Lec 9: Estimation of Average Treatment Effect

Yanjun Han

Nov 12, 2024

Potential outcome model.

For a binary treatment $W \in \{0,1\}$, an individual i has two potential outcomes $Y_i(1)$ and $Y_i(0)$ ← the outcome individual i would have experienced had he/she received the treament or not, respectively

Average Treatment Effect (ATE):

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

A typical dataset: $\{(X_i, W_i, Y_i)\}_{i=1}^{n}$:

- $W_i \in \{0,1\}$: indicator of treatment / control
- $Y_i \in \mathbb{R}$: observed outcome $Y_i = Y_i(W_i)$
- $X_i \in \mathbb{R}^p$ (optional) : feature of individual i

(Optional material: SUTVA — stable unit treatment value assumption

"the potential outcomes for any unit do not vary with the treatments assigned to each other unit, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes", e.g.

- you taking the aspirin cannot have an affect on my headache
- different aspirins should have the same strength)

Randomized control trials (RCT) (no X_i)

Assumption: $\begin{cases} W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) & \text{(random treatment assignment)} \\ \text{each } i \text{ has the same marginal prob. of getting treated} \end{cases}$

Difference-in-mean estimation:

$$\boxed{\hat{\tau}_{DM} = \frac{1}{n_1}\sum_{W_i=1} Y_i - \frac{1}{n_0}\sum_{W_i=0} Y_i, \quad \text{where } n_j = \#\{i : W_i = j\}}$$

Unbiasedness of $\hat{\tau}_{DM}$:

Recap: properties of conditional expectation

1) Tower property:

$$\mathbb{E}[\mathbb{E}[Y \mid X_1, X_2] \mid X_1] = \mathbb{E}[Y \mid X_1]$$

2) Take out what's known:

$$\mathbb{E}[Y f(X) \mid X] = f(X)\,\mathbb{E}[Y \mid X]$$

3) independence: if $Y_1 \perp Y_2 \mid X$,

$$\mathbb{E}[f(Y_1) g(Y_2) \mid X] = \mathbb{E}[f(Y_1) \mid X] \cdot \mathbb{E}[g(Y_2) \mid X]$$

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n_1}\sum_{W_i=1} Y_i\right] &= \mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n} W_i Y_i\right] \\
&= \mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n} W_i Y_i(1)\right] \quad \text{(SUTVA)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n} W_i Y_i(1) \,\Big|\, \{Y_i(0), Y_i(1)\}_{i=1}^{n}, n_1\right]\right] \\
&= \mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n} Y_i(1) \cdot \mathbb{E}[W_i \mid n_1]\right] \quad \text{(random treatment assignment)} \\
&= \mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n} Y_i(1) \cdot \frac{n_1}{n}\right] \quad \text{(same marginal prob.)} \\
&= \mathbb{E}[Y_i(1)]
\end{aligned}$$

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

Propensity score.

Question. What happens if we combine two RCTs, but with different treatment probabilities?

Failure of $\hat{\tau}_{DM}$: Simpson's Paradox

Example: discourage teenagers from smoking in Palo Alto, CA (5%) & NYC (20%)

Palo Alto	Non-S.	Smoker
Treat.	95	5
Control	1700	200

19 : 1 vs. 8.5 : 1

treatment effect: +

$+$

NYC	Non-S.	Smoker
Treat.	255	145
Control	800	800

1.76 : 1 vs. 1 : 1

treatment effect: +

$=$

All	Non-S.	Smoker
Treat.	350	150
Control	2500	1000

2.33 : 1 vs. 2.5 : 1

treatment effect: − (!!)

Implication: propensity score plays a central role !

$$\boxed{\text{Propensity score:} \quad e(x) = \mathbb{P}(W_i = 1 \mid X_i = x)}$$

Assumptions: 1. unconfoundedness: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i$

(no unexplained feature affects both W_i & $(Y_i(0), Y_i(1))$)

2. overlap: $\eta \le e(x) \le 1-\eta$ for all x.

<u>Inverse-propensity weighting (IPW)</u>.

<u>Theorem</u>. $$\mathbb{E}\Big[\underbrace{\frac{WY}{e(X)} - \frac{(1-W)Y}{1-e(X)} - \tau}_{f_{\tau,e}(W,X,Y):\ \text{estimating function}}\Big] = 0$$

<u>Pf</u>.
$$\begin{aligned}
\mathbb{E}\left[\frac{WY}{e(X)}\right] &= \mathbb{E}\left[\frac{WY(1)}{e(X)}\right] && \text{(SUTVA)} \\
&= \mathbb{E}\left\{\mathbb{E}\left[\frac{WY(1)}{e(X)} \,\Big|\, X\right]\right\} \\
&= \mathbb{E}\left\{\frac{1}{e(X)}\mathbb{E}[W \mid X]\,\mathbb{E}[Y(1) \mid X]\right\} && \text{(unconfoundedness)} \\
&= \mathbb{E}\left\{\mathbb{E}[Y(1) \mid X]\right\} && (e(X) = \mathbb{P}(W=1 \mid X)) \\
&= \mathbb{E}[Y(1)]. && \square
\end{aligned}$$

IPW estimator: given an estimate $\hat{e}(x)$ for $e(x)$, then

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

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

Pros: consistent ($\hat{\tau}_{IPW} \to \tau$ as $n \to \infty$)

Cons: potentially large variance ;
not robust to nuisance estimation error $\hat{e}(x) - e(x)$

Double robust estimation: Augmented IPW (AIPW).

Model.
$$\begin{cases} Y = \mu_W(X) + \varepsilon_W, & \mathbb{E}[\varepsilon_0 | W, X] = 0, \ \mathbb{E}[\varepsilon_1 | W, X] = 0. \\ W \sim \mathrm{Bern}(e(X)) \end{cases}$$

Target parameter: $\tau = \mathbb{E}[\mu_1(X) - \mu_0(X)]$

Nuisance parameter: mean outcomes $\mu_0(X)$, $\mu_1(X)$
propensity score $e(X)$

AIPW estimator. Given nuisance estimates $(\hat{\mu}_1(x), \hat{\mu}_0(x), \hat{e}(x))$:

$$\boxed{\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + W_i\frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_0(X_i)}{1-\hat{e}(X_i)}\right)}$$

Interpretation: 1. from IPW, subtract the mean outcomes $(\hat{\mu}_0(X_i), \hat{\mu}_1(X_i))$ from Y_i ;

2. from $\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$, debias using IPW applied to the regression residuals.

Double machine learning in practice.

1. Split the dataset into K folds ;
2. For $k=1, \cdots, K$, use all data but the k-th fold to estimate $(\hat{\mu}_1^{(-k)}(x), \hat{\mu}_0^{(-k)}(x), \hat{e}^{(-k)}(x))$, possibly via machine learning ;
3. Estimate ATE by

i belongs to k_i-th fold

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\mu}_1^{(-k_i)}(X_i) - \hat{\mu}_0^{(-k_i)}(X_i) + W_i \frac{Y_i - \hat{\mu}_1^{(-k_i)}(X_i)}{\hat{e}^{(-k_i)}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_0^{(-k_i)}(X_i)}{1-\hat{e}^{(-k_i)}(X_i)} \Big)$$

Theoretical properties.

$$f_{(\mu_1,\mu_0,e,\tau)}(W,X,Y) = \mu_1(X) - \mu_0(X) + W\frac{Y-\mu_1(X)}{e(X)} - (1-W)\frac{Y-\mu_0(X)}{1-e(X)} - \tau$$

Claim 1: f is an estimating function, i.e.

$$\mathbb{E}[f_{(\mu_1,\mu_0,e,\tau)}(W,X,Y)] = 0.$$

Pf.
$$\mathbb{E}\Big[W\frac{Y-\mu_1(X)}{e(X)}\Big]$$
$$= \mathbb{E}\Big[W\frac{Y(1)-\mu_1(X)}{e(X)}\Big] \quad \text{(SUTVA)}$$
$$= \mathbb{E}\Big[\frac{W\varepsilon_1}{e(X)}\Big]$$
$$= \mathbb{E}\Big\{\mathbb{E}\Big[\frac{W\varepsilon_1}{e(X)} \Big| W, X\Big]\Big\}$$
$$= 0 \quad (\mathbb{E}[\varepsilon_1 | W, X] = 0, \text{ or unconfoundedness})$$

$$\Rightarrow \mathbb{E}[f] = \mathbb{E}[\mu_1(X) - \mu_0(X)] - \tau$$
$$= \mathbb{E}[\varepsilon_0 - \varepsilon_1] = 0. \qquad \square$$

Claim 2: f is Neyman orthogonal, i.e.

$$\mathbb{E}[\nabla_g f_{(\mu_1,\mu_0,e,\tau)}(W,X,Y)] = 0, \quad \forall g \in \{\mu_0, \mu_1, e\}.$$

(Implication: nuisance estimation errors only have second-order effects on the estimation of τ)

Pf. (1) $g=\mu_1$:
$$\mathbb{E}[\nabla_{\mu_1} f] = \mathbb{E}\left[1 - \frac{W}{e(X)} \,\middle|\, X\right]$$
$$= 1 - \mathbb{E}\left[\frac{W}{e(X)} \,\middle|\, X\right]$$
$$= 0 \quad (P(W=1 \mid X) = e(X))$$

(2) $g=\mu_0$:
$$\mathbb{E}[\nabla_{\mu_0} f] = \mathbb{E}\left[-1 + \frac{1-W}{1-e(X)} \,\middle|\, X\right]$$
$$= -1 + \mathbb{E}\left[\frac{1-W}{1-e(X)} \,\middle|\, X\right]$$
$$= 0 \quad (P(W=0 \mid X) = 1-e(X))$$

(3) $g=e$:
$$\mathbb{E}[\nabla_e f] = \mathbb{E}\left[-\frac{W(Y-\mu_1(X))}{e(X)^2} + \frac{(1-W)(Y-\mu_0(X))}{(1-e(X))^2} \,\middle|\, X\right]$$
$$= \mathbb{E}\left[-\frac{W\varepsilon_1}{e(X)^2} + \frac{(1-W)\varepsilon_0}{(1-e(X))^2} \,\middle|\, X\right] \text{ (SUTVA)}$$
$$= \mathbb{E}\left\{\mathbb{E}\left[-\frac{W\varepsilon_1}{e(X)^2} + \frac{(1-W)\varepsilon_0}{(1-e(X))^2} \,\middle|\, W,X\right] \,\middle|\, X\right\}$$
$$= 0 \quad (\mathbb{E}[\varepsilon_1, \varepsilon_0 \mid W, X] = 0, \text{ or unconfoundedness}) \quad \square$$

Claim 3: f is (weakly) double robust, i.e.

$$\mathbb{E}[f_{(\hat{\mu}_1,\hat{\mu}_0,\hat{e},\tau)}(W,X,Y)] = 0 \quad \text{if} \quad (\hat{\mu}_1,\hat{\mu}_0) = (\mu_1,\mu_0) \text{ } \underline{\text{OR}} \text{ } \hat{e} = e.$$

(Implication: AIPW is consistent if either $(\hat{\mu}_1(x), \hat{\mu}_0(x))$ are consistent, or $\hat{e}(x)$ is consistent)

Pf. (1) If $(\hat{\mu}_1, \hat{\mu}_0) = (\mu_1, \mu_0)$: same argument in Claim 1

(2) If $\hat{e} = e$, rewrite

$$f_{(\hat{\mu}_1, \hat{\mu}_0, e, \tau)}(W, X, Y) = \underbrace{\frac{WY}{e(X)} - \frac{(1-W)Y}{1-e(X)} - \tau}_{\mathbb{E}[\cdot]=0 \text{ by IPW analysis}}$$

$$\underbrace{- (W - e(X)) \left(\frac{\hat{\mu}_1(X)}{e(X)} - \frac{\hat{\mu}_0(X)}{1-e(X)} \right)}_{\mathbb{E}[\cdot] = \mathbb{E}\{\mathbb{E}[\cdot \mid X]\} = 0 \text{ since } \mathbb{P}(W=1 \mid X) = e(X).}$$

$\square$

Derivation of AIPW (Optional)

First derivation : use efficient influence

(see J. Hahn, "On the role of propensity score in efficient semiparametric estimation of average treatment effects", Econometrica, 1998)

Second derivation : find the projection of IPW

$$f_{\tau,e}(W, X, Y) = \frac{WY}{e(X)} - \frac{(1-W)Y}{1-e(X)} - \tau$$

to the orthogonal complement of L, where

$$L = \{ g(W, X, Y) : \mathbb{E}[g \mid X, Y(0), Y(1)] = 0 \}.$$

Lemma 1 $L = \{ (W - e(X)) h(X) \text{ for general } h \}$

Pf. Obviously $\mathbb{E}[(W - e(X)) h(X) \mid X, Y(0), Y(1)] = h(X)\, \mathbb{E}[W - e(X) \mid X] = 0.$

Now we show that any $g(W, X, Y) \in L$ must take this form.

$$\mathbb{E}[g \mid X, Y(0), Y(1)] = e(X) \underbrace{g(1, X, Y(1))}_{\equiv g_1(X)} + (1 - e(X)) \underbrace{g(0, X, Y(0))}_{\equiv g_0(X)} \equiv 0$$

$$\Rightarrow \frac{g_1(X)}{1-e(X)} = -\frac{g_0(X)}{e(X)} =: h(X)$$

$$\Rightarrow g(W,X,Y) = \begin{cases} g_1(X) & \text{if } W=1 \\ g_0(X) & \text{if } W=0 \end{cases} = (W-e(X))\,h(X) \qquad \square.$$

Lemma 2. $\mathrm{Proj}_{L^\perp}(f_{\tau,e}(W,X,Y)) = f_{(\mu_0,\mu_1,e,\tau)}(W,X,Y)$,

the estimating function of AIPW.

Pf. Aim to find $h_0(X)$ s.t.

$$\mathbb{E}\left[\left(\frac{WY}{e(X)} - \frac{(1-W)Y}{1-e(X)} - \tau - (W-e(X))h_0(X)\right) \times (W-e(X))h(X)\right] = 0, \ \forall h$$

$$\Rightarrow 0 = \mathbb{E}\left[\left(\frac{WY}{e(X)} - \frac{(1-W)Y}{1-e(X)} - \tau - (W-e(X))h_0(X)\right)(W-e(X)) \,\Big|\, X\right]$$

$$= \mu_1(X)(1-e(X)) - \mu_0(X)(-e(X)) - e(X)(1-e(X))h_0(X)$$

$$\Rightarrow h_0(X) = \frac{\mu_1(X)}{e(X)} + \frac{\mu_0(X)}{1-e(X)}$$

Therefore, $\mathrm{Proj}_{L^\perp}(f_{\tau,e}(W,X,Y))$

$$= \frac{WY}{e(X)} - \frac{(1-W)Y}{1-e(X)} - \tau - (W-e(X))\left(\frac{\mu_1(X)}{e(X)} + \frac{\mu_0(X)}{1-e(X)}\right)$$

$$= \mu_1(X) - \mu_0(X) - \tau + W\frac{Y-\mu_1(X)}{e(X)} - (1-W)\frac{Y-\mu_0(X)}{1-e(X)}$$

$$= f_{(\mu_0,\mu_1,e,\tau)}(W,X,Y) \qquad \square$$