

# STA 235 - Causal Inference: Introduction to Observational Studies

Spring 2021

McCombs School of Business, UT Austin

# Last class

- **Randomized controlled trials**
  - Why is it considered the gold standard?
  - How to analyze an RCT in practice?
  - Assumptions and limitations.
- **Statistical power:**
  - How do sample sizes play a role?
  - Randomized inference.



# Today



- **Introduction to Observational Studies:**
  - Can we identify causal effects without RCTs?
  - Assumptions
  - Matching vs OLS

No more chance[s]

# Introduction to observational studies

- Most times, we will not be able to randomize, and we need to work with **existing data**

## Observational data

- Data for which we can't manipulate the treatment assignment, e.g. data in its "natural state".

**Can we reasonably assume that the ignorability assumption holds?**

# Introduction to observational studies (cont.)



- Moving away from the core assumption of RCTs: that **"the probability of treatment assignment is a known function"** (Imbens & Rubin, 2015).

# Introduction to observational studies (cont.)



- Moving away from the core assumption of RCTs: that **"the probability of treatment assignment is a known function"** (Imbens & Rubin, 2015).
- We will maintain the assumption of **unconfoundedness** (to a certain extent).

What is that?

# Calling in the CIA

- **Unconfoundedness** means that the treatment assignment is independent from the potential outcomes.
- If you recall, the ignorability assumption assumes that:

$$Y(0), Y(1) \perp\!\!\!\perp Z$$

- What if you could assume that this holds **conditional on some covariates**?

**Conditional Independence Assumption (CIA)**

$$Y(0), Y(1) \perp\!\!\!\perp Z|X$$



# The assignment mechanism

- **Key component** in causal analysis:
  - In RCTs, **assignment mechanism** is *known*.
  - But in **observational studies**?



# The assignment mechanism (cont.)

- For now, we will consider 3 assumptions:

**Individualistic assignment**

**Probabilistic assignment**

**Unconfounded assignment**

# Individualistic assignment

- Limits the dependence of treatment assignment for unit  $i$  on the outcomes and assignments for other units.
- For some function  $q(\cdot) \in [0, 1]$ :

$$p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = q(X_i, Y_i(0), Y_i(1))$$

From last class, can you think of an example where this doesn't hold?

# Probabilistic assignment

- Nonzero probability for each treatment value, for each unit.

$$0 < p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) < 1 \quad \forall i = 1, \dots, N$$

,for each possible  $\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)$

**When could this assumption fail?**

# Unconfounded assignment

- The assignment mechanism does not depend on the potential outcomes:

$$Pr(\mathbf{Z}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = Pr(\mathbf{Z}|\mathbf{X}, \mathbf{Y}'(0), \mathbf{Y}'(1))$$

,for all possible  $\mathbf{Z}$ ,  $\mathbf{X}$ ,  $\mathbf{Y}(0)$ ,  $\mathbf{Y}(1)$ ,  $\mathbf{Y}'(0)$ , and  $\mathbf{Y}'(1)$

# Selection on observables

- Units select into treatment based on characteristics **I can observe**.



$$\begin{aligned} &(Y_1(0), Y_1(1)) \\ &Z = 1 \\ &Y = y_1 \end{aligned}$$



$$\begin{aligned} &(Y_2(0), Y_2(1)) \\ &Z = 0 \\ &Y = y_2 \end{aligned}$$

# Selection on observables

- Units select into treatment based on characteristics **I can observe**.

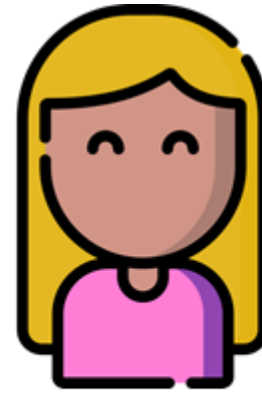


$$(Y_1(0), Y_1(1))$$

$$Z = 1$$

$$Y = y_1$$

$$\mathbf{X} = \mathbf{X}_F$$



$$(Y_2(0), Y_2(1))$$

$$Z = 0$$

$$Y = y_2$$

$$\mathbf{X} = \mathbf{X}_F$$

# How do we adjust for observables?

- One way we have seen so far is **regression adjustment**

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \varepsilon_i$$

**Under unconfoundedness, how would we interpret  $\beta_1$ ?**



# How do we adjust for observables?

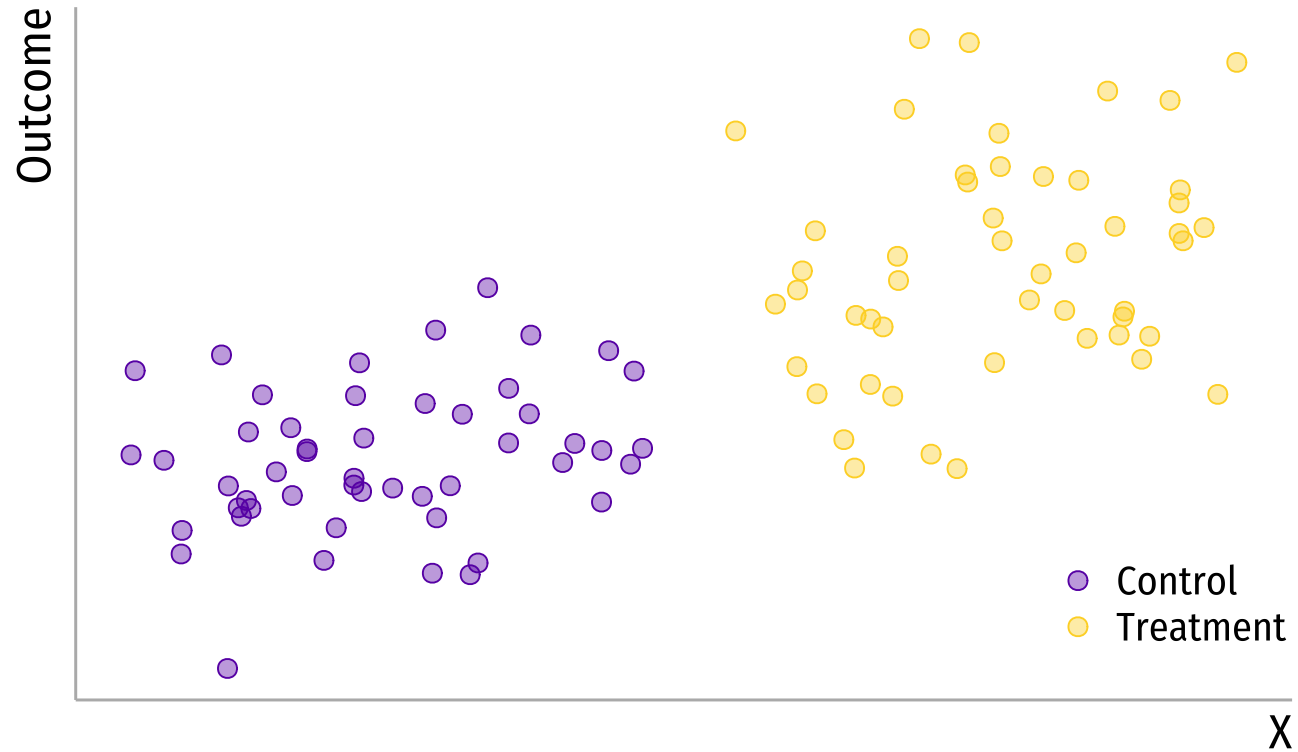
- One way we have seen so far is **regression adjustment**

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \varepsilon_i$$

**$\beta_1$  is the effect of Z on Y, holding X constant**

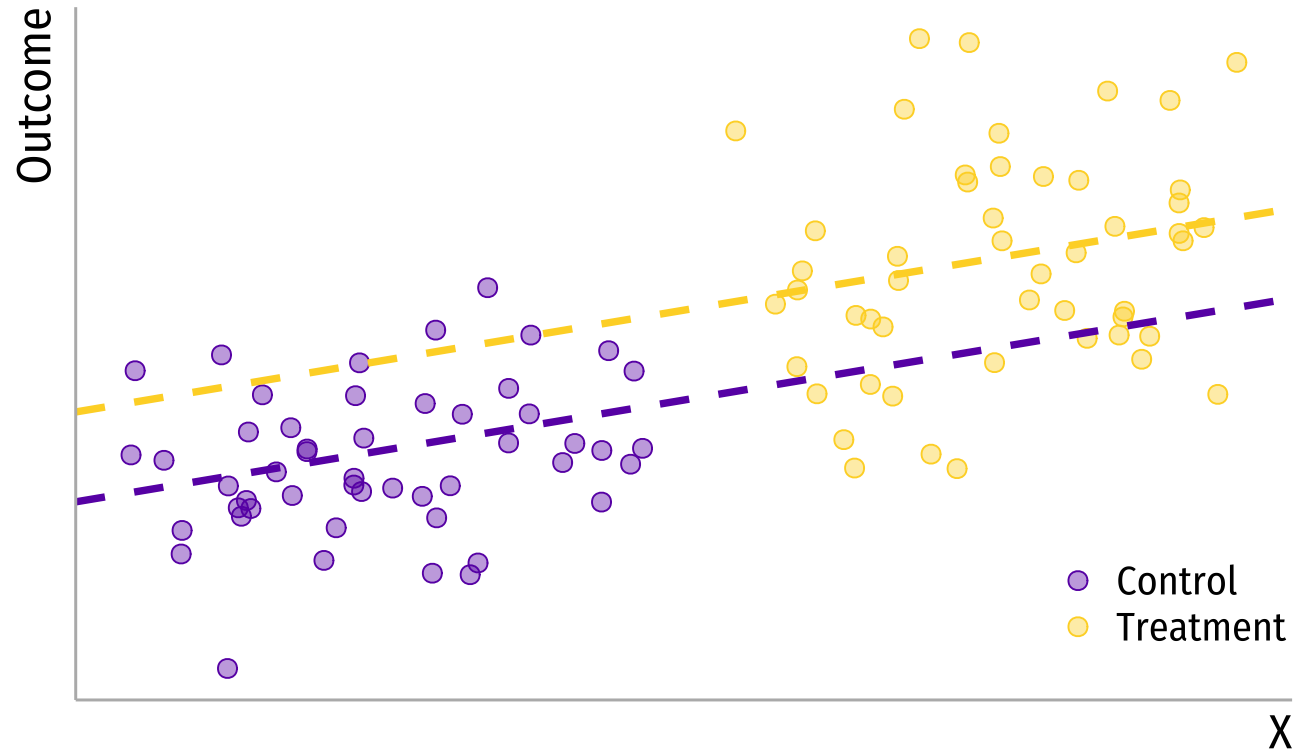
# How do we adjust for observables?

- But what if our data looks like this? Do you see a problem?



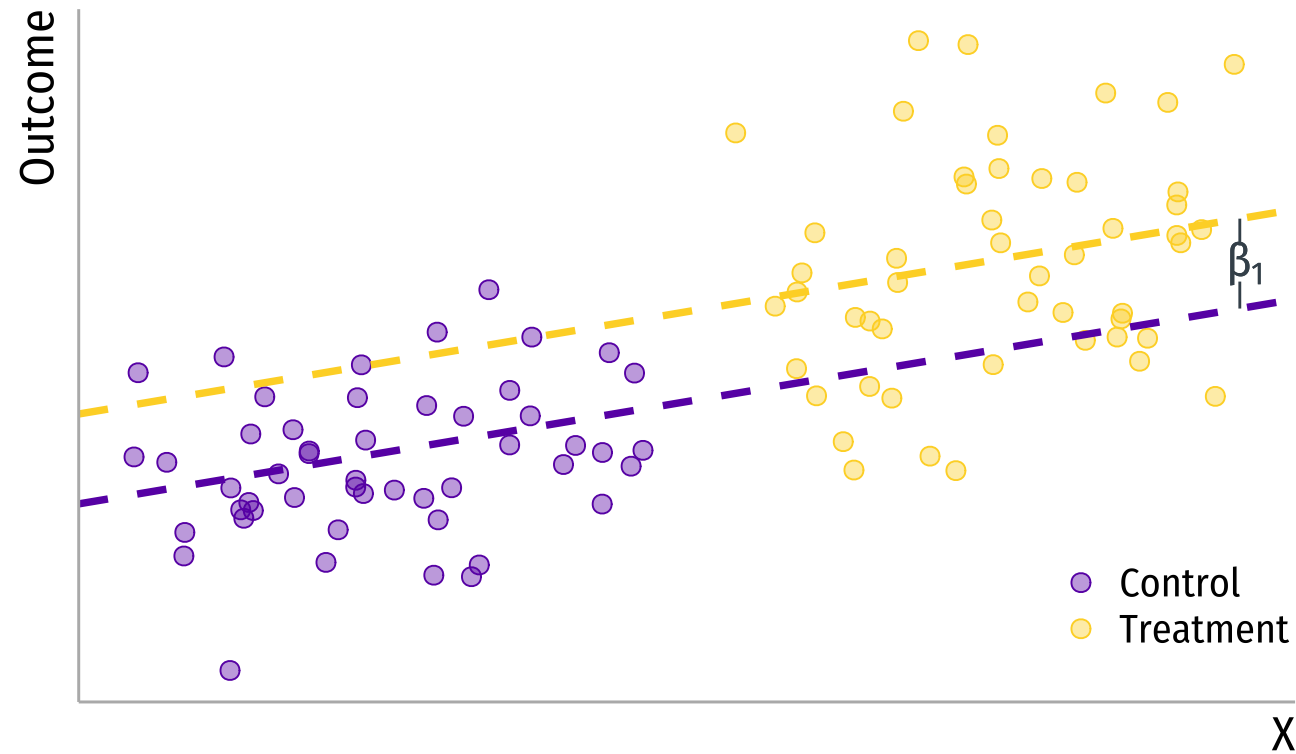
# How do we adjust for observables?

- But what if our data looks like this? Do you see a problem?



# How do we adjust for observables?

- But what if our data looks like this? Do you see a problem?



Finding your perfect match...

# Two peas in a pod

- One other route we could take is to **find similar units** in our sample and **group them together**.
- There are different ways to do it:
  - E.g. subclassification, matching.



# Two peas in a pod

- One other route we could take is to **find similar units** in our sample and **group them together**.
- There are different ways to do it:
  - E.g. subclassification, matching.

What do we gain?



# Advantages of matching methods

**Reduce model dependence**

Imbalance → model dependence → researcher discretion → bias

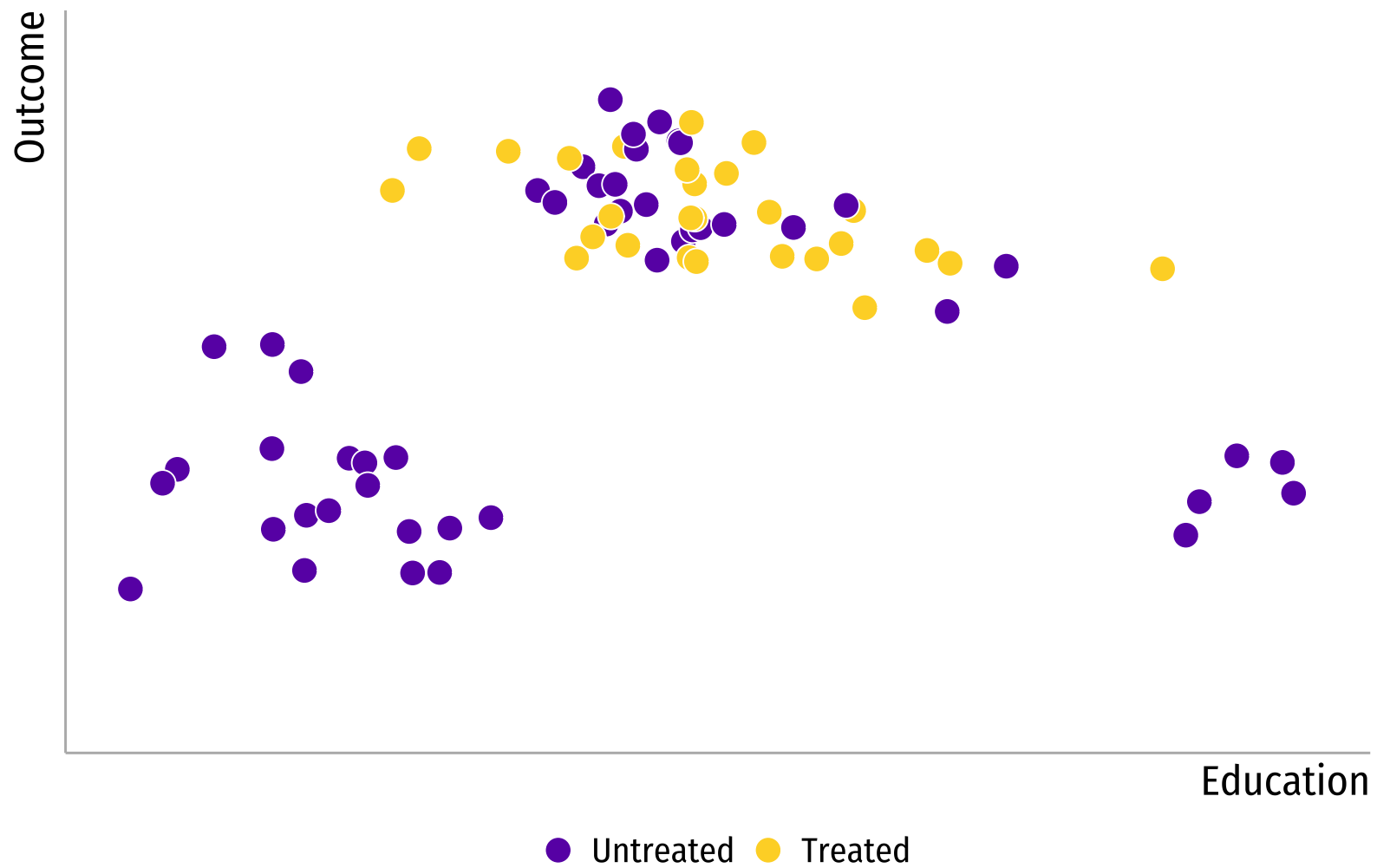
**Compare like to like**

**No extrapolation!**

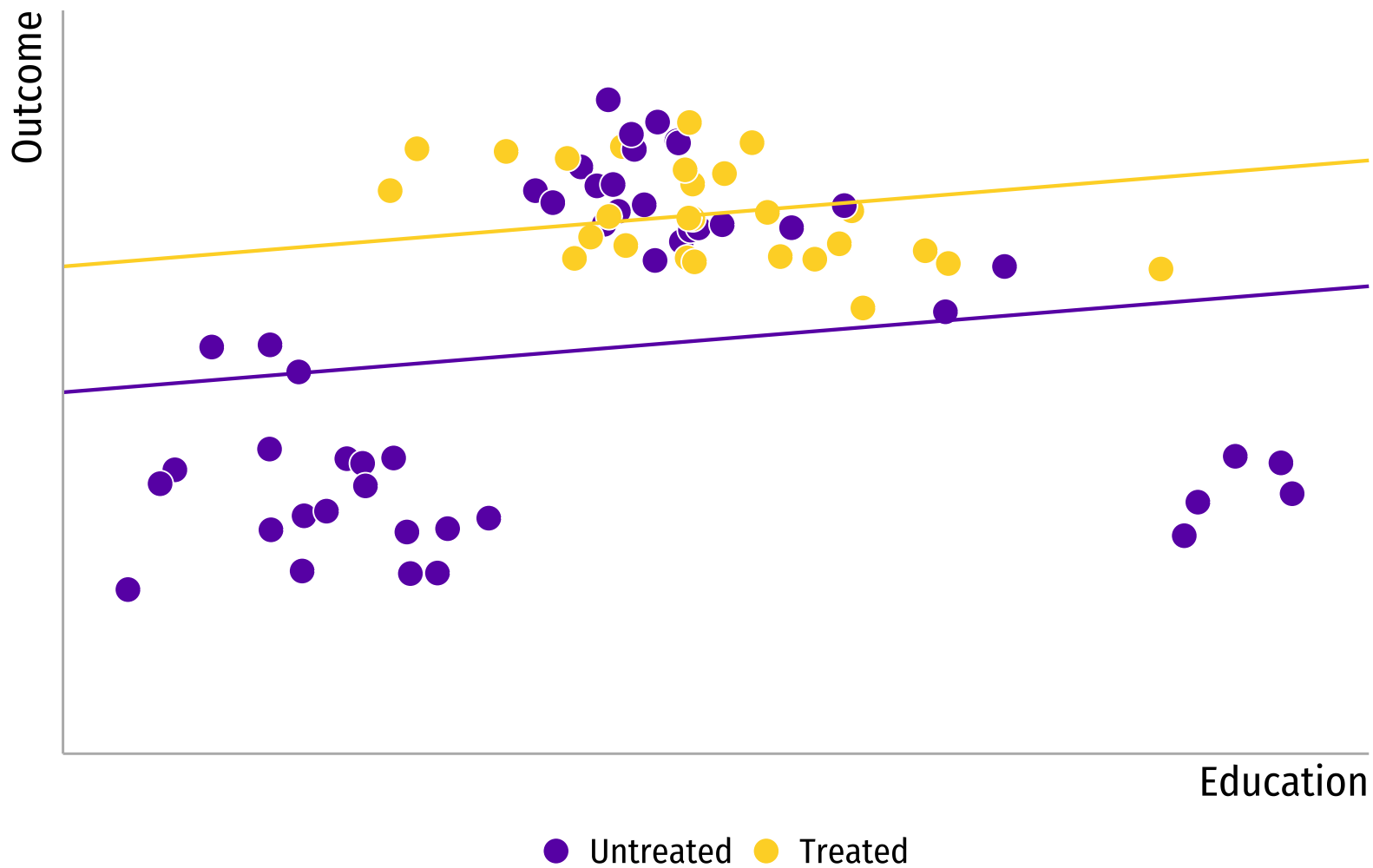
**Can adjust closely by covariates**

Exact matching, coarsened exact matching, fine balance..

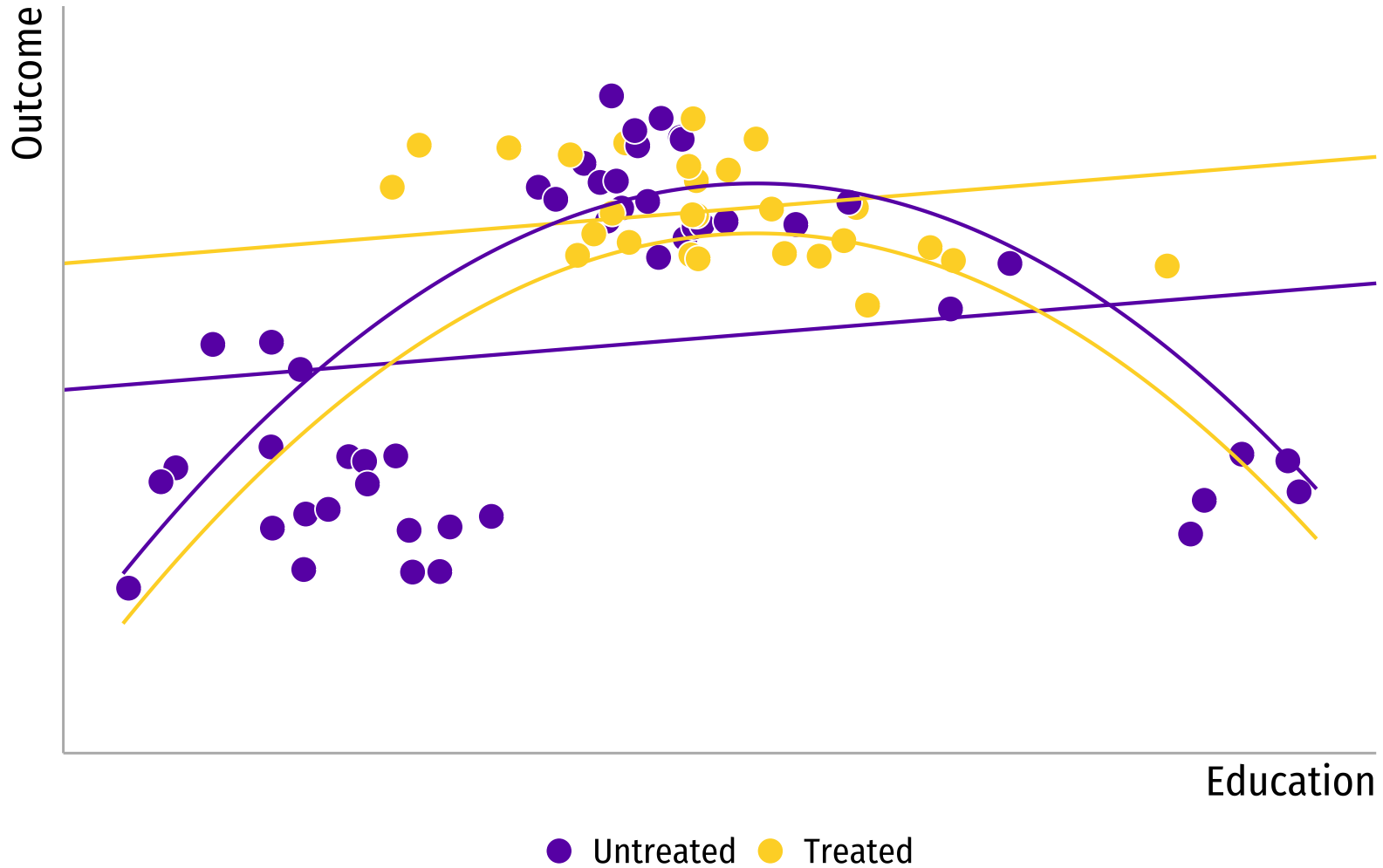


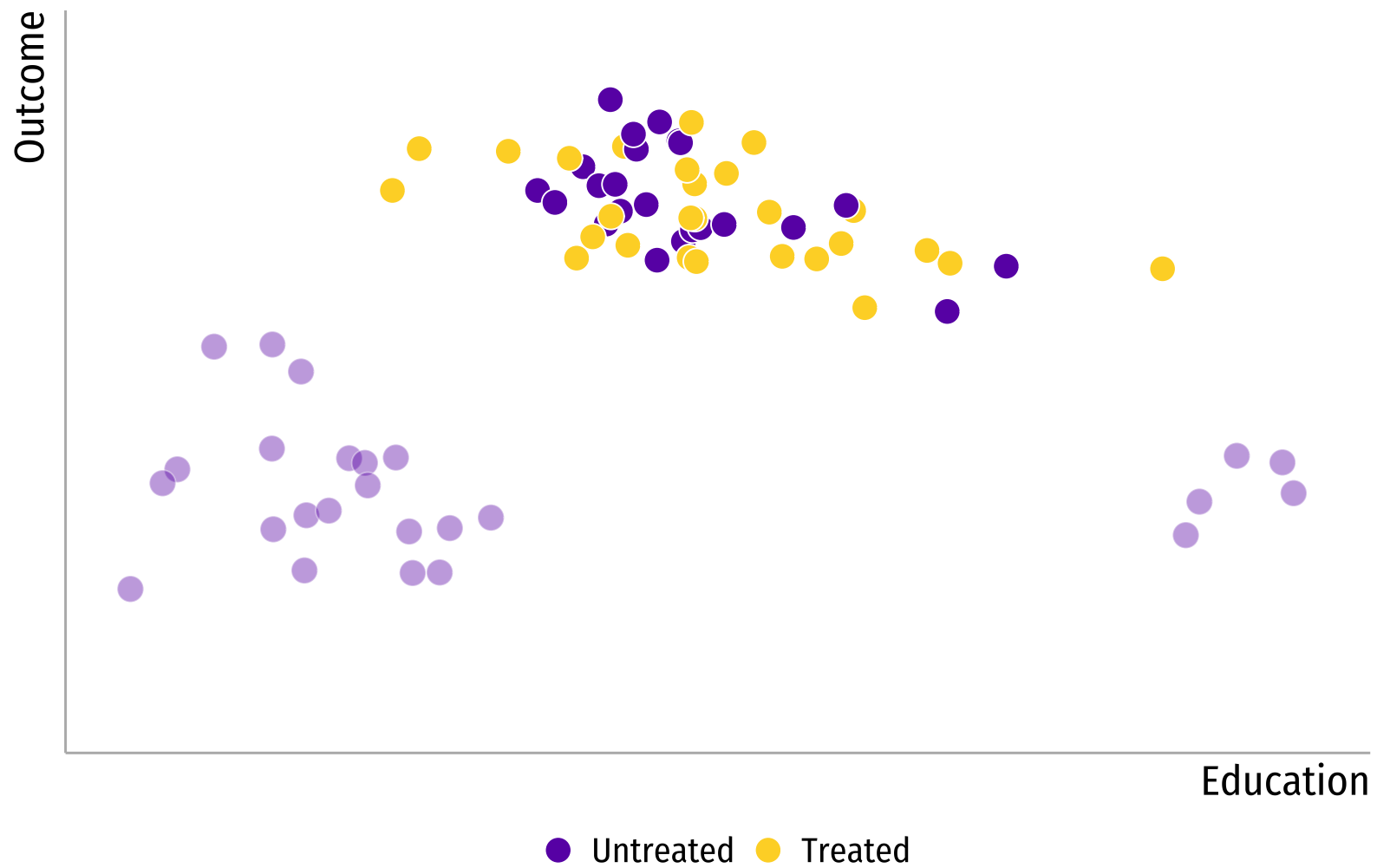


$$\text{Outcome} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Treatment}$$

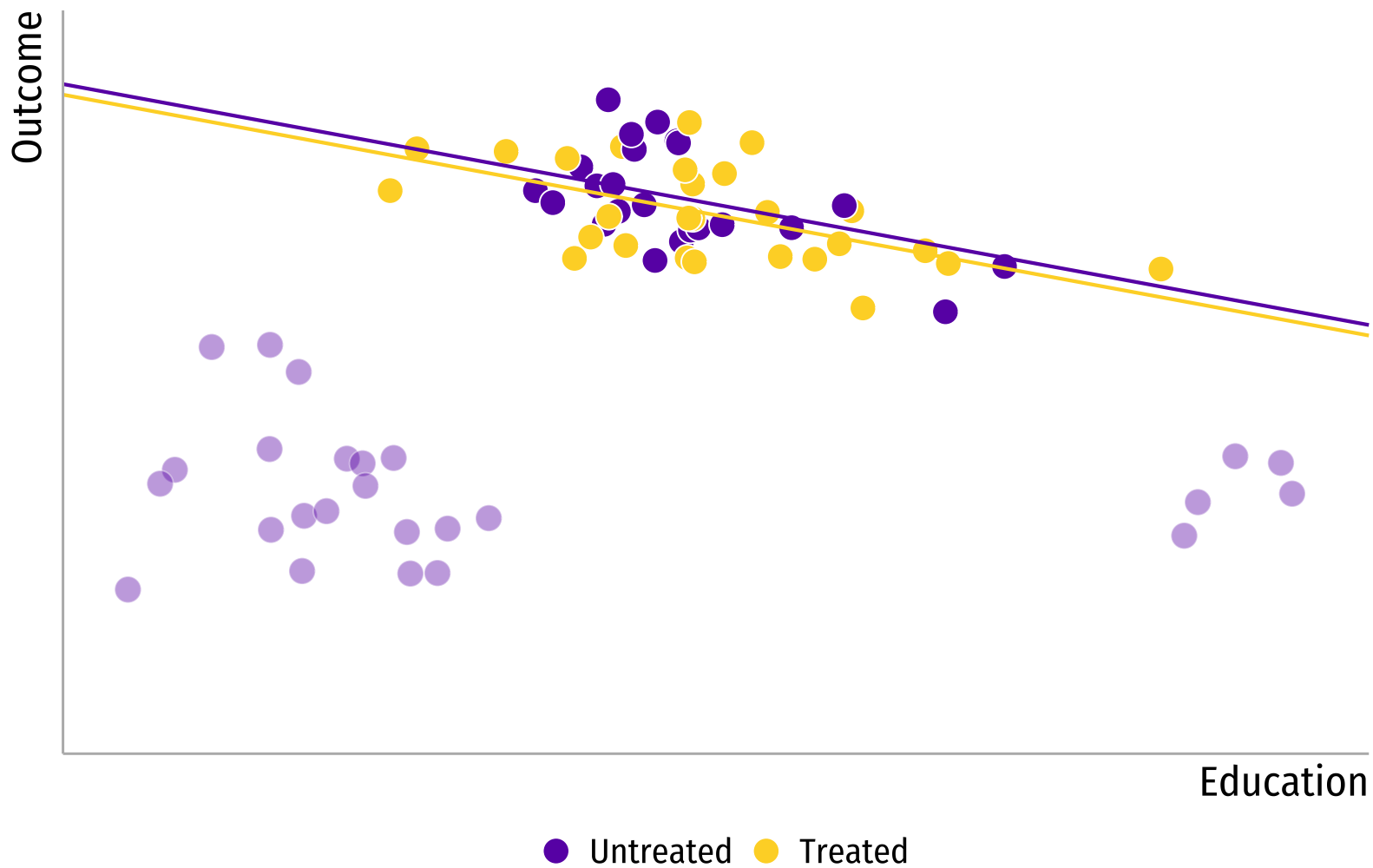


$$\text{Outcome} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Education}^2 + \beta_3 \text{Treatment}$$

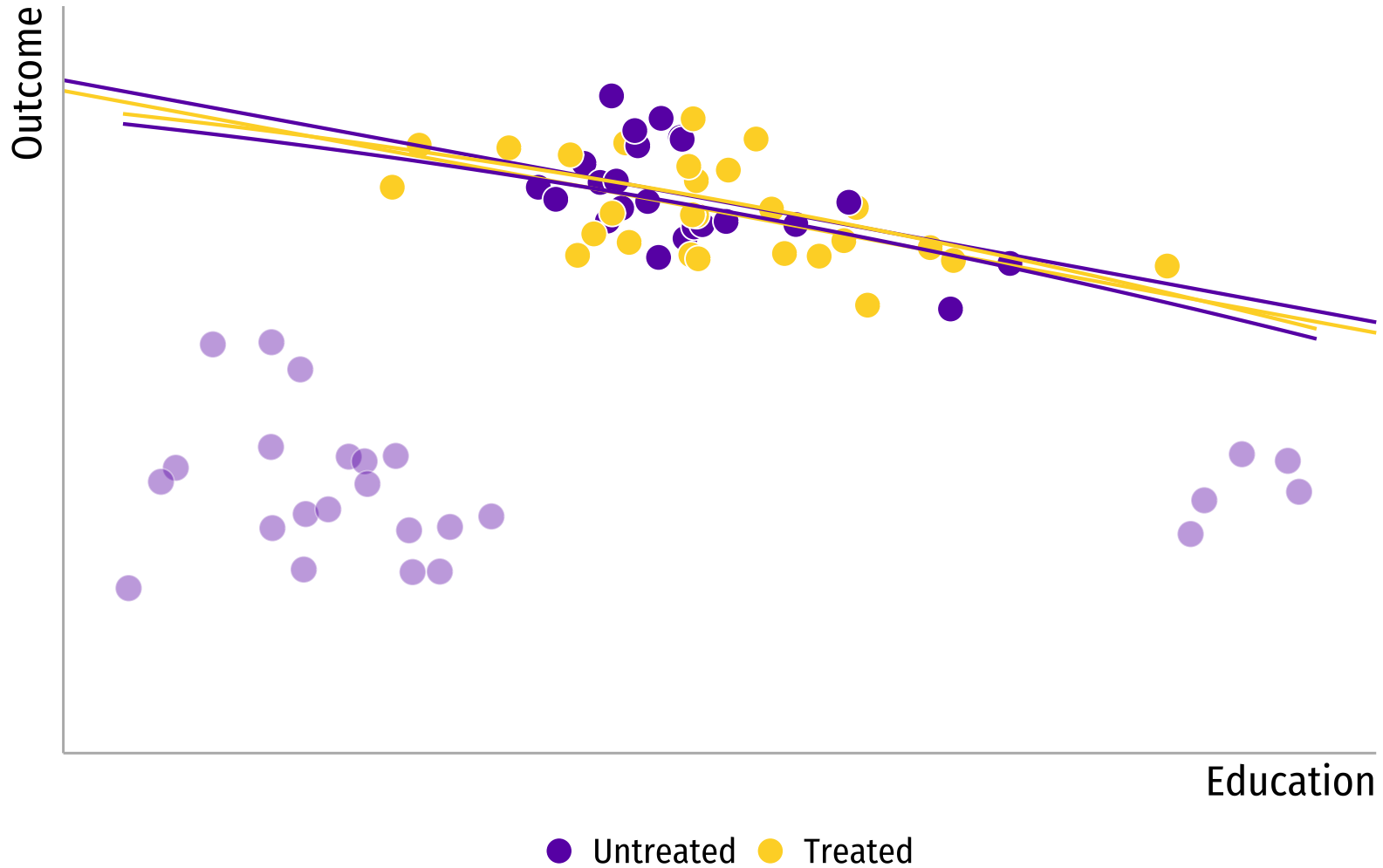




$$\text{Outcome} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Treatment}$$



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

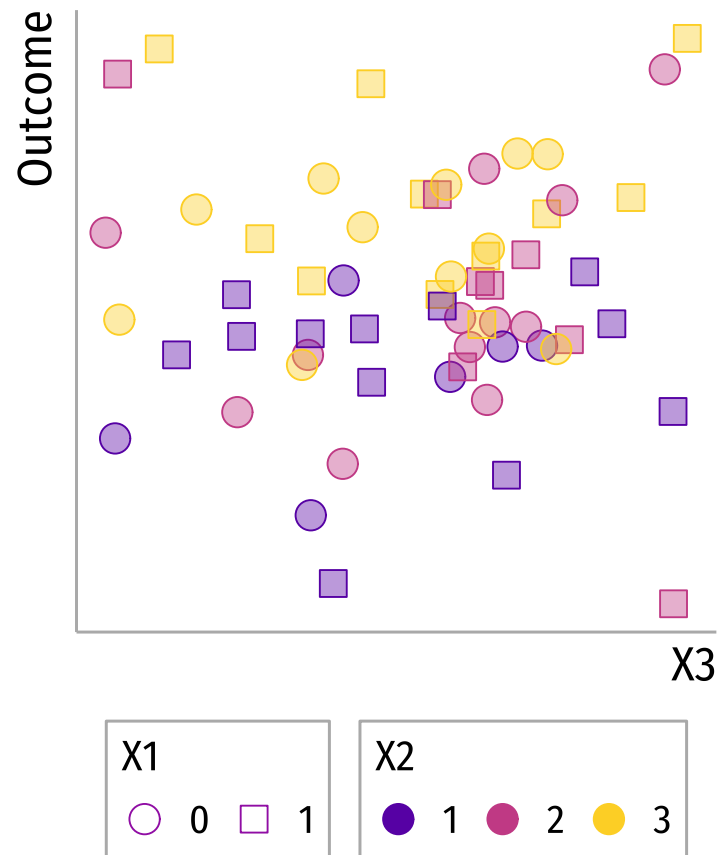


## How do we know we can remove those observations?



# Subclassification

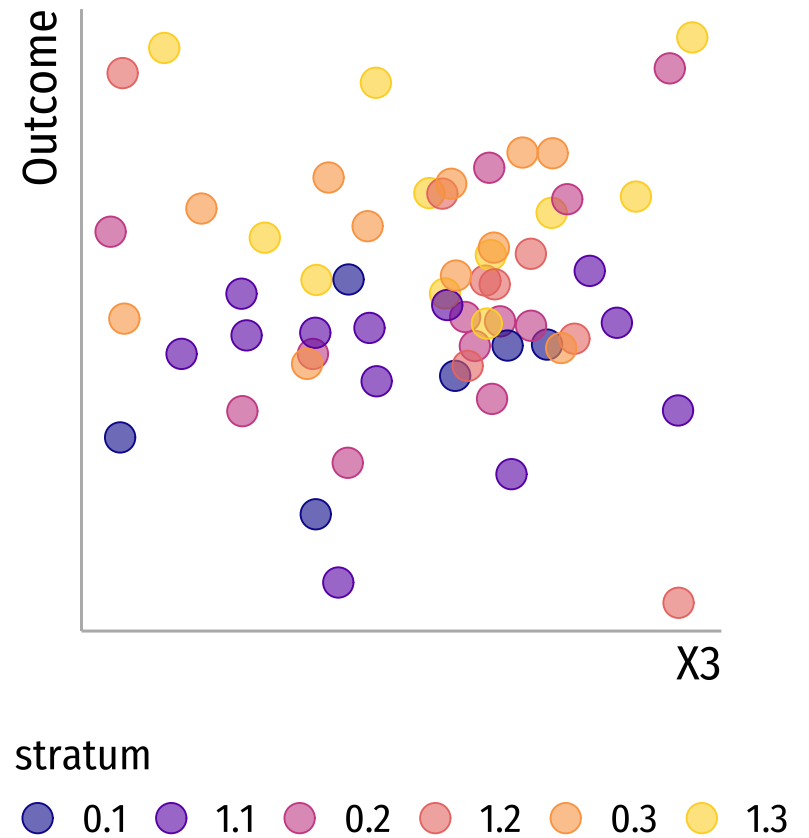
- Very similar to **stratifying**.





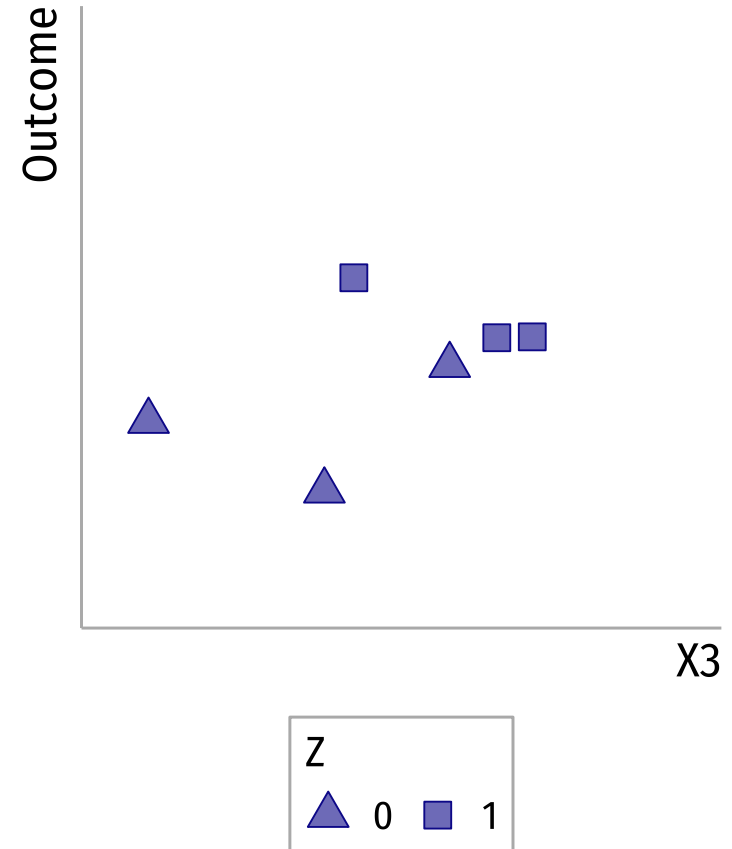
# Subclassification

- Very similar to **stratifying**.
- Build **combination** of X1 and X2 (strata).



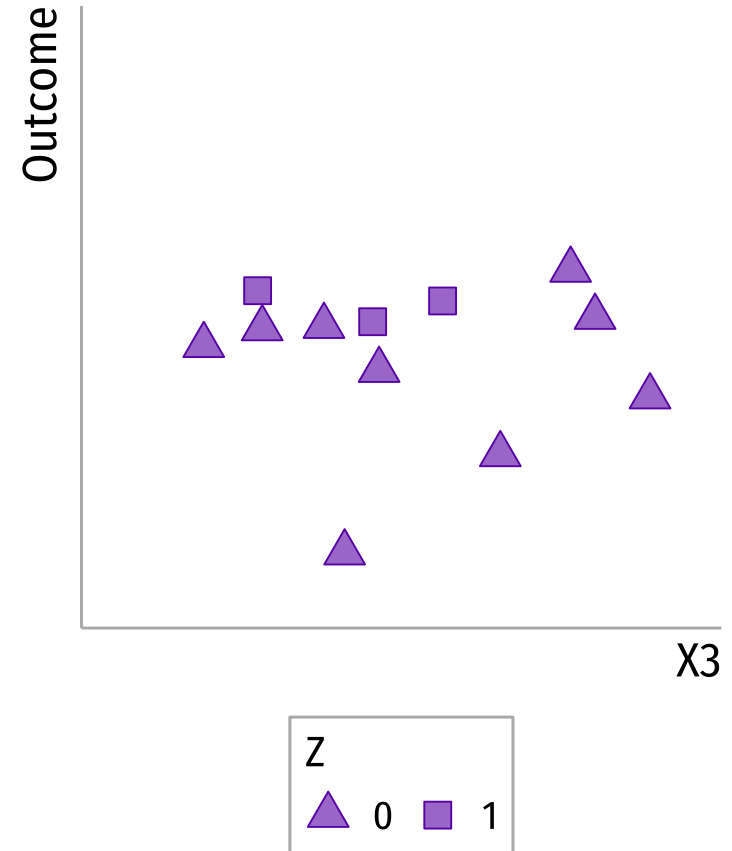
# Subclassification

- Very similar to **stratifying**.
- Build **combination** of X1 and X2 (strata).
- Compare **within stratum**.



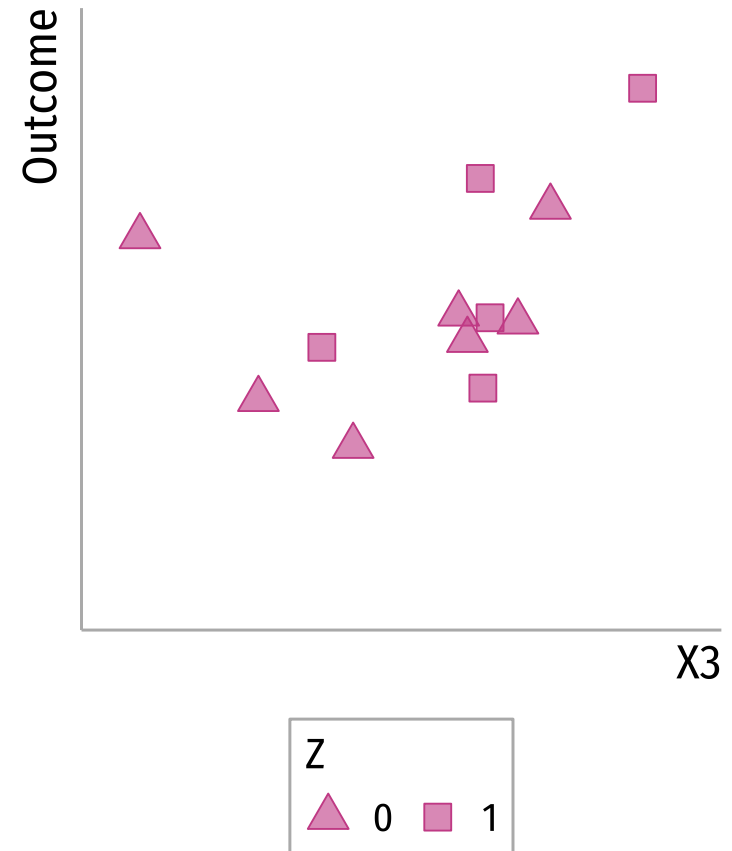
# Subclassification

- Very similar to **stratifying**.
- Build **combination** of X1 and X2 (strata).
- Compare **within stratum**.



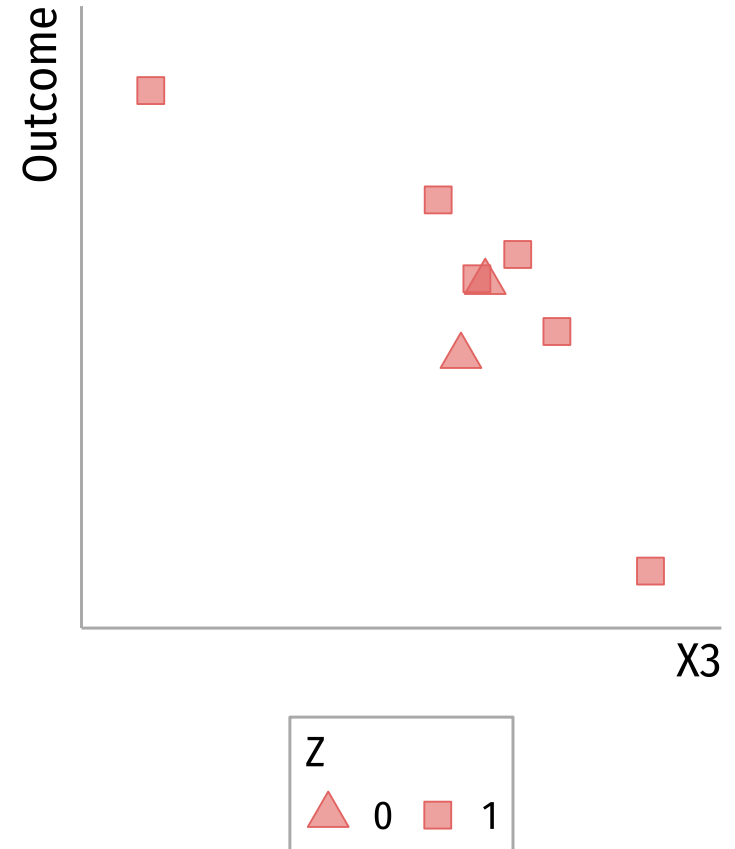
# Subclassification

- Very similar to **stratifying**.
- Build **combination** of X1 and X2 (strata).
- Compare **within stratum**.



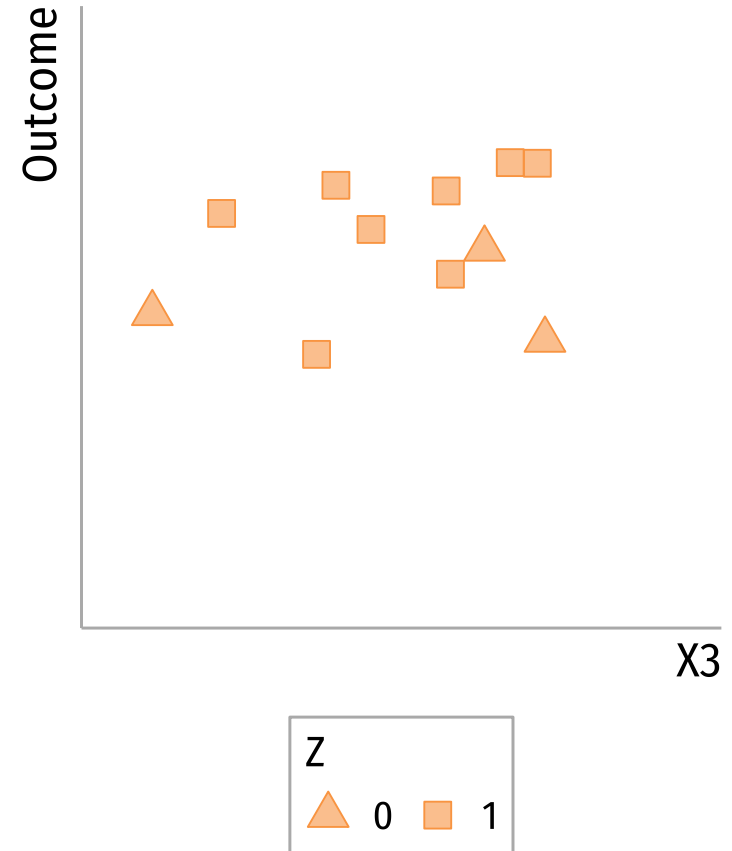
# Subclassification

- Very similar to **stratifying**.
- Build **combination** of X1 and X2 (strata).
- Compare **within stratum**.



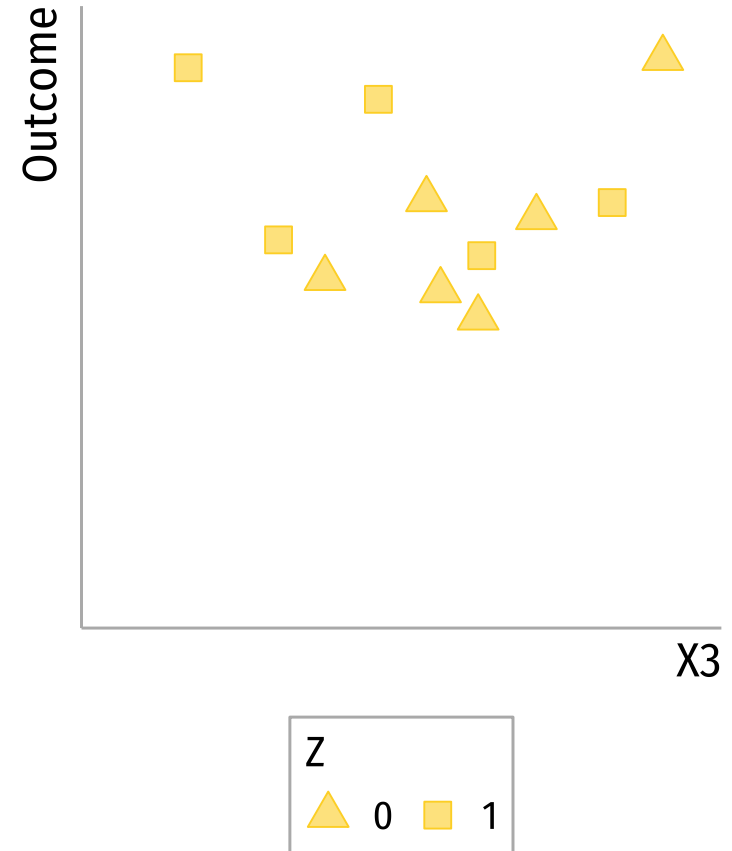
# Subclassification

- Very similar to **stratifying**.
- Build **combination** of X1 and X2 (strata).
- Compare **within stratum**.



# Subclassification

- Very similar to **stratifying**.
- Build **combination** of X1 and X2 (strata).
- Compare **within stratum**.



# Subclassification

- To estimate the Average Treatment Effect, we take a **weighted average**:

$$\hat{ATE} = \sum_{s=1}^S \frac{N_s}{N} (\bar{Y}_{1s} - \bar{Y}_{0s})$$

**What happens when we have too many variables to build strata?**



# The curse of dimensionality

- When we have too many covariates, the number of strata or groups grow **exponentially!**
  - E.g. with 4 covariates, each with 5 categories, we have **625 combinations!**
- Very possible that a stratum only has treatment or control units.



# The curse of dimensionality

- When we have too many covariates, the number of strata or groups grow **exponentially!**
  - E.g. with 4 covariates, each with 5 categories, we have **625 combinations!**
- Very possible that a stratum only has treatment or control units.

What to do?



# Breaking the curse: Balancing scores

- Want to **reduce the dimensionality** of our covariates
- A balancing score  $b(x)$  is a function of the covariates such that:

$$Z_i \perp\!\!\!\perp X_i | b(X_i)$$

- This means that conditioning on the balancing score is **enough to remove bias** associated to the covariates.
- Under unconfoundedness:

$$Z_i \perp\!\!\!\perp Y_i(0), Y_i(1) | b(X_i)$$

- There are different balancing scores:
  - E.g. propensity scores, mahalanobis distance.

# Estimating balancing scores

## Propensity score

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

where  $p = \Pr(Z = 1)$

```
e <- predict(glm(z ~ x1 + x2 + x3, data = d, family = binomial(link="logit")),  
             type="response")
```

# Estimating balancing scores

## Mahalanobis Distance

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

where  $\mathbf{x}$  is the covariate vector for observation  $i$ ,  $\boldsymbol{\mu}$  is the mean vector of covariates for the sample, and  $\mathbf{S}$  is the covariance matrix.

```
Sx <- cov(x)
D <- mahalanobis(x, colMeans(x), Sx)

# We can also use a rank-based Mahalanobis distance matrix
library(designmatch)

D <- distmat(z, x)
```

# How matchy-matchy

- Using the previous balancing scores (or covariates directly!) we can **match observations between the treatment and control group**

## Step 1: Preprocessing

Try to model the treatment assignment

## Step 2: Estimation

Use the new trimmed/preprocessed data to build a model, calculate difference in means, etc.

# How matchy-matchy

- There are different matching methods (and different ways to use them!)

**Nearest neighbor (NN)**

Use balancing scores; Greedy algorithm

**Optimal matching**

Solves an optimization problem; slow on large samples

**Mixed Integer Programming (MIP) matching**

Balances covariates directly; can generate smaller samples

**Let's go to R**



# Weighting, the cousin of matching

- We can also use **weights** to make two samples look alike.

## Inverse Probability Weighting (IPW)

- Make some observations more important than others.

# Weighting, the cousin of matching

- You can make the treatment and the control group look more like each other:

$$w_{ATE} = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i}$$

- Observations in the control group will have a weight of  $\frac{1}{1-e_i}$ , while observations in the treatment group will have weights of  $\frac{1}{e_i}$

**What happens with obs that are very likely to be in the treatment group?**

# Weighting, the cousin of matching

- You can make the treatment and the control group look more like each other:

$$w_{ATT} = \frac{e_i \cdot Z_i}{e_i} + \frac{e_i(1 - Z_i)}{1 - e_i}$$

- Observations in the control group will have a weight of  $\frac{e_i}{1-e_i}$ , while observations in the treatment group will have weights of 1

**Why do observations in the treatment group get a weight of 1?**

# Weighting for approximating a population

- If we assume that our sample was selected on observables, and we want it to **look more like a population of interest**, we can also do that!
- We can use the same formulas, but now  $e_i$  is not the probability of being assigned to treatment, but the probability of **being in the population sample**.
- For this type of weighting, we only need population covariates.

**Let's go to R**

# The shortcomings of matching

- Many researchers misuse matching and **confuse it with an identification strategy**
- In terms of identification, **matching still relies on selection on observables**

**You need other source of exogenous variation!**

- Claiming that you can identify a **causal effect** just by using matching is almost the same as claiming this using a regression approach.

**Not a good idea...**

# Don't get it twisted

- Matching works great as an **adjustment method**.
- Combined with **other identification strategies**, it can improve results!



# Main takeaways



- Matching and weighting methods can be great tools for your analysis.
  - Create more similar groups of comparisons.
  - Reduce model dependence
  - Even help with external validity (under assumptions)



# Next week

- We will look at some **identification strategies** for observational studies:
  - Natural experiments and differences-in-differences.
- What **assumptions** need to hold?
- How do we **identify a natural experiment**?
- What does **DD** buy us?

# References

- Angrist, J. and S. Pischke. (2015). "Mastering Metrics". *Chapter 2*.
- Heiss, A. (2020). "Program Evaluation for Public Policy". *Class 7: Randomization and Matching, Course at BYU*
- Imbens, G. and D. Rubin. (2015). "Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction". *Chapter 3*
- Cunningham, S. (2021). "Causal Inference: The Mixtape". *Chapter 5*