Heterogeneous Treatment Effects

Kosuke Imai

Harvard University

Spring 2021

Heterogeneous Treatment Effects

- Same treatment may affect different individuals differently
- Conditional Average Treatment Effect (CATE)

$$\tau(\mathbf{x}) \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}) \quad \text{where} \quad \mathbf{x} \in \mathcal{X}$$

  - 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 $\mathbf{X}_i$

- 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

Machine Learning for Heterogeneous Causal Effects

- Motivation:
  1. avoid strong modeling assumptions $\rightsquigarrow$ data-driven approach
  2. avoid false discoveries $\rightsquigarrow$ avoid over-fitting via regularization

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

- Mean squared error decomposition:

$$\begin{aligned} & \mathbb{E}[(\tau_i - \hat{\tau}(\mathbf{x}))^2 \mid \mathbf{X}_i = \mathbf{x}] \\ = \; & \mathbb{E}[(\tau_i - \tau(\mathbf{x}))^2 \mid \mathbf{X}_i = \mathbf{x}] + \mathbb{E}[(\tau(\mathbf{x}) - \hat{\tau}(\mathbf{x}))^2 \mid \mathbf{X}_i = \mathbf{x}] \end{aligned}$$

- Inference of heterogenous treatment effects depends on
  1. How predictive $\mathbf{X}_i$ is of τ_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 $\mu_t(\mathbf{x}) = \mathbb{E}(Y_i \mid T_i = t, \mathbf{X}_i = \mathbf{x})$ using a single model
  2. compute $\hat{\tau}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x})$

  $\rightsquigarrow$ modeling interactions between T_i and $\mathbf{X}_i$ can be challenging

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

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

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

  $\rightsquigarrow$ more robust but less efficient

Penalized Maximum Likelihood Estimator

- PMLE:
$$\hat{\boldsymbol{\theta}} \;=\; \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \;\log \mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}) + P(\lambda, \boldsymbol{\theta})$$
  - Ridge: $P(\lambda, \boldsymbol{\theta}) = \lambda \sum_{j=1}^{p} \beta_j^2$
  - Lasso: $P(\lambda, \boldsymbol{\theta}) = \lambda \sum_{j=1}^{p} |\beta_j|$
- Sample splitting:
  1. training data: estimate $\boldsymbol{\theta}$ given λ
  2. test data: choose $\hat{\lambda}$
  3. validation data: estimate CATE given $\hat{\lambda}$
- S-learner (Imai and Ratkovic. 2013. *Ann. Appl. Stat.*)
  - Lasso with support vector machine
  - separate tuning parameters λ 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

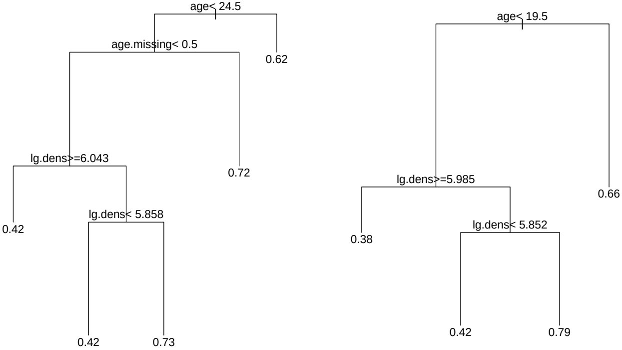
Job Training Program (Imai and Ratkovic. 2013. *Ann. Appl. Stat.*)

- 44 covariates including some square and interaction terms
- 44 interactions between the treatment and covariates
- sparcity of the model helps with interpretation

Groups most helped or hurt by the treatment	**Average effect**	**Age**	**Educ.**	**Race**	**Married**	**Highschool degree**	**Earnings (1975)**	**Unemp. (1975)**
Positive effects								
Low education, Non-Hispanic	53	31	4	White	No	No	10,700	No
High Earning	50	31	4	Black	No	No	4020	No
	40	28	15	Black	No	Yes	0	Yes
Unemployed, Black,	38	30	14	Black	Yes	Yes	0	Yes
Some College	37	22	16	Black	No	Yes	0	Yes
	45	33	5	Hisp	No	No	0	Yes
	39	50	10	Hisp	No	No	0	Yes
Unemployed, Hispanic	37	33	9	Hisp	Yes	No	0	Yes
	37	28	11	Hisp	Yes	No	0	Yes
	37	32	12	Hisp	Yes	Yes	0	Yes
Negative effects								
Older Blacks,	−17	43	10	Black	No	No	4130	No
No HS Degree	−20	50	8	Black	Yes	No	5630	No
	−17	29	12	White	No	Yes	12,200	No
Unmarried Whites,	−17	31	13	White	No	Yes	5500	No
HS Degree	−19	31	12	White	No	Yes	495	No
	−19	31	12	White	No	Yes	2610	No
	−20	36	12	Hisp	No	Yes	11,500	No
High earning Hispanic	−21	34	11	Hisp	No	No	4640	No
	−21	27	12	Hisp	No	Yes	24,300	No
	−21	36	11	Hisp	No	No	3060	No

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



- *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
- Random forest (Wager and Athey. 2018. *J. Amer. Stat. Asoc.*)

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

- Assumption: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid \mathbf{X}_i = \mathbf{x}$ and $0 < \pi(\mathbf{x}) < 1$ for all $\mathbf{x}$
- A motivating model for potential outcomes:

$$Y_i(t) \;=\; \underbrace{\mathbb{E}(Y_i(0) \mid \mathbf{X}_i)}_{\mu_0(\mathbf{X}_i)} + t \times \underbrace{\tau(\mathbf{X}_i)}_{\mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)} + \epsilon_i(t) \quad \text{for } t = 0, 1$$

- Partial linear regression for (residualized) observed data:

$$Y_i - \underbrace{\mathbb{E}(Y_i \mid \mathbf{X}_i)}_{\mu(\mathbf{X}_i)} \;=\; \{T_i - \pi(\mathbf{X}_i)\}\tau(\mathbf{X}_i) + \epsilon_i$$

  where $\mu(\mathbf{X}_i) = \mu_0(\mathbf{X}_i) + \pi(\mathbf{X}_i)\tau(\mathbf{X}_i)$ and $\epsilon_i = \epsilon_i(T_i)$
- Estimation procedure based on cross-validation
  1. Train models for $\pi(\mathbf{x})$ and $\mu(\mathbf{x})$
  2. Obtain the CATE estimate via

$$\hat{\tau} \;=\; \underset{\tau}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} [\{Y_i - \hat{\mu}(\mathbf{X}_i)\} - \{T_i - \hat{\pi}(\mathbf{X}_i)\}\tau(\mathbf{X}_i)]^2 + \underbrace{\Lambda_n(\tau)}_{\text{regularization}}$$

Individualized Treatment Rule (ITR)

- Two-step procedure:
  1. estimate the CATE $\hat{\tau}(\mathbf{x})$
  2. construct an ITR as $f(\mathbf{x}) = \mathbf{1}\{\hat{\tau}(\mathbf{x}) > 0\}$
- One-step procedure: outcome weighted learning (Zhao *et al*. 2012. *J. Am. Stat. Assoc.*) $\rightsquigarrow$ optimal classification
  - randomized experiment

$$\begin{aligned}
\underset{f}{\operatorname{argmax}}\, \mathbb{E}\{Y_i(f(\mathbf{X}_i))\} &= \underset{f}{\operatorname{argmin}}\, \mathbb{E}\{Y_i(1 - f(\mathbf{X}_i))\} \\
= \underset{f}{\operatorname{argmin}} \quad & \underbrace{\mathbb{E}[\mathbf{1}\{f(\mathbf{X}_i) = 0\} Y_i \mid T_i = 1]}_{\text{treated units who are assigned to control}} \\
+ \quad & \underbrace{\mathbb{E}[\mathbf{1}\{f(\mathbf{X}_i) = 1\} Y_i \mid T_i = 0]}_{\text{control units who are assigned to treatment}}
\end{aligned}$$

  - classification problem $\rightsquigarrow$ weighted support vector machine:

$$\underset{\tau}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{Y_i}{A_i \pi + (1 - A_i)/2}}_{\textit{weights}} \mathbf{1}\{A_i \neq \operatorname{sign}(\tau(\mathbf{X}_i))\}$$

where $A_i = 2T_i - 1 \in \{-1, 1\}$ and $\pi = \Pr(T_i = 1)$