EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

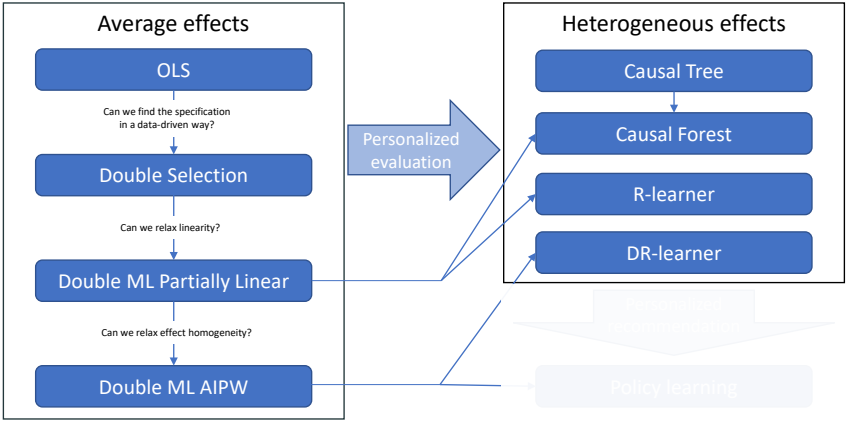
Causal Machine Learning

Heterogeneous effects with inference

Michael Knaus

March 13, 2023

State of the journey



We learned how to predict heterogeneous effects applying concepts developed for average effect estimation

Plan of this morning

What to do with all these predicted effects?

1. How to validate predicted CATEs?

2. How to describe/understand CATEs?

3. Can more aggregated effects be interesting? (GATE vs. IATE)

4. Even more on effect heterogeneity

Too much information

After applying Causal **insert here your favourite supervised ML** or **insert here your favourite letter (combination)**-learner, we end up with (at least) *N* flexibly estimated CATEs

These can be useful for decision making, but how to communicate them in papers/reports or to decision makers?

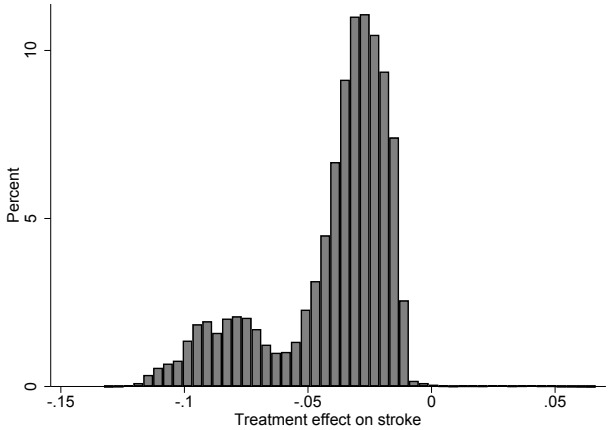
Nobody can digest a table with *N* effects

The first step is usually to plot the distribution

Many different ways...

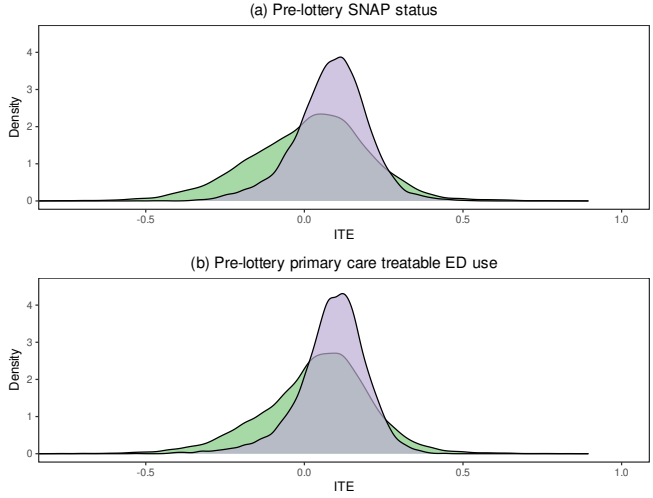
As histograms

Figure 4: Distribution of Treatment Effects Across VHA Patients



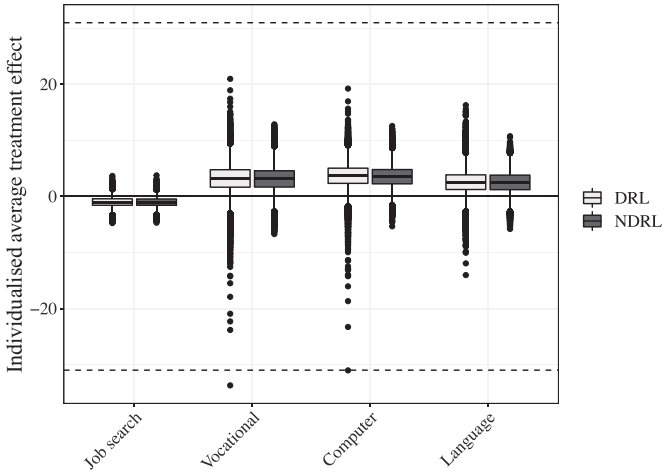
Source: Abaluck et al. (2020)

As density plots



Source: Denteh & Liebert, 2022

As boxplots



Source: Knaus (2022)

And now?

Okay, these are nice pictures and there seems to be heterogeneity

But what are we really looking at 🤔

In the next steps we ...

1. ... investigate if we found systematic effect heterogeneity or just noise
2. ... investigate how to understand what drives the heterogeneous effects
3. ... take a step back and look at alternatives that do not aim for the most individualized effects but allow for statistical inference

How to validate predicted CATEs?

Again space for clever ideas

Challenges:

- Due to the missing counterfactual, we can not benchmark our predicted effect against the true effect ⇒ no classic out-of-sample testing possible (unique to causal ML)
- Statistical inference for predicted CATE is not available or at least challenging (shared with supervised ML)

First steps: The literature on validation of CATEs is still in an early stage, especially when it comes to practice, but I want to introduce the ideas of Chernozhukov, Demirer, Duflo & Fernandez-Val (2017-2022) in this direction:

- Best Linear Predictor
- Sorted Group Average Treatment Effect

Best Linear Predictor - definition

Chernozhukov et al. (2017-2022) propose to look at the Best Linear Predictor (BLP) which is defined as the solution of the hypothetical regression of the true CATE on the demeaned predicted CATE:

Definition BLP

$$(\beta_1, \beta_2) = \underset{b_1, b_2}{\arg\min}\, \mathbb{E}\{[\tau(X) - b_1 - b_2 \underbrace{(\hat{\tau}(X) - \mathbb{E}[\hat{\tau}(X)])}_{\text{demeaned prediction}}]^2\}$$

where

- $\beta_1 = \mathbb{E}[\tau(X)] = ATE$ because of the demeaning
- $\beta_2 = \frac{Cov[\tau(X), \hat{\tau}(X)]}{Var[\hat{\tau}(X)]}$

Paradigm shift from most individualized effects to low dimensional summary statistics of effects

Best Linear Predictor - interpretation

$\beta_2 = 1$ if $\hat{\tau}(X) = \tau(X)$ (what we would like to see)

$\beta_2 = 0$ if $Cov[\tau(X), \hat{\tau}(X)] = 0$ this can have two reasons

1. $\tau(X)$ is constant (no heterogeneity to detect)
2. $\tau(X)$ is not constant but the estimator is not capable of finding it (bad estimator and/or not enough observations)

Therefore, testing $H_0 : \beta_2 = 0$ is a joint test of

(i) existence of heterogeneity

(ii) the estimators capability to find it

Best Linear Predictor - identification A

Chernozhukov et al. (2017-2022) show that the BLP parameters are identified in randomized experiments (known propensity score $e(X)$)

Strategy A: Weighted residual BLP

$$(\beta_1, \beta_2) = \underset{b_1, b_2}{\arg\min}\, \mathbb{E}\{\omega(X)[Y - b_1(W - e(X)) - b_2(W - e(X))(\hat{\tau}(X) - \mathbb{E}[\hat{\tau}(X)]) - a\tilde{X}]^2\}$$

where $\omega(X) = [e(X)(1 - e(X))]^{-1}$

$\tilde{X}$ is not required for identification, but contains optional functions of X to reduce estimation noise, e.g $[1, \hat{m}(0, X), e(X), e(X)\hat{\tau}(X)]$

See Appendix A of the paper for a detailed derivation

Best Linear Predictor - identification B

Chernozhukov et al. (2017-2022) show that the BLP parameters are identified in randomized experiments (known propensity score $e(X)$)

Strategy B: Horvitz-Thompson BLP

$$(\beta_1, \beta_2) = \underset{b_1, b_2}{\arg\min}\, \mathbb{E}\{[HY - b_1 - b_2(\hat{\tau}(X) - \mathbb{E}[\hat{\tau}(X)]) - aH\tilde{X}]^2\}$$

where $H = \frac{D - e(X)}{e(X)(1 - e(X))}$ are the familiar IPW weights

$\tilde{X}$ is not required for identification, but may be used to reduce estimation noise

See Appendix A of the paper for a detailed derivation

Best Linear Predictor - implementation

1. Split your sample in training and test set
2. (optional) Learn a model for $m(0, X) = E[Y \mid W = 0, X]$ in the training sample
3. Learn a model for $\tau(X)$ in the training sample
4. Predict $\hat{\tau}(X)$ (and $\hat{m}(0, X)$) in the test sample
5. Run the regression of strategy A and/or B in the test sample
6. Test $H_0 : \beta_2 = 0$ as usual

This will give you the BLP of one specific training/test split

Chernozhukov et al. (2017-2022) also show how to aggregate results from repeated splits

Best Linear Predictor - example

Chernozhukov et al. (2017-2022) evaluate the effect of an intervention in Indian villages that analyzes the effect of a bundle of immunization incentives (W) on "Number of children who completed the immunization schedule" (Y)

TABLE 3. BLP of Immunization Incentives Using Causal Proxies

Elastic Net		Neural Network	
ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
2.814	1.047	2.441	0.899
(1.087,4.506)	(0.826,1.262)	(0.846,3.979)	(0.685,1.107)
[0.004]	[0.000]	[0.004]	[0.000]

Notes: Medians over 250 splits. Median Confidence Intervals ($\alpha = .05$) in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternative in brackets.

Sorted Group Average Treatment Effect - definition

Chernozhukov et al. (2017-2022) also propose to look at the Sorted Group Average Treatment Effects (GATES) which are defined as:

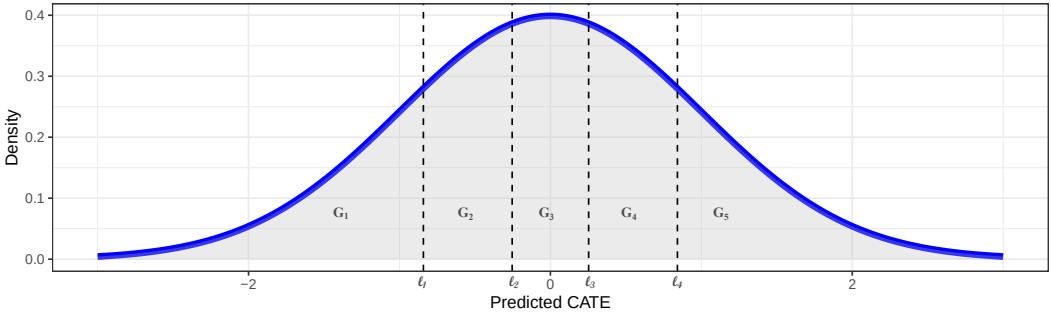
Definition GATES

$$\gamma_k = E[\tau(X) \mid G_k], \ k = 1, ..., K$$

where $G_k = \{\hat{\tau}(X) \in I_k\}$ with $I_k = [l_{k-1}, l_k)$ and $-\infty = l_0 < l_1 < ... < l_K = \infty$

In words: We slice the distribution of $\hat{\tau}(X)$ into K parts and are interested in the average effect of individuals within each slice

Sorted Group Average Treatment Effect - illustration



$\hat{\gamma}_k$ is then the mean of the predicted CATEs in the respective group G_k

Sorted Group Average Treatment Effect - interpretation

If the slices are build on the true CATE, we would observe the following monotonicity

$$\gamma_1^* \leq ... \leq \gamma_K^*$$

where the * indicates that the parameter builds on the true CATE

$\Rightarrow$ We expect to see the same monotonicity if $\hat{\tau}(X)$ provides a good approximation of $\tau(X)$

Situations where $\gamma_1, ..., \gamma_K$ are similar indicate that no systematic heterogeneity was detected

Sorted Group Average Treatment Effect - identification

Similar to the BLP, the GATES are identified using two different strategies

Strategy A: Weighted residual GATES

$$(\gamma_1, ..., \gamma_K) = \arg\min_g \mathbb{E}\{\omega(X)[Y - \sum_k g_k(W - e(X))\mathbb{1}[G_k] - a\tilde{X}]^2\}$$

Strategy B: Horvitz-Thompson GATES

$$(\gamma_1, ..., \gamma_K) = \arg\min_g \mathbb{E}\{[HY - \sum_k g_k\mathbb{1}[G_k] - aH\tilde{X}]^2\}$$

Again covariates $\tilde{X}$ are not required for identification, but reduce estimation noise

See Appendix A of the paper for a detailed derivation

Sorted Group Average Treatment Effect - implementation

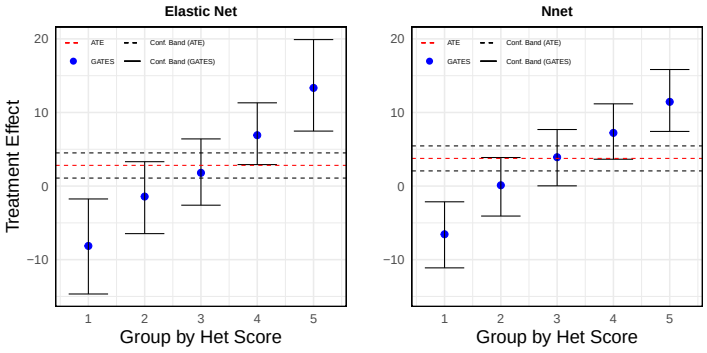
1. Split your sample in training and test set
2. (optional) Learn a model for $m(0, X) = E[Y \mid W = 0, X]$ in the training sample
3. Learn a model for $\tau(X)$ in the training sample
4. Predict $\hat{\tau}(X)$ (and $\hat{m}(0, X)$) in the test sample
5. Sort $\hat{\tau}(X)$ and create K slices
6. Run the regression of strategy A and/or B in the test sample
7. Test for example $H_0 : \hat{\gamma}_K - \hat{\gamma}_1 = 0$

This will give you the BLP of one specific training/test split

Chernozhukov et al. (2017-2022) also show how to aggregate results from repeated splits, but we focus on single splits

Sorted Group Average Treatment Effect - example

FIGURE 5. GATES of Immunization Incentives



Notes: GATES of Immunization Incentives, based upon Causal Learners. Median point estimates and Median confidence interval ($\alpha = .05$) in parenthesis, over 250 splits.

Sorted Group Average Treatment Effect - example

TABLE 4. GATES of 20% Most and Least Affected Groups

	Elastic Net			Nnet		
	20% Most (G_5)	20% Least (G_1)	Difference	20% Most (G_5)	20% Least (G_1)	Difference
GATE	13.230	-8.000	21.60	11.210	-6.551	18.13
$\gamma_k := \widehat{\mathrm{E}}[s_0(Z) \mid G_k]$	(8.219,18.67)	(-13.41,-2.574)	(13.70,29.74)	(7.721,14.47)	(-10.37,-2.786)	(12.84,23.52)
	[0.000]	[0.009]	[0.000]	[0.000]	[0.002]	[0.000]
Control Mean	2.19	12.68	-10.56	1.19	10.32	-9.17
$:= \widehat{\mathrm{E}}[b_0(Z) \mid G_k]$	(1.27,3.06)	(11.73,13.59)	(-11.84,-9.38)	(0.44,1.87)	(9.65,11.02)	(-10.17,-8.14)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]

Notes: Medians over 250 splits. Median confidence interval ($\alpha = .05$) in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternative in brackets.

Generalizations

Chernozhukov et al. (2017-2022) focus on experiments

However, the ideas of BLP and GATES extend to any case where we have access to a pseudo-outcome that is an unbiased signal of the parameter of interest

For example, we can use $\tilde{Y}_{ATE}$ because $\mathbb{E}[\tilde{Y}_{ATE} \mid X] = \tau(X)$ to evaluate CATE under unconfoundedness in the following regressions

- BLP: $(\beta_1, \beta_2) = \arg\min_{b_1, b_2} \mathbb{E}\{[\tilde{Y}_{ATE} - b_1 - b_2(\hat{\tau}(X) - \mathbb{E}[\hat{\tau}(X)])]^2\}$
- GATES: $(\gamma_1, ..., \gamma_K) = \arg\min_g \mathbb{E}\{[\tilde{Y}_{ATE} - \sum_k g_k \mathbb{1}[G_k]]^2\}$

GATES are even more general, we can apply any suitable estimator for the average effect within each data slice

How to describe/understand CATEs?

What ”drives” heterogeneity?

If BLP and GATES indicate that our algorithms detect systematic effect heterogeneity, we usually would like to understand which covariates are most predictive

However, the causal ML methods do not produce seemingly easy to interpret OLS outputs or something similar

I am currently aware of two options:

1. Rely on classic variable importance measures from the supervised ML literature inheriting all their strengths and weaknesses
2. Run a Classification Analysis of Chernozhukov et al. (2017-2022)

Classification Analysis

Definition BLP

Classification Analysis (CLAN) compares the covariate values of the "least affected group" G_1 with the "most affected group" G_K defined for the GATES:

$$\delta_K - \delta_1$$

where $\delta_1 = \mathbb{E}[X \mid G_1]$ and $\delta_K = \mathbb{E}[X \mid G_K]$

This can be achieved by simple mean comparison in the test sample for a single train/test split

Again Chernozhukov et al. show how to aggregate results from repeated splits

Classification Analysis - example

TABLE 5. CLAN of Immunization Incentives

	Elastic Net			Nnet		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
Number of vaccines	2.187	2.277	-0.081	2.174	2.285	-0.112
to pregnant mother	(2.115,2.259)	(2.212,2.342)	(-0.180,0.015)	(2.111,2.234)	(2.224,2.345)	(-0.202,-0.028)
	-	-	[0.190]	-	-	[0.019]
Number of vaccines	4.077	4.639	-0.562	4.264	4.734	-0.490
to child since birth	(3.858,4.304)	(4.444,4.859)	(-0.863,-0.260)	(4.091,4.434)	(4.549,4.900)	(-0.739,-0.250)
	-	-	[0.001]	-	-	[0.000]
Fraction of children	0.998	1.000	-0.002	1.000	1.000	0.000
received polio drops	(0.995,1.001)	(0.997,1.003)	(-0.006,0.002)	(1.000,1.000)	(1.000,1.000)	(0.000,0.000)
	-	-	[0.683]	-	-	[0.943]
Number of polio	2.955	2.993	-0.037	2.965	2.998	-0.032
drops to child	(2.935,2.974)	(2.976,3.010)	(-0.063,-0.010)	(2.953,2.977)	(2.985,3.010)	(-0.049,-0.016)
	-	-	[0.013]	-	-	[0.000]
Fraction of children	0.803	0.926	-0.121	0.908	0.927	-0.027
received immunization card	(0.754,0.851)	(0.882,0.969)	(-0.187,-0.054)	(0.881,0.932)	(0.898,0.959)	(-0.059,0.007)
	-	-	[0.001]	-	-	[0.217]
Fraction of children received	0.133	0.243	-0.106	0.126	0.260	-0.131
Measles vaccine	(0.097,0.169)	(0.209,0.276)	(-0.153,-0.056)	(0.095,0.159)	(0.228,0.291)	(-0.176,-0.085)
by 15 months of age	-	-	[0.000]	-	-	[0.000]
Fraction of children received	0.293	0.399	-0.110	0.289	0.433	-0.142
Measles vaccine	(0.246,0.338)	(0.358,0.444)	(-0.174,-0.045)	(0.246,0.331)	(0.391,0.475)	(-0.206,-0.084)
at credible locations	-	-	[0.002]	-	-	[0.000]

Can more aggregated effects be interesting? (GATE vs. IATE)

Maybe some middle-ground

We jumped from the most aggregated effect τ_{ATE} directly to the most disaggregated effect $\tau(X)$

However, there might be some interesting middle-ground

To conceptualize this, it is useful to distinguish between two types of CATEs:

- Group ATE (GATE) for some groups G that are defined using X

  $\Rightarrow \tau(G) = E[Y(1) - Y(0) \mid G]$
- Individualized ATE (IATE) what we called CATE so far

  $\Rightarrow \tau(X) = E[Y(1) - Y(0) \mid X]$

Groups for GATEs are usually chosen by researchers based on substantive grounds, while IATEs provide the most personalized effect possible

Examples of GATEs

Classic mutually exclusive subgroups, like $G = \{female, male, ...\}$, $G = \{citizen, foreigner\}$, $G = \{age < 50, age \geq 50\}$, $G = \{G_1, ..., G_K\}$, ...

Single continuous variable, like $G = age$, $G = income$, $G = \hat{\tau}(X)$, ...

Or any function or subset of X

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

Estimation of GATEs

Remember the DR-learner that used the pseudo-outcome $\tilde{Y}_{ATE}$ in a generic regression?

GATEs can be estimated as a special case of the DR-learner where we 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

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

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)

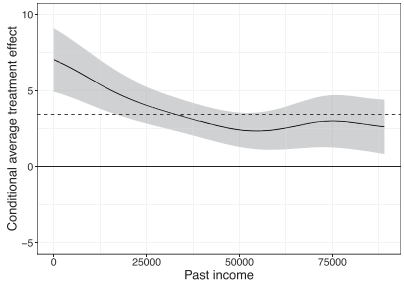
GATE - example - OLS

	Job search	Vocational	Computer	Language
Constant	-0.48	4.46*	4.93**	6.07***
	(0.70)	(2.40)	(2.37)	(2.09)
Female	0.25	-2.12**	1.89**	-1.61**
	(0.27)	(0.92)	(0.90)	(0.80)
Age	0.02	0.09*	0.05	-0.07
	(0.01)	(0.05)	(0.05)	(0.05)
Foreigner	0.50*	1.60*	-1.20	-2.77***
	(0.27)	(0.90)	(0.96)	(0.74)
Medium employability	-0.65*	-1.64	-2.36*	-0.78
	(0.37)	(1.18)	(1.22)	(0.99)
High employability	-1.03**	-3.13**	-3.15*	-0.47
	(0.51)	(1.55)	(1.71)	(1.53)
Past income in CHF 10,000	-0.26***	-0.62***	-0.39**	0.31*
	(0.06)	(0.23)	(0.19)	(0.18)
F-statistic	6.95***	4.12***	3.35***	5.22***

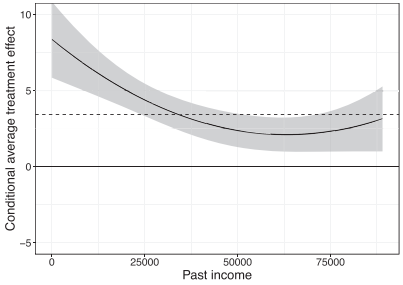
GATE - example - kernel/splines

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

Even more on effect heterogeneity

More methods

- BART (Hill, 2011; Hahn, Murray & Carvalho, 2020)
- Causal Boosting/MARS, ... (Powers, Qian, Jung, Schuler, Shah, Hastie & Tibshirani, 2019)
- Dragonnet (Shi, Blei & Veitch 2019)
- Modified Causal Forest (Lechner & Mareckova, 2022)
- Orthogonal Random Forest (Oprescu, Syrgkanis & Wu, 2019)
- TARNet (Shalit, Johansson & Sontag 2019)
- X-learner (Künzel, Sekhon, Bickel & Yu, 2019)

and many many more

More on validation

- Rank-weighted average treatment effect (RATE) (Yadlowsky, Fleming, Shah, Brunskill & Wager, 2021)
- Calibration Error for Heterogeneous Treatment Effects (Xu & Yadlowsky, 2022)
- More on GATES in experiments (Imai & Li, 2022)

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

$\Rightarrow$ separate identification and estimation