Biometrika (2020), **xx**, x, *pp.* 1–13
Printed in Great Britain

Regression adjustment in completely randomized experiments with a diverging number of covariates

BY LIHUA LEI
Departments of Statistics, Stanford University
lihualei@stanford.edu

PENG DING
Departments of Statistics, University of California, Berkeley
pengdingpku@berkeley.edu

SUMMARY

Randomized experiments have become important tools in empirical research. In a completely randomized treatment-control experiment, the simple difference in means of the outcome is unbiased for the average treatment effect, and covariate adjustment can further improve the efficiency without assuming a correctly specified outcome model. In modern applications, experimenters often have access to many covariates, motivating the need for a theory of covariate adjustment under the asymptotic regime with a diverging number of covariates. We study the asymptotic properties of covariate adjustment under the potential outcomes model and propose a bias-corrected estimator that is consistent and asymptotically normal under weaker conditions. Our theory is purely randomization-based without imposing any parametric outcome model assumptions. To prove the theoretical results, we develop novel vector and matrix concentration inequalities for sampling without replacement.

Some key words: causal inference; average treatment effect; high-dimensional covariates; model misspecification

1. INTRODUCTION

1.1. *Randomized experiment and Neyman's randomization model*

Randomized experiments have been powerful tools in agricultural, industrial, biomedical, and social sciences (e.g., Fisher, 1935; Kempthorne, 1952; Box et al., 2005; Rosenberger & Lachin, 2015; Duflo et al., 2007; Gerber & Green, 2012; Imbens & Rubin, 2015). In a treatment-control experiment, let $Y_i(1)$ and $Y_i(0)$ be the potential outcomes if unit $i \in \{1, \ldots, n\}$ receives the treatment and control, respectively (Neyman, 1923/1990). Define the parameter of interest as the average treatment effect $\tau = n^{-1} \sum_{i=1}^{n} \tau_i$, where $\tau_i = Y_i(1) - Y_i(0)$ is the individual treatment effect for unit i. In a completely randomized experiment, the experimenter randomly assigns n_1 units to the treatment group and n_0 units to the control group, with $n = n_1 + n_0$. Let T_i denote the assignment of the i-th unit where $T_i = 1$ corresponds to the treatment and $T_i = 0$ corresponds to the control. For unit i, only $Y_i^{\mathrm{obs}} = Y_i(T_i)$ is observed while the other potential outcome $Y_i(1 - T_i)$ is missing.

Neyman (1923/1990) assumed that all potential outcomes are fixed and the randomness comes solely from the treatment indicators. This finite-population perspective has a long history for analyzing randomized experiments (e.g. Kempthorne, 1952; Imbens & Rubin, 2015; Dasgupta et al., 2015; Middleton & Aronow, 2015; Mukerjee et al., 2018; Fogarty, 2018; Li et al., 2018; Li & Ding, 2020). It clarifies the role of the study design in the analysis without postulating a hypothetical outcome generating process. By contrast, the super-population perspective (e.g. Tsiatis et al., 2008; Berk et al., 2013; Negi & Wooldridge, 2020) assumes that the potential outcomes and other individual characteristics are independent and identically distributed draws from some distribution. These two perspectives are both popular in the literature, but they are different in the source of randomness: the finite-population perspective conditions on the potential outcomes and quantifies the uncertainty from the treatment assignment, and the super-population averages over the potential outcomes and quantifies the uncertainty from the independent sampling process. We focus on the former throughout the paper.

Let $\mathbb{1}$ denote the vector with all entries 1, $\mathbb{I}$ the identity matrix, and $\mathbb{V} = \mathbb{I} - (\mathbb{1}^\mathsf{T}\mathbb{1})^{-1}\mathbb{1}\mathbb{1}^\mathsf{T}$ the projection matrix orthogonal to $\mathbb{1}$, with appropriate dimensions depending on the context. Let $\|\cdot\|_q$ be the vector q-norm, i.e. $\|\alpha\|_q = (\sum_{i=1}^n |\alpha_i|^q)^{1/q}$ and $\|\alpha\|_\infty = \max_{1\le i\le n}|\alpha_i|$. Let $\|\cdot\|_{\mathrm{op}}$ denote the matrix operator norm. Let $N(0,1)$ denote the standard normal distribution, and $t(\nu)$ denote standard t distribution with degrees of freedom ν.

1.2. *Average treatment effect estimates with and without regression adjustment*

Let $\mathcal{T}_t = \{i : T_i = t\}$ be the indices and $n_t = |\mathcal{T}_t|$ be the fixed sample size for the treatment arm $t \in \{0,1\}$. Consider a completely randomized experiment in which $\mathcal{T}_1$ is a random size-n_1 subset of $\{1,\dots,n\}$ uniformly over all $n!/(n_1!n_0!)$ subsets. The simple difference in means

$$\hat{\tau}_{\mathrm{unadj}} = n_1^{-1}\sum_{i\in\mathcal{T}_1} Y_i^{\mathrm{obs}} - n_0^{-1}\sum_{i\in\mathcal{T}_0} Y_i^{\mathrm{obs}} = n_1^{-1}\sum_{i\in\mathcal{T}_1} Y_i(1) - n_0^{-1}\sum_{i\in\mathcal{T}_0} Y_i(0)$$

is unbiased for τ with variance $S_1^2/n_1 + S_0^2/n_0 - S_\tau^2/n$ (Neyman, 1923/1990), where S_1^2, S_0^2 and S_τ^2 are the finite-population variances of the $Y_i(1)$'s, $Y_i(0)$'s and τ_i's, respectively.

The experimenter usually collects pre-treatment covariates. If they are predictive of the potential outcomes, incorporating them in the analysis can improve the estimation efficiency. Suppose unit i has a p-dimensional vector of pre-treatment covariates $x_i \in \mathbb{R}^p$. Early works on the analysis of covariance assumed a constant treatment effect (Fisher, 1935; Kempthorne, 1952; Hinkelmann & Kempthorne, 2007), under which a commonly-used estimate is the coefficient of the treatment indicator of the ordinary least squares fit of the Y_i^{obs}'s on T_i's and x_i's. Freedman (2008) criticized this approach, showing that it can be even less efficient than $\hat{\tau}_{\mathrm{unadj}}$ in the presence of treatment effect heterogeneity, and the estimated standard error based on the ordinary least squares can be inconsistent for the true standard error under the randomization model.

Lin (2013) proposes a simple solution. Without loss of generality, we center the covariates at $n^{-1}\sum_{i=1}^n x_i = 0$. His estimator for τ is the coefficient of the treatment indicator in the ordinary least squares fit of the Y_i^{obs}'s on T_i's, x_i's and the interaction terms $T_i x_i$'s. His estimator is consistent, asymptotically normal, and more efficient than $\hat{\tau}_{\mathrm{unadj}}$. He further shows that the Eicker–Huber–White standard error is consistent for the true standard error. His results hold under the finite-population randomization model, without assuming that the linear model is correct.

We use an alternative formulation of the regression adjustment and consider the following family of covariate-adjusted estimators:

$$\hat{\tau}(\gamma_1,\gamma_0) = n_1^{-1}\sum_{i\in\mathcal{T}_1}(Y_i^{\mathrm{obs}} - x_i^\mathsf{T}\gamma_1) - n_0^{-1}\sum_{i\in\mathcal{T}_0}(Y_i^{\mathrm{obs}} - x_i^\mathsf{T}\gamma_0). \tag{1}$$

Because $n_t^{-1}\sum_{i\in\mathcal{T}_t} x_i^\mathsf{T}\gamma_t$ has expectation zero over all possible randomizations, the estimator in (1) is unbiased for any fixed coefficient vectors $\gamma_t \in \mathbb{R}^p$ $(t = 0,1)$. It is the difference-in-means estimator with potential outcomes replaced by $\{Y_i(1) - x_i^\mathsf{T}\gamma_1, Y_i(0) - x_i^\mathsf{T}\gamma_0\}_{i=1}^n$.

Let $Y(t) = (Y_1(t), \ldots, Y_n(t))^{\mathsf{T}} \in \mathbb{R}^n$ denote the vector of potential outcomes under treatment t and $X = (x_1, \ldots, x_n)^{\mathsf{T}}$ denote the covariate matrix. Without loss of generality, we assume $\mathbb{1}^{\mathsf{T}} X = 0$ and $\mathrm{rank}(X) = p$, i.e., the covariate matrix has centered columns and full column rank. Otherwise, we transform X to $\mathbb{V}X$ and remove the redundant columns to ensure the full column rank condition. This operation does not affect inferential validity because X is fixed, or, equivalently, our inference conditions on X.

Let β_t be the population ordinary least squares coefficient of regressing $Y(t)$ on $(\mathbb{1}, X)$:

$$(\mu_t, \beta_t) = \underset{\mu\in\mathbb{R},\beta\in\mathbb{R}^p}{\arg\min} \|Y(t) - \mu\mathbb{1} - X\beta\|_2^2 = \left(n^{-1}\sum_{i=1}^{n} Y_i(t), (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y(t)\right), \quad (2)$$

which holds because X is orthogonal to $\mathbb{1}$. Li & Ding (2017, Example 9) show that the ordinary least squares coefficients (β_1, β_0) in (2) minimize the variance of the estimator in (1). We emphasize that β_1 and β_0 are both unobserved population quantities.

The classical analysis of covariance chooses $\gamma_1 = \gamma_0 = \hat{\beta}$, the coefficient of the covariates in the ordinary least squares fit of the Y_i^{obs}'s on T_i's and x_i's with an intercept. This strategy implicitly assumes away treatment effect heterogeneity, and can lead to inferior properties when $\beta_1 \neq \beta_0$ (Freedman, 2008). Lin (2013) chooses $\gamma_1 = \hat{\beta}_1$ and $\gamma_0 = \hat{\beta}_0$, the coefficients of the covariates in the ordinary least squares fit of Y_i^{obs}'s on x_i's with an intercept, in the treatment and control groups, respectively. Numerically, this is identical to the estimator obtained from the regression with interactions discussed before. Throughout the rest of the paper, we refer to it as the regression-adjusted estimator.

1.3. *Our contributions*

In practice, experiments often have many covariates. Therefore, it is important to approximate the sampling distribution with p growing with the sample size n at a certain rate. Under the finite-population randomization model, Bloniarz et al. (2016) discussed a high dimensional regime with possibly larger p than n but assumed that the potential outcomes could be well approximated by a sparse linear combination of the covariates, under the regime with the number of non-zero coefficients being much smaller than $n^{1/2}/\log p$. Under a super-population framework, Wager et al. (2016) discussed covariate adjustment using the ordinary least squares and some other machine learning techniques.

We study the regression-adjusted estimator under the finite-population perspective in the regime where $p < n$ but p grows with n at a certain rate. We argue that this type of large-n-moderate-p asymptotics is more important than the large-n-fixed-p asymptotics to analyze completely randomized experiments when p is not a negligible number compared to n. For instance, the study on pulmonary artery catheter in Bloniarz et al. (2016) has 1013 subjects with 59 covariates. In this case, p is approximately $n^{0.6}$ and thus the inferential guarantees based on fix-p asymptotics are questionable.

We focus on this estimator because it is widely used in practice thanks to its simplicity, and it does not require any tuning parameter, unlike other high dimensional or machine learning methods. As in the classic linear regression, the asymptotic properties depend crucially on the maximum leverage score

$$\kappa = \max_{1\leq i\leq n} H_{ii},$$

where the i-th leverage score H_{ii} is i-th diagonal entry of the hat matrix $H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$. Under the regime $\kappa \log p \to 0$, we prove the consistency of the regression-adjusted estimator under mild moment conditions on the population ordinary least squares residuals. In the favorable case where all leverage scores are close to their average p/n, the consistency holds if $p = o(n/\log n)$.

In addition, we prove that the regression-adjusted estimator is asymptotically normal under $\kappa p \to 0$ and extra mild conditions, with the same variance formula as in the fixed-p regime. Furthermore, we propose a debiased estimator, which is asymptotically normal under an even weaker assumption $\kappa^2 p \log p \to 0$, with the same variance as before. Therefore, this new estimator reduces the asymptotic bias without inflating the asymptotic variance. In the favorable case where all leverage scores are close to their average p/n, the regression-adjusted estimator is asymptotically normal when $p = o(n^{1/2})$, but the debiased estimator is asymptotically normal when $p = o\{n^{2/3}/(\log n)^{1/3}\}$. The regression-adjusted estimator may also be asymptotically normal in the latter regime, but it requires an extra condition; see Theorem 3. In our simulation, the debiased estimator indeed yields better finite-sample inferences.

For statistical inference, we propose several asymptotically conservative variance estimators, which yield valid asymptotic Wald-type confidence intervals for the average treatment effect. We prove the results under the same conditions as required for the asymptotic normality. To prove these results, we also make some technical contributions by proving novel vector and matrix concentration inequalities for sampling without replacement.

2. Regression Adjustment

2.1. *Point estimators*

We reformulate the regression-adjusted estimator. The average treatment effect τ is the difference between the two intercepts of the population ordinary least squares coefficients in (2): $\tau = n^{-1}\sum_{i=1}^{n} Y_i(1) - n^{-1}\sum_{i=1}^{n} Y_i(0) = \mu_1 - \mu_0$. Therefore, we focus on estimating μ_1 and μ_0. Let $X_t \in \mathbb{R}^{n_t \times p}$ denote the sub-matrix formed by the rows of X, and $Y_t^{\mathrm{obs}} \in \mathbb{R}^{n_t}$ the subvector of $Y^{\mathrm{obs}} = (Y_1^{\mathrm{obs}}, \ldots, Y_n^{\mathrm{obs}})^{\mathsf{T}}$, with indices in $\mathcal{T}_t$ $(t = 0, 1)$. The regression-adjusted estimator follows two steps. First, for $t \in \{0, 1\}$, we regress Y_t^{obs} on X_t with an intercept, and obtain the fitted intercept $\hat{\mu}_t \in \mathbb{R}$ and coefficient of the covariate $\hat{\beta}_t \in \mathbb{R}^p$. Second, we estimate τ by

$$\hat{\tau}_{\mathrm{adj}} = \hat{\mu}_1 - \hat{\mu}_0. \tag{3}$$

In general, $\hat{\tau}_{\mathrm{adj}}$ is biased in finite samples. Correcting the bias gives stronger theoretical guarantees as our later asymptotic analysis confirms. Here we propose a bias-corrected estimator. Define the potential residuals based on the population ordinary least squares as

$$e(t) = Y(t) - \mu_t - X\beta_t, \quad (t = 0, 1). \tag{4}$$

The property of the ordinary least squares guarantees that $e(t)$ is orthogonal to $\mathbb{1}$ and X:

$$\mathbb{1}^{\mathsf{T}} e(t) = 0, \quad X^{\mathsf{T}} e(t) = 0, \quad (t = 0, 1). \tag{5}$$

Let $\hat{e} \in \mathbb{R}^n$ be the vector residuals from the sample ordinary least squares, where $\hat{e}_i = Y_i^{\mathrm{obs}} - \hat{\mu}_1 - x_i^{\mathsf{T}}\hat{\beta}_1$ for the treated units and $\hat{e}_i = Y_i^{\mathrm{obs}} - \hat{\mu}_0 - x_i^{\mathsf{T}}\hat{\beta}_0$ for the control units. For any vector $\alpha \in \mathbb{R}^n$, let α_t denote the subvector of α with indices in $\mathcal{T}_t$, e.g., $Y_t(1)$ and $e_t(1)$ are the subvectors of $Y(1)$ and $e(1)$ corresponding to the units in treatment arm t, respectively. Let

$$H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}, \quad H_t = X_t(X_t^{\mathsf{T}}X_t)^{-1}X_t^{\mathsf{T}}$$

be the hat matrices of X and X_t, respectively. Let H_{ii} be the i-th diagonal element of H, also termed as the leverage score, and let $H_{t,ii}$ be the diagonal element of H_t corresponding to unit i.

From the higher order asymptotic expansion, the bias of $\hat{\tau}_{\text{adj}}$ depends on

$$\Delta_t = n^{-1}\sum_{i=1}^{n} e_i(t)H_{ii}, \quad \Delta = \max\{|\Delta_1|, |\Delta_0|\}. \tag{6}$$

With the empirical analogs $\hat{\Delta}_t = n_t^{-1}\sum_{i\in\mathcal{T}_t}\hat{e}_i H_{ii}$, we introduce the following debiased estimator:

$$\hat{\tau}_{\text{adj}}^{\text{de}} = \hat{\tau}_{\text{adj}} - \left(\frac{n_1}{n_0}\hat{\Delta}_0 - \frac{n_0}{n_1}\hat{\Delta}_1\right). \tag{7}$$

When $p = 1$, (7) reduces to the bias formula in Lin (2013, Section 6 point (iv)). Thus (7) is an extension to the multivariate case. With some algebraic manipulations, we can show that $\hat{\tau}_{\text{adj}}^{\text{de}}$ is a finite-population analog of Tan (2014)'s bias-corrected regression estimator in the context of survey sampling with a fixed p.

2.2. *Variance estimators*

With a fixed p, Lin (2013) proved that $n^{1/2}(\hat{\tau}_{\text{adj}} - \tau)$ is asymptotically normal with variance

$$\sigma_n^2 = n_1^{-1}\sum_{i=1}^{n} e_i^2(1) + n_0^{-1}\sum_{i=1}^{n} e_i^2(0) - n^{-1}\sum_{i=1}^{n}\{e_i(1) - e_i(0)\}^2. \tag{8}$$

Formula (8) motivates conservative variance estimators since the third term in (8) has no consistent estimator without further assumptions on $e(1)$ and $e(0)$. Ignoring it and estimating the first two terms in (8) by their sample analogs, we have the following variance estimator:

$$\hat{\sigma}^2 = \frac{n}{n_1(n_1-1)}\sum_{i\in\mathcal{T}_1}\hat{e}_i^2 + \frac{n}{n_0(n_0-1)}\sum_{i\in\mathcal{T}_0}\hat{e}_i^2. \tag{9}$$

Although (9) appears to be conservative due to the neglect of the third term in (8), we find in numerical experiments that it typically underestimates σ_n^2 if the number of covariates is large. The classic linear regression literature (e.g. MacKinnon, 2013) suggests rescaling the residual as $\tilde{e}_i = \zeta_i\hat{e}_i$, where $\zeta_i = 1$ for HC0, $\zeta_i = \{(n_t - 1)/(n_t - p)\}^{1/2}$ for HC1, $\zeta_i = 1/(1 - H_{t,ii})^{1/2}$ for HC2 and $\zeta_i = 1/(1 - H_{t,ii})$ for HC3, for $i \in \mathcal{T}_t$. HC0 corresponds to the estimator (9) without corrections. Previous literature has shown that the above corrections, especially HC3, are effective in improving the finite sample performance of variance estimator in linear regression under independent super-population sampling. More interestingly, it is also beneficial to use these rescaled residuals in the context of a completely randomized experiment, motivating

$$\hat{\sigma}_{\text{HC}j}^2 = \frac{n}{n_1(n_1-1)}\sum_{i\in\mathcal{T}_1}\tilde{e}_{i,j}^2 + \frac{n}{n_0(n_0-1)}\sum_{i\in\mathcal{T}_0}\tilde{e}_{i,j}^2 \tag{10}$$

with residual $\tilde{e}_{i,j}$ corresponding to HCj for $j = 0, 1, 2, 3$. Based on the normal approximations, we can construct Wald-type confidence intervals for τ based on point estimators $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}_{\text{adj}}^{\text{de}}$ with estimated standard errors $\hat{\sigma}_{\text{HC}j}/n^{1/2}$.

3. Main Results

3.1. *Regularity conditions*

We embed the finite-population quantities $\{x_i, Y_i(1), Y_i(0)\}_{i=1}^n$ into a sequence, and impose regularity conditions on this sequence. The first condition is on the sample sizes.

Assumption 1. $n/n_1 = O(1)$ and $n/n_0 = O(1)$.

Assumption 1 holds automatically if treatment and control groups have fixed proportions, e.g., $n_1/n = n_0/n = 1/2$ for balanced experiments. It is not essential and can be removed at the cost of complicating the statements.

The second condition is on $\kappa = \max_{1\le i\le n} H_{ii}$, the maximum leverage score, which also plays a crucial role in the theory of classic linear models (e.g. Huber, 1973; Mammen, 1989).

Assumption 2. $\kappa \log p = o(1)$.

The maximum leverage score satisfies

$$p/n = \mathrm{tr}(H)/n \le \kappa \le \|H\|_{\mathrm{op}} = 1 \Longrightarrow \kappa \in [p/n, 1]. \tag{11}$$

Assumption 2 permits influential observations as long as $\kappa = o(1/\log p)$. In the favorable case with $\kappa = O(p/n)$, it reduces to $p \log p/n \to 0$, which permits p to grow as fast as n^{γ} for any $0 \le \gamma < 1$. Moreover, it implies

$$p/n \le \kappa = o\left(1/\log p\right) = o(1) \Longrightarrow p = o(n). \tag{12}$$

Assumptions 1 and 2 are useful for establishing consistency. The following two extra conditions are useful for the variance estimation and asymptotic normality. The third condition is on the correlation between the potential residuals from the population ordinary least squares in (4).

Assumption 3. There exists a constant $\eta > 0$ independent of n such that

$$\rho_e = e(1)^{\mathsf{T}} e(0)/\{\|e(1)\|_2 \|e(0)\|_2\} > -1 + \eta.$$

Assumption 3 is mild because it is unlikely to have a perfectly negative sample correlation between the treatment and control potential residuals in practice.

The fourth condition is on the following two measures of the potential residuals:

$$\mathcal{E}_2 = n^{-1} \max\left\{\|e(0)\|_2^2, \|e(1)\|_2^2\right\}, \quad \mathcal{E}_\infty = \max\left\{\|e(0)\|_\infty, \|e(1)\|_\infty\right\}.$$

Assumption 4. $\mathcal{E}_\infty^2/(n\mathcal{E}_2) = o(1)$.

Assumption 4 is a Lindeberg–Feller-type condition requiring that no single residual dominates the others. A similar form appeared in Hájek (1960)'s finite-population central limit theorem. Previous works require more stringent assumptions on the fourth moment (Lin, 2013; Bloniarz et al., 2016) while Assumption 4 allows for heavy-tailed outcomes with $\mathcal{E}_2$ growing with n.

These assumptions are weaker than those in previous works (e.g. Lin, 2013; Bloniarz et al., 2016; Li & Ding, 2020). Supplementary Material II provides further discussions.

3.2. *Asymptotic expansions and consistency*

We start with the asymptotic expansions of $\hat{\tau}_{\mathrm{adj}}$ and $\hat{\tau}_{\mathrm{adj}}^{\mathrm{de}}$.

THEOREM 1. *Under Assumptions* 1 *and* 2,

$$\hat{\tau}_{\mathrm{adj}} - \tau = \hat{\tau}_e + O_{\mathbb{P}}\left[\Delta + \left\{\mathcal{E}_2(\kappa^2 p \log p + \kappa)/n\right\}^{1/2}\right], \tag{13}$$

$$\hat{\tau}_{\mathrm{adj}}^{\mathrm{de}} - \tau = \hat{\tau}_e + O_{\mathbb{P}}\left[\left\{\mathcal{E}_2(\kappa^2 p \log p + \kappa)/n\right\}^{1/2}\right], \tag{14}$$

where $\hat{\tau}_e = \mathbb{1}^{\mathsf{T}} e_1(1)/n_1 - \mathbb{1}^{\mathsf{T}} e_0(0)/n_0$ *is the difference in means of the potential residuals.*

In (13) and (14), $\hat{\tau}_e$ has mean 0 and variance σ_n^2/n (Neyman, 1923/1990), which is $O_{\mathbb{P}}\{(\sigma_n^2/n)^{1/2}\}$ by Chebyshev's inequality. Based on the definitions in (6), we further have

$$\Delta^2 = \max_{t=0,1} \Delta_t^2 \leq \left(n^{-1} \sum_{i=1}^n H_{ii} \right) \left\{ \max_{t=0,1} n^{-1} \sum_{i=1}^n e_i^2(t) H_{ii} \right\} \leq \mathcal{E}_2 \kappa p/n \tag{15}$$

by the Cauchy–Schwarz inequality and the facts that $\sum_{i=1}^n H_{ii} = p$ and $H_{ii} \leq \kappa$. Since $\kappa \leq 1$ and $\sigma_n^2 = O(\mathcal{E}_2)$, Theorem 1 implies

$$\hat{\tau}_{\text{adj}} - \tau = O_{\mathbb{P}} \left[\{\mathcal{E}_2(\kappa p + 1)/n\}^{1/2} \right], \quad \hat{\tau}_{\text{adj}}^{\text{de}} - \tau = O_{\mathbb{P}} \left[\left\{\mathcal{E}_2(\kappa^2 p \log p + 1)/n\right\}^{1/2} \right],$$

which further imply the following consistency results by requiring the right-hand sides to vanish.

THEOREM 2. *Under Assumptions* 1 *and* 2, $\hat{\tau}_{\text{adj}}$ *is consistent if* $\mathcal{E}_2 = o\{n/(\kappa p + 1)\}$, *and* $\hat{\tau}_{\text{adj}}^{\text{de}}$ *is consistent if* $\mathcal{E}_2 = o\{n/(\kappa^2 p \log p + 1)\}$.

Theorem 2 implies the following consistency results for a fixed or diverging p.

COROLLARY 1. *Under Assumptions* 1 *and* 2, *both* $\hat{\tau}_{\text{adj}}$ *and* $\hat{\tau}_{\text{adj}}^{\text{de}}$ *are consistent if either (i)* p *is fixed and* $\mathcal{E}_2 = o(n)$ *or (ii)* p *is diverging with* n *and* $\mathcal{E}_2 = O(n/p)$.

We can prove Corollary 1 by verifying the more stringent condition for the consistency of $\hat{\tau}_{\text{adj}}$ in Theorem 2. With a fixed p and $\mathcal{E}_2 = o(n)$, we have $\mathcal{E}_2/\{n/(\kappa p + 1)\} \leq \mathcal{E}_2/n \times (p+1) \to 0$ because $\kappa \leq 1$; with a diverging p and $\mathcal{E}_2 = O(n/p)$, we have $\mathcal{E}_2/\{n/(\kappa p+1)\} = \mathcal{E}_2/(n/p) \times (\kappa + 1/p) \to 0$ because Assumption 2 implies $\kappa = o(1/\log p)$.

3.3. *Asymptotic normality and variance estimation*

In (13) and (14), $\hat{\tau}_e$ is asymptotically normal with mean 0 and variance σ_n^2/n. Therefore, the asymptotic normalities of $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}_{\text{adj}}^{\text{de}}$ hold if the the remainders vanish after being multiplied by $n^{1/2}/\sigma_n$. We first present the result for $\hat{\tau}_{\text{adj}}$.

THEOREM 3. *Under Assumptions* 1–4, $n^{1/2}(\hat{\tau}_{\text{adj}} - \tau)/\sigma_n \rightsquigarrow N(0,1)$ *if* $\kappa^2 p \log p = o(1)$ *and* $n\Delta^2 = o(\mathcal{E}_2)$.

The term $n\Delta^2$ is the squared bias of $n^{1/2}\hat{\tau}_{\text{adj}}$. If it vanishes, $\hat{\tau}_{\text{adj}}$ has the same asymptotic normality as $\hat{\tau}_e$. We can use Theorem 3 to find more interpretable sufficient conditions to replace $n\Delta^2 = o(\mathcal{E}_2)$. An upper bound on Δ is in (15). So an obvious sufficient condition is $\kappa p = o(1)$, which also implies $\kappa^2 p \log p = (\kappa p)(\kappa \log p) = o(1)$ under Assumption 2. On the other hand, because $e(t)$ has mean zero, we have $\Delta_t = n^{-1} \sum_{i=1}^n e_i(t)\,(H_{ii} - p/n)$, which helps to derive another upper bound on Δ. Define the maximum absolute deviation of the H_{ii}'s from their average as

$$\kappa_0 = \max_{1 \leq i \leq n} |H_{ii} - p/n|,$$

and then we can use the Cauchy–Schwarz inequality to obtain

$$\Delta = \max_{t=0,1} |\Delta_t| \leq \kappa_0 \max_{t=0,1} n^{-1} \sum_{i=1}^n |e_i(t)| \leq \kappa_0 \mathcal{E}_2^{1/2}.$$

So another sufficient condition is $\kappa_0 = o(n^{-1/2})$. This condition implies that $\kappa \leq \kappa_0 + p/n = o(n^{-1/2}) + p/n$, which, coupled with $p = o\{n^{2/3}/(\log n)^{1/3}\}$, implies $\kappa^2 p \log p = o(1)$. The following corollary summarizes the results from the above discussion.

COROLLARY 2. *Under Assumptions* 1–4*,* $n^{1/2}(\hat{\tau}_{\mathrm{adj}} - \tau)/\sigma_n \rightsquigarrow N(0, 1)$ *if either (i)* $\kappa p = o(1)$ *or (ii)* $p = o\{n^{2/3}/(\log n)^{1/3}\}$ *and* $\kappa_0 = o(n^{-1/2})$*.*

Consider the favorable case with $\kappa = O(p/n)$. Condition (i) reduces to $p = o(n^{1/2})$, so Corollary 2 extends Lin (2013)'s result to $p = o(n^{1/2})$ without any further assumptions. Condition (ii) states that when all the leverage scores are within an $o(n^{-1/2})$ neighborhood of their average p/n, the requirement on p can be relaxed to $o\{n^{2/3}/(\log n)^{1/3}\}$. Supplementary Material II shows that when the x_is are realizations of multivariate normal vectors as assumed by Wager et al. (2016), the leverage score conditions hold with high probability.

Although we can relax the constraint on the dimension p under condition (ii), it is not ideal to impose an extra condition on the leverage scores. When $p > n^{1/2}$, the leverage score condition is more stringent than that in the favorable case. By contrast, the debiased estimator is asymptotically normal without any additional condition.

THEOREM 4. *Under Assumptions* 1–4*,* $n^{1/2}(\hat{\tau}^{\mathrm{de}}_{\mathrm{adj}} - \tau)/\sigma_n \rightsquigarrow N(0, 1)$ *if* $\kappa^2 p \log p = o(1)$*.*

In the favorable case with $\kappa = O(p/n)$, the condition in Theorem 4 reduces to $p^3 \log p/n^2 = o(1)$, which permits p to grow as fast as $o\{n^{2/3}/(\log n)^{1/3}\}$, verifying the claim in Section 1. In general, it is strictly weaker than the condition in Theorem 3, which relies on an extra assumption that $n\Delta^2 = o(\mathcal{E}_2)$. In the favorable case, as shown in Corollary 2, Theorem 4 removes the condition on κ_0.

The variance estimators $\hat{\sigma}^2_{\mathrm{HC}j}$'s are all asymptotically equivalent because the correction terms are negligible under our asymptotic regime. They are asymptotically conservative estimators of σ_n^2, so the Wald-type confidence intervals for τ are all asymptotically conservative.

THEOREM 5. *Under Assumptions* 1–4*, there exists a non-negative sequence* $a_n = o_{\mathbb{P}}(1)$ *which depends on* $\{x_i, Y_i(1), Y_i(0)\}_{i=1}^n$ *such that* $\hat{\sigma}^2_{\mathrm{HC}j}/\sigma_n^2 \geq 1 - a_n$ *for all* $j \in \{0, 1, 2, 3\}$*.*

3.4. *Comparison with existing results*

Theoretical analyses under the finite-population randomization model are challenging due to the lack of probability tools. The closest work to ours is Bloniarz et al. (2016), which allows p to grow with n and potentially exceed n. However, they assume that the potential outcomes have sparse linear representations based on the covariates, and require $s = o(n^{1/2}/\log p)$ where s is a measure of sparsity. Under additional regularities conditions, they show that $\hat{\tau}(\hat{\beta}_1^{\mathrm{lasso}}, \hat{\beta}_0^{\mathrm{lasso}})$ is consistent and asymptotically normal with $(\hat{\beta}_1^{\mathrm{lasso}}, \hat{\beta}_0^{\mathrm{lasso}})$ being the LASSO coefficients of the covariates. Although the LASSO-adjusted estimator can handle ultra-high dimensional case where $p \gg n$, it has three limitations. First, the requirement $s \ll n^{1/2}/\log p$ is stringent. Second, the penalty level of the LASSO depends on unobserved quantities. Although they use the cross-validation to select the penalty level, the theoretical properties of this procedure is still unclear. Third, their "restrictive eigenvalue condition" imposes certain non-singularity on the submatrices of the covariate matrix. However, the covariate matrix can be ill-conditioned especially when interaction terms of the basic covariates are included in practice. In addition, this condition is computationally challenging to check. Although our results cannot deal with the case of $p > n$, we argue that $p < n$ without sparsity is an important regime in many applications.

Due to the numerical equivalence of the regression-adjusted estimator to the ordinary least squares estimator, it is attempting to view our theory as a special case of the existing literature on high dimensional linear models (e.g. Huber, 1973; Portnoy, 1985; Mammen, 1989; Lei et al., 2018; Cattaneo et al., 2018). However, the two approaches are fundamentally different. They assume a linear model for the observed outcomes $Y_i^{\mathrm{obs}} = \alpha + T_i\tau + x_i^{\mathsf{T}}\beta + \epsilon_i$, where T_i denotes the treatment indicator, x_i denotes the covariates to be adjusted for, and ϵ_i denotes the random error for unit i. Under their framework, the linear model must be correctly-specified with the random error ϵ_i being an important component in statistical inference. Moreover, a linear model implicitly assumes treatment-unit additivity, that is, the treatment effect is either constant or uncorrelated with covariates. By contrast, we do not assume any correctly specified linear model for the potential outcomes but treat them as fixed quantities instead. Neyman (1923/1990)'s model allows for arbitrary treatment effect heterogeneity which suggests that the additive linear model is an inadequate specification (Freedman, 2008). Therefore, the results in this paper are distinct from those assuming linear models; they are not directly comparable. Similarly, although Wager et al. (2016) relax the assumption of the constant treatment effect in linear models and can handle the high dimensional case with sparsity level $s = o(n/\log p)$ or $p/n \to \gamma \in (0, \infty)$, their theory requires X to be normal and $Y(t)$ to be a homoskedastic linear model of X. By contrast, our analysis needs none of these assumptions.

4. NUMERICAL EXPERIMENTS

4.1. *Data generating process*

To confirm and complement our theory, we use extensive numerical experiments to examine the finite-sample performance of the estimators $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}^{\text{de}}_{\text{adj}}$ as well as the variance estimators $\hat{\sigma}^2_{\text{HC}j}$ for $j = 0, 1, 2, 3$. To save space, we only present the results for one synthetic data and relegate the results for other synthetic data to Supplementary Material III.

We set $n = 2000, n_1 = n\pi_1$ with $\pi_1 = 0.2$ and generate a matrix $\mathcal{X} \in \mathbb{R}^{n \times n}$ with independent and identically distributed entries from $t(2)$. We only generate one copy of X per experiment and keep it fixed. For each exponent $\gamma \in \{0, 0.05, \ldots, 0.7\}$, we let $p = \lceil n^{\gamma} \rceil$ and take the first p columns of $\mathcal{X}$ as the covariate matrix. In Supplementary Material III, we also simulate X with $N(0, 1)$ and $t(1)$ entries with both $\pi_1 \in \{0.2, 0.5\}$. We select $t(2)$ distribution for presentation because it is neither too idealized as $N(0, 1)$, for which $\kappa \sim p/n$, nor too irregular as $t(1)$. It is helpful to illustrate and complement our theory.

With X, we construct the potential outcomes from $Y(1) = X\beta^*_1 + \epsilon(1)$ and $Y(0) = X\beta^*_0 + \epsilon(0)$. Because $\hat{\beta}_t - \beta^*_t = (X_t^{\mathsf{T}} X_t)^{-1} X_t^{\mathsf{T}} \epsilon(t)$ does not depend on β^*_t, we take $\beta^*_1 = \beta^*_0 = 0 \in \mathbb{R}^p$ without changing the bias, variance and coverage properties of the estimates $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}^{\text{de}}_{\text{adj}}$. We generate $\{\epsilon(1), \epsilon(0)\}$ as realizations of random vectors with independent and identically distributed entries from $N(0, 1)$, $t(2)$, or $t(1)$. We also consider another case with $\epsilon(1) = \epsilon(0)$ that corresponds to the sharp null hypothesis in Supplementary Material III. Given $X \in \mathbb{R}^{n \times p}$ and potential outcomes $Y(1), Y(0) \in \mathbb{R}^n$, we generate 5000 binary vectors $T \in \mathbb{R}^n$, and for each T, we observe half of the potential outcomes.

4.2. *Repeated sampling evaluations*

Based on the observed data, we obtain two estimates $\hat{\tau}_{\text{adj}}$ and $\hat{\tau}^{\text{de}}_{\text{adj}}$, as well as four variance estimates $\hat{\sigma}^2_{\text{HC}j}$ $(j = 0, 1, 2, 3)$ and the theoretical asymptotic variance σ^2_n. Below $\hat{\tau}$ can be either $\hat{\tau}_{\text{adj}}$ or $\hat{\tau}^{\text{de}}_{\text{adj}}$, and $\hat{\sigma}^2$ can be any of the five estimates. Let $\hat{\tau}_1, \ldots, \hat{\tau}_R$ denote the estimates in $R = 5000$ replicates, and τ denote the true average treatment effect. The empirical relative absolute bias is $n^{1/2}|R^{-1}\sum_{k=1}^{R} \hat{\tau}_k - \tau|/\sigma_n$. Similarly, let $\hat{\sigma}^2_1, \ldots, \hat{\sigma}^2_R$ denote the variance estimates obtained in R replicates, and $\hat{\sigma}^2_*$ denote the empirical variance of $(n^{1/2}\hat{\tau}_1, \ldots, n^{1/2}\hat{\tau}_R)$. We compute the standard deviation inflation ratio $R^{-1}\sum_{k=1}^{R} \hat{\sigma}_k/\hat{\sigma}_*$. Note that $\hat{\sigma}^2_*$ is an unbiased estimate of true sampling variance of $n^{1/2}\hat{\tau}$, which can be different from the theoretical asymptotic variance σ^2_n. For each estimate and variance estimate, we compute the t-statistic $n^{1/2}(\hat{\tau} - \tau)/\hat{\sigma}$. For each

t-statistic, we estimate the empirical 95% coverage rate by the proportion within $[-1.96, 1.96]$, the 95% quantile range of $N(0,1)$.

In summary, we compute three measures defined above: the relative bias, standard deviation inflation ratio, and 95% coverage rate. We repeat 50 times using different random seeds and record the medians of each measure. Fig. 1 summarizes the results.

4.3. *Results*

From Figure 1a, $\hat{\tau}_{\text{adj}}^{\text{de}}$ does reduce the bias regardless of the distribution of potential outcomes, especially for moderately large p. For standard deviation inflation ratios, the true sampling variances of $n^{1/2}\hat{\tau}_{\text{adj}}$ and $n^{1/2}\hat{\tau}_{\text{adj}}^{\text{de}}$ are almost identical and thus we set the sampling variance of $n^{1/2}\hat{\tau}_{\text{adj}}$ as the baseline variance $\hat{\sigma}_*^2$. Figure 1b shows an interesting phenomenon that the theoretical asymptotic variance σ_n^2 tends to underestimate the true sampling variance for large p. Theorem 1 partially suggests this. The theoretical asymptotic variance is simply the variance of $\hat{\tau}_e$ while the finite sample variance also involves the remainder, which can be large in the presence of high dimensional or influential observations. All variance estimators overestimate σ_n^2 because they all ignore the third term of σ_n^2. However, all estimators, except the HC3 estimator, tend to underestimate the true sampling variance for large p. By contrast, the HC3 estimator does not suffer from anti-conservatism in this case.

Figures 1b shows that HC0 and HC1 variance estimates lie between the theoretical asymptotic variance and the HC2 variance estimate. For better visualization, Figures 1c only shows the 95% coverage rates of t-statistics computed from $\sigma_n^2, \hat{\sigma}_{\text{HC2}}^2$ and $\hat{\sigma}_{\text{HC3}}^2$, based on which we draw the following conclusions. First, as we pointed out previously, the coverage rates based on two estimates are almost identical because the relative bias is small in these scenarios. Second, as Figures 1b suggests, the t-statistic with HC3 variance estimate has the best coverage rate, which is robust with covariates of an increasing dimension. By contrast, the theoretical asymptotic variance and the HCj $(j = 0, 1, 2)$ variance estimates yield significantly lower coverage rates for large p. We recommend $\hat{\sigma}_{\text{HC3}}^2$.

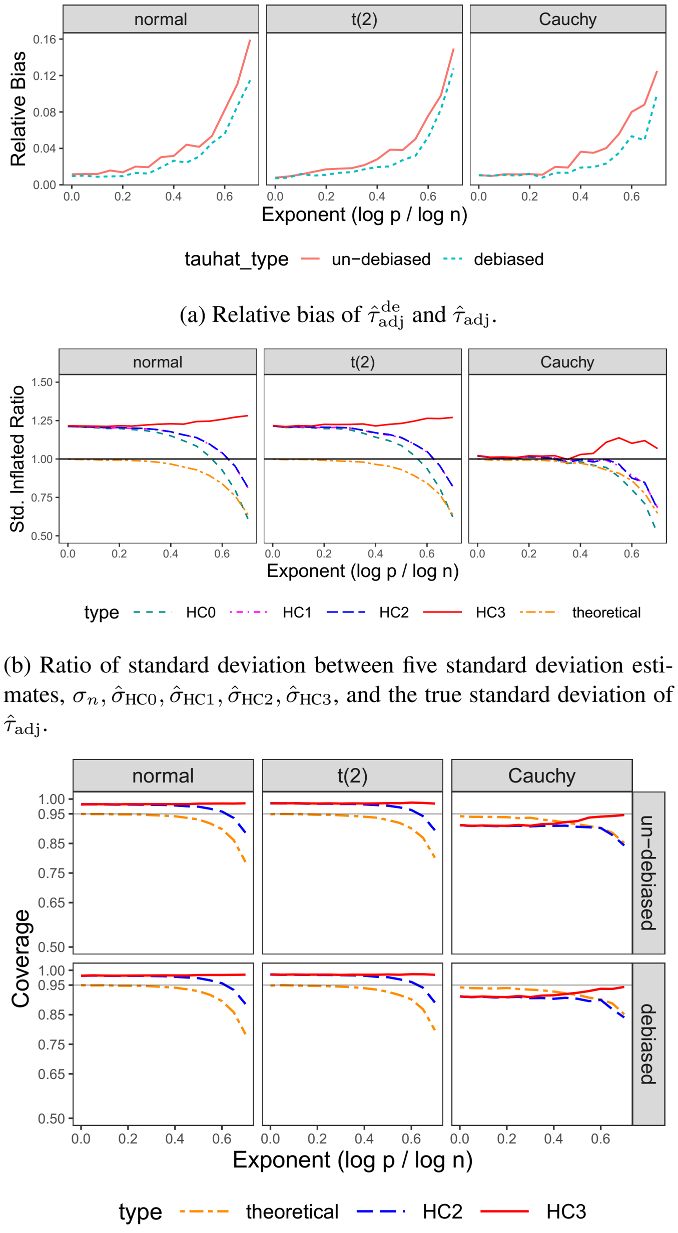
4.4. *Effectiveness of debiasing*

In the aforementioned settings, $\hat{\tau}_{\text{adj}}^{\text{de}}$ yields almost identical inference as $\hat{\tau}_{\text{adj}}$. This is not surprising because in the above scenarios the potential outcomes are generated from linear models and thus the regression-adjusted estimator has bias close to zero. However, in practice, the potential outcomes might not have prefect linear relationships with the covariates. To illustrate the potential benefits of debiasing, we consider the worst-case situation which maximizes the bias. Specifically, we consider the case where $\epsilon(0) = \epsilon$ and $\epsilon(1) = 2\epsilon$ for some vector ϵ that satisfies (5) with sample variance 1. To maximize the bias term, we take ϵ as the solution of

$$\max_{\epsilon \in \mathbb{R}^n} \left| \frac{n_1}{n_0}\Delta_0 - \frac{n_0}{n_1}\Delta_1 \right| = \max_{\epsilon \in \mathbb{R}^n} \left(\frac{2n_0}{n_1} - \frac{n_1}{n_0} \right) \left| \sum_{i=1}^n H_{ii}\epsilon_i \right|, \tag{16}$$

such that $\|\epsilon\|_2^2/n = 1$ and $X^{\mathsf{T}}\epsilon = \mathbb{1}^{\mathsf{T}}\epsilon = 0$. Supplementary Material III gives more details of constructing ϵ. From (16), the bias is amplified when the group sizes are unbalanced, and it effectively imposes a non-linear relationship between potential outcomes and covariates.

We perform simulation detailed in Section 4.2 based on potential outcomes in (16) and report the relative bias and coverage rate to demonstrate the effectiveness of debiasing. To save space, we only report the coverage rates based on $\hat{\sigma}_{\text{HC2}}^2$ and $\hat{\sigma}_{\text{HC3}}^2$. Fig. 2 summarizes the results. Unlike the previous settings, the relative bias in this setting is large enough to affect the coverage rate. The debiased estimator reduces a fair proportion of bias and improves the coverage

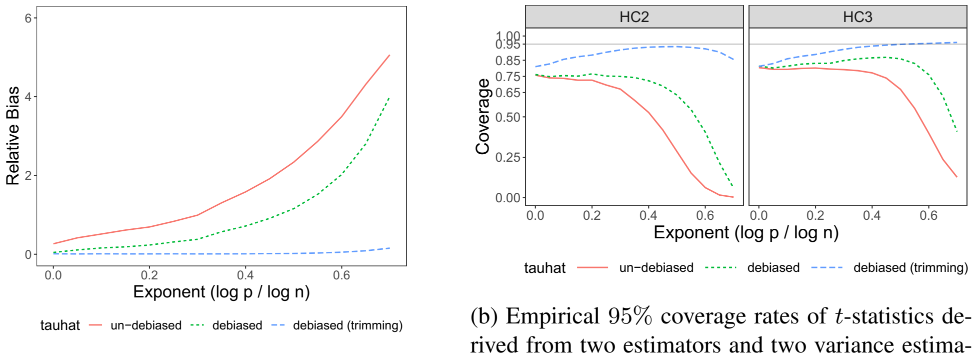


(a) Relative bias of $\hat{\tau}_{\text{adj}}^{\text{de}}$ and $\hat{\tau}_{\text{adj}}$.

(b) Ratio of standard deviation between five standard deviation estimates, $\sigma_n, \hat{\sigma}_{\text{HC0}}, \hat{\sigma}_{\text{HC1}}, \hat{\sigma}_{\text{HC2}}, \hat{\sigma}_{\text{HC3}}$, and the true standard deviation of $\hat{\tau}_{\text{adj}}$.

(c) Empirical 95% coverage rates of t-statistics derived from two estimators and four variance estimators ("theoretical" for σ_n^2, "HC2" for $\hat{\sigma}_{\text{HC2}}^2$ and "HC3" for $\hat{\sigma}_{\text{HC3}}^2$)

Fig. 1: Simulation with $\pi_1 = 0.2$. X is a realization of a random matrix with $t(2)$ entries, and $\epsilon(t)$ is a realization of a random vector with entries from a distribution corresponding to each column.



(a) Relative bias of $\hat{\tau}^{\mathrm{de}}_{\mathrm{adj}}$ and $\hat{\tau}_{\mathrm{adj}}$.

(b) Empirical 95% coverage rates of t-statistics derived from two estimators and two variance estimators ("HC2" for $\hat{\sigma}^2_{\mathrm{HC2}}$ and "HC3" for $\hat{\sigma}^2_{\mathrm{HC3}}$)

Fig. 2: Simulation. X is a realization of a random matrix with $t(2)$ entries, $\pi_1 = 0.2$ and $\epsilon(t)$ is defined in (16).

rate especially when the dimension is high. We provide experimental results in more settings in Supplementary Material III.

4.5. *Trimming covariates*

Because our theory holds even for mis-specified linear models, we can preprocess the covariate matrix X arbitrarily without changing the estimand, provided that the preprocessing step does not involve T or Y^{obs}. This is a feature of our finite-population theory. Moreover, our asymptotic theory suggests that the maximum leverage score of the design matrix affects the properties of $\hat{\tau}_{\mathrm{adj}}$ and $\hat{\tau}^{\mathrm{de}}_{\mathrm{adj}}$. When there are many influential observations, it is beneficial to reduce κ by trimming the values of covariates before regression adjustment. Importantly, trimming covariates should not use any information of T or Y^{obs}.

For the cases considered in previous subsections, we consider trimming each covariate at its 2.5% and 97.5% quantiles. For the 50 design matrices used in Section 4 with $p = \lceil n^{2/3} \rceil$ and $n = 2000$, the average of κ is 0.9558 with standard error 0.0384. After trimming, the average of κ reduces dramatically to 0.0704 with standard error 0.0212. Fig. 2 shows that the bias is significantly reduced and the coverage rate gets drastically improved after trimming covariates.

ACKNOWLEDGMENT

We thank the editor, associate editor, two reviewers, Cheng Gao, and Luke Miratrix for helpful suggestions. Peng Ding was partially supported by a grant from the U.S. National Science Foundation.

SUPPLEMENTARY MATERIAL

The supplementary material contains the technical details and additional numerical examples.

REFERENCES

BERK, R., PITKIN, E., BROWN, L., BUJA, A., GEORGE, E. & ZHAO, L. (2013). Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review* **37**, 170–196.

BLONIARZ, A., LIU, H., ZHANG, C.-H., SEKHON, J. S. & YU, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences* **113**, 7383–7390.

BOX, G. E. P., HUNTER, J. S. & HUNTER, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. New York: Wiley-Interscience.

CATTANEO, M. D., JANSSON, M. & NEWEY, W. K. (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association* **113**, 1350–1361.

DASGUPTA, T., PILLAI, N. S. & RUBIN, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B* **77**, 727–753.

DUFLO, E., GLENNERSTER, R. & KREMER, M. (2007). Using randomization in development economics research: A toolkit. In *Handbook of Development Econmics*, J. A. S. T. P. Schultz, ed., vol. 4, chap. 61. Elsevier, pp. 3895–3962.

FISHER, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver And Boyd, 1st ed.

FOGARTY, C. B. (2018). Regression assisted inference for the average treatment effect in paired experiments. *Biometrika* **105**, 994–1000.

FREEDMAN, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics* **40**, 180–193.

GERBER, A. S. & GREEN, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York: WW Norton.

HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science* **5**, 361–74.

HINKELMANN, K. & KEMPTHORNE, O. (2007). *Design and Analysis of Experiments, Introduction to Experimental Design*, vol. 1. New York: John Wiley & Sons.

HUBER, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics* **1**, 799–821.

IMBENS, G. W. & RUBIN, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments.* New York: Wiley.

LEI, L., BICKEL, P. J. & EL KAROUI, N. (2018). Asymptotics for high dimensional regression M-estimates: Fixed design results. *Probability Theory and Related Fields* **172**, 983–1079.

LI, X. & DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association* **112**, 1759–1769.

LI, X. & DING, P. (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society, Series B* **82**, 241–268.

LI, X., DING, P. & RUBIN, D. B. (2018). Asymptotic theory of rerandomization in treatment-control experiments. *Proceedings of the National Academy of Sciences* **115**, 9157–9162.

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics* **7**, 295–318.

MACKINNON, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In *Recent advances and future directions in causality, prediction, and specification analysis*. Springer, pp. 437–461.

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics* **17**, 382–400.

MIDDLETON, J. A. & ARONOW, P. M. (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy* **6**, 39–75.

MUKERJEE, R., DASGUPTA, T. & RUBIN, D. B. (2018). Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association* **113**, 868–881.

NEGI, A. & WOOLDRIDGE, J. M. (2020). Robust and efficient estimation of potential outcome means under random assignment. *arXiv:2010.01800* .

NEYMAN, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated by Dabrowska, D. M. and Speed, T. P. *Statistical Science* **5**, 465–472.

PORTNOY, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; II. Normal approximation. *The Annals of Statistics* **13**, 1403–1417.

ROSENBERGER, W. F. & LACHIN, J. M. (2015). *Randomization in Clinical Trials: Theory and Practice*. Hoboken, NJ: John Wiley & Sons.

TAN, Z. (2014). Second-order asymptotic theory for calibration estimators in sampling and missing-data problems. *Journal of Multivariate Analysis* **131**, 240–253.

TSIATIS, A., DAVIDIAN, M., ZHANG, M. & LU, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine* **27**, 4658–4677.

WAGER, S., DU, W., TAYLOR, J. & TIBSHIRANI, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences* **113**, 12673–12678.