# Causal Machine Learning

Week 3

Booth School Of Business
University of Chicago

## Applications

# Causality

> *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)

# Causal Constructs

- What types of outcomes do economists care about?
- What types of treatments?
- What objects do they wish to estimate?
- *Why* do they wish to estiamte these?

# Why estimate causal effects?

- **Hypotesis testing**
- **Policy descriptors**
- **Counterfactual Policy evaluation**
- **Policy design**

# Models and Causality

- **Q: Do we need models to get causal effects?**
- **Q: What exactly is a model?**
- **Q: Do we need a structural model? What's that?**

# Parameteric Models

- Consider the standard difference in means estimator
- One can equivalently write this as

$$Y_i^{\text{obs}} \mid \mathbf{W}, \tilde{\theta} \sim \mathcal{N}\left(\mu_{\text{c}} + W_i \cdot \tau, \sigma^2\right)$$

- In this case (as in the difference in means) the ATE is simple $\tau$

# Parameteric Models: Extended Example

- Now consider the following:

$$\begin{pmatrix} \ln\left(Y_i(0)\right) \\ \ln\left(Y_i(1)\right) \end{pmatrix} \mid \theta \sim \mathcal{N}\left( \begin{pmatrix} \mu_{\mathrm{c}} \\ \mu_{\mathrm{t}} \end{pmatrix}, \begin{pmatrix} \sigma_{\mathrm{c}}^2 & 0 \\ 0 & \sigma_{\mathrm{t}}^2 \end{pmatrix} \right)$$

- What is the ATE here?

# Parameteric Models: Extended Example

- Now consider the following:

$$\begin{pmatrix} \ln\left(Y_i(0)\right) \\ \ln\left(Y_i(1)\right) \end{pmatrix} \mid \theta \sim \mathcal{N}\left( \begin{pmatrix} \mu_{\mathrm{c}} \\ \mu_{\mathrm{t}} \end{pmatrix}, \begin{pmatrix} \sigma_{\mathrm{c}}^2 & 0 \\ 0 & \sigma_{\mathrm{t}}^2 \end{pmatrix} \right)$$

- The ATE is:

$$\tau = \tau(\theta) = \exp\left(\mu_{\mathrm{t}} + \frac{1}{2} \cdot \sigma_{\mathrm{t}}^2\right) - \exp\left(\mu_{\mathrm{c}} + \frac{1}{2} \cdot \sigma_{\mathrm{c}}^2\right)$$

# Parameteric Models: Inference

- Define $\theta = \{\mu_c, \mu_t, \sigma_c, \sigma_t\}$
- Set up likelihood $\ell(\mathbb{D}|\theta)$ and obtain $\widehat{\theta}$
- Compute the Hessian at $\widehat{\theta}$ : $\widehat{\mathbf{H}} = \left\{ \frac{\partial^2 \ell}{\partial \theta_{jk}} \right\}$
- Use Delta method

$$\text{se}\left(\tau\left(\widehat{\theta}\right)\right) = \left[\frac{\partial \tau(\theta)}{\partial \theta}\Big|_{\theta=\widehat{\theta}}\right]' \widehat{\mathbf{H}}^{-1} \left[\frac{\partial \tau(\theta)}{\partial \theta}\Big|_{\theta=\widehat{\theta}}\right]$$

# Discussion Problem I

- Imagine a firm that sends customers catalogs
- They randomize the treatment so that 90% get the catalog and 10% are held out
- A firm wishes to figure out the causal effect of the catalog ($x$) on buying behavior ($y$)
- What should the model be?
- What decisions do we make?

# Discussion Problem II

- Imagine a firm that charges a price for a product. Consumer buys a single unit or not at all.
- A firm wishes to set optimal prices. To do so they need to figure out the causal effect of prices ($x$) on purchase decision ($y$)
- **What should the model be?**
- **How should the firm set prices?**

# Discussion Problem III

- Detailing refers to the act of pharma reps calling on physicians
- A firm wishes to figure out the causal effect of detailing ($x$) on prescribing behavior ($y$)
- Currently the avearge number of calls is 10.
- **What should the model be?**
- **Should the firm increase the number of calls?**

# Discussion Problem I

- Imagine a firm that sends customers catalogs
- They randomize the treatment so that 90% get the catalog and 10% are held out
- A firm wishes to figure out the causal effect of the catalog ($x$) on buying behavior ($y$)
- What should the model be?
- What decisions do we make?

# Conditional average treatment effect (CATE)

- Main object of interest:

$$\tau(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$$

- Causal effect for subpopulation of customers with identical features $X_i = x$
- In our application: Incremental effect on spending when being targeted

# Profits

- Profit contribution depending on targeting status

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

  - $m$ is the margin percentage
  - $c$ is the targeting cost
  - Easily generalizable for heterogeneous margins and costs

# Targeting policy evaluation

- Targeting policy $d : \mathbb{X} \to \{0, 1\}$
- Goal: Evaluate the expected profit from any targeting policy, $d$

$$\begin{aligned}\mathbb{E}[\Pi(d, (X_i))] &= \sum_{i=1}^{N} \mathbb{E}\left[1\{d(X_i) = 0\} \cdot \pi_i(0) + 1\{d(X_i) = 1\} \cdot \pi_i(1) | X_i\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[(1 - d(X_i)) \cdot \pi_i(0) + d(X_i) \cdot \pi_i(1) | X_i\right]\end{aligned}$$

  - Conditional on observed customer features, $(X_i) = (X_i)_{i=1}^{N}$

# Optimal targeting policy

- $d^*$ is an optimal policy if it maximizes $\mathbb{E}[\Pi(d, (X_i))]$
- Assume: $W_i$ does not affect the behavior of any other customer $i' \neq i$ (SUTVA)
- Then $d^*$ is optimal if and only if it maximizes the expected profit from each individual customer with features $X_i$,

$$\mathbb{E}\left[(1 - d(X_i)) \cdot \pi_i(0) + d(X_i) \cdot \pi_i(1) | X_i\right]$$

- Optimal policy, $d^*$: Target a customer if and only if

$$\mathbb{E}\left[\pi_i(1)|X_i\right] > \mathbb{E}\left[\pi_i(0)|X_i\right]$$

- Equivalently:

$$\begin{aligned}\mathbb{E}\left[\pi_i(1) - \pi_i(0)|X_i\right] > 0 &\Leftrightarrow \mathbb{E}\left[(mY_i(1) - c) - (mY_i(0))|X_i\right] > 0 \\ &\Leftrightarrow m\mathbb{E}\left[Y_i(1) - Y_i(0)|X_i\right] - c > 0 \\ &\Leftrightarrow m\tau(x) > c\end{aligned}$$

- Lessons
  - An optimal targeting policy is based on the incremental effect of targeting
  - Predict optimal policy based on estimate of the CATE

# Data

- Observed outcome:

$$Y_i = \begin{cases} Y_i(0) & \text{if } W_i = 0 \\ Y_i(1) & \text{if } W_i = 1 \end{cases}$$

  - Only one of the potential outcomes is observed

- Data $\mathcal{D} = (Y_i, X_i, W_i)_{i=1}^N$

# The fundamental problem of causal inference

- Because only one of the potential outcomes is observed, the individual treatment effect $Y_i(1) - Y_i(0)$ is not observed
- Using the observed outcomes $Y_i$ only the CATE is not generally identified,

$$\tau(x) \neq \mathbb{E}[Y_i|X_i, W_i = 1] - \mathbb{E}[Y_i|X_i, W_i = 0] \quad \text{in general}$$

  - Example: Selection on unobservables in online advertising ([**Gordon-Zettelmeyer-Bhargava-etal-2017**])
- An estimate of $\tau(x)$ cannot even be calculated if targeting policy in data either targets all or no customers with features $X_i = x$

# Identification of conditional average treatment effect

- Assumptions
  1. *Unconfoundedness*
  $$Y_i(0), Y_i(1) W_i \mid X_i$$
  2. *Overlap*: The propensity score $e(x) \equiv \Pr\{W_i = 1 | X_i = x\}$ satisfies
  $$0 < e(x) < 1$$
  3. Stable unit treatment value assumption (*SUTVA*)
     - Rules out social or equilibrium effects

- Under these assumptions $\tau(x)$ is identified
$$\tau(x) = \mathbb{E}[Y_i | X_i, W_i = 1] - \mathbb{E}[Y_i | X_i, W_i = 0]$$

- In our application
  - Unconfoundedness and overlap can be satisfied by (experimental) design
  - SUTVA likely satisfied

# Linear model

- Assumption: Conditional expectation function well approximated using a linear function of the features and interactions of the features with the treatment

$$\mathbb{E}[Y_i|X_i = x, W_i = w] = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \delta_0 w + \sum_{k=1}^{p} \delta_k x_{ik} w_i$$

- It follows that the CATE is a linear function of the features

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y_i|X_i = x, 1] - \mathbb{E}[Y_i|X_i = x, 0] \\ &= \delta_0 + \sum_{k=1}^{p} \delta_k x_{ik}\end{aligned}$$

- Estimation using OLS

- Equivalent to estimating two separate linear models for the treated and untreated units

# Aside: Transformed outcome

- Defined as

$$\begin{aligned} Y_i^* &= W_i \cdot \frac{Y_i(1)}{e(X_i)} - (1 - W_i) \cdot \frac{Y_i(0)}{1 - e(X_i)} \\ &= \frac{W_i - e(X_i)}{e(X_i)(1 - e(X_i))} Y_i \end{aligned}$$

  - $Y_i^*$ is observed if the propensity score is known

- If unconfoundedness holds, then

$$\mathbb{E}[Y_i^* | X_i = x] = \tau(x)$$

- Hence $Y_i^*$ is a proxy for the CATE:

$$Y_i^* = \tau(X_i) + \nu_i$$

  - $\mathbb{E}[\nu_i | X_i] = 0$ and $\nu_i$ is orthogonal to any function of $X_i$

# Targeting policy evaluation

- Back to the ultimate goal: Evaluate the expected profit from a targeting policy, $d$

$$\mathbb{E}[\Pi(d,(X_i))] = \sum_{i=1}^{N} \mathbb{E}\left[1\{d(X_i)=0\}\cdot\pi_i(0) + 1\{d(X_i)=1\}\cdot\pi_i(1)|X_i\right]$$

- We will mostly compare optimal targeting policies predicted based on the CATE estimates,

$$d(x) = \begin{cases} 0 & \text{if } m\hat{\tau}(x) - c \leq 0 \\ 1 & \text{if } m\hat{\tau}(x) - c > 0 \end{cases}$$

- Comparison straightforward field experiments that implement each policy

# Targeting policy evaluation in a randomized sample

- Problem:
  - The intended targeting status is not generally equal to the randomized treatment assignment
  - However, for some customers $i$, the intended and realized treatment assignment will agree,

$$d(X_i) = W_i$$

- Proposed profit estimator:

$$\hat{\Pi}(d, (X_i)) = \sum_{i=1}^{N} \left( \frac{1 - W_i}{1 - e(X_i)} (1 - d(X_i)) \cdot \pi_i(0) + \frac{W_i}{e(X_i)} d(X_i) \cdot \pi_i(1) \right)$$

# Inverse probability-weighted targeting profit estimator

$$\hat{\Pi}(d,(X_i)) = \sum_{i=1}^{N} \left( \frac{1-W_i}{1-e(X_i)}(1-d(X_i)) \cdot \pi_i(0) + \frac{W_i}{e(X_i)} d(X_i) \cdot \pi_i(1) \right)$$

- The estimator sums all profits when the intended and realized treatment assignment agree
- Why the inverse probability weights? — Consider all customers who should be targeted, $d(X_i) = 1$. If the targeting probability in the randomized sample is $e(X_i) = \frac{2}{3}$, there is a $\frac{2}{3}$ chance that these customers were actually targeted. However, $\frac{1}{3}$ of all observations are "missing," i.e. were not actually targeted. Weighting by the inverse of $\frac{2}{3}$ restores the correct scale of profits.

$\hat{\Pi}(d,(X_i))$ is an unbiased estimator for the expected profit $\mathbb{E}[\Pi(d,(X_i))]$:

$$\begin{aligned}
\mathbb{E}[\hat{\Pi}(d,(X_i))] &= \sum_{i=1}^{N} \mathbb{E}\left[\frac{1-W_i}{1-e(X_i)}(1-d(X_i))\cdot\pi_i(0) + \frac{W_i}{e(X_i)}d(X_i)\cdot\pi_i(1)|X_i\right] \\
&= \sum_{i=1}^{N}\left(\frac{1-e(X_i)}{1-e(X_i)}(1-d(X_i))\cdot\mathbb{E}[\pi_i(0)|X_i] + \frac{e(X_i)}{e(X_i)}d(X_i)\cdot\mathbb{E}[\pi_i(1)|X_i]\right) \\
&= \sum_{i=1}^{N}\left((1-d(X_i))\cdot\mathbb{E}[\pi_i(0)|X_i] + d(X_i)\cdot\mathbb{E}[\pi_i(1)|X_i]\right) \\
&= \sum_{i=1}^{N}\mathbb{E}\left[(1-d(X_i))\cdot\pi_i(0) + d(X_i)\cdot\pi_i(1)|X_i\right] \\
&= \mathbb{E}[\Pi(d,(X_i))]
\end{aligned}$$

# Comparison to "classical" CRM approach

- Classical CRM is based on a scoring method, where customers are scored according to their predicted spending or profit *level*
- Firm that we collaborate with predicts $\mathbb{E}[Y_i|X_i = x, W_i = 1]$ using a logit/OLS model
- Naive targeting approach: Target customer if

$$m \cdot \mathbb{E}[Y_i|X_i = x, W_i = 1] > c$$

# Empirical application

- Data stems from collaboration with a U.S. company that sells durable consumer products
- Multi-channel retailer
  - Direct channel: Phone, mail, website
  - Brick-and-mortar stores in some U.S. regions
- Sophisticated data science team, plans targeting efforts including
  - Mail/catalog
  - Display ads, Facebook
  - E-mails
- Evaluation of targeting efforts using randomized samples

# Data

- Catalogs mailed in spring of 2015 and 2016
  - Instances of the same campaign that is repeated annually at the same time within the calendar year
- Two randomized samples
  - 2015: 293 thousand observations
  - 2016: 148 thousand observations
  - Targeting probability is $\frac{2}{3}$

# Observed spending

- Phone/mail orders and website transactions using product codes in catalog during three-month period after catalog was received
- Web purchases not using the product codes in thirty-day period after catalog was received
- According to company most product code transactions occur in the month after receipt of the catalog
  - $\Rightarrow$ spending measure allow us to estimate *short-run* causal effect of a targeting effort
  - May under or overestimate long-run effect

# Customer attributes — features

- 472 features
  - Information on customer a few weeks before catalog mailing
- Demographics
- RFM (recency, frequency, and monetary value) variables
  - For different time periods, e.g. last six months, last 7-12 months, etc.
  - For different product types
  - For different sales channels
  - Most granular level: Transactions at time period/product type/channel level
    - E.g. number of purchases placed online during the last twelve months for products in a specific category
- Website browsing behavior
  - Page views, clicks on product pages
- Promotional e-mails
  - Number e-mails received, viewed, and resulting click-throughs

# Summary statistics

| | Mean | SD | N | Percentiles | | | | |
|---|---|---|---|---|---|---|---|---|
| **2015** | | | | 1% | 5% | 50% | 95% | 99% |
| Spending | 7.311 | 43.549 | 292657 | 0.000 | 0.000 | 0.000 | 39.950 | 182.990 |
| Spending if purchased | 117.725 | 132.445 | 18174 | 17.950 | 27.950 | 79.900 | 322.753 | 605.727 |
| Purchase | 0.062 | 0.241 | 292657 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| Treatment | 0.669 | 0.471 | 292657 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 |
| **2016** | | | | | | | | |
| Spending | 6.461 | 39.565 | 148200 | 0.000 | 0.000 | 0.000 | 36.981 | 168.000 |
| Spending if purchased | 115.066 | 124.021 | 8322 | 19.950 | 27.950 | 79.900 | 309.748 | 585.282 |
| Purchase | 0.056 | 0.230 | 148200 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| Treatment | 0.663 | 0.473 | 148200 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 |

Average treatment effect and differences among treated and untreated customers

| | Mean | SE |
|---|---|---|
| **2015** | | |
| ATE spending | 2.561 | 0.166 |
| ATE purchase | 0.022 | 0.001 |
| $E[Y_i \mid Y_i > 0, W_i = 1] - E[Y_i \mid Y_i > 0, W_i = 0]$ | -0.798 | 2.456 |
| **2016** | | |
| ATE spending | 2.377 | 0.210 |
| ATE purchase | 0.022 | 0.001 |
| $E[Y_i \mid Y_i > 0, W_i = 1] - E[Y_i \mid Y_i > 0, W_i = 0]$ | -2.170 | 3.383 |

- Based on margin and cost data customer should be targeted if ATE $> 2.003$
- Hence, if the treatment effects were homogenous, company should use a blanket targeting strategy