Randomized Control Trials

Md Moshi Ul Alam

April 1, 2025

Clark University

(Recap) Treatment effects: Potential Outcomes and selection bias

Treatment D_i for each unit i with outcome Y_i

We observe only *one of the two* **potential outcomes** $Y_i(0)$ or $Y_i(1)$

We do not observe the counterfactual for each unit which leads to selection bias

ATT $= E\left[\mathrm{Y}_i(1) - \mathrm{Y}_i(0) \mid \mathrm{D}_i = 1\right]$ cannot be estimated directly from the data

We showed that observed differences between groups = ATT + Selection bias

Selection bias

$$\underbrace{E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right]}_{\text{Observed difference in avg outcomes}} = \quad E\left[Y_{1i} \mid D_i = 1\right] - E\left[Y_{0i} \mid D_i = 0\right]$$

$$= \quad \underbrace{E\left[Y_i(1) \mid D_i = 1\right] - E\left[Y_i(0) \mid D_i = 1\right]}_{\text{average treatment effect on the treated}}$$

$$\underbrace{+E\left[Y_i(0) \mid D_i = 1\right] - E\left[Y_i(0) \mid D_i = 0\right]}_{\text{selection bias}}$$

Avg differences in group outcomes is not causal evidence, because of selection bias

Selection bias

$$
\begin{aligned}
\underbrace{E\left[\mathrm{Y}_i \mid \mathrm{D}_i=1\right]-E\left[\mathrm{Y}_i \mid \mathrm{D}_i=0\right]}_{\text{Observed difference in avg outcomes}} = & \quad E\left[\mathrm{Y}_{1i} \mid \mathrm{D}_i=1\right]-E\left[\mathrm{Y}_{0i} \mid \mathrm{D}_i=0\right] \\
= & \quad \underbrace{E\left[\mathrm{Y}_i(1) \mid \mathrm{D}_i=1\right]-E\left[\mathrm{Y}_i(0) \mid \mathrm{D}_i=1\right]}_{\text{average treatment effect on the treated}} \\
& \quad \underbrace{+E\left[\mathrm{Y}_i(0) \mid \mathrm{D}_i=1\right]-E\left[\mathrm{Y}_i(0) \mid \mathrm{D}_i=0\right]}_{\text{selection bias}}
\end{aligned}
$$

Avg differences in group outcomes is not causal evidence, because of selection bias

Let us think of the drivers of selection bias

Other variables also could impact outcome

Q: Are people with health insurance more healthy than those without?

Q: Does having health insurance make people healthy?

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
			A. Health			
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
			B. Characteristics			
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Observed and unobserved differences

- Other differences between those who have and do not have insurance could drive the selection bias

Observed and unobserved differences

- Other differences between those who have and do not have insurance could drive the selection bias

- We do not have *ceteris paribus* situation of ALL ELSE EQUAL

Observed and unobserved differences

- Other differences between those who have and do not have insurance could drive the selection bias
- We do not have *ceteris paribus* situation of ALL ELSE EQUAL
- Part of the selection bias is driven by *observable differences*

Observed and unobserved differences

- Other differences between those who have and do not have insurance could drive the selection bias
- We do not have *ceteris paribus* situation of ALL ELSE EQUAL
- Part of the selection bias is driven by *observable differences*
- Have data, so we can *control for those differences*

Observed and unobserved differences

- Other differences between those who have and do not have insurance could drive the selection bias
- We do not have *ceteris paribus* situation of ALL ELSE EQUAL
- Part of the selection bias is driven by *observable differences*
- Have data, so we can *control for those differences*
- **Challenge**: The selection bias driven by *unobservables* !

Observed and unobserved differences

- Other differences between those who have and do not have insurance could drive the selection bias
- We do not have *ceteris paribus* situation of ALL ELSE EQUAL
- Part of the selection bias is driven by *observable differences*
- Have data, so we can *control for those differences*
- **Challenge**: The selection bias driven by *unobservables* !
  – Always!

Randomized Control Trials

Random assignment eliminates selection bias: Example

- Select an uninsured group
- Randomly assign health insurance (coin tosses)
- Compare health outcomes of those with and without insurance

Random assignment eliminates selection bias: Example

- Select an uninsured group
- Randomly assign health insurance (coin tosses)
- Compare health outcomes of those with and without insurance
- Random assignment makes the comparison *ceteris paribus*

Random assignment eliminates selection bias: Example

- Select an uninsured group
- Randomly assign health insurance (coin tosses)
- Compare health outcomes of those with and without insurance
- Random assignment makes the comparison *ceteris paribus*
- A coin toss is independent of all observable or unobservable that could have driven the choice to buy health insurance and impacted health outcomes.

Let's formalize this!

Randomized Control Trials (RCT)

- Imagine a situation where the treatment $D_i = \{0, 1\}$ is randomly assigned.
- Randomization $\implies D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$

Mathematically working out

$$\underbrace{E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right]}_{\text{Observed difference in avg outcomes}} = E\left[Y_i(1) \mid D_i = 1\right] - E\left[Y_i(0) \mid D_i = 0\right]$$

$$= \underbrace{E\left[Y_i(1) \mid D_i = 1\right] - E\left[Y_i(0) \mid D_i = 1\right]}_{\text{ATT}}$$

$$\underbrace{+E\left[Y_i(0) \mid D_i = 1\right] - E\left[Y_i(0) \mid D_i = 0\right]}_{\text{selection bias}}$$

Mathematically working out

$$\underbrace{E\left[\mathrm{Y}_i \mid \mathrm{D}_i=1\right]-E\left[\mathrm{Y}_i \mid \mathrm{D}_i=0\right]}_{\text{Observed difference in avg outcomes}} = E\left[\mathrm{Y}_i(1) \mid \mathrm{D}_i=1\right]-E\left[\mathrm{Y}_i(0) \mid \mathrm{D}_i=0\right]$$

$$= \underbrace{E\left[\mathrm{Y}_i(1) \mid \mathrm{D}_i=1\right]-E\left[\mathrm{Y}_i(0) \mid \mathrm{D}_i=1\right]}_{\text{ATT}}$$

$$\underbrace{+E\left[\mathrm{Y}_i(0) \mid \mathrm{D}_i=1\right]-E\left[\mathrm{Y}_i(0) \mid \mathrm{D}_i=0\right]}_{\text{selection bias}}$$

- In a RCT: selection bias $=0$ by design
- Randomly assigned $D_i=1$ and $D_i=0$ come from the same population -> should have the same average $Y_i(0)$
- Thus, in a RCT
$E[Y_i \mid D_i=1]-E[Y_i \mid D_i=0]=E[Y_i(1) \mid D_i=1]-E[Y_i(0) \mid D_i=1] \equiv ATT$

When can we estimate these expectations by using corresponding sample average?

Estimation

RCT and regressions

$$ATT = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right] \tag{1}$$

In a RCT regressing Y_i on D_i gives us an unbiased and consistent estimate of the avg causal effect of D_i on Y_i

$$\mathbb{E}\left[Y_i \mid D_i = 1\right] - \mathbb{E}\left[Y_i \mid D_i = 0\right] = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right] \tag{2}$$

Running a regression of Y on D:

$$\begin{aligned} & Y_i = \alpha + \beta D_i + u_i \\ \implies\ & \beta = \underbrace{E\left[Y_i \mid D_i = 1\right]}_{\alpha+\beta+E[u_i|D_i=1]} - \underbrace{E\left[Y_i \mid D_i = 0\right]}_{\alpha+E[u_i|D_i=0]} \quad \text{if} \quad E\left[u_i \mid D_i\right] = 0 \end{aligned}$$

RCT and regressions

$$ATT = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right] \tag{1}$$

In a RCT regressing Y_i on D_i gives us an unbiased and consistent estimate of the avg causal effect of D_i on Y_i

$$\mathbb{E}\left[Y_i \mid D_i = 1\right] - \mathbb{E}\left[Y_i \mid D_i = 0\right] = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right] \tag{2}$$

Running a regression of Y on D:

$$\begin{aligned} & Y_i = \alpha + \beta D_i + u_i \\ \implies \; & \beta = \underbrace{E[Y_i \mid D_i = 1]}_{\alpha+\beta+E[u_i|D_i=1]} - \underbrace{E[Y_i \mid D_i = 0]}_{\alpha+E[u_i|D_i=0]} \quad \text{if} \quad E[u_i \mid D_i] = 0 \end{aligned}$$

In other words, selection bias = $E[u_i \mid D_i = 1] - E[u_i \mid D_i = 0]$

This works because of the **random** assignment of D_i

This works because of the **random** assignment of D_i

- The random assignment of D_i makes the TG & CG identical on average including unobservables

- Since D_i is random $\implies cov(D_i, u_i) = 0 \implies E[u_i \mid D_i] = 0$

  - $E[Y_i(1) \mid D_i = 1] = E[Y_i(1) \mid D_i = 0]$
    Potential outcomes are independent of treatment assignment

Why control for other covariates X?

- If D_i is randomly assigned then estimating $Y_i = \alpha + \beta D_i + e_i$ gives unbiased and consistent estimate of ATT of D_i on Y_i
- Then why estimate $Y_i = \alpha + \beta D_i + \gamma X_i + u_i$?

To see unbiasedness of β in $Y_i = \alpha + \beta D_i + \gamma X_i + u_i$

$$E[Y_i \mid D_i = 1, X_i] - E[Y_i \mid D_i = 0, X_i]$$

To see unbiasedness of β in $Y_i = \alpha + \beta D_i + \gamma X_i + u_i$

$$E[Y_i \mid D_i = 1, X_i] - E[Y_i \mid D_i = 0, X_i]$$
$$= E[Y_i(1) \mid D_i = 1, X_i] - E[Y_i(0) \mid D_i = 0, X_i]$$

To see unbiasedness of β in $Y_i = \alpha + \beta D_i + \gamma X_i + u_i$

$$E[Y_i \mid D_i = 1, X_i] - E[Y_i \mid D_i = 0, X_i]$$

$$= E[Y_i(1) \mid D_i = 1, X_i] - E[Y_i(0) \mid D_i = 0, X_i]$$

$$= \underbrace{E[Y_i(1) \mid D_i = 1, X_i] - E[Y_i(0) \mid D_i = 1, X_i]}_{ATT} \quad \underbrace{+E[Y_i(0) \mid D_i = 1, X_i] - E[Y_i(0) \mid D_i = 0, X_i]}_{\text{selection bias}}$$

- Random assignment of $D_i \implies E[Y_i(0) \mid D_i = 1, X_i] = E[Y_i(0) \mid D_i = 0, X_i]$

To see unbiasedness of β in $Y_i = \alpha + \beta D_i + \gamma X_i + u_i$

$$E[Y_i \mid D_i = 1, X_i] - E[Y_i \mid D_i = 0, X_i]$$

$$= E[Y_i(1) \mid D_i = 1, X_i] - E[Y_i(0) \mid D_i = 0, X_i]$$

$$= \underbrace{E[Y_i(1) \mid D_i = 1, X_i] - E[Y_i(0) \mid D_i = 1, X_i]}_{ATT} \quad \underbrace{+E[Y_i(0) \mid D_i = 1, X_i] - E[Y_i(0) \mid D_i = 0, X_i]}_{\text{selection bias}}$$

- Random assignment of $D_i \implies E[Y_i(0) \mid D_i = 1, X_i] = E[Y_i(0) \mid D_i = 0, X_i]$
- So far we had: $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$
- Now we have implied by the above: $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) \mid X_i$

What happens to the SE of β?

Estimating $Y_i = \alpha + \beta D_i + e_i$ VS $Y_i = \alpha + \beta D_i + \gamma X_i + u_i$

Let us code and work out the math too!

Things to take care of in RCTs

1. Was the randomization successful?

1. Description of the implementation of random assignment [Give eg.]

1. Was the randomization successful?

1. Description of the implementation of random assignment [Give eg.]
2. BALANCE between TG and CG

1. Was the randomization successful?

1. Description of the implementation of random assignment [Give eg.]
2. BALANCE between TG and CG
   - Test for any diff in pre-treatment outcomes and covariates
   - Test for any standardized diff in pre-treatment outcomes and covariates (Why? Draw example)

1. Was the randomization successful?

1. Description of the implementation of random assignment [Give eg.]
2. BALANCE between TG and CG
   - Test for any diff in pre-treatment outcomes and covariates
   - Test for any standardized diff in pre-treatment outcomes and covariates (Why? Draw example)
3. Sample size

1. Was the randomization successful?

1. Description of the implementation of random assignment [Give eg.]
2. BALANCE between TG and CG
   - Test for any diff in pre-treatment outcomes and covariates
   - Test for any standardized diff in pre-treatment outcomes and covariates (Why? Draw example)
3. Sample size
4. Fractions treated in the treatment group and in the control group *[We will revisit this after we have learnt IV]*. For now we will assume perfect compliance.

2. Design: Stratification and Balance

E.g. treatment to only boys in a class

2. Design: Stratification and Balance

E.g. treatment to only boys in a class

- Stratify to improve balance b/w TG and CG
- Create strata and then randomize. [Give example and depth]

The idea here is that conditional on X_i the assignment of D_i is random.

$$D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) \mid X_i$$

3. Spillovers

- The pure treatment effect may not be correctly estimated if there are spillovers from treated units to untreated units

- Positive spillovers will lead to underestimate treatment effects

- Depends on the level at which randomization is done

- Eg. treating kids for de-worming in Africa at the level of schools instead at the individual level (Miguel and Kremer, 2004 Econometrica)

- $3.50 -> one additional year of attendance. Much cheaper intervention to improve absenteeism than other policies like better teachers, free books etc.

Issues with implementing RCT

Primarily logistic

- Long duration
- Political economy interacts with implementation
- Very high costs and thus scale issues
- Sometimes completely infeasible

Scaling up is quite a different problem, at times because of equilibrium effects etc.

While trying to answer a causal question it is best to think about the ideal RCT

Other Data Contexts and Identification Strategies

- Experimental:
  - Randomized Control Trials

Other Data Contexts and Identification Strategies

- Experimental:
  - Randomized Control Trials
- Non-experimental:
  - Regressions *Already covered*
  - Matching *(won't cover in this course)*

Other Data Contexts and Identification Strategies

- Experimental:
  - Randomized Control Trials
- Non-experimental:
  - Regressions *Already covered*
  - Matching *(won't cover in this course)*
- Quasi-experimental / Natural experiments: "As good as randomly assigned"
  - Instrumental Variables
  - Regression Discontinuity Design
  - Fixed effects / Random Effects (panel data)
  - Difference-in-differences *(will cover depending on time)*

Podcasts on RCT:

How do we know what really works in healthcare? – Freakonomics (April 2, 2015)
Amy Finkelstein (MIT) on the impact of health insurance (Oregon Medicaid expansion experiment). Steve Levitt here initially gives a small introduction to RCT.

Whats not on the test – Hidden Brain (May 13, 2019)
Jim Heckman (U-Chicago) on the impact of early childhood investments on long run outcomes

When you start to miss Tony from Accounting – Hidden Brain (Nov 16, 2020)
Nicholas Bloom (Stanford) on the impact of working from home on productivity

The Price of Doing Business with John List – People I (Mostly) Admire (Dec 9, 2022)
John List (U-Chicago) on the rise of RCT and worries of scaling up experiments (SUTVA violations)