Journal of Econometrics 241 (2024) 105740

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Covariate adjustment in experiments with matched pairs[☆]

Yuehao Bai [a], Liang Jiang [b,*], Joseph P. Romano [c], Azeem M. Shaikh [d], Yichong Zhang [e]

[a] *Department of Economics, University of Southern California, USA*
[b] *International School of Finance, Fudan University, China*
[c] *Departments of Economics & Statistics, Stanford University, USA*
[d] *Department of Economics, University of Chicago, USA*
[e] *School of Economics, Singapore Management University, Singapore*

ARTICLE INFO

JEL classification:
C12
C14

Keywords:
Experiment
Matched pairs
Covariate adjustment
Randomized controlled trial
Treatment assignment
LASSO

ABSTRACT

This paper studies inference for the average treatment effect (ATE) in experiments in which treatment status is determined according to "matched pairs" and it is additionally desired to adjust for observed, baseline covariates to gain further precision. By a "matched pairs" design, we mean that units are sampled i.i.d. from the population of interest, paired according to observed, baseline covariates, and finally, within each pair, one unit is selected at random for treatment. Importantly, we presume that not all observed, baseline covariates are used in determining treatment assignment. We study a broad class of estimators based on a "doubly robust" moment condition that permits us to study estimators with both finite-dimensional and high-dimensional forms of covariate adjustment. We find that estimators with finite-dimensional, linear adjustments need not lead to improvements in precision relative to the unadjusted difference-in-means estimator. This phenomenon persists even if the adjustments interact with treatment; in fact, doing so leads to no changes in precision. However, gains in precision can be ensured by including fixed effects for each of the pairs. Indeed, we show that this adjustment leads to the minimum asymptotic variance of the corresponding ATE estimator among all finite-dimensional, linear adjustments. We additionally study an estimator with a regularized adjustment, which can accommodate high-dimensional covariates. We show that this estimator leads to improvements in precision relative to the unadjusted difference-in-means estimator and also provides conditions under which it leads to the "optimal" nonparametric, covariate adjustment. A simulation study confirms the practical relevance of our theoretical analysis, and the methods are employed to reanalyze data from an experiment using a "matched pairs" design to study the effect of macroinsurance on microenterprise.

1. Introduction

This paper studies inference for the average treatment effect in experiments in which treatment status is determined according to "matched pairs". By a "matched pairs" design, we mean that units are sampled i.i.d. from the population of interest, paired according to observed, baseline covariates and finally, within each pair, one unit is selected at random for treatment. This method

[☆] We thank the Editor, Elie Tamer, the Associate Editor and two anonymous referees for many useful comments that helped to improve this paper. Yichong Zhang acknowledges the financial support from the National Natural Science Foundation of China under the grant No. 72133002 and a Lee Kong Chian fellowship, Singapore. Any and all errors are our own.

[*] Corresponding author.

E-mail addresses: yuehao.bai@usc.edu (Y. Bai), jiangliang@fudan.edu.cn (L. Jiang), romano@stanford.edu (J.P. Romano), amshaikh@uchicago.edu (A.M. Shaikh), yczhang@smu.edu.sg (Y. Zhang).

https://doi.org/10.1016/j.jeconom.2024.105740
Received 12 February 2023; Received in revised form 18 October 2023; Accepted 26 January 2024

is used routinely in all parts of the sciences. Indeed, commands to facilitate its implementation are included in popular software packages, such as `sampsi` in Stata. References to a variety of specific examples can be found, for instance, in the following surveys of various field experiments: Donner and Klar (2000), Glennerster and Takavarasha (2013), and Rosenberger and Lachin (2015). See also Bruhn and McKenzie (2009), who, based on a survey of selected development economists, report that 56% of researchers have used such a design at some point. Bai et al. (2022) develop methods for inference for the average treatment effect in such experiments based on the difference-in-means estimator. In this paper, we pursue the goal of improving upon the precision of this estimator by exploiting observed, baseline covariates that are not used in determining treatment status.

To this end, we study a broad class of estimators for the average treatment effect based on a "doubly robust" moment condition. The estimators in this framework are distinguished via different "working models" for the conditional expectations of potential outcomes under treatment and control given the observed, baseline covariates. Importantly, because of the double-robustness, these "working models" need not be correctly specified for the resulting estimator to be consistent. In this way, the framework permits us to study both finite-dimensional and high-dimensional forms of covariate adjustment without imposing unreasonable restrictions on the conditional expectations themselves. Under high-level conditions on the "working models" and their corresponding estimators and a requirement that pairs are formed so that units within pairs are suitably "close" in terms of the baseline covariates, we derive the limiting distribution of the covariate-adjusted estimator of the average treatment effect. We further construct an estimator for the variance of the limiting distribution and provide conditions under which it is consistent for this quantity.

Using our general framework, we first consider finite-dimensional, linear adjustments. For this class of estimators, our main findings are summarized as follows. First, we find that estimators with such adjustments are not guaranteed to be weakly more efficient than the unadjusted difference-in-means estimator. This finding echoes similar findings by Yang and Tsiatis (2001) and Tsiatis et al. (2008) in settings in which treatment is determined by i.i.d. coin flips, and Freedman (2008) in a finite population setting in which treatment is determined according to complete randomization. See Negi and Wooldridge (2021) for a succinct treatment of that literature. Moreover, we find that this phenomenon persists even if the adjustments are interacted with treatment. In fact, doing so leads to no changes in precision. In this sense, our results diverge from those in settings with complete randomization and treated fraction one half, where adjustments based on the uninteracted and interacted linear adjustments both guarantee gains in precision. Last, we show that estimators with both uninteracted and interacted linear adjustments with pair fixed effects are guaranteed to be weakly more efficient than the unadjusted difference-in-means estimator.

We then use our framework to consider high-dimensional adjustments based on ℓ_1 penalization. Specifically, we first obtain an intermediate estimator by using the LASSO to estimate the "working model" for the relevant conditional expectations. When the treatment is determined according to "matched pairs", however, this estimator need not be more precise than the unadjusted difference-in-means estimator. Therefore, following Cohen and Fogarty (2024), we consider, in an additional step, an estimator based on the finite-dimensional, linear adjustment described above that uses the predicted values for the "working model" as the covariates and includes fixed effects for each of the pairs. We show that the resulting estimator improves upon both the intermediate estimator and the unadjusted difference-in-means estimator in terms of precision. Moreover, we provide conditions under which the refitted adjustments attain the relevant efficiency bound derived by Armstrong (2022).

Concurrent with our paper, Cytrynbaum (2023) considers covariate adjustment in experiments in which units are grouped into tuples with possibly more than two units, rather than pairs. Both our paper and Cytrynbaum (2023) find that finite-dimensional, linear regression adjustments with pair fixed effects are guaranteed to improve precision relative to the unadjusted difference-in-means estimator, and show that such adjustments are indeed optimal among all linear adjustments. However, Cytrynbaum (2023) does not pursue more general forms of covariate adjustments, including the regularized adjustments described above. Such results permit us to study nonparametric adjustments as well as high-dimensional adjustments using covariates whose dimension diverges rapidly with the sample size.

The remainder of our paper is organized as follows. In Section 2, we describe our setup and notation. In particular, there we describe the precise sense in which we require that units in each pair are "close" in terms of their baseline covariates. In Section 3, we introduce our general class of estimators based on a "doubly robust" moment condition. Under certain high-level conditions on the "working models" and their corresponding estimators, we derive the limiting behavior of the covariate-adjusted estimator. In Section 4, we use our general framework to study a variety of estimators with finite-dimensional, linear covariate adjustment. In Section 5, we use our general framework to study covariate adjustment based on regularized regression. In Section 6, we examine the finite-sample behavior of tests based on these different estimators via a small simulation study. We find that covariate adjustment can lead to considerable gains in precision. Finally, in Section 7, we apply our methods to reanalyze data from an experiment using a "matched pairs" design to study the effect of macroinsurance on microenterprise. Proofs of all results and some details for simulations are given in the Online Supplement.

2. Setup and notation

Let $Y_i \in \mathbf{R}$ denote the (observed) outcome of interest for the ith unit, $D_i \in \{0,1\}$ be an indicator for whether the ith unit is treated, and $X_i \in \mathbf{R}^{k_x}$ and $W_i \in \mathbf{R}^{k_w}$ denote observed, baseline covariates for the ith unit; X_i and W_i will be distinguished below through the feature that only the former will be used in determining treatment assignment. Further, denote by $Y_i(1)$ the potential outcome of the ith unit if treated and by $Y_i(0)$ the potential outcome of the ith unit if not treated. The (observed) outcome and potential outcomes are related to treatment status by the relationship

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i). \tag{1}$$

For a random variable indexed by i, A_i, it will be useful to denote by $A^{(n)}$ the random vector $(A_1, \dots, A_{2n})$. Denote by P_n the distribution of the observed data $Z^{(n)}$, where $Z_i = (Y_i, D_i, X_i, W_i)$, and by Q_n the distribution of $U^{(n)}$, where $U_i = (Y_i(1), Y_i(0), X_i, W_i)$. Note that P_n is determined by (1), Q_n, and the mechanism for determining treatment assignment. We assume throughout that $U^{(n)}$ consists of $2n$ i.i.d. observations, i.e., $Q_n = Q^{2n}$, where Q is the marginal distribution of U_i. We therefore state our assumptions below in terms of assumptions on Q and the mechanism for determining treatment assignment. Indeed, we will not refer P_n in the sequel, and all operations are understood to be under Q and the mechanism for determining the treatment assignment. Our object of interest is the average effect of the treatment on the outcome of interest, which may be expressed in terms of this notation as

$$\Delta(Q) = E[Y_i(1) - Y_i(0)]. \tag{2}$$

We now describe our assumptions on Q. We restrict Q to satisfy the following mild requirement:

Assumption 2.1. The distribution Q is such that

(a) $0 < E[\mathrm{Var}[Y_i(d)|X_i]]$ for $d \in \{0,1\}$.
(b) $E[Y_i^2(d)] < \infty$ for $d \in \{0,1\}$.
(c) $E[Y_i(d)|X_i = x]$ and $E[Y_i^2(d)|X_i = x]$ are Lipschitz for $d \in \{0,1\}$.

Next, we describe our assumptions on the mechanism determining treatment assignment. To describe these assumptions more formally, we require some further notation to define the relevant pairs of units. The n pairs may be represented by the sets

$$\{\pi(2j-1), \pi(2j)\} \text{ for } j = 1, \dots, n,$$

where $\pi = \pi_n(X^{(n)})$ is a permutation of $2n$ elements. Because of its possible dependence on $X^{(n)}$, π encompasses a broad variety of different ways of pairing the $2n$ units according to the observed, baseline covariates $X^{(n)}$. Given such a π, we assume that treatment status is assigned as described in the following assumption:

Assumption 2.2. Treatment status is assigned so that $(Y^{(n)}(1), Y^{(n)}(0), W^{(n)}) \perp\!\!\!\perp D^{(n)}|X^{(n)}$ and, conditional on $X^{(n)}$, $(D_{\pi(2j-1)}, D_{\pi(2j)})$, $j = 1, \dots, n$ are i.i.d. and each uniformly distributed over the values in $\{(0,1),(1,0)\}$.

Following Bai et al. (2022), our analysis will additionally require some discipline on how pairs are formed. Let $\|\cdot\|_2$ denote the Euclidean norm. We will require that units in each pair are "close" in the sense described by the following assumption:

Assumption 2.3. The pairs used in determining treatment status satisfy

$$\frac{1}{n} \sum_{1 \le j \le n} \|X_{\pi(2j)} - X_{\pi(2j-1)}\|_2^r \xrightarrow{P} 0$$

for $r \in \{1,2\}$.

It will at times be convenient to require further that units in consecutive pairs are also "close" in terms of their baseline covariates. One may view this requirement, which is formalized in the following assumption, as "pairing the pairs" so that they are "close" in terms of their baseline covariates.

Assumption 2.4. The pairs used in determining treatment status satisfy

$$\frac{1}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} \|X_{\pi(4j-k)} - X_{\pi(4j-\ell)}\|_2^2 \xrightarrow{P} 0$$

for any $k \in \{2,3\}$ and $\ell \in \{0,1\}$.

Bai et al. (2022) provide results to facilitate constructing pairs satisfying Assumptions 2.3–2.4 under weak assumptions on Q. In particular, given pairs satisfying Assumption 2.3, it is frequently possible to "re-order" them so that Assumption 2.4 is satisfied. See Theorem 4.3 in Bai et al. (2022) for further details. As in Bai et al. (2022), we highlight the fact that Assumption 2.4 will only be used to enable consistent estimation of relevant variances.

Remark 2.1. Under this setup, Bai et al. (2022) consider the unadjusted difference-in-means estimator

$$\hat{\Delta}_n^{\mathrm{unadj}} = \frac{1}{n} \sum_{1 \le i \le 2n} D_i Y_i - \frac{1}{n} \sum_{1 \le i \le 2n} (1 - D_i) Y_i \tag{3}$$

and show that it is consistent and asymptotically normal with limiting variance

$$\sigma^2_{\mathrm{unadj}}(Q) = \frac{1}{2} \mathrm{Var}[E[Y_i(1) - Y_i(0)|X_i]] + E[\mathrm{Var}[Y_i(1)|X_i]] + E[\mathrm{Var}[Y_i(0)|X_i]].$$

We note that $\hat{\Delta}_n^{\mathrm{unadj}}$ is the unadjusted estimator because it does not use information in W_i in either the design or analysis stage. If both X_i and W_i are used to form pairs in the "matched pairs" design, then the difference-in-means estimator, which we refer to as $\hat{\Delta}_n^{\mathrm{ideal}}$, has limiting variance

$$\sigma^2_{\mathrm{ideal}}(Q) = \frac{1}{2} \mathrm{Var}[E[Y_i(1) - Y_i(0)|X_i, W_i]] + E[\mathrm{Var}[Y_i(1)|X_i, W_i]] + E[\mathrm{Var}[Y_i(0)|X_i, W_i]].$$

In this case, $\hat{\Delta}_n^{\text{ideal}}$ achieves the efficiency bound derived by Armstrong (2022), and we can see that

$$\sigma^2_{\text{unadj}}(Q) - \sigma^2_{\text{ideal}}(Q) = \frac{1}{2}E[\text{Var}[E[Y_i(1) + Y_i(0)|X_i, W_i]|X_i]] \geq 0.$$

For related results for parameters other than the average treatment effect, see Bai et al. (2023a). We note, however, that it is not always practical to form pairs using both X_i and W_i for two reasons. First, the covariate W_i may only be collected along with the outcome variable and therefore may not be available at the design stage. Second, the quality of pairing decreases with the dimension of matching variables. Indeed, it is common in practice to match on some but not all baseline covariates. Such considerations motivate our analysis below. ■

3. Main results

To accommodate various forms of covariate-adjusted estimators of $\Delta(Q)$ in a single framework, it is useful to note that it follows from Assumption 2.2 that for any $d \in \{0, 1\}$ and any function $m_{d,n} : \mathbf{R}^{k_x} \times \mathbf{R}^{k_w} \to \mathbf{R}$ such that $E[|m_{d,n}(X_i, W_i)|] < \infty$,

$$E\left[2I\{D_i = d\}(Y_i - m_{d,n}(X_i, W_i)) + m_{d,n}(X_i, W_i)\right] = E[Y_i(d)]. \tag{4}$$

We note that (4) is just the augmented inverse propensity score weighted moment for $E[Y_i(d)]$ in which the propensity score is $1/2$ and the conditional mean model is $m_{d,n}(X_i, W_i)$. Such a moment is also "doubly robust". As the propensity score for the "matched pairs" design is exactly one half, we do not require the conditional mean model to be correctly specified, i.e., $m_{d,n}(X_i, W_i) = E[Y_i(d)|X_i, W_i]$. See, for instance, Robins et al. (1995). Intuitively, $m_{d,n}$ is the "working model" which researchers use to estimate $E[Y_i(d)|X_i, W_i]$, and can be arbitrarily misspecified because of (4). Although $m_{d,n}$ will be identical across $n \geq 1$ for the examples in Section 4, the notation permits $m_{d,n}$ to depend on the sample size n in anticipation of the high-dimensional results in Section 5. Based on the moment condition in (4), our proposed estimator of $\Delta(Q)$ is given by

$$\hat{\Delta}_n = \hat{\mu}_n(1) - \hat{\mu}_n(0), \tag{5}$$

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

$$\hat{\mu}_n(d) = \frac{1}{2n}\sum_{1 \leq i \leq 2n}(2I\{D_i = d\}(Y_i - \hat{m}_{d,n}(X_i, W_i)) + \hat{m}_{d,n}(X_i, W_i)) \tag{6}$$

and $\hat{m}_{d,n}$ is a suitable estimator of the "working model" $m_{d,n}$ in (4).

By some simple algebra, we have[1]

$$\hat{\Delta}_n = \frac{1}{n}\sum_{1 \leq i \leq 2n} D_i\tilde{Y}_i - \frac{1}{n}\sum_{1 \leq i \leq 2n}(1 - D_i)\tilde{Y}_i, \tag{7}$$

where

$$\tilde{Y}_i = Y_i - \frac{1}{2}(\hat{m}_{1,n}(X_i, W_i) + \hat{m}_{0,n}(X_i, W_i)). \tag{8}$$

It means our regression adjusted estimator can be viewed as a difference-in-means estimator, but with the "adjusted" outcome $\tilde{Y}_i$.

We require some new discipline on the behavior of $m_{d,n}$ for $d \in \{0, 1\}$ and $n \geq 1$:

Assumption 3.1. The functions $m_{d,n}$ for $d \in \{0, 1\}$ and $n \geq 1$ satisfy

(a) For $d \in \{0, 1\}$,

$$\liminf_{n \to \infty} E\left\{\text{Var}\left[Y_i(d) - \frac{1}{2}(m_{1,n}(X_i, W_i) + m_{0,n}(X_i, W_i))\middle|X_i\right]\right\} > 0.$$

(b) For $d \in \{0, 1\}$,

$$\lim_{\lambda \to \infty}\limsup_{n \to \infty} E[m^2_{d,n}(X_i, W_i)I\{|m_{d,n}(X_i, W_i)| > \lambda\}] = 0.$$

(c) $E[m_{d,n}(X_i, W_i)|X_i = x]$, $E[m^2_{d,n}(X_i, W_i)|X_i = x]$, $E[m_{d,n}(X_i, W_i)Y_i(d)|X_i = x]$ for $d \in \{0, 1\}$, and $E[m_{1,n}(X_i, W_i)m_{0,n}(X_i, W_i)|X_i = x]$ are Lipschitz uniformly over $n \geq 1$.

Assumption 3.1(a) is an assumption to rule out degenerate situations. Assumption 3.1(b) is a mild uniform integrability assumption on the "working models". If $m_{d,n}(\cdot) \equiv m_d(\cdot)$ for $d \in \{0, 1\}$, then it is satisfied as long as $E[m^2_d(X_i, W_i)] < \infty$. Assumption 3.1(c) ensures that units that are "close" in terms of the observed covariates are also "close" in terms of potential outcomes, uniformly across $n \geq 1$.

Theorem 3.1 below establishes the limit in distribution of $\hat{\Delta}_n$. We note that the theorem depends on high-level conditions on $m_{d,n}(\cdot)$ and $\hat{m}_{d,n}(\cdot)$. In the sequel, these conditions will be verified in several examples.

[1] We thank the referee for this excellent point.

Theorem 3.1. *Suppose Q satisfies Assumption 2.1, the treatment assignment mechanism satisfies Assumptions 2.2–2.3, and $m_{d,n}(\cdot)$ for $d \in \{0,1\}$ and $n \geq 1$ satisfy Assumption 3.1. Further suppose $\hat{m}_{d,n}(\cdot)$ satisfies*

$$\frac{1}{\sqrt{2n}} \sum_{1 \leq i \leq 2n} (2D_i - 1)(\hat{m}_{d,n}(X_i, W_i) - m_{d,n}(X_i, W_i)) \xrightarrow{P} 0. \tag{9}$$

Then, $\hat{\Delta}_n$ defined in (5) satisfies

$$\frac{\sqrt{n}(\hat{\Delta}_n - \Delta(Q))}{\sigma_n(Q)} \xrightarrow{d} N(0,1), \tag{10}$$

where $\sigma_n^2(Q) = \sigma_{1,n}^2(Q) + \sigma_{2,n}^2(Q) + \sigma_{3,n}^2(Q)$ with

$$\sigma_{1,n}^2(Q) = \frac{1}{2} E[\text{Var}[E[Y_i(1) + Y_i(0)|X_i, W_i] - (m_{1,n}(X_i, W_i) + m_{0,n}(X_i, W_i))|X_i]]$$

$$\sigma_{2,n}^2(Q) = \frac{1}{2} \text{Var}[E[Y_i(1) - Y_i(0)|X_i, W_i]]$$

$$\sigma_{3,n}^2(Q) = E[\text{Var}[Y_i(1)|X_i, W_i]] + E[\text{Var}[Y_i(0)|X_i, W_i]].$$

In order to facilitate the use of Theorem 3.1 for inference about $\Delta(Q)$, we next provide a consistent estimator of $\sigma_n(Q)$. Define

$$\hat{\tau}_n^2 = \frac{1}{n} \sum_{1 \leq j \leq n} (\tilde{Y}_{\pi(2j-1)} - \tilde{Y}_{\pi(2j)})^2$$

$$\hat{\lambda}_n = \frac{2}{n} \sum_{1 \leq j \leq \lfloor \frac{n}{2} \rfloor} (\tilde{Y}_{\pi(4j-3)} - \tilde{Y}_{\pi(4j-2)})(\tilde{Y}_{\pi(4j-1)} - \tilde{Y}_{\pi(4j)})(D_{\pi(4j-3)} - D_{\pi(4j-2)})(D_{\pi(4j-1)} - D_{\pi(4j)}),$$

where $\tilde{Y}_i$ is defined in (8). The variance estimator is given by

$$\hat{\sigma}_n^2 = \hat{\tau}_n^2 - \frac{1}{2}(\hat{\lambda}_n + \hat{\Delta}_n^2). \tag{11}$$

The variance estimator in (11), in particular its component $\hat{\lambda}_n$, is analogous to the "pairs of pairs" variance estimator in Bai et al. (2022). Such a variance estimator has also been used in Abadie and Imbens (2008) in a related setting. Note that it can be shown similarly as in Remark 3.9 of Bai et al. (2022) that $\hat{\sigma}_n^2$ in (11) is nonnegative.

Theorem 3.2 below establishes the consistency of this estimator and its implications for inference about $\Delta(Q)$. In the statement of the theorem, we make use of the following notation: for any scalars a and b, $[a \pm b]$ is understood to be $[a - b, a + b]$.

Theorem 3.2. *Suppose Q satisfies Assumption 2.1, the treatment assignment mechanism satisfies Assumptions 2.2–2.4, and $m_{d,n}(\cdot)$ for $d \in \{0,1\}$ and $n \geq 1$ satisfy Assumption 3.1. Further suppose $\hat{m}_{d,n}(\cdot)$ satisfies (9) and*

$$\frac{1}{2n} \sum_{1 \leq i \leq 2n} (\hat{m}_{d,n}(X_i, W_i) - m_{d,n}(X_i, W_i))^2 \xrightarrow{P} 0. \tag{12}$$

Then,

$$\frac{\hat{\sigma}_n}{\sigma_n(Q)} \xrightarrow{P} 1.$$

Hence, (10) holds with $\hat{\sigma}_n$ in place of $\sigma_n(Q)$. In particular, for any $\alpha \in (0,1)$,

$$P\left\{\Delta(Q) \in \left[\hat{\Delta}_n \pm \hat{\sigma}_n \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]\right\} \to 1 - \alpha,$$

where Φ is the standard normal c.d.f.

Remark 3.1. Based on (7), it is natural to estimate $\sigma_n^2(Q)$ using the usual estimator of the limiting variance of the difference-in-means estimator, i.e.,

$$\hat{\sigma}_{\text{diff},n}^2 = \frac{1}{n} \sum_{1 \leq i \leq 2n} D_i \left(\tilde{Y}_i - \left(\frac{1}{n} \sum_{1 \leq i \leq 2n} D_i \tilde{Y}_i\right)\right)^2 + \frac{1}{n} \sum_{1 \leq i \leq 2n} (1 - D_i) \left(\tilde{Y}_i - \left(\frac{1}{n} \sum_{1 \leq i \leq 2n} (1 - D_i) \tilde{Y}_i\right)\right)^2.$$

However, it can be shown that $\hat{\sigma}_{\text{diff},n}^2 = \sigma_{\text{diff},n}^2(Q) + o_P(1)$, where

$$\sigma_{\text{diff},n}^2(Q) = \text{Var}\left[Y_i(1) - \frac{1}{2}(m_{1,n}(X_i, W_i) + m_{0,n}(X_i, W_i))\right] + \text{Var}\left[Y_i(0) - \frac{1}{2}(m_{1,n}(X_i, W_i) + m_{0,n}(X_i, W_i))\right].$$

Furthermore,

$$\sigma_{\text{diff},n}^2(Q) - \sigma_n^2(Q) = \frac{1}{2} \text{Var}\left[E[Y_i(1) + Y_i(0) - (m_{1,n}(X_i, W_i) + m_{0,n}(X_i, W_i))|X_i]\right] \geq 0,$$

where the inequality is strict unless

$$E[Y_i(1) + Y_i(0) - (m_{1,n}(X_i, W_i) + m_{0,n}(X_i, W_i))|X_i] = E[Y_i(1) + Y_i(0) - (m_{1,n}(X_i, W_i) + m_{0,n}(X_i, W_i))]$$

with probability one. In this sense, the usual estimator of the limiting variance of the difference-in-means estimator is conservative. ■

Remark 3.2. An important and immediate implication of Theorem 3.1 is that $\sigma_n^2(Q)$ is minimized when

$$\begin{aligned}&E[Y_i(0)+Y_i(1)|X_i,W_i]-E[Y_i(0)+Y_i(1)|X_i]=\\&m_{0,n}(X_i,W_i)+m_{1,n}(X_i,W_i)-E[m_{0,n}(X_i,W_i)+m_{1,n}(X_i,W_i)|X_i]\end{aligned}$$

with probability one. In other words, the "working model" for $E[Y_i(0)+Y_i(1)|X_i,W_i]$ given by $m_{0,n}(X_i,W_i)+m_{1,n}(X_i,W_i)$, need only be correct "on average" over the variables that are not used in determining the pairs. For such a choice of $m_{0,n}(X_i,W_i)$ and $m_{1,n}(X_i,W_i)$, $\sigma_n^2(Q)$ in Theorem 3.1 becomes simply

$$\frac{1}{2}\operatorname{Var}[E[Y_i(1)-Y_i(0)\Big|X_i,W_i]]+E[\operatorname{Var}[Y_i(1)|X_i,W_i]]+E[\operatorname{Var}[Y_i(0)|X_i,W_i]],$$

which agrees with the variance obtained in Bai et al. (2022) when both X_i and W_i are used in determining the pairs. Such a variance also achieves the efficiency bound derived by Armstrong (2022). ■

Remark 3.3. Following Bai et al. (2023b), it is straightforward to extend the analysis in this paper to the case with multiple treatment arms and where treatment status is determined using a "matched tuples" design, but we do not pursue this further in this paper. ■

Remark 3.4. Following Bai et al. (2022), we conjecture it is possible to establish the validity of a randomization test based on the test statistic studentized by a randomized version of (11). We emphasize that the validity of the randomization test depends crucially on the choice of studentization in the test statistic. See, for instance, Remark 3.16 in Bai et al. (2022). Such tests have been studied in finite-population settings with covariate adjustments by Zhao and Ding (2021). We leave a detailed analysis of randomization tests for future work. ■

4. Linear adjustments

In this section, we consider linearly covariate-adjusted estimators of $\Delta(Q)$ based on a set of regressors generated by $X_i \in \mathbf{R}^{k_x}$ and $W_i \in \mathbf{R}^{k_w}$. To this end, define $\psi_i = \psi(X_i, W_i)$, where $\psi : \mathbf{R}^{k_x} \times \mathbf{R}^{k_w} \to \mathbf{R}^p$. We impose the following assumptions on the function ψ:

Assumption 4.1. The function ψ is such that

(a) no component of ψ is constant and $E[\operatorname{Var}[\psi_i|X_i]]$ is non-singular.
(b) $\operatorname{Var}[\psi_i] < \infty$.
(c) $E[\psi_i|X_i=x]$, $E[\psi_i\psi_i'|X_i=x]$, and $E[\psi_i Y_i(d)|X_i=x]$ for $d \in \{0,1\}$ are Lipschitz.

Assumption 4.1 is analogous to Assumption 2.1. Note, in particular, that Assumption 4.1(a) rules out situations where ψ_i is a function of X_i only. See Remark 4.3 for a discussion of the behavior of the covariate-adjusted estimators in such situations.

4.1. Linear adjustments without pair fixed effects

Consider the following linear regression model:

$$Y_i = \alpha + \Delta D_i + \psi_i'\beta + \epsilon_i. \tag{13}$$

Let $\hat\alpha_n^{\text{naive}}$, $\hat\Delta_n^{\text{naive}}$, and $\hat\beta_n^{\text{naive}}$ denote the OLS estimators of α, Δ, and β in (13). We call these estimators naïve because the corresponding regression adjustment is subject to Freedman's critique and can lead to an adjusted estimator that is less efficient than the simple difference-in-means estimator $\hat\Delta_n^{\text{unadj}}$.

It follows from a direct calculation that

$$\hat\Delta_n^{\text{naive}} = \frac{1}{n}\sum_{1\le i\le 2n}(Y_i - \psi_i'\hat\beta_n^{\text{naive}})(2D_i-1).$$

Therefore, $\hat\Delta_n^{\text{naive}}$ satisfies (5)–(6) with

$$\hat m_{d,n}(X_i,W_i) = \psi_i'\hat\beta_n^{\text{naive}}.$$

Theorem 4.1 establishes (9) and (12) for a suitable choice of $m_{d,n}(X_i,W_i)$ for $d \in \{0,1\}$ and, as a result, the limiting distribution of $\hat\Delta_n^{\text{naive}}$ and the validity of the variance estimator.

Theorem 4.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.3. Further suppose ψ satisfies Assumption 4.1. Then, as $n \to \infty$,*

$$\hat\beta_n^{\text{naive}} \xrightarrow{P} \beta^{\text{naive}} = \operatorname{Var}[\psi_i]^{-1}\operatorname{Cov}[\psi_i, Y_i(1)+Y_i(0)].$$

Moreover, (9), (12), and Assumption 3.1 are satisfied with

$$m_{d,n}(X_i, W_i) = \psi_i' \beta^{\text{naive}}$$

for $d \in \{0, 1\}$ and $n \geq 1$.

Remark 4.1. Freedman (2008) studies regression adjustment based on (13) when treatment is assigned by complete randomization instead of a "matched pairs" design. In such settings, Lin (2013) proposes adjustment based on the following linear regression model:

$$Y_i = \alpha + \Delta D_i + (\psi_i - \bar{\psi}_n)'\gamma + D_i(\psi_i - \bar{\psi}_n)'\eta + \epsilon_i, \tag{14}$$

where

$$\bar{\psi}_n = \frac{1}{2n} \sum_{1 \leq i \leq 2n} \psi_i .$$

Let $\hat{\alpha}_n^{\text{int}}, \hat{\Delta}_n^{\text{int}}, \hat{\gamma}_n^{\text{int}}, \hat{\eta}_n^{\text{int}}$ denote the OLS estimators for $\alpha, \Delta, \gamma, \eta$ in (14). It is straightforward to show $\hat{\Delta}_n^{\text{int}}$ satisfies (5)–(6) with

$$\hat{m}_{1,n}(X_i, W_i) = (\psi_i - \hat{\mu}_{\psi,n}(1))'(\hat{\gamma}_n^{\text{int}} + \hat{\eta}_n^{\text{int}})$$
$$\hat{m}_{0,n}(X_i, W_i) = (\psi_i - \hat{\mu}_{\psi,n}(0))'\hat{\gamma}_n^{\text{int}},$$

where

$$\hat{\mu}_{\psi,n}(d) = \frac{1}{n} \sum_{1 \leq i \leq 2n} I\{D_i = d\}\psi_i .$$

It can be shown using similar arguments to those used to establish Theorem 4.1 that (9) and Assumption 3.1 are satisfied with

$$m_{d,n}(X_i, W_i) = (\psi_i - E[\psi_i])' \operatorname{Var}[\psi_i]^{-1} \operatorname{Cov}[\psi_i, Y_i(d)]$$

for $d \in \{0, 1\}$ and $n \geq 1$. It thus follows by inspecting the expression for $\sigma_n^2(Q)$ in Theorem 3.1 that the limiting variance of $\hat{\Delta}_n^{\text{int}}$ is the same as that of $\hat{\Delta}_n^{\text{naive}}$ based on (13). ■

Remark 4.2. Note that $\hat{\Delta}_n^{\text{naive}}$ is the ordinary least squares estimator for Δ in the linear regression

$$Y_i - \psi_i' \hat{\beta}_n^{\text{naive}} = \alpha + \Delta D_i + \epsilon_i .$$

Furthermore, Theorem 4.1 implies that its limiting variance is $\sigma_{\text{naive}}^2(Q)$, given by $\sigma_n^2(Q)$ in Theorem 3.1 with $m_d(X_i, W_i) = \psi_i' \beta^{\text{naive}}$. The usual heteroskedasticity-robust estimator of the limiting variance of $\hat{\Delta}_n^{\text{naive}}$ is, however, simply $\hat{\sigma}_{\text{diff},n}^2$ defined in Remark 3.1 with $\hat{m}_{d,n}(X_i, W_i) = \psi_i' \hat{\beta}_n^{\text{naive}}$. It thus follows that $\hat{\sigma}_{\text{diff},n}^2$ is conservative for $\sigma_{\text{naive}}^2(Q)$ in the sense described therein. It is, of course, possible to estimate $\sigma_{\text{naive}}^2(Q)$ consistently using $\hat{\sigma}_n^2$ proposed in Theorem 3.2 with $\hat{m}_{d,n}(W_i, X_i) = \psi_i' \hat{\beta}_n^{\text{naive}}$, but $\sigma_{\text{naive}}^2(Q)$ is not guaranteed to be smaller than the limiting variance of the unadjusted estimator, i.e., $\sigma_{\text{unadj}}^2(Q)$, so the linear adjustment without pair fixed effects can harm the precision of the estimator. Evidence of this phenomenon is provided in our simulations in Section 6. ■

4.2. Linear adjustments with pair fixed effects

Remark 4.1 implies that in "matched pairs" designs, including interaction terms in the linear regression does not lead to an estimator with lower limiting variance than the one based on the linear regression without interaction terms. It is therefore interesting to study whether there exists a linearly covariate-adjusted estimator with lower limiting variance than the ones based on (13) and (14) as well as the difference-in-means estimator. To that end, consider instead the following linear regression model:

$$Y_i = \Delta D_i + \psi_i' \beta + \sum_{1 \leq j \leq n} \theta_j I\{i \in \{\pi(2j-1), \pi(2j)\}\} + \epsilon_i . \tag{15}$$

Let $\hat{\Delta}_n^{\text{pfe}}$, $\hat{\beta}_n^{\text{pfe}}$, and $\hat{\gamma}_{j,n}$, $1 \leq j \leq n$ denote the OLS estimators of Δ, β, θ_j, $1 \leq j \leq n$ in (15), where "pfe" stands for pair fixed effect. It follows from the Frisch-Waugh-Lovell theorem that

$$\hat{\Delta}_n^{\text{pfe}} = \frac{1}{n} \sum_{1 \leq i \leq 2n} (Y_i - \psi_i' \hat{\beta}_n^{\text{pfe}})(2D_i - 1).$$

Therefore, $\hat{\Delta}_n^{\text{pfe}}$ satisfies (5)–(6) with

$$\hat{m}_{d,n}(X_i, W_i) = \psi_i' \hat{\beta}_n^{\text{pfe}} .$$

Theorem 4.2 establishes (9) and (12) for a suitable choice of $m_{d,n}(X_i, W_i), d \in \{0, 1\}$ and, as a result, the limiting distribution of $\hat{\Delta}_n^{\text{pfe}}$ and the validity of the variance estimator.

Theorem 4.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.3. Then, as $n \to \infty$,*

$$\hat{\beta}_n^{\mathrm{pfe}} \xrightarrow{P} \beta^{\mathrm{pfe}} = (2E[\mathrm{Var}[\psi_i|X_i]])^{-1}E[\mathrm{Cov}[\psi_i, Y_i(1)+Y_i(0)|X_i]].$$

Moreover, (9), (12), and Assumption 3.1 are satisfied with

$$m_{d,n}(X_i, W_i) = \psi_i'\beta^{\mathrm{pfe}}$$

for $d \in \{0,1\}$ and $n \geq 1$.

Remark 4.3. When ψ is restricted to be a function of X_i only, $\hat{\Delta}_n^{\mathrm{pfe}}$ coincides to first order with the unadjusted difference-in-means estimator $\hat{\Delta}_n^{\mathrm{unadj}}$ defined in (3). To see this, suppose further that ψ is Lipschitz and that $\mathrm{Var}[Y_i(d)|X_i = x], d \in \{0,1\}$ are bounded. The proof of Theorem 4.2 reveals that $\hat{\Delta}_n^{\mathrm{pfe}}$ and $\hat{\beta}_n^{\mathrm{pfe}}$ coincide with the OLS estimators of the intercept and slope parameters in a linear regression of $(Y_{\pi(2j)} - Y_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)})$ on a constant and $(\psi_{\pi(2j)} - \psi_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)})$. Using this observation, it follows by arguing as in Section S.1.1 of Bai et al. (2022) that

$$\sqrt{n}(\hat{\Delta}_n^{\mathrm{pfe}} - \Delta(Q)) = \sqrt{n}(\hat{\Delta}_n^{\mathrm{unadj}} - \Delta(Q)) + o_P(1).$$

See also Remark 3.8 of Bai et al. (2022). ■

Remark 4.4. The regression-adjusted estimators studied in this section are also examined in Imbens and Rubin (2015) and Fogarty (2018) in a finite population setting. Their regularity conditions, however, only permit regression adjustment using W_i. By contrast, we employ a super-population framework and permit regression adjustment using ψ_i, which is a function of both W_i and X_i. ■

Remark 4.5. Note in the expression of $\sigma_n^2(Q)$ in Theorem 3.1 only depends on $m_{d,n}(X_i, W_i), d \in \{0,1\}$ through $\sigma_{1,n}^2(Q)$. With this in mind, consider the class of all linearly covariate-adjusted estimators based on ψ_i, i.e., $m_{d,n}(X_i, W_i) = \psi_i'\beta(d)$. For this specification of $m_{d,n}(X_i, W_i), d \in \{0,1\}$,

$$\sigma_{1,n}^2(Q) = E[(E[Y_i(1)+Y_i(0)|X_i, W_i] - E[Y_i(1)+Y_i(0)|X_i] - (\psi_i - E[\psi_i|X_i])'(\beta(1)+\beta(0)))^2].$$

It follows that among all such linear adjustments, $\sigma_n^2(Q)$ in (10) is minimized when

$$\beta(1) + \beta(0) = 2\beta^{\mathrm{pfe}}.$$

This observation implies that the linear adjustment with pair fixed effects, i.e., $\hat{\Delta}_n^{\mathrm{pfe}}$, yields the optimal linear adjustment in the sense of minimizing $\sigma_n^2(Q)$. Its limiting variance is, in particular, weakly smaller than the limiting variance of the unadjusted difference-in-means estimator defined in (3). On the other hand, the covariate-adjusted estimators based on (13) or (14), i.e., $\hat{\Delta}_n^{\mathrm{naive}}$ and $\hat{\Delta}_n^{\mathrm{int}}$, are in general not optimal among all linearly covariate-adjusted estimators based on ψ_i. In fact, the limiting variances of these two estimators may even be larger than that of the unadjusted difference-in-means estimator. ■

Remark 4.6. "Matched pairs" design is essentially a non-parametric way to adjust for X_i. Projecting ψ_i on the pair dummies in (15) is equivalent to pair-wise demeaning, which effectively removes $E[\psi_i|X_i]$ from ψ_i. This is key to the optimality of $\hat{\Delta}_n^{\mathrm{pfe}}$ over all linearly adjusted estimators. Following the same logic, we expect that by replacing the pair dummies with sieve bases of X_i in (15), the linear regression can still effectively remove $E[\psi_i|X_i]$ from ψ_i so that the new adjusted estimator is asymptotically equivalent to $\hat{\Delta}_n^{\mathrm{pfe}}$, and thus, linearly optimal. ■

Remark 4.7. Remark 4.2 also applies here with β^{naive} replaced by β^{pfe}. Even though $\hat{\Delta}_n^{\mathrm{pfe}}$ can be computed via OLS estimation of (15), we emphasize that the usual heteroskedascity-robust standard errors that naïvely treats the data (including treatment status) as if it were i.i.d. need not be consistent for the limiting variance derived in our analysis. ■

Remark 4.8. One can also consider the estimator based on the following linear regression model:

$$Y_i = \Delta D_i + (\psi_i - \bar{\psi}_n)'\gamma + D_i(\psi_i - \hat{\mu}_{\psi,n}(1))'\eta + \sum_{1\leq j\leq n} \theta_j I\{i \in \{\pi(2j-1), \pi(2j)\}\} + \epsilon_i. \tag{16}$$

Let $\hat{\Delta}_n^{\mathrm{int-pfe}}, \hat{\gamma}_n^{\mathrm{int-pfe}}, \hat{\eta}_n^{\mathrm{int-pfe}}$ denote the OLS estimators for Δ, γ, η in (16). It is straightforward to show $\hat{\Delta}_n^{\mathrm{int-pfe}}$ satisfies (5)–(6) with

$$\hat{m}_{1,n}(X_i, W_i) = (\psi_i - \hat{\mu}_{\psi,n}(1))'\hat{\eta}_n^{\mathrm{int-pfe}}$$
$$\hat{m}_{0,n}(X_i, W_i) = (\psi_i - \hat{\mu}_{\psi,n}(0))'(\hat{\eta}_n^{\mathrm{int-pfe}} - \hat{\gamma}_n^{\mathrm{int-pfe}}).$$

Following similar arguments to those used in the proof of Theorem 4.1, we can establish that (9) and Assumption 3.1 are satisfied with

$$m_{1,n}(X_i, W_i) = (\psi_i - E[\psi_i])'\eta^{\mathrm{int-pfe}}$$
$$m_{0,n}(X_i, W_i) = (\psi_i - E[\psi_i])'(\eta^{\mathrm{int-pfe}} - \gamma^{\mathrm{int-pfe}}),$$

where

$$\gamma^{\text{int-pfe}} = (E[\text{Var}[\psi_i|X_i]])^{-1} E[\text{Cov}[\psi_i, Y_i(1) - Y_i(0)|X_i]],$$
$$\eta^{\text{int-pfe}} = (E[\text{Var}[\psi_i|X_i]])^{-1} E[\text{Cov}[\psi_i, Y_i(1)|X_i]].$$

Because $2\eta^{\text{int-pfe}} - \gamma^{\text{int-pfe}} = 2\beta^{\text{pfe}}$, it follows from Remark 4.5 that the limiting variance of $\hat{\Delta}_n^{\text{int-pfe}}$ is identical to the limiting variance of $\hat{\Delta}_n^{\text{pfe}}$. ■

Remark 4.9. Wu and Gagnon-Bartsch (2021) consider the covariate adjustment for paired experiments under the design-based framework, where the covariates are treated as deterministic, and thus, the cross-sectional dependence between units in the same pair due to the closeness of their covariates is not counted in their analysis. We differ from them by considering the sampling-based framework in which the covariates are treated as random and the pairs are formed by matching, and thus, have an impact on statistical inference. Under their framework, Wu and Gagnon-Bartsch (2021) point out that covariate adjustments may have a positive or negative effect on the estimation accuracy depending on how they are estimated. This is consistent with our findings in this section. Specifically, we show that when the regression adjustments are estimated by linear regression with pair fixed effects, the resulting ATE estimator is guaranteed to weakly improve upon the difference-in-means estimator in terms of efficiency. However, this improvement is not guaranteed if the adjustments are estimated without pair fixed effects. ■

Remark 4.10. If we choose ψ_i as a set of sieve basis functions with increasing dimension, then under suitable regularity conditions, the linear adjustments both with and without pair fixed effects achieve the same limiting variance as $\hat{\Delta}_n^{\text{ideal}}$, and thus, the efficiency bound. In fact, if ψ_i contains sieve bases, then the linear adjustment without pair fixed effects can approximate the true specification $E[Y_i(1) + Y_i(0)|X_i, W_i]$ in the sense that $E[Y_i(1) + Y_i(0)|X_i, W_i] = \psi_i'\beta^{\text{naive}} + R_i$ and $E[R_i^2] = o(1)$. This property implies $\sigma_{1,n}^2(Q)$ in Theorem 3.1 equals zero. Similarly, the linear adjustment with pair fixed effects can approximate the true specification $E[Y_i(1)+Y_i(0)|X_i, W_i] - E[Y_i(1)+Y_i(0)|X_i]$ in the sense that $E[Y_i(1)+Y_i(0)|X_i, W_i] - E[Y_i(1)+Y_i(0)|X_i] = \tilde{\psi}_i'\beta^{\text{naive}} + \tilde{R}_i$ and $E[\tilde{R}_i^2] = o(1)$. This property again implies $\sigma_{1,n}^2(Q)$ in Theorem 3.1 equals zero. Therefore, in both cases, the adjusted estimator achieves the minimum variance. In the next section, we consider ℓ_1-regularized adjustments which may be viewed as providing a way to choose the relevant sieve bases in a data-driven manner. ■

5. Regularized adjustments

In this section, we study covariate adjustments based on the ℓ_1-regularized linear regression. Such settings can arise if the covariates W_i are high-dimensional or if the dimension of W_i is fixed but the regressors include many sieve basis functions of X_i and W_i. To accommodate situations where the dimension of W_i increases with n, we add a subscript and denote it by $W_{n,i}$ instead. Let $k_{w,n}$ denote the dimension of $W_{n,i}$. For $n \geq 1$, let $\psi_{n,i} = \psi_n(X_i, W_{n,i})$, where $\psi_n : \mathbf{R}^{k_x} \times \mathbf{R}^{k_{w,n}} \to \mathbf{R}^{p_n}$ and p_n will be permitted below to be possibly much larger than n.

In what follows, we consider a two-step method in the spirit of Cohen and Fogarty (2024). In the first step, an intermediate estimator, $\hat{\Delta}_n^{\text{r}}$, is obtained using (5) with a "working model" obtained through a ℓ_1-regularized linear regression adjustments $m_{d,n}(X_i, W_i)$ for $d \in \{0, 1\}$. As explained further below in Theorem 5.1, when $m_{d,n}(X_i, W_i)$ is approximately correctly specified, such an estimator is optimal in the sense that it minimizes the limiting variance in Theorem 3.1. When this is not the case, however, for reasons analogous to those put forward in Remark 4.2, it needs not to have a limiting variance weakly smaller than the unadjusted difference-in-means estimator. In a second step, we therefore consider an estimator by refitting a version of (15) in which the covariates ψ_i are replaced by the regularized estimates of $m_{d,n}(X_i, W_i)$ for $d \in \{0, 1\}$. The resulting estimator, $\hat{\Delta}_n^{\text{refit}}$, has the limiting variance weakly smaller than that of the intermediate estimator and thus remains optimal under approximately correct specification in the same sense. Moreover, it has limiting variance weakly smaller than the unadjusted difference-in-means estimator. Wager et al. (2016) also consider high-dimensional regression adjustments in randomized experiments using LASSO. We differ from their work by considering the "matched pairs" design, and more importantly, discussing when and how regularized adjustments can improve estimation efficiency upon the difference-in-means estimator.

Before proceeding, we introduce some additional notation that will be required in our formal description of the methods. We denote by $\psi_{n,i,l}$ the lth components of $\psi_{n,i}$. For a vector $a \in \mathbf{R}^k$ and $0 \leq p \leq \infty$, recall that

$$\|a\|_p = \Big(\sum_{1 \leq l \leq k} |a_l|^p \Big)^{1/p},$$

where it is understood that $\|a\|_0 = \sum_{1 \leq l \leq k} I\{a_k \neq 0\}$ and $\|a\|_\infty = \sup_{1 \leq l \leq k} |a_l|$. Using this notation, we further define

$$\Xi_n = \sup_{(x,w) \times \text{supp}(X_i) \times \text{supp}(W_i)} \|\psi_{n,i}(x, w)\|_\infty.$$

For $d \in \{0, 1\}$, define

$$(\hat{\alpha}_{d,n}^{\text{r}}, \hat{\beta}_{d,n}^{\text{r}}) \in \operatorname*{argmin}_{a \in \mathbf{R}, b \in \mathbf{R}^{p_n}} \frac{1}{n} \sum_{1 \leq i \leq 2n : D_i = d} (Y_i - a - \psi_{n,i}' b)^2 + \lambda_{d,n}^{\text{r}} \|\hat{\Omega}_n(d) b\|_1, \tag{17}$$

where $\lambda_{d,n}^{\text{r}}$ is a penalty parameter that will be disciplined by the assumptions below, $\hat{\Omega}_n(d) = \text{diag}(\hat{\omega}_1(d), \ldots, \hat{\omega}_{p_n}(d))$ is a diagonal matrix, and $\hat{\omega}_{n,l}(d)$ is the penalty loading for the lth regressor. Let $\hat{\Delta}_n^{\text{r}}$ denote the estimator in (5) with $\hat{m}_{d,n}(X_i, W_{n,i}) = \psi_{n,i}' \hat{\beta}_{d,n}^{\text{r}}$ for $d \in \{0, 1\}$.

We now proceed with the statement of our assumptions. The first assumption collects a variety of moment conditions that will be used in our formal analysis:

Assumption 5.1.

(a) There exist nonrandom quantities $(\alpha^{\mathrm{r}}_{d,n}, \beta^{\mathrm{r}}_{d,n})$ such that with $\epsilon^{\mathrm{r}}_{n,i}(d)$ defined as

$$\epsilon^{\mathrm{r}}_{n,i}(d) = Y_i(d) - \alpha^{\mathrm{r}}_{d,n} - \psi'_{n,i}\beta^{\mathrm{r}}_{d,n},$$

we have

$$\|\Omega_n(d)^{-1} E[\psi_{n,i}\epsilon^{\mathrm{r}}_{n,i}(d)]\|_\infty + |E[\epsilon^{\mathrm{r}}_{n,i}(d)]| = o\left(\lambda^{\mathrm{r}}_{d,n}\right), \tag{18}$$

where $\Omega_n(d) = \mathrm{diag}(\omega_{n,1}(d), \dots, \omega_{n,p_n}(d))$ and $\omega^2_{n,l}(d) = \mathrm{Var}[\psi_{n,i,l}\epsilon^{\mathrm{r}}_{n,i}(d)]$.

(b) For some $q > 2$ and constant C_1,

$$\sup_{n\ge1} \max_{1\le l\le p_n} E[|\psi^q_{n,i,l}||X_i] \le C_1,$$
$$\sup_{n\ge1} |\psi'_{n,i}\beta^{\mathrm{r-pd}}_{d,n}| \le C_1,$$
$$\sup_{n\ge1} |E[Y_i(a)|X_i, W_{n,i}]| \le C_1,$$

with probability one.

(c) For some $\underline{c}$ and $\bar{c}$, we require that

$$0 < \underline{c} \le \liminf_{n\to\infty} \min_{1\le l\le p_n} \hat{\omega}_{n,l}(d)/\omega_{n,l}(d) \le \limsup_{n\to\infty} \max_{1\le l\le p_n} \hat{\omega}_{n,l}(d)/\omega_{n,l}(d) \le \bar{c} < \infty. \tag{19}$$

(d) For some c_0, $\underline{\sigma}$, $\bar{\sigma}$, the following statements hold with probability one:

$$0 < \underline{\sigma}^2 \le \liminf_{n\to\infty} \min_{d\in\{0,1\}, 1\le l\le p_n} \omega^2_{n,l}(d) \le \limsup_{n\to\infty} \max_{d\in\{0,1\},1\le l\le p_n} \omega^2_{n,l}(d) \le \bar{\sigma}^2 < \infty,$$

$$\sup_{n\ge1} \max_{d\in\{0,1\}} E[(\psi'_{n,i}\beta^{\mathrm{r}}_{d,n})^2] \le c_0 < \infty,$$
$$\max_{d\in\{0,1\},1\le l\le p_n} \frac{1}{2n} \sum_{1\le i\le 2n} E[\epsilon^4_{n,i}(d)|X_i] \le c_0 < \infty,$$
$$\sup_{n\ge1} \max_{d\in\{0,1\}} E[\epsilon^4_{n,i}(d)] \le c_0 < \infty,$$
$$\min_{d\in\{0,1\}} \mathrm{Var}[Y_i(d) - \psi'_{n,i}(\beta^{\mathrm{r}}_{1,n} + \beta^{\mathrm{r}}_{0,n})/2] \ge \underline{\sigma}^2 > 0,$$
$$\min_{1\le l\le p_n} \frac{1}{n} \sum_{1\le i\le 2n} I\{D_i = d\}\, \mathrm{Var}[\psi_{n,i,l}\epsilon_{n,i}(d)|X_i] \ge \underline{\sigma}^2 > 0,$$
$$\min_{1\le l\le p_n} \mathrm{Var}[E[\psi_{n,i,l}\epsilon_{n,i}(d)|X_i]] \ge \underline{\sigma}^2 > 0.$$

Remark 5.1. It is instructive to note that (18) in Assumption 5.1(a) is the subgradient condition for a ℓ_1-penalized regression of the outcome $Y_i(d)$ on $\psi_{n,i}$ when the penalty is of order $o(\lambda^{\mathrm{r}}_n)$. Specifically, if $p_n \ll n$, then this condition holds automatically for the $\beta^{\mathrm{r}}_{d,n}$ equal to the coefficients of a linear projection of $Y_i(d)$ onto $(1, \psi'_{n,i})$. When $p_n \gg n$, but $E[Y_i(d)|X_i, W_i]$ is approximately correctly specified in the sense that the approximation error $R_{n,i}(d) = E[Y_i(d)|X_i, W_i] - \alpha^{\mathrm{r}}_{d,n} - \psi'_{n,i}\beta^{\mathrm{r}}_{d,n}$ is sufficiently small, then (18) also holds. However, the approximately correct specification is not necessary for (18). For example, suppose $W_{n,i} = (W_{n,i,1}, \dots, W_{n,i,p_n})$ is a p_n vector of independent standard normal random variables, $W_{n,i}$ is independent of X_i, $\psi_{n,i} = (X'_i, W'_{n,i})'$, and

$$Y_i(d) = \alpha^{\mathrm{r}}_{d,n} + \psi'_{n,i}\beta^{\mathrm{r}}_{d,n} + \sum_{l=1}^{p_n} \frac{W^2_{n,i,l} - 1}{\sqrt{p_n}} + u_{n,i}(d),$$

where $E(u_{n,i}(d)|X_i, W_{n,i}) = 0$. Then, Assumption 5.1(a) holds with $\epsilon^{\mathrm{r}}_{n,i}(d) = \sum_{l=1}^{p_n} \frac{W^2_{n,i,l}-1}{\sqrt{p_n}} + u_{n,i}(d)$. We can impose a sparse restriction on $\beta^{\mathrm{r}}_{d,n}$ so that it further satisfies Assumption 5.3(b) below. On the other hand, the linear regression adjustment is not approximately correctly specified because $R_{n,i}(d) = E(Y_i(d)|X_i, W_{n,i}) - (\alpha^{\mathrm{r}}_{d,n} + \psi'_{n,i}\beta^{\mathrm{r}}_{d,n}) = \sum_{l=1}^{p_n} \frac{W^2_{n,i,l}-1}{\sqrt{p_n}}$, and we have $ER^2_{n,i}(d) = 2 \nrightarrow 0$. ■

Remark 5.2. Assumption 5.1(b) and 5.1(d) are standard in the high-dimensional estimation literature; see, for instance, Belloni et al. (2017). The last four inequalities in Assumption 5.1(d), in particular, permit us to apply the high-dimensional central limit theorem in Chernozhukov et al. (2017, Theorem 2.1). ■

Remark 5.3. The penalty loadings in Assumption 5.1(c) can be computed by an iterative procedure proposed by Belloni et al. (2017). We provide more detail in Algorithm 5.1 below. We can then verify (19) under "matched pairs" designs following arguments similar to those in Belloni et al. (2017). ■

Our analysis will, as before, also require some discipline on how pairs are formed. For this purpose, Assumption 2.3 will suffice, but we will need an additional Lipshitz-like condition:

Assumption 5.2. For some $L > 0$ and any x_1 and x_2 in the support of X_i, we have

$$|(\Psi(x_1) - \Psi(x_2))'\beta^{\mathrm{r}}_{d,n}| \le L\|x_1 - x_2\|_2.$$

We next specify our restrictions on the penalty parameter $\lambda^{\mathrm{r}}_{d,n}$.

Assumption 5.3.

(a) For some $\ell\ell_n \to \infty$,

$$\lambda^{\mathrm{r}}_{d,n} = \frac{\ell\ell_n}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{0.1}{2\log(n)p_n}\right).$$

(b) $\Xi_n^2(\log p_n)^7/n \to 0$ and $(\ell\ell_n s_n \log p_n)/\sqrt{n} \to 0$, where

$$s_n = \max_{d\in\{0,1\}} \|\beta^{\mathrm{r}}_{d,n}\|_0. \tag{20}$$

We note that Assumption 5.3(b) permits p_n to be much greater than n. It also requires sparsity in the sense that $s_n = o(\sqrt{n})$.

Finally, as is common in the analysis of ℓ_1-penalized regression, we require a "restricted eigenvalue" condition. This assumption permits us to apply Bickel et al. (2009, Lemma 4.1) and establish the error bounds for $|\hat{\alpha}^{\mathrm{r}}_{d,n} - \alpha^{\mathrm{r}}_{d,n}| + \|\hat{\beta}^{\mathrm{r}}_{d,n} - \beta^{\mathrm{r}}_{d,n}\|_1$ and $\frac{1}{n}\sum_{1\le i\le 2n} I\{D_i = d\}\left(\hat{\alpha}^{\mathrm{r}}_{d,n} - \alpha^{\mathrm{r}}_{d,n} + \psi'_{n,i}(\hat{\beta}^{\mathrm{r}}_{d,n} - \beta^{\mathrm{r}}_{d,n})\right)^2$.

Assumption 5.4. For some $\kappa_1 > 0, \kappa_2$ and $\ell_n \to \infty$, the following statements hold with probability approaching one:

$$\inf_{d\in\{0,1\}, v\in\mathbf{R}^{p_n+1}:\|v\|_0\le(s_n+1)\ell_n} (\|v\|_2^2)^{-1}v'\left(\frac{1}{n}\sum_{1\le i\le 2n} I\{D_i = d\}\breve{\psi}_{n,i}\breve{\psi}'_{n,i}\right)v \ge \kappa_1$$

$$\sup_{d\in\{0,1\}, v\in\mathbf{R}^{p_n+1}:\|v\|_0\le(s_n+1)\ell_n} (\|v\|_2^2)^{-1}v'\left(\frac{1}{n}\sum_{1\le i\le 2n} I\{D_i = d\}\breve{\psi}_{n,i}\breve{\psi}'_{n,i}\right)v \le \kappa_2$$

$$\inf_{d\in\{0,1\}, v\in\mathbf{R}^{p_n+1}:\|v\|_0\le(s_n+1)\ell_n} (\|v\|_2^2)^{-1}v'\left(\frac{1}{n}\sum_{1\le i\le 2n} I\{D_i = d\}E[\breve{\psi}_{n,i}\breve{\psi}'_{n,i}|X_i]\right)v \ge \kappa_1$$

$$\sup_{d\in\{0,1\}, v\in\mathbf{R}^{p_n+1}:\|v\|_0\le(s_n+1)\ell_n} (\|v\|_2^2)^{-1}v'\left(\frac{1}{n}\sum_{1\le i\le 2n} I\{D_i = d\}E[\breve{\psi}_{n,i}\breve{\psi}'_{n,i}|X_i]\right)v \le \kappa_2,$$

where $\breve{\psi}_{n,i} = (1, \psi'_{n,i})'$.

Using these assumptions, the following theorem characterizes the behavior of $\hat{\Delta}^{\mathrm{r}}_n$:

Theorem 5.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.3. Further suppose Assumptions 5.1–5.4 hold. Then, (9), (12), and Assumption 3.1 are satisfied with* $\hat{m}_{d,n}(X_i, W_{n,i}) = \hat{\alpha}^{\mathrm{r}}_{d,n} + \psi'_{n,i}\hat{\beta}^{\mathrm{r}}_{d,n}$ *and*

$$m_{d,n}(X_i, W_{n,i}) = \alpha^{\mathrm{r}}_{d,n} + \psi'_{n,i}\beta^{\mathrm{r}}_{d,n}$$

for $d \in \{0,1\}$ *and* $n \ge 1$. *Denote the variance of* $\hat{\Delta}^{\mathrm{r}}_n$ *by* $\sigma_n^{\mathrm{r},2}$. *If the regularized adjustment is approximately correctly specified, i.e.,* $E[Y_i(d)|X_i, W_{n,i}] = \alpha^{\mathrm{r}}_{d,n} + \psi'_{n,i}\beta^{\mathrm{r}}_{d,n} + R_{n,i}(d)$ *and* $\max_{d\in\{0,1\}} E[R^2_{n,i}(d)] = o(1)$, *then* $\sigma_n^{\mathrm{r},2}$ *achieves the minimum variance, i.e.,*

$$\lim_{n\to\infty} \sigma_n^{\mathrm{r},2} = \sigma_2^2(Q) + \sigma_3^2(Q).$$

Remark 5.4. We recommend employing an iterative estimation procedure outlined by Belloni et al. (2017) to estimate $\hat{\beta}^{\mathrm{r}}_{d,n}$, in which the mth step's penalty loadings are estimated based on the $(m-1)$th step's LASSO estimates. Formally, this iterative procedure is described by the following algorithm:

Algorithm 5.1.

Step 0: Set $\hat{\epsilon}^{\mathrm{r},(0)}_{n,i}(d) = Y_i$ if $D_i = d$.

⋮

Step m: Compute $\hat{\omega}^{(m)}_{n,l}(d) = \sqrt{\frac{1}{n}\sum_{1\le i\le 2n} I\{D_i = d\}\psi^2_{n,i,l}(\hat{\epsilon}^{\mathrm{r},(m-1)}_{n,i}(d))^2}$ and compute $(\hat{\alpha}^{\mathrm{r},(m)}_{d,n}, \hat{\beta}^{\mathrm{r},(m)}_{d,n})$ following (17) with $\hat{\omega}^{(m)}_{n,l}$ as the penalty loadings, and $\hat{\epsilon}^{\mathrm{r},(m)}_{n,i}(d) = Y_i - \hat{\alpha}^{\mathrm{r},(m)}_{d,n} - \psi_i'\hat{\beta}^{\mathrm{r},(m)}_{d,n}$ if $D_i = d$.

$\vdots$

Step M: …

Step $M+1$: Set $\hat{\beta}^{\mathrm{r}}_{d,n} = \hat{\beta}^{\mathrm{r},(M)}_{d,n}$.

As suggested by Belloni et al. (2017), we set M to be 15. We note that R package **hdm** has a built-in option for this iterative procedure. For this choice of penalty loadings, arguments similar to those in Belloni et al. (2017) can be used to verify (19) under "matched pairs" designs. ■

Remark 5.5. When the ℓ_1-regularized adjustment is approximately correctly specified, Theorem 5.1 shows $\hat{\Delta}^{\mathrm{r}}_n$ achieves the minimum variance derived in Remark 3.2, and thus, is guaranteed to be weakly more efficient than the difference-in-means estimator ($\hat{\Delta}^{\mathrm{unadj}}_n$). When $W_{n,i}$ is fixed dimensional and $\psi_{n,i}$ consists of sieve basis functions of $(X_i, W_{n,i})$, the approximately correct specification usually holds. Specifically, under regularity conditions such as the smoothness of $E(Y_i(d)|X_i, W_{n,i})$, we can approximate $E(Y_i(d)|X_i, W_{n,i})$ by $\alpha^{\mathrm{r}}_{d,n} + \psi_{n,i}'\beta^{\mathrm{r}}_{d,n}$ and $\beta^{\mathrm{r}}_{d,n}$ is automatically sparse in the sense that $\|\beta^{\mathrm{r}}_{d,n}\|_0 \ll n$. This means our regularized regression adjustment can select relevant sieve bases in nonparametric regression adjustments in a data-driven manner and automatically minimize the limiting variance of the corresponding ATE estimator. ■

Remark 5.6. When the dimension of $\psi_{n,i}$ is ultra-high (i.e., $p_n \gg n$) and the regularized adjustment is not approximately correctly specified, $\hat{\Delta}^{\mathrm{r}}_n$ suffers from Freedman (2008)'s critique that, theoretically, it is possible to be less efficient than $\hat{\Delta}^{\mathrm{unadj}}_n$. To overcome this problem, we consider an additional step in which we treat the regularized adjustments $(\psi_{n,i}'\hat{\beta}^{\mathrm{r}}_{1,n}, \psi_{n,i}'\hat{\beta}^{\mathrm{r}}_{0,n})$ as a two-dimensional covariate and refit a linear regression with pair fixed effects. Such a procedure has also been studied by Cohen and Fogarty (2024) in the setting with low-dimensional covariates and complete randomization. In fact, this strategy can improve upon general initial regression adjustments as long as (9), (12), and Assumption 3.1 are satisfied. ■

Theorem 5.2 below shows the "refit" estimator for the ATE is weakly more efficient than both $\hat{\Delta}^{\mathrm{unadj}}_n$ and $\hat{\Delta}^{\mathrm{r}}_n$. To state the results, define $\Gamma_{n,i} = (\psi_{n,i}'\beta^{\mathrm{r}}_{1,n}, \psi_{n,i}'\beta^{\mathrm{r}}_{0,n})'$, $\hat{\Gamma}_{n,i} = (\psi_{n,i}'\hat{\beta}^{\mathrm{r}}_{1,n}, \psi_{n,i}'\hat{\beta}^{\mathrm{r}}_{0,n})$, and $\hat{\Delta}^{\mathrm{refit}}_n$ as the estimator in (15) with ψ_i replaced by $\hat{\Gamma}_{n,i}$. Note that $\hat{\Delta}^{\mathrm{refit}}_n$ remains numerically the same if we include the intercept $\hat{\alpha}^{\mathrm{r}}_{d,n}$ in the definition of $\hat{\Gamma}_{n,i}$. Following Remark 4.3, $\hat{\Delta}^{\mathrm{refit}}_n$ is the intercept in the linear regression of $(D_{\pi(2j-1)} - D_{\pi(2j-1)})(Y_{\pi(2j-1)} - Y_{\pi(2j)})$ on constant and $(D_{\pi(2j-1)} - D_{\pi(2j-1)})(\hat{\Gamma}_{n,\pi(2j-1)} - \hat{\Gamma}_{n,\pi(2j)})$. Replacing $\hat{\Gamma}_{n,i}$ by $\hat{\Gamma}_{n,i} + (\hat{\alpha}^{\mathrm{r}}_{1,n}, \hat{\alpha}^{\mathrm{r}}_{0,n})'$ will not change the regression estimators.

The following assumption will be employed to control $\Gamma_{n,i}$ in our subsequent analysis:

Assumption 5.5. For some $\kappa_1 > 0$ and κ_2,

$$\inf_{n\ge 1}\inf_{v\in\mathbf{R}^2} \|v\|_2^{-2} v' E[\mathrm{Var}[\Gamma_{n,i}|X_i]]v \ge \kappa_1$$

$$\sup_{n\ge 1}\sup_{v\in\mathbf{R}^2} \|v\|_2^{-2} v' E[\mathrm{Var}[\Gamma_{n,i}|X_i]]v \le \kappa_2.$$

The following theorem characterizes the behavior of $\hat{\Delta}^{\mathrm{refit}}_n$:

Theorem 5.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.3. Further suppose Assumptions 5.1–5.5 hold. Then, (9), (12), and Assumption 3.1 are satisfied with $\hat{m}_{d,n}(X_i, W_{n,i}) = \hat{\Gamma}_{n,i}'\hat{\beta}^{\mathrm{refit}}_n$ and*

$$m_{d,n}(X_i, W_{n,i}) = \Gamma_{n,i}'\beta^{\mathrm{refit}}_n$$

for $d \in \{0,1\}$ and $n \ge 1$, where $\beta^{\mathrm{refit}}_n = (2E[\mathrm{Var}[\Gamma_{n,i}|X_i]])^{-1}E[\mathrm{Cov}[\Gamma_{n,i}, Y_i(1) + Y_i(0)|X_i]]$. In addition, denote the asymptotic variance of $\hat{\Delta}^{\mathrm{refit}}_n$ as $\sigma^{\mathrm{refit},2}_n$. Then, $\sigma^{\mathrm{unadj},2}_n \ge \sigma^{\mathrm{refit},2}_n$ and $\sigma^{\mathrm{r},2}_n \ge \sigma^{\mathrm{refit},2}_n$.

Remark 5.7. It is possible to further relax the full rank condition in Assumption 5.5 by running a ridge regression or truncating the minimum eigenvalue of the gram matrix in the refitting step. ■

6. Simulations

In this section, we conduct Monte Carlo experiments to assess the finite-sample performance of the inference methods proposed in the paper. In all cases, we follow Bai et al. (2022) to consider tests of the hypothesis that

$$H_0 : \Delta(Q) = \Delta_0 \text{ versus } H_1 : \Delta(Q) \neq \Delta_0.$$

with $\Delta_0 = 0$ at nominal level $\alpha = 0.05$.

6.1. Data generating processes

We generate potential outcomes for $d \in \{0,1\}$ and $1 \leq i \leq 2n$ by the equation

$$Y_i(d) = \mu_d + m_d(X_i, W_i) + \sigma_d(X_i, W_i)\epsilon_{d,i}, \ d = 0, 1, \tag{21}$$

where $\mu_d, m_d\left(X_i, W_i\right), \sigma_d\left(X_i, W_i\right)$, and $\epsilon_{d,i}$ are specified in each model as follows. In each of the specifications, $(X_i, W_i, \epsilon_{0,i}, \epsilon_{1,i})$ are i.i.d. across i. The number of pairs n is equal to 100 and 200. The number of replications is 10,000.

Model 1 $\left(X_i, W_i\right)^\top = \left(\Phi\left(V_{i1}\right), \Phi\left(V_{i2}\right)\right)^\top$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function and

$$V_i \sim N\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array}\right)\right),$$

$m_0\left(X_i, W_i\right) = \gamma\left(W_i - \frac{1}{2}\right)$; $m_1\left(X_i, W_i\right) = m_0\left(X_i, W_i\right)$; $\epsilon_{d,i} \sim N(0,1)$ for $d = 0, 1$; $\sigma_0\left(X_i, W_i\right) = \sigma_1\left(X_i, W_i\right) = 1$. We set $\gamma = 4$ and $\rho = 0.2$.

Model 2 $\left(X_i, W_i\right)^\top = \left(\Phi\left(V_{i1}\right), V_{1i}V_{i2}\right)^\top$, where V_i is the same as in Model 1. $m_0\left(X_i, W_i\right) = m_1\left(X_i, W_i\right) = \gamma_1\left(W_i - \rho\right) + \gamma_2\left(\Phi^{-1}\left(X_i\right)^2 - 1\right)$. $\epsilon_{d,i} \sim N(0,1)$ for $d = 0, 1$; $\sigma_0\left(X_i, W_i\right) = \sigma_1\left(X_i, W_i\right) = 1$. $\left(\gamma_1, \gamma_2\right)^\top = (1,2)^\top$ and $\rho = 0.2$.

Model 3 The same as in Model 2, except that $m_0\left(X_i, W_i\right) = m_1\left(X_i, W_i\right) = \gamma_1\left(W_i - \rho\right) + \gamma_2\left(\Phi\left(W_i\right) - \frac{1}{2}\right) + \gamma_3\left(\Phi^{-1}\left(X_i\right)^2 - 1\right)$ with $\left(\gamma_1, \gamma_2, \gamma_3\right)^\top = \left(\frac{1}{4}, 1, 2\right)^\top$.

Model 4 $\left(X_i, W_i\right)^\top = \left(V_{i1}, V_{1i}V_{i2}\right)^\top$, where V_i is the same as in Model 1. $m_0\left(X_i, W_i\right) = m_1\left(X_i, W_i\right) = \gamma_1\left(W_i - \rho\right) + \gamma_2\left(\Phi\left(W_i\right) - \frac{1}{2}\right) + \gamma_3\left(X_i^2 - 1\right)$. $\epsilon_{d,i} \sim N(0,1)$ for $d = 0, 1$; $\sigma_0\left(X_i, W_i\right) = \sigma_1\left(X_i, W_i\right) = 1$. $\left(\gamma_1, \gamma_2, \gamma_3\right)^\top = (2,1,2)^\top$ and $\rho = 0.2$.

Model 5 The same as in Model 4, except that $m_1\left(X_i, W_i\right) = m_0\left(X_i, W_i\right) + \left(\Phi\left(X_i\right) - \frac{1}{2}\right)$.

Model 6 The same as in Model 5, except that $\sigma_0\left(X_i, W_i\right) = \sigma_1\left(X_i, W_i\right) = \left(\Phi\left(X_i\right) + 0.5\right)$.

Model 7 $X_i = \left(V_{i1}, V_{i2}\right)^\top$ and $W_i = \left(V_{i1}V_{i3}, V_{i2}V_{i4}\right)^\top$, where $V_i \sim N(0, \Sigma)$ with $\dim(V_i) = 4$ and Σ consisting of 1 on the diagonal and ρ on all other elements. $m_0\left(X_i, W_i\right) = m_1\left(X_i, W_i\right) = \gamma_1'\left(W_i - \rho\right) + \gamma_2'\left(\Phi\left(W_i\right) - \frac{1}{2}\right) + \gamma_3\left(X_{i1}^2 - 1\right)$ with $\gamma_1 = (2,2)^\top, \gamma_2 = (1,1)^\top, \gamma_3 = 1$. $\epsilon_{d,i} \sim N(0,1)$ for $d = 0, 1$; $\sigma_0\left(X_i, W_i\right) = \sigma_1\left(X_i, W_i\right) = 1$. $\rho = 0.2$.

Model 8 The same as in Model 7, except that $m_1\left(X_i, W_i\right) = m_0\left(X_i, W_i\right) + \left(\Phi\left(X_{i1}\right) - \frac{1}{2}\right)$.

Model 9 The same as in Model 8, except that $\sigma_0\left(X_i, W_i\right) = \sigma_1\left(X_i, W_i\right) = \left(\Phi\left(X_{i1}\right) + 0.5\right)$

Model 10 $X_i = \left(\Phi\left(V_{i1}\right), \ldots, \Phi\left(V_{i4}\right)\right)^\top$ and $W_i = \left(V_{i1}V_{i5}, V_{i2}V_{i6}\right)^\top$, where $V_i \sim N(0, \Sigma)$ with $\dim(V_i) = 6$ and Σ consisting of 1 on the diagonal and ρ on all other elements. $m_0\left(X_i, W_i\right) = m_1\left(X_i, W_i\right) = \gamma_1'\left(W_i - \rho\right) + \gamma_2'\left(\Phi\left(W_i\right) - \frac{1}{2}\right) + \gamma_3'\left(\left(\Phi^{-1}\left(X_{i1}\right)^2, \Phi^{-1}\left(X_{i2}\right)^2\right)^\top - 1\right)$ with $\gamma_1 = (1,1)^\top, \gamma_2 = \left(\frac{1}{2}, \frac{1}{2}\right)^\top, \gamma_3 = \left(\frac{1}{2}, \frac{1}{2}\right)^\top$. $\epsilon_{d,i} \sim N(0,1)$ for $d = 0, 1$; $\sigma_0\left(X_i, W_i\right) = \sigma_1\left(X_i, W_i\right) = 1$. $\rho = 0.2$

Model 11 The same as in Model 10, except that $m_1\left(X_i, W_i\right) = m_0\left(X_i, W_i\right) + \frac{1}{4}\sum_{j=1}^{4}\left(X_{ij} - \frac{1}{2}\right)$.

Model 12 $X_i = \left(\Phi\left(V_{i1}\right), \ldots, \Phi\left(V_{i4}\right)\right)^\top$ and $W_i = \left(V_{i1}V_{i41}, \ldots, V_{i40}V_{i80}\right)^\top$, where $V_i \sim N(0, \Sigma)$ with $\dim(V_i) = 80$. Σ is the Toeplitz matrix

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.5^2 & \cdots & 0.5^{79} \\ 0.5 & 1 & 0.5 & \cdots & 0.5^{78} \\ 0.5^2 & 0.5 & 1 & \cdots & 0.5^{77} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.5^{79} & 0.5^{78} & 0.5^{77} & \cdots & 1 \end{pmatrix}.$$

$m_0\left(X_i, W_i\right) = m_1\left(X_i, W_i\right) = \gamma_1'W_i + \gamma_2'\left(\Phi^{-1}\left(X_i\right)^2 - 1\right)$, $\gamma_1 = \left(\frac{1}{1^2}, \frac{1}{2^2}, \ldots, \frac{1}{40^2}\right)^\top$ with $\dim(\gamma_1) = 40$, and $\gamma_2 = \frac{1}{2}\left(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)^\top$ with $\dim(\gamma_2) = 4$. $\epsilon_{d,i} \sim N(0,1)$ for $d = 0, 1$; $\sigma_0\left(X_i, W_i\right) = \sigma_1\left(X_i, W_i\right) = 1$.

Model 13 The same as in Model 12, except that $m_0\left(X_i, W_i\right) = m_1\left(X_i, W_i\right) = \gamma_1'W_i + \gamma_2'\left(\Phi\left(W_i\right) - \frac{1}{2}\right) + \gamma_3'\left(\Phi^{-1}\left(X_i\right)^2 - 1\right)$, $\gamma_1 = \left(\frac{1}{1^2}, \ldots, \frac{1}{40^2}\right)^\top$, $\gamma_2 = \frac{1}{8}\left(\frac{1}{1^2}, \ldots, \frac{1}{40^2}\right)^\top$, and $\gamma_3 = \frac{1}{2}\left(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)^\top$ with $\dim(\gamma_1) = \dim(\gamma_2) = 40$ and $\dim(\gamma_3) = 4$.

Table 1
Rejection probabilities for Models 1–11 when $n = 100$.

Model	H_0: $\Delta = 0$					H_1: $\Delta = 1/4$				
	unadj	naïve	naïve2	pfe	refit	unadj	naïve	naïve2	pfe	refit
1	5.47	5.57	5.63	5.76	5.84	22.48	43.89	43.95	43.91	43.92
2	4.96	5.26	5.30	5.47	5.32	23.32	28.02	27.96	37.21	33.12
3	4.99	5.28	5.24	5.48	5.27	32.19	27.88	27.96	37.34	36.29
4	5.31	5.28	5.28	5.48	5.79	11.78	27.88	28.03	37.34	43.28
5	5.43	5.09	5.08	5.49	5.78	11.87	27.72	27.88	36.69	43.08
6	5.28	5.43	5.41	5.58	5.79	11.78	26.67	26.72	34.71	40.29
7	5.64	5.63	5.62	5.98	6.04	9.24	34.55	34.65	37.96	42.08
8	5.63	5.54	5.51	6.03	6.17	9.28	34.11	34.42	37.22	41.29
9	5.74	5.69	5.76	6.19	5.89	8.99	32.39	32.30	35.42	38.75
10	5.24	5.78	5.73	6.05	6.04	14.27	30.80	30.75	32.02	32.51
11	5.19	5.78	5.72	6.07	5.95	14.36	30.60	30.49	32.21	32.81

Table 2
Rejection probabilities for Models 12–15 when $n = 100$.

	H_0: $\Delta = 0$		H_1: $\Delta = 1/4$	
	unadj	refit	unadj	refit
12	5.39	6.14	22.32	43.90
13	5.36	6.22	21.82	43.98
14	5.43	6.31	21.58	42.63
15	5.73	6.56	20.90	39.71

Model 14 The same as in Model 13, except that $m_1\left(X_i, W_i\right) = m_0\left(X_i, W_i\right) + \sum_{j=1}^{4} \frac{1}{j^2}\left(X_{ij} - \frac{1}{2}\right)$.

Model 15 The same as in Model 14, except that $\sigma_0\left(X_i, W_i\right) = \sigma_1\left(X_i, W_i\right) = \left(X_{i1} + 0.5\right)$.

It is worth noting that Models 1, 2, 3, 4, 7, 10, 12, and 13 imply homogeneous treatment effects because $m_1\left(X_i, W_i\right) = m_0\left(X_i, W_i\right)$. Among them, $E[Y_i(d)|X_i, W_i] - E[Y_i(d)|X_i]$ is linear in W_i in Models 1, 2, and 12. Models 5, 8, 11, and 14 have heterogeneous but homoscedastic treatment effects. In Models 6, 9, and 15, however, the implied treatment effects are both heterogeneous and heteroscedastic. Models 12–15 contain high-dimensional covariates.

We follow Bai et al. (2022) to match pairs. Specifically, if $\dim\left(X_i\right) = 1$, we match pairs by sorting $X_i, i = 1, \ldots, 2n$. If $\dim\left(X_i\right) > 1$, we match pairs by the permutation π calculated using the R package *nbpMatching*. For more details, see Bai et al. (2022, Section 4). After matching the pairs, we flip coins to randomly select one unit within each pair for treatment and another for control.

6.2. Estimation and inference

We set $\mu_0 = 0$ and $\mu_1 = \Delta$, where $\Delta = 0$ and $1/4$ are used to illustrate the size and power, respectively. Rejection probabilities in percentage points are presented. To further illustrate the efficiency gains obtained by regression adjustments, in Fig. 1, we plot the average standard error reduction in percentage relative to the standard error of the estimator without adjustments for various estimation methods.

Specifically, we consider the following adjusted estimators.

(i) unadj: the estimator with no adjustments. In this case, our standard error is identical to the adjusted standard error proposed by Bai et al. (2022).
(ii) naïve: the linear adjustments with regressors W_i but without pair dummies.
(iii) naïve2: the linear adjustments with X_i and W_i regressors but without pair dummies.
(iv) pfe: the linear adjustments with regressors W_i and pair dummies.
(v) refit: refit the ℓ_1-regularized adjustments by linear regression with pair dummies.

See Section C in the Online Supplement for the regressors used in the regularized adjustments.

For Models 1–11, we examine the performance of estimators (i)–(v). For Models 12–15, we assess the performance among estimators (i) and (v) in high-dimensional settings. Note that the adjustments are misspecified for almost all the models. The only exception is Model 1, for which the linear adjustment in W_i is correctly specified because $m_d(X_i, W_i)$ is just a linear function of W_i.

6.3. Simulation results

Tables 1 and 3 report rejection probabilities at the 0.05 level and power of the different methods for Models 1–11 when n is 100 and 200, respectively. Several patterns emerge. First, for all the estimators, the rejection rates under H_0 are close to the nominal

Table 3
Rejection probabilities for Models 1–11 when $n = 200$.

Model	H_0: $\Delta = 0$					H_1: $\Delta = 1/4$				
	unadj	naïve	naïve2	pfe	refit	unadj	naïve	naïve2	pfe	refit
1	5.08	5.04	5.10	5.21	5.31	38.94	70.35	70.36	70.32	70.30
2	5.69	5.28	5.28	5.24	5.40	40.31	49.25	49.32	65.36	57.87
3	5.44	5.29	5.30	5.35	5.41	56.89	49.43	49.51	64.96	62.42
4	5.45	5.29	5.29	5.35	5.20	18.55	49.43	49.67	64.96	69.96
5	5.45	5.24	5.18	5.19	5.29	18.41	48.65	48.80	64.11	69.09
6	5.62	5.32	5.31	5.35	5.43	18.19	46.71	46.67	61.09	65.98
7	5.24	5.51	5.46	5.34	5.49	11.86	60.73	60.63	65.14	69.24
8	5.23	5.49	5.47	5.35	5.65	11.84	60.00	60.10	64.93	68.02
9	5.30	5.58	5.57	5.66	5.81	11.90	57.25	57.28	61.61	64.88
10	5.34	5.19	5.15	5.25	5.31	23.95	55.49	55.44	56.64	56.43
11	5.41	5.36	5.32	5.34	5.41	23.88	55.01	55.05	56.31	56.18

Table 4
Rejection probabilities for Models 12–15 when $n = 200$.

	H_0: $\Delta = 0$		H_1: $\Delta = 1/4$	
	unadj	refit	unadj	refit
12	4.99	5.84	39.56	69.66
13	5.01	5.80	38.58	69.54
14	5.05	5.72	37.94	68.08
15	5.03	5.64	37.20	65.04

level even when $n = 100$ and with misspecified adjustments. This result is expected because all the estimators take into account the dependence structure arising in the "matched pairs" design, consistent with the findings in Bai et al. (2022).

Second, in terms of power, "pfe" is higher than "unadj", "naïve", and "naïve2" for all eleven models, as predicted by our theory. This finding confirms that "pfe" is the optimal linear adjustment and will not degrade the precision of the ATE estimator. In contrast, we observe that "naïve" and "naïve2" in Model 3 are even less powerful than the unadjusted estimator "unadj". Fig. 1 further confirms that these two methods inflate the estimation standard error. This result echoes Freedman's critique (Freedman, 2008) that careless regression adjustments may degrade the estimation precision. Our "pfe" addresses this issue because it has been proven to be weakly more efficient than the unadjusted estimator.

Third, the improvement of power for "pfe" is mainly due to the reduction of estimation standard errors, which can be more than 50% as shown in Fig. 1 for Models 4–9. This means that the length of the confidence interval of the "pfe" estimator is just half of that for the "unadj" estimator. Note the standard error of the "unadj" estimator is the one proposed by Bai et al. (2022), which has already been adjusted to account for the cross-sectional dependence created in pair matching. The extra 50% reduction is therefore produced purely by the regression adjustment. For Models 10–11, the reduction of standard errors achieved by "pfe" is more than 40% as well. For Model 1, the linear regression is correctly specified so that all three methods achieve the global minimum asymptotic variance and maximum power. For Model 2, $m_d(X_i, W_i) - E[m_d(X_i, W_i)|X_i] = \gamma(W_i - E[W_i|X_i])$ so that the linear adjustment γW_i satisfies the conditions in Theorem 3.1. Therefore, "pfe", as the best linear adjustment, is also the best adjustment globally, achieving the global minimum asymptotic variance and maximum power. In contrast, "naïve" and "naïve2" are not the best linear adjustment and therefore less powerful than "pfe" because of the omitted pair dummies.

Finally, the "refit" method has the best power for most models as they automatically achieve the global minimum asymptotic variance when the dimension of W_i is fixed.

Tables 2 and 4 report the size and power for the "refit" adjustments when both W_i and X_i are high-dimensional. We see that the size under the null is close to the nominal 5% while the power for the adjusted estimator is higher than the unadjusted one. Fig. 1 further illustrates the reduction of the standard error is more than 30% for all high-dimensional models.

7. Empirical illustration

In this section, we revisit the randomized experiment with a matched pairs design conducted in Groh and McKenzie (2016). In the paper, they examined the impact of macroinsurance on microenterprises. Here, we apply the covariate adjustment methods developed in this paper to their data and reinvestigate the average effect of macroinsurance on three outcome variables: the microenterprises' monthly profits, revenues, and investment.

The subjects in the experiment are microenterprise owners, who were the clients of the largest microfinance institution in Egypt. In the randomization, after an exact match of gender and the institution's branch code, those clients were grouped into pairs by applying an optimal greedy algorithm to additional 13 matching variables. Within each pair, a macroinsurance product was then offered to one randomly assigned client, and the other acted as a control. Based on the pair identities and all the matching variables,

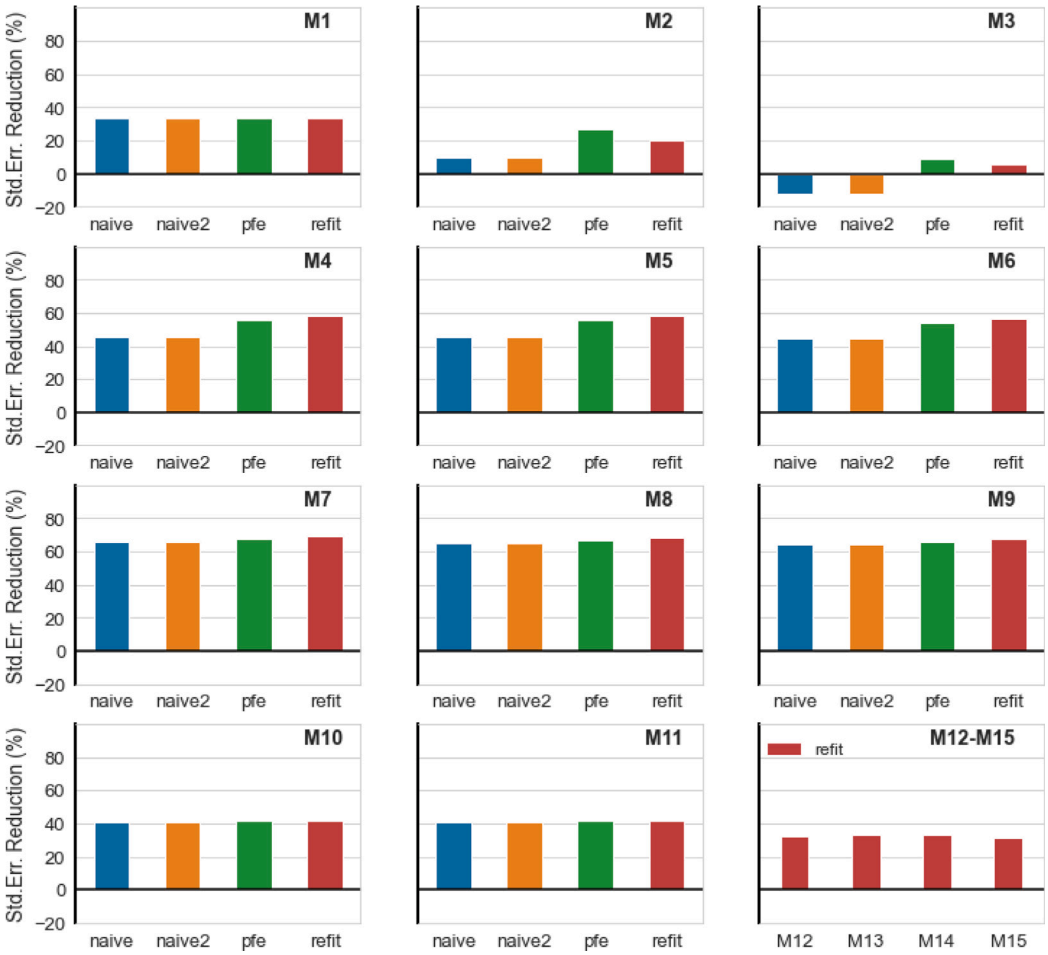

Fig. 1. Average Standard Error Reduction in Percentage under H_1 when $n = 200$.
Notes: The figure plots average standard error reduction in percentage achieved by regression adjustments relative to "unadj" under H_1 for Models 1–15 when $n = 200$.

we re-order the pairs in our sample according to the procedure described in Section 5.1 of Jiang et al. (2022). The resulting sample contains 2824 microenterprise owners, that is, 1412 pairs of them.[2]

Table 5 reports the ATEs with the standard errors (in parentheses) estimated by different methods. Among them, "GM" corresponds to the method used in Groh and McKenzie (2016). The description of other methods is similar to that in Section 6.2.[3] The results in this table prompt the following observations.

First, aligning with our theoretical and simulation findings, we observe that the standard errors associated with the covariate-adjusted ATEs, particularly those for the "naïve2" and "pfe" estimates, are generally lower compared to the ATE estimate without any adjustment. This pattern is consistent across nearly all the outcome variables. To illustrate, when examining the revenue outcome, the standard errors for the "pfe" estimates are 10.2% smaller than those for the unadjusted ATE estimate.

Second, the standard errors of the "refit" estimates are consistently smaller than those of the unadjusted ATE estimate across all the outcome variables. For example, when profits are the outcome variable, the "refit" estimates exhibit standard errors 7.5% smaller than those of the unadjusted ATE estimate. Moreover, compared with those of the "pfe" estimates, the standard errors of "refit" are slightly smaller.

8. Conclusion

This paper considers covariate adjustment for the estimation of average treatment effect in "matched pairs" designs when covariates other than the matching variables are available. When the dimension of these covariates is low, we suggest estimating the

[2] See Groh and McKenzie (2016) and Jiang et al. (2022) for more details.

[3] See Section D in the Online Supplement for the details of the regressors used in all these methods.

Table 5
Impacts of macronsurance for microenterprises.

Y	n	unadj	GM	naïve	naïve2	pfe	refit
Profits	1322	−85.65	−50.88	−41.69	−50.97	−51.60	−55.13
		(49.43)	(46.46)	(47.22)	(45.49)	(46.94)	(45.71)
Revenue	1318	−838.60	−660.16	−611.75	−610.80	−635.80	−600.97
		(319.02)	(284.02)	(286.93)	(282.93)	(286.50)	(284.60)
Investment	1410	−66.60	−66.60	−49.37	−50.72	−67.31	−58.77
		(118.93)	(118.66)	(119.23)	(118.97)	(118.88)	(118.84)

Notes: The table reports the ATE estimates of the effect of macroinsurance for microenterprises. Standard errors are in parentheses.

average treatment effect by linear regression of the outcome on treatment status and covariates, controlling for pair fixed effects. We show that this estimator is no worse than the simple difference-in-means estimator in terms of efficiency. When the dimension of these covariates is high, we suggest a two-step estimation procedure: in the first step, we run ℓ_1-regularized regressions of outcome on covariates for the treated and control groups separately and obtain the fitted values for both potential outcomes, and in the second step, we estimate the average treatment effect by refitting a linear regression of outcome on treatment status and regularized adjustments from the first step, controlling for the pair fixed effects. We show that the final estimator is no worse than the simple difference-in-means estimator in terms of efficiency. When the conditional mean models are approximately correctly specified, this estimator further achieves the minimum variance as if all relevant covariates are used to form pairs in the experiment design stage. We take the choice of variables to use in forming pairs as given and focus on how to obtain more efficient estimators of the average treatment effect in the analysis stage. Our paper is therefore silent on the important question of how to choose the relevant matching variables in the design stage. This topic is left for future research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2024.105740.

References

Abadie, A., Imbens, G.W., 2008. Estimation of the conditional variance in paired experiments. Annales d'É conomie et de Statistique (91/92), 175–187.
Armstrong, T.B., 2022. Asymptotic efficiency bounds for a class of experimental designs. arXiv:2205.02726, [stat], URL http://arxiv.org/abs/2205.02726.
Bai, Y., Liu, J., Shaikh, A.M., Tabord-Meehan, M., 2023a. On the efficiency of finely stratified experiments. arXiv:2307.15181, [econ, math, stat], URL http://arxiv.org/abs/2307.15181.
Bai, Y., Liu, J., Tabord-Meehan, M., 2023b. Inference for matched tuples and fully blocked factorial designs. arXiv:2206.04157, [econ, math, stat], URL http://arxiv.org/abs/2206.04157.
Bai, Y., Romano, J.P., Shaikh, A.M., 2022. Inference in experiments with matched pairs. J. Amer. Statist. Assoc. 117 (540), 1726–1737.
Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2017. Program evaluation and causal inference with high-dimensional data. Econometrica 85 (1), 233–298.
Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. 37 (4), 1705–1732, URL https://projecteuclid.org/journals/annals-of-statistics/volume-37/issue-4/Simultaneous-analysis-of-Lasso-and-Dantzig-selector/10.1214/08-AOS620.full.
Bruhn, M., McKenzie, D., 2009. In pursuit of balance: Randomization in practice in development field experiments. Am. Econ. J.: Appl. Econ. 1 (4), 200–232.
Chernozhukov, V., Chetverikov, D., Kato, K., 2017. Central limit theorems and bootstrap in high dimensions. Ann. Probab. 45 (4), 2309–2352, URL https://projecteuclid.org/journals/annals-of-probability/volume-45/issue-4/Central-limit-theorems-and-bootstrap-in-high-dimensions/10.1214/16-AOP1113.full.
Cohen, P.L., Fogarty, C.B., 2024. No-harm calibration for generalized oaxaca-blinder estimators. Biometrika 111 (1), 331–338.
Cytrynbaum, M., 2023. Covariate adjustment in stratified experiments.
Donner, A., Klar, N., 2000. Design and Analysis of Cluster Randomization Trials in Health Research. Vol. 27, Arnold London.
Fogarty, C.B., 2018. Regression-assisted inference for the average treatment effect in paired experiments. Biometrika 105 (4), 994–1000.
Freedman, D.A., 2008. On regression adjustments to experimental data. Adv. in Appl. Math. 40 (2), 180–193, URL http://www.sciencedirect.com/science/article/pii/S019688580700005X.
Glennerster, R., Takavarasha, K., 2013. Running Randomized Evaluations: a Practical Guide. Princeton University Press.
Groh, M., McKenzie, D., 2016. Macroinsurance for microenterprises: A randomized experiment in post-revolution Egypt. J. Dev. Econ. 118, 13–25.
Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.
Jiang, L., Liu, X., Phillips, P.C., Zhang, Y., 2022. Bootstrap inference for quantile treatment effects in randomized experiments with matched pairs. Rev. Econ. Stat. 106 (2), 542–556.
Lin, W., 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. Ann. Appl. Stat. 7 (1), 295–318, URL https://projecteuclid.org/euclid.aoas/1365527200.
Negi, A., Wooldridge, J.M., 2021. Revisiting regression adjustment in experiments with heterogeneous treatment effects. Econometric Rev. 40 (5), 504–534.
Robins, J.M., Rotnitzky, A., Zhao, L.P., 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J. Amer. Statist. Assoc. 90 (429), 106–121, URL https://www.jstor.org/stable/2291134.
Rosenberger, W.F., Lachin, J.M., 2015. Randomization in Clinical Trials: Theory and Practice. John Wiley & Sons.
Tsiatis, A.A., Davidian, M., Zhang, M., Lu, X., 2008. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Stat. Med. 27 (23), 4658–4677.
Wager, S., Du, W., Taylor, J., Tibshirani, R.J., 2016. High-dimensional regression adjustments in randomized experiments. Proc. Natl. Acad. Sci. 113 (45), 12673–12678.
Wu, E., Gagnon-Bartsch, J.A., 2021. Design-based covariate adjustments in paired experiments. J. Educ. Behav. Stat. 46 (1), 109–132.
Yang, L., Tsiatis, A.A., 2001. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. Amer. Statist. 55 (4), 314–321.
Zhao, A., Ding, P., 2021. Covariate-adjusted Fisher randomization tests for the average treatment effect. J. Econometrics 225 (2), 278–294, URL https://www.sciencedirect.com/science/article/pii/S0304407621001457.