OIT 661: Lecture 2
Unconfoundedness and the Propensity Score
(for class use only, please do not distribute)

Beyond a single randomized controlled trial. 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. We assume SUTVA, $Y_i = Y_i(W_i)$, and want to estimate the average treatment effect

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

In the first lecture, we assumed random treatment assignment, $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i$, and studied several $\sqrt{n}$-consistent estimators for the ATE.

The simplest way to move beyond one RCT is to consider two RCTs. As a concrete example, supposed that we are interested in giving teenagers cash incentives to discourage them from smoking. A random subset of $\sim 5\%$ of teenagers in Palo Alto, CA, and a random subset of $\sim 20\%$ of teenagers in Geneva, Switzerland are eligible for the study.

Palo Alto	Non-S.	Smoker
Treat.	152	5
Control	2362	122

Geneva	Non-S.	Smoker
Treat.	581	350
Control	2278	1979

Within each city, we have a randomized controlled study, and in fact readily see that the treatment helps. However, looking at aggregate data is misleading, and it looks like the treatment hurts.

Palo Alto + Geneva	Non-Smoker	Smoker
Treatment	733	401
Control	4640	2101

Once we aggregate the data, this is no longer an RCT because Genevans are both more likely to get treated, and more likely to smoke whether or not they get treated. In order to get a consistent estimate of the ATE, we need to estimate treatment effects in each city separately:

$$\hat{\tau}_{\text{PA}} = \frac{5}{152+5} - \frac{122}{2362+122} \approx -1.7\%, \quad \hat{\tau}_{\text{GVA}} = \frac{350}{350+581} - \frac{1979}{2278+1979} \approx -8.9\%$$

$$\hat{\tau} = \frac{2641}{2641+5188}\hat{\tau}_{\text{PA}} + \frac{5188}{2641+5188}\hat{\tau}_{\text{GVA}} \approx -6.5\%.$$

How does this idea generalize to continuous x? How should we build a confidence interval?

Aggregating difference-in-means estimators. Suppose that we have covariates X_i that take values in a discrete space $X_i \in \mathcal{X}$, with $|\mathcal{X}| = p < \infty$. Suppose moreover that the treatment assignment is random conditionally on X_i, (i.e., we have an RCT in each bucket):

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i = x, \quad \text{for all} \quad x \in \mathcal{X}.$$

Define the bucket-wise average treatment effect as

$$\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right].$$

Then, as above, we can estimate the ATE τ as

$$\hat{\tau} = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \hat{\tau}(x), \quad \hat{\tau}(x) = \frac{1}{n_{x1}} \sum_{\{X_i = x,\, W_i = 1\}} Y_i - \frac{1}{n_{x0}} \sum_{\{X_i = x,\, W_i = 0\}} Y_i, \tag{1}$$

where $n_x = |\{i : X_i = x\}|$ and $n_{xw} = |\{i : X_i = x,\, W_i = w\}|$. How good is this estimator? Intuitively, we have needed to estimate $|\mathcal{X}| = p$ "parameters" so we might expect the variance to scale linearly with p?

To study this estimator it is helpful to write it as follows. First, for any bucket x, define $e(x)$ as the probability of getting treated in that bucket, $e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$, and note that

$$\sqrt{n_x}\,(\hat{\tau}(x) - \tau(x)) \Rightarrow \mathcal{N}\left(0,\, \frac{\mathrm{Var}\left[Y(0) \,\middle|\, X = x\right]}{1 - e(x)} + \frac{\mathrm{Var}\left[Y(1) \,\middle|\, X = x\right]}{e(x)}\right).$$

Furthermore, under the simplifying assumption that $\mathrm{Var}\left[Y(w) \,\middle|\, X = x\right] = \sigma^2(x)$ does not depend on w, we get

$$\sqrt{n_x}\,(\hat{\tau}(x) - \tau(x)) \Rightarrow \mathcal{N}\left(0,\, \frac{\sigma^2(x)}{e(x)(1 - e(x))}\right).$$

Next, for the aggregated estimator, defining $\hat{\pi}(x) = n_x/n$ as the fraction of observations with $X_i = x$ and $\pi(x) = \mathbb{P}\left[X_i = x\right]$ as its expectation, we have

$$\begin{aligned}\hat{\tau} = \sum_{x \in \mathcal{X}} \hat{\pi}(x)\hat{\tau}(x) = \underbrace{\sum_{x \in \mathcal{X}} \pi(x)\tau(x)}_{=\tau} + \underbrace{\sum_{x \in \mathcal{X}} \pi(x)\left(\hat{\tau}(x) - \tau(x)\right)}_{\approx \mathcal{N}\left(0,\, \sum_{x \in \mathcal{X}} \pi^2(x)\,\mathrm{Var}[\hat{\tau}(x)]\right)} \\ + \underbrace{\sum_{x \in \mathcal{X}} \left(\hat{\pi}(x) - \pi(x)\right)\tau(x)}_{\approx \mathcal{N}(0,\, n^{-1}\,\mathrm{Var}[\tau(X_i)])} + \underbrace{\sum_{x \in \mathcal{X}} \left(\hat{\pi}(x) - \pi(x)\right)\left(\hat{\tau}(x) - \tau(x)\right)}_{= \mathcal{O}_P(1/n)}.\end{aligned}$$

Putting the pieces together, we get $\sqrt{n}\,(\hat{\tau} - \tau) \Rightarrow \mathcal{N}\,(0,\, V_{BUCKET})$

$$\begin{aligned} V_{BUCKET} &= \text{Var}\left[\tau(X)\right] + \sum_{x \in \mathcal{X}} \pi^2(x) \frac{1}{\pi(x)} \frac{\sigma^2(x)}{e(x)(1-e(x))} \\ &= \text{Var}\left[\tau(X)\right] + \mathbb{E}\left[\frac{\sigma^2(X)}{e(X)(1-e(X))}\right]. \end{aligned} \tag{2}$$

Note that this does not depend on $|\mathcal{X}| = p$, the number of buckets(!)

The average treatment effect on the treated. In some applications, we may have a bucketed design where some buckets have no treated individuals (i.e., $n_{x1} = 0$). For example, in a medical trial, some buckets may consist of "healthy" individuals, none of who needed the medicine. In this case, the above estimator $\hat{\tau}_{BUCKET}$ cannot be run, and in fact the ATE isn't identified.

A common strategy is to change the question and ask whether the treatment helped those individuals who actually got treated, and estimate the average treatment effect on the treated (ATT):

$$\tau_{ATT} = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i\right]. \tag{3}$$

Often the ATT is scientifically more relevant than the ATE. For example, are you more interested in the effect of chemotherapy on random samples from the population, or on cancer patients who were prescribed chemotherapy by a doctor?

In the bucketed setup, a natural estimator for the ATT is

$$\hat{\tau} = \sum_{\{x \in \mathcal{X}: n_{x1} > 0\}} \frac{n_{x1}}{n_1} \hat{\tau}(x), \quad \hat{\tau}(x) = \frac{1}{n_{x1}} \sum_{\{X_i = x,\, W_i = 1\}} Y_i - \frac{1}{n_{x0}} \sum_{\{X_i = x,\, W_i = 0\}} Y_i, \tag{4}$$

where n_1 is the total number of treated units. Note that we avoid having to consider buckets with $n_{x1} = 0$.

Continuous X and the propensity score. We found that if $\mathcal{X}$ is discrete with a finite number of buckets, then the exact number of buckets $|\mathcal{X}| = p$ does not affect the accuracy of inference. However, if $\mathcal{X}$ is continuous, this result does not apply—unless we can create buckets of individuals who are "like" each other for the purpose of average treatment effect estimation.

As a first step in generalizing our analysis to the continuous-X case, we need to generalize the "RCT assumption in each bucket" assumption. Cosmetically, we just write the same thing,

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i, \tag{5}$$

but now interpret it differently: It means that we have measured enough covariates to capture any dependence between W_i and the potential outcomes. We call this assumption **unconfoundedness**.

Next, we need to find an analogue to "buckets". In doing so, the propensity score $e(x)$ plays a central role[1]

$$e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X = x\right].$$

We may suspect that this object is important because, with constant $\sigma^2(x)$, V_{BUCKET} only varies with $e(x)$. Mathematically, the key property of the propensity score is that it is a balancing score: If (5) holds, then in fact

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, e(X_i), \tag{6}$$

i.e., it actually suffices to control for $e(X)$ rather than X to remove biases associated with a non-random treatment assignment. We can verify this claim as follows:

$$\begin{aligned}
&\mathbb{P}\left[W_i = w \,\middle|\, \{Y_i(0),\, Y_i(1)\}, e(X_i)\right] \\
&\quad = \int_{\mathcal{X}} \mathbb{P}\left[W_i = w \,\middle|\, \{Y_i(0),\, Y_i(1)\}, X_i = x\right] \mathbb{P}\left[X_i = x \,\middle|\, e(X_i)\right]\; dx \\
&\quad = \int_{\mathcal{X}} \mathbb{P}\left[W_i = w \,\middle|\, X_i = x\right] \mathbb{P}\left[X_i = x \,\middle|\, e(X_i)\right]\; dx && \text{(unconf.)} \\
&\quad = e(X_i).
\end{aligned}$$

The upshot is that if we knew that our data fell into a finite number $j = 1, ..., J$ of strata S_j such that $e(x) = e_j$ is constant within each stratum S_j, then we could consistently estimate the ATE as

$$\hat{\tau} = \sum_{j=1}^{j} \frac{n_j}{n} \hat{\tau}_j, \quad \hat{\tau}_j = \frac{1}{n_{j1}} \sum_{\{X_i \in S_j,\, W_i = 1\}} Y_i - \frac{1}{n_{j0}} \sum_{\{X_i \in S_j,\, W_i = 0\}} Y_i. \tag{7}$$

To verify consistency of this estimator, we note that $\mathbb{E}\left[\hat{\tau}_j\right] = \tau_j$ for $\tau_j = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i \in S_j\right]$ because

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n_{jw}} \sum_{\{X_i \in S_j,\, W_i = w\}} Y_i\right] &= \mathbb{E}\left[Y_i \,\middle|\, X_i \in S_j,\, W_i = w\right] && \text{(IID)} \\
&= \mathbb{E}\left[Y_i(w) \,\middle|\, X_i \in S_j,\, W_i = w\right] && \text{(SUTVA)} \\
&= \mathbb{E}\left[Y_i(w) \,\middle|\, X_i \in S_j\right] && \text{(unconf.)}
\end{aligned}$$

[1]Note that we can define this object without any issues even when x is continuous; however, we can no longer trivially estimate it via $\hat{e}(x) = n_{x1}/n_x$ is this case.

Inverse-propensity weighting. The propensity-bucketed estimator (7) is conceptually nice, but relies on the often-not-true assumption that there is a finite number of known strata with a constant propensity score. Luckily, we can generalize this estimator. We start by re-writing it:

$$\begin{aligned}\hat{\tau} &= \sum_{j=1}^{j} \frac{n_j}{n} \left(\frac{1}{n_{j1}} \sum_{\{X_i \in S_j, W_i = 1\}} Y_i - \frac{1}{n_{j0}} \sum_{\{X_i \in S_j, W_i = 0\}} Y_i \right) \\ &= \frac{1}{n} \sum_{j=1}^{J} \left(\frac{1}{\hat{e}_j} \sum_{\{X_i \in S_j, W_i = 1\}} Y_i - \frac{1}{1 - \hat{e}_j} \sum_{\{X_i \in S_j, W_i = 0\}} Y_i \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right),\end{aligned}$$

where $\hat{e}_j = n_{j1}/n_j$ and $\hat{e}(x) = \hat{e}_j$ for all $x \in S_j$. Now, this functional form appears to suggest a much more general strategy for estimating the ATE:

1. Estimate the propensity score $\hat{e}(x)$ via any method (logistic regression, a forest, a deep net), and

2. Use it for inverse-propensity weighted estimation of the average treatment effect:

$$\hat{\tau}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right). \tag{8}$$

How good is this estimator? The simplest way to analyze it is by comparing it to an oracle that actually knows the propensity score:

$$\hat{\tau}^*_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right). \tag{9}$$

Suppose that we have "overlap", i.e., $\eta \leq e(x) \leq 1 - \eta$ for all $x \in \mathcal{X}$. Suppose moreover that $|Y_i| \leq M$, and that we know that $\sup_{x \in \mathcal{X}} |e(x) - \hat{e}(x)| = \mathcal{O}_P(a_n) \to 0$. Then, we can check that

$$|\hat{\tau}_{IPW} - \hat{\tau}^*_{IPW}| = \mathcal{O}_P \left(\frac{a_n M}{\eta} \right),$$

and so if $\hat{\tau}^*_{IPW}$ is consistent, then so is $\hat{\tau}_{IPW}$.

It remains to analyze the behavior of the oracle IPW estimator $\hat{\tau}^*_{IPW}$. First,

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}^*_{IPW}\right] &= \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right] && \text{(IID)} \\
&= \mathbb{E}\left[\frac{W_i Y_i(1)}{e(X_i)} - \frac{(1-W_i)Y_i(0)}{1-e(X_i)}\right] && \text{(SUTVA)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{W_i Y_i(1)}{e(X_i)} \,\middle|\, e(X_i)\right] - \mathbb{E}\left[\frac{(1-W_i)Y_i(0)}{1-e(X_i)} \,\middle|\, e(X_i)\right]\right] \\
&= \mathbb{E}\left[Y_i(1) - Y_i(0)\right] && \text{(unconf.)},
\end{aligned}$$

and so the oracle estimator is in fact unbiased. To study its variance, it is helpful to write (without loss of generality)

$$\begin{aligned}
Y_i(0) &= m(X_i) - (1-e(X_i))\tau(X_i) + \varepsilon_i(0), \quad \mathbb{E}\left[\varepsilon_i(0) \,\middle|\, X_i\right] = 0 \\
Y_i(1) &= m(X_i) + e(X_i)\tau(X_i) + \varepsilon_i(1), \quad \mathbb{E}\left[\varepsilon_i(1) \,\middle|\, X_i\right] = 0,
\end{aligned}$$

and assume for simplicity that $\text{Var}\left[\varepsilon_i(w) \,\middle|\, X_i = x\right] = \sigma^2(x)$ does not depend on w. Then, we can verify that (on the second line, the fact that the variances separate is non-trivial, and is a result of the parametrization)

$$\begin{aligned}
n\,\text{Var}\left[\hat{\tau}^*_{IPW}\right] &= \text{Var}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right] \\
&= \text{Var}\left[\frac{W_i m(X_i)}{e(X_i)} - \frac{(1-W_i)m(X_i)}{1-e(X_i)}\right] + \text{Var}\left[\tau(X_i)\right] \\
&\quad + \text{Var}\left[\frac{W_i \varepsilon_i}{e(X_i)} - \frac{(1-W_i)\varepsilon_i}{1-e(X_i)}\right] \\
&= \mathbb{E}\left[\frac{m^2(X_i)}{e(X_i)(1-e(X_i))}\right] + \text{Var}\left[\tau(X)\right] + \mathbb{E}\left[\frac{\sigma^2(X)}{e(X)(1-e(X))}\right].
\end{aligned}$$

Pulling everything together, we see that

$$\begin{aligned}
&\sqrt{n}\left(\hat{\tau}^*_{IPW} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V_{IPW^*}\right), \\
&V_{IPW^*} = \mathbb{E}\left[\frac{m^2(X_i)}{e(X_i)(1-e(X_i))}\right] + \text{Var}\left[\tau(X)\right] + \mathbb{E}\left[\frac{\sigma^2(X)}{e(X)(1-e(X))}\right].
\end{aligned} \tag{10}$$

Perhaps disappointingly, even the *oracle* IPW estimator is worse than the motivating "bucketed" estimator. And our bounds for the feasible IPW estimator (8) are strictly worse than those for the oracle estimator.

IPW versus bucketing Let's now go back to the simplest case, where $\mathcal{X}$ is discrete. We have discussed two estimators:

$$\hat{\tau}_{BUCKET} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)\,Y_i}{1-\hat{e}(X_i)}\right), \quad \hat{e}(x) = \frac{n_{x1}}{n_1},$$

$$\hat{\tau}^*_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)\,Y_i}{1-e(X_i)}\right).$$

Perhaps surprisingly, the "feasible" IPW estimator is actually better than the "oracle" IPW estimator. At a high level, the estimated propensity score corrects for local variability in the sampling distribution of the W_i (i.e., it accounts for the number of units that were *actually* treated in each bucket). This suggests that our above analysis is very loose, and that the IPW construction proposed at (8) is missing some important piece of the puzzle.

Recap. As discussed today,

- If we combine data from many small randomized trials, we obtain an observational study (where features $X_i \in \mathcal{X}$) indicate which randomized trial every subject belongs to.
- In this case, we can efficiently estimate the average treatment effect estimate by averaging bucket-wise treatment effect estimates—and the asymptotic variance (2) of doing so does not depend on the number of buckets.
- With continuous X, we can estimate average treatment effects by stratifying based on the propensity score, provided we assume unconfoundedness as in (5).
- A formal generalization of this argument also motivates inverse-propensity weighting (8). However, at least using a naive analysis, IPW does not appear to be as accurate as other considered methods.

For further reading, see the following:

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. **Efficient estimation of average treatment effects using the estimated propensity score.** *Econometrica*, 71(4), 2003.

Rosenbaum, Paul R., and Donald B. Rubin. **The central role of the propensity score in observational studies for causal effects.** *Biometrika*, 70(1), 1983.