Today

Nicholas Ray

2024-11-18

Potential Outcomes

The potential outcomes model is:

$$\begin{aligned} Y_i &= \begin{cases} Y_{0i} \text{ if } D_i = 0 \\ Y_{1i} \text{ if } D_i = 1 \end{cases} \\ &= Y_{0i} + (Y_{1i} - Y_{0i})D_i. \end{aligned}$$

The fundamental problem of causal inference is that we only observe one of the potential outcomes $\{Y_0, Y_1\}$ for unit i. Unit i either gets the treatment or they don't. If they got the treatment, then we are observing what would have happened to them if they got the treatment (potential outcome Y_{1i}). If they didn't get the treatment, then we are observing what would have happened to them if they didn't get the treatment (potential outcome Y_{0i}).

We can never go back in time and give the exact same unit i a different treatment assignment. If we could, we'd be able to compare what would have happened to them if they got treatment and what would have happened to them if they didn't: $Y_{1i} - Y_{0i}$. This is the treatment effect of D for unit i.

However, the point is that under some basic assumptions (i.i.d. (Y_i, D_i), SUTVA [stable or fixed potential outcomes and you're only affected by your own treatment]) we can estimate the average treatment effect (ATE: $\mathbb{E}[Y_{1i} - Y_{0i}]$) if we have 1) some units that got treatment and some that didn't and 2) a way to eliminate selection bias (loosely meaning an identification strategy). Essentially, selection bias is the bias in the average treatment effect from the treatment group being "special" and receiving treatment for a reason (although there are different kinds of selection bias). With selection bias it becomes impossible to say how much of the observed average effect is from the treatment and how much of it is from the characteristic(s) that made units get treatment.

On a fundamental level, unless you are randomly forcing units to receive treatment there is always the possibility that there is something "special" about the units that willingly selected into receiving treatment. As just implied, one way to eliminate selection bias is to randomly assign treatment. Absent true randomization, we will rely on clever identification strategies that sometimes leverage the unexpectedness of treatment to claim it was assigned "as if" random. You will learn these identification strategies later, but it is perhaps useful to remember that many of these strategies fail if units did indeed expect treatment and changed their behavior accordingly.

Why the observed difference in means (ODIM) fails to be a "good" (i.e., unbiased) estimator for the ATE:

- decomposition of ODIM into ATT
- decomposition of ODIM into ATC

Regression

Consider our expression for the potential outcomes model: $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$. We can rewrite this as a linear projection:

$$
\begin{aligned}
Y_i &= Y_{0i} + (Y_{1i} - Y_{0i})D_i \\
&= Y_{0i} + (Y_{1i} - Y_{0i})D_i + \{\text{prediction error of outcome to be minimized}\} \\
&= \mathbb{E}[Y_{0i}] + \mathbb{E}\big[(Y_{1i} - Y_{0i})D_i\big] + \mathbb{E}\big[\{\text{prediction error of outcome to be minimized}\}\big] \\
&= \underbrace{\mathbb{E}[Y_{0i}]}_{\beta_0} \underbrace{+\mathbb{E}\big[(Y_{1i} - Y_{0i})\big]D_i}_{+ D_i\beta_1} \underbrace{+\{D_i(Y_{1i} - \mathbb{E}[Y_{1i}]) + (1 - D_i)(Y_{0i} - \mathbb{E}[Y_{0i}])\}}_{+ e_i}
\end{aligned}
$$

Random assignment ensures that $\mathbb{E}[e_i]$ and $\mathbb{E}[D_i e_i]$.

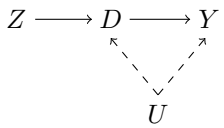
Note that this is just the way I make sense of this. We already have a model, which says that an outcome for an individual is a function of their potential outcomes and treatment. To apply the machinery of OLS, we just "create" an error term by explicitly recognizing that an individual's outcome may deviate from an average (like prediction error earlier).

Matt would say the following, which you should always default to:

> The point here is not that we're aiming to minimize the prediction error. The point of this decomposition is just to show: if we take the potential outcomes, and rearrange (adding and substracting the same terms) we end up with a linear function of D_i, with an error term e_i and we can show that $\mathbb{E}[e_i] = \mathbb{E}[D_i e_i] = 0$. The fact that we have a linear function with $\mathbb{E}[e_i] = \mathbb{E}[D_i e_i] = 0$ means that the linear function is a linear projection and thus τ can be obtained by linear regression.

Instrumental Variables (IV)

Consider the following directed acyclic graph (DAG) (Cunningham 2021, 319):



where Z is a binary treatment assignment indicator, D is a binary treatment uptake indicator, Y is the observed outcome, and U is a set of unobservables that confounds the relationship between D and Y. Because of U (and our inability to "control" for it), we cannot say exactly what the effect of the treatment is on Y.

However, we can use these observed relationships and some assumptions to learn about a "local" or more specific (less general) treatment effect.

- A1) **independence**: Z (treatment assignment) is randomly assigned and so independent of potential treatment uptake (D_{0i} or D_{1i}) and potential outcomes (Y_{0i} or Y_{1i})
- A2) **exclusion**: Z only affects Y through D (D fully mediates Z)

- A3) **relevance** or **"First stage"**: Although Z is independent of potential treatment uptake (D_{0i} or D_{1i}), once treatment is assigned it has an effect on uptaking treatment (D_i)
- A4) **monotonicity**: Z's effect on treatment uptake is weakly positive (getting assigned to treatment can only make treatment uptake more likely, not less). In other words, there are no "defiers"

Under these assumptions, we can back out the effect of Z on Y. However, this effect is contextualized by the mediator D. It is a "local" average treatment effect ($LATE$) in the sense that it is an average treatment effect specific to the subpopulation of compliers. We only know the effect of Z on Y through those who complied with uptaking the treatment D, and this can be a way to model heterogeneous treatment effects.

$$\mathbb{E}[Y_{1i} - Y_{0i} | D_{1i} > D_{0i}] = \frac{\mathbb{E}[Y_i | Z_i = 1] - \mathbb{E}[Y_i | Z_i = 0]}{\mathbb{E}[D_i | Z_i = 1] - \mathbb{E}[D_i | Z_i = 0]}$$

- Walking through derivation on the slides

Basics of Using Logs

If we have $\ln(Y_i) = \beta_0 + \beta_1 X_i + e_i$ (and $\mathbb{E}[Xe] = 0$), then a δ change in X (e.g., $X_2 - X_1$) corresponds to a $100(\delta \times \beta_1)$ percent change in the conditional geometric mean of Y. This follows from the "natural log approximation property," with the approximation getting worse as the purported percent change gets larger.

- really should not log variables that are not strictly greater than 0

Bibliography

Cunningham, Scott. 2021. *Causal Inference: The Mixtape.* Yale University Press.