $\mathbf{X}_i$ is the vector of covariates and β is the vector of coefficients.

- The **estimated propensity scores** $\hat{p}(\mathrm{X}_i)$ can be obtained by plugging in the estimated coefficients $\hat{\beta}$.

$$p(\mathrm{X}_i) = \frac{1}{1 + e^{(-\mathbf{X}_i\hat{\beta})}}$$

Propensity-Score Matching

Estimation: Logit Regression

- However, for the *nonlinearity* of the model, the **marginal effect** of coviarites on the propensity score is not constant.
    - It means that *the same change of the covariates will NOT have the same effect* on the propensity score for all the values of the covariates.
- Therefore, a more common way to estimate the propensity score is to use the **log odds ratio**,

$$ln\left(\frac{p(\mathrm{X}_i)}{1-p(\mathrm{X}_i)}\right) = \mathbf{X}_i\beta$$

- **Recall**: We claimed that matching is over regression as it is **non-parametric**, don't need to specify the functional form of the model.
- However, in the propensity score method, we still need to **specify the functional form of the model and estimate the coefficients**.

Propensity-Score Matching

Estimation: Predicted instead of Explained

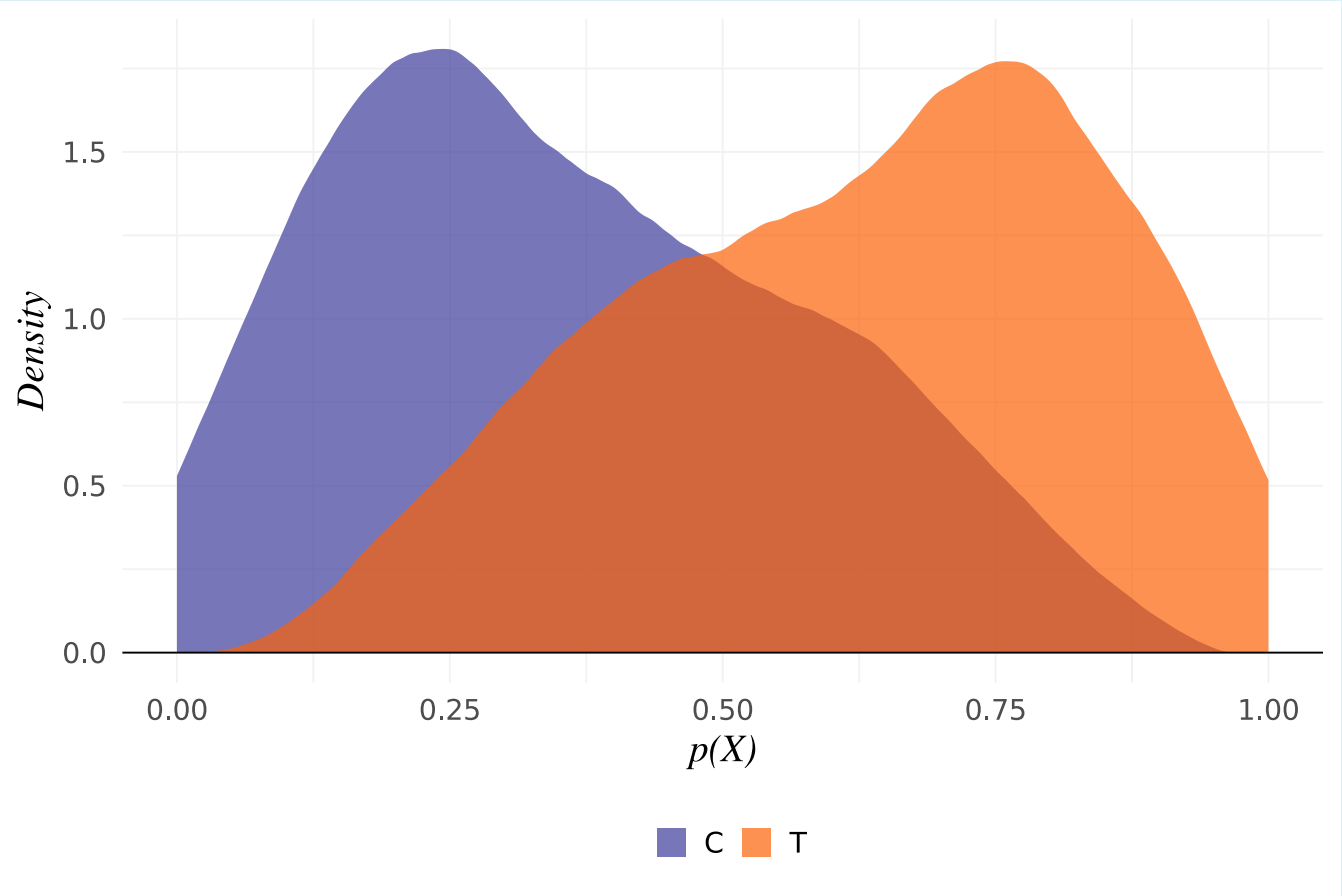
- **Note**: The focus in the model here is a little bit different from the one we learned in the **binary dependent variable regression**.
    - Here we focus on the predicted probability of being treated, which is the **propensity score**, and the covariates are the explanatory variables.
- While in the **binary dependent variable regression**, we focus on the **explanatory coefficient** of the covariates(only one or two in most cases) on the treated variable(which actually is the dependent variable).
- Therefore, when we estimate the propensity score by the logit model, the function form should be **as flexible as possible** to capture the relationship between the covariates and the treatment variable.
    - **Polynomial terms and interaction terms** are often included in the model.
    - Even **ML methods** can be used to estimate the propensity score as well.

Overlap Assumption in Propensity Score Methods

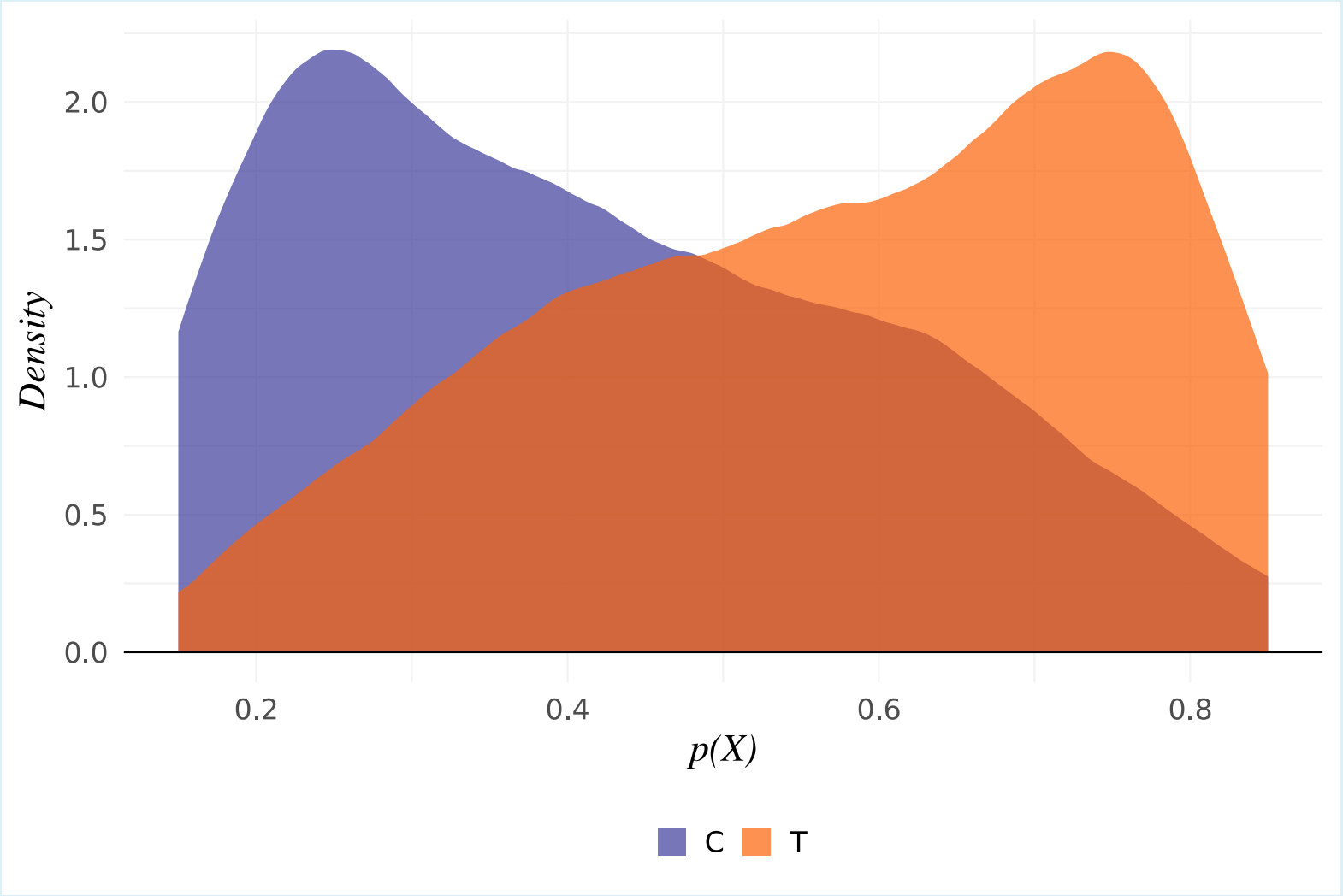
- **Recall**: The **Overlap Assumption**

$$0 < \Pr(\mathrm{D}_i = 1 \mid \mathrm{X}_i) < 1$$

- Which is to ensure that *we can find a matched untreated sample for each treated sample*, or *the distribution of X for the treated and control groups should overlap*.
- In P-score methods, the overlap assumption is about the **distribution of the propensity score** rather than the covariates.
- The easiest way to check the **overlap assumption** is to **plot the distribution of covariates before and after matching**.
  - As we did it in the training example, in which we plotted the distribution of only one covariate.
  - Apparently when X is a vector which can be tough as the dimensions of $\mathbf{X}$ expand.



- **Example**: The plot below shows the distribution of the estimated propensity score $p(\mathbf{X}_i)$ for the treated and control groups.
  - The pscore in the treated group is in $[0.15, 1]$
  - The pscore in the control group is in $[0, 0.85]$



- Trimming samples to overlap in $p(\mathrm{X}_i)$, thus we only keep the samples if $0.15 \leq p(X) \leq 0.85$

Regression and Propensity Scores Reweight

Regression with Propensity Scores

- Based on the **Propensity Score Theorem**, conditional on the propensity score, the treatment is as good as random.
- Then, the simple idea is to use **propensity scores as a control variable** instead of the raw covariates in the regression model

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 p(X_i) + u_i$$

- **Assumption**: the relationship between the outcome and the propensity score is **linear**.
- To consider the **non-linearity**, we can add the **polynomial terms** or **interaction terms** between the propensity score and the treatment to make a more flexible model.

$$\begin{aligned} Y_i = \beta_0 + \beta_1 D_i + \beta_2 p(X_i) + \beta_3 D_i \cdot p(X_i) + \beta_4 p(X_i)^2 + \\ \beta_5 D_i \cdot p(X_i)^2 + \beta_6 p(X_i)^3 + \beta_7 D_i \cdot p(X_i)^3 + \ldots + u_i \end{aligned}$$

- Normally, the cubic term is enough for the flexibility.

Inverse Probability Weighting

- The **inverse probability weighting (IPW)** is an alternative way to use the propensity score to control the bias due to the selection on observables.
  - The idea is to weight the treated and control units by the inverse of the propensity score.
- The Average Treatment Effect (ATE) can be derived by the following formula:

$$\delta_{ATE} = E(Y_{1i} - Y_{0i}) = E\left[\frac{D_i}{p(X_i)}Y_i\right] - E\left[\frac{1-D_i}{1-p(X_i)}Y_i\right]$$

- Under the **CIA** and **Overlap Assumption**, we could show that

$$E[Y_{1i}] = E\left[\frac{D_i}{p(X_i)}Y_i\right]$$
$$E[Y_{0i}] = E\left[\frac{1-D_i}{1-p(X_i)}Y_i\right]$$

Inverse Probability Weighting

$$E\left(\frac{D_i Y_i}{p(\mathbf{X}_i)}\right) = E\left(\frac{D_i Y_{1i}}{p(\mathbf{X}_i)}\right) \because D_i Y_i = D_i Y_{1i} \text{ for all i and } D_i = 1 \text{ and } D_i = 0$$

$$= E\left[E\left(\frac{D_i Y_{1i}}{p(\mathbf{X}_i)}\middle|\mathbf{X}_i\right)\right] \because \text{LIE}$$

$$= E\left[\frac{1}{p(\mathbf{X}_i)} E\left(D_i Y_{1i}\middle|\mathbf{X}_i\right)\right] \because \text{LIE and } p\left(\mathbf{X}_i\right) \text{ is a function of X}$$

$$= E\left[\frac{1}{p(\mathbf{X}_i)} E(D_i|\mathbf{X}_i)\, E(Y_{1i}|\mathbf{X}_i)\right] \because \text{CIA, thus } (\mathrm{Y}_{0i},\ \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i|\mathrm{X}_i$$

$$= E[E(Y_{1i}|\mathbf{X}_i)] \because \text{P-Scores' definition: } p\left(\mathbf{X}_i\right) = E(D_i|\mathbf{X}_i)$$

$$= E(Y_{1i}) \because \text{LIE}$$

- Thus, we have the following result:

$$E[Y_{1i}] = E\left[\frac{D_i}{p(X_i)} Y_i\right]$$

IPW Estimator for ATE

- Similarly, we could show that

$$E[Y_{0i}] = E\left[\frac{1 - D_i}{1 - p(X_i)} Y_i\right]$$

- Then, we could get the ATE by the following

$$\delta_{ATE} = E\left[\frac{D_i}{p(X_i)} Y_i\right] - E\left[\frac{1-D_i}{1-p(X_i)} Y_i\right]$$

$$= \mathrm{E}\left[\frac{(D_i - D_i \cdot p(\mathbf{X}_i) - p(\mathbf{X}_i) + D_i \cdot p(\mathbf{X}_i))}{p(\mathbf{X}_i)(1-p(\mathbf{X}_i))} Y_i\right]$$

$$= \mathrm{E}\left[\frac{(D_i - p(\mathbf{X}_i))}{p(\mathbf{X}_i)(1-p(\mathbf{X}_i))} Y_i\right]$$

- This is the *Horvitz-Thompson* **IPW estimator** for the ATE.

IPW Estimator for ATE

- Then the IPW estimator for ATE is given by

$$\hat{\delta}_{ATE}^{HW} = \frac{1}{N} \sum_{i=1}^{N} \frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))} Y_i$$

- The IPW weights here are the **inverse of the propensity score**.

$$\frac{D_i - \hat{p}(\mathbf{X}_i)}{\hat{p}(\mathbf{X}_i)(1 - \hat{p}(\mathbf{X}_i))} = \begin{cases} \frac{1}{\hat{p}(\mathbf{X}_i)} & \text{if } D_i = 1 \\ -\frac{1}{1-\hat{p}(\mathbf{X}_i)} & \text{if } D_i = 0 \end{cases}$$

- HW weights do not necessarily average to 1,which can be a problem.

A more general IPW Estimator for ATE

- The **standardization** means dividing each group's weights by the sum of all weights within that group.
- A more general IPW estimand is given by

$$\delta_{ATE} = E(Y_{1i} - Y_{0i}) = \frac{E\left[\frac{D_i}{p(X_i)} Y_i\right]}{E\left[\frac{D_i}{p(X_i)}\right]} - \frac{E\left[\frac{1-D_i}{1-p(X_i)} Y_i\right]}{E\left[\frac{1-D_i}{1-p(X_i)}\right]}$$

  - Where $E(\frac{D_i}{p(X_i)})$ can be seen as the **average weight** for the treated group and $E(\frac{1-D_i}{1-p(X_i)})$ can be seen as the **average weight** for the control group.

- Then corresponding IPW estimator for ATE is given by

$$\hat{\delta}_{ATE}^{IPW} = \left[\sum_{i=1}^{N} \frac{Y_i D_i}{p(X_i)}\right] / \left[\sum_{i=1}^{N} \frac{D_i}{p(X_i)}\right] - \left[\sum_{i=1}^{N} \frac{Y_i (1 - D_i)}{(1 - p(X_i))}\right] / \left[\sum_{i=1}^{N} \frac{(1 - D_i)}{(1 - p(X_i))}\right]$$

Practical Implications

- IPW provides a way to estimate causal effects without explicitly modeling the outcome process like matching.
- **Practical challenges**:
  - **Extreme weights** when $p(X)$ is close to 0 or 1
  - Need for careful diagnostics (covariate balance, weight distribution)
- Some Extensions: Combine IPW with outcome regression
  - **Double Robustness method**:
  - **Consistent if either the propensity score is correctly specified or the outcome regression is correctly specified**.

Matching in practice

Matching in Practice

Introduction

- Although matching is a simple concept, it can be more difficult to implement in practice.
- There are many decisions to make when matching units. The questions are as follows:

  1. How to choose variables as the matching covariates?
  2. Which matching methods should be used? **distances** and **weights**: Matching/Propensity Score Matching
  3. How many control units should be matched to each treatment unit?: **one-to-one** or **many-to-one**?
  4. The sample is matched **with or without replacement**?
  5. The order of matching: **greedy** or **optimal**?

Matching in Practice

Choosing Variables

- **Question**: Which variables should be used for matching treatment and control units?
- **Answer**: Include all variables that are likely to be **confounders**. (*Recall the "good and bad controls" framework*)
    - Irrelevant variables
    - Relevant variables
    - Omitted variables
    - Colliders
    - Confounders
- Selecting matching covariates follows similar principles as in regression analysis.
- As with OLS regression, comparing results across different sets of variables serves as a **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.
- **Shortcomings of matching with replacement**:
  - reduces the effective sample size when the same control units are used multiple times.
  - may lead to estimates being overly influenced by a small number of frequently used control units.

Matching in practice

Greedy v.s Optimal Matching

- The **greedy matching** is a simple and fast algorithm that matches each treated unit to the control unit with the closest distance.
- However, the closest control units for every single sample may not be the best match for the treated unit as a whole.
    - Thus the **local** optimal solution may not be the **global** optimal solution.
- The **optimal matching** is a more complex algorithm that finds the best possible match for each treated unit simultaneously.
- It is often computationally expensive because it have to consider all possible matches for all treated units.

Matching in practice

1:1 v.s 1:m Matching

- **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.
- Normally, matching procedures need a relatively large number of samples to be able to find a good match.

Matching in practice

In a Summary

- 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, matching with vs. without replacement.
- 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.

Wrap up

- 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.
- Most importantly, **matching is essentially as the same as regression**, only different in the weight of estimating the CEF function.
- **Question**: Why we still need matching?
- **Answer**: Matching is over regression in the following aspects:
    1. Due to its non-parametric characteristics, matching does not impose any restrictions on empirical specification or estimate specific parameters of the CEF function.
    2. Regression does not account for the **common support** issue **explicitly**, while matching does.
- In practice, using matching alone as main identification strategy is **less common** in economics, more frequently combined with other methods like DID and SCM, which we will discuss later on.

Appendix: Matching vs Regression

Matching in essential is Regression

- Although matching is a non-parametric or semi-parametric method, it is essentially as the same as regression.
- Suppose the ATT, thus the average treatment effect on the treated, is the parameter of interest.

$$ATT = \delta_{ATT} = E[Y_{i1} - Y_{i0}|D = 1]$$

- Under the CIA, if we can control/balance some covariates X_i, then we have the selection bias equals to zero, thus

$$E[Y_{0i}|X_i, D = 1] = E[Y_{i0}|X_i, D = 0]$$

Matching in essential is Regression

- Using the **LIE** and **CIA**,

$$\delta_{ATT} = E\Big[(E[Y_{i1}|X_i, D=1] - E[Y_{i0}|X_i, D=1])|D_i = 1\Big]$$

$$= E\Big[(E[Y_{i1}|X_i, D=1] - E[Y_{i0}|X_i, D=0])|D_i = 1\Big]$$

$$= E\Big[(E[Y_i|X_i, D=1] - E[Y_i|X_i, D=0])|D_i = 1\Big]$$

$$= E\Big[\delta_X|D_i = 1\Big]$$

- Where δ_X is the average outcomes gaps between two groups within observed covariates X_i.

$$\delta_X = E[Y_i|X_i, D=1] - E[Y_i|X_i, D=0]$$

Matching in essential is a regression

- Recall: the Bayes' rule

$$P\left(\mathrm{X}_i = x \mid \mathrm{D}_i = 1\right) = \frac{P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right) \cdot P\left(\mathrm{X}_i = x\right)}{P\left(\mathrm{D}_i = 1\right)}$$

- Then if X_i is discrete, then the matching estimator can be written as

$$\delta_M = E\big[\delta_X | D_i = 1\big] = \Sigma_x \delta_X P(X_i = x | D_i = 1)$$

$$= \sum_x \delta_x \frac{P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right) P\left(\mathrm{X}_i = x\right)}{P\left(\mathrm{D}_i = 1\right)}$$

$$= \sum_x \delta_x \left[\frac{P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right) P\left(\mathrm{X}_i = x\right)}{\sum_x P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right) P\left(\mathrm{X}_i = x\right)} \right]$$

- Where the $P(X_i = x | D_i = 1)$ is the probability mass function for X_i given the treatment status $D_i = 1$.

Regression in essential is a matching

- Suppose we have a saturated regression model

$$Y_i = \sum_x \mathbb{I}(X_i = x)\beta_x + \delta_R D_i + u_i$$

- The $\mathbb{I}(X_i = x)$ is the indicator function for $X_i = x$,which means that

$$\mathbb{I}(X_i = x) = \{ 1 \quad \text{if } X_i = x \; 0 \quad \text{otherwise}$$

- D is the treatment status, u_i is the error term.
- The β_x is the coefficient of $X_i = x$ in the regression model.
- The δ_R is the regression estimator of the treatment effect.
- **Note**: Saturating X means allowing a separate intercept for each unique value of X_i, which makes the conditional expectation function $E[Y|X]$ to be linear in X.

Regression in essential is a matching

- We can prove that the regression estimator δ_R can be expressed as follows†

$$\delta_R = E\left[\frac{Var(D_i|X_i)}{E\left[Var(D_i|X_i)\right]}\delta_X\right]$$

- Where $Var(D_i|X_i)$ is the conditional variance of D_i given X_i, thus

$$Var(D_i|X_i) = E\left[(D_i - E[D_i|X_i])^2 \mid X_i\right]$$

- Let $p(x) = P(D_i = 1|X_i = x)$, then

$$Var(D_i|X_i) = p(x)(1-p(x)) = P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right)\left(1 - P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right)\right)$$

†The detailed proof is somewhat complex, so it's placed in the appendix for those interested. You can also refer to the **Mostly Harmless Econometrics** (MHE) textbook for the proof (pp54-55).

Regression in essential is a matching

- Then the regression estimator δ_R

$$\delta_R = E\left[\frac{Var(D_i|X_i)}{E\left[Var(D_i|X_i)\right]}\delta_X\right]$$

$$= \frac{E[Var(D_i|X_i)\delta_X]}{E\left[Var(D_i|X_i)\right]}$$

$$= \frac{\sum_x \delta_X \cdot Var(D_i|X_i) \cdot P(X_i = x)}{\sum_x \cdot Var(D_i|X_i) \cdot P(X_i = x)}$$

$$= \sum_x \delta_X \left[\frac{P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right)\left(1 - P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right)\right)\ P\left(\mathrm{X}_i = x\right)}{\sum_x P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right)\left(1 - P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right)\right)\ P\left(\mathrm{X}_i = x\right)}\right]$$

Matching vs. Regression in essential

- The **matching estimator** δ_M

$$\delta_M = \sum_x \delta_X \left[\frac{P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right) P\left(\mathrm{X}_i = x\right)}{\sum_x P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right) P\left(\mathrm{X}_i = x\right)} \right]$$

- The **regression estimator** δ_R

$$\delta_R = \sum_x \delta_X \left[\frac{P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right) \left(1 - P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right)\right) \; P\left(\mathrm{X}_i = x\right)}{\sum_x P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right) \left(1 - P\left(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x\right)\right) \; P\left(\mathrm{X}_i = x\right)} \right]$$

- The difference between the two estimators is **the weight** of the treatment effect δ_X for each unique value of X_i.
  - The matching estimator δ_M uses the weight $P(D_i = 1|X_i = x)$, which is larger for more treated samples.
  - The regression estimator δ_R uses the weight $P(D_i = 1|X_i = x)(1 - P(D_i = 1|X_i = x))$, which is the largest when $P(D_i = 1|X_i = x) = 0.5$, thus half treated and half untreated observations.

Appendix

- Then the regression model of D on X is

$$D_i = \sum_x \mathbb{I}(X_i = x)\gamma_x + v_i$$

  - The $\mathbb{I}(X_i = x)$ is still the indicator function for $X_i = x$
  - The γ_x is the coefficient of $X_i = x$ and v_i is the error term.

- Then the population regression function (PRF) in terms of conditional expectation function (CEF) as

$$E[D|X] = \sum_x \mathbb{I}(X_i = x)\gamma_x = \gamma$$

- Then the residuals of the regression model of D on X,$\tilde{\mathrm{D}}_i$, is

$$\tilde{D}_i = D - E[D|X_i = x]$$

Proof of the regression estimator δ_R

- Because our regression model is saturated as follows

$$Y_i = \sum_x \mathbb{I}(X_i = x)\beta_x + \delta_R D_i + u_i$$

- Then the key coefficient of interest is δ_R. Based on FWL theorem, we have

$$\delta_R = \frac{\mathrm{Cov}\left(\tilde{\mathrm{D}}_i, \mathrm{Y}_i\right)}{V\left(\tilde{\mathrm{D}}_i\right)}$$

Proof of the regression estimator δ_R

$$\delta_R = \frac{\operatorname{Cov}\left(\tilde{\mathrm{D}}_i, \mathrm{Y}_i\right)}{V\left(\tilde{\mathrm{D}}_i\right)}$$

$$= \frac{E[\tilde{D}_i, Y_i]}{\tilde{D}_i^2} \quad \because \operatorname{Cov}\left(\tilde{\mathrm{D}}_i, \mathrm{Y}_i\right) = E[\tilde{D}_i, Y_i] - E[\tilde{D}_i]E[\mathrm{Y}_i]$$

$$= \frac{E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)\mathrm{Y}_i\right]}{E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)^2\right]} \quad \because \tilde{D}_i = D - E[D|X_i = x]$$

$$= \frac{E\left\{\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)E\left[\mathrm{Y}_i \mid \mathrm{D}_i, \mathrm{X}_i\right]\right\}}{E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)^2\right]} \because ILE$$

Proof of the regression estimator δ_R

- Because

$$E\left[\mathrm{Y}_i \mid \mathrm{D}_i, \mathrm{X}_i\right] = E\left[\mathrm{Y}_i \mid \mathrm{D}_i = 0, \mathrm{X}_i\right] + \delta_X \mathrm{D}_i$$

- Then the numerator of δ_R is

$$\begin{aligned}
&E\left\{\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right) E\left[\mathrm{Y}_i \mid \mathrm{D}_i, \mathrm{X}_i\right]\right\} \\
&= E\left\{\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right) E\left[\mathrm{Y}_i \mid \mathrm{D}_i = 0, \mathrm{X}_i\right]\right\} + E\left\{\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right) \mathrm{D}_i \delta_X\right\} \\
&= 0 + E\left\{\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right) \mathrm{D}_i \delta_X\right\} \because Cov(\tilde{D}_i, X) = 0 \\
&= E\left\{\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)^2 \delta_X\right\} \because Cov(\tilde{D}_i, D) = \tilde{D}_i^2
\end{aligned}$$

Proof of the regression estimator δ_R

$$\delta_R = \frac{E\left\{\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right) E\left[\mathrm{Y}_i \mid \mathrm{D}_i, \mathrm{X}_i\right]\right\}}{E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)^2\right]}$$

$$= \frac{E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)^2 \delta_X\right]}{E\left[\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)^2 \mid X_i\right]\right]}$$

$$= \frac{E\left[\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)^2 \mid X_i\right] \delta_X\right]}{E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)^2\right]}$$

$$= \frac{E\left[Var^2\left(\mathrm{X}_i \mid D_i\right) \delta_X\right]}{E\left[Var^2\left(\mathrm{X}_i \mid D_i\right)\right]}$$

- Where $Var^2\left(\mathrm{X}_i \mid D_i\right) = E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i \mid \mathrm{X}_i\right]\right)^2 \mid X_i\right]$, thus the conditional variance of D given X_i.