Appendix to "Optimality of Matched-Pair Designs in Randomized Controlled Trials" (For Online Publication)

Yuehao Bai
Department of Economics
University of Michigan
yuehaob@umich.edu

October 12, 2022

Abstract

This document contains proofs of the results in the author's paper "Optimality of Matched-Pair Designs in Randomized Controlled Trials," as well as some additional results.

KEYWORDS: Matched-pair design, baseline outcome, stratified randomization, experiment, randomized controlled trial

JEL CLASSIFICATION CODES: C12, C13, C14, C90

A Proofs of Main Results

In the appendix, Q denotes the distribution of $((Y_i(1), Y_i(0), X_i))'$. I denote the observed quantities by $W_i = (Y_i, X_i', D_i)'$ and the pilot data by $\tilde{W}^{(m)} = ((\tilde{Y}_j, \tilde{X}_j', \tilde{D}_j)' : 1 \leq j \leq m)$. $\dim(X_i)$ denotes the dimension of X_i and $\text{supp}(X_i)$ denotes the support of X_i. I use $a \lesssim b$ to denote there exists $c \geq 0$ such that $a \leq cb$.

A.1 Proof of Lemma II.2

Let $\lambda = (\lambda_1, \ldots, \lambda_S)$ be a stratification and recall $n_s = |\lambda_s|$. Let $(d_1, \ldots, d_{2n})$ be a vector of values that $(D_1, \ldots, D_{2n})$ may take under λ, and for every s let $(d_1^s, \ldots, d_{n_s}^s)$ denote treatment status of the units in stratum s. For every s, there are $(n_s/2)!$ matched-pair designs in stratum s that could lead to $(d_1^s, \ldots, d_{n_s}^s)$. For each of such designs, $(d_1^s, \ldots, d_{n_s}^s)$ is realized with probability $2^{-n_s/2}$. Accordingly, if instead of implementing λ, I implement a matched-pair design in each stratum, uniformly across all matched-pair designs within each stratum, and independently across strata, the probability that $(d_1, \ldots, d_{2n})$ is realized is

$$\prod_{1 \leq s \leq S} \frac{1}{\binom{n_s}{n_s/2}(n_s/2)!/2^{n_s/2}} (n_s/2)! \frac{1}{2^{n_s}} = \prod_{1 \leq s \leq S} \frac{1}{\binom{n_s}{n_s/2}} ,$$

which is the probability that $(d_1, \ldots, d_{2n})$ is realized under λ. To see the number of matched-pair designs in a stratum with n_s units is

$$\binom{n_s}{n_s/2} (n_s/2)!/2^{n_s/2} ,$$

consider the following thought experiment: First, choose $n_s/2$ units and fix their positions; next, permute the rest $n_s/2$ units and match them to the fixed positions, and note each permutation leads to a matched-pair design; finally, note I have counted each matched-pair design repeatedly, and precisely $2^{n_s/2}$ times, because I could flip the positions of the two units within each pair. ■

A.2 Proof of Theorem IV.1

Follows immediately from Lemma B.3 with $\tau = \frac{1}{2}$. Note condition (c) on h in Theorem C.2 is satisfied because of Lemma B.5. ■

A.3 Proof of Theorem IV.2

To begin with, note $\hat{\mu}_n(d) \xrightarrow{P} E[Y_i(d)]$ and $\hat{\sigma}_n^2(d) \xrightarrow{P} \text{Var}[Y_i(d)]$ for $d \in \{0, 1\}$, by Lemma 6.5 in Bai, Romano and Shaikh (2021). Next, I show

$$E[\hat{\rho}_n | h^{(n)}] \xrightarrow{P} E[(E[Y_i(1) + Y_i(0) | h(X_i)])^2] . \tag{S.1}$$

For convenience, I define $\mu_d(h_i) = E[Y_i(d)|h(X_i) = h_i]$ for $d \in \{0,1\}$ and $g_h(h_i) = \mu_1(h_i) + \mu_0(h_i)$. To see this, note

$$
\begin{aligned}
&E[(Y_{\pi^h(4j-3)} + Y_{\pi^h(4j-2)})(Y_{\pi^h(4j-1)} + Y_{\pi^h(4j)})|h^{(n)}] \\
&= \frac{1}{4}(\mu_1(h_{\pi^h(4j-3)}) + \mu_0(h_{\pi^h(4j-2)}))(\mu_1(h_{\pi^h(4j-1)}) + \mu_0(h_{\pi^h(4j)})) \\
&+ \frac{1}{4}(\mu_1(h_{\pi^h(4j-3)}) + \mu_0(h_{\pi^h(4j-2)}))(\mu_1(h_{\pi^h(4j)}) + \mu_0(h_{\pi^h(4j-1)})) \\
&+ \frac{1}{4}(\mu_1(h_{\pi^h(4j-2)}) + \mu_0(h_{\pi^h(4j-3)}))(\mu_1(h_{\pi^h(4j-1)}) + \mu_0(h_{\pi^h(4j)})) \\
&+ \frac{1}{4}(\mu_1(h_{\pi^h(4j-2)}) + \mu_0(h_{\pi^h(4j-3)}))(\mu_1(h_{\pi^h(4j)}) + \mu_0(h_{\pi^h(4j-1)})) \\
&= \frac{1}{4}(g_h(h_{\pi^h(4j-3)}) + g_h(h_{\pi^h(4j-2)}))(g_h(h_{\pi^h(4j-1)}) + g_h(h_{\pi^h(4j)})) \\
&= \frac{1}{4} \sum_{k\in\{2,3\},l\in\{0,1\}} g_h^2(h_{\pi^h(4j-k)}) + g_h^2(h_{\pi^h(4j-l)}) - (g_h(h_{\pi^h(4j-k)}) - g_h(h_{\pi^h(4j-l)}))^2 \ .
\end{aligned}
$$

As a result,

$$
\begin{aligned}
E[\hat{\rho}_n|h^{(n)}] &= \sum_{1\le j\le \lfloor \frac{n}{2} \rfloor} E[(Y_{\pi^h(4j-3)} + Y_{\pi^h(4j-2)})(Y_{\pi^h(4j-1)} + Y_{\pi^h(4j)})|h^{(n)}] \\
&= \frac{1}{2n} \sum_{1\le i\le 2n} g_h^2(h(X_i)) - \frac{1}{4n} \sum_{1\le j\le \lfloor \frac{n}{2} \rfloor} \sum_{k\in\{2,3\},l\in\{0,1\}} (g_h(h_{\pi^h(4j-k)}) - g_h(h_{\pi^h(4j-l)}))^2 \ .
\end{aligned}
$$

By the assumption that $E[h^2(X_i)] < \infty$, (S.15) holds. (S.1) then follows from the assumption that $E[Y_i^r(d)|h(X_i) = z]$ is Lipschitz in z for $r = 1, 2$ and $d = 0, 1$, (S.15), the fact that

$$
\begin{aligned}
E[g_h^2(h(X_i))] &\lesssim E[E[Y_i(1)|h(X_i)]^2] + E[E[Y_i(0)|h(X_i)]^2] \\
&\le E[E[Y_i^2(1)|h(X_i)]] + E[E[Y_i^2(0)|h(X_i)]] = E[Y_i^2(1) + Y_i^2(0)] < \infty
\end{aligned}
$$

because of Assumption IV.1, and the weak law of large numbers.

It remains to show $\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}] \xrightarrow{P} 0$. I will prove

$$
\frac{2}{n} \sum_{1\le j\le \lfloor \frac{n}{2} \rfloor} (Y_{\pi^h(4j-2)} Y_{\pi^h(4j)} - E[Y_{\pi^h(4j-2)} Y_{\pi^h(4j)}|h^{(n)}]) \xrightarrow{P} 0 \ ,
$$

and the others follow similarly. I will repeatedly use the following elementary inequalities for any $a, b \in \mathbf{R}$ and $\lambda > 0$:

$$
\begin{aligned}
|a+b|I\{|a+b| > \lambda\} &\le 2|a|I\{|a| > \lambda/2\} + 2|b|I\{|b| > \lambda/2\} \\
|ab|I\{|ab| > \lambda\} &\le |a|^2 I\{|a| > \sqrt{\lambda}\} + |b|^2 I\{|b| > \sqrt{\lambda}\} \ .
\end{aligned}
$$

To begin with,

$$
E[Y_{\pi^h(4j-2)} Y_{\pi^h(4j)}|h^{(n)}] = \frac{1}{2}\mu_1(h_{\pi^h(4j-2)})\mu_0(h_{\pi^h(4j)}) + \frac{1}{2}\mu_1(h_{\pi^h(4j)})\mu_0(h_{\pi^h(4j-2)})
$$

Next, note

$$
\begin{aligned}
&\frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E[|Y_{\pi^h(4j-2)}Y_{\pi^h(4j)} - E[Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|h^{(n)}]| \\
&\qquad I\{|Y_{\pi^h(4j-2)}Y_{\pi^h(4j)} - E[Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|h^{(n)}]| > \lambda\}|h^{(n)}] \\
&\lesssim \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E[|Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|I\{|Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}| > \sqrt{\lambda/2}\}|h^{(n)}] \\
&\qquad + E[|E[Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|h^{(n)}]|I\{|E[Y_{\pi^h(4j-2)}Y_{\pi^h(4j)}|h^{(n)}]| > \sqrt{\lambda/2}\}|h^{(n)}] \\
&\lesssim \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E[Y^2_{\pi^h(4j-2)}I\{|Y_{\pi^h(4j-2)}| > \sqrt{\lambda/2}\}|h^{(n)}] + E[Y^2_{\pi^h(4j)}I\{|Y_{\pi^h(4j)}| > \sqrt{\lambda/2}\}|h^{(n)}] \\
&\qquad + |\mu_1(h_{\pi^h(4j-2)})\mu_0(h_{\pi^h(4j)})|I\{\mu_1(h_{\pi^h(4j-2)})\mu_0(h_{\pi^h(4j)}) > \lambda/2\} \\
&\qquad + |\mu_1(h_{\pi^h(4j)})\mu_0(h_{\pi^h(4j-2)})|I\{\mu_1(h_{\pi^h(4j)})\mu_0(h_{\pi^h(4j-2)}) > \lambda/2\} \\
&\lesssim \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E[Y^2_{\pi^h(4j-2)}(1)I\{|Y_{\pi^h(4j-2)}(1)| > \sqrt{\lambda/2}\}|h^{(n)}] \\
&\qquad + E[Y^2_{\pi^h(4j-2)}(0)I\{|Y_{\pi^h(4j-2)}(0)| > \sqrt{\lambda/2}\}|h^{(n)}] \\
&\qquad + E[Y^2_{\pi^h(4j)}(1)I\{|Y_{\pi^h(4j)}(1)| > \sqrt{\lambda/2}\}|h^{(n)}] + E[Y^2_{\pi^h(4j)}(0)I\{|Y_{\pi^h(4j)}(0)| > \sqrt{\lambda/2}\}|h^{(n)}] \\
&\qquad + \mu_1^2(h_{\pi^h(4j-2)})I\{|\mu_1(h_{\pi^h(4j-2)})| > \sqrt{\lambda/2}\} + \mu_0^2(h_{\pi^h(4j)})I\{|\mu_0(h_{\pi^h(4j)})| > \sqrt{\lambda/2}\} \\
&\qquad + \mu_1^2(h_{\pi^h(4j)})I\{|\mu_1(h_{\pi^h(4j)})| > \sqrt{\lambda/2}\} + \mu_0^2(h_{\pi^h(4j-2)})I\{|\mu_0(h_{\pi^h(4j-2)})| > \sqrt{\lambda/2}\} \\
&\lesssim \frac{1}{2n}\sum_{1\le i\le 2n} E[Y_i^2(1)I\{|Y_i(1) > \sqrt{\lambda/2}|\}|h(X_i)] + E[Y_i^2(0)I\{|Y_i(1) > \sqrt{\lambda/2}|\}|h(X_i)] \\
&\qquad + E[Y_i^2(1)|h(X_i)]I\{E[Y_i^2(1)|h(X_i)] > \sqrt{\lambda/2}\} + E[Y_i^2(0)|h(X_i)]I\{E[Y_i^2(0)|h(X_i)] > \sqrt{\lambda/2}\} \\
&\xrightarrow{P} E[Y_i^2(1)I\{|Y_i(1) > \sqrt{\lambda/2}|\}] + E[Y_i^2(0)I\{|Y_i(1) > \sqrt{\lambda/2}|\}] \\
&\qquad + E[E[Y_i^2(1)|h(X_i)]I\{E[Y_i^2(1)|h(X_i)] > \sqrt{\lambda/2}\}] \\
&\qquad + E[E[Y_i^2(0)|h(X_i)]I\{E[Y_i^2(0)|h(X_i)] > \sqrt{\lambda/2}\}] \,, \qquad \text{(S.2)}
\end{aligned}
$$

where the last line follows from WLLN and the law of iterated expectation. Since by Assumption IV.1 I have $E[Y_i^2(d)] < \infty$ and hence $E[E[Y_i(d)|h(X_i)]^2] < E[Y_i^2(d)]$ by Jensen's inequality, the limit as $\lambda \to \infty$ of the last line is 0, by the dominated convergence theorem. I finish the proof by arguing by contradiction. Suppose

$$\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}]$$

does not converge in probability to 0. There must then exist $\epsilon > 0$ and $\delta > 0$ and a subsequence, which for simplicity I again denote by $\{n\}$, such that

$$P\{|\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}]| > \epsilon\} \to \delta \qquad \text{(S.3)}$$

along this subsequence. But because of (S.2), there exists a further subsequence along which the condition in Lemma 6.3 of Bai, Romano and Shaikh (2021) holds with probability one for $h^{(n)}$, but then along this subsequence $\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}] \xrightarrow{P} 0$ conditional on $h^{(n)}$ with probability one for $h^{(n)}$,

i.e., for any $\epsilon > 0$, with probability one for $h^{(n)}$,

$$P\{|\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}]| > \epsilon|h^{(n)}\} \to 0 \ .$$

Since probabilities are bounded and hence uniformly integrable,

$$P\{|\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}]| > \epsilon\} \to 0$$

along the chosen subsequence, which implies a contradiction to (S.3). ■

A.4 Proof of Theorem IV.3

By repeating the arguments in the proof of Lemma B.3, I write

$$\sqrt{n}(\hat{\theta}_n - \theta(Q)) = A_n - B_n + C_n - D_n \ ,$$

where

$$\begin{aligned}
A_n &= \frac{1}{\sqrt{n}} \sum_{1 \le i \le 2n} (Y_i(1)D_i - E[Y_i(1)D_i|h^{(n)}, D^{(n)}]) \\
B_n &= \frac{1}{\sqrt{n}} \sum_{1 \le i \le 2n} (Y_i(0)(1-D_i) - E[Y_i(0)(1-D_i)|h^{(n)}, D^{(n)}]) \\
C_n &= \frac{1}{\sqrt{n}} \sum_{1 \le i \le 2n} (E[(Y_i(1)+Y_i(0))D_i|h^{(n)}, D^{(n)}] - D_i E[Y_i(1)+Y_i(0)]) \\
D_n &= \frac{1}{\sqrt{n}} \sum_{1 \le i \le 2n} (E[Y_i(0)|h^{(n)}, D^{(n)}] - E[Y_i(0)]) \ .
\end{aligned}$$

Note by the assumption that $E[h^2(X_i)] < \infty$, Assumption IV.3, and Lemma B.6, (S.16) holds. Since $0 < E[\text{Var}[Y_i(d)|h(X_i)]]$ for $d \in \{0,1\}$, $E[Y_i^r(d)|h(X_i) = z]$ is Lipschitz in z for $r = 1, 2$ and $d = 0, 1$, and (S.16) holds, by repeating the arguments in the proof of Lemma B.3 with $\tau = \frac{1}{2}$, the desired convergence in distribution holds. In fact, instead of requiring $E[Y_i(d)|h(X_i) = z]$ and $E[Y_i^2(d)|h(X_i) = z]$ to both be Lipschitz continuous, it suffices to require $\text{Var}[Y_i(d)|h(X_i) = z]$ to be Lipschitz continuous.

Next, I show $\hat{\varsigma}^2_{\tilde{g}_m,n} \xrightarrow{P} \varsigma^2_g$ as $m, n \to \infty$. Similar arguments as those used in Theorem IV.2 go through if (S.16) and (S.17) hold. Since (S.17) follows from Assumptions IV.3 by Lemma B.6, the conclusion follows. ■

B Auxiliary Lemmas

B.1 Auxiliary Lemmas for Main Results

Lemma B.1. *Suppose $m \geq 2$, and $x_1, \ldots, x_{2m}$ are real number such that $x_1 \leq \cdots \leq x_{2m}$. Then, for any $\pi \in \Pi_n$,*

$$\sum_{k=1}^{m} x_{\pi(2k-1)} x_{\pi(2k)} \leq \sum_{k=1}^{m} x_{2k-1} x_{2k} \ .$$

Proof of Lemma B.1. I start by considering π which only permutes the indices $\{k_1, k_2, k_3, k_4\}$ and leaves the other entries intact. I need only consider the case where there exists $k_1 < k_2 < k_3 < k_4$ such that at least one of $\pi(k_1), \pi(k_2)$ is greater than at least one of $\pi(k_3), \pi(k_4)$ because the lemma trivially holds otherwise. Suppose without loss of generality that $\pi(k_2) < \pi(k_3) < \pi(k_4) < \pi(k_1)$, then it is easy to verify

$$x_{\pi(k_1)} x_{\pi(k_2)} + x_{\pi(k_3)} x_{\pi(k_4)} \leq x_{\pi(k_2)} x_{\pi(k_3)} + x_{\pi(k_1)} x_{\pi(k_4)}$$

so that by interchanging two indices I decrease the sum weakly. To conclude the proof, note a finite number of those interchanges maps π back to the identity operator, and the conclusion follows. ■

Lemma B.2. *Let X_n, Y_n, Z_n be random variables. Suppose $Y_n = g(Z_n) \xrightarrow{d} Y$ as $n \to \infty$, where $g : \mathbf{R} \to \mathbf{R}$ is measurable and $X_n \xrightarrow{d} X$ conditional on Z_n, with probability one for Z_n. Furthermore, suppose the distributions of both X and Y are continuous everywhere. Then*

$$(X_n, Y_n) \xrightarrow{d} (X, Y) \ ,$$

where $X \perp\!\!\!\perp Y$.

Proof of Lemma B.2. Since X and Y both have continuous distribution functions and they are independent, I need only show for any $x, y \in \mathbf{R}$,

$$P\{X_n \leq x, Y_n \leq y\} \to P\{X \leq x\} P\{Y \leq y\} \ .$$

To this end, note

$$\begin{aligned}
&P\{X_n \leq x, Y_n \leq y\} - P\{X \leq x\} P\{Y \leq y\} \\
&= E[E[I\{X_n \leq x\} I\{Y_n \leq y\} | Z_n]] - P\{X \leq x\} P\{Y \leq y\} \\
&= E[E[I\{X_n \leq x\} | Z_n] I\{Y_n \leq y\}] - P\{X \leq x\} P\{Y \leq y\} \\
&= E[(E[I\{X_n \leq x\} | Z_n] - P\{X \leq x\}) I\{Y_n \leq y\}] + E[P\{X \leq x\} (I\{Y_n \leq y\} - P\{Y \leq y\})] \\
&= E[(P\{X_n \leq x | Z_n\} - P\{X \leq x\}) I\{Y_n \leq y\}] + (P\{Y_n \leq y\} - P\{Y \leq y\}) P\{X \leq x\} \ .
\end{aligned}$$

Since

$$P\{X_n \le x|Z_n\} - P\{X \le x\} \to 0$$

with probability one for Z_n, and hence the product inside the expectation converges to 0 with probability one as well, which in turn implies the expectation converges to 0 by the dominated convergence theorem because probabilities are bounded. The second term converges to 0 because of the definition of convergence in distribution and the fact that the distribution function of Y is continuous everywhere. ■

Lemma B.3. *Suppose the sample size is kn for $k \in \mathbf{Z}$ and the treatment assignment scheme satisfies $\tau_s \equiv \tau = \frac{l}{k}$, where $l \in \mathbf{Z}$, $0 < l < k$, and they are mutually prime. Suppose Q satisfies Assumption IV.1 and h satisfies the assumptions in Theorem C.2. Then, under $\lambda^{\tau,h}(X^{(n)})$ defined in* (S.34)*, as $n \to \infty$,*

$$\sqrt{kn}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\tau,h}) \ ,$$

where

$$\varsigma^2_{\tau,h} = \frac{\mathrm{Var}[Y_i(1)]}{\tau} + \frac{\mathrm{Var}[Y_i(0)]}{1-\tau} - \tau(1-\tau)E\Big[\Big(E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] - \Big(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\Big)\Big)^2\Big] \ . \tag{S.4}$$

Proof of Lemma B.3. To begin with, note

$$\sqrt{kn}(\hat{\theta}_n - \theta(Q)) = A_n - B_n + C_n - D_n \ ,$$

where

$$\begin{aligned}
A_n &= \frac{1}{\sqrt{kn}} \sum_{1\le i\le kn} \Big(\frac{Y_i(1)D_i}{\tau} - E\Big[\frac{Y_i(1)D_i}{\tau}\Big|h^{(n)}, D^{(n)}\Big]\Big) \\
B_n &= \frac{1}{\sqrt{kn}} \sum_{1\le i\le kn} \Big(\frac{Y_i(0)(1-D_i)}{1-\tau} - E\Big[\frac{Y_i(0)(1-D_i)}{1-\tau}\Big|h^{(n)}, D^{(n)}\Big]\Big) \\
C_n &= \frac{1}{\sqrt{kn}} \sum_{1\le i\le kn} \Big(E\Big[\frac{Y_i(1)D_i}{\tau}\Big|h^{(n)}, D^{(n)}\Big] - E[Y_i(1)]\Big) \\
D_n &= \frac{1}{\sqrt{kn}} \sum_{1\le i\le kn} \Big(E\Big[\frac{Y_i(0)(1-D_i)}{1-\tau}\Big|h^{(n)}, D^{(n)}\Big] - E[Y_i(0)]\Big) \ .
\end{aligned}$$

Note conditional on $h^{(n)}$ and $D^{(n)}$, A_n and B_n are independent and C_n and D_n are constant.

I first study the limiting behavior of A_n. Conditional on $h^{(n)}$ and $D^{(n)}$, the terms in the sum are independent but not identically distributed. Therefore, I proceed to verify the Lindeberg condition holds in probability conditional on $h^{(n)}$ and $D^{(n)}$. To that end, define

$$s_n^2 = s_n^2(h^{(n)}, D^{(n)}) = \sum_{1\le i\le kn} \mathrm{Var}\Big[\frac{Y_i(1)D_i}{\tau}\Big|h^{(n)}, D^{(n)}\Big]$$

and note

$$\begin{aligned} s_n^2 &= \sum_{1\le i\le kn} \mathrm{Var}\Big[\frac{Y_i(1)D_i}{\tau}\Big|h^{(n)}, D^{(n)}\Big] \\ &= \frac{1}{\tau^2}\sum_{1\le i\le kn} D_i \,\mathrm{Var}[Y_i(1)|h^{(n)}] \\ &= \frac{1}{\tau^2}\sum_{1\le i\le kn: D_i=1} \mathrm{Var}[Y_i(1)|h(X_i)] \ , \end{aligned}$$

where the second equality follows from (1) and the third follows from the fact that units are i.i.d. It follows that

$$\begin{aligned} \tau\frac{s_n^2}{kn} = \frac{1}{kn}\sum_{1\le i\le kn} \mathrm{Var}[Y_i(1)|h(X_i)] + \Big(\frac{1-\tau}{\tau kn}\sum_{1\le i\le kn: D_i=1} \mathrm{Var}[Y_i(1)|h(X_i)] \\ - \frac{1}{kn}\sum_{1\le i\le kn: D_i=0} \mathrm{Var}[Y_i(1)|h(X_i)]\Big) \ . \end{aligned} \tag{S.5}$$

By Assumption IV.1,

$$\frac{1}{kn}\sum_{1\le i\le kn} \mathrm{Var}[Y_i(1)|h(X_i)] \xrightarrow{P} E[\mathrm{Var}[Y_i(1)|h(X_i)]] < E[Y_i(1)] < \infty \ . \tag{S.6}$$

Meanwhile,

$$\begin{aligned} &\Big|\frac{1-\tau}{\tau kn}\sum_{1\le i\le kn: D_i=1} \mathrm{Var}[Y_i(1)|h(X_i)] - \frac{1}{kn}\sum_{1\le i\le kn: D_i=0} \mathrm{Var}[Y_i(1)|h(X_i)]\Big| \\ &\lesssim \Big|\frac{1-\tau}{\tau kn}\sum_{1\le i\le kn: D_i=1} h_i - \frac{1}{kn}\sum_{1\le i\le kn: D_i=0} h_i\Big| \\ &= \frac{1}{\tau kn}\Big|\sum_{1\le s\le n}\sum_{(s-1)k+1\le j\le sk: D_{\pi^{\tau,h}(j)}=1} (h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau)\Big| \\ &\le \frac{1}{\tau kn}\sum_{1\le s\le n}\sum_{(s-1)k+1\le j\le sk: D_{\pi^{\tau,h}(j)}=1} |h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau| \\ &\lesssim \frac{1}{n}\sum_{1\le s\le n}\sum_{(s-1)k+1\le j\le sk} |h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau| \\ &\le k^{1/2}\Big(\frac{1}{n}\sum_{1\le s\le n}\sum_{(s-1)k+1\le j\le sk} |h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau|^2\Big)^{1/2} \xrightarrow{P} 0 \ , \end{aligned} \tag{S.7}$$

where the first inequality holds because $E[Y_i^r(d)|h(X_i)=z]$ is Lipschitz for $r=1,2$ and $d=0,1$, the second holds by assumption, the third holds by inspection, the second to last holds by the Cauchy-Schwarz inequality, the last holds by condition (c) in Theorem C.2, and the equality holds by inspection. Combining (S.5), (S.6), and (S.7), I have

$$\frac{s_n^2}{kn} \xrightarrow{P} \frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]]}{\tau} > 0 \ , \tag{S.8}$$

where the inequality holds by assumption.

I now argue the Lindeberg condition holds in probability conditional on $h^{(n)}$ and $D^{(n)}$, i.e., for any $\epsilon > 0$,

$$E_n = \frac{1}{s_n^2 \tau^2} \sum_{1 \leq i \leq kn} E[|Y_i(1)D_i - E[Y_i(1)D_i|h^{(n)}, D^{(n)}]|^2 \\ I\{|Y_i(1)D_i - E[Y_i(1)D_i|h^{(n)}, D^{(n)}]| > \epsilon \tau s_n\}|h^{(n)}, D^{(n)}] \xrightarrow{P} 0 \ . \quad \text{(S.9)}$$

To this end, first note for any $M > 0$,

$$P\{\epsilon \tau s_n > M\} \to 1 \quad \text{(S.10)}$$

because of (S.8). Next, note

$$E[Y_i(1)D_i|h^{(n)}, D^{(n)}] = E[Y_i(1)|h(X_i)]D_i$$

because of (1). As a result, for any $M > 0$

$$\begin{aligned} E_n &= \frac{1}{s_n^2 \tau^2} \sum_{1 \leq i \leq kn : D_i = 1} E[|Y_i(1) - E[Y_i(1)|h^{(n)}, D^{(n)}]|^2 I\{|Y_i(1) - E[Y_i(1|h^{(n)}, D^{(n)}]| > \epsilon \tau s_n\}|h^{(n)}, D^{(n)}] \\ &\leq \frac{1}{s_n^2 \tau^2} \sum_{1 \leq i \leq kn} E[|Y_i(1) - E[Y_i(1)|h^{(n)}, D^{(n)}]|^2 I\{|Y_i(1) - E[Y_i(1)|h^{(n)}, D^{(n)}]| > \epsilon \tau s_n\}|h^{(n)}, D^{(n)}] \\ &\leq \frac{1}{s_n^2 \tau^2} \sum_{1 \leq i \leq kn} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)] > M\}|h^{(n)}, D^{(n)}] + o_p(1) \\ &= \frac{kn}{s_n^2 \tau^2} \frac{1}{kn} \sum_{1 \leq i \leq kn} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}|h^{(n)}, D^{(n)}] + o_p(1) \quad \text{(S.11)} \\ &\xrightarrow{P} (E[\text{Var}[Y_i(1)|h(X_i)]])^{-1} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}] \ , \quad \text{(S.12)} \end{aligned}$$

where the first inequality holds by inspection, the second holds because of (S.10) and the equality follows because (1) and $Q_n = Q^{kn}$, and the convergence in probability follows from (S.8) and the fact that Assumption IV.1 implies

$$E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}] \\ \leq E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2] = E[\text{Var}[Y_i(1)|h(X_i)]] \leq E[Y_i^2(1)] < \infty \ .$$

In addition, by the dominated convergence theorem,

$$\lim_{M \to \infty} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}] = 0 \ .$$

To show $E_n \xrightarrow{P} 0$, fix any subsequence $\{n(j)\}$, and I argue there is a further subsequence $\{n(j(k(l)))\}$ along which $E_{n(j(k(l)))}$ converges to 0 almost surely. Indeed, for the subsequence $\{n(j)\}$, for any fixed M, $E_{n(j)}$ is bounded by (S.11), which I define as $U_{n(j)}(M)$. I know from above that $U_{n(j)}(M) \xrightarrow{P} U(M)$, where $U(M)$ is defined as (S.12). Hence, there exists a further subsequence $\{n(j(k))\}$ along which $U_{n(j(k))}(M) \to U(M)$ almost surely. I then choose a sequence $\{M_{n(j(k))}\}_{n\geq 1}$ such that $M_{n(j(k))} \to \infty$. By the dominated convergence theorem, $\lim_{n\to\infty} U(M_{n(j(k))}) = 0$. By a diagonalization argument, I could construct a further subsequence $\{n(j(k(l)))\}$ along which $U_{n(j(k(l)))}(M_{n(j(k(l)))}) \to 0$. Along this subsequence, because $E_n \leq U_n(M_n)$ for each n, the almost sure limit of E_n must be zero because it is non-negative.

I now argue

$$\sup_{t\in\mathbf{R}} |P\{A_n \leq t | h^{(n)}, D^{(n)}\} - \Phi(t/\sqrt{E[\mathrm{Var}[Y_i(1)|h(X_i)]]/\tau})| \xrightarrow{P} 0 \ .$$

Fix any subsequence. Since $E_n \xrightarrow{P} 0$, there exists a further subsequence along which $E_n \to 0$ with probability one for $h^{(n)}, D^{(n)}$. Because of the Lindeberg condition and (S.8), it follows that with probability one for $h^{(n)}, D^{(n)}$, $A_n \xrightarrow{d} N(0, E[\mathrm{Var}[Y_i(1)|h(X_i)]]/\tau)$ conditional on $h^{(n)}, D^{(n)}$. But then the left-hand side of the preceding display must converge almost surely to 0 by Pólya's theorem. Since for any subsequence there exists a further subsequence along which it converges to 0 almost surely, it must converge to 0 in probability.

A similar argument establishes

$$\sup_{t\in\mathbf{R}} |P\{B_n \leq t | h^{(n)}, D^{(n)}\} - \Phi(t/\sqrt{E[\mathrm{Var}[Y_i(0)|h(X_i)]]/(1-\tau)})| \xrightarrow{P} 0 \ .$$

Since A_n and B_n are independent conditional on $h^{(n)}$ and $D^{(n)}$, it follows by a similar subsequencing argument that

$$\sup_{t\in\mathbf{R}} |P\{A_n - B_n \leq t | h^{(n)}, D^{(n)}\} - \Phi(t/\sqrt{E[\mathrm{Var}[Y_i(1)|h(X_i)]/\tau + E[\mathrm{Var}[Y_i(0)|h(X_i)]]/(1-\tau)})| \xrightarrow{P} 0 \ . \tag{S.13}$$

To study C_n, note by (1),

$$C_n = \frac{1}{\sqrt{kn}} \sum_{1\leq i\leq kn} \Big(E\Big[\frac{Y_i(1)}{\tau}\Big| h(X_i)\Big] D_i - E[Y_i(1)] \Big) \ .$$

So I have

$$E[C_n | h^{(n)}] = \frac{1}{\sqrt{kn}} \sum_{1\leq i\leq kn} (E[Y_i(1)|h(X_i)] - E[Y_i(1)]) \ .$$

Furthermore, by the assumptions that $E[Y_i^r(d)|h(X_i) = z]$ is Lipschitz for $r = 1, 2$ and $d = 0, 1$ and

$\frac{1}{n}\sum_{1\le s\le n}\sum_{(s-1)k+1\le j\le sk}|h_{\pi^{\tau,h}(j)}-\bar{h}_s^\tau|^2\xrightarrow{P}0$, I have

$$\mathrm{Var}[C_n|h^{(n)}]\propto\frac{1}{kn}\sum_{1\le s\le n}(h_{\pi^{\tau,h}(i)}-\bar{h}_\tau^s)^2\xrightarrow{P}0\ ,$$

where the first relation can be established by repeating the arguments used in the last step of establishing Theorem C.1. It therefore follows by Chebyshev's inequality that for any $\epsilon>0$,

$$P\{|C_n-E[C_n|h^{(n)}]|>\epsilon|h^{(n)}\}\xrightarrow{P}0\ ,$$

and because probabilities are bounded and hence uniformly integrable,

$$P\{|C_n-E[C_n|h^{(n)}]|>\epsilon\}\xrightarrow{P}0\ ,$$

and hence

$$C_n=\frac{1}{\sqrt{kn}}\sum_{1\le i\le kn}(E[Y_i(1)|h(X_i)]-E[Y_i(1)])+o_p(1)\ .$$

A similar proof shows

$$D_n=\frac{1}{\sqrt{kn}}\sum_{1\le i\le kn}(E[Y_i(0)|h(X_i)]-E[Y_i(0)])+o_p(1)\ .$$

and therefore

$$\begin{aligned}C_n-D_n&=\frac{1}{\sqrt{kn}}\sum_{1\le i\le kn}(E[Y_i(1)|h(X_i)]-E[Y_i(1)]-(E[Y_i(0)|h(X_i)]-E[Y_i(0)]))+o_p(1)\\&\xrightarrow{d}N\Big(0,E\Big[(E[Y_i(1)|h(X_i)]-E[Y_i(1)]-(E[Y_i(0)|h(X_i)]-E[Y_i(0)]))^2\Big]\Big)\ .\end{aligned}$$

I now show by contradiction that

$$\sup_{t\in\mathbf{R}}|P\{\sqrt{n}(\hat{\theta}_n-\theta(Q))\le t\}-\Phi(t/\varsigma_h)|\to0\ .$$

Suppose not, then there must exist a subsequence along which the left-hand side of the above display converges to some $\delta>0$. Along this subsequence, I could find a further subsequence along which the left-hand side of (S.13) converges to 0 with probability one for $h^{(n)}$ and $D^{(n)}$, i.e.,

$$A_n-B_n\xrightarrow{d}N\Big(0,\frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]}{\tau}+\frac{E[\mathrm{Var}[Y_i(0)|h(X_i)]]}{1-\tau}\Big)$$

with probability one for $h^{(n)}$ and $D^{(n)}$. Since C_n-D_n is constant for each $h^{(n)}$ and $D^{(n)}$, Lemma B.2 establishes

$$A_n-B_n+C_n-D_n\xrightarrow{d}N\Big(0,\frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]}{\tau}+\frac{E[\mathrm{Var}[Y_i(0)|h(X_i)]]}{1-\tau}+$$

$$E\Big[(E[Y_i(1)|h(X_i)] - E[Y_i(1)] - (E[Y_i(0)|h(X_i)] - E[Y_i(0)]))^2\Big]\Big) \ ,$$

which, by Pólya's Theorem, implies a contradiction.

Finally, note

$$\begin{aligned}
&\frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]}{\tau} + \frac{E[\mathrm{Var}[Y_i(0)|h(X_i)]]}{1-\tau} \\
&\qquad + E\Big[(E[Y_i(1)|h(X_i)] - E[Y_i(1)] - (E[Y_i(0)|h(X_i)] - E[Y_i(0)]))^2\Big] \\
&= \frac{\mathrm{Var}[Y_i(1)]}{\tau} + \frac{\mathrm{Var}[Y_i(0)]}{1-\tau} - \frac{\mathrm{Var}[E[Y_i(1)|h(X_i)]]}{\tau} - \frac{\mathrm{Var}[E[Y_i(0)|h(X_i)]]}{1-\tau} + \\
&\qquad E\Big[(E[Y_i(1)|h(X_i)] - E[Y_i(1)] - (E[Y_i(0)|h(X_i)] - E[Y_i(0)]))^2\Big] \\
&= \frac{\mathrm{Var}[Y_i(1)]}{\tau} + \frac{\mathrm{Var}[Y_i(0)]}{1-\tau} - \frac{1-\tau}{\tau} E[(E[Y_i(1)|h(X_i)] - E[Y_i(1)])^2] \\
&\qquad - \frac{\tau}{1-\tau} E[(E[Y_i(0)|h(X_i)] - E[Y_i(0)])^2] \\
&\qquad - 2E\Big[(E[Y_i(1)|h(X_i)] - E[Y_i(1)])(E[Y_i(0)|h(X_i)] - E[Y_i(0)])\Big] \\
&= \frac{\mathrm{Var}[Y_i(1)]}{\tau} + \frac{\mathrm{Var}[Y_i(0)]}{1-\tau} - \tau(1-\tau) E\Big[\Big(E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] - \Big(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\Big)\Big)^2\Big] \ ,
\end{aligned}$$

and the result follows. ■

Lemma B.4. *Suppose U_i, $1 \le i \le n$ are i.i.d. random variables where $E|U_i|^r < \infty$. Then*

$$n^{-1/r} \max_{1 \le i \le n} |U_i| \xrightarrow{P} 0 \ .$$

Proof of Lemma B.4. Note for all $\epsilon > 0$,

$$\begin{aligned}
P\Big\{n^{-1/r} \max_{1 \le i \le n} |U_i| > \epsilon\Big\} &= P\Big\{ \max_{1 \le i \le n} |U_i|^r > n\epsilon^r\Big\} \\
&\le nP\{|U_i|^r > n\epsilon^r\} \le \frac{n}{n\epsilon^r} E[|U_i|^r I\{|U_i|^r > n\epsilon^r\}] = \frac{1}{\epsilon^r} E[|U_i|^r I\{|U_i|^r > n\epsilon^r\}] \to 0 \ ,
\end{aligned}$$

where the convergence follows because of the dominated convergence theorem and that $E|U_i|^r < \infty$. ■

Lemma B.5. *Suppose $E[h^2(X_i)] < \infty$. Then, as $n \to \infty$,*

$$\frac{1}{n} \sum_{1 \le s \le n} |h_{\pi^h(2s-1)} - h_{\pi^h(2s)}|^2 \xrightarrow{P} 0 \ , \tag{S.14}$$

and

$$\frac{1}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} |h_{\pi^h(4j-k)} - h_{\pi^h(4j-l)}|^2 \xrightarrow{P} 0 \text{ for } k \in \{2,3\} \text{ and } l \in \{0,1\} \ . \tag{S.15}$$

Proof of Lemma B.5. Note

$$\sum_{1\le s\le n} |h_{\pi^h(2s-1)} - h_{\pi^h(2s)}|^2 \le |h_{\pi^h(2n)} - h_{\pi^h(1)}|^2 \le 4 \max_{1\le i\le 2n} h^2(X_i) \ ,$$

where the first inequality follows from the definition of π^h and the second inequality follows by inspection, and therefore it follows from Lemma B.4 that

$$\frac{1}{n}\sum_{1\le s\le n} |h_{\pi^h(2s-1)} - h_{\pi^h(2s)}|^2 \le \frac{4}{n} \max_{1\le i\le 2n} h^2(X_i) \xrightarrow{P} 0 \ .$$

(S.14) thus holds. To see Assumption S.15 holds, note

$$\frac{1}{n}\sum_{1\le j\le \lfloor \frac{n}{2} \rfloor} |h_{\pi^h(4j-k)} - h_{\pi^h(4j-l)}|^2 \lesssim \frac{1}{n}|h_{\pi^h(2n)} - h_{\pi^h(1)}|^2 \ ,$$

and the result follows similarly. ■

Lemma B.6. *Suppose $E[h^2(X_i)] < \infty$ and $\tilde{h}_m$ satisfies Assumption IV.3. Then, as $m, n \to \infty$,*

$$\frac{1}{n}\sum_{1\le s\le n} |h_{\pi^{\tilde{h}_m}(2s-1)} - h_{\pi^{\tilde{h}_m}(2s)}|^2 \xrightarrow{P} 0 \ , \tag{S.16}$$

and

$$\frac{1}{n}\sum_{1\le j\le \lfloor \frac{n}{2} \rfloor} |h_{\pi^{\tilde{h}_m}(4j-k)} - h_{\pi^{\tilde{h}_m}(4j-l)}|^2 \xrightarrow{P} 0 \text{ for } k \in \{2,3\} \text{ and } l \in \{0,1\} \ . \tag{S.17}$$

Proof of Lemma B.6. I only prove the first conclusion as the second can be shown by similar arguments. I first show Assumption IV.3 implies

$$\frac{1}{n}\sum_{1\le i\le 2n} |\tilde{h}_i - h_i|^2 \xrightarrow{P} 0 \ . \tag{S.18}$$

Suppose Assumption IV.3 holds. For any $\epsilon > 0$, $\delta > 0$, there exists $M > 0$ such that for $m > M$,

$$P\Big\{ \int_{\mathrm{supp}(X_i)} |\tilde{h}_m(x) - h(x)|^2 \, Q_X(dx) > \frac{\epsilon\delta}{2} \Big\} \le \frac{\delta}{2} \ . \tag{S.19}$$

By Chebyshev's inequality again, if

$$\int_{\mathrm{supp}(X_i)} |\tilde{h}_m(x) - h(x)|^2 \, Q_X(dx) \le \frac{\epsilon\delta}{2} \ ,$$

then by the independence of $\tilde{W}^{(m)}$ and $W^{(n)}$,

$$\begin{aligned} P\Big\{ \frac{1}{2n}\sum_{1\le i\le 2n} |\tilde{h}_i - h_i|^2 > \epsilon \Big| \tilde{W}^{(m)} \Big\} &\le \frac{1}{\epsilon} E\Big[\frac{1}{2n}\sum_{1\le i\le 2n} |\tilde{g}_i - g_i|^2 \Big| \tilde{W}^{(m)} \Big] \\ &= \frac{1}{\epsilon}\int_{\mathrm{supp}(X_i)} |\tilde{h}_m(x) - h(x)|^2 \, Q_X(dx) \le \frac{\delta}{2} \ . \end{aligned} \tag{S.20}$$

Then,

$$P\Big\{\frac{1}{2n}\sum_{1\leq i\leq 2n}|\tilde{h}_i - h_i|^2 > \epsilon\Big\} \leq P\Big\{\frac{1}{2n}\sum_{1\leq i\leq 2n}|\tilde{h}_i - h_i|^2 > \epsilon\Big|\tilde{W}^{(m)}\Big\}$$
$$\times P\Big\{\int_{\text{supp}(X_i)}|\tilde{h}_m(x) - h(x)|^2\, Q_X(dx) \leq \frac{\epsilon\delta}{2}\Big\}$$
$$+ P\Big\{\int_{\text{supp}(X_i)}|\tilde{h}_m(x) - h(x)|^2\, Q_X(dx) > \frac{\epsilon\delta}{2}\Big\}$$
$$\leq \frac{\delta}{2}\Big(1 - \frac{\delta}{2}\Big) + \frac{\delta}{2} \leq \delta\ ,$$

where the first inequality follows by definition, and the second inequality follows from (S.19) and (S.20).

Next, note because $|a+b|^2 \leq 2(a^2+b^2)$ for any $a, b \in \mathbf{R}$,

$$\frac{1}{n}\sum_{1\leq s\leq n}|h_{\pi^{\tilde{h}_m}(2s-1)} - h_{\pi^{\tilde{h}_m}(2s)}|^2 \lesssim \frac{1}{n}\sum_{1\leq s\leq n}|\tilde{h}_{\pi^{\tilde{h}_m}(2s-1)} - \tilde{h}_{\pi^{\tilde{h}_m}(2s)}|^2 + \frac{1}{n}\sum_{1\leq i\leq 2n}|\tilde{h}_i - h_i|^2\ . \quad \text{(S.21)}$$

Next, note

$$\frac{1}{n}\sum_{1\leq s\leq n}|\tilde{h}_{\pi^{\tilde{h}_m}(2s-1)} - \tilde{h}_{\pi^{\tilde{h}_m}(2s)}|^2 \lesssim \frac{1}{n}\max_{1\leq i\leq 2n}|\tilde{h}_i|^2$$
$$\lesssim \frac{1}{n}\max_{1\leq i\leq 2n}|h_i|^2 + \frac{1}{n}\max_{1\leq i\leq 2n}|\tilde{h}_i - h_i|^2 \lesssim \frac{1}{n}\max_{1\leq i\leq 2n}|h_i|^2 + \frac{1}{n}\sum_{1\leq i\leq 2n}|\tilde{h}_i - h_i|^2\ . \quad \text{(S.22)}$$

The conclusion then follows from (S.18), (S.21), (S.22), the assumption that $E[h^2(X_i)] < \infty$ and an application of Lemma B.4. ■

B.2 Sufficient Conditions for Lipschitz Continuity

Let f denote the density function of X. Recall $C^{(r)}$ is the class of functions which are rth continuously differentiable. I impose the following assumption on h and f.

Assumption B.1. *The function h and density function f satisfy the following conditions.*

(a) $h \in C^{(2)}$.

(b) $\frac{\partial h(x)}{\partial x_p} \neq 0$ *Lebesgue a.e.*

(c) $f \in C^{(2)}$.

Lemma B.7 (Theorem 24.4 of Munkres (1991))**.** *Let O be open in $\mathbf{R}^p$ and $f : O \to \mathbf{R}$ be of class $C^{(r)}$ for $r \geq 1$. Let M be the set of points x for which $f(x) = 0$ and N be the set of points x for which $f(x) \geq 0$. Suppose M is non-empty and $Df(x)$ has rank 1 at each point of M. Then N is a p-manifold in $\mathbf{R}^p$ and $\partial N = M$.*

Lemma B.8. *Suppose Assumption B.1(a)–(b) hold. Then $M = \{x : h(x) = z\}$ is a $(p-1)$-manifold in $\mathbf{R}^p$.*

PROOF OF LEMMA B.8. For each $x \in M$, I aim at providing a coordinate patch on M about x. Indeed, by Assumption B.1(a)–(b) and Theorem 9.2 (implicit function theorem) of Munkres (1991), there exists an open set U containing $u = (x_1, \ldots, x_{p-1})$, an open ball $B(z)$ containing z and an open set O in $\mathbf{R}$ containing x_p, and a function $k : U \times B(z) \to \mathbf{R}^p$ of class $C^{(2)}$ such that $h(u, k(u, z')) = z'$ for all $u \in U$, $z' \in B(z)$ and $x \in O$. Moreover, $k(U \times B(z)) = O$. Define the coordinate patch $\alpha(u; z) = (u, k(u, z))$. The conclusion follows by Theorem 5-2 of Spivak (1965). ■

Note $M = \{x : h(x) = z\}$ is a $(p-1)$-manifold by Lemmas B.7 and B.8. In what follows, I will need the definition of the integral of a function g over the manifold M. In order to do so, note there exists a coordinate patch as $\{\alpha_j : U_j \subseteq \mathbf{R}^{p-1} \to V_j \subseteq M, j \in \mathcal{J}\}$, where $\alpha_j(u) = \alpha_j(u, z)$, and each $\alpha_j(u) = (u, k_j(u))$ for some function $k_j : U \to \mathbf{R}$ which is of class C^2, as shown in the proof of Lemma B.8, and $\alpha_j(U_j) = V_j$. Next, there exists a partition of unity $\{\phi_i : i \in \mathcal{I}\}$ dominated by the $\{V_j : j \in \mathcal{J}\}$. Moreover, both $\mathcal{I}$ and $\mathcal{J}$ can be chosen to be countable, according to Section 25 of Munkres (1991). The integral of a scalar function g over the manifold is written as

$$\int_M g \, \mathrm{d}V = \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{I}} \int_{U_j} [(g\phi_i) \circ \alpha_j] V(D\alpha_j) \ ,$$

where $V(A) = \sqrt{\det(A'A)}$ is the volume. I have

$$D\alpha_j = \begin{bmatrix} I_{p-1} & \dfrac{\partial k_j(u,z)}{\partial u} \end{bmatrix} \ ,$$

so that

$$V(D\alpha_j) = \sqrt{1 + \frac{\partial k_j(u,z)}{\partial u'} \frac{\partial k_j(u,z)}{\partial u}} = \frac{\|\nabla h(u, k_j(u,z))\|}{|D_p h(u, k_j(u,z))|} \ ,$$

where $D_p = \frac{\partial}{\partial x_p}$, by the implicit function theorem and matrix determinant lemma. Note on one hand, for each $j \in \mathcal{J}$, only a finite number of ϕ_i is positive, and on the other hand, $\{\phi_i : i \in \mathcal{I}\}$ is dominated by the coordinate patch, which means each ϕ_i is supported on a compact set inside a single V_j. As a result, the order of the above double sum can be interchanged.

By p.345 of Bogachev (2007), the conditional expectation of a function g on the manifold M is defined as

$$E[g(X)|M] = \lim_{t \to 0} \frac{E[g(X) I\{z \le h(X) \le z + t\}]}{P\{z \le h(X) \le z + t\}} \ .$$

Lemma B.9. *Suppose Assumption B.1(a)–(c) hold. Then*

$$E[g(X)|M] = \frac{\displaystyle\int_M \frac{fg}{\|\nabla h\|} \, \mathrm{d}V}{\displaystyle\int_M \frac{f}{\|\nabla h\|} \, \mathrm{d}V} \ . \tag{S.23}$$

For a continuously differentiable function $h : \mathbf{R}^p \to \mathbf{R}$, $x \in \mathbf{R}^p$ is a critical point of h if $\nabla h(x) = 0$, where $\nabla h(x)$ is the gradient of h at x; otherwise x is a regular point of h. A value z is a critical value of h if the set $\{x : h(x) = z\}$ contains at least one critical point; otherwise z is a regular value of h.

Proof of Lemma B.9. By L'Hospital's rule,

$$E[g(X)|M] = \frac{\lim_{t\to 0} \dfrac{E[g(X)I\{z \le h(X) \le z+t\}]}{t}}{\lim_{t\to 0} \dfrac{P\{z \le h(X) \le z+t\}}{t}} ,$$

and the lemma follows from Lemma A.1 of Chernozhukov, Fernández-Val and Luo (2018). In particular, the denominator equals the one in (S.23) directly by the same lemma, while for the numerator I merely need to redefine the 'density' function as fg and the same proof goes through. ■

Lemma B.10. *Suppose Assumption B.1(a)–(b) hold. Let $M = \{x : h(x) = z\}$, where z is a regular value of h on $\mathbf{R}^p$. Then for any $g \in C^{(2)}$,*

$$\frac{\partial}{\partial z}\int_M g \,\mathrm{d}V = \int_M \frac{D_p g}{D_p h}\,\mathrm{d}V + \int_M g \frac{1}{\|\nabla h\|^2}\sum_{1\le i\le p}\frac{D_i h D_{ip} h}{D_p h}\,\mathrm{d}V - \int_M g\frac{D_{pp}h}{D_p^2 h}\,\mathrm{d}V . \tag{S.24}$$

Proof of Lemma B.10. To begin with, note

$$\begin{aligned}
&\frac{\partial}{\partial z}\int_{U_j}[(g\phi_i)\circ\alpha_j]V(D\alpha_j) \\
&\quad = \int_{U_j} D_p(g\phi_i)\frac{\partial k_j(u,z)}{\partial z}\frac{\|\nabla h\|}{|D_p h|} \\
&\qquad + \int_{U_j} g\phi_i \frac{|D_p h|}{\|\nabla h\|}\frac{\partial k_j(u,z)}{\partial z}\frac{1}{D_p^4 h}\Big(D_p^2 h\sum_{1\le i\le p} D_i h D_{ip} h - D_p h D_{pp} h\sum_{1\le i\le p} D_i^2 h\Big) ,
\end{aligned} \tag{S.25}$$

where $D_{ij}h = \partial_i\partial_j h$ for any function $h \in C^{(2)}$. I have suppressed the arguments of h, being $(u, k_j(u,z))$. Note it is legitimate to pass differentiation inside the integral by the dominated convergence theorem. By the Implicit Function Theorem again,

$$\frac{\partial k_j(u,z)}{\partial z} = \frac{1}{D_p h(u, k_j(u,z))} . \tag{S.26}$$

By Theorem 7.17 of Rudin (1976), I know $\frac{\partial}{\partial z}\int_M g(x)\,\mathrm{d}V$ is the sum over $i \in \mathcal{I}, j \in \mathcal{J}$ of the two terms in (S.25). Using (S.26), the sum of the first term is

$$\begin{aligned}
&\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}}\int_{U_j}(\phi_i D_p g + g D_p\phi_i)\frac{1}{D_p h}\frac{\|\nabla h\|}{|D_p h|} \\
&= \sum_j \int_{U_j}\frac{D_p g}{D_p h}V(D\alpha_j)
\end{aligned}$$

$$= \int_M \frac{D_p g}{D_p h} \, \mathrm{d}V \ , \tag{S.27}$$

because $\sum_{i \in \mathcal{I}} \phi_i = 1$ and hence $\sum_{i \in \mathcal{I}} D_p \phi_i = D_p \sum_{i \in \mathcal{I}} \phi_i = 0$. Again, the interchange of differentiation and sum is allowed because the sum is actually over a finite number of terms, by definition of a partition of unity. The sum of the second term is

$$\begin{aligned}
&\sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{I}} \int_{U_j} g \phi_i \frac{|D_p h|}{\|\nabla h\|} \frac{1}{D_p^4 h} \sum_{1 \le i \le p} (D_i h D_p h D_{ip} h - D_i^2 h D_{pp} h) \\
&= \sum_{j \in \mathcal{J}} \int_{U_j} g \frac{D_p^2 h}{\|\nabla h\|^2} \frac{1}{D_p^4 h} \sum_{1 \le i < p} (D_i h D_p h D_{ip} h - D_i^2 h D_{pp} h) V(D\alpha) \\
&= \int_M g \frac{1}{\|\nabla h\|^2 D_p^2 h} \sum_{1 \le i \le p} (D_i h D_p h D_{ip} h - D_i^2 h D_{pp} h) \, \mathrm{d}V \\
&= \int_M g \frac{1}{\|\nabla h\|^2} \sum_{1 \le i \le p} \frac{D_i h D_{ip} h}{D_p h} \, \mathrm{d}V - \int_M g \frac{D_{pp} h}{D_p^2 h} \, \mathrm{d}V \ .
\end{aligned} \tag{S.28}$$

(S.24) now follows from (S.27) and (S.28). ■

Theorem B.1. *Suppose Assumption B.1 holds. If z is a regular value of h, then*

$$\frac{\partial}{\partial z} E[g(X)|M] = \frac{\int_M \frac{D_p(fg/D_p h)}{\|\nabla h\|} \, \mathrm{d}V \int_M \frac{f}{\|\nabla h\|} \, \mathrm{d}V - \int_M \frac{D_p(f/D_p h)}{\|\nabla h\|} \, \mathrm{d}V \int_M \frac{fg}{\|\nabla h\|} \, \mathrm{d}V}{\Big[\int_M \frac{f}{\|\nabla h\|} \, \mathrm{d}V\Big]^2} \ . \tag{S.29}$$

Proof of Theorem B.1. To begin with, replace g in Lemma B.10 with $\frac{f}{\|\nabla h\|}$. I then have

$$\begin{aligned}
&\frac{\partial}{\partial z} \int_M \frac{f}{\|\nabla h\|} \, \mathrm{d}V \\
&= \int_M \frac{\|\nabla h\| D_p f - \frac{f \sum_{1 \le i \le p} D_i h D_{ip} h}{\|\nabla h\|}}{\|\nabla h\|^2 D_p h} \, \mathrm{d}V \\
&\qquad + \int_M \frac{f}{\|\nabla h\|^3} \sum_{1 \le i \le p} \frac{D_i h D_{ip} h}{D_p h} \, \mathrm{d}V - \int_M \frac{f D_{pp} h}{\|\nabla h\| D_p^2 h} \, \mathrm{d}V \\
&= \int_M \frac{D_p f D_p h - f D_{pp} h}{\|\nabla h\| D_p^2 h} \, \mathrm{d}V \\
&= \int_M \frac{D_p(f/D_p h)}{\|\nabla h\|} \, \mathrm{d}V \ .
\end{aligned} \tag{S.30}$$

By the same arguments,

$$\frac{\partial}{\partial z} \int_M \frac{fg}{\|\nabla h\|} \, \mathrm{d}V = \int_M \frac{D_p(fg/D_p h)}{\|\nabla h\|} \, \mathrm{d}V \ . \tag{S.31}$$

(S.29) now follows from (S.30) and (S.31) together with the quotient rule. ■

In general, by the Law of Iterated Expectation

$$E[Y_i^r(d)|h(X)=z] = E[E[Y_i^r(d)|X]|h(X)=z] \ .$$

Suppose h and the density function of X, $f(X)$ satisfy the smoothness conditions in Assumption B.1, the derivative

$$\frac{\partial}{\partial z} E[g(X)|h(X)=z]$$

is given in Theorem B.1, where $g(x) = E[Y_i^r(d)|X=x]$ for $r=1,2$ and $d=0,1$. In particular, it is equal to

$$\begin{aligned} &E\Big[\frac{D_p g}{D_p h} + \frac{g D_p f}{f D_p h} - \frac{g D_{pp} h}{D_p^2 h}\Big| h(X)=z\Big] - E\Big[\frac{D_p f}{f D_p h} - \frac{D_{pp} h}{D_p^2 h}\Big| h(X)=z\Big] E\Big[g\Big| h(X)=z\Big] \\ &= E\Big[\frac{D_p g}{D_p h}\Big| h(X)=z\Big] + \mathrm{Cov}\Big[\frac{D_p f}{f D_p h} - \frac{D_{pp} h}{D_p^2 h}, g\Big| h(X)=z\Big] \ . \end{aligned} \tag{S.32}$$

Lemma B.11. *Each of the following conditions imply the boundedness of* (S.32).

1. *h is linear, $\|D_p g\|_\infty < \infty$, $\|g\|_\infty < \infty$ and $\|D_p(\ln f)\|_\infty < \infty$.*
2. *h is linear, $\sup_{z\in\mathbf{R}} |E[D_p g|h(X)=z]| < \infty$, $\sup_{z\in\mathbf{R}} |E[g^2|h(X)=z]| < \infty$ and $\sup_{z\in\mathbf{R}} |E[D_p^2(\ln f)|h(X)=z]| < \infty$.*
3. *h includes linear and interaction terms, $\left\|\frac{D_p g}{D_p h}\right\|_\infty < \infty$, $\|g\|_\infty < \infty$ and $\left\|\frac{D_p(\ln f)}{D_p h}\right\|_\infty < \infty$.*

Proof of Lemma B.11. Follows from inspection. ■

C Supplementary Theoretical Results

C.1 Optimal Stratification for General Treated Fractions

The next theorem shows the infeasible optimal stratification has a similar structure to (6) when $\tau \neq \frac{1}{2}$.

Theorem C.1. *Assume $\tau = \frac{l}{k}$ where $l, k \in \mathbf{N}$, $0 < l < k$, and that the sample size is kn. Let π^{τ,g^τ} be a permutation of $\{1,\ldots,kn\}$ such that $g^\tau_{\pi^{\tau,g^\tau}(1)} \leq \cdots \leq g^\tau_{\pi^{\tau,g^\tau}(kn)}$ for g^τ defined in* (7). *Then,* (3) *is solved by*

$$\lambda^{\tau,g}(X^{(n)}) = \{\{\pi^{\tau,g^\tau}((s-1)k+1), \ldots \pi^{\tau,g^\tau}(sk)\} : 1 \leq s \leq n\} \ . \tag{S.33}$$

Proof of Theorem C.1. First, note

$$\hat{\theta}_n = \frac{1}{kn} \sum_{1\leq i\leq kn} \Big(\frac{1}{\tau} Y_i(1) D_i - \frac{1}{1-\tau} Y_i(0)(1-D_i)\Big) \ .$$

Next,

$$\mathrm{MSE}(\lambda|X^{(n)}) = \mathrm{Var}_\lambda[\hat{\theta}_n|X^{(n)}] \ ,$$

so that I need only consider conditional variances of $\hat{\theta}$ given $X^{(n)}$ which can be decomposed as in (5). By repeating the arguments in the proof of Lemma II.1, for any $\lambda \in \Lambda_n$, the first term of the right-hand side of (5) equals

$$\frac{1}{k^2 n^2} \sum_{1 \le i \le kn} \Big(\frac{\mathrm{Var}[Y_i(1)|X_i]}{\tau} + \frac{\mathrm{Var}[Y_i(0)|X_i]}{1-\tau} \Big) \ ,$$

again identical across all $\lambda \in \Lambda_n$. Therefore, I need only consider

$$\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \ .$$

By repeating the arguments in the proof of Lemma II.2, a stratum of size kl where $l > 1$ is a convex combination of stratifications with strata only of size k. In particular, let Λ_n^k denote the set of all stratifications for which each stratum is of size k. Then, I have $\Lambda_n \in \mathrm{co}(\Lambda_n^k)$. I could therefore focus on the case where each stratum is of size k. For any stratification of the form $\lambda = \{\{\pi((s-1)k+1, \dots \pi(sk)\} : 1 \le s \le n\}$,

$$\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \propto \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk} (g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2 \ ,$$

where g_i^τ is defined in (7) and

$$\bar{g}^\tau_s = \frac{1}{k} \sum_{(s-1)k+1 \le j \le sk} g^\tau_{\pi(j)} \ .$$

To see this, first note units are independent across strata, so that by repeating the arguments in the proof of Lemma II.1,

$$\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \propto \sum_{1 \le s \le n} \mathrm{Var}_\lambda \Big[\sum_{(s-1)k+1 \le j \le sk} g^\tau_{\pi(j)} D_{\pi(j)} \Big] \ .$$

Next,

$$\begin{aligned}
&\mathrm{Var}_\lambda \Big[\sum_{(s-1)k+1 \le j \le sk} g^\tau_{\pi(j)} D_{\pi(j)} \Big] \\
&= \frac{1}{\binom{k}{l}} \sum_{(s-1)k+1 \le j_1 < \cdots < j_l \le sk} \Big(\sum_{1 \le \iota \le l} g^\tau_{\pi(j_\iota)} - l\bar{g}^\tau_s \Big)^2 \\
&= \frac{l}{k} \sum_{(s-1)k+1 \le j \le sk} (g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2 + \frac{1}{\binom{k}{l}} \sum_{(s-1)k+1 \le j_1 < \cdots < j_l \le sk} \sum_{1 \le \iota_1 \ne \iota_2 \le l} (g_{\pi(j_{\iota_1})} - \bar{g}^\tau_s)(g_{\pi(j_{\iota_2})} - \bar{g}^\tau_s) \\
&= \frac{l}{k} \sum_{(s-1)k+1 \le j \le sk} (g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2 + \frac{\binom{k-2}{l-2}}{\binom{k}{l}} \Big[\Big(\sum_{(s-1)k+1 \le j \le sk} g^\tau_{\pi(j)} - k\bar{g}^\tau_s \Big)^2 - \sum_{(s-1)k+1 \le j \le sk} (g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2 \Big] \\
&\propto \sum_{(s-1)k+1 \le j \le sk} (g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2 \ ,
\end{aligned}$$

where the first equality holds by definition, the second holds by expanding the square, the third holds by accounting for cross product terms, and the fourth holds because the first term inside the square bracket on the fourth line is 0. The conclusion follows from similar arguments to those used in the proof of Lemma B.1. ■

Remark C.1. Researchers are sometimes faced with the the situation where both l and k in Theorem C.1 are large and they are mutually prime. For example, suppose there are 52 participants and 31 seats for treatment. In that case, because the treated fraction is close to 3/5, our recommendation is to split the sample into 6 strata of size 10, 10, 10, 10, 10, 2, treat 6 of the 10 units in each of the first five strata, and 1 of the 2 units in the last stratum. ■

The next theorem is the limiting counterpart to Theorem C.1. It shows the asymptotic variance of $\hat{\theta}_n$ is minimized by choosing $h = g^\tau$ defined in (7).

Theorem C.2. *Suppose $\tau \in (0,1)$. Let $h : \text{supp}(X_i) \to \mathbf{R}$ be a measurable function, and $\pi^{\tau,h}$ be a permutation of $\{1,\ldots,kn\}$ such that $h_{\pi^{\tau,h}(1)} \leq \cdots \leq h_{\pi^{\tau,h}(kn)}$. Define*

$$\lambda^{\tau,h}(X^{(n)}) = \{\{\pi^{\tau,h}((s-1)k+1), \ldots \pi^{\tau,h}(sk)\} : 1 \leq s \leq n\} \ . \tag{S.34}$$

Further define $\bar{h}^\tau_s = \frac{1}{k}\sum_{(s-1)k+1\leq j\leq sk} h_{\pi^{\tau,h}(j)}$. Suppose h satisfies

(a) $0 < E[\text{Var}[Y_i(d)|h(X_i)]]$ *for* $d \in \{0,1\}$.

(b) $E[Y_i^r(d)|h(X_i) = z]$ *is Lipschitz for* $r = 1, 2$ *and* $d = 0, 1$.

(c) $\dfrac{1}{n}\displaystyle\sum_{1\leq s\leq n}\ \sum_{(s-1)k+1\leq j\leq sk} |h_{\pi^{\tau,h}(j)} - \bar{h}^\tau_s|^2 \xrightarrow{P} 0$.

Then,

$$\varsigma^2_{\tau,g^\tau} \leq \varsigma^2_{\tau,h} \ ,$$

for $\varsigma^2_{\tau,g^\tau}$ and $\varsigma^2_{\tau,h}$ defined in (S.4) *and g^τ defined in* (7). *Moreover, the inequality is strict unless $E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] = g^\tau(X_i)$ with probability one under Q.*

Proof of Theorem C.2. By the definition of $\varsigma^2_{\tau,h}$ in (S.4), minimizing $\varsigma^2_{\tau,h}$ with respect to h is equivalent to maximizing

$$E\Big[\Big(E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] - \Big(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\Big)\Big)^2\Big] \ .$$

Next, note

$$\begin{aligned}
&E\Big[\Big(g^\tau(X_i) - E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big]\Big)\Big(E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] - \Big(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\Big)\Big)\Big] \\
&= E\Big[E\Big[g^\tau(X_i) - E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big]\Big|h(X_i)\Big] \\
&\qquad \Big(E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] - \Big(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\Big)\Big)\Big]
\end{aligned}$$

$$= 0 \ , \tag{S.35}$$

where the second equality holds because

$$E[g^\tau(X_i)|h(X_i)] = E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big]$$

by the law of iterated expectation. Therefore,

$$\begin{aligned}
&E\Big[\Big(g^\tau(X_i) - \Big(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\Big)\Big)^2\Big] \\
&= E\Big[\Big(g^\tau(X_i) - E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] + E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] - \Big(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\Big)\Big)^2\Big] \\
&= E\Big[\Big(g^\tau(X_i) - E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big]\Big)^2\Big] \\
&\qquad + E\Big[\Big(E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] - \Big(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\Big)\Big)^2\Big] \ , \\
&\geq E\Big[\Big(E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] - \Big(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\Big)\Big)^2\Big] \ .
\end{aligned}$$

where the second equality follows from (S.35) and the last inequality is strict except unless $E\Big[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\Big|h(X_i)\Big] = g^\tau(X_i)$ with probability one under Q. ■

C.2 Unequal Treated Fractions Across Subpopulations

In this section, I consider settings in which treated fractions are allowed to vary across subpopulations. Let $1 \leq r \leq R$ index the subpopulations, where $R \geq 1$ is an integer. I assume the subpopulations are determined by the covariates according to a function $f : \text{supp}(X_i) \to \{1, \ldots, R\}$. I replace Assumption I.1 by

Assumption C.1. *For $1 \leq r \leq R$, exactly τ_r fraction of the units of each stratum in the rth subpopulation are treated.*

Moreover, I assume treatment status is assigned independently across subpopulations. Under Assumption (C.1), $\hat{\theta}_n$ is generally inconsistent for θ. In such settings researchers often use the estimator from the fully saturated regression in Bugni, Canay and Shaikh (2019). For $1 \leq r \leq R$, let n_r denote the total number of observations in the rth subpopulation. For $1 \leq r \leq R$ and $d \in \{0, 1\}$, define

$$\hat{\mu}_{n,r}(1) = \frac{1}{n_r \tau_r} \sum_{i:f(X_i)=r} Y_i D_i$$

and

$$\hat{\mu}_{n,r}(0) = \frac{1}{n_r(1-\tau_r)} \sum_{i:f(X_i)=r} Y_i(1 - D_i) \ .$$

The estimator for the ATE from the fully saturated regression is

$$\hat{\theta}_n^{\text{sat}} = \sum_{1\le r\le R} \frac{n_r}{n}(\hat{\mu}_{n,r}(1) - \hat{\mu}_{n,r}(0)) \ . \tag{S.36}$$

Note $\hat{\theta}_n^{\text{sat}}$ and $\hat{\theta}_n$ coincide whenever $\tau_r \equiv \tau \in (0,1)$. See Bugni, Canay and Shaikh (2018), Tabord-Meehan (2022), and Bugni, Canay and Shaikh (2019) for more details. By repeating the arguments used in the proof of Theorem II.1 and Theorem C.1, I could find the stratification that minimizes the conditional MSE of $\hat{\theta}_n^{\text{sat}}$. The solution is as follows: I first calculate the stratification defined in (S.33) with τ, g, and $X^{(n)}$ defined separately for each subpopulation, and then take the union of those stratifications. Moreover, the next theorem enables us to derive feasible procedures when treated fractions are allowed to vary across subpopulations. In particular, it reveals any plug-in estimator that satisfies the regularity conditions in Theorem C.2 leads to a stratification under which the asymptotic variance of $\hat{\theta}_n^{\text{sat}}$ is no greater than and typically strictly less than that under procedures with each subpopulation as a stratum.

Theorem C.3. *Suppose the sample size is n. Define $N_r = \{i : f(X_i) = r\}$, $X^{N_r} = (X_i : i \in N_r)$, $n_r = |N_r|$, and $p(r) = Q\{f(X_i) = r\}$. Define $\lambda^{\text{large}} = \bigcup_{1\le r\le R} N_r$. For $1 \le r \le R$, let τ_r be the treated fraction in N_r. Define functions $h^r : \text{supp}(X_i) \to \mathbf{R}$ for $1 \le r \le R$. Define $\lambda^{\text{small}} = \bigcup_{1\le r\le R} \lambda^{\tau_r,h^r}(X^{N_r})$, where $\lambda^{\tau_r,h^r}(X^{N_r})$ is defined in* (S.34)*. Suppose Q satisfies Assumption IV.1 and the treatment assignment scheme satisfies Assumption C.1. Then, under λ^{large}, for $\hat{\theta}_n^{\text{sat}}$ defined in* (S.36)*, as $n \to \infty$,*

$$\sqrt{n}(\hat{\theta}_n^{\text{sat}} - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\text{large}}) \ ,$$

where

$$\varsigma^2_{\text{large}} = \sum_{1\le r\le R} p(r)\Big(\frac{\text{Var}[Y_i(1)|f(X_i)=r]}{\tau_r} + \frac{\text{Var}[Y_i(0)|f(X_i)=r]}{1-\tau_r}\Big) + \sum_{1\le r\le R} p(r)(E[Y_i(1)-Y_i(0)|f(X_i)=r] - E[Y_i(1)-Y_i(0)])^2 \ .$$

Suppose in addition that $h^r, 1 \le r \le R$ satisfy the assumption in Theorem C.2, under Q restricted to $\{x \in \text{supp}(X_i) : f(x) = r\}$. Then, under λ^{small}, for $\hat{\theta}_n^{\text{sat}}$ defined in (S.36)*, as $n \to \infty$,*

$$\sqrt{n}(\hat{\theta}_n^{\text{sat}} - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\text{small}}) \ ,$$

where

$$\varsigma^2_{\text{small}} = \sum_{1\le r\le R} p(r)\Big(\frac{\text{Var}[Y_i(1)|f(X_i)=r]}{\tau_r} + \frac{\text{Var}[Y_i(0)|f(X_i)=r]}{1-\tau_r} - \tau_r(1-\tau_r)E\Big[\Big(E\Big[\frac{Y_i(1)}{\tau_r} + \frac{Y_i(0)}{1-\tau_r}\Big|h^r(X_i)\Big] - E\Big[\frac{Y_i(1)}{\tau_r} + \frac{Y_i(0)}{1-\tau_r}\Big|f(X_i)=r\Big]\Big)^2\Big|f(X_i)=r\Big]\Big)$$

$$+ \sum_{1 \leq r \leq R} p(r)(E[Y_i(1) - Y_i(0)|f(X_i) = r] - E[Y_i(1) - Y_i(0)])^2 \ .$$

In addition, $\varsigma^2_{\text{small}} \leq \varsigma^2_{\text{large}}$*, where the inequality is strict unless for all* $1 \leq r \leq R$*,*

$$E\Big[\frac{Y_i(1)}{\tau_r} + \frac{Y_i(0)}{1-\tau_r}\Big|h^r(X_i)\Big] = E\Big[\frac{Y_i(1)}{\tau_r} + \frac{Y_i(0)}{1-\tau_r}\Big|f(X_i) = r\Big]$$

with probability one under

$$Q^r(A) = \frac{Q(A \cap \{f(X_i) = r\})}{Q\{f(X_i) = r\}} \ .$$

Moreover, among all choices of $(h^r : 1 \leq r \leq R)$*,* $\varsigma^2_{\text{small}}$ *is minimized by setting* $h^r = g^{\tau_r}$*, where* g^{τ_r} *is defined in* (7)*.*

Proof of Theorem C.3. The first convergence holds by Theorem 3.1 of Bugni, Canay and Shaikh (2019). Define $\theta_r = E[Y_i(1) - Y_i(0)|f(X_i) = r]$. For the second convergence, note

$$\sqrt{n}(\hat{\theta}^{\text{sat}}_n - \theta(Q)) = \sum_{1 \leq r \leq R} \Big(\Big(\frac{n_r}{n}\Big)^{1/2} \sqrt{n_r}(\hat{\mu}_{n,r}(1) - \hat{\mu}_{n,r}(0) - \theta_r) + \sqrt{n}\Big(\frac{n_r}{n} - p(r)\Big)\theta_r\Big) \ . \tag{S.37}$$

Define

$$L^1_n = (\sqrt{n_r}(\hat{\mu}_{n,r}(1) - \hat{\mu}_{n,r}(0) - \theta_r) : 1 \leq r \leq R)$$
$$L^2_n = \Big(\sqrt{n}\Big(\frac{n_r}{n} - p(r)\Big) : 1 \leq r \leq R\Big) \ .$$

It follows from the coupling argument in Lemma C.1 of Bugni, Canay and Shaikh (2019) that

$$(L^1_n, L^2_n) = (L^{*1}_n, L^2_n) + o_P(1)$$

where $L^{*1}_n \perp\!\!\!\perp L^2_n$ and $L^{*1}_n \xrightarrow{d} N(0, \text{diag}(\varsigma^2_{r,\text{small}} : 1 \leq r \leq R))$ with

$$\varsigma^2_{r,\text{small}} = \frac{E[(Y_i(1) - E[Y_i(1)|h^r(X_i)])^2|f(X_i) = r]}{\tau_r} + \frac{E[(Y_i(0) - E[Y_i(0)|h^r(X_i)])^2|f(X_i) = r]}{1-\tau_r}$$
$$+ E[(E[Y_i(1) - Y_i(0)|h^r(X_i)] - \theta_r)^2|f(X_i) = r] \ .$$

Meanwhile, the central limit theorem implies

$$L^2_n \xrightarrow{d} N(0, \text{diag}(p(r) : 1 \leq r \leq R) - (p(r) : 1 \leq r \leq R)(p(r) : 1 \leq r \leq R)') \ .$$

In addition, it follows from the weak law of large numbers that $\frac{n_r}{n} \xrightarrow{P} p(r)$ for $1 \leq r \leq R$. By (S.37) and Slutsky's lemma, I have that $\sqrt{n}(\hat{\theta}^{\text{sat}}_n - \theta(Q))$ converges to a normal distribution with zero mean and variance

$$\sum_{1 \leq r \leq R} p(r)\varsigma^2_{r,\text{small}} + \sum_{1 \leq r \leq R} p(r)(\theta_r - \theta(Q))^2 \ ,$$

where I use the fact that

$$\sum_{1 \leq r \leq R} p(r)\theta_r = \theta(Q) \ .$$

The first result then follows from a similar calculation to that at the end of the proof of Lemma B.3. The last two results can be shown by similar arguments to those used in the proof of Theorem C.2. ■

Remark C.2. Tabord-Meehan (2022) considers stratification trees, which leads to a small number of large strata, with different treated fractions in each stratum. Using results from Theorem C.3, it is straightforward to combine his procedure with procedures in this paper. The asymptotic variance of $\hat{\theta}_n^{\text{sat}}$ under the combined procedure is no greater than and typically strictly less than that under his procedure alone. The combined procedure is as follows: First, implement the procedure in Tabord-Meehan (2022), which produces a finite number of strata with a target treated fraction for each stratum. Second, I view each stratum as a subpopulation and calculate the stratification in (S.34) either with a fixed function h or some plug-in estimate, with τ equal the target treated fraction. Finally, I take the union of these stratifications. The desired properties mentioned above now follow from Theorem C.3. ■

C.3 Formal Justification of Results with Attrition

Let A_i be a binary variable such that $A_i = 1$ if and only if the unit does not attrite. The difference-in-means estimator for the non-attritors is

$$\hat{\theta}_n^{\text{A}} = \frac{\frac{1}{n}\sum_{i:D_i=1} A_i Y_i(1)}{\frac{1}{n}\sum_{i:D_i=1} A_i} - \frac{\frac{1}{n}\sum_{i:D_i=0} A_i Y_i(0)}{\frac{1}{n}\sum_{i:D_i=0} A_i} \ .$$

By repeating the arguments in the proof of Theorem S.1.5 of Bai, Romano and Shaikh (2021), under the assumption that $((Y^{(n)}(0), Y^{(n)}(1), A^{(n)}) \perp\!\!\!\perp D^{(n)} | X^{(n)}$, I have

$$\begin{aligned}
\frac{1}{n}\sum_{i:D_i=1} A_i Y_i(1) &\xrightarrow{P} E[A_i Y_i(1)] \\
\frac{1}{n}\sum_{i:D_i=1} A_i &\xrightarrow{P} E[A_i] \\
\frac{1}{n}\sum_{i:D_i=0} A_i Y_i(0) &\xrightarrow{P} E[A_i Y_i(0)] \\
\frac{1}{n}\sum_{i:D_i=0} A_i &\xrightarrow{P} E[A_i] \ .
\end{aligned}$$

As a result,

$$\hat{\theta}_n^{A} \xrightarrow{P} \frac{E[A_i(Y_i(1) - Y_i(0))]}{E[A_i]} \ .$$

If $A_i \perp\!\!\!\perp (Y_i(1) - Y_i(0))$, then the right hand side is $\theta(Q)$. ■

C.4 Nonnegativity of the Variance Estimator

In this subsection, I show $\hat{\varsigma}^2_{h,n}$ in (14) is nonnegative. For convenience of notation, I suppress h in the subscripts. I have

$$\begin{aligned}
\hat{\varsigma}^2_n &= \hat{\sigma}^2_n(1) + \hat{\sigma}^2_n(0) - \frac{1}{2}\hat{\rho}_n + \frac{1}{2}(\hat{\mu}_n(1) + \hat{\mu}_n(0))^2 \\
&= \frac{1}{n}\sum_{1\le i\le 2n: D_i=1} Y_i^2 - \hat{\mu}^2_n(1) + \frac{1}{n}\sum_{1\le i\le 2n: D_i=0} Y_i^2 - \hat{\mu}^2_n(0) - \frac{1}{2}\hat{\rho}_n + \frac{1}{2}(\hat{\mu}_n(1) + \hat{\mu}_n(0))^2 \\
&= \frac{1}{n}\sum_{1\le i\le 2n} Y_i^2 - \frac{1}{2}\hat{\rho}_n - \frac{1}{2}(\hat{\mu}_n(1) - \hat{\mu}_n(0))^2 \\
&= \frac{1}{2n}\sum_{1\le i\le 2n} Y_i^2 - \frac{1}{n}\sum_{1\le s\le n} Y_{\pi(2s-1)}Y_{\pi(2s)} + \frac{1}{2n}\sum_{1\le s\le n}(Y_{\pi(2s-1)} + Y_{\pi(2s)})^2 \\
&\quad - \frac{1}{n}\sum_{1\le j\le n/2}(Y_{\pi(4j-3)} + Y_{\pi(4j-2)})(Y_{\pi(4j-1)} + Y_{\pi(4j)}) \\
&\quad - \frac{1}{2}\Big(\frac{1}{n}\sum_{1\le s\le n}(D_{\pi(2s-1)} - D_{\pi(2s)})(Y_{\pi(2s-1)} - Y_{\pi(2s)})\Big)^2 \\
&= \frac{1}{2}\Big(\frac{1}{n}\sum_{1\le s\le n}((D_{\pi(2s-1)} - D_{\pi(2s)})(Y_{\pi(2s-1)} - Y_{\pi(2s)}))^2 \\
&\quad - \Big(\frac{1}{n}\sum_{1\le s\le n}(D_{\pi(2s-1)} - D_{\pi(2s)})(Y_{\pi(2s-1)} - Y_{\pi(2s)})\Big)^2\Big) \\
&\quad + \frac{1}{2n}\sum_{1\le j\le n/2}(Y_{\pi(4j-3)} + Y_{\pi(4j-2)} - (Y_{\pi(4j-1)} + Y_{\pi(4j)}))^2 \ .
\end{aligned}$$

In the last expression, the second term is obviously nonnegative, while the first term is nonnegative because it is one half the sample variance of $\{(D_{\pi(2s-1)} - D_{\pi(2s)})(Y_{\pi(2s-1)} - Y_{\pi(2s)}) : 1 \le s \le n\}$. ■

C.5 Details of the Penalized Procedure

In this section, I discuss the details of the penalized procedure. For $d \in \{0,1\}$, let $\tilde{\beta}_m(d)$ denote the least-square estimators of the linear regression coefficients among the treated or untreated units in the pilot experiment:

$$\tilde{\beta}_m(d) = \Big(\sum_{1\le j\le m:\tilde{D}_j=d} \tilde{X}_j\tilde{X}_j'\Big)^{-1} \sum_{1\le j\le m:\tilde{D}_j=d} \tilde{X}_j\tilde{Y}_j \ ,$$

and let $\tilde{\Omega}_m(d)$ denote the variance estimators assuming homoskedasticity:

$$\tilde{\Omega}_m(d) = \tilde{\nu}^2_m(d)\Big(\sum_{1\le j\le m:\tilde{D}_j=d} \tilde{X}_j\tilde{X}_j'\Big)^{-1} \ ,$$

where

$$\tilde{\nu}^2_m(d) = \frac{\sum_{1\le j\le m}(\tilde{Y}_j - \tilde{X}_j'\tilde{\beta}_m(d))^2 I\{\tilde{D}_j = d\}}{\sum_{1\le j\le m} I\{\tilde{D}_j = d\}} \ .$$

For d^{pen} defined in (10), let π^{pen} denote the solution to

$$\min_{\pi \in \Pi} \sum_{1 \leq s \leq n} d^{\text{pen}}(X_{\pi(2s-1)}, X_{\pi(2s)}) \ . \tag{S.38}$$

Units are then paired to solve (S.38), so the stratification is given by

$$\lambda^{\text{pen}}(X^{(n)}) = \{\{\pi^{\text{pen}}(2s-1), \pi^{\text{pen}}(2s)\} : 1 \leq s \leq n\} \ . \tag{S.39}$$

I start with a further justification for (S.39) by discussing its optimality in a Bayesian framework, in the sense that it minimizes the integrated risk in a Bayesian framework with a diffuse normal prior, where the conditional expectations of potential outcomes are linear. With some abuse of notation, denote the conditional MSE in (3) by $\text{MSE}(\lambda|g, X^{(n)})$, where I make explicit the dependence on g. Suppose I have a prior distribution of g, denoted by $F(dg)$. Let $Q_X^n(dx^{(n)})$ denote the distribution of $X^{(n)}$ and $Q_{\tilde{W}}^m(d\tilde{w}^{(m)})$ denote the distribution of $\tilde{W}^{(m)}$. Consider the solution to following problem of minimizing the integrated risk across all measurable functions of the form $u : (\tilde{w}^{(m)}, x^{(n)}) \mapsto \lambda \in \Lambda_n$:

$$\min_u \iiint \text{MSE}(u(\tilde{w}^{(m)}, x^{(n)})|g, x^{(n)}) Q_X^n(dx^{(n)}) Q_{\tilde{W}}^m(d\tilde{w}^{(m)}) F(dg) \ . \tag{S.40}$$

I focus on the special case under which and $Y_i(d) \sim N(X_i'\beta(d), \sigma^2)$ for $d \in \{0, 1\}$. Note the potential outcomes are homoskedastic conditional on the covariates. Define $\beta = \beta(1) + \beta(0)$, and I have $g(x) = x'\beta$. As before, I suppose $\tilde{W}^{(m)} = ((\tilde{Y}_j, \tilde{X}_j', \tilde{D}_j)' : 1 \leq j \leq m)$ is available from a pilot experiment. Suppose the prior on $\beta(d)$ is $G_d \overset{d}{=} N(\eta(d), \Omega(d))$ for $d \in \{0, 1\}$, being independent across $d \in \{0, 1\}$. The prior distribution of β is then $G(d\beta) \overset{d}{=} N(\eta(1) + \eta(0), \Omega(1) + \Omega(0))$. I could show the posterior distribution of $\beta(d)$ conditional on $\tilde{W}^{(m)}$ is

$$\bar{G}_d(d\beta|\tilde{W}^{(m)}) \overset{d}{=} N(\bar{\eta}, \bar{\Omega}) \ ,$$

where for $d \in \{0, 1\}$,

$$\begin{aligned}
\bar{\eta}(d) &= \Big((\sigma^2)^{-1} \sum_{j:\tilde{D}_j = d} \tilde{X}_j \tilde{X}_j' + \Omega^{-1}(d)\Big)^{-1} \Big((\sigma^2)^{-1} \sum_{j:\tilde{D}_j = d} \tilde{X}_j \tilde{Y}_j + \Omega^{-1}(d)\eta(d)\Big) \\
\bar{\Omega}(d) &= \Big((\sigma^2)^{-1} \sum_{j:\tilde{D}_j = d} \tilde{X}_j \tilde{X}_j' + \Omega^{-1}(d)\Big)^{-1} \ .
\end{aligned}$$

Define $\bar{\eta} = \bar{\eta}(1) + \bar{\eta}(0)$ and $\bar{\Omega} = \bar{\Omega}(1) + \bar{\Omega}(0)$. The posterior distribution for β is

$$\bar{G}(d\beta|\tilde{W}^{(m)}) \overset{d}{=} (\bar{\eta}, \bar{\Omega}) \ ,$$

because $G_d(d\beta)$'s are independent across $d \in \{0, 1\}$.

The next lemma provides the solution to the Bayesian problem in (S.40), where the choice set is over all measurable functions $u : (\tilde{w}^{(m)}, x^{(n)}) \mapsto \lambda \in \Lambda_n$.

Lemma C.1. *The solution to* (S.40) *maps each* $(\tilde{w}^{(m)}, x^{(n)})$ *to* $\lambda = \{\{\pi(2s-1), \pi(2s)\} : 1 \leq s \leq n/2\}$*, where* π *solves*

$$\min_{\pi \in \Pi_n} \sum_{1 \leq s \leq n} \bar{d}\Big(x_{\pi(2s-1)}, x_{\pi(2s)}\Big) \ ,$$

where

$$\bar{d}(x_1, x_2) = (x_1'\bar{\eta} - x_2'\bar{\eta})^2 + (x_1 - x_2)'\bar{\Omega}(x_1 - x_2) \ . \tag{S.41}$$

Proof. First note by similar calculations to those leading to Lemma II.1, (S.40) is equivalent to

$$\min_u \iiint L(u(\tilde{w}^{(m)}, x^{(n)})|\beta, x^{(n)}) Q_X^n(dx^{(n)}) Q_{\tilde{W}}^m(d\tilde{w}^{(m)}) G(d\beta) \ , \tag{S.42}$$

where

$$L(u|\beta, x^{(n)}) = \beta'(x^{(n)})' \operatorname{Var}_\lambda[D^{(n)}|X^{(n)}] x^{(n)} \beta \ .$$

Next, note I could solve the problem pointwise for $\tilde{w}^{(m)}$ and $x^{(n)}$ because (S.42) is equivalent to

$$\min_u \bar{R}(u|\tilde{W}^{(m)}) \ , \tag{S.43}$$

where

$$\bar{R}(u|\tilde{W}^{(m)}) = \int L(u(\tilde{W}^{(m)}, x^{(n)})|\beta, x^{(n)}) \bar{G}(d\beta|\tilde{W}^{(m)}) \ .$$

To solve (S.43), first note because $\bar{R}(u|\tilde{W}^{(m)})$ is linear in u, by Lemma II.2, it is solved by a matched-pair design. Next,

$$\bar{R}(u|\tilde{W}^{(m)}) = \sum_{1 \leq s \leq n} ((x'_{\pi(2s-1)}\bar{\eta} - x'_{\pi(2s)}\bar{\eta})^2 + (x_{\pi(2s-1)} - x_{\pi(2s)})'\bar{\Omega}(x_{\pi(2s-1)} - x_{\pi(2s)})) \ .$$

As a result, minimizing it is equivalent to minimizing the sum of the distances defined in (S.41). ■

Lemma C.1 shows the solution to (S.40) is not to naïvely pair units according to the values of $X_i'\bar{\eta}$, where $\bar{\eta}$ is posterior mean of β. Instead, the solution to (S.40) depends not only on the posterior mean of β, but also on the posterior variance of it. The posterior variance serves as a penalty to matching naïvely on the posterior mean of β: the larger the variance, the more it penalizes matching on the posterior mean.

Finally, I make the prior irrelevant. For this purpose, suppose $\Omega = cI$ where I is an identity matrix. I let the constant $c \to \infty$, so that the prior diverges to a diffuse (uninformative) one. Then, $\bar{\eta}(d)$ converges to $\tilde{\beta}_m(d)$ and $\bar{\Omega}(d)$ converges to $\tilde{\Omega}_m(d)$. The metric then (S.41) converges to the metric defined in (10).

In practice, (S.38) can be solved as follows. Define R_m as the result of the following Cholesky decomposition:

$$R_m'R_m = \tilde{\beta}_m\tilde{\beta}_m' + \tilde{\Omega}_m \ ,$$

and define

$$Z_i = R_m X_i \ .$$

To see (S.38) is equivalent to

$$\min_{\pi \in \Pi_n} \frac{1}{n} \sum_{1 \le s \le n} \|Z_{\pi(2s-1)} - Z_{\pi(2s)}\|^2 \ , \tag{S.44}$$

note

$$d^{\text{pen}}(x_1, x_2) = (x_1 - x_2)' \tilde{\beta}_m \tilde{\beta}_m' (x_1 - x_2) + (x_1 - x_2)' \tilde{\Omega}_m (x_1 - x_2) = (x_1 - x_2)' R_m' R_m (x_1 - x_2) \ .$$

The penalized stratification pairs units to minimize the sum of distances in terms of Z_i within pairs. When $\dim(X_i)$ is not too large, the problem can be solved quickly by the package `nbpMatching` in `R`.

Because the penalized matched-pair design can be viewed as pairing to minimize the Euclidean distances of Z as in (S.44), inference can be implemented by "pairing the pairs" as in Section 4 of Bai, Romano and Shaikh (2021). I refer the readers to that paper for details.

The next theorem establishes the behavior of the difference-in-means estimator under the penalized procedure as the sample sizes of the pilot and the main experiment both increase. Not surprisingly, as the sample size of the pilot experiment goes to infinity, the penalized procedure behaves similarly to pairing units according to h, where h is a linear function. In particular, if the selection on observable assumption holds in the pilot data, then the conclusion of the next theorem holds with $\beta = \beta(1) + \beta(0)$, where $\beta(d) = E[XX']^{-1} E[XY(d)]$ for $d \in \{0, 1\}$.

Theorem C.4. *Suppose Q satisfies Assumption IV.1, $h(x) = x'\beta$ satisfies Assumption IV.2 for some $\beta \in \mathbf{R}^{\dim(X_i)}$. Suppose $E[XX'] < \infty$. Further suppose $\tilde{\beta}_m \xrightarrow{P} \beta$ and $\tilde{\Omega}_m \xrightarrow{P} 0$ as $m \to \infty$. Then, under λ^{pen} defined in* (S.39)*, as $m, n \to \infty$, $\sqrt{n}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma_h^2)$ for ς_h^2 in* (11)*. Furthermore, $(\hat{\varsigma}_n^{\text{pen}})^2 \xrightarrow{P} \varsigma_h^2$.*

Proof. I only prove the convergence in distribution because the convergence of the standard error follows from similar arguments to those used in the proof of Theorem 4.3 in Bai, Romano and Shaikh (2021). Define $\tilde{h}_m(x) = x'\tilde{\beta}_m$. Let $\|\tilde{\Omega}_m\|_{\text{op}}$ denote the operator norm of $\tilde{\Omega}_m$. Note

$$\begin{aligned}
&\frac{1}{n} \sum_{1 \le s \le n} ((X'_{\pi^{\text{pen}}(2s-1)} \tilde{\beta}_m - X'_{\pi^{\text{pen}}(2s)} \tilde{\beta}_m)^2 + (X_{\pi^{\text{pen}}(2s-1)} - X_{\pi^{\text{pen}}(2s)})' \tilde{\Omega}_m (X_{\pi^{\text{pen}}(2s-1)} - X_{\pi^{\text{pen}}(2s)})) \\
&\le \frac{1}{n} \sum_{1 \le s \le n} ((X'_{\pi^{\tilde{h}_m}(2s-1)} \tilde{\beta}_m - X'_{\pi^{\tilde{h}_m}(2s)} \tilde{\beta}_m)^2 + (X_{\pi^{\tilde{h}_m}(2s-1)} - X_{\pi^{\tilde{h}_m}(2s)})' \tilde{\Omega}_m (X_{\pi^{\tilde{h}_m}(2s-1)} - X_{\pi^{\tilde{h}_m}(2s)})) \\
&= \frac{1}{n} \sum_{1 \le s \le n} ((X'_{\pi^{\tilde{h}_m}(2s-1)} \tilde{\beta}_m - X'_{\pi^{\tilde{h}_m}(2s)} \tilde{\beta}_m)^2 + \|\tilde{\Omega}_m\|_{\text{op}} |X_{\pi^{\tilde{h}_m}(2s-1)} - X_{\pi^{\tilde{h}_m}(2s)}|^2) \\
&\le \frac{1}{n} \sum_{1 \le s \le n} (X'_{\pi^{\tilde{h}_m}(2s-1)} \tilde{\beta}_m - X'_{\pi^{\tilde{h}_m}(2s)} \tilde{\beta}_m)^2 + \|\tilde{\Omega}_m\|_{\text{op}} \frac{2}{n} \sum_{1 \le i \le 2n} |X_i|^2
\end{aligned}$$

$$= \frac{1}{n} \sum_{1 \leq s \leq n} (X'_{\pi^{\tilde{h}_m}(2s-1)} \tilde{\beta}_m - X'_{\pi^{\tilde{h}_m}(2s)} \tilde{\beta}_m)^2 + o_P(1) \ ,$$

where the first inequality follows because π^{pen} solves (S.38), the first equality follows from the definition of the operator norm, the second inequality follows from the fact that $|a+b|^2 \leq 2(|a|^2+|a|^2)$ for $a, b \in \mathbf{R}^{\dim(X_i)}$, and the last equality follows from the assumptions that $E[XX'] < \infty$ and $\tilde{\Omega}_m \xrightarrow{P} 0$ and the weak law of large numbers. Because I also know

$$\frac{1}{n} \sum_{1 \leq s \leq n} (X'_{\pi^{\text{pen}}(2s-1)} \tilde{\beta}_m - X'_{\pi^{\text{pen}}(2s-1)} \tilde{\beta}_m)^2 \geq \frac{1}{n} \sum_{1 \leq s \leq n} (X'_{\pi^{\tilde{h}_m}(2s-1)} \tilde{\beta}_m - X'_{\pi^{\tilde{h}_m}(2s-1)} \tilde{\beta}_m)^2 \ ,$$

I have

$$\frac{1}{n} \sum_{1 \leq s \leq n} (X'_{\pi^{\text{pen}}(2s-1)} \tilde{\beta}_m - X'_{\pi^{\text{pen}}(2s-1)} \tilde{\beta}_m)^2 = \frac{1}{n} \sum_{1 \leq s \leq n} (X'_{\pi^{\tilde{h}_m}(2s-1)} \tilde{\beta}_m - X'_{\pi^{\tilde{h}_m}(2s-1)} \tilde{\beta}_m)^2 + o_P(1) \ .$$

The rest of the proof follows from similar arguments to those used in the proof of Lemma B.6. ■

C.6 Inference with Pooled Data

So far I have disregarded data from the pilot experiment for inference except when computing $\tilde{g}_m$. I end this section by describing a test that combines data from the pilot and the main experiments. Define

$$\tilde{\theta}_m = \tilde{\mu}_m(1) - \tilde{\mu}_m(0) \ ,$$

where

$$\tilde{\mu}_m(d) = \frac{\sum_{1 \leq j \leq m} \tilde{Y}_j I\{\tilde{D}_j = d\}}{\sum_{1 \leq j \leq m} I\{\tilde{D}_j = d\}}$$

for $d \in \{0, 1\}$. I define the new estimator for $\theta(Q)$ as

$$\hat{\theta}_n^{\text{combined}} = \frac{m}{m+2n} \tilde{\theta}_m + \frac{2n}{m+2n} \hat{\theta}_n \ .$$

Let $\tilde{\varsigma}^2_{\text{pilot},m}$ denote the variance estimator of $\tilde{\theta}_m$ in the pilot experiment. I define the test as

$$\phi_n^{\text{combined}}(W^{(n)}, \tilde{W}^{(m)}) = I\{|T_n^{\text{combined}}(W^{(n)}, \tilde{W}^{(m)})| > \Phi^{-1}(1 - \frac{\alpha}{2})\} \ , \tag{S.45}$$

where

$$T_n^{\text{combined}}(W^{(n)}, \tilde{W}^{(m)}) = \frac{\sqrt{m+2n}(\hat{\theta}_n^{\text{combined}} - \theta_0)}{\sqrt{\frac{m}{m+2n} \tilde{\varsigma}^2_{\text{pilot},m} + \frac{2n}{m+2n} 2\hat{\varsigma}^2_{\tilde{h}_m,n}}} \ ,$$

and $\Phi^{-1}(1 - \frac{\alpha}{2})$ denotes the $(1 - \frac{\alpha}{2})$-th quantile of the standard normal distribution.

The following theorem shows the test defined in (S.45) is asymptotically exact as the sample sizes of both the pilot and the main experiments increase. The main additional requirement is as $m \to \infty$, $\sqrt{m}(\tilde{\theta}_m - \theta(Q))$ converges in distribution to a normal distribution whose variance is consistently

estimable. The assumption is satisfied by many treatment assignment schemes, including i.i.d. coin flips and covariate-adaptive randomization. See Bugni, Canay and Shaikh (2018) and Bugni, Canay and Shaikh (2019) for details.

Theorem C.5. *Suppose the treatment assignment scheme satisfies Assumption I.1, Q satisfies Assumptions IV.1, h satisfies Assumption IV.2, and $\tilde{h}_m$ satisfies Assumption IV.3. Suppose in addition that as $m \to \infty$, $\sqrt{m}(\tilde{\theta}_m - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\text{pilot}})$, $\tilde{\varsigma}^2_{\text{pilot},m} \xrightarrow{P} \varsigma^2_{\text{pilot}}$, and that as $m, n \to \infty$,*

$$\frac{m}{m+2n} \to \nu \in [0,1] \ .$$

Then, under $\lambda^{\tilde{g}_m}(X^{(n)})$ for $h = \tilde{g}_m$, as $m, n \to \infty$,

$$\frac{\sqrt{m+2n}(\hat{\theta}^{\text{combined}}_n - \theta(Q))}{\sqrt{\frac{m}{m+2n}\tilde{\varsigma}^2_{\text{pilot},m} + \frac{2n}{m+2n}2\hat{\varsigma}^2_{\tilde{h}_m,n}}} \xrightarrow{d} N(0,1) \ .$$

Thus, for the problem of testing (12) *at level $\alpha \in (0,1)$, $\phi^{\text{combined}}_n(W^{(n)}, \tilde{W}^{(m)})$ in* (S.45) *satisfies*

$$\lim_{m,n\to\infty} E[\phi^{\text{combined}}_n(W^{(n)}, \tilde{W}^{(m)})] = \alpha \ ,$$

whenever Q additionally satisfies the null hypothesis, i.e. $\theta(Q) = \theta_0$.

Proof. To begin with, note I need only establish as $m, n \to \infty$,

$$\sqrt{m+2n}(\hat{\theta}^{\text{combined}}_n - \theta(Q)) \xrightarrow{d} N(0, \nu\varsigma^2_{\text{pilot}} + (1-\nu)2\varsigma^2_h) \ , \tag{S.46}$$

and the rest follows from Slutsky's lemma. I prove (S.46) by contradiction. Suppose (S.46) does not hold. Then, there exists a subsequence still denoted by $\{m, n\}$ for notational simplicity, along which as $m, n \to \infty$,

$$\sup_{t\in\mathbf{R}} \left| \sqrt{m+2n}(\hat{\theta}^{\text{combined}}_n - \theta(Q)) - \Phi(z/\sqrt{\nu\varsigma^2_{\text{pilot}} + (1-\nu)2\varsigma^2_h}) \right| \to c \ , \tag{S.47}$$

where $c > 0$, and

$$\frac{m}{m+2n} \to \nu \in [0,1] \ .$$

Now consider this subsequence. Since the two convergences in the Lemma B.6 hold in probability, there exists a further subsequence along which they hold with probability one. By Theorem IV.1, I could see along this subsequence, as $m, n \to \infty$, with probability one for $\tilde{W}^{(m)}$,

$$\sup_{t\in\mathbf{R}} \left| Q\{\sqrt{n}(\hat{\theta}_n - \theta(Q)) \leq t | \tilde{W}^{(m)}\} - \Phi(z/\varsigma_h) \right| \to 0 \ . \tag{S.48}$$

Along the subsequence I construct, because $\frac{m}{m+2n} \to \nu$, by (S.48), Slutsky's lemma, and Lemma B.2,

$$\sqrt{m+2n}(\hat{\theta}^{\text{combined}}_n - \theta(Q)) \xrightarrow{d} N(0, \nu\varsigma^2_{\text{pilot}} + (1-\nu)2\varsigma^2_h) \ ,$$

which is a contradiction to (S.47). The theorem therefore holds. ■

C.7 Adjusted Standard Error With Four Units

Still suppose the sample size is $2n$, in order to be consistent with the notation in the main text. Suppose units are matched into sets of four units according to a function $h \in \mathbf{H}$. Let π^h be such that $h_{\pi^h(1)} \leq \cdots \leq h_{\pi^h(2n)}$. The stratification is given by

$$\{\{\pi^h(4s-3), \pi^h(4s-2), \pi^h(4s-1), \pi^h(4s)\} : 1 \leq s \leq n/2\} \ . \tag{S.49}$$

By Lemma B.3, I still have $\sqrt{n}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma_h^2)$, for ς_h^2 in (11). The variance estimator is

$$(\hat{\varsigma}_{h,n}^{\text{four}})^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\hat{\rho}_n^{\text{four}} + \frac{1}{2}(\hat{\mu}_n(1) + \hat{\mu}_n(0))^2 \ , \tag{S.50}$$

where

$$\hat{\rho}_n^{\text{four}} = \frac{2}{n} \sum_{1 \leq s \leq n/2} \frac{1}{2} \sum_{i,j,k,l \in \lambda_s, i<j, k<l : D_i = D_j = 1, D_k = D_l = 0} (Y_i + Y_k)(Y_j + Y_l) \ . \tag{S.51}$$

To establish

$$\hat{\rho}_n^{\text{four}} \xrightarrow{P} E[E[g(X_i)|h(X_i)]^2] \ ,$$

note

$$\begin{aligned}
&E\Big[\frac{1}{2} \sum_{i,j,k,l \in \lambda_s, i<j, k<l : D_i = D_j = 1, D_k = D_l = 0} (Y_i + Y_k)(Y_j + Y_l) | h^{(n)}\Big] \\
&= \frac{1}{12} \sum_{i,j,k,l \in \{0,1,2,3\}, i<j,k<l} (\mu_1(h_{\pi^h(4s-i)}) + \mu_0(h_{\pi^h(4s-k)}))(\mu_1(h_{\pi^h(4s-j)}) + \mu_0(h_{\pi^h(4s-l)})) \\
&\qquad + (\mu_1(h_{\pi^h(4s-i)}) + \mu_0(h_{\pi^h(4s-l)}))(\mu_1(h_{\pi^h(4s-j)}) + \mu_0(h_{\pi^h(4s-k)})) \\
&= \frac{1}{12}(g_h(h_{\pi^h(4j-3)}) + g_h(h_{\pi^h(4j-2)}))(g_h(h_{\pi^h(4j-1)}) + g_h(h_{\pi^h(4j)})) \\
&\qquad + \frac{1}{12}(g_h(h_{\pi^h(4j-3)}) + g_h(h_{\pi^h(4j-1)}))(g_h(h_{\pi^h(4j-2)}) + g_h(h_{\pi^h(4j)})) \\
&\qquad + \frac{1}{12}(g_h(h_{\pi^h(4j-3)}) + g_h(h_{\pi^h(4j)}))(g_h(h_{\pi^h(4j-2)}) + g_h(h_{\pi^h(4j-1)})) \ ,
\end{aligned}$$

where the coefficient $\frac{1}{12} = \frac{1}{2} \times \frac{1}{6}$ appears in the first equality because there are $\binom{4}{2} = 6$ ways to choose 2 units among 4 units to be treated. The consistency of $(\hat{\varsigma}_{h,n}^{\text{four}})^2$ in (S.50) then follows from similar arguments to those used in the proof of Theorem IV.2. By repeating the arguments in Section C.4, I can also show the variance estimator is nonnegative.

Finally, I discuss the variance estimator on p.100 of in Athey and Imbens (2017). Suppose $\lambda_s = \{i(s), j(s), k(s), l(s)\}$, $D_{i(s)} = D_{j(s)} = 1$, and $D_{k(s)} = D_{l(s)} = 0$. The variance estimator is constructed as

$$\sum_{1 \leq s \leq n/2} n\Big(\frac{4}{2n}\Big)^2 \Big(\frac{1}{4}(Y_{i(s)} - Y_{j(s)})^2 + \frac{1}{4}(Y_{k(s)} - Y_{l(s)})^2\Big) \ .$$

It follows from similar arguments to those used above that the variance estimator converges in probability to

$$E[\text{Var}[Y_i(1)|h(X_i)]] + E[\text{Var}[Y_i(1)|h(X_i)]] \ ,$$

which is less than the asymptotic variance of $\hat{\theta}_n$, i.e., ς_h^2 in (11), and strictly so unless (9) holds. Therefore, unless (9) holds, the test in Athey and Imbens (2017) fails to control size. The failure of size control for Athey and Imbens (2017) also arises in settings with a fixed number of strata, as discussed in Section 5 of Bugni, Canay and Shaikh (2019). ■

D Additional Simulation Results

This section contains the tables with the raw numbers for the main text and some additional simulation results.

Table S.1–S.2 contain the raw numbers for Table 2 in the main text.

Table S.3 contains the summary statistics for the following stratifications in addition to the ones in the main text:

(i") None-reg-int: No stratification with the estimator in Lin (2013), i.e., the OLS estimator of the coefficient on D in the linear regression of Y on a constant, D, $X - \bar{X}_n$, and $D(X - \bar{X}_n)$, where $\bar{X}_n$ is the sample average of X_i's, together with White's heteroskedasticity-robust standard error.

(j) MS X2: Matched sets of four to minimize the sum of the Mahalanobis distances of X_2, namely, all covariates in the main regression specification except the baseline outcome.

(k) MS pilot: Matched sets of four according to $\tilde{g}_m$ from the pilot.

(l) MS pen: Matched sets of four to minimize the sum of the distances in (10) of all covariates.

Tables S.4–S.5 contain the raw numbers for stratifications (j)–(l).

Table S.6 includes the size of the test in Athey and Imbens (2017) for matched sets of four. Athey and Imbens (2017) assume a finite-population setting, in which the potential outcomes and the covariates are fixed, and the parameter of interest is the sample ATE, defined as the average difference between the treated and untreated potential outcomes of all units in the sample. My paper instead focuses on the ATE and assumes that units are drawn from a superpopulation, so potential outcomes and covariates are random. I have shown in Section C.7 that because of the differences in sampling frameworks, the test in Athey and Imbens (2017) does not control size in my setting unless (9) holds, which only happens in Model 1. Therefore, in Model 1, the size of the test is around the nominal level of 5%. In Models 2 and 3, the size of the test is larger than 5%.

Table S.1: MSEs, size, and standard errors for stratifications (a)–(i) across papers 1–5

Paper	Model	θ		(a) MP X	(b) MS X	(c) MP base	(d) MS base	(e) MP X2	(f) MP pilot	(g) MP pen	(h) Origin	(i) None	(i') None-reg
1	1	0	MSE	0.00048	0.00070	-	-	0.00045	0.00066	0.00046	-	0.001	0.00099
			size (adj/adj4)	1.4	5.9	-	-	0.5	3.0	0.7	-	5.1	4.7
			size (MPt)	5.1	-	-	-	4.7	5.5	5.4	-	-	-
			s.e. (adj/adj4)	0.028	0.026	-	-	0.028	0.029	0.028	-	0.032	0.031
			s.e. (MPt)	0.021	-	-	-	0.021	0.025	0.021	-	-	-
	2	0.033	MSE	0.00055	0.00073	-	-	0.00052	0.00071	0.00054	-	0.00092	0.00092
			size (adj/adj4)	2.5	5.1	-	-	2.3	3.4	1.6	-	3.6	3.2
			size (MPt)	2.7	-	-	-	2.2	3.7	1.5	-	-	-
			s.e. (adj/adj4)	0.028	0.027	-	-	0.028	0.029	0.028	-	0.032	0.031
			s.e. (MPt)	0.028	-	-	-	0.028	0.029	0.028	-	-	-
2	1	0	MSE	0.038	0.053	0.070	0.07	0.063	0.053	0.033	-	0.087	0.079
			size (adj/adj4)	1.2	4.8	4.4	4.6	3.9	2.6	0.60	-	5.7	5.3
			size (MPt)	4.8	-	4.6	-	5.4	5.3	5.0	-	-	-
			s.e. (adj/adj4)	0.25	0.23	0.27	0.27	0.26	0.26	0.25	-	0.29	0.28
			s.e. (MPt)	0.19	-	0.27	-	0.24	0.22	0.19	-	-	-
	2	0.18	MSE	0.047	0.053	0.073	0.070	0.062	0.057	0.049	-	0.081	0.073
			size (adj/adj4)	1.7	3.6	4.9	4.4	3.9	3.3	2.0	-	4.9	5.2
			size (MPt)	1.7	-	4.7	-	3.7	2.7	1.9	-	-	-
			s.e. (adj/adj4)	0.25	0.24	0.27	0.27	0.26	0.26	0.25	-	0.29	0.28
			s.e. (MPt)	0.26	-	0.28	-	0.27	0.27	0.26	-	-	-
	3	0.012	MSE	0.060	0.069	0.058	0.062	0.076	0.070	0.058	-	0.091	0.090
			size (adj/adj4)	2.5	5.2	4.6	4.8	4.7	3.4	2.5	-	4.9	4.8
			size (MPt)	2.5	-	4.2	-	3.5	3.4	2.4	-	-	-
			s.e. (adj/adj4)	0.28	0.27	0.25	0.25	0.29	0.28	0.28	-	0.3	0.3
			s.e. (MPt)	0.28	-	0.25	-	0.29	0.29	0.28	-	-	-
3	1	0	MSE	0.079	0.13	0.15	0.17	0.15	0.11	0.075	0.22	0.22	0.17
			size (adj/adj4)	0.3	5.2	4.8	4.9	2.9	2.7	0.4	5.2	4.4	4.5
			size (MPt)	5.2	-	4.9	-	4.1	5.3	4.0	-	-	-
			s.e. (adj/adj4)	0.38	0.35	0.41	0.41	0.42	0.38	0.38	0.47	0.47	0.41
			s.e. (MPt)	0.28	-	0.40	-	0.39	0.33	0.28	-	-	-
	2	0.41	MSE	0.11	0.15	0.16	0.16	0.16	0.13	0.11	0.20	0.21	0.16
			size (adj/adj4)	2.6	6.6	4.9	5.4	4.1	4.2	2.50	4	4.2	4.8
			size (MPt)	2.2	-	4.9	-	3.6	3.1	1.9	-	-	-
			s.e. (adj/adj4)	0.38	0.36	0.40	0.40	0.43	0.39	0.38	0.46	0.46	0.41
			s.e. (MPt)	0.39	-	0.40	-	0.44	0.40	0.39	-	-	-
	3	0.60	MSE	0.084	0.10	0.098	0.10	0.11	0.092	0.078	0.13	0.13	0.11
			size (adj/adj4)	3.0	6.2	4.7	4.6	4.8	5.1	2.7	5.3	5	4.7
			size (MPt)	2.3	-	4.1	-	3.8	3.7	1.9	-	-	-
			s.e. (adj/adj4)	0.32	0.30	0.31	0.31	0.34	0.31	0.31	0.36	0.36	0.33
			s.e. (MPt)	0.33	-	0.32	-	0.35	0.33	0.33	-	-	-
4	1	0	MSE	0.045	0.069	0.089	0.084	0.086	0.058	0.042	0.14	0.15	0.14
			size (adj/adj4)	0.9	4.7	5.9	4.7	2.0	2.2	1.0	3.7	4.5	4.7
			size (MPt)	5.5	-	6	-	5.4	5.6	5.6	-	-	-
			s.e. (adj/adj4)	0.28	0.26	0.29	0.29	0.35	0.27	0.26	0.37	0.39	0.38
			s.e. (MPt)	0.21	-	0.29	-	0.29	0.23	0.20	-	-	-
	2	-1.31	MSE	0.058	0.072	0.075	0.078	0.089	0.064	0.052	0.14	0.15	0.14
			size (adj/adj4)	2.2	4.8	3.9	4.6	2.3	3.1	2.1	5.4	4.4	3.9
			size (MPt)	2.2	-	3.6	-	3.0	2.7	1.5	-	-	-
			s.e. (adj/adj4)	0.28	0.27	0.29	0.29	0.35	0.27	0.27	0.37	0.39	0.38
			s.e. (MPt)	0.29	-	0.29	-	0.34	0.28	0.28	-	-	-
	3	-1.78	MSE	0.075	0.086	0.070	0.070	0.13	0.072	0.062	0.18	0.17	0.17
			size (adj/adj4)	2.2	4.3	5	5.9	2.7	4.8	2.7	5.7	4.3	3.6
			size (MPt)	1.9	-	4.7	-	3.1	3.8	1.3	-	-	-
			s.e. (adj/adj4)	0.32	0.30	0.27	0.27	0.40	0.28	0.27	0.41	0.43	0.42
			s.e. (MPt)	0.32	-	0.27	-	0.39	0.30	0.30	-	-	-
5	1	0	MSE	0.55	0.82	0.84	1.02	1.03	0.79	0.59	-	1.24	1.25
			size (adj/adj4)	1.4	5.2	3.4	6.1	6.0	2.9	1.8	-	6.9	5.4
			size (MPt)	4.6	-	5.6	-	5.4	5.1	5.5	-	-	-
			s.e. (adj/adj4)	0.99	0.92	1.04	1.01	1.01	1.03	1.01	-	1.09	1.13
			s.e. (MPt)	0.75	-	0.90	-	1	0.88	0.74	-	-	-
	2	-0.35	MSE	0.82	0.90	1.07	1.12	1.17	1.01	0.90	-	1.26	1.28
			size (adj/adj4)	3.8	5.1	5.4	6.1	6.9	4.6	4.6	-	5.6	5.1
			size (MPt)	2.8	-	3.9	-	6.3	4.5	3.3	-	-	-
			s.e. (adj/adj4)	1.00	0.97	1.05	1.04	1.07	1.05	1.02	-	1.13	1.16
			s.e. (MPt)	1.07	-	1.11	-	1.07	1.08	1.07	-	-	-
	3	-0.54	MSE	0.92	1.01	0.94	1.08	1.11	1.03	0.94	-	1.20	1.21
			size (adj/adj4)	3.5	5.3	3.2	5	4.8	4.2	3.80	-	5.5	4.6
			size (MPt)	2.6	-	2.8	-	5.3	4.2	2.7	-	-	-
			s.e. (adj/adj4)	1.05	1.01	1.04	1.02	1.06	1.05	1.04	-	1.09	1.13
			s.e. (MPt)	1.11	-	1.10	-	1.06	1.09	1.11	-	-	-

For each stratification, I report (1) the MSE, (2) the size of testing (12) for $\theta_0 = \theta$ at the 5% level, in percentage, and (3) the average standard error. The tests used in this table are as follows: for matched-pair designs, the adjusted t-test with the variance estimator in (14) (adj) and the test in Imbens and Rubin (2015) (MPt); for matched sets of four, the adjusted t-test with the variance estimator in (S.50) in the supplement (adj4); for the original stratifications, the test in (23) of Bugni, Canay and Shaikh (2018); for no stratification, the two-sample t-test; for the regression-adjusted estimator, the t-test with White's heteroskedasticity-robust standard error. Rows are labeled according to the papers, models, and metrics. Columns are labeled according to the stratifications. The definitions of the stratifications can be found in the main text.

Table S.2: MSEs, size, and standard errors for stratifications (a)–(i) across papers 6–10

Paper	Model	θ		(a) MP X	(b) MS X	(c) MP base	(d) MS base	(e) MP X2	(f) MP pilot	(g) MP pen	(h) Origin	(i) None	(i') None-reg
6	1	0	MSE	0.000046	0.000072	-	-	0.000045	0.000072	0.000044	0.00012	0.00012	0.00011
			size (adj/adj4)	0.7	5.1	-	-	0.6	3.0	0.8	5.5	4.4	4.2
			size (MPt)	5.0	-	-	-	5.1	5.3	3.7	-	-	-
			s.e. (adj/adj4)	0.0094	0.0086	-	-	0.0094	0.0097	0.0094	0.011	0.011	0.011
			s.e. (MPt)	0.0068	-	-	-	0.0068	0.0084	0.0068	-	-	-
	2	0.01	MSE	0.000082	0.000098	-	-	0.000081	0.000094	0.000081	0.00013	0.00013	0.00013
			size (adj/adj4)	3.1	4.8	-	-	2.2	3.5	2.8	5.3	6.1	6.7
			size (MPt)	2.7	-	-	-	1.7	3.0	2.0	-	-	-
			s.e. (adj/adj4)	0.010	0.0098	-	-	0.010	0.010	0.010	0.011	0.011	0.011
			s.e. (MPt)	0.011	-	-	-	0.011	0.011	0.011	-	-	-
7	1	0	MSE	0.014	0.024	-	-	0.013	0.020	0.015	0.033	0.031	0.029
			size (adj/adj4)	0.3	5.3	-	-	0.5	2.6	0.7	4.6	4.2	3.8
			size (MPt)	4.6	-	-	-	3.4	5.2	5.3	-	-	-
			s.e. (adj/adj4)	0.17	0.15	-	-	0.16	0.17	0.16	0.18	0.19	0.18
			s.e. (MPt)	0.12	-	-	-	0.12	0.14	0.12	-	-	-
	2	0.21	MSE	0.024	0.028	-	-	0.022	0.028	0.025	0.033	0.036	0.033
			size (adj/adj4)	2.6	4.9	-	-	2.3	3.8	3.6	4.7	5	4.5
			size (MPt)	1.7	-	-	-	1.3	3.7	3.0	-	-	-
			s.e. (adj/adj4)	0.17	0.17	-	-	0.17	0.18	0.17	0.19	0.19	0.19
			s.e. (MPt)	0.18	-	-	-	0.18	0.18	0.18	-	-	-
8	1	0	MSE	0.19	0.28	0.41	0.41	0.20	0.27	0.18	-	0.46	0.47
			size (adj/adj4)	0.5	4.5	5.7	4.8	0.9	2.1	1.5	-	5.4	5.9
			size (MPt)	6.4	-	5.6	-	4.2	4.6	5.4	-	-	-
			s.e. (adj/adj4)	0.59	0.54	0.63	0.63	0.60	0.59	0.58	-	0.67	0.67
			s.e. (MPt)	0.42	-	0.63	-	0.45	0.51	0.42	-	-	-
	2	0.041	MSE	0.24	0.32	0.43	0.38	0.26	0.32	0.25	-	0.45	0.45
			size (adj/adj4)	1.9	5.5	5.4	4.7	2.3	3.1	2.0	-	4.6	4.9
			size (MPt)	2.7	-	5.7	-	3.2	3.1	3.0	-	-	-
			s.e. (adj/adj4)	0.59	0.56	0.63	0.63	0.61	0.61	0.59	-	0.67	0.66
			s.e. (MPt)	0.56	-	0.63	-	0.58	0.60	0.57	-	-	-
	3	1.07	MSE	0.42	0.47	0.40	0.41	0.42	0.52	0.41	-	0.59	0.59
			size (adj/adj4)	3.5	4.9	5.7	5.3	3.0	5.5	3.6	-	5.6	5.0
			size (MPt)	2.3	-	4.8	-	2.4	4.0	3.0	-	-	-
			s.e. (adj/adj4)	0.71	0.69	0.63	0.63	0.72	0.72	0.71	-	0.75	0.76
			s.e. (MPt)	0.76	-	0.64	-	0.76	0.76	0.75	-	-	-
9	1	0	MSE	0.0042	0.0065	0.0082	0.0090	0.0044	0.0057	0.0038	0.011	0.0095	0.0094
			size (adj/adj4)	1.0	5.2	3.6	5.1	0.8	2.2	0.5	5.7	5.5	6.0
			size (MPt)	5.0	-	3.9	-	5.5	4.8	4.8	-	-	-
			s.e. (adj/adj4)	0.086	0.079	0.094	0.094	0.089	0.087	0.085	0.099	0.10	0.098
			s.e. (MPt)	0.063	-	0.094	-	0.065	0.075	0.062	-	-	-
	2	-0.10	MSE	0.0065	0.0077	0.0097	0.0089	0.0068	0.0075	0.0065	0.010	0.0094	0.0093
			size (adj/adj4)	2.9	6.1	5.8	5.2	3.3	4.3	3.0	7.3	5.9	5.7
			size (MPt)	2.2	-	6.2	-	2.6	3.6	2.9	-	-	-
			s.e. (adj/adj4)	0.09	0.087	0.096	0.096	0.091	0.092	0.09	0.097	0.098	0.097
			s.e. (MPt)	0.094	-	0.096	-	0.094	0.095	0.093	-	-	-
	3	-0.012	MSE	0.0065	0.0073	0.0076	0.0077	0.0064	0.0068	0.0065	0.0080	0.0078	0.0079
			size (adj/adj4)	4.8	6.4	5.6	6.2	4.3	5.4	4.8	6.4	5.2	5.7
			size (MPt)	3.4	-	6	-	2.2	3.0	2.3	-	-	-
			s.e. (adj/adj4)	0.083	0.081	0.085	0.085	0.083	0.083	0.083	0.085	0.086	0.087
			s.e. (MPt)	0.094	-	0.086	-	0.094	0.091	0.094	-	-	-
10	1	0	MSE	38.26	44.42	44.68	51.34	48.29	37.41	35.58	-	56.29	52.54
			size (adj/adj4)	2.9	5.2	4.2	5.5	3.8	2.6	2.3	-	5.7	5.0
			size (MPt)	5.1	-	4.6	-	3.8	3.3	3.3	-	-	-
			s.e. (adj/adj4)	6.73	6.56	6.91	6.86	7.1	6.65	6.63	-	7.44	7.25
			s.e. (MPt)	6.16	-	6.72	-	6.99	6.32	6.1	-	-	-
	2	5.61	MSE	53.10	57.04	78.45	86.84	76.40	57.85	53.09	-	91.23	84.21
			size (adj/adj4)	4.4	4.9	4.9	6.2	4.4	5.1	4.0	-	5.3	5.0
			size (MPt)	4.8	-	4.8	-	4.9	5.3	3.5	-	-	-
			s.e. (adj/adj4)	7.67	7.58	8.9	8.87	8.79	7.65	7.59	-	9.5	9.23
			s.e. (MPt)	7.61	-	8.92	-	8.79	7.65	7.57	-	-	-
	3	1.91	MSE	49.23	55.22	46.99	52.37	71.83	51.18	50.39	-	83.09	79.53
			size (adj/adj4)	3.4	4.2	4.2	6.3	4.2	3.9	3.4	-	5.5	5.4
			size (MPt)	3.4	-	4.8	-	4.2	3.4	3.8	-	-	-
			s.e. (adj/adj4)	7.76	7.59	7.21	7.16	8.5	7.69	7.68	-	8.69	8.63
			s.e. (MPt)	7.68	-	7.14	-	8.55	7.7	7.66	-	-	-

For each stratification, I report (1) the MSE, (2) the size of testing (12) for $\theta_0 = \theta$ at the 5% level, in percentage, and (3) the average standard error. The tests used in this table are as follows: for matched-pair designs, the adjusted t-test with the variance estimator in (14) (adj) and the test in Imbens and Rubin (2015) (MPt); for matched sets of four, the adjusted t-test with the variance estimator in (S.50) in the supplement (adj4); for the original stratifications, the test in (23) of Bugni, Canay and Shaikh (2018); for no stratification, the two-sample t-test; for the regression-adjusted estimator, the t-test with White's heteroskedasticity-robust standard error. Rows are labeled according to the papers, models, and metrics. Columns are labeled according to the stratifications. The definitions of the stratifications can be found in the main text.

Table S.3: Summary statistics for MSEs, size, and standard errors for additional methods not in the main text across all papers and models

Stratification	MSE (ratio vs. None)	size (%)	s.e. (ratio vs. None)
MS X2	0.798	5.207	0.890
	[0.551, 0.938]	[3.200, 6.300]	[0.770, 0.975]
MS pilot	0.749	5.256	0.857
	[0.444, 0.939]	[3.900, 6.600]	[0.644, 0.957]
MS pen	0.693	4.604	0.844
	[0.402, 0.966]	[3.100, 7.200]	[0.690, 0.940]
None-reg-int	0.949	4.870	0.984
	[0.782, 1.014]	[3.300, 6.800]	[0.890, 1.041]

For each stratification, I report summary statistics across all papers and models of (1) the ratio between the MSE under the particular stratification and the MSE under no stratification, (2) the size of testing (12) for $\theta_0 = \theta$ at the 5% level, in percentage, and (3) the ratio between the average standard error under the particular stratification and the average standard error under no stratification. The tests used in this table are: for MS, the adjusted t-test with the variance estimator in (S.50); for the regression-adjusted estimator, the t-test with White's heteroskedasticity-robust standard error. For each metric, I show the average and $[\min, \max]$ across all papers and models. Rows are labeled according to the stratifications. Columns are labeled according to the metrics. The definitions of the stratifications can be found in the text.

Table S.4: MSEs, size, and standard errors for stratifications (j)–(l) across papers 1–5

Paper	Model	θ		(j) MS X2	(k) MS pilot	(l) MS pen	(i") None-reg-int
1	1	0	MSE	0.00070	0.00075	0.00067	0.00099
			size (adj4)	6.3	6.1	3.1	4.7
			s.e. (adj4)	0.026	0.028	0.027	0.032
	2	0.033	MSE	0.00076	0.00081	0.00072	0.00092
			size (adj4)	6.3	4.7	4.9	3.3
			s.e. (adj4)	0.027	0.029	0.027	0.031
2	1	0	MSE	0.063	0.057	0.048	0.079
			size (adj4)	4.3	4.6	3.2	5.3
			s.e. (adj4)	0.26	0.25	0.25	0.28
	2	0.18	MSE	0.067	0.067	0.052	0.073
			size (adj4)	4.8	6.0	3.3	5.3
			s.e. (adj4)	0.26	0.26	0.25	0.28
	3	0.012	MSE	0.076	0.074	0.069	0.090
			size (adj4)	4.3	3.9	4.4	4.7
			s.e. (adj4)	0.28	0.28	0.28	0.30
3	1	0	MSE	0.17	0.14	0.13	0.17
			size (adj4)	4.6	5.1	3.9	4.3
			s.e. (adj4)	0.41	0.37	0.37	0.42
	2	0.41	MSE	0.17	0.16	0.13	0.16
			size (adj4)	4.7	6.2	4.3	4.6
			s.e. (adj4)	0.42	0.38	0.37	0.41
	3	0.60	MSE	0.12	0.099	0.090	0.11
			size (adj4)	5.4	5.9	4.5	5.1
			s.e. (adj4)	0.33	0.31	0.31	0.33
4	1	0	MSE	0.11	0.067	0.063	0.14
			size (adj4)	5.2	4.6	3.3	4.7
			s.e. (adj4)	0.33	0.26	0.27	0.38
	2	-1.31	MSE	0.11	0.070	0.065	0.14
			size (adj4)	5.5	3.9	3.4	3.9
			s.e. (adj4)	0.34	0.27	0.27	0.38
	2	-1.78	MSE	0.15	0.077	0.070	0.17
			size (adj4)	4.8	5.0	3.2	3.6
			s.e. (adj4)	0.38	0.28	0.30	0.42
5	1	0	MSE	1.01	0.93	0.99	1.25
			size (adj4)	5.3	4.9	6.9	5.2
			s.e. (adj4)	1.01	0.99	0.96	1.13
	2	-0.35	MSE	1.11	1.08	1.00	1.28
			size (adj4)	6.1	6.3	4.4	5.0
			s.e. (adj4)	1.07	1.03	0.99	1.17
	3	-0.54	MSE	1.07	1.03	1.12	1.21
			size (adj4)	5.6	4.7	6.3	4.5
			s.e. (adj4)	1.06	1.04	1.02	1.13

For each stratification, I report (1) the MSE, (2) the size of testing (12) for $\theta_0 = \theta$ at the 5% level, in percentage, and (3) the average standard error. The tests used in this table are: for MS, the adjusted t-test with the variance estimator in (S.50); for the regression-adjusted estimator, the t-test with White's heteroskedasticity-robust standard error. Rows are labeled according to the papers and models. Columns are labeled according to the stratifications. The definitions of the stratifications can be found in the text.

Table S.5: MSEs, size, and standard errors for stratifications (j)–(l) across papers 6–10

Paper	Model	θ		(j) MS X2	(k) MS pilot	(l) MS pen	(i") None-reg-int
6	1	0	MSE	0.000068	0.000093	0.000079	0.00011
			size (adj4)	3.2	5.7	4.7	4.2
			s.e. (adj4)	0.0086	0.0093	0.00091	0.011
	2	0.010	MSE	0.000096	0.00012	0.00010	0.00013
			size (adj4)	5.4	6.6	5.4	6.8
			s.e. (adj4)	0.0098	0.010	0.0099	0.011
7	1	0	MSE	0.025	0.026	0.023	0.029
			size (adj4)	6.3	4.9	5.1	3.7
			s.e. (adj4)	0.15	0.16	0.15	0.19
	2	0.21	MSE	0.028	0.031	0.029	0.033
			size (adj4)	5.0	5.2	5.1	4.6
			s.e. (adj4)	0.17	0.17	0.17	0.19
8	1	0	MSE	0.31	0.34	0.28	0.47
			size (adj4)	5.2	5.4	3.4	5.8
			s.e. (adj4)	0.56	0.57	0.56	0.67
	2	0.041	MSE	0.34	0.35	0.31	0.45
			size (adj4)	5.6	4.8	3.6	4.8
			s.e. (adj4)	0.59	0.59	0.58	0.67
	3	1.07	MSE	0.46	0.52	0.48	0.59
			size (adj4)	4.5	6.0	4.7	5.1
			s.e. (adj4)	0.70	0.71	0.69	0.75
9	1	0	MSE	0.0067	0.0068	0.0065	0.0094
			size (adj4)	4.8	4.8	5.2	6.0
			s.e. (adj4)	0.082	0.083	0.082	0.010
	2	-0.10	MSE	0.0080	0.0083	0.0085	0.0094
			size (adj4)	5.4	6.0	6.2	5.6
			s.e. (adj4)	0.088	0.090	0.088	0.097
	3	-0.012	MSE	0.0068	0.0074	0.0076	0.0079
			size (adj4)	5.8	6.5	7.2	5.6
			s.e. (adj4)	0.082	0.082	0.081	0.088
10	1	0	MSE	52.82	39.16	40.30	52.49
			size (adj4)	5.1	4.4	3.7	4.8
			s.e. (adj4)	7.07	6.55	6.56	7.29
	2	5.61	MSE	85.45	59.39	56.69	84.05
			size (adj4)	6.3	4.4	6.1	5.0
			s.e. (adj4)	8.78	7.62	7.55	9.26
	3	1.91	MSE	71.99	56.51	54.92	79.53
			size (adj4)	4.8	5.3	4.8	5.3
			s.e. (adj4)	8.47	7.60	7.57	8.63

For each stratification, I report (1) the MSE, (2) the size of testing (12) for $\theta_0 = \theta$ at the 5% level, in percentage, and (3) the average standard error. The tests used in this table are: for MS, the adjusted t-test with the variance estimator in (S.50); for the regression-adjusted estimator, the t-test with White's heteroskedasticity-robust standard error. Rows are labeled according to the papers and models. Columns are labeled according to the stratifications. The definitions of the stratifications can be found in the text.

Table S.6: The size of the test in Athey and Imbens (2017) for matched sets of four

Paper	Model	(b) MS X	(d) MS base
1	1	6.0	-
	2	8.2	-
2	1	5.1	4.8
	2	6.3	5.1
	3	6.3	5.1
3	1	5.7	5.2
	2	8.7	5.8
	3	8.0	5.7
4	1	5.0	5.0
	2	7.6	4.2
	3	6.4	5.9
5	1	6.0	6.6
	2	7.7	9.6
	3	7.7	9.0
6	1	5.2	-
	2	7.7	-
7	1	5.5	-
	2	8.5	-
8	1	4.4	4.8
	2	7.1	5.0
	3	8.5	5.6
9	1	6.2	5.3
	2	8.7	5.2
	3	10.4	6.2
10	1	5.3	5.8
	2	6.1	6.3
	3	5.2	6.6

This table shows the size of the test in Athey and Imbens (2017) for testing (12) for $\theta_0 = \theta$ at the 5% level, in percentage. Rows are labeled according to papers and models. Columns are labeled according to the stratifications. The definitions of the stratifications can be found in the main text.

Here are the details of all the data used in simulation:

1. Herskowitz (2021):

   I re-implement the analysis on p.93 of the original paper and estimate the effect of lumpy prime on demand. I use Wave 2 data from "Panel-Clean.dta". There are 997 observations for simulation. I fill in the missing values and choose the covariates following the "OVERALL SPECIFICATION COVARIATES" part in Analysis-3.do. 20% of the original data are sampled with replacement and fixed throughout the replications to be used as the pilot data, which contains 199 observations. I consider Model 1 and Model 2 for data imputation. When drawing from the empirical distribution, 996 observations are sampled with replacement so the sample size is a multiple of four.

   dependent: lmp_matrix (an indicator for whether the maximum number of tickets was demanded.)

   covariates: meaninc (mean income for duration of study), betmean_prop (mean amount spent on betting / mean income during study), lmp_purchased, lumpy_incprop

   treatment: lumpyprime (lumpy expenditure prime treatment group)

2. Lee et al. (2021):

   I re-implement the analysis in (1) on p.49 of the original paper and estimate the effect of treatment on total remittances sent from migrants. I rerun the 1–3 code files in folder "Migrant-Survey" and get the migrants data "Endline-Data-Combined-Status-Merged-Ready-18-Active-Acc.dta" for regression. There are 809 observations. 20% of the original data are sampled with replacement and fixed throughout the replications to be used as the pilot data, which contains 161 observations. I consider Models 1–3 for data imputation. When drawing from the empirical distribution, 808 observations are sampled with replacement so the sample size is a multiple of four.

   dependent: log_total_remittances (missing values generated because of log transformation and are filled using 0)

   covariates: log_total_remittances_b (baseline outcome), household_size (missing values are filled using baseline outcomes), hohh_age, hohh_female, and hohh_completed_primary

   treatment: treatment

3. Abel, Burger and Piraino (2020):

   I re-implement the analysis in (4) on p.56 of the original paper and estimated the effect of reference letters on employment. I used experiment 2 data from "experiment2_employment.dta". There are 1000 observations. 20% of the original data are sampled with replacement and fixed throughout the replications to be used as the pilot data, which contains 200 observations. I consider Models 1–3 for data imputation. When drawing from the empirical distribution, 1000 observations are sampled with replacement so the sample size is a multiple of four.

dependent: f2_b3 (number of jobs applied in the last four months, 246 observations with missing values are dropped. the treatment percentage did not change much, about 0.55)

covariates: bs_c3_jobs_applied (baseline outcome), age_yr, female_d, educ_yr (missing values are filled using the mean), married_d, lang_zulu_d, lang_xhosa_d, lang_venda_d (there are 4 languages and here we used 3 dummy variables)

treatment: reference_d

original stratification: gender, total 2

4. Gerber et al. (2020):

   I re-perform the OLS on p.303 of the original paper and estimat the effect of the close poll treatment on vote margin predictions using data from 2010 RCT experiment and following lines 677–710 in code file "Main_e2010.do". Because the observations in the control group do not have any experiment records, I only consider the two treatment groups "close" and "not close." There are 6650 observations. 20% of the original data are sampled with replacement and fixed throughout the replications to be used as the pilot data, which contains 1330 observations. I consider Models 1–3 for data imputation. When drawing from the empirical distribution, 6648 observations are sampled with replacement so the sample size is a multiple of four.

   dependent: votemarg_post (post-treat vote margin prediction)

   covariates: int_gov_scale (interest in politics, 1-5 scale), pelosi (identify Nancy Pelosi as speaker), vote_admin_past (share voted previous 5 elections), male, race (4 dummy variables), schooling (schooling years), inc(1-5 scale), age (1-7 scale)

   treatment: t_close

   original stratification: ppstaten, in each replication, only consider states with more than 5 observations.

5. Deserranno, Stryjan and Sulaiman (2019):

   I re-implement the group-level analysis on p.263 of the original paper, which is also shown in Table A.15 of the online appendix, and estimate the effect of treatment on wealth score inequality. I follow lines 1250–1289 in code file "AEJApp-2018-0248_Tables-and-Figures.do" and transfer member-level panel data into group-level using IQR function. After discarding observations with missing group indices, I finally get 92 groups. 20% of the original data are sampled with replacement and fixed throughout the replications to be used as the pilot data, which contains 18 observations. I consider Models 1–3 for data imputation. When drawing from the empirical distribution, 92 observations are sampled with replacement so the sample size is a multiple of four.

   dependent: iqrwealth (endline group-level wealth scores, generated from wealth_endline)

   covariates: branch (scale from 1 to 9, numeric)

   treatment: vote

6. Barrera-Osorio, Linden and Saavedra (2019):

   I re-implement the analysis on p.268, table 3, column (1) of the original paper and estimate the impacts of the basic treatment on on-time secondary enrollment outcomes. I follow lines 144–151 in code file "Final_Tables_Journal.do." Observations with missing values on the dependent variable are filtered out. Moreover, variables ending in "_missing" recorded the missing status of correspondent variables ending in "_fill" and I impute "_fill" variables using the median. Running one replication using full data takes 6 hours, so I randomly sample 7880 out of the 15759 observations to reduce the running time. 20% of the original data are sampled with replacement and fixed throughout the replications to be used as the pilot data, which contains 1576 observations. I consider Models 1–2 for data imputation. When drawing from the empirical distribution, 7880 observations are sampled with replacement so the sample size is a multiple of four.

   dependent: on_time (binary, whether enrolled or not)

   covariates: s_teneviv_fill (indicator of house ownership), s_utilities_fill (number of utilities in the house), s_durables_fill (number of durable goods), s_infraest_hh_fill (infrastructure in the household, scale 0-22) , s_age_sorteo_fill (age at the moment of lottery), s_sexo_fill (gender of student), s_yrs_fill (years of education), grade_fill (grade at baseline), s_single_fill (if the household is single headed), s_edadhead_fill (age of the head of the household), s_yrshead_fill (years of education of the head), s_tpersona_fill (number of individuals in the house, scale 0-22), s_num18_fill (number of people under 18 in the house), s_estrato_fill (strata of the household, scale 0-2), s_puntaje_fill (SISBEN score), s_ingtotal_fill (income, from 0-4000).

   treatment: treatment

   original stratification: grader(6-11), gender, total 12

7. Himmler, Jäckle and Weinschenk (2019):

   I re-implement the analysis of table 2, column (9) on p.130 of the original paper and estimate the effect of commitment treatment on the number of exams passed. I use data in "soft_commitments_AEJ.dta" and followed lines 74–77 in code file "soft_commitments_AEJ.do." There are 392 observations, and 32.91% are assigned to the treatment group in the original paper. Then, 20% of the original data are sampled with replacement and fixed throughout the replications to be used as the pilot data, which contains 78 observations. There are no baseline outcomes. Models 1–2 are considered for data imputation. When drawing from the empirical distribution, 392 observations are sampled with replacement so the sample size is a multiple of four.

   dependent: pass_all (number of exams passed)

   covariates: male, c_HSGPA (centered high school GPA), age (scale 1-21, generated from original dummy variables dage1 - dage21), dschooltype1 , dschooltype2 (two binary school type

variables), nongerman (foreign citizership), c_app_day (centered application day), freshdeg (fresh HS degree).

treatment: commitment

original stratification: gpa, 4 strata

8. Abel et al. (2019):

   I re-implement the analysis of table 3, column (2) on p.292 of the original paper and estimate the effects of WorkshopPlus on search hours. I use data in file "final_data2.dta" and follow lines 159–163 in code file "AP_final_AEJ.do." I only consider the control group vs the "workshopPlus" group and filter out observations originally assigned to the "workshop" group. There are two outcomes in the experiment. In order to be consistent with the second simulation regarding to the second outcome "application number," I discard the observations missing any "search hours" or "application numbers" values. I get 1479 observations, 45.57% of which are assigned to the treatment group originally. 20% of the original data are sampled with replacement and fixed throughout the replications to be used as pilot data, which contains 295 observations. Models 1-3 are considered for data imputation. Model 2 used baseline1 + covariates. When drawing from the empirical distribution, 1476 observations are sampled with replacement so the sample size is a multiple of four.

   dependent: b2_t (search hours)

   covariates: educ_yr (education years), age_yr (age in years), female_d, lang_ (three dummy variables indicating spoken languages), location_f_ (two dummy variables indicating location), round (follow-up 2)

   baseline: nomiss_bs_c2 (variable miss_bs_c2 is an indicator for missing values, and we filled in missing baseline outcomes using the median)

   treatment: ws_plus_d

   secondary outcome: b3_t (application numbers)

9. de Mel, McKenzie and Woodruff (2019):

   I re-implement the analysis in table 5 panel A "Number of paid workers," column "After subsidy Year 3+" on p.220 of the original paper and estimated the effect of treatment on employment. I follow lines 59–76 to define treatment status and lines 585–606 to perform regression in code file "AEJreplicationfile_LaborDrops.do" using data in "Sri-Lanka-Panel-Experiment-Paper.dta". There are 12 rounds of experiments and the means of rounds 10-12 outcomes are treated as endline (year 3+) outcomes. I filter out observations whose outcomes are missing in any of these three rounds and impute the covariates with missing values using the median. I finally got 454 observations. 20% of the original data are sampled with replacement and fixed throughout the replications to be used as the pilot data, which contains 90 observations. I consider Models 1–3 for data imputation. When drawing from the empirical distribution, 452 observations are sampled with replacement so the sample size is a multiple of four.

dependent: allpaid_trunc (number of all paid works, calculated using mean of round 10-12 grouped by the key variable sheno).

baseline: baseallpaid

covariates: basetotalscore, booster, raven, digitspan, baselK_noland, baseedn, baseprofits. (Did not find descriptions. Variables baselowassets, basehighassets, baselowprofits, basehighprofits are dismissed because they make the matrix not invertible)

treatment: voucheronly

original stratification: there are 6 variables (strata1-strata6)

10. Lafortune, Riutort and Tessada (2018):

    I re-implement the analysis of interactions in Panel A, Column "Income per capita," Table 5 on p.242 of the original paper and estimate the effect of role models on income per capita. I run the provided ".do" files to generate the dataset "Base_Analisis_SEG1.dta" according to the "Read-me.pdf". I then perform data processing and regression analysis on this dataset following the codes in "Interacciones.do". After removing observations with missing outcomes and imputing missing gender and age values using the median, I finally get 979 observations. 20% of the original data are sampled with replacement and fixed throughout the replications to be used as the pilot data, which contains 195 observations. I consider Models 1–3 for data imputation. When drawing from the empirical distribution, 976 observations are sampled with replacement so the sample size is a multiple of four.

    dependent: IncomePC_seg1 (income per capita)

    baseline: lb_IncomePC_seg1

    covariates: Edad (age), mujer (gender), Ed2, Ed3 (education dummy variables, ignore Ed1), NO_info_Educ (indicating missing values of these three education variables), NegocioBsico030_fi, NegocioIntermedio300M1M_fi (description written in foreign language, but are dummy variables indicating different amount of money, ignore NegocioDesarrollado1MM_fi), NO_info_Size (indicating missing values of former three "Nego_" variables

    treatment: asignado_role_Model

E Matched-Pair Designs in the AEA RCT Registry

The following experiments in the AEA RCT Registry use matched-pair designs: AEARCTR-0000086, 0000171, 0000293, 0000443, 0000481, 0000550, 0000578, 0000587, 0000644, 0000688, 0000721, 0000983, 0000986, 0001034, 0001097, 0001218, 0001370, 0001591, 0001607, 0001712, 0001714, 0001778, 0001992, 0001995, 0002010, 0002125, 0002132, 0002282, 0002585, 0002622, 0002664, 0002750, 0002776, 0003056, 0003076, 0003524, 0003581, 0003629, 0003648, 0003779, 0003814, 0003933, 0003994, 0004024, 0004042, 0004022, 0006706.

References

Abel, Martin, Rulof Burger, and Patrizio Piraino. 2020. "The Value of Reference Letters: Experimental Evidence from South Africa." *American Economic Journal: Applied Economics*, 12(3): 40–71.

Abel, Martin, Rulof Burger, Eliana Carranza, and Patrizio Piraino. 2019. "Bridging the Intention-Behavior Gap? The Effect of Plan-Making Prompts on Job Search and Employment." *American Economic Journal: Applied Economics*, 11(2): 284–301.

Athey, Susan, and Guido W Imbens. 2017. "The Econometrics of Randomized Experiments." In *Handbook of Economic Field Experiments*. Vol. 1, 73–140. Elsevier.

Bai, Yuehao, Joseph P. Romano, and Azeem M. Shaikh. 2021. "Inference in Experiments with Matched Pairs*." *Journal of the American Statistical Association*, 0(ja): 1–37. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2021.1883437.

Barrera-Osorio, Felipe, Leigh L. Linden, and Juan E. Saavedra. 2019. "Medium- and Long-Term Educational Consequences of Alternative Conditional Cash Transfer Designs: Experimental Evidence from Colombia." *American Economic Journal: Applied Economics*, 11(3): 54–91.

Bogachev, Vladimir I. 2007. *Measure Theory.* Berlin Heidelberg:Springer-Verlag.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh. 2018. "Inference Under Covariate-Adaptive Randomization." *Journal of the American Statistical Association*, 113(524): 1784–1796. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1375934.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh. 2019. "Inference under covariate-adaptive randomization with multiple treatments." *Quantitative Economics*, 10(4): 1747–1785. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE1150.

Chernozhukov, Victor, Iván Fernández-Val, and Ye Luo. 2018. "The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages." *Econometrica*, 86(6): 1911–1938. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA14415.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2019. "Labor Drops: Experimental Evidence on the Return to Additional Labor in Microenterprises." *American Economic Journal: Applied Economics*, 11(1): 202–235.

Deserranno, Erika, Miri Stryjan, and Munshi Sulaiman. 2019. "Leader Selection and Service Delivery in Community Groups: Experimental Evidence from Uganda." *American Economic Journal: Applied Economics*, 11(4): 240–267.

Gerber, Alan, Mitchell Hoffman, John Morgan, and Collin Raymond. 2020. "One in a Million: Field Experiments on Perceived Closeness of the Election and Voter Turnout." *American Economic Journal: Applied Economics*, 12(3): 287–325.

Herskowitz, Sylvan. 2021. "Gambling, Saving, and Lumpy Liquidity Needs." *American Economic Journal: Applied Economics*, 13(1): 72–104.

Himmler, Oliver, Robert Jäckle, and Philipp Weinschenk. 2019. "Soft Commitments, Reminders, and Academic Performance." *American Economic Journal: Applied Economics*, 11(2): 114–142.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press.

Lafortune, Jeanne, Julio Riutort, and José Tessada. 2018. "Role Models or Individual Consulting: The Impact of Personalizing Micro-entrepreneurship Training." *American Economic Journal: Applied Economics*, 10(4): 222–245.

Lee, Jean N., Jonathan Morduch, Saravana Ravindran, Abu Shonchoy, and Hassan Zaman. 2021. "Poverty and Migration in the Digital Age: Experimental Evidence on Mobile Banking in Bangladesh." *American Economic Journal: Applied Economics*, 13(1): 38–71.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *Annals of Applied Statistics*, 7(1): 295–318. Publisher: Institute of Mathematical Statistics.

Munkres, James R. 1991. *Analysis on manifolds.* Addison-Wesley Publishing Company, Advanced Book Program, Redwood City, CA.

Rudin, Walter. 1976. *Principles of mathematical analysis.* . Third ed., McGraw-Hill Book Co., New York-Auckland-Düsseldorf.

Spivak, Michael. 1965. *Calculus on manifolds. A modern approach to classical theorems of advanced calculus.* W. A. Benjamin, Inc., New York-Amsterdam.

Tabord-Meehan, Max. 2022. "Stratification Trees for Adaptive Randomization in Randomized Controlled Trials." *The Review of Economic Studies*. Forthcoming.