Causal Machine Learning – Fall 2023

Week 3: Two-step estimation

Max H. Farrell & Sanjog Misra

Topics to cover

1. Causal identification in observational data
2. (Parametric) two step estimation
3. How the first step estimation impacts second step inference
4. Doubly robust estimation

Observational Data - Binary Treatment

- Recall the selection bias problem: $\mathbb{E}[Y(0) \mid T=1] \neq \mathbb{E}[Y(0) \mid T=0]$
- Randomization made this go away
- Key idea with observational data: X captures why people select
  $\Rightarrow \mathbb{E}[Y(0) \mid T = 1, X = x] = \mathbb{E}[Y(0) \mid X = x]$
- Intuition: need an RCT for each $X = x$
- CIA, unconfoundedness, missing at random, ...
  - Strong version: $Y(1), Y(0) \perp\!\!\!\perp T \mid X$
  - Weak version: $\mathbb{E}[Y(t) \mid T, X] = \mathbb{E}[Y(t) \mid X]$
- Also still need overlap, consistency, SUTVA

Two step estimation

- Our goal is to estimate $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ **and** provide inference
- $Y = \alpha(X) + \beta(X)T + \varepsilon$ is w.l.o.g. (last class)
- $\tau = \mathbb{E}[\beta(X)]$
- In an RCT you recover the average of heterogeneous effects:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}T \quad \longrightarrow \quad \hat{\beta} \to_p \mathbb{E}[\beta(X)]$$

- But in general this fails
  - Need to account for heterogeneity, but we also want to exploit it
  - Need to get the CATE correct
- Two step estimation:
  1. Estimate $\alpha(x)$ and $\beta(x)$
  2. Use these to estimate $\tau = \mathbb{E}[\beta(X)]$ and do inference

Example: Linear models

- Assume a correctly specified linear (or other parametric) model:

$$\mu_t(x) = \mathbb{E}[Y(t) \mid X = x] = x'\beta_t$$

- $\mathrm{CATE} = \beta(x) = \tau(x) = x'\beta_1 - x'\beta_0$
- Run a regression in treatment and control groups separately, then project everywhere (or run a saturated model).
- Then $\hat{\tau} = \widehat{\mathbb{E}[Y(1)]} - \widehat{\mathbb{E}[Y(0)]} = \frac{1}{n}\sum_{i=1}^{n} x_i\hat{\beta}_1 - \frac{1}{n}\sum_{i=1}^{n} x_i\hat{\beta}_0$.

Big questions for today:

- How do we do inference for τ even though we estimate β_t first
- How can we change our approach to make this easier/better?
- Where do influence functions fit in?

Example: Linear models

Identification is **constructive**, motivates estimator:

- Identification: $\mathbb{E}[Y(1)] = \mathbb{E}[\mu_1(x)] = \mathbb{E}[X'\beta_1]$
- Estimation: $\widehat{\mathbb{E}[Y(1)]} = \dfrac{1}{n}\sum_{i=1}^{n} x_i\hat{\beta}_1$

Intuition for the problem

When we estimate $\mathbb{E}[Y(1)]$ there are two sources of uncertainty:

1. Usual frequentist parameter uncertainty: when the data changes the numbers change

   If we knew β_1 or $\hat{\beta}_1$ was fixed, we'd have a standard CLT:

   $$\sqrt{n}\left(\widehat{\mathbb{E}[Y(1)]} - \mathbb{E}[Y(1)]\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{x_i\hat{\beta}_1 - \mathbb{E}[Y(1)]\right\} \to_d \mathcal{N}(0, \Sigma),$$

   Data changes $\to$ x_i changes $\to$ $x_i\hat{\beta}_1$ changes $\to$ $\widehat{\mathbb{E}[Y(1)]}$ changes

2. Model uncertainty – when the data changes the function(al) $\hat{\beta}_1(F_n)$ changes

$$\begin{array}{ccccc} & \nearrow & x_i \text{ changes} & \searrow & \\ \text{data changes} & & & & x_i\hat{\beta}_1 \text{ changes twice} \\ & \searrow & \hat{\beta}_1 \text{ changes} & \nearrow & \end{array}$$

Example: Linear models

- Derive IF for $\widehat{\mathbb{E}[Y(1)]} = \dfrac{1}{n}\sum_{i=1}^{n} x_i \hat{\beta}_1 \ldots$
- Use IF for $\hat{\beta}_1$

Key idea: use the IF for estimation

- $\hat{\beta}_1$ is a function of the data: $\hat{\beta}_1(F_n) \to_p \beta_1(F) = \beta_1$
- $\widehat{\mathbb{E}[Y(1)]} = \frac{1}{n}\sum_{i=1}^{n} x_i \hat{\beta}_1$ is **also** a function of the data, twice in fact

$$\widehat{\mathbb{E}[Y(1)]} = \widehat{\mathbb{E}[Y(1)]}(F_n) = \widehat{\mathbb{E}[Y(1)]}(F_n, \hat{\beta}_1(F_n))$$

$$\widehat{\mathbb{E}[Y(1)]} \to_p \mathbb{E}[Y(1)] = \mathbb{E}_F[X\beta(F)]$$

- Can we find a **different** function of the data that still estimates $\mathbb{E}[Y(1)] = \mathbb{E}_F[X\beta(F)]$, but without this two step estimation problem?
- Yes! We use the influence function

$$\widetilde{\mathbb{E}[Y(1)]} = \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}(z_i) = \frac{1}{n}\sum_{i=1}^{n} \left\{ x_i' \hat{\beta}_1 + \mathbb{E}_n[x_i'] \hat{M}_1^{-1} t_i x_i \hat{\varepsilon}_i \right\}$$

Doubly robust estimation

Similar idea, but from a different angle.

- We already saw two ways to identify $\mathbb{E}[Y(1)]$

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\mathbb{E}\left[Y \mid T = 1, X\right]\right] = \mathbb{E}\left[\frac{TY}{p(X)}\right]$$

- So we can use one or the other estimator:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(x_i) \qquad\qquad \frac{1}{n}\sum_{i=1}^{n} \frac{t_i y_i}{\hat{p}(x_i)}$$

- Each relies on a first step estimator: $\hat{\mu}_1(x) = \widehat{\mathbb{E}}\left[Y \mid T = 1, X = x\right]$ and $\hat{p}(x_i) = \widehat{\mathbb{P}}[T = 1 \mid X = x]$

Doubly robust estimation

Basic idea of doubly robust estimation:

- Two chances to get the right answer
- Cost: do **both** first step estimators
- Benefit: ATE is right if **either** first step is right

$$\widehat{\mathbb{E}[Y(1)]}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n} \frac{t_i\left(y_i - \hat{\mu}_1(x_i)\right)}{\hat{p}(x_i)} + \hat{\mu}_1(x_i)$$

- What if only $\hat{\mu}_1(x_i)$ is right? What if only $\hat{p}(x_i)$ is right?
- What if both are “close”?