Causal Inference + Doubly Robust Estimation

Recap:

Covariate shift: $p_t(y|x) = p_s(y|x)$

Importance weighting: $L(\theta) = \mathbb{E}_{p_s}\left[\frac{p_t(x)}{p_s(x)} \ell(\theta; x)\right]$

p^* $\tilde{p}$

Causal inference:

$X, T, Y(0), Y(1)$

Unconfoundedness assumption: $Y(0), Y(1) \perp\!\!\!\perp T \mid X$

$\rightarrow$ IPW estimator: $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}[T_i=1]}{p(T=1|X_i)} - \frac{\mathbb{I}[T_i=0]}{p(T=0|X_i)}\right) Y_i$

Unbiased estimate of ATE $= \mathbb{E}[Y(1) - Y(0)]$

if treated, divide by treatment prob

if not treated, divide by non-treatment prob

This time: Compute variance of IPW estimator

Introduce variance reduction technique to improve estimator

Compute bias $\Rightarrow$ show how to approximately de-bias estimator to make it more robust.

Variance of IPW estimator:

$$\mathbb{E}_p\left[\left(\frac{\mathbb{I}[T=1]}{p(T|x)} - \frac{\mathbb{I}[T=0]}{p(T|x)}\right) Y(T)\right]$$

at most one term is non-zero

Variance of avg. of n samples is $\frac{1}{n} \cdot \mathrm{Var}_p\Big[\underbrace{\left(\frac{\mathbb{I}[T=1]}{p(T|x)} - \frac{\mathbb{I}[T=0]}{p(T|x)}\right) Y(T)}_{Q}\Big]$

$$\mathrm{Var}_p[Q] = \underbrace{\mathbb{E}_p[Q^2]}_{2} - \underbrace{\mathbb{E}_p[Q]^2}_{1}$$

$$\left(\mathbb{E}_p[Y(1)-Y(0)]\right)^2$$

$$\mathbb{E}_p\left[\left(\frac{\mathbb{I}[T=1]}{p(T|X)^2} + \frac{\mathbb{I}[T=0]}{p(T|X)^2}\right) Y(T)^2\right]$$

$$= \mathbb{E}_{X\sim p}\left[\frac{p(T=1|X)}{p(T=1|X)^2} Y(1)^2 + \frac{p(T=0|X)}{p(T=0|X)^2} Y(0)^2\right]$$

$$= \mathbb{E}_{X\sim p}\left[\frac{Y(1)^2}{p(T=1|X)} + \frac{Y(0)^2}{p(T=0|X)}\right].$$

dominant term in variance of IPW estimate.

Issues:

1) Variance can explode if treatment probability is very close to 0 or 1

2) Assumes $p(T|X)$ is known, but in reality need to fit a model. Modeling error can introduce additional error (typically bias).

See how to partially mitigate these

Basic idea: Variance reduction.

Assume we have some prediction $\bar{Y}_1(X)$ of what happens if $T=1$ and $\bar{Y}_0(X)$ of what happens if $T=0$

Hope: $\bar{Y}_1(X) \approx Y(1)$

$\bar{Y}_0(X) \approx Y(0)$

Key observation: $\bar{Y}$ only depends on X, so don't need T

$$ATE = \mathbb{E}_p[Y(1) - Y(0)]$$

Think of as new "treatment effect" and apply IPW

$$= \mathbb{E}_p\Big[(\bar{Y}_1(x) - \bar{Y}_0(x)) + \underbrace{(Y(1) - \bar{Y}_1(x))}_{\text{"treatment effect"}} - \underbrace{(Y(0) - \bar{Y}_0(x))}_{}\Big]$$

$$= \mathbb{E}_p[\bar{Y}_1(x) - \bar{Y}_0(x)] + \mathbb{E}_p\left[\left(\frac{\mathbb{I}[T=1]}{p(T=1|x)} - \frac{\mathbb{I}[T=0]}{p(T=0|x)}\right)(Y(T) - \bar{Y}_T(x))\right]$$

only requires X

"doubly robust estimator" (DRE)

Both can be estimated from observed data.

Can be much smaller than $Y(T)$ itself

Bias of doubly robust estimator.

Suppose we have estimate $q(T|x)$ of propensity to treat.

Claim. DRE is unbiased if either $q = p$ or $\bar{Y}_T(x) = \mathbb{E}[Y(T)|X,T]$

Better claim. Bias of DRE is "product of biases" in q and $\bar{Y}$.

Computing bias?

$$ATE = \mathbb{E}_p[Y(1) - Y(0)]$$

$$DRE = \mathbb{E}_p[\bar{Y}_1(x) - \bar{Y}_0(x)] + \mathbb{E}_p\left[\left(\frac{\mathbb{I}[T=1]}{q(T=1|x)} - \frac{\mathbb{I}[T=0]}{q(T=0|x)}\right)(Y(T) - \bar{Y}_T(x))\right]$$

$$ATE - DRE = \mathbb{E}_p\left[\boxed{(Y(1) - \bar{Y}_1(x))} - \boxed{(Y(0) - \bar{Y}_0(x))}\right] - \mathbb{E}_p\left[\frac{\mathbb{I}[T=1]}{q(T=1|x)}\boxed{(Y(1) - \bar{Y}_1(x))} - \frac{\mathbb{I}[T=0]}{q(T=0|x)}\boxed{(Y(0) - \bar{Y}_0(x))}\right]$$

Focus on $T=1$ terms

$$\mathbb{E}_p\left[(Y(1) - \bar{Y}_1(x))\left(1 - \frac{\mathbb{I}[T=1]}{q(T=1|x)}\right)\right]$$

$$= \mathbb{E}_{x\sim p}\left[(\mathbb{E}[Y(1)|x] - \bar{Y}_1(x))\left(1 - \frac{p(T=1|x)}{q(T=1|x)}\right)\right]$$

$$\leq \mathbb{E}_{X\sim p}\Big[\underbrace{\big(\mathbb{E}[Y(1)|X] - \tilde{Y}_1(X)\big)^2}_{\text{"bias in } \tilde{Y}_1\text{"}}\Big]^{1/2} \underbrace{\mathbb{E}_{X\sim p}\Big[\Big(1 - \frac{p(T=1|X)}{q(T=1|X)}\Big)^2\Big]}_{\text{"bias in } q\text{"}}^{1/2}$$

$\Longrightarrow$ Formalizes idea that bias of DRE is $\leq$ product of biases of $\tilde{Y}$ and q.

Variance of estimate,

Focus on $T=1$ term:

$$\mathrm{Var}\left[(Y(1) - \tilde{Y}_1(X))\,\frac{\mathbb{I}[T=1]}{q(T=1|X)}\right]$$

$$\leq \mathbb{E}\left[(Y(1) - \tilde{Y}_1(X))^2\,\frac{\mathbb{I}[T=1]^2}{q(T=1|X)^2}\right]$$

$$= \mathbb{E}_{X\sim p}\left[\mathbb{E}\left[(Y(1) - \tilde{Y}_1(X))^2 \cdot \frac{\mathbb{I}[T=1]}{q(T=1|X)^2}\,\middle|\, X\right]\right]$$

these two terms are independent given X

$$= \mathbb{E}_{X\sim p}\Big[\underbrace{\mathbb{E}[(Y(1) - \tilde{Y}_1(X))^2 | X]}_{\text{term for squared error}} \cdot \underbrace{\frac{p(T=1|X)}{q(T=1|X)^2}}_{\text{term for } q(\cdot|X)}\Big]$$

Error of empirical DRE, i.e. avg. over n samples:

- $\text{Bias}^2 + \frac{1}{n}\text{Variance}$

$$\leq \left(\sum_T \overbrace{\mathbb{E}_{x\sim p}\left[\left(\mathbb{E}[Y(T)\mid X] - \hat{Y}_T(X) \right)^2 \right]^{1/2}}^{\text{"Bias}^2\text{ of } \hat{Y}\text{"}} \cdot \overbrace{\mathbb{E}_{x\sim p}\left[\left(1 - \frac{p(T\mid x)}{q(T\mid x)} \right)^2 \right]^{1/2}}^{\text{"Bias}^2\text{ of } q\text{"}} \right)^2$$

$$+ \frac{1}{n}\mathbb{E}_{x\sim p}\left[\mathbb{E}\left[\left(Y(T) - \hat{Y}_T(X) \right)^2 \mid X \right] \cdot \frac{p(T=1\mid X)}{q(T=1\mid X)^2} \right]$$

This term could be non-zero even as $n \to \infty$.

Solution: Semiparametric estimation.

- Use non-parametric estimators (i.e. kernel ridge regression) to estimate q and $\hat{Y}$.
- Suppose each of those converges to truth at rate $n^{-\beta}$
- Then Bias^2 converge at rate of

$n^{-2\beta}$. As long as $\beta > \frac{1}{2}$, get parametric $\frac{1}{n}$ rate.

Nie and Wager, 2017

Practical recipe.

- Collect your data
- Fit neural network model to g and $\vec{\gamma}$
- Apply DRE w/ g and $\vec{\gamma}$.

Other variations.

- Leverage additional unlabeled data
- Two-phase survey design

Examples of variance reduction.

- Predict GPA from previous year's GPA
- Temperature
- Chronic disease.

- Effect of NPIs: previous cases, extrapolated forward