A review of “On the Failure of the Bootstrap for Matching Estimators” (Abadie and Imbens; 2008)

Andrew Do, Kellie Ottoboni, Simon Walter

April 8, 2016

Introduction

Abadie and Imbens (2008)

- Notation and Assumptions
- The Bootstrap

Simulations

- Example 1
- Example 2

A theoretical view

Rescuing the bootstrap

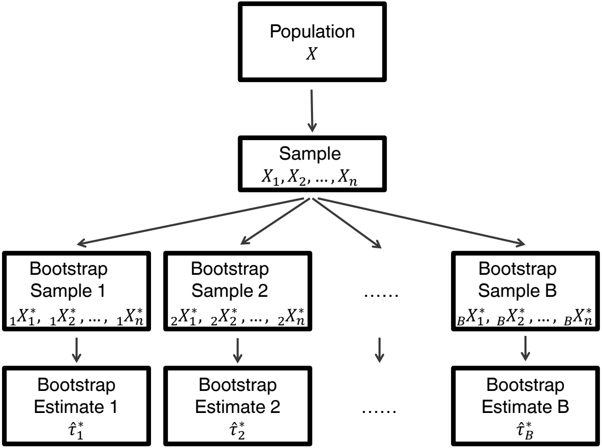
- Example 3

Problem

- Matching is sometimes done to control for pretreatment covariates in observational studies
- Matching estimators are nonlinear functions of the data and do not follow any nice known distribution
- **Problem:** how do you find standard errors for matching estimators?
- Two common ways: asymptotic approximations and resampling methods

The bootstrap

Sample with replacement from the observed data, pretending it is the population, to approximate the distribution of the statistic



On the failure of the bootstrap

The bootstrap estimate of the variance of the matching estimator $\hat{\tau}$ is given by

$$\hat{V}^B = \frac{1}{B}\sum_{b=1}^{B}(\hat{\tau}_b - \hat{\tau})^2$$

Abadie and Imbens show that $\hat{V}^B$ is not generally valid for matching estimators.

They focus on the case of one-to-one matching on a single continuous covariate.

Notation and Assumptions

- Suppose we have a random sample of N_0 units from the control population and a random sample of N_1 units from the treated population, with $N = N_0 + N_1$
- Each unit has a pair of potential outcomes, $Y_i(0)$ and $Y_i(1)$, under the control and active treatments
- Let W_i indicate treatment: we observe $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$
- In addition to the outcome, we observe a (scalar) covariate X_i for each individual

We're interested in the **average treatment effect for the treated** (ATT):

$$\tau = \mathbb{E}(Y_i(1) - Y_i(0) \mid W_i = 1)$$

Notation and Assumptions

We make the usual assumptions for matching:

- Unconfoundedness: For almost all x,

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i = x \text{ almost surely}$$

- Overlap: For some $0 < c < 1$ and almost all x,

$$c \leq \mathbb{P}(W_i = 1 \mid X_i = x) \leq 1 - c$$

Notation and Assumptions

D_i is the distance between the covariate values for observation i and the closest control group match:

$$D_i = \min_{j=1,\ldots,N : W_j = 0} \|X_i - X_j\|$$

$\mathcal{J}(i)$ is the set of closest matches for treated unit i.

$$\mathcal{J}(i) = \begin{cases} \{j \in \{1, \ldots, N\} : W_j = 0, \|X_i - X_j\| = D_i\} & \text{if } W_i = 1 \\ \emptyset & \text{if } W_i = 0 \end{cases}$$

If X is continuous, this set will consist of one unit with probability 1. In bootstrap samples, units may appear more than once.

Notation and Assumptions

Estimate the counterfactual for each treated unit as:

$$\hat{Y}_i(0) = \frac{1}{\#\mathcal{J}(i)} \sum_{j \in \mathcal{J}(i)} Y_i$$

The matching estimator of τ is then

$$\hat{\tau} = \frac{1}{N_1} \sum_{i:W_i=1} \left(Y_i - \hat{Y}_i(0) \right)$$

Notation and Assumptions

An alternative way of writing the estimator is

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} (W_i - (1 - W_i) K_i) Y_i$$

where K_i is the weighted number of times that unit i is used as a match:

$$K_i = \begin{cases} 0 & \text{if } W_i = 1 \\ \sum_{j: W_j = 1} 1\{i \in \mathcal{J}(j)\} \frac{1}{\# \mathcal{J}(j)} & \text{if } W_i = 0 \end{cases}$$

Bootstrap

- We obtain a **bootstrap sample** Z_b by taking a random sample with replacement from $Z = (X, W, Y)$.
- Let $\hat{\tau}_b = t(Z_b)$ be the matching statistic computed on bootstrap sample b.
- The bootstrap variance of $\hat{\tau}$ is the variance of $\hat{\tau}_b$ conditional on the original data Z:

$$V^B = \mathbb{E}\left[(\hat{\tau}_b - \hat{\tau})^2 \mid Z\right]$$

- We estimate it by generating B bootstrap samples from Z and taking the following average:

$$\hat{V}^B = \frac{1}{B}\sum_{b=1}^{B}(\hat{\tau}_b - \hat{\tau})^2$$

Bootstrap

- The bootstrap works for linear statistics that are asymptotically normal
- For nonlinear statistics, the bootstrap requires additional smoothness conditions.
- We will address each of these points in detail later.
- Earlier work also by Abadie and Imbens show that $\hat{\tau}$ is asymptotically normal and provides a consistent estimate of the variance.
- Why does the bootstrap fail here?

Bootstrap

Issue: the bootstrap fails to replicate the distribution of K_i, even in large samples.

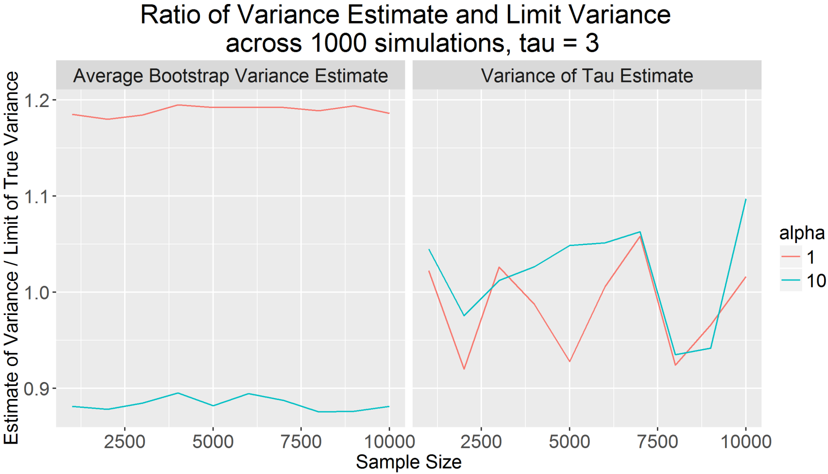
Example:

- Suppose the ratio N_1/N_0 is small (i.e. there are many more controls than treated)
- In the original sample, few controls are used as a match more than once
- In bootstrap samples, treated units may appear multiple times, creating situations where $\mathbb{P}(K_{b,i} > 1) > \mathbb{P}(K_i > 1)$

Basic Example from Abadie and Imbens

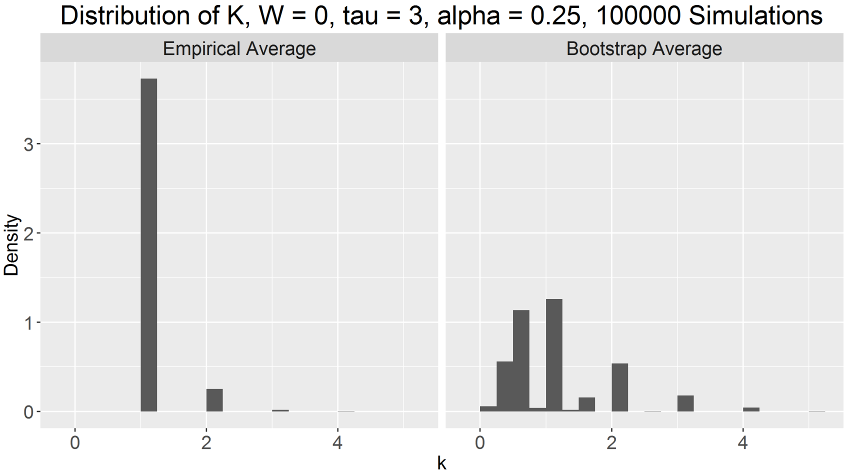
- The marginal distribution of X is uniform on $[0, 1]$
- The ratio of treated and control units is $\frac{N_1}{N_0} = \alpha$, fixed
- The propensity score is $e(x) = Pr(W_i = 1, X_i = x)$ is constant
- The last two items imply $e(x) = \frac{\alpha}{1+\alpha}$
- The distribution of $Y_i(i)$ with $Pr(Y_i(1) = \tau) = 1$ and the conditional distribution of $Y_i(0)$ given $X_i = x$ is standard normal.

Results



The bootstrap estimate of variance does not converge to the limiting value of $var(\hat{\tau})$. This is not fixed with increasing sample

Distribution of K_i

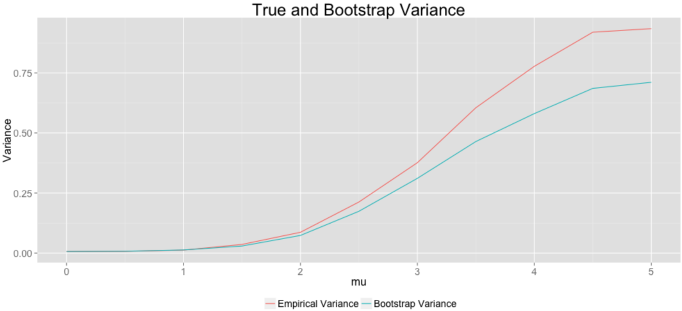


Note that the tail of the bootstrap distribution is much fatter due to the multiplicity of treatment units.

Effect of covariate distributions

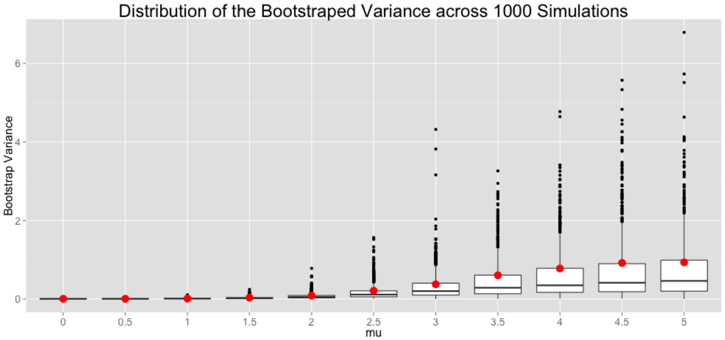
- Potential outcomes $Y(1)$ and $Y(0)$ defined as before
- Treatment assigned at random with $\frac{N_1}{N_0} = \alpha = 2$ fixed
- Change the covariate distributions: $X_i \sim N(0,1)$ if $W_i = 1$ and $X_i \sim N(\mu,1)$ if $W_i = 0$.
- We vary μ from 0 to 5

Results



The bootstrap tends to underestimate the variance of $\hat{\tau}$. The bias increases as the distance μ between treatment and control groups increases.

Results



Red points indicate the observed variance of the test statistic. The distribution of bootstrap variances has a long right tail. The skew worsens as μ grows.

Previous work on the bootstrap

- The bootstrap has been with us in its modern form since 1977.
- The procedure works in a very wide variety of situations but there are several famous examples where the bootstrap fails:
  - $X_i \overset{iid}{\sim} \mathscr{U}(0, \theta)$ and $\hat{\theta} = X_{(n)}$
  - $X_i \overset{iid}{\sim} \mathscr{N}(\mu, 1)$ where $\mu \in [0, \infty)$ and $\hat{\mu} = \max(0, \bar{X})$;
  - $X_i \overset{iid}{\sim} \mathscr{N}(\mu, 1)$ and $\hat{\mu} = \begin{cases} b\bar{X}_n \text{ if } |\bar{X}_n| < n^{-1/4} \\ \bar{X}_n \text{ if } |\bar{X}_n| \geq n^{-1/4} \end{cases} \quad b \in (0, 1)$
- There are some big theorems guaranteeing the consistency of the bootstrap in general situations.

When does the bootstrap work?

Csörgő and Mason (1989) prove a result regular linear statistics summarized by Mammen 1992.

Theorem

If $\theta(F) = \frac{1}{n}\sum_{i-1}^{n} h_n(X_i)$ *for some arbitrary function* h_n *then the bootstrap works in the sense that*

$$d_\infty\left[\mathscr{L}(\theta(\hat{F}) - \hat{t}_n), \mathscr{L}(\theta(F) - t_n)\right] \underset{p}{\rightarrow} 0$$

if and only if there exists σ_n *and* t_n *such that*

$$d_\infty\left[\mathscr{L}(\theta(F) - t_n), N(0, \sigma_n^2)\right] \underset{p}{\rightarrow} 0$$

Where $\hat{t}_n$ *is some function of our sample (typically the statistic itself) and* $t_n = \mathbb{E}(\hat{t}_n)$

The next theorem is from Politis, Romano and Wolf (1999). A statistic is Fréchet differentiable if

$$\theta(G) = \theta(F) + L_F(G - F) + o(\|G - F\|)$$

as $\|G - F\| \to 0$ for some linear functional L_F.

Theorem

Let $\mathscr{F}$ be the class of all distributions with finite support. Assume that F is drawn from $\mathscr{F}$ and the statistic $\theta(\cdot)$ is Frechet differentiable at F and L_F satisfies a certain condition. Then $\theta(F)$ is asymptotically normal and the bootstrap works in the sense of the previous theorem.

High level intuition for this theorem is that we want to have $\hat{F} \to F \Rightarrow \theta(\hat{F}) \to \theta(F)$. This is very similar to the usual definition of continuity from analysis.

Is there an if and only if result for smooth functionals

- There are cases of statistics that are not Frèchet differentiable for which the bootstrap works (eg the sample median) but ...
- According to Mammen (1992), van Zwet (1989) suggests smoothness is not only sufficient but necessary.
- van Zwet (1989) was a seminar given at a conference in 1989 in Germany and unfortunately I can find no record of it.
- Cursory comments in Mammen (1992) suggest the argument is by way of the Hoeffding decomposition theorem, but I don't currently understand this theorem or the proof strategy.

Placing Abadie and Imbens in the literature

- Some of this theoretical work on the bootstrap is not well known by practitioners, and it may be that early versions of Abadie and Imbens (2008) were written without detailed knowledge of this work.
- Arguably, the results of Abadie and Imbens are not surprising and not entirely novel to statisticians familiar with the theoretical work of the 1980s.
- The statistic of the matching estimator investigated is not smooth because if a treated X_i falls exactly at the midpoint between two control X_j, X_k the statistic changes discontinuously if we shift X_i infinitesimally closer to one of X_j or X_k.
- But the paper is very valuable because it highlights for practitioners one of the (many) limitations of the vanilla bootstrap.

Contribution to theoretical understanding of the bootstrap

- In a longer version of their paper Abadie and Imbens suggest this is the first case for which the bootstrap is inconsistent for a statistic that is asymptotically normal and $\sqrt{n}$-consistent.
- But this is not quite true. An example we have already discussed due to Beran (1982) also exhibits bootstrap failure and is asymptotically normal and asymptotically unbiased:

  $$\theta(X_1, \ldots, X_n) = \begin{cases} b\bar{X}_n \text{ if } |\bar{X}_n| < n^{-1/4} \\ \bar{X}_n \text{ if } |\bar{X}_n| \geq n^{-1/4} \end{cases}$$

  Although in this case bootstrap failure only occurs on a set of Lebesgue measure zero
- The proof of this fact is not easy.

An example of bootstrap inconsistency for an unbiased statistic

- It is also not too difficult to construct other simpler examples where the bootstrap fails on a parameter set with positive measure that are unbiased for finite samples and not just asymptotically; although they seem not to have previously appeared in the literature.
- We give one here: suppose X_i are drawn iid from $\{N(\mu, 1)\}_{\mu \in \mathbb{R}}$.
- Our estimate for μ is

$$\theta(\hat{F}) = \theta(X_1, \ldots, X_n) = \bar{X} + \#\{(i,j) : X_i = X_j, i \neq j\}$$

- Under the true sampling distribution the second summand is almost surely zero.
- But under the bootstrap distribution the second summand is at least one with probability $1 - n!/n^n$

An example of bootstrap inconsistency for an unbiased statistic

- The previous example is unnatural and unlikely arise in practice.
- But it does not seem hard to extend this to cases of practical interest involving ties.
- For example the critical value of Wilcoxon rank-sum and signed rank tests might be approximated using the bootstrap when ties are present in the data.
- Alternatively an analyst might try to approximate the distribution of the Kolmogorov-Smirnov test statistic using the bootstrap.

Should we follow Abadie and Imbens recommendation?

- Main conclusion is only their prior work based on asymptotic normality or subsampling have formal justification.
- Subsampling is seldom used in practice because it is sensitive to the choice of subsample size, and it is typically difficult to find the optimal subsample size from data.
- Using asymptotic normality is not satisfying either because it is only 'first order' correct; on the other hand the bootstrap is often 'second order' correct.
- What does this mean ... ?

- For many statistics of interest we can form Edgeworth expansions of the distribution function:

$$\mathbb{P}(\sqrt{n}(\hat{\theta} - \theta_0)/\sigma \leq x) = \Phi(x) + n^{-1/2} p_1(x)\phi(x) + \dots$$

- The $p_j(x)$ are polynomials with coefficients determined by low order moments of the statistic.
- The bootstrap distribution has a similar expansion:

$$\mathbb{P}^*(\sqrt{n}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma} \leq x) = \Phi(x) + n^{-1/2} \hat{p}_1(x)\phi(x) + \dots$$

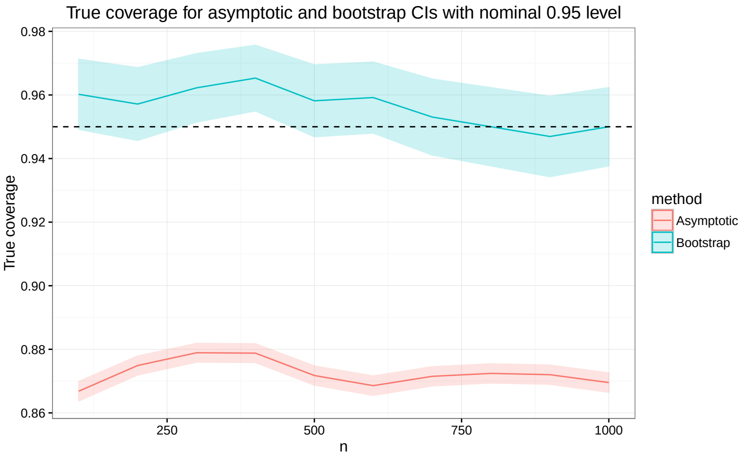
- Here the $\hat{p}_j(x)$ are $p_j(x)$ with population moments substituted for their empirical counterparts.
- Bootstrap is second order correct if

$$n^{-1/2}\hat{p}_1(x)\phi(x) = n^{-1/2} p_1(x)\phi(x) + o(n^{-1/2})$$

Can we rescue Efron's bootstrap for matching estimators?

- If we can make our statistic asymptotically smooth in an appropriate sense then we should be OK.
- A natural way to do this is to let the number of matches, m, increase with N_0.
- Tentative empirical results suggest the consistency of the bootstrap depends on the rate at which m grows.
  - $m \asymp \sqrt{N_0}$ yields intervals with incorrect coverage.
  - $m \asymp \log(N_0)$ yields intervals with asymptotically correct coverage.
- But in both cases the 2006 result of Abadie and Imbens using asymptotic normality seems to fail.

Simulation results for fixing the bootstrap



Conclusions

- Efron's bootstrap will not work for this kind of matching estimator with a single match on one covariate when the distribution of Y(0) is not degenerate.
- We can save the bootstrap by making the estimator smoother asymptotically or by subsampling.
- Although we did not discuss it there might be other fixes too including the m-out-of-n bootstrap and the wild bootstrap.
- There is also relatively recent work that suggests that Efron's bootstrap may be preferable even when it is not consistent because it outperforms other (possibly consistent) methods in finite samples.