Statistical Inference for Heterogeneous Treatment Effects Discovered by Generic Machine Learning in Randomized Experiments

Kosuke Imai
Harvard University

January 8, 2024
JSQPS Winter Meeting

Joint work with Michael Lingzhi Li (Harvard Business School)
Motivation and Overview

- Two methodological revolutions over the past two decades
  1. randomized experiments (field/lab/survey)
  2. machine learning

- Causal machine learning (causal ML)
  1. estimation of heterogeneous treatment effects
  2. development of individualized treatment rules

- Experimental evaluation of causal ML
  1. ML algorithms may not work well in practice
  2. assumption-free uncertainty quantification is essential

- I will show how to experimentally evaluate heterogeneous treatment effects discovered by generic causal ML
Setup

- **Notation:**
  - $n$ experimental units
  - $T_i \in \{0, 1\}$: binary treatment
  - $Y_i(t)$ where $t \in \{0, 1\}$: potential outcomes
  - $Y_i = Y_i(T_i)$: observed outcome
  - $X_i$: moderator of interest

- **Assumptions:**
  1. no interference between units:
     \[
     Y_i(T_1 = t_1, \ldots, T_n = t_n) = Y_i(T_i = t_i)
     \]
  2. randomization of treatment assignment:
     \[
     \{Y_i(1), Y_i(0)\} \perp \perp T_i
     \]
  3. random sampling of units:
     \[
     \{Y_i(1), Y_i(0)\} \overset{\text{i.i.d.}}{\sim} P
     \]
Exploration of Heterogeneous Treatment Effects

- Two commonly used treatment prioritization scores
  1. Conditional average treatment effect (CATE):
     \[
     \tau(x) = \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x)
     \]
  2. Baseline risk:
     \[
     \lambda(x) = \mathbb{E}(Y_i(0) \mid X_i = x)
     \]

- Estimate a score with ML algorithm using an external data set
  \[
  f : \mathcal{X} \rightarrow S \subset \mathbb{R}
  \]

- Group Average Treatment Effect (GATES; Chernozhukov et al. 2019)
  \[
  \tau_k = \mathbb{E}(Y_i(1) - Y_i(0) \mid p_{k-1} \leq S_i = f(X_i) < p_k)
  \]
  for \( k = 1, 2, \ldots, K \) where \( p_k \) is a cutoff (\( p_0 = -\infty, p_K = \infty \))
How can we make valid statistical inference for GATES without assuming that the scores are correctly estimated by ML algorithm?

A natural difference-in-means estimator for GATES:

\[
\hat{\tau}_k = \frac{K}{n_1} \sum_{i=1}^{n} Y_i T_i \hat{f}_k(X_i) - \frac{K}{n_0} \sum_{i=1}^{n} Y_i (1 - T_i) \hat{f}_k(X_i),
\]

where \( \hat{f}_k(X_i) = 1\{S_i \geq \hat{p}_k(s)\} - 1\{S_i \geq \hat{p}_{k-1}\} \)

Bias bound and exact variance are derived, accounting for the estimation uncertainty of cutoffs

Under mild regularity conditions (e.g., continuity of CATE at thresholds), the distribution of \( \hat{\tau}_k \) is asymptotically normal
Statistical Hypothesis Tests for Subgroups

1. Nonparametric test of treatment effect homogeneity:
   - Null hypothesis:
     \[ H_0 : \tau_1 = \tau_2 = \cdots = \tau_K. \]
   - Test statistic:
     \[ \hat{\tau}^\top \Sigma^{-1} \hat{\tau} \xrightarrow{d} \chi^2_K \]
     where \( \hat{\tau} = (\hat{\tau}_1 - \hat{\mu}, \cdots, \hat{\tau}_K - \hat{\mu})^\top \)

2. Nonparametric test of rank-consistent treatment effect heterogeneity:
   - Null hypothesis:
     \[ H_0^* : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_K. \]
   - Test statistic:
     \[ (\hat{\tau} - \mu^*(\hat{\tau}))^\top \Sigma^{-1} (\hat{\tau} - \mu^*(\hat{\tau})) \xrightarrow{d} \chi^2_K. \]
     where \( \mu^*(x) = \arg\min_{\mu} \|\mu - x\|^2_2 \) subject to \( \mu_1 \leq \mu_2 \leq \cdots \leq \mu_K \).
Estimation and Evaluation Using the Same Data

- **Cross-fitting procedure:**
  1. randomly split the data into $L$ folds: $\mathcal{Z}_1, \ldots, \mathcal{Z}_L$
  2. estimate the score using $L-1$ folds: $\hat{f}_{-\ell}$
  3. estimate GATES with the hold-out set: $\hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})$
  4. repeat the process for each $\ell$ and average

\[
\hat{\tau}_k(F; n-m) = \frac{1}{L} \sum_{\ell=1}^{L} \hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})
\]

where $F : \mathcal{Z} \rightarrow \mathcal{F}$ is a generic but stable ML algorithm with $\mathcal{Z}_{\text{train}} \in \mathcal{Z}$ and $\hat{f}_{\mathcal{Z}_{\text{train}}} = F(\mathcal{Z}_{\text{train}}) \in \mathcal{F}$

- **Estimand:** average performance of $F$

\[
\tau_k(F; n-m) = \mathbb{E}[\mathbb{E}\{Y_i(1) - Y_i(0) \mid p_{k-1}(\hat{f}_{\mathcal{Z}_{\text{train}}}^{n-m}) \leq \hat{f}_{\mathcal{Z}_{\text{train}}}^{n-m}(X_i) < p_k(\hat{f}_{\mathcal{Z}_{\text{train}}}^{n-m})\}].
\]

- **Unbiasedness:** $\mathbb{E}(\hat{\tau}_k(F; n-m)) = \tau_k(F; n-m)$
- **Finite-sample (conservative) variance estimator** (Imai and Li, *JASA*, 2023)
Simulation Study

- A highly nonlinear specification from the 2016 ACIC competition
  - 58 covariates (3 categorical, 5 binary, 27 counts, 13 continuous)
  - sample size: \( n = 4802 \)
  - use empirical distribution of \( X_i \) as true distribution

- Machine learning algorithms
  - Causal forest and Lasso
  - \( L = 5 \) and also use 5-fold cross validation for tuning

- Fixed score (see the paper) and estimated one with cross-fitting
Simulation Results: Bias and Coverage

<table>
<thead>
<tr>
<th></th>
<th>( n = 100 )</th>
<th>( n = 500 )</th>
<th>( n = 2500 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bias</td>
<td>s.d.</td>
<td>coverage</td>
</tr>
<tr>
<td><strong>Causal Forest</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{\tau}_1 )</td>
<td>-0.05</td>
<td>2.97</td>
<td>94.0%</td>
</tr>
<tr>
<td>( \hat{\tau}_2 )</td>
<td>-0.06</td>
<td>2.58</td>
<td>95.9</td>
</tr>
<tr>
<td>( \hat{\tau}_3 )</td>
<td>-0.01</td>
<td>2.56</td>
<td>96.7</td>
</tr>
<tr>
<td>( \hat{\tau}_4 )</td>
<td>-0.12</td>
<td>2.87</td>
<td>97.4</td>
</tr>
<tr>
<td>( \hat{\tau}_5 )</td>
<td>0.14</td>
<td>3.45</td>
<td>94.1</td>
</tr>
<tr>
<td><strong>LASSO</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{\tau}_1 )</td>
<td>-0.13</td>
<td>3.20</td>
<td>97.6%</td>
</tr>
<tr>
<td>( \hat{\tau}_2 )</td>
<td>0.04</td>
<td>2.28</td>
<td>97.5</td>
</tr>
<tr>
<td>( \hat{\tau}_3 )</td>
<td>-0.13</td>
<td>2.35</td>
<td>96.6</td>
</tr>
<tr>
<td>( \hat{\tau}_4 )</td>
<td>-0.00</td>
<td>2.54</td>
<td>96.8</td>
</tr>
<tr>
<td>( \hat{\tau}_5 )</td>
<td>0.11</td>
<td>3.62</td>
<td>96.2</td>
</tr>
</tbody>
</table>

- Reduction in standard errors compared with fixed \( F \) of the same evaluation size is more than 50% in some cases
Simulation Results: Size and Power of Tests

<table>
<thead>
<tr>
<th></th>
<th>( n = 100 )</th>
<th></th>
<th>( n = 500 )</th>
<th></th>
<th>( n = 2500 )</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rejection rate</td>
<td>median ( p )-value</td>
<td>rejection rate</td>
<td>median ( p )-value</td>
<td>rejection rate</td>
<td>median ( p )-value</td>
</tr>
<tr>
<td>Causal Forest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogeneity</td>
<td>1.4%</td>
<td>0.79</td>
<td>4.6%</td>
<td>0.71</td>
<td>51.4%</td>
<td>0.04</td>
</tr>
<tr>
<td>Rank-consistency</td>
<td>1.4%</td>
<td>0.70</td>
<td>0.8%</td>
<td>0.85</td>
<td>0.0%</td>
<td>0.98</td>
</tr>
<tr>
<td>LASSO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogeneity</td>
<td>0.6%</td>
<td>0.88</td>
<td>1.8%</td>
<td>0.85</td>
<td>9.0%</td>
<td>0.66</td>
</tr>
<tr>
<td>Rank-consistency</td>
<td>1.0%</td>
<td>0.72</td>
<td>0.6%</td>
<td>0.77</td>
<td>0.2%</td>
<td>0.89</td>
</tr>
</tbody>
</table>

- Heterogeneous but rank-consistent effects
- More conservative and lower power than fixed case
- When sample size is large, cross-fitting yields higher power
Empirical Application

- National Supported Work Demonstration Program (LaLonde 1986)
- Temporary employment program to help disadvantaged workers by giving them a guaranteed job for 9 to 18 months

Data
- sample size: $n_1 = 297$ and $n_0 = 425$
- outcome: annualized earnings in 1978 (36 months after the program)
- 7 pre-treatment covariates: demographics and prior earnings

Setup
- ML algorithms: Causal Forest, BART, and LASSO
- Sample-splitting: 2/3 of the data as training data
- Cross-fitting: 3 folds
### GATES Estimates (in 1,000 US Dollars)

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\tau}_1$</th>
<th>$\hat{\tau}_2$</th>
<th>$\hat{\tau}_3$</th>
<th>$\hat{\tau}_4$</th>
<th>$\hat{\tau}_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample-splitting</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BART</td>
<td>2.90</td>
<td>-0.73</td>
<td>-0.02</td>
<td>3.25</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td>[-2.25, 8.06]</td>
<td>[-5.05, 3.58]</td>
<td>[-3.47, 3.43]</td>
<td>[-1.53, 8.03]</td>
<td>[-3.82, 8.97]</td>
</tr>
<tr>
<td>Causal Forest</td>
<td>3.40</td>
<td>0.13</td>
<td>-0.85</td>
<td>-1.91</td>
<td>7.21</td>
</tr>
<tr>
<td></td>
<td>[-1.29, 3.40]</td>
<td>[-5.37, 5.63]</td>
<td>[-5.22, 3.52]</td>
<td>[-5.16, 1.34]</td>
<td>[1.22, 13.19]</td>
</tr>
<tr>
<td>LASSO</td>
<td>1.86</td>
<td>2.62</td>
<td>-2.07</td>
<td>1.39</td>
<td>4.17</td>
</tr>
<tr>
<td></td>
<td>[-3.59, 7.30]</td>
<td>[-1.69, 6.93]</td>
<td>[-5.39, 1.26]</td>
<td>[-2.95, 5.73]</td>
<td>[-2.30, 10.65]</td>
</tr>
<tr>
<td><strong>Cross-fitting</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BART</td>
<td>0.40</td>
<td>-0.15</td>
<td>-0.40</td>
<td>2.52</td>
<td>2.19</td>
</tr>
<tr>
<td></td>
<td>[-3.79, 4.59]</td>
<td>[-2.54, 2.23]</td>
<td>[-3.37, 2.56]</td>
<td>[-0.99, 6.03]</td>
<td>[-0.73, 5.11]</td>
</tr>
<tr>
<td>Causal Forest</td>
<td>-3.72</td>
<td>1.05</td>
<td>5.32</td>
<td>-2.64</td>
<td>4.55</td>
</tr>
<tr>
<td></td>
<td>[-6.52, -0.93]</td>
<td>[-2.28, 4.37]</td>
<td>[2.63, 8.01]</td>
<td>[-5.07, -0.22]</td>
<td>[1.14, 7.96]</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.65</td>
<td>0.45</td>
<td>-2.88</td>
<td>1.32</td>
<td>5.02</td>
</tr>
<tr>
<td></td>
<td>[-3.65, 4.94]</td>
<td>[-3.28, 4.18]</td>
<td>[-5.38, -0.38]</td>
<td>[-1.83, 4.48]</td>
<td>[-0.14, 10.18]</td>
</tr>
</tbody>
</table>

- Greater statistical power with cross-fitting
- ML algorithms are not necessarily reliable
## Results of Hypothesis Tests

<table>
<thead>
<tr>
<th></th>
<th>Causal Forest</th>
<th></th>
<th>BART</th>
<th></th>
<th>LASSO</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>stat</td>
<td>p-value</td>
<td>stat</td>
<td>p-value</td>
<td>stat</td>
<td>p-value</td>
</tr>
<tr>
<td><strong>Sample-splitting</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogeneity</td>
<td>9.78</td>
<td>0.08</td>
<td>2.76</td>
<td>0.74</td>
<td>5.26</td>
<td>0.36</td>
</tr>
<tr>
<td>Rank-consistency</td>
<td>3.07</td>
<td>0.32</td>
<td>1.13</td>
<td>0.66</td>
<td>3.14</td>
<td>0.30</td>
</tr>
<tr>
<td><strong>Cross-fitting</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogeneity</td>
<td>30.29</td>
<td>0.00</td>
<td>2.32</td>
<td>0.80</td>
<td>10.79</td>
<td>0.06</td>
</tr>
<tr>
<td>Rank-consistency</td>
<td>0.06</td>
<td>0.69</td>
<td>0.04</td>
<td>0.89</td>
<td>0.45</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)

- Safe deployment of causal ML requires uncertainty quantification
  - experimental evaluation of HTEs and ITRs
  - no modeling assumption
  - no resampling (computationally efficient)
  - applicable to any complex causal ML algorithms
  - good small sample performance

- Open source software: evalITR: Evaluating Individualized Treatment Rules at CRAN https://CRAN.R-project.org/package=evalITR

- More information: https://imai.fas.harvard.edu/research/