Causal Machine Learning – Autumn Quarter 2024–2025

Slides Set #1: Causal Inference Basics

Max H. Farrell

(version 1.1, compiled October 3, 2024)

Class Goals/Overview

Overall goal

1. Be able to intelligently use some ML in your research/job
2. Be able to read a paper and understand what's going on

Subjects to cover

1. Identification & Causality
2. Nonparametric/Flexible Models
   - Including high dimensional models (lasso), ML (deep nets, forests)
3. Semiparametric Inference
   - Including influence functions, double machine learning
4. More if we have time

Work to be done

1. Learn some theory
2. Do some coding
3. Homeworks & final project
4. Readings?

Some Motivation

Important Questions/Issues to Keep in Mind

- What types of outcomes do social scientists care about?
- What types of treatments?
- What objects do they wish to estimate?
- Identification & assumptions
- Do they take the model seriously?
- *Why* do they wish to estimate these?
- What are we doing?
  - Hypothesis testing
  - Policy description
  - Counterfactual policy evaluation
  - Policy design

Starting with Some Basics

Binary Treatment

1. One of our workhorse models
   - Often work with simple cases/models
   - Everything we learn carries over
2. Difference in means
3. Identification & assumptions
4. Regression with and without covariates
   - Linear model (and variants) is our other workhorse
5. Heterogeneity will raise issues and questions

The Basic Set Up

The Fundamental Problem of Causal Inference

- Never see the same person treated and untreated
- Missing data problem
- **Every single** causal inference methods "solve" this problem by assuming a valid comparison group (one way or another)

Random Variables

- Treatment: $T \in \{0, 1\}$
- Outcomes: $Y(1)$, $Y(0)$
  - Already made assumptions. What?
- Covariates: X

Estimands

- ITE $= Y_i(1) - Y_i(0)$. Impact of the treatment on person i. Not identified, not consistently estimable (i.e. DID vs Synthetic Controls)
- CATE $= \mathbb{E}[Y(1) - Y(0)|X = x] := \tau(x)$. Average treatment effect for individuals with a specific realization of observables, i.e. people of "type", $X = x$
- ATE $= \mathbb{E}[Y(1) - Y(0)] := \tau = \mathbb{E}[\tau(X)]$
- ATT $= \mathbb{E}[Y(1) - Y(0) \mid T = 1]$
- Things we won't address in this class: SATE, QTE, MTE, LATE, ...
  - Not unimportant, just not our main focus

Key questions

- Which one do we care about?
- What's identified? What is estimable and how fast?
- What's the difference between ATE vs. ATT?

Start Simple: Difference in Means

Goal is to estimate ATE $= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] := \tau$

Plug-in / Sample Analog $\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1} \sum_{\text{treated}} Y_i - \frac{1}{n_0} \sum_{\text{control}} Y_i$

What does this estimate? a.k.a. What is identified?

- Define identification?
- Law of Large Numbers
  - Boring assumptions: regularity conditions
- Causal Effect
  - Real Assumptions: Consistency, SUTVA, Exclusion/Independence

Step 1: Law of Large Numbers

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

Regularity conditions

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

Step 2: Causal Effect

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

Now real assumptions:

- Positivity/Overlap: $\mathbb{P}[T = 1] > 0$
- SUTVA: Only your treatment matters
- Consistency: Observed outcome matches treatment "assignment": $Y = TY(1) + (1 - T)Y(0)$
- These get us to $\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0 \to_p \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0]$

To get the ATT and ATT we need exclusion/independence

- Randomization implies $T \perp\!\!\!\perp Y(1), Y(0)$
- $\mathbb{E}[Y(1) \mid T = 1] = \mathbb{E}[Y(1) \mid T = 0] = \mathbb{E}[Y(1)]$

Using Regression

Just run a regression of Y on T?

Start with the estimator:

- $Y = b_0 + b_1 T + e$
- b_1 is exactly $\bar{Y}_1 - \bar{Y}_0 = \widehat{\tau}$

More interesting to 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?

- Very common practice
- Two main goals: improve precision & test/estimate heterogeneity
- Results:
  - OLS consistently estimates the ATE & improves precision
  - Even if misspecified
- Pre-treatment X is crucial. Watch for bad controls.

Model

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

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 OLS
- Still consistent, always more efficient
- But also yields **heterogeneity** in limited form

Bad controls

- Pre-treatment covariates crucial

Bad Controls Example

- $T = \{0, 1\}$ is randomized
- Outcome is $Y = TY(1) + (1 - T)Y(0)$ (SUTVA, consistency, etc)
- Covariate is binary $X = \{0, 1\}$
- X is not pre-treatment, so just like Y we have $X = TX(1) + (1 - T)X(0)$.
- E.g., $X(1) = 0$ means that the value X would take in the treatment group is zero.

Estimation: include de-meaned interaction term:

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

- Fully saturated, so recovers differences in means per X group
- Does the difference in means for $X = 1$ recover the ATE for $X = 1$?

Bad Controls Example

With the same logic as before (slide 8) randomized treatment means the difference in means actually estimates the following:

$$\begin{aligned}\bar{Y}_1, 1 - \bar{Y}_{0,1} \to_p \mathbb{E}[Y \mid X=1, T=1] - \mathbb{E}[Y \mid X=1, T=0] \\ = \mathbb{E}[Y(1) \mid X(1)=1, T=1] - \mathbb{E}[Y(0) \mid X(0)=1, T=0] \\ = \mathbb{E}[Y(1) \mid X(1)=1] - \mathbb{E}[Y(0) \mid X(0)=1]\end{aligned}$$

- The first line is the LLN, the second is the definition of potential outcomes and potential covariates, and the third is because treatment is randomized, so conditioning on it does not matter for the expectation of $Y(t)$.
- The $T \leftrightarrow Y$ selection bias is gone by randomization, but there is now an $X \leftrightarrow Y$ selection problem. X comes after T, so X being 1 under treatment $(X(1)=1)$ occurs for different reasons than under control $(X(0)=1)$, and this could be causing the difference in Y.
- Add/subtract to see the selection bias:

$$\begin{aligned}\mathbb{E}[Y(1) \mid X(1)=1] - \mathbb{E}[Y(0) \mid X(0)=1] = \quad & \mathbb{E}[Y(1) \mid X(1)=1] - \mathbb{E}[Y(0) \mid X(0)=1] \\ & \underbrace{-\mathbb{E}[Y(0) \mid X(1)=1]}_{\text{ATE for } X(1)=1} \underbrace{+\mathbb{E}[Y(0) \mid X(1)=1]}_{\text{Selection into } X=1}\end{aligned}$$

Wrapping up Binary T RCT

The paradigmatic assertion in causal relationships is that manipulation of a cause will result in the manipulation of an effect. ... Causation implies that by varying one factor I can make another vary.

(Cook & Campbell 1979: 36, emphasis in original)

Does being a woman **cause** lower pay?

$$\text{wages} = \alpha + \beta \text{ gender} + \gamma' X + \epsilon$$

- What is the interpretation of β?
- What would you include in X? Does this change the interpretation of β?
- How should we think of (e.g.) Bertrand & Mullainathan (AER 2004)?
- Do we like this model?

Observational Data

When is a regression causal?

1. Overlap
2. Selection on observables
   (*aka* Unconfoundedness / Conditional independence / Missing at random)
3. Correct specification

Selection on observables

- Recall selection bias: $\mathbb{E}[Y(0) \mid T = 1] \neq \mathbb{E}[Y(0) \mid T = 0]$
- People choose their own treatment based on expected benefits
- The *observed* covariates have to capture the entire *selection* problem
- T randomized $\Rightarrow$ T randomized for every $X = x$
- Need an RCT for every $X = x$
- $Y(1), Y(0) \perp\!\!\!\perp T \mid X$ or $\mathbb{E}[Y(t) \mid T = s, X = x] = \mathbb{E}[Y(t) \mid X = x]$

Correct specification

- Work backwards to see what the assumptions have to be:
- Estimator is OLS: $\hat{Y}_i = \widehat{\alpha} + \widehat{\tau} T_i + \widehat{\gamma}_1' X_i + \widehat{\gamma}_2' (X_i - \bar{X}) T_i$
- So it must be that $Y = \alpha + \tau T + \gamma_1' X + \gamma_2 (X - \mathbb{E}[X]) T + \varepsilon$
- So it must be that $Y(t) = \mu(t, X) + \varepsilon_t$, with $\mu(t, X) = \alpha + \gamma_1' X + T(\tau + \gamma_2' (X - \mathbb{E}[X]))$
- Then $\widehat{\tau} \to_p \text{ATE}$.
- We even get the CATEs: $\tau(x) = \tau + \gamma_2' x$.

But we need the believe the model!

- The model is actually a model for the CATE, i.e. the heterogeneity
- The problem is that if $\tau(x) \neq \tau$, then $\widehat{\tau} \not\to_p \tau = \mathbb{E}[\tau(X)]$

Heterogeneous Effects

Start General: $Y_i = \alpha_i + \beta_i T + \varepsilon_i$

- What is ε_i?
- What is β_i?
- What is CIA here?

The standard model: $Y_i = \alpha(X_i) + \beta(X_i)T + \varepsilon_i$

- Observed and Unobserved heterogeneity
- What even is unobserved heterogeneity?
- Fixed and random effects
- Random coefficients
- Targeting/personalization

For Later: 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: $\mathbb{E}[Y(t+1, X) - Y(t, X)] = \mathbb{E}[\beta(X)]$

Identification Without Correct Specification

Just need to identify the CATE

- If we have $\tau(x)$, then we can average to get $\tau = \mathbb{E}[\tau(X)]$
- If we have $\mathbb{E}[Y(1) \mid X]$, we can average to get $\mathbb{E}[Y(1)]$

Two strategies

- Imputation: $\mathbb{E}[Y \mid T = 1, X = x]$
- Inverse weighting: $\mathbb{E}[YT \mid X = x] / \mathbb{E}[T \mid X = x]$

Estimation

- How can we take our model to data?
- Something like this?

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta}(X_i), \qquad \text{with} \qquad \widehat{Y}_i = \widehat{\alpha}(X_i) + \widehat{\beta}(X_i)T_i$$

- It'll take us most of the quarter to do this completely:
  - ML/nonparametrics for flexible estimators for $\alpha(x)$, $\beta(x)$
  - Semiparametrics, and "double machine learning", for inference on τ

Welfare

- Consider **assigning** people to treatment:

$$d(X) : \mathcal{X} \to \mathcal{T} = \{0, 1\}$$

- What is the right $d(X)$?
- Now we don't care about ATE, we care about **welfare**

$$\mathbb{E}[d(X)Y(1) + (1 - d(X))Y(0)] \overset{\text{Why?}}{=} \mathbb{E}[Y(0)] + \mathbb{E}[d(X)\tau(X)]$$

- $\Rightarrow$ Ideally $d(X) = \mathbb{1}\{\tau(x) > 0\}$.
- Slightly more general, social welfare could have margins and costs:

$$\pi(T_i) = \begin{cases} mY_i(0) & \text{if } T_i = 0 \\ mY_i(1) - c & \text{if } T_i = 1 \end{cases}$$

- Welfare is now $\mathbb{E}[\pi(0)] + \mathbb{E}[d(X)(\pi(1) - \pi(0))]$
- $\Rightarrow$ Ideally $d(X) = \mathbb{1}\{m\tau(x) > c\}$.