FACULTY OF SCIENCES

Inserm Workshop 282, Best practices and recent advances in causal analyses, May 5-7, 2025, Bordeaux, France

Challenges and solutions in high-dimensional settings

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.
  - Such adjustment is not required in randomized experiments, but nonetheless desirable to boost precision.
- Two popular approaches are standardisation and inverse probability weighting.

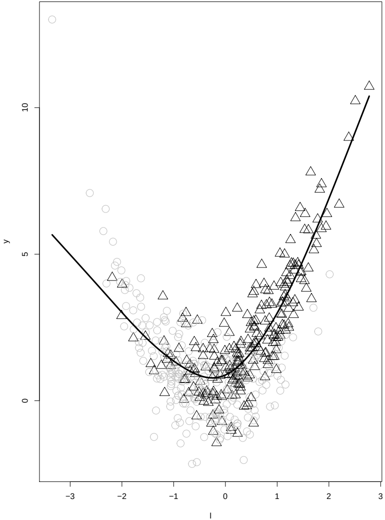
STANDARDISATION / G-COMPUTATION

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.

This can be based on statistical model building.

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.
  - Even models that fit the observed data well,
    may cause large extrapolation bias.
- 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.

Plan for this Lecture

- How best adjust for high-dimensional confounding when the aim is valid inference for $E(Y^1 - Y^0)$?
  - I will not aim for estimates with e.g. minimal mean squared error.
- I will do this, assuming we have access to a collection of variables L that suffices to adjust for confounding.
  - I will sidestep the difficulty of excluding post-treatment variables, colliders, ... which should precede the analysis and be based on subject-matter knowledge.

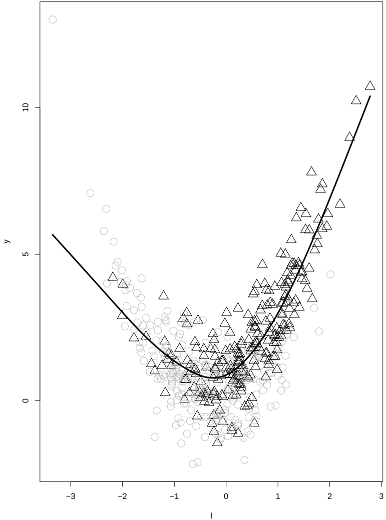
Why Naïve ML-based G-computation does not work

GHENT UNIVERSITY

Problem 1: Nailing the Wrong Target...

- To avoid overfitting, we inevitably rely on modeling — e.g., learning $E(Y \mid A = 1, L)$.
- Standard statistical and ML methods aim to optimize a bias-variance tradeoff.
- But this tradeoff is tuned for prediction accuracy, not for causal effect estimation.
- The result? Impressive predictions, biased effects.

Why Bias?



- Bias often comes from oversmoothing — even when models fit observed data well, they can fail badly where predictions matter.
- Bias can also arise from dropping key confounders: variables strongly predicting treatment are at risk to be trimmed away.
- This is known as plug-in bias: bias introduced when naïve predictions are plugged into causal effect estimators.

PLUG-IN BIAS

- When using flexible machine learning algorithms,
  or consistent model selection over a rich class,
  plug-in bias shrinks with sample size...
- ...but painfully slowly.
- The bias can dominate the estimator's standard deviation,
  leading to confidence intervals with poor coverage and biased p-values.

Problem 2: Underestimating Uncertainty

- Statistical and ML models readily produce predictions of $E(Y \mid A = 1, L)$, but we often have no clue how precise they are...
- ...or how this uncertainty propagates into the causal effect estimate.
- Standard error formulas from statistical model-fitting routines account for estimation uncertainty but not model uncertainty — so they underestimate variability.
- Even sample splitting or the bootstrap fail to capture this properly.

(see e.g., Samworth, 2011)

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 debiased plug-in estimators based on machine learning,
  which he called Targeted Maximum Likelihood Estimators.

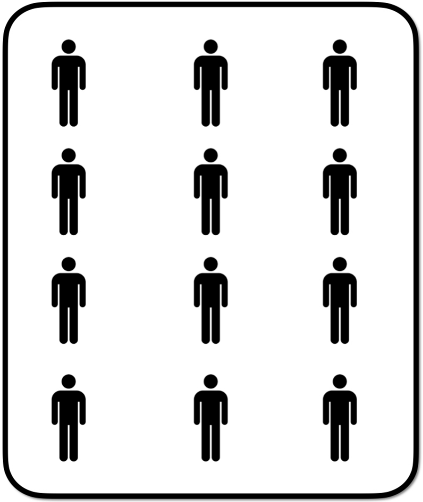
  (van der Laan and Rubin, 2008; van der Laan and Rose, 2014)

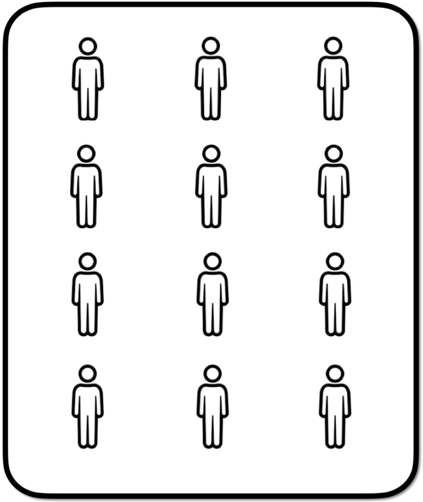
- His approach is now called targeted learning.
- Chernozhukov, Newey, Robins, ... popularised this theory,
  under weaker conditions by invoking sample splitting.

  (Robins et al., 2008; Zheng and van der Laan, 2010; Chernozhukov et al., 2018)

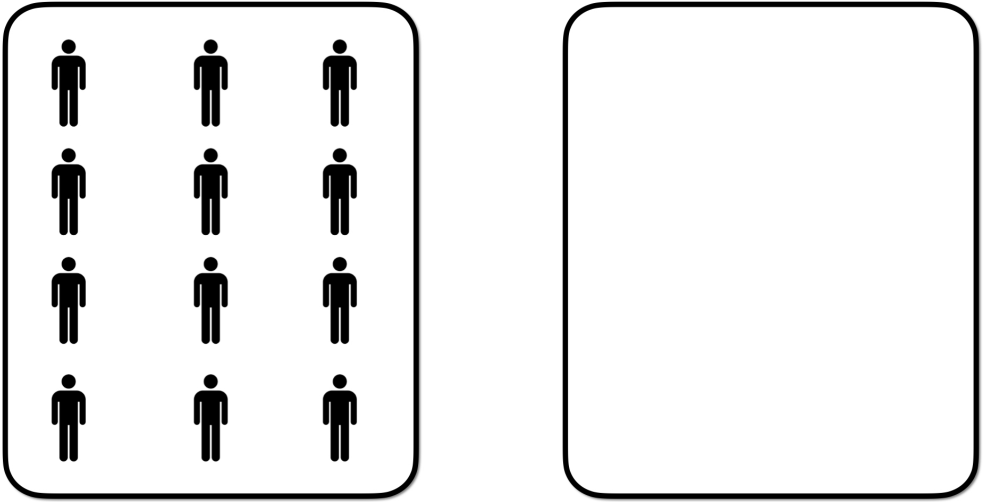
- They refer to their approach as double / debiased machine learning.

Observational Data

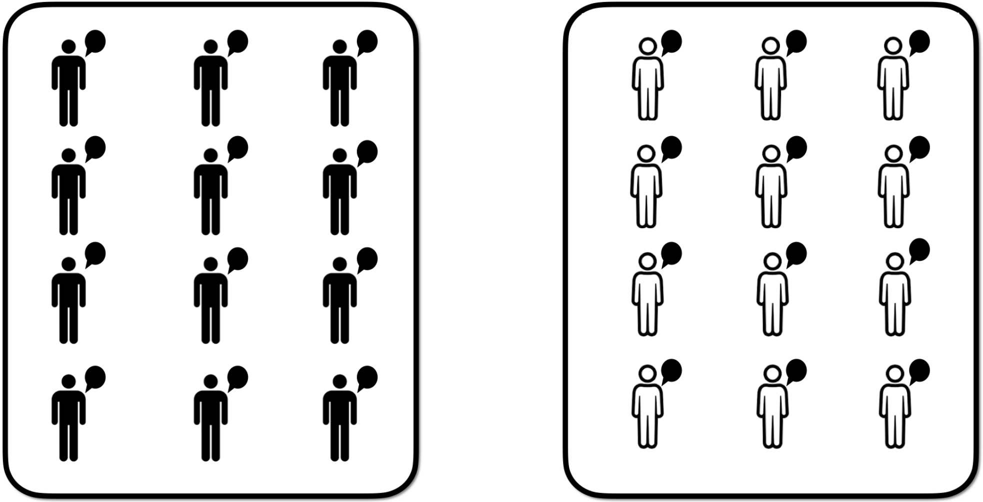




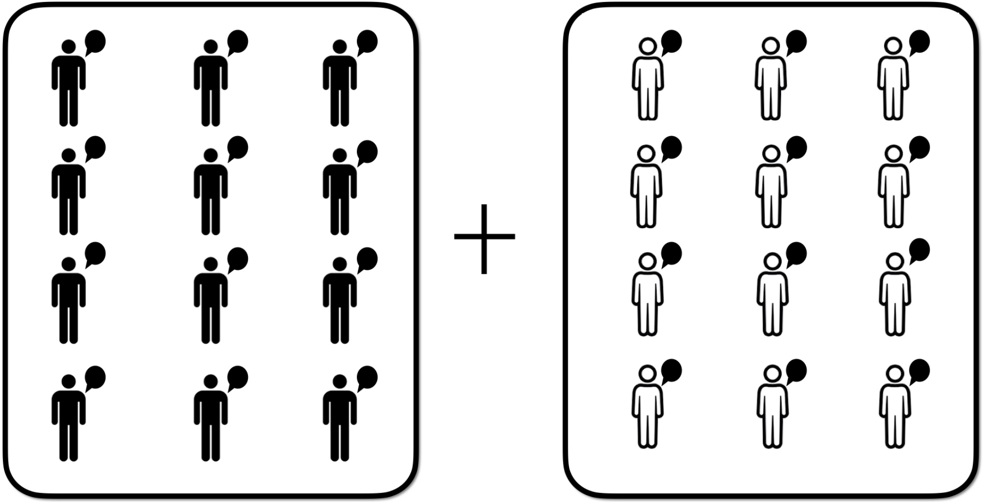
TRAIN IN TREATED, USING CONFOUNDERS



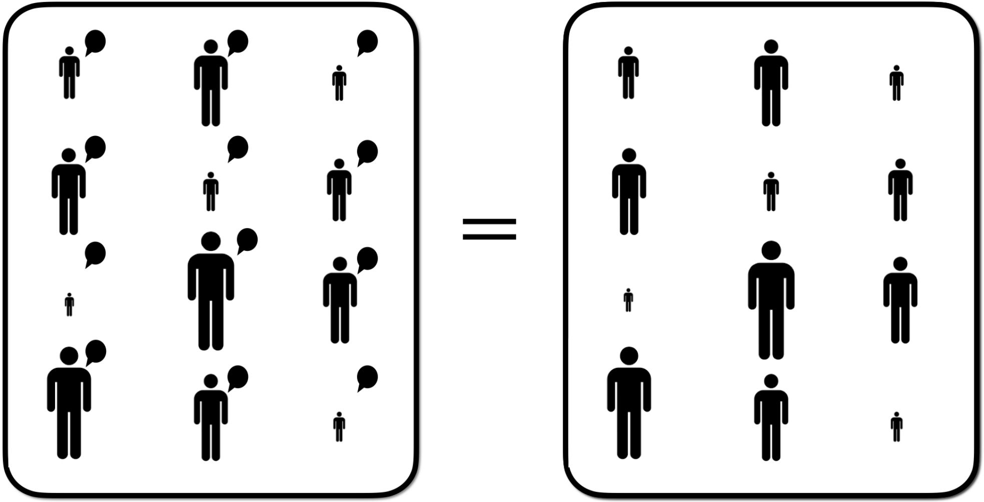
PREDICT OUTCOME ON TREATMENT FOR ALL



AVERAGE PREDICTED TREATMENT OUTCOME OVER ALL



How to Debias Outcome Mean on Treatment?



How to Debias Outcome Mean on Treatment? (cont'd)

- 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:

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\hat{P}(A_i=1\mid L_i)}\left\{Y_i-\hat{E}(Y_i\mid A_i=1,L_i)\right\}$$

- Debiased machine learning subtracts this bias from the estimate:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{E}(Y_i\mid A_i=1,L_i)+\frac{A_i}{\hat{P}(A_i=1\mid L_i)}\left\{Y_i-\hat{E}(Y_i\mid A_i=1,L_i)\right\}$$

- This delivers the popular augmented IPW estimator,
  which uses standard ML for both the propensity score and outcome predictions.

Targeted Learning

- Instead of subtracting bias from a plug-in estimate, targeted learning updates outcome predictions to remove bias.
- E.g., for a binary outcome, it does this by building a logistic regression model around initial predictions $\hat{E}^{(0)}(Y_i \mid A_i, L_i)$:

$$\operatorname{logit} E(Y_i \mid A_i, L_i) = \underbrace{\operatorname{logit} \hat{E}^{(0)}(Y_i \mid A_i, L_i)}_{\texttt{qlogis(pred0)}} + \delta \underbrace{\frac{A_i}{\hat{P}(A_i = 1 \mid L_i)}}_{\texttt{C = a/ps}}$$

- In R:

```
model <- glm(
  y ~ offset(qlogis(pred0)) + C,
  family = binomial(),
  data = dataset)
```

Targeted Learning (cont'd)

- When this model is fitted using maximum likelihood, the updated estimator

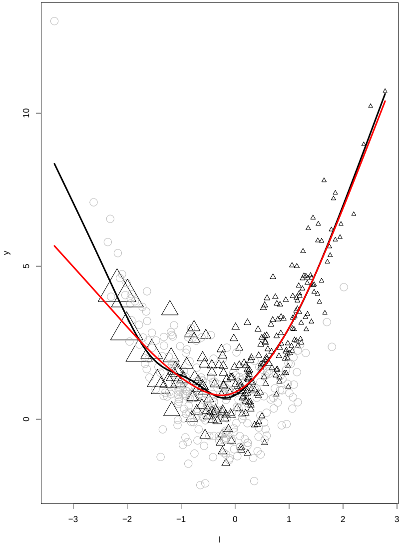
$$\frac{1}{n}\sum_{i=1}^{n}\hat{E}^{(1)}(Y_i \mid A_i = 1, L_i)$$

  is free of plug-in bias.
- This estimator is equivalent to the AIPW estimator in large samples, but may have finite-sample benefits.

Problem 1 — How Did We Remove Plug-in Bias?

- In variable selection procedures, we eliminate plug-in bias because confounders get two chances to be selected — for both treatment and outcome models.
- In machine learning procedures, we reduce plug-in bias by targeting prediction performance over the full covariate distribution, not just where data are dense.

TARGETED / DEBIASED LEARNING



Problem 2 — How Can We Recover Uncertainty?

- Once plug-in bias is removed, estimator enjoys an oracle property: it behaves like an AIPW estimator based on the true propensity score and outcome regression:

$$\frac{1}{n}\sum_{i=1}^{n}\left[E(Y_i \mid A_i = 1, L_i) + \frac{A_i}{P(A_i = 1 \mid L_i)}\{Y_i - E(Y_i \mid A_i = 1, L_i)\}\right]$$

- As a result, we can validly compute standard errors as if no machine learning or model selection had been used.
- This works the same for debiased and targeted learning.

Implementation on ACTG175 in R

```
> install.packages("tmle")
> library(tmle)
> W = cbind(age,wtkg,hemo,homo,drugs,karnof,oprior,z30,zprior,preanti,race,gender,str2,
   strat,symptom,cd40)
> lib = c("SL.glm", "tmle.SL.dbarts.k.5", "SL.gam", "SL.glmnet", "SL.glm.interaction",
   "SL.ranger", "SL.nnet")
> tmle_est <- tmle(Y=Y,A=A,W=W,Q.SL.library=lib)
> summary(tmle_est)

 Initial estimation of Q
 Procedure: cv-SuperLearner, ensemble
 Model: Y ~  SL.glm_All + [...]
 Coefficients:
      SL.glm_All    0
 tmle.SL.dbarts.k.5_All    0.1071496
      SL.gam_All    0.3746302
   SL.glmnet_All    0.3434137
 SL.glm.interaction_All    0
   SL.ranger_All    0.1748065
     SL.nnet_All    0

 Cross-validated R squared :  0.393
```

Implementation on ACTG175 in R (cont'd)

```
Estimation of g (treatment mechanism)
Procedure: SuperLearner, ensemble  Empirical AUC = 0.6177

Model:
A ~  SL.glm_All + tmle.SL.dbarts.k.5_All + SL.gam_All

Coefficients:
     SL.glm_All    0.7703587
tmle.SL.dbarts.k.5_All    0.2296413
     SL.gam_All    0

Additive Effect
  Parameter Estimate:  49.514
  Estimated Variance:  25.768
             p-value:  <2e-16
   95% Conf Interval: (39.565, 59.463)
```

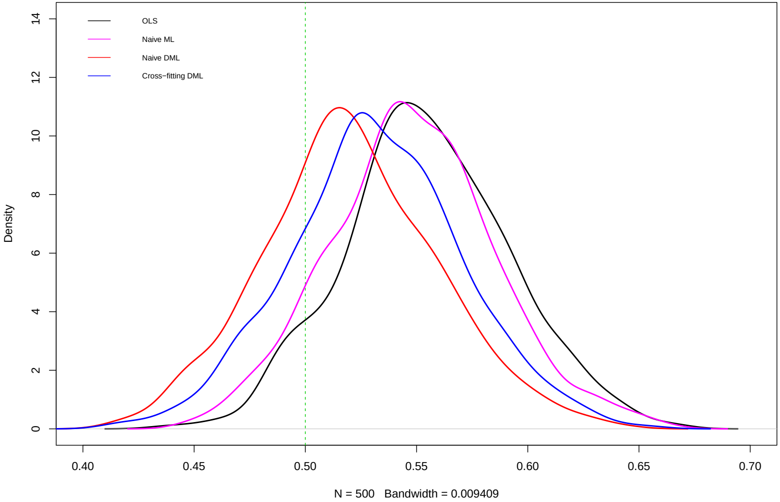
Implementation on ACTG175 in R (cont'd)

```
> tl0 = lm(cd420~treat,data=ACTG175)
> summary(tl0)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  336.139      6.210  54.125  < 2e-16 ***
treat         46.810      7.165   6.533 8.03e-11 ***

> confint(tl0)

                2.5 %    97.5 %
(Intercept) 323.95990 348.31830
treat        32.75921  60.86179
```

Caveat: Sample Splitting

- Sample splitting helps avoid overfitting bias when ML is used.
- Information loss is ‘minimized’ via cross-fitting procedures,
  but these can introduce some finite-sample bias and excess variability.

An impression from simulation studies



Caveat (cont'd): Residual Bias

- Debiased (targeted learning) estimators are not completely unbiased, but their bias can be bounded (in absolute value) by:

$$E\left[\left\{\frac{P(A=1\mid L)-\hat{P}(A=1\mid L)}{\hat{P}(A=1\mid L)}\right\}^2\right]^{1/2} E\left[\left\{E(Y\mid A=1,L)-\hat{E}(Y\mid A=1,L)\right\}^2\right]^{1/2}$$

- If this product shrinks faster than $1/\sqrt{n}$, then the residual bias is asymptotically negligible.

Caveat (cont'd): Rate Double Robustness

- This is the case for many machine learning algorithms
  and consistent model selection over rich model classes,
  but is not guaranteed in general.
- It expresses rate double robustness:
  errors in one nuisance estimate can be offset by accuracy in the other.
- In randomized experiments, this is guaranteed:
  - if the true propensity score is used, or
  - if the sample proportion is used as propensity score,
    and the outcome predictions are consistent.
  - For targeted learning estimators, even this outcome consistency can be relaxed.

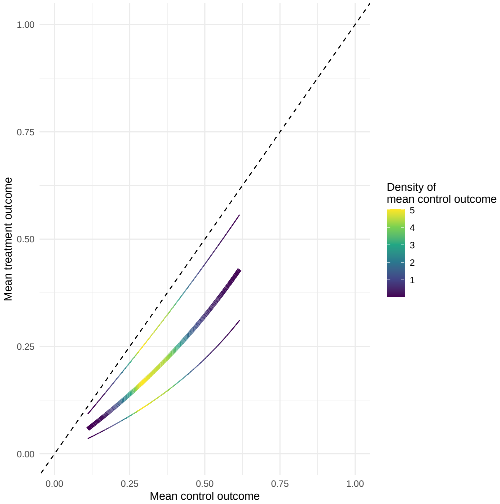
    (Van Lancker, Díaz and Vansteelandt, 2025)

Bridging modeling and ML...

GHENT UNIVERSITY

Bridging statistical modeling and ML...

- Debiased ML can also be connected to statistical modeling.
- ML for confounder adjustment.
- Results 'projected' onto a simple model.
- This is the focus of assumption-lean modeling.
- It alleviates concerns about modeling e.g., MSMs, Cox models in target trials, ...
- It may also provide more refined insight.



Assumption-lean modeling

- For a dichotomous, randomized exposure A and baseline covariates L, consider

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

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

(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 of the Approach

- Prevents model misspecification bias by incorporating flexible models and ML—
  it guarantees valid causal effect estimation, even if the model is wrong.
- Overcomes Occam's dilemma by separating:
  - modeling for interpretation via

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

  - from data-adaptive modeling to tackle high-dimensionality.

    (Breiman, 2001)
- Delivers valid post-selection inference—
  even after using ML or model/variable selection.
- Supports (near) full pre-specification of the analysis plan.
- Offers the flexibility and simplicity of regression,
  e.g., it readily accommodates continuous exposures.

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.

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/`