Applied Machine Learning Methods in Adjusting for Population Differences

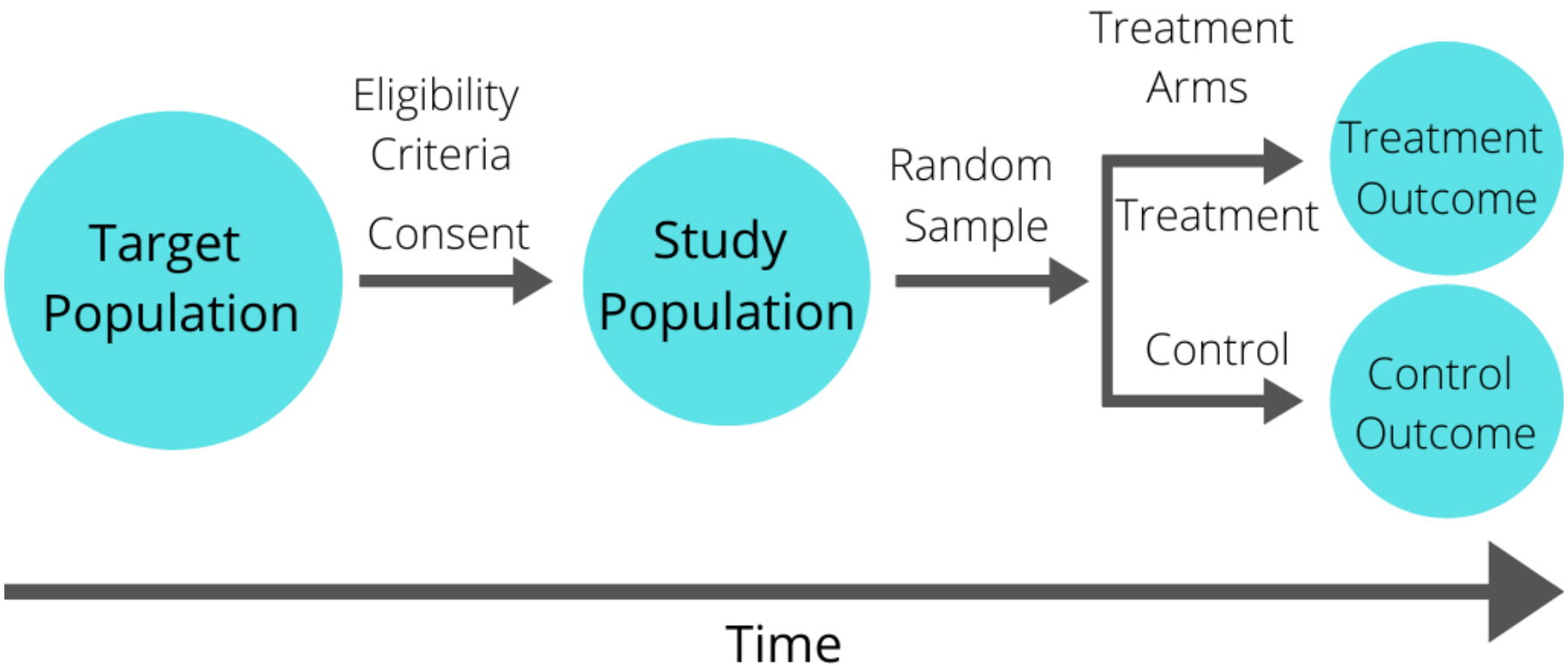
Lauren Cappiello

February 3, 2020

Acknowledgements

Dr. Zhiwei Zhang, National Cancer Institute
Dr. Xinping Cui, UCR Statistics Department
Dr. Changyu Shen, Harvard Medical School

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 mean outcomes and treatment effects in 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.
- Increase flexibility of existing (parametric) methods.

Confounding Adjustment in Clinical Settings

There are two quantities of interest:

- Mean treatment outcome.
  - Adjust a mean outcome from one population to another.
- Average treatment effect.
  - Adjust a treatment effect for population differences.

Adjusting a Mean Outcome

- This is necessary when a random sample of the clinical outcome for the target population is unavailable.
- Example:
  - A one-armed trial identifies the mean clinical outcome for the treated.
  - Historical data gives some information about the clinical outcome for the placebo.

Adjusting a Treatment Effect

- Extend treatment effect estimation from study population to target population.
  - Example: estimate a treatment effect given a study comparing two treatments (no placebo) and a historical study.
- Could use mean outcome adjustment on both treatment settings.
  - Requires stronger assumptions.

Notation

Adjusting a Mean Outcome

- Let Y^* be the outcome variable of interest.
- Let $\boldsymbol{X}^*$ be the associated covariates in the target population.
- Let $(\boldsymbol{X}, Y)$ be the counterparts of $(\boldsymbol{X}^*, Y^*)$ in the study population.

Adjusting a Treatment Effect

- Let $Y^*(t)$ be the potential outcome for randomized treatment $t \in \{0, 1\}$.
- Let $\boldsymbol{X}^*$ be a vector of baseline covariates in the target population.
- Let $Y(t)$, $t = 0, 1$ and $\boldsymbol{X}$ be the study population counterparts.

Notation

Adjusting a Mean Outcome

The data consist of

- $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$, a random sample of $(\boldsymbol{X}, Y)$.
- $\{\boldsymbol{X}_i^*, i = 1, \ldots, n^*\}$, a random sample of $\boldsymbol{X}^*$.

Target: $\mu^* = \mathrm{E}(Y^*)$ for some fixed treatment.

Adjusting a Treatment Effect

The data consist of

- $\{(\boldsymbol{X}_i, T_i, Y_i), i = 1, \ldots, n\}$ a random sample of $(\boldsymbol{X}, T, Y)$.
- $\{\boldsymbol{X}_i^*, i = 1, \ldots, n^*\}$ a random sample of $\boldsymbol{X}^*$.

Target: $\delta^* = \mu_1^* - \mu_0^*$ where $\mu_t^* = \mathrm{E}\{Y^*(t)\}$.

Assumptions

First, assume $\mathcal{X}^* = \mathcal{X}$ where $\mathcal{X}$ ($\mathcal{X}^*$) denotes the support of $\boldsymbol{X}$ ($\boldsymbol{X}^*$).

Adjusting a Mean Outcome

Then assume

$$m(\boldsymbol{x}) := \mathrm{E}(Y^*|\boldsymbol{X}^* = \boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}^*$$

where m is known as the outcome regression (OR) function.

Adjusting a Treatment Effect

Then assume

$$\begin{aligned} d(\boldsymbol{x}) :&= \mathrm{E}[Y^*(1) - Y^*(0)|\boldsymbol{X}^* = \boldsymbol{x}] \\ &= \mathrm{E}(Y|T = 1, \boldsymbol{X} = \boldsymbol{x}) - E(Y|T = 0, \boldsymbol{X} = \boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}. \end{aligned}$$

Imputation Method

Adjusting a Mean Outcome

The imputation approach to estimating μ^* is

$$\hat{\mu^*}_{IM} = \frac{1}{n^*} \sum_{i=1}^{n^*} \hat{m}(\boldsymbol{X}^*)$$

where $\hat{m}$ is some generic estimate of m based on $(\boldsymbol{X}, Y)$.

Adjusting a Treatment Effect

The imputation approach to treatment effect adjustment is

$$\hat{\delta}^*_{IM} = \frac{1}{n^*} \sum_{i=1}^{n^*} \hat{d}(\boldsymbol{X}^*_i)$$

where $\hat{d}$ is some generic estimate of d.

Adjusting a Mean Outcome

We may also write

$$\mu^* = \int m(\boldsymbol{x}) f^*(\boldsymbol{x}) d\nu(\boldsymbol{x}) = \int m(\boldsymbol{x}) \frac{f^*(\boldsymbol{x})}{f(\boldsymbol{x})} f(\boldsymbol{x}) d\nu(\boldsymbol{x}) = \mathrm{E}\left[Y \frac{f^*(\boldsymbol{X})}{f(\boldsymbol{X})} \right],$$

where f and f^* are the densities of $\boldsymbol{X}$ and $\boldsymbol{X}^*$, respectively, with respect to some common measure ν.

Adjusting a Treatment Effect

Another representation of δ^* is given by

$$\delta^* = \int d(\boldsymbol{x}) f^*(\boldsymbol{x}) d\nu(\boldsymbol{x}) = \int d(\boldsymbol{x}) \frac{f^*(\boldsymbol{x})}{f(\boldsymbol{x})} f(\boldsymbol{x}) d\nu(\boldsymbol{x}) = \mathrm{E}\left[D \frac{f^*(\boldsymbol{X})}{f(\boldsymbol{X})} \right],$$

noting that $d(\boldsymbol{X}) = E(D|\boldsymbol{X})$ where $D = \frac{TY}{\pi} - \frac{(1-T)Y}{1-\pi}$.

Weighting Method

Adjusting a Mean Outcome

This motivates the following weighted estimator:

$$\hat{\mu}^*_{WT} = \frac{1}{n}\sum_{i=1}^{n} Y_i \hat{r}(\boldsymbol{X}_i)$$

where $\hat{r}$ is some generic estimate of $r = f^*(\boldsymbol{X})/f(\boldsymbol{X})$.

Adjusting a Treatment Effect

The weighted estimator for treatment effect adjustment is

$$\hat{\delta}^*_{WT} = \frac{1}{n}\sum_{i=1}^{n} Y_i \hat{r}(\boldsymbol{X}_i)\left(\frac{T_i}{\pi} - \frac{1-T_i}{1-\pi}\right).$$

Weighting Method

Estimation of r can be based on

$$\hat{r}(x) = \exp\{\text{logit}[\hat{p}(x)] - \log(n^*/n)\} \tag{1}$$

where $\hat{p}$ is a generic binary regression estimate of the propensity score function

$$p(x) = \text{E}(T|X = x).$$

Doubly Robust Approach

Adjusting a Mean Outcome

A doubly robust estimator of μ^* is given by

$$\begin{aligned}\hat{\mu}^*_{DR} &= \hat{\mu}^*_{IM} + \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}(\boldsymbol{X}_i)]\hat{r}(\boldsymbol{X}_i) \\ &= \hat{\mu}^*_{WT} - \frac{1}{n}\sum_{i=1}^{n}\hat{m}(\boldsymbol{X}_i)\hat{r}(\boldsymbol{X}_i) + \frac{1}{n^*}\sum_{i=1}^{n^*}\hat{m}(\boldsymbol{X}^*_i)\end{aligned}$$

Estimation of m and r is typically based on parametric methods.

Doubly Robust Approach

Adjusting a Treatment Effect

A doubly robust estimator of δ^* is given by

$$\hat{\delta}^*_{DR} = \hat{\delta}^*_{IM} + \frac{1}{n}\sum_{i=1}^{n} \hat{r}(\boldsymbol{X}_i)\left[D_i - \hat{d}(\boldsymbol{X}_i) - (T_i - \pi)\hat{h}(\boldsymbol{X}_i)\right]$$

where h is some generic estimate of

$$h(x) = \frac{m_1(x)}{\pi} + \frac{m_0(x)}{1-\pi} = \mathrm{E}\left[\frac{TY}{\pi^2} + \frac{(1-T)Y}{(1-\pi)^2}\middle|\boldsymbol{X} = \boldsymbol{x}\right].$$

Estimation of d, r, and h is typically based on parametric methods.

Machine Learning in DR Methods

We consider estimating the nuisance functions using statistical machine learning methods.

- Let f represent a nuisance function: m, d, r, or h.
- Assume that there exists a limit function f_∞ such that, with probability 1, $\hat{f}(x) \to f_\infty(x)$ for all $x \in \boldsymbol{X}$.
- $\hat{\mu}^*_{DR}$ is consistent for μ^* if $m_\infty = m$ or $r_\infty = r$.
- $\hat{\delta}^*_{DR}$ is consistent for δ^* if $d_\infty = d$ or $r_\infty = r$.

Adjusting a Mean Outcome

For $\sqrt{n}$-consistency and asymptotic normality, we assume

$$m_\infty = m, \quad r_\infty = r, \quad \text{and} \quad ||\hat{m} - m||_2 ||\hat{r} - r||_2 = o_p(n^{-1/2}),$$

where $||.||_2$ denotes the L_2-norm with respect to the distribution of $\boldsymbol{X}$.

Adjusting a Treatment Effect

For $\sqrt{n}$-consistency and asymptotic normality, we assume

$$d_\infty = d, \quad r_\infty = r, \quad \text{and} \quad ||\hat{d} - d||_2 ||\hat{r} - r||_2 = o_p(n^{-1/2}).$$

Under the aforementioned assumptions as well as some regularity conditions including a Donsker condition

Adjusting a Mean Outcome

$\sqrt{n}(\hat{\mu}^*_{DR} - \mu^*)$ converges to a normal distribution with mean 0 and variance

$$\text{var}\{[Y - m(\boldsymbol{X})]r(\boldsymbol{X})\} + \sqrt{\frac{n}{n^*}}\text{var}[m(\boldsymbol{X}^*)]$$

Adjusting a Treatment Effect

$\sqrt{n}(\hat{\delta}^*_{DR} - \delta^*)$ converges to a normal distribution with mean 0 and variance

$$\text{var}\{r(\boldsymbol{X})[D - d(\boldsymbol{X}) - (T - \pi)h_\infty(\boldsymbol{X})]\} + \sqrt{\frac{n}{n^*}}\text{var}[d(\boldsymbol{X}^*)].$$

Adjusting a Mean Outcome

- This is the nonparametric variance bound for estimating μ^*.
- Thus, $\hat{\mu}^*_{DR}$ is asymptotically efficient in the nonparametric sense.

Adjusting a Treatment Effect

- When $h_\infty = h$, the asymptotic variance becomes the nonparametric variance bound for estimating δ^*.
- Then $\hat{\delta}^*_{DR}$ is asymptotically efficient in the nonparametric sense.

The Super Learner

How does one choose the optimal machine learning approach?

Consider the principle of super learning.

- Combines several candidate learners to create one "super learner".
- Involves the use of cross-validation to select among many candidate methods to compute a single learner.
- The final learner is a weighted combination of the candidates.

Sample Splitting

- Even with the super learner, efficiency and $\sqrt{n}$-consistency of $\hat{\mu}^*_{DR}$ and $\hat{\delta}^*_{DR}$ depend on a Donsker condition.
  - This imposes a limitation on the class of algorithms that can be included in the super learner.
- Sample splitting, or cross-fitting, may be used to remove the Donsker condition while retaining efficiency and $\sqrt{n}$-consistency.

Sample Splitting

- The entire sample $\{(\boldsymbol{X}_i, T_i, Y_i), i = 1, \ldots, n\} \cup \{\boldsymbol{X}_i^*, i = 1, \ldots, n^*\}$ is partitioned randomly into L roughly equally-sized subsamples.
- Let S_i and S_i^* be independent and uniformly distributed on $\{1, \ldots, L\}$.
- The lth subsample consists of $\{(\boldsymbol{X}_i, T_i, Y_i) : S_i = l\} \cup \{\boldsymbol{X}_i^*, S_i^* = l\}$.
- For every $l \in \{1, \ldots, L\}$, temporarily exclude the lth subsample.
- Obtain nuisance functions (e.g., $\hat{m}^{(-l)}$) from the rest of the sample.

Doubly Robust Estimator Based on Sample Splitting

Adjusting a Mean Outcome

Then μ^* is estimated using

$$\hat{\mu}^*_{DR2} = \frac{1}{n}\sum_{i=1}^{n^*} \hat{m}^{(-S^*_i)}(X^*_i) + \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}^{(-S_i)}(X_i)]\hat{r}^{(-S_i)}(X_i).$$

Adjusting a Treatment Effect

And δ^* is estimated as

$$\begin{aligned}\delta^*_{DR2} =& \frac{1}{n^*}\sum_{i=1}^{n^*} \hat{d}^{(-S^*_i)}(X^*_i) \\ &+ \frac{1}{n}\sum_{i=1}^{n} \hat{r}^{(-S_i)}(X_i)\left[D_i - \hat{d}^{(-S_i)}(X_i) - (T_i - \pi)\hat{h}^{(-S_i)}(X_i)\right].\end{aligned}$$

Doubly Robust Estimator Based on Sample Splitting

Adjusting a Mean Outcome

- $\hat{\mu}^*_{DR2}$ is consistent for μ^* if $m_\infty = m$ or $r_\infty = r$ or both.
- $\hat{\mu}^*_{DR2}$ is $\sqrt{n}$-consistent, asymptotically normal, and asymptotically efficient under our assumptions as well as some basic regularity conditions.
  - $\mathcal{X}^* = \mathcal{X}$
  - $\mathrm{E}(Y^*|\boldsymbol{X}^* = \boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x})$
  - $||\hat{m} - m||_2||\hat{r} - r||_2 = o_p(n^{-1/2})$
- These regularity conditions no longer include a Donsker condition.

Doubly Robust Estimator Based on Sample Splitting

Adjusting a Treatment Effect

- $\hat{\delta}^*_{DR2}$ is consistent for δ^* if $d_\infty = d$, $r_\infty = r$, or both.
- $\hat{\delta}^*_{DR2}$ is $\sqrt{n}$-consistent, asymptotically normal, and asymptotically equivalent to $\hat{\delta}^*_{DR}$ under our assumptions as well as some basic regularity conditions.
  - $\mathcal{X}^* = \mathcal{X}$
  - $\mathrm{E}\{Y^*(1) - Y^*(0)|\boldsymbol{X}^* = \boldsymbol{x}\}$
    $= \mathrm{E}(Y|T = 1, \boldsymbol{X} = \boldsymbol{x}) - E(Y|T = 0, \boldsymbol{X} = \boldsymbol{x})$
  - $||\hat{d} - d||_2||\hat{r} - r||_2 = o_p(n^{-1/2})$
- These conditions no longer include a Donkser condition.
- If also $h_\infty = h$, then $\hat{\delta}^*_{DR2}$ is asymptotically efficient in the nonparametric sense.

Simulation Study

Generate W from the trivariate standard normal distribution and generate Z according to

$$\text{logit}[P(Z=1|W)] = \begin{cases} W_1 - W_2 + W_3 & \text{(PS0)} \\ W_1 - W_2 + W_3 + 0.25 W_1 \text{sign}(W_2) & \text{(PS1)} \end{cases}.$$

Then we take a random sample of X from the conditional distribution $(W|Z=0)$ and a random sample of X^* from $(W|Z=1)$.

Simulation Study

Adjusting a Mean Outcome

Generate Y as

$$Y = \begin{cases} -0.5 + X_1 + X_3 + \epsilon & \text{(OR0)} \\ -1 + (X_1 \vee 0)^2 + X_3 + \epsilon & \text{(OR1)} \end{cases},$$

where $\vee$ denotes maximum and $\epsilon \sim N(0, 1)$.

Adjusting a Treatment Effect

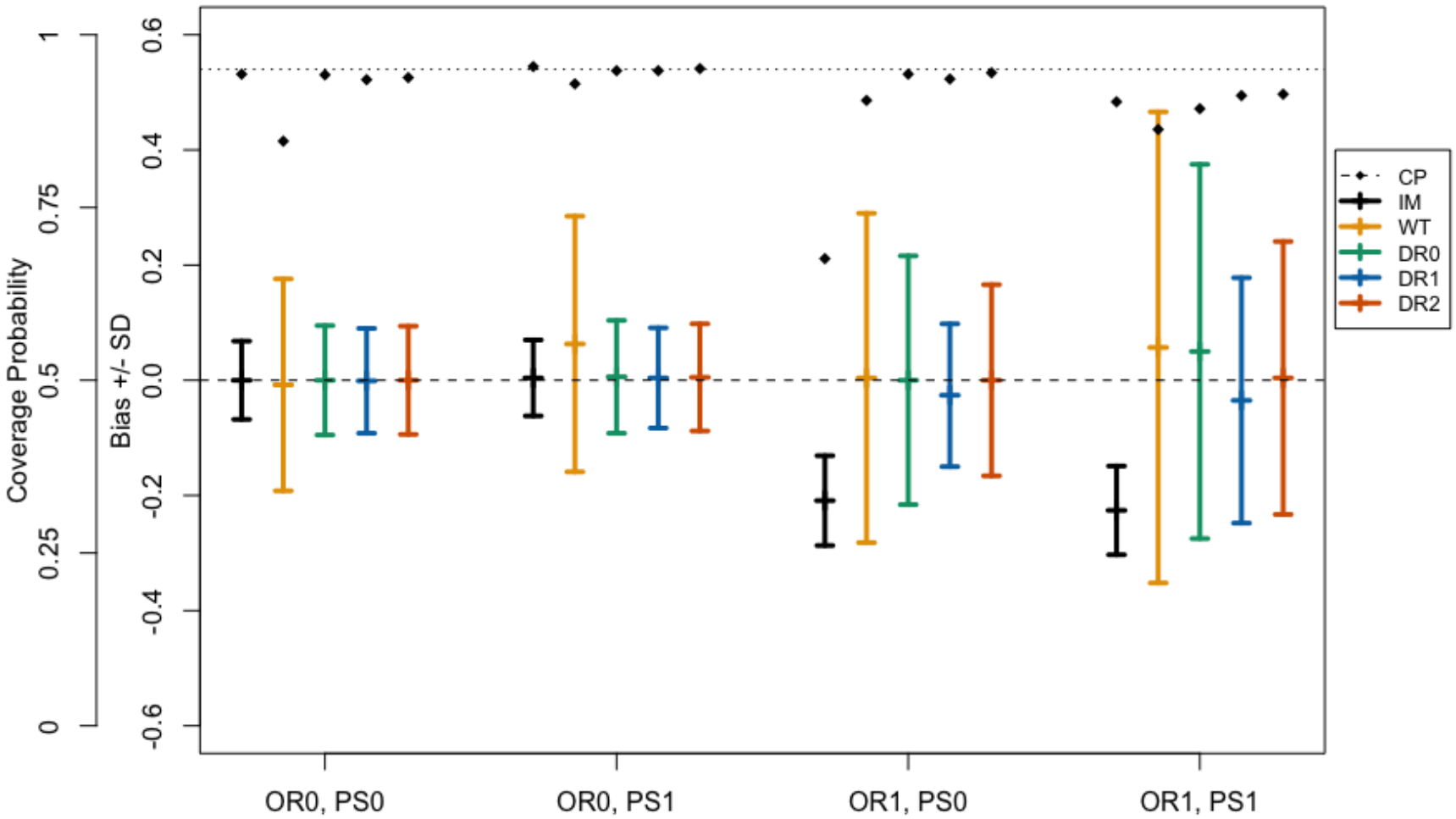
Generate a treatment indicator $T \sim \text{Bernoulli}(\pi)$, $\pi = P(T = 1) = 1/2$.

$$Y = \begin{cases} -0.5 + X_1 + X_3 + T - 0.5TX_3 + \epsilon & \text{(OR0)} \\ -1 + (X_1 \vee 0)^2 + X_3 + T - 0.5TX_3 + 0.25TX_3^2 + \epsilon & \text{(OR1)} \end{cases},$$

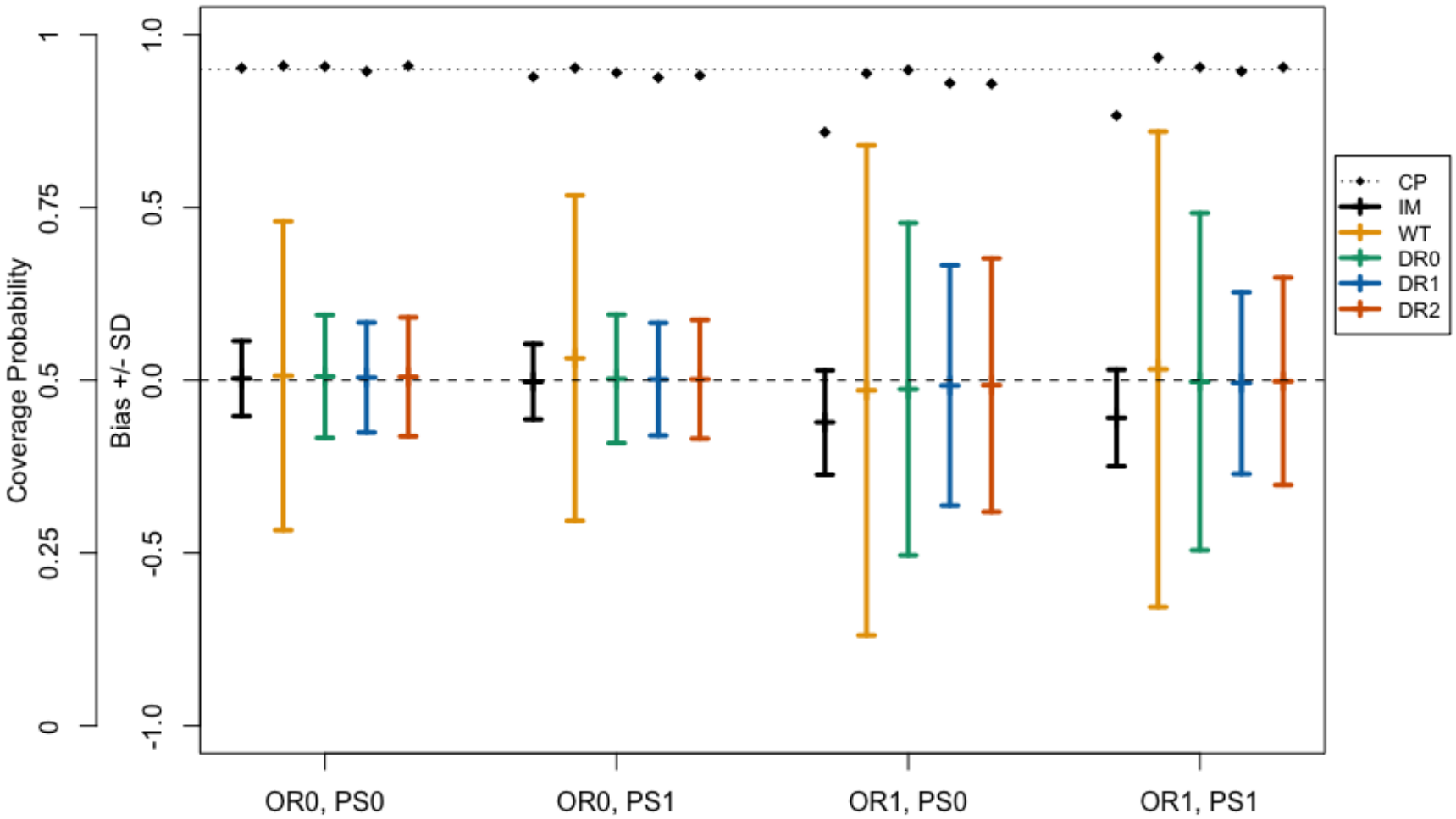
Simulation Study

- These methods are applied to 1000 replicate samples with $n = n^* = 1000$.
- The super learner library is based on
  - `glm` (generalized linear model)
  - `gam` (generalized addititve model)
  - `rpart` (recursive partitioning and regression tree)
- For the parametric methods, bootstrap standard errors are obtained from 200 bootstrap samples.
- Analytical standard errors are obtained for the nonparametric methods.

Simulation Results: Mean Outcome



Simulation Results: Treatment Effect



Ongoing and Future Research

Ongoing:

- Application to data on implantable cardioverter-defibrillators.
- Sensitivity analysis for the ignorability assumption.

Future:

- Machine learning methods applied to other estimators.
- Application to high dimensional data.
- Machine learning in causal inference for longitudinal data.

Thank you!

Removing the Donsker Condition

We use the following lemma to exploit the independence implied by sample splitting.

Lemma: Let $\hat{f}(\boldsymbol{o})$ be a function estimated from a sample $\boldsymbol{O}^N = (\boldsymbol{O}_{n+1}, \ldots, \boldsymbol{O}_N)$, and let $\mathbb{P}_n$ denote the empirical measure over $(\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n)$, which is independent of $\boldsymbol{O}^N$. Then

$$(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) = O_{\mathbb{P}}\left(\frac{||\hat{f} - f||}{\sqrt{n}}\right).$$

Source: E H Kennedy, S Balakrishnan, and M G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. arXiv:1801.03635. 2018.

The Super Learner

We want to estimate $m_0(\boldsymbol{X}) = \mathrm{E}_0(Y|\boldsymbol{X})$ for some $Y \in \mathcal{Y}$, $\boldsymbol{X} \in \mathcal{X}$. Define the regression as the minimizer of the expected squared loss,

$$m_0 = \arg\min_{\alpha_0} \mathrm{E}_0 L(O, \alpha),$$

where $L(O, \alpha) = [Y - m(\boldsymbol{X})]^2$.

Given candidate learners $\hat{m}_k$, $k = 1, \ldots, K$, the super learner is a linear combination of the candidates with coefficients determined via cross-validation.

Source: M Van der Laan, E Polley, and A Hubbard. Super Learner. U.C. Berkeley Division of Biostatistics Working Paper Series, 2007.

The Super Learner

1. Randomly partition the sample $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ into J roughly equally sized subsamples.
2. For each $j \in \{1, \ldots, J\}$, use the jth subsample as a validation sample and combine the other subsamples into a training sample.
3. Obtain $\hat{m}_k^{(-j)}$ from this training sample using the same method used for obtaining $\hat{m}_k$.
4. Find he coefficients for the training sample using

$$(\hat{\alpha_1}, \ldots, \hat{\alpha_K}) = \arg \min_{(\alpha_1, \ldots, \alpha_K)} \sum_{i=1}^{n} \left[Y_i - \sum_{k=1}^{K} \alpha_k \hat{m}_k^{(-j_i)}(X_i) \right]^2,$$

   such that $\sum_{k=1}^{K} \alpha_k = 1$ and $\alpha_k \geq 0\ \forall k$.
5. The super learner estimate of m is $\hat{m}_{SL} = \sum_{k=1}^{K} \hat{\alpha}_k \hat{m}_k$.

Simulation Results: Mean Outcome

OR0-PS0, $\mu^* \approx 0.16$					OR0-PS1, $\mu^* \approx 0.14$				
Method	Bias	SD	RMSE	CP	Method	Bias	SD	RMSE	CP
IM	0.000	0.068	0.068	0.943	IM	0.004	0.066	0.066	0.954
WT	-0.008	0.184	0.185	0.846	WT	0.063	0.222	0.231	0.929
DR0	0.000	0.095	0.095	0.942	DR0	0.006	0.098	0.099	0.948
DR1	-0.001	0.091	0.091	0.935	DR1	0.004	0.087	0.087	0.948
DR2	0.000	0.094	0.094	0.938	DR2	0.005	0.093	0.093	0.951
OR1-PS0, $\mu^* \approx 0.08$					OR1-PS1, $\mu^* \approx 0.08$				
Method	Bias	SD	RMSE	CP	Method	Bias	SD	RMSE	CP
IM	-0.209	0.078	0.223	0.676	IM	-0.226	0.077	0.239	0.903
WT	0.004	0.286	0.286	0.905	WT	0.057	0.409	0.413	0.863
DR0	0.000	0.216	0.216	0.943	DR0	0.050	0.325	0.329	0.893
DR1	-0.026	0.124	0.127	0.936	DR1	-0.035	0.213	0.216	0.912
DR2	0.000	0.166	0.166	0.945	DR2	0.004	0.237	0.237	0.914

Simulation results for estimating a mean outcome: empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP). DR0 is the parametric and DR1 is the nonparametric DR method.

Simulation Results: Average Treatment Effect

OR0-PS0, $\delta^* \approx 0.84$					OR0-PS1, $\delta^* \approx 0.84$				
Method	Bias	SD	RMSE	CP	Method	Bias	SD	RMSE	CP
IM	0.005	0.109	0.109	0.952	IM	-0.004	0.109	0.109	0.939
WT	0.013	0.447	0.447	0.955	WT	0.064	0.471	0.475	0.952
DR0	0.011	0.178	0.178	0.954	DR0	0.004	0.186	0.186	0.945
DR1	0.008	0.159	0.159	0.947	DR1	0.003	0.163	0.163	0.938
DR2	0.010	0.172	0.172	0.955	DR2	0.003	0.172	0.172	0.941
OR1-PS0, $\delta^* \approx 1.09$					OR1-PS1, $\delta^* \approx 1.09$				
Method	Bias	SD	RMSE	CP	Method	Bias	SD	RMSE	CP
IM	-0.122	0.151	0.194	0.859	IM	-0.109	0.140	0.177	0.883
WT	-0.039	0.709	0.710	0.944	WT	0.032	0.688	0.678	0.967
DR0	-0.026	0.481	0.481	0.939	DR0	-0.004	0.488	0.488	0.953
DR1	-0.015	0.348	0.348	0.930	DR1	-0.008	0.263	0.263	0.947
DR2	-0.014	0.367	0.367	0.929	DR2	-0.003	0.300	0.300	0.953

Simulation results for estimating an average treatment effect: empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP). DR0 is the parametric and DR1 is the nonparametric DR method.