Average Treatment Effects

David S. Rosenberg

NYU: CDS

February 19, 2021

Contents

1. Neyman-Rubin potential outcome framework
2. Randomized control trials (RCTs)
3. Average treatment effect and RCTs
4. Randomization on the basis of a covariate

Neyman-Rubin potential outcome framework

Treatments

- Suppose we want to know
  - whether a new medicine improves outcomes
  - whether a new webpage layout keeps people on the page longer
  - whether sending somebody a particular postcard increases their probability of donating money
- We can think of each of these as a **treatment**.
- Our **goal** is to understand the effect of a treatment on an outcome measure.

Individual Treatment Effects

- Suppose we have an individual i
- What's the effect of giving a treatment to i?
- Let $Y_i(1) \in \mathbb{R}$ be the "**potential outcome**" if we **give** the treatment.
- Let $Y_i(0) \in \mathbb{R}$ be the "**potential outcome**" if we **do not give** the treatment
- The **individual treatment effect** for individual i is defined as

$$D_i = Y_i(1) - Y_i(0).$$

- The problem is, we never observe $Y_i(1)$ and $Y_i(0)$ for the same person!
- Some call this the **fundamental problem of causal inference**.
- We're going to think about this as a missing data problem.

Treatment assignment indicator

- $W_i \in \{0,1\}$ is the **treatment indicator** for individual i:

$$W_i = \begin{cases} 0 & \text{if individual } i \text{ does \textbf{not} receive the treatment} \\ 1 & \text{if individual } i \text{ receives the treatment} \end{cases}$$

- When $W_i = 1$, we observe $Y_i(1)$ but not $Y_i(0)$.
- When $W_i = 0$, we observe $Y_i(0)$ but not $Y_i(1)$.
- The group of individuals with $W_i = 0$ is called the **control group**.
- The group of individuals with $W_i = 1$ is called the **treatment group**.

What we observe

- We'll write the **observed data** $\mathcal{D}$ as

$$(W_1, Y_1), \ldots, (W_n, Y_n),$$

where

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0 \\ Y_i(1) & \text{if } W_i = 1 \end{cases}.$$

What we observe

Id	W	$Y(0)$	$Y(1)$	$D = Y(1) - Y(0)$
1	0	1.2	?	?
2	0	2.3	?	?
3	1	?	8.6	?
4	0	.7	?	?
5	1	?	3.4	?

Variation on missing data problem

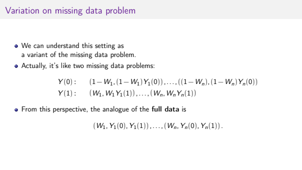
- We can understand this setting as a variant of the missing data problem.
- Actually, it's like two missing data problems:

$$\begin{array}{ll} Y(0): & (1-W_1,(1-W_1)Y_1(0)),\ldots,((1-W_n),(1-W_n)Y_n(0)) \\ Y(1): & (W_1,W_1Y_1(1)),\ldots,(W_n,W_nY_n(1)) \end{array}$$

- From this perspective, the analogue of the **full data** is

$$(W_1,Y_1(0),Y_1(1)),\ldots,(W_n,Y_n(0),Y_n(1)).$$

└Variation on missing data problem



Variation on missing data problem

- We can understand this setting as
  a variant of the missing data problem.
- Actually, it's like two missing data problems:

$$\begin{aligned} Y(0): &\quad (1-W_1,(1-W_1)Y_1(0)),\ldots,((1-W_n),(1-W_n)Y_n(0)) \\ Y(1): &\quad (W_1,W_1Y_1(1)),\ldots,(W_n,W_nY_n(1)) \end{aligned}$$

- From this perspective, the analogue of the **full data** is

$$(W_1,Y_1(0),Y_1(1)),\ldots,(W_n,Y_n(0),Y_n(1)).$$

- We'll refer to the two missing data problems as "$Y(0)$" and "$Y(1)$".
- The response indicators for these problems are $1-W$ and W, respectively.

Randomized control trials (RCTs)

Time for probabilistic assumptions

- So far, we've made no probabilistic assumptions.
- First we'll introduce the **randomized control trial (RCT)**.
    - This will be the experiment analogue of MCAR.
- Next we'll introduce the **"unconfoundedness**" or "**ignorability**" assumption
    - This will be the experiment analogue of MAR.

General idea

- Randomly sample individuals $i = 1, \ldots, n$ from a population.
- Each individual i is assigned to one of two groups:

  control group: individuals do not receive the treatment

  treatment group: individuals do receive the treatment
- Individuals **assigned randomly** to treatment and control groups
  - Simplest: Individuals assigned by a flipping a **fair coin** (RCT)
  - Simple: Individuals assigned by a flipping a **biased coin** (RCT)
  - Less simple: bias of coin depends on **covariates** (unconfoundedness)

Randomized control trial (RCT)

In a randomized control trial (RCT), we assume
$(W, Y(0), Y(1)), (W_1, Y_1(0), Y_1(1)), \ldots, (W_n, Y_n(0), Y_n(1))$
are i.i.d. subject to the following assumption:

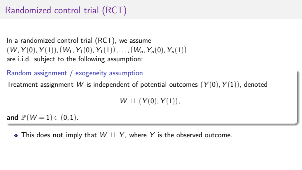
Random assignment / exogeneity assumption

Treatment assignment W is independent of potential outcomes $(Y(0), Y(1))$, denoted

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

and $\mathbb{P}(W = 1) \in (0, 1)$.

- This does **not** imply that $W \perp\!\!\!\perp Y$, where Y is the observed outcome.



Randomized control trial (RCT)

In a randomized control trial (RCT), we assume
$(W, Y(0), Y(1)), (W_1, Y_1(0), Y_1(1)), \ldots, (W_n, Y_n(0), Y_n(1))$
are i.i.d. subject to the following assumption:

Random assignment / exogeneity assumption

Treatment assignment W is independent of potential outcomes $(Y(0), Y(1))$, denoted

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

and $\mathbb{P}(W = 1) \in (0,1)$.

- This does **not** imply that $W \perp\!\!\!\perp Y$, where Y is the observed outcome.

- Note that we've added a generic individual $(W, Y(0), Y(1))$ that has the same distribution as any one in our sample. This is a common trick in probability and statistics to clean up notation. Since individuals are i.i.d., we could talk about $\mathbb{E}Y_1(0)$ or $\mathbb{E}Y_i(0)$ and they're both the same. So by introducing $(W, Y(0), Y(1))$ we can just drop the subscript.

- Recall that the **observed outcome** is $Y = (1-W)Y(0) + WY(1)$.
- Should be clear why we want $\mathbb{P}(W = 1) \in (0,1)$?
- Making the connection to the missing data setting, the exogeneity assumption implies that the $Y(0)$ and $Y(1)$ missing data problems are both MCAR, since exogeneity implies $(1-W) \perp\!\!\!\perp Y(0)$ and $W \perp\!\!\!\perp Y(1)$ and .
- In words, exogeneity says that even if we know an individual's potential outcomes to treatment and control (i.e. $Y(0)$ and $Y(1)$), this would give no information on whether the individual was assigned the treatment. This precludes doctors giving treatments to individuals who they believe are more likely to benefit (assuming the doctors' predictions are not completely independent of reality).

Connection to MCAR

- Consider the exogeneity assumption $W \perp\!\!\!\perp (Y(0), Y(1))$.
- Exogeneity implies
  - $Y(0)$ missing data problem is MCAR (i.e. $(1-W) \perp\!\!\!\perp Y(0)$)
  - $Y(1)$ missing data problem is MCAR (i.e. $W \perp\!\!\!\perp Y(1)$)
- Where $1-W$ and W are the respective response indicators.

Average treatment effect and RCTs

Average treatment effect

- Define the **average treatment effect** as

$$\text{ATE} := \mathbb{E}\left[Y(1) - Y(0)\right].$$

- If we had full data, we could use the natural estimator:

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i(1) - Y_i(0)\right).$$

- Unfortunately, either $Y_i(1)$ or $Y_i(0)$ is missing in every summand.

Difference-in-means estimator

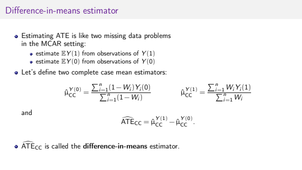
- Estimating ATE is like two missing data problems in the MCAR setting:
  - estimate $\mathbb{E}Y(1)$ from observations of $Y(1)$
  - estimate $\mathbb{E}Y(0)$ from observations of $Y(0)$
- Let's define two complete case mean estimators:

$$\hat{\mu}_{\text{CC}}^{Y(0)} = \frac{\sum_{i=1}^{n}(1-W_i)Y_i(0)}{\sum_{i=1}^{n}(1-W_i)} \qquad \hat{\mu}_{\text{CC}}^{Y(1)} = \frac{\sum_{i=1}^{n}W_iY_i(1)}{\sum_{i=1}^{n}W_i}$$

  and

$$\widehat{\text{ATE}}_{\text{CC}} = \hat{\mu}_{\text{CC}}^{Y(1)} - \hat{\mu}_{\text{CC}}^{Y(0)}.$$

- $\widehat{\text{ATE}}_{\text{CC}}$ is called the **difference-in-means** estimator.



Difference-in-means estimator

- Estimating ATE is like two missing data problems in the MCAR setting:
  - estimate $\mathbb{E}Y(1)$ from observations of $Y(1)$
  - estimate $\mathbb{E}Y(0)$ from observations of $Y(0)$
- Let's define two complete case mean estimators:

$$\hat{\mu}_{\mathrm{CC}}^{Y(0)} = \frac{\sum_{i=1}^{n}(1-W_i)Y_i(0)}{\sum_{i=1}^{n}(1-W_i)} \qquad \hat{\mu}_{\mathrm{CC}}^{Y(1)} = \frac{\sum_{i=1}^{n}W_iY_i(1)}{\sum_{i=1}^{n}W_i}$$

and

$$\widehat{\mathrm{ATE}}_{\mathrm{CC}} = \hat{\mu}_{\mathrm{CC}}^{Y(1)} - \hat{\mu}_{\mathrm{CC}}^{Y(0)}.$$

- $\widehat{\mathrm{ATE}}_{\mathrm{CC}}$ is called the **difference-in-means** estimator.

- We only use $Y_i(1)$ when $W_i = 1$ and $Y_i(0)$ and $W_i = 0$. So we only need the observed data to use these estimators.
- We could also have written these estimators by replacing both $Y_i(1)$ and $Y_i(0)$ with the observed outcome Y_i.

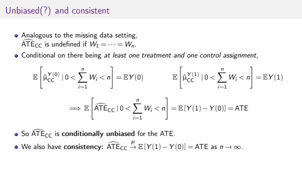
Unbiased(?) and consistent

- Analogous to the missing data setting, $\widehat{\text{ATE}}_{\text{CC}}$ is undefined if $W_1 = \cdots = W_n$.
- Conditional on there being *at least one treatment and one control assignment*,

$$\mathbb{E}\left[\hat{\mu}_{\text{CC}}^{Y(0)} \mid 0 < \sum_{i=1}^{n} W_i < n\right] = \mathbb{E}Y(0) \qquad \mathbb{E}\left[\hat{\mu}_{\text{CC}}^{Y(1)} \mid 0 < \sum_{i=1}^{n} W_i < n\right] = \mathbb{E}Y(1)$$

$$\implies \mathbb{E}\left[\widehat{\text{ATE}}_{\text{CC}} \mid 0 < \sum_{i=1}^{n} W_i < n\right] = \mathbb{E}\left[Y(1) - Y(0)\right] = \text{ATE}$$

- So $\widehat{\text{ATE}}_{\text{CC}}$ is **conditionally unbiased** for the ATE.
- We also have **consistency:** $\widehat{\text{ATE}}_{\text{CC}} \xrightarrow{P} \mathbb{E}\left[Y(1) - Y(0)\right] = \text{ATE}$ as $n \to \infty$.



Unbiased(?) and consistent

- Analogous to the missing data setting, $\widehat{\text{ATE}}_{\text{CC}}$ is undefined if $W_1 = \cdots = W_n$.
- Conditional on there being *at least one treatment and one control assignment*,

$$\mathbb{E}\left[\hat{\mu}_{\text{CC}}^{Y(0)} \mid 0 < \sum_{i=1}^{n} W_i < n\right] = \mathbb{E}Y(0) \qquad \mathbb{E}\left[\hat{\mu}_{\text{CC}}^{Y(1)} \mid 0 < \sum_{i=1}^{n} W_i < n\right] = \mathbb{E}Y(1)$$

$$\implies \mathbb{E}\left[\widehat{\text{ATE}}_{\text{CC}} \mid 0 < \sum_{i=1}^{n} W_i < n\right] = \mathbb{E}[Y(1) - Y(0)] = \text{ATE}$$

- So $\widehat{\text{ATE}}_{\text{CC}}$ is **conditionally unbiased** for the ATE.
- We also have **consistency:** $\widehat{\text{ATE}}_{\text{CC}} \xrightarrow{P} \mathbb{E}[Y(1) - Y(0)] = \text{ATE}$ as $n \to \infty$.

- In practice, if n is so small that there's a real chance of an empty treatment or control group, one would probably want to take a different approach to doing the treatment assignments. Rather than flipping a coin to determine each treatment assignment, one would split the n individuals into groups of equal size. The mathematics of this approach is a bit different, and it's not as relevant to our other applications. That's the perspective taken by (Rosenbaum 2017) and (Rosenbaum 2002), for example.

- From our study of the complete case estimator in MCAR, we know that $\hat{\mu}_{\text{CC}}^{Y(0)} \xrightarrow{P} \mathbb{E}Y(0)$ and $\hat{\mu}_{\text{CC}}^{Y(1)} \xrightarrow{P} \mathbb{E}Y(1)$ as $n \to \infty$. Note that this requires $\mathbb{P}(W = 0) \in (0, 1)$. Therefore, by Slutsky's theorem, $\widehat{\text{ATE}}_{\text{CC}} = \hat{\mu}_{\text{CC}}^{Y(1)} - \hat{\mu}_{\text{CC}}^{Y(0)} \xrightarrow{P} \mathbb{E}[Y(1) - Y(0)]$ as $n \to \infty$. So the estimator is consistent.

Randomization on the basis of a covariate

Relaxing random assignment / exogeneity

- Rubin 1977 speaks of
  "assignment to treatment group on the basis of a covariate."
- Think of assigning an individual to treatment or control by a coin toss
  - but the coin has a different bias depending on the covariates / features of the individual
- In econometrics, this is known as
  - **conditional exogeneity** or the
  - **conditional independence assumption**.
- We'll refer to it as **ignorability** or **unconfoundedness.**

Introducing a covariate

- For each individual i,
  - we'll associate a covariate $X_i \in \mathcal{X}$.
- Then our **full data** is

$$(X, W, Y(0), Y(1)), (X_1, W_1, Y_1(0), Y_1(1)), \ldots, (X_n, W_n, Y_n(0), Y_n(1)),$$

which we assume are i.i.d.

Ignorability / unconfoundedness assumption

Ignorability / unconfoundedness assumption

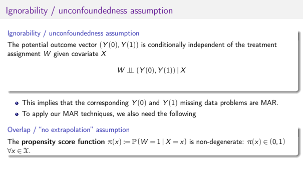
The potential outcome vector $(Y(0), Y(1))$ is conditionally independent of the treatment assignment W given covariate X

$$W \perp\!\!\!\perp (Y(0), Y(1)) \mid X$$

- This implies that the corresponding $Y(0)$ and $Y(1)$ missing data problems are MAR.
- To apply our MAR techniques, we also need the following

Overlap / "no extrapolation" assumption

The **propensity score function** $\pi(x) := \mathbb{P}(W = 1 \mid X = x)$ is non-degenerate: $\pi(x) \in (0,1)$ $\forall x \in \mathcal{X}$.



Ignorability / unconfoundedness assumption

Ignorability / unconfoundedness assumption

The potential outcome vector $(Y(0), Y(1))$ is conditionally independent of the treatment assignment W given covariate X

$$W \perp\!\!\!\perp (Y(0), Y(1)) \mid X$$

- This implies that the corresponding $Y(0)$ and $Y(1)$ missing data problems are MAR.
- To apply our MAR techniques, we also need the following

Overlap / "no extrapolation" assumption

The **propensity score function** $\pi(x) := \mathbb{P}(W = 1 \mid X = x)$ is non-degenerate: $\pi(x) \in (0,1)$ $\forall x \in \mathcal{X}$.

- In words: are once we observe X, knowing the potential outcomes $Y(0)$ and $Y(1)$ would give no additional information about treatment assignment W.
- By analogy with the MAR terminology, we might want to call the ignorability assumption "assignment at random", and by analogy with MCAR we might call exogeneity / random assignment "assignment completely at random." Unfortunately, nobody uses these terms, so we won't either.
- The overlap assumptions will imply that the propensity score is strictly positive for both the $Y(0)$ and the $Y(1)$ missing data problems.

Implications

- Under the assumptions of
  1. ignorability / unconfoundedness and
  2. overlap / “no extrapolation”
- We can treat ATE estimation as two missing data problems under MAR.
- All of our estimators in the missing data setting can be applied as ATE estimators.

The IPW estimator

- Let's estimate $\mathbb{E}Y(1)$ and $\mathbb{E}Y(0)$ using missing data strategies.
- Let's try the IPW mean estimator:

$$\hat{\mu}_{\text{ipw}}^{Y(1)} := \sum_{i=1}^{n} \frac{W_i Y(1)}{\pi(X_i)} \qquad \hat{\mu}_{\text{ipw}}^{Y(0)} := \sum_{i=1}^{n} \frac{(1-W_i) Y(0)}{1-\pi(X_i)}$$

- Putting this together gives us

$$\widehat{\text{ATE}}_{\text{ipw}} := \hat{\mu}_{\text{ipw}}^{Y(1)} - \hat{\mu}_{\text{ipw}}^{Y(0)}.$$

- This is **unbiased** for ATE since the estimators for $\mathbb{E}Y(1)$ and $\mathbb{E}Y(0)$ are unbiased.
- This is **consistent** for ATE since the estimators for $\mathbb{E}Y(1)$ and $\mathbb{E}Y(0)$ are consistent.

And so on...

- We can build ATE estimators using
  - self-normalized IPW estimators (haven't seen this in the literature, but surely somebody has done it)
  - regression imputation estimators
  - augmented IPW estimators (informally referred to as "the doubly robust estimator")
- To summarize, the common assumptions made in ATE estimation allow us to reduce ATE estimation to two missing data problems.
- This works well for us, since we now have a pretty thorough understanding of missing data problems :).

References

Resources

- Terminology was based primarily on course notes from Chernozhukov and Fernández-Val 2017 and Chapters 1 and 2 from Stefan Wager's course notes for Stats 361: Causal Inference (Stanford, Spring 2020) Wager 2020, Ch 1, 2.
- A lot of doing analysis of experiments correctly isn't really about hard math. It's about finding the right math for the situation at hand. There is much more to this than we can teach in a week. To get into it, I highly recommend Paul Rosenbaum's Observation & Experiment Rosenbaum 2017. It's a wonderful, couch-readable, book about randomized experiments and observational studies, with basically no math you can't do in your head. As mentioned in the note above, one subtle difference between his treatment and ours is that he assumes treatment and control groups are assigned by randomly splitting the subjects into two [equally sized] groups, while we assume treatment assignment by independent coin flips. The mathematics are slightly different, but the important principles are the same.

References I

Chernozhukov, Victor and Iván Fernández-Val (2017). "Treatment effects". In: *Econometrics—MIT Course 14.382*. MIT OpenCourseWare. Cambridge MA. URL: https://ocw.mit.edu/courses/economics/14-382-econometrics-spring-2017/lecture-notes/MIT14_382S17_lec12.pdf.

Rosenbaum, Paul R. (2002). *Observational Studies*. Springer Series in Statistics. Springer New York. DOI: 10.1007/978-1-4757-3692-2. URL: https://doi.org/10.1007/978-1-4757-3692-2.

– (2017). *Observation and experiment: an introduction to causal inference*. Harvard University Press.

Rubin, Donald B. (1977). "Assignment To Treatment Group on the Basis of a Covariate". In: *Journal of Educational Statistics* 2.1, p. 1. DOI: 10.2307/1164933. URL: https://doi.org/10.2307/1164933.

References II

Wager, Stefan (2020). *STATS 361: Causal Inference (Lecture notes)*. https://web.stanford.edu/~swager/stats361.pdf. [Online; accessed 14-February-2020].