Causal Inference Crash Course

Heterogeneous Treatment Effect Models and Inference

Julian Hsu

Overview

- Use cases for heterogeneous treatment effects (HTE) models.
- Additional challenges compared to non-HTE models
- Showcase a few families of HTE models:
    1. Flexible learners and "filtering"
    2. OLS and DML
    3. ML-driven weights: GRF, Neural Nets, and others

    - An on-going theme will be the benefits of making structural or functional form assumptions
    - We will only cover cross-sectional models here for simplicity
- Conclude with fun extensions and a lot of citations.

When do we care about hetergeneous treatment effects (HTE)?

- We can use data to recommend a universal policy:
    1. What's the federal minimum wage?
    2. Product return policy
    3. Product pricing
    - These are not good use cases for heterogeneous treatment effects (HTE)
- We also want to make targeted policies or take customized actions:

1. Which customers should be defaulted to faster delivery options?
2. How do we match sellers with the best support or representatives?
3. Which users see which sort of ads?
4. Which orders should we scrutize and delay for fraud investigation?

- These are use cases for HTE.

HTE use cases

- Just like non-HTE use cases, like Average Treatment Effect (ATE) or Average Treatment Effect on the Treated (ATET), HTE remains a *causal question*.
- Like all causal questions, the foundational problem is that we not observe the outcomes under both the treatment and control conditions.
- This means it inherits all the causal complexities from its ATE/ATET cousins, and more.
    - Cross-sectional: overlap, unconfoundedness, etc.

    - Panel: parallel/overlaping trends
    - Your ears perk up when you can do a randomized experiment
- You should **always** first estimate ATE/ATETs before trying HTE

HTE modeling and notation

- Suppose you have a dataset with:
  - Y_i, outcome
  - X_i, features or confounders
  - Z_i, features you want to explore heterogeneity in
  - W_i, treatment indicator
- When we think about ATE/ATET, we can estimate by taking the difference between the outcomes under treatment and control:

$$\tau = E[Y_1(X_i) - Y_0(X_i)]$$

- Recall that the causal problem is that we only observe Y_1 for treated ($W_i = 1$) units or Y_0 for control ($W_i = 0$) units, not both.
- Therefore I am controlling for X_i. Again, this doesn't guarantee you have the correct estimate, but is common practice when assumptions hold.

- For HTE, we are interested in variation across Z_i:

$$\tau^{hte}(Z_i) = E[Y_1(X_i) - Y_0(X_i)|Z_i]$$

HTE interpretations

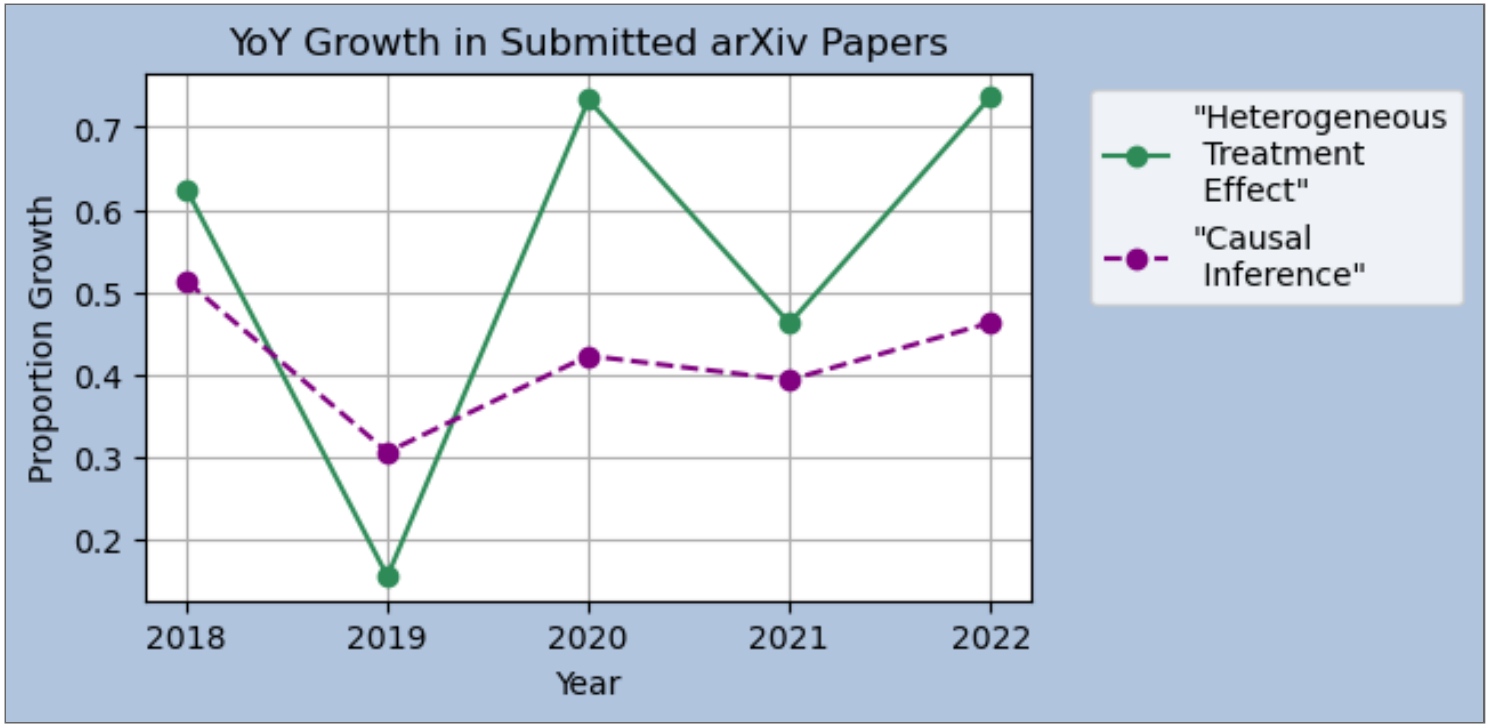
$$\tau^{hte}(Z_i) = E[Y_1 - Y_0|Z_i]$$

- We can also call this the conditional average treatment effect (CATE)
- Since our estimate is no longer a single scalar number, but a function that inputs Z_i, we are now estimating an HTE function.
- Your use case can allow different interpretations.
    - "How being offered a five-day versus two-day delivery options impacts different customers."

- "How the average seller reacts to being treated with different services."

Some HTE Models

This is a very active literative, so consider this just a brief summary of broad classes of HTE models



Class	categorical treatment	continuous treatment	cross-sectional data	panel data
T+X (T2X) Learner	Y	N	Y	Sometimes
OLS	Y	Y	Y	Y
ML-Weights	Y	N	Y	Sometimes

Common ingredients across models

- Estimated counterfactual outcome, $\hat{Y}_1(X_i), \hat{Y}_0(X_i)$
    - How would treated units perform if they were control units instead, and vice versa?
- Estimated outcome, $\hat{Y}(X_i)$
- Propensity score, $\hat{P}(X_i)$
    - Probability that a given unit is treated.
    - We will rely on this with the unconfoundedness/exogeneity assumption common in causal models

T+X Learners - the big idea

$$\tau^{hte}(Z_i) = E[Y_1(X_i) - Y_0(X_i)|Z_i]$$

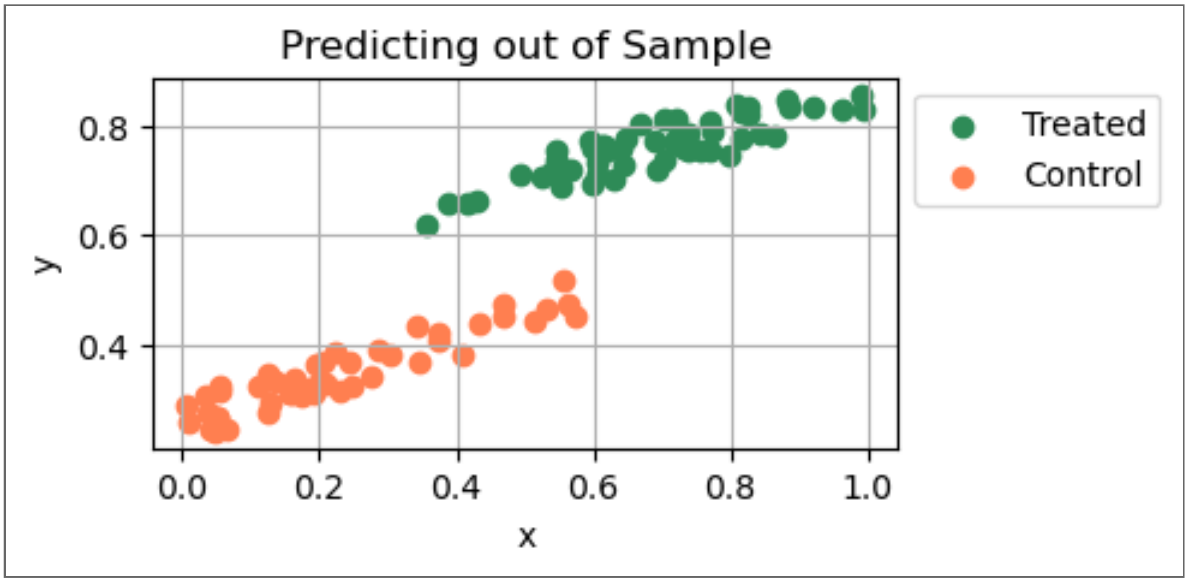
- Let's treat it as a prediction problem ("T-Learner").
  - For each observation, predict $Y_1(X_i)$ and $Y_0(X_i)$ values
  - Use your favorite ML model to train two models 1: $Y_1(X_i)$; 2: $Y_0(X_i)$.
- Calculate observation-level differences:

  $\tau_i^{hte} = \hat{Y}_1(X_i) - \hat{Y}_0(X_i)$
- We look at variation in τ_i^{hte} across Z_i.

- We can train a third prediction model of τ_i^{hte} as a function of Z_i for interpretability and reduce noise in τ_i^{hte}. We will come back to this in a few slides.

T+X Learners - one question

1. In the figure below, how are we sure that our $\hat{Y}_1(X_i)$ model would do a good job predicting the outcome of the control units?
    - We are extrapolating into an unknown space. We are training a model on treated units to predict the outcome of control units, so we do not have a ground truth to validate our $\hat{Y}_1(X_i)$ model.


Predicting out of Sample
Treated
Control
0.8
0.6
0.4
y
0.0
0.2
0.4
0.6
0.8
1.0
x

T+X Learners - relying on propensity scores

- We can rely on additional information, the propensity score $\hat{P}(X_i)$
- Relying on the unconfoundedness assumption, we can compare treated and control units with similar propensity scores to have the correct estimate.
- A common way of doing this would to take a weighted average of the estimate for control and treated units ("X-Learner").
- Künzel et al. (2017) propose taking the weighted average after predicting variation in $\hat{\tau}_i^{hte}$ across Z_i.

We will revisit this in the next slide.

$$\hat{\tau}^{hte,x} = \hat{P}(X_i)\hat{\tau}_i^0 + (1 - \hat{P}(X_i))\hat{\tau}_i^1$$

- Where $\hat{\tau}_i^0$ ($\hat{\tau}_i^1$) is the impact for control (treated) units.

T+X Learners - a second question

1. Independent of the above question, how do we know that variation across Z_i is real or due to noise?
    - This is a big question we should ask of *every HTE model* and is an HTE-specific complexity.
- Chernozhukov et al. (2017) propose using the propensity score and variation over Z_i in a single equation, where you estimate how much variation in $\hat{\tau}_i^{hte}$ is driven by selected Z_i while simultaneously controlling for the propensity score.

- Kennedy (2020) propose estimating $\hat{\tau}_i^{hte}$ using $\hat{Y}_1(X_i)$, $\hat{Y}_0(X_i)$, and $\hat{P}(X_i)$ in a single equation, and then estimating variation across Z_i.

OLS

- Where T2X Learner HTE models allow a lot of flexibility, we see that leveraging additional information, specifically the propensity score yields improvements.
- We can simultaneously leverage both predictions of the outcome and propensity score with a common causal model, ordinary least squares (OLS)
- We will improve on its framework with a double-debiased machine learning (DML) approach

OLS - simple model

- Recall that under the same assumptions as before, we can estimate the ATE/ATET with an OLS model:

$$Y_i = \beta X_i + \hat{\tau} W_i + \epsilon_i$$

- We can incorporate HTE by including additional features

$$Y_i = \beta X_i + \hat{\tau} W_i + \hat{\tau}^{hte,z} W_i \times Z_i + \epsilon_i$$

- Therefore, the HTE is a baseline treatment is a combination fo the baseline treatment $\hat{\tau}$ with $\hat{\tau}^{hte,z} Z_i$.

$$\hat{\tau}^{\text{hte}} = \hat{\tau} + \hat{\tau}^{\text{hte,z}} Z_i$$

OLS - drawbacks of the simple model

$$Y_i = \beta X_i + \hat{\tau} W_i + \hat{\tau}^{hte,z} W_i \times Z_i + \epsilon_i$$

- This approach yields unbiased estimates for $\hat{\tau}^{hte}$. Interpretation is also very straight forward. However, we can face difficulties when:
    1. The functions that determine Y_i or W_i cannot be well modeled linearly
    2. Z_i has high dimensionality; or
- We will solve both by incorporating approaches from double/debiased machine learning (DML) from Chernozhukov (2016).

OLS - DML applied to HTE, Part 1

- Semenova et al (2017) approach borrows the residualizing approach from DML, where we predict the observed outcome $\hat{Y}(X_i)$ and propensity score $\hat{P}(X_i)$.
    - This is the "first stage" in DML. We then run the "second stage":

$$\tilde{Y}_i = \hat{\tau}\tilde{W}_i + \hat{\tau}^{hte,z}\tilde{W}_i \times Z_i + \eta_i$$

    - Where the $\tilde{Y}_i$ and $\tilde{W}_i$ are the difference between the predicted and observed outcome and treatment status, respectively.

- If we run this "second stage" as is, then we can still have problems if Z_i is high-dimensional. We can think of this as a feature selection problem.
    - We can have high dimensionality over different transformations of a variable. For example: $Z_i = [x_{1i}, x_{1i}^2, x_{1i}^3, \log(x_{1i}), 1\{x_{1i} > 0\})]$.

OLS - DML applied to HTE, Part 2

- Semenova et al (2017) incorporates selection over Z_i by adapting a sample-splitted LASSO regression.
    - In general, LASSO regression coefficients do not have a causal interpretation.
    - We get around this by doing sample-splitting

$$\widetilde{Y}_i = \hat{\tau}\widetilde{W}_i + \hat{\tau}^{hte,z}\widetilde{W}_i \times Z_i + \eta_i$$

- We will split the sample into training and test samples. We select Z_i on the training set using

LASSO, and then estimate OLS on the selected Z_i on the test set.

- We then average "selected" OLS coefficients across test samples.

ML-Weights

- When we estimating differences between treatment and control units, we want to compare treatment and control units that are otherwise similar.
    - If we compare the outcomes of treatment/control that are similar on observable characteristics and we assume exogeneity, then the difference is causal.
- We can do matching based on propensity scores. For example below. However, we may run into

situations where a control unit can be a good match for multiple treated units (ie rows 1 and 3)

	Matched Set	Treatment Unit	Control Unit
1.	1	0.20	0.21
2.	1		0.19
3.	2	0.23	0.22
4.	2		0.24

Weighting Units

	Matched Set	Treatment Unit	Control Unit
1.	1	0.20	0.21
2.	1		0.19
3.	2	0.23	0.22
4.	2		0.24

- We can weight control units based on how close they are in propensity scores or other distributional assumptions.
- The more similar control and treated units are, the larger the weight should be.

- Athey et al. (2018) use a Generalized Random Forest to determine weights that:
    1. Compares similar treatment and control units; and
    2. Explores variation across Z_i

Causal Tree is a building block for Generalized Random Forest (GRF)

- A Generalized Random Forest (GRF) is a forest of Causal Tree (CT), rather than standard Decision Trees (DT).
- We'll distinguish how standard Random Forests and Causal Forests would estimate τ^{hte}.
- Both split the data sample based on having similar features X_i to predict τ^{hte}. Splits are evaluated based on the variation in τ^{hte} (ie entropy). This splitting

process has two components. For a given potential split of a node:

1. How good is this bifurcation compared to other potential splits?
2. What's the predicted τ^{hte} in each of the splits?

- DT uses the same data to answer 1. and 2. In contrast, CT uses different data for each. Think of this as having one training sample determines splits, and the other training sample estimates τ^{hte}.
    - You can think of 1. as a way to maximize prediction power, and 2. as an evaluation (for causal inference, estimating the effect).

- This allows us to calculate confidence intervals over CT and GRF's estimates. Note this is the same high-level approach as doing inference with Lasso regressions in the previous section.

GRF approach

- GRF calculates weights so that:
    1. Control and treatment units with similar X_i are compared to estimate $\hat{\tau}^{hte}$; and
    2. It maximizes variation in $\hat{\tau}^{hte}$.
- Note that this forces variation in $\hat{\tau}^{hte}$ across X_i, not necessarily Z_i. This is because the weighting is used to estimate the treatment effect based on matching and estimate heterogeneity.
- We can improve upon this by taking the residualization concept from DML. That is, you first

calculate $\widetilde{Y}_i$ and $\widetilde{W}_i$ and then train GRF to do variation across Z_i.

Papers

T2X Learners:

- Künzel, Sekhon, Bickel, Yu. *Meta-learners for estimating heterogeneous treatment effects using machine learning*
  http://arxiv.org/abs/1706.03461
- Semenova, Chernozhukov. *Debiased Machine Learning of Conditional Average Treatment Effects and Other Causal Functions*
  https://arxiv.org/abs/1702.06240

- Chernozhukov, Demirer, Duflo, Fernández-Val. *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments* **https://arxiv.org/abs/1712.04802**
- Kennedy. *Towards optimal doubly robust estimation of heterogeneous causal effects* **https://arxiv.org/abs/2004.14497**
- Sant'Anna, Zhao *Doubly Robust Difference-in-Differences Estimators* **https://arxiv.org/abs/1812.01723**

OLS:

- Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins. *Double/Debiased Machine Learning*

for Treatment and Causal Parameters
https://arxiv.org/abs/1608.00060

- Semenova, Goldman, Chernozhukov, Taddy. *Estimation and Inference on Heterogeneous Treatment Effects in High-Dimensional Dynamic Panels under Weak Dependence*
  https://arxiv.org/abs/1712.09988

ML-Weights

- Athey, Tibshirani, Wager. *Generalized Random Forests* **https://arxiv.org/abs/1610.01271**
- Friedberg, Tibshirani, Athey, Wager. *Local Linear Forests* **https://arxiv.org/abs/1807.11408**

- Wager, Athey. *Estimation and Inference of Heterogeneous Treatment Effects using Random Forests* **https://arxiv.org/abs/1510.04342**
- Farrell, Liang, Misra. *Deep Neural Networks for Estimation and Inference* **https://arxiv.org/abs/1809.09953**

In [ ]: