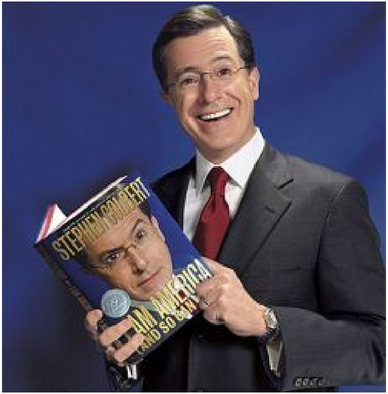
Lecture 5: Matching

Naijia Liu

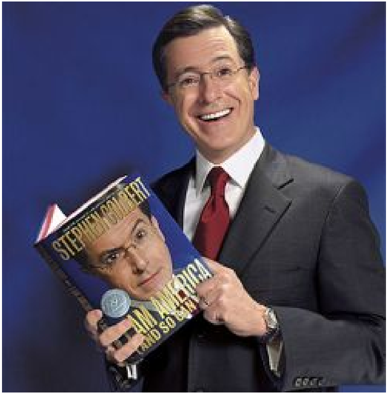
Feb. 6, 2024

Donation and Stephen Colbert



- **Colbert Bump**: Does a legislator's appearance on Stephen's show cause an increase in donations?

Donation and Stephen Colbert



- **Colbert Bump**: Does a legislator's appearance on Stephen's show cause an increase in donations?
- Appearances are not random.

Example: The Colbert Effect

- Does appearing on the Colbert Report cause an increase in donations? (Fowler, 2008)
- Colbert selects certain kinds of representatives.
- Potential ways to identify a **causal** effect:

Example: The Colbert Effect

- Does appearing on the Colbert Report cause an increase in donations? (Fowler, 2008)
- Colbert selects certain kinds of representatives.
- Potential ways to identify a **causal** effect:
  - Experiment: Randomly draw names to appear on the show.
  - DID: Assume parallel trend between selected and those who were left out.
  - IV: ???

Can we create a counterfactual case?

- Say Obama was invited to the show before 2008 election.
- We are interested in Colbert effect on Obama: what if he was not invited?
- A clone of Obama: **same** in every other aspects but the Colbert show appearance.

Can we create a counterfactual case?

- Say Obama was invited to the show before 2008 election.
- We are interested in Colbert effect on Obama: what if he was not invited?
- A clone of Obama: **same** in every other aspects but the Colbert show appearance.
- Same exercise for every other representative who appeared on the show.

$$\frac{\sum_i^N Y_i(\text{Appeared})}{N} - \frac{\sum_i^N Y_i(\text{Cloned})}{N}$$

Notation

- Treatment: T_i
- Potential Outcome: $Y_i(T_i)$
- Covariate: X_i

  For now we only consider pre-treatment covariates, such as gender, age, income.

Observed Data

	Pre-treatment	Treatment	Potential Outcomes	
Unit	Covariate	Indicator	Treated	Control
1				
2				
3				
4				
$\vdots$				
N				

Observed Data

	Pre-treatment	Treatment	Potential Outcomes	
Unit	Covariate	Indicator	Treated	Control
1				
2				
3				
4				
$\vdots$				
N				

Observed Data

	Pre-treatment	Treatment	Potential Outcomes	
Unit	Covariate	Indicator	Treated	Control
1	X_1			
2	X_2			
3	X_3			
4	X_4			
$\vdots$	$\vdots$			
N	X_N			

Eg: Gender of the Reps

Observed Data

	Pre-treatment	Treatment	Potential Outcomes	
Unit	Covariate	Indicator	Treated	Control
1	X_1	1		
2	X_2	0		
3	X_3	0		
4	X_4	1		
$\vdots$	$\vdots$	$\vdots$		
N	X_N	1		

Appeared on the show: 1

Observed Data

	Pre-treatment	Treatment	Potential Outcomes	
Unit	Covariate	Indicator	Treated	Control
1	X_1	1	$Y_1(1)$	
2	X_2	0	?	
3	X_3	0	?	
4	X_4	1	$Y_4(1)$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
N	X_N	1	$Y_N(1)$	

Donation amount for those who appeared on the show.

Observed Data

	Pre-treatment	Treatment	Potential Outcomes	
Unit	Covariate	Indicator	Treated	Control
1	X_1	1	$Y_1(1)$	$?$
2	X_2	0	$?$	$Y_2(0)$
3	X_3	0	$?$	$Y_3(0)$
4	X_4	1	$Y_4(1)$	$?$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X_N	1	$Y_N(1)$	$?$

Donation amount for those who did not appear on the show.

Matching to Better Approximate Counterfactuals

- Instead of clones, we can find similar **matches** of those who appeared on the show.

Matching to Better Approximate Counterfactuals

- Instead of clones, we can find similar **matches** of those who appeared on the show.
- Authors matched on incumbency, party, and donations in the previous 20 days.
- Potential concerns?

More Notations

- Potential Outcome of the matched: $Y_i^M(T_i)$

  The donation amount of the clone we found for representative i.

- Covariate of the matched: X_i^M

  Gender the clone we found for representative i.

Observed Data

Start with all those who appeared on the show.

	Pre-treatment Covariates	Treatment	Observed Outcomes	
Unit	Treated	Indicator	Treated	**Matched**
1	$X_1,$	1		
2	$X_2,$	1		
3	$X_3,$	1		
4	$X_4,$	1		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$X_N,$	1		

Observed Data

Find matches among those who did not appear.

	Pre-treatment Covariates	Treatment	Observed Outcomes	
Unit	Treated	Indicator	Treated	**Matched**
1	$X_1, \quad \mathbf{X_1^M}$	$1 \,, \mathbf{0}$		
2	$X_2, \quad \mathbf{X_2^M}$	$1 \,, \mathbf{0}$		
3	$X_3, \quad \mathbf{X_3^M}$	$1 \,, \mathbf{0}$		
4	$X_4, \quad \mathbf{X_4^M}$	$1 \,, \mathbf{0}$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$X_N, \quad \mathbf{X_N^M}$	$1 \,, \mathbf{0}$		

Observed Data

	Pre-treatment Covariates	Treatment	Observed Outcomes	
Unit	Treated, **Matched**	Indicator	Treated	**Matched**
1	$X_1, \ \mathbf{X_1^M}$	1, **0**	$Y_1(1)$	$Y_1^M(0)$
2	$X_2, \ \mathbf{X_2^M}$	1, **0**	$Y_1(1)$	$Y_2^M(0)$
3	$X_3, \ \mathbf{X_3^M}$	1, **0**	$Y_1(1)$	$Y_3^M(0)$
4	$X_4, \ \mathbf{X_4^M}$	1, **0**	$Y_1(1)$	$Y_4^M(0)$
⋮	⋮	⋮	⋮	⋮
N	$X_N, \ \mathbf{X_N^M}$	1, **0**	$Y_N(1)$	$Y_N^M(0)$

Observed Data

	Pre-treatment Covariates	Treatment	Observed Outcomes	
Unit	Treated, **Matched**	Indicator	Treated	**Matched**
1	$X_1, \ \mathbf{X_1^M}$	1, **0**	$Y_1(1)$	$Y_1^M(0)$
2	$X_2, \ \mathbf{X_2^M}$	1, **0**	$Y_1(1)$	$Y_2^M(0)$
3	$X_3, \ \mathbf{X_3^M}$	1, **0**	$Y_1(1)$	$Y_3^M(0)$
4	$X_4, \ \mathbf{X_4^M}$	1, **0**	$Y_1(1)$	$Y_4^M(0)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$X_N, \ \mathbf{X_N^M}$	1, **0**	$Y_N(1)$	$Y_N^M(0)$

Matching gives the Average Treatment Effect on the Treated (ATT)

Observed Data

	Pre-treatment Covariates	Treatment	Observed Outcomes	
Unit	Treated, **Matched**	Indicator	Treated	**Matched**
1	$X_1, \ \mathbf{X_1^M}$	1, **0**	$Y_1(1)$	$Y_1^M(0)$
2	$X_2, \ \mathbf{X_2^M}$	1, **0**	$Y_1(1)$	$Y_2^M(0)$
3	$X_3, \ \mathbf{X_3^M}$	1, **0**	$Y_1(1)$	$Y_3^M(0)$
4	$X_4, \ \mathbf{X_4^M}$	1, **0**	$Y_1(1)$	$Y_4^M(0)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$X_N, \ \mathbf{X_N^M}$	1, **0**	$Y_N(1)$	$Y_N^M(0)$

Matching gives the Average Treatment Effect on the Treated (ATT)
Becase we used the "clones" for those treated.

When Does Matching Work?

The argument with matching:

- If $X_i \approx \mathbf{X_i^M}$, then $Y_i(0) \approx \mathbf{Y_i^M(0)}$

When Does Matching Work?

The argument with matching:

- If $X_i \approx \mathbf{X_i^M}$, then $Y_i(0) \approx \mathbf{Y_i^M(0)}$
- To estimate the treatment effect

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\, Y_i(1) - \underbrace{Y_i(0)}_{\text{Unobserved}} \,\right\} \approx \frac{1}{N}\sum_{i=1}^{N}\left\{\, Y_i(1) - \mathbf{Y_i^M(0)} \right\}$$

Since attendance on the Report is not fully at random

- Estimate the counterfactual with the matched subset of the data
- May not work exactly for each observation; may work on average over the matched subset

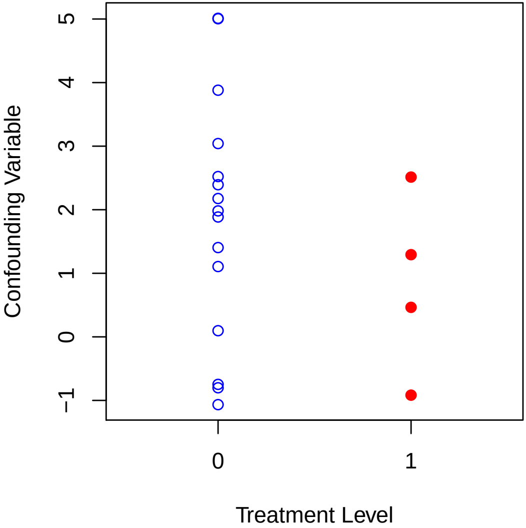
Types of Matching

- Exact Match

Types of Matching

- Exact Match
- Obama: African American, Democrats, in his 40s, male, law degree, married.
- Matched representative for Obama: African American, Democrats, in his 40s, male, law degree, married.

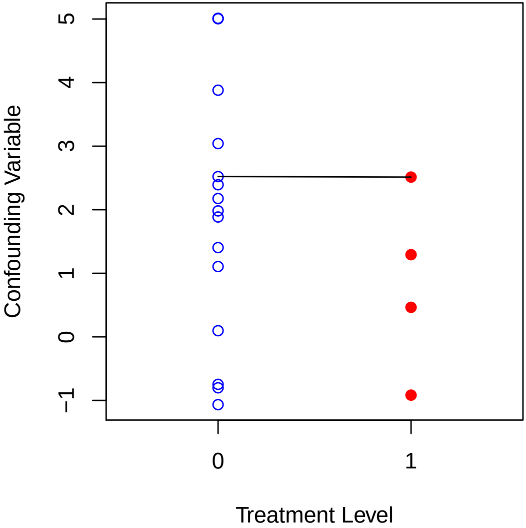
Matching on a Single Variable



In raw data

- Data not directly comparable

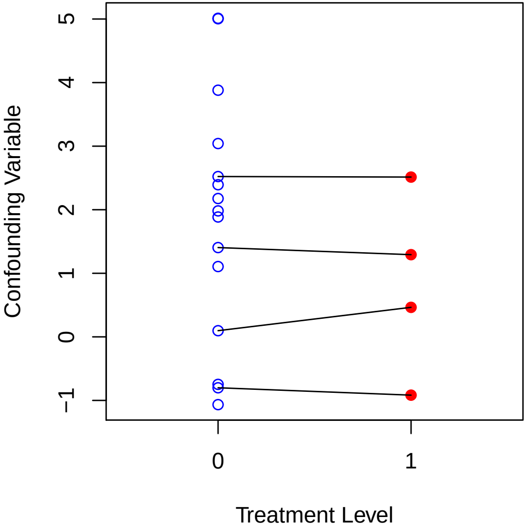
Matching on a Single Variable



In raw data

- Match to most similar observation

Matching on a Single Variable



In raw data

- Match to most similar observation
- Repeat
- Compare outcomes in matched subset

Exact Matching

- Almost perfect clones for the treated group.

Exact Matching

- Almost perfect clones for the treated group.
- Relies heavily on data quality.

  Almost impossible to find a clone Obama in reality.

Types of Matching

- Distance Matching

Types of Matching

- Distance Matching
  - Definition of distance under a multivariate context.

    Absolute value distance

    Mahalanobis distance matching

Types of Matching

- Distance Matching
  - Definition of distance under a multivariate context.

    Absolute value distance

    Mahalanobis distance matching
  - Pick a threshold

Types of Matching

- Distance Matching
  - Definition of distance under a multivariate context.

    Absolute value distance

    Mahalanobis distance matching
  - Pick a threshold
  - Matched if the distance between treated unit and control unit is less or equal to the threshold.

Distance Matching

Unit	GPA	Age	Treatment
1	3	21	1
2	3.3	19	0
3	2.9	17	0

Absolute Distance Matching

- Distance between unit 1 and 2:

$$|3 - 3.3| + |21 - 19| = 0.3 + 2 = 2.3$$

- Distance between unit 1 and 3:

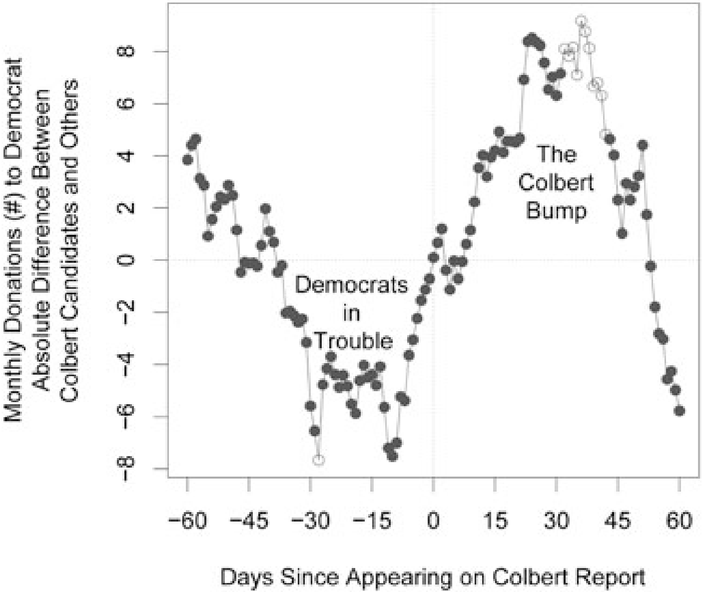
$$|3 - 2.9| + |21 - 17| = 0.1 + 4 = 4.1$$

- Unit 2 is a better match than unit 3.

Distance Matching

- More flexibility in finding matched units.
- What is the optimal distance measure + threshold?
  - Absolute distance is sensitive to outlier variables.
  - Mahalanobis distance is more well-utilized.

The Colbert Bounce?



Assumptions

- Probabilistic treatment: **Not deterministic that you will (or will never) be invited to the show.**

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

See more of these reviews for lecture 1 slides.

Assumptions

- Probabilistic treatment: **Not deterministic that you will (or will never) be invited to the show.**

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

- Ignorability: **If we were to know all covariates of candidates, invitation is almost randomly sent.**

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

See more of these reviews for lecture 1 slides.

Assumptions

- Probabilistic treatment: **Not deterministic that you will (or will never) be invited to the show.**

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

- Ignorability: **If we were to know all covariates of candidates, invitation is almost randomly sent.**

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

- SUTVA: **Other people's donation does not affect your donation amount. One single version of treatment level.**

$$Y_i(T_i, Y_i(T_i')) = Y_i(T_i) \quad \forall i \neq i'$$
$$Y_i(T_i) = Y_i(T_i') \text{ if } T_i = T_i'$$

See more of these reviews for lecture 1 slides.

Identifying assumptions

- **Valid matched set**: **On average, donation for those invited would have been the same to those in the matched set (not invited), if they were not invited.**

$$\underbrace{\mathbb{E}(Y_i(0) \mid T_i = 1)}_{\text{unobserved}} = \mathbb{E}(Y_i(0) \mid T_i = 0, i \in \mathcal{M})$$

More on Assumptions

- Critique on above assumptions.

More on Assumptions

- Critique on above assumptions.
  - What if observables are not enough?

    Aside from incumbency, party, previous donations, what are other variables that might matter?

More on Assumptions

- Critique on above assumptions.
  - What if observables are not enough?

    Aside from incumbency, party, previous donations, what are other variables that might matter?
  - Mediators?

    Next Lecture