Statistical Inference for Subgroups Discovered by Machine Learning

Kosuke Imai

Harvard University

The 14th Annual Conference on Statistical Issues in Clinical Trials
University of Pennsylvania
April 12, 2022

Approaches to Subgroup Identification

1. Adaptive experimental design (Simon)
    - Goal: identify a subgroup with a positive average effect
    - Pre-specify strata and then drop those with little promise
2. Multi-period crossover trial (Ivanova)
    - Goal: identify the subgroup that maximizes the product of the average treatment effect and prevalence
    - Inference based on cross-validation and bootstrap
3. Estimation of the conditional average treatment effect (Lipkovich)
    - Goal: use machine learning to estimate the CATE
    - Identify a subgroup with large CATE estimates
4. Non-exchangeable subgroups (Schnell)
    - Goal: test consistency or heterogeneity among subgroups
    - Challenges of multiple comparisons in subgroup analysis

Subgroup Identification with Machine Learning (ML)

- What if we use an ML algorithm to identify subgroups?
- Can we make proper statistical inference for discovered subgroups?
  - ML algorithms can be blackbox or even adhoc
  - cannot assume ML algorithms converge uniformly
  - avoid a computationally intensive procedure
- Joint work with Michael Lingzhi Li (MIT)
- Setup:
  - Conditional Average Treatment Effect (CATE):

$$\tau(\mathrm{x}) \; = \; \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathrm{X}_i = \mathrm{x})$$

  - CATE estimation based on a generic ML algorithm

$$s : \mathcal{X} \longrightarrow \mathcal{S} \subset \mathbb{R}$$

  - Sorted Group Average Treatment Effect (GATE; Chernozhukov et al. 2019)

$$\tau_k \; := \; \mathbb{E}(Y_i(1) - Y_i(0) \mid c_{k-1}(s) \leq s(\mathrm{X}_i) < c_k(s))$$

    for $k = 1, 2, \ldots, K$ where c_k represents the cutoff between the $(k-1)$th and kth groups

Statistical Inference for Subgroups

- An unbiased GATE estimator (within-subgroup difference-in-means):

$$\hat{\tau}_k \;=\; \frac{K}{n_1}\sum_{i=1}^{n} Y_i T_i \hat{f}_k(\mathrm{X}_i) - \frac{K}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)\hat{f}_k(\mathrm{X}_i),$$

where $\hat{f}_k(\mathrm{X}_i) \;=\; 1\{s(\mathrm{X}_i) \geq \hat{c}_k(s)\} - 1\{s(\mathrm{X}_i) \geq \hat{c}_{k-1}(s)\}$

- Statistical inference based on Neyman's repeated sampling framework
  - random assignment of treatment
  - random sampling of units
  - random splits for cross-fitting
- Standard error and confidence intervals, etc. for each τ_k
- No assumption about the properties of ML algorithms

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:
     $$\hat{\boldsymbol{\tau}}^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\tau}} \ \xrightarrow{d} \ \chi^2_K$$
     where $\hat{\boldsymbol{\tau}} \ = \ (\hat{\tau}_1 - \hat{\tau}, \cdots, \hat{\tau}_K - \hat{\tau})^\top$

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:
     $$(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}})) \xrightarrow{d} \bar{\chi}^2_K.$$
     where $\boldsymbol{\mu}^*(\boldsymbol{x}) = \mathrm{argmin}_{\boldsymbol{\mu}} \|\boldsymbol{\mu} - \boldsymbol{x}\|_2^2$ subject to $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_K$.

Simulation Study

Estimator	truth	$n_{\text{test}} = 100$ bias	coverage	$n_{\text{test}} = 500$ bias	coverage	$n_{\text{test}} = 2500$ bias	coverage
Causal Forest							
$\hat{\tau}_1$	2.164	0.034	93.8%	0.041	95.0%	0.007	96.0%
$\hat{\tau}_2$	4.001	0.011	93.7	−0.060	94.4	−0.002	95.3
$\hat{\tau}_3$	4.583	−0.018	94.0	−0.003	96.4	0.020	95.8
$\hat{\tau}_4$	4.931	−0.077	94.6	0.001	94.3	0.003	95.6
$\hat{\tau}_5$	5.728	−0.058	96.0	−0.010	95.0	−0.009	95.2
BART							
$\hat{\tau}_1$	2.092	0.016	94.0%	−0.014	96.2%	0.009	95.8%
$\hat{\tau}_2$	3.913	0.127	95.1	0.028	94.0	−0.003	95.3
$\hat{\tau}_3$	4.478	−0.077	94.3	−0.041	95.0	−0.001	95.1
$\hat{\tau}_4$	5.042	−0.154	94.2	0.014	95.8	0.015	95.4
$\hat{\tau}_5$	5.881	−0.019	94.7	−0.019	94.4	−0.000	95.0
LASSO							
$\hat{\tau}_1$	3.243	0.028	94.1%	0.049	95.1%	0.003	95.1%
$\hat{\tau}_2$	3.817	−0.012	93.6	−0.013	94.5	−0.000	95.4
$\hat{\tau}_3$	4.318	−0.013	94.2	−0.002	94.5	0.010	95.0
$\hat{\tau}_4$	4.788	−0.041	94.0	−0.015	94.6	−0.001	94.6
$\hat{\tau}_5$	5.241	−0.046	94.4	0.021	95.1	0.002	95.3

Concluding Remarks

- Statistical inference for subgroups is challenging especially when they are discovered by complex machine learning algorithms
- The proposed methodology
    - no modeling assumption is required
    - any machine learning algorithms can be used
    - design-based: random sampling, random assignments, random splits
    - applicable to cross-fitting estimators
    - simulations: good small sample performance
- Ongoing extension: dynamic treatment regime settings
- Papers:
    - https://arxiv.org/pdf/2203.14511.pdf
    - Experimental Evaluation of Individualized Treatment Rules (*Journal of the American Statistical Association*)
- Open-source software (R package): evalITR: Evaluating Individualized Treatment Rules