Selection on Observables

Francis J. DiTraglia

University of Oxford

Core Empirical Research Methods

Remainder of the Course: Causal Inference

- So far: causality in **linear models** with **homogeneous effects**.
  1. Linear Regression
  2. Instrumental variables
  3. Fixed effects
- Now: **heterogenous effects** and weaker modeling assumptions.
  1. Selection on Observables (Today, chapter 4)
  2. Directed Acyclic Graphs and Bad Controls (Tomorrow)
  3. Regression Discontinuity (chapter 7)
  4. Local Average Treatment Effects (chapter 5)
  5. Difference-in-differences (chapter 8)

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

- ATE $= \mathbb{E}(Y_1 - Y_0)$ is the average effect of *forcing* a randomly-chosen person to attend university.
- 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

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

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

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

[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.

Does education cause political participation?[4]

- College graduates are more likely to vote, volunteer for campaigns, contact elected representatives, participate in demonstrations.
- Y = index of political participation: $\uparrow Y$ means $\uparrow$ participation.
- $D = 0$ is **no college**; $D = 1$ is **college**
- It seems implausible that $D \perp\!\!\!\perp (Y_0, Y_1)$ in this example.
- E.g. family background may cause both education and political participation.

An Idea

If we *condition* on family background, income, sex, race, and other observed variables, perhaps we can break the dependence between D and (Y_0, Y_1).

[4] Kam and Palmer (2008)

Assumptions

Propensity Score $p(\boldsymbol{X})$

Treatment probability given observed covariates: $p(\boldsymbol{X}) \equiv \mathbb{P}(D = 1|\boldsymbol{X}) = \mathbb{E}(D|\boldsymbol{X})$

Selection on Observables Assumption[5]

$$\mathbb{E}(Y_0|\boldsymbol{X}, D) = \mathbb{E}(Y_0|\boldsymbol{X}), \quad \text{and} \quad \mathbb{E}(Y_1|\boldsymbol{X}, D) = \mathbb{E}(Y_1|\boldsymbol{X}).$$

- Conditional on $\boldsymbol{X}$, Y_0 and Y_1 are **mean independent** of D.
- People with the same observed characteristics have the same potential outcomes, on average, regardless of whether they were *actually* treated or not.

Overlap Assumption

- $0 < p(\boldsymbol{x}) < 1$ for all $\boldsymbol{x}$ in the support of $\boldsymbol{X}$.
- Among people with given characteristics, some but not all are treated.

[5]See my blog post for a discussion of what this assumption does *not mean*.

How can we evaluate these 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 auxiliary 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 deserves a lecture of its own. We'll discuss in more detail next time.

Simulation Example

```
set.seed(5672349)

n <- 5000
x1 <- rbinom(n, 1, 0.25)
x2 <- rnorm(n, 4)

p <- plogis(-3 + 0.4 * x1 + 0.5 * x2 + 0.3 * x1 * x2)
d <- rbinom(n, 1, p)

y0 <- 0.05 * x1 + 0.15 * x2 + 0.25 * x1 * x2 + rnorm(n, 0.1)
y1 <- 0.1 * x1 + 0.1 * x2 + 0.35 * x1 * x2 + rnorm(n, 0.1)
y <- (1 - d) * y0 + d * y1
```

ATE, TOT, and Selection Bias in Simulation Example

```
c(ATE = mean(y1 - y0),
  TOT = mean(y1[d == 1]) - mean(y0[d == 1]),
  selection_bias = mean(y0[d == 1]) - mean(y0[d == 0]),
  naive = mean(y[d == 1]) - mean(y[d == 0])) |>
  round(2)
```

```
##            ATE            TOT selection_bias          naive
##          -0.11           0.01           0.38           0.39
```

First Approach: 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 ATE($\boldsymbol{x}$), the stratum-specific ATE.
- Average the stratum-specific ATEs, weighting by the number of people in each.

Theorem

Under the selection on observables and overlap assumptions:

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

By iterated expectations, $\text{ATE} = \mathbb{E}[\text{ATE}(\boldsymbol{X})]$ so the ATE is identified.

Regression Adjustment Derivation[6]

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 $\boldsymbol{X}$.

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

Regression Adjustment: Simulation Example

```r
reg0 <- lm(y ~ x1 * x2, subset = (d == 0))
reg1 <- lm(y ~ x1 * x2, subset = (d == 1))

y0_pred <- predict(reg0, data.frame(x1 = x1, x2 = x2))
y1_pred <- predict(reg1, data.frame(x1 = x1, x2 = x2))

c(ATE = mean(y1 - y0),
  reg_adj = mean(y1_pred - y0_pred),
  naive = mean(y[d == 1]) - mean(y[d == 0])) |>
  round(2)
```

```
##      ATE reg_adj   naive
##    -0.11   -0.08    0.39
```

Another way to carry out regression adjustment...

Unlike the previous approach, this one provides a standard error automatically.[7]

```
library(tidyverse); library(broom)
x1_tilde <- x1 - mean(x1)
x2_tilde <- x2 - mean(x2)
reg_combined <- lm(y ~ d + (x1 * x2) + d:(x1_tilde * x2_tilde))

reg_combined |> tidy() |> filter(term == 'd') |>
  select(estimate, std.error) |> round(2)
```

```
## # A tibble: 1 x 2
##   estimate std.error
##      <dbl>     <dbl>
## 1    -0.08      0.03
```

[7]Technically we should account for estimation uncertainty in $\bar{\boldsymbol{X}}$.

Second Approach: Propensity Score Weighting

Intuition

- Suppose that biological sex causes D and that potential outcomes vary with sex.
- Women more likely to be treated than men $\Rightarrow$ too few men among the treated and too few women among the untreated.
- To compensate: **upweight** treated men and untreated women when computing the average outcomes for treated and untreated groups.

Theorem

Under the selection on observables and overlap assumptions:

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

Propensity Score Weighting Derivation[8]

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

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

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

[8]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

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

since we can bring any function of $\boldsymbol{X}$ inside the conditional expectations.[9] 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.

[9] This is simply “taking out what is known” in reverse.

Propensity Score Weighting: Simulation Example

```r
lreg <- glm(d ~ x1 * x2, family = binomial())

p_scores <- predict(lreg, data.frame(x1 = x1, x2 = x2),
                    type = 'response') # CRUCIAL!

psw <- mean(d * y / p_scores) - mean((1 - d) * y / (1 - p_scores))

c(psw = psw,
  reg_adj = mean(y1_pred - y0_pred),
  ATE = mean(y1 - y0),
  naive = mean(y[d == 1]) - mean(y[d == 0])) |>
  round(2)
```

```
##      psw reg_adj     ATE   naive
##    -0.09   -0.08   -0.11    0.39
```

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*.[10]

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.

[10] 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...

TOT 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,

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

Regression Adjustment for the TOT

```r
x_treated <- tibble(x1, x2, d) |>
  filter(d == 1)

y0_pred_treated <- predict(reg0, x_treated)

c(TOT = mean(y1[d == 1]) - mean(y0[d == 1]),
  reg_adj = mean(y[d == 1]) - mean(y0_pred_treated)) |>
  round(2)
```

```
##      TOT reg_adj
##     0.01   -0.03
```

Next year I'll add some slides about matching!

- An alternative to propensity score weighting and regression adjustment.
- Relies on effectively identical assumptions, but computed differently.
- Simplest version: use $\boldsymbol{X}$ to find "most similar" control for each treated unit, then subtract outcomes for the resulting matched pairs and average.
- For more, see Stuart (2010) and Dehejia & Wahba (2002).
- The Kam and Palmer (2008) paper mentioned above also uses matching.