Machine Learning and Heterogeneous Effects

Mixtape Session

Prof. Brigham Frandsen

Allow me to introduce myself

- Economics professor at Brigham Young University in Utah

Allow me to introduce myself

- Economics professor at Brigham Young University in Utah
- 4 biological kids, 3 foster daughters, most of whom can now run and mountain bike faster than me

Allow me to introduce myself

- Economics professor at Brigham Young University in Utah
- 4 biological kids, 3 foster daughters, most of whom can now run and mountain bike faster than me
- A big fan of causal inference in observational settings:
  - Quasi-experimental evaluations of the effects of unions (Frandsen 2016, 2017, 2021; Chen, Frandsen, Grabowski, Town, Sojourner 2015)
  - Distributional effects (Frandsen and Lefgren 2018, 2021; Frandsen, Froelich, Melly 2012)
- And of exploring machine learning in applied economics:
  - Teach Machine Learning for Economists at BYU
  - Research on the power of ML in empirical strategies (Angrist and Frandsen 2022)

Effects Ex Machina: Where we're going

Machine Learning + Heterogeneous Treatment Effects

- Causality primer/review
- Machine learning (ML) prediction primer/review
- Heterogeneous treatment effects
  - When they matter
  - Conceptual framework
  - Using ML to predict treatment effects: Random Causal Forests
  - Python/R implementation

(Prequel to this course: Machine Learning and Causal Inference)

Potential outcomes and treatment effects

Potential outcomes and treatment effects

$43M

Potential outcomes and treatment effects

$43M

Potential outcomes and treatment effects

$43M

$700K

Potential outcomes and treatment effects

$$$

$

$43M
D = 1
Y(1)

$700K
D = 0
Y(0)

Potential outcomes

Potential outcomes and treatment effects

\$43M	-	\$700K	=	\$42.3M
Y(1)	-	Y(0)	=	Treatment effect

Potential outcomes and treatment effects

$$$ $

\$43M - \$700K = \$42.3M

Y(1) - Y(0) = Treatment effect

counterfactual

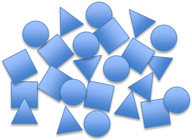
Potential outcomes and treatment effects

$$$ $

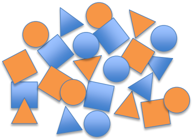
$43M - $700K = $42.3M

Y(1) - Y(0) = Treatment effect

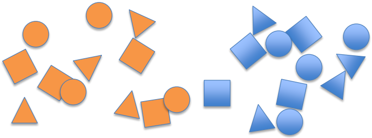
Potential outcomes and treatment effects



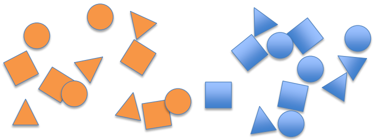
Potential outcomes and treatment effects



Potential outcomes and treatment effects

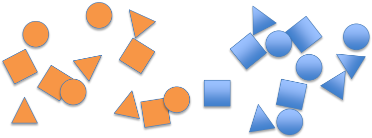


Potential outcomes and treatment effects



E[Y(1)] E[Y(0)]

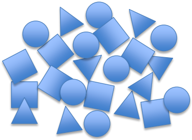
Potential outcomes and treatment effects



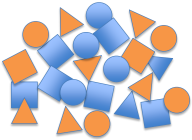
E[Y(1)] - E[Y(0)] = E[Y(1) – Y(0)]

ATE

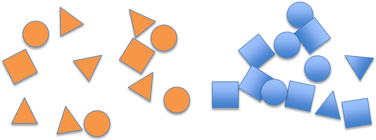
Potential outcomes and treatment effects



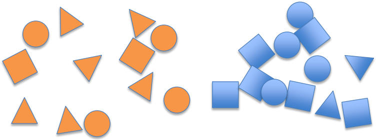
Potential outcomes and treatment effects



Potential outcomes and treatment effects



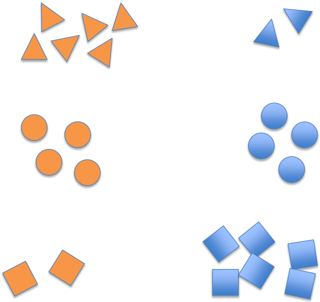
Potential outcomes and treatment effects



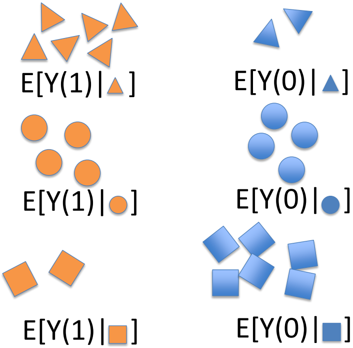
$$\underbrace{E[Y|D=1] - E[Y|D=0]}_{} = \underbrace{E[Y(1) - Y(0)|D=1]}_{\text{ATT}}$$

$$+ \underbrace{E[Y(0)|D=1] - E[Y(0)|D=0]}_{\text{selection bias}}$$

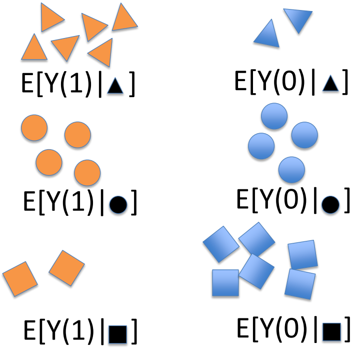
Potential outcomes and treatment effects



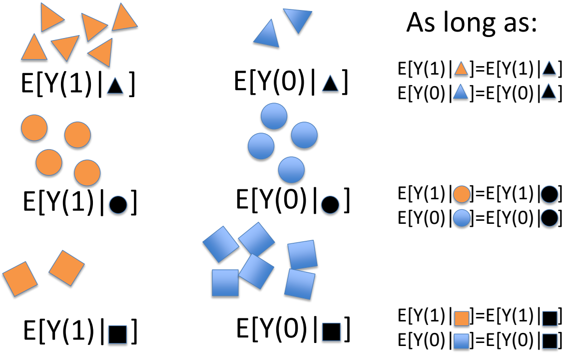
Potential outcomes and treatment effects



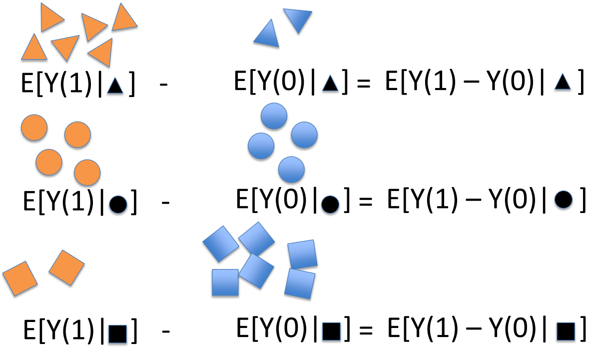
Potential outcomes and treatment effects



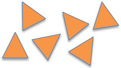
Potential outcomes and treatment effects



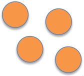
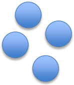
Potential outcomes and treatment effects



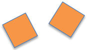
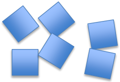
Potential outcomes and treatment effects




E[Y(1)|▲] - E[Y(0)|▲] = E[Y(1) – Y(0)| ▲]

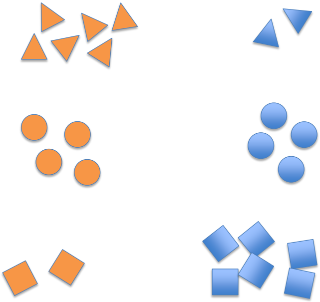
E[Y(1)|●] - E[Y(0)|●] = E[Y(1) – Y(0)| ●]

E[Y(1)|■] - E[Y(0)|■] = E[Y(1) – Y(0)| ■]

E[Y(1) – Y(0)]

Potential outcomes and treatment effects

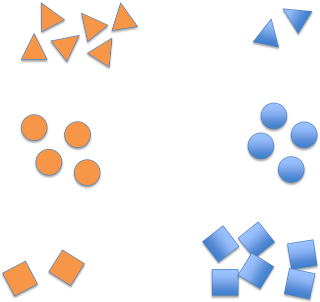


$Y(1),Y(0) \perp\!\!\!\perp D \,|\, \blacktriangle$

$Y(1),Y(0) \perp\!\!\!\perp D \,|\, \bullet$

$Y(1),Y(0) \perp\!\!\!\perp D \,|\, \blacksquare$

Potential outcomes and treatment effects



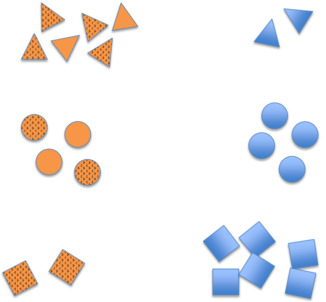
$Y(1), Y(0) \perp\!\!\!\perp D \mid$ ▲

$Y(1), Y(0) \perp\!\!\!\perp D \mid$ ●

$Y(1), Y(0) \perp\!\!\!\perp D \mid$ ■

Selection on observables, "unconfoundedness": $Y(1), Y(0) \perp\!\!\!\perp D \mid X$

Potential outcomes and treatment effects

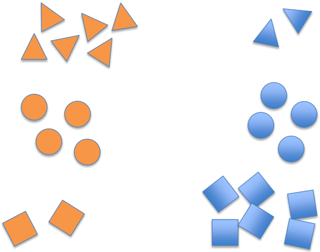


$Y(1),Y(0) \not\perp D \mid \blacktriangle$

$Y(1),Y(0) \not\perp D \mid \bullet$

$Y(1),Y(0) \not\perp D \mid \blacksquare$

Potential outcomes and treatment effects

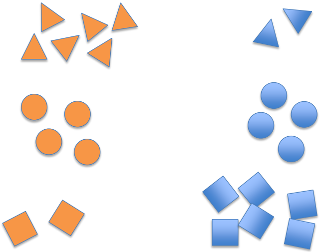


Pr(D=1|▲) = .75

Pr(D=1|●) = .5

Pr(D=1|■) = .25

Potential outcomes and treatment effects



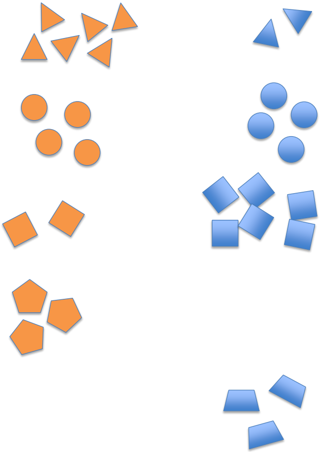
Pr(D=1|▲) = .75

Pr(D=1|●) = .5

Pr(D=1|■) = .25

Common support,
“overlap”: $0 < Pr(D=1|X) < 1$

Potential outcomes and treatment effects

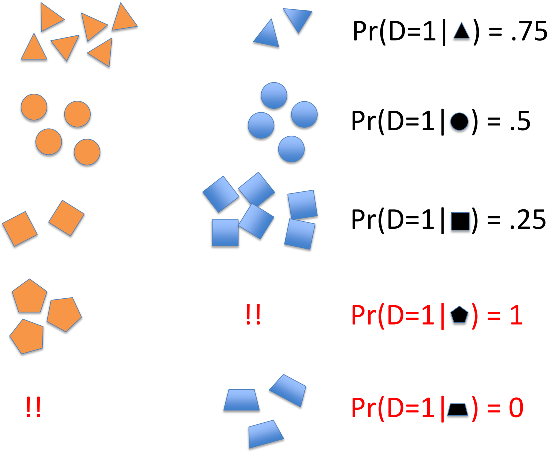


Pr(D=1|▲) = .75

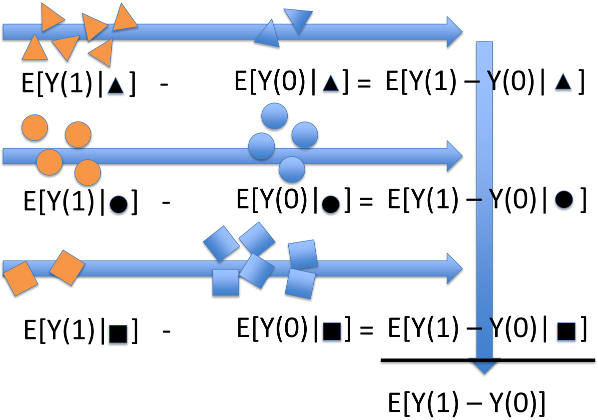
Pr(D=1|●) = .5

Pr(D=1|■) = .25

Potential outcomes and treatment effects



Potential outcomes and treatment effects



Basic causal inference summary

- Target (for now!):

$$ATE = E\left[Y_i(1) - Y_i(0)\right] = E\left[\tau_i\right]$$

- Key identifying assumption:

$$\left(Y_i(0), Y_i(1)\right) \perp\!\!\!\perp D_i | X_i$$

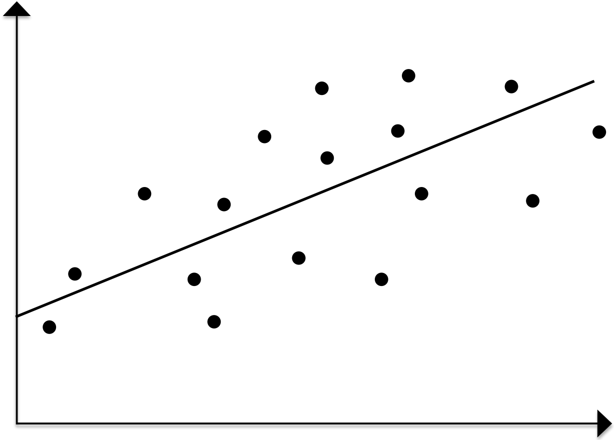
- Estimation:
  - Multiple linear regression (OLS)

$$Y_i = \beta_0 + \tau D_i + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon$$

  - Matching
  - Propensity score methods
  - Machine-assisted:
    - Post-Double Selection Lasso
    - Double/De-biased Machine Learning
- Go to python!

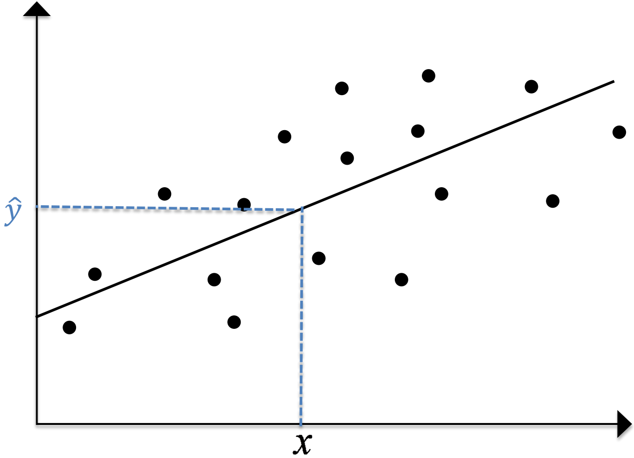
Prediction Target

$$y_i = \alpha + \beta x_i + \varepsilon_i$$



Prediction Target

$$y_i = \underbrace{\alpha + \beta x_i}_{\hat{y}} + \varepsilon_i$$



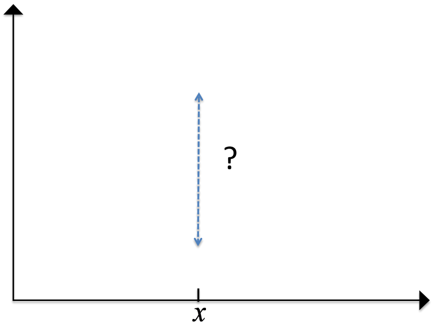
Prediction Methods

Supervised machine learning algorithms:

- Decision trees
- Random forests
- Penalized regression (ridge, lasso)
- Support vector machines

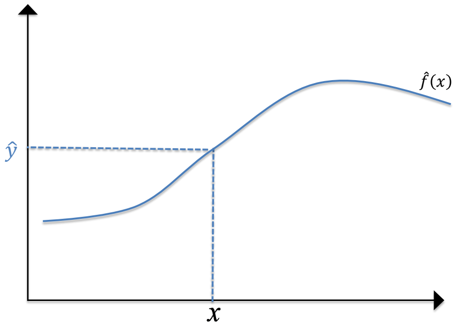
Prediction mechanics

- **Goal:** Predict an out-of-sample outcome Y



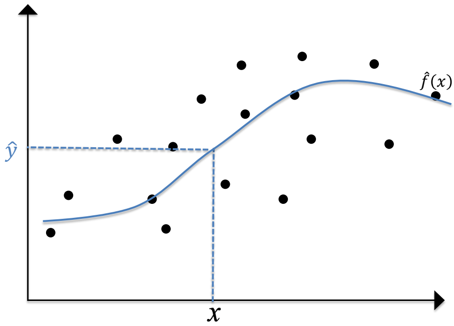
Prediction mechanics

- **Goal:** Predict an out-of-sample outcome Y
- as a function, $\hat{f}(X)$, of **features** $X = (1, X_1, X_2, \ldots, X_K)'$.



Prediction mechanics

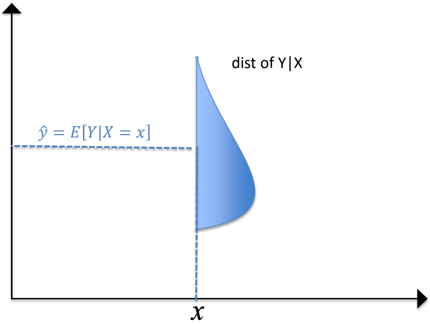
- **Goal:** Predict an out-of-sample outcome Y
- as a function, $\hat{f}(X)$, of **features** $X = (1, X_1, X_2, \ldots, X_K)'$.
- Estimate the function $\hat{f}$ (aka "train the model") based on **training sample** $\{(Y_i, X_i); i = 1, \ldots, N\}$



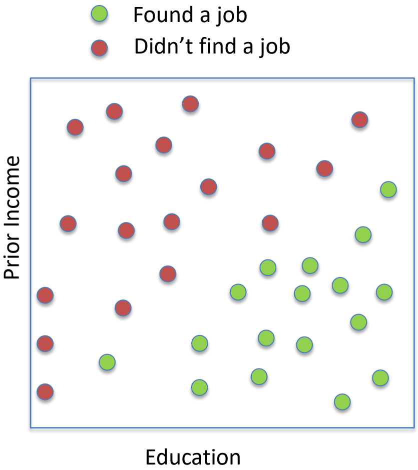
What's a "good" prediction?

- Want our prediction to be "close," i.e. minimize the expected **mean squared error**:

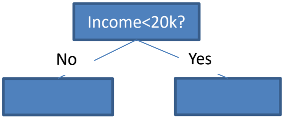
$$\min_{f(x)} E\left[\left(f(x) - Y\right)^2 \middle| X = x\right]$$

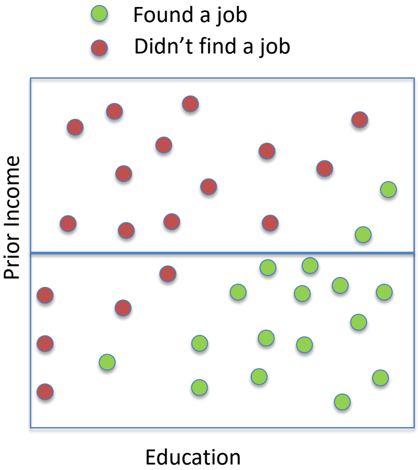


Decision Trees

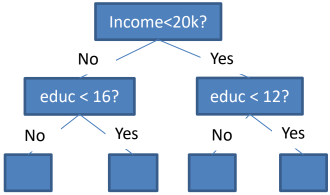


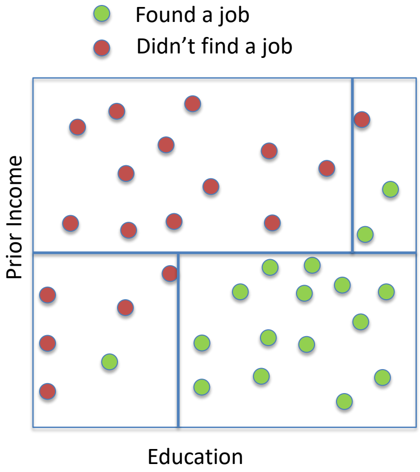
Decision Trees



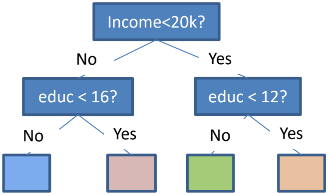


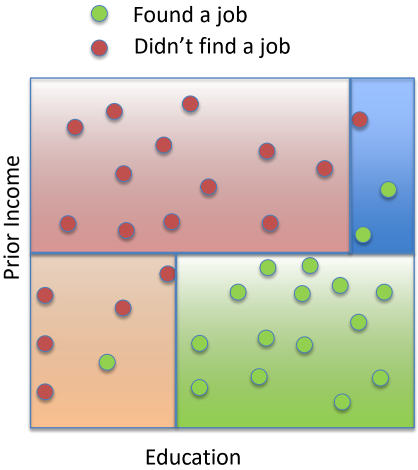
Decision Trees



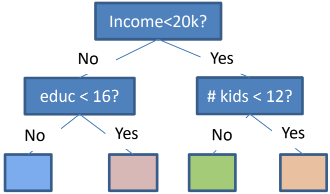


Decision Trees



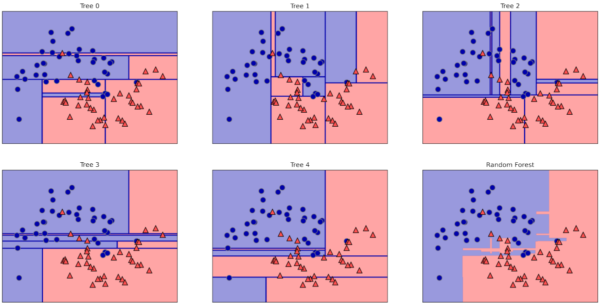


Decision Trees



- Where to split:
  Choose the feature from $\{x_1, \ldots, x_p\}$ and the value of that feature to minimize MSE in the resulting child nodes
- Tuning parameters
  - Max depth
  - Min training obs per leaf
  - Min improvement in fit in order to go ahead with a split

Forest for the Trees



- Value proposition: reduce variance by averaging together multiple predictions
- The catch: individual trees need to be **de-correlated**
- Algorithm:
  - Grow B trees, each on a different bootstrapped sample
  - At each split, consider only a random subset of features
  - Average together the individual predictions
- Let's grow some trees in python!

Combining causal effects and ML: predicting heterogeneous treatment effects

- What is the effect of job training on the probability of finding a job . . .
  - for more-educated vs. less-educated individuals?
  - for men vs. women?
  - for married vs. single?
  - for high-earning vs. low-earning (prior to training)?
  - for minorities vs. non-minorities?
- Why does it matter?
- Other examples where heterogeneity in treatment effects matter?

Traditional heterogeneity analysis: Interacted regression

To estimate the overall average effect:

$$Y_i = \tau D_i + \varepsilon_i, \quad i \in \{1, \ldots, n\}$$

Traditional heterogeneity analysis: Interacted regression

To estimate the overall average effect:

$$Y_i = \tau D_i + \varepsilon_i, \quad i \in \{1, \dots, n\}$$

To explore heterogeneity by sex:

$$\begin{aligned} Y_i &= \tau^{female} D_i + \varepsilon_i, \qquad i : Female_i = 1 \\ Y_i &= \tau^{male} D_i + \varepsilon_i, \qquad i : Female_i = 0, \end{aligned}$$

Traditional heterogeneity analysis: Interacted regression

To estimate the overall average effect:

$$Y_i = \tau D_i + \varepsilon_i, \quad i \in \{1, \ldots, n\}$$

To explore heterogeneity by sex:

$$\begin{aligned} Y_i &= \tau^{female} D_i + \varepsilon_i, \qquad i : Female_i = 1 \\ Y_i &= \tau^{male} D_i + \varepsilon_i, \qquad i : Female_i = 0, \end{aligned}$$

or, equivalently:

$$\begin{aligned} Y_i &= \tau^{male} D_i + \beta Female_i + \gamma D_i \times Female_i + \varepsilon_i \\ \tau^{female} &= \tau^{male} + \gamma. \end{aligned}$$

Traditional heterogeneity analysis: Interacted regression

To estimate the overall average effect:

$$Y_i = \tau D_i + \varepsilon_i, \quad i \in \{1, \ldots, n\}$$

To explore heterogeneity by sex:

$$\begin{aligned} Y_i &= \tau^{female} D_i + \varepsilon_i, \qquad i : Female_i = 1 \\ Y_i &= \tau^{male} D_i + \varepsilon_i, \qquad i : Female_i = 0, \end{aligned}$$

or, equivalently:

$$\begin{aligned} Y_i &= \tau^{male} D_i + \beta Female_i + \gamma D_i \times Female_i + \varepsilon_i \\ \tau^{female} &= \tau^{male} + \gamma. \end{aligned}$$

More generally,

$$\begin{aligned} Y_i &= \tau D_i + X_i' \beta + D_i X_i' \gamma + \varepsilon_i, \\ \tau(x) &= \tau + x' \gamma \end{aligned}$$

Challenges with traditional heterogeneity analysis

$$Y_i = \tau D_i + X_i'\beta + D_i X_i'\gamma + \varepsilon_i$$

- Functional form: treatment effects may not vary linearly with X_i
- Curse of dimensionality: when X_i includes many variables, OLS impractical or infeasible
- These are problems ML was born to solve!

Predicting outcomes vs. treatment effects

Predicting outcomes	Predicting treatment effects
Target: $\hat{y}(x) = E[Y_i \mid X_i = x]$	Target: $\tau(x) = E[\tau_i \mid X_i = x]$
Criterion:	Criterion:
$\min E\left[(\hat{y}(x) - Y_i)^2 \mid X_i = x\right]$	$\min E\left[(\tau(x) - \tau_i)^2 \mid X_i = x\right]$
Training data: $\{Y_i, X_i\}_{i=1}^n$	Training data: $\{\tau_i, X_i\}_{i=1}^n$

Predicting outcomes vs. treatment effects

Predicting outcomes	Predicting treatment effects
Target: $\hat{y}(x) = E[Y_i \mid X_i = x]$	Target: $\tau(x) = E[\tau_i \mid X_i = x]$
Criterion:	Criterion:
$\min E\left[\left(\hat{y}(x) - Y_i\right)^2 \mid X_i = x\right]$	$\min E\left[\left(\tau(x) - \tau_i\right)^2 \mid X_i = x\right]$
Training data: $\{Y_i, X_i\}_{i=1}^n$	Training data: $\{\tau_i, X_i\}_{i=1}^n$

Why is training data a problem for predicting treatment effecs?

Predicting outcomes vs. treatment effects

Predicting outcomes	Predicting treatment effects
Target: $\hat{y}(x) = E[Y_i \mid X_i = x]$	Target: $\tau(x) = E[\tau_i \mid X_i = x]$
Criterion:	Criterion:
$\min E\left[(\hat{y}(x) - Y_i)^2 \mid X_i = x\right]$	$\min E\left[(\tau(x) - \tau_i)^2 \mid X_i = x\right]$
Training data: $\{Y_i, X_i\}_{i=1}^n$	Training data: $\{\tau_i, X_i\}_{i=1}^n$

Why is training data a problem for predicting treatment effecs?

- Consequence: can't apply ML directly to predicting treatment effects; have to adapt them

Adapting ML to predict treatment effects

- Break it up:

$$\begin{aligned} E\left[\tau_i | X_i\right] &:= E\left[Y_i\left(1\right) - Y_i\left(0\right)|X_i\right] \\ &= E\left[Y_i | X_i, D_i = 1\right] - E\left[Y_i | X_i, D_i = 0\right] \end{aligned}$$

(by what assumption?)

Adapting ML to predict treatment effects

- Break it up:

$$\begin{aligned} E\left[\tau_i | X_i\right] &:= E\left[Y_i\left(1\right) - Y_i\left(0\right)|X_i\right] \\ &= E\left[Y_i | X_i, D_i = 1\right] - E\left[Y_i | X_i, D_i = 0\right] \end{aligned}$$

  (by what assumption?)

- Adjust the criterion: (why?)

$$\min \sum_{i=1}^{n} \left(\tau\left(X_i\right) - \tau_i\right)^2 \iff \max \sum_{i=1}^{n} \tau\left(X_i\right)^2$$

Adapting ML to predict treatment effects

- Break it up:

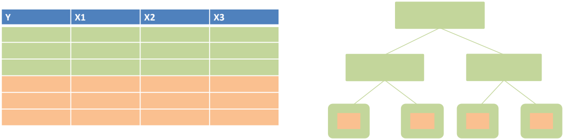
$$
\begin{aligned}
E\left[\tau_i | X_i\right] &:= E\left[Y_i(1) - Y_i(0) | X_i\right] \\
&= E\left[Y_i | X_i, D_i = 1\right] - E\left[Y_i | X_i, D_i = 0\right]
\end{aligned}
$$

  (by what assumption?)

- Adjust the criterion: (why?)

$$
\min \sum_{i=1}^{n} \left(\tau\left(X_i\right) - \tau_i\right)^2 \iff \max \sum_{i=1}^{n} \tau\left(X_i\right)^2
$$

- Be honest: use one set of observations to select the tree structure, and another to generate predictions



Predicting treatment effects using ML: Summary

- Target:

$$CATE := \tau(x) = E[\tau_i | X_i = x]$$

- Key identifying assumption:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | X_i$$

- Estimation: Random Causal Forest
  - Grow decision trees on many bootstrapped samples
  - Choose splits using the training set to max $\sum_{i=1}^{n} \tau(X_i)^2$
  - Generate predictions in each leaf using the estimation set
  - Average predictions over the trees in the forest
- Go to python!

Thank you!