Quantitative Economics 11 (2020), 957–982 1759-7331/20200957

Quantile treatment effects and bootstrap inference under covariate-adaptive randomization

YICHONG ZHANG
School of Economics, Singapore Management University

XIN ZHENG
School of Economics, Singapore Management University

In this paper, we study the estimation and inference of the quantile treatment effect under covariate-adaptive randomization. We propose two estimation methods: (1) the simple quantile regression and (2) the inverse propensity score weighted quantile regression. For the two estimators, we derive their asymptotic distributions uniformly over a compact set of quantile indexes, and show that, when the treatment assignment rule does not achieve strong balance, the inverse propensity score weighted estimator has a smaller asymptotic variance than the simple quantile regression estimator. For the inference of method (1), we show that the Wald test using a weighted bootstrap standard error underrejects. But for method (2), its asymptotic size equals the nominal level. We also show that, for both methods, the asymptotic size of the Wald test using a covariate-adaptive bootstrap standard error equals the nominal level. We illustrate the finite sample performance of the new estimation and inference methods using both simulated and real datasets.

KEYWORDS. Bootstrap inference, quantile treatment effect.

JEL CLASSIFICATION. C14, C21.

1. INTRODUCTION

The randomized control trial (RCT), as pointed out by Angrist and Pischke (2008), is one of the five most common methods (along with instrumental variable regressions, matching estimations, differences-in-differences, and regression discontinuity designs) for causal inference. Researchers can use the RCT to estimate not only average treatment effects (ATEs) but also quantile treatment effects (QTEs), which capture the heterogeneity of the sign and magnitude of treatment effects, varying depending on their place

Yichong Zhang: yczhang@smu.edu.sg
Xin Zheng: xin.zheng.2015@phdecons.smu.edu.sg

We are grateful to Federico Bugni, Qu Feng, Sukjin Han, Yu-Chin Hsu, Shakeeb Khan, Frank Kleibergen, Michael Leung, Jia Li, Wenjie Wang, and seminar participants at NTU, Financial Econometrics and New Finance Conference at Zhejiang University, SH3 Conference on Econometrics at SMU, 2019 Shanghai Workshop of Econometrics, and Asian Meeting of the Econometric Society for their valuable comments. We also thank the two anonymous referees for their valuable comments which greatly improve our paper. Zhang acknowledges the financial support from Singapore Ministry of Education Tier 2 grant under grant MOE2018-T2-2-169 and the Lee Kong Chian fellowship. Any and all errors are our own.

 https://doi.org/10.3982/QE1323

in the overall distribution of outcomes. For example, Muralidharan and Sundararaman (2011) estimated the QTE of teacher performance pay program on student learning via the difference of empirical quantiles of test scores between treatment and control groups. Duflo, Greenstone, Pande, and Ryan (2013) and Banerjee, Duflo, Glennerster, Kinnan (2015) estimated the QTEs of audits on endline pollution and a group-lending microcredit program on informal borrowing, respectively, via linear quantile regressions (QRs). Crépon, Devoto, Duflo, and Parienté (2015) estimated the QTE of microcredit on various household outcomes via a minimum distance method. Byrne, Nauze, and Martin (2018) estimated the QTE of being informed on energy use via the inverse propensity score weighted (IPW) QR. With the exception of Crépon et al. (2015), the other four papers all use the bootstrap to construct confidence intervals for their QTE estimates. However, RCTs have also been routinely implemented with covariate-adaptive randomization. Individuals are first stratified based on some baseline covariates, and then, within each stratum, the treatment status is assigned (independent of covariates) to achieve some balance between the sizes of treatment and control groups; as examples, see Imbens and Rubin (2015, Chapter 9) for a textbook treatment of the topic, and Duflo, Glennerster, and Kremer (2007) and Bruhn and McKenzie (2009) for two excellent surveys on implementing RCTs in development economics. To achieve such balance, treatment status for different individuals usually exhibits a (negative) cross-sectional *dependence*. The standard inference procedures that rely on cross-sectional *independence* are usually conservative and lacking power. How do we consistently estimate QTEs under covariate-adaptive randomization? What are the asymptotic distributions for the QTE estimators, and how do we conduct proper bootstrap inference? These questions are as yet unaddressed.

We propose two ways to estimate QTEs: (1) the simple quantile regression (SQR) and (2) the IPW QR. We establish the weak limits for both estimators uniformly over a compact set of quantile indexes and show that the IPW estimator has a smaller asymptotic variance than the SQR estimator when the treatment assignment rule does not achieve strong balance.[1] If strong balance is achieved, then the two estimators are asymptotically first-order equivalent. For inference, we show that the Wald test combined with weighted bootstrap based critical values can lead to underrejection for method (1), but its asymptotic size equals the nominal level for method (2). We also study the covariate-adaptive bootstrap which respects the cross-sectional dependence when generating the bootstrap sample. The estimator based on the covariate-adaptive bootstrap sample can mimic that of the original sample in terms of the standard error. Thus, using proper covariate-adaptive bootstrap based critical values, the asymptotic size of the Wald test equals the nominal level for both estimators.

As originally proposed by Doksum (1974), the QTE, for a fixed quantile index, corresponds to the horizontal difference between the marginal distributions of the potential outcomes for treatment and control groups. Firpo (2007) studied the identification and estimation of QTE under unconfoundedness. Our estimators (1) and (2) directly follow those in Doksum (1974) and Firpo (2007), respectively.

[1] We will define "strong balance" in Section 2.

Shao, Yu, and Zhong (2010) first pointed out that, under covariate-adaptive randomization, the usual two-sample t-test for the ATE is conservative. They then proposed a covariate-adaptive bootstrap which can produce the correct standard error. Shao and Yu (2013) extended the results to generalized linear models. However, both groups of researchers parametrized the (transformed) conditional mean equation by a specific linear model and focused on a specific randomization scheme (covariate-adaptive biased coin method). Ma, Qin, Li, and Hu (2018) derived the theoretical properties of ATE estimators based on general covariate-adaptive randomization under the linear model framework. Bugni, Canay, and Shaikh (2018) substantially generalized the framework to a fully nonparametric setting with a general class of randomization schemes. However, they mainly focused on the ATE and showed that the standard two-sample t-test and the t-test based on the linear regression with strata fixed effects are conservative. They then obtained analytical estimators for the correct standard errors and studied the validity of permutation tests. Hahn, Hirano, and Karlan (2011) studied the IPW estimator for the ATE under adaptive randomization. However, they assumed the treatment status is assigned completely independently across individuals. More recently, Bugni, Canay, and Shaikh (2019) studied the estimation of ATE with multiple treatments and proposed a fully saturated estimator. Tabord-Meehan (2018) studied the estimation of ATE under an adaptive randomization procedure.

Our paper complements the above papers in four aspects. First, we consider the estimation and inference of the QTE, which is a function of quantile index τ. We rely on the empirical processes theories developed by van der Vaart and Wellner (1996) and Chernozhukov, Chetverikov, and Kato (2014) to obtain uniformly weak convergence of our estimators over a compact set of τ. Based on the uniform convergence, we can construct not only point-wise but also uniform confidence bands. Second, we study the asymptotic properties of the IPW estimator under covariate-adaptive randomization. When the treatment assignment rule does not achieve strong balance, the IPW estimator is more efficient than the SQR estimator. Third, we investigate the weighted bootstrap approximation to the asymptotic distributions of the SQR and IPW estimators. We show that the weighted bootstrap ignores the (negative) cross-sectional dependence due to the covariate-adaptive randomization and overestimates the asymptotic variance for the SQR estimator. However, the asymptotic variance for the IPW estimator does not rely on the randomization scheme implemented. Thus, the asymptotic size of the Wald test using the IPW estimator paired with the weighted bootstrap based critical values equals the nominal level. Fourth, we investigate the covariate-adaptive bootstrap approximation to the asymptotic distributions of the SQR and IPW estimators. We establish that, using either estimator paired with its corresponding covariate-adaptive bootstrap based critical values, the asymptotic size of the Wald test equals the nominal level. Shao, Yu, and Zhong (2010) first proposed the covariate-adaptive bootstrap and establish its validity for the ATE in a linear regression model under the null hypothesis that the treatment effect is not only zero but also homogeneous.[2] We modify the covariate-adaptive bootstrap and establish its validity for the QTE in the nonparametric setting proposed by

[2]We say the average treatment effect is homogeneous if the conditional average treatment effect given covariates is the same as the unconditional one.

Bugni, Canay, and Shaikh (2018). In addition, our results do not rely on the homogeneity of the treatment effect. Compared with the analytical inference, the two bootstrap inferences for QTEs we study in this paper avoid estimating the infinite-dimensional nuisance parameters such as the densities of the potential outcomes, and thus, the choices of tuning parameters. In addition, unlike the permutation tests studied in Bugni, Canay, and Shaikh (2018), the validity of bootstrap inferences does not require either strong balance condition or studentization. In particular, such studentization is cumbersome in the QTE context.

As the asymptotic variance for the IPW estimator does not depend on the treatment assignment rule implemented in RCTs, this estimator (and equivalently, the fully saturated estimator for the ATE) is suitable for settings where the knowledge of the exact treatment assignment rule is not available. Such scenario occurs when researchers are using an experiment that was run in the past and the randomization procedure may not have been fully described. It also occurs in subsample analysis, where subgroups are defined using variables that may have not been used to form the strata and the treatment assignment rule for each subgroup becomes unknown. We illustrate this fact in the subsample analysis of the empirical application in Section 8.

The rest of the paper is organized as follows. In Section 2, we describe the model setup and notation. In Sections 3.1 and 3.2, we discuss the asymptotic properties of estimators (1) and (2), respectively. In Sections 4 and 5, we investigate the weighted and covariate-adaptive bootstrap approximations to the asymptotic distributions of estimators (1) and (2), respectively. In Section 6, we examine the finite-sample performance of the estimation and inference methods. In Section 7, we provide recommendations for practitioners. In Section 8, we apply the new methods to estimate and infer the average and quantile treatment effects of iron efficiency on educational attainment. In Section 9, we conclude. We provide proofs for all results in an Appendix in the Online Supplemental Material (Zhang and Zheng (2020)). We study the strata fixed effects quantile regression estimator and provide additional simulation results in the second online supplement located in the replication file.

2. Setup and notation

First, denote the potential outcomes for treated and control groups as $Y(1)$ and $Y(0)$, respectively. The treatment status is denoted as A, where $A = 1$ means treated and $A = 0$ means untreated. The researcher can only observe $\{Y_i, Z_i, A_i\}_{i=1}^n$ where $Y_i = Y_i(1)A_i + Y_i(0)(1 - A_i)$, and Z_i is a collection of baseline covariates. Strata are constructed from Z using a function $S : \text{Supp}(Z) \mapsto \mathcal{S}$, where $\mathcal{S}$ is a finite set. For $1 \leq i \leq n$, let $S_i = S(Z_i)$ and $p(s) = \mathbb{P}(S_i = s)$. Throughout the paper, we maintain the assumption that $p(s)$ is fixed w.r.t. n and is positive for every $s \in \mathcal{S}$.[3] We make the following assumption for the data generating process (DGP) and the treatment assignment rule.

[3]We can also allow for the DGP to depend on n so that $p_n(s) = \mathbb{P}_n(S_i = s)$ and $p(s) = \lim p_n(s)$. All the results in this paper still hold as long as $n(s) \to \infty$ a.s. Interested readers can refer to the previous version of this paper on arXiv for more details.

ASSUMPTION 1. (i) $\{Y_i(1), Y_i(0), S_i\}_{i=1}^n$ *is i.i.d.*

(ii) $\{Y_i(1), Y_i(0)\}_{i=1}^n \perp\!\!\!\perp \{A_i\}_{i=1}^n | \{S_i\}_{i=1}^n$.

(iii) $\{\{\frac{D_n(s)}{\sqrt{n}}\}_{s\in\mathcal{S}} | \{S_i\}_{i=1}^n\} \rightsquigarrow N(0, \Sigma_D)$ *a.s., where*

$$D_n(s) = \sum_{i=1}^n (A_i - \pi) 1\{S_i = s\} \quad \textit{and} \quad \Sigma_D = \operatorname{diag}\big\{p(s)\gamma(s) : s \in \mathcal{S}\big\}$$

with $0 \le \gamma(s) \le \pi(1-\pi)$.

(iv) $\frac{D_n(s)}{n(s)} = o_p(1)$ *for* $s \in \mathcal{S}$, *where* $n(s) = \sum_{i=1}^n 1\{S_i = s\}$.

Several remarks are in order. First, Assumptions 1(i)–1(iii) are exactly the same as Bugni, Canay, and Shaikh (2018, Assumption 2.2). We refer interested readers to Bugni, Canay, and Shaikh (2018) for more discussion of these assumptions. Second, note that in Assumption 1(iii) the parameter π is the target proportion of treatment for each stratum and $D_n(s)$ measures the imbalance. Bugni, Canay, and Shaikh (2019) studied the more general case that π can take distinct values for different strata. Third, we follow the terminology in Bugni, Canay, and Shaikh (2018), which follows that of Efron (1971) and Hu and Hu (2012), saying a treatment assignment rule achieves strong balance if $\gamma(s) = 0$. Fourth, we do not require that the treatment status is assigned independently. Instead, we only require Assumption 1(iii) or Assumption 1(iv), which condition is satisfied by several treatment assignment rules such as simple random sampling (SRS), biased-coin design (BCD), adaptive biased-coin design (WEI), and stratified block randomization (SBR). Bugni, Canay, and Shaikh (2018, Section 3) provided an excellent summary of these four examples. For completeness, we briefly repeat their descriptions below. Note that both BCD and SBR assignment rules achieve strong balance. Last, as $p(s) > 0$, Assumption 1(iii) implies Assumption 1(iv).

EXAMPLE 1 (SRS). Let $\{A_i\}_{i=1}^n$ be drawn independently across i and of $\{S_i\}_{i=1}^n$ as Bernoulli random variables with success rate π, that is, for $k = 1, \ldots, n$,

$$\mathbb{P}\big(A_k = 1 | \{S_i\}_{i=1}^n, \{A_j\}_{j=1}^{k-1}\big) = \mathbb{P}(A_k = 1) = \pi.$$

Then Assumption 1(iii) holds with $\gamma(s) = \pi(1-\pi)$.

EXAMPLE 2 (WEI). The design is first proposed by Wei (1978). Let $n_{k-1}(S_k) = \sum_{i=1}^{k-1} 1\{S_i = S_k\}$, $D_{k-1}(s) = \sum_{i=1}^{k-1}(A_i - \frac{1}{2})1\{S_i = s\}$, and

$$\mathbb{P}\big(A_k = 1 | \{S_i\}_{i=1}^k, \{A_i\}_{i=1}^{k-1}\big) = \phi\left(\frac{D_{k-1}(S_k)}{n_{k-1}(S_k)}\right),$$

where $\phi(\cdot) : [-1, 1] \mapsto [0, 1]$ is a prespecified nonincreasing function satisfying $\phi(-x) = 1 - \phi(x)$. Here, $\frac{D_0(S_1)}{0}$ is understood to be zero. Then Bugni, Canay, and Shaikh (2018) showed that Assumption 1(iii) holds with $\pi = \frac{1}{2}$ and $\gamma(s) = \frac{1}{4}(1 - 4\phi'(0))^{-1}$.

EXAMPLE 3 (BCD). The treatment status is determined sequentially for $1 \le k \le n$ as

$$\mathbb{P}\big(A_k = 1 | \{S_i\}_{i=1}^k, \{A_i\}_{i=1}^{k-1}\big) = \begin{cases} \frac{1}{2} & \text{if } D_{k-1}(S_k) = 0, \\ \lambda & \text{if } D_{k-1}(S_k) < 0, \\ 1-\lambda & \text{if } D_{k-1}(S_k) > 0, \end{cases}$$

where $D_{k-1}(s)$ is defined as above and $\frac{1}{2} < \lambda \le 1$. Then Bugni, Canay, and Shaikh (2018) showed that Assumption 1(iii) holds with $\pi = \frac{1}{2}$ and $\gamma(s) = 0$.

EXAMPLE 4 (SBR). For each stratum, $\lfloor \pi n(s) \rfloor$ units are assigned to treatment and the rest is assigned to control. Bugni, Canay, and Shaikh (2018) then showed that Assumption 1(iii) holds with $\gamma(s) = 0$.

Our parameter of interest is the τth QTE defined as

$$q(\tau) = q_1(\tau) - q_0(\tau),$$

where $\tau \in (0,1)$ is a quantile index and $q_j(\tau)$ is the τth quantile of random variable $Y(j)$ for $j = 0, 1$. For inference, although we mainly focus on the Wald test for the null hypothesis that $q(\tau)$ equals some particular value, our method can also be used to test hypotheses involving multiple or even a continuum of quantile indexes. The following regularity conditions are common in the literature of quantile estimations.

ASSUMPTION 2. *For $j = 0, 1$, denote $f_j(\cdot)$ and $f_j(\cdot|s)$ as the PDFs of $Y_i(j)$ and $Y_i(j)|S_i = s$, respectively.*

(i) *$f_j(q_j(\tau))$ and $f_j(q_j(\tau)|s)$ are bounded and bounded away from zero uniformly over $\tau \in \Upsilon$ and $s \in \mathcal{S}$, where Υ is a compact subset of $(0,1)$.*

(ii) *$f_j(\cdot)$ and $f_j(\cdot|s)$ are Lipschitz over $\{q_j(\tau) : \tau \in \Upsilon\}$.*

3. ESTIMATION

3.1 *Simple quantile regression*

In this section, we propose to estimate $q(\tau)$ by a QR of Y_i on A_i. Denote $\beta(\tau) = (\beta_0(\tau), \beta_1(\tau))'$, $\beta_0(\tau) = q_0(\tau)$, and $\beta_1(\tau) = q(\tau)$. We estimate $\beta(\tau)$ by $\hat{\beta}(\tau)$, where

$$\hat{\beta}(\tau) = \underset{b=(b_0,b_1)' \in \Re^2}{\arg\min} \sum_{i=1}^n \rho_\tau\big(Y_i - \dot{A}_i' b\big),$$

$\dot{A}_i = (1, A_i)'$, and $\rho_\tau(u) = u(\tau - 1\{u \le 0\})$ is the standard check function. We refer to $\hat{\beta}_1(\tau)$, the second element of $\hat{\beta}(\tau)$, as our SQR estimator for the τth QTE. As A_i is a dummy variable, $\hat{\beta}_1(\tau)$ is numerically the same as the difference between the τth empirical quantiles of Y in the treatment and control groups.

THEOREM 3.1. *If Assumptions* 1(i)–1(iii) *and* 2 *hold, then uniformly over* $\tau \in \Upsilon$,

$$\sqrt{n}\big(\hat{\beta}_1(\tau) - q(\tau)\big) \rightsquigarrow \mathcal{B}_{\mathrm{sqr}}(\tau), \quad \text{as } n \to \infty,$$

where $\mathcal{B}_{\mathrm{sqr}}(\cdot)$ *is a Gaussian process with covariance kernel* $\Sigma_{\mathrm{sqr}}(\cdot,\cdot)$. *The expression for* $\Sigma_{\mathrm{sqr}}(\cdot,\cdot)$ *can be found in the Appendix.*

The asymptotic variance for $\sqrt{n}(\hat{\beta}_1(\tau) - \beta_1(\tau))$ is $\zeta_Y^2(\pi,\tau) + \zeta_A^2(\pi,\tau) + \zeta_S^2(\tau)$, where

$$\zeta_Y^2(\pi,\tau) = \frac{\tau(1-\tau) - \mathbb{E}m_1^2(S,\tau)}{\pi f_1^2\big(q_1(\tau)\big)} + \frac{\tau(1-\tau) - \mathbb{E}m_0^2(S,\tau)}{(1-\pi) f_0^2\big(q_0(\tau)\big)},$$

$$\zeta_A^2(\pi,\tau) = \mathbb{E}\gamma(S)\left(\frac{m_1(S,\tau)}{\pi f_1\big(q_1(\tau)\big)} + \frac{m_0(S,\tau)}{(1-\pi) f_0\big(q_0(\tau)\big)}\right)^2,$$

$$\zeta_S^2(\tau) = \mathbb{E}\left(\frac{m_1(S,\tau)}{f_1\big(q_1(\tau)\big)} - \frac{m_0(S,\tau)}{f_0\big(q_0(\tau)\big)}\right)^2,$$

and $m_j(s,\tau) = \mathbb{E}(\tau - 1\{Y(j) \le q_j(\tau)\}|S = s)$. Note that, if the treatment assignment rule achieves strong balance or the stratification is irrelevant[4] then $\zeta_A^2(\pi,\tau) = 0$.

3.2 *Inverse propensity score weighted quantile regression*

Denote $\hat{\pi}(s) = n_1(s)/n(s)$, $n_1(s) = \sum_{i=1}^n A_i 1\{S_i = s\}$, and $n(s) = \sum_{i=1}^n 1\{S_i = s\}$. Note $\hat{\pi}(S_i)$ is an estimator for the propensity score, that is, π. In addition, Assumption 1(ii) implies that the unconfoundedness condition holds. Thus, following the lead of Firpo (2007), we can estimate $q_j(\tau)$ by the IPW QR. Let

$$\hat{q}_1(\tau) = \arg\min_q \frac{1}{n}\sum_{i=1}^n \frac{A_i}{\hat{\pi}(S_i)}\rho_\tau(Y_i - q) \quad \text{and} \quad \hat{q}_0(\tau) = \arg\min_q \frac{1}{n}\sum_{i=1}^n \frac{1-A_i}{1-\hat{\pi}(S_i)}\rho_\tau(Y_i - q).$$

We then estimate $q(\tau)$ by $\hat{q}(\tau) = \hat{q}_1(\tau) - \hat{q}_0(\tau)$.

THEOREM 3.2. *If Assumptions* 1(i), 1(ii), 1(iv), *and* 2 *hold, then uniformly over* $\tau \in \Upsilon$,

$$\sqrt{n}\big(\hat{q}(\tau) - q(\tau)\big) \rightsquigarrow \mathcal{B}_{\mathrm{ipw}}(\tau), \quad \text{as } n \to \infty,$$

where $\mathcal{B}_{\mathrm{ipw}}(\cdot)$ *is a scalar Gaussian process with covariance kernel* $\Sigma_{\mathrm{ipw}}(\cdot,\cdot)$. *The expression for* $\Sigma_{\mathrm{ipw}}(\cdot,\cdot)$ *can be found in the Appendix.*

Two remarks are in order. First, the asymptotic variance for $\hat{q}(\tau)$ is

$$\zeta_Y^2(\pi,\tau) + \zeta_S^2(\tau).$$

[4]It means $\mathbb{P}(Y(j) \le q_j(\tau)|S = s) = \tau$ for $s \in \mathcal{S}$, $j = 0, 1$.

When strong balance is not achieved and the stratification is relevant, we have $\zeta_A^2(\pi,\tau) > 0$. Thus, $\hat{q}(\tau)$ is more efficient than $\hat{\beta}_1(\tau)$ in the sense that

$$\Sigma_{\text{ipw}}(\tau,\tau) < \Sigma_{\text{sqr}}(\tau,\tau).$$

When strong balance is achieved ($\gamma(s) = 0$), we have $\zeta_A^2(\pi,\tau) = 0$. Thus, the two estimators are asymptotically first-order equivalent. Based on the same argument, one can potentially prove that, when strong balance is not achieved and the stratification is relevant, the IPW estimator for ATE has strictly smaller asymptotic variance than the simple two-sample difference and strata fixed effects estimators studied by Bugni, Canay, and Shaikh (2018), and is asymptotically equivalent to the fully saturated linear regression estimator proposed by Bugni, Canay, and Shaikh (2019). Second, since the amount of "balance" of the treatment assignment rule does not play a role in the limiting distribution of the IPW estimator, Assumption 1(iii) is replaced by Assumption 1(iv).

4. Weighted bootstrap

In this section, we approximate the asymptotic distributions of the SQR and IPW estimators via the weighted bootstrap. Let $\{\xi_i\}_{i=1}^n$ be a sequence of bootstrap weights which will be specified later. Further denote $n_1^w(s) = \sum_{i=1}^n \xi_i A_i 1\{S_i = s\}$, $n^w(s) = \sum_{i=1}^n \xi_i 1\{S_i = s\}$, and $\hat{\pi}^w(s) = n_1^w(s)/n^w(s)$. The weighted bootstrap counterparts for the two estimators we study in this paper can then be written respectively as

$$\hat{\beta}^w(\tau) = \arg\min_b \sum_{i=1}^n \xi_i \rho_\tau\big(Y_i - \dot{A}_i' b\big)$$

and

$$\hat{q}^w(\tau) = \hat{q}_1^w(\tau) - \hat{q}_0^w(\tau),$$

where

$$\hat{q}_1^w(\tau) = \arg\min_q \sum_{i=1}^n \frac{\xi_i A_i}{\hat{\pi}^w(S_i)} \rho_\tau(Y_i - q) \quad \text{and} \quad \hat{q}_0^w(\tau) = \arg\min_q \sum_{i=1}^n \frac{\xi_i (1 - A_i)}{1 - \hat{\pi}^w(S_i)} \rho_\tau(Y_i - q).$$

The second element $\hat{\beta}_1^w(\tau)$ of $\hat{\beta}^w(\tau)$ and $\hat{q}^w(\tau)$ are the SQR and IPW bootstrap estimators for the τth QTE, respectively. Next, we specify the bootstrap weights.

Assumption 3. *Suppose $\{\xi_i\}_{i=1}^n$ is a sequence of nonnegative i.i.d. random variables with unit expectation and variance and a subexponential upper tail.*

The nonnegativity is required to maintain the convexity of the quantile regression objective function. The other conditions in Assumption 3 are common for the weighted bootstrap approximation. In practice, we generate $\{\xi_i\}_{i=1}^n$ by the standard exponential distribution. The corresponding weighted bootstrap is also known as the Bayesian bootstrap.

THEOREM 4.1. *If Assumptions* 1(i)–1(iii), 2, *and* 3 *hold, then uniformly over* $\tau \in \Upsilon$ *and conditionally on data,*

$$\sqrt{n}\big(\hat{\beta}_1^w(\tau) - \hat{\beta}_1(\tau)\big) \rightsquigarrow \tilde{\mathcal{B}}_{\mathrm{sqr}}(\tau), \quad \text{as } n \to \infty,$$

where $\tilde{\mathcal{B}}_{\mathrm{sqr}}(\tau)$ *is a Gaussian process. In addition,* $\tilde{\mathcal{B}}_{\mathrm{sqr}}(\tau)$ *shares the same covariance kernel with* $\mathcal{B}_{\mathrm{sqr}}(\tau)$ *defined in Theorems* 3.1 *with* $\gamma(s)$ *there replaced by* $\pi(1-\pi)$.

If Assumptions 1(i), 1(ii), 1(iv), 2, *and* 3 *hold, then uniformly over* $\tau \in \Upsilon$ *and conditionally on data,*

$$\sqrt{n}\big(\hat{q}^w(\tau) - \hat{q}(\tau)\big) \rightsquigarrow \mathcal{B}_{\mathrm{ipw}}(\tau), \quad \text{as } n \to \infty,$$

where $\mathcal{B}_{\mathrm{ipw}}(\tau)$ *is the same Gaussian process defined in Theorem* 3.2.

Four remarks are in order. First, the weighted bootstrap sample does not preserve the negative cross-sectional dependence in the original sample. Asymptotic variances of the weighted bootstrap estimators equal those of their original sample counterparts as if SRS is applied. In fact, the asymptotic variance for $\hat{\beta}_1^w(\tau)$ is

$$\zeta_Y^2(\pi,\tau) + \tilde{\zeta}_A^2(\pi,\tau) + \zeta_S^2(\tau),$$

where

$$\tilde{\zeta}_A^2(\pi,\tau) = \mathbb{E}\pi(1-\pi)\left(\frac{m_1(S,\tau)}{\pi f_1\big(q_1(\tau)\big)} + \frac{m_0(S,\tau)}{(1-\pi) f_0\big(q_0(\tau)\big)}\right)^2.$$

This asymptotic variance is intuitive as the weight ξ_i is independent with each other, which implies that, conditionally on data, the bootstrap sample observations are independent. As $\gamma(s) \le \pi(1-\pi)$, we have

$$\zeta_A^2(\pi,\tau) \le \tilde{\zeta}_A^2(\pi,\tau).$$

If the inequality is strict, then the weighted bootstrap overestimates the asymptotic variance of the SQR estimator, and thus, the Wald test constructed using the SQR estimator and its weighted bootstrap standard error is conservative.

Second, the asymptotic distribution of the weighted bootstrap IPW estimator coincides with that of the original estimator. The asymptotic size of the Wald test constructed using the IPW estimator and its weighted bootstrap standard error then equals the nominal level. Theorem 3.2 shows that the asymptotic variance for $\hat{q}(\tau)$ is invariant in the treatment assignment rule applied. Thus, even though the weighted bootstrap sample ignores the cross-sectional dependence and behaves as if the treatment status is generated randomly, the asymptotic variance for $\hat{q}^w(\tau)$ is still

$$\zeta_Y^2(\pi,\tau) + \zeta_S^2(\tau).$$

Third, the validity of weighted bootstrap for the IPW estimator only requires Assumption 1(iv) instead of 1(iii), for the same reason mentioned after Theorem 3.2.

Fourth, it is possible to consider the conventional nonparametric bootstrap which generates the bootstrap sample from the empirical distribution of the data. If the observations are i.i.d., van der Vaart and Wellner (1996, Section 3.6) showed that the conventional bootstrap is first-order equivalent to a weighted bootstrap with Poisson(1) weights. However, in the current setting, $\{A_i\}_{i\geq 1}$ is dependent. It is technically challenging to rigorously show that the above equivalence still holds. We leave it as an interesting topic for future research.

5. Covariate-adaptive bootstrap

In this section, we consider the covariate-adaptive bootstrap procedure as follows:

(i) Draw $\{S_i^*\}_{i=1}^n$ from the empirical distribution of $\{S_i\}_{i=1}^n$ with replacement.

(ii) Generate $\{A_i^*\}_{i=1}^n$ based on $\{S_i^*\}_{i=1}^n$ and the treatment assignment rule.

(iii) For $A_i^* = a$ and $S_i^* = s$, draw Y_i^* from the empirical distribution of Y_i given $A_i = a$ and $S_i = s$ with replacement.

First, Step (i) is the conventional nonparametric bootstrap. The bootstrap sample $\{S_i^*\}_{i=1}^n$ is obtained by drawing from the empirical distribution of $\{S_i\}_{i=1}^n$ with replacement n times. Second, Step (ii) follows the treatment assignment rule, and thus preserves the cross-sectional dependence structure in the bootstrap sample, even after conditioning on data. The weighted bootstrap sample, by contrast, is cross-sectionally independent given data. Third, Step (iii) applies the conventional bootstrap procedure to the outcome Y_i in the cell $(S_i, A_i) = (s, a) \in \mathcal{S} \times \{0, 1\}$. Given that the original data contain $n_a(s)$ observations in this cell, in this step, the bootstrap sample $\{Y_i^*\}_{i:A_i^*=a,S_i^*=s}$ is obtained by drawing from the empirical distribution of these $n_a(s)$ outcomes with replacement $n_a^*(s)$ times, where $n_a^*(s) = \sum_{i=1}^n 1\{A_i^* = a, S_i^* = s\}$. Unlike the conventional bootstrap, here both $n_a(s)$ and $n_a^*(s)$ are random and are not necessarily the same. Last, to implement the covariate-adaptive bootstrap, researchers need to know the treatment assignment rule for the original sample. Unlike observational studies, such information is usually available for RCTs. If one only knows that the treatment assignment rule achieves strong balance, then Theorem 5.1 below still holds, provided that the bootstrap sample is generated from any treatment assignment rule that achieves strong balance. Even worse, if no information on the treatment assignment rule is available, then one cannot implement the covariate-adaptive bootstrap inference. In this case, the weighted bootstrap for the IPW estimator can still provide a nonconservative Wald test, as shown in Theorem 4.1.

Using the bootstrap sample $\{Y_i^*, A_i^*, S_i^*\}_{i=1}^n$, we can estimate QTE by the two methods considered in the paper. Let $n_1^*(s) = \sum_{i=1}^n A_i^* 1\{S_i^* = s\}$, $n^*(s) = \sum_{i=1}^n 1\{S_i^* = s\}$, $\hat{\pi}^*(s) = \frac{n_1^*(s)}{n^*(s)}$, and $\dot{A}_i^* = (1, A_i^*)'$. Then the two bootstrap estimators can be written respectively as

$$\hat{\beta}^*(\tau) = \arg\min_b \sum_{i=1}^n \rho_\tau\big(Y_i^* - \dot{A}_i^* b\big)$$

and

$$\hat{q}^*(\tau) = \hat{q}_1^*(\tau) - \hat{q}_0^*(\tau),$$

where

$$\hat{q}_1^* = \arg\min_q \sum_{i=1}^n \frac{A_i^*}{\hat{\pi}^*(S_i^*)} \rho_\tau(Y_i^* - q) \quad \text{and} \quad \hat{q}_0^* = \arg\min_q \sum_{i=1}^n \frac{1 - A_i^*}{1 - \hat{\pi}^*(S_i^*)} \rho_\tau(Y_i^* - q).$$

The second element $\hat{\beta}_1^*(\tau)$ of $\hat{\beta}^*(\tau)$ and $\hat{q}^*(\tau)$ are the SQR and IPW bootstrap estimators for the τth QTE, respectively. Parallel to Assumption 1, we make the following assumption for the bootstrap sample.

ASSUMPTION 4. *Let* $D_n^*(s) = \sum_{i=1}^n (A_i^* - \pi) 1\{S_i^* = s\}$.

(i) $\{\{\frac{D_n^*(s)}{\sqrt{n}}\}_{s\in\mathcal{S}} | \{S_i^*\}_{i=1}^n\} \rightsquigarrow N(0, \Sigma_D)$ *a.s., where* $\Sigma_D = \operatorname{diag}\{p(s)\gamma(s) : s \in \mathcal{S}\}$.

(ii) $\sup_{s\in\mathcal{S}} \frac{|D_n^*(s)|}{\sqrt{n^*(s)}} = O_p(1)$, $\sup_{s\in\mathcal{S}} \frac{|D_n(s)|}{\sqrt{n(s)}} = O_p(1)$.

Assumption 4(i) is a high-level assumption. Obviously, it holds for SRS. For WEI, this condition holds by the same argument in Bugni, Canay, and Shaikh (2018, Lemma B.12) with the fact that $\frac{n^*(s)}{n(s)} \xrightarrow{p} 1$. For BCD, as shown in Bugni, Canay, and Shaikh (2018, Lemma B.11),

$$D_n^*(s) | \{S_i^*\}_{i=1}^n = O_p(1).$$

Therefore, $D_n^*(s)/\sqrt{n^*(s)} \xrightarrow{p} 0$ and Assumption 4(i) holds with $\gamma(s) = 0$. For SBR, it is clear that $|D_n^*(s)| \le 1$. Thus, Assumption 4(i) holds with $\gamma(s) = 0$ as well. In addition, as $p(s) > 0$, based on the standard bootstrap results, we have $n^*(s)/n \xrightarrow{p} p(s)$ and $n(s)/n \xrightarrow{p} p(s)$. Therefore, Assumption 4(i) is sufficient for Assumption 4(ii). Last, note that Assumption 4(ii) implies Assumption 1(iv).

THEOREM 5.1. *Suppose Assumptions* 1(i), 1(ii), 2, *and* 4(ii) *hold. Then, uniformly over* $\tau \in \Upsilon$ *and conditionally on data,*

$$\sqrt{n}\big(\hat{q}^*(\tau) - \hat{q}(\tau)\big) \rightsquigarrow \mathcal{B}_{\text{ipw}}(\tau), \quad \text{as } n \to \infty.$$

If, in addition, Assumptions 1(iii) *and* 4(i) *hold, then*

$$\sqrt{n}\big(\hat{\beta}_1^*(\tau) - \hat{q}(\tau)\big) \rightsquigarrow \mathcal{B}_{\text{sqr}}(\tau), \quad \text{as } n \to \infty.$$

Here, $\mathcal{B}_{\text{sqr}}(\tau)$ *and* $\mathcal{B}_{\text{ipw}}(\tau)$ *are two Gaussian processes defined in Theorems* 3.1 *and* 3.2, *respectively.*

Several remarks are in order. First, unlike the usual bootstrap estimator, the covariate-adaptive bootstrap SQR estimator is not centered around its corresponding counterpart from the original sample, but rather $\hat{q}(\tau)$. The reason is that the treatment

status A_i^* is not generated by bootstrap. In the linear expansion for the bootstrap estimator $\hat{\beta}_1^*(\tau)$, the part of the influence function that accounts for the variation generated by A_i^* need not be centered. We also know from the proof of Theorem 3.2 that $\hat{q}(\tau)$ do not have an influence function that represents the variation generated by A_i. Thus, $\hat{q}(\tau)$ can be used to center $\hat{\beta}_1^*(\tau)$.

Second, the choice of $\hat{q}(\tau)$ as the center is somehow ad-hoc. In fact, any estimator $\tilde{q}(\tau)$ that is first-order equivalent to $\hat{q}(\tau)$ in the sense that

$$\sup_{\tau\in Y}\left|\tilde{q}(\tau)-\hat{q}(\tau)\right| = o_p(1/\sqrt{n})$$

can serve as the center for the bootstrap estimators $\hat{q}^*(\tau)$ and $\hat{\beta}_1^*(\tau)$.

Third, when the treatment assignment rule achieves strong balance, $\hat{\beta}_1(\tau)$ and $\hat{q}(\tau)$ are first-order equivalent. In this case, $\hat{\beta}_1(\tau)$ can serve as the center for $\hat{\beta}_1^*(\tau)$ and various bootstrap inference methods are valid. On the other hand, when the treatment assignment rule does not achieve strong balance, $\hat{\beta}_1(\tau)$ and $\hat{q}(\tau)$ are not first-order equivalent. In this case, the asymptotic size of the percentile bootstrap for the SQR estimator using the quantiles of $\hat{\beta}_1^*(\tau)$ does not equal the nominal level. In the next section, we propose a way to compute the bootstrap standard error which does not depend on the choice of the center. Based on the bootstrap standard error, researchers can construct t-statistics and use standard normal critical values for inference.

Fourth, for ATE, we can use the same bootstrap sample to compute the standard errors for the simple and strata fixed effects estimators proposed in Bugni, Canay, and Shaikh (2018) as well as the IPW estimator. We expect that all the results in this paper hold for the ATE as well.

6. Simulation

We can summarize four bootstrap scenarios from the analysis in Sections 4 and 5: (i) the SQR estimator with the weighted bootstrap, (ii) the IPW estimator with either the weighted or covariate-adaptive bootstrap, (iii) the SQR estimator with the covariate-adaptive bootstrap when the assignment rule achieves strong balance, and (iv) the SQR estimator with the covariate-adaptive bootstrap when the assignment rule does not achieve strong balance. The results of Sections 4 and 5 imply that the bootstrap in scenario (i) produces conservative Wald-tests when the treatment assignment rule is not SRS. For scenarios (ii) and (iii), various bootstrap based inference methods are valid. However, for scenario (iv), researchers should be careful about the centering issue. In particular, the percentile bootstrap inference using the quantiles of $\hat{\beta}_1^*$ is invalid. In the following, we propose one single bootstrap inference method that works for scenarios (ii)–(iv). In addition, the proposed method does not require the knowledge of the centering.

We take the IPW estimator as an example. We can repeat the bootstrap estimation[5] B times and obtain B bootstrap IPW estimates, denoted as $\{\hat{q}_b^*(\tau)\}_{b=1}^B$. Further denote

[5]For the IPW estimator, we can use either the weighted or covariate-adaptive bootstrap. For the SQR estimator, we can only use the covariate-adaptive bootstrap.

$\hat{Q}(\alpha)$ as the α-th empirical quantile of $\{\hat{q}_b^*(\tau)\}_{b=1}^B$. We can test the null hypothesis that $q(\tau) = q^0(\tau)$ via $1\{|\frac{\hat{q}(\tau)-q^0(\tau)}{\hat{\sigma}_n^*}| > z_{1-\alpha/2}\}$, where $\hat{q}(\tau)$, $z_{1-\alpha/2}$, and $\hat{\sigma}_n^*$ are the IPW estimator, the $(1-\alpha/2)$-th quantile of the standard normal distribution, and

$$\hat{\sigma}_n^* = \frac{\hat{Q}(0.975) - \hat{Q}(0.025)}{z_{0.975} - z_{0.025}},$$

respectively. In scenarios (ii)–(iv), the asymptotic size of such test equals the nominal level α. In scenarios (ii) and (iii), we recommend the t-statistic and confidence interval using this particular bootstrap standard error (i.e., $\hat{\sigma}_n^*$) over other bootstrap inference methods (e.g., bootstrap confidence interval, percentile bootstrap confidence interval, etc.) because based on unreported simulations, they have better finite sample performance.

6.1 *Data generating processes*

We consider two DGPs with parameters $\gamma = 4$, $\sigma = 2$, and μ which will be specified later.

(i) Let Z be standardized Beta$(2, 2)$ distributed, $S_i = \sum_{j=1}^4 1\{Z_i \le g_j\}$, and $(g_1, \ldots, g_4) = (-0.25\sqrt{20}, 0, 0.25\sqrt{20}, 0.5\sqrt{20})$. The outcome equation is

$$Y_i = A_i\mu + \gamma Z_i + \eta_i,$$

where $\eta_i = \sigma A_i \varepsilon_{i,1} + (1 - A_i)\varepsilon_{i,2}$ and $(\varepsilon_{i,1}, \varepsilon_{i,2})$ are jointly standard normal.

(ii) Let Z be uniformly distributed on $[-2, 2]$, $S_i = \sum_{j=1}^4 1\{Z_i \le g_j\}$, and $(g_1, \ldots, g_4) = (-1, 0, 1, 2)$. The outcome equation is

$$Y_i = A_i\mu + A_i\nu_{i,1} + (1 - A_i)\nu_{i,0} + \eta_i,$$

where $\nu_{i,0} = \gamma Z_i^2 1\{|Z_i| \ge 1\} + \frac{\gamma}{4}(2 - Z_i^2)1\{|Z_i| < 1\}$, $\nu_{i,1} = -\nu_{i,0}$, $\eta_i = \sigma(1 + Z_i^2)A_i\varepsilon_{i,1} + (1 + Z_i^2)(1 - A_i)\varepsilon_{i,2}$, and $(\varepsilon_{i,1}, \varepsilon_{i,2})$ are mutually independent $T(3)/3$ distributed.

When $\pi = \frac{1}{2}$, for each DGP, we consider four randomization schemes:

(i) SRS: Treatment assignment is generated as in Example 1.

(ii) WEI: Treatment assignment is generated as in Example 2 with $\phi(x) = (1 - x)/2$.

(iii) BCD: Treatment assignment is generated as in Example 3 with $\lambda = 0.75$.

(iv) SBR: Treatment assignment is generated as in Example 4.

When $\pi \neq 0.5$, BCD is not defined while WEI is not defined in the original paper (Wei (1978)). Recently, Hu (2016) generalized the adaptive biased-coin design (i.e., WEI) to multiple treatment values and unequal target treatment ratios. Here, for $\pi \neq 0.5$, we only consider SRS and SBR as in Bugni, Canay, and Shaikh (2018). We conduct the simulations with sample sizes $n = 200$ and 400. The numbers of simulation replications and bootstrap samples are 1000. Under the null, $\mu = 0$ and we compute the true parameters of interest using simulations with 10^6 sample size and 10^4 replications. Under the alternative, we perturb the true values by $\mu = 1$ and $\mu = 0.75$ for $n = 200$ and 400, respectively.

We report the results for the median QTE. The second online supplement contains additional simulation results for ATE and QTEs with $\tau = 0.25$ and 0.75. All the observations made in this section still apply.

6.2 *QTE,* $\pi = 0.5$

We consider the Wald test with six t-statistics and 95% nominal rate. We construct the t-statistics using one of our two point estimates and some estimate of the standard error. We will reject the null hypothesis when the absolute value of the t-statistic is greater than 1.96. The details about the point estimates and standard errors are as follows:

(i) "s/naive": the point estimator is computed by the SQR and its standard error $\hat{\sigma}_{\text{naive}}(\tau)$ is computed as

$$\begin{aligned}\hat{\sigma}^2_{\text{naive}}(\tau) = {} & \frac{\tau(1-\tau) - \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \hat{m}_1^2(S_i,\tau)}{\pi \hat{f}_1^2\big(\hat{q}_1(\tau)\big)} + \frac{\tau(1-\tau) - \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \hat{m}_0^2(S_i,\tau)}{(1-\pi)\hat{f}_0^2\big(\hat{q}_0(\tau)\big)} \\ & + \frac{1}{n}\sum_{i=1}^{n} \pi(1-\pi)\bigg(\frac{\hat{m}_1(S_i,\tau)}{\pi \hat{f}_1\big(\hat{q}_1(\tau)\big)} + \frac{\hat{m}_0(S_i,\tau)}{(1-\pi)\hat{f}_0\big(\hat{q}_0(\tau)\big)}\bigg)^2 \\ & + \frac{1}{n}\sum_{i=1}^{n}\bigg(\frac{\hat{m}_1(S_i,\tau)}{\hat{f}_1\big(\hat{q}_1(\tau)\big)} - \frac{\hat{m}_0(S_i,\tau)}{\hat{f}_0\big(\hat{q}_0(\tau)\big)}\bigg)^2, \end{aligned} \tag{6.1}$$

where $\hat{q}_j(\tau)$ is the τth empirical quantile of $Y_i | A_i = j$,

$$\hat{m}_{i,1}(s,\tau) = \frac{\displaystyle\sum_{i=1}^{n} A_i 1\{S_i = s\}\big(\tau - 1\big\{Y_i \le \hat{q}_1(\tau)\big\}\big)}{n_1(s)},$$

$$\hat{m}_{i,0}(s,\tau) = \frac{\displaystyle\sum_{i=1}^{n} (1-A_i) 1\{S_i = s\}\big(\tau - 1\big\{Y_i \le \hat{q}_0(\tau)\big\}\big)}{n(s) - n_1(s)}.$$

For $j = 0, 1$, $\hat{f}_j(\cdot)$ is computed by the kernel density estimation using the observations Y_i provided that $A_i = j$, bandwidth $h_j = 1.06\hat{\sigma}_j n_j^{-1/5}$, Gaussian kernel function, standard deviation $\hat{\sigma}_j$ of the observations Y_i provided that $A_i = j$, and $n_j = \sum_{i=1}^{n} 1\{A_i = j\}$.

(ii) "s/adj": exactly the same as the "s/naive" method with one difference: replacing $\pi(1-\pi)$ in (6.1) by $\gamma(S_i)$.

(iii) "s/W": the point estimator is computed by the SQR and its standard error $\hat{\sigma}_W(\tau)$ is computed by the weighted bootstrap procedure. The bootstrap weights $\{\xi_i\}_{i=1}^n$ are generated from the standard exponential distribution. Denote $\{\hat{\beta}^w_{1,b}\}_{b=1}^B$ as the collection of

B weighted bootstrap SQR estimates. Then

$$\hat{\sigma}_W(\tau) = \frac{\hat{Q}(0.975) - \hat{Q}(0.025)}{z_{0.975} - z_{0.025}},$$

where $\hat{Q}(\alpha)$ is the αth empirical quantile of $\{\hat{\beta}^w_{1,b}(\tau)\}_{b=1}^B$.

(iv) "ipw/W": the same as above with one difference: the estimation method for both the original and bootstrap samples is the IPW QR.

(v) "s/CA": the point estimator is computed by the SQR and its standard error $\hat{\sigma}_{\mathrm{CA}}(\tau)$ is computed by the covariate-adaptive bootstrap procedure. Denote $\{\hat{\beta}^*_{1,b}\}_{b=1}^B$ as the collection of B estimates obtained by the SQR applied to the samples generated by the covariate-adaptive bootstrap procedure. Then

$$\hat{\sigma}_{\mathrm{CA}}(\tau) = \frac{\hat{Q}(0.975) - \hat{Q}(0.025)}{z_{0.975} - z_{0.025}},$$

where $\hat{Q}(\alpha)$ is the αth empirical quantile of $\{\hat{\beta}^*_{1,b}(\tau)\}_{b=1}^B$.

(vi) "ipw/CA": the same as above with one difference: the estimation method for both the original and bootstrap samples is the IPW QR.

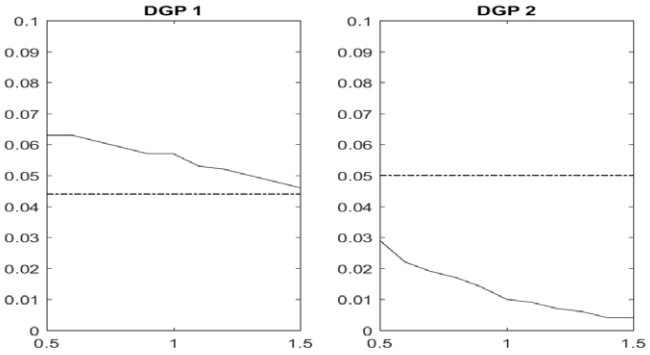
Tables 1 and 2 present the rejection probabilities (multiplied by 100) for the six t-tests under both the null hypothesis and the alternative hypothesis, with sample sizes $n = 200$ and 400, respectively. In these two tables, columns M and A represent DGPs and treatment assignment rules, respectively. From the rejection probabilities under the null, we can make five observations. First, the naive t-test ("s/naive") is conservative for WEI, BCD, and SBR, which is consistent with the findings for ATE estimators by Shao, Yu, and Zhong (2010) and Bugni, Canay, and Shaikh (2018). Second, although the asymptotic size of the adjusted t-test ("s/adj") is expected to equal the nominal level, it does not perform well for DGP2. The main reason is that, in order to analytically compute the standard error, we must compute nuisance parameters such as the unconditional densities of $Y(0)$ and $Y(1)$, which requires tuning parameters. We further compute the standard errors following (6.1) with $\pi(1-\pi)$ and the tuning parameter h_j replaced by $\gamma(S_i)$ and $1.06 C_f \hat{\sigma}_j n_j^{-1/5}$, respectively, for some constant $C_f \in [0.5, 1.5]$. Figure 1 plots the rejection probabilities of the "s/adj" t-tests against C_f for the BCD assignment rule with $n = 200$, $\tau = 0.5$, and $\pi = 0.5$. We see that (i) the rejection probability is sensitive to the choice of bandwidth, (ii) there is no universal optimal bandwidth across two DGPs, and (iii) the covariate-adaptive bootstrap t-tests ("s/CA") represented by the dotted dash lines are quite stable across different DGPs and their rejection probabilities are close to the nominal rate of rejection. Third, the weighted bootstrap t-test for the SQR estimator ("s/W") is conservative, especially for the BCD and SBR assignment rules which achieve strong balance. Fourth, the rejection probabilities of the weighted bootstrap t-test for the IPW estimator ("ipw/W") are close to the nominal rate even for sample size $n = 200$, which is consistent with Theorem 4.1. Last, the rejection rates for the two covariate-adaptive bootstrap t-tests ("s/CA" and "ipw/CA") are close to the nominal rate, which is consistent with Theorem 5.1.

TABLE 1. $n = 200$, $\tau = 0.5$, $\pi = 0.5$.

		H_0						H_1					
M	A	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA
1	SRS	4.5	4.5	4.7	4.4	4.4	3.9	18.3	18.3	19.3	44.1	20.0	42.9
	WEI	1.2	4.0	1.4	4.3	3.7	3.5	11.6	29.5	13.8	44.7	29.8	43.6
	BCD	0.2	5.7	0.3	4.1	4.4	3.9	7.2	47.2	9.5	45.3	43.4	44.8
	SBR	0.1	5.7	0.1	4.6	4.5	4.4	8.5	48.5	9.9	46.0	45.7	44.8
2	SRS	0.4	0.4	4.7	5.2	5.2	5.3	79.7	79.7	90.4	91.6	90.2	91.3
	WEI	0.6	0.6	4.5	5.8	5.2	5.7	80.2	80.7	90.7	90.9	91.3	90.6
	BCD	1.0	1.0	4.5	5.1	5.0	5.3	79.6	80.4	90.2	91.1	90.8	90.6
	SBR	0.8	1.1	4.8	5.3	4.6	4.7	77.1	77.4	89.7	90.1	89.9	89.9

TABLE 2. $n = 400$, $\tau = 0.5$, $\pi = 0.5$.

		H_0						H_1					
M	A	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA
1	SRS	4.2	4.2	5.4	4.0	4.6	4.1	21.8	21.8	23.2	50.2	23.5	50.2
	WEI	1.0	4.9	0.8	4.7	4.6	4.2	14.7	35.6	16.0	50.3	35.0	50.7
	BCD	0.3	4.5	0.2	4.3	3.5	4.0	8.9	52.6	11.7	50.2	49.3	49.6
	SBR	0.2	4.6	0.0	3.7	3.6	3.7	8.9	55.0	10.9	51.8	52.4	51.9
2	SRS	1.2	1.2	4.3	4.8	4.6	5.0	89.7	89.7	95.6	95.6	95.7	95.7
	WEI	1.4	1.6	5.7	6.0	5.5	5.7	89.2	89.2	95.4	94.8	95.1	94.8
	BCD	1.3	1.3	5.5	6.1	5.1	5.2	88.7	88.9	95.2	95.4	95.7	95.6
	SBR	0.6	0.6	4.0	3.9	3.8	3.8	90.0	90.2	95.4	95.4	95.8	95.7



Note: Rejection probabilities for BCD assignment rule with $n = 200$, $\pi = 0.5$, and $\tau = 0.5$. The X-axis is C_f. The solid lines are the rejection probabilities for "s/adj". The densities of Y_j is computed using the tuning parameters $h_j = 1.06 C_f \hat{\sigma}_j n_j^{-1/5}$, for $j = 0, 1$. The dotted dash lines are the rejection probability for "s/CA".

FIGURE 1. Rejection probabilities across different bandwidth values.

TABLE 3. $n = 200$, $\tau = 0.5$, $\pi = 0.7$.

		H_0						H_1					
M	A	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA
1	SRS	4.8	4.8	5.2	4.7	3.4	4.4	17.0	17.0	17.2	42.5	16.7	40.6
	SBR	0.1	0.7	0.2	4.0	4.4	3.7	4.3	21.2	6.0	45.5	45.7	43.4
2	SRS	1.6	1.6	5.2	5.4	5.1	5.3	77.1	77.1	89.1	90.3	89.5	89.4
	SBR	0.4	0.5	3.9	4.8	4.5	4.8	76.0	76.9	89.2	91.1	90.1	90.0

Turning to the rejection rates under the alternative in Tables 1 and 2, we can make two additional observations. First, for BCD and SBR, the rejection probabilities (power) for "ipw/W," "s/CA," and "ipw/CA" are close. This is because both BCD and SBR achieve strong balance. In this case, the two estimators we propose are asymptotically first-order equivalent. Second, for DGP1 with SRS and WEI assignment rules, "ipw/CA" is more powerful than "s/CA." This confirms our theoretical finding that the IPW estimator is *strictly* more efficient than the SQR estimator when the treatment assignment rule does *not* achieve strong balance. For DGP2 the three t-tests, that is, "ipw/W," "s/CA," and "ipw/CA," have similar power.

6.3 *QTE, $\pi = 0.7$*

Tables 3 and 4 show the similar results with $\pi = 0.7$. The same comments for Tables 1 and 2 still apply.

6.4 *Difference between two QTEs*

Last, we consider to infer $q(0.25) - q(0.75)$ when $\pi = 0.5$:

$$H_0 : q(0.25) - q(0.75) = \text{the true value} \quad \text{v.s.} \quad H_1 : q(0.25) - q(0.75) = \text{the true value} + \mu,$$

where $\mu = 1$ and 0.75 for sample sizes 200 and 400, respectively. The two estimators for QTEs at $\tau = 0.25$ and 0.75 are correlated. We can compute the naive and adjusted standard errors for the SQR estimator by taking this covariance structure into account.[6]

TABLE 4. $n = 400$, $\tau = 0.5$, $\pi = 0.7$.

		H_0						H_1					
M	A	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA
1	SRS	4.4	4.4	5.1	3.9	4.8	3.7	18.4	18.4	18.7	47.9	19.4	46.6
	SBR	0.1	0.2	0	3.9	3.5	4	4.2	22	5.9	49.8	50.5	48.2
2	SRS	0.7	0.7	3.9	4.2	4.2	4.7	86.7	86.7	93.9	93.3	94.1	93.6
	SBR	0.6	0.6	3.5	3.6	3.7	3.7	88.3	88.8	94.8	95.2	95.5	95.2

[6]The formulas for the covariances can be found in the proofs of Theorems 3.1 and 3.2.

TABLE 5. $n = 200$, $q(0.25) - q(0.75)$.

		H_0						H_1					
M	A	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA
1	SRS	4.0	4.0	3.6	3.8	3.5	3.5	15.6	15.6	14.9	19.4	16.0	19.4
	WEI	2.3	4.9	2.0	4.0	5.1	3.9	11.3	17.9	11.0	19.0	16.0	18.6
	BCD	1.0	4.1	1.1	4.4	3.7	4.2	9.9	20.7	10.1	22.0	20.6	21.4
	SBR	1.1	4.3	0.9	4.1	4.1	4.2	9.4	21.8	8.7	17.3	20.0	17.2
2	SRS	5.0	5.0	3.1	3.1	3.1	3.1	53.7	53.7	47.1	48.4	47.8	48.2
	WEI	3.6	3.6	2.1	2.8	2.9	2.9	57.0	57.7	47.6	49.8	50.3	50.0
	BCD	4.2	4.8	2.4	2.5	3.6	2.7	58.0	59.4	49.1	52.0	52.8	50.8
	SBR	5.1	5.3	2.4	3.4	4.1	3.4	55.5	57.0	46.5	46.5	50.5	45.6

On the other hand, in addition to avoiding the tuning parameters, another advantage of the bootstrap inference is that it does not require the knowledge of this complicated covariance structure. Researchers may construct the t-statistic using the difference of two QTE estimators with the corresponding weighted and covariate-adaptive bootstrap standard errors, which are calculated using the exact same procedure as in Sections 4 and 5. Taking the SQR estimator as an example, we estimate $q(0.25) - q(0.75)$ via $\hat{\beta}_1(0.25) - \hat{\beta}_1(0.75)$ and the corresponding covariate-adaptive bootstrap standard error is

$$\hat{\sigma}_{\mathrm{CA}} = \frac{\hat{Q}(0.975) - \hat{Q}(0.025)}{z_{0.975} - z_{0.025}},$$

where $\hat{Q}(\alpha)$ is the αth empirical quantile of $\{\hat{\beta}^*_{1,b}(0.25) - \hat{\beta}^*_{1,b}(0.75)\}_{b=1}^B$.

Based on the rejection rates reported in Tables 5 and 6, the general observations for the previous simulation results still apply. Although under the null, the rejection rates for "ipw/W," "S/CA," "ipw/CA" in DGP2 are below the nominal 5%, they gradually increase as the sample size increases from 200 to 400.

TABLE 6. $n = 400$, $q(0.25) - q(0.75)$.

		H_0						H_1					
M	A	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA	s/naive	s/adj	s/W	ipw/W	s/CA	ipw/CA
1	SRS	3.8	3.8	3.9	5.1	3.7	5.0	17.2	17.2	15.9	21.5	16.8	21.2
	WEI	2.0	4.2	2.4	3.3	4.4	3.5	11.8	20.2	11.5	21.4	20.2	20.7
	BCD	1.4	4.4	1.4	4.3	4.4	4.1	10.5	21.8	10.2	20.7	21.5	20.6
	SBR	0.8	3.8	0.8	3.9	3.7	3.8	12.1	25.0	12.6	21.8	23.7	22.3
2	SRS	5.3	5.3	3.9	4.7	4.3	4.8	63.2	63.2	55.7	57.7	56.8	57.6
	WEI	5.4	5.8	3.4	3.7	4.1	3.5	63.6	64.4	55.6	58.0	58.0	58.5
	BCD	4.0	4.3	2.6	2.8	3.1	3.1	62.1	63.3	54.7	55.7	57.4	56.0
	SBR	5.1	5.7	4.0	4.5	4.4	4.5	61.1	62.0	52.4	51.3	56.0	53.0

7. Guidance for practitioners

We recommend employing the t-statistic (or equivalently, the confidence interval) constructed using the IPW estimator and its weighted bootstrap standard error for inference in covariate-adaptive randomization, for the following four reasons. First, its asymptotic size equals the nominal level. Second, the IPW estimator has a smaller asymptotic variance than the SQR estimator when the treatment assignment rule does not achieve strong balance and the stratification is relevant.[7] Third, compared with the covariate-adaptive bootstrap, the validity of the weighted bootstrap requires a weaker condition that $\sup_{s\in\mathcal{S}} |D_n(s)/n(s)| = o_p(1)$. Fourth, this method does not require the knowledge of the exact treatment assignment rule, thus is suitable in settings where such information is lacking, for example, using someone else's RCT or subsample analysis. When the treatment assignment rule achieves strong balance, SQR estimator can also be used. But in this case, only the covariate-adaptive bootstrap standard error is valid. Last, the Wald test using SQR estimator and the weighted bootstrap standard error is not recommended, as it is conservative when the treatment assignment rule introduces negative dependence (i.e., $\gamma(s) < \pi(1-\pi)$) such as WEI, BCD, and SBR.

8. Empirical application

We illustrate our methods by estimating and inferring the average and quantile treatment effects of iron efficiency on educational attainment. The dataset we use is the same as the one analyzed by Chong, Cohen, Field, Nakasone, and Torero (2016) and Bugni, Canay, and Shaikh (2018).

8.1 *Data description*

The dataset consists of 215 students from one Peruvian secondary school during the 2009 school year. About two-thirds of students were assigned to the treatment group ($A = 1$ or $A = 2$). The other one-third of students were assigned to the control group ($A = 0$). One-half of the students in the treatment group were shown a video in which a physician encouraged iron supplements ($A = 1$) and the other half were shown the same encouragement from a popular soccer player ($A = 2$). Those assignments were stratified by the number of years of secondary school completed ($\mathcal{S} = \{1, \ldots, 5\}$). The field experiment used a stratified block randomization scheme with fractions $(1/3, 1/3, 1/3)$ for each group, which achieves strong balance ($\gamma(s) = 0$).

In the following, we focus on the observations with $A = 0$ and $A = 1$, and estimate the treatment effect of the exposure to a video of encouraging iron supplements by a physician only. This practice was also implemented in Bugni, Canay, and Shaikh (2018). In this case, the target proportions of treatment is $\pi = 1/2$. As in Chong et al. (2016), it is also possible to combine the two treatment groups, that is, $A = 1$ and $A = 2$ and compute the treatment effects of exposure to a video of encouraging iron supplements

[7] In this case, for ATE, the IPW estimator also has a strictly smaller asymptotic variance than the strata fixed effects estimator studied in Bugni, Canay, and Shaikh (2018).

by either a physician or a popular soccer player. Last, one can use the method developed in Bugni, Canay, and Shaikh (2019) to estimate the ATEs under multiple treatment status. However, in this setting, the estimation of QTE and the validity of bootstrap inference have not been investigated yet and are interesting topics for future research.

For each observation, we have three outcome variables: number of pills taken, grade point average, and cognitive ability measured by the average score across different Nintendo Wii games. For more details about the outcome variables, we refer interested readers to Chong et al. (2016). In the following, we focus on the grade point average only as the other two outcomes are discrete.

8.2 *Computation*

We consider three pairs of point estimates and their corresponding nonconservative standard errors: (i) the SQR estimator with the covariate-adaptive bootstrap standard error, (ii) the IPW estimator with the covariate-adaptive bootstrap standard error, and (iii) the IPW estimator with the weighted bootstrap standard error. We denote them as "s/CA," "ipw/CA," and "ipw/W," respectively. For comparison, we also compute the SQR estimator with its weighted bootstrap standard error, which is denoted as "s/W." The SQR estimator for the τth QTE refers to $\hat{\beta}_1(\tau)$ as the second element of $\hat{\beta}(\tau) = (\hat{\beta}_0(\tau), \hat{\beta}_1(\tau))$, where

$$\hat{\beta}(\tau) = \underset{b=(b_0,b_1)'\in\Re^2}{\arg\min} \sum_{i=1}^{n} \rho_\tau\big(Y_i - \dot{A}_i' b\big),$$

$\dot{A}_i = (1, A_i)'$, and $\rho_\tau(u) = u(\tau - 1\{u \le 0\})$ is the standard check function. It is also just the difference between the τth empirical quantiles of treatment and control groups. The IPW estimator refers to $\hat{q}(\tau) = \hat{q}_1(\tau) - \hat{q}_0(\tau)$, where

$$\hat{q}_1(\tau) = \arg\min_q \frac{1}{n}\sum_{i=1}^{n} \frac{A_i}{\hat{\pi}(S_i)}\rho_\tau(Y_i - q), \qquad \hat{q}_0(\tau) = \arg\min_q \frac{1}{n}\sum_{i=1}^{n} \frac{1 - A_i}{1 - \hat{\pi}(S_i)}\rho_\tau(Y_i - q),$$

$\hat{\pi}(\cdot)$ denotes the propensity score estimator, $\hat{\pi}(s) = n_1(s)/n(s)$, $n_1(s) = \sum_{i=1}^n A_i 1\{S_i = s\}$, and $n(s) = \sum_{i=1}^n 1\{S_i = s\}$. The covariate-adaptive bootstrap standard error ("CA") refers to the standard error computed in Section 5. In particular, we can draw the covariate-adaptive bootstrap sample $(Y_i^*, A_i^*, S_i^*)_{i=1}^n$ following the procedure in Section 5. We then recompute the SQR and IPW estimates using the bootstrap sample. We repeat the bootstrap estimation B times, and obtain $\{\hat{\beta}_{b,1}^*(\tau), \hat{q}_b^*(\tau)\}_{b=1}^B$. The standard errors for SQR and IPW estimates are computed as

$$\hat{\sigma}_{\text{sqr}}(\tau) = \frac{\hat{Q}_{\text{sqr}}(0.975) - \hat{Q}_{\text{sqr}}(0.025)}{z_{0.975} - z_{0.025}} \quad \text{and} \quad \hat{\sigma}_{\text{ipw}}(\tau) = \frac{\hat{Q}_{\text{ipw}}(0.975) - \hat{Q}_{\text{ipw}}(0.025)}{z_{0.975} - z_{0.025}},$$

respectively, where $\hat{Q}_{\text{sqr}}(\alpha)$ and $\hat{Q}_{\text{ipw}}(\alpha)$ are the α-th empirical quantiles of $\{\hat{\beta}_{b,1}^*(\tau)\}_{b=1}^B$ and $\{\hat{q}_b^*(\tau)\}_{b=1}^B$, respectively, and z_α is the αth percentile of the standard normal distribution, that is, $z_{0.975} \approx 1.96$ and $z_{0.025} \approx -1.96$. The weighted bootstrap standard error

TABLE 7. Grades points average.

	s/adj	s/W	s/CA	ipw/W	ipw/CA
ATE	0.35	0.35	0.35	0.37	0.37
	(0.16)	(0.16)	(0.17)	(0.16)	(0.17)
QTE, 25%		0.43	0.43	0.43	0.43
		(0.15)	(0.15)	(0.15)	(0.15)
QTE, 50%		0.29	0.29	0.29	0.29
		(0.22)	(0.23)	(0.22)	(0.24)
QTE, 75%		0.35	0.35	0.36	0.36
		(0.25)	(0.24)	(0.25)	(0.25)

for the IPW estimate can be computed in the same manner with only one difference, the covariate-adaptive bootstrap estimator $\{\hat{q}_b^*(\tau)\}_{b=1}^B$ is replaced by the weighted bootstrap estimator $\{\hat{q}_b^w(\tau)\}_{b=1}^B$, where for the bth replication, $\hat{q}_b^w(\tau) = \hat{q}_{b,1}^w(\tau) - \hat{q}_{b,0}^w(\tau)$,

$$\hat{q}_{b,1}^w(\tau) = \arg\min_q \frac{1}{n}\sum_{i=1}^n \frac{\xi_i^b A_i}{\hat{\pi}^w(S_i)}\rho_\tau(Y_i - q),$$

$$\hat{q}_{b,1}^w(\tau) = \arg\min_q \frac{1}{n}\sum_{i=1}^n \frac{\xi_i^b (1-A_i)}{1-\hat{\pi}^w(S_i)}\rho_\tau(Y_i - q),$$

$\{\xi_i^b\}_{i=1}^n$ is a sequence of i.i.d. standard exponentially distributed random variables, $\hat{\pi}^w(s) = n_1^w(s)/n^w(s)$, $n_1^w(s) = \sum_{i=1}^n \xi_i A_i 1\{S_i = s\}$, and $n^w(s) = \sum_{i=1}^n \xi_i 1\{S_i = s\}$. Similarly, we compute the weighted bootstrap SQR estimates $\{\beta_{b,1}^w(\tau)\}_{b=1}^B$ as the second element of $\{\beta_b^w(\tau)\}_{b=1}^B$, where

$$\beta_b^w(\tau) = \arg\min_{b=(b_0,b_1)'\in\Re^2} \frac{1}{n}\sum_{i=1}^n \xi_i^b \rho_\tau\big(Y_i - \dot{A}_i' b\big).$$

For the ATEs, we also compute the SQR estimator with the adjusted standard error based on the analytical formula derived by Bugni, Canay, and Shaikh (2018), that is, "s/adj." For QTE estimates, we consider quantile indexes $\{0.1, 0.15, \ldots, 0.90\}$. The number of replications for the two bootstrap methods is $B = 1000$.

8.3 *Main results*

Table 7 shows the estimates with the corresponding standard errors in parentheses. From the table, we can make several remarks. First, for both ATE and QTE, the SQR and IPW estimates are very close to each other and so do their standard errors computed via the analytical formula, weighted bootstrap, and covariate-adaptive bootstrap. This is consistent with our theory that, under strong balance, the two estimators are first-order equivalent. Second, although in theory, the weighted bootstrap standard errors for the SQR estimators should be larger than those computed via the covariate-adaptive bootstrap, in this application, they are very close. This is consistent with the finding in

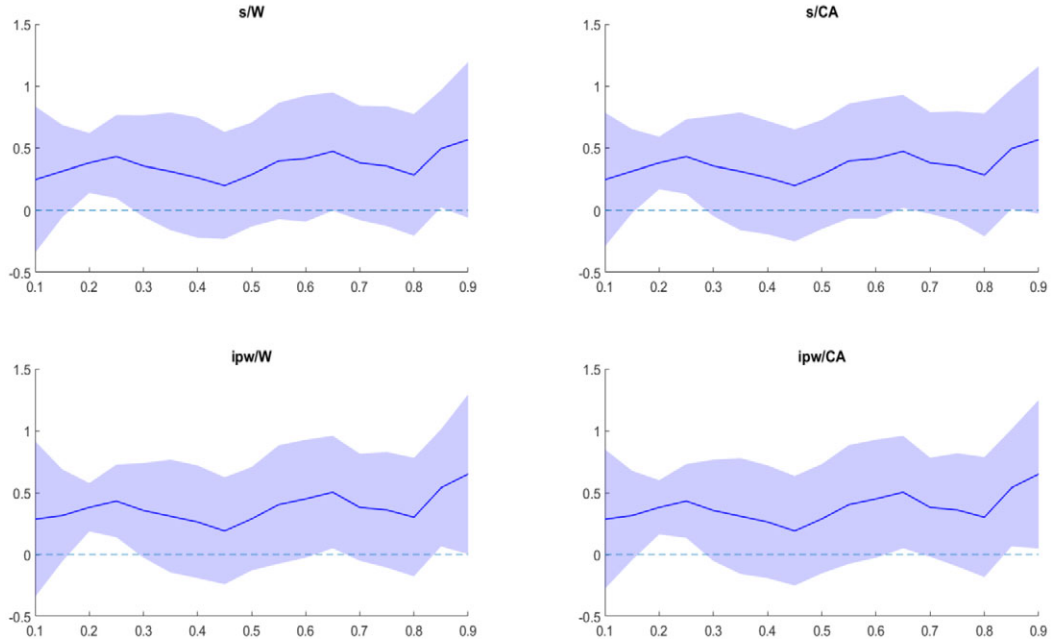


FIGURE 2. 95% Pointwise confidence interval for quantile treatment effects.

Bugni, Canay, and Shaikh (2018) that their adjusted p-value for the ATE estimate is close to the naive one. It implies the stratification may be irrelevant for the full-sample analysis. Third, we do not compute the adjusted standard error for the QTEs as it requires tuning parameters. Fourth, the QTEs provide us a new insight that the impact of supplementation on grade promotion is only significantly positive at 25% among the three quantiles. This may imply that the policy of reducing iron deficits is more effective for lower-ranked students.

In order to provide more details on the QTE estimates, we plot the 95% pointwise confidence band in Figure 2 with quantile index ranging from 0.1 to 0.9. The solid line and the shadow area represent the point estimate and its 95% pointwise confidence interval, respectively. The confidence interval is constructed by

$$\big[\hat{\beta} - 1.96\hat{\sigma}(\hat{\beta}), \hat{\beta} + 1.96\hat{\sigma}(\hat{\beta})\big],$$

where $\hat{\beta}$ and $\hat{\sigma}(\hat{\beta})$ are the point estimates and the corresponding standard errors described above. As we expected, all the four findings look the same and the estimates are only significantly positive at low quantiles (15%–30%).

8.4 *Subsample results*

Following Chong et al. (2016), we further split the sample into two based on whether the student is anemic, that is, $\text{Anem}_i = 0$ or 1. We anticipate that there is no treatment effect for the nonanemic students and positive effects for anemic ones. In this subsample

TABLE 8. Grades points average for subsamples.

	Anemic		Nonanemic	
	s/W	ipw/W	s/W	ipw/W
ATE	0.67	0.69	0.13	0.19
	(0.23)	(0.20)	(0.23)	(0.20)
QTE, 25%	0.74	0.76	0.14	0.22
	(0.24)	(0.22)	(0.28)	(0.26)
QTE, 50%	1.05	1.05	−0.14	−0.14
	(0.29)	(0.27)	(0.29)	(0.27)
QTE, 75%	0.71	0.76	0.14	0.14
	(0.36)	(0.32)	(0.39)	(0.37)

analysis, the covariate-adaptive bootstrap is infeasible, as in each subgroup, the strong-balance condition may be lost and the treatment assignment rule is not necessarily SBR and is generally unknown.[8] However, the weighted bootstrap is still feasible as it does not require the knowledge of the treatment assignment rule. According to Theorem 4.1, the IPW estimator paired with the weighted bootstrap standard error is valid if

$$\sup_{s\in\mathcal{S}}\left|\frac{D_n^{(1)}(s)}{n^{(1)}(s)}\right| \equiv \sup_{s\in\mathcal{S}}\left|\frac{\sum_{i=1}^{n}(A_i-\pi)1\{S_i=s\}1\{\text{Anem}_i=1\}}{\sum_{i=1}^{n}1\{S_i=s\}1\{\text{Anem}_i=1\}}\right| = o_p(1) \tag{8.1}$$

and

$$\sup_{s\in\mathcal{S}}\left|\frac{D_n^{(0)}(s)}{n^{(0)}(s)}\right| \equiv \sup_{s\in\mathcal{S}}\left|\frac{\sum_{i=1}^{n}(A_i-\pi)1\{S_i=s\}1\{\text{Anem}_i=0\}}{\sum_{i=1}^{n}1\{S_i=s\}1\{\text{Anem}_i=0\}}\right| = o_p(1). \tag{8.2}$$

We maintain this mild condition in this section. In our sample,

$$\sup_{s\in\mathcal{S}}\left|\frac{D_n^{(1)}(s)}{n^{(1)}(s)}\right| = 0 \quad \text{and} \quad \sup_{s\in\mathcal{S}}\left|\frac{D_n^{(0)}(s)}{n^{(0)}(s)}\right| = 0.071,$$

which indicate that (8.1) and (8.2) are plausible.

From Table 8 and Figure 3, we see that the QTE estimates are significantly positive for the anemic students when the quantile index is between around 20%–75%, but are insignificant for nonanemic students. The lack of significance at low and high quantiles

[8]As the anonymous referee pointed out, it is possible to implement the covariate-adaptive bootstrap on the full sample and pick out the observations in the subsample to construct a bootstrap subsample. The analysis can then be repeated on this covariate-adaptive bootstrap subsample. Establishing the validity of this procedure is left as a topic for future research.

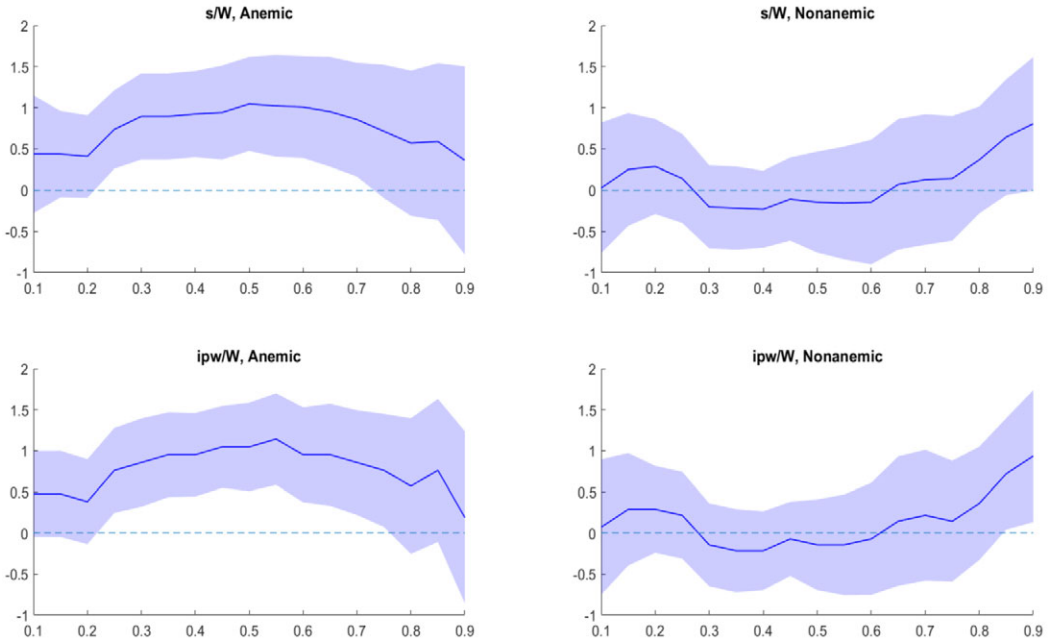


FIGURE 3. 95% Pointwise confidence interval for anemic and nonanemic students.

for the anemic subsample may be due to a poor asymptotic normal approximation at extreme quantiles. To extend the inference of extremal QTEs in Zhang (2018) to the context of covariate-adaptive randomization is an interesting topic for future research. We also note that for both subsamples, the weighted bootstrap standard errors for the SQR estimators are larger than those for the IPW estimators, which is consistent with Theorem 4.1. It implies, for both subgroups, the stratification is relevant.

9. Conclusion

This paper studies the estimation and bootstrap inference for QTEs under covariate-adaptive randomization. We show that the weighted bootstrap standard error is only valid for the IPW estimator while the covariate-adaptive bootstrap standard error is valid for both SQR and IPW estimators. In the empirical application, we find that the QTE of iron supplementation on grade promotion is trivial for nonanemic students, while the impact is significantly positive for middle-ranked anemic students.

References

Angrist, J. D. and J.-S. Pischke (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. [957]

Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan (2015), "The miracle of microfinance? Evidence from a randomized evaluation." *American Economic Journal: Applied Economics*, 7 (1), 22–53. [958]

Bruhn, M. and D. McKenzie (2009), "In pursuit of balance: Randomization in practice in development field experiments." *American Economic Journal: Applied Economics*, 1 (4), 200–232. [958]

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018), "Inference under covariate-adaptive randomization." *Journal of the American Statistical Association*, 113 (524), 1741–1768. [959, 960, 961, 962, 964, 967, 968, 969, 971, 975, 977, 978]

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2019), "Inference under covariate-adaptive randomization with multiple treatments." *Quantitative Economics*, 10 (4), 1747–1785. [959, 961, 964, 976]

Byrne, D. P., A. L. Nauze, and L. A. Martin (2018), "Tell me something I don't already know: Informedness and the impact of information programs." *Review of Economics and Statistics*, 100 (3), 510–527. [958]

Chernozhukov, V., D. Chetverikov, and K. Kato (2014), "Gaussian approximation of suprema of empirical processes." *The Annals of Statistics*, 42 (4), 1564–1597. [959]

Chong, A., I. Cohen, E. Field, E. Nakasone, and M. Torero (2016), "Iron deficiency and schooling attainment in Peru." *American Economic Journal: Applied Economics*, 8 (4), 222–255. [975, 976, 978]

Crépon, B., F. Devoto, E. Duflo, and W. Parienté (2015), "Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco." *American Economic Journal: Applied Economics*, 7 (1), 123–150. [958]

Doksum, K. (1974), "Empirical probability plots and statistical inference for nonlinear models in the two-sample case." *The Annals of Statistics*, 2 (2), 267–277. [958]

Duflo, E., R. Glennerster, and M. Kremer (2007), "Using randomization in development economics research: A toolkit." *Handbook of Development Economics*, 4, 3895–3962. [958]

Duflo, E., M. Greenstone, R. Pande, and N. Ryan (2013), "Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India." *The Quarterly Journal of Economics*, 128 (4), 1499–1545. [958]

Efron, B. (1971), "Forcing a sequential experiment to be balanced." *Biometrika*, 58 (3), 403–417. [961]

Firpo, S. (2007), "Efficient semiparametric estimation of quantile treatment effects." *Econometrica*, 75 (1), 259–276. [958, 963]

Hahn, J., K. Hirano, and D. Karlan (2011), "Adaptive experimental design using the propensity score." *Journal of Business & Economic Statistics*, 29 (1), 96–108. [959]

Hu, Y. (2016), "Generalized Efron's biased coin design and its theoretical properties." *Journal of Applied Probability*, 53 (2), 327–340. [969]

Hu, Y. and F. Hu (2012), "Asymptotic properties of covariate-adaptive randomization." *The Annals of Statistics*, 40 (3), 1794–1815. [961]

Imbens, G. W. and D. B. Rubin (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press. [958]

Ma, W., Y. Qin, Y. Li, and F. Hu (2018), "Statistical inference of covariate-adjusted randomized experiments." arXiv:1807.09678. [959]

Muralidharan, K. and V. Sundararaman (2011), "Teacher performance pay: Experimental evidence from India." *Journal of Political Economy*, 119 (1), 39–77. [958]

Shao, J., X. Yu, and B. Zhong (2010), "A theory for testing hypotheses under covariate-adaptive randomization." *Biometrika*, 97 (2), 347–360. [959, 971]

Shao, J. and X. Yu (2013), "Validity of tests under covariate-adaptive biased coin randomization and generalized linear models." *Biometrics*, 69 (4), 960–969. [959]

Tabord-Meehan, M. (2018), "Stratification trees for adaptive randomization in randomized controlled trials." arXiv:1806.05127. [959]

van der Vaart, A. and J. A. Wellner (1996), *Weak Convergence and Empirical Processes*. Springer, New York, NY. [959, 966]

Wei, L. (1978), "An application of an urn model to the design of sequential controlled clinical trials." *Journal of the American Statistical Association*, 73 (363), 559–563. [961, 969]

Zhang, Y. (2018), "Extremal quantile treatment effects." *The Annals of Statistics*, 46 (6B), 3707–3740. [980]

Zhang, Y., and X. Zheng (2020), "Supplement to 'Quantile treatment effects and bootstrap inference under covariate-adaptive randomization'." *Quantitative Economics Supplemental Material*, 11, https://doi.org/10.3982/QE1323. [960]

Co-editor Andres Santos handled this manuscript.

Manuscript received 6 April, 2019; final version accepted 10 February, 2020; available online 18 February, 2020.