# STA 235H - Introduction to Observational Studies I

## Fall 2021

McCombs School of Business, UT Austin

# Announcements

## Homework 3 will be posted on Thursday

- Remember: Group collaboration is part of the assignment
    - If there are unsolvable issues, contact the instruction team (on the earlier side!).
- Homework 2 answer key has been added to the course website (check out the rubric!)
- There will be office hours today (if anyone needs to meet Wednesday, let me know)
- I'll send out a poll for a review session for the midterm
    - If you are interested, please respond.

# Warning

## COLLABORATION, STUDY PARTNERS/GROUPS, AND ACADEMIC INTEGRITY

In addition to the general UT policies regarding academic integrity that are described in the syllabus (and in the UT Course Catalogue), this course has a few other specific policies:

- You are encouraged to form study groups. Collaboration is key for learning! However, you are not allowed to copy directly from another student or let someone else copy from you (this includes copying between groups).
- These same rules apply to R code. You are encouraged to discuss potential problems, but you (your group) need to write your own R code. In any case where we suspect cheating, we will compare both R scripts and homework write-ups, and all students involved will receive an F in this course and be referred to the Dean's office for further disciplinary proceedings (and further potential academic consequences).
- To avoid any potential conflicts, please do not share your files with another student/group. This is also considered cheating and you will be subject to the same disciplinary actions stated above.
- All students in this course assume responsibility for abiding by these policies. If you are unsure about whether a specific type of collaboration crosses the line into copying, then just ask us.

# Last week

- Randomized controlled trials
    - Why is it considered the gold standard?
    - How to analyze an RCT in practice?
    - Assumptions and limitations.

# Last week

- Randomized controlled trials
  - Assumptions for RCTs?

# Last week

- Randomized controlled trials
  - Limitations?

# Today, we're moving forward...


- 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

- Unconfoundnedness 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)

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

# An example about the CIA

- Let's think about the fake CV example and a real life application.
- Causal question: How does getting an internship affect your probability of being in the film industry 5 years later?
- A firm needs to hire interns ASAP, no time for interviews. What would this firm look at in a CV?
    - e.g. level of education, experience, name?
- *Could we assume that conditional on education, experience, name characteristics, etc. receiving an internship is independent from your potential outcomes?*

# The assignment mechanism

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

# Selection on observables

- Units select into treatment based on characteristics **I can observe**.
- What this means in practice is that **all confounders are observable and I can adjust for them**.
    - **Overt bias:** Bias caused by observed confounders. I can remove it by adjusting by these variables.
    - **Hidden bias:** Bias caused by unobserved confounders. I can't directly remove it (I need to rely on other assumptions).

# 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?

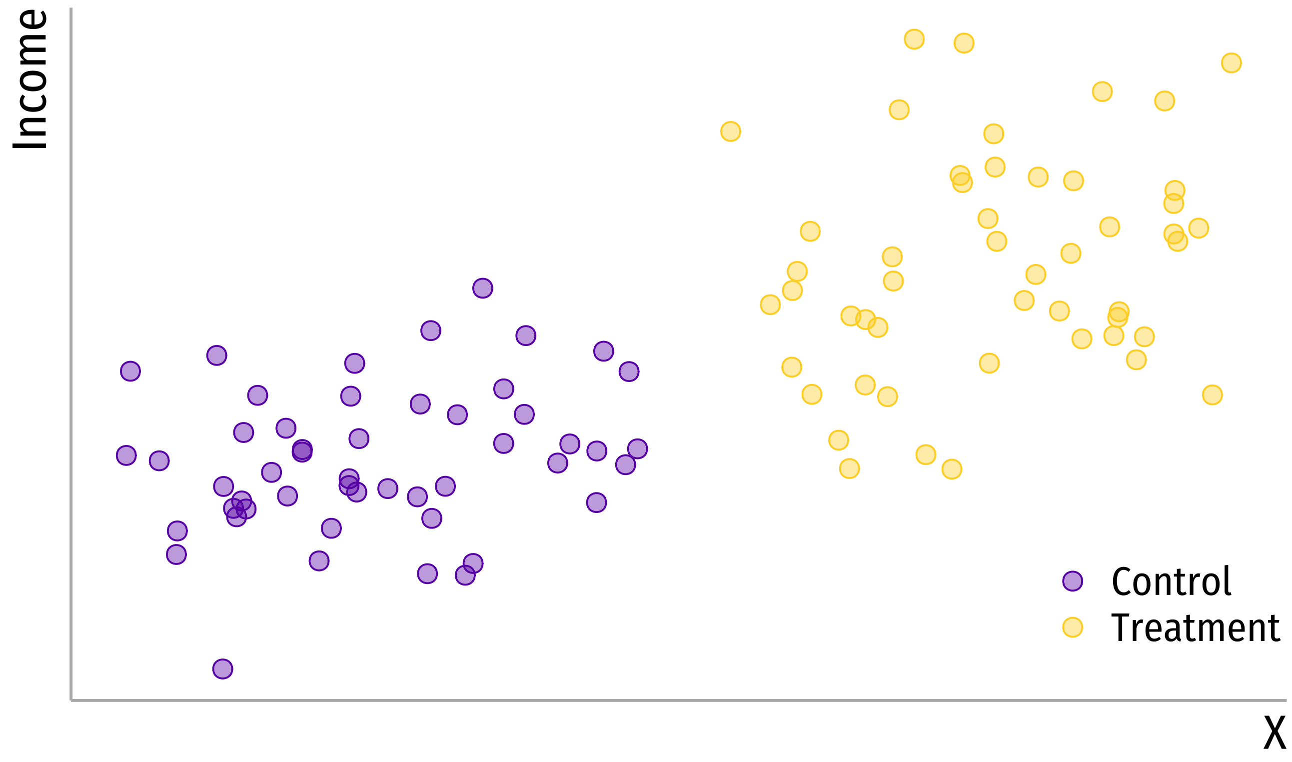
- One way we have seen so far is **regression adjustment**

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

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

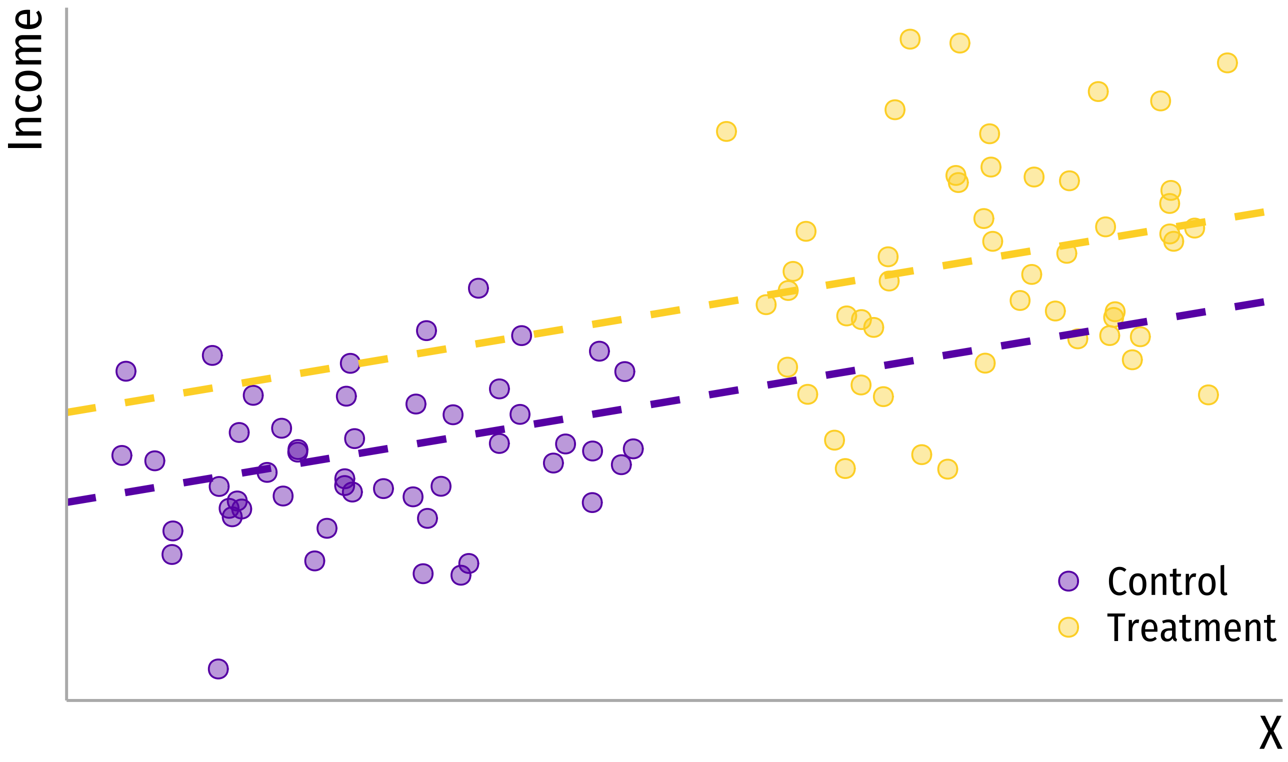
# How do we adjust for observables?

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



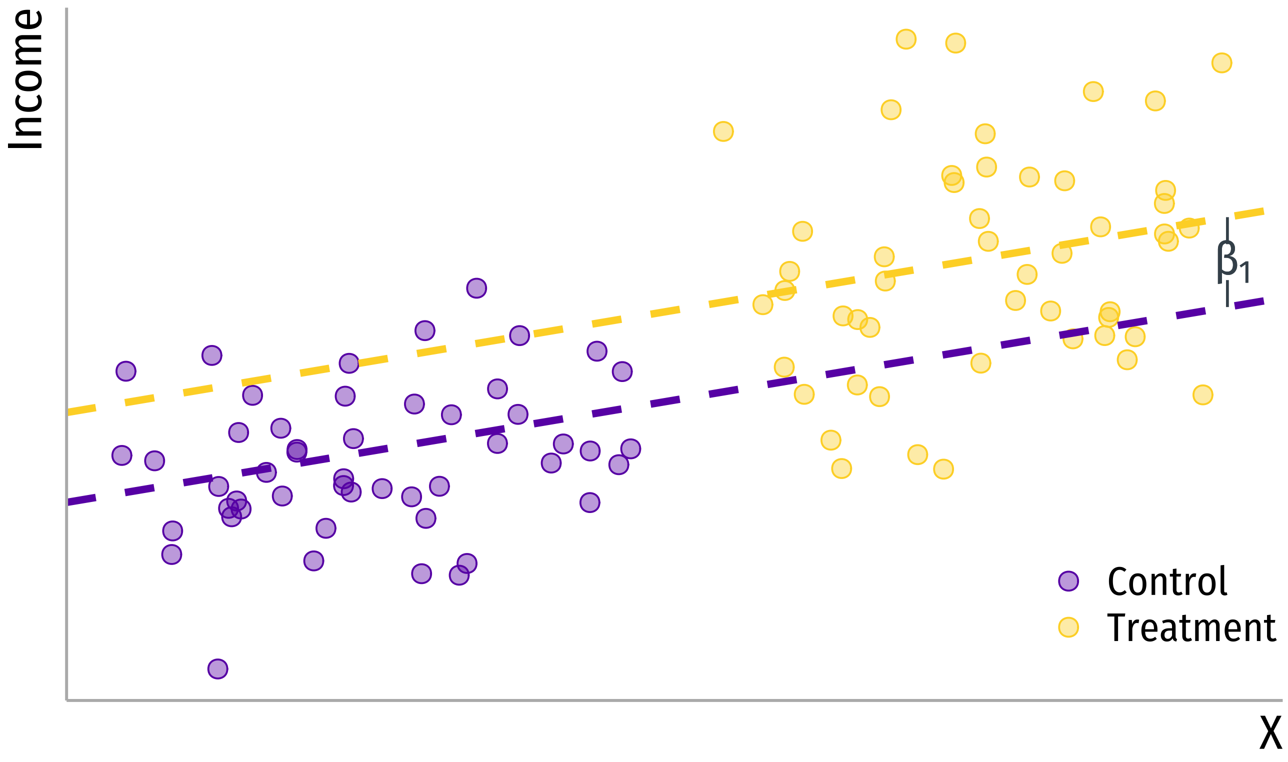
# How do we adjust for observables?

- 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**

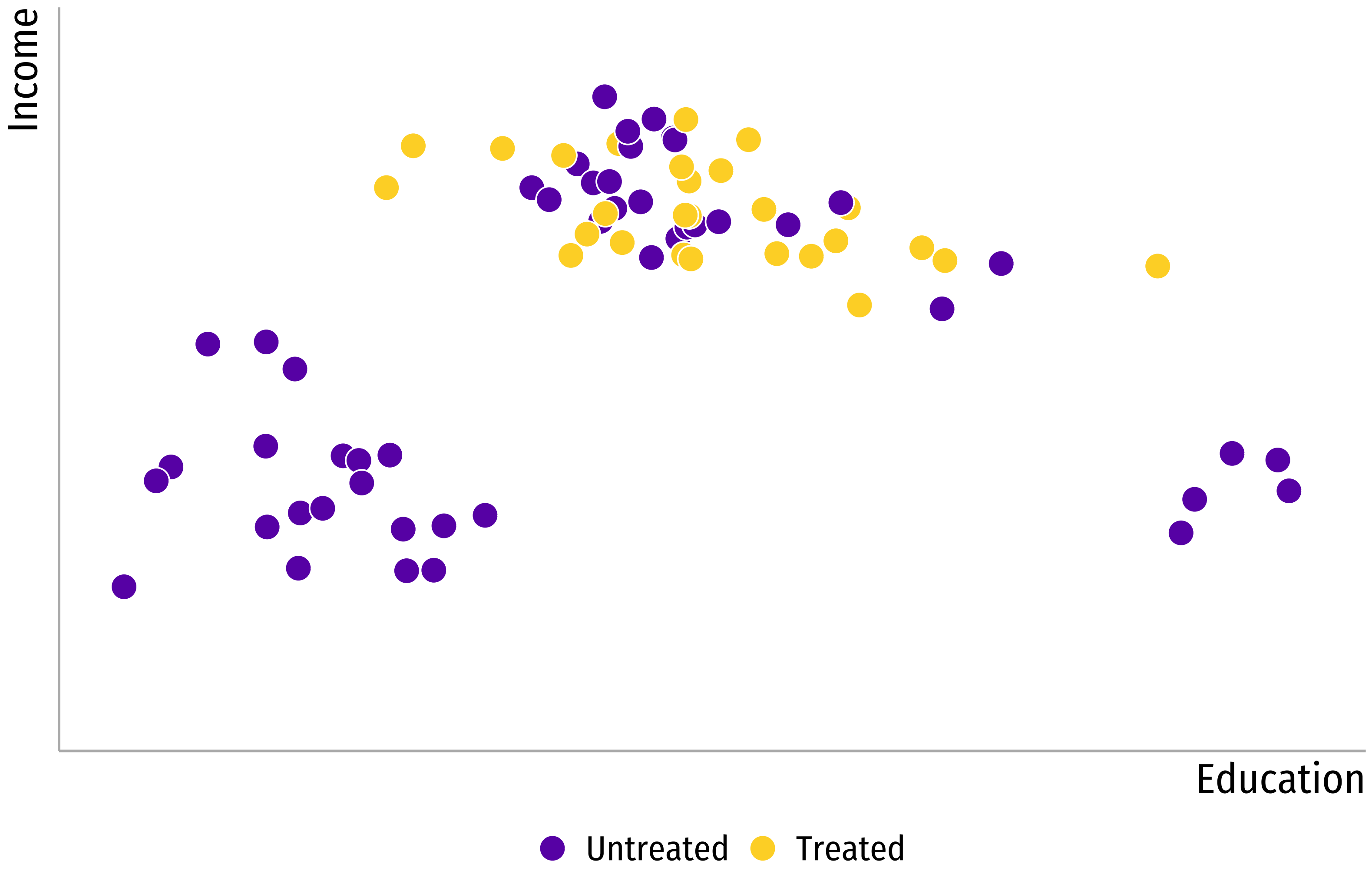
Imbalance ⟶ model dependence ⟶ researcher discretion ⟶ bias

**Compare like to like**

No extrapolation!

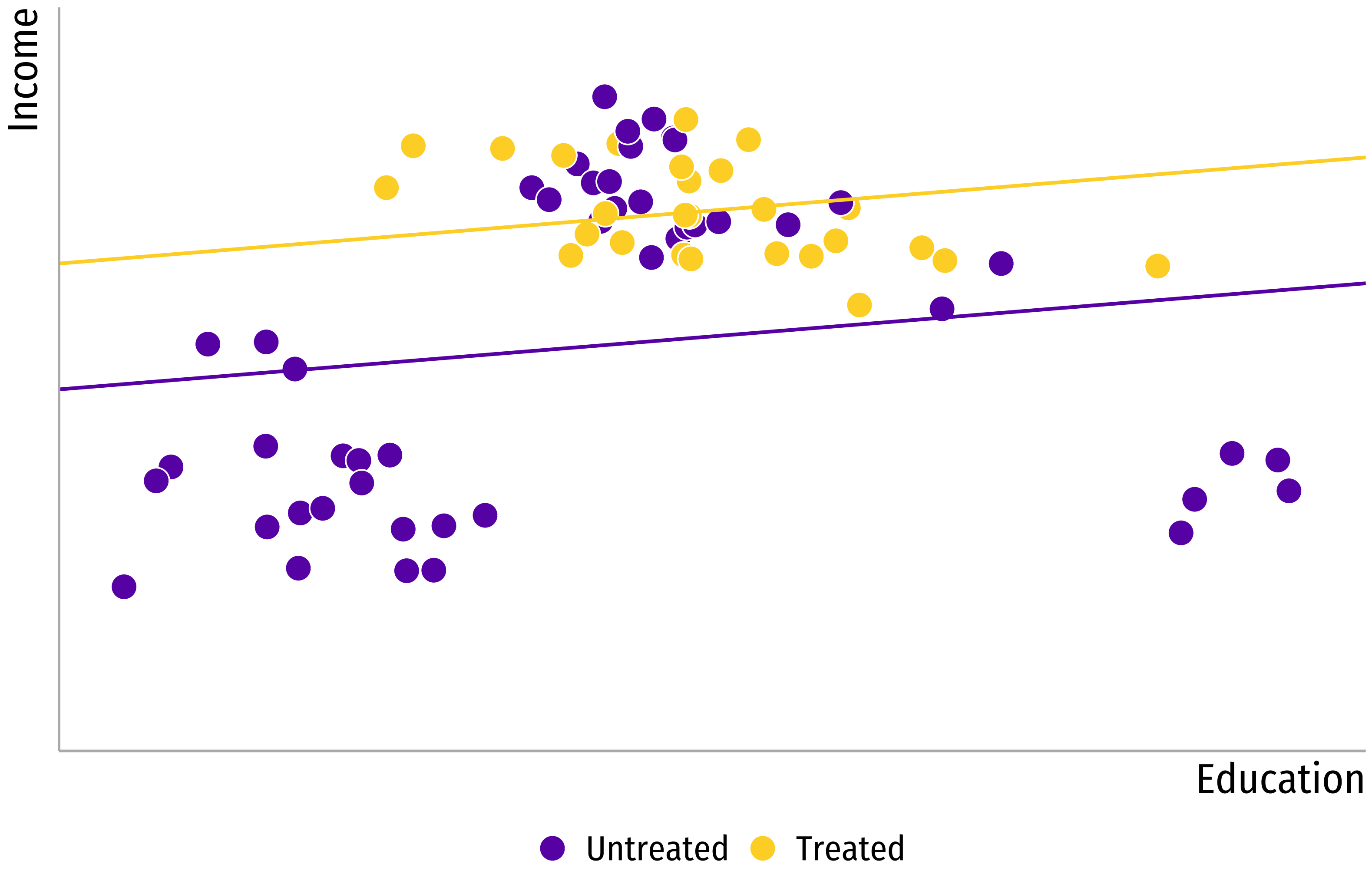
**Can adjust closely by covariates**

Exact matching, coarsened exact matching, fine balance..

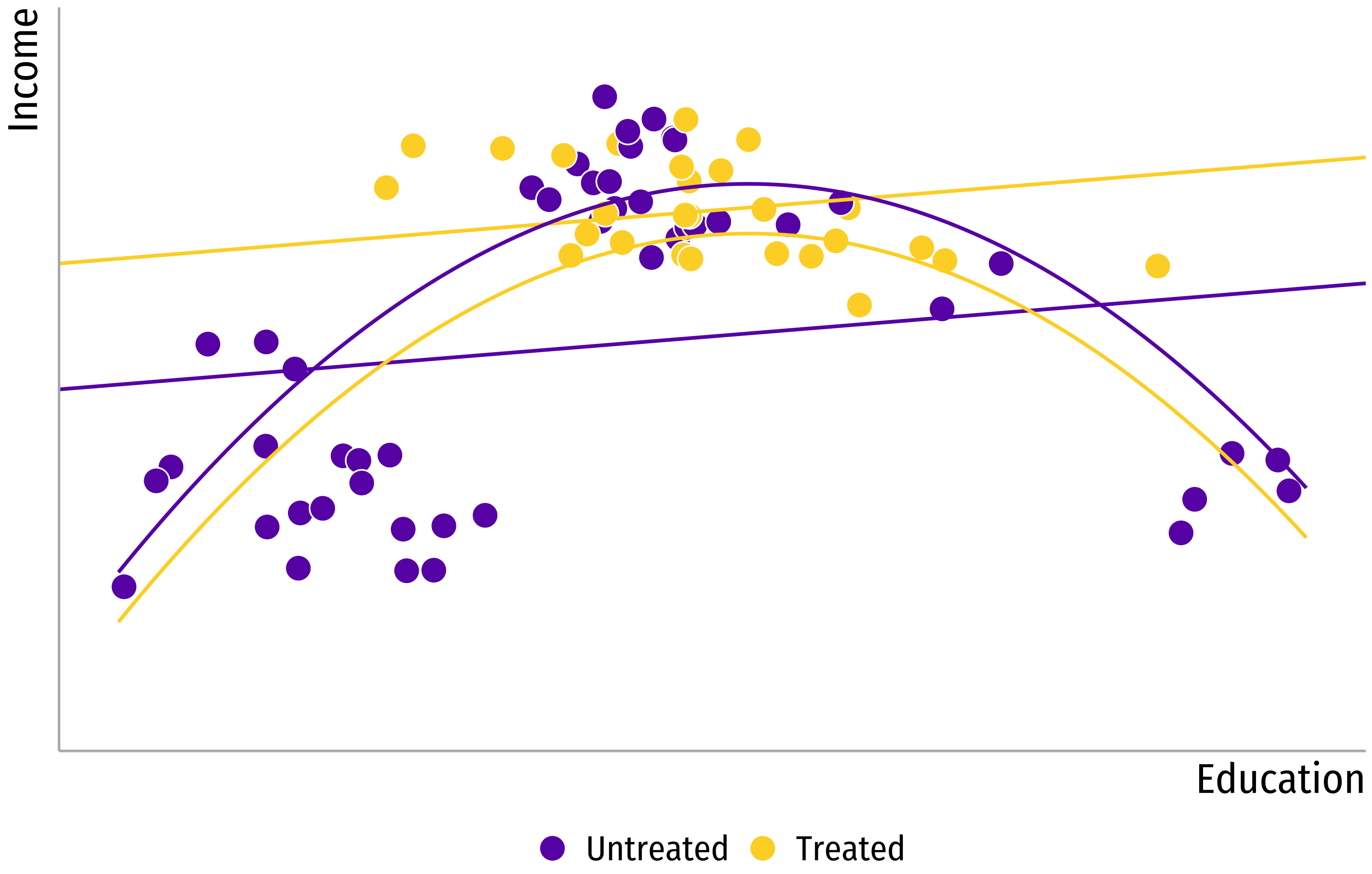


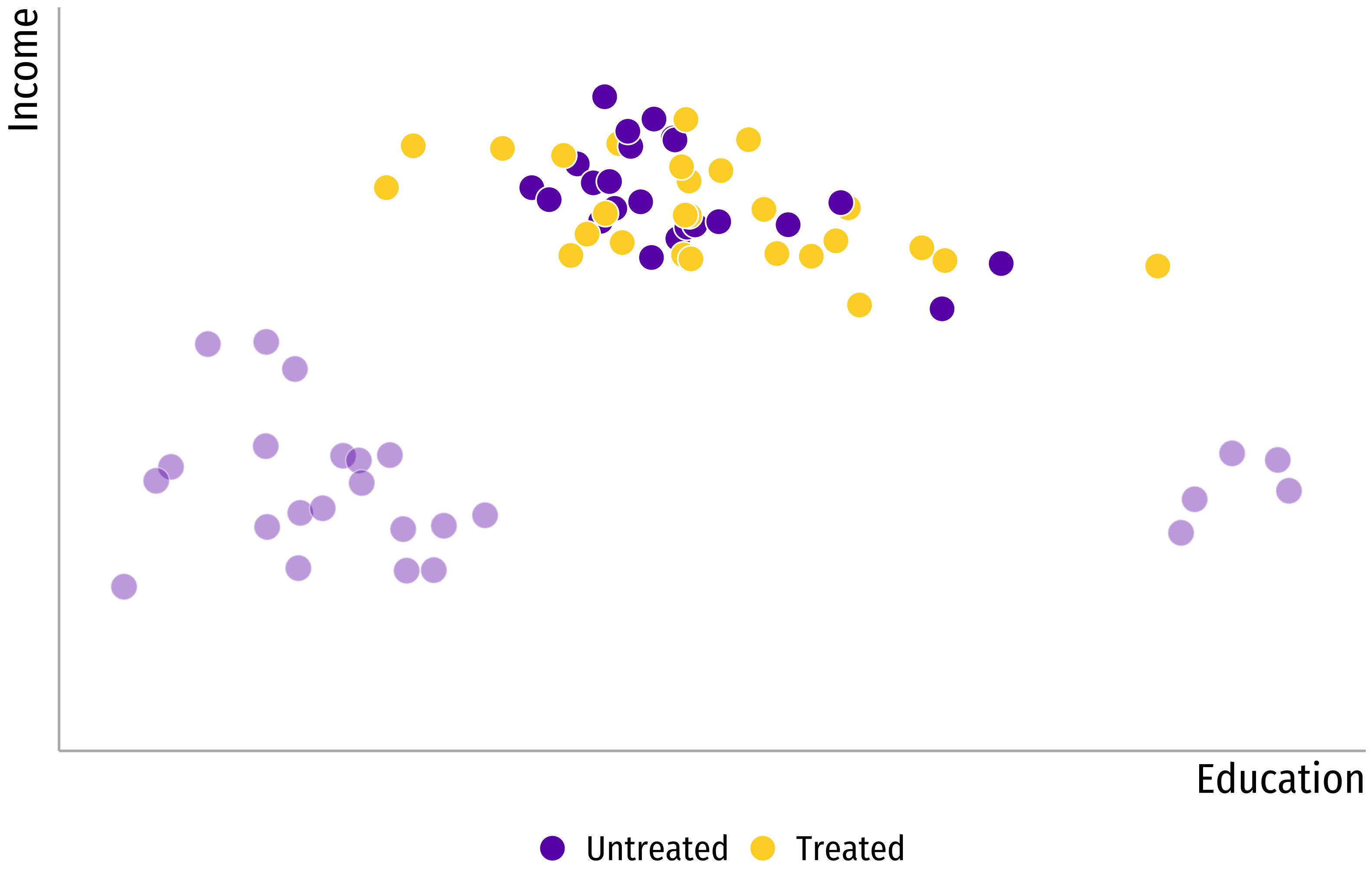
Income
Education
Untreated
Treated

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

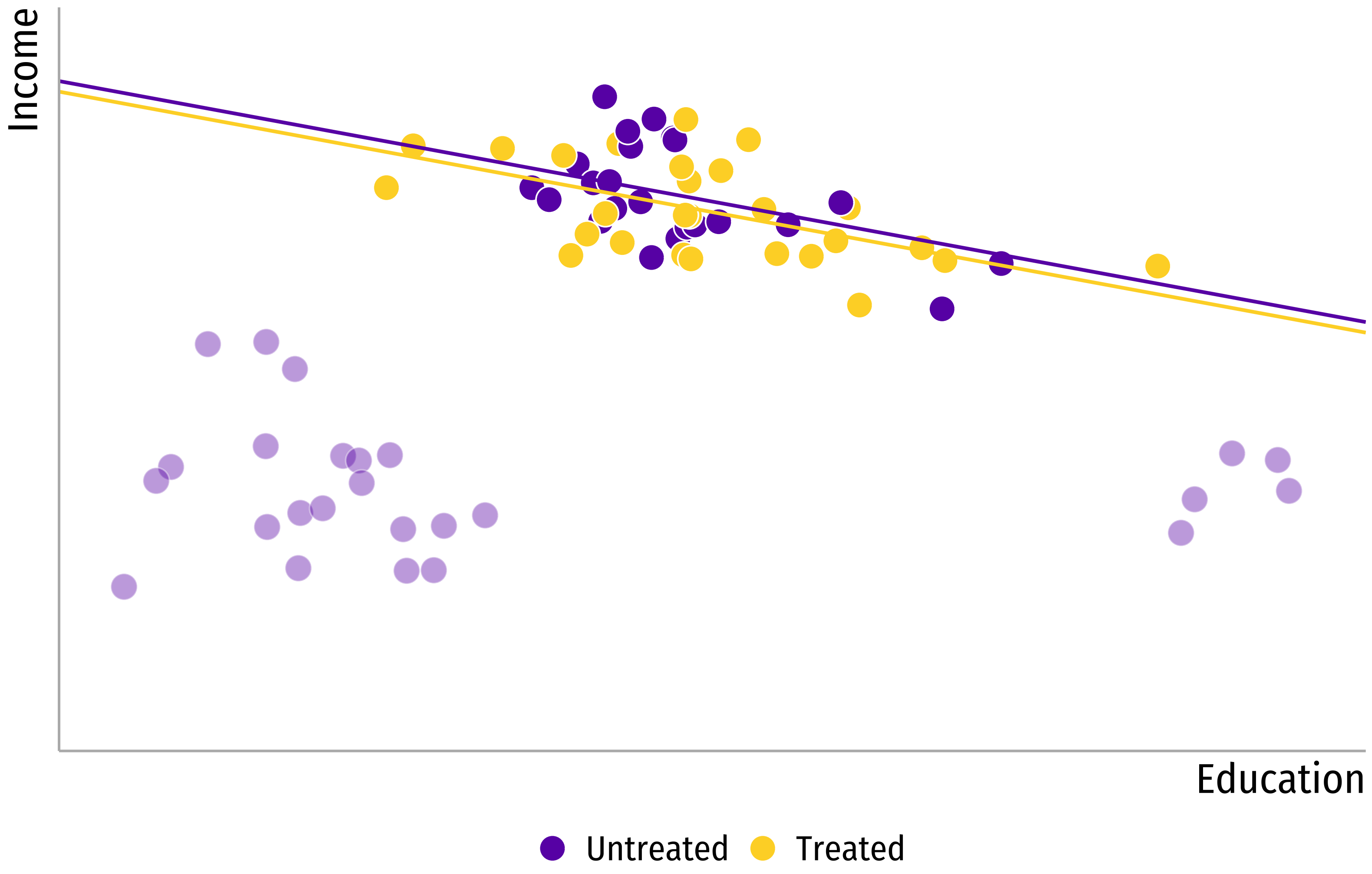


$$\text{Income} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Education}^2 + \beta_3 \text{Treatment}$$

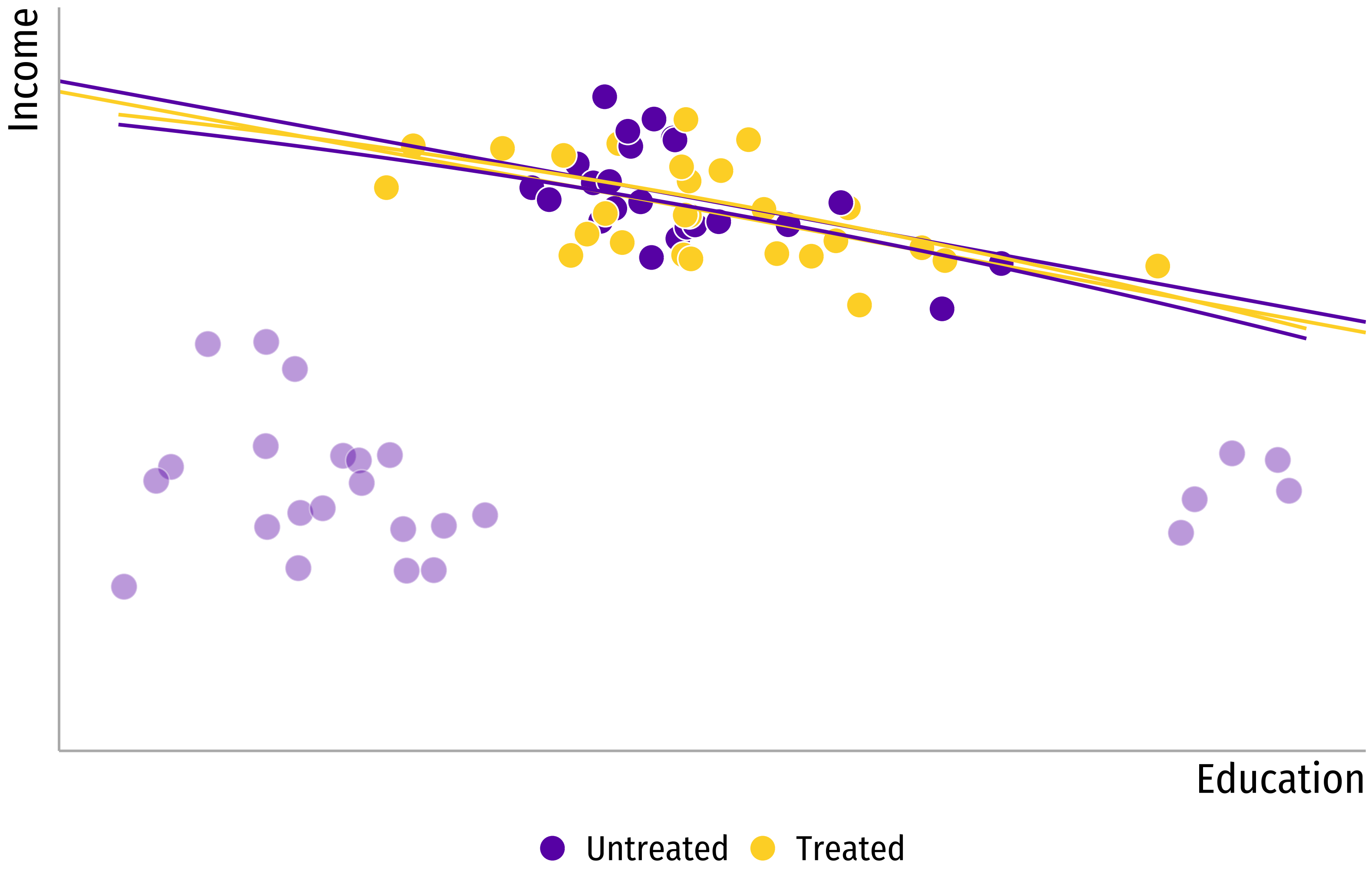



Income
Education
Untreated
Treated

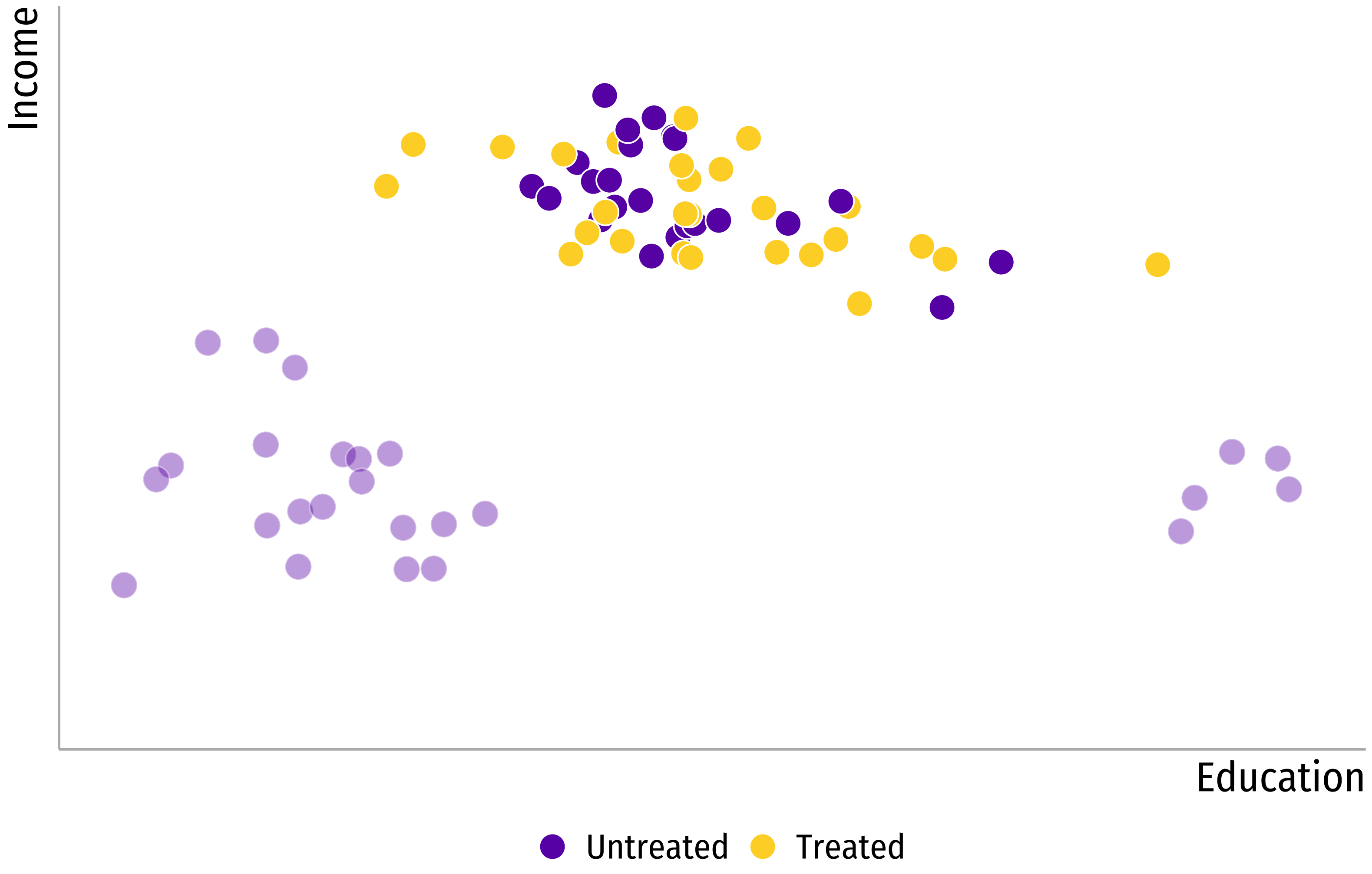
$$\text{Income} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Treatment}$$



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

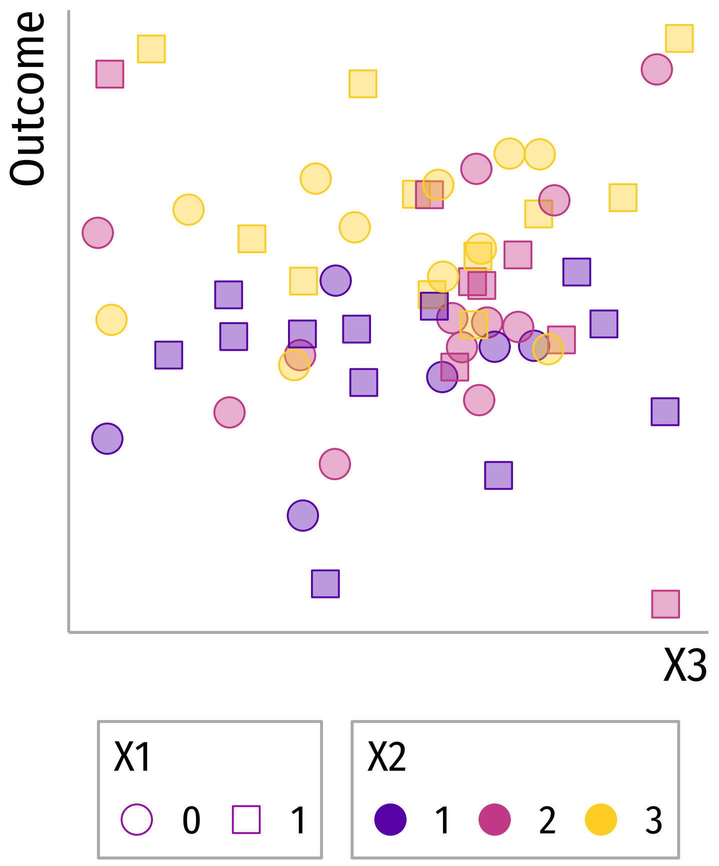


# How do we know we can remove those observations?



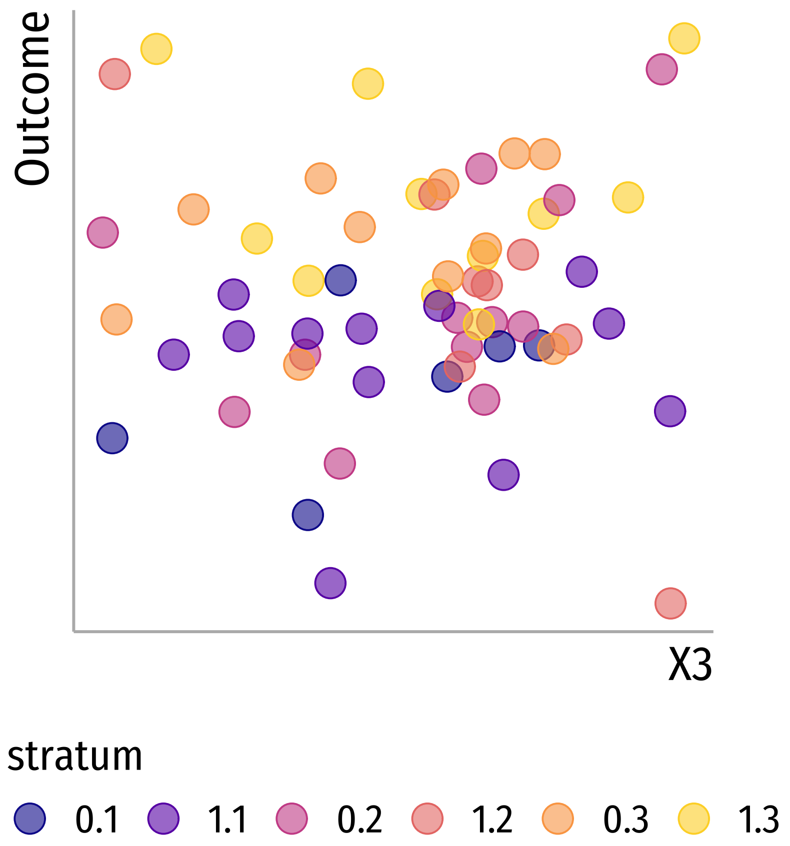
# Subclassification

- Very similar to **stratifying**.



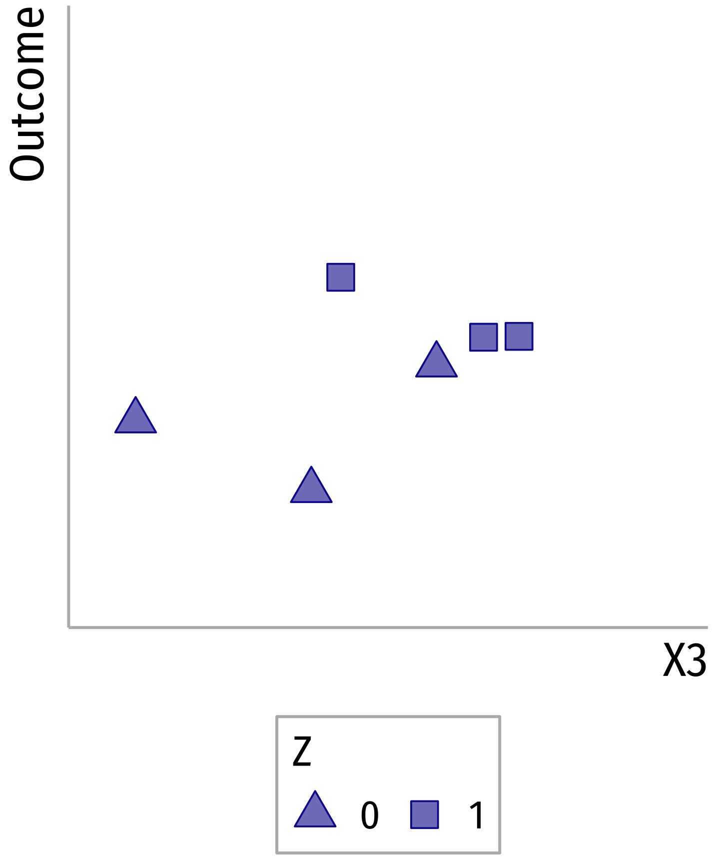
# Subclassification

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



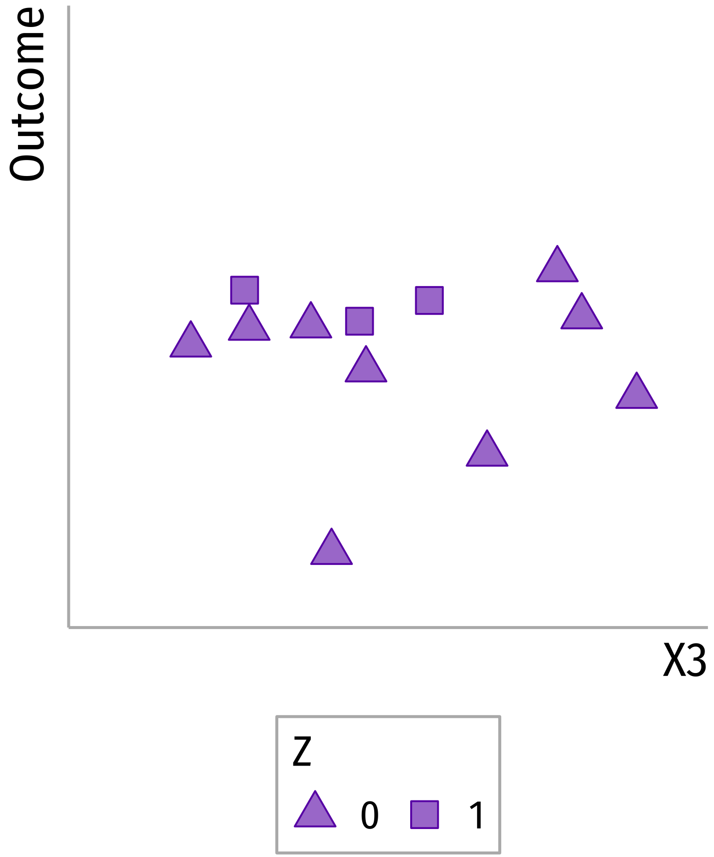
# Subclassification

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



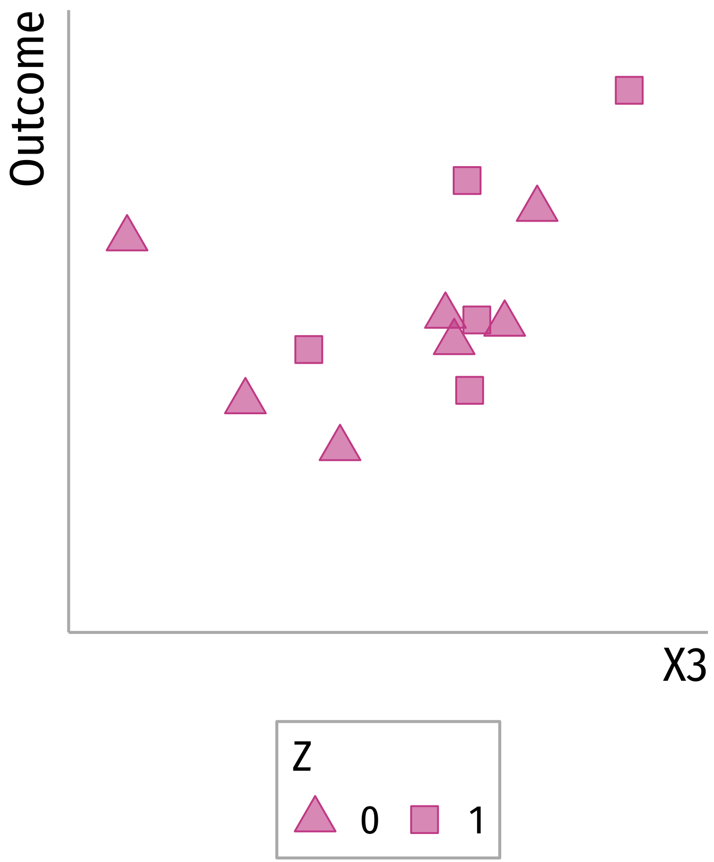
# Subclassification

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



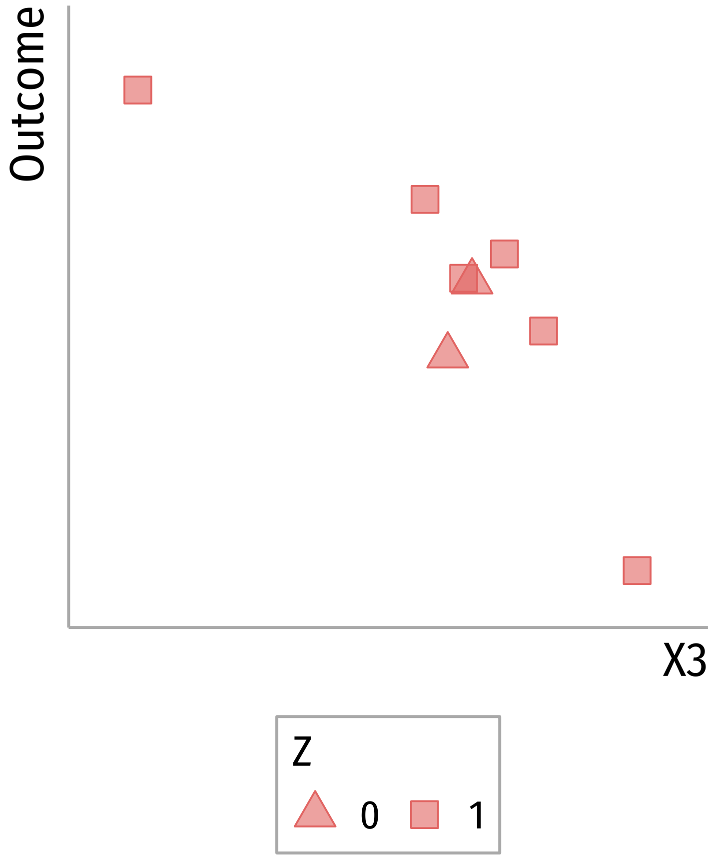
# Subclassification

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



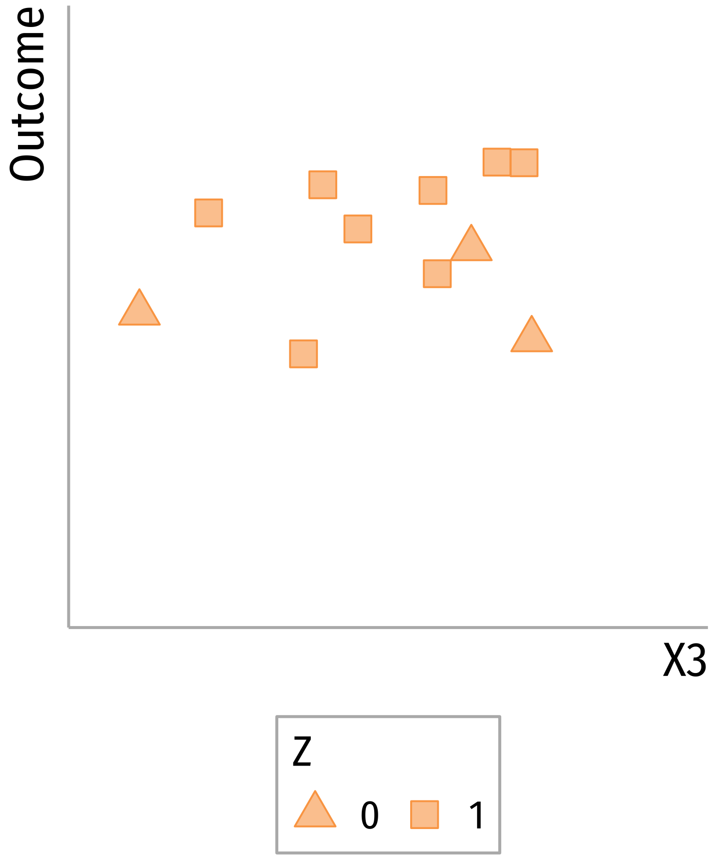
# Subclassification

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



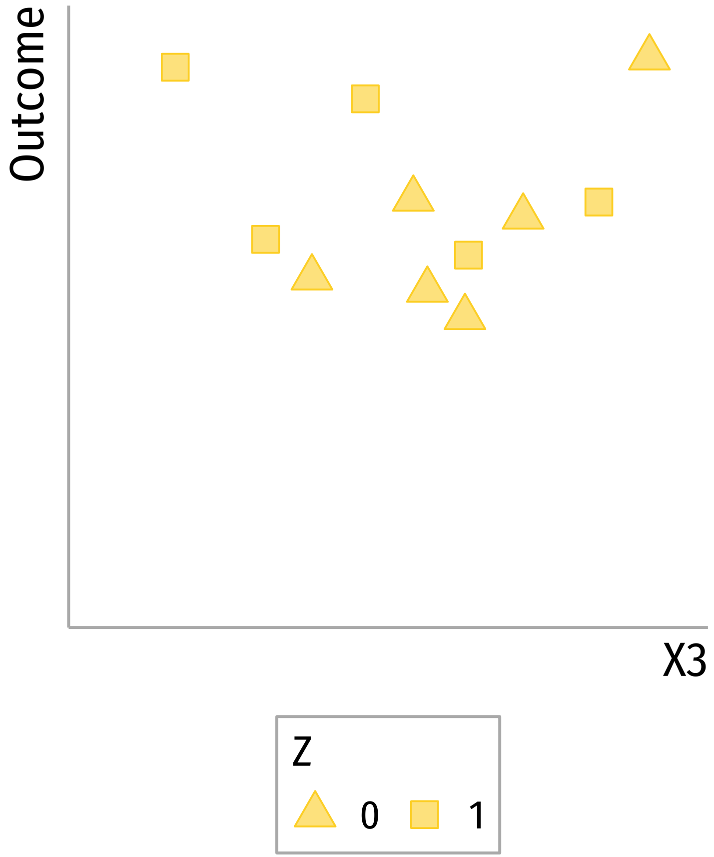
# Subclassification

- 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 dimentionality

- 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 dimentionality

- 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 dimentionality 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:

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

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

# Estimating balancing scores

**Propensity score**

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

where $p = Pr(Z = 1)$

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

# Estimating balancing scores

**Propensity score**

- Importance of overlap region

# Making groups comparable

- 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

# The shortcomings of matching

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

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

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

**Usually 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 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*