Effect or Treatment Heterogeneity?

Policy Evaluation with Aggregated and Disaggregated Treatments

Michael Knaus
with **Phillip Heiler** (Aarhus University)

Causal Data Science Meeting 2022

Motivation

Researchers and practitioners in business, economics, medicine and beyond are often interested in heterogeneous effects of ads/policies/treatments/...

$\Rightarrow$ Who benefits or loses by how much $\Rightarrow$ personalized ads, policy, medicine, ...

$\Rightarrow$ Much is known about **how** to estimate heterogeneous effects

- Classic subgroup analysis or interaction terms (standard)
- A-/DR-/R-/S-/T-/X-/...-learner or Causal BART/Boosting/Forest/Tree/... (new)

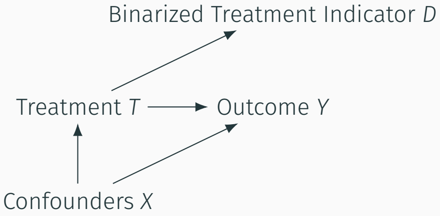
Less work on the conceptual side: **What** do we estimate?

This paper focuses on the common case where an evaluated binary treatment can be conceived as being itself heterogeneous

Scenario 1: binarized treatment

Multi-valued treatments are aggregated into a binary indicator

- different smoking intensities subsumed in ”smoking yes/no”
- pollution intensity binarized into ”pollution high/low”



Scenario 2: Multiple treatment versions

Binary treatment can be disaggregated in taking multiple versions

- marketing measures or surgeries delivered by different people
- different spezializations within job training program

$$
\begin{tikzcd}
\text{Binary Treatment } D \arrow[r] & \text{Treatment Version } T \arrow[r] & \text{Outcome } Y \\
& \text{Confounders } X \arrow[u] \arrow[ur] &
\end{tikzcd}
$$

RQ: What do heterogeneous effects mean if the treatment is heterogeneous? And what can we do about it?

Illustrating toy example - average effect

Evaluation of marketing measure to increase customer satisfaction using a randomized controlled trial (A/B testing)

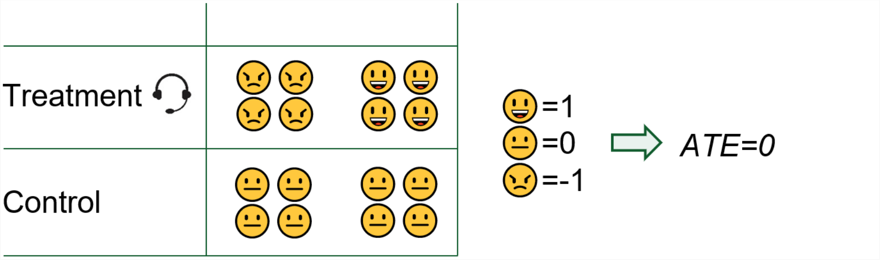
The measure is to call customers and to inform about new exciting products

Illustrating toy example - average effect

Evaluation of marketing measure to increase customer satisfaction using a randomized controlled trial (A/B testing)

The measure is to call customers and to inform about new exciting products

First check the AVERAGE TREATMENT EFFECT (*ATE*)



Illustrating toy example - heterogeneous effects

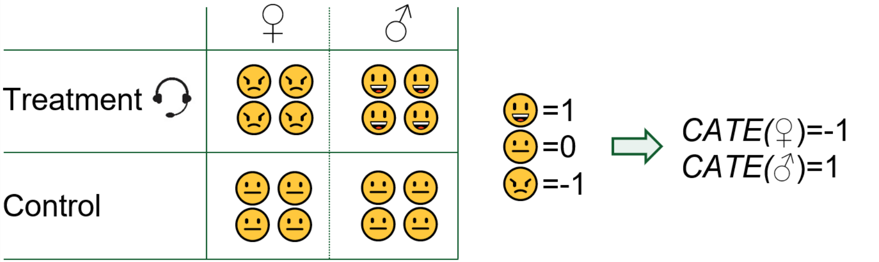
The goal of canonical effect heterogeneity analysis is to figure out whether at least some groups respond positively to treatment

We are interested in the CONDITIONAL AVERAGE TREATMENT EFFECT (*CATE*)

Illustrating toy example - heterogeneous effects

The goal of canonical effect heterogeneity analysis is to figure out whether at least some groups respond positively to treatment

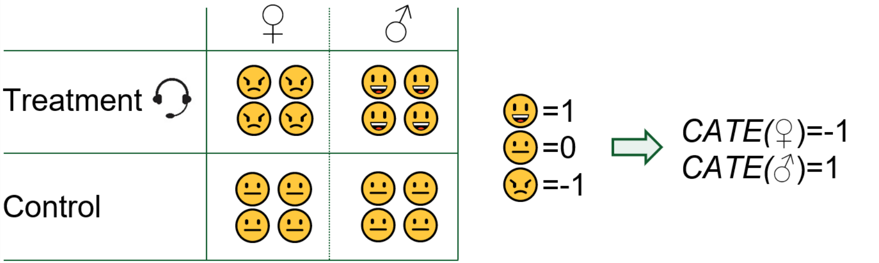
We are interested in the Conditional Average Treatment Effect (*CATE*)



Illustrating toy example - heterogeneous effects

The goal of canonical effect heterogeneity analysis is to figure out whether at least some groups respond positively to treatment

We are interested in the CONDITIONAL AVERAGE TREATMENT EFFECT (*CATE*)



⇒ Only men should be treated

Causal Machine Learning does this personalization nowadays in a data-driven way

Illustrating toy example - treatment heterogeneity

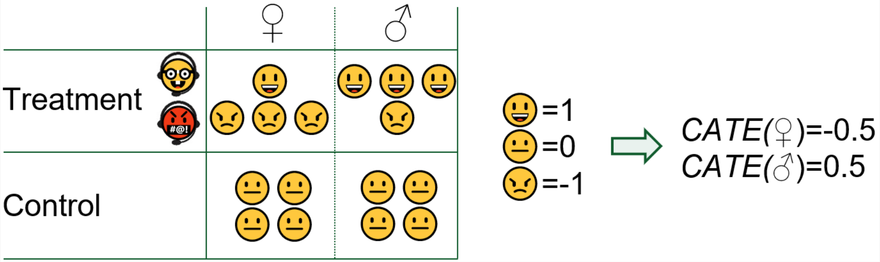
It is unlikely that all customers receive the same (homogeneous) treatment

Example: Different employees deliver the call, but not random across groups

Illustrating toy example - treatment heterogeneity

It is unlikely that all customers receive the same (homogeneous) treatment

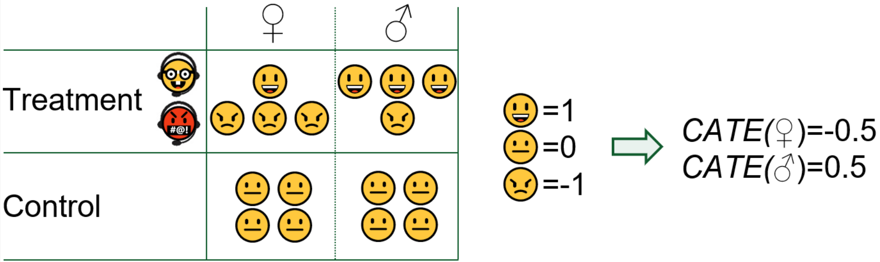
Example: Different employees deliver the call, but not random across groups



Illustrating toy example - treatment heterogeneity

It is unlikely that all customers receive the same (homogeneous) treatment

Example: Different employees deliver the call, but not random across groups



"Effect heterogeneity" is driven by different groups receiving different treatments

NOT that different groups respond differently to the same treatment

Consequences

The standard conclusion would be to target male customers

However, this ignores the mechanism behind the heterogeneity

The correct conclusion would be that women receive a worse treatment mix and this should be fixed

We propose a decomposition to disentangle effect and treatment heterogeneity

Illustrating toy example - decomposition

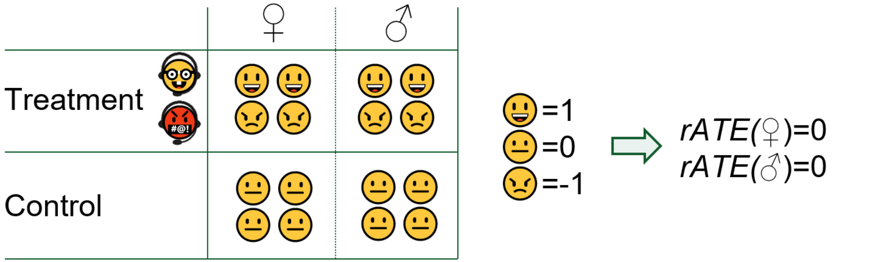
Decomposition enforces the same distribution of treatment versions along subgroups

We call this benchmark parameter the RANDOMIZED CATE (*rATE*)

Illustrating toy example - decomposition

Decomposition enforces the same distribution of treatment versions along subgroups

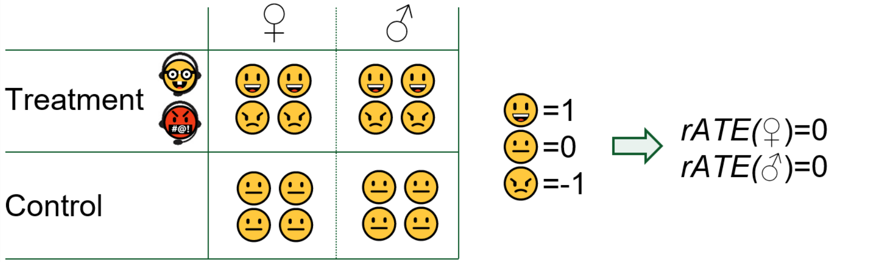
We call this benchmark parameter the RANDOMIZED CATE (*rATE*)



Illustrating toy example - decomposition

Decomposition enforces the same distribution of treatment versions along subgroups

We call this benchmark parameter the RANDOMIZED CATE (*rATE*)



$\Rightarrow$ Gender gap in effectiveness disappears under same treatment composition

$\Rightarrow$ It was completely driven by treatment heterogeneity

Illustrating toy example - decomposition

Difference between canonical *CATE* and *rATE* shows how much of the canonical *CATE* is driven by treatment heterogeneity:

- $\Delta(♀) = CATE(♀) - rATE(♀) = -0.5$
- $\Delta(♂) = CATE(♂) - rATE(♂) = 0.5$

Positive values indicate that the assignment of treatment versions is better than random

Heterogeneity in $\Delta(x)$ indicates that the selection quality of versions varies across different groups

This can be used e.g. to understand the fairness of the current assignment mechanism of treatment versions

Application: Job Corps

Job Corps is the largest training program for disadvantaged youth (16–24) in the US

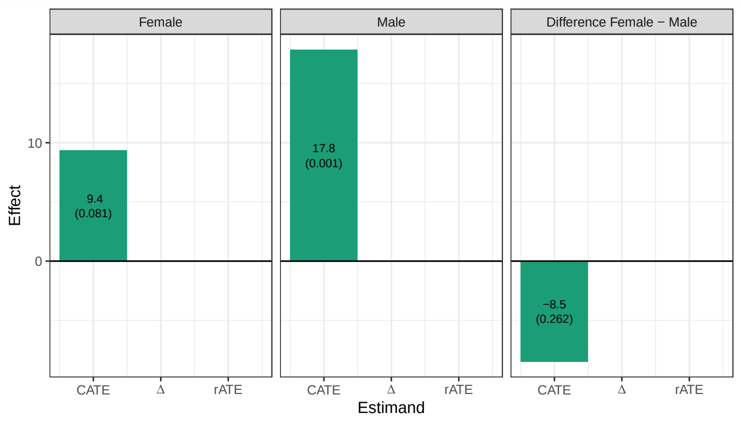
Experimental assessment from 1994–1996 (Schochet et al., 2001, 2008)

Common finding: women profit less in terms of earnings than men

Potential explanation: Vocational training of men in JC focuses more on craft jobs, for women more on service oriented $\Rightarrow$ Lower returns to vocational training

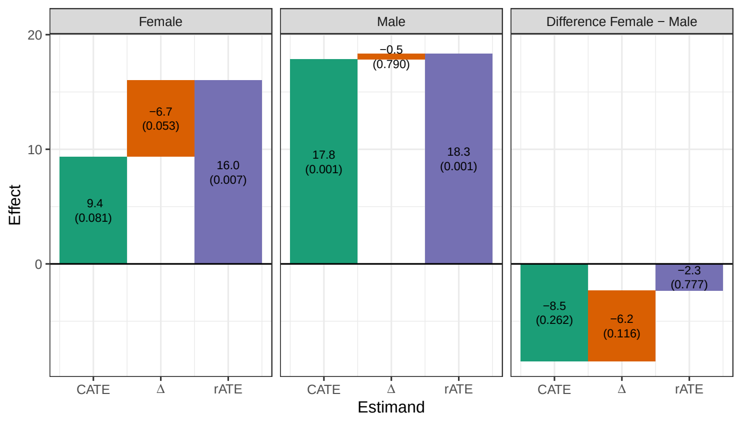
Outcome: weekly earnings four years after program start

Application: Job Corps



Increase in women's earnings half of men's

Application: Job Corps



Imposing the same vocational training nearly removes the gender gap

⇒ Targeting of vocational training for women worse than random

Conclusion

Interpretation of seemingly easy heterogeneous effects might be more complicated if analyzed treatment is itself heterogeneous

Our method provides a first way to address this issue

More in the paper:

- Double ML based estimator allowing for small p-scores and many treatments
- Application on ethnicity and age gaps in the effect of smoking on birthweight
- We provide an R implementation of the method (`causalDML`) and replication notebooks

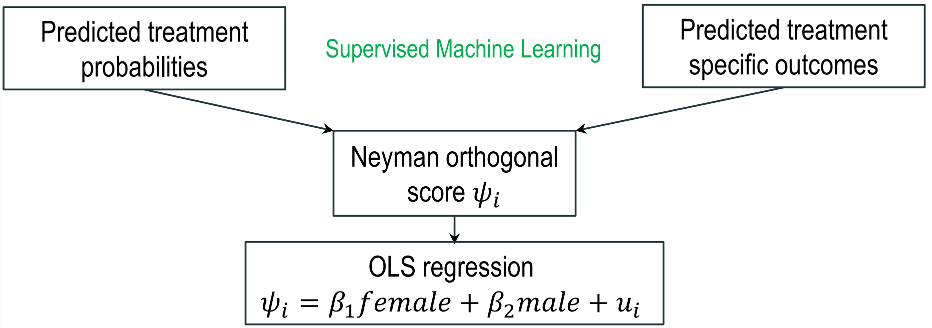
Thank you for your attention!

paper on arXiv:2110.01427

michael.knaus@uni-tuebingen.de
mcknaus.github.io
@MC_Knaus

How does the estimation procedure work?

Double Machine Learning based estimator building on Semenova & Chernozhukov (2021):



ψ_i is unbiased signal of decomposition parameters $\Rightarrow$ pseudo-outcome ▸ formal