Machine Learning for Precision Optimization and Bias Correction in Randomized Trials with Missing Data

Amos O. Okutse[1], Joseph W. Hogan[1]
[1]Brown University, School of Public Health, Providence, RI, USA

Overview

We compare efficiency gains using machine learning (ML) for the adjustment model relative to linear regression in randomized trials under missing data and model misspecification. The potential of these methods is demonstrated using a simulation study and an application to a cluster randomized trial. We then suggest guidelines for using ML adjustment for bias correction and precision optimization.

Background

- Baseline covariate adjustment can improve the precision of treatment effect estimates in randomized trials.
- The correct adjustment model is usually unknown.
- With complete outcome data, linear regression adjustment improves precision even if it is not the right model.
- With incomplete data, covariate adjustment using an incorrect model can induce bias. We investigate whether:

(i) ML adjustment results in an approximately correct model and thus maximizes precision under complete data.

(ii) with incomplete data, ML adjustment alleviates bias attributable to model misspecification.

Simulation methods

- Simulations mimic randomized trials with sample sizes between 500 and 2000.
- Estimate the effect of a treatment, A on the outcome, Y using baseline covariates, X.
- We assume intention-to-treat (ITT) analyses and estimate

$$ATE = E(Y_1 - Y_0).$$

Table 1: Data generating mechanism and definition of notations used to define the correct outcome model.

Notation	Definition
$\boldsymbol{Z} = (\boldsymbol{Z_1}, \cdots, \boldsymbol{Z_4})^T$	p× 1 vector of true outcome-generating variables $Z_i \sim N(0,1)$.
$\boldsymbol{X} = (\boldsymbol{X_1}, \cdots, \boldsymbol{X_4})^T$	The p × 1 vector of actual observed baseline covariates derived from $\boldsymbol{Z}$ and which are used in adjustment models.
$\boldsymbol{A} = \boldsymbol{0}, \boldsymbol{1}$	A binary treatment variable denoting control and treatment arms, respectively, and assume $A \perp\!\!\!\perp \mathrm{X}$.
$\boldsymbol{Y_{a=0,1}}$	Potential outcomes under treatment and control status. Assume, $Y_a \perp\!\!\!\perp \mathrm{A}\vert\mathrm{X} \Rightarrow \mathrm{P}(\mathrm{Y_a}\vert\mathrm{X}) = \mathrm{P}(\mathrm{Y_a}\vert\mathrm{A},\mathrm{X}) = \mathrm{P}(\mathrm{Y}\vert\mathrm{A}=\mathrm{a},\mathrm{X})$
$\boldsymbol{R} \in (\boldsymbol{0}, \boldsymbol{1})$	Missingness indicator: $R = 1$ when Y is observed. $R \sim Ber(\hat{\pi})$ and $\hat{\pi}$ is the probability of observing a unit i. Assume $R \perp\!\!\!\perp Y\vert X$.
$\boldsymbol{Y}$	The continuous primary outcome; defined as $Y_i = \beta_0 + \theta A + \sum_{i=1}^{4} \beta Z_i + \epsilon_i$ and $\epsilon_i \sim N(0,\ \sigma \in (1, 45)$.

- Models for *E(Y|A, Z)* are often only rough approximations.
- We misspecify our adjustment model by using *E(Y|A, X)*.
- Impacts inference when $P(X|A=1) \neq P(X|A=0)$.
- Correct adjustment model terms given *X* would be: $\log(X_1), X_2, X_1^2 X_2, \frac{1}{\log(X_2)}, \frac{X_3}{\log(X_1)}, \sqrt{X_4}$, and A

Estimators

(i) Multiple Linear Regression (MLR)
- Predict potential outcomes under each treatment as a linear combination of covariates.

(ii) Bayesian Additive Regression Trees (BART)
- 'sum of trees' model + regularization prior.

(iii) Random Forests (RF)
- Reduce correlation and tree variance by introducing split randomness.
- Based on bagging.

(iv) eXtreme Gradient Boosted Tree Ensemble (XGBoost)
- Weak learners are boosted to enhance their performance.
- Regularized objective function to reduce overfitting.

(v) Super Learner (SL)
- Weighted linear combination of base learners minimizing the cross-validated risk.
- RF, MLR, and XGBoost used as base learners.
- All modeling in R using `*tidymodels*` package in R.

Note: Performance evaluations were based on relative % increase in efficiency defined as 100*(ratio of variances - 1), standard errors, and bias.

Results

- Gains in efficiency using ML adjustment relative to MLR under complete data and misspecification of the adjustment model are directly proportional to the proportion variance explained.
- Performance improves with increasing n (Table 2).

Table 2: Relative percentage increase in efficiency comparing ML to MLR adjustment by sample size and proportion variance explained under complete data.

	n = 500		n = 2000	
	Low R^2	High R^2	Low R^2	High R^2
Estimator	**RE**	**RE**	**RE**	**RE**
Correct model	1.01%	99.65%	9.47%	99.64%
BART	1.38%	85.28%	7.03%	90.35%
SL	-18.85%	70.91%	1.86%	89.51%
XGBoost	-4.23%	65.19%	3.63%	87.83%
RF	-57.36%	53.50%	1.55%	81.02%

RE: Relative efficiency; R^2: Proportion variance explained.

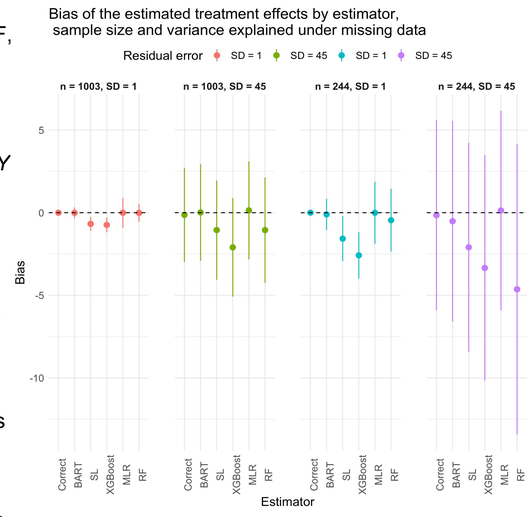
- For missing data, we explored whether ML enhances efficiency while correcting bias attributable to model misspecification.
- **87% gain in precision** using BART relative to LM under high variance explained and up to **3% gain** under low variance explained even under adjustment model misspecification (Table 3).

Table 3: Relative percentage increase in efficiency comparing ML to MLR adjustment by sample size and proportion variance explained under missing data.

	n = 258		n = 1003	
	Low R^2	High R^2	Low R^2	High R^2
Estimator	**RE**	**RE**	**RE**	**RE**
Correct model	9.50%	99.55%	8.07%	99.55%
BART	-0.86%	74.63%	3.12%	86.96%
SL	-9.66%	47.46%	-2.29%	78.01%
XGBoost	-27.46%	44.09%	-0.70%	75.62%
RF	-111.15%	-2.32%	-14.88%	64.02%

RE: Relative efficiency; R^2: Proportion variance explained

- Figure 1: Bias using RF, XGBoost, and SL remained high, especially under low sample size and proportion variance in *Y* explained the lines are SE of the effect estimates.
- Improvements in performance with large n.
- BART enhanced efficiency and kept bias low, unlike MLR.



Conclusion

- Covariate adjustment is almost always recommended in RCTs.
- With complete data, ML adjustment can improve efficiency relative to simpler models, e.g., MLR.
- Need to exercise causation under small sample sizes.
- With missing outcome data, ML can potentially reduce bias relative to misspecified.
- Strength: Basis for using ML covariate adjustment to enhance precision.
- This ensures that vain interventions are ruled out while ensuring efficient treatments are available in time to handle public health emergencies.
- Limitation: Relied on Missing At Random (MAR) assumption.
- Extensions of these methods could evaluate cross-fitted ML with non-continuous primary outcomes.

Research supported by NIH Grant D43 TW010050