FACULTY OF SCIENCES

JUNE 29, 2021

TARGETED LEARNING: WHAT, AND WHY YOU SHOULD CARE

Stijn Vansteelandt, Ghent University and the London School of Hygiene and Tropical Medicine

GHENT UNIVERSITY

INTRODUCTION

GHENT
UNIVERSITY

UNADJUSTED ANALYSES ARE NOT ENTIRELY SATISFACTORY

- The primary analysis of RCTs is typically unadjusted or adjusted for only a limited number of discrete stratification factors.
- This is not entirely satisfactory: covariate adjustment
  - can lead to drastic gains in power,

    (see Kelly Van Lancker)
  - and may even be needed to control for informative censoring (or dropout).

    (see Alex Luedtke, Oliver Dukes)
- The default strategy for covariate adjustment focuses on coefficients indexing regression models.
- It is also not entirely satisfactory.

Standard adjusted analyses are not entirely satisfactory

- Typical regression parameters (e.g., odds ratios, hazard ratios) can be subtle to interpret and even change magnitude depending on which covariates are adjusted.

  (see Rhian Daniel)

- Models may be misspecified,
  leading to bias in effect estimates and standard errors.

  (e.g., Freedman, 2001; Robins and Rotnitzky, 2001; van der Laan, 2015)

  (see Kelly Van Lancker, Alex Luedtke, Oliver Dukes)

  - This concern is made worse because of trade-offs between correctness and simplicity.

    (e.g., Breiman, 2001)

- Model-based analyses can be difficult to pre-specify.
- Model building algorithms aim to prevent misspecification,
  but may induce model uncertainty.
  - This may inflate Type I errors, and invalidate standard inference.

    (Leeb and Pötscher, 2006; van der Laan and Rose, 2011; Dukes and Vansteelandt, 2020)

CAN WE DO BETTER?

A SIMPLE TRY...

- Suppose we aim to learn the treatment effect on a dichotomous outcome (e.g. 'disease').
- Let's use a simple imputation procedure:
  - Estimate disease risk on treatment, $\hat{P}^1$, for all trial participants
    based on a logistic regression in the treated, in function of baseline covariates.

Age	Trt	Y	Y^1	$\hat{P}^1$
40	1	1	1	0.8
50	1	0	0	0.6
60	1	1	1	0.7
50	0	0	?	0.7
30	0	1	?	0.6
40	0	0	?	0.5

  - Average these risks for all trial participants
    to obtain an estimate of population disease risk on treatment.

A SIMPLE TRY...

- Next,
  - Estimate disease risk on control, $\hat{P}^0$, for all trial participants
    based on a logistic regression in the controls, in function of baseline covariates.

Age	Trt	Y	Y^1	$\hat{P}^1$	Y^0	$\hat{P}^0$
40	1	1	1	0.8	?	0.7
50	1	0	0	0.6	?	0.55
60	1	1	1	0.7	?	0.6
50	0	0	?	0.7	0	0.6
30	0	1	?	0.6	1	0.5
40	0	0	?	0.5	0	0.45

  - Average these risks for all trial participants
    to obtain an estimate of population disease risk on control.
- We can contrast these estimates as differences, ratios, ...

Some immediate advantages

- Simple analysis
- Simple interpretation

  no matter how complex the logistic regression model is.

  (thus no need for making trade-offs)
- By contrasting disease risks for the same participants with and without treatment, we gain precision.
  - This is because we can contrast people with the same age, with vs without treatment.

Some magic

- Model misspecification does not induce bias in effect estimates.
- Standard errors easy to calculate

  (with 1 line of code)

  and are valid (in simple randomised experiments)
  - even when (standard) variable selection is used;

    (van der Laan and Rose, 2011)
  - even when the model is misspecified.

    (Vermeulen and Vansteelandt, 2015; Avagyan and Vansteelandt, 2021)
- These properties are the result of exploiting knowledge that randomisation happens independently of covariates.
  - This knowledge is ignored by likelihood-based approaches.

TARGETED LEARNING

GHENT
UNIVERSITY

More flexible modelling strategies

- This simple imputation procedure
  happens to be an example of targeted learning.
- It appears to lend itself easily to more general prediction strategies
  and even machine learning.
  - This is useful because more accurate modelling can lead to power gains
    and becomes essential when adjustment is needed for confounding or selection bias.
  - However, it is not guaranteed to have these desirable properties more generally,
    because these strategies are aimed at small prediction error
    and not at accurate treatment effect estimates.

TARGETED LEARNING

- Targeted learning strategies therefore update initial predictions and target them towards the estimand of interest.

  (van der Laan and Rubin, 2006; Moore and van der Laan, 2009; van der Laan and Rose, 2011)

  (see Alex Luedtke)

- It is therefore essential that the starting point of the analysis is the choice of an estimand (rather than the choice of a model).
- This updating does not require advanced methods: it is usually based on a specific single-parameter model built around initial predictions, which is then fitted using maximum likelihood.
- There are parallel developments, known as debiased machine learning.

  (Chernozhukov et al., 2018)

TARGETED LEARNING

- Targeted learning is transforming the way how we will do data analysis in the future.
- It brings data analysis back to its essence:
  translating a scientific question into an estimands, doing sanity checks, ...
  with automated model building strategies running in the background.
- This renders pre-specification of the analysis accessible.
- It makes the data analysis more honest, by acknowledging model uncertainty.
- That this is feasible, is quite impressive!

What sample sizes are needed?

- Reliance on asymptotic theory
  and experience with nonparametric regression procedures may make one concerned that enormous sample sizes will be needed to make this work.
- This intuition is misleading.
- The focus here is on population-averaged effects,

  (cfr. the simple imputation strategy)

  which usually do not demand large sample sizes.

Is Targeted Learning not too complicated for my data?

- An analogy...
- Also martingale theory underlying Cox regression is complex,
  but it does not make Cox regression less popular.
- Targeted learning relies on theory on nonparametric influence functions,
  which is likewise not known to many.
- But it need not stop one,
  from using principled analyses that target the treatment effect of interest,
  while acknowledging 'all' uncertainties.
- See Targeted Learning Webinar series on YouTube.

  tinyurl.com/youtube-PDS

  www.youtube.com/channel/UC6Cg1XjzX-MlyxKlWfHezFQ

KEY REFERENCES

- Bartlett, J. W. (2018). Covariate adjustment and estimation of mean response in randomised trials. Pharmaceutical statistics, 17(5), 648-666.
- Moore, K. L., & van der Laan, M. J. (2009). Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. Statistics in medicine, 28(1), 39-64.
- Rosenblum, M., & Van Der Laan, M. J. (2009). Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. Biometrics, 65(3), 937-945.
- Steingrimsson, J. A., Hanley, D. F., & Rosenblum, M. (2017). Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions. Contemporary clinical trials, 54, 18-24.
- Tsiatis, A. A., Davidian, M., Zhang, M., & Lu, X. (2008). Covariate adjustment for two?sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Statistics in medicine, 27(23), 4658-4677.
- Vermeulen, K., Thas, O., & Vansteelandt, S. (2015). Increasing the power of the Mann-Whitney test in randomized experiments through flexible covariate adjustment. Statistics in medicine, 34(6), 1012-1030.