

# 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

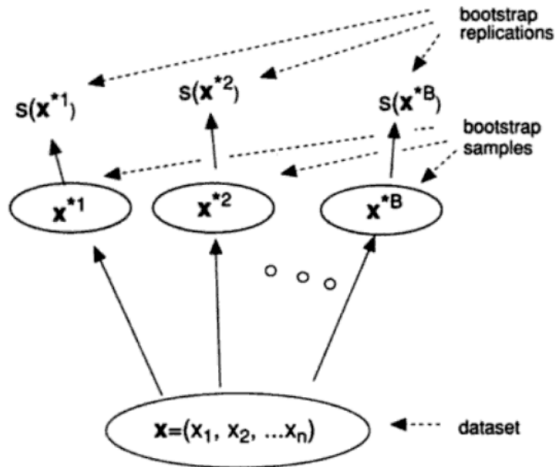
Example 1

Example 2

# 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 **TO DO: ANYONE KNOW OF A BETTER GRAPHIC? TO DO: CITE**

# 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:

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

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

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

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:

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

The matching estimator of  $\tau$  is then

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

# 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  $Z$ :

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

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

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

- ▶ The bootstrap works for linear statistics that are consistent and asymptotically normal
- ▶ For nonlinear statistics, the bootstrap requires additional smoothness conditions **TO DO: CITE?**
- ▶ **TO DO: ABADIE AND IMBENS SHOW THAT  $\hat{\tau}$  IS ASYMPTOTICALLY NORMAL WITH UGLY VARIANCE**
- ▶ Why does the bootstrap fail here?

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

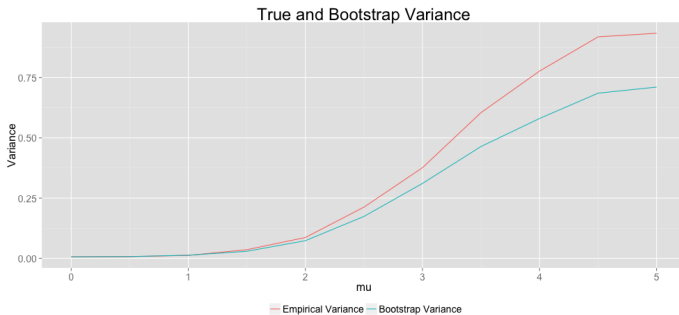
Example:

- ▶ 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  $\mathbb{P}(K_{b,i} > 1) > \mathbb{P}(K_i > 1)$  **TO DO: IS THIS TECHNICALLY CORRECT? IS THERE A BETTER WAY TO PUT THIS?**

# Effect of covariate distributions

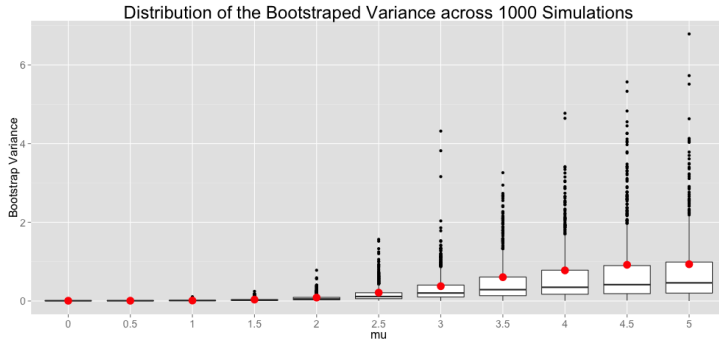
- ▶ Potential outcomes  $Y(1)$  and  $Y(0)$  defined as before
- ▶ Treatment assigned at random with  $\frac{N_1}{N_0} = \alpha = 2$  fixed
- ▶ Change the covariate distributions:  $X_i \sim N(0, 1)$  if  $W_i = 1$  and  $X_i \sim N(\mu, 1)$  if  $W_i = 0$ .
- ▶ We vary  $\mu$  from 0 to 5
- ▶ TO DO: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

# Results



The bootstrap tends to underestimate the variance of  $\hat{\tau}$ . The bias increases as the distance  $\mu$  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  $\mu$  grows.



## Placing Abadie and Imbens in the literature

- ▶ Csörgő and Mason (1989) established that linear statistics are consistently estimated by the bootstrap if and only if they are asymptotically normal.
- ▶ **TO DO: X (198?)** suggests rigorous results for non-linear statistics require in addition to asymptotic normality that the statistic is a smooth function of the data.
- ▶ Formalizing this requires a notion of smoothness of random quantities called Fréchet differentiability, but we will elide it.

## Placing Abadie and Imbens in the literature

- ▶ The revision history of the manuscript suggests that at least initially Abadie and Imbens were not particularly familiar with this prior work and some vestiges of this position remain.
- ▶ The results of Abadie and Imbens are not surprising and not novel to specialists familiar with the theoretical work of the 1980s.
- ▶ But it is valuable because it informs practitioners one of the (many) limitations of the vanilla bootstrap.

# Contribution to theoretical understanding of the bootstrap

- ▶ In a 2006 version of their paper Abadie and Imbens claim 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 true. Beran (1982) establishes that a Hodges-type estimator for the mean:

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

is not consistently estimated by the bootstrap when the true mean is zero.

- ▶ The proof of this fact is not easy and requires some knowledge of random measures.

# Contribution to theoretical understanding of the bootstrap

- ▶ In the final version of the paper, Abadie and Imbens emphasize the novelty of an example for which the bootstrap is inconsistent for a statistic that is asymptotically normal,  $\sqrt{n}$ -consistent and asymptotically unbiased.
- ▶ This is not the first example either because Beran's 1982 example is also asymptotically unbiased.
- ▶ It is also easy to construct simpler examples where the bootstrap fails that are unbiased in finite samples too, although they seem not to have previously appeared in the literature.

# An example of bootstrap inconsistency for an unbiased statistic

- ▶ We give one here: suppose  $X$  is drawn from the location family  $\{N(\mu, 1)\}_{\mu \in \mathbb{R}}$ .
- ▶ Our estimate for  $\mu$  is

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

- ▶ 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 was bizarre and unnatural.
- ▶ 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.

## Should we follow Abadie and Imbens recommendation?

- ▶ Main conclusion is only their prior work based on asymptotic normality or subsampling have formal justification.
- ▶ This is not satisfying: this is only 'first order' correct but the bootstrap is used because it 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 the bootstrap for matching estimators

- ▶ Yes! We think...

- ▶