Heterogeneous Treatment Effects

Kosuke Imai

Harvard University

Spring 2021
Heterogeneous Treatment Effects

- Same treatment may affect different individuals differently
- Conditional Average Treatment Effect (CATE)
  \[ \tau(x) = \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x) \text{ where } x \in \mathcal{X} \]
  - who benefits from and is harmed by the treatment?

- Individualized treatment rule (ITR)
  \[ f : \mathcal{X} \longrightarrow \{0, 1\} \]
  - We can never identify an individual causal effect
  \[ \tau_i = Y_i(1) - Y_i(0) \]
  - ITR depends on the choice of \(X_i\)

- Use of machine learning methods
Subgroup Analysis and Pre-registration

- If we have a hypothesis about the some group-specific effects:
  - stratify the data and estimate the ATE within each strata
  - compare the ATE between groups

- Problem: multiple testing, data snooping, “p-hacking”, “fishing”
- Solution: Pre-register hypotheses and analyses
  - standard in medicine, has become a norm in social sciences
  - repositories
    - Evidence in Governance and Politics (EGAP)
    - American Economic Association (AEA)
    - Registry for International Development Impact Evaluations (RIDIE)

- Pre-registration solves commitment and transparency problems

- It does not solve the statistical problem of multiple testing
  - FWER (family-wise error rate): probability of making any type I error
  - FDR (false discovery rate): expected proportion of type I error among all rejections
Motivation:
1. avoid strong modeling assumptions $\leadsto$ data-driven approach
2. avoid false discoveries $\leadsto$ avoid over-fitting via regularization

Difference between prediction and causality
- prediction $\leadsto$ use $\mathbf{X}_i$ to predict $Y_i$
- causality $\leadsto$ use $\mathbf{X}_i$ to predict $\tau_i = Y_i(1) - Y_i(0)$

Mean squared error decomposition:
\[
\mathbb{E}[(\tau_i - \hat{\tau}(\mathbf{x}))^2 \mid \mathbf{X}_i = \mathbf{x}] = \mathbb{E}[(\tau_i - \tau(\mathbf{x}))^2 \mid \mathbf{X}_i = \mathbf{x}] + \mathbb{E}[(\tau(\mathbf{x}) - \hat{\tau}(\mathbf{x}))^2 \mid \mathbf{X}_i = \mathbf{x}]
\]

Inference of heterogeneous treatment effects depends on
1. How predictive $\mathbf{X}_i$ is of $\tau_i$
2. How good your model is for estimating $\tau(\mathbf{x})$
Estimation of the CATE (Künzel et al. 2018. PNAS)

- **S-learner**
  1. estimate $\mu_t(x) = \mathbb{E}(Y_i \mid T_i = t, X_i = x)$ using a single model
  2. compute $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$

  $\Rightarrow$ modeling interactions between $T_i$ and $X_i$ can be challenging

- **T-learner**
  1. estimate $\mu_t(x) = \mathbb{E}(Y_i \mid T_i = t, X_i)$ separately for each $t$
  2. compute $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$

  $\Rightarrow$ difficult if the treatment assignment is lopsided, $\hat{\tau}$ may not be smooth

- **X-learner**
  1. estimate $\mu_t(x) = \mathbb{E}(Y_i \mid T_i = t, X_i)$ separately for each $t$
  2. impute missing potential outcomes as $\hat{\mu}_{1-T_i}(X_i)$ and compute $\hat{\tau}_i$
  3. model estimated individual treatment effects $\hat{\tau}_i$ using $X_i$

  $\Rightarrow$ more robust but less efficient
Penalized Maximum Likelihood Estimator

- **PMLE:**
  \[
  \hat{\theta} = \arg\max \log \mathcal{L}(\theta; Y, X) + P(\lambda, \theta)
  \]
  - **Ridge:** \( P(\lambda, \theta) = \lambda \sum_{j=1}^{p} \beta_j^2 \)
  - **Lasso:** \( P(\lambda, \theta) = \lambda \sum_{j=1}^{p} |\beta_j| \)

- **Sample splitting:**
  1. training data: estimate \( \theta \) given \( \lambda \)
  2. test data: choose \( \hat{\lambda} \)
  3. validation data: estimate CATE given \( \hat{\lambda} \)

  - Lasso with support vector machine
  - separate tuning parameters \( \lambda \) for main terms and interactions \( \rightsquigarrow \)
    two-dimensional grid search

- **T-learner** (Qian and Murphy. 2011. *Ann. Stat.*)
  - Lasso with least squares
  - separately fitted for the treatment and control groups
  - uses S-learner when the treatment has more than 2 categories
44 covariates including some square and interaction terms
44 interactions between the treatment and covariates
sparcity of the model helps with interpretation

<table>
<thead>
<tr>
<th>Groups most helped or hurt by the treatment</th>
<th>Average effect</th>
<th>Age</th>
<th>Educ.</th>
<th>Race</th>
<th>Married</th>
<th>Highschool degree</th>
<th>Earnings (1975)</th>
<th>Unemp. (1975)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low education, Non-Hispanic</td>
<td>53</td>
<td>31</td>
<td>4</td>
<td>White</td>
<td>No</td>
<td>No</td>
<td>10,700</td>
<td>No</td>
</tr>
<tr>
<td>High Earning</td>
<td>50</td>
<td>31</td>
<td>4</td>
<td>Black</td>
<td>No</td>
<td>No</td>
<td>4020</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>28</td>
<td>15</td>
<td>Black</td>
<td>No</td>
<td>Yes</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>Unemployed, Black, Some College</td>
<td>38</td>
<td>30</td>
<td>14</td>
<td>Black</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>32</td>
<td>12</td>
<td>Hisp</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>33</td>
<td>9</td>
<td>Hisp</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>28</td>
<td>11</td>
<td>Hisp</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>32</td>
<td>12</td>
<td>Hisp</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Negative effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Older Blacks, No HS Degree</td>
<td>−17</td>
<td>43</td>
<td>10</td>
<td>Black</td>
<td>No</td>
<td>No</td>
<td>4130</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>−20</td>
<td>50</td>
<td>8</td>
<td>Black</td>
<td>Yes</td>
<td>No</td>
<td>5630</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>−17</td>
<td>29</td>
<td>12</td>
<td>White</td>
<td>No</td>
<td>Yes</td>
<td>12,200</td>
<td>No</td>
</tr>
<tr>
<td>Unmarried Whites, HS Degree</td>
<td>−17</td>
<td>31</td>
<td>13</td>
<td>White</td>
<td>No</td>
<td>Yes</td>
<td>5500</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>−19</td>
<td>31</td>
<td>12</td>
<td>White</td>
<td>No</td>
<td>Yes</td>
<td>495</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>−19</td>
<td>31</td>
<td>12</td>
<td>White</td>
<td>No</td>
<td>Yes</td>
<td>2610</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>−20</td>
<td>36</td>
<td>12</td>
<td>Hisp</td>
<td>No</td>
<td>Yes</td>
<td>11,500</td>
<td>No</td>
</tr>
<tr>
<td>High earning Hispanic</td>
<td>−21</td>
<td>34</td>
<td>11</td>
<td>Hisp</td>
<td>No</td>
<td>No</td>
<td>4640</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>−21</td>
<td>27</td>
<td>12</td>
<td>Hisp</td>
<td>No</td>
<td>Yes</td>
<td>24,300</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>−21</td>
<td>36</td>
<td>11</td>
<td>Hisp</td>
<td>No</td>
<td>No</td>
<td>3060</td>
<td>No</td>
</tr>
</tbody>
</table>

Note: Each row represents the estimated treatment effect given the characteristics of workers. The most effective treatment would target low-education, high-income Non-Hispanics; unemployed blacks with some college, and unemployed Hispanics. The treatment would be least effective when administered to older, employed recipients; unmarried whites with a high school degree but no college; and high earning Hispanics with no college. The last column represents the PSID weights, which are the inverse of the estimated probability of being in the NSW sample, standardized to have mean one. Weights marked with an asterisk indicate the groups which are not identified as having highest or lowest treatment effects when generalizing the results to the PSID sample (see Table 3 for those results).
Classification and Regression Trees (CART)

- CART is flexible and interpretable
- **T-learner** (Imai and Strauss. 2011. *Political Anal.*)
  - GOTV experiment with text messaging
  - separately fitted to the treatment (right) and control (left) groups

![Diagram of CART tree](image)

- **S-learner** (Athey and Imbens. 2016. *PNAS*)
  - target the MSE of CATE rather than the MSE of prediction
  - 3-way sample splitting: growing a tree, pruning, estimating CATE

8 / 10
**R-Learner** (Nie and Wager. 2021. *Biometrika*)

- Assumption: \( \{ Y_i(0), Y_i(1) \} \perp \perp T_i \mid X_i = x \) and \( 0 < \pi(x) < 1 \) for all \( x \)

- A motivating model for potential outcomes:
  \[
  Y_i(t) = \underbrace{\mathbb{E}(Y_i(0) \mid X_i)}_{\mu_0(X_i)} + t \times \underbrace{\tau(X_i)}_{\mu_1(X_i) - \mu_0(X_i)} + \epsilon_i(t) \quad \text{for } t = 0, 1
  \]

- Partial linear regression for (residualized) observed data:
  \[
  Y_i - \underbrace{\mathbb{E}(Y_i \mid X_i)}_{\mu(X_i)} = \left\{ T_i - \pi(X_i) \right\} \tau(X_i) + \epsilon_i
  \]
  where \( \mu(X_i) = \mu_0(X_i) + \pi(X_i)\tau(X_i) \) and \( \epsilon_i = \epsilon_i(T_i) \)

- Estimation procedure based on cross-validation
  1. Train models for \( \pi(x) \) and \( \mu(x) \)
  2. Obtain the CATE estimate via
  \[
  \hat{\tau} = \arg\min_{\tau} \frac{1}{n} \sum_{i=1}^{n} \left[ \{ Y_i - \hat{\mu}(X_i) \} - \left\{ T_i - \hat{\pi}(X_i) \right\} \tau(X_i) \right]^2 + \Lambda_n(\tau)
  \]
  regularization
Individualized Treatment Rule (ITR)

- **Two-step procedure:**
  1. estimate the CATE $\hat{\tau}(x)$
  2. construct an ITR as $f(x) = 1\{\hat{\tau}(x) > 0\}$

- **One-step procedure:** outcome weighted learning (Zhao et al. 2012. *J. Am. Stat. Assoc.*) $\sim\rightarrow$ optimal classification
  - randomized experiment
    \[
    \arg\max_f \mathbb{E}\{Y_i(f(X_i))\} = \arg\min_f \mathbb{E}\{Y_i(1 - f(X_i))\} \\
    = \arg\min_f \mathbb{E}[1\{f(X_i) = 0\} Y_i \mid T_i = 1] \\
    + \mathbb{E}[1\{f(X_i) = 1\} Y_i \mid T_i = 0]
    \]
    treated units who are assigned to control
    control units who are assigned to treatment

- classification problem $\sim\rightarrow$ weighted support vector machine:
  \[
  \arg\min_\tau \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{A_i \pi + (1 - A_i)/2} 1\{A_i \neq \text{sign}(\tau(X_i))\}
  \]
  where $A_i = 2T_i - 1 \in \{-1, 1\}$ and $\pi = \Pr(T_i = 1)$