Causal Effect Estimation and Transportation Using Modified Bootstrap

Tao Liu, PhD

Joint work with:
Whitney Su & Zeyuan Pei

Updated Nov 4, 2021

Outline

Background

- Observational Studies
- Causal Estimation
- Transportation to new population

Modified Bootstrap Method

- Assumptions
- Bootstrap procedure
- Asymptotic properties

Simulation study

Extension

- Double robustness
- Unmeasured confounding
- Unknown weights w_2

Discussion

Observational Data

Observational studies offer an important alternative to RCTs for studying the effect of a treatment on study subjects.

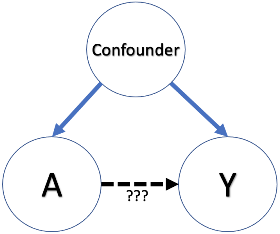
Observational data, such as electronic health records (EHR) and insurance claim data, represent one of the richest data sources for clinical research.

Analysis challenges:

- Non-randomized treatments –> confounding/selection bias.
- Source data population not representative of research (target) population –> covariate shift.

Confounding

Causal diagram



Structural model

$$A = \alpha_0 + \alpha_1 X + \epsilon_1$$
$$Y = \theta_0 + \theta_1 A + \theta_2 X + \epsilon_2$$

- Observational DB often includes a rich set of covariates hoping to capture most confounders.
- Commonly used methods: Regressions, inverse probability of treatment weighting (IPTW), stratification, matching, etc.

Transportation to new population

- Data (source) population and target study population are often different.
- Estimate from one population cannot be directly used on the other when treatment effects are heterogeneous.
- Setting inclusion/exclusion criteria are generally inadequate.

- Issues to consider:
  - There may exist many X's that differentiate the two populations.
  - How to make inference if estimation and transportation are accomplished in separate steps.

Goal: Look for a flexible and efficient method that can jointly solve these problems (Causal estimation, transportation, and inference).

Notations

Let $\mathcal{P}_O$ and $\mathcal{P}_T$ denote the observed (source) data population and study (target) population, respectively.

For the observed data population $\mathcal{P}_O$:

- X: p-vector of pre-treatment characteristics.
- A: Observed treatments taking values from discrete set $\{a\}$.
- Y: Observed outcome.
- $\mathbf{D}_n$: Observed dataset of n subjects,

$$\mathbf{D}_n = \{D_i = (X_i, A_i, Y_i)\}_{i=1}^{n}. \tag{1}$$

For both populations:

- $Y(a)$: denote the potential outcome of a subject if treatment set to a.

Notations (cont')

Let $\mathbb{E}_{\mathcal{P}_z}(\cdot)$ denote the expectation of a random variable from population $z \in \{O, T\}$.

Denote estimand of interest by Δ.

- Average causal effect:

$$\mathrm{ACE}(a, a') = \mathbb{E}_{\mathcal{P}_T} Y(a) - \mathbb{E}_{\mathcal{P}_T} Y(a').$$

- Causal risk ratio:

$$\mathrm{CRR}(a, a') = \frac{\mathbb{E}_{\mathcal{P}_T} Y(a)}{\mathbb{E}_{\mathcal{P}_T} Y(a')}.$$

- Causal odds ratio:

$$\mathrm{COR}(a, a') = \frac{\mathbb{E}_{\mathcal{P}_T} Y(a)/\{1 - \mathbb{E}_{\mathcal{P}_T} Y(a)\}}{\mathbb{E}_{\mathcal{P}_T} Y(a')/\{1 - \mathbb{E}_{\mathcal{P}_T} Y(a')\}}.$$

Covariate shift

Suppose that the difference between $\mathcal{P}_O$ and $\mathcal{P}_T$ that is related to Δ is captured by X.

Denote the distribution functions of X in the two populations by

$$G_O(x), \quad G_T(x)$$

and their corresponding density (mass) functions by

$$g_O(x), \quad g_T(x).$$

Exponential Tilting (ET):

$$g_T(x) \propto \exp\{h(x;\alpha)\}g_O(x).$$

Assumptions

A1: Stable unit treatment value assumption ($SUTVA$).

A2: *Strong ignorability*:

$$\{Y(a)\} \perp A \mid X.$$

A3: *Positivity of treatments*: For all $a \in \mathscr{A}$,

$$\Pr(A = a \mid X) > 0.$$

A4: *Absolute continuity:* G_T is absolutely continuous on G_O, so that $\mathrm{d}G_T/\mathrm{d}G_O$ is well defined.

A5: *Conditional exchangeability*: For $z \in \{O, T\}$,

$$\mathbb{E}_{\mathcal{P}_z}\{Y(a) - Y(a') \mid X\} =: \Delta_{\mathcal{P}_z}(X; a, a') = \Delta(X; a, a').$$

Bootstrap

Standard bootstrap method (Efron 1982):

- Approximate distributional properties of a *given* estimator.
- Re-sample data with equal probabilities of $1/n$.
- Popular method for estimating variances and constructing confidence intervals.

Modified bootstrap method

We propose to bootstrap $\mathbf{D}_n$ with re-sampling weights

$$w_i := w(X_i) \propto w_1(X_i) w_2(X_i)$$

where

- $w_1(X) = \sum_a \frac{\mathbf{1}(A=a)}{\Pr(A=a|X)}$: weight used in the IPTW method;
- $w_2(X) = \mathrm{d}G_T / \mathrm{d}G_O$: a version of the Radon–Nikodym derivative that captures the difference between the source and target study populations.
- If the ET assumption holds,

$$w_2(X) = \exp\{h(X; \alpha)\}.$$

Estimation:

- M-bootstrap m data points from $\mathbf{D}_n$, $m \gg n$.

$$\mathbf{D}^*_{m(n)} = \{D^*_i = (X^*_i, A^*_i, Y^*_i)\}_{i=1}^m. \tag{2}$$

- Estimate $\mathbb{E}_{\mathcal{P}_T}\{Y(a)\}$ by

$$\mu_n(Y(a)) = \frac{\sum_{\mathbf{D}^*_{m(n)}} \mathbf{1}(A^*_i = a) Y^*_i}{\sum_{\mathbf{D}^*_{m(n)}} \mathbf{1}(A^*_i = a)},$$

and plugin in the estimands, e.g. we estimate ACE by

$$\Delta_n := \Delta_n(a, a'; \mathbf{D}^*_{m(n)}) = \mu_n\{Y(a)\} - \mu_n\{Y(a')\}.$$

Inference:

- M-bootstrap n data points from $\mathbf{D}_n$; repeat for B times.

$$\mathbf{D}^*_{n(n),1}, \ldots, \mathbf{D}^*_{n(n),B}. \tag{3}$$

- From each bootstrapped dataset, calculate

$$\Delta^*_j = \Delta_n(a, a'; \mathbf{D}^*_{n(n),j}), \quad j = 1, \ldots, B.$$

  - We can approximate the standard error of Δ_n by the bootstrapped sample standard deviation

$$S^*_n = \frac{1}{B} \sum_{j=1}^{B} (\Delta^*_j - \Delta_n)^2.$$

  - Can use the empirical distribution of $\{\Delta^*_j\}$ to construct (e.g. percentile) CI of Δ_n as usual.

Large-sample properties

Theorem 1. (Consistency) As m and n tend to infinity,

$$\Delta_n(\cdot; \mathbf{D}^*_{m(n)}) \longrightarrow_{a.s.} \Delta.$$

Theorem 2. Let F^*_n and F_n denote the distributions functions of the following pivotal quantities,

$$\sqrt{n}(\Delta^*_j - \Delta_n) \mid \mathbf{D}_n \sim F^*_n,$$
$$\sqrt{n}(\Delta_n - \Delta) \sim F_n.$$

Then, F^*_n is “close” to F_n, in the sense that their distance in d_2 (the "Mallows") metric

$$d_2(F_n, F^*_n) \longrightarrow 0$$

as n tend to infinity.

Simulation study

Target population $\mathcal{P}_T$:

Suppose that

- Covariates

$$X = \begin{bmatrix} X_1 \\ \cdots \\ X_5 \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} 1 & -.2 & .3 & 0 & 0 \\ & 1 & .1 & 0 & 0 \\ & & 1 & 0 & 0 \\ & & & 1 & 0 \\ & & & 0 & 1 \end{bmatrix}\right).$$

- Potential outcomes under two treatments: Conditional on X,

$$\begin{bmatrix} Y(0) \\ Y(1) \end{bmatrix} \sim N\left(\begin{bmatrix} .1X_1 - .2X_2 - .1X_4 - .2(X_5^2 - 1) \\ 1 + .1X_1 + .1X_3 + .1X_4 + .2X_5 \end{bmatrix}, \begin{bmatrix} 1 & .2 \\ & 1 \end{bmatrix}\right).$$

- Estimand: $\text{ACE} = \mathbb{E}_{\mathcal{P}_T}\{Y(1) - Y(0)\} = 1$.

Observed data of $\mathcal{P}_O$:

We simulate

- Covariates X with a density

$$g_O(x) \propto \exp(\gamma x) g_T(x),$$

  where $\gamma = \mathbf{0}$ for Scenario 1; $\gamma = (.1, .1, -.1, .3, -.2)$ for Scenario 2.

- Treatment received

$$A \mid X \sim Bernoulli(\pi_X),$$

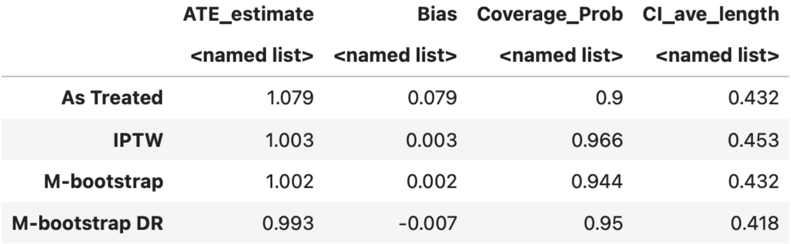
  where $\text{logit}(\pi_X) = .1X_1 - .1X_2 + .3X_3 - .2X_4 + .1X_5$.

- Potential outcomes $[Y(0), Y(1)]^\top \mid X$: same model as $\mathcal{P}_T$.
- Observed outcome: $Y = Y(A)$.

- Sample size: $n = 500$
- Bootstrap: $B = 1000$, $m = 500K$
- Experiments: 1000.

Simulation results

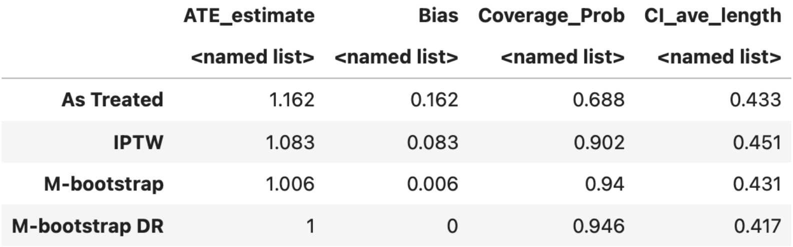
Scenario 1: No covariate shift



	ATE_estimate	Bias	Coverage_Prob	CI_ave_length
	<named list>	<named list>	<named list>	<named list>
As Treated	1.079	0.079	0.9	0.432
IPTW	1.003	0.003	0.966	0.453
M-bootstrap	1.002	0.002	0.944	0.432
M-bootstrap DR	0.993	-0.007	0.95	0.418

- CIs for IPTW are obtained using the "Huber-White" robust standard error. Overestimated; Reifeis & Hudgens (2020).
- CIs for M-Bootstrap obtained using the percentile method.

Scenario 2: When target and source populations differ



	ATE_estimate	Bias	Coverage_Prob	CI_ave_length
	<named list>	<named list>	<named list>	<named list>
As Treated	1.162	0.162	0.688	0.433
IPTW	1.083	0.083	0.902	0.451
M-bootstrap	1.006	0.006	0.94	0.431
M-bootstrap DR	1	0	0.946	0.417

Double robustness

- IPTW "removes" the causal pathway of the selection process: $X \longrightarrow A$ by imposing a propensity score model

$$e(a; x) = \Pr(A = a \mid X).$$

- We can also try to block the confounding pathway of $X \longrightarrow Y$ by imposing a model of the scientific process:

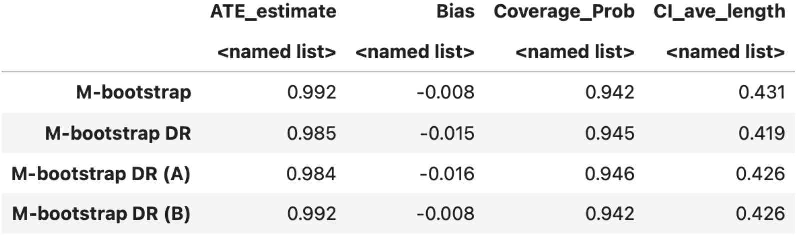
$$Y = \theta_0 + \theta_1 A + Q(A, X; \theta_2) + \epsilon.$$

with $\mathbb{E}\, Q(\cdot) = 0$, $\mathbb{E}\, \epsilon = 0$.

- The resulting estimator has a property of double robustness (DR).

Scenario 3: Double robustness

- DR (A): The PS model is incorrect.
- DR (B): The outcome model is incorrect.
- DR: Both models are correct.



	ATE_estimate	Bias	Coverage_Prob	CI_ave_length
	<named list>	<named list>	<named list>	<named list>
M-bootstrap	0.992	-0.008	0.942	0.431
M-bootstrap DR	0.985	-0.015	0.945	0.419
M-bootstrap DR (A)	0.984	-0.016	0.946	0.426
M-bootstrap DR (B)	0.992	-0.008	0.942	0.426

Unmeasured confounding

- The strong ignorability can be plausible in certain cases but more likely violated more or less.

- Let U denote the confounding effect of other factors beyond X.

- Assume a structural model:

$$\begin{aligned} h\{\Pr(A=a)\} &= \alpha_0 + \alpha_1 X + \alpha_2 U \\ Y &= \theta_0 + \theta_1 A + \theta_2 X + \theta_3 U + \epsilon \end{aligned}$$

- Some thoughts (not verified):
  - Calculate residuals $R = Y - (\widehat{\theta}_0 + \widehat{\theta}_1 A + \widehat{\theta}_2 X)$.
  - Include R in the PS model to conduct a sensitivity analysis (by varying its coefficient).

When data of $\mathcal{P}_T$ are available...

- So far, we assume that $w_2(X) = \mathrm{d}G_T/\mathrm{d}G_O$ is given.
- When w_2 is unknown but a data set of X of $\mathcal{P}_T$ is available:
  - Merge the data of X from $\mathcal{P}_O$ and $\mathcal{P}_T$ with proper labels.
  - Fit a logistic model or generalized additive model (GAM) with a logit link to predict the target population labels (source population as reference).
  - Calculate the "linear" predictor and use its exponentiated value as w_2.

Discussion

- Unlike standard BT which treats sampling variation as the sole source of uncertainty for inference, M-BT views uncertainty from both sampling as well as stochastic nature of treatment selection.
- M-bootstrap is computationally straightforward and simple.