Do this at beginning of class

- Download lecture 4 Stata "do-file" link
  - Save to a folder you can easily find
  - Note: you will have to change filepaths ("cd" command) for do-file to run
- Dataset we will be working with today
  - ELS 2002 dataset [different than last week] link
  - Save to folder you can easily find; do not change file names

EDUC 263: Introduction to Econometrics, Lecture 4

Introduction to matching (part 1 of matching unit)

Ozan Jaquette
ozanj@ucla.edu

University of California, Los Angeles
Higher Education & Organizational Change

Goals for three-week unit on matching

Today: Conceptual introduction to matching

- Gain deeper understanding of Rubin's Causal Model/ potential outcomes framework, which is core to matching
- Introduce the parts of matching, so that you can understand matching as a whole

Following two weeks: Deeper understanding of matching

- Next week:
  - Brief overview of logistic regression
  - In-depth discussion of matching assumptions
  - In-depth discussion of implementation and alternative matching methods
- Following week:
  - Inverse probability weighting: preferred matching approach
  - More practice conducting matching

What we will do today

1. Introduce research question: What is the effect of attending community college on BA attainment
2. Rubin's Causal Model (RCM) and treatment effects
   - Potential outcomes
   - Treatment effects
3. Overview of matching
   - Comparing matching to OLS
   - Simple example of matching by hand
4. The propensity score as the cure for the curse of dimensionality
   - The curse of dimensionality
   - The propensity score
5. Practical example of implementing matching using "nearest neighbor" approach

Introduce research question: What is the effect of attending community college on BA attainment

Research question we will use for matching unit

Research question:

- What is the effect of attending a community college (rather than 4-year institution) on BA attainment (Y)?

Treatment variable T_i

- $T_i = 1$: first institution is public community college (CC)
- $T_i = 0$: first institution is four-year institution

Data: Education longitudinal study of 2002 (ELS)

- 10th graders in 2002

Sample:

- 12th grade in 2004; expect to obtain BA; attend a CC or 4-year by 2006

Run code together in Stata

Rubin's Causal Model (RCM) and treatment effects

Review Rubin causal model (RCM) of potential outcomes

RQ: What is the effect of attending a community college ($T_i = 1$) vs. 4-year ($T_i = 0$) on salary (Y)?

- sample=high school grad in 2004; salary measured in 2011

Each person i has two **potential outcomes** Y_i

- $Y_i(1)$ salary when assigned to treatment (community college)
- $Y_i(0)$ salary when assigned to control (4-year institution)

How to think about potential outcomes

- for each person i, the treated potential outcome $Y_i(1)$ and the untreated potential outcome $Y_i(0)$ already exist
- Treatment variable T_i just determines which of the two potential outcomes we get to observe

Review potential outcomes notation

What is effect of attending community college (CC) vs. 4-year on salary (Y)?

- Treatment, T_i
  - 1=treated (CC); 0= untreated/control (4-year)
- Four combinations of potential outcome and treatment assignment:
  - $Y_i(1) \mid T_i = 1$ ➡ salary after treatment for those actually assigned to treatment (**observed**)
  - $Y_i(0) \mid T_i = 1$ ➡ salary after no treatment for those actually assigned to treatment (**not observed**)
  - $Y_i(1) \mid T_i = 0$ ➡ salary after treatment for those assigned to control (**not observed**)
  - $Y_i(0) \mid T_i = 0$ ➡ salary after no treatment for those assigned to control (**observed**)

Review potential outcomes notation

What is effect of attending community college (CC) vs. 4-year on salary (Y)?

	Actually treated (T=1)	Actually not treated (T=0)
$Y_i(1)$	$Y_i(1) \mid T_i = 1$	$Y_i(1) \mid T_i = 0$
$Y_i(0)$	$Y_i(0) \mid T_i = 1$	$Y_i(0) \mid T_i = 0$

The two blue cells show the notation for the *actual* outcomes

- E.g., for those who were actually treated (attended CC), their salary in 2011 is represented by $Y_i(1) \mid T_i = 1$
- "The treated potential outcome, for those who underwent treatment"

Red cells are *counterfactual* outcomes, which we never see

- Need counterfactual to estimate "true" causal effect
- Need some substitute for the counterfactual to estimate causal effect

Table of potential outcomes

What is effect of attending community college (CC) vs. 4-year on salary (Y)?

Imagine we know treated $Y_i(1)$ and untreated $Y_i(0)$ potential outcomes for all i

i	$Y_i(1)$ **Treated**	$Y_i(0)$ **Untreated**	τ_i **Unit effect**
1	65	60	5
2	30	35	-5
3	55	60	-5
4	25	30	-5
5	50	50	0
6	80	70	10
7	45	45	0
Avg	**50**	**50**	**0**

True causal effect for each i

- Unit treatment effect: for each person, compare their treated salary to untreated salary
- $UTE_i = Y_i(1) - Y_i(0)$

Types of treatment effects

1. Unit treatment effect (UTE)
   - For each person i, compare their treated to untreated Y
   - $UTE_i = Y_i(1) - Y_i(0)$
2. Average treatment effect (ATE)
   - For all people ($T_i = 1$ and $T_i = 0$), what is the average effect of receiving treatment rather control?
   - $ATE = E[Y_i(1)] - E[Y_i(0)]$
3. Average treatment effect on the treated (ATT)
   - For people assigned to treatment ($T_i = 1$), what is the average effect of receiving treatment rather control?
   - $ATT = E(Y_i(1)|T = 1) - E(Y_i(0)|T = 1)$

Table of potential outcomes, calculating ATE

What is effect of attending community college (CC) vs. 4-year on salary (Y)?

i	T_i **Assignment**	$Y_i(1)$ **Treated**	$Y_i(0)$ **Untreated**	τ_i **Unit effect**
1	1	65	60	5
2	0	30	35	-5
3	1	55	60	-5
4	0	25	30	-5
5	1	50	50	0
6	0	80	70	10
7	1	45	45	0
Avg		**50**	**50**	**0**

Calculating average treatment effect (ATE)

- $ATE = E[Y_i(1)] - E[Y_i(0)] =$
- $ATE = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0)$
- $ATE = 50 - 50 = 0$

Table of potential outcomes, calculating ATT

What is effect of attending community college (CC) vs. 4-year on salary (Y)?

i	T_i Assignment	$Y_i(1)$ Treated	$Y_i(0)$ Untreated	τ_i Unit effect
1	**1**	**65**	**60**	**5**
2	0	30	35	-5
3	**1**	**55**	**60**	**-5**
4	0	25	30	-5
5	**1**	**50**	**50**	**0**
6	0	80	70	10
7	**1**	**45**	**45**	**0**
Avg [treated]		**53.75**	**53.75**	**0**

Calculating average treatment effect on the treated (ATT)

- $ATT = E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1] =$
- $ATE = 53.75 - 53.75 = 0$

Should you prefer average treatment effect (ATE) or average treatment effect on treated (ATT?)

- ATE: for all people, what is the average effect of receiving treatment rather control?
- ATT: for people assigned to treatment ($T_i = 1$), what is the average effect of receiving treatment rather control?

Conceptually, depends and on policy debate/implications around your research question (RQ)

- RQ: effect of CC vs. 4-year on BA; policy debate about whether CC students would be better served at 4-year, suggests ATT
- RQ: effect of MAS on graduation; interested in whether all students (not just MAS participants) would benefit from MAS, suggests ATE

Should you prefer average treatment effect (ATE) or average treatment effect on treated (ATT?)

- ATE: for all people, what is the average effect of receiving treatment rather control?
- ATT: for people assigned to treatment ($T_i = 1$), what is the average effect of receiving treatment rather control?

Methodological considerations

- ATT requires fewer assumptions (next slide)
- If participation voluntary, treated likely differ from untreated in ways difficult to measure.
  - Cleaner to estimate ATT:
  - ATE requires you to estimate effect of treated getting treatment rather than control (the ATT) **and** of untreated getting treatment rather than control

Estimating treatment effects is a missing data problem; we don't know counterfactual(s)

- Unit treatment effect ($UTE_i = Y_i(1) - Y_i(0)$)
  - Missing data: we observe one potential outcome for each i
- Average treatment effect (ATE)
  - For all people ($T_i = 1$ and $T_i = 0$), what is the average effect of receiving treatment rather control?
  - $ATE = E[Y_i(1)] - E[Y_i(0)]$
  - Two pieces of missing data:
    1. Don't know $E[Y_i(1)]$ because don't observe $Y_i(1)$ for all i
    2. Don't know $E[Y_i(0)]$ because don't observe $Y_i(0)$ for all i
- Average treatment effect on the treated (ATT)
  - For people assigned to treatment ($T_i = 1$), what is the average effect of receiving treatment rather control?
  - $ATT = E(Y_i(1)|T = 1) - E(Y_i(0)|T = 1)$
  - Only one piece of missing data:
    - We know $E(Y_i(1)|T = 1)$
    - Don't know $E(Y_i(0)|T = 1)$ because don't observe $Y_i(0)$ for i assigned to treatment

Random assignment experiments solve the missing data problem

Average treatment effect on the treated (ATT)

- Average treatment effect on the treated (ATT)
  - For people assigned to treatment ($T_i = 1$), what is the average effect of receiving treatment rather control?
  - $ATT = E(Y_i(1)|T = 1) - E(Y_i(0)|T = 1)$
  - One piece of missing data:
    - We know $E(Y_i(1)|T = 1)$
    - Don't know $E(Y_i(0)|T = 1)$ because don't observe $Y_i(0)$ for i assigned to treatment
- If value of T_i randomly assigned:
  - Treatment assignment unrelated to values of potential outcomes $Y_i(1)$ and $Y_i(0)$
  - Can substitute in known quantity $E[Y_i(0)|T_i = 0]$ for missing counterfactual $E[Y_i(0)|T_i = 1]$
  - $E[Y_i(0)|T_i = 0] = E[Y_i(0)|T_i = 1]$
  - $ATT = E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0]$

Random assignment solve missing data problem
Average treatment effects (ATE)

- Average treatment effect ($ATE = E[Y_i(1)] - E[Y_i(0)]$)
  - ATE: For all people, what is the average effect of receiving treatment rather control?
  - Two pieces of missing data:
    1. Don't know $E[Y_i(1)]$ because don't observe $Y_i(1)$ for all i
    2. Don't know $E[Y_i(0)]$ because don't observe $Y_i(0)$ for all i
- If value of T_i randomly assigned:
  - Substitute known quantities for missing counterfactuals
- Formula for ATE more complicated than formula for ATT
  - Must substitute for two counterfactuals
  - Formula must include $Pr(T_i = 1)$ and $Pr(T_i = 0)$
- ATE formula [don't worry about this]:
  - $ATE = E(Y_i(1)) - E(Y_i(0))$
  - $ATE = [Pr(T = 1)E(Y_i(1)|T = 1) + Pr(T = 0)E(Y_i(1)|T = 0)] - [Pr(T = 1)E(Y_i(0)|T = 1) + Pr(T = 0)E(Y_i(0)|T = 0)]$
  - $ATE = Pr(T = 1)[E(Y_i(1)|T = 1) - E(Y_i(0)|T = 1)] + Pr(T = 0)[E(Y_i(1)|T = 0) - E(Y_i(0)|T = 0)]$

Overview of matching

Estimating treatment effects when treatment not randomly assigned

- Average treatment effect on the treated (ATT)
  - For people assigned to treatment, what is the average effect of receiving treatment rather control?
  - $ATT = E(Y_i(1)|T = 1) - E(Y_i(0)|T = 1)$
  - Problem: we don't know counterfactual $E(Y_i(0)|T = 1)$
  - Experiments substitute $E(Y_i(0)|T = 0)$ for $E(Y_i(0)|T = 1)$
- If value of T_i not randomly assigned:
  - Treatment assignment likely related to values of potential outcomes $Y_i(1)$ and $Y_i(0)$
  - $E(Y_i(0)|T = 0)$ not a substitute for $E(Y_i(0)|T = 1)$
- Matching tries to replicate random assignment
  - find untreated units that look just like treated units
  - Assume that outcome for these untreated units $(Y_i(0)|T = 0)$ is the counterfactual for treated units $(Y_i(0)|T = 1)$

Conceptually, matching is like estimating causal effects when you know both potential outcomes

Imagine that bolded i actually receive the treatment, $T_i = 1$

i	T_i **Assignment**	$Y_i(1)$ **Treated**	$Y_i(0)$ **Untreated**	τ_i **Unit effect**
1	1	**65**	?	?
2	0	?	**35**	?
3	0	?	60	?
4	1	**25**	?	?
5	0	?	**50**	?
6	1	**80**	?	?
7	0	?	**45**	?

What matching does, Using $i = 1$ as example

- $T_i = 1$ for $i = 1$
- actual: $(Y_i(1)|T_i = 1) = 65$
- counterfactual: $(Y_i(0)|T_i = 1) =?$
- In order to estimate unit treatment effect, we need counterfactual
- Matching: choose an untreated observation that is similar to $i = 1$; substitute $(Y_i(0)|T_i = 0)$ for the counterfactual

To calculate ATT

- Follow same steps for $i = 4$, $i = 6$
- ATT is avg of unit treatment effects from $i = 1, i = 4, i = 6$

So what is matching?

Think of matching as **data preprocessing** rather than statistical analysis

Matching [average effect of treatment on treated (ATT)]

- Goal: for each treated unit, we have $Y_i(1)|T = 1$ but we need to find counterfactual $Y_i(0)|T = 1$
- Matching identifies untreated units that are similar to each treated unit
- Use $Y_i(0)|T = 0$ for the untreated unit as a substitute for $Y_i(0)|T = 1$ of the treated unit
  - Substitute for counterfactual may contain data from several untreated *i*s
- For each treated unit, calculate the difference between actual $Y_i(1)|T = 1$ and substitute for the counterfactual
- ATT is just the average of this difference

Matching to calculate average treatment effect ATE

1. Calculate ATT (as above)
2. For each untreated i, find similar treated i and follow steps above

Crucial assumption underlying matching

Matching estimates causal effects [ATT] by

1. identify untreated i that is **"similar"** to each treated i
2. substituting $Y_i(0)|T = 0$ for the untreated unit for the treated unit counterfactual $Y_i(0)|T = 1$

What does it mean for untreated and treated i to be similar?

- The matched treated and untreated is have the same values on variables that determine the outcome and selection into the treatment

Critical assumption: conditional independence

- When identifying untreated i similar to treated i, you include all variables that drive both the outcome **and** selection into the treatment
- This is the same assumption underlying multiple regression (no omitted variables)

Difference between matching and OLS regression

How regression estimates causal effects

Ordinary least squares (OLS) regression

- Population regression: $Y_i = \alpha T_i + \beta X_i + \epsilon_i$
  - α and β are **population parameters**
  - α is ATE: average effect (for all) of receiving treatment rather than control
  - T_i =treatment; $X_i =$ "control" variables
- OLS regression line: $\hat{Y}_i = \hat{\alpha} T_i + \hat{\beta} X_i$
  - $\hat{\alpha}$ is our estimate of ATE
  - Reynolds and DesJardins (2009) describe regression as "parametric" in that we are estimating parameters

Assumption to claim $\hat{\alpha}$ is unbiased estimate of α (the ATE):

- $E(u_i|T_i) = 0$; Independent variable X_i is unrelated to the "other factors", u_i, not included in model
- i.e., no omitted variables that both (i) affect Y_i **and** (ii) systematically related to T_i
- Also assumes we correctly specified the functional form linking Y_i to control variables X_i

Difference between matching and OLS regression

How regression estimates causal effects

Parametric approach to estimating treatment effects (e.g., OLS)

- Use regression to estimate the value of the missing counterfactual
- Statistically control for observable characteristics

Matching approach

- Substitute for counterfactual is not a parameter estimate; rather, choose an observation that represents best counterfactual for each treated i
- Instead of controlling for characteristics, group together treated and untreated is that have same value on observable characteristics

Difference between matching and OLS regression
How regression estimates causal effects

How regression constructs the counterfactual, $Y_i(0)|T_i = 1$

1. Estimate OLS line:
   - $\hat{Y}_i = \hat{\alpha} T_i + \hat{\beta} X_i$
2. Predicted outcome given treatment:
   - $\hat{Y}_{1i} = \hat{\alpha}(T = 1) + \hat{\beta}(X_i) = \hat{\alpha} + \hat{\beta}(X_i)$
3. Estimate counterfactual $(Y_i(0)|T_i = 1)$ by T=1 with T=0
   - $\hat{Y}_{0i} = \hat{\alpha}(T = 0) + \hat{\beta}(X_i) = \hat{\beta}(X_i)$

How matching constructs the counterfactual, $Y_i(0)|T_i = 1$

- Matching methods do not use estimated parameters to construct counterfactual
- Missing counterfactual $(Y_i(0)|T_i = 1)$ constructed from actual outcomes of untreated units that are matched to treated units

“Nearest neighbor” matching example by hand

No mystery with matching, we simply create a control group that looks like the treated

- So we could actually do this by hand

Let's try! Do this exercise with person next to you:

- Open the “NN matching example” spreadsheet link
- Fictitious data
- You identify matches and calculate ATE, ATT
- Instructions at bottom

The propensity score as the cure for the curse of dimensionality

Curse of dimensionality

Goal: match treated i to "similar" untreated i

- "Similar" untreated units have same values on variables that drive both the outcome and selection into treatment
- Assumption: when identifying untreated i similar to treated i, you include **all** variables that drive both the outcome and selection into the treatment

For each possible combination of Xs, must have at least one untreated i to match to treated i

- Not an issue when sample size is large and small number of variables drive Y and selection into treatment, T
- For example, imagine "first-gen" is only X that affects both Y and T

	first-gen=1	first-gen=0
Start at CC ($T_i = 1$)	40	10
Start at 4-yr ($T_i = 0$)	10	40

Table: Number of observations in each cell, selection on one X variable

Curse of dimensionality

For each possible combination of Xs, must have at least one untreated i to match to treated i

- Often have empty cells as number of Xs increases
- Below, there are no 4-yr students ($T_i = 0$) to match with the 25 first-gen, low-income CC students ($T_i = 1$)

	first-gen=1			first-gen=0		
Family income	low	med	high	low	med	high
Start at CC ($T_i = 1$)	**25**	10	5	5	5	**0**
Start at 4-yr ($T_i = 0$)	**0**	5	5	5	10	**25**

Table: Number of observations in each cell, selection on two X variables

Reyrnolds and DesJardins (2009, p. 59):

> *"To eliminate the selection problem, we need to match on a large number of characteristics, some of which may be continuous. Even if we categorize a continuous variable, we would like to make the categories as fine as possible to reduce selection bias. As X becomes larger in dimension there is a higher probability of missing cells, a problem known as the curse of dimensionality."*

Propensity score cures the curse of dimensionality

Rosenbaum and Rubin (1983) show that:

- Matching on predicted probability of treatment given X is equivalent to matching on X
- Matching on probability of treatment balances distribution of observable characteristics across treated and untreated groups
  - Probability that any treated unit has specific trait (e.g., high income) will be same as probability untreated unit has that trait

Solution to the curse of dimensionality

- Propensity score ($\hat{P}_i(X)$): probability unit i treated given their observable characteristics X
- Instead of matching on individual characteristics (X), match on probability of treatment given X
  - e.g., match treated i with ($\hat{P}_i(X) = .8$) to untreated i with ($\hat{P}_i(X) = .8$)
- Matching on propensity score reduces matching problem to a single dimension

How to calculate and utilize propensity score

1. Run regression to estimate probability of treatment
   - Y= receipt of treatment (0/1)
   - X=variables that affect outcome of interest and selection into treatment
   - Usually, this is "logistic" or "probit" regression, rather than OLS
   - $\hat{Y}_i$ from this regression is the propensity score (probability of treatment)
2. Match treated units with untreated units that have similar propensity scores
3. For each treated i, calculate difference between Y_i and Y_i for matched untreated unit
4. Calculate ATT as average of these differences

Practical steps in propensity score analysis

Not necessarily in this order

1. Decide which treatment effect to estimate: ATE, ATT?
2. Identify Xs that drive Y and selection into treatment
   - Assess assumption 1: conditional independence
3. Calculate propensity scores from logit/probit regression
4. Choose method to match treated and untreated
   - e.g., "nearest neighbor"(NN); NN with caliper; kernal; inverse propensity weights
5. Assess assumptions 2 and 3 (will cover later)
   - Assumption 2 ("common support"): for each treated *i* is there untreated *i* with same propensity?
   - Assumption 3 ("balance"): after matching, are untreated similar to treated on observed characteristics?
6. Estimate causal effect (ATT, ATE)

Practical example of implementing matching using “nearest neighbor” approach

Choose method (algorithm) to match controls to treated

Generally, take a treated unit and find one or more control units with a similar propensity

- Key: face a tradeoff between bias and efficiency (standard errors)
- Limit matches to very good matches (very similar propensities)
  - End up with small sample size (high SEs) and little bias
- Allow matches to differ
  - Larger sample size (low SEs), but probably some bias, because matches are not as good

All depends on your particular application, the number of treated and control units, and the amount of common support

Nearest neighbor method to match controls to treated

- "Nearest neighbor"
  - Pick a treated unit, find control whose propensity score is closest
- With or without replacement
  - Without replacement: once a treated and control unit are matched, the control is pulled out of the pool
  - With replacement: single control can be matched to multiple treated units
  - Tradeoff between bias and efficiency
- Can also oversample (more than one control per treated); recommended if your sample is larger enough
- Generally not used by itself
  - Nearest neighbor could be very far away (i.e., a bad match)
  - Use of a caliper solves this problem

Nearest neighbor with caliper [SKIP/SKIM]

What is a caliper in real life?

- Define a caliper c, usually $c \leq .25\sigma$, where σ is calculated from the propensity score
- Search for nearest neighbor only among C's within the the caliper
  - Out of these C's, choose the nearest
- Avoids bad matches, but may result in fewer matches
- Caliper width is atheoretical
  - Should be convincing
  - Consider a caliper of .01 versus .10

Assess match quality [SKIP/SKIM]

Remember: goal of propensity model and matching process is to yield two groups that look alike for a given set of covariates

Several ways to see if this is the case

- Difference in means
- Standardized bias - difference between sample means as a percentage of square root of variances

$$100 * \frac{\bar{X}_T - \bar{X}_C}{\sqrt{.5(V(X_C) + V(X_C))}}$$

Can use tables or figures to display information

- For many covariates, plot distribution of bias before and after

Practical example: What is the effect of mother being a smoker (X) on birth weight (Y)?

Work through example in lecture 4 do-file:

- Will have to install some commands
- Dataset:
  - 4,642 observations
  - Each observation represents a mother

	Full sample	Smokers	Non-smokers	Δ
Prenatal visit 1st tri.	.80	.69	0.83	-.14
Married	.70	.47	0.75	-.28
Mother's age	26.5	25.2	26.8	-1.60
Mother's education	12.70	11.60	12.90	-1.30
n	4,642	864	3,778	

Mothers who smoke are less likely to see the doctor, be married, are younger, and have less education.