# Causal Inference Crash Course Part 5: Heterogeneous Treatment Effect Models and Inference

Julian Hsu

# Causal Inference Series

1) Foundations
2) Defining Some ATE/ATET Causal Models
3) ATE/ATET Inference, Asymptotic Theory, and Bootstrapping
4) Best Practices: Outliers, Class Imbalance, Feature Selection, and Bad Control [skipped for now]
5) **Heterogeneous Treatment Effect Models and Inference**
6) Difference-in-Difference Models for Panel Data
7) Regression Discontinuity Models
8) Arguable Validation

# 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:
  - Double Machine Learning following Semenova et al. (2021)
  - Heterogeneous Residuals
  - Causal Forests / Local Linear Forests
  - Doubly Robust models following Kennedy (2020)
- Wrap up with a simulation demonstration

# HTE Overview

- Average treatment effect (ATE) and average treatment effect on the treated (ATET) models want to know aggregate treatment effects.
- Instead, HTE model want to estimate the distribution of treatment effects.

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

- So that $\hat{\tau}(Z_i)$ is the HTE and varies over $Z_i$. We keep $X_i$ different from $Z_i$ for more flexible notation.
- You also $\hat{\tau}(Z_i)$ denoted as the conditional average treatment effect: $\mathrm{E}[\tau(Z_i)|Z_i]$

# 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 can assume linearity as well.
  - We average over more granular treatment effects.
- Estimating more granular treatment effects means there are additional challenges.

# How much variation do we want in HTE?

- 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

2. Segmented estimates are the least inflexible, with the least risk of noise driving estimates.

- 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 – Semenova et al. | Y | Y | Y |
| DML – Heterogeneous Residuals | Y | Y | Y |
| Generalized Random Forests | Y | N | N |
| Doubly Robust – Kennedy (2020) | Y | N | N |
| | | | |

# DML-Style Models

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

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

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

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

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

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

- Continuing this example, the model is:

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

# SGCT uses residualization

- Now how do we estimate this equation?

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

- 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

$$\widetilde{Y}_i = \hat{\tau}_0 \widetilde{T}_i + \hat{\tau}_1 z_{1i} \widetilde{T}_i + \hat{\tau}_2 z_{1i}^2 \widetilde{T}_i + \eta_i$$

- Where $\widetilde{Y}_i$ and $\widetilde{T}_i$ are the residualized outcome and treatment.
- This works via Frisch-Waugh-Lovell, which will come up again when we look at the Heterogeneous Residuals model.

# SGCT – HTE and inference

- We now need to do inference for individual treatment effects from
$$\widetilde{Y}_i = \hat{\tau}_0 \widetilde{T}_i + \hat{\tau}_1 z_{1i} \widetilde{T}_i + \hat{\tau}_2 z_{1i}^2 \widetilde{T}_i + \eta_i$$
- 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 account tenure)

# SGCT – inference with post-LASSO regression

$$\widetilde{Y}_i = \hat{\tau}_0 \tilde{T}_i + \hat{\tau}_1 z_{1i} \tilde{T}_i + \hat{\tau}_2 z_{1i}^2 \tilde{T}_i + \eta_i$$

- 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, 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.

# Heterogeneous residuals (HR)

- Based on SGCT and does more flexible residualization.

$$\widetilde{Y}_i = \hat{\tau}_0 \widetilde{T}_i + \hat{\tau}_1 z_{1i}\, \widetilde{T}_i + \hat{\tau}_2 z_{1i}^2 \widetilde{T}_i + \eta_i$$

- From SGCT, the estimating equation uses the residualized outcome ($\widetilde{Y}_i$) and treatment ($\widetilde{T}_i$).
- By Frisch-Waugh-Lovell:

$$\hat{\tau}_1 = \frac{cov(\, \tilde{Y}_i, z_{1i}\, \tilde{T}_i - E[z_{1i}\, \tilde{T}_i \,|z_{1i}^2 \tilde{T}_i])}{var(z_{1i}\, \tilde{T}_i - E[z_{1i}\, \tilde{T}_i \,|z_{1i}^2 \tilde{T}_i])}$$

- The problem is that $E[z_{1i}\, \tilde{T}_i \,|z_{1i}^2 \tilde{T}_i]$ is a linear expectation and could be more flexible

# HR – flexible residualization

$$\hat{\tau}_1 = \frac{cov(\tilde{Y}_i, z_{1i}\,\tilde{T}_i - E[z_{1i}\,\tilde{T}_i\,|z_{1i}^2\tilde{T}_i])}{var(z_{1i}\,\tilde{T}_i - E[z_{1i}\,\tilde{T}_i\,|z_{1i}^2\tilde{T}_i])}$$

- Let's use ML models to calculate the expectations.
- Therefore, we need to residualize $T_i, z_{1i}T_i,$ and $z_{1i}^2T_i$.
- This treats HTE as a multiple treatments problem. Instead of estimating how the treatment effect varies over features, we estimate separate treatments.
- This gives us additional flexibility to better apply Frisch-Waugh-Lovell.

# Generalized Random Forests

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

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

- 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):
  $$\left(\hat{\tau}(Z_i), \hat{g}(X_i)\right) = argmin\{E[\alpha_i(z)(Y\ - \hat{g}(X_i)\ - \hat{\tau}(Z_i)W_i |\, Z_i = z]^2\}$$
- We have already motivated this part. So what is the purpose of this?
- $\alpha_i(z)$ is a weight used to allow flexibility in $\hat{\tau}(Z_i)$.
  - Can be estimated to kernel methods (ie. Localized DSI model), 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 the flexibility to variation in $\hat{\tau}(Z_i)$ across different points $z$.

# 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.
  - See GRF / Appendix for the technical definition of $\alpha_i(z)$ .
- 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 $\hat{\tau}(z)$ when $z$ is removed from the sample

# Doubly Robust – Kennedy (2020)

- Recall the interactive regression model from DML:

$$\hat{\tau}_{ATE} = E[\,\left(\hat{Y}_{1,i} - \hat{Y}_{0,i}\right) + \frac{T_i\left(Y_i - \hat{Y}_{1,i}\right)}{\hat{p}_i} - \frac{(1 - T_i)\left(Y_i - \hat{Y}_{0,i}\right)}{1 - \hat{p}_i}]$$

- 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

$$\left(\hat{Y}_{1,i} - \hat{Y}_{0,i}\right) + \frac{T_i\left(Y_i - \hat{Y}_{1,i}\right)}{\hat{p}_i} - \frac{(1 - T_i)\left(Y_i - \hat{Y}_{0,i}\right)}{1 - \hat{p}_i}$$

# 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.

$$hte_i = \left(\hat{Y}_{1,i} - \hat{Y}_{0,i}\right) + \frac{T_i\left(Y_i - \hat{Y}_{1,i}\right)}{\hat{p}_i} - \frac{(1 - T_i)\left(Y_i - \hat{Y}_{0,i}\right)}{1 - \hat{p}_i}$$

# Refining estimates with a second stage

- Kennedy (2020) proposes applying a regression model to the “noisy” estimates of the true HTE, $hte_i$.
  - OLS
  - Kernel regression
  - Cross-splitted LASSO regressions

# Simulation Study

# Context

- Recall, HTE models are not about estimating the <u>average</u> 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

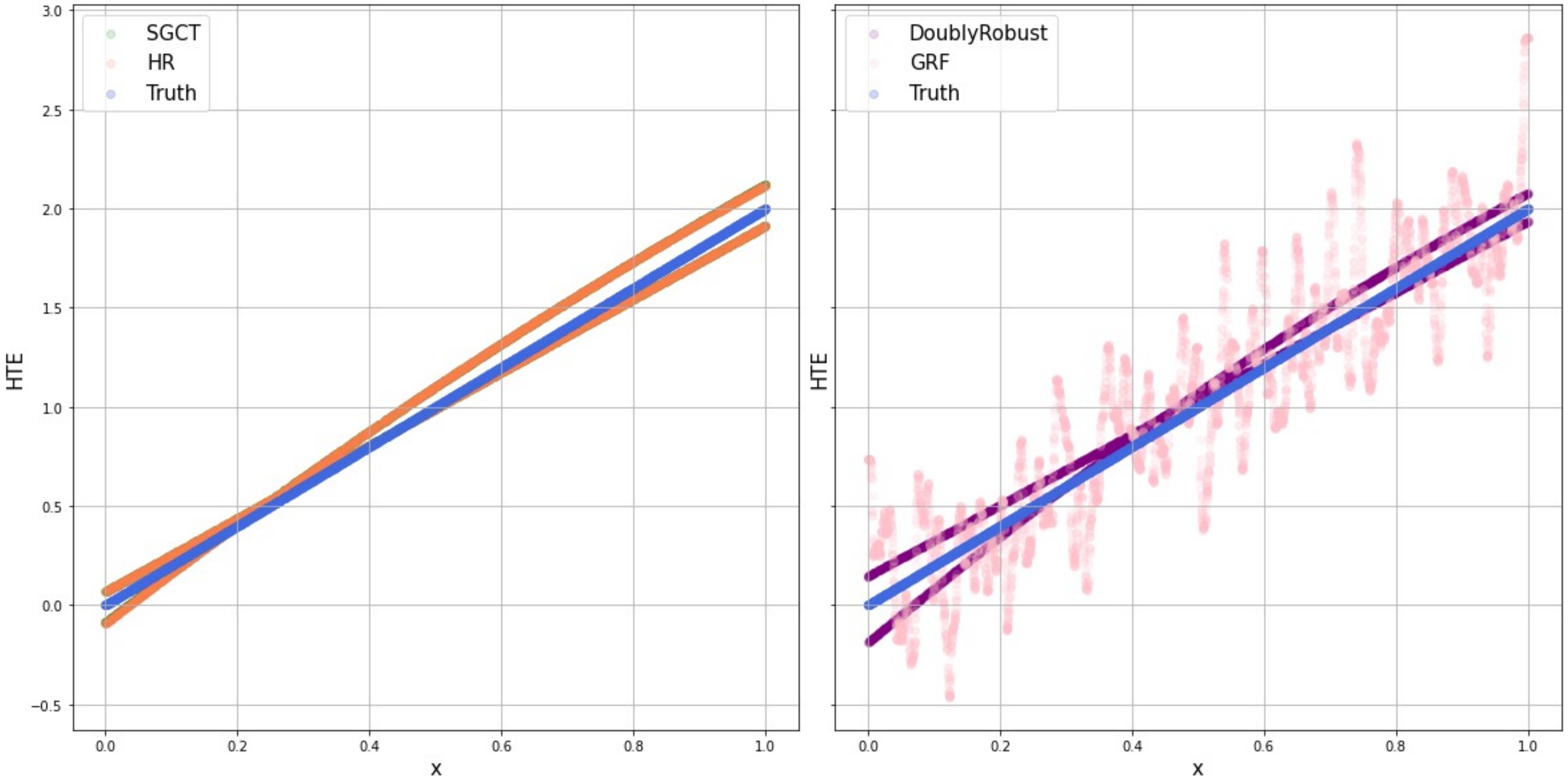
$$x \sim U[0,1]$$
$$y = 10 + 2 * \ln(1 + x) + \epsilon, \epsilon \sim N(0,1)$$
$$W = 1\{\frac{\exp(x)}{1+\exp(x)} + \eta > 0 \}, \eta \sim N(0,1)$$

- We want to know the HTE of $W$.
- We show three examples, with different HTE functions

# First Example – Linearity

$$HTE(x) = 2x$$

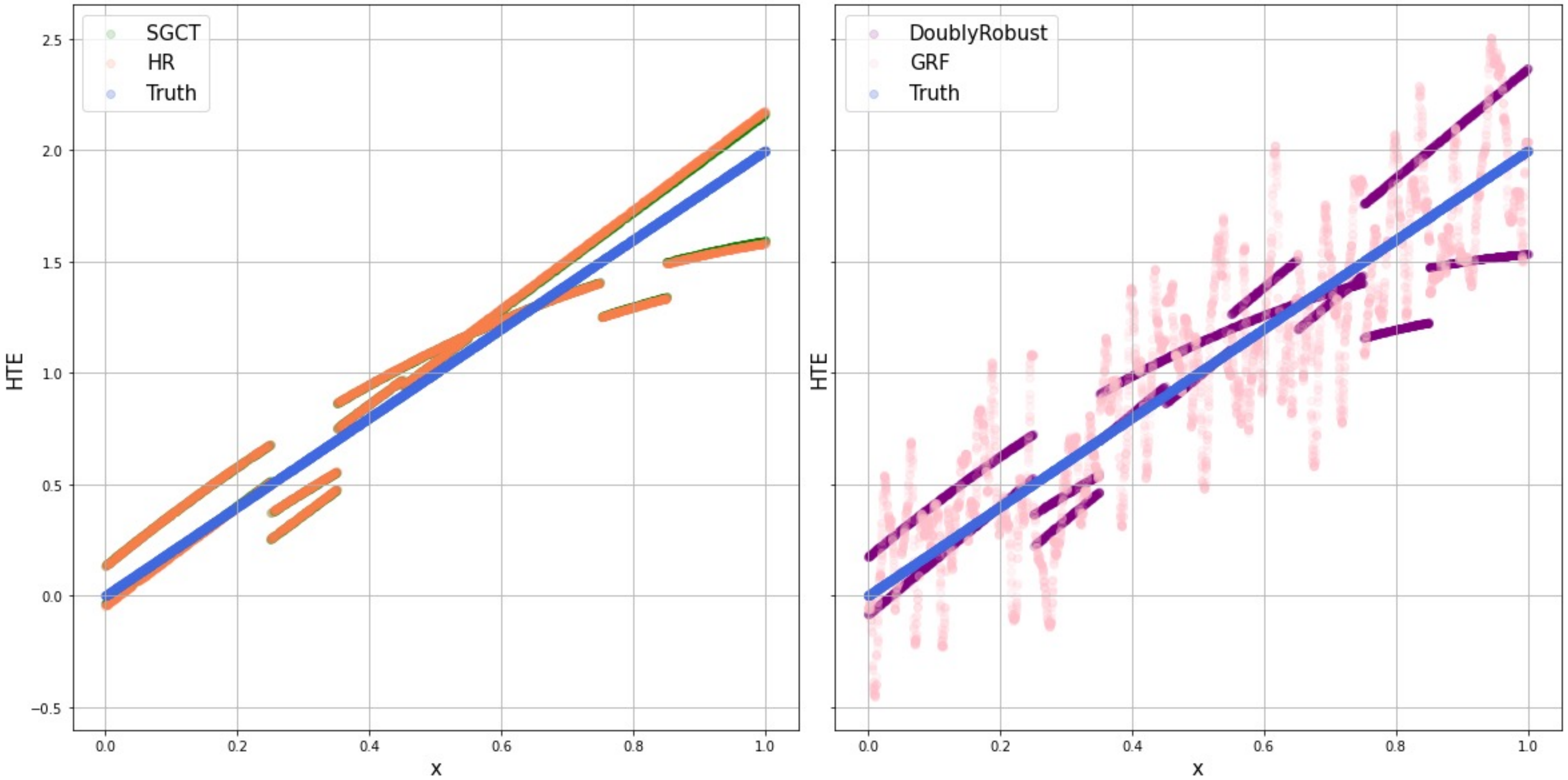
Model controls: $x, x^2$



# First Example – Linearity with more controls

$$HTE(x) = 2x$$

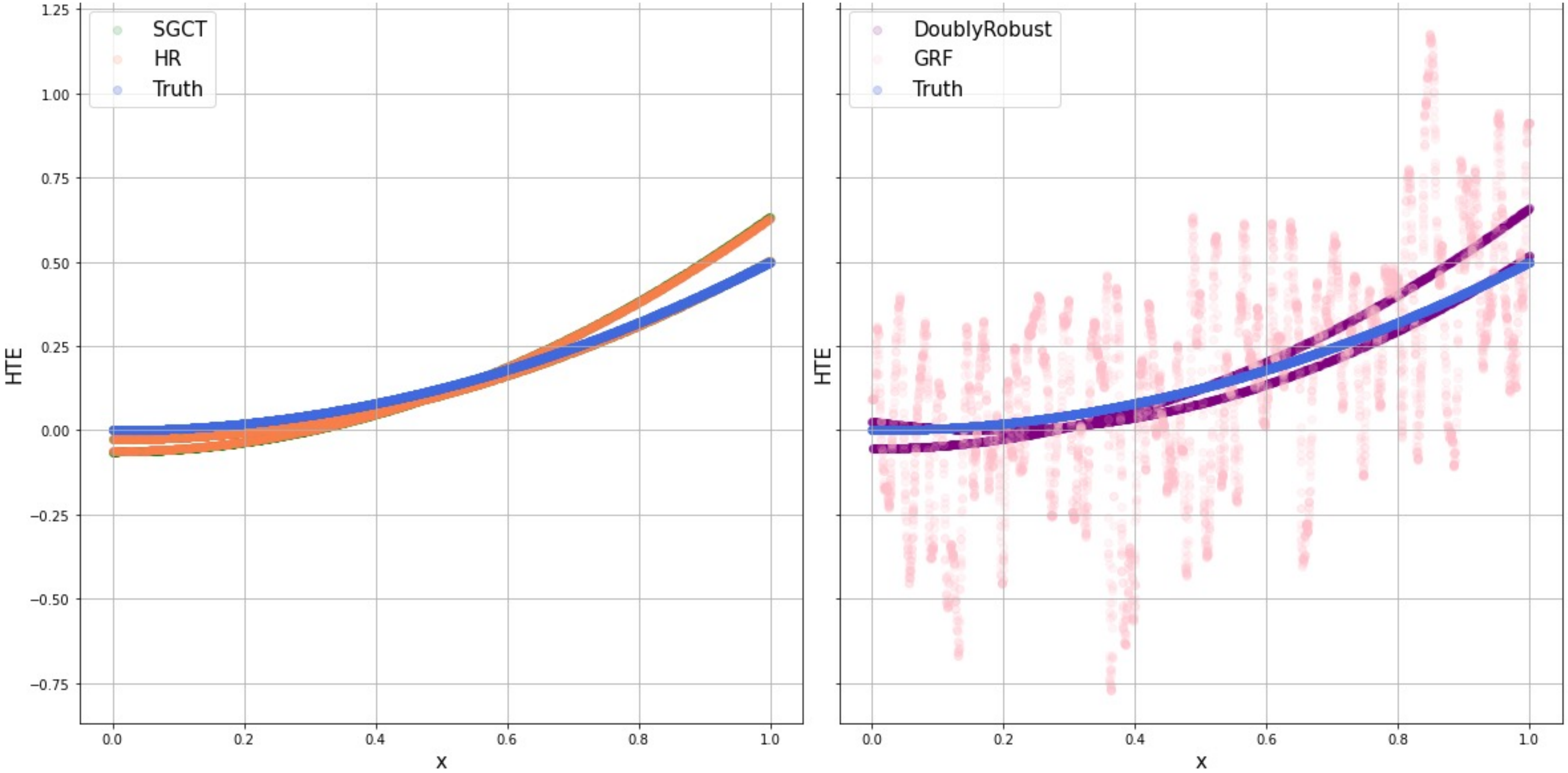
Model controls: $x, x^2, 1\{0 \geq x > 0.2\}, 1\{0.2 \geq x > 0.4\}, \ldots\ 1\{0.8 \geq x > 1\},$



# Second Example – Quadratics

$HTE(x) = \frac{1}{2}x^2$

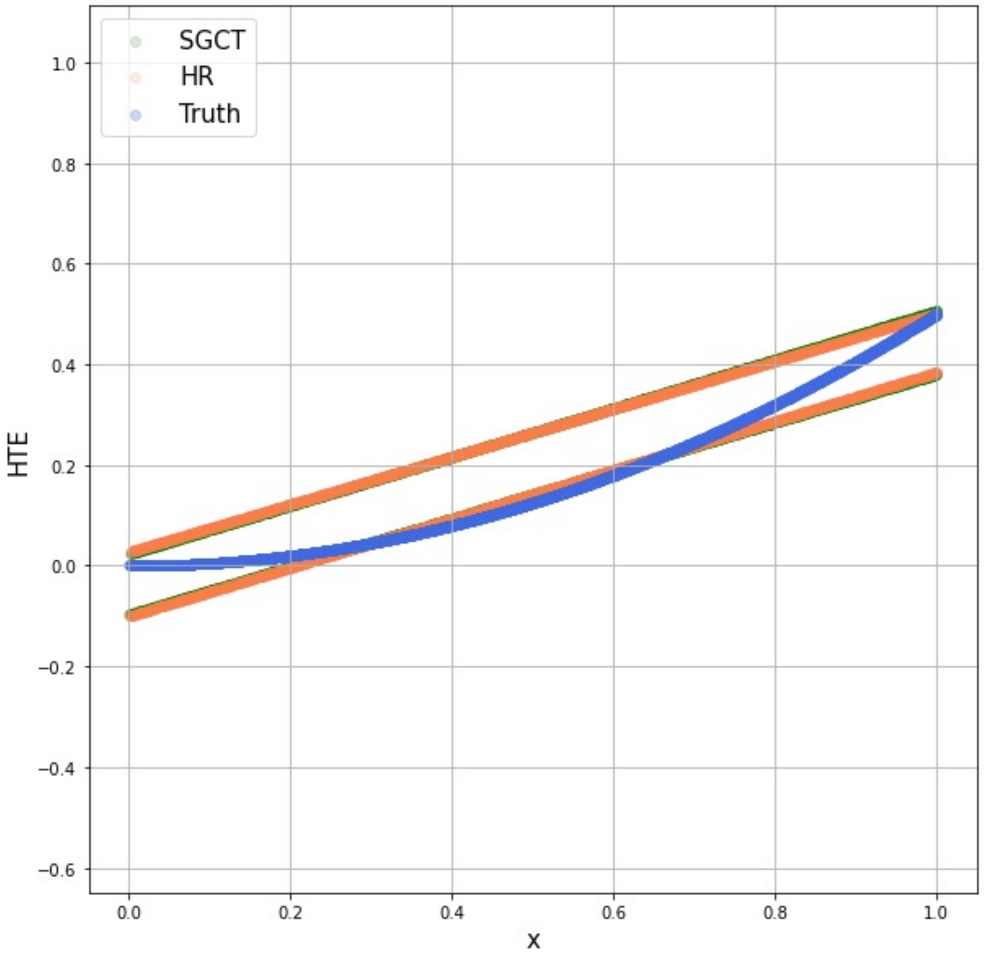
Model controls: $x, x^2$

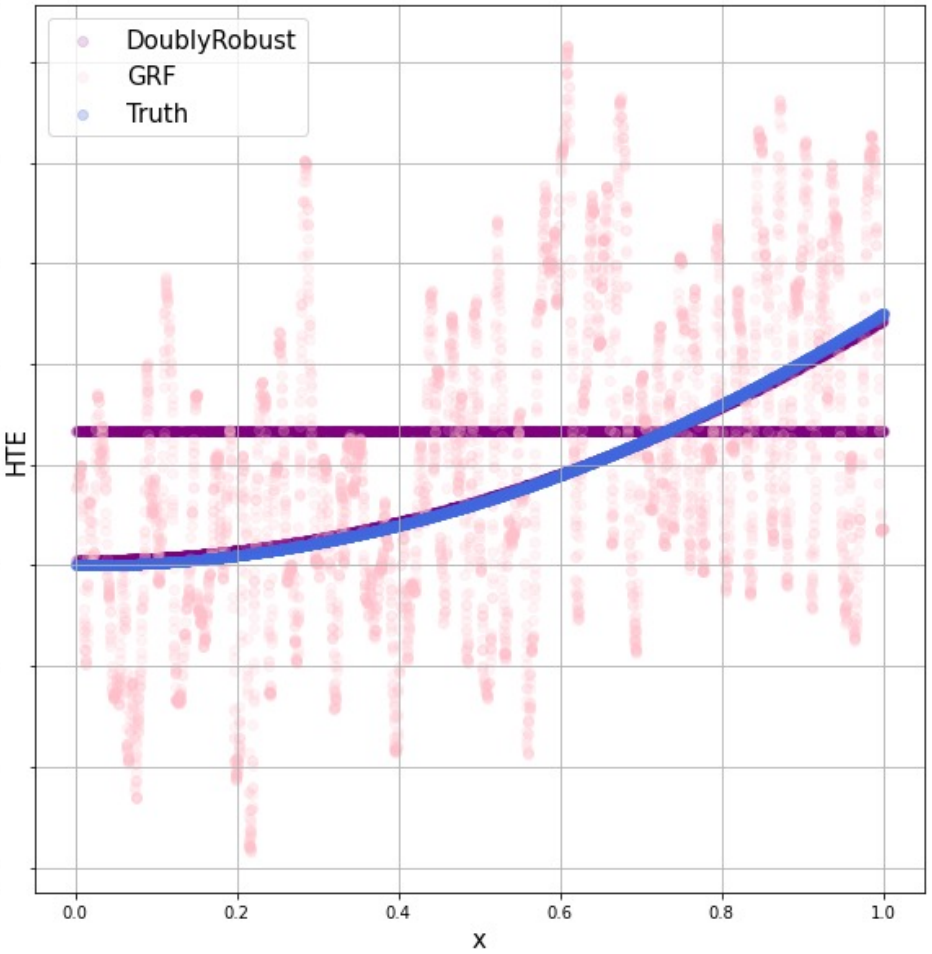


# Second Example – Quadratics with more controls

$HTE(x) = 2x$

Model controls: $x, x^2, 1\{0 \geq x > 0.2\}, 1\{0.2 \geq x > 0.4\}, \ldots 1\{0.8 \geq x > 1\}$,

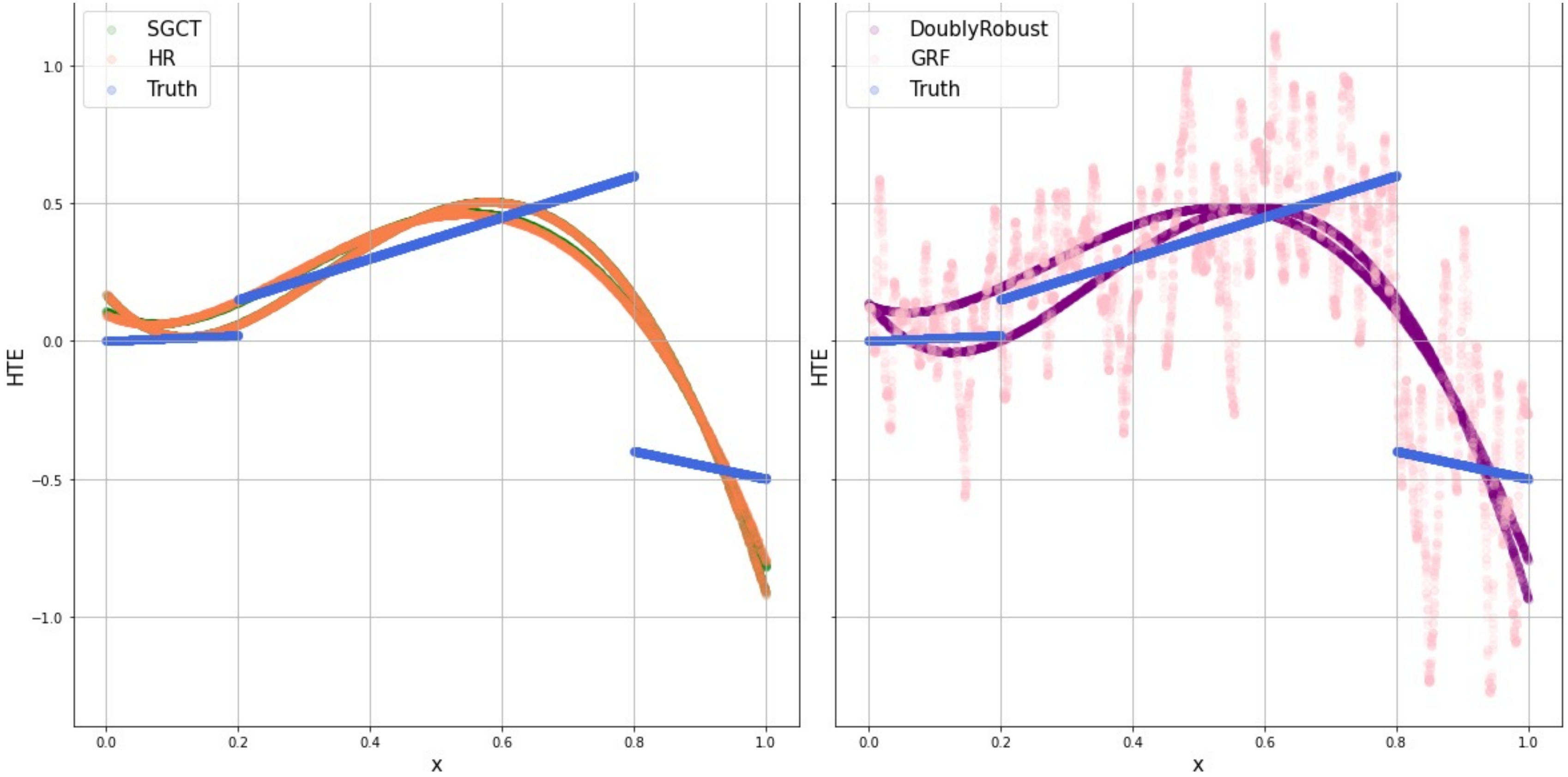


# Third Example – Piece-wise

$$HTE(x)\begin{cases} 0.10x, x < 0.20 \\ 0.75x, 0.20 \le x < 0.80 \\ -0.50x, 0.80 \le x \end{cases}$$

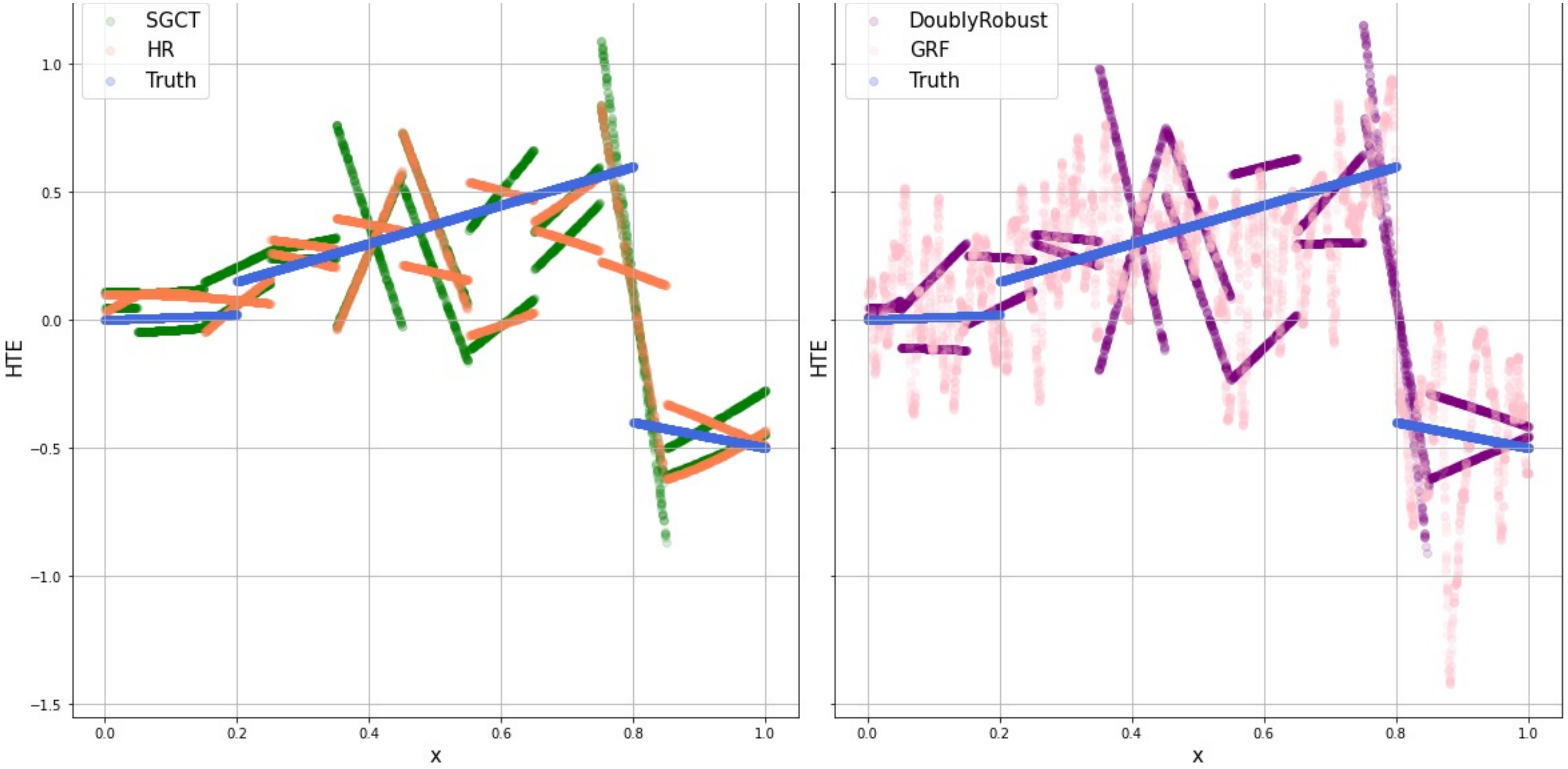
Model controls: $x, x^2$



# Third Example – Piece-wise with more controls

$$HTE(x)\begin{cases} 0.10x, x < 0.20 \\ 0.75x, 0.20 \leq x < 0.80 \\ -0.50x, 0.80 \leq x \end{cases}$$

Model controls: $x, x^2, 1\{0 \geq x > 0.2\}, 1\{0.2 \geq x > 0.4\}, \ldots 1\{0.8 \geq x > 1\}$



# 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

# Causal Inference Series

1) Foundations
2) Defining Some ATE/ATET Causal Models
3) ATE/ATET Inference, Asymptotic Theory, and Bootstrapping
4) Best Practices: Outliers, Class Imbalance, Feature Selection, and Bad Control [skipped for now]
5) **Heterogeneous Treatment Effect Models and Inference**
6) Difference-in-Difference Models for Panel Data
7) Regression Discontinuity Models
8) Arguable Validation

# Appendix Slides