Heterogeneous Causal Effects with Machine Learning

Apoorva Lal

July 24, 2022

Stanford

Heterogeneous Treatment Effects - Setup

- For i.i.d. observations $i \in \{1, .., N\}$, we observe $\{Y_i, \mathbf{X}_i, W_i\}_i^N$ where:
  - $Y_i \in \mathbb{R}$ is the **outcome**
  - $W_i \in \{0, \ldots, K\}$ is the **treatment assignment**
  - $\mathbf{X}_i \in \mathbb{R}^k$ is the **feature vector**
- We posit the existence of **potential outcomes** $Y^0, \ldots, Y^k$ for each unit. Append them into a 'science table' that is $N \times K$.

Heterogeneous Treatment Effects - Setup

- For i.i.d. observations $i \in \{1, .., N\}$, we observe $\{Y_i, \mathbf{X}_i, W_i\}_i^N$ where:
  - $Y_i \in \mathbb{R}$ is the **outcome**
  - $W_i \in \{0, \ldots, K\}$ is the **treatment assignment**
  - $\mathbf{X}_i \in \mathbb{R}^k$ is the **feature vector**
- We posit the existence of **potential outcomes** $Y^0, \ldots, Y^k$ for each unit. Append them into a 'science table' that is $N \times K$.
- Treatment effects (*estimands*) are defined as functions of *potential outcomes*, and since $(K-1)/K$ of them are unobserved, we need assumptions to use *estimators* to compute them using data.

Identifying counterfactual means and friends

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i=k} Y_i^k$. What does this assume?

Identifying counterfactual means and friends

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i = k} Y_i^k$. What does this assume?
  - i's outcome is only affected by i's treatment status. This may not be the case in many settings, e.g. with peer effects/interference/spillovers/contagion.
  - In such settings, the potential outcomes are indexed by $Y^{\mathbf{W}}$. In the extreme case of unrestricted interference, the 'science table' has width K^n. Need new assumptions / different estimands.

Identifying counterfactual means and friends

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i=k} Y_i^k$. What does this assume?
  - i's outcome is only affected by i's treatment status. This may not be the case in many settings, e.g. with peer effects/interference/spillovers/contagion.
  - In such settings, the potential outcomes are indexed by $Y^{\mathbf{W}}$. In the extreme case of unrestricted interference, the 'science table' has width K^n. Need new assumptions / different estimands.
- *Unconfoundedness*: $Y^1, Y^0 \perp\!\!\!\perp W_i | \mathbf{X}_i$. Treatment is as good as random given covariates.
- *Overlap*: $0 < \pi^w(\mathbf{X}) < 1$. Each unit has positive probability of treatment.

Identifying counterfactual means and friends

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i=k} Y_i^k$. What does this assume?
  - i's outcome is only affected by i's treatment status. This may not be the case in many settings, e.g. with peer effects/interference/spillovers/contagion.
  - In such settings, the potential outcomes are indexed by $Y^{\mathbf{W}}$. In the extreme case of unrestricted interference, the 'science table' has width K^n. Need new assumptions / different estimands.
- *Unconfoundedness*: $Y^1, Y^0 \perp\!\!\!\perp W_i | \mathbf{X}_i$. Treatment is as good as random given covariates.
- *Overlap*: $0 < \pi^w(\mathbf{X}) < 1$. Each unit has positive probability of treatment.

Then, the *Counterfactual mean* is non-parametrically identified, as are causal contrasts. AIPW estimator:

$$\widehat{\Gamma}_i^{(w)} = \underbrace{\widehat{\mu}^w(\mathbf{X})}_{\text{Outcome Model}} + \underbrace{\frac{\mathbb{1}_{W_i=w}}{\widehat{\pi}^w(\mathbf{X})}}_{\text{(Inv) Propensity score}} \left(Y_i - \widehat{\mu}^w(\mathbf{X})\right)$$

Identifying counterfactual means and friends

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i = k} Y_i^k$. What does this assume?
  - i's outcome is only affected by i's treatment status. This may not be the case in many settings, e.g. with peer effects/interference/spillovers/contagion.
  - In such settings, the potential outcomes are indexed by $Y^{\mathbf{W}}$. In the extreme case of unrestricted interference, the 'science table' has width K^n. Need new assumptions / different estimands.
- *Unconfoundedness*: $Y^1, Y^0 \perp\!\!\!\perp W_i | \mathbf{X}_i$. Treatment is as good as random given covariates.
- *Overlap*: $0 < \pi^w(\mathbf{X}) < 1$. Each unit has positive probability of treatment.

Then, the *Counterfactual mean* is non-parametrically identified, as are causal contrasts. AIPW estimator:

$$\widehat{\Gamma}_i^{(w)} = \underbrace{\widehat{\mu}^w(\mathbf{X})}_{\text{Outcome Model}} + \underbrace{\frac{\mathbb{1}_{W_i = w}}{\widehat{\pi}^w(\mathbf{X})}}_{\text{(Inv) Propensity score}} \left(Y_i - \widehat{\mu}^w(\mathbf{X})\right)$$

- $\widehat{\mu}^w(\cdot), \widehat{\pi}^w(\cdot)$ are *nuisance functions* (potentially) high-dim quantities incidental to low-dim target (marginal mean, causal contrast).
- All nuisance functions are henceforth *cross-fit*

Heterogeneous Effects

- Focus (w.log) on binary treatment case
- We are interested in the **Conditional Average Treatment Effect (CATE)**:
  $\tau(\mathbf{X}) = E[Y^{(1)} - Y^{(0)} | \mathbf{X} = \mathbf{x}]$
    - This is a *function*, not a number, so we may want to summarise
        - projecting imputed effects linearly on covariates (BLP)
        - binning estimates (GATE)

Parametric Outcome Modeling: Estimate OLS with interactions

- $Y_i = \beta_0 + \beta_1 W_i + \beta_2 X_i + \beta_3 W_i X_i + \epsilon_i$
  - Implicit outcome models: $Y_i^0 = \beta_2 X_i, Y_i^1 = Y_i^0 + \beta_1 + \beta_3 X_i$
- $\widehat{\text{CATE}}_X = \hat{\beta}_1 + \hat{\beta}_3 X_i$
- Why do we need machine learning / regularization to do this?

Parametric Outcome Modeling: Estimate OLS with interactions

- $Y_i = \beta_0 + \beta_1 W_i + \beta_2 X_i + \beta_3 W_i X_i + \epsilon_i$
  - Implicit outcome models: $Y_i^0 = \beta_2 X_i, Y_i^1 = Y_i^0 + \beta_1 + \beta_3 X_i$
- $\widehat{\text{CATE}}_X = \hat{\beta}_1 + \hat{\beta}_3 X_i$
- Why do we need machine learning / regularization to do this?
  - **Overfitting**: We know that in general, when $k \approx N$, traditional OLS methods will badly overfit
  - **Unknown Functional Form**: The analyst does not know what the underlying heterogeneity looks like
  - **fishing**: Why should the reader believe that this specification fell from the sky?

T-Learner

- fits separate models on the treated and controls.
- Learn $\hat{\mu}_{(0)}(x)$ by predicting Y_i from X_i on the subset of observations with $W_i = 0$.
- Learn $\hat{\mu}_{(1)}(x)$ by predicting Y_i from X_i on the subset of observations with $W_i = 1$.
- Report $\hat{\tau}(x) = \hat{\mu}_{(1)}(x) - \hat{\mu}_{(0)}(x)$.

Pure ML Ideas

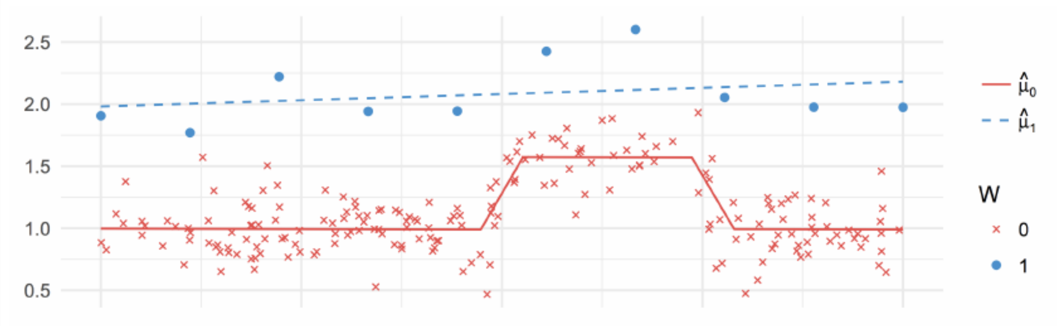
T-Learner

- fits separate models on the treated and controls.
- Learn $\hat{\mu}_{(0)}(x)$ by predicting Y_i from X_i on the subset of observations with $W_i = 0$.
- Learn $\hat{\mu}_{(1)}(x)$ by predicting Y_i from X_i on the subset of observations with $W_i = 1$.
- Report $\hat{\tau}(x) = \hat{\mu}_{(1)}(x) - \hat{\mu}_{(0)}(x)$.

S-Learner

- fits a single model to all the data.
- Learn $\hat{\mu}(z)$ by predicting Y_i from $Z_i := (X_i, W_i)$ on all the data.
- Report $\hat{\tau}(x) = \hat{\mu}((x, 1)) - \hat{\mu}((x, 0))$.

They were bad: Regularization Bias



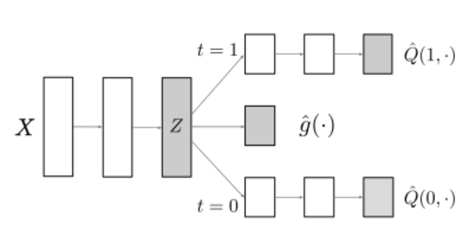
- Differential shrinkage across treatment levels leads to ‘hallucinated’ heterogeneity
- Problem is generic for any regression learner. Need some kind of ‘joint’ modelling for potential outcomes.

Dragonnet, Tarnet, etc.

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \widehat{R}(\theta; \mathbf{X}) \text{ where}$$

$$\widehat{R}(\theta; \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} ((Q^{nn}(w_i, \mathbf{X}_i, \theta) - y_i)^2 +$$

$$\alpha \text{CrossEntropy}(g^{nn}(\mathbf{X}_i; \theta), w_i))$$



https://arxiv.org/pdf/1906.02120.pdf

Sidestepping Regularisation Bias: X, R Learners

X-Learner

- Fit $\hat{\mu}^{(0)}(x), \hat{\mu}^{(1)}(x)$ using nonparametric regression
- Define pseudo-effects $\widetilde{D}_i^1 := Y_i - \widehat{\mu}^{(0)}(\mathbf{X}_i)$ and use them to fit $\widehat{\tau}^1(\mathbf{X}_i)$ on $\{i : W_i = 1\}$
- Define pseudo-effects $\widetilde{D}_i^0 := \widehat{\mu}^{(1)}(\mathbf{X}_i) - Y_i$ and use them to fit $\widehat{\tau}^0(\mathbf{X}_i)$ on $\{i : W_i = 0\}$
- Aggregate them as $\widehat{\tau}(x) = (1 - \widehat{\pi}(x))\widehat{\tau}^1(\mathbf{x}) + \widehat{\pi}(x)\widehat{\tau}^0(\mathbf{x})$

https://arxiv.org/abs/1706.03461

R-Learner

- Minimise Robinson (R) Loss

$$\widehat{\tau} = \underset{\tau}{\text{argmin}} \left\{ \widehat{L}_n(\tau(\cdot)) + \Lambda_n(\tau(\cdot)) \right\}$$

$$\widehat{L}(\tau(\cdot)) = \frac{1}{n} \sum_{i=1}^{n} ((Y_i - \widehat{\mu}(\mathbf{X}_i)) - (W_i - \widehat{\pi}(\mathbf{X}_i))\, \tau(\mathbf{X}_i))^2$$

- IOW, Regress pseudo outcome $\frac{Y - \mu(\mathbf{X})}{W - \widehat{\pi}(X)}$ on covariates $\psi(\mathbf{X}_i)$
- weights $(W - \widehat{\pi}(\mathbf{X}))^2$

https://arxiv.org/abs/1712.04912

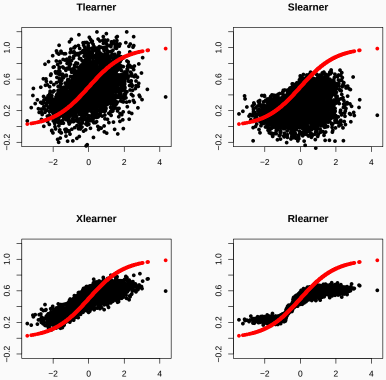
DR-Learner

- Construct pseudo-outcomes $\widehat{\varphi}(Z) := \widehat{\Gamma}_i^1 - \widehat{\Gamma}_i^0$ using AIPW score function
- Regress it on covariates $\psi(\mathbf{X}_i)$

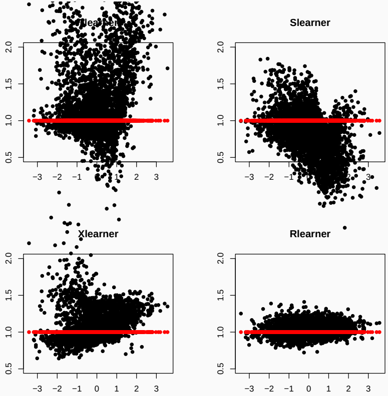
https://arxiv.org/abs/2004.14497

- Simulation + Implementation

Experiment



Observational



Summary of Generic Approaches [Knaus et al 2021]

Table 1. Summary of generic approaches to estimate IATEs.

Approach	w_i	Y_i^*
MOM IPW	1	$Y_{i,IPW}^*$
MOM DR	1	$Y_{i,DR}^*$
MCM	$T_i \frac{D_i - p(X_i)}{4p(X_i)(1-p(X_i))}$	$2T_iY_i$
MCM with EA	$T_i \frac{D_i - p(X_i)}{4p(X_i)(1-p(X_i))}$	$2T_i(Y_i - \mu(X_i))$
Orthogonal Learning	$(D_i - p(X_i))^2$	$\frac{Y_i - \mu(X_i)}{D_i - p(X_i)}$

https://arxiv.org/abs/1810.13237

- $D_i = W_i \in \{0, 1\}$
- $T_i = 2D_i - 1 \in \{-1, 1\}$
- $Y^*_{\text{IPW}} = \frac{W_i - \pi(\mathbf{X}_i)}{\pi(\mathbf{X}_i)(1-\pi(\mathbf{X}_i))}$
- $Y^*_{DR} = \widehat{\Gamma}_i^1 - \widehat{\Gamma}_i^0$
- All problems solve weighted least squares

$$\min_{\tau} \left(\frac{1}{n} \sum_{i=1}^{n} w_i (Y_i^* - \tau(\mathbf{X}_i))^2 \right)$$

Evaluating HTE Estimators

Stratification

- Since Het-FX estimators produce estimates of $\widehat{\tau}_i$, a gut-check for how well this works is to then stratify on $\widehat{\tau}_i$ (say, J bins), and compute $\widehat{\text{ATE}}^j$ in each bin using say AIPW
- If $\widehat{\text{ATE}}^j$s are sorted along their bin indices, this increases confidence that $\widehat{\tau}_i$s aren't all noise

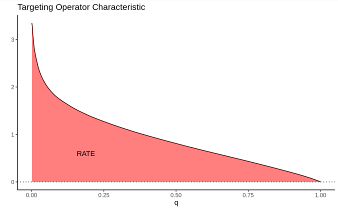
Best linear predictor method

- Create synthetic predictors $C_i = \overline{\tau}(W_i - \widehat{\pi}^{-i}(\mathbf{X}_i))$ and $D = (\widehat{\tau}^{-i}(\mathbf{X}_i) - \overline{\tau})(W_i - \widehat{\pi}(\mathbf{X}_i))$
- Regress $Y_i - \widehat{\mu}^{-i}(\mathbf{X}_i) \sim \alpha C_i + \beta D_i$
- $\alpha \approx 1$ indicates quality of ATE
- $\beta \approx 1$ indicates quality of CATE estimates (p.value is an omnibus test of heterogeneity fit by $\widehat{\tau}_i$)

- https://datascience.quantecon.org/applications/heterogeneity.html
- https://grf-labs.github.io/grf/articles/diagnostics.html

Rank Average Treatment Effects (RATE)

- Define a targeting rule $S(\mathbf{X}_i)$ which may be based on $\widehat{\tau}$s, risk scores, costs (typical S is simply τ_i)
- Define the Targeting Operator Characteristic (TOC) given distribution $\mathbb{F}(S(\mathbf{X}_i))$ and $q \in (0, 1]$



$$\text{TOC} = \mathbb{E}\left[Y_i^1 - Y_i^0 | S(\mathbf{X}_i) \geq \mathbb{F}^{-1}_{S(\mathbf{X}_i)}(1-q)\right] - \mathbb{E}\left[Y_i^1 - Y_i^0\right]$$

- This is largest for small qs and decays down to the ATE. If RATE ≈ 0, not much gain from prioritisation

https://grf-labs.github.io/grf/articles/rate.html

Summary

- Model-free estimation of ATE and friends largely settled : DML
- In contrast, CATE estimation is a very active area of research
- No silver bullets; good estimators typically depend on substantive knowledge of DGP [Smooth v sparse, etc]
  - prefer estimators that don't bake in function form (e.g. X,R,DR)
- Also prefer estimators that account for confounding (even in RCTs) because of incidental imbalance
- What to do with estimates? Optimal assignment policy learning, AUTOC, etc.