Causal Inference: Heterogeneous Treatment Effect Models

Hsu (**@hsujulia**)

Delivery Experience (DEX) Economics

You can find these slides here and accompanying notebook here.

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*
- Conclude with some recommendations and best practices.
- Other HTE presentations at Amazon:
  - *Common Mistakes in Estimating Treatment Effects: Heterogeneous and otherwise (@duncang / @yuwhsieh / @dizeng)*
  - *CATE Meta Learners (@shyurya)*

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

- Making universal policies are not good use cases:
  1. Product return policy
  2. Product pricing
- Making targeted policies or taking customized actions are good use cases:
  1. Which customers should be defaulted to faster delivery options?
  2. How do we match sellers with the best support or representatives?
  3. Which customers spend more when we expand ASIN selection?
  4. Which orders should we scrutize and delay for fraud investigation?

HTE use cases

- Just like non-HTE causal use cases, like Average Treatment Effect (ATE) or Average Treatment Effect on the Treated (ATET), HTE remains a *causal question*.
- The problem is that we not observe the individual outcomes under both the treatment and control conditions.
- HTE has all the causal complexities from its ATE/ATET cousins, and more.

HTE modeling and notation

- We can estimate ATE/ATET by taking the average difference between the outcomes under treatment and control:

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

- I am controlling for X_i, assuming that the unconfoundedness assumption holds.
- For HTE, we are interested in variation across Z_i:

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

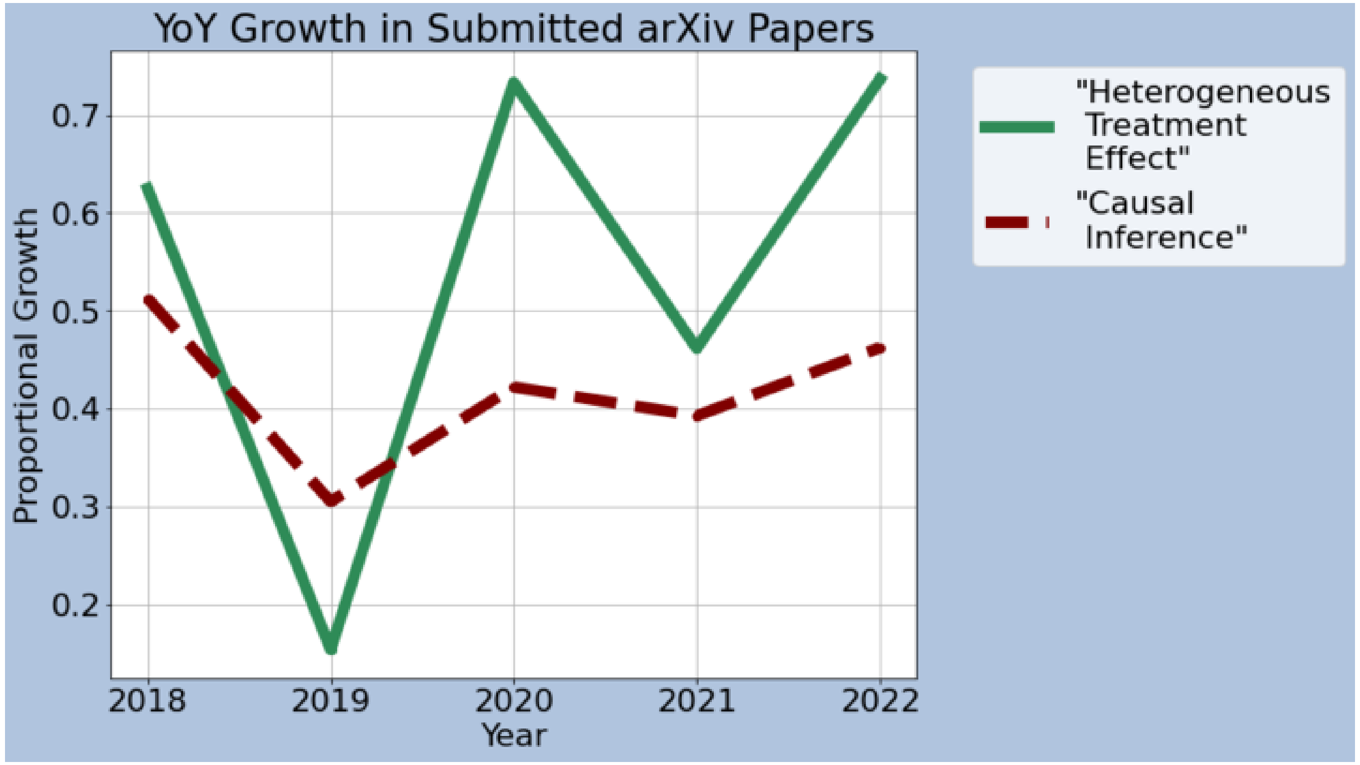
HTE interpretations

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

- We can also call this the conditional average treatment effect (CATE) or individualized treatment estimate (ITE).
- Since our estimate is no longer a single scalar number, but a function that inputs Z_i, we are now estimating a function.
- Z_i can be individual features, segment definitions, or other treatments.

Some HTE Models

This is a very active literative, so consider this a brief summary of broad classes of HTE 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: $Y_1(X_i)$; $Y_0(X_i)$.
- 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}.

T&X Learners - using propensity scores

- We can use propensity scores $\hat{P}(X_i)$ to remove noise from "T-Learners"
- 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").
- As an example, Künzel et al. (2017) propose taking the weighted average after predicting variation in $\hat{\tau}_i^{hte}$ across Z_i.

T&X Learners - other ways to use propensity scores

- 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.

- **Here at Amazon...**
  - Perfect Order Experience (POE) team uses Orthogonal Signal Smoothing to flexibility predict variation in HTE while maintaining interpretability (paper and wiki). (POCs: @dizeng, @chhrajul)
  - Ads Economics uses Gaussian processes to estimate counterfactual models (paper and poster), using them to measure the DSI of completing queries at the advertiser level (wiki and paper). (POC: @grezgerm)
  - Ads Economics estimates CATE ads incrementality by estimating
  - Probability of Treatment given Outcome (POT) and leverages Bayes rule
  - (paper) .(POC: @shyurya)

OLS

- Ordinary least squares (OLS) is a common causal models that we can also use to estimate HTE. It's simple approach does not require separately predicting outcomes or estimating propensity scores.
- OLS allows us to study continuous treatments as well.
- In low dimensional scenarios, OLS estimates unbiased HTE.
- In higher dimensional scenarios, we can use double-debiased machine learning (DML) approaches to estimate HTE.

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 linear combination fo the baseline treatment $\hat{\tau}$ with $\hat{\tau}^{hte,z} Z_i$.

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

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; or
    2. Z_i has high dimensionality.
- 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 residualized outcome and treatment statuses.
- 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: $Z_i = [z_{1i}, z_{1i}^2, z_{1i}^3, log(x_{1i}))]$.

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 with sample-splitting

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

- 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.

OLS and DML: Takeaways

- We can get HTE through OLS, where we interact the treatment indicator with Z_i.
- This works great under linearity assumptions and when Z_i is low dimensional.
- Otherwise, we can leverage concepts from DML. We residualize the outcome and treatment features to allow non-linearity in the components, and use a sample-splitted LASSO regression to select elements of Z_i.
- **Here at Amazon...**:
  - "Heterogeneous Residuals" model does more flexible residualization by treating HTE as a multiple treatment effects problem (paper / slides) (POC: @hsujulia)
  - HTE Generalized Additive Model (GAM) allows more flexible but interpretable transformations of Z_i (paper) (POC: @jinyangn)
  - DSI 3.0 uses Principal Component Analysis (PCA) to reduce dimensionality of Z_i (paper)

ML-Weights

- When we estimating treatment efects, we want to compare treatment and control units that are otherwise similar.
- Relying on standard unconfoundedness assumptions, we can match based on propensity scores. However, matching requires a lot of manual tuning and could overfit.
- Athey et al. (2018) use a Generalized Random Forest where each Causal Tree matches similar control and treatment units.

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

- A Generalized Random Forest (GRF) is a forest of Causal Trees (CT), rather than standard Decision Trees (DF).
- Both find similar treatment and control units X_i to predict τ^{hte}. Splits are evaluated based on the variation in τ^{hte} (ie entropy).
- DT uses the same data to evaluate the split and predict τ^{hte}.
- CT uses half the data to evaluate the split and the second half to predict τ^{hte}.
- This allows us to calculate confidence intervals over CT and GRF's estimates.

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 $\hat{\tau}^{hte}$.
- GRF also accommodates the residualization concept from DML. You first calculate $\tilde{Y}_i$ and $\tilde{W}_i$ and then train GRF to do variation across Z_i.
- **Here at Amazon...**:
    - Localized DSI through the Pareto tool uses K-means clustering with kernel methods to form clusters rather than a forest approach (paper / website). (POC: @ywenting)
    - CIV estimates HTE using a similar forest approach (CIV Review in 2020 Doc). (POC: @sufangli)

Other HTE Applications across Amazon

- We can apply HTE to DSE-style surrogate models to predict long-term impacts. HTE DSE Models paper (POC: @yaxic)
- HTE models can also be applied to panel models. CIV application (POC: @gitalo)

Guidelines for Estimating HTE

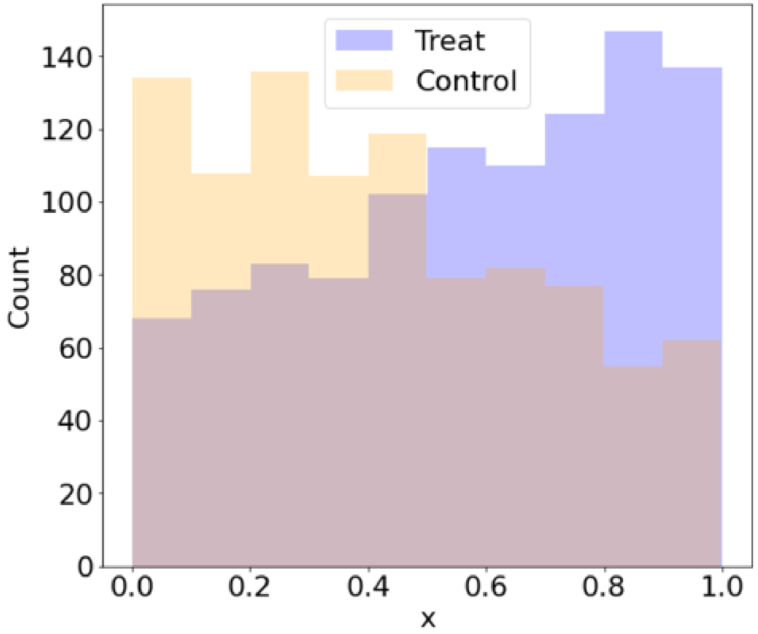
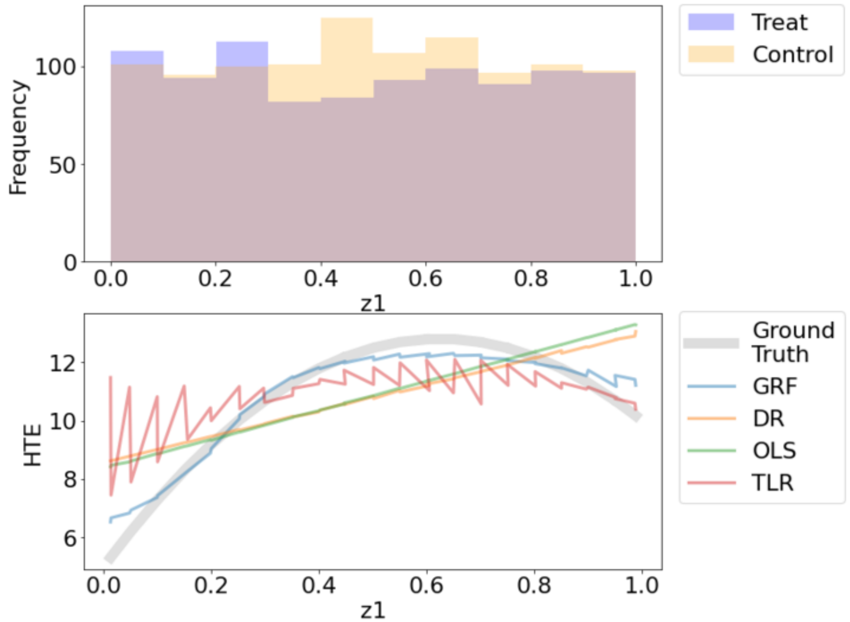
- We will use simulation data where we know exactly what $\hat{\tau}^{hte}(Z_i)$ looks like to make two points:

1. Like any causal problem, make sure you have overlap in propensity scores; and
2. Make sure you have overlap in features Z_i that drive heterogeneity.

Simple Simulation Setup

- We will have only two input features:
  - x is an input of the propensity score and outcome;
  - z is the only input for HTE.
- Often we will consider x, z to affect the propensity score, outcomes, and HTE. I separate them for illustrative purposes.
- We will study four models:
  - Generalized Random Forests (GRF)
  - Doubly Robust (DR)
  - Ordinary Least Squares (OLS)
  - T-Learner (TLR)

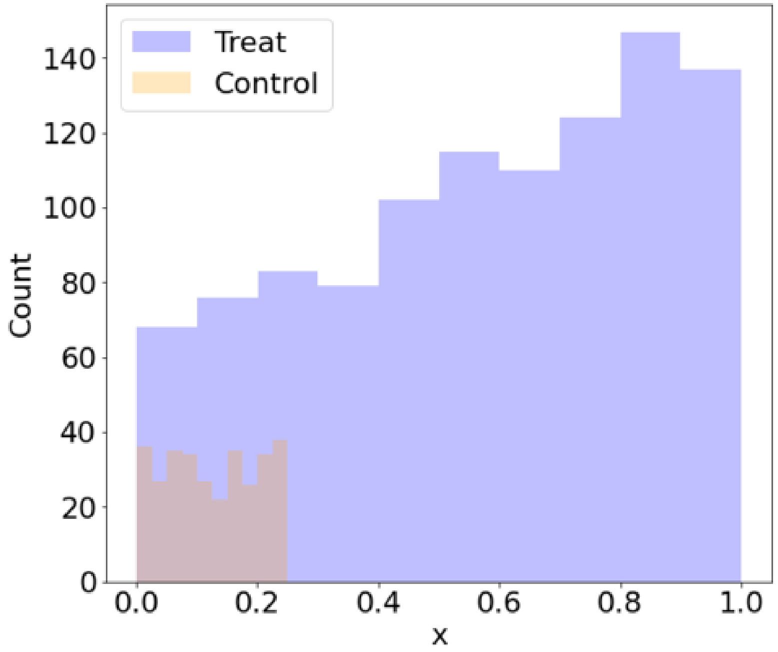
Baseline Simulation: Nothing is Wrong

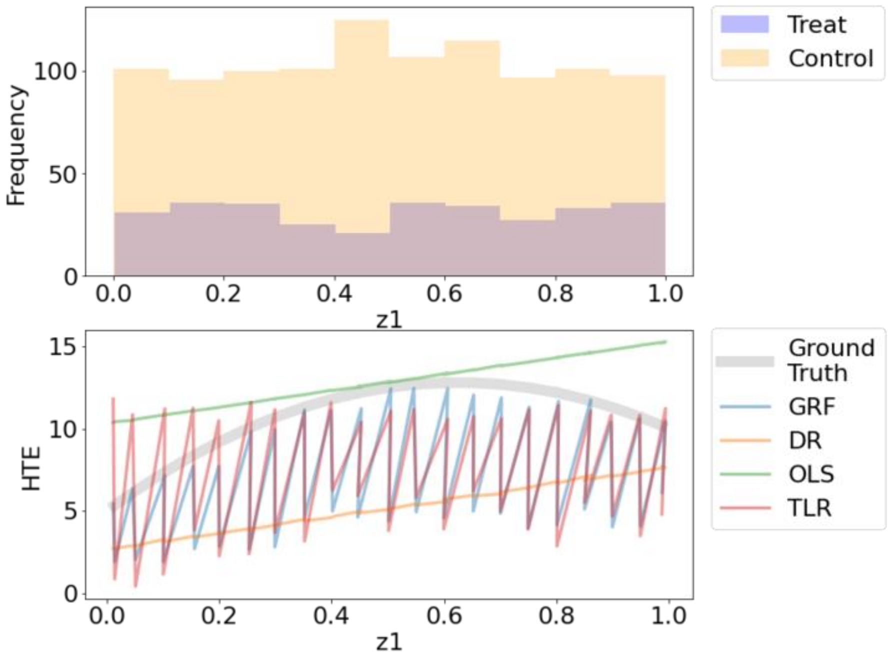
Propensity Score Overlap	HTE
	

- Note that even in the ideal setting, TLR has noisy HTE estimates.
- We are only entering z linearly into the models, and GRF is only one able to pick up on the non-linearity. We can achieve similar results by including z^2 as an input for the model.

No Propensity Overlap

Propensity Score Overlap   HTE

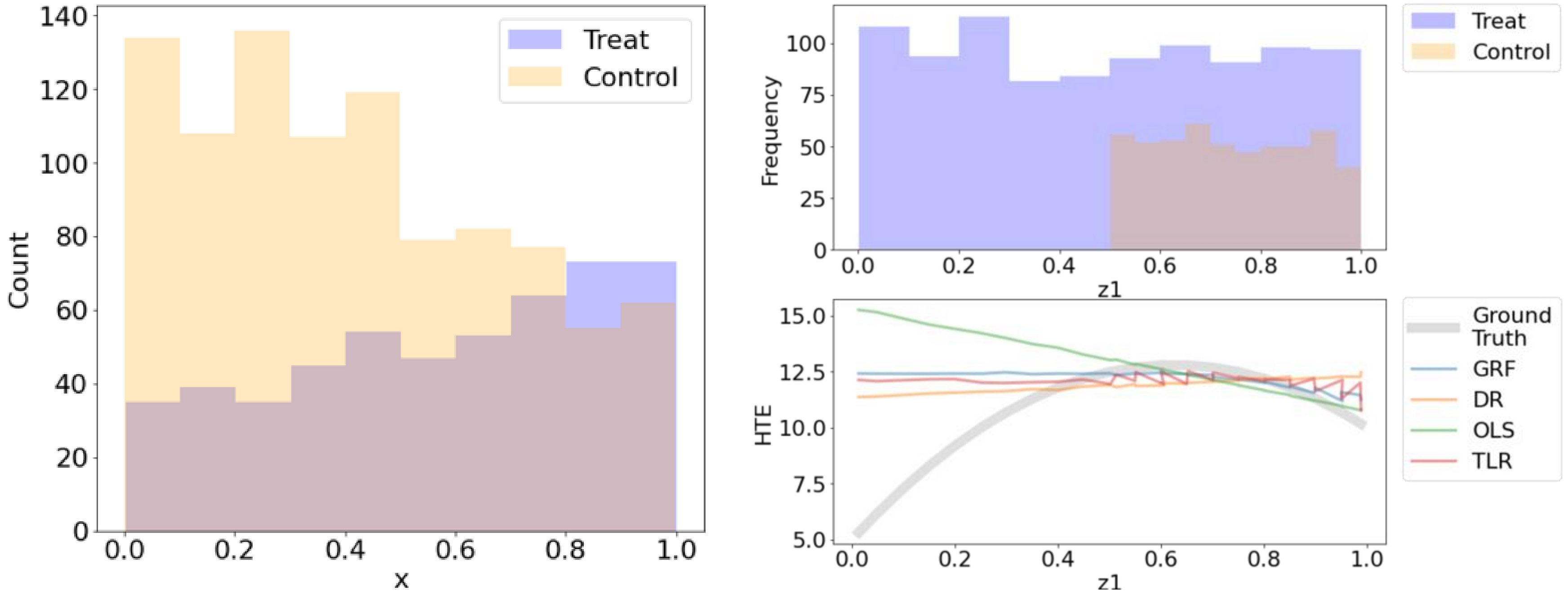




- Now all models are under performing

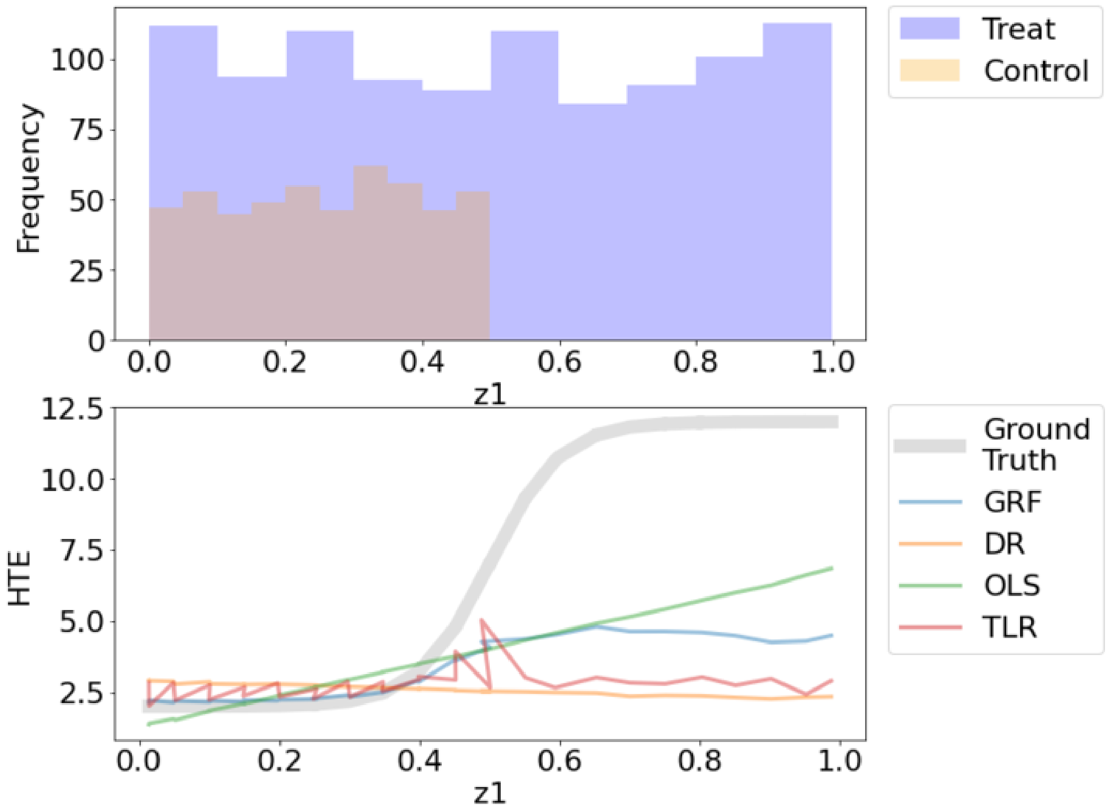
No HTE Overlap

Propensity Score Overlap HTE

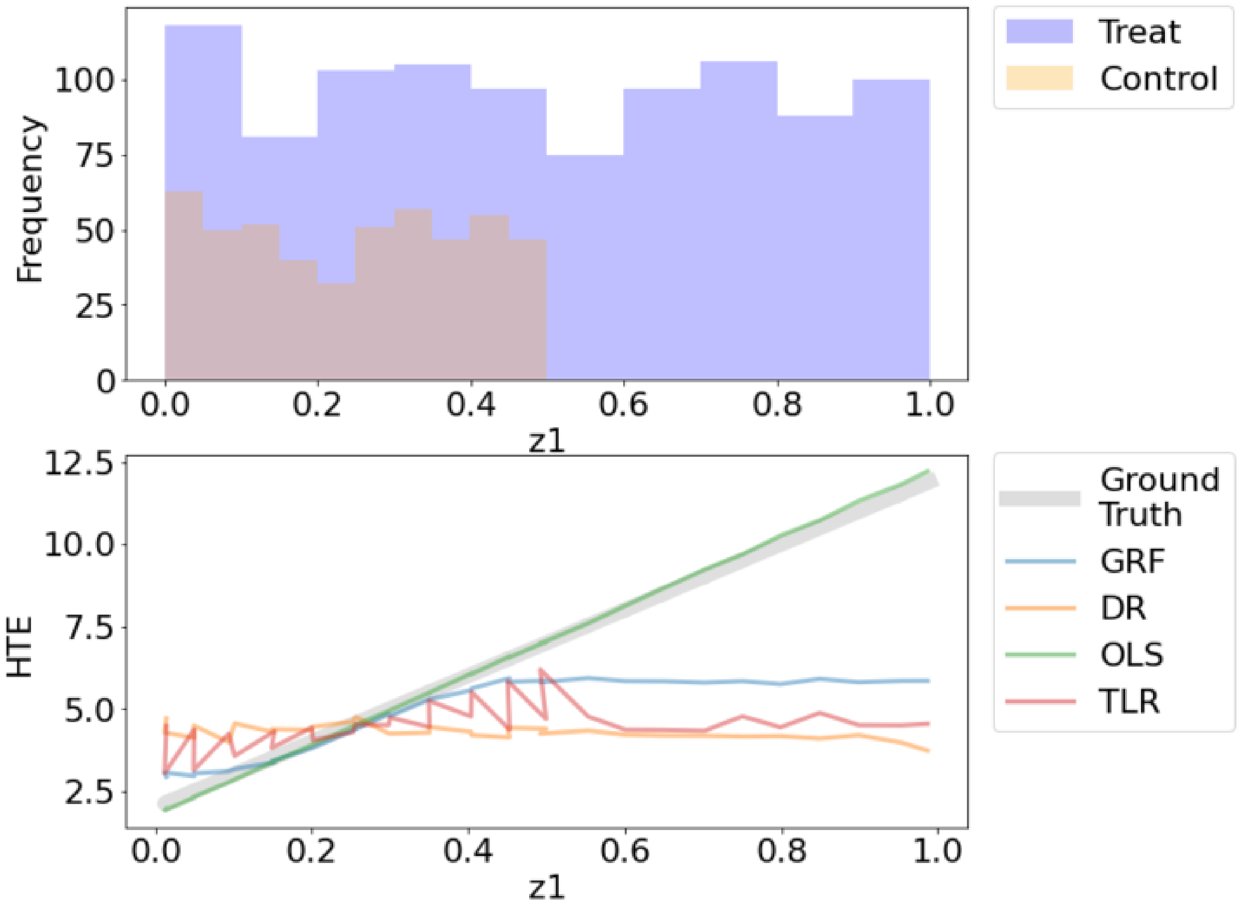


- Even though we have propensity score overlap, we only have overlap in z for high values of z.
- All models are extrapolating an HTE model estimated for high values of z to low values of z.
- This includes the GRF model which could pick up on non-linearity when there was overlap.

No HTE overlap means you are extrapolating



HTE is a Logit Function



HTE is a Linear Function

Conclusion

- Showed that HTE problems are still causal problems, where we do not observe ground truth.
- Focusing on individual ground truth makes it even harder.
- We should always make sure we have a valid design to estimating the average causal estimate before looking for HTE.
- HTE also requires sufficient coverage in dimensions of heterogeneity.
- Briefly overviewed some HTE models: T-Learners, X-Learners, OLS, DML, and GRF models.

Thank You For Viewing!

& Happy Model Running ☺

Papers - T&X 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

Papers - 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 [ ]:
```