Adjusting for Population Differences in Treatment Effect Estimation Using Machine Learning Methods

Zhiwei Zhang

Biometric Research Program
National Cancer Institute
zhiwei.zhang@nih.gov

Joint work with Lauren Cappiello, Changyu Shen, Neel Butala, Xinping Cui, and Robert Yeh

Outline

- Introduction
- Methodology
- Application
- Summary

Introduction

- Adjusting for population differences is a major challenge in utilizing real-world data to evaluate medical treatments
  - We will focus on generalizing clinical trial results to a target population
- Existing statistical methods:
  - Imputation method based on an outcome regression (OR) model
  - Methods based on a propensity score (PS) model:
    - weighting
    - matching
    - stratification
  - Doubly robust (DR) methods
    - involve both OR and PS models
    - consistent if either model is correct
    - efficient if both models are correct
- Will explore the use of machine learning (ML) methods in this work
  - ML refers to any method/algorithm that can be used to estimate a regression function (OR or PS)
  - Nonparametric ML methods are generally more flexible, though slower in convergence, than parametric models

Some Notation

- Trial population/cohort
  - Baseline covariates: W
  - Potential outcomes: $Y(a)$, $a = 0$ (control) or 1 (experimental)
  - Randomized treatment: $A = 0$ or 1 (independent of all baseline variables)
  - Actual outcome: $Y = Y(A)$ (assuming consistency)
  - Observed data: (W_i, A_i, Y_i), $i = 1, \ldots, n$
- Target population/cohort
  - Baseline covariates: W^*
  - Potential outcomes: $Y^*(a)$, $a = 0$ (control) or 1 (experimental)
  - Observed data: W_i^*, $i = 1, \ldots, n^*$
- Interested in $\delta^* = \mathrm{E}\{Y^*(1) - Y^*(0)\}$, which is generally different from $\delta = \mathrm{E}\{Y(1) - Y(0)\}$
- Assumptions are needed for identification

Key Assumptions

- Any difference between δ and δ^* can be explained through baseline covariates, so that

$$\mathrm{E}\{Y(1) - Y(0)|W = w\} = \mathrm{E}\{Y^*(1) - Y^*(0)|W^* = w\} =: d(w)$$

- Any patient in the target population could potentially be included in the trial cohort
  - Formally, we assume $\mathcal{W}^* \subset \mathcal{W}$
  - For simplicity, we further assume $\mathcal{W}^* = \mathcal{W}$
  - This is equivalent to

$$0 < r(w) := f^*(w)/f(w) < \infty,$$

    where f and f^* are the density functions of W and W^*, respectively, with respect to a common measure

- Together, these two assumptions are sufficient for identifying δ^*.

Existing Methods

- Imputation
  - Estimate $\delta^* = \mathrm{E}\{d(W^*)\}$ with

$$\widehat{\delta}^*_{\text{imp}} = \frac{1}{n^*} \sum_{i=1}^{n^*} \widehat{d}(W_i^*),$$

    where $\widehat{d}$ is a generic estimate of d based on $\{(W_i, A_i, Y_i), i = 1, \ldots, n\}$
- Weighting
  - Estimate $\delta^* = \mathrm{E}\{Dr(W)\}$, where $D = AY/\pi - (1-A)Y/(1-\pi)$, with

$$\widehat{\delta}^*_{\text{wt}} = \frac{1}{n} \sum_{i=1}^{n} D_i \widehat{r}(W_i),$$

    where $\widehat{r}$ is an estimate of r based on $\{W_i, i = 1, \ldots, n\}$ and $\{W_i^*, i = 1, \ldots, n^*\}$

- DR method
  - Motivated by semiparametric theory, a DR estimator of δ^* may be constructed as

$$\widehat{\delta}^*_{\text{dr}} = \frac{1}{n^*}\sum_{i=1}^{n^*}\widehat{d}(W_i^*) + \frac{1}{n}\sum_{i=1}^{n}\widehat{r}(W_i)\Big\{D_i - \widehat{d}(W_i) - (A_i - \pi)\widehat{h}(W_i)\Big\},$$

    where $\widehat{h}(w)$ is an estimate of

$$h(w) = \frac{m_1(w)}{\pi} + \frac{m_0(w)}{1-\pi} = \mathrm{E}\left(\frac{AY}{\pi^2} + \frac{(1-A)Y}{(1-\pi)^2}\,\middle|\, W = w\right)$$

  - $\widehat{\delta}^*_{\text{dr}}$ is consistent if either $\widehat{d}$ or $\widehat{r}$ is consistent
- In these methods, estimation of (d, r, h) is usually based on parametric models

Proposed Methods

- $\widehat{\delta}^*_{\text{dr}}$ with $(\widehat{d}, \widehat{r}, \widehat{h})$ obtained using statistical ML methods
- Questions:
  - $\sqrt{n}$-consistent?
  - Asymptotically normal?
  - Asymptotically efficient?
- We assume there exist limit functions d_∞, r_∞ and h_∞ such that, with probability 1, $\widehat{d}(w) \to d_\infty(w)$, $\widehat{r}(w) \to r_\infty(w)$ and $\widehat{h}(w) \to h_\infty(w)$ for all $w \in \mathcal{W}$
  - So $\widehat{\delta}^*_{\text{dr}}$ is consistent for δ^* if $d_\infty = d$ or $r_\infty = r$ (or both), regardless of h_∞
- For $\sqrt{n}$-consistency and asymptotic normality, we assume
  - $d_\infty = d$ and $r_\infty = r$
  - $\|\widehat{d} - d\|_2 \|\widehat{r} - r\|_2 = o_p(n^{-1/2})$
  - $(\widehat{d}, \widehat{r}, \widehat{h})$ belong to a Donsker class with probability tending to 1
  - $n^*/n \to \lambda \in (0, \infty)$ as $n \to \infty$

- Under these conditions, it can be shown that $\sqrt{n}(\widehat{\delta}^*_{\text{dr}} - \delta^*)$ converges to a normal distribution with mean 0 and variance

$$\text{var}[r(W)\{D - d(W) - (A - \pi)h_\infty(W)\}] + \lambda^{-1}\,\text{var}\{d(W^*)\}$$

- When $h_\infty = h$, the above asymptotic variance becomes the nonparametric variance bound for estimating δ^*, and $\widehat{\delta}^*_{\text{dr}}$ is then asymptotically efficient in the nonparametric sense
- The rate condition $\|\widehat{d} - d\|_2 \|\widehat{r} - r\|_2 = o_p(n^{-1/2})$ can be satisfied in a variety of ways
- There is a great variety of ML methods available, some of which may perform better than others in a given application
- Multiple candidate ML methods can be combined through cross-validation into a super learner with a desirable oracle property

- The Donsker condition on $(\widehat{d}, \widehat{r}, \widehat{h})$ may limit the collection of ML methods that can be included in the super learner
- Sample splitting (aka cross-fitting) has been suggested as a way to remove the Donsker condition while retaining $\sqrt{n}$-consistency and asymptotic normality
  - Partition the sample (trial & target cohorts) into L subsamples that are roughly equal in size
  - For each $l \in \{1, \ldots, L\}$, exclude the lth subsample and obtain $(\widehat{d}^{(-l)}, \widehat{r}^{(-l)}, \widehat{h}^{(-l)})$ from the rest of the sample using the same methods for obtaining $(\widehat{d}, \widehat{r}, \widehat{h})$
  - Estimate δ^* with

$$\widehat{\delta}^*_{\text{dr.ss}} = \frac{1}{n^*} \sum_{i=1}^{n^*} \widehat{d}^{(-l_i^*)}(W_i^*) + \frac{1}{n} \sum_{i=1}^{n} \widehat{r}^{(-l_i)}(W_i) \left\{ D_i - \widehat{d}^{(-l_i)}(W_i) - (A_i - \pi) \widehat{h}^{(-l_i)}(W_i) \right\},$$

where l_i (l_i^*) is the index of the subsample that includes the ith subject in the study (target) cohort

Application (Aortic Stenosis)

- Surgical aortic valve replacement (SAVR) has been the standard of care
- Transcatheter aortic valve replacement (TAVR) is a newer and less invasive treatment option
- A randomized clinical trial (CoreValve) in high-risk patients found an absolute reduction of 4.9% (95% CI: -0.4 to 10.2%) for TAVR versus SAVR in the all-cause mortality rate at one year after treatment
- There have been questions about the generalizability of the trial results to the target population of high-risk patients
- The target population is better described by the Medicare Provider and Review (MedPAR) database of the US Centers for Medicare and Medicaid Services
- We are interested in utilizing the MedPAR database together with the CoreValve trial data to estimate the treatment difference in the target population

Table: Summary of baseline characteristics for the trial cohort (by treatment and overall) and the target cohort in the cardiology example: mean (standard deviation) for continuous variables and percentage (%) for binary ones.

Patient Characteristic	Trial Cohort			Target Cohort
	TAVR	SAVR	Overall	
	$n_1 = 314$	$n_0 = 286$	$n = 600$	$n^* = 49,591$
age in years	83.6 (6.5)	83.4 (6.2)	83.5 (6.4)	82.7 (7.4)
male sex	52.9	52.1	52.5	51.6
white race	97.1	94.8	96.0	92.9
congestive heart failure	69.7	61.5	65.8	75.2
pulmonary circulation disorder	19.7	18.9	19.3	23.7
chronic pulmonary disease	28.7	26.2	27.5	27.4
hypothyroidism	17.8	16.4	17.1	22.0
renal failure	32.8	29.4	31.2	37.6
frailty percentile	46.2 (27.3)	45.7 (28.9)	46.0 (28.1)	50.8 (28.9)

Table: Data analysis for the cardiology example: point estimates (standard errors) of the one-year mortality rates (as percentages) of TAVR and SAVR as well as their difference (TAVR – SAVR) in the target population.

Method	TAVR	SAVR	Difference
Imputation	13.7 (2.0)	17.0 (2.3)	−3.3 (3.2)
Weighting	13.6 (2.0)	17.0 (2.5)	−3.3 (3.4)
DR.par	13.6 (2.1)	17.0 (2.4)	−3.4 (3.1)
DR.np	13.7 (1.8)	17.3 (2.2)	−3.9 (2.9)
DR.np.ss	14.5 (2.1)	18.2 (2.5)	−3.3 (2.9)

Summary

- The problem of adjusting for population differences is becoming increasingly important with increasing use of real-world data for treatment evaluation
- Existing statistical methods for this purpose generally rely on correct specification of parametric regression models
- In this work, we have investigated the use of nonparametric ML methods to estimate nuisance functions in adjusting for population differences
- Incorporating such ML estimates into the DR method leads to nonparametric DR estimators that are theoretically justified and perform well in simulation studies
- Future research may address possible violations of key assumptions