Kevin's Notes on Matching Estimators for Nik

Kevin D. Duncan

July 7, 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 $u_i^* = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability} 1/2 \end{cases}$. I.e. the Rademacher distribution
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 observe an outcome variable Y_i^{obs}, a set of covariates, X_i, and whether or not an individual was treated, W_i, which is either 1, or 0. For both the treated and untreated groups, we can define the conditional means, $\mu(0, x) = \mathrm{E}[Y_i(0) \mid X_i = x] \quad \mu(1, x) = \mathrm{E}[Y_i(1) \mid X_i = x]$. Then, for each individual i, the Average Treatment Effect on the Treated (ATET) is

$$\tau_i^t = \mu(1, x) - \mu(0, x)$$

For each group, we have $Y_i^{obs} = \mu(W_i, X_i) + e_i$, whether the error structure may dependend on whether or not an individual was treated or not. We can also define $\hat{Y}_i(0)$ to be the "local average" of the M closest individuals who were in the control group to the treated individual (see [3] or [1] for a more explicit deviation of this). Assuming both groups have mean zero errors, we have the sample individual ATET,

$$\hat{\tau}_i^t = Y_i^{obs} - \hat{Y}_i(0) = Y_i(1) - \hat{Y}_i(0)$$

For finite samples the matching may not be perfect even under Assumptions (1.1. When the matching is perfect both units of this pair have covariate values equal to that of the matched unit, e.g. $X_i = X_{m_i^c}$. Assumptions 1.1 imply that as $n \to \infty$ with scalar covariates, the matching becomes precise. In finite samples, or with discrete covriates, with inexact matching, this isn't the case and $X_i \neq X_{m_i^c}$. Then,

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

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 that assume that

$$B_n^t = 0$$

Both [5] and [1] 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 $\hat{\tau}^t$.

Under assumptions 1.1 we know there are 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. $e_i = Y_i - \mu(W_i.X_i)$, and $\epsilon_i = \mu(1,X_i) - \hat{\mu}(0,X_i)$. Then,

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

This expression is the same as in [5] equation (6). 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.[1]

2 Variance

Now we return to the proposed bootstrap procedure. A first test of its validity is matching the results of [2] for their test case, their data generating process is outlined below

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

[1] Ideally the distribution should match the moments of the underlying process, so mean zero, with second, third, and fourth moments being zero.

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,

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

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

For the proposed bootstrap estimator, we have

$$\begin{aligned}
\hat{\epsilon}_i &= Y_i(1)^{obs} - \hat{Y}_i(0) - \hat{\tau}^t \\
&= \mu(x_i, 1) + e_i - \frac{1}{M} \sum_j (\mu(0, x_j) + e_j) - \hat{\tau}^t
\end{aligned}$$

Then, $\mathrm{Var}(\hat{\epsilon}_i \mid X, Y, W) = \sum_j \mathrm{Var}(e_j \mid X, Y, W) + \mathrm{Var}(\hat{\tau}^t \mid X, Y, W) + 2 \sum_j \mathrm{Cov}(\hat{\tau}^t, e_j \mid X, Y, W)$. As, under the assumptions, we know

$$\mathrm{Var}(e_i \mid X, Y, W) = \begin{cases} 0 & W_i = 1 \\ 1 & W_i = 1 \end{cases}$$

$$\mathrm{Var}(\hat{\tau}^t \mid X, Y, W) = N_1^{-2} \sum_i K_i^2$$

and $\mathrm{Cov}(\hat{\tau}^t, e_j \mid X, Y, W) = N_1^{-1} K_j$, and finally, that $\sum_{W_i=1} \sum_{j \in J_m(i)} 1 = \sum_{W_i=0} K_i$. As a result, we have

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

We can compare this to the conditional variance of $\hat{\tau}^t$,

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

A result of this is we can show conditions where the proposed estimator will have both larger and smaller variance than the actual conditional variance. In particular, the proposed estimator will have smaller conditional variance if,

$$
\begin{aligned}
N_1^{-2} \sum_{W_i=0} K_i^2 &\geq \frac{1}{N_1} + 3N_1^{-3} \sum_{W_i=0} K_i^2 \\
\sum_{W_i=0} K_i^2 \frac{N_1 - 3}{N_1^3} &\geq \frac{1}{N_1}
\end{aligned}
$$

But under this data generating process $\sum_{W_i=0} K_i = N_1$, as for each treated individual we only have one match, so we get that the proposed estimator will have smaller conditoinal variance than the DGP when

$$
\sum_{W_i=0} K_i^2 \geq \frac{N_1^2}{N_1 - 3}
$$

And larger conditional variance otherwise. From the left hand term, we see that the estimator is sensitive to both the relative size of N_0 to N_1 as well as the number of matches. For a better example of this, let us look at the unconditional variance. From [2] we know that

$$
\text{E}(K_i^2 \mid W_i = 0) = \frac{N_1}{N_0} + \frac{3}{2} \frac{N_1(N_1 - 1)(N_0 + 8/3)}{N_0(N_0 + 1)(N_0 + 2)}
$$

Thus, the above inequality for the UNconditional variance becomes

$$
\begin{aligned}
N_0\left(\frac{N_1}{N_0}+\frac{3}{2}\frac{N_1(N_1-1)(N_0+8/3)}{N_0(N_0+1)(N_0+2)}\right) &\geq \frac{N_1^2}{N_1-3}\\
\frac{N_0}{N_1^2}\left(\frac{N_1}{N_0}+\frac{3}{2}\frac{N_1(N_1-1)(N_0+8/3)}{N_0(N_0+1)(N_0+2)}\right) &\geq (N_1-3)^{-1}\\
\frac{1}{N_1}+\frac{3}{2}\frac{(N_1-1)(N_0+8/3)}{N_1(N_0+1)(N_0+2)} &\geq (N_1-3)^{-1}\\
\frac{N_1-3}{N_1}+\frac{3}{2}\frac{(N_1-3)(N_1-1)(N_0+8/3)}{N_1(N_0+1)(N_0+2)} &\geq 1
\end{aligned}
$$

for "large N" we have

$$
\begin{aligned}
1+\frac{3}{2}\frac{N_1N_1N_0}{N_1N_0N_0} &\geq 1\\
1+3/2\alpha &\geq 1
\end{aligned}
$$

It is important to remember that under fixed M asymptotics, that $\hat{Y}_i(0)$ is a random variable, and including it in the wild bootstrap estimator when trying to recover the variance of Y_i^{obs} for is going to add additonal terms into the conditional variance. In this case, there are two sources of bias. The first, by adding in $Y_i(0)$ over the preffered $\hat{\mu}(1,x_i)$, where $\hat{\mu}(\cdot)$ is a consistent estimator for $\mu(\cdot)$ (see [1] for conditions for this to hold, or [3] for parameetric estimators of this mean). Secondly, we know from the large sample theory that $\hat{\tau}^t$ is also a random variable, and, since we are not imposing a "wild bootstrap" component onto either of these terms, the covariance between them further biases the conditional variance. The clear solutions to these problems has already been covered in [5].

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 ().