Causal Inference Crash Course

Heterogeneous Treatment Effect Models and Inference

Julian Hsu

Causal Inference Series

1. Foundations
2. Defining some Causal Models
3. Inference, Asymptotic Theory, and Bootstrapping
4. Best Practices: Outliers, Feature Selection, and Bad Control
5. Heterogeneous Treatment Effect Models
6. Arguable Validation
7. Panel Data
8. Regression Discontinuity Models

Overview

- This presentation covers the general problem of estimating heterogeneous treatment effects (HTE) and how it differs from ATE/ATET estimation.
- Covers a few models:
    - Linear models via Double Machine Learning
    - Causal Forests / Local Linear Forests
    - Doubly Robust models following Kennedy (2020)
- Wrap up with a simulation demonstration
- Mentions about extensions to panel models

HTE Overview

- "What's the effect of faster shipping?" Average treatment effect (ATE) and average treatment effect on the treated (ATET) models want to know aggregate treatment effects.
- "What's the effect of faster shipping on customer in rural areas? / Which customers benefit the most from faster shipping?" HTE model want to estimate the distribution of treatment effects or for a specific subset or individuals

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

- $\hat{\tau}(Z_i)$ is the HTE and it varies of Z_i. We keep X_i different from Z_i for more flexible notation.
- We can also call this the conditional average tratment effect

HTE as an estimated function

- We want to estimate the functional form of HTE.
- When estimating ATE/ATET, we are only concerned with the average.
    - We average over more granular treatment effects.
- Estimating more granular treatment effects means there are additional challenges.

How much variation do we want in our HTE function?

- Two extremes:

1. Individualized treatment estimates allow more flexibility, but can demand large sample sizes and variation in data.

- Increases the risk of noise driving estimates

1. Segmented estimates are the least flexible, with the least risk of noise driving estimates.

- They can also be more interpretable for other applciations. • In-between case is to allow treatment effects to vary across some dimensions, but not others

HTE ideal experiment

- We can understand these two extreme based on what the ideal experiment is to estimate unbiased HTE.
- For individualized HTE, the ideal is to randomize treatment for each individual. (impossible)
- For segmented HTE, the ideal is to randomize treatment for each segment. (stratified randomization)
- The more individualized HTE is, the more data and assumptions are needed to distinguish between real patterns and statistical noise in the data.

HTE inference challenge

- Statistical inference for ATE/ATET estimates is based on the distribution of error around the average estimate.
- The challenge is getting a distribution around an individual estimate.
- The solution is to rely on either model specifications or bootstrapping-esque methods.

Some HTE Models

Support across use cases

	Cross Sectional Data	Panel Data	Continuous Treatment
DML	Y	Y	Y
Generalized Random Forests	Y	N	N
Doubly Robust	Y	N	N

DML-Style models

- Semenova, Goldman, Chernozhukov, Taddy (2021) - SGCT
- Let's start with linearity assumptions, which allows interpretability:

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

- SGCT decomposes $\hat{\tau}(Z_i)$ into a functionl form:

$$\hat{\tau}(Z_i) \rightarrow \hat{\tau} \cdot g(Z_i)$$

- where $g(Z_i)$ is different functions of Z_i. For example:

$$\hat{\tau} \cdot g(Z_i) = \hat{\tau}_0 + \hat{\tau}_1 z_{1i} + \hat{\tau}_2 z_{1i}^2$$

- Continuing this example, the model would be:

$$Y_i = \hat{\beta} X_i + \hat{\tau}_0 W_i + \hat{\tau}_1 z_{1i} W_i + \hat{\tau}_2 z_{1i}^2 W_i + \epsilon_i$$

SGCT users residualization

- Now how do we estimate this equation?

$$Y_i = \hat{\beta} X_i + \hat{\tau}_0 + \hat{\tau}_1 z_{1i} + \hat{\tau}_2 z_{1i}^2 + \epsilon_i$$

- At first glance we can just do OLS, but we can improve that approach with double machine learning (DML; aka residualization).
    - Recall DML works through the Frisch-Waugh-Lovell theorem
- SGCT estimates this equation

$$\tilde{Y}_i = \hat{\tau}_0 \tilde{W}_i + \hat{\tau}_1 z_{1i} \tilde{W}_i + \hat{\tau}_2 z_{1i}^2 \tilde{W}_i + \epsilon_i$$

- Where $\tilde{Y}_i$ and $\tilde{W}_i$ are the residualized outcome and treatment

SGCT – HTE and inference

- We now need to do inference for individual treatment effects from

$$\tilde{Y}_i = \hat{\tau}_0 \tilde{W}_i + \hat{\tau}_1 z_{1i} \tilde{W}_i + \hat{\tau}_2 z_{1i}^2 \tilde{W}_i + \epsilon_i$$

- HTE is $\hat{\tau}_1 z_{1i} + \hat{\tau}_2 z_{1i}^2$, where the standard error is calculated via the Delta method.
- We can use OLS to estimate the above equation if:
    - • There are few dimensions of heterogeneity (ie $g(Z_i)$ is low dimensional); or
    - We are interested in specific dimensions of heterogeneity (ie we only want to know HTE across user tenure)

SGCT – inference with post-LASSO regression

- A problem is if we estimate with all possible transformations of z_{1i}. In other words, overfitting.
- We can select the relevant transformations of z_{1i} with LASSO regression, but then we cannot do inference.
- Get around this with a sample-splitted LASSO for inference. Select features with LASSO on one half of the dataset, and then estimate HTE using those selected features on the other half.

Generalized Random Forests

- Causal forests are a special class of generalized random forests, which we will discuss here.
- As motiviation, note that under the unconfoundedness assumption, we assume that treatment is random:

$$E[(Y - \hat{g}(X_i) - \hat{\tau}(Z_i)W_i)W_i] = 0$$

- In other words, $\hat{\tau}(Z_i)$ satisfy the orthogonality assumption similar to Frisch-Waugh-Lovell.
- The problem is that we do not know what $\hat{\tau}(Z_i)$ looks like, so we want a flexible specification. Ideally, something non-parametric.

GRF – Causal Forest Objective Function

- The estimating equation (with simplified notation is):

$$\left(\hat{\tau}(Z_i), \hat{g}(X_i)\right) = argmin\left\{E[\alpha_i(z)(Y_i - \hat{g}(X_i) - \hat{\tau}(Z_i)W_i | Z_i\right.$$

- What's the purpose of $\alpha_i(z)$?
- $\alpha_i(z)$ is a weight used to allow flexibility in $\hat{\tau}(Z_i)$.
  - Can be estimated to kernel methods but performance suffers under high dimensions
  - Estimate $\alpha_i(z)$ with a random forest to deal with high dimensionality of Z_i!
- This weight gives us enough variation in $\hat{\tau}(Z_i)$.

GRF - Weights $\alpha_i(z)$

- $\alpha_i(z)$ represents the probability that a training sample i falls into the same leaf as sample z, across different trees in a random forest
- Splits in the random forest used to estimate $\alpha_i(z)$ are determined to maximize variation $\hat{\tau}(Z_i)$ across splits

GRF - Inference

- Athey, Tibshirani, and Wager (2019) show that $\hat{\tau}(Z_i)$ is asymptotically normal.
- This is because $\alpha_i(z)$ is estimated in an "honest" (Athey and Wager, 2018) fashion, where different samples are used to determine splits in $\alpha_i(z)$ and $\hat{\tau}(Z_i)$.
- Standard errors and confidence intervals are available based on a bootstrap/jackknife approach.
- Intuitively, estimate the distribution in $\hat{\tau}(Z_i)$ when z is removed from the sample

Doubly Robust - Kennedy (2020)

- Recall the interactive regression model from DML:

$$\hat{\tau}_{ATE} = E\Big[(\hat{Y}_{1,i} - \hat{Y}_{0,i}) + \frac{W_i(Y_i - \hat{Y}_{1,i})}{\hat{P}_i} - \frac{(1 - W_i)(Y_i - \hat{Y}}{1 - \hat{P}_i}$$

- Recall that we can intuitively understand this as a individual-level comparison from a regression adjustment model, correcting for prediction errors.
- Removing the expectation, we can see that these are individual-level treatment effect estimates

$$(\hat{Y}_{1,i} - \hat{Y}_{0,i}) + \frac{W_i(Y_i - \hat{Y}_{1,i})}{\hat{P}_i} - \frac{(1 - W_i)(Y_i - \hat{Y}_{0,i})}{1 - \hat{P}_i}$$

Applying inference to the individual estimates

- The problems are that these estimates:

1. Are meant to be averaged to get the ATE/ATET; and
2. Do not have inference properties.

- Kennedy (2020) frames these as “noisy” estimates of the true HTE, and proposes “refining” them with a second stage.

$$hte_i = (\hat{Y}_{1,i} - \hat{Y}_{0,i}) + \frac{W_i(Y_i - \hat{Y}_{1,i})}{\hat{P}_i} - \frac{(1 - W_i)(Y_i - \hat{Y}_{0,i})}{1 - \hat{P}_i}$$

```
In [ ]:
```

```
In [ ]:
```

Simulation Study

Context

- Recall, HTE models are not about estimating the average effect, but rather the functional form of treatment effects.
- The additional complexity of functional form can make this very difficult.
- We will demonstrating using simulation evidence, where we can change the **true HTE function**

General Simulation Context

- For simplicity, there is only one feature \align \align

$$x \sim U[0, 1]$$

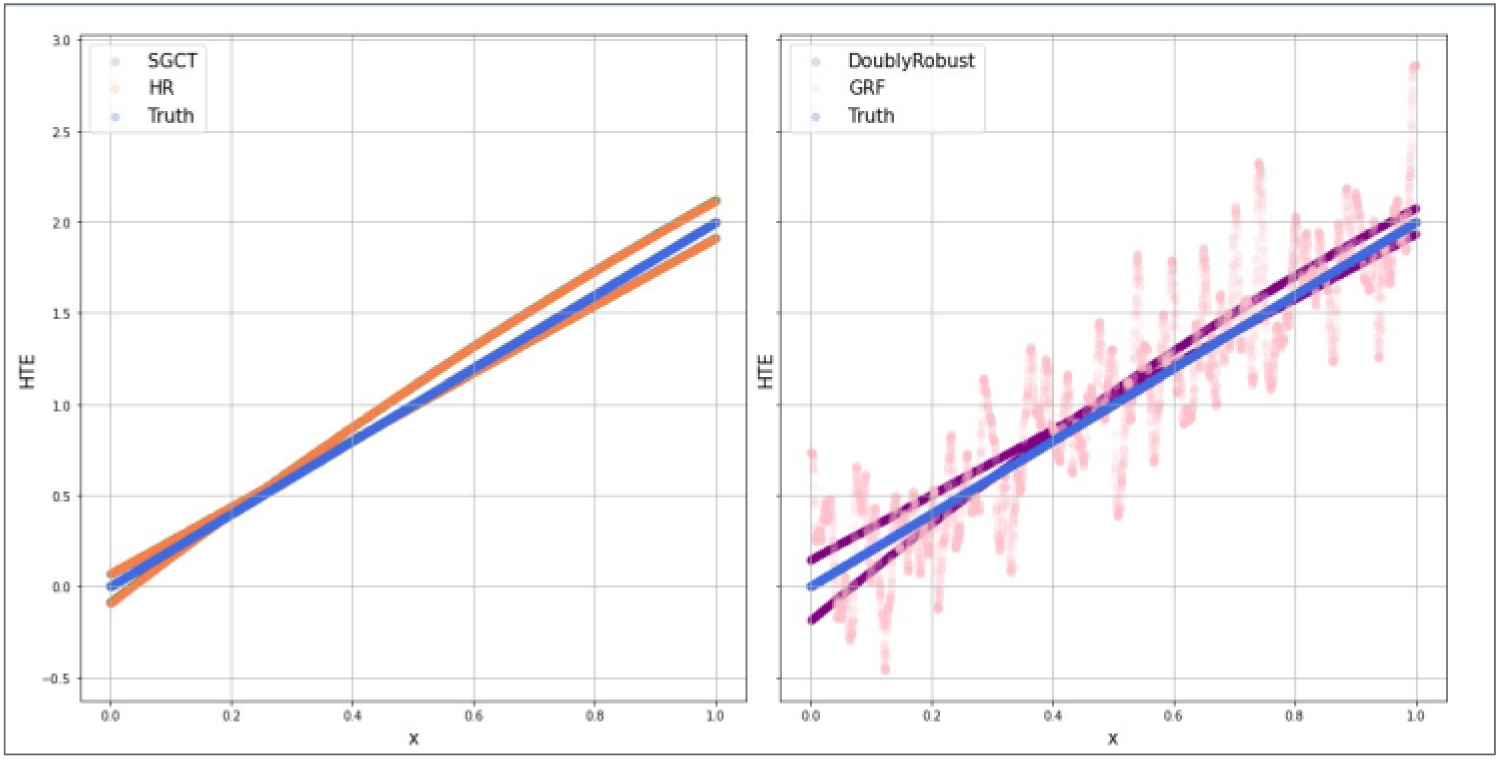
$$y = 10 + 2ln(1 + x) + \epsilon, \epsilon \sim N(0, 1)$$

$$W = 1\{\frac{exp(x)}{1 + exp(x)} + \eta > 0\}, \eta \sim N(0, 1)$$

- We want to know the HTE of W_i
- We show three examples, with different HTE functions

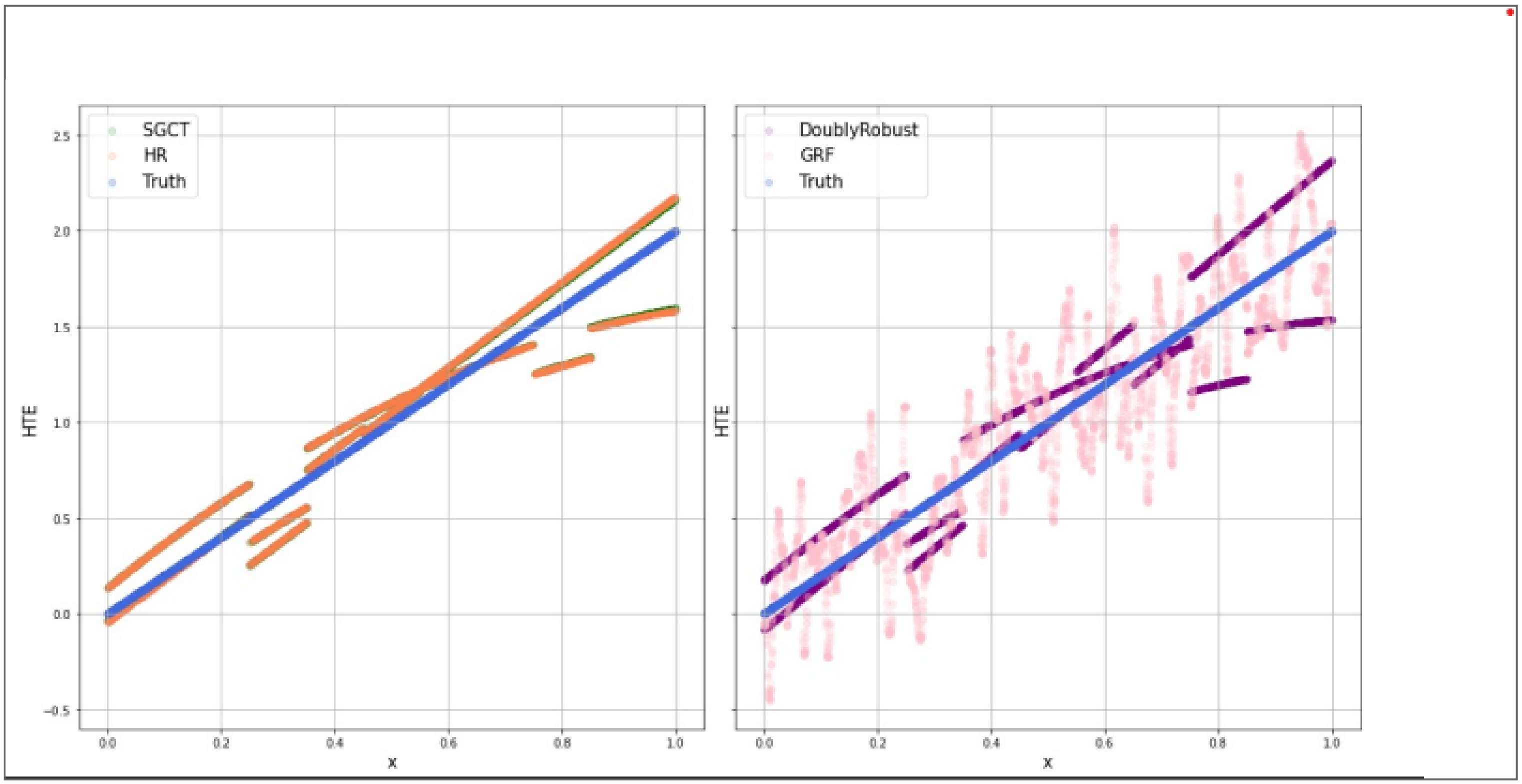
First Example - Linearity

- HTE $= 2x$
- Model controls: x, x^2



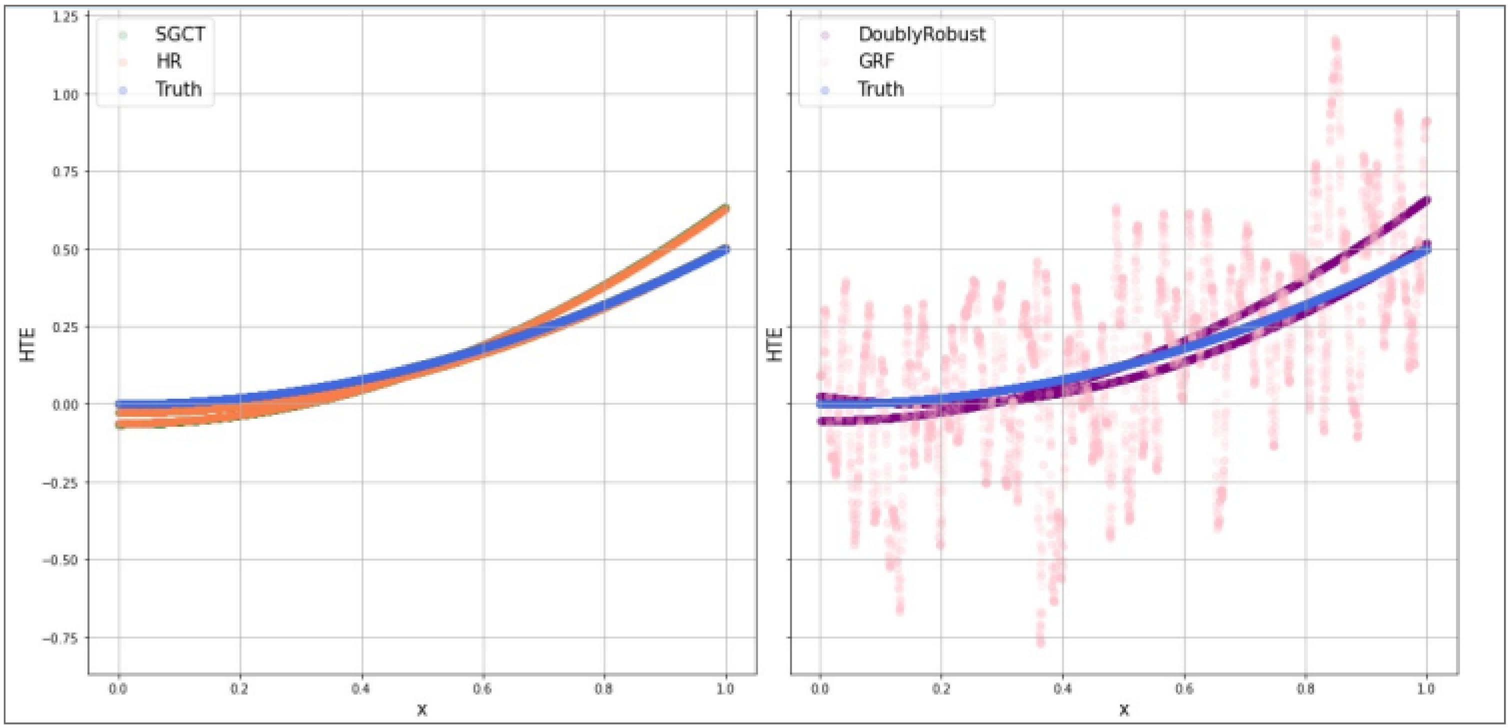
First Example - Linearity with more controls

- HTE $= 2x$
- Model controls: $x, x^2, 1\{0 \leq x > 0.2\}, \ldots, 1\{0.8 \leq x > 1\}$



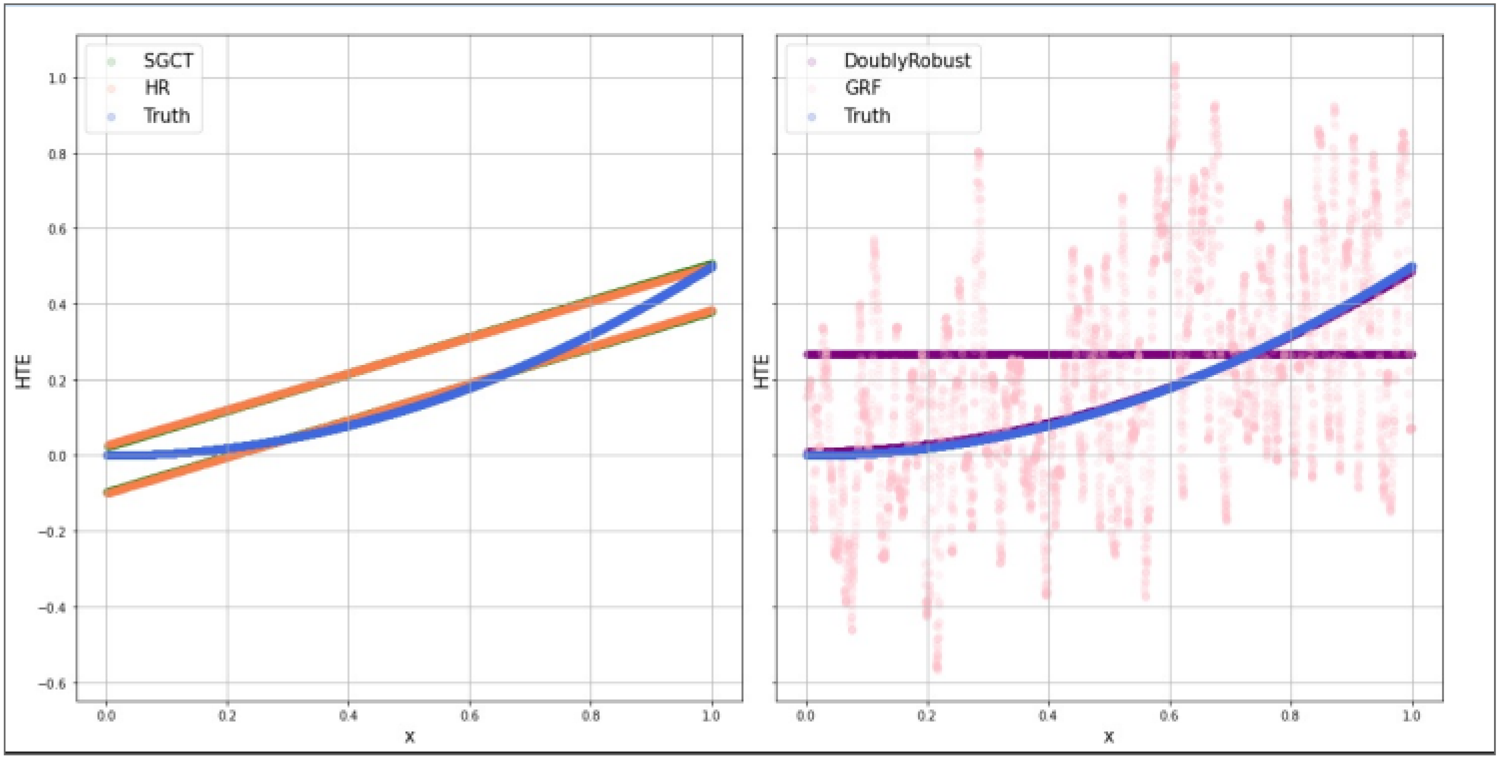
Second Example - Quadratics

- HTE $= 2x$
- Model controls: x, x^2



Second Example - Quadratics with more controls

- HTE $= 2x$
- Model controls: $x, x^2, 1\{0 \leq x > 0.2\}, \ldots, 1\{0.8 \leq x > 1\}$

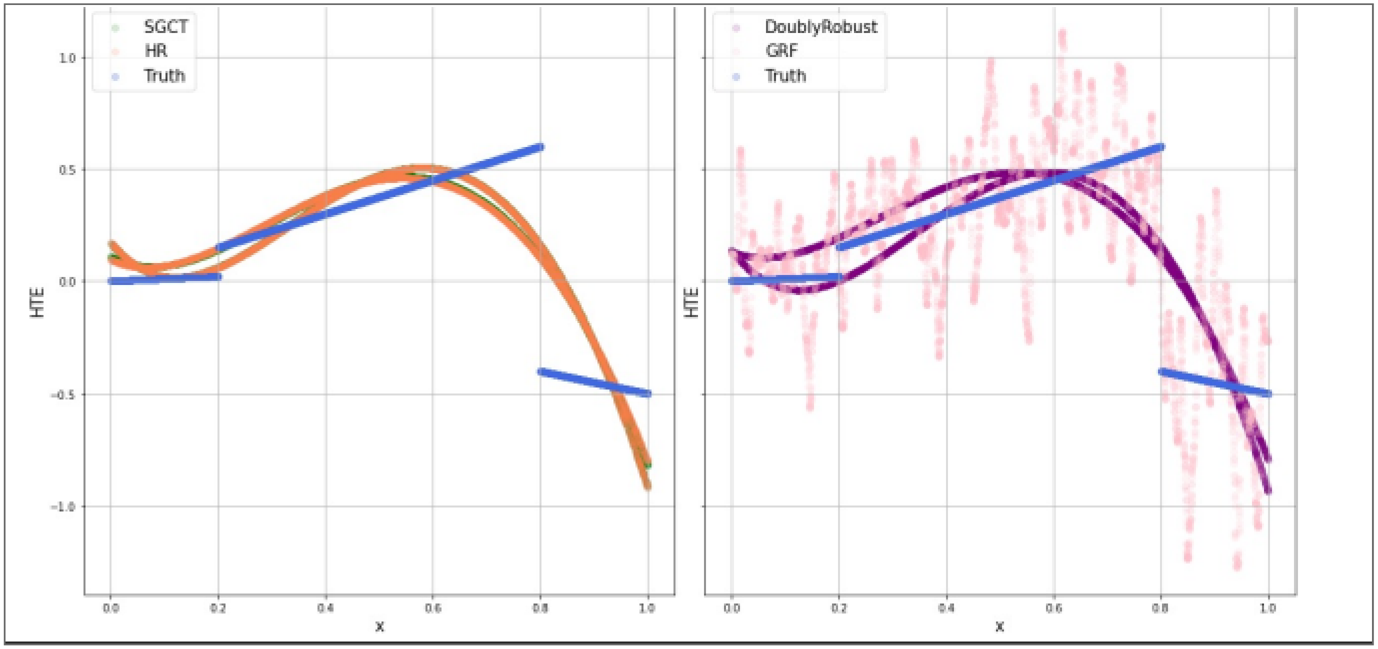


Third Example - Piece-wise

- $ HTE(x) = \Big{ \begin{array}{rl}

```
0.10x, & x < 0.20 \\
0.75x ,& 0.20 \leq x < 0.80 \\
-0.50x, & 0.80 \leq x \\
```

  \end{array} $
- Model controls: x, x^2

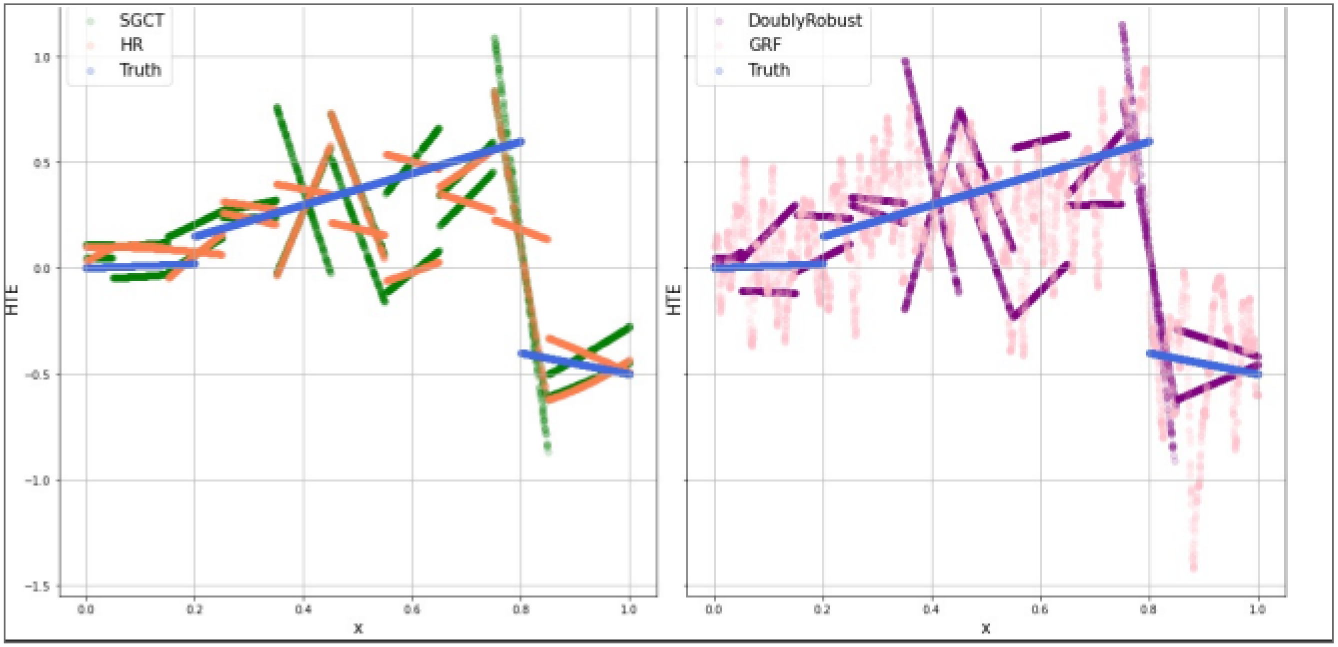


Third Example - Piece-wise with more controls

- $ HTE(x) = \Big{ \begin{array}{rl}

```
0.10x, & x < 0.20 \\
0.75x ,& 0.20 \leq x < 0.80 \\
-0.50x, & 0.80 \leq x \\
```

  \end{array} $
- Model controls: $x, x^2, 1\{0 \leq x > 0.2\}, \ldots, 1\{0.8 \leq x > 1\}$


SGCT
HR
Truth
DoublyRobust
GRF
Truth
HTE
x
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0

Takeaways

- Including more features to estimate a more flexible HTE may not necessarily increase performance.
- The more complicated, or more fine-grained, you want HTE estimates to be, the more data you need.

Review and Conclusion

- Covered the additional complexities and challenges of estimating HTE
- Covered a parametric (DML, HR) and non-parametric (forests) models
    - Deep neural network models (Farrell et. al 2020) not covered because of code availability
- Demonstration with simulated data

```
In [ ]:
```