EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

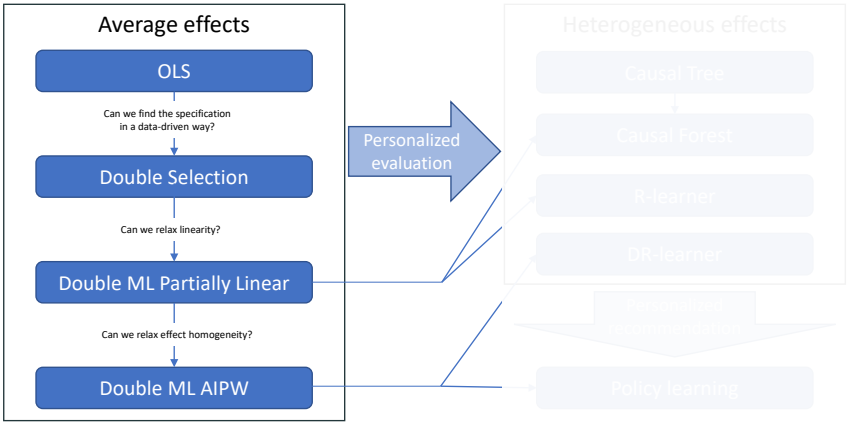
Causal Machine Learning

Heterogeneous effects

Michael Knaus

WiSe 23/24

Today we move away from average effects



Plan of this morning

1. Conditional target parameters
2. Definition and identification
3. Group Average Treatment Effects
4. Predicting effects: The challenge
5. Causal trees/forests
6. Meta-learner

Conditional target parameters

Effect heterogeneity

So far we focused either on constant effects or allowed for heterogeneous effects in broad groups like ATE, ATT or LATE

Such aggregate effect measures are the starting point of most analyses as they provide an excellent summary of the effectiveness of the treatment

$\Rightarrow$ Causal ML provided flexible tools for standard problems

However, often we want to go beyond average effects and

- estimate the effect in pre-specified subgroups (already standard)
- flexibly estimate heterogeneous effects (only recently via Causal ML)

$\Rightarrow$ Causal ML can increase the scope of our analyses

Why is this interesting?

More comprehensive evaluation: who wins or loses and by how much?

This is useful along at least two dimensions:

- **Informs action:** More efficient allocation of public and private resources via targeting in the future

  $\Rightarrow$ Personalized policies, ads, medicine, ...
- **Understanding**: Heterogeneous effects can be suggestive for underlying mechanisms

Beyond that I think it is just exciting to think about how we can predict individualized effects, though we cannot observe them

Definition and identification

There are plenty of heterogeneous effects

Recall that we defined the Conditional Average Treatment Effect (CATE):

$$\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}[\Delta \mid X = x]$$

- Expected treatment effect in the subpopulation with characteristics X?

So far we did not think about where X comes from but there are numerous options to define the "target subpopulation"

To fix ideas it is useful to define $X = H \cup C$, i.e. as the union of

- H: Variables for which we want to understand effect heterogeneity
  - Usually motivated by the research question
- C: Confounders that are required in the anticipated identification step

Special cases

RCTs: Life is relatively simple

- There are no confounders $\Rightarrow X = H \Rightarrow$ CATE defined with respect to considered heterogeneity variables

Measured confounding: It is useful to distinguish two types of CATEs:

- Group ATE (GATE) for some groups G that are defined using H
  $\Rightarrow \tau(g) := E[Y(1) - Y(0) \mid G = g]$
- Individualized ATE (IATE)
  $\Rightarrow \tau(x) := E[Y(1) - Y(0) \mid X = x]$

We will see in the following that the estimation step is affected by whether we are interested in GATEs or IATEs

Recycling identification

From an identification point of view, we do not have to establish new results

In what follows all target parameters can be thought of as special cases of conditioning ITE on some function $f(X)$ such that by the tower property

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0) \mid f(X) = f(x)] &= \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid X = x, f(X) = f(x)] \mid f(X) = f(x)] \\ &= \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid X = x] \mid f(X) = f(x)]\end{aligned}$$

As $X = H \cup C$ is assumed to contain all confounders, we know how to identify the inner expectation $\mathbb{E}[Y(1) - Y(0) \mid X = x]$ in RCTs or under measured confounding

$\Rightarrow$ All aggregations with respect to a function of X are also identified

Group Average Treatment Effects

Examples of GATEs

First, focus on estimation **and inference** for GROUP ATE (GATE)

$$\tau(g) = E[Y(1) - Y(0) \mid G = g]$$

Examples for subgroups of interest:

- Classic mutually exclusive subgroups, like $G = \{female, male, ...\}$, $G = \{age < 50, age \geq 50\}$, $G = \{age < 50 \ \& \ female, age < 50 \ \& \ male, age \geq 50 \ \& \ female, age \geq 50 \ \& \ male, ...\}$, ...
- Single or low-dimensional continuous variable, like $G = age$, $G = income$, ...
- Or other functions or small subsets of X

These groups should be pre-determined or at least out-of-sample and not be the result of data snooping

Current practice (1/2)

Such GATEs are the focus of numerous "heterogeneous effects" sections of empirical papers

There are basically two common strategies:

1. Stratify the data and rerun the analysis for female vs. male, young vs. old, native vs. foreign, ...
2. Specify an interaction term in a regression model of the general form
   $Y = \alpha + \theta W + \rho_0 G + \rho_1 W * G + X\beta + \epsilon$

   where ρ_0 represents the "main effect" of the heterogeneity variable and ρ_1 the "interaction effect"

Current practice (2/2)

Both strategies have their drawbacks:

1. Rerunning the analysis several times might be computationally intensive if you use methods beyond OLS
2. OLS with interaction relies on correct model specification with all the consequences we discussed in the previous part

Elegant solution based on Double ML and OLS (1/3)

Recall from previous lectures that the ATE can be estimated as the mean of a pseudo-outcome:

$$\hat{\tau}_{ATE}^{AIPW} = \frac{1}{N} \sum_i \tilde{Y}_{i,ATE}$$

where

$$\tilde{Y}_{ATE} = \hat{m}(1,X) - \hat{m}(0,X) + \frac{W(Y - \hat{m}(1,X))}{\hat{e}(X)} - \frac{(1-W)(Y - \hat{m}(0,X))}{1 - \hat{e}(X)}$$

This is equivalent to specifying a "linear" regression model with pseudo-outcome and constant:

$$\tilde{Y}_{ATE} = \alpha + \epsilon$$

Estimating α by OLS yields $\hat{\alpha} = \hat{\tau}_{ATE}^{AIPW}$

Elegant solution based on Double ML and OLS (2/3)

Why should we stop at the model with only a constant?

Indeed, we can specify a linear model using our heterogeneity variable(s) G

$$\tilde{Y}_{ATE} = \alpha + \rho G + \epsilon$$

The cool thing is that we can interpret the coefficients in this model like we are used to (including hypothesis tests, Semenova and Chernozhukov, 2021)

The twist is that **we model the level of the effect, not the level of the outcome**

Elegant solution based on Double ML and OLS (3/3)

This has several practical advantages:

- Computationally less expensive than subgroup analyses (only one additional OLS, no new nuisance parameters)
- It is more flexible than specifying interaction terms in a linear model, as we flexibly adjust for confounding by ML methods

As $\tilde{Y}_{ATE}$ is an unbiased signal, i.e. $\mathbb{E}[\tilde{Y}_{ATE} \mid G = g] = \tau(g)$, we can either use

- OLS or series regression (Semenova and Chernozhukov, 2021)
- or kernel regression (Fan et al., 2022; Zimmert & Lechner, 2019)

to regress the pseudo-outcome $\tilde{Y}_{ATE}$ on low-dimensional G

Most importantly, the Neyman-orthogonality of $\tilde{Y}_{ATE}$ allows to apply standard statistical inference

Show that $\mathbb{E}[\tilde{Y}_{ATE} \mid G = g] = \tau(g)$

Recall from the AIPW identification results of the CAPO in the AIPW slide deck that $\mathbb{E}[Y(w) \mid X = x] = \frac{\mathbb{E}[D(w)Y \mid X=x]}{e_w(x)} = \mathbb{E}\left[\frac{D(w)Y}{e_w(x)} \middle| X = x\right]$ and appy this in line 3 to four:

$$
\begin{aligned}
&\mathbb{E}[\tilde{Y}_{ATE} \mid G = g] \\
&= \mathbb{E}\left[m(1,X) + \frac{W(Y - m(1,X))}{e(X)} - m(0,X) - \frac{(1-W)(Y - m(0,X))}{1 - e(X)} \middle| G = g\right] \\
&\overset{LIE}{=} \mathbb{E}\left[\underbrace{\mathbb{E}\left[m(1,x) + \frac{W(Y - m(1,x))}{e(x)} \middle| X = x\right]}_{\mathbb{E}[Y(1)|X=x]} - \underbrace{\mathbb{E}\left[m(0,x) + \frac{(1-W)(Y - m(0,x))}{1 - e(x)} \middle| X = x\right]}_{\mathbb{E}[Y(0)|X=x]} \middle| G = g\right] \\
&= \mathbb{E}[\mathbb{E}[Y(1) \mid X = x] - \mathbb{E}[Y(0) \mid X = x] \mid G = g] \\
&\overset{LIE}{=} \mathbb{E}[Y(1) - Y(0) \mid G = g] \\
&= \tau(g) \quad \square
\end{aligned}
$$

The law of iteration applications use that G is a function of X

GATE - example - subgroup analysis

Question: Do the effects of job training programs on employment differ for men and women?

Ingredients: $\tilde{Y}_{ATE}$ & OLS: $\tilde{Y}_{ATE} = \beta_0 + \beta_1 female + error$

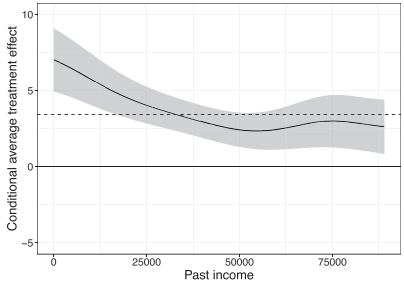
	Job search	Vocational	Computer	Language
Constant	-1.29***	3.82***	2.33***	3.40***
	(0.17)	(0.55)	(0.60)	(0.46)
Female	0.60**	-1.27	2.49***	-1.97**
	(0.25)	(0.87)	(0.85)	(0.77)

Source: Knaus (2022)

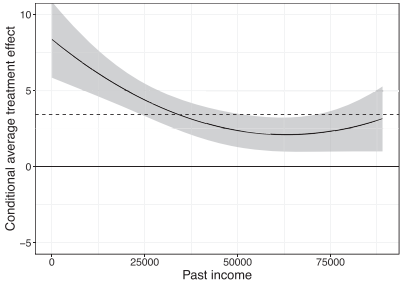
GATE - example - kernel/splines

Question: Do the effects of job training programs differ by past income?

Ingredients: $\tilde{Y}_{ATE}$ & kernel/spline regression: $\tilde{Y}_{ATE} = f(income) + error$



(e) Computer (Kernel)



(f) Computer (Spline)

Source: Knaus (2022)

Simulation notebook: Group Average Treatment Effects

Application notebook: Double ML for group average treatment effects

Predicting effects: The challenge

Predicting effects

The heterogeneity variables in the previous chapter were hand-crafted

Now we focus on the case where we aim for the most flexibel/ personalized/ individualized effect prediction CATE/IATE (only CATE from now on):

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

This is a conditional expectation and we know how to approximate that using machine learning 🥳

BUT the outcome is unobserved

$\Rightarrow$ Standard supervised ML cannot be applied directly 😞

$\Rightarrow$ Clever ideas needed

Captain obvious

The most straightforward idea is to use the identification result that we established for experiments and under IA3 (strong ignorability)

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] = \underbrace{\mathbb{E}[Y \mid W = 1, X = x]}_{\text{CEF in treated group}} - \underbrace{\mathbb{E}[Y \mid W = 0, X = x]}_{\text{CEF in control group}}$$

We know how to approximate such CEFs and this motivates two ”simple” algorithms

S-learner and T-learner

The S-learner proceeds as follows:

1. Use ML estimator of your choice to fit outcome model using X AND W in the full sample: $\hat{m}(W, X)$
2. Estimate CATE as $\hat{\tau}(x) = \hat{m}(1, x) - \hat{m}(0, x)$

The T-learner proceeds as follows:

1. Use ML estimator of your choice to fit model $\hat{m}(1, X)$ in treated subsample
2. Use ML estimator of your choice to fit model $\hat{m}(0, X)$ in control subsample
3. Estimate CATE as $\hat{\tau}(x) = \hat{m}(1, x) - \hat{m}(0, x)$

Why is this not the best idea? (1/2)

The prediction problems do not know of joint goal to approximate a difference

$\Rightarrow \hat{m}(1,x)$ minimizes $MSE(\hat{m}(1,x)) = \mathbb{E}[(\hat{m}(1,x) - m(1,x))^2]$

$\Rightarrow \hat{m}(0,x)$ minimizes $MSE(\hat{m}(0,x)) = \mathbb{E}[(\hat{m}(0,x) - m(0,x))^2]$

BUT what they should aim to minimize is

$$\begin{aligned} MSE(\hat{\tau}(x)) &= \mathbb{E}[(\hat{\tau}(x) - \tau(x))^2] \\ &= \mathbb{E}[(\hat{m}(1,x) - \hat{m}(0,x) - (m(1,x) - m(0,x)))^2] \\ &= \mathbb{E}[(\hat{m}(1,x) - m(1,x))^2] + \mathbb{E}[(\hat{m}(0,x) - m(0,x))^2] \\ &\quad - 2\,\mathbb{E}[(\hat{m}(1,x) - m(1,x))(\hat{m}(0,x) - m(0,x))] \\ &= MSE(\hat{m}(1,x)) + MSE(\hat{m}(0,x)) - 2MCE(\hat{m}(1,x), \hat{m}(0,x)) \end{aligned}$$

Lechner (2018) calls the additional term Mean Correlated Error (MCE)

Why is this not the best idea? (2/2)

The MCE tells us that positively correlated errors matter less

Example

$\hat{m}(1,x) = m(1,x) + 2$ and $\hat{m}(0,x) = m(0,x) + 2 \Rightarrow$ both make same error

$\Rightarrow MSE(\hat{m}(1,x)) = 4$ and $MSE(\hat{m}(0,x)) = 4$

But their CATE would be just on point: $MSE(\hat{\tau}(x)) = 4 + 4 - 2(2 \times 2) = 0$

On the other hand if errors go in different direction $\hat{m}(0,x) = m(0,x) - 2$

$\Rightarrow MSE(\hat{m}(1,x)) = 4$, $MSE(\hat{m}(0,x)) = 4$ but $MSE(\hat{\tau}(x)) = 16$

$\Rightarrow$ Not so clever yet, methods that are aware of the joint goal could provide improvements

Two strategies

Different ways to teach ML to target causal parameters:

1. **Modify** supervised ML methods to target causal effect estimation
   - Causal tree
   - Causal forest
2. **Combine** supervised ML methods to target causal effect estimation
   - R-learner
   - DR-learner

The first are method specific, the second are generic approaches/meta-learner

Many more methods, but too many to cover in one day

$\Rightarrow$ focus on those building on familiar ideas

Causal trees/forests

Causal Trees (1/3)

Recall from the supervised ML lecture that one representation of the splitting criterion for regression trees is to maximize squared predictions $\max \sum_i \hat{m}^{tree}(X_i)^2$

This is very helpful if we want to model effect heterogeneity

There is no need to observe the outcome we are predicting

Adapted to the causal setting our splitting criterion becomes $\max \sum_i \hat{\tau}^{tree}(X_i)^2$

It suffices to be able to calculate the average treatment effect in each candidate leaf

We know how to do that...

Causal Trees (2/3)

Parent node can be split along variable j at split point s in a left leaf $L(j, s) = \{X \mid X_j \leq s\}$ and a right leaf $R(j, s) = \{X \mid X_j > s\}$

We seek j and s that maximize the sum of squared leaf specific ATEs:

$$\max_{j,s} \left[\sum_{i:x_i \in L(j,s)} \hat{\tau}^2_{L(j,s)} + \sum_{i:x_i \in R(j,s)} \hat{\tau}^2_{R(j,s)} \right] \tag{1}$$

where we have different ways to calculate the leaf specific ATEs $\hat{\tau}_{L(j,s)}$ and $\hat{\tau}_{R(j,s)}$:

- In experiments, difference in outcome means between treated and controls
- Using, e.g. AIPW in case of confounding

Causal Trees (3/3)

This is the basic idea behind Athey and Imbens (2016)

The paper also extends the splitting criterion to anticipate honest estimation

They use an honest approach to guarantee valid inference after tree building

Cross-validation can be applied for pruning as in the standard case

Implemented in causalTree package for experiments

Causal Forest(s): a longer journey

The development of Causal Forests went through different stages:

- Causal Forest of Wager and Athey (2018) is an ensemble of Causal Trees (CT) and focused on the experimental setting
- Causal Forest of Athey, Tibshirani & Wager (2019) uses an approximation of the CT splitting rule for a binary random treatment but extends also to
  - observational settings
  - continuous treatments

The former might be considered as an interim technology

Athey et al. provide an excellent infrastructure for causal ML around the Causal Forest in the R package `grf`

Causal Forest and partially linear model

Today we think about Causal Forest (CF) in the following way:

CF estimates CATEs as a localized/individualized residual-on-residual regression

$$\hat{\tau}^{cf}(x) = \underset{\tau}{\arg\min} \left\{ \sum_{i=1}^{N} \alpha_i(x) \left[(Y_i - \hat{m}(X_i)) - \tau(x)(W_i - \hat{e}(X_i)) \right]^2 \right\}, \tag{2}$$

where $\alpha(x)$ are x-specific weights

This should look familiar, it estimates a localized version of the partially linear estimator

$\Rightarrow$ nuisance parameters $\hat{m}(X)$ and $\hat{e}(X)$ estimated in first step using cross-fitting or out-of-bag (a random forest specific way to ensure out-of-sample predictions)

Recall: Random Forest as weights

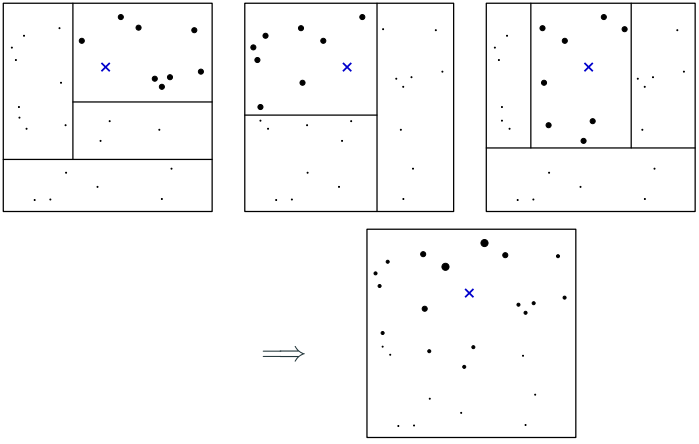


FIG. 1. *Illustration of the random forest weighting function. Each tree starts by giving equal (pos-itive) weight to the training examples in the same leaf as our test point x of interest, and zero weight to all the other training examples. Then the forest averages all these tree-based weightings, and effectively measures how often each training example falls into the same leaf as x.*

Source: Athey, Tibshirani & Wager (2019)

Splitting criterion uses influence function

What is the splitting criterion of causal forests that is applied in the tree building and eventually results in the weights for the RORR?

Causal forests use the following pseudo-outcome to place regression tree splits:

$$\rho = \left[\sum_i (W - \hat{e}(X))^2\right]^{-1} \left[(Y - \hat{m}(X)) - \hat{\tau}(W - \hat{e}(X))\right](W - \hat{e}(X))$$

This is the influence function of the partially linear estimator (see last lecture)

The splitting along the influence function is a general recipe to estimate heterogeneous parameters, e.g. in IV settings (Biewen & Kugler, 2021)

$\Rightarrow$ Generalized Random Forest

A note on statistical inference

Athey et al. (2019) show that the CATEs estimated using their CF are asymptotically normal and propose and implement an inference procedure

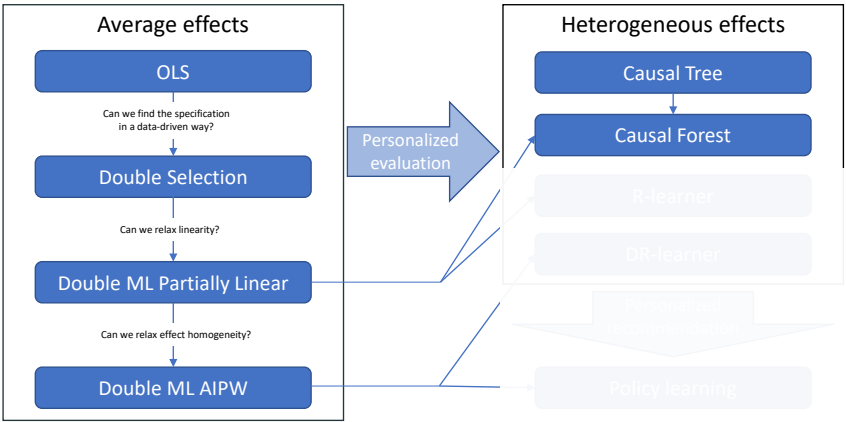
This is really nice and works theoretically when X is relatively small

You might see papers or presentations claiming that CF is so cool because it allows for high-dimensional X

Be aware that when it comes to the standard errors this is not in line with the theoretical analysis of the paper

The results require low-dimensional variables and even there we may need a lot of observations until the ”asymptotics kick in”

Causal botanics unlocked



Simulation notebook: Causal Tree and Causal Forest

Application notebook: Predicting effects (part 1)

Causal Forest Fun Shinyapp

Meta-learner

What are Meta-learner?

Meta-learner combine multiple supervised ML steps in a pipeline that outputs predicted CATEs

The common ones require the following steps:

1. Estimate nuisance parameters using suitable ML method
2. Plug them into a clever minimization problem targeting CATE
3. Solve the minimization problem using suitable ML method
4. Predict CATE using the model learned in 3

Most popular ML methods are suitable and can be applied in steps 1, 3 and 4

Like for standard prediction methods, statistical inference is usually not available

R-learner: idea

Recall the partially linear model, but now allowing for x-specific treatment effects:

$$Y(w) = \tau(X)w + g(X) + U_{Y(w)}; \;\; \mathbb{E}[U_{Y(w)} \mid W, X] = 0 \tag{3}$$

$$\Rightarrow Y = W\tau(X) + g(X) + U_{Y(W)} \tag{4}$$

$$\Rightarrow Y - m(X) = \tau(X)(W - e(X)) + U_{Y(W)} \tag{5}$$

This motivates the R-learner of Nie and Wager (2020):

$$\hat{\tau}^{rl}(X) = \underset{\tau}{\arg\min} \sum_{i=1}^{N} \left(Y_i - \hat{m}(X_i) - \tau(X_i)(W_i - \hat{e}(X_i))\right)^2 \tag{6}$$

with cross-fitted high-quality nuisance parameters from first step

But how to estimate it 🤔

R-learner: linear ML methods

An interesting option arises if we model the CATE as linear function: $\tau(X) = X'\beta$

$$\hat{\beta}^{rl} = \underset{\beta}{\arg\min} \sum_{i=1}^{N} \big(Y_i - \hat{m}(X_i) - \underbrace{(W_i - \hat{e}(X_i))X_i'}_{=\tilde{X}_i'} \beta\big)^2 = \underset{\beta}{\arg\min} \sum_{i=1}^{N} \big(Y_i - \hat{m}(X_i) - \tilde{X}_i'\beta\big)^2 \quad (7)$$

where $\tilde{X} = (W - \hat{e}(X))X$ are modified/pseudo-**covariates**

and $\hat{\tau}^{rl}(x) = x\hat{\beta}^{rl} \neq \tilde{x}\hat{\beta}^{rl}$ is the estimated CATE for a specific x

$\Rightarrow$ All the linear shrinkage estimators (Lasso and friends) can be applied

Remark: The nuisance parameters can still be estimated with non-linear ML

R-learner: generic ML methods

We rewrite (6) differently if we are not willing to impose linearity of the CATE:

$$\hat{\tau}^{rl}(X) = \arg\min_{\tau} \sum_{i=1}^{N} (W_i - \hat{e}(X_i))^2 \left(\frac{Y_i - \hat{m}(X_i)}{W_i - \hat{e}(X_i)} - \tau(X_i) \right)^2 \tag{8}$$

Every supervised ML model that is capable of dealing with weighted minimization problems can be used (neural nets, random forest, boosting, ...) with

- weights $(W - \hat{e}(X))^2$
- pseudo-outcome $\frac{Y - \hat{m}(X)}{W - \hat{e}(X)}$
- the unmodified covariates

Rewrite R-learner

$$
\begin{aligned}
\hat{\tau}^{rl}(X) &= \arg\min_{\tau} \sum_{i=1}^{N} \left(Y_i - \hat{m}(X_i) - \tau(X_i)(W_i - \hat{e}(X_i))\right)^2 \\
&= \arg\min_{\tau} \sum_{i=1}^{N} \frac{(W_i - \hat{e}(X_i))^2}{(W_i - \hat{e}(X_i))^2} \left(Y_i - \hat{m}(X_i) - \tau(X_i)(W_i - \hat{e}(X_i))\right)^2 \\
&= \arg\min_{\tau} \sum_{i=1}^{N} (W_i - \hat{e}(X_i))^2 \left(\frac{Y_i - \hat{m}(X_i) - \tau(X_i)(W_i - \hat{e}(X_i))}{W_i - \hat{e}(X_i)}\right)^2 \\
&= \arg\min_{\tau} \sum_{i=1}^{N} (W_i - \hat{e}(X_i))^2 \left(\frac{Y_i - \hat{m}(X_i)}{W_i - \hat{e}(X_i)} - \tau(X_i)\right)^2
\end{aligned}
$$

DR-learner

Recall or note that

$$\tau(x) = \mathbb{E}\Bigg[\underbrace{m(1,x) - m(0,x) + \frac{(W - e(X))(Y - m(W,X))}{e(X)(1 - e(X))}}_{\tilde{Y}_{ATE}} \Bigg| X = x\Bigg]$$

$\Rightarrow \mathbb{E}[\tilde{Y}_{ATE} \mid X]$ is a CEF of a (admittedly fancy looking) random variable

and we know how to approximate CEFs of random variables

$\Rightarrow$ The DR-learner of Kennedy (2020) uses $\tilde{Y}_{ATE}$ in a generic ML problem

$$\hat{\tau}^{dr}(X) = \arg\min_{\tau} \sum_{i=1}^{N} \left(\tilde{Y}_{i,ATE} - \tau(X_i)\right)^2 \tag{9}$$

We used the same "trick" already for GATE estimation

Discussion Meta-learner

Advantages:

- Very flexible
- Very individualized
- Predicted effects can be used for policy assignment (treat if $\hat{\tau}(x) > 0$)
- DR-learner naturally extends to multiple treatments

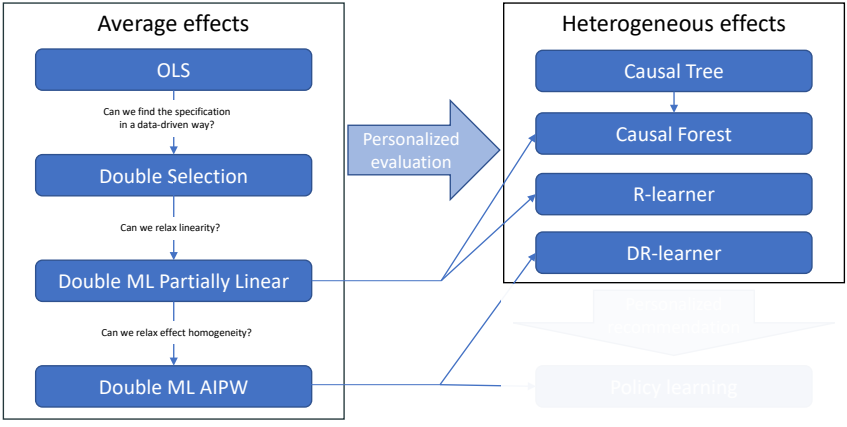
Disadvantages:

- No/little inference results (same problem as with standard ML)
- Hard to interpret (same problem as with standard ML)

Simulation notebook: Meta-learner

Application notebook: Predicting effects (part 2)

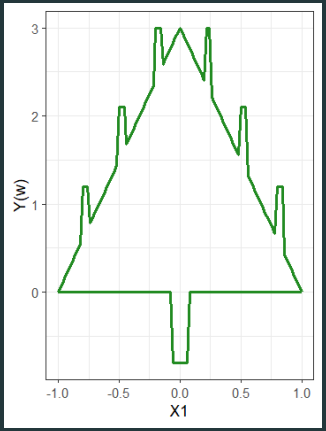
Heterogeneous effects unlocked



We could leverage concepts from previous lectures to make fast progress

Bonus assignment: Causal Christmas Tree Challenge

Send me your solutions for 5 bonus points until 22.12. via mail



Ceterum censeo a fancy method alone is not a credible identification strategy

$\Rightarrow$ separate identification and estimation