Selection on Observables

Francis J. DiTraglia

University of Oxford

Core Empirical Research Methods

Potential Outcomes Framework[1]

- Binary **Treatment** $D \in \{0, 1\}$
- **Observed Outcome** Y depends on **Potential Outcomes** (Y_0, Y_1) via

$$Y = (1 - D)Y_0 + DY_1 = Y_0 + D(Y_1 - Y_0)$$

- Only one of (Y_0, Y_1) is observed for any given person at any given time.
- The unobserved potential outcome is a **counterfactual**, i.e. a **what if?**
- **Average Treatment Effect:** $\text{ATE} \equiv \mathbb{E}(Y_1 - Y_0)$.
- **Treatment on the Treated:** $\text{TOT} \equiv \mathbb{E}(Y_1 - Y_0 | D = 1)$.

[1] Videos: https://expl.ai/QHUAVRV and https://expl.ai/DWVNRZU.

Example: Y is Wage, D is Attend University

Counterfactuals

- $D = 1 \implies Y_0$ is the wage you *would have earned* if you *hadn't* attended.
- $D = 0 \implies Y_1$ is the wage you *would have earned* if you *had* attended.

Treatment Effects

- $\text{ATE} = \mathbb{E}(Y_1 - Y_0)$ is the average effect of *forcing* a randomly-chosen person to attend university.
- $\text{TOT} = \mathbb{E}(Y_1 - Y_0 | D = 1)$ is the average effect of attending university *for the sort of people who choose to attend voluntarily*.

Problem: Selection Bias

- We don't force randomly-chosen people to attend university!
- People who choose to attend are likely different in *many ways*

Why do we study *average* treatment effects?

Fundamental Problem of Causal Inference

- Never observe both Y_0 and Y_1 at the same time for the same person.
- This means we *cannot* learn the joint distribution of the potential outcomes.[2]
- Treatment effect depends on *both* potential outcomes: $(Y_1 - Y_0)$. What to do?

Linearity of Expectation

- $\mathbb{E}[X - Z] = \mathbb{E}[X] - \mathbb{E}[Z]$ *regardless* of the joint distribution of (X, Z).
- **Very special** property. It doesn't hold, e.g., for variance, quantiles, etc.
- Replace infeasible within-person comparison with between-person comparison:

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$$

[2] The joint distribution is not point identified, but it can be *bounded*. See chapter 3 of the notes.

Selection Bias

Naïve Comparison of Means

$$\mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0) = \mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=0)$$

$$= \mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=0) + \mathbb{E}(Y_0|D=1) - \mathbb{E}(Y_0|D=1)$$

$$= \underbrace{\mathbb{E}(Y_1 - Y_0|D=1)}_{\text{TOT}} + \underbrace{[\mathbb{E}(Y_0|D=1) - \mathbb{E}(Y_0|D=0)]}_{\text{Selection Bias}}$$

How does selection matter?

1. TOT is probably different from ATE: selection on gains.
2. Average value of Y_0 ("outside option") probably varies with D.

Randomization eliminates selection bias.

Independence[3]

- $X \perp\!\!\!\perp Z$ is shorthand for "X is **statistically independent** of Z."
- $X \perp\!\!\!\perp Z \iff f(x,z) = f(x)f(z)$ for all x and z.
- Statistical independence implies **conditional mean independence**

$$\mathbb{E}[X|Z=z] \equiv \int_{-\infty}^{\infty} x \cdot f(x|z)\,\mathrm{d}x = \int_{-\infty}^{\infty} x \cdot \frac{f(x)f(z)}{f(z)}\,\mathrm{d}x = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x \equiv \mathbb{E}[X]$$

Random Assignment: $D \perp\!\!\!\perp (Y_0, Y_1)$

$$\begin{aligned}\text{TOT} &= \mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=1) = \mathbb{E}(Y_1) - \mathbb{E}(Y_0) \equiv \text{ATE} \\ \text{Selection Bias} &\equiv \mathbb{E}[Y_0|D=1] - \mathbb{E}[Y_0|D=0] = \mathbb{E}[Y_0] - \mathbb{E}[Y_0] = 0\end{aligned}$$

[3] See chapter 2 of the notes, https://expl.ai/LXPVDDN and my blog post for more on independence.

But randomization may be impossible, impractical, or unethical.

Returns to Education

Tempting though it may be during admissions season, I would face some serious consequences if I randomly admitted students to Oxford!

Women's Labor Supply

We wouldn't randomly assign different numbers of children to different women to test the causal effect on their labor supply.

Fox News and Voting Behavior

We can't force some people to watch Fox news and others to watch CNN and then keep track of who they voted for.

Causal inference from observational data is challenging, but it's often the best we can do.

Disease Example: $Y = 1$ means survive, $Y = 0$ means perish

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Exercise – Calculate the Following

1. ATE
2. $\mathbb{E}(Y|D = 1) - \mathbb{E}(Y|D = 0)$
3. TOT
4. Selection Bias

```r
library(tidyverse)

people <- c("Aiden", "Bella", "Carter", "Dakota", "Ethel", "Floyd",
            "Gladys", "Herbert", "Irma", "Julius")

x <- c("young", "young", "young", "young", "old", "old",
       "old", "old", "old", "old")

y0 <- c(1, 1, 1, 1, 0, 0, 0, 0, 0, 0)
y1 <- c(1, 1, 1, 1, 1, 0, 0, 1, 0, 0)
d <- c(0, 0, 0, 1, 0, 0, 0, 1, 1, 1)
y <- (1 - d) * y0 + d * y1

tbl <- tibble(name = people, d, y, y0, y1, x)
rm(y0, y1, d, y, x, people)
```

```r
ATE <- tbl |>
  summarize(mean(y1 - y0)) |>
  pull()

TOT <- tbl |>
  filter(d == 1) |>
  summarize(mean(y1 - y0)) |>
  pull()

SB <- tbl |>
  group_by(d) |>
  summarize(y0_mean = mean(y0)) |>
  pull(y0_mean) |>
  diff()
```

```r
# E(Y|D=1) and E(Y|D=0)
means <- tbl |>
  group_by(d) |>
  summarize(y_mean = mean(y))

means
```

```
## # A tibble: 2 x 2
##       d y_mean
##   <dbl>  <dbl>
## 1     0    0.5
## 2     1    0.5
```

```r
# Naive difference of means
naive <- means |>
  pull(y_mean) |>
  diff()
```

Solution

```
# Everything we've calculated
c(ATE = ATE, naive = naive, TOT = TOT, SB = SB)
```

```
##   ATE naive   TOT    SB
##  0.20  0.00  0.25 -0.25
```

- Sanity Check: results satisfy the "Fundamental Decomposition"

$$\underbrace{\mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0)}_{\text{Observed Difference of Means}} = \underbrace{\mathbb{E}(Y_1 - Y_0|D=1)}_{\text{TOT}} + \underbrace{[\mathbb{E}(Y_0|D=1) - \mathbb{E}(Y_0|D=0)]}_{\text{Selection Bias}}$$

- **Selection into treatment:** the treated are **older** on average.

Conditional Average Treatment Effects (CATEs)

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Intuition

How do treatment effects vary with observed characteristics X?

Definition

$\text{CATE}(x) \equiv \mathbb{E}(Y_1 - Y_0 | X = x)$

Exercise

1. Compute CATE(Young)
2. Compute CATE(Old)
3. Relate these to the *overall* ATE.

Solution: No treatment effect for Young; positive effect for Old.

```
# Conditional ATEs
tbl |>
  group_by(x) |>
  summarize(CATE = mean(y1 - y0))
```

```
## # A tibble: 2 x 2
##   x      CATE
##   <chr> <dbl>
## 1 old   0.333
## 2 young 0
```

Solution: No treatment effect for Young; positive effect for Old.

```
group_stats <- tbl |>
  group_by(x) |>
  summarize(CATE_x = mean(y1 - y0), count = n()) |>
  mutate(p_x = count / sum(count))

group_stats
```

```
## # A tibble: 2 x 4
##   x     CATE_x count   p_x
##   <chr>  <dbl> <int> <dbl>
## 1 old    0.333     6   0.6
## 2 young  0         4   0.4
```

Solution: ATE $= \mathbb{E}[\text{CATE}(\boldsymbol{X})]$ by the Law of Iterated Expectations

```
# E[E(Y1 - Y0 | X)]
group_stats |>
  summarize(sum(CATE_x * p_x)) |>
  pull()
```

```
## [1] 0.2
```

```
# E(Y1 - Y0)
tbl |>
  summarize(mean(y1 - y0)) |>
  pull()
```

```
## [1] 0.2
```

Wait, what is this lecture supposed to be about again?

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Disease Example

Selection into treatment: naive comparison of means doesn't give ATE.

Iterated Expectations

If we learn the CATEs, we can average them to get the ATE.

Idea

Maybe if we **adjust for age**, we can address the selection problem.

Selection-on-observables

A pair of assumptions that shows us when this idea will work out.

Propensity Score: Who is more likely to be treated?

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Propensity Score $p(x)$

- $p(x) \equiv \mathbb{P}(D = 1|X = x)$
- Share treated by age group.

Exercise

Calculate $p(\text{Young})$ and $p(\text{Old})$

Propensity Score: Who is more likely to be treated?

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Propensity Score $p(x)$

- $p(x) \equiv \mathbb{P}(D = 1|X = x)$
- Share treated by age group.

Exercise

Calculate $p(\text{Young})$ and $p(\text{Old})$

Solution

$p(\text{Young}) = 1/4, \quad p(\text{Old}) = 1/2$

Old people are more likely to take treatment and more likely to die with or without it!
Age *confounds* the relationship between D and Y.

Wishful Thinking

Wouldn't it be great if $\text{CATE}(x) = \mathbb{E}(Y|D=1, X=x) - \mathbb{E}(Y|D=0, X=x)$?

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Stratify by Age

- Perhaps *within* age groups there is no selection problem.
- If so, learn the CATE for each group.

Exercise

Check if this claim holds in our example.

Stratifying by age works in this example

$$\text{CATE}(x) = \mathbb{E}(Y|D=1, X=x) - \mathbb{E}(Y|D=0, X=x)$$

```
tbl |>
  group_by(x) |>
  summarize(CATE = mean(y1-y0)) |>
  knitr::kable(digits = 2)
```

x	CATE
old	0.33
young	0.00

```
tbl |>
  group_by(x, d) |>
  summarize(y_mean = mean(y)) |>
  knitr::kable(digits = 2)
```

x	d	y_mean
old	0	0.00
old	1	0.33
young	0	1.00
young	1	1.00

Final Step: Iterated Expectations to Get ATE

$$\text{ATE} = \text{CATE}(\text{Young})\mathbb{P}(\text{Young}) + \text{CATE}(\text{Old})\mathbb{P}(\text{Old}) = 0 \times 2/5 + 1/3 \times 3/5 = 0.2$$

This worked because our example satisfies two key assumptions.

Definition: Conditional Independence

- $W \perp\!\!\!\perp Z|R \iff \mathbb{P}(W, Z|R) = \mathbb{P}(W|R) \cdot \mathbb{P}(Z|R)$.
- See chapter 2 of the lecture notes and this video for more details.

Assumption 1 – Selection on Observables:[4] $D \perp\!\!\!\perp (Y_0, Y_1)|\boldsymbol{X}$

- Implies that people with the same observed characteristics have the same potential outcomes, on average, regardless of whether they were *actually* treated or not.
- See my blog post for more discussion of this assumption.

Assumption 2 – Overlap: $0 < p(\boldsymbol{x}) < 1$ for all values of $\boldsymbol{x}$.

- Recall that $p(\boldsymbol{x}) \equiv \mathbb{P}(D = 1|\boldsymbol{X} = \boldsymbol{x})$.
- Among people with given characteristics $\boldsymbol{x}$, some but not all are treated.

[4] This can be weakened to $\mathbb{E}(Y_d|D, \boldsymbol{X}) = \mathbb{E}(Y_d|\boldsymbol{X})$ for $d = 0, 1$, i.e. *mean* independence.

The approach we used above is called "Regression Adjustment"

Intuition

- Form **strata** based on common value $\boldsymbol{x}$ of covariates.
- Within each stratum, compute the average outcome among treated and untreated.
- Subtract these to estimate CATE($\boldsymbol{x}$), the stratum-specific ATE.
- Average the stratum-specific ATEs, weighting by the fraction of people in each.

Main Result[5]

Under the selection on observables and overlap assumptions:

$$\text{CATE}(\boldsymbol{x}) \equiv \mathbb{E}(Y_1 - Y_0 | \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(Y | D = 1, \boldsymbol{X} = \boldsymbol{x}) - \mathbb{E}(Y | D = 0, \boldsymbol{X} = \boldsymbol{x}).$$

By iterated expectations, $\text{ATE} = \mathbb{E}[\text{CATE}(\boldsymbol{X})]$ so we can learn the ATE.

[5] See my video for the proof: https://expl.ai/BJWTFKG

How to implement with a regression? (Assumes all covariates binary)

1. Center all covariates $\boldsymbol{X}$ around their means: $\widetilde{\boldsymbol{X}} \equiv \boldsymbol{X} - \bar{\boldsymbol{X}}$
2. Regress Y on D, $\widetilde{\boldsymbol{X}}$ and all interactions.
3. The coefficient on D is the ATE and its standard error is correct.[6]

```
library(broom)
tbl |>
  mutate(old = (x == 'old'), xtilde = old - mean(old)) |>
  lm(y ~ d * xtilde, data = _) |>
  tidy() |> filter(term == 'd')
```

```
## # A tibble: 1 x 5
##   term  estimate std.error statistic p.value
##   <chr>    <dbl>     <dbl>     <dbl>   <dbl>
## 1 d          0.2     0.224     0.891   0.407
```

[6] Technically this ignores estimation error in the mean of $\boldsymbol{X}$.

Alternative Approach: Propensity Score Weighting

Intuition

- Disease example: older people are more likely to be treated and more likely die regardless of whether they are treated.
- *Too few* young people among the treated and *too few* old people among the untreated relative to what we'd have in a randomized experiment.
- To compensate: **upweight** treated young people untreated old people when computing average outcomes for the treated and untreated groups.

Main Result[7]

Under the selection on observables and overlap assumptions:

$$\text{ATE} = \mathbb{E}\left[w_1(\boldsymbol{X}) \cdot Y\right] - \mathbb{E}\left[w_0(\boldsymbol{X}) \cdot Y\right], \quad w_1(\boldsymbol{X}) = \frac{D}{p(\boldsymbol{X})}, \quad w_0(\boldsymbol{X}) = \frac{1-D}{1-p(\boldsymbol{X})}$$

[7] See my video for the proof: https://expl.ai/BASRRGX.

Propensity Score Weighting in Our Example

```
psw <- tbl |>
  group_by(x) |>
  mutate(pscore = mean(d)) |>
  ungroup() |>
  mutate(weight1 = d / pscore,
         weight0 = (1 - d) / (1 - pscore))
```

Propensity Score Weighting in Our Example

```
psw |> select(-y0, -y1)
```

```
## # A tibble: 10 x 7
##    name         d     y x     pscore weight1 weight0
##    <chr>    <dbl> <dbl> <chr>  <dbl>   <dbl>   <dbl>
##  1 Aiden        0     1 young   0.25       0    1.33
##  2 Bella        0     1 young   0.25       0    1.33
##  3 Carter       0     1 young   0.25       0    1.33
##  4 Dakota       1     1 young   0.25       4    0
##  5 Ethel        0     0 old     0.5        0    2
##  6 Floyd        0     0 old     0.5        0    2
##  7 Gladys       0     0 old     0.5        0    2
##  8 Herbert      1     1 old     0.5        2    0
##  9 Irma         1     0 old     0.5        2    0
## 10 Julius       1     0 old     0.5        2    0
```

Propensity Score Weighting in Our Example

```
psw |> summarize(sum(weight1), sum(weight0))
```

```
## # A tibble: 1 x 2
##   `sum(weight1)` `sum(weight0)`
##            <dbl>          <dbl>
## 1             10             10
```

```
psw |>
  summarize(mean(weight1 * y) - mean(weight0 * y)) |>
  pull()
```

```
## [1] 0.2
```

```
ATE
```

```
## [1] 0.2
```

How can we evaluate the assumptions?

Overlap

- Since D and $\boldsymbol{X}$ are observed, we can check this directly.
- The more characteristics we put into $\boldsymbol{X}$, the harder it becomes to satisfy overlap.

Selection on Observables

- Without outside data or extra assumptions, there's no way to check this.
- Else equal, the more characteristics we put into $\boldsymbol{X}$, the more plausible this becomes.

Bad Controls

- More is **not always better**. Some characteristics definitely **shouldn't** go into $\boldsymbol{X}$.
- This is the topic of our next lecture!

ATE or TOT?

Maybe we don't *want* the ATE

- ATE is the average effect of *forcing* a randomly chosen person to be treated.
- But in real life we can't usually force anyone to be treated; only offer treatment.
- TOT is the average benefit of treatment for people who will *voluntarily take it*.[8]

Maybe we can't *get* the ATE

- Models of rational choice assume that agents compare costs and benefits of choices.
- Benefit of treatment is equal (or at least related to) to $Y_1 - Y_0$.
- Selection on observables implies $\mathbb{E}(Y_1 - Y_0|D, \boldsymbol{X}) = \mathbb{E}(Y_1 - Y_0|\boldsymbol{X})$.
- I.e. agents lack (or don't act on) private information about gains from treatment.

[8] Another angle: TOT is the forgone benefit per person of *discontinuing* a program.

Identifying the TOT with Weaker Assumptions

Assumptions

1. $\mathbb{E}(Y_0|D, \boldsymbol{X}) = \mathbb{E}(Y_0|\boldsymbol{X})$
2. $p(\boldsymbol{x}) < 1$ for all $\boldsymbol{x}$ in the support of $\boldsymbol{X}$.

Why are these assumptions weaker?

- Places *no restrictions* on relationship between $(Y_1 - Y_0)$ and D.
- It's fine if people select into treatment based on private info about $(Y_1 - Y_0)$.
- Overlap condition is also weaker: it's fine if there are no treated people for some $\boldsymbol{x}$

Theorem

$\text{TOT} = \mathbb{E}[Y|D = 1] - \mathbb{E}_{\boldsymbol{X}|D=1}\left[\mathbb{E}(Y|D = 0, \boldsymbol{X})\right]$.

There's also a version for propensity score weighting...

Appendix

Regression Adjustment Derivation[9]

Since $Y = (1 - D)Y_0 + DY_1 = Y_0 + D(Y_1 - Y_0)$, taking expectations of both sides:

$$\begin{aligned}\mathbb{E}(Y|\boldsymbol{X}, D) &= \mathbb{E}(Y_0|\boldsymbol{X}, D) + D\left[\mathbb{E}(Y_1|\boldsymbol{X}, D) - \mathbb{E}(Y_0|\boldsymbol{X}, D)\right] \\ &= \mathbb{E}(Y_0|\boldsymbol{X}) + D\left[\mathbb{E}(Y_1|\boldsymbol{X}) - \mathbb{E}(Y_0|\boldsymbol{X})\right]\end{aligned}$$

by the selection on observables assumption. Substituting $D = 0$ and $D = 1$ in turn,

$$\mathbb{E}(Y|\boldsymbol{X}, D = 0) = \mathbb{E}(Y_0|\boldsymbol{X}), \quad \mathbb{E}(Y|\boldsymbol{X}, D = 1) = \mathbb{E}(Y_1|\boldsymbol{X}).$$

Therefore,

$$\text{ATE}(\boldsymbol{X}) = \mathbb{E}(Y_1|\boldsymbol{X}) - \mathbb{E}(Y_0|\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X}, D = 1) - \mathbb{E}(Y|\boldsymbol{X}, D = 0).$$

The overlap assumption ensures that $\text{ATE}(\boldsymbol{X})$ is well-defined for all $\boldsymbol{X}$.

[9] Video: https://expl.ai/BJWTFKG

Propensity Score Weighting Derivation[10]

Since D is binary, $D^2 = D$, $(1-D)^2 = (1-D)$, and $D(1-D) = 0$. Hence,

$$\begin{aligned} DY &= D\left[(1-D)Y_0 + DY_1\right] \\ &= D^2 Y_1 + D(1-D)Y_0 \\ &= DY_1 \end{aligned}$$

$$\begin{aligned} (1-D)Y &= (1-D)\left[(1-D)Y_0 + DY_1\right] \\ &= (1-D)DY_1 + (1-D)^2 Y_0 \\ &= (1-D)Y_0. \end{aligned}$$

[10] Video: https://expl.ai/BASRRGX

Propensity Score Weighting Derivation Continued

Since $DY = DY_1$,

$$\begin{aligned}
\mathbb{E}[DY|\boldsymbol{X}] &= \mathbb{E}[DY_1|\boldsymbol{X}] = \mathbb{E}_{D|\boldsymbol{X}}\left[D\,\mathbb{E}(Y_1|D,\boldsymbol{X})\right] && \text{(Iterated Expectations)}\\
&= \mathbb{E}_{D|\boldsymbol{X}}\left[D\,\mathbb{E}(Y_1|\boldsymbol{X})\right] && \text{(Selection on Observables)}\\
&= \mathbb{E}(D|\boldsymbol{X})\mathbb{E}(Y_1|\boldsymbol{X}) && \text{(Take out what is known)}\\
&= p(\boldsymbol{X})\mathbb{E}(Y_1|\boldsymbol{X}). && \text{(Defn. of Propensity Score)}
\end{aligned}$$

Since $(1-D)Y = (1-D)Y_0$, an effectively identical argument gives:

$$\mathbb{E}[(1-D)Y|\boldsymbol{X}] = \mathbb{E}[(1-D)Y_0|\boldsymbol{X}] = [1-p(\boldsymbol{X})]\mathbb{E}(Y_0|\boldsymbol{X}).$$

Propensity Score Weighting Derivation Continued Again

Previous slide:

$$\mathbb{E}[DY|\boldsymbol{X}] = p(\boldsymbol{X})\mathbb{E}(Y_1|\boldsymbol{X}), \quad \mathbb{E}[(1-D)Y|\boldsymbol{X}] = [1-p(\boldsymbol{X})]\mathbb{E}(Y_0|\boldsymbol{X})$$

Dividing through by $p(\boldsymbol{X})$ and $[1-p(\boldsymbol{X})]$, respectively, gives

$$\mathbb{E}\left[\frac{DY}{p(\boldsymbol{X})}\middle|\boldsymbol{X}\right] = \mathbb{E}(Y_1|\boldsymbol{X}), \quad \mathbb{E}\left[\frac{(1-D)Y}{1-p(\boldsymbol{X})}\middle|\boldsymbol{X}\right] = \mathbb{E}(Y_0|\boldsymbol{X})$$

since we can bring any function of $\boldsymbol{X}$ inside the conditional expectations.[11] Finally, by iterated expectations:

$$\mathbb{E}\left[\frac{DY}{p(\boldsymbol{X})}\right] = \mathbb{E}(Y_1), \quad \mathbb{E}\left[\frac{(1-D)Y}{1-p(\boldsymbol{X})}\right] = \mathbb{E}(Y_0)$$

and the difference of these is the ATE.

[11] This is simply "taking out what is known" in reverse.

TOT via Regression Adjustment Derivation

Since $Y = Y_1$ when $D = 1$,

$$\text{TOT} \equiv \mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=1) = \mathbb{E}(Y|D=1) - \mathbb{E}(Y_0|D=1).$$

Now, by iterated expectations

$$\mathbb{E}(Y_0|D=1) = \mathbb{E}_{\boldsymbol{X}|D=1}\left[\mathbb{E}(Y_0|D=1,X)\right] = \mathbb{E}_{\boldsymbol{X}|D=1}\left[\mathbb{E}(Y_0|D=0,\mathbf{X})\right]$$

since $\mathbb{E}(Y_0|D,\boldsymbol{X}) = \mathbb{E}(Y_0|\boldsymbol{X})$ by assumption. But since $Y = Y_0$ given $D = 0$,

$$\mathbb{E}(Y_0|D=0,\boldsymbol{X}) = \mathbb{E}(Y|D=0,\boldsymbol{X}).$$

Therefore,

$$\mathbb{E}(Y_0|D=1) = \mathbb{E}_{\boldsymbol{X}|D=1}\left[\mathbb{E}(Y|D=0,\boldsymbol{X})\right] = \int_{\mathcal{X}} \mathbb{E}(Y|D=0,\boldsymbol{X}=\boldsymbol{x}) f(\boldsymbol{x}|D=1)\,\mathrm{d}\boldsymbol{x}.$$