Causal Machine Learning – Fall 2023

Week 1: Introduction to Causal Inference

Max H. Farrell & Sanjog Misra

Topics to cover

1. Set up: notation, estimands, etc
2. Difference in means in an RCT
3. Assumptions & Identification
4. Regression
5. Observational data

The Basic Set Up

The Fundamental Problem of Causal Inference

- Never see the same person treated and untreated
- Missing data problem
- Every causal inference method "solves" this by finding a comparison group in one way or another

Potential Outcomes

- $T \in \{0, 1\}$
- $Y(1)$, $Y(0)$

Estimands

- ITE $= Y_i(1) - Y_i(0)$. Impact of the treatment on person i. Not observed, not estimable without very strong assumptions.
- CATE $= \mathbb{E}[Y(1) - Y(0)|X = x] := \tau(x)$. Average treatment effect for individuals with a specific realization of observables, i.e. type, $X = x$.
- ATE $= \mathbb{E}[Y(1) - Y(0)] := \tau$
- ATT $= \mathbb{E}[Y(1) - Y(0) \mid T = 1]$
    - What's the difference between ATE vs. ATT? Which one do we care about?
- Things we won't address in this class: QTE, MTE, LATE, ...

Difference in means

$$\hat{\tau} = \bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - T_i)$$

What does this estimate? AKA What is identified?

- What is identification?
- Law of Large Numbers
  - Assumptions: regularity conditions
- Causal Effect
  - Assumptions have substance: Consistency, SUTVA, Exclusion/Independence

Step 1: Law of Large Numbers

$$\bar{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n} Y_i T_i \to_p \mathbb{P}[T = 1]^{-1} \mathbb{E}[YT] = \mathbb{E}[Y \mid T = 1]$$

- Positivity/Overlap: $\mathbb{P}[T = 1] > 0$
- How do we get this?
- What do we learn?

Step 2: Causal Effect

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

- SUTVA: Only your treatment matters
- Consistency: Observed outcome matches treatment "assignment": $Y = TY(1) + (1 - T)Y(0)$

Only yields $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0 \to_p \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0]$

To get the ATT and ATT we need exclusion/independence

- Randomization

Using Regression

Just run a regression of Y on T?

Start with the estimator:

- $Y = b_0 + b_1 T + e$
- Projection: b_1 (always) estimates cov$(T, Y)/\mathbb{V}[T]$
- $b_1 \equiv \bar{Y}_1 - \bar{Y}_0 = \hat{\tau}$
- Need assumptions beyond typical OLS

Start with the assumptions

- *Define* μ_t and ε_t via $Y(t) = \mu_t + \varepsilon_t$
- $Y = TY(1) + (1 - T)Y(0) = \alpha + \beta T + \varepsilon$
- OLS assumptions: rank $\leftrightarrow$ overlap, orthogonality $\leftrightarrow$ randomization
- By construction $\beta \equiv \tau$

Using Regression With Covariates

Just run a regression of Y on T **and** X?

Start with the estimator:

- Data set includes a vector of covariates X_i
- $Y = b_0 + b_1 T + b_2 X + e$
- Why bother?
- $b_1 \rightarrow_p \tau$
- But with more assumptions this time
- Pre-treatment variables and bad controls

Start with the assumptions

- $Y(t) = \mu_t(X) + \varepsilon_t$
- $\tau(x) = \mu_1(x) - \mu_0(x)$
- $Y = TY(1) + (1-T)Y(0) = \alpha + \beta T + \gamma X + \varepsilon$?
- $b_1 \rightarrow_p \beta$?
- $\beta = \tau = \mathbb{E}[\tau(x)]$?

Using Regression With Covariates

Just run a regression of Y on T **and** X?

Include de-meaned interaction term:

$$Y = b_0 + b_1 T + b_2 X + b_3 T(X - \bar{X}) + e$$

- Best practice for getting τ from b_1
- But also yields **heterogeneity**

Heterogeneous Effects

- Useful thought experiment: $Y_i = \alpha_i + \beta_i T + \varepsilon_i$
- CATE
- Targeting/personalization

Beyond Binary Treatment

Model:

$$Y = \alpha(X) + \beta(X)T + \varepsilon$$

- wlog for binary T (w or w/o heterogeneity)
- Multivalued T the same: $T = (\mathbb{1}\{T=0\}, \mathbb{1}\{T=1\}, \ldots, \mathbb{1}\{T=J\})'$
- But **structural** for other T

Potential Outcomes

- Fully general: $Y(t) = f(t, X, \varepsilon)$,
  ATE becomes ASF: $\mathbb{E}[Y(t)] = \int_\varepsilon \int_X f(t, x, e) dF_{X\varepsilon}(x, e)$
- $Y(t, x) = \alpha(x) + \beta(x)t + \varepsilon(t)$
- Average Partial Effect: $\mathbb{E}[Y(t+1, X) - Y(t, X)] = \mathbb{E}[\beta(X)]$

Observational Data

Binary Treatment

- Need an RCT for each $X = x$
- $Y(1), Y(0) \perp\!\!\!\perp T \mid X$
- CIA, unconfoundedness, missing at random, $\ldots +$ Overlap
- $\mathbb{E}[Y(0) \mid T = 1, X = x] = \mathbb{E}[Y(0) \mid T = 0, X = x]$?

Continuous Treatment

- How does $Y = \alpha + \beta T + \varepsilon$ have a causal interpretation?
- Exclusion/independence