Kevin's Notes on Matching Estimators for Nik

Kevin D. Duncan

June 21, 2017

Proposed Procedure

1. Estimate $\hat{\tau}^t$
2. Get residuals $\hat{\epsilon}_i = Y_i^{obs} - \hat{Y_i(0)} - \hat{\tau}^t$ for each treated observation i. (you constantly forget the hat on the epsilons, its important to denote that its a sample residual and not from the population!)
3. For each i, draw $u_i^* = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability} 1/2 \end{cases}$
4. Create N_1 bootstrap outcomes, $Y_i^* = \hat{Y_i(0)} + \hat{\tau}^t + u_i^* \hat{\epsilon}_i$
5. Estimate $\hat{\tau}_b^t$ on the bootstrap sample $Z = \{Y_i^*, W, X\}$
6. Repeat the steps 3-5 for $b = 1, ..., B$ bootstraps, and estimate $\hat{\tau}$ as the sample variance of $\hat{\tau}_b$ over the b bootstraps.

1 A Brief Rederivation of the ATET

Some Assumptions from [5] and [2]

Assumption 1.1. *1. Conditional on $W_i = w$ the sample consistss of independent draws from $Y, X \mid W = w$ for $w \in \{0, 1\}$. For some $r \leq 1$, $N_1^r / N_0 \to \theta \in (0, \infty)$*

2. X is continuously distributed on compact and convex support $X \subset R^k$. The density of X is bounded and bounded away from zero on $\mathbb{X}$.

3. W is independent of $Y_i(0)$ conditional on $X = x$ for almost every x. There exists a positive constant c such that $P(W = 1 \mid X = x) \leq 1 - c$ for almost every x

4. For $w = 0, 1$, $\mu(w, x)$ and $\sigma^2(w, x)$ are Lipschitz in $\mathbb{X}$, $\sigma^2(w, x)$ is bounded away from zero on $\mathbb{X}$, and $\mathrm{E}[Y^4 \mid W = w, X = x]$ is bounded uniformly on $\mathbb{X}$.

For each individual we can write the Average Treatment Effect on the Treated (ATET) as,[1]

$$\hat{\tau}_i^t = Y_i^{obs} - Y_{m_i^C}^{obs} = Y_i(1) - Y_{m_i^c}(0)$$

[1] The following notation is a bastard mix from [3] pp 415-416 merged with notation from [2]- sorry if its not perfectly clear.

When the matching is perfect both units of this pair have covariate values equal to that of the matched unit, e.g. $X_i = X_{m_i^c}$, with inexact matching, this isn't the case and $X_i \neq X_{m_i^c}$. Taking expectations over the population, define

$$\mu(0,x) = \mathrm{E}[Y_i(0) \mid X_i = x] \quad \mu(1,x) = \mathrm{E}[Y_i(1) \mid X_i = x]$$

Then taking expectations,

$$\mathrm{E}(Y_i^{obs} - Y_{m_i^c} \mid W_i = 1, X_i = X_{m_i^c} = x) = \mathrm{E}(Y_i^{obs} - Y_{m_i^c} \mid X_i = X_{m_i^c} = x) = \mu(1,x) - \mu(0,x) = \tau(x)$$

Now, let M be a bandwidth such that for every $j \in \{1, ..., N_c\}$ such that j is "closer" to i than M, then, $j \in J_M(i)$, being the set of matched pairs. Then, when we have $X_i \neq X_{m_i^c}$

$$\begin{aligned}\mathrm{E}_{sp}(Y_i^{obs} - Y_{m_i^c} \mid W_i = 1, X_i, X_{m_i^c}) = \mathrm{E}_{sp}(Y_i^{obs} - Y_{m_i^c} \mid X_i, X_{m_i^c}) &= \mu(1,X_i) - \mu(0,X_{m_i^c}) \\ &= \tau(x) + \mu(0,X_i) - \mu(0,X_{m_i^c})\end{aligned}$$

Thus we see in finite samples, when our matching isn't perfect, that there will be a residual bias term that differs from the ATET. [1] show that the bias term is stochastically bounded, in particular for the ATET it is $O_p(N^{-r/k})$, such that if $k > 2r$ the bias term maw disrupt the convergence in distribution. Regardless, for fixed n, for finite samples by summing across treated individuals,

$$\begin{aligned}\hat{\tau} = N_1^{-1} \sum_{W_i=1} \hat{\tau}_i^t &= \tau + N_1^{-1} \sum_{W_i=1} \mu(0,X_i) - \mu(0,X_{m_i^c}) \\ &= \tau + N_1^{-1} \sum_i^N W_i \left(M^{-1} \sum_{j \in J_M(i)} (\mu(0,X_i) - \mu(0,X_j)) \right)\end{aligned}$$

Now define,

$$B_n^t = N_1^{-1} \sum_i^N W_i \left(M^{-1} \sum_{j \in J_M(i)} (\mu(0,X_i) - \mu(0,X_j)) \right)$$

Then, we have no reason in finite samples to assume that

$$\mathrm{E}(B_n^t) = 0$$

Even when the expectation holds over the population.

The current proposed bootstrap outlined in Table **??** estimates $\hat{\tau}$, however, [5] and [1] both calculate $\tilde{\tau}^t = \hat{\tau}^t - B_n^t$. [1] show that this bias term is $O_p(N_1^{-r/k})$. As a result, $\sqrt{N_1} B_n^T$ is not $O_p(1)$ in general, and if k is large enough, the asymptotic distribution of $\sqrt{N_1}(\hat{\tau}^t - \tau^t)$ is dominated by the bias term and the matching estimator in general is not $\sqrt{N_1}$-consistent. In the case where $k = r = 1$, then we get that $\sqrt{N_1}(\hat{\tau}^t - \tau^t)$ will be asymptotically normal. But under Assumptions 1.1 we've tried to generalize this requirement, and thus (generally) require a bias corrected estimator for $\hat{\tau}^t$. Naturally this suggestsk,

$$\sqrt{(\hat{\tau}^t - B_n^t - \tau)}/\sigma_n^t \Rightarrow N(0,1)$$

But now note that this is NOT what the proposed estimator is currently doing. You are not taking the bias corrected form, and estimates still include the bias term for finite samples. The large n approximations required to make B_n^t be $o(1)$ aren't present in your simulations, and repeatedly sampling from the data generating process doesn't fix this as far as I can see.

Instead, let us assume we have a consistent nonparametric estimators $\hat{\mu}(1,x) \to^p \mu(1,x)$ and $\hat{\mu}(0,x) \to^p \mu(0,x))$ for all x. Then, the natural estimator that arises is,

$$\begin{aligned}
\tilde{\tau}^t &= \hat{\tau}^t - \hat{B}_n^t \\
&= N_1^{-1} \sum_{W_i=1} \left((Y_i^{obs} - M^{-1} \sum_{j \in J_M(i)} Y_j^{obs}) - \left(M^{-1} \sum_{j \in J_M(i)} (\hat{\mu}(0,X_i) - \hat{\mu}(0,X_j)) \right) \right) \\
&= N_1^{-1} \sum_{W_i=1} \left((Y_i^{obs} - \hat{\mu}(0,X_i) - M^{-1} \sum_{j \in J_M(i)} Y_j^{obs} - \hat{\mu}(0,X_j)) \right) \\
&= N_1^{-1} \sum_{W_i=1} \left((Y_i^{obs} - \hat{\mu}(1,X_i) - M^{-1} \sum_{j \in J_M(i)} (Y_j^{obs} - \hat{\mu}(0,X_j)) + \hat{\mu}(1,X_i) - \hat{\mu}(0,X_i) \right)
\end{aligned}$$

Moreover, for later notation, we can define $K_m(i) = \sum_{l=1}^n I\{i \in J_m(l)\}$ to be the number of times a particular observation appears in the match. Then,

$$\begin{aligned}
\tilde{\tau}^t &= N_1^{-1} \sum_{W_i=1} \left((Y_i^{obs} - \hat{\mu}(1,X_i) - M^{-1} \sum_{j \in J_M(i)} (Y_j^{obs} - \hat{\mu}(0,X_j)) + \hat{\mu}(1,X_i) - \hat{\mu}(0,X_i) \right) \\
&= N_1^{-1} \sum_{i=1}^N (W_i(\hat{e}_i + \hat{\epsilon}_i) + (1 - W_i) M^{-1} K_m(i) \hat{e}_i) \eta_i^*
\end{aligned}$$

This expression is the same as in [5] equation (6) (from their working paper, whose notation I find much easier to read). Then define $\hat{e}_i = Y_i^{obs} - \hat{\mu}(W_i, X_i)$, and $\epsilon_i = \hat{\mu}(1,X_i) - \hat{\mu}(0,X_i)$, such that we get

$$\tilde{\tau}^t = N_1^{-1} \sum_{W_i=1} \left((\hat{e}_i - M^{-1} \sum_j \hat{e}_j) + \epsilon_i \right)$$

[5] then draw a an iid sequence $\{\eta_i^*\}_{i=1}^N$ such that $\mathrm{E}[\eta_i \mid Y, X < W] = 0,\ \mathrm{E}[\eta_i^2 \mid Y, X < W] = 1,\ \mathrm{E}[\eta_i^4 \mid Y, X < W] < \infty$ (e.g. [4]) for each individual in the sample and then applies it to the above form to generate a wild bootstrap.

Your bootstrap estimator still might have the right variance, and I'll try to help with that below.

2 Variance

Assumption 2.1. *1. The marginal distribution of X is uniform on the interval* $[0,1]$

2. *The ratio of treated and control units is* $N_1/N_0 = \alpha$ *for some* $\alpha > 0$

3. *The propensity score* $e(x) = P(W_i = 1 \mid X_i = x)$ *is constant*

4. *The distribution of* $Y_i(1)$ *is degenerate with* $P(Y_i(1) = \tau) = 1$ *and the conditional distribution of* $Y_i(0)$ *given* $X_i = x$ *is* $N(0, 1)$

Under this framework we know that

$$\hat{\tau}^t = N_1^{-1} \sum_{W_i=1} Y_i^{obs} - \sum_{W_i=0} K_i Y_i \tag{1}$$

$$= \tau - N_1^{-1} \sum_{W_i=0} K_i Y_i \tag{2}$$

Conditioning on X, W we know that $\mathrm{E}(Y_i(0) \mid W_i = 0, X) = 0$ and has conditional variance 1. And that $\mathrm{E}(Y_i(1) \mid W_i = 1, X) = \tau$. Then the estimatiion errors are,

$$\begin{aligned}
\hat{\epsilon}_i &= Y_i^{obs} - \hat{Y_i(0)} - \hat{\tau}^t \\
&= Y_i^{obs} - \hat{Y_i(0)} - (N_1^{-1} \sum_i (W_i - (1 - W_i)K_i)Y_i \\
&= Y_i^{obs} - \hat{Y_i(0)} - N_1^{-1} \sum_{W_i=1} Y_i^{obs} + N_1^{-1} \sum_{W_i=0} K_i)Y_i
\end{aligned}$$

In your current draft you have a minus on this last term. Am I missing something? Then,

$$\begin{aligned}
\mathrm{E}(\hat{\epsilon}_i \mid X, W) &= \mathrm{E}(Y_i^{obs} - \hat{Y_i(0)} - N_1^{-1} \sum_{W_i=1)} Y_i^{obs} + N_1^{-1} \sum_{W_i=0} K_i Y_i) \\
&= \tau - 0 - \tau = 0
\end{aligned}$$

Note that under more general conditions this is not true, and the errors are no longer mean zero! Regardless, as a result we get,

$$\begin{aligned}
\mathrm{E}(\hat{\epsilon}_i^2 \mid X, W) &= \mathrm{E}((Y_i^{obs} - \hat{Y_i(0)} - N_1^{-1} \sum_{W_i=1)} Y_i^{obs} + N_1^{-1} \sum_{W_i=0} K_i)Y_i)^2 \mid X, W) \\
&= \mathrm{E}((\tau^t - \hat{Y_i(0)} - \tau^t + N_1^{-1} \sum_{W_i=0} K_i Y_i)^2 \mid X, W) \\
&= \mathrm{E}((-\hat{Y_i(0)} + N_1^{-1} \sum_{W_i=0} K_i Y_i)^2 \mid X, W) \\
&= \mathrm{E}((N_1^{-1} \sum_{W_i=0} K_i Y_i - M^{-1} \sum_{j \in J_m(i)} Y_j(0))^2 \mid X, W)
\end{aligned}$$

Since we know $Y_i(1) = \tau^t$, and $Y_i(0)$ is iid with mean 0 and variance 1. Moreover, by conditioning on X, W we know K_i, so the only stochastic element remains Y_i. Therefore,

$$\begin{aligned}\text{Var}(\hat{\epsilon}_i^2 \mid X, W) &= \text{E}(((N_1^{-1} \sum_{W_i=0} K_i Y_i)^2 - 2(N_1^{-1} \sum_{W_i=0} K_i Y_i) M^{-1} \sum_{j \in J_m(i)} Y_j(0))^2 + (M^{-1} \sum_{j \in J_m(i)} Y_j(0))^2 \mid X, W) \\ &= N_1^{-2} \sum_{W_i=0} K_i^2 - 2(N_1 M)^{-1} \sum_{j \in J_m(i)} K_j + M^{-1}\end{aligned}$$

We can still be constructive about the asymptotic distribution of (now pulling from your pdf on the github) about $\hat{\tau}_b$. Then, as u_i^* is iid with variance 1

$$\begin{aligned}\text{Var}^*[\hat{\tau}_b^t \mid X, W] &= N_1^{-2} \sum_i^{N_1} \text{E}(\hat{\epsilon}_i^2 \mid X, W) \\ &= N_1^{-2} \sum_{W_i=1} \left(N_1^{-2} \sum_{W_i=0} K_i^2 - 2(N_1 M)^{-1} \sum_{j \in J_m(i)} K_j + M^{-1} \right) \\ &= N_1^{-3} \sum_{W_i=0} K_i^2 - 2N_1^{-3} M^{-1} \sum_{W_i=1} \sum_{j \in J_m(i)} K_j + (MN_1)^{-1} \\ &= N_1^{-3} \sum_{W_i=0} K_i^2 + (MN_1)^{-1} (1 - 2N_1^{-2} \sum_{W_i=1} \sum_{j \in J_m(i)} K_j) \\ &= N_1^{-3} \sum_{W_i=0} K_i^2 + (MN_1)^{-1} (1 - 2N_1^{-2} \sum_i^{N_0} K_i)\end{aligned}$$

This feels wrong to me, but I'm still leaving it here. It appears as if the major problem is the construction of the errors $\hat{\epsilon}_i$ for $i = 1, ..., N_1$. We can compare this to the conditional variance of $\hat{\tau}^t$,

$$\text{Var}(\hat{\tau}^t \mid X, W) = \text{E}(N_1^{-1} \sum_i K_i Y_i) = N_1^{-2} \sum_i K_i^2$$

Under this framework, we never have that the variances of the two estimators converge. [5] do a wild bootstrap procedure that samples individual errors from the sample means, e.g. $\hat{\epsilon}_i = Y_i - \mu(W_i, x)$, which has better variance properties.

$$\text{Var}(\sqrt{N_1} T_n \mid Y, X, W) =$$

References

[1] A.L. Abadie and G.W Imbens. "Large Sample Properties of Matching Estimators for Average Treatment Effects". In: *Econometrica* 74 (1 2006), pp. 235–267.

[2] A.L. Abadie and G.W Imbens. "On the failure of the bootstrap for matching estimators". In: *Econometrica* 76 (6 2008), pp. 1537–1557.

[3] G.W. Imbens and D.B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[4] E. Mammen. "Bootstrap, wild bootstrap, and asymptotic normality". In: *Probability Theory Related Fields* 83 (1992), pp. 439–455.

[5] Taisuke Otsu and Yoshiyasu Rai. “Bootstrap inference of matching estimators for average treatment effects”. In: *Journal of the American Statistical Association* Forthcoming ().