Econ 512 Project Report

Yuanxing Long

March 2019

1 Motivation

This project is a replication of Adusumilli's job market paper "Bootstrap inference for propensity score matching". Before Adusumilli's work, there is no literature concerning the bootstrap methods on the treatment effects. The goal of this paper is to invent a valid bootstrap inference method for treatment effects when using the prevalent matching approach.

It has been a long history that treatment effects are studied in econometric literature. There are two basic approaches to estimating the average treatment effect, the weighting approach and the matching approach. The weighting approach is based on the fact that the conditional independence assumption will lead to a simple analytic expression for average treatment effect. It has good asymptotic properties and is favored by theoretical econometricans. However, it is not the case among empirical economists. Most empirical work uses the matching approach which is intuitively sensible.

Rosenbaum and Rubin (1983), a seminal paper on matching approach to estimating treatment effects, shows that the conditional independence assumption can result in an expression for average treatment effect where only the propensity score is concerned instead of the full set of covariates. This expression of average treatment effect indicates that the average treatment effect can be estimated on matching the propensity scores for each observation. Specifically, one can match each observation with the observation with the closet propensity score from the opposite treatment arm and take average over all observation. This matching estimator is usually used in empirical applications.

Another paper, Abadie and Imbens (2012) derives the asymptotic results for this matching estimator. A well known result by Hirano, Imbens and

Ridder (2001) is that the asymptotic variance of treatment effects will be smaller if the estimator propensity score instread of true propensity score is used in estimating treatment effects in the weighting approach. This fact still holds in the matching approach. Using estimated propensity score will also reduce the asymptotic variance of this matching estimator. Those findings make the studies of estimating treatment effects more complete. However, Abadie and Imbens (2008) shows that naive bootstrap inference method for average treatment is not valid.

What Adusumilli (2018) contributes to the treatment effect literature is a valid bootstrap inference method for propensity score matching estimator. The bootstrap procedure is consistent and can be used for inference on treatment effects. Monte Carlo Simulation results also show that the bootstrap inference method performs better in finite sample especially when the sample size is small.

This report will discuss the bootstrap inference method and replicate the Monte Carlo Simulation results of Adusumilli (2018).

2 Bootstrap Procedure

2.1 Setup

I will first introduce the basic setup and then illustrate the bootstrap procedure. The goal is to estimate the treatment W which takes 0 or 1 on an outcome Y. The potential outcome when $W = w \in \{0, 1\}$ is $Y(w)$ which cannot be always obeserved. What is observed is $Y = W * Y(1) + (1 - W) * Y(0)$. There are a set of covariates X to be controlled for. The data we have in hand is $(W_i, X_i, Y_i)_{i=1}^{N}$. Two assumptions usually made are:

Assumption 1: $(Y(0), Y(1))$ are independent of W conditional on covariates X.
Assumption 2: (Y_i, W_i, X_i) are i.i.d. realizations from joint distribution of (Y, W, X).

What we are interested in is estimating the average treatment effect, $\tau = \mathbb{E}(Y(1) - Y(0))$. An important role in this estimation is played by the propensity score, which is defined as $p(X) = Pr(W = 1|X)$. It is interpreted as the propensity to be treated. Introduce the notations $\mu(w, X) = \mathbb{E}(Y|W = w, X)$

and $\mu(w, p(X)) = \mathbb{E}(Y|W = w, p(X))$.

The average treatment effect can be written as

$$\tau = \mathbb{E}(\mu(1, p(X)) - \mu(0, p(X))).$$

This result is by Rosenbaum and Rubin (1983), which shows that assumption 1 implies that $(Y(0), Y(1))$ are independent of W conditional on the propensity score $p(X)$. This expression leads to an approach of estimating peer effect based on propensity score. The matching estimator considered is:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\Big(Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i;p(X))} Y_j\Big)$$

where $\mathcal{J}_M(i; p(X))$ is the observation i's M matches from the opposite treatment group, whose propensity score is closet to i's propensity score.

The propensity scores are estimated in the following. First, assume that propensity score is correctly specified, $p(X) = F(X'\theta)$ where θ is a parameter and function F is known. Then estimate the parameter as:

$$\hat{\theta} = \arg\max_{\theta} L(\theta|\mathbf{W}, \mathbf{X})$$

where $L(\theta|\mathbf{W}, \mathbf{X})$ is the associated maximum likelihood function.

Based on $\hat{\theta}$, one can get $\mathcal{J}_M(i; \hat{\theta})$ and thus $\hat{\tau}$. Abadie and Imbens (2016) shows that

$$\sqrt{N}(\hat{\tau} - \tau) \xrightarrow{d} N(0, \sigma^2 - c' I_{\theta_0}^{-1} c)$$

where σ^2 is the asymptotic variance for this matching estimator when true propensity score is used. $c' I_{\theta_0}^{-1} c$ is the deduction in the asymptotic variance when estimated propensity score is used. This project will also do several Monte Carlo Simulations using this asymptotic approach based on estimated propensity scores.

2.2 Estimation

Next, I will introduce the bootstrap procedure. First, define two variances:

$$e_{1i}(\theta) = \mu(1, F(X_i'\theta)) - \mu(0, F(X_i'\theta)) - \tau;$$
$$e_{2i}(w, \theta) = Y_i - \mu(w, F(X_i'\theta)).$$

And also define the potential errors for each observation i:

$$\epsilon_i(w, \theta) = e_{1i}(\theta) + (2w - 1)\Big(1 + \frac{\tilde{K}_M(i, w, \theta)}{M}\Big)e_{2i}(w, \theta); w = 0, 1,$$

where $\tilde{K}_M(i, w, \theta)$ is the number of time observation i is used as a match when he/she is treated $w = 1$ or not $w = 0$. After defining these items, the bootstrap procedure.

$\mu(w, F(X_i'\theta))$ is estimated by non-parametric methods, such as kernel estimation, series regression or smoothing splines. $\hat{e}_{2i}(w, \theta)$ is set equal to $\hat{e}_{2i}(W_i, \theta)$ if $w = W_i$, or $\hat{e}_{2\mathcal{J}_{NN}(i)}(W_{\mathcal{J}_{NN}(i)}, \theta)$ if $w \neq W_i$. $\mathcal{J}_{NN}(i)$ is i's match from the opposite treatment group, whose covariates are closet to i's covariates based on a metric for the full set of covariates.

The author follows a similar idea to construct the matching function $\tilde{K}_M(i, w, \theta)$. It is set equal to $K_M(i, \theta)$ if $w = W_i$ where $K_M(i, \theta)$ is the number of times i used as a match in the process of propensity score matching. In the case of $w \neq W_i$, the paper first divide the sample into q_N blocks based on q_N-quantiles of estimated propensity scores and then match observation i with the observations from the other treatment arm in the same block. $\tilde{K}_M(i, w, \theta)$ at $w \neq W_i$ is set to a weighted average of the matching functions for i's matches. Denote the weight vector as $\mathbf{M}(i)$ for each i. It completes the estimation of all objects that appear in the bootstrap procedure.

2.3 Bootstrap alogrithm

The bootstrap algorithm are implemented in the following steps.

Step 0: For each observation i, compute $\mathcal{J}_{NN}(i)$, the match based on distance between covariates and also obtain the weight vector $\mathbf{M}(i)$ which can be draw from a standardized multinominal distribution. These two objects are fixed throughout the bootstrap procedure.

Step 1: Draw N independent index random variables $\mathbf{S}^* = (S_1^*, ..., S_N^*)$ which corresponds to the non-parametric bootstrap draw sample $\mathbf{X}^* = (X_1^*, ..., X_N^*)$ which is drawn subsequently. Note that $X_i^* = X_{S_i^*}$.

Step 2: Draw the bootstrap treatment variables $\mathbf{W}^* = (W_1^*, ..., W_N^*)$ from the distribution $Bernoulli(F(X_i^{*\prime}\hat{\theta})$ for each i where $\hat{\theta}$ is the propensity score estimated from the original data.

Step 3: Discard bootstrap samples whose numbers of treatment or control observations are less than $M + 1$. And then obtain a new estimator $\hat{\theta}^*$ based on bootstrap sample $(\mathbf{W}^*, \mathbf{X}^*)$ by maximum likelihood estimation.

Step 4: Based on $\hat{\theta}^*$ and the original sample $\mathbf{W}, \mathbf{X}$, compute multiple objects $\tilde{K}_M(i, w, \theta)$, $(\hat{e}_{1i}(\hat{\theta}^*), \hat{e}_{2i}(w, \hat{\theta}^*))$ by using the matching techniques and non-parametric estimation methods discussed before. Then one can obtain the potential errors $\hat{\epsilon}_i(w, \hat{\theta}^*)$.

The bootstrap realized error is $\hat{\epsilon}_i^*(\hat{\theta}^*) = \hat{\epsilon}_{S_i^*}(W_i^*, \hat{\theta}^*)$. The center of bootstrap realized error is $C^*(\hat{\theta}^*) = \frac{1}{N}\sum_{i=1}^{N} \hat{\epsilon}_i(W_i, \hat{\theta}^*)$

The bootstrap statistic is

$$T_N^*(\hat{\theta}^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \{\hat{\epsilon}_i^*(\hat{\theta}^*) - C^*(\hat{\theta}^*)$$

Step 5: Compute the critical value by $c_{B,}^* = inf\{t : F_B^*(t) \geq 1 - \alpha\}$ where $F_B^*(t)$ is the empirical distribution of $T_N^*(\hat{\theta}^*)$ based on B bootstrap repetitions.

As shown in the original paper, under a set of assumptions, the bootstrap procedure is consistent in the sense that the statistic $T_N^*(\hat{\theta}^*)$ has the same asymptotic distribution as $\sqrt{N}(\hat{\tau} - \tau)$ and therefore the critical value generated from the bootstrap procedure can be used in the inference on the average treatment effect τ.

3 Simulation Design

This project intends to replicate the simulation results for four data generating processes.

The first DGP works as follows. Draw X_1, X_2 from a uniform distribution over $[-1/2, 1/2]$ and they independent. Potential outcomes are generated as

$Y(0) = 3X_1 - 3X_2 + U_0$ and $Y(1) = 5 + 5X_1 + 5X_2 + U_1$, where U_1 and U_0 are independent standard normal random variables. Propensity scores are simulated as:

$$p(X) = Pr(W = 1|X) = \frac{exp(X_1 + 2X_2)}{1 + exp(X_1 + 2X_2)}$$

and treatment W are generated from $Bernoulli(p(X))$. And the outcome variables are $Y = WY(1) + (1 - W)Y(0)$.

The second DGP is the same as the first one with the only difference that $Y(0) = -3X_1 + 3X_2 + U_0$ and $Y(1) = 5 + 7X_1 + 12X_2^2 + U_1$. It involves some non-linearity in the potential outcomes.

The third DGP is similar to the first one with the only difference that propensity scores are generated as

$$Pr(W = 1|X) = \frac{exp(X_1 + 7X_2)}{1 + exp(X_1 + 7X_2)}$$

This DGP will greatly reduce the balance of propensity score between the treated and control groups, compared with the first DGP.

The fourth DGP works differently. Four covariates X_1, X_2, X_3, X_4 are independent independent standarad normal random variables. The potential outcomes are $Y(1) = 210 + 27.4X_1 + 13.7X_2 + 13.7X_2 + 13.7X_3 + 13.7X_4 + U_1$ and $Y(0) = U_0$. And the propensity scores are generated as

$$Pr(W = 1|X) = \frac{exp(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)}{1 + exp(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)}$$

.

In all the simulations, M is set as 1 and the weight when computing the potential matching function is fixed for each bootstrap procedure. This simulation exercise will use the number of bootstrap repetition $B = 400$ and the number of Monte-Carlo simulations $R = 1000$. And different sample size $N = 100, 200, 500, 1000$ are used in the bootstrap procedure.

4 Issue

When implementing the bootstrap procedure, one need to match observations from the opposite arm based on propensity score, covaraites. It is complicated to construct the matching function which involves dividing the propensity score into several blocks. When the sample is large, each block might have a large number of observations, the storage of weights used in constructing $\tilde{K}_M(i, w, \theta)$ is of large size which is $O(N^2)$.

Another issue is about the selection of bandwidth when kernel estimation is used. Especially when the treated or control group is too small, a small bandwidth will cause problems. Therefore, I use a relatively larger bandwidth in this case.

The third issue is that when simulating the original sample, if the sample size is small, say $N = 100$, some DGPs especially DGP3, does not generate a balanced sample. The size of control or treated group might be too small to compute the potential matching function for some observations.

The fourth issue is about the choice of non-parametric estimation method. When doing the asymptotic inference, using kernel estimation for $\mu(w, X_i)$ will cause problems if the dimension of X_i exceeds 4. However, Using series estimation for $\mu(w, p(X_i))$ cause serious problems. Therefore, the choice of non-parametric estimation method is one big topic here.

5 Techniques

I overcome the issues with different techniques.

To address the issue of a large weight matrix of matching functions, I use a small sample size, i.e., a sample size which does not exceed 1000. If the sample size was as large as 10000, I would expect that computation would be costly.

To Address the problems caused by a small bandwidth when using kernel estimation, I employ a large bandwidth.

To tackle the issue that some samples are highly unbalanced, I discard these samples and draw from the data generating process again if it is an original sample or generate a new bootstrap sample based on the empirical distribution of the original if it is a bootstrap sample.

To overcome the issue that kernel estimation does not work well in high dimension when estimating $\mu(w, X_i)$, I use series estimation instead. It works well in this case. The reason is that the potential outcomes are linear in covaites in all DGPs except DGP3. The parametric specification for $\mu(w, X_i)$ is correct in the simulations. The estimates are much more reasonable. However, when estimating $\mu(w, p(X_i))$, I have tried both series estimations and kernel estimations. Kernel estimations work much better than series estimations. But it puzzles me that Adusumulli (2018) used series estimation and got reasonable rejection probabilities.

6 Results

Table 1: Rejection probabilities when the size of test is 0.05

DGP	Estimation Method	$N = 100$	$N = 200$	N=500	N=1000
DGP1	Bootstrap	0.087			
DGP1	Asymptotic	0.056	0.049	0.053	0.052
DGP2	Bootstrap	0.088			
DGP2	Asymptotic	0.059	0.047	0.044	0.053
DGP3	Bootstrap	0.089			
DGP3	Asymptotic	0.079	0.087	0.084	0.065
DGP4	Bootstrap	0.094			
DGP4	Asymptotic	0.030	0.030	0.040	0.045