OIT 661: Lecture 3
Balancing and Double Robustness
(for class use only, please do not distribute)

Recap: Inverse propensity weighting. 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 a single RCT is to consider many RCTs with different treatment probabilities. 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 unconfounded,

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

Then, defining the bucket-wise average treatment effect as

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

and assuming that $\text{Var}\left[Y(w) \,\big|\, X = x\right] = \sigma^2(x)$ does not depend on w, we can estimate the ATE τ as

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

where $n_x = |\{i : X_i = x\}|$ and $n_{xw} = |\{i : X_i = x,\, W_i = w\}|$, and verify that

$$\sqrt{n}\left(\hat{\tau}_{BUCKET} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V\right), \quad V = \text{Var}\left[\tau(X)\right] + \mathbb{E}\left[\frac{\sigma^2(X)}{e(X)(1 - e(X))}\right],$$

where $e(x) = \mathbb{P}\left[W_i = 1 \,\big|\, X_i = x\right]$ is the propensity score.

More generally, we found that the above estimator is a special case of an inverse-propensity weighted estimator,

$$\hat{\tau}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)}\right), \quad \text{with} \;\; \hat{e}(x) = \frac{n_{x1}}{n_x}.$$

This suggests a generalization beyond discrete-valued X. If X is continuous, we can no longer use $\hat{e}(x) = n_{1x}/n_x$; however, we can instead estimate $\hat{e}(x)$ using any method (such as logistic regression), and then use those propensities for computing $\hat{\tau}_{IPW}$. However, despite a promising motivation, this approach has its problems. First of all, we might think that the "oracle" IPW estimator that uses the true $e(\cdot)$ is a good idea,

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

But actually, we found that this oracle IPW estimator is asymptotically less accurate than the bucketed estimator with estimated propensities:

$$\lim_{n \to \infty} n \operatorname{Var}\left[\hat{\tau}_{IPW-O}\right] = V + \mathbb{E}\left[\frac{\big((1 - e(x))\mu_{(0)}(x) + e(x)\mu_{(1)}(x)\big)^2}{e(x)(1 - e(x))} \right],$$

where $\mu_{(w)}(x) = \mathbb{E}\left[Y(w) \,\middle|\, X_i = x\right]$. Moreover, for generically chosen $\hat{e}(x)$-estimators, we would expect the estimation error in the propensity score to make the variance even worse.

The above discussion strongly suggests that our understanding of inverse-propensity weighting is "missing" something. In the bucketed case (which we understand more fully), IPW with smart estimated propensities does better than oracle IPW, which again is better than IPW with non-smart estimated propensities.

How to estimate a propensity score. Our goal now is to develop a more general framework for obtaining "smart" $\hat{e}(x)$-estimates that inherit the nice properties of the bucketed estimator. To do so, we consider general estimators of the form

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\gamma}_{(1)}(X_i) W_i Y_i - \hat{\gamma}_{(0)}(X_i)(1 - W_i) Y_i \right),$$

where the $\hat{\gamma}_{(w)}(x)$ are some weighting functions (IPW uses $\hat{\gamma}_{(1)}(x) = 1/\hat{e}(x)$ and $\hat{\gamma}_{(0)}(x) = 1/(1 - \hat{e}(x))$.

Suppose, for now, that we have a linear-logistic specification (we assume that the covariates x include an intercept),

$$e(x) = 1/(1 + e^{-x \cdot \alpha}), \quad \text{and} \quad \mu_{(w)}(x) = x \cdot \beta_{(w)} \text{ for } w = 0,\, 1,$$

and write $Y_i = \mu_{(W_i)}(X_i) + \varepsilon_i$. Given this setup, we can verify that

$$\begin{aligned}\hat{\tau} &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\gamma}_{(1)}(X_i)W_i\left(X_i\cdot\beta_{(1)}+\varepsilon_i\right)+\hat{\gamma}_{(0)}(X_i)(1-W_i)\left(X_i\cdot\beta_{(0)}+\varepsilon_i\right)\right)\\ &= \underbrace{\overline{X}\cdot\left(\beta_{(1)}-\beta_{(0)}\right)}_{\text{a very good ATE estimator}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\gamma}_{(1)}(X_i)W_i\varepsilon_i-\hat{\gamma}_{(0)}(X_i)(1-W_i)\varepsilon_i\right)}_{\text{the noise we have to pay}}\\ &\quad+\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_{(1)}(X_i)W_iX_i-\overline{X}\right)\cdot\beta_{(1)}}_{\text{bad news}_{(1)}}-\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_{(0)}(X_i)(1-W_i)X_i-\overline{X}\right)\cdot\beta_{(0)}}_{\text{bad news}_{(0)}}.\end{aligned}$$

Wouldn't it be great if the "bad news" terms didn't exist?

Now, thanks to the linear outcome model, the first "bad news" term disappears whenever the γ-weighted average of the treated samples matches $\overline{X}$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_{(1)}(X_i)W_iX_i-\overline{X}=0.$$

A natural idea is to use γ-weights obtained from some logistic regression model, i.e., $\hat{\gamma}_{(1)}(x) = 1/\hat{e}(x) = 1+\exp(-x\cdot\hat{\alpha}_{(1)})$, but to solve for the parameters of the logistic regression by moment matching:

$$\frac{1}{n}\sum_{i=1}^{n}W_iX_i\left(1+e^{-X_i\cdot\hat{\alpha}_{(1)}}\right)-\overline{X}=0.$$

We call the resulting propensity estimate a covariate balancing propensity score (CBPS) to emphasize the fact that $\hat{\alpha}_{(1)}$ achieves moment balance between the features X_i in full sample and the γ-weighted features X_i in the treated sample.

Similarly, we set $\hat{\gamma}_{(0)}(x) = \exp(-x\cdot\hat{\alpha}_{(0)})/(1+\exp(-x\cdot\hat{\alpha}_{(0)})$ where $\hat{\alpha}_{(0)}$ is chosen such as to make the second bad-news term go away. We note that $\hat{\gamma}_{(0)}(x)$ and $\hat{\gamma}_{(1)}(x)$ do not use the same propensity model, but is that a problem?

We will verify below that the $\hat{\alpha}_{(w)}$ are $\sqrt{n}$-consistent, i.e.,

$$\left\|\hat{\alpha}_{(0)}-\alpha\right\|_2,\quad\left\|\hat{\alpha}_{(1)}-\alpha\right\|_2=\mathcal{O}_P\left(\frac{1}{\sqrt{n}}\right).$$

Then, under regularity assumptions (including overlap, $\eta\leq\mathbb{P}\left[W=1\,\middle|\,X=x\right]\leq 1-\eta$), we see that

$$\hat{\tau}_{CBPS}=\overline{X}\cdot\left(\beta_{(1)}-\beta_{(0)}\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i\varepsilon_i}{e(X_i)}-\frac{(1-W_i)\varepsilon_i}{e(X_i)}\right)+\mathcal{O}_P\left(\frac{1}{n}\right),$$

and so

$$\sqrt{n}\left(\hat{\tau}_{CBPS} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V\right), \quad V = \mathrm{Var}\left[\tau(X)\right] + \mathbb{E}\left[\frac{\sigma^2(X)}{e(X)(1-e(X))}\right].$$

This is the same asymptotic variance as for bucketing! Our use of balancing propensity score estimates enables us to to be more accurate than oracle IPW.

Consistency of the propensity model. It remains to check that $\left\|\hat{\alpha}_{(1)} - \alpha\right\|_2 = \mathcal{O}_P\left(1/\sqrt{n}\right)$. To do so, it is helpful to write the moment condition identifying $\hat{\alpha}_{(1)}$ as a loss-minimization problem:

$$\hat{\alpha}_{(1)} = \mathrm{argmin}_\alpha \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(W_i e^{-X_i \cdot \alpha} - (W_i - 1) X_i \cdot \alpha \right) \right\}.$$

Here the loss function is convex, and the previous moment condition arises by setting (the negative of) its derivative to 0, so the two expressions for $\hat{\alpha}_{(1)}$ are in fact the same. Now, noting standard facts about convex loss minimization, we know that $\hat{\alpha}_{(1)}$ must converge to a limit $\bar{\alpha}_{(1)}$ that solves the population score moment condition:

$$\left\|\hat{\alpha}_{(1)} - \bar{\alpha}_{(1)}\right\|_2 = \mathcal{O}_P\left(\frac{1}{\sqrt{n}}\right), \quad \mathbb{E}\left[-W_i X_i \left(1 + e^{-X_i \cdot \alpha}\right) + X_i\right] = 0.$$

Finally, we know that the parameter α appearing in the true propensity score $e(x) = 1/(1 + e^{-x \cdot \alpha})$ solves the population moment condition, and moreover it is the unique solution (because the moment condition is the derivative of the convex population loss with a unique minimum), and so $\bar{\alpha}_{(1)} = \alpha$ as desired.

Double robustness of CBPS. So far, we have showed that under the linear-logistic specification for the outcome and propensity models, $\hat{\tau}_{CBPS}$ attains the asymptotic variance V of bucketing. It is also interesting to examine the behavior of $\hat{\tau}_{CBPS}$ under the condition that only one of these two specifications holds:

(1) The outcome model is linear, but the propensity score $e(x)$ is not logistic. In this case it is no longer true that $\hat{\alpha}_{(1)}$ is consistent for α (because there is no α!). However, thanks to our solution strategy for $\hat{\alpha}_{(1)}$, we still have balance,

$$\frac{1}{n} \sum_{i=1}^{n} W_i X_i \left(1 + e^{-X_i \cdot \hat{\alpha}_{(1)}}\right) - \overline{X} = 0,$$

etc., and so by the same argument as above,

$$\hat{\tau}_{CBPS} = \underbrace{\overline{X} \cdot \left(\beta_{(1)} - \beta_{(0)}\right)}_{\text{a very good ATE estimator}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\gamma}_{(1)}(X_i)W_i\varepsilon_i - \hat{\gamma}_{(0)}(X_i)(1-W_i)\varepsilon_i\right)}_{\mathcal{O}_P(1/\sqrt{n}) \text{ error term}},$$

and so $\hat{\tau}_{CBPS}$ is still consistent, although it may not achieve the asymptotic variance V. Note that now, since we did not assume anything about the propensity scores, we don't have a clean characterization for what the $\hat{\gamma}_{(w)}(X_i)$ converge to anymore—but this doesn't matter for consistency, since they get multiplied by mean-zero noise.

(2) The propensity score $e(x)$ is logistic, but the outcome model is not linear. In this case, we can no longer expand out the error of $\hat{\tau}_{CPBS}$ in terms of $\beta_{(0)}$ and $\beta_{(1)}$, and need to stick with the more abstract form:

$$\hat{\tau}_{CBPS} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_iY_i}{\hat{e}\left(X_i;\, \hat{\alpha}_{(1)}\right)} - \frac{(1-W_i)Y_i}{1-\hat{e}\left(X_i;\, \hat{\alpha}_{(0)}\right)}\right),$$

where the parameters $\hat{\alpha}_{(w)}$ are solved for by moment matching. Now, even if we can't use linearity to leverage the balancing property, we still know that the $\hat{e}(x;\, \hat{\alpha}^{(w)})$ are consistent propensity estimates. Thus, we can analyze $\hat{\tau}_{CBPS}$ as an IPW estimator to show that it is consistent (although, again, it may not achieve the bucket variance V).

Thus, although we motivated CBPS in a setting with a linear outcome model and a logistic propensity model, the estimator will actually be consistent if either of these assumptions hold (although accuracy may suffer). This property is called double robustness, and was introduced by Jamie Robins in the 1990's.

Another doubly robust estimator. The construction of the CBPS estimator and the resulting double robustness property may have appeared ad-hoc. Here is another, seemingly completely different estimator that is also doubly robust. Write

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

and suppose we have access to estimators $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ to these quantities. We then write

$$\hat{\tau}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1-\hat{e}(X_i)}\right),$$

where DR stands for doubly robust. Claim: this estimator is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

(1) Suppose that $\hat{\mu}_{(w)}(x)$ is consistent, i.e., $\hat{\mu}_{(w)}(x) \approx \mu_{(w)}(x)$. Then,

$$\hat{\tau}_{DR} \approx \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right)}_{\text{a consistent treatment effect estimator}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(W_i\frac{Y_i - \mu_{(1)}(X_i)}{\hat{e}(X_i)} - (1-W_i)\frac{Y_i - \mu_{(0)}(X_i)}{1-\hat{e}(X_i)}\right)}_{\mathcal{O}_P(1/\sqrt{n})\text{ mean-zero noise}},$$

because $\mathbb{E}\left[Y_i - \mu_{(1)}(X_i) \,\middle|\, X_i, W_i\right] = 0$, and so the "garbage" propensity score weight $1/\hat{e}(X_i)$ is multiplied by mean-zero noise that makes it go away. Thus $\hat{\tau}_{DR}$ is consistent.

(2) Suppose that $\hat{e}(x)$ is consistent, i.e., $\hat{e}(x) \approx e(x)$. Then,

$$\hat{\tau}_{DR} \approx \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)}_{\text{the IPW estimator}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i)\left(1 - \frac{W_i}{e(X_i)}\right) - \hat{\mu}_{(0)}(X_i)\left(1 - \frac{1-W_i}{1-e(X_i)}\right)\right)}_{\mathcal{O}_P(1/\sqrt{n})\text{ mean-zero noise}},$$

because $\mathbb{E}\left[1 - W_i/e(X_i) \,\middle|\, X_i\right] = 0$, and so the "garbage" regression adjustment $\hat{\mu}_{(1)}(X_i)$ is multiplied by mean-zero noise that makes it go away. Thus $\hat{\tau}_{DR}$ is consistent.

Finally, what if both models are consistent? In the next class, we will study how the estimator errors in $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ interact. For now, let's just consider the oracle DR estimator

$$\hat{\tau}_{DR^*} = \frac{1}{n}\sum_{i=1}^{n}\left(\mu_{(1)}(X_i) - \mu_{(0)}(X_i) + W_i\frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)} - (1-W_i)\frac{Y_i - \mu_{(0)}(X_i)}{1-e(X_i)}\right).$$

We can readily verify that this estimator is unbiased, and that the terms with $Y_i - \mu_{(1)}(X_i)$ are mean-zero conditionally on X_i. This implies that $\sqrt{n}(\hat{\tau}_{DR^*} - \tau)$ is asymptotically normal with variance

$$\underbrace{\text{Var}\left[\mu_{(1)}(X_i) - \mu_{(0)}(X_i)\right]}_{=\text{Var}[\tau(X)]} + \underbrace{\text{Var}\left[W_i\frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)} - (1-W_i)\frac{Y_i - \mu_{(0)}(X_i)}{1-e(X_i)}\right]}_{=\mathbb{E}[\sigma^2(X)/(e(X)(1-e(X)))]},$$

and so $\hat{\tau}_{DR}$ also asymptotically achieves the same accuracy as the bucketed estimator.

Closing thoughts. In this lecture, we have discussed how to build "good" average treatment effect estimators in observational studies, that generalize desirable properties of the simple bucketed estimator for discrete X. We found the popular inverse-propensity estimator to be disappointing in terms of its asymptotic accuracy; however, we exhibited other estimators that do better. Intriguingly, both estimators tie together two seemingly unrelated properties: double robustness, and achieving the same asymptotic variance as the bucketed estimator.

For further reading, see the following:

Graham, Bryan S., Cristine Campos de Xavier Pinto, and Daniel Egel. **Inverse probability tilting for moment condition models with missing data.** *The Review of Economic Studies*, 79(3), 2012.

Imai, Kosuke, and Marc Ratkovic. **Covariate balancing propensity score.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 2014.

Kang, Joseph D. Y., and Joseph L. Schafer. **Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.** *Statistical Science*, 2007.