OIT 661: Lecture 1
Treatment effect estimation in randomized experiments
(for class use only, please do not distribute)

The average treatment effect. We define the causal effect of a treatment via potential outcomes. For a binary treatment $w \in \{0, 1\}$, we define potential outcomes $Y_i(1)$ and $Y_i(0)$ corresponding to the outcome the i-th subject would have experienced had they respectively received the treatment or not. The causal effect of the treatment is then

$$\Delta_i = Y_i(1) - Y_i(0).$$

The fundamental problem in causal inference is that only one treatment can be assigned to a given individual, and so only one of $Y_i(0)$ and $Y_i(1)$ can ever be observed. Thus, Δ_i can never be observed.

Now, although Δ_i itself is fundamentally unknowable, we can (perhaps remarkably) used randomized experiments to learn certain properties of the Δ_i. In particular, large randomized experiments let us recover the average treatment effect (ATE)

$$\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right].$$

To do so, assume that we observe n independent and identically distributed samples (Y_i, W_i) satisfying the following two properties:

$$\begin{aligned} &Y_i = Y_i(W_i) && \text{(SUTVA)} \\ &W_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} && \text{(random treatment assignment)} \end{aligned}$$

Then, the difference-in-means estimator

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{W_i=1} Y_i - \frac{1}{n_0} \sum_{W_i=1} Y_i, \qquad n_w = |\{i : W_i = w\}|$$

is unbiased and consistent for the average treatment effect

Difference-in-means estimation. The statistical properties of $\hat{\tau}_{DM}$ can readily be established. Noting that, for $w \in \{0, 1\}$

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n_w} \sum_{W_i=w} Y_i\right] &= \mathbb{E}\left[Y_i \,\middle|\, W_i = w\right] && \text{(IID)} \\ &= \mathbb{E}\left[Y_i(w) \,\middle|\, W_i = w\right] && \text{(SUTVA)} \\ &= \mathbb{E}\left[Y_i(w)\right] && \text{(random assignment)}, \end{aligned}$$

we find that the difference-in-means estimator is unbiased[1]

$$\mathbb{E}\left[\hat{\tau}_{DM}\right] = \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right] = \tau.$$

Moreover, we can write the variance as

$$\text{Var}\left[\hat{\tau}_{DM}\right] = \frac{1}{n_0}\text{Var}\left[Y_i(0)\right] + \frac{1}{n_1}\text{Var}\left[Y_i(1)\right].$$

A standard central limit theorem can be used to verify that

$$\begin{aligned}&\sqrt{n}\left(\hat{\tau}_{DM} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V_{DM}\right),\\ &V_{DM} = \text{Var}\left[Y_i(0)\right] \,/\, \mathbb{P}\left[W_i = 0\right] + \text{Var}\left[Y_i(1)\right] \,/\, \mathbb{P}\left[W_i = 1\right].\end{aligned}$$

Finally, note that we can estimate V_{DM} via routine plug-in estimators to build valid Gaussian confidence intervals for τ:

$$\lim_{n\to\infty} \mathbb{P}\left[\tau \in \left(\hat{\tau}_{DM} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_{DM}/n}\right)\right] = 1-\alpha,$$

where Φ denotes the standard Gaussian cumulative distribution function.

From a certain perspective, the above is all that is needed to estimate average treatment effects in randomized trials. The difference in means estimator $\hat{\tau}_{DM}$ is consistent and allows for valid asymptotic inference; moreover, the estimator is very simple to implement, and hard to "cheat" with (there is little room for an unscrupulous analyst to try different estimation strategies and report the one that gives the answer closest to the one they want). On the other hand, it is far from clear that $\hat{\tau}_{DM}$ is the "optimal" way to use the data, in the sense that it provides the most accurate value of τ for a given sample size. Below, we try to see if/when we can do better.

Randomized trials in the linear model. To better understand the behavior of $\hat{\tau}_{DM}$, it is helpful to look at special cases. First, we consider the linear model: We observe $(X_i,\, Y_i,\, W_i)$ generated as

$$Y_i(w) = c_{(w)} + X_i\beta_{(w)} + \varepsilon_i(w), \quad \mathbb{E}\left[\varepsilon_i(w) \,\middle|\, X_i\right] = 0, \quad \text{Var}\left[\varepsilon_i(w) \,\middle|\, X_i\right] = \sigma^2. \quad (1)$$

Here, $\hat{\tau}_{DM}$ does not use the X_i; however, we will characterize its behavior in terms of the distribution of the X_i. Throughout our analysis, we assume for simplicity that we are in a balanced randomized trial, with

$$\mathbb{P}\left[W_i = 0\right] = \mathbb{P}\left[W_i = 1\right] = \frac{1}{2}.$$

[1]For a precise statement, one would need to worry about the case where n_0 or n_1 is 0.

Moreover, we assume (without loss of generality) that

$$\mathbb{E}[X] = 0, \quad \text{and define} \quad A = \text{Var}[X].$$

The assumption that $\mathbb{E}[X] = 0$ is without loss of generality because all estimators we will consider today are translation invariant (but of course the analyst cannot be allowed to make use of knowledge that $\mathbb{E}[X] = 0$).

Given this setup, we can write the asymptotic variance of $\hat{\tau}_{DM}$ as

$$\begin{aligned} V_{DM} &= \text{Var}[Y_i(0)] \,/\, \mathbb{P}[W_i = 0] + \text{Var}[Y_i(1)] \,/\, \mathbb{P}[W_i = 1] \\ &= 2\left(\text{Var}\left[X_i\beta_{(0)}\right] + \sigma^2\right) + 2\left(\text{Var}\left[X_i\beta_{(1)}\right] + \sigma^2\right) \\ &= 4\sigma^2 + 2\left\|\beta_{(0)}\right\|_A^2 + 2\left\|\beta_{(1)}\right\|_A^2 \\ &= 4\sigma^2 + \left\|\beta_{(0)} + \beta_{(1)}\right\|_A^2 + \left\|\beta_{(0)} - \beta_{(1)}\right\|_A^2, \end{aligned}$$

where we used the notation

$$\|v\|_A^2 = v'Av.$$

Is this the best possible estimator for τ?

OLS for randomized trials. If we assume the linear model (1), it is natural to want to use it for better estimation. Note that, given this model, we can write that ATE as

$$\tau = \mathbb{E}[Y(1) - Y(0)] = c_{(1)} - c_{(0)} + \mathbb{E}[X]\left(\beta_{(1)} - \beta_{(0)}\right).$$

This suggests an ordinary least-squares estimator

$$\hat{\tau}_{OLS} = \hat{c}_{(1)} - \hat{c}_{(0)} + \overline{X}\left(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}\right), \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

where the $(\hat{c}_{(w)},\, \hat{\beta}_{(w)})$ are obtained by running OLS on those observations with $W_i = w$. Standard results about OLS imply that (recall that, wlog, we work with $\mathbb{E}[X] = 0$)

$$\sqrt{n_w}\left(\begin{pmatrix}\hat{c}_{(w)}\\ \hat{\beta}_{(w)}\end{pmatrix} - \begin{pmatrix}c_{(w)}\\ \beta_{(w)}\end{pmatrix}\right) \Rightarrow \mathcal{N}\left(0,\, \sigma^2\begin{pmatrix}1 & 0\\ 0 & A^{-1}\end{pmatrix}\right).$$

In particular, we find that $\hat{c}_{(0)}$, $\hat{c}_{(1)}$, $\hat{\beta}_{(0)}$, $\hat{\beta}_{(1)}$ and $\overline{X}$ are all asymptotically independent. Then, we can write

$$\hat{\tau}_{OLS} - \tau = \underbrace{\hat{c}_{(1)} - c_{(1)}}_{\approx\mathcal{N}(0,\,\sigma^2/n_1)} - \underbrace{\hat{c}_{(0)} - c_{(0)}}_{\approx\mathcal{N}(0,\,\sigma^2/n_0)} + \underbrace{\overline{X}\left(\beta_{(1)} - \beta_{(0)}\right)}_{\approx\mathcal{N}\left(0,\,\left\|\beta_{(1)}-\beta_{(0)}\right\|_A^2/n\right)} + \underbrace{\overline{X}\left(\hat{\beta}_{(1)} - \hat{\beta}_{(0)} - \beta_{(1)} + \hat{\beta}_{(0)}\right)}_{\mathcal{O}_P(1/n)},$$

which leads us to the central limit theorem

$$\sqrt{n}\,(\hat{\tau}_{OLS} - \tau) \Rightarrow \mathcal{N}\,(0,\, V_{OLS})\,, \qquad V_{OLS} = 4\sigma^2 + \left\|\beta_{(0)} - \beta_{(1)}\right\|_A^2.$$

In particular, note that $V_{DM} = V_{OLS} + \left\|\beta_{(0)} + \beta_{(1)}\right\|_A^2$, and so OLS in fact helps reduce asymptotic error in the linear model.

OLS without linearity. The above result is perhaps not so surprising: If we assume a linear model, than using an estimator that leverages linearity ought to help. However, it is possible to prove a much stronger result for OLS in randomized trials: OLS is never worse that the difference-in-means methos in terms of its asymptotic variance, and usually improves on it (even in misspecified models).

Replace our linearity assumption with the following generic assumption:

$$Y_i(w) = \mu_{(w)}(X_i) + \varepsilon_i(w), \quad \mathbb{E}\left[\varepsilon_i(w) \,\middle|\, X_i\right] = 0, \quad \operatorname{Var}\left[\varepsilon_i(w) \,\middle|\, X_i\right] = \sigma^2,$$

for some arbitrary function $\mu_{(w)}(x)$. As before, we can check that (recall that we assume that $\mathbb{P}\left[W_i = 1\right] = 0.5$)

$$\sqrt{n}\,(\hat{\tau}_{DM} - \tau) \Rightarrow \mathcal{N}\,(0,\, V_{DM}) = 4\sigma^2 + 2\operatorname{Var}\left[\mu_{(0)}(X_i)\right] + 2\operatorname{Var}\left[\mu_{(1)}(X_i)\right],$$

and so $\hat{\tau}_{DM}$ provides a simple way of getting consistent estimates of τ.

In order to analyze OLS, we need to use the Huber-White analysis of linear regression. Without any assumption on $\mu_{(w)}(x)$, the OLS estimates $(\hat{c}_{(w)},\, \hat{\beta}_{(w)})$ converge to a limit characterized as

$$\left(c^*_{(w)},\, \beta^*_{(w)}\right) = \operatorname{argmin}_{c,\,\beta}\left\{\mathbb{E}\left[\left(Y_i(w) - X_i\beta - c\right)^2\right]\right\}.$$

If the linear model is misspecified, $(c^*_{(w)},\, \beta^*_{(w)})$ can be understood as those parameters that minimize the expected mean-squared error of any linear model. Given this notation, it is well known that (recall that we still assume wlog that $\mathbb{E}\left[X\right] = 0$)

$$\sqrt{n_w}\left(\begin{pmatrix}\hat{c}_{(w)}\\ \hat{\beta}_{(w)}\end{pmatrix} - \begin{pmatrix}c^*_{(w)}\\ \beta^*_{(w)}\end{pmatrix}\right) \Rightarrow \mathcal{N}\left(0,\, \begin{pmatrix}MSE^*_{(w)} & 0\\ 0 & \cdots\end{pmatrix}\right)$$
$$c^*_{(w)} = \mathbb{E}\left[Y_i(w)\right], \qquad MSE^*_{(w)} = \mathbb{E}\left[\left(Y_i(w) - X_i\beta^*_{(w)} - \hat{c}^*_{(w)}\right)^2\right]$$

Then, following the line of argumentation in the previous section, we can derive a central limit theorem

$$\sqrt{n}\,(\hat{\tau}_{OLS} - \tau) \Rightarrow \mathcal{N}\,(0,\, V_{OLS})\,,$$

with asymptotic variance[2]

$$
\begin{aligned}
V_{OLS} &= 2MSE^*_{(0)} + 2MSE^*_{(1)} + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|^2_A \\
&= 4\sigma^2 + 2\,\mathrm{Var}\left[\mu_{(0)}(X) - X\beta^*_{(0)}\right] \\
&\quad + 2\,\mathrm{Var}\left[\mu_{(1)}(X) - X\beta^*_{(1)}\right] + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|^2_A \\
&= 4\sigma^2 + 2\left(\mathrm{Var}\left[\mu_{(0)}(X)\right] - \mathrm{Var}\left[X\beta^*_{(0)}\right]\right) \\
&\quad + 2\left(\mathrm{Var}\left[\mu_{(1)}(X)\right] - \mathrm{Var}\left[X\beta^*_{(1)}\right]\right) + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|^2_A \\
&= 4\sigma^2 + 2\left(\mathrm{Var}\left[\mu_{(0)}(X)\right] + \mathrm{Var}\left[\mu_{(1)}(X)\right]\right) \\
&\quad + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|^2_A - 2\left\|\beta^*_{(0)}\right\|^2_A - 2\left\|\beta^*_{(1)}\right\|^2_A \\
&= 4\sigma^2 + 2\left(\mathrm{Var}\left[\mu_{(0)}(X)\right] + \mathrm{Var}\left[\mu_{(1)}(X)\right]\right) - \left\|\beta^*_{(0)} + \beta^*_{(1)}\right\|^2_A \\
&= V_{DM} - \left\|\beta^*_{(0)} + \beta^*_{(1)}\right\|^2_A .
\end{aligned}
$$

In other words, whether or not the true effect function $\mu_w(x)$ is linear, OLS always reduces the asymptotic variance of DM. Moreover, the amount of variance reduction scales by the amount by which OLS in fact chooses to fit the training data. A worst case for OLS is when $\beta^*_{(0)} = \beta^*_{(1)} = 0$, i.e., when OLS asymptotically just does nothing, and $\hat{\tau}_{OLS}$ reduces to $\hat{\tau}_{DM}$.

Recap. We discussed the following:

- A central object of interest in causal inference is the individual treatment effect $\Delta_i = Y_i(1) - Y_i(0)$.
- These effects Δ_i are fundamentally unknowable.
- However, a large randomized controlled trial lets us recover the average treatment effect $\tau = \mathbb{E}\left[\Delta_i\right]$.
- And, even without assuming linearity, OLS regression adjustments generally improve on the performance of the simple difference in means estimator.

Some closing thoughts:

- Throughout our analysis, we defined the target estimand $\tau = \mathbb{E}\left[\Delta_i\right]$ *before* making any modeling assumptions. Linear modeling was only used as a tool to estimate τ, but did not inform the scientific question we tried to answer.

[2]For the third equality, we use the fact that $X\beta^*_{(w)}$ is the projection of $\mu_{(w)}(X)$ on to the linear span of the features X, and so $\mathrm{Cov}[\mu_{(w)}(X),\, X\beta^*_{(w)}] = \mathrm{Var}[X\beta^*_{(w)}]$.

- In particular, we did *not* try to estimate τ via a linear regression model $Y_i \sim X_i\beta + W_i\tau + \varepsilon_i$. This approach has the vice of tying our scientific question to our modeling strategy: τ appears to just have become a coefficient in our linear model, not a fact of nature that's conceptually prior to modeling decisions. (Here, we ran two separate regressions, not just one regression with a specialized "treatment effect" coefficient.)

- Note that our OLS estimator can effectively be viewed as

$$\hat{\tau}_{OLS} = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\left(\hat{c}_{(1)} + X_i\hat{\beta}_{(1)}\right)}_{\hat{\mu}_{(1)}(X_i)} - \underbrace{\left(\hat{c}_{(0)} + X_i\hat{\beta}_{(0)}\right)}_{\hat{\mu}_{(0)}(X_i)}\right),$$

  where $\hat{\mu}_{(w)}(x)$ denotes OLS predictions at x. Could we use other methods to estimate $\hat{\mu}_{(w)}(x)$ rather than OLS (e.g., deep nets, forests)? How would this affect asymptotic variance? More on this in the homework.

Finally, for further reading, see the following:

Lin, Winston. **Agnostic notes on regression adjustments to experimental data: Reexamining Freedmans critique.** *The Annals of Applied Statistics*, 7(1), 2013.

Wager, Stefan, Wenfei Du, Jonathan Taylor, and Robert Tibshirani. **High-dimensional regression adjustments in randomized experiments.** *Proceedings of the National Academy of Sciences*, 113(45), 2016.