Part 3: The propensity score

The propensity score

The propensity score is an **ubiquitous concept** in causal inference. Its definition is simple,

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

i.e., the propensity score measures the **probability of being treated** conditionally on X_i.

In a **randomized trial**, the propensity score is constant $e(x) = e_0 \in (0,\, 1)$.

- At least qualitatively, the variability of the propensity score gives a measure of how far we are from a randomized trial.

The propensity score

The key fact is that under **unconfoundedness**

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

the average treatment effect can be characterized as

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

This implies that the following **inverse-propensity weighted** estimator is unbiased for the average treatment effect:

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

The same idea underlies **importance weighting**, **Horvitz-Thompson sampling**, etc.

Inverse-propensity weighting

Inverse-propensity weighting is unbiased because:

$$
\begin{aligned}
\tau &= \mathbb{E}\left[Y_i(1) - Y_i(0)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i(1) \,\middle|\, X_i\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, X_i\right]\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[W_i \,\middle|\, X_i\right] \mathbb{E}\left[Y_i(1) \,\middle|\, X_i\right]}{e(X_i)} - \frac{\mathbb{E}\left[1 - W_i \,\middle|\, X_i\right] \mathbb{E}\left[Y_i(0) \,\middle|\, X_i\right]}{1 - e(X_i)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[W_i Y_i(1) \,\middle|\, X_i\right]}{e(X_i)} - \frac{\mathbb{E}\left[(1 - W_i)\, Y_i(0) \,\middle|\, X_i\right]}{1 - e(X_i)}\right] \\
&= \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)}\right].
\end{aligned}
$$

The 5-th equality depends on consistency of the **potential outcomes**, and the 4-th equality relies on **unconfoundedness**,

$$
\left[\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i\right] \,\big|\, X_i.
$$

Inverse-propensity weighting

We know that the **inverse-propensity weighted** estimator,

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

is unbiased for τ if we **know the propensity scores** $e(\cdot)$ a-priori,

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

When we don't know them, a natural idea is to first **estimate** $\hat{e}(\cdot)$ via some machine learning method (e.g., a forest), and then use

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

Is this any good?

Estimating propensity scores

We know that the **oracle IPW estimator** that gets to use the true propensity scores is unbiased:

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

It converges at a $\sqrt{n}$ rate, and satisfies a **central limit theorem**

$$\sqrt{n} \left(\hat{\tau}^*_{IPW} - \tau \right) \Rightarrow \mathcal{N} \left(0,\ \text{Var} \left[\frac{W_i Y_i}{e(X_i)} - \frac{1 - W_i Y_i}{1 - e(X_i)} \right] \right).$$

We can then re-express our feasible estimator (i.e., with estimated propensity scores) as

$$\hat{\tau}_{IPW} = \underbrace{\hat{\tau}^*_{IPW}}_{\text{a good estimator}} + \underbrace{\hat{\tau}_{IPW} - \hat{\tau}^*_{IPW}}_{\text{due to errors in } \hat{e}(\cdot)}.$$

Hope: Is the second (error) term "small"? Specifically, is the error term **lower order**, i.e., $\ll 1/\sqrt{n}$?

Estimating propensity scores

Let's try to **bound the error** using Cauchy-Schwarz:

$$
\begin{aligned}
&\hat{\tau}_{IPW} - \hat{\tau}^*_{IPW} \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{W_i}{\hat{e}(X_i)} - \frac{(1-W_i)}{1-\hat{e}(X_i)}\right) - \left(\frac{W_i}{e(X_i)} - \frac{(1-W_i)}{1-e(X_i)}\right)\right) Y_i \\
&\quad \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{W_i}{\hat{e}(X_i)} - \frac{(1-W_i)}{1-\hat{e}(X_i)}\right) - \left(\frac{W_i}{e(X_i)} - \frac{(1-W_i)}{1-e(X_i)}\right)\right)^2} \\
&\qquad \times \sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2} \\
&\quad \asymp \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{e}(X_i) - e(X_i))^2}.
\end{aligned}
$$

Estimating propensity scores

We've shown the error term to be on the same scale as the **root-mean square error** of $\hat{e}(X_i)$.

$$\hat{\tau}_{IPW} - \hat{\tau}^*_{IPW} \lesssim \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{e}(X_i) - e(X_i))^2}.$$

This is **not good enough**. We'd want error $\ll 1/\sqrt{n}$, but

- In a **parametric** problem, you get $RMSE \sim 1/\sqrt{n}$.
- In a **non-parametric** problem (the ones where we'd use machine learning), you get $RMSE \gg 1/\sqrt{n}$
- We essentially **never** get $RMSE \ll 1/\sqrt{n}$.

The error from replacing the true propensity scores $e(X_i)$ with estimates $\hat{e}(X_i)$ swamps the sampling error of the oracle estimator.

- The reason to use **machine learning** methods is that we hope for flexible **consistency** results in large samples.
- But even though you get consistency under flexible conditions, you're still left with **non-negligible errors** in finite samples.
- When using machine learning methods, one **cannot ignore** bias. A naïve approach that ignores bias will generally result in invalid confidence intervals.