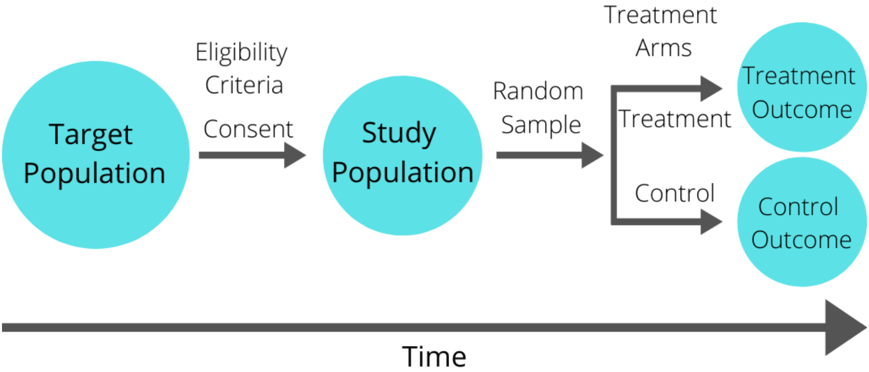
Applied Machine Learning Methods in Adjusting for Population Differences

Lauren Cappiello

February 11, 2020

Randomized Clinical Trials



Example: Chronic Heart Disease

- Target population: US patients with chronic heart disease.
- Study population limited by
  - Eligibility criteria.
  - Consent.
- Patients randomly assigned to treatment or placebo.
- Results represent study population.

Objective

Estimate key quantities for the target population:

- Adjust for population differences in some target population.
- Use clinical data from a study population along with observational data from the target population.
- Improve existing (parametric) methods.

Confounding Adjustment

There are two quantities of interest:

- Mean treatment outcome.
  - Example: A one-armed trial identifies the mean clinical outcome for the treated.
- Average treatment effect.
  - Example: Estimate a treatment effect given a study comparing two treatments (no placebo) and a historical study.

Confounding Adjustment

- There are several approaches to these problems.
- Each approach uses some combination of
  - A model relating treatment and covariates to outcome.
  - A model for the propensity scores, or the probability of a specific treatment group given the covariates.
- Current approaches utilize parametric models.

The Parametric Approach

- These parametric models can be difficult to specify correctly.
- If the parametric models are incorrect, the estimates are inconsistent.
  - An estimator is consistent if, as the sample size gets larger, the estimate gets closer to the true value.
- We want to avoid these issues by applying machine learning methods.

The Super Learner

How does one choose the optimal machine learning approach?

Consider the principle of super learning.

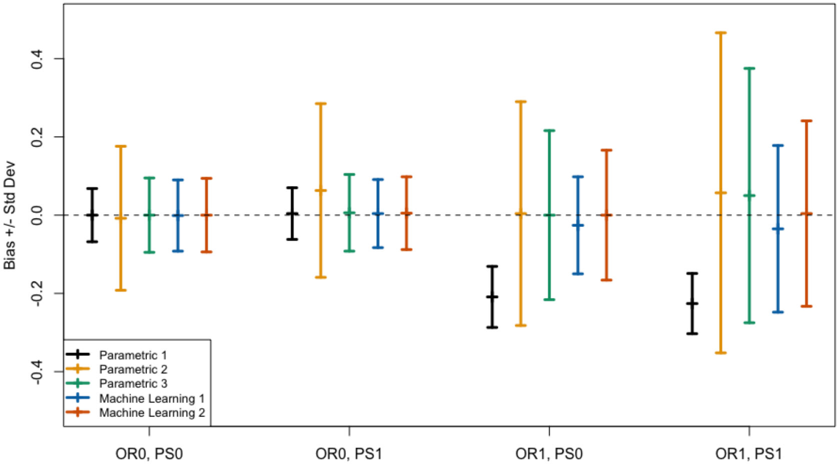
- Combines several machine learning methods to create one ”super learner”.
- Runs each machine learning method on a subset of the data and uses the rest of the data to examine performance.
- This process is repeated several times with different subsets.
- The final learner is a weighted combination of the different methods based on how each performed.

The Machine Learning Approach

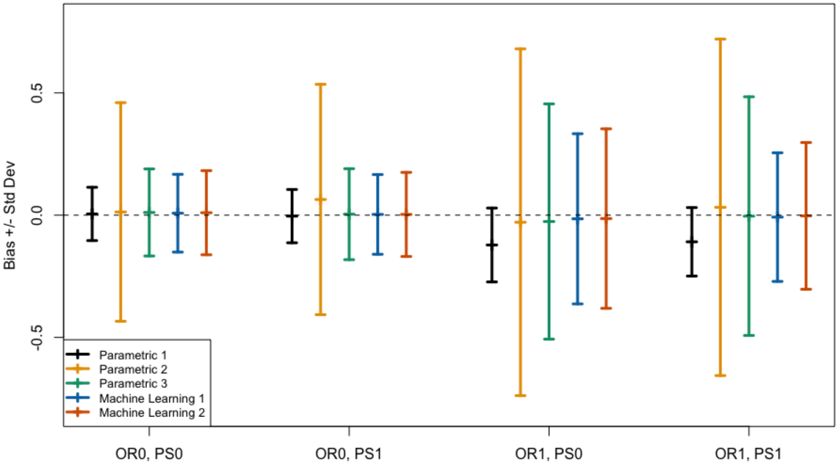
Under mild assumptions, the estimators based on a machine learning approach are:

- Consistent.
  - As the sample size gets larger, the estimate gets closer to the true value.
- Asymptotically efficient.
  - As the sample size gets larger, the quantity is estimated in the "best possible manner" (based on bias and variability).
- Asymptotically normal.
  - As the sample size gets larger, estimates from repeated samples form a normal distribution.

Simulation Results: Mean Outcome



Simulation Results: Treatment Effect



Ongoing Research

Application to data on implantable cardioverter-defibrillators:

- Several RCTs demonstrate the effectiveness of implantable cardioverter-defibrillators (ICDs) in reducing the risk of Sudden Cardiac Death (SCD).
- Results from these trials led to current care guidelines on who should receive ICDs.
- Post-hoc analyses raise concerns with guidelines.

Ongoing Research

Sensitivity analysis for the ignorability assumption:

- This work examines what happens when one of our key assumptions is violated.
- This involves building a violation into the assumption and re-deriving all of the methods.
- This will allow us to vary the level of violation to examine how "sensitive" the methods are to these changes.

Future Research: Machine Learning in Causal Inference

- Machine learning methods applied to other estimators.
- Adjusting for population differences for other desired quantities.
- Application of new methodology to high dimensional data.
- Machine learning in causal inference for longitudinal data.

Future Research: Collaboration and Applied Statistics

- Applied machine learning.
- Causal inference.
- General statistical consulting.

Thank you!