FACULTY OF SCIENCES

STRATOS LORENTZ WORKSHOP, LEIDEN, SEPTEMBER 2024

CAUSAL MACHINE LEARNING

Stijn Vansteelandt

Ghent University, Belgium

GHENT UNIVERSITY

Introduction

GHENT
UNIVERSITY

Evaluating treatment effects

- Evaluation of the effect of a treatment A on an outcome Y is commonly based on contrasts

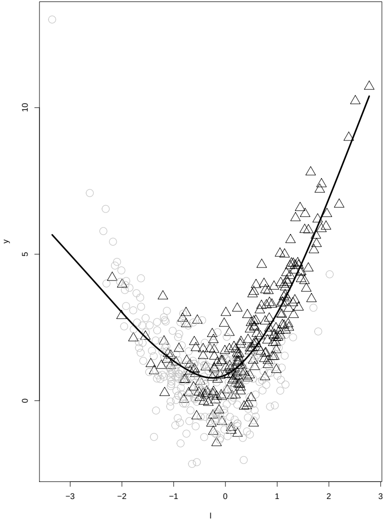
$$E(Y^1 - Y^0)$$

  of the expected outcome with (Y^1) versus without (Y^0) treatment.
- In observational studies, this demands adjustment for potentially high-dimensional confounders.
- Two popular approaches are standardisation and inverse probability weighting.

STANDARDISATION

To estimate the mean outcome under treatment,

- train a prediction model for outcome in the treated, using confounders;
- use this to predict outcome for all;
- average these predictions.
- The use of machine learning is increasingly popular.



WHY MACHINE LEARNING?

- Model misspecification is likely,
  and difficult to diagnose
  when treated and untreated subjects have limited overlap.
- The analysis can be made more objective
  by pre-specifying the machine learning algorithms.
  - In contrast, the human process of building a model is time-consuming
    and even more black box; pre-specifying it is difficult.
- If a more statistical approach is deemed preferable,
  then stacking statistical and machine learners allows one to do at least as good.

But…

Two caveats

Caveat 1: no valid uncertainty margins

machine learning 'easily' produces estimates,
but we have 'no clue' how precise these are...

- Even sample splitting or the bootstrap does not work.

(e.g. Samworth, 2011)

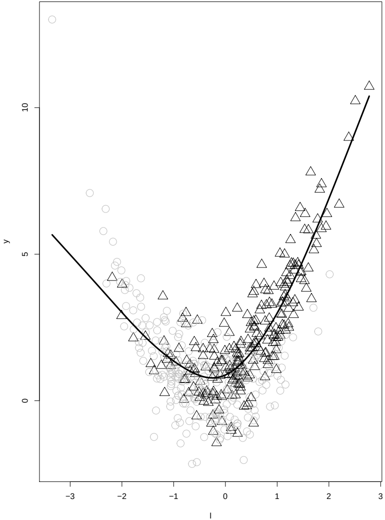
Caveat 2: plug-in bias

plugging machine learning predictions into a statistical analysis,
typically induces plug-in bias.

- The bias-variance tradeoff is so heavily optimized towards minimal prediction error, that machine learning algorithms underperform when used for other purposes.
- It leads to biased estimates, p-values and confidence intervals.

What is plug-in bias?

- Plug-in bias is the result of oversmoothing in the range of the data where predictions are needed,
- or due to mistakenly throwing out important confounders.



DEBIASED MACHINE LEARNING

GHENT
UNIVERSITY

A BIT OF HISTORY...

- Foundations for a solution have been laid in the 80's - 90's.

  (e.g. Pfanzagl, 1982; Bickel et al., 1998; Newey, 1990; Robins and Rotnitzky, 1995; van der Vaart, 1991)

- van der Laan made use of this theory
  to construct plug-in estimators based on machine learning,
  which he called Targeted Maximum Likelihood Estimators.

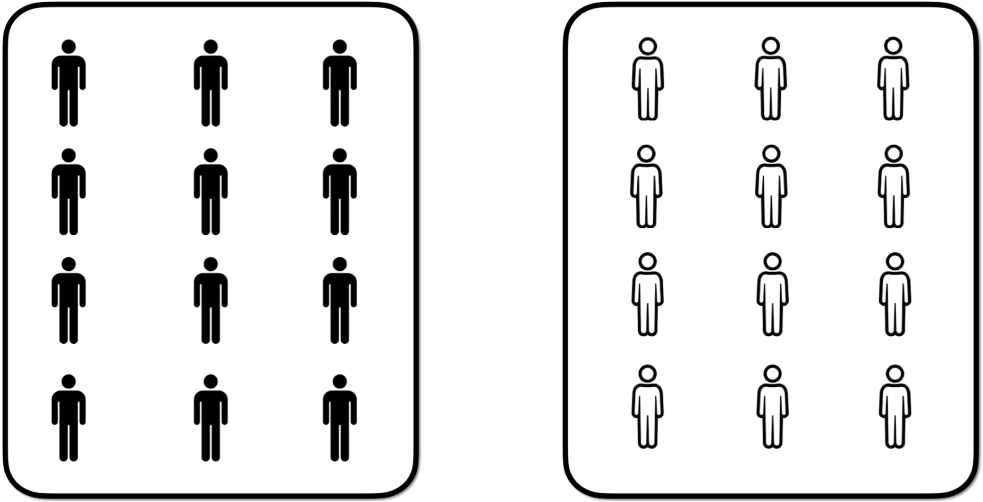
  (van der Laan and Rubin, 2008; van der Laan and Rose, 2014)

- His approach is now called targeted learning.
- Chernozhukov, Newey, Robins, ... popularised this theory,
  under weaker conditions by invoking sample splitting.

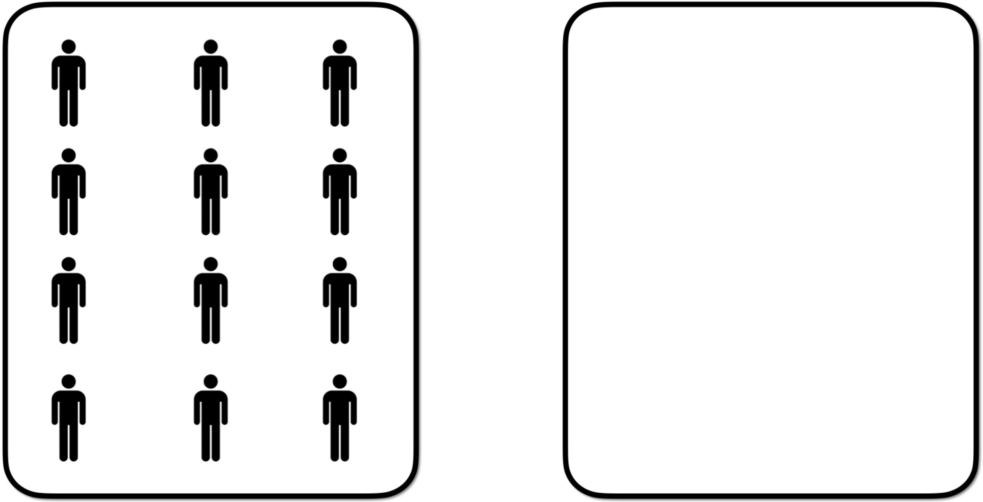
  (Robins et al., 2008; Chernozhukov et al., 2018)

- They refer to their approach as double / debiased machine learning.

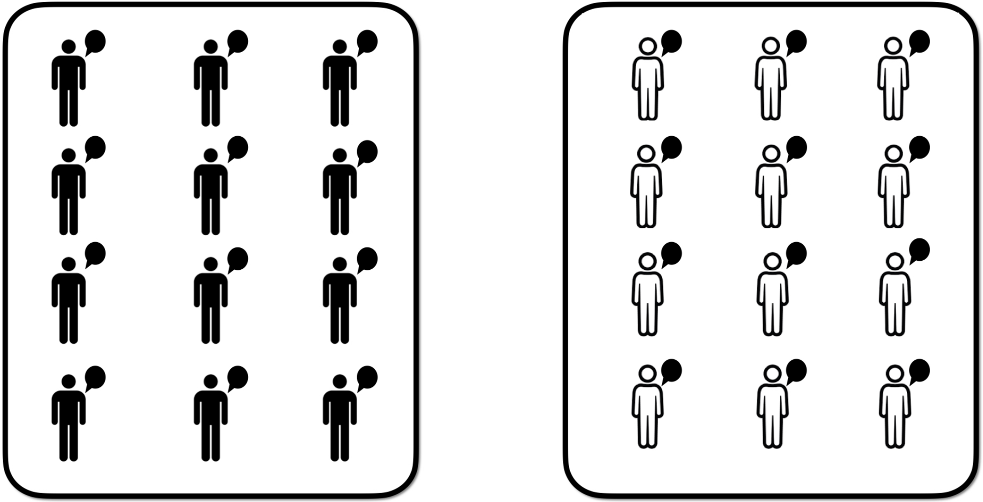
Observational data



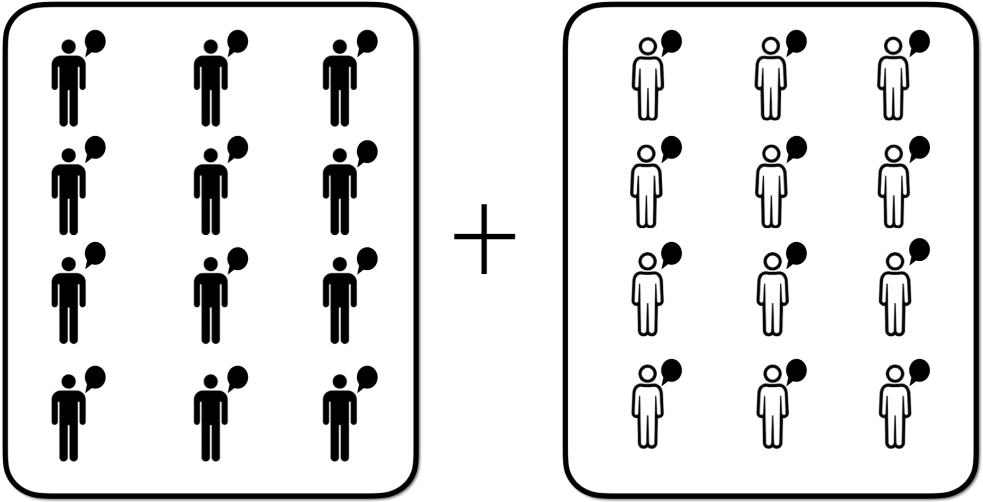
TRAIN IN TREATED, USING CONFOUNDERS



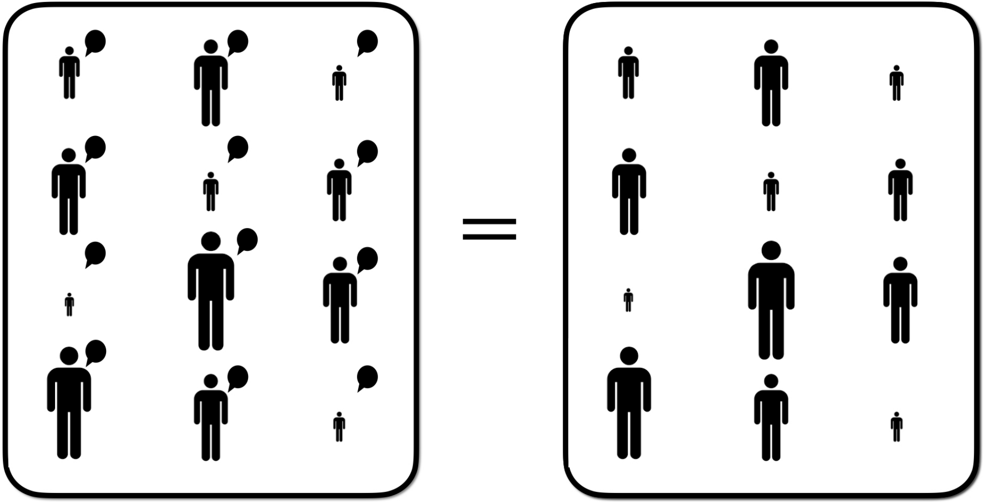
PREDICT OUTCOME ON TREATMENT FOR ALL



AVERAGE PREDICTED TREATMENT OUTCOME OVER ALL



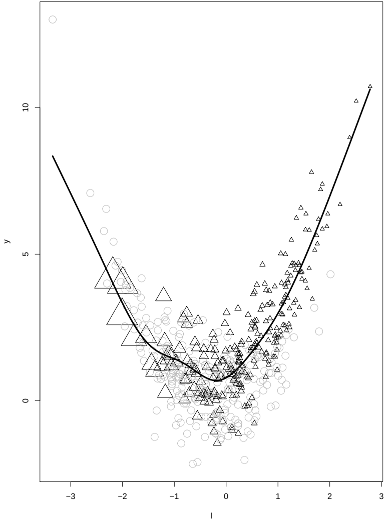
How to debias outcome mean on treatment?

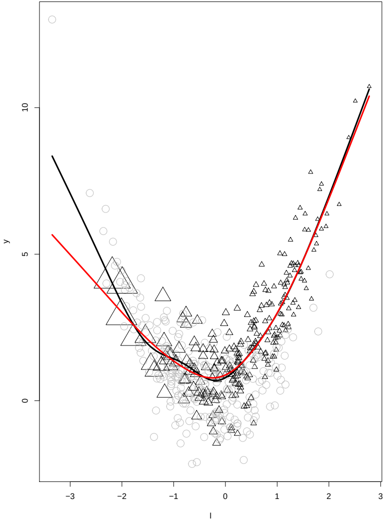


A SKETCH HOW TO DEBIAS OUTCOME MEAN ON TREATMENT

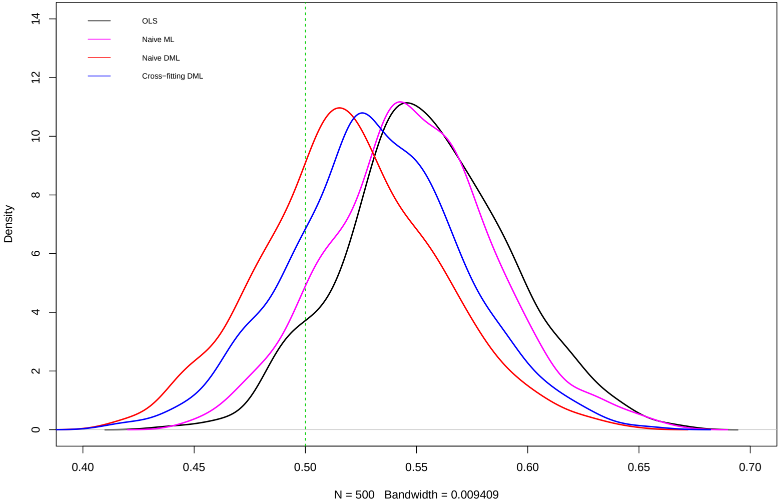
- To learn the amount of plug-in bias,
  we evaluate prediction errors in the treated,
  but weigh them (inversely to the propensity score) to approximate bias in the full sample.
- Debiased machine learning subtracts this bias from the estimate.
- Targeted learning updates predictions to be free of bias.
- Sample splitting is used to prevent overfitting bias,
  but may also induce finite-sample bias and excess variability.

TARGETED / DEBIASED LEARNING





AN IMPRESSION FROM SIMULATION STUDIES



Ditch the statistical model?

GHENT UNIVERSITY

Ditch the statistical model?

- Developments on debiased machine learning are centered around efficient influence curves for model-free estimands.

  (Hines et al., 2021)
- This can be useful, to target simplicity.
- But by giving up on models to summarize, developments are largely limited to 'simple' causal queries.
- Compromises are therefore made 'to fit the framework',
  - E.g., 'What if all had above median levels of glycoprotein acetyls at all times'?

  or recourse is made to modeling, bringing back the earlier critiques.
  - E.g., marginal structural models, incompatible Cox models in target trials, ...

Bridging stats and ML...

GHENT
UNIVERSITY

Assumption-lean modeling

- For a dichotomous, randomized exposure A and baseline covariates L, we consider 'assumption-lean' models of the form

$$g\left\{E(Y^a|L)\right\} = \alpha(L) + \beta(L)a$$

for a known link $g(.)$ and $a = 0, 1$.

(JRSS-B discussion paper on assumption-lean regression by Vansteelandt and Dukes (2022))

- In generalised (partially) linear models / SMMs, we would assume that

$$\beta(L) = \beta \quad \text{and/or} \quad \alpha(L) = \alpha' L.$$

- We will avoid such assumptions

and learn the mean and variance (or other summaries) of $\beta(L)$ instead

(Vansteelandt and Dukes, 2022)

or quantify what components of L explain the variance of $\beta(L)$ the most.

(Hines, Diaz-Ordaz and Vansteelandt, 2022)

Assumption-lean loglinear modeling algorithm

1. Predict A based on L to obtain predictions $\hat{p}_i$.
2. Predict Y based on A and L to obtain predictions $\hat{Y}_i$.
3. Predict $\log\left(\hat{Y}\right)$ based on L to obtain predictions $\hat{q}_i$.
4. Linearly regress (using least squares)

$$\log\left(\hat{Y}_i\right) - \hat{q}_i + \frac{Y_i}{\hat{Y}_i} - 1$$

on $A_i - \hat{p}_i$ to obtain an estimate for β and a robust standard error.

When using variable selection in a loglinear model, this debiases the naïve estimate $\hat{\beta}$ as

$$\hat{\beta} + \frac{\sum_{i=1}^{n}(A_i - \hat{p}_i)\left(Y_i e^{-\hat{\beta}A_i - \hat{\gamma}' L_i} - 1\right)}{\sum_{i=1}^{n}(A_i - \hat{p}_i)^2}$$

and delivers valid post-selection inference.

Features

- The flexibility of standard regression

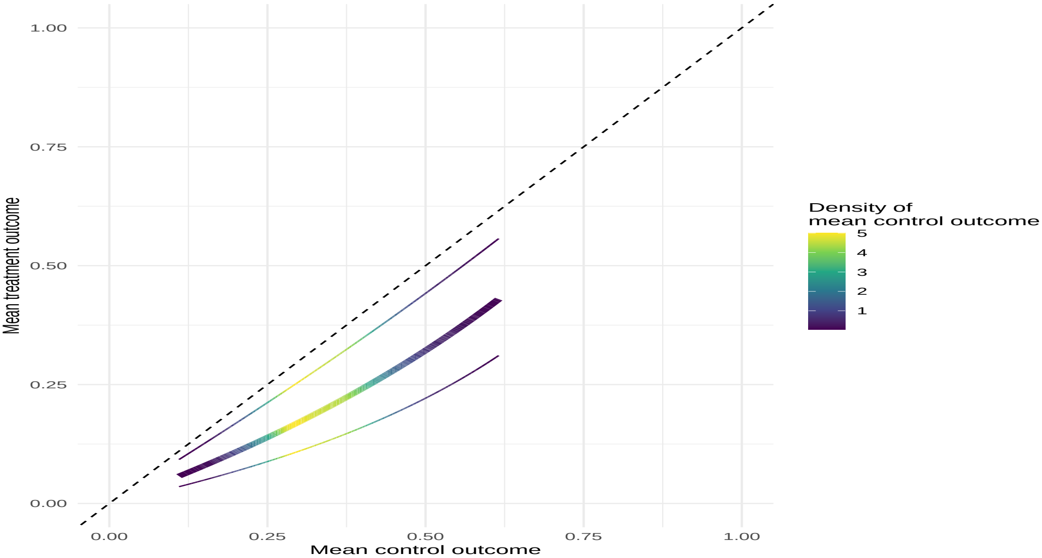
  (e.g., it readily handles continuous exposures).

- It overcomes Occam's dilemma by separating modeling to summarise from (data-adaptive) modeling to handle the curse of dimensionality.

  (Breiman, 2001)

- It prevents model misspecification bias by incorporating flexible modeling, machine learning, and is clear on what is being estimated, even when the model is wrong.
- It avoids to extract information from modeling assumptions by working under the nonparametric model.
- It delivers valid (post-selection) inference after using ML, variable / model selection.
- It enables (near) pre-specification of the entire analysis.
- It is 'simple' to obtain.

Percentiles of $E(Y^1|L)$ vs $E(Y^0|L)$ in ACTG175



SUMMARY

GHENT
UNIVERSITY

SUMMARY

- Standard statistical analyses
  - ignore model uncertainty,
  - leave residual confounding bias due to model misspecification,
  - and complicate pre-specification of the analysis.
- Debiased / targeted learning overcome these concerns.
- These techniques are essential for any data-adaptive analysis,
  in particular enabling valid use of variable selection in parametric models.

SUMMARY

- Causal machine learning = machine learning for evaluating treatment effects as opposed to prediction.
- This is much harder: we can compare predictions with observed outcomes, but cannot compare estimated with true treatment effects.
- This is why results from asymptotic statistics are essential.

  Hines, O., Dukes, O., Diaz-Ordaz K., and Vansteelandt, S. (2021). Demystifying statistical learning based on efficient influence functions. The American Statistician, 1-48.
- Most existing works have focused on the average effect of a binary treatment, leading to lack of flexibility and oversimplification.
- Assumption-lean modeling bridges traditional modeling with debiased machine learning.

  Vansteelandt, S., & Dukes, O. (2022). Assumption-lean inference for generalised linear model parameters (with discussion). JRSS - B, 84, 657-685.
- Orthogonal learning targets prediction of counterfactuals, causal effects, ...

  (e.g., Athey and Imbens, 2016; Wager and Athey, 2018; Künzel et al., 2019; Kennedy, 2020; Nie and Wager, 2021; Foster and Syrgkanis, 2023; Vansteelandt and Morzywolek, 2023, van der Laan et al., 2024)

Selected references

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21, C1–C68.

Hines, O., Dukes, O., Diaz-Ordaz K., and Vansteelandt, S. (2021). Demystifying statistical learning based on efficient influence functions. The American Statistician, 76, 292-304..

van der Laan, M. J., & Rose, S. (2011). Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media.

Vansteelandt, S., & Dukes, O. (2022). Assumption-lean inference for generalised linear model parameters (with discussion). Journal of the Royal Statistical Society - B, 84, 657-685.

Vansteelandt, S. (2021). Statistical modelling in the age of data science. Observational Studies, 7, 217-228.

Slides: `users.ugent.be/~svsteela/`