Lecture 5: Causal Estimation

Tianpei Xie

Sep. 22nd., 2022

Contents

1 Recall what we learned

- Given a popuduation, we can compute the ***average treatment effect (ATE)***, or, *average causal effect (ACE)* [Imbens and Rubin, 2015, Rosenbaum, 2017, Neal, 2020] by taking an average over the ITEs:

  $$\tau := \mathbb{E}\left[Y_i(1) - Y_i(0)\right] = \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right] \tag{1}$$

  where the average is over the individuals i if $Y_i(t)$ is deterministic. If $Y_i(t)$ is random, the average is also over any other randomness.

- We have the following assumptions that are commonly used

  - **Assumption 1.1** ***(Ignorability / Exchangeability)*** *[Neal, 2020]*

    $$(Y(1), Y(0)) \perp\!\!\!\perp T \tag{2}$$

  - **Assumption 1.2** ***(Conditional Exchangeability / Unconfoundedness)*** *[Neal, 2020]*

    $$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X \tag{3}$$

  - **Assumption 1.3** ***(Positivity / Overlap / Common Support)*** *[Neal, 2020]*
    For all values of covariates x present in the population of interest (i.e. x such that $P(X = x) > 0$),

    $$0 < P(T = 1|X = x) < 1 \tag{4}$$

  - **Assumption 1.4** ***(No Interference)*** *[Neal, 2020]*

    $$Y_i(t_1, \ldots, t_i, \ldots, t_n) = Y_i(t_i) \tag{5}$$

  - **Assumption 1.5** ***(Consistency)*** *[Neal, 2020]*
    If the treatment is T, then the observed outcome Y is the potential outcome under treatment T. Formally,

    $$T = t \Rightarrow Y = Y(t) \tag{6}$$

    We could write this equivalently as follow:

    $$Y = Y(T) \tag{7}$$

  - ***Stable unit-treatment value assumption (SUTVA)*** is satisfied if unit (individual) i's outcome is simply a **function** of unit i's treatment. Therefore, SUTVA is a combination of consistency and no interference (and also ***deterministic*** potential outcomes).

- **Theorem 1.6** ***(Adjustment Formula)*** *[Neal, 2020]*
  Given the assumptions of ***unconfoundedness****,* ***positivity****,* ***consistency****, and* ***no interference****, we can identify the average treatment effect:*

  $$\tau = \mathbb{E}\left[Y(1) - Y(0)\right] = \mathbb{E}_X\left[\mathbb{E}\left[Y \,|T = 1,\, X\right] - \mathbb{E}\left[Y \,|T = 0,\, X\right]\right]$$

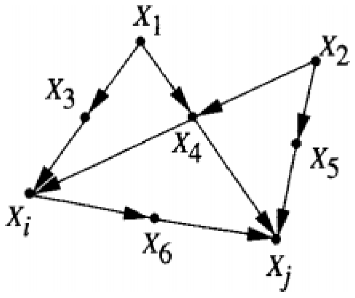

Figure 3.4 A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j \mid \hat{x}_i)$. Adjusting for $\{X_4\}$ or $\{X_6\}$ would yield a biased estimate.

Figure 1: The back-door adjustment [Pearl, 2000]

- **Definition** (***Structural causal models (SCMs)***) [Peters et al., 2017]
  A **SCM** $\mathfrak{C} := (S, P_{\boldsymbol{N}})$ with graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a collection S of d ***(structural) assignments***:

  $$X_s := f_s(X_{\pi(s)}, N_s), \quad s = 1, \ldots, d \tag{8}$$

  where $X_{\pi(s)}$ are called **parents** of X_s; and a joint distribution $P_{\boldsymbol{N}} = \prod_{s=1}^{d} P_{N_s}$ over the noise variables, which we require to be ***jointly independent***.

  The $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of a SCM is obtained by creating one vertex for each X_s and drawing **directed edges** from each parent in $X_{\pi(s)}$ to X_s, that is, from each variable X_k occurring on the right-hand side of equation (8) to X_s. $\mathcal{G}$ is a **directed acyclic graph (DAG)**.

  We sometimes call the elements of $X_{\pi(s)}$ not only parents but also ***direct causes*** of X_s, and we call X_s a ***direct effect*** of each of its direct causes. SCMs are also called (***nonlinear***) ***SEMs***.

- **Definition** (***Back-Door***) [Pearl, 2000]
  A set of variables Z satisfies the ***back-door criterion*** relative to an ***ordered*** pair of variables $(X_i \rightarrow X_j)$ in a DAG $\mathcal{G}$ if:

  1. ***no*** node in Z is a ***descendant*** of X_i; and
  2. Z ***blocks every path*** between X_i and X_j that contains an arrow ***into*** X_i.

  Similarly, if X and Y are two **disjoint** subsets of nodes in $\mathcal{G}$, then Z is said to satisfy ***the back-door criterion*** relative to (X, Y) if it satisfies the criterion relative to *any pair* (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

- Satisfying the back-door criterion makes Z a ***sufficient adjustment set***. The main insight of the graphical approach to covariate adjustment is that the adjustment set must **block all *noncausal* paths without blocking** any ***causal*** paths between X and Y.

- **Theorem 1.7** *(**Back-Door Adjustment**) [Pearl, 2000, Neal, 2020]*
  *If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is **identifiable** and is given by the formula*

  $$P(y \mid do(x)) = \sum_{z} P(y \mid x, z) P(z) \tag{9}$$

  To see why this works we need to know that $P(z|do(x)) = P(z)$ since by back-door criterion,

Z has no descendant of X. Also $P(y \mid do(x), z) = P(y \mid x, z)$ since Z blocks all paths from X to Y, so by modularity

- **Definition** (***Front-Door***) [Pearl, 2000]
  A set of variables M is said to satisfy the ***front-door criterion*** relative to an ***ordered*** pair of variables (T, Y) if:

  1. M **intercepts all** directed paths from T to Y;
  2. there is ***no unblocked back-door path*** from T to M; and
  3. ***all back-door paths*** from M to Y are ***blocked*** by T.

- A set of variables M ***completely mediates*** the effect of T on Y if all causal (directed) paths from T to Y go through M. If M satisfies the front-door criterion, M is a set of complete mediators.

- **Theorem 1.8** *(**Front-Door Adjustment**) [Pearl, 2000, Neal, 2020]*
  *If M satisfies the front-door criterion relative to (T, Y) and if $P(t, m) > 0$, then the causal effect of T on Y is **identifiable** and is given by the formula*

$$P(y| \, do(t)) = \sum_{m} P(m| \, t) \sum_{t'} P(y| \, m, t') \; P(t') \tag{10}$$

- **Proposition 1.9** *(**Rules of do Calculus**)[Pearl, 2000]*
  *Let $\mathcal{G}$ be the directed acyclic graph associated with a causal model as defined in (**??**), (**??**), and let $P(\cdot)$ stand for the probability distribution induced by that model. For any **disjoint** subsets of variables X, Y, Z, and W, we have the following rules.*

  1. *(**Insertion/deletion of observations**):*

$$p(y|\hat{x}, z, w) = p(y|\hat{x}, w) \quad \textit{if } (Y \perp\!\!\!\perp Z|X, W)_{\widehat{\mathcal{G}}_X} \tag{11}$$

  where $\hat{x} := do(X = x)$ and $\widehat{\mathcal{G}}_X$ is induced sub-graph under intervention $\hat{x}$.

  2. *(**Action/observation exchange**):*

$$p(y|\hat{x}, \hat{z}, w) = p(y|\hat{x}, z, w) \quad \textit{if } (Y \perp\!\!\!\perp Z|X, W)_{\widehat{\mathcal{G}}_{X,Z}} \tag{12}$$

  where $\widehat{\mathcal{G}}_{X,Z}$ is induced sub-graph under intervention $\hat{x}, \hat{z}$.

  3. *(**Insertion/deletion of actions**):*

$$p(y|\hat{x}, \hat{z}, w) = p(y|\hat{x}, w) \quad \textit{if } (Y \perp\!\!\!\perp Z|X, W)_{\widehat{\mathcal{G}}_{X,Z(W)}} \tag{13}$$

  *where $Z(W)$ is the set of Z-nodes that are **not ancestors** of any W-node in $\widehat{\mathcal{G}}_X$.*

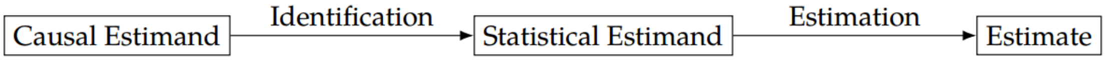

Figure 2.5: The Identification-Estimation Flowchart – a flowchart that illustrates the process of moving from a target causal estimand to a corresponding estimate, through identification and estimation.

Figure 2: The basic process of identification and estimation in causal inference. [Neal, 2020]

2 Conditional Outcome Modeling (COM)

2.1 Basic concepts

- From Adjustment Formula, we see that

  $$\tau = \mathbb{E}\left[Y(1) - Y(0)\right] = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}\left[Y\,|T=1,\, \boldsymbol{X}\right] - \mathbb{E}\left[Y\,|T=0,\, \boldsymbol{X}\right]\right]$$

  On the left hand side, it is a causal estimand and on the right hand side, it is a statistical estimand.

- To estimate the ATE τ, we need to fit a model for $\mathbb{E}\left[Y\,|T,\, \boldsymbol{X}\right]$ and then approximate $\mathbb{E}_{\boldsymbol{X}}\left[\delta_{\boldsymbol{X}}\right]$, with an empirical mean over the n data points.

  $$\begin{aligned} \delta_X &:= \mu(1, \boldsymbol{x}) - \mu(0, \boldsymbol{x}) = \mathbb{E}\left[Y\,|T=1,\, \boldsymbol{X}\right] - \mathbb{E}\left[Y\,|T=0,\, \boldsymbol{X}\right] \\ &\text{where } \mu(t, \boldsymbol{x}) := \mathbb{E}\left[Y\,|T=t,\, \boldsymbol{X}=\boldsymbol{x}\right] \end{aligned} \tag{14}$$

  Here the estimated model $\hat{\mu}(t, \boldsymbol{x})$ is called ***conditional outcome model (COM)***. Then the estimated ATE is

  $$\hat{\tau} = \sum_{i=1}^{n} \left(\hat{\mu}(1, \boldsymbol{x}_i) - \hat{\mu}(0, \boldsymbol{x}_i)\right) \tag{15}$$

  We will refer to estimators that take this form as ***conditional outcome model (COM) estimators***.

- For each individual covariate, we may be interested in the ***conditional average treatment effect (CATE)*** $\tau(\boldsymbol{x})$:

  $$\tau(\boldsymbol{x}) = \mathbb{E}\left[Y(1) - Y(0)\,|\,\boldsymbol{X} = \boldsymbol{x}\right] \tag{16}$$

  The X that is conditioned on does not need to consist of all of the observed covariates, but this is often the case when people refer to CATEs. For each individual subject, we call that ***individualized average treatment effects (IATEs)***.

- When we are interested in the CATE, we can split the covariate set as $W \cup X$, where X are observed covariates, then the COM model

  $$\mu(t, \boldsymbol{w}, \boldsymbol{x}) := \mathbb{E}\left[Y|\,T=t,\, W=\boldsymbol{w},\, \boldsymbol{X}=\boldsymbol{x}\right]$$

  Then CATE estimator is

  $$\hat{\tau}(\boldsymbol{x}) = \sum_{i \in \{i : \boldsymbol{x}_i = \boldsymbol{x}\}} \left(\hat{\mu}(1, \boldsymbol{w}_i, \boldsymbol{x}) - \hat{\mu}(0, \boldsymbol{w}_i, \boldsymbol{x})\right) \tag{17}$$

and the IATE estimator is

$$\hat{\tau}(\boldsymbol{x}_i) = \hat{\mu}(1, \boldsymbol{w}_i, \boldsymbol{x}_i) - \hat{\mu}(0, \boldsymbol{w}_i, \boldsymbol{x}_i) \tag{18}$$

Even, though IATEs are different from ITEs ($\tau(\boldsymbol{x}_i) \neq \tau_i$), if we really want to give estimates for ITEs, it is relatively common to take this estimator as our estimator of the ITE τ_i as well.

- COM estimators have many different names in the literature. For example, they are often called ***G-computation estimators***, ***parametric G-formula***, or ***standardization*** in epidemiology and biostatistics [Imbens and Rubin, 2015]. Because we are fitting a *single statistical model* for μ here, "COM estimator" is sometimes referred to as an *S-learner*, where the S stands for single.

2.2 Grouped Conditional Outcome Modeling (GCOM)

- COM above train a single model $\mu(t, \boldsymbol{x}) = \mathbb{E}\left[Y \,|T = t,\, \boldsymbol{X} = \boldsymbol{x}\right]$ based on $(\boldsymbol{X}, T, Y)$, where T is ***1-dimensional*** and $\boldsymbol{X}$ is *multi-dimensional.* It is likely that models would ignore variable T and focus on the rest of covariates. But T is only thing changing between two terms. This would result in an ATE estimate of zero.
- The solution is to train **two** models, one for treatment group and one for control group:

$$\mu_1(\boldsymbol{x}) := \mathbb{E}\left[Y \,|\, T = 1,\, \boldsymbol{X} = \boldsymbol{x}\right], \quad \mu_0(\boldsymbol{x}) := \mathbb{E}\left[Y \,|\, T = 0,\, \boldsymbol{X} = \boldsymbol{x}\right] \tag{19}$$

Using two separate models for the values of treatment ensures that T cannot be ignored. These models are called ***grouped conditional outcome model (GCOM)***.

- The ***grouped conditional outcome model (GCOM) estimator*** for ATE is

$$\hat{\tau} = \sum_{i=1}^{n} \left(\hat{\mu}_1(\boldsymbol{x}_i) - \hat{\mu}_0(\boldsymbol{x}_i)\right) \tag{20}$$

- Similarly, for CATE,

$$\hat{\tau}(\boldsymbol{x}) = \sum_{i \in \{i : \boldsymbol{x}_i = \boldsymbol{x}\}} \left(\hat{\mu}_1(\boldsymbol{w}_i, \boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{w}_i, \boldsymbol{x})\right) \tag{21}$$

- The GCOM may fix the zero estimate for ATE but it has a **major drawback**: each model is trained on **a sub-population of data** (i.e. treatment group or control group). This cause a significant drop in ***data efficiency***.

3 Increasing Data Efficiency: TARNet and X-learning

- One way to mitigate the data efficiency issue is to learn a ***treatment-agnostic representation (TAR)*** of covariates $\boldsymbol{X}$, using all of the data while still forcing the model to not ignore T by **branching** into ***two heads*** for the different values of T. In other words, ***TARNet*** [Shalit et al., 2017] uses the knowledge we have about T (as a uniquely important variable) in its architecture. See details in [Neal, 2020].

- Another solution is to use ***X-learners*** [Künzel et al., 2019, Neal, 2020], which is neither COM or GCOM. These models use all of the data for both models that are part of the estimators.
- Note that the CATE has ***Observational-Counterfactual Decomposition*** [Neal, 2020].

$$\begin{aligned} \mathbb{E}\left[Y(1) - Y(0)\,|\,\boldsymbol{X} = \boldsymbol{x}\right] = e(\boldsymbol{x})\tau_1(\boldsymbol{x}) + (1 - e(\boldsymbol{x}))\tau_0(\boldsymbol{x}) & \\ \tau_1(\boldsymbol{x}) = \mathbb{E}\left[Y\,|\,T = 1,\, \boldsymbol{X} = \boldsymbol{x}\right] - \mathbb{E}\left[Y(0)|T = 1,\, \boldsymbol{X} = \boldsymbol{x}\right] & \quad (22) \\ := \mathbb{E}\left[Y - \mu_0(\boldsymbol{X})\,|\,T = 1,\, \boldsymbol{X} = \boldsymbol{x}\right] & \\ \tau_0(\boldsymbol{x}) = \mathbb{E}\left[Y(1)\,|\,T = 0,\, \boldsymbol{X} = \boldsymbol{x}\right] - \mathbb{E}\left[Y\,|\,T = 0,\, \boldsymbol{X} = \boldsymbol{x}\right] & \quad (23) \\ := \mathbb{E}\left[\mu_1(\boldsymbol{X}) - Y\,|\,T = 0,\, \boldsymbol{X} = \boldsymbol{x}\right] & \end{aligned}$$

where $e(\boldsymbol{x}) = P(T = 1|\boldsymbol{X} = \boldsymbol{x})$ is the ***propensity score***. $\tau_1(\boldsymbol{x})$ and $\tau_0(\boldsymbol{x})$ are causal estimands.

- Like GCOM, X-learner first estimates the ***conditional potential outcomes***

$$\mu_1(\boldsymbol{x}) = \mathbb{E}\left[Y(1)|\boldsymbol{X} = \boldsymbol{x}\right] = \mathbb{E}\left[Y|\, do(T = 1),\, \boldsymbol{X} = \boldsymbol{x}\right], \quad (24)$$

$$\mu_0(\boldsymbol{x}) = \mathbb{E}\left[Y(0)|\boldsymbol{X} = \boldsymbol{x}\right] = \mathbb{E}\left[Y|\, do(T = 0),\, \boldsymbol{X} = \boldsymbol{x}\right] \quad (25)$$

using

$$\hat{\mu}_1(\boldsymbol{x}) = \widehat{\mathbb{E}}_{model}\left[Y|\,T = 1,\, \boldsymbol{X} = \boldsymbol{x}\right]$$

$$\hat{\mu}_0(\boldsymbol{x}) = \widehat{\mathbb{E}}_{model}\left[Y|\,T = 0,\, \boldsymbol{X} = \boldsymbol{x}\right].$$

The main difference is that $\hat{\mu}_1(\boldsymbol{x})$ and $\hat{\mu}_0(\boldsymbol{x})$ are served as a ***counterfactual term*** in the target, where they are used to train treatment responses $\hat{\tau}_0(\boldsymbol{x})$ and $\hat{\tau}_1(\boldsymbol{x})$ in ***alternative*** group as compared to their training set .

There are three steps to X-learning:

1. Learning **two models** $\hat{\mu}_0(\boldsymbol{x})$ and $\hat{\mu}_1(\boldsymbol{x})$ from using *control group data* and *treatment group data*, respectively.
2. Then we define the ***target*** in the treatment and control group using the estimated ITE

$$\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(\boldsymbol{x}_i) \quad (26)$$

$$\hat{\tau}_{0,i} = \hat{\mu}_1(\boldsymbol{x}_i) - Y_i(0) \quad (27)$$

   Here, $\hat{\tau}_{1,i}$ is estimated using the ***treatment group outcomes*** and the ***imputed counterfactual*** in $\hat{\mu}_0(\boldsymbol{x}_i)$, which was learned from the ***control group data***. Similarly, $\hat{\tau}_{0,i}$ is estimated using the ***control group outcomes*** and the ***imputed counterfactual*** $\hat{\mu}_1(\boldsymbol{x}_i)$ using ***treatment group data***.

   This method is called X**-learning** since the observational and counterfactual terms are arranged in "X" shaped in (26) and (27).

   We can fit model $\hat{\tau}_1(\boldsymbol{x}_i)$ using data $\{(\boldsymbol{x}_i, \hat{\tau}_{1,i}), i \in \mathcal{T}\}$ from the ***treatment group*** subjects. Similarly, we fit model $\hat{\tau}_0(\boldsymbol{x}_i)$ using data $\{(\boldsymbol{x}_j, \hat{\tau}_{0,j}), j \in \mathcal{C}\}$ from the ***control group*** subjects.
3. Finally, the estimated CATE is

$$\hat{\tau}(\boldsymbol{x}) = \hat{e}(\boldsymbol{x})\hat{\tau}_1(\boldsymbol{x}) + (1 - \hat{e}(\boldsymbol{x}))\hat{\tau}_0(\boldsymbol{x}) \quad (28)$$

   where $\hat{e}(\boldsymbol{x}) \in [0, 1]$ is an estimate of propensity score $e(\boldsymbol{x})$.

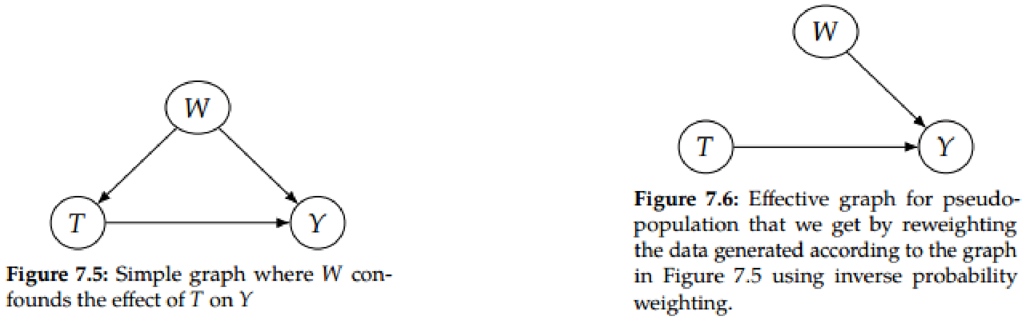

Figure 7.5: Simple graph where W confounds the effect of T on Y

(a)

Figure 7.6: Effective graph for pseudo-population that we get by reweighting the data generated according to the graph in Figure 7.5 using inverse probability weighting.

(b)

Figure 3: The effective causal graph for pseudo-population generated via inverse probability weighting

4 Inverse Probability Weighting (IPW)

- ***Matching perspective***: Remember during matching with propensity score, we match subjects in treatment and control group with the same propensity score. Suppose there is a subset of subjects in both groups whose propensity scores are matched. Instead of choosing only a pair of them for matching, we can **choose all of them as matched** and ***downweight*** the contribution of each sample ***uniformly***. There are in total $P(T = 1|\boldsymbol{w}_r) \times N_{\boldsymbol{w}_r}$ matched treated subjects. So we need to reweighting these subjects by the ***inverse probability of treatment*** $\frac{1}{P(T=1|\boldsymbol{w}_r)}$. Similarly, the matched control subjects is reweighted by $\frac{1}{P(T=0|\boldsymbol{w}_r)}$.

  By reweighting, each matched sample contributed the same, which is equivalent to selecting just one pair of samples from matched subsets.

- ***Oversampling perspective***: In an *observational study*, certain groups may be ***oversampled*** relative to the hypothetical sample from a ***randomized trial*** due to existence of **confounder $\boldsymbol{W}$**.

  Specifically, in Figure 3 (a), the group of treated data with confounder $\boldsymbol{W} = \boldsymbol{w}$ are sampled according to the propensity score $P(T = 1|\, \boldsymbol{w})$, which is different from the hypothetical sampling probability $P(T = 1)$ under randomized trial. $P(T = 1|\, \boldsymbol{w}) \neq P(T = 1)$.

  We can ***resample*** the population with probability $\frac{1}{P(T=1|\, \boldsymbol{W})}$ for the treatment group and $\frac{1}{P(T=0|\, \boldsymbol{W})}$ for the control group. By this way, we get a ***pseudo-population*** where $P(T = 1|\, \boldsymbol{W}) = P(T = 1)$ or equals some constant; the important part is that we make T independent of $\boldsymbol{W}$. In this pseudo-population, the treatment group and control group are *balanced*, i.e. ***everyone is equally likely to be treated***. See Figure 3 (b). Large weights from IPW means that the given subject was likely to be treated, given their covariates, *but wasn't*.

- We can derive the ***Inverse Probability Weighting (IPW)** estimator*

$$\mathbb{E}\left[Y(t)\right] = \mathbb{E}_{\boldsymbol{W},Y}\left[\frac{\mathbb{1}\left\{T = t\right\}\, Y}{P(T = t\,|\,\boldsymbol{W}\,)}\right] \tag{29}$$

  This is a ***causal identification equation***, where the left hand side is a causal estimand and the right hand side is a statistical estimand. The proof follows from the adjustment formula

$$\mathbb{E}\left[Y(t)\right] = \mathbb{E}_{\boldsymbol{W}}\left[\mathbb{E}\left[Y\,|\,t, \boldsymbol{W}\right]\right]$$

$$
\begin{aligned}
&= \sum_{\boldsymbol{w}} \sum_{y} y\, P(y|\, t, \boldsymbol{w}) P(\boldsymbol{w}) \\
&= \sum_{\boldsymbol{w}} \sum_{y} y\, \frac{P(y|\, t, \boldsymbol{w}) P(t \,|\, \boldsymbol{w}\,) P(\boldsymbol{w})}{P(t \,|\, \boldsymbol{w}\,)} \\
&= \sum_{\boldsymbol{w}} \sum_{y} y\, P(y, t, \boldsymbol{w})\, \frac{1}{P(t \,|\, \boldsymbol{w}\,)} \\
&= \sum_{\boldsymbol{w}} \mathbb{E}\left[\mathbb{1}\left\{\boldsymbol{W} = \boldsymbol{w} \wedge T = t\right\} Y\right]\, \frac{1}{P(t \,|\, \boldsymbol{w}\,)} \\
&= \mathbb{E}_{\boldsymbol{W},Y}\left[\frac{\mathbb{1}\left\{T = t\right\} Y}{P(t \,|\, \boldsymbol{W}\,)}\right] \qquad \blacksquare
\end{aligned}
$$

- Assuming binary treatment (and ***exchangeablity*** and ***positivity*** so that $1/P(T = t \,|\, \boldsymbol{W}\,)$ exists), the following identification equation for the ATE follows from (29):

$$
\tau = \mathbb{E}\left[Y(1) - Y(0)\right] = \mathbb{E}_{\boldsymbol{W},Y}\left[\frac{\mathbb{1}\left\{T = 1\right\}\, Y}{e(\boldsymbol{W})}\right] - \mathbb{E}_{\boldsymbol{W},Y}\left[\frac{\mathbb{1}\left\{T = 0\right\}\, Y}{1 - e(\boldsymbol{W})}\right]. \tag{30}
$$

  We have the sample IPW estimator for ATE:

$$
\begin{aligned}
\hat{\tau} &= \frac{1}{n} \sum_{i} \left(\frac{\mathbb{1}\left\{T_i = 1\right\}\, y_i}{\hat{e}(\boldsymbol{w}_i)} - \frac{\mathbb{1}\left\{T_i = 0\right\}\, y_i}{1 - \hat{e}(\boldsymbol{w}_i)} \right) \\
&= \frac{1}{n_T} \sum_{i:T_i=1} \frac{y_i}{\hat{e}(\boldsymbol{w}_i)} - \frac{1}{n_C} \sum_{i:T_i=0} \frac{y_i}{1 - \hat{e}(\boldsymbol{w}_i)}, \qquad (31)
\end{aligned}
$$

  where n_T and n_C are number of treated subjects and control subjects, respectively.

- We can also find the IPW estimator of CATE:

$$
\hat{\tau}(\boldsymbol{x}) = \frac{1}{n_{\boldsymbol{x}}} \sum_{i:\boldsymbol{x}_i=\boldsymbol{x}} \left(\frac{\mathbb{1}\left\{T_i = 1\right\}\, y_i}{\hat{e}(\boldsymbol{w}_i)} - \frac{\mathbb{1}\left\{T_i = 0\right\}\, y_i}{1 - \hat{e}(\boldsymbol{w}_i)} \right), \tag{32}
$$

  where $n_{\boldsymbol{x}}$ is the number of data points with $\boldsymbol{x}_i = \boldsymbol{x}$.

  However, with limited sample size, this estimator may suffer high variance issue.

- We want to **trim the tails** on the **propensity score distribution** so that the IPWs would **not** be **too high** and the **positivity assumption** is not violated. But it would change the distribution. We can also **truncate the scores**, which increase bias but decrease variance.

- We may **normalize** the inverse probability weights so that the estimator is **bounded** if Y_i is bounded

$$
\begin{aligned}
\hat{\tau}_1 &= \frac{\sum_i \hat{p}_i \mathbb{1}\left\{T_i = 1\right\}\, y_i,}{\sum_i \hat{p}_i \mathbb{1}\left\{T_i = 1\right\}}, \\
\hat{\tau}_0 &= \frac{\sum_i \hat{q}_i \mathbb{1}\left\{T_i = 0\right\}\, y_i,}{\sum_i \hat{q}_i \mathbb{1}\left\{T_i = 0\right\}},
\end{aligned}
$$

  where $\hat{p}_i = \frac{1}{\hat{e}(\boldsymbol{w}_i)}$ and $\hat{q}_i = \frac{1}{1 - \hat{e}(\boldsymbol{w}_i)}$.

- Note that this IPW is essentially an ***importance sampling*** method.

5 Doubly Robust Methods

- We can estimate causal effects by modeling $\mu(t, \boldsymbol{x}) = \mathbb{E}\left[Y|t, \boldsymbol{x}\right]$, or by modeling $e(\boldsymbol{x}) = P(T = 1|\boldsymbol{x})$. What if we modeled both $\mu(t, \boldsymbol{x})$ and $e(\boldsymbol{x})$?
- A ***doubly robust estimator*** has the property that it is a ***consistent estimator*** of τ if ***either*** the COM $\hat{\mu}$ is a consistent estimator of μ or the propensity score $\hat{e}$ is a consistent estimate of e. See survey in [Seaman and Vansteelandt, 2018]. A related topic is known as ***targeted maximum likelihood estimation (TMLE)*** [Van Der Laan and Rubin, 2006, Van der Laan et al., 2011, Schuler and Rose, 2017].
- An example is the ***Augmented Inverse Probability Weighting (AIPW)*** [Kurz, 2022]:

$$\mathbb{E}\left[Y(1)\right] = \mathbb{E}_{\boldsymbol{X},Y}\left[\frac{\mathbb{1}\left\{T=1\right\}Y}{e(\boldsymbol{X})} - \frac{\mathbb{1}\left\{T=1\right\} - e(\boldsymbol{X})}{e(\boldsymbol{X})}\mu_1(\boldsymbol{X})\right] \tag{33}$$

$$= \mathbb{E}_{\boldsymbol{X},Y}\left[\frac{\mathbb{1}\left\{T=1\right\}(Y - \mu_1(\boldsymbol{X}))}{e(\boldsymbol{X})} + \mu_1(\boldsymbol{X})\right] \tag{34}$$

$$\mathbb{E}\left[Y(0)\right] = \mathbb{E}_{\boldsymbol{X},Y}\left[\frac{(1-\mathbb{1}\left\{T=1\right\})Y}{1-e(\boldsymbol{X})} - \frac{\mathbb{1}\left\{T=1\right\} - e(\boldsymbol{X})}{1-e(\boldsymbol{X})}\mu_0(\boldsymbol{X})\right] \tag{35}$$

$$= \mathbb{E}_{\boldsymbol{X},Y}\left[\frac{\mathbb{1}\left\{T=0\right\}(Y - \mu_0(\boldsymbol{X}))}{1-e(\boldsymbol{X})} + \mu_0(\boldsymbol{X})\right] \tag{36}$$

  where $e(\boldsymbol{X})$ is the propensity score and

$$\mu_1(\boldsymbol{x}) = \mathbb{E}\left[Y(1)|\boldsymbol{X} = \boldsymbol{x}\right], \quad \mu_0(\boldsymbol{x}) = \mathbb{E}\left[Y(0)|\boldsymbol{X} = \boldsymbol{x}\right]$$

  are the ***conditional outcome models (COMs)***.
- The sample estimator for ATE τ is

$$\begin{aligned}\hat{\tau} = &\frac{1}{n}\sum_i\left(\frac{\mathbb{1}\left\{T_i=1\right\}\, y_i}{\hat{e}(\boldsymbol{x}_i)} - \frac{\mathbb{1}\left\{T_i=1\right\} - \hat{e}(\boldsymbol{x}_i)}{\hat{e}(\boldsymbol{x}_i)}\hat{\mu}_1(\boldsymbol{x}_i)\right)\\ &- \frac{1}{n}\sum_i\left(\frac{(1-\mathbb{1}\left\{T_i=1\right\})\, y_i}{1-\hat{e}(\boldsymbol{x}_i)} - \frac{\mathbb{1}\left\{T_i=1\right\} - \hat{e}(\boldsymbol{x}_i)}{1-\hat{e}(\boldsymbol{x}_i)}\hat{\mu}_0(\boldsymbol{x}_i)\right)\end{aligned} \tag{37}$$

- We see that either $\hat{\mu}_1 \to \mu_1$ and $\hat{\mu}_2 \to \mu_2$ **or** $\hat{e} \to e$, we will have $\hat{\tau} \to \tau$.
  - If the COM model estimator is consistent $\hat{\mu}_1 \to \mu_1$ and $\hat{\mu}_2 \to \mu_2$. Then from (34) the first term $\mathbb{1}\left\{T=1\right\}(Y - \hat{\mu}_1) \to 0$; similarly from (36), $\mathbb{1}\left\{T=0\right\}(Y - \hat{\mu}_0) \to 0$. Then we have $\hat{\tau} \to \mathbb{E}\left[\mu_1 - \mu_0\right] = \tau$. Thus $\hat{\tau}$ is consistent;
  - If the propensity score estimator is consistent $\hat{e} \to e$. Then from (33) the second term $\widehat{\mathbb{E}}\left[\widehat{\mathbb{E}}\left[\frac{\mathbb{1}\{T=1\}-\hat{e}(\boldsymbol{X})}{\hat{e}(\boldsymbol{X})}|\boldsymbol{X}\right]\hat{\mu}_1(\boldsymbol{X})\right] \to \mathbb{E}\left[\frac{\mathbb{E}[\mathbb{1}\{T=1\}|\boldsymbol{X}]-e(\boldsymbol{X})}{e(\boldsymbol{X})}\mu_1(\boldsymbol{X})\right] \to 0$ since $\mathbb{E}\left[\mathbb{1}\left\{T=1\right\}|\boldsymbol{X}\right] = e(\boldsymbol{X})$. Similarly, (35) the second term approach to zero. Therefore $\hat{\tau} \to \mathbb{E}\left[\frac{\mathbb{1}\{T=1\}Y}{e(\boldsymbol{X})} - \frac{\mathbb{1}\{T=0\}Y}{1-e(\boldsymbol{X})}\right] = \tau$. Thus $\hat{\tau}$ is consistent;
- AIPW is analyzed using *semiparametric theory.*

References

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Christoph F Kurz. Augmented inverse probability weighting and the double robustness property. *Medical Decision Making*, 42(2):156–167, 2022.

Brady Neal. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Paul Rosenbaum. *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press, 2017.

Megan S Schuler and Sherri Rose. Targeted maximum likelihood estimation for causal inference in observational studies. *American journal of epidemiology*, 185(1):65–73, 2017.

Shaun R Seaman and Stijn Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(2):184, 2018.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 10. Springer, 2011.