

Lec 9: Estimation of Average Treatment Effect

Yanjin Han

Nov 7, 2023



Potential outcome model.

For a binary treatment $W \in \{0, 1\}$, an individual i has two potential outcomes $Y_i(1)$ and $Y_i(0)$ \leftarrow the outcome individual i would have experienced had he/she received the treatment 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 effect on my headache
- different aspirins should have the same strength

Randomized control trials (RCT) (no X_i)

Assumption: $\left\{ \begin{array}{l} W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \text{ (random treatment assignment)} \\ \text{each } i \text{ has the same marginal probab of getting treated} \end{array} \right.$

Difference-in-mean estimation:

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

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

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

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

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		NYC	Non-S.	Smoker		All	Non-S.	Smoker
Treat.	95	5	+	Treat.	255	145	=	Treat.	350	150
Control	1700	200		Control	800	800		Control	2500	1000
19:1 vs. 8.5:1				1.76:1 vs. 1:1				2.33:1 vs. 2.5:1		
treatment effect: +				treatment effect: +				treatment effect: - (!!)		

Implication: propensity score plays a central role!

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

- Assumptions:
- unconfoundedness: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i$
(no unexplained feature affects both W_i & $(Y_i(0), Y_i(1))$)
 - overlap: $\eta \leq e(x) \leq 1 - \eta$ for all x .

Inverse-propensity weighting (IPW).

Theorem. $\mathbb{E} \left[\underbrace{\frac{WY}{e(x)} - \frac{(1-W)Y}{1-e(x)}}_{f_{\tau, e}(W, X, Y): \text{estimating function}} - \tau \right] = 0$

Pf. $\mathbb{E} \left[\frac{WY}{e(x)} \right] = \mathbb{E} \left[\frac{WY(1)}{e(x)} \right] \quad (\text{SUTVA})$

Side note:

- estimating function (sometimes called score function; we won't use this name in the course):

$$f_{\theta}(x) \text{ s.t. } \mathbb{E}_{X \sim p_0} [f_{\theta}(x)] = 0;$$

- estimating equation:

$$\frac{1}{n} \sum_{i=1}^n f_{\theta}(x_i) = 0.$$

$$= \mathbb{E} \left\{ \mathbb{E} \left[\frac{WY(1)}{e(x)} \mid X \right] \right\}$$

$$= \mathbb{E} \left\{ \frac{1}{e(x)} \mathbb{E}[W \mid X] \mathbb{E}[Y(1) \mid X] \right\} \quad (\text{unconfoundedness})$$

$$= \mathbb{E} \left\{ \mathbb{E}[Y(1) \mid X] \right\} \quad (e(x) = \mathbb{P}(W=1 \mid X))$$

$$= \mathbb{E}[Y(1)].$$

□

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

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

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

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

Double robust estimation: Augmented IPW (AIPW).

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

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 ($\hat{\mu}_1(x)$, $\hat{\mu}_0(x)$, $\hat{e}(x)$):

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

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 $k=1, \dots, K$, 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

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

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

$$\begin{aligned} \text{Pf. } & \mathbb{E} \left[W \frac{Y - \mu_1(X)}{e(X)} \right] \\ &= \mathbb{E} \left[W \frac{Y(1) - \mu_1(X)}{e(X)} \right] \quad (\text{SUTVA}) \\ &= \mathbb{E} \left[\frac{W \varepsilon_1}{e(X)} \right] \\ &= \mathbb{E} \left\{ \mathbb{E} \left[\frac{W \varepsilon_1}{e(X)} \mid W, X \right] \right\} \\ &= 0 \quad (\mathbb{E}[\varepsilon_1 \mid W, X] = 0, \text{ or unconfoundedness}) \end{aligned}$$

$$\begin{aligned} \Rightarrow \mathbb{E}[f] &= \mathbb{E}[\mu_1(X) - \mu_0(X)] - \tau \\ &= \mathbb{E}[\varepsilon_0 - \varepsilon_1] = 0. \end{aligned}$$

□

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

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

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

$$\begin{aligned} \text{Pf. (1) } g = \mu_1: \quad \mathbb{E}[\nabla_{\mu_1} f] &= \mathbb{E}\left[1 - \frac{W}{e(X)} \mid X\right] \\ &= 1 - \mathbb{E}\left[\frac{W}{e(X)} \mid X\right] \\ &= 0 \quad (\mathbb{P}(W=1 \mid X) = e(X)) \end{aligned}$$

$$\begin{aligned} \text{(2) } g = \mu_0: \quad \mathbb{E}[\nabla_{\mu_0} f] &= \mathbb{E}\left[-1 + \frac{1-W}{1-e(X)} \mid X\right] \\ &= -1 + \mathbb{E}\left[\frac{1-W}{1-e(X)} \mid X\right] \\ &= 0 \quad (\mathbb{P}(W=0 \mid X) = 1 - e(X)) \end{aligned}$$

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

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

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

(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 $\mathbb{E}[\cdot] = 0$ by IPW analysis

$$f(\hat{\mu}_1, \hat{\mu}_0, e, \tau)(W, X, Y) = \frac{WY}{e(X)} - \frac{(1-W)Y}{1-e(X)} - \tau$$

$$- (W - e(X)) \left(\frac{\hat{\mu}_1(X)}{e(X)} - \frac{\hat{\mu}_0(X)}{1-e(X)} \right)$$

$$\mathbb{E}[\cdot] = \mathbb{E}\{\mathbb{E}[\cdot | X]\} = 0$$

Since $P(W=1 | X) = e(X)$.

□

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

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

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

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

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

$$\mathbb{E}[g | X, Y(0), Y(1)] = e(X) \underbrace{g(1, X, Y(1))}_{\equiv j_1(X)} + (1-e(X)) \underbrace{g(0, X, Y(0))}_{\equiv j_0(X)} \equiv 0$$

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

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

Lemma 2. $\text{Prj}_{L^\perp}(f_{\tau, e}(W, X, Y)) = f_{(\mu_0, \mu_1, e, \tau)}(W, X, Y)$,
the estimating function of AIPW.

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

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

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

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

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

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

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

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

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