Heterogeneous Treatment Effects

Kosuke Imai

Harvard University

STAT 186 / GOV 2002 CAUSAL INFERENCE

Fall 2019

Causal Heterogeneity

1. Heterogeneous Treatment Effects
   - Same treatment may affect different individuals differently
   - Conditional Average Treatment Effect (CATE)

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

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

2. Causal interaction
   - Different combinations of treatments may have different effects
   - Interaction among treatment variables instead of interaction between a treatment and covariates
   - Factorial designs, e.g., conjoint analysis

Subgroup Analysis and Pre-registration

- Stratify the data and estimate the ATE within each strata
- Most straightforward and popular unbiased estimation of the CATE

- Problem: false discovery (data snooping, "p-hacking", "fishing")
- Solution: Pre-register hypotheses and analyses
  - standard in medicine, becoming 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

Machine Learning for Heterogeneous Causal Effects

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

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

  - in a randomized experiment, we have

$$\tau(\mathbf{x}) \ = \ \mathbb{E}(Y_i \mid T_i = 1, \mathbf{X}_i = \mathbf{x}) - \mathbb{E}(Y_i \mid T_i = 0, \mathbf{X}_i = \mathbf{x})$$

- 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 $\hat{\tau}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x})$

  $\rightsquigarrow$ may be difficult if the treatment assignment is lopsided

- *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 $\mathbf{X}_i$

  $\rightsquigarrow$ may be more robust but less efficient than *T*-learner

Penalized Maximum Likelihood Estimator

- Recall the 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: $P(\lambda, \boldsymbol{\theta}) = \lambda \sum_{j=1}^{p} |\beta_j|$

- *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

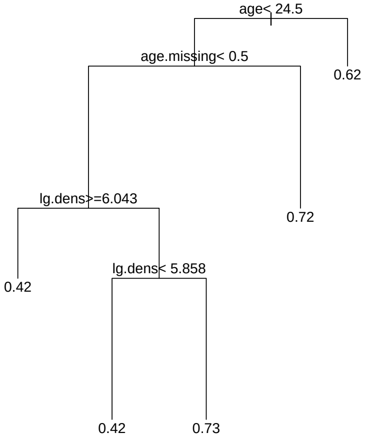
- *T*-learner (Qian and Murphy. 2011. *Ann. Stat.*)
  - Lasso with least squares
  - separately fitted for the treatment and control groups
  - uses *S*-learner when the treatment has more than 2 categories

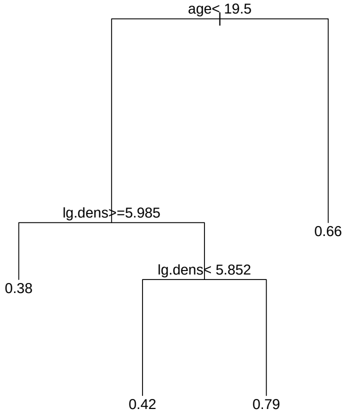
- 44 covariates including some square and interaction terms
- 44 interactions between the treatment and covariates
- sparcity of the model helps with interpretation

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





CART for Causal Effects (Athey and Imbens. 2016. *PNAS*)

- Predictive criteria for a tree Π:

$$\begin{aligned}\text{MSE}_\mu &= \frac{1}{N_{\text{test}}} \sum_{i \in \mathcal{S}_{\text{test}}} \left[\{Y_i - \hat{\mu}(\mathbf{X}_i; \mathcal{S}_{\text{train}}, \Pi)\}^2 - Y_i^2\right] \\ &= \frac{1}{N_{\text{test}}} \sum_{i \in \mathcal{S}_{\text{test}}} \left\{\hat{\mu}(\mathbf{X}_i; \mathcal{S}_{\text{train}}, \Pi)^2 - 2\hat{\mu}(\mathbf{X}_i; \mathcal{S}_{\text{test}}, \Pi)\hat{\mu}(\mathbf{X}_i; \mathcal{S}_{\text{train}}, \Pi)\right\}\end{aligned}$$

  where $\hat{\mu}(\mathbf{x}; \mathcal{S}, \Pi) = \sum_{i \in \mathcal{S}: \mathbf{X}_i \in \ell(\mathbf{x}; \Pi)} Y_i / \#\{i \in \mathcal{S} : \mathbf{X}_i \in \ell(\mathbf{x}; \Pi)\}$

- Causal criteria for a tree Π:

$$\begin{aligned}\text{MSE}_\tau &= \frac{1}{N_{\text{test}}} \sum_{i \in \mathcal{S}_{\text{test}}} \left[\{\tau_i - \hat{\tau}(\mathbf{X}_i; \mathcal{S}_{\text{train}}, \Pi)\}^2 - \tau_i^2\right] \\ &= \frac{1}{N_{\text{test}}} \sum_{i \in \mathcal{S}_{\text{test}}} \left\{\hat{\tau}(\mathbf{X}_i; \mathcal{S}_{\text{train}}, \Pi)^2 - 2\hat{\tau}(\mathbf{X}_i; \mathcal{S}_{\text{test}}, \Pi)\hat{\tau}(\mathbf{X}_i; \mathcal{S}_{\text{train}}, \Pi)\right\}\end{aligned}$$

  where $\hat{\tau}(\mathbf{X}_i; \mathcal{S}, \Pi) = \hat{\mu}(1, \mathbf{X}_i; S, \Pi) - \hat{\mu}(0, \mathbf{X}_i; S, \Pi)$

- Honest inference $\rightsquigarrow$ use separate samples for splitting, estimating, and validating

Outcome Weighted Learning (Zhao *et al.* 2012. *J. Am. Stat. Assoc.*)

- So far, we used a two-step procedure:
  1. estimate the CATE
  2. construct an optimal treatment rule using the estimated CATE
- An alternative approach: directly estimate the optimal treatment rule that maximizes the outcome
  - randomized experiment: $A_i = 1$ (treated) and $= -1$ (control)
  - individualized treatment rule: $D(\mathbf{X}_i) \in \{-1, 1\}$

$$D^* = \underset{D}{\operatorname{argmax}} \mathbb{E}\{Y_i(D(\mathbf{X}_i))\} = \underset{D}{\operatorname{argmin}} \mathbb{E}\left[\frac{\mathbf{1}\{A_i \neq D(\mathbf{X}_i)\}}{A_i\pi + (1 - A_i)/2} Y_i\right]$$

  where $\pi = \Pr(A_i = 1)$
  - classification problem $\rightsquigarrow$ weighted support vector machine:

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

  where $D(\mathbf{X}_i) = \operatorname{sign}(f(\mathbf{X}_i))$

Causal Interaction

- Another type of causal heterogeneity:
  - What combination of treatments is efficacious?
  - Interaction among multiple treatment variables
- Factorial experiments: e.g., conjoint analysis
- Example: Immigration preference (Hopkins and Hainmueller 2014)
  - representative sample of 1,407 American adults
  - each respondent evaluates 5 pairs of immigrant profiles
  - `gender`^{2}, `education`^{7}, `origin`^{10}, `experience`^{4}, `plan`^{4}, `language`^{4}, `profession`^{11}, `application reason`^{3}, `prior trips`^{5}
  - What combinations of immigrant characteristics do Americans prefer?
  - High dimension: over 1 million treatment combinations

Factorial Experiments with Two Treatments

- Two factorial treatments (e.g., gender and race):

$$
\begin{aligned}
A &\in \mathcal{A} = \{a_0, a_1, \ldots, a_{L_A-1}\} \\
B &\in \mathcal{B} = \{b_0, b_1, \ldots, b_{L_B-1}\}
\end{aligned}
$$

- Assumption: Full factorial design
  1. Randomization of treatment assignment

$$
\{Y(a_\ell, b_m)\}_{a_\ell \in \mathcal{A}, b_m \in \mathcal{B}} \perp\!\!\!\perp \{A, B\}
$$

  2. Non-zero probability for all treatment combination

$$
\Pr(A = a_\ell, B = b_m) > 0 \quad \text{for all } a_\ell \in \mathcal{A} \quad \text{and} \quad b_m \in \mathcal{B}
$$

Causal Estimands in Factorial Experiments

1. Average Combination Effect (ACE):
   - Average effect of treatment combination $(A, B) = (a_\ell, b_m)$ relative to the baseline condition $(A, B) = (a_0, b_0)$

$$\tau_{AB}(a_\ell, b_m; a_0, b_0) \quad = \quad \mathbb{E}\{Y(a_\ell, b_m) - Y(a_0, b_0)\}$$

   - Effect of being Asian male

2. Average Marginal Effect (AME; Hainmueller *et al.* 2014; Dasgupta *et al.* 2015):
   - Average effect of treatment $A = a_\ell$ relative to the baseline condition $A = a_0$ averaging over the other treatment B

$$\psi_A(a_\ell, a_0) \quad = \quad \int \mathbb{E}\{Y(a_\ell, B) - Y(a_0, B)\} dF(B)$$

   - Effect of being male averaging over race

Both can be estimated using the difference-in-means estimators

Causal Interaction Effects (Egami and Imai. 2019. *J. Am. Stat. Assoc.*)

- Average Marginal Interaction Effect (AMIE):

$$\pi_{AB}(a_\ell, b_m; a_0, b_0) = \underbrace{\tau_{AB}(a_\ell, b_m; a_0, b_0)}_{\text{ACE of } (a_\ell, b_m)} - \underbrace{\psi_A(a_\ell, a_0)}_{\text{AME of } a_\ell} - \underbrace{\psi_B(b_m, b_0)}_{\text{AME of } b_m}$$

- Interpretation: additional effect induced by $A = a_\ell$ and $B = b_m$ together beyond the separate effect of $A = a_\ell$ and that of $B = b_m$
- Additional effect of being Asian male beyond the sum of separate effects for being male and being Asian

- Invariance: *relative magnitude* of AMIE does not depend on the choice of baseline condition
- Generalizable to higher-order interaction

- ANOVA with direct regularization on AMEs and AMIEs

Conjoint Analysis of Ethnic Voting in Africa

- Ethnic voting and accountability (Carlson 2015, *World Politics*)
- Do voters prefer candidates of same ethnicity regardless of their prior performance? Do ethnicity and performance interact?

- Conjoint analysis in Uganda: 547 voters from 32 villages
- Each voter evaluates 3 pairs of hypothetical candidates
- 5 factors: `Coethnicity`^{2}, `Prior record`^{2}, `Prior office`^{4}, `Platform`^{3}, `Education`^{8}

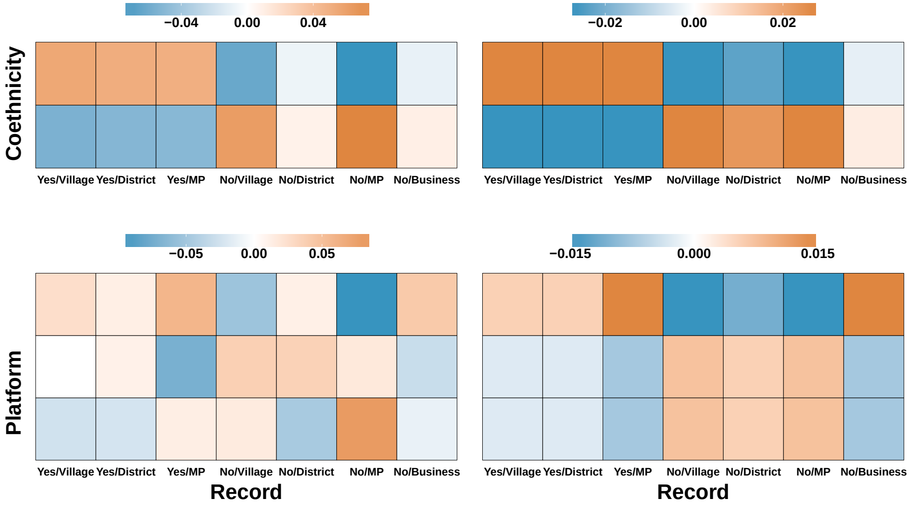
- Linear probability model with ANOVA constraints and direct regularization on AMEs and AMIEs

Ranges of Estimated AMEs and AMIEs

	Range	Selection prob.
AME		
Record	0.122	1.00
Coethnicity	0.053	1.00
Platform	0.023	0.93
Degree	0.000	0.33
AMIE		
Coethnicity × Record	0.053	1.00
Record × Platform	0.030	0.92
Platform × Coethnic	0.008	0.64
Coethnicity × Degree	0.000	0.62
Platform × Degree	0.000	0.35
Record × Degree	0.000	0.09

- Factor selection probability based on bootstrap

Effect of Regularization on AMIEs



Without Regularization **With Regularization**

Concluding Remarks

- Interaction effects play an essential role in causal heterogeneity
  1. causal moderation = treatment $\times$ covariate
  2. causal interaction = treatment $\times$ treatment

- Problems of multiple testing
  - pre-registration
  - controlling family-wise error and false discovery rates
  - regularization

- Role of machine learning methods
  - causal inference = counterfactual prediction
  - machine learning plays a role in estimation rather than identification (but see the growing literature on causal discovery)
  - key = optimize causal quantities of interest