Machine Learning
for Economists

Effect Heterogeneity

Anthony Strittmatter

Outline

- Conditional Average Treatment Effects
- Application: Swiss Job Search Program
- Causal Forest
- Application: Summer Jobs
- Useful References

Conditional Average Treatment Effects (CATEs)

- CATEs:

$$\delta(x) = E[Y_i(1) - Y_i(0)|X_i = x]$$

- ATEs:

$$\delta = E[E[Y_i(1) - Y_i(0)|X_i = x]]$$

- Group Average Treatment Effects (GATEs):

$$\delta(g) = E[\delta(x)|G_i = g]$$

where the groups g can be defined based on exogenous or endogenous variables

- Examples of GATEs:
  - Aggregate by gender: $\delta(m) = E[\delta(x)|Male]$ and $\delta(f) = E[\delta(x)|Female]$
  - Aggregate by earnings-quantile-range $[y_{floor}(\tau), y_{ceil}(\tau)]$:

$$\delta(\tau) = E[\delta(x)|y_{floor}(\tau) \leq Y_i < y_{ceil}(\tau)]$$

Application: Swiss Job Search Program

- Content: Learn how to search and apply for a job
- Goal: Improve matching process
- Duration approximately 3 weeks
- Class room training
- Private providers
- Participants should continue active job search during the programme
- Yearly expenditures approximately 100 million CHF

Reference: Knaus, Lechner, and Strittmatter (2017)

Linked Unemployed-Caseworker Data

- Combination of social security, caseworker questionnaire, and regional data
- All registered unemployed in the year 2003 (12,000 participants and 72,000 controls)
- Outcome: Months employed after participation begins
- Treatment: First participation in a job search programme during the first six months of unemployment
- Caseworkers can assign unemployed persons to job search programmes (mostly) based on subjective measures

$\Rightarrow$ Non-random selection into participation

Rich Controls

- Unemployed: residence status, qualification, gender, education, language skills, employment history, profession, job position, industry of last job, desired occupation and industry, subjective employability rate by their caseworker, etc.
- Caseworker: age, gender, tenure, education, employment history, cooperativeness, etc.
- Region: language, population size of municipalities, the cantonal unemployment rate, etc.

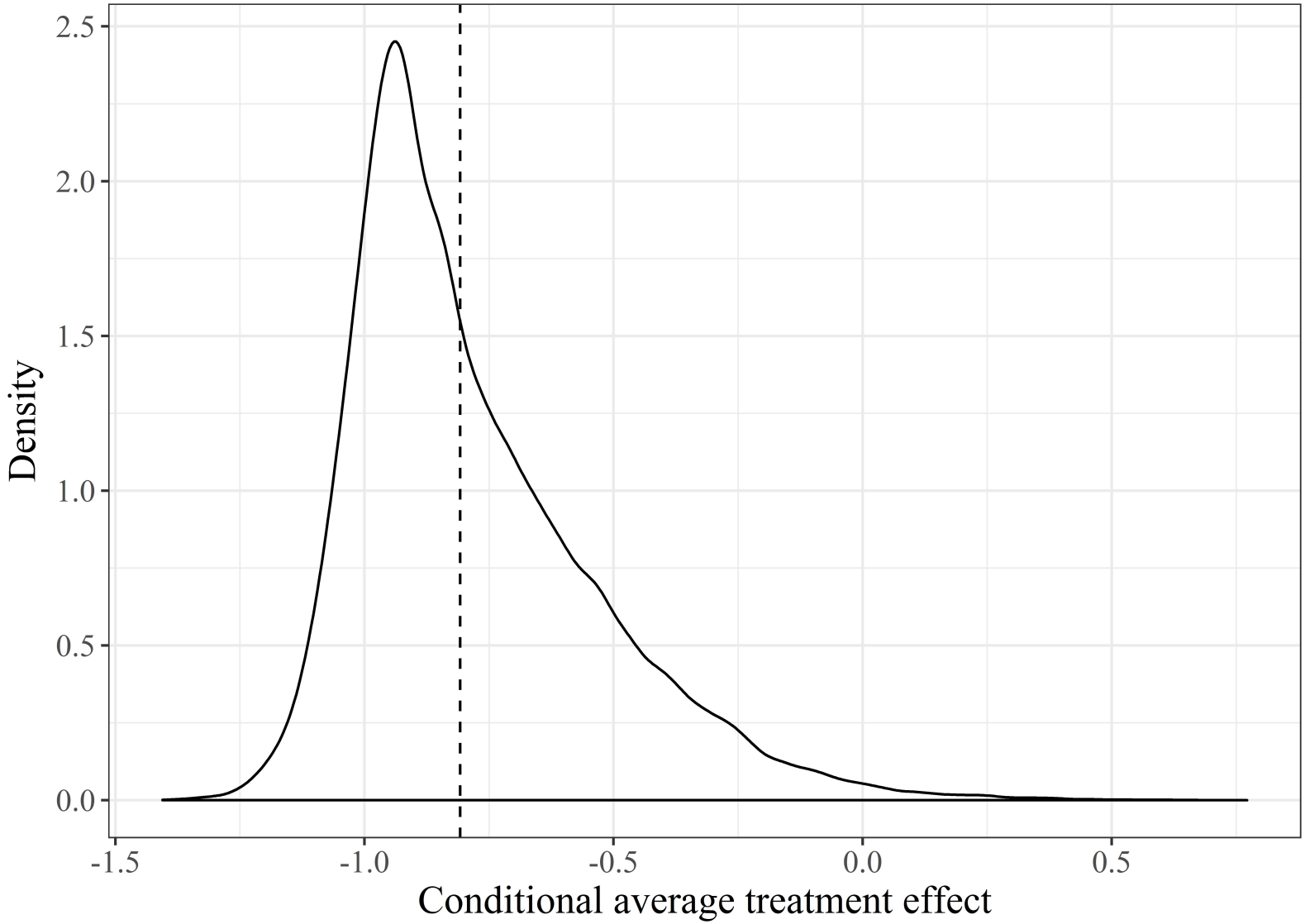
$\rightarrow$ Approximately 120 covariates and additionally interactions and polynomials

Aggregated Average Effects

Months employed since start of participation	ATE		ATET	
	Coef.	S.E.	Coef.	S.E.
	(1)		(2)	
During first 6 months	-0.80***	(0.02)	-0.82***	(0.02)
During first 12 months	-1.10***	(0.05)	-1.13***	(0.04)
During first 31 months	-1.14***	(0.14)	-1.20***	(0.13)
During months 25-31	-0.007	(0.03)	-0.011	(0.03)

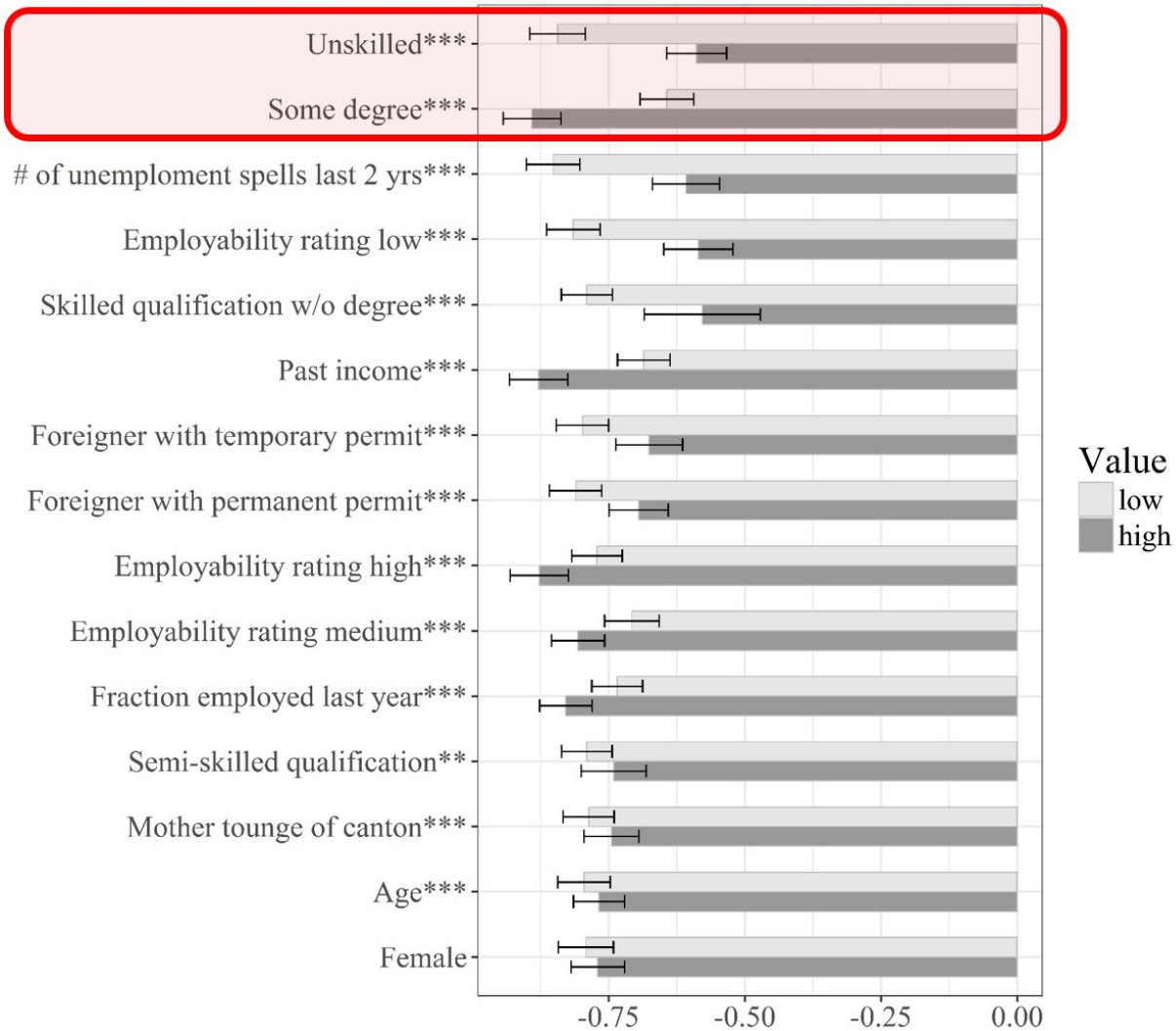
We obtain standard errors (S.E.) from a clustered bootstrap at caseworker level with 4,999 replications. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively.

Distribution of Predicted CATEs



Kernel smoothed distribution of average predicted individual effects. Gaussian kernel with bandwidth 0.02, chosen by Silverman's rule-of-thumb. The dashed vertical line shows the ATE.

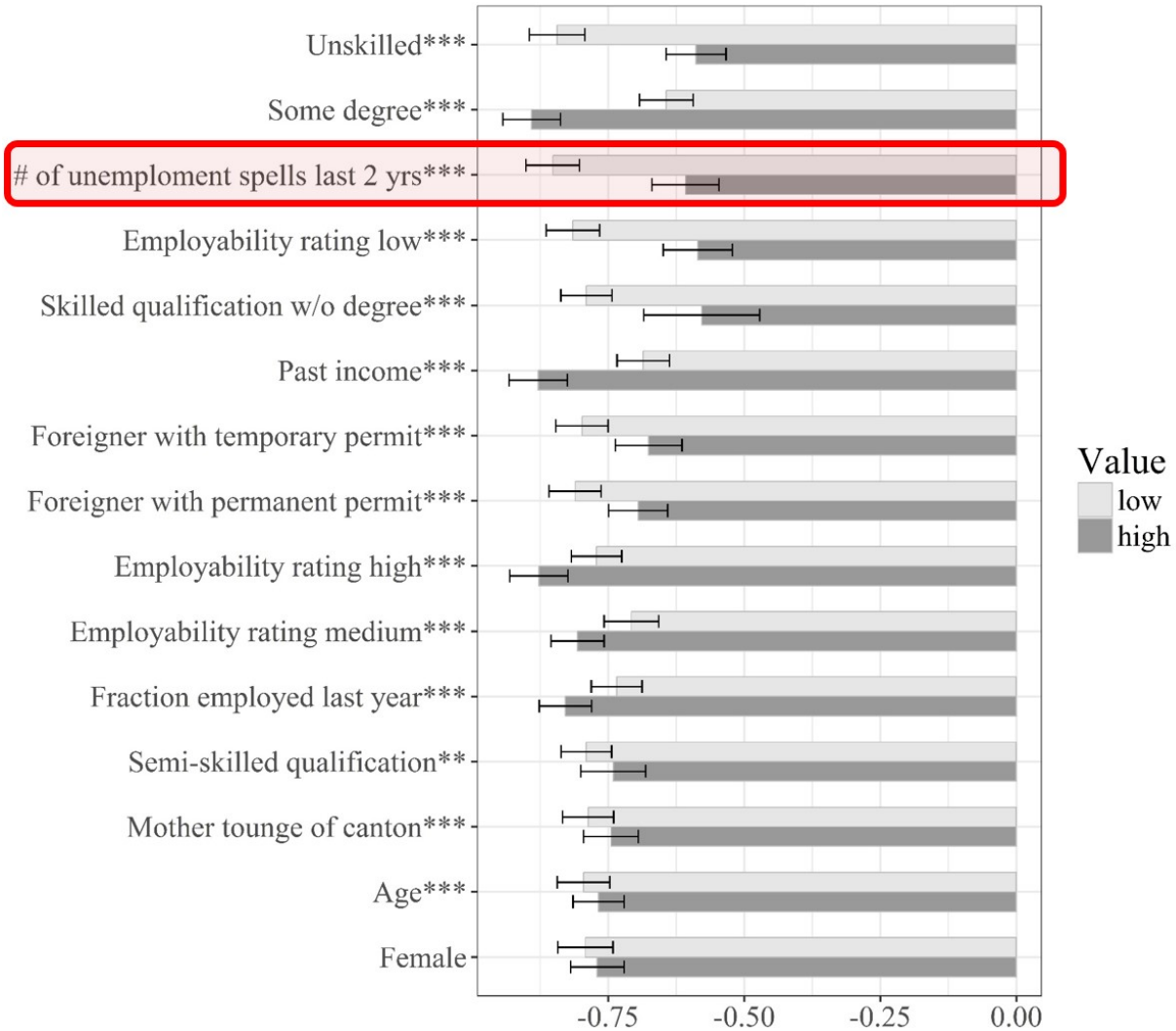
Characteristics of Unemployed Persons



Large heterogeneity by education:
Highly educated suffer much more

CATEs by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary.

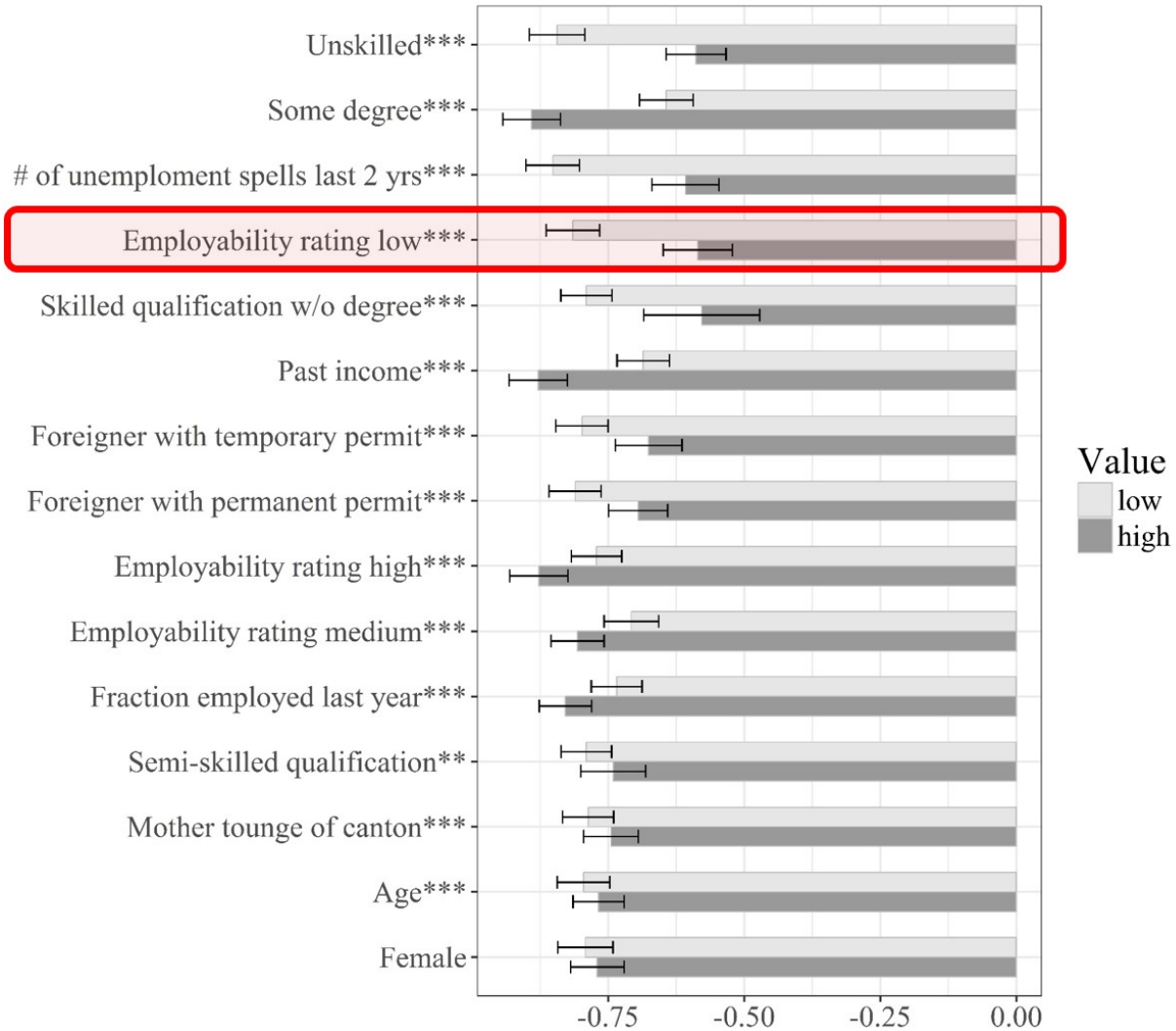
Characteristics of Unemployed Persons



Large heterogeneity by previous labor market success:
Never unemployed suffer much more

CATEs by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary.

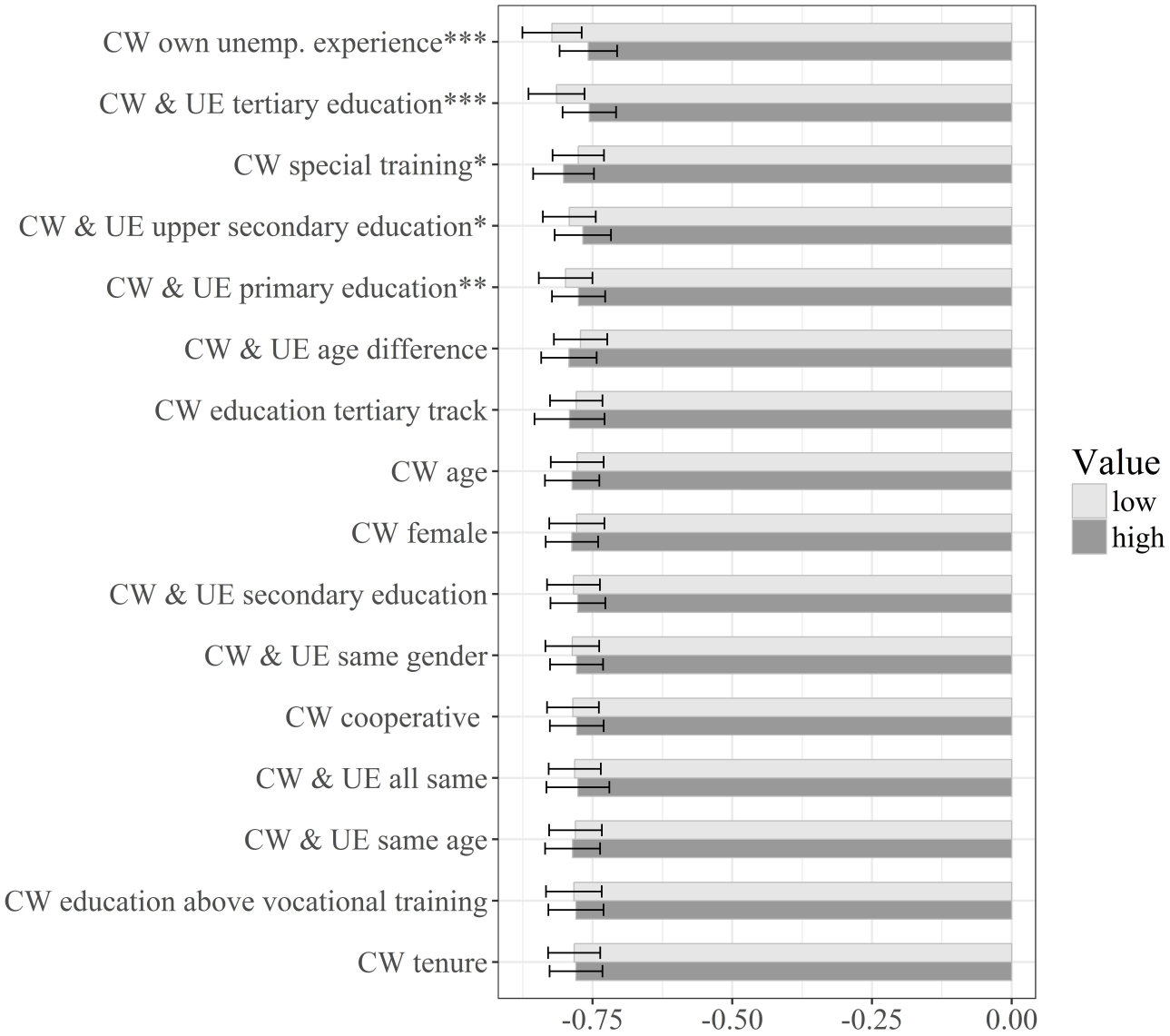
Characteristics of Unemployed Persons



Large heterogeneity by employability rating: Unemployed with low employability rating suffer much less

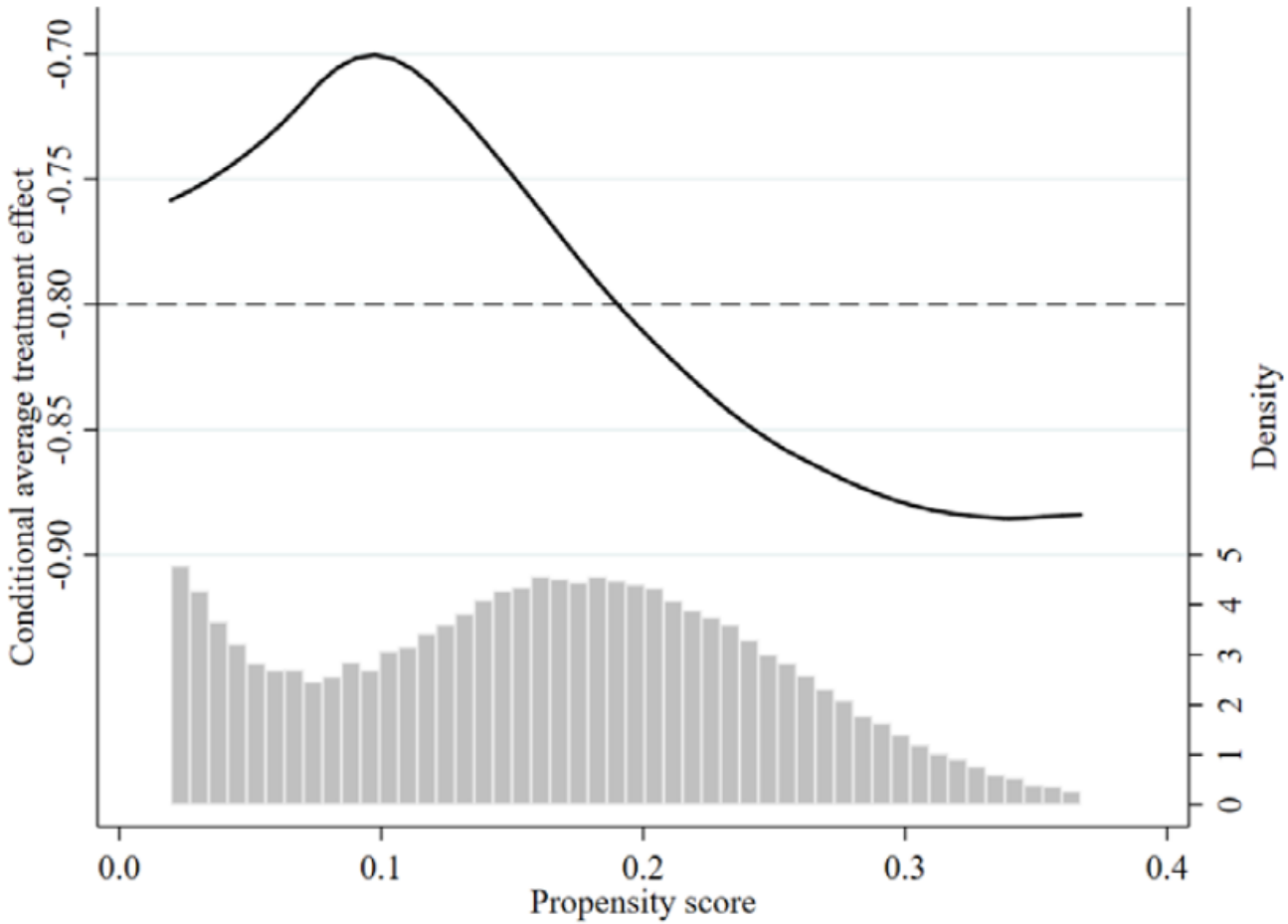
CATEs by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary.

Caseworker Characteristics



CATEs by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary.

Correlation of the propensity score and the CATEs



Kernel smoothed regression of propensity score and CATEs. Local constant kernel regression used with Gaussian kernel and bandwidth 0.02. The dashed horizontal line shows the ATE. The grey bars show the histogram of the propensity score.
See Zimmert and Lechner (2019) for valid inference procedure.

Generalised Causal Forest (GCF)

- Consider the random effects model $Y_i = D_i\delta(X_i) + \varepsilon_i$ with

$$\delta(x) = Var(D_i|X_i = x)^{-1} Cov(Y_i, D_i|X_i = x)$$

- Estimator

$$\hat{\delta}(x) = \left(\sum_{i=1}^{N} \alpha_i(x)(D_i - \bar{D}_\alpha)^2 \right)^{-1} \sum_{i=1}^{N} \alpha_i(x)(D_i - \bar{D}_\alpha)(Y_i - \bar{Y}_\alpha)$$

  with $\bar{D}_\alpha = \sum_{i=1}^{N} \alpha_i(x) D_i$ and $\bar{Y}_\alpha = \sum_{i=1}^{N} \alpha_i(x) Y_i$

- The weights $\alpha_i(x)$ are obtained with the generalised causal forest algorithm
  - Weights for causal tree g:

$$\alpha_{ig}(x) = \frac{1\{X_i \in l_j(x, d, \pi_g)\}}{\sum_{i=1}^{N} 1\{X_i \in l_j(x, d, \pi_g)\}}$$

  - Causal forest weights:

$$\alpha_i(x) = \frac{1}{G} \sum_{i=1}^{G} \alpha_{ig}(x)$$

Reference: Athey, Tibshirani, and Wager (2018)

Algorithm

- Optimal optimisation criteria:

$$\max \Delta(C_1, C_2) = \frac{N_{C_1} N_{C_2}}{N_P} \left(\hat{\delta}_{C_1} - \hat{\delta}_{C_2} \right)^2$$

  $\rightarrow$ Requires the estimation of $\hat{\delta}_{C_1}$ and $\hat{\delta}_{C_2}$ at each candidate split

- Computational efficient optimisation algorithm

  (1) **Labelling step:** Calculate $\hat{\delta}_P$, $A_P = Var_P(D_i)$, and the pseudo-outcome

$$p_i = A_P^{-1}(D_i - \bar{D}_P)(Y_i - \bar{Y}_P - (D_i - \bar{D}_P)\hat{\delta}_P)$$

  (2) **Regression step:**

$$\max \tilde{\Delta}(C_1, C_2) = \frac{1}{N_{C_1}} \left(\sum_{i:X_i \in C_1} p_i \right)^2 + \frac{1}{N_{C_2}} \left(\sum_{i:X_i \in C_2} p_i \right)^2$$

  (3) Relabel child nodes to parent nodes and repeat (1) and (2) until stopping criteria of tree is reached

  (4) Calculate weights $\alpha_{ig}(x)$ and build forest by repeating (1)-(3) with different subsamples and covariates

Pseudo-Outcome

- Assume D_i is binary and randomly assigned and denote $p = \bar{D}_P = Pr(D_i = 1) = Pr(D_i = 1|C_j)$

$$
\begin{aligned}
E[p_i|C_j] =& E\left[\frac{D_i - p}{p(1-p)}(Y_i - \bar{Y}_P - (D_i - p)\hat{\delta}_P)\middle| C_j\right] \\
=& E\left[\frac{1}{p}(Y_i - \bar{Y}_P - (1-p)\hat{\delta}_P)\middle| C_j, D_i = 1\right] p \\
& - E\left[\frac{1}{(1-p)}(Y_i - \bar{Y}_P + p\hat{\delta}_P)\middle| C_j, D_i = 0\right](1-p) \\
=& E\left[Y_i - \bar{Y}_P - (1-p)\hat{\delta}_P\middle| C_j, D_i = 1\right] - E\left[Y_i - \bar{Y}_P + p\hat{\delta}_P\middle| C_j, D_i = 0\right] \\
=& E\left[Y_i(1) - \bar{Y}_P\middle| C_j, D_i = 1\right] - E\left[Y_i(0) - \bar{Y}_P\middle| C_j, D_i = 0\right] - \hat{\delta}_P \\
=& E\left[Y_i(1) - Y_i(0)|C_j\right] - \hat{\delta}_P
\end{aligned}
$$

- Difference between approximated causal effect at child and parent node
- Pseudo-outcome is updated at each parent node

Local Centering

- Generalised causal forest is targeted to find maximum heterogeneity in pseudo-outcome
- But not specifically designed to account for selection into treatment (even though deep causal forests correct automatically for some extent of selection)
- Define centred variables

$$\tilde{Y}_i = Y_i - \hat{\mu}(X_i)$$

and

$$\tilde{D}_i = Y_i - \hat{p}(X_i)$$

with $\hat{\mu}(x) = \hat{E}[Y_i|X_i = x]$ and $\hat{p}(x) = \hat{E}[D_i|X_i = x]$

- Apply generalised causal forest algorithm to $\tilde{Y}_i$ and $\tilde{D}_i$ instead of Y_i and D_i

$\rightarrow$ Orthogonalisation

Local Centering (cont.)

MSE from Simulation

Confounding	Heterogeneity	K	N	GCF	Centred GCF
No	No	10	800	0.85	0.87
No	No	10	1,600	0.58	0.59
No	No	20	800	0.92	0.93
No	No	20	1,600	0.52	0.52
Yes	No	10	800	1.12	0.27
Yes	No	10	1,600	0.80	0.20
Yes	No	20	800	1.17	0.17
Yes	No	20	1,600	0.95	0.11
Yes	Yes	10	800	1.92	0.91
Yes	Yes	10	1,600	1.51	0.62
Yes	Yes	20	800	1.92	0.93
Yes	Yes	20	1,600	1.55	0.57

Note: K is the number of covariates in the simulation

Source: Athey, Tibshirani, and Wager (2018)

Asymptotic Properties for Causal Forest

- Minimum subsample size S is scaled N^{β} with $\beta_{min} < \beta < 1$
- CATEs are consistent and asymptotically normal

$$(\hat{\delta}(x) - \delta(x)) / \sqrt{Var(\delta(x))} \xrightarrow{d} N(0,1)$$

- Infinitesimal Jackknife
  - $Q_{ig} = 1$ when observation i is used to built tree g and $Q_{ig} = 0$ otherwise
  - Calculate the covariance $Cov(\hat{\delta}(x), Q_{ig})$ across all trees $g = 1, ..., G$
  - Variance estimator:

$$\hat{V}(x) = \frac{N-1}{N} \left(\frac{N}{N-S} \right)^2 \sum_{i=1}^{N} Cov(\hat{\delta}(x), Q_{ig})^2$$

  - $(N/(N-S)))^2$ is a finite sample correction for subsampling
  - $\hat{V}(x) \xrightarrow{p} Var(\delta(x))$

Reference: Wager and Athey (2018)

Advantages and Disadvantages of Causal Forest

Advantages:

- D_i can be binary or continuous
- Asymptotic properties for CATEs available
- Variance estimator available
- Extensions to IV and quantiles available

Disadvantages:

- Best suited for experiments
- We have to assume that subsample sizes do not get too small, because otherwise asymptotic results break down

Application: Summer Jobs

Davis and Heller (2017) : “Using Causal Forests to Predict Treatment Heterogeneity:An Application to Summer Jobs”, AER P&P

- Chicago’s One Summer Plus (OSP) program conducted in 2012 and 2013
- OSP provides disadvantaged youth ages 14-22 with 25 hours a week of employment, an adult mentor, and some other programming
- Participants are paid Chicago’s minimum wage ($8.25 at the time)
- Previous study finds on average 43 percent reduction in violent-crime arrests in the 16 months after random assignment (Heller, 2014)
- **Outcomes**:
  - Violent-crime arrests within two years of random assignment
  - Employment during the six quarters after the program
- **Covariates:**
  age, gender, education, ethnicity, criminal history, employment history, regional unemployment rate, regional median income

Reference: Davis and Heller (2017)

Application: Summer Jobs (cont.)

Estimation target GATEs:

$$\hat{\delta}(+) = E[\hat{\delta}(X_i)|\hat{\delta}_{tr}(X_i) > 0] \text{ and } \hat{\delta}(-) = E[\hat{\delta}(X_i)|\hat{\delta}_{tr}(X_i) \leq 0]$$

	No. arrests	Employment
	In-sample	
$\hat{\delta}(+)$	0.22	0.19
	(0.05)	(0.03)
$\hat{\delta}(-)$	-0.05	-0.14
	(0.02)	(0.03)
H_0: p-val	0.00	0.00
	Out-of-sample	
$\hat{\delta}(+)$	-0.01	0.08
	(0.05)	(0.03)
$\hat{\delta}(-)$	-0.02	-0.01
	(0.02)	(0.03)
H_0: p-val	0.77	0.02

Source: Davis and Heller (2017)

Application: Summer Jobs (cont.)

Estimation target GATEs:

$$\hat{\delta}(+) = E[\hat{\delta}(X_i)|\hat{\delta}_{tr}(X_i) > 0] \text{ and } \hat{\delta}(-) = E[\hat{\delta}(X_i)|\hat{\delta}_{tr}(X_i) \leq 0]$$

	No. arrests	Employment
	Adjusted In-sample	
$\hat{\delta}(+)$	0.06	0.05
	(0.04)	(0.03)
$\hat{\delta}(-)$	-0.02	-0.04
	(0.02)	(0.03)
H_0: p-val	0.41	0.02

Source: Davis and Heller (2017)

Additional References

- Chernozhukov, Demirer, Duflo, Fernández-Val (2019): “Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments”, download
- Chernozhukov, Fernández-Val, Luo (2018): “The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages”, download
- Nie and Wager (2019): “Quasi-Oracle Estimation of Heterogeneous Treatment Effects”, download
- Lee, Okui, Whang (2017): “Doubly robust uniform confidence band for the conditional average treatment effect function”, download