Statistica Sinica Preprint No: SS-2025-0214	
Title	Statistical Inference on Quantile Treatment Effect Under Covariate-adaptive Randomization
Manuscript ID	SS-2025-0214
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0214
Complete List of Authors	Yuhang Tao and Li-Xin Zhang
Corresponding Authors	Li-Xin Zhang
E-mails	stazlx@mail.zjgsu.edu.cn
Notice: Accepted author version.	

Statistical Inference on Quantile Treatment Effect under Covariate-Adaptive Randomization

Yuhang Tao[1] and Li-Xin Zhang[2,1,*]

[1] *Zhejiang University and* [2] *Zhejiang Gongshang University*

Abstract:

Covariate-adaptive randomization (CAR) is commonly used to enhance covariate balance in randomized experiments, yet the development of statistical inference procedures on quantile treatment effects (QTEs) under CAR is still limited, with most existing work focusing on stratified CAR procedures. In this paper, we establish a unified theoretical framework for statistical inference on QTE under CAR procedures, encompassing a wide range of commonly used CAR procedures. First, we introduce a simple regression estimator for QTE and derive its asymptotic properties. To further enhance estimation efficiency, we propose an imbalance vector adjusted estimator. Additionally, we develop methods for estimating the asymptotic variance, including the moving block estimator and the covariate-adaptive randomized bootstrap. Simulation studies and a real data example demonstrate the effectiveness of the proposed methods, highlighting their practical utility in various settings.

Key words and phrases: Bootstrap, covariate-adaptive randomization, quantile

*Corresponding author. E-mail: stazlx@mail.zjgsu.edu.cn

treatment effect.

1. Introduction

Nowadays, a growing number of researchers tend to choose covariate-adaptive randomization (CAR) procedures to balance baseline covariates in clinical trials (Rosenberger and Lachin, 2015; Taves, 2010; Lin et al., 2015) and experiments in economics research (Chong et al., 2016; Burchardi et al., 2019; Anderson and McKenzie, 2022). For categorical covariates or discrete covariates, many CAR procedures including stratified permuted-block randomization (Zelen, 1974), Pocock and Simon's procedure (Taves, 1974; Pocock and Simon, 1975) and Hu and Hu's procedure (Hu and Hu, 2012) have been proposed to balance treatment assignments within strata and over covariate margins. These procedures can also be applied to continuous covariates through discretization, though this may result in a loss of information. To address this issue, numerous approaches have been proposed that directly utilize continuous covariates, such as Atkinson's D_A-biased coin design (Atkinson, 1982), the method based on the Kullback-Leibler divergence between empirical cumulative distribution functions (Lin and Su, 2012), the method based on kernel density (Ma and Hu, 2013) and adaptive randomization via Mahalanobis distance (Qin et al., 2022). Additionally,

Ma et al. (2024) recently proposed a new and unified family of CAR procedures, which are applicable to discrete or continuous covariates or their combinations.

In recent years, statistical inference under CAR procedures have attracted many researchers interests, especially parameter estimation and hypothesis testing problem of average treatment effects (ATEs). A large number of literature has explored the statistical inference issues concerning the ATEs under stratified randomization. Shao et al. (2010) proved that the two-sample t-test is conservative under a covariate-adaptive biased coin design. Ma et al. (2015) studied hypotheses testing under the linear model framework for stratified CAR, which assumes that the overall and marginal imbalances across covariates are bounded in probability. Bugni et al. (2018, 2019) studied the statistical inference for ATE under the framework of stratified CAR, which requires that the fraction of units within each stratum tends to be normally distributed. Moreover, in order to further improve the estimation efficiency, some researchers include additional auxiliary baseline covariates for regression adjustment (Ye et al., 2022, Ma et al., 2022). Lasso adjustment and machine learning methods have also been developed, which is applicable for high dimensional baseline covariates (Liu et al., 2023). For more general CAR procedure, Ma et al.

(2020) derived the theoretical properties of statistical methods based on general CAR under the linear model framework and extended the framework to logistic regression models.

Although ATE provides a valuable overview of the general impact of a treatment by quantifying the mean difference in outcomes between treated and control groups, it falls short in capturing the heterogeneity of treatment effects across different positions of the outcome distribution. For example, in clinical trials for rheumatoid arthritis treatment, as noted by (van der Heijde et al., 2006, Kremer et al., 2006), the efficacy in preventing structural damage is notably consistent across approximately 75% of the patient population. However, in the case of the most challenging 25% of patients, there is a notable divergence in treatment efficacy, with the less effective treatment showing a substantial decline in effectiveness, which is unable to be captured by ATE.

In order to solve the problem, some researchers arise their interests in the statistical properties of quantile treatment effects (QTEs) under CAR procudure. Under the framework of Bugni et al. (2018), Zhang and Zheng (2020) proposed the simple quantile regression and the inverse propensity score weighted quantile regression to estimate unconditional QTE and Jiang et al. (2023) proposed regression-adjusted estimation to further improve es-

timation efficiency. Besides, Liu et al. (2025) investigated inference for conditional quantile treatment effects (cQTEs) under stratified CAR procedure. However, to the best of our knowledge, the methods for estimating QTE under CAR procedure that directly incorporate continuous covariates remain largely unexplored. In the past, progress in this direction was limited by the lack of theoretical understanding of such designs. In recent years, however, the theoretical developments of CAR procedures that accommodate continuous covariates have made this line of research increasingly feasible (Qin et al., 2022, Ma et al., 2024, Zhang, 2023). In this article, we investigate the estimation of quantile treatment effects within a more general and flexible framework. This framework not only encompasses most stratified randomization procedures commonly used in practice but also extends to methods that directly utilize continuous covariates without requiring discretization. By addressing a broader class of covariate-adaptive randomization procedures, our approach enhances the applicability and relevance of QTE estimation in diverse experimental settings.

The main contributions of this article are as follows. First, we propose a unified framework that encompasses a wide range of CAR procedures and derive the asymptotic properties of the simple regression estimator within our general framework, providing theoretical support for quantile treatment

effect estimation across a broader family of covariate-adaptive randomization procedures. Besides, we propose the imbalance vector adjusted estimator which includes additional baseline covariates to improve estimation efficiency. Third, we propose two methods to estimate asymptotic variance of our estimator.

The rest of the paper is organized as follows. In Section 2, we describe the model setup and notation. In Section 3, we propose simple regression estimator and imbalance vector adjusted estimator. In Section 4, we introduce some common CAR procedures. In Section 5, we propose a method to estimate the asymptotic variance of simple regression estimator. In Section 6 and 7, simulations and a real data example are presented to illustrate the effectiveness of the proposed theory. The technical proofs are given in Supplementary Material.

2. Set up and Notation

Suppose that n units are to be assigned to two treatment groups using a CAR procedure. Let T_i be the assignment of the ith unit, where $T_i = 1$ indicates assignment to treatment group and $T_i = 0$ indicates assignment to control group. Define $n_1 = \sum_{i=1}^{n} T_i$ and $n_0 = \sum_{i=1}^{n}(1 - T_i)$ as the number of units in the treatment and control groups, respectively. Let $Y_i(1)$ and

$Y_i(0)$ represent the potential outcomes for unit i under treatment group and control group, respectively. The researcher observes $\{Y_i, X_i, T_i\}_{i=1}^n$, where $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$ is the observed outcome and $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\mathsf{T}$ denotes the covariate vector for unit i. We next introduce the assumptions that characterize the data-generating process (DGP) and the randomization schemes.

Assumption 1. (i) For $W_i = (Y_i(1), Y_i(0), X_i^\mathsf{T})^\mathsf{T}$, $\{W_i\}_{i=1}^n$ are independent and identically distributed as $W = (Y(1), Y(0), X^\mathsf{T})^\mathsf{T}$ with $\mathbb{E}\|W\|_2^2 < \infty$.
(ii) $\{Y_i(1), Y_i(0)\}_{i=1}^n \perp\!\!\!\perp \{T_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n$.

Assumption 2. If $m(X)$ is a function of X with $\mathbb{E}\{m^2(X)\} < \infty$, then there is a $\sigma_m \geq 0$, such that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (2T_i - 1)m(X_i) \xrightarrow{d} \mathcal{N}(0, \sigma_m^2). \tag{2.1}$$

Furthermore, if $f(X)$ is a function of X with $\mathbb{E}[f(X)] = 0$ and $\mathrm{Var}\{f(X)\} = \sigma_f^2 < \infty$; $\{Z_i(1), Z_i(0), X_i\}_{i=1}^n$ are independent copies of $\{Z(1), Z(0), X\}$ with $\mathbb{E}[Z(a)|X] = 0$ and $\mathrm{Var}\{Z(a)\} < \infty$, $a = 0, 1$; and that, conditional on $\{X_i\}_{i=1}^n$, $\{Z_i(1), Z_i(0)\}_{i=1}^n$ are independent of $\{T_i\}_{i=1}^n$. Let

$$\xi_i = (2T_i - 1)m(X_i) + f(X_i) + Z_i(1)T_i + Z_i(0)(1 - T_i), \tag{2.2}$$

then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i \xrightarrow{d} \mathcal{N}(0,\sigma_\xi^2),$$

where

$$\sigma_\xi^2 = \sigma_m^2 + \sigma_f^2 + \frac{1}{2}\operatorname{Var}\{Z(0)\} + \frac{1}{2}\operatorname{Var}\{Z(1)\}. \tag{2.3}$$

Assumption 1 is a regularity condition that is frequently adopted in the literature on CAR and holds for a broad class of CAR procedures. In contrast, Assumption 2 serves as a key condition in our framework.

Although Assumption 2 may appear complex, Ma et al. (2024) demonstrated that it holds for the unified family of CAR procedures they proposed, which encompasses many commonly used CAR methods such as Hu and Hu's procedure and COV procedure. Moreover, Assumption 2 also holds for adaptive randomization via Mahalanobis distance (Qin et al., 2022), with the proof provided in the Supplementary Material.

Equation (2.1) can be viewed as central limit theorem for CAR procedures. Nevertheless, for the unified family of CAR procedures, the closed-form expression for σ_m^2 remains an open problem (Ma et al., 2024, Zhang, 2023).

In addition, we introduce the following assumption that widely used in the literature of quantile estimations.

Assumption 3. For $a = 0, 1$, denote $f_a(\cdot)$ and $f_a(\cdot|\boldsymbol{x})$ as the probability density functions of $Y_i(a)$ and $Y_i(a)|X_i = \boldsymbol{x}$, respectively. $f_a(\cdot)$ and $f_a(\cdot|\boldsymbol{x})$ are continuous, bounded and bounded away from zero for each $\tau \in (0, 1)$.

3. Estimation of Quantile Treatment Effect

Our parameter of interest is the τth quantile treatment effect defined as $q(\tau) = q_1(\tau) - q_0(\tau)$, where $\tau \in (0, 1)$ is a quantile index and $q_a(\tau)$ is the τth quantile of random variable $Y(a)$ for $a = 0, 1$. In this section, we propose a simple regression estimator to estimate $q(\tau)$, and an imbalance vector adjusted estimator to further improve its efficiency.

3.1 Simple Regression Estimator

In this subsection, we propose the simple regression estimator to estimate $q(\tau)$, which is widely used in the estimation of quantile treatment effect (Zhang and Zheng, 2020). We obtain simple regression estimation by a quantile regression of Y_i on T_i, that is

$$(\widehat{q}_1(\tau), \widehat{q}_0(\tau))^{\mathsf{T}} = \arg\min_{\delta_1, \delta_0} \sum_{i=1}^{n} \rho_\tau(Y_i - T_i\delta_1 - (1 - T_i)\delta_0),$$

where $\rho_\tau(m) = m(\tau - I\{m \leq 0\})$ is the standard check function. The simple regression estimator is defined as $\widehat{q}(\tau) = \widehat{q}_1(\tau) - \widehat{q}_0(\tau)$ and the following theorem establishes its asymptotic property.

Theorem 1. *If Assumptions 1, 2 and 3 hold, then for fixed $\tau \in (0,1)$, we have*

$$\sqrt{n}(\widehat{q}(\tau) - q(\tau)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\tau)),$$

where

$$\sigma^2(\tau) = \sigma_1^2(\tau) + \sigma_2^2(\tau) + \sigma_3^2(\tau),$$

$$\sigma_2^2(\tau) = \mathrm{Var}\left\{ \frac{1}{f_1(q_1(\tau))} \left[\tau - \mathbb{P}(Y_i(1) \le q_1(\tau)|X_i)\right] - \frac{1}{f_0(q_0(\tau))} \left[\tau - \mathbb{P}(Y_i(0) \le q_0(\tau)|X_i)\right] \right\},$$

$$\begin{aligned} \sigma_3^2(\tau) = &\mathrm{Var}\left\{ \frac{1}{f_1(q_1(\tau))} \left[\mathbb{P}(Y_i(1) \le q_1(\tau)|X_i) - I\{Y_i(1) \le q_1(\tau)\}\right] \right\} \\ &+ \mathrm{Var}\left\{ \frac{1}{f_0(q_0(\tau))} \left[\mathbb{P}(Y_i(0) \le q_0(\tau)|X_i) - I\{Y_i(0) \le q_0(\tau)\}\right] \right\}, \end{aligned}$$

and $\sigma_1^2(\tau)$ satisfies that $n^{-1/2}\sum_{i=1}^n (2T_i - 1)g(X_i) \xrightarrow{d} \mathcal{N}(0, \sigma_1^2(\tau))$ with

$$g(X_i) = \frac{1}{f_1(q_1(\tau))} \left[\tau - \mathbb{P}(Y_i(1) \le q_1(\tau)|X_i)\right] + \frac{1}{f_0(q_0(\tau))} \left[\tau - \mathbb{P}(Y_i(0) \le q_0(\tau)|X_i)\right].$$

Note that $\sigma_2^2(\tau)$ and $\sigma_3^2(\tau)$ are invariant across different CAR procedures, we therefore refer to them as the design-invariant variance components. In contrast, the randomization-specific contribution to the asymptotic variance of $q(\tau)$ is captured by $\sigma_2^2(\tau)$, which we refer to as the randomization-specific variance component. If there exists a CAR procedure such that $\sigma_1^2(\tau) = 0$, then the corresponding $\widehat{q}(\tau)$ has a possibly minimum asymptotic variance.

3.2 Imbalance Vector Adjusted Estimator

The simple regression estimator is constructed based solely on the observed outcomes $\{Y_i\}_{i=1}^n$ and treatment assignments $\{T_i\}_{i=1}^n$, without incorporating any covariate information $\{X_i\}_{i=1}^n$. However, it is now widely recognized that incorporating the covariates used in CAR procedures into the inference stage can substantially improve estimation efficiency (Ma et al., 2022, Ye et al., 2022, Liu et al., 2023, Jiang et al., 2023). Motivated by this, we extend our analysis to include covariates in the estimation procedure. Under certain randomization procedures, such as complete randomization and Hu and Hu's procedure, the term $\sigma_1^2(\tau)$ admits a closed-form expression. In these cases, conventional covariate-adjusted estimators can be applied, as discussed in detail in Zhang and Zheng (2020) and Jiang et al. (2023). In contrast, when $\sigma_1^2(\tau)$ does not admit an explicit expression, the effect of conventional regression adjustment on this randomization-specific component becomes nontransparent. In particular, it is unclear whether the induced change increases or decreases $\sigma_1^2(\tau)$, and hence whether efficiency is improved. Consequently, traditional covariate-adjusted methods cannot, in general, guarantee efficiency improvement at the theoretical level. This observation highlights the importance of selecting covariates that enable a clean variance decomposition. The key requirement for covariate selection

is to leave the randomization-specific component $\sigma_1^2(\tau)$ unaffected, while reducing the design-invariant components $\sigma_2^2(\tau)$ and $\sigma_3^2(\tau)$, thereby improving estimation efficiency. To achieve this goal, we impose the following assumption on the choice of covariates.

Assumption 4 (Balance of imbalance vector)**.** Suppose that there exists a design-based r-dimensional vector $\boldsymbol{\psi}(X_i) = (\psi_1(X_i), \cdots, \psi_r(X_i))^\mathsf{T}$ such that $\mathbb{E}[\boldsymbol{\psi}(X_i)] = \mathbf{0}_r$ and $\sum_{i=1}^n (2T_i - 1)\boldsymbol{\psi}(X_i) = o_P(\sqrt{n})$.

Remark 1. Assumption 4 is satisfied by a broad class of CAR procedures. In particular, Ma et al. (2024) proposed a unified framework that encompasses many widely used CAR methods and establishes that Assumption 4(iii) holds under mild conditions. Consequently, Assumption 4 is fully satisfied provided that the overall treatment imbalance satisfies $\sum_{i=1}^n (2T_i - 1) = o_P(\sqrt{n})$.

Remark 2. This assumption indicates that only highly balanced covariates can be used for covariate adjustment. The results in Zhang (2023) show that only design-based covariates, which correspond to the covariates used to define the imbalance measure in CAR procedures and their linear combinations, satisfy this assumption. Consequently, additional covariates that are not design-based cannot be incorporated into the covariate adjustment

proposed in this paper. For this reason, we refer to the proposed adjustment method as the imbalance vector adjusted estimator.

First, we consider initial imbalance vector adjusted estimator of $q_1(\tau)$ and $q_0(\tau)$ with respect to $\boldsymbol{\psi}(X)$, which are defined as

$$\widehat{q}^{\star}_{a,adj}(\tau, \boldsymbol{\lambda}_a, \boldsymbol{\psi}) = \arg\min_{q} \sum_{i=1}^{n} I\{T_i = a\} \left[\rho_\tau(Y_i - q) + q\boldsymbol{\lambda}_a^{\mathsf{T}}\boldsymbol{\psi}(X_i)\right], a = 0, 1,$$

where $\boldsymbol{\lambda}_a = (\lambda_{1,a}, \ldots, \lambda_{r,a})^{\mathsf{T}}$ are constant r-dimensional vectors. Then, the initial imbalance vector adjusted QTE estimator is

$$\widehat{q}^{\star}_{adj}(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_0, \boldsymbol{\psi}) = \widehat{q}^{\star}_{1,adj}(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\psi}) - \widehat{q}^{\star}_{0,adj}(\tau, \boldsymbol{\lambda}_0, \boldsymbol{\psi}).$$

Specifically, when $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_0 = \mathbf{0}$, it reduces to the simple regression estimator. The following theorem establishes the asymptotic property of $\widehat{q}^{\star}_{adj}(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_0, \boldsymbol{\psi})$.

Theorem 2. *If Assumptions 1, 2, 3 and 4 hold. Then, for fixed $\tau \in (0, 1)$, we have*

$$\sqrt{n}(\widehat{q}^{\star}_{adj}(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_0, \psi) - q(\tau)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\tau) + V(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_0, \boldsymbol{\psi})),$$

where

$$\begin{aligned} V(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_0, \boldsymbol{\psi}) = &\left(\frac{\boldsymbol{\lambda}_1}{f_1(q_1(\tau))} - \frac{\boldsymbol{\lambda}_0}{f_0(q_0(\tau))}\right)^{\mathsf{T}} \boldsymbol{\Sigma}_{\boldsymbol{\psi}\boldsymbol{\psi}} \left(\frac{\boldsymbol{\lambda}_1}{f_1(q_1(\tau))} - \frac{\boldsymbol{\lambda}_0}{f_0(q_0(\tau))}\right) \\ &- 2\left(\frac{\boldsymbol{\lambda}_1}{f_1(q_1(\tau))} - \frac{\boldsymbol{\lambda}_0}{f_0(q_0(\tau))}\right)^{\mathsf{T}} \left(\frac{\boldsymbol{b}_1(\tau, \boldsymbol{\psi})}{f_1(q_1(\tau))} - \frac{\boldsymbol{b}_0(\tau, \boldsymbol{\psi})}{f_0(q_0(\tau))}\right), \end{aligned}$$

$\boldsymbol{\Sigma}_{\boldsymbol{\psi}\boldsymbol{\psi}} = \mathrm{Cov}(\boldsymbol{\psi}(X_i))$ *and* $\boldsymbol{b}_a(\tau, \boldsymbol{\psi}) = \mathbb{E}[(\tau - I\{Y_i(a) \leq q_a(\tau)\})\boldsymbol{\psi}(X_i)]$ *for* $a = 0, 1$. *In addition,* $V(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_0, \boldsymbol{\psi})$ *attains its minimal value when*

$$\frac{\boldsymbol{\lambda}_1}{f_1(q_1(\tau))} - \frac{\boldsymbol{\lambda}_0}{f_0(q_0(\tau))} = \boldsymbol{\Sigma}^{+}_{\boldsymbol{\psi}\boldsymbol{\psi}} \left(\frac{\boldsymbol{b}_1(\tau, \boldsymbol{\psi})}{f_1(q_1(\tau))} - \frac{\boldsymbol{b}_0(\tau, \boldsymbol{\psi})}{f_0(q_0(\tau))} \right),$$

where $\boldsymbol{\Sigma}^{+}_{\boldsymbol{\psi}\boldsymbol{\psi}}$ *denote the Moore-Penrose pseudoinverse of* $\boldsymbol{\Sigma}_{\boldsymbol{\psi}\boldsymbol{\psi}}$. *Furthermore, let* $V_{min}(\tau, \boldsymbol{\psi})$ *denote the minimal value of* $V(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_0, \boldsymbol{\psi})$, *then*

$$V_{min}(\tau, \boldsymbol{\psi}) = -\left(\frac{\boldsymbol{b}_1(\tau, \boldsymbol{\psi})}{f_1(q_1(\tau))} - \frac{\boldsymbol{b}_0(\tau, \boldsymbol{\psi})}{f_0(q_0(\tau))} \right)^{\mathsf{T}} \boldsymbol{\Sigma}^{+}_{\boldsymbol{\psi}\boldsymbol{\psi}} \left(\frac{\boldsymbol{b}_1(\tau, \boldsymbol{\psi})}{f_1(q_1(\tau))} - \frac{\boldsymbol{b}_0(\tau, \boldsymbol{\psi})}{f_0(q_0(\tau))} \right).$$

This theorem shows that Assumption 4 helps isolate the contribution of the design from that of the adjustment, thereby preventing potential cancellation between these two effects in the asymptotic expansion.

Note that the function $V(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_0, \boldsymbol{\psi})$ may be positive for some choices of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_0$, but $V_{min}(\tau, \boldsymbol{\psi}) \leq V(\tau, \mathbf{0}, \mathbf{0}, \boldsymbol{\psi}) = 0$. Therefore, choosing suitable values of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_0$ is crucial for enhancing the efficiency of the estimator. In the following, we determine the optimal values of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_0$ that minimize $V(\tau, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_0, \boldsymbol{\psi})$, and subsequently discuss their estimation. Due to the challenges associated with accurately estimating $f_a(q_a(\tau))$, it is preferable to avoid direct estimation of $f_a(q_a(\tau))$. A practical and effective choice is

$$\boldsymbol{\lambda}_{a,opt}(\tau, \boldsymbol{\psi}) = \boldsymbol{\Sigma}^{+}_{\boldsymbol{\psi}\boldsymbol{\psi}} \boldsymbol{b}_a(\tau, \boldsymbol{\psi}), a = 0, 1,$$

which can be estimated by

$$\widehat{\boldsymbol{\lambda}}_{a,opt}(\tau,\boldsymbol{\psi}) = \widehat{\boldsymbol{\Sigma}}^{+}_{\boldsymbol{\psi}\boldsymbol{\psi}}\widehat{\boldsymbol{b}}_a(\tau,\boldsymbol{\psi}), a = 0, 1,$$

where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\psi}\boldsymbol{\psi}}$ are the sample covariance matrix of $\boldsymbol{\psi}$ and

$$\widehat{\boldsymbol{b}}_a(\tau,\boldsymbol{\psi}) = \frac{1}{n_a}\sum_{i=1}^{n}\boldsymbol{\psi}(X_i)\left(\tau - I\{Y_i \le \widehat{q}_a(\tau)\}\right) I\{T_i = a\}, a = 0, 1.$$

Then, the imbalance vector adjusted estimator with respect to $\boldsymbol{\psi}(X)$ is defined as

$$\widehat{q}_{adj}(\tau,\boldsymbol{\psi}) = \widehat{q}_{1,adj}(\tau,\boldsymbol{\psi}) - \widehat{q}_{0,adj}(\tau,\boldsymbol{\psi}),$$

where

$$\widehat{q}_{a,adj}(\tau,\boldsymbol{\psi}) = \arg\min_{q}\sum_{i=1}^{n} I\{T_i = a\}\left[\rho_\tau(Y_i - q) + q\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}_{a,opt}(\tau,\boldsymbol{\psi})\boldsymbol{\psi}(X_i)\right], a = 0, 1.$$

The asymptotic property of $\widehat{q}_{adj}(\tau,\boldsymbol{\psi})$ is summarized in the following corollary, which follows immediately from Theorem 2.

Corollary 1. *If Assumptions 1, 2, 3 and 4 hold, then for fixed* $\tau \in (0,1)$, *we have* $\sqrt{n}(\widehat{q}_{adj}(\tau,\boldsymbol{\psi}) - q(\tau)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\tau) + V_{min}(\tau,\boldsymbol{\psi}))$.

Corollary 1 indicates that the efficiency of the imbalance vector adjusted estimator is at least as efficient as the simple regression estimator. Furthermore, if $\boldsymbol{\Sigma}_{\boldsymbol{\psi}\boldsymbol{\psi}}$ is non-singular, then as long as $\boldsymbol{b}_1(\tau,\boldsymbol{\psi})/f_1(q_1(\tau)) \neq \boldsymbol{b}_0(\tau,\boldsymbol{\psi})/f_0(q_0(\tau))$, we have $V_{\min}(\tau,\boldsymbol{\psi}) < 0$.

3.3 Selection of ψ for Imbalance Vector Adjusted Estimator

In the previous subsection, we proposed the imbalance vector adjusted estimator under the condition that $\boldsymbol{\psi}$ satisfies Assumption 4. In what follows, we explore the optimal choice of $\boldsymbol{\psi}$. Under a CAR procedure, there exists a design-based imbalance vector, which we denote by $\boldsymbol{\psi}_{DB}$. Let $w(X)$ be a function of X. In most cases, the existing theoretical results can only guarantee that $\sum_{i=1}^{n}(2T_i - 1)w(X_i) = o_P(\sqrt{n})$ if $w(X) \in \text{Span}\{\boldsymbol{\psi}_{DB}(X)\}$ (Ma et al., 2024, Zhang, 2023). Therefore, all admissible functions can be written as a linear combination $\boldsymbol{\psi}(X) = H\boldsymbol{\psi}_{DB}(X)$, for some matrix H. The following theorem provides the theoretical foundation for the optimal choice of $\boldsymbol{\psi}$.

Theorem 3. *If Assumptions 1, 2, 3 and 4 hold, then we have the following conclusions:*

(i) If H and $\text{Cov}(\boldsymbol{\psi}_{DB})$ are non-singular and $\boldsymbol{\psi}^* = H\boldsymbol{\psi}_{DB}$, then $V_{min}(\tau, \boldsymbol{\psi}_{DB}) = V_{min}(\tau, \boldsymbol{\psi}^*)$.

(ii) Let $\Psi = \{\boldsymbol{\psi}^* : \boldsymbol{\psi}^* = H\boldsymbol{\psi}_{DB}$ for some matrix $H\}$ be the set of all admissible imbalance vector functions. Then, $\min_{\boldsymbol{\psi}\in\Psi} V_{min}(\tau, \boldsymbol{\psi}) = V_{min}(\tau, \boldsymbol{\psi}_{DB})$.

This theorem shows that the design-based imbalance vector is optimal. For convenience, it can sometimes be multiplied by an non-singular matrix

without affecting the optimality. We will further illustrate this point with specific examples in the next section.

4. Properties of Several Randomization Procedures

In this section, we present several common randomization procedures that satisfy Assumptions 1 and 2 and provide the form of $\boldsymbol{\psi}(X_i)$ in imbalance measure adjusted estimator for each CAR procedures.

Example 1 (Complete Randomization)**.** Complete randomization (CR) assigns units to each treatment group with the equal probability 1/2. It can be seen as a special case of CAR procedure. Since the treatment assignment is independent and does not depend on covariates, it follows from central limit theorem that $n^{-1/2}\sum_{i=1}^{n}(2T_i-1)m(X_i) \xrightarrow{d} \mathcal{N}(0,\sigma^2_{CR,m})$, where $\sigma^2_{CR,m} = \mathrm{Var}(m(X))$ and $n^{-1/2}\sum_{i=1}^{n}\xi_i \xrightarrow{d} \mathcal{N}(0,\sigma^2_{CR,\xi})$, where $\sigma^2_{CR,\xi} = \mathrm{Var}\{(2T_i-1)m(X_i)+f(X_i)+Z_i(1)T_i+Z_i(0)(1-T_i)\} = \sigma^2_{CR,m}+\sigma^2_f+\frac{1}{2}\mathrm{Var}\{Z(0)\}+\frac{1}{2}\mathrm{Var}\{Z(1)\}$. Thus, Assumption 2 is satisfied under complete randomization and $\sigma^2_{1,CR} = \mathrm{Var}(g(X_i))$. However, Assumption 4 does not hold, thus the imbalance measure adjusted estimator is not applicable for complete randomization.

Example 2 (Hu and Hu's Procedure)**.** Hu and Hu's procedure (HH) is proposed by Hu and Hu (2012), which pursuits to simultaneously decrease

overall, marginal, and within-stratum imbalance, denoted by D_n, $D_n(l; s_l)$ and $D_n(\boldsymbol{s})$, respectively. The procedure is designed for discrete (categorical) covariates. In practice, continuous covariates must be discretized into categorical variables with a finite number of levels. Let $\widetilde{X}_i = (\widetilde{x}_{i1}, \cdots, \widetilde{x}_{ip})^{\mathsf{T}}$ denote the discrete version of X_i, where $\widetilde{x}_{ij} \in \{0, 1, \cdots, t_j\}$. Suppose the kth unit falls into stratum $\boldsymbol{s}^* = (s_1^*, \cdots, s_p^*)$. The imbalance measure is then defined as

$$\text{Imb}_{k,HH} = w_o D_n^2 + \sum_{l=1}^{p} w_{m,l} \{D_n(l; s_l^*)\}^2 + w_s \{D_n(\boldsymbol{s}^*)\}^2, \tag{4.4}$$

where w_o, $w_{m,l}$ and $w_s \geq 0$ denote the weights for D_n, $D_n(l; s_l^*)$, and $D_n(\boldsymbol{s}^*)$, respectively. The treatment status is determined sequentially for $1 \leq k \leq n$ as

$$\mathbb{P}(T_k = 1 | \{X_i\}_{i=1}^k, \{T_i\}_{i=1}^{k-1}) = \begin{cases} \rho & \text{if } \text{Imb}_{k,HH}^{(1)} < \text{Imb}_{k,HH}^{(0)}, \\ 1-\rho & \text{if } \text{Imb}_{k,HH}^{(1)} > \text{Imb}_{k,HH}^{(0)}, \\ 0.5 & \text{if } \text{Imb}_{k,HH}^{(1)} = \text{Imb}_{k,HH}^{(0)}, \end{cases} \tag{4.5}$$

where $0.5 < \rho \leq 1$, $\text{Imb}_k^{(1)}$ and $\text{Imb}_k^{(0)}$ denote the "potential" imbalance measures corresponding to $T_k = 1$ and $T_k = 0$, respectively. The procedure provides a general framework for CAR procedures, under which many existing designs can be viewed as specific instances. When $w_o = w_s = 0$, the procedure corresponds to Pocock and Simon's procedure (Pocock and

Simon, 1975). When $w_o = w_{m,l} = 0$, the procedure corresponds to stratified biased coin design (Shao et al., 2010). Hu and Hu's procedure can be regarded as a special case of the unified family of CAR procedures (Ma et al., 2024), which ensures that Assumption 2 holds. In particular, if $w_s > 0$, then $\sigma^2_{1,HH} = \mathbb{E}[\mathrm{Var}(g(X_i)|\widetilde{X}_i)] \leq \mathrm{Var}(g(X_i))$, which implies that simple regression estimator of quantile treatment effect under Hu and Hu's procedures is more efficient than that under CR. Based on the established theoretical properties in Hu and Zhang (2020), Assumption 4 holds for Pocock and Simon's procedure with

$$\boldsymbol{\psi}_{PS}(X_i) = (I\{\widetilde{x}_{ij} = s_j\} - \mathbb{P}(\widetilde{x}_{ij} = s_j) : j = 1, \cdots, p; s_j = 0, \cdots, t_j)^{\mathsf{T}}, \tag{4.6}$$

and Assumption 4 holds for stratified biased coin design with

$$\boldsymbol{\psi}_{SB}(X_i) = (I\{\widetilde{X}_i = \boldsymbol{s}\} - \mathbb{P}(\widetilde{X}_i = \boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S})^{\mathsf{T}}, \tag{4.7}$$

where $\mathcal{S}$ denotes the set of all strata. For Hu and Hu's procedure with $w_{m,l} > 0$ and $w_s > 0$, (4.7) can be used as the imbalance measure.

Remark 3. Under Hu and Hu's procedure, the asymptotic variance admits an explicit expression. This facilitates the incorporation of additional covariates to further improve estimation efficiency; see Zhang and Zheng (2020) and Jiang et al. (2023) for a detailed analysis.

Example 3 (COV Procedure)**.** COV procedure, originally proposed by Ma et al. (2024), incorporates the balancing of overall differences, covariate means, and covariance matrices. The imbalance measure is defined as

$$\mathrm{Imb}_{n,COV} = w_0 D_n^2 + w_1 \left\| \bar{X}_1 - \bar{X}_0 \right\|_2^2 + w_2 \left\| S_1 - S_0 \right\|_{\mathrm{F}}^2 ,$$

where $\bar{X}_a$ and S_a denote the covariate sample mean and covariance matrix under treatment ($a = 1$) and control ($a = 0$), respectively, $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm and w_0, w_1 and w_2 are the corresponding weights. The sequential treatment assignment rule for the kth unit follows the same form as Equation (4.4) with $\mathrm{Imb}_{k,HH}$ replaced by $\mathrm{Imb}_{k,COV}$. It has been established in Ma et al. (2024) that Assumption 2 is satisfied. Moreover, Assumption 4 holds with $\boldsymbol{\psi}_{COV}(X_i) = \left(X_i^\mathsf{T} - \mathbb{E}[X_i^\mathsf{T}], \mathrm{vec}(X_i X_i^\mathsf{T} - \mathbb{E}[X_i X_i^\mathsf{T}])_{unique}^\mathsf{T}\right)^\mathsf{T}$, where $\mathrm{vec}(X_i X_i^\mathsf{T} - \mathbb{E}[X_i X_i^\mathsf{T}])_{unique}$ denotes the vectorization of the unique (non-redundant) elements in the matrix $X_i X_i^\mathsf{T} - \mathbb{E}[X_i X_i^\mathsf{T}]$.

Example 4 (Adaptive randomization via Mahalanobis distance)**.** Adaptive randomization via Mahalanobis distance (ARM) is proposed by Qin et al. (2022), which allocates treatment sequentially by minimizing the Mahalanobis distance between covariate means of treatment group and control group, thereby achieving covariate balance without discretization. For convenience, we assume that n is even; otherwise, the last unit is assigned to

the treatment and control groups with equal probability. The imbalance measure for the first $2k$ units is defined as

$$\mathrm{Imb}_{2k,ARM} = \left(\bar{X}_{1,2k} - \bar{X}_{0,2k}\right)^{\mathsf{T}} \widehat{\mathbf{\Sigma}}^{-1} \left(\bar{X}_{1,2k} - \bar{X}_{0,2k}\right),$$

where $\widehat{\mathbf{\Sigma}}$ is the sample covariance matrix, $\bar{X}_{1,2k}$ and $\bar{X}_{0,2k}$ denote the covariate means of the first $2k$ units in the treatment and control groups, respectively. The sequential treatment assignment rule for the $(2k-1)$th unit is given by

$$\mathbb{P}(T_{2k-1} = 1 | \{X_i\}_{i=1}^{2k}, \{T_i\}_{i=1}^{2k-2}) = \begin{cases} \rho & \text{if } \mathrm{Imb}_{2k,ARM}^{(1,0)} < \mathrm{Imb}_{2k,ARM}^{(0,1)}, \\ 1-\rho & \text{if } \mathrm{Imb}_{2k,ARM}^{(1,0)} > \mathrm{Imb}_{2k,ARM}^{(0,1)}, \\ 0.5 & \text{if } \mathrm{Imb}_{2k,ARM}^{(1,0)} = \mathrm{Imb}_{2k,ARM}^{(0,1)}, \end{cases}$$

where $0.5 < \rho \leq 1$, $\mathrm{Imb}_{k,ARM}^{(1,0)}$ and $\mathrm{Imb}_{k,ARM}^{(0,1)}$ denote the "potential" imbalance measures corresponding to $(T_{2k-1}, T_{2k}) = (1,0)$ and $(T_{2k-1}, T_{2k}) = (0,1)$, respectively. The $(2k)$th unit is then assigned as $T_{2k} = 1 - T_{2k-1}$ to maintain the equal proportions. As shown in Qin et al. (2022), Assumption 4 is satisfied with $\boldsymbol{\psi}_{ARM}(X_i) = X_i - \mathbb{E}[X_i]$. In addition, Assumption 2 is satisfied under mild conditions, with the proof provided in the Supplementary Material.

Remark 4. For ARM, the design-based imbalance measure is $\boldsymbol{\psi}_{ARM,DB}(X_i) = \Sigma_{XX}^{-1/2} \boldsymbol{\psi}_{ARM}(X_i)$, where $\Sigma_{XX} = \mathrm{Cov}(X_i)$ and $\Sigma_{XX}^{-1/2}$ denotes its inverse

square root. For convenience, we tend to select $\boldsymbol{\psi}_{ARM}(X_i)$ as the imbalance measure in practice, which is also optimal from the perspective of efficiency.

5. Estimation of Asymptotic Variance

In this section, we consider the estimation of asymptotic variance of $\widehat{q}(\tau)$ and $\widehat{q}_{adj}(\tau, \boldsymbol{\psi})$. First, it is easy to note that the consistent estimator of $V_{min}(\tau, \boldsymbol{\psi})$ can be obtained by

$$\widehat{V}_{min}(\tau, \boldsymbol{\psi}) = -\left(\frac{\widehat{\boldsymbol{b}}_1(\tau, \boldsymbol{\psi})}{\widehat{f}_1(\widehat{q}_1(\tau))} - \frac{\widehat{\boldsymbol{b}}_0(\tau, \boldsymbol{\psi})}{\widehat{f}_0(\widehat{q}_0(\tau))}\right)^{\mathsf{T}} \widehat{\boldsymbol{\Sigma}}^{+}_{\boldsymbol{\psi}\boldsymbol{\psi}} \left(\frac{\widehat{\boldsymbol{b}}_1(\tau, \boldsymbol{\psi})}{\widehat{f}_1(\widehat{q}_1(\tau))} - \frac{\widehat{\boldsymbol{b}}_0(\tau, \boldsymbol{\psi})}{\widehat{f}_0(\widehat{q}_0(\tau))}\right),$$

where $\widehat{f}_a(\cdot)$ is estimated by the kernel density estimation using the observations Y_i provided that $T_i = a$. Therefore, we only need to consider the consistent estimation of $\sigma^2(\tau)$.

Multiplier bootstrap and covariate-adaptive randomized bootstrap are two widely used approach to estimate the asymptotic variance of quantile treatment effect estimators under covariate-adaptive randomization (Zhang and Zheng, 2020, Jiang et al., 2023, Liu et al., 2025). Unfortunately, under our framework, multiplier bootstrap estimator is not consistent, since the multiplier bootstrap sample does not preserve the negative cross-sectional dependence in the original sample. In this section, we propose moving block estimator and covariate-adaptive randomized bootstrap to estimate $\sigma^2(\tau)$.

5.1 Moving Block Estimator

In this subsection, we propose the moving block estimator (MB) to estimate the asymptotic variance. For each $i = 0, 1, \cdots, n-\ell$, we select a block of data $\{Y_j, X_j, T_j\}_{j=i+1}^{i+\ell}$ and obtain simple regression estimator of $q(\tau)$ based on the block of data, which is denoted as $\widehat{q}_{i,\ell,mb}(\tau)$. Then, the moving block estimator of the asymptotic variance of $\sqrt{n}(\widehat{q}(\tau) - q(\tau))$ is given by the sample variance of $\{\widehat{q}_{i,\ell,mb}(\tau)\}_{i=0}^{n-\ell}$ multiplied by ℓ. That is,

$$\widehat{\sigma}^2_{mb}(\tau) = \frac{\ell}{n-\ell+1} \sum_{i=0}^{n-\ell} \left(\widehat{q}_{i,\ell,mb}(\tau) - \bar{q}_{mb}(\tau)\right)^2,$$

where $\bar{q}_{mb}(\tau)$ denotes sample mean of $\{\widehat{q}_{i,\ell,mb}(\tau)\}_{i=0}^{n-\ell}$.

To prove $\widehat{\sigma}^2_{mb}(\tau)$ is consistent, we introduce the following assumption.

Assumption 5. Suppose that $\ell \to \infty$ and $\ell/n \to 0$, as $n \to \infty$, we have

$$\frac{1}{n\ell} \sum_{i=0}^{n-\ell} \left(\sum_{j=i+1}^{i+\ell} \xi_i \right)^2 \xrightarrow{P} \sigma^2_\xi,$$

where the definition of ξ_i and σ^2_ξ are provided in equations (2.2) and (2.3), respectively.

Remark 5. It can be proved that Assumption 5 is satisfied for the new and unified family of CAR procedures proposed by Ma et al. (2024). The proof follows a similar approach to that of Theorem 3.6 in Zhang (2023) and we will not go into further details.

Theorem 4. *Suppose that Assumptions 1, 2, 3 and 5 hold. If* $\ell \to \infty$ *and* $\ell/n \to 0$ *as* $n \to \infty$*, then for fixed* $\tau \in (0,1)$, $\widehat{\sigma}^2_{mb}(\tau) \xrightarrow{P} \sigma^2(\tau)$.

Remark 6. Although the moving block estimator is excellent in theoretical properties, the converge rate is very slow, which implies that it is not useful when the sample size is not very large. For moderate sample sizes, the estimator's performance is highly sensitive to the choice of ℓ, yet its optimal selection remains an open question. A similar dilemma is discussed in Zhang (2023) when estimating the asymptotic variance of the estimator for average treatment effect. Further investigation into this issue constitutes an important direction for future research. In practical applications, Zhang (2023) recommends setting ℓ to integer$(n^{1/3})$ or integer$(n^{1/2})$.

5.2 Covariate-adaptive Randomized Bootstrap

Inspired by the bootstrap methods proposed by Shao et al. (2010), Zhang and Zheng (2020) and Liu et al. (2025), we propose the covariate-adaptive randomized bootstrap (CAB) for our framework. The procedure is as follow.

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

(ii) Generating treatment assignments: Generate $\{T_i^*\}_{i=1}^n$ based on $\{X_i^*\}_{i=1}^n$

and the predefined treatment assignment rule.

(iii) Resampling outcomes: For $T_i^* = a$ and $X_i^* = \boldsymbol{x}$, we randomly select k^* from the set $\arg\min_{k\in\{j:T_j=a\}} ||X_k - X_i^*||_2$, which consists of indices k corresponding to units assigned to treatment a and having covariates X_k closest to X_i^* in terms of the Euclidean distance. Since this set is guaranteed to be non-empty, we can always draw k^*. The outcome Y_i^* is assigned as Y_{k^*}.

(iv) Computing bootstrap estimates: Compute $\widehat{q}^*(\tau)$ based on the bootstrap sample $\{Y_i^*, T_i^*, X_i^*\}_{i=1}^n$ with the same working model as that for $\widehat{q}(\tau)$.

(v) Repeating the process: Repeat (i) - (iv) for B times and denote the B bootstrap estimates as $\{\widehat{q}^{*b}(\tau)\}_{b=1}^B$.

Then the covariate-adaptive randomized bootstrap estimator is given by

$$\widehat{\sigma}^2_{cab} = \frac{n}{B-1}\sum_{b=1}^{B}\left(\widehat{q}^{*b}(\tau) - \bar{q}^B(\tau)\right)^2,$$

where $\bar{q}^B(\tau)$ denote the sample mean of $\{\widehat{q}^{*b}(\tau)\}_{b=1}^B$.

Remark 7. The theoretical properties of the covariate-adaptive randomized bootstrap have not yet been fully established. However, extensive simulation studies demonstrate that the method performs effectively in practice, providing reliable variance estimates and supporting accurate inference in various settings.

6. Simulation

To assess and compare the performance of our proposed estimators, $\widehat{q}(\tau)$ and $\widehat{q}_{adj}(\tau, \boldsymbol{\psi})$, we conduct a series of simulation studies. Specifically, we consider the following data-generating process (DGP). For each treatment level $a = 0, 1$, the potential outcome $Y(a)$ is generated according to

$$Y_i(a) = \mu_a + g_a(X_i) + \sigma_a(X_i)\varepsilon_i(a),$$

where X_i, $g_a(X_i)$, $\sigma_a(X_i)$ and $\varepsilon_i(a)$ are specified below for two different models. In each of the models, $\{X_i, \varepsilon_i(0), \varepsilon_i(1)\}_{i=1}^n$ are i.i.d. Besides, $\varepsilon_i(0)$ and $\varepsilon_i(1)$ are independent of X_i.

Model 1: X_i is a two-dimensional vector satisfying $X_i \sim \mathcal{N}_2(\mathbf{0}, 0.5 \times \boldsymbol{I}_2 + 0.5 \times \mathbf{1}_{2\times 2})$, where $\mathbf{1}_{2\times 2}$ denotes a 2×2 all-ones matrix. The model components are specified as follows: $\mu_1 = 2.5$, $\mu_0 = 0$, $g_1(X_i) = 5(x_{i1} + x_{i2})$, $g_0(X_i) = x_{i1}^2 + x_{i2}^2 + x_{i1}x_{i2} + x_{i1} + x_{i2}$, $\sigma_1(X_i) = 5(1 + x_{i1}^2)$, $\sigma_0(X_i) = 1 + x_{i1}^2$ and $(\varepsilon_i(1), \varepsilon_i(0))$ are jointly standard normal.

Model 2: X_i is a four-dimensional vector, where $x_{i1} \sim \mathcal{N}(0, 1)$, $x_{i2} \sim$ Beta$(2, 2)$, $x_{i3} \sim$ Unif$[0, 4]$ and $x_{i4} \sim$ Exp(1), and they are independent of each other. The model components are specified as follows: $\mu_1 = 17.8$, $\mu_0 = 0$, $\sigma_1(X_i) = 1$, $\sigma_0(X_i) = x_{i1}^2 + x_{i2}^2 + x_{i3}^2 + x_{i4}^2$, $g_1(X_i) = e^{x_{i,1}+x_{i,2}} + x_{i,3} + x_{i,4}$, $g_0(X_i) = 2e^{x_{i,1}+x_{i,2}} + 6(x_{i,3} + x_{i,4})$ and $(\varepsilon_i(1), \varepsilon_i(0))$ are mutually indepen-

dently $T(4)$ distributed.

For each DGP, we employ the following randomization procedures:

(i) Complete randomization (CR);

(ii) Hu and Hu's procedure (HH) with weights $w_o = 2/5$, $w_s = 2/5$ and $w_{m,j} = 1/(5p)$, $j = 1, \cdots, p$;

(iii) COV procedure (COV) with weights $w_0 = 1$, $w_1 = p$ and $w_2 = 1$;

(iv) Adaptive Randomization via Mahalanobis distance (ARM).

When applying HH, discretization of continuous variables is required. For model 1, we discretize x_{i1} and x_{i2} into four levels and for $k = 1, 2$, $\widetilde{x}_{ik} = \sum_{j=1}^{3} I\{x_{ik} \leq s_j\}$, where $(s_1, s_2, s_3) = (-0.6, 0, 0.6)$. For model 2, we discretize x_{i1}, x_{i2}, x_{i3} and x_{i4} into two levels and the discretizing rule is given by $\widetilde{x}_{i1} = I\{x_{i1} \leq 0\}$, $\widetilde{x}_{i2} = I\{x_{i2} \leq 0.5\}$, $\widetilde{x}_{i3} = I\{x_{i3} \leq 2\}$ and $\widetilde{x}_{i4} = I\{x_{i4} \leq 1\}$.

To approximate the true value of $q(\tau)$, which is difficult to compute exactly, we compute an empirical estimate based on simulations with a sample size of 50,000 and 10,000 replications. The simulation results are presented in Table 1. In addition, we also report the ATE as a benchmark for comparison, which helps to illustrate how QTEs provide a more detailed characterization of treatment effects across the outcome distribution. Both Model 1 and Model 2 demonstrate the advantage of using QTEs

over the ATE in characterizing treatment effects. For instance, in Model 1, the ATE is approximately zero, suggesting no overall difference between the two treatment groups. However, the estimated QTEs reveal substantial heterogeneity across the outcome distribution. This contrast clearly illustrates that while the ATE masks distributional differences, QTEs capture how treatment effects vary across individuals with different outcome levels.

Table 1: Empirical true values of ATE and QTEs obtained from simulations

	ATE	QTE($\tau = 0.25$)	QTE($\tau = 0.5$)	QTE($\tau = 0.75$)
Model 1	0.000	-1.758	1.140	3.094
Model 2	0.010	9.039	2.872	-6.284

We report the results for $\tau = 0.5$ under sample sizes $n = 1,000$ in Table 2 of the main text. Additional results for $\tau = 0.25$ and $\tau = 0.75$ are presented in the Supplementary Material.

Table 2 presents the bias, standard deviations (SD), root mean square error (RMSE), average estimated standard deviation (SE) and coverage probability (CP) of asymptotic 95% confidence intervals of simple regression estimator and imbalance vector adjusted estimator under different data

generating models and randomization procedures. The SEs and CPs are based on variance estimators from moving block estimator and covariate-adaptive randomized bootstrap estimators, respectively. Specifically, we set $\ell = \lceil \sqrt{n} \rceil = 32$ for moving block estimator and use 2,000 bootstrap replications for covariate-adaptive randomized bootstrap estimators. The confidence interval based on the simple regression estimator and imbalance vector adjusted estimator are given by

$$\left[\widehat{q}(\tau) \pm z_{0.975}\sqrt{\widehat{\sigma}^2_{\dagger}(\tau)/n}\right] \text{ and } \left[\widehat{q}_{adj}(\tau, \boldsymbol{\psi}) \pm z_{0.975}\sqrt{(\widehat{\sigma}^2_{\dagger}(\tau) - \widehat{V}_{min}(\tau, \boldsymbol{\psi}))/n}\right],$$

where $\widehat{\sigma}^2_{\dagger}(\tau)$ represents the estimated asymptotic variance obtained from the moving block estimator or the covariate-adaptive randomized bootstrap estimator ($\dagger \in \{mb, cab\}$). In addition, every scenario is evaluated with 2,000 simulation runs.

First, from Table 2, it is evident that all the estimators exhibit sufficiently small biases and their RMSEs are almost identical to their SDs, indicating their overall reliability in parameter estimation. Second, the SDs and RMSEs of the simple regression estimator under various CAR procedures are consistently smaller than those under CR. This suggests that CAR procedures enhance the estimation efficiency of the quantile treatment effect. Third, imbalance vector adjusted estimator outperforms the simple regression estimator in terms of SDs and RMSEs under the same model

and CAR procedure. This improvement is particularly illustrated in Model 2 under COV procedure. Specifically, the SD of $\widehat{q}(\tau)$ is reduced by 11.7% compared to $\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{COV})$. Fourth, the values of SEs are generally close to those of SDs and RMSEs, indicating that the estimated variances are reasonably accurate. Moreover, the SEs obtained from the covariate-adaptive randomized bootstrap estimators are more consistent with the corresponding SD and RMSE values than those from the moving block estimators, suggesting that the covariate-adaptive randomized bootstrap estimator is more accurate. Finally, the CPs are close to the nominal level of 95%, demonstrating that the constructed confidence intervals achieve the desired coverage and the inference is reliable.

7. Real Data Example

In this section, we analyze the well-known Lalonde experimental dataset (lalonde.exp), which originates from the National Supported Work (NSW) Demonstration study conducted by LaLonde (1986). The dataset includes information on individuals who participated in a job training program and a control group of nonparticipants. The total number of units was 445 and the outcome was post-treatment earnings in 1978. We use the dataset solely for the purpose of generating synthetic data to illustrate the efficiency of simple

Table 2: Simulated bias, SD, RMSE, SE and CP of the QTE estimates under different models and randomization procedures ($n = 1000$, $\tau = 0.5$)

						SE		CP	
M.	Ran.	Estimator	Bias	SD	RMSE	CAB	MB	CAB	MB
1	CR	$\widehat{q}(\tau)$	0.006	0.383	0.383	0.382	0.371	0.945	0.935
	HH	$\widehat{q}(\tau)$	-0.016	0.338	0.339	0.339	0.354	0.942	0.949
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{SB})$	-0.015	0.322	0.323	0.320	0.336	0.941	0.953
	COV	$\widehat{q}(\tau)$	-0.004	0.359	0.359	0.355	0.343	0.938	0.937
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{COV})$	-0.003	0.344	0.344	0.337	0.325	0.935	0.928
	ARM	$\widehat{q}(\tau)$	0.005	0.344	0.345	0.342	0.344	0.942	0.941
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{ARM})$	0.010	0.328	0.329	0.335	0.328	0.945	0.941
2	CR	$\widehat{q}(\tau)$	-0.034	0.798	0.799	0.791	0.773	0.942	0.938
	HH	$\widehat{q}(\tau)$	-0.005	0.716	0.716	0.719	0.758	0.945	0.960
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{SB})$	0.010	0.680	0.680	0.673	0.715	0.938	0.958
	COV	$\widehat{q}(\tau)$	-0.007	0.725	0.724	0.724	0.757	0.944	0.954
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{COV})$	-0.005	0.640	0.639	0.663	0.701	0.940	0.956
	ARM	$\widehat{q}(\tau)$	-0.005	0.695	0.695	0.706	0.694	0.940	0.945
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{ARM})$	-0.008	0.658	0.658	0.658	0.663	0.939	0.942

Abbreviations: M, Model; Ran, Randomization.

regression estimator under various CAR procedures and the capability of imbalance vector adjusted estimators to further improve efficiency.

To generate the synthetic data, we employed the random forest approach to complete the lalonde.exp dataset. To further improve the accuracy of the imputation, we incorporated auxiliary information from the lalonde.psid dataset, which contains similar variables collected from a different sample.

Then, we reassign units into treatment and control groups using various randomization procedures, including CR, HH, COV, and ARM. The HH procedure is implemented by stratifying subjects according to three binary covariates (black, hispan, and married), resulting in eight strata. The COV and ARM procedures are implemented using six covariates: age, education, black, hispan, married, and nodegree.

Table 3 reports QTEs evaluated at $\tau \in \{0.25, 0.5, 0.75\}$, their associated 95% confidence intervals and the variance reduction. The variance reduction for the simple regression estimator is measured against complete randomization, while that for the imbalance vector adjusted estimator is measured against the simple regression estimator at the same quantile level under the same CAR procedure. The confidence intervals are constructed based on the variance estimators from covariate-adaptive randomized boot-

strap estimator with 2,000 bootstrap replications.

In most cases, the confidence intervals constructed from the simple regression estimator under CAR procedures are shorter than those under CR. The efficiency of the simple regression estimator varies across different CAR procedures and quantile levels. Specifically, the HH procedure performs best at $\tau = 0.25$ and $\tau = 0.75$, whereas the ARM procedure achieves the highest efficiency at $\tau = 0.5$.

Across different CAR procedures, the estimated quantile treatment effects exhibit similar patterns. There is substantial heterogeneity in the job training program on the earnings in 1978. At the upper quantile ($\tau = 0.75$), the confidence interval includes zero, suggesting that the job training program does not have a statistically significant impact on income among higher-income individuals. In contrast, at the lower and median quantiles ($\tau = 0.25$ and $\tau = 0.5$), the confidence intervals exclude zero, indicating significant income gains for low-income and middle-income groups from the job training program. Notably, the estimated QTE is largest at $\tau = 0.25$, implying that the job training program is particularly effective in improving the earnings of low-income individuals.

Table 3: QTEs on the earnings in 1978

Ran	τ	Estimator	Estimate	95% CI	VR
CR	0.25	$\widehat{q}(\tau)$	2.278	$(0.321, 4.235)$	-
	0.5	$\widehat{q}(\tau)$	0.837	$(0.286, 1.388)$	-
	0.75	$\widehat{q}(\tau)$	1.221	$(-0.343, 2.784)$	-
HH	0.25	$\widehat{q}(\tau)$	1.958	$(0.549, 3.366)$	0.482
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{HH})$	1.958	$(0.552, 3.364)$	0.004
	0.5	$\widehat{q}(\tau)$	0.666	$(0.177, 1.155)$	0.212
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{HH})$	0.645	$(0.165, 1.125)$	0.037
	0.75	$\widehat{q}(\tau)$	0.734	$(-0.586, 2.154)$	0.287
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{HH})$	0.729	$(-0.569, 2.027)$	0.033
COV	0.25	$\widehat{q}(\tau)$	3.168	$(1.534, 4.803)$	0.214
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{COV})$	3.062	$(1.448, 4.676)$	0.026
	0.5	$\widehat{q}(\tau)$	0.995	$(0.435, 1.475)$	-0.098
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{COV})$	0.809	$(0.347, 1.271)$	0.211
	0.75	$\widehat{q}(\tau)$	1.155	$(-0.035, 2.345)$	0.135
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{COV})$	0.960	$(-0.170, 2.090)$	0.098
ARM	0.25	$\widehat{q}(\tau)$	2.205	$(0.496, 3.915)$	0.237
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{ARM})$	1.893	$(0.185, 3.600)$	0.018
	0.5	$\widehat{q}(\tau)$	0.642	$(0.265, 1.019)$	0.532
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{ARM})$	0.629	$(0.263, 0.996)$	0.161
	0.75	$\widehat{q}(\tau)$	1.254	$(-0.288, 2.796)$	0.027
		$\widehat{q}_{adj}(\tau, \boldsymbol{\psi}_{ARM})$	0.761	$(-0.756, 2, 278)$	0.076

Abbreviations: Ran, Randomization; CI, confidence interval; VR, variance reduction.

8. Conclusion

This paper develops statistical inference methods for QTE within a broad class of CAR procedures, significantly generalizing existing approaches that are largely limited to stratified CAR procedures. We show that simple regression estimator is asymptotic normal under our framework and propose imbalance vector adjusted estimator to further improve the efficiency. Besides, we establish two methods to estimate the asymptotic variance of the simple regression estimator. In simulations, both simple regression estimator and imbalance vector adjusted estimator performs well. This work lays a theoretical foundation for the statistical inference on quantile treatment effect under CAR experiments.

The findings of this paper can be extended in several directions. First, this paper focuses on QTE. In contrast, cQTE provide more precise policy guidance by capturing the heterogeneity of treatment effects across subpopulations with different covariate profiles. Statistical inference on cQTE has been studied under stratified CAR procedures (Liu et al., 2025). It is therefore desirable to extend these inference methods to the more general CAR framework considered in this paper. Second, multi-armed experiments and unequal allocations are ubiquitous in both clinical trials and economic studies. The theoretical properties of the unified family of multi-armed CAR

procedures and unequal allocation CAR procedures have been well studied by Zhang (2023) and Liu et al. (2025), respectively. It would thus be valuable to extend our framework to accommodate the multi-armed setting and unequal allocation setting.

Supplementary Materials

The supplementary materials contain proofs and additional numerical studies.

Acknowledgements

The authors are grateful for the detailed comments and suggestions of the reviewers. This work was supported by grants from National Key R&D Program of China (No. 2024YFA1013502), NSF of China (Grant Nos. U23A2064 and 12031005) and the Summit Advancement Disciplines of Zhejiang Province (Zhejiang Gongshang University - Statistics).

References

Anderson, S. J. and D. McKenzie (2022). Improving business practices and the boundary of the entrepreneur: A randomized experiment comparing training, consulting, insourcing, and outsourcing. *Journal of Political Economy 130*(1), 157–209.

Atkinson, A. C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika 69*(1), 61–67.

Baldi Antognini, A. and M. Zagoraiou (2011). The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika 98*(3), 519–535.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association 113*(524), 1784–1796.

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.

Burchardi, K. B., S. Gulesci, B. Lerva, and M. Sulaiman (2019). Moral hazard: Experimental evidence from tenancy contracts. *The Quarterly Journal of Economics 134*(1), 281–347.

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.

He, X., Y.-H. Hsu, and M. Hu (2010). Detection of treatment effects by covariate-adjusted expected shortfall. *The Annals of Applied Statistics 4*(4), 2114–2125.

Hjort, N. L. and D. Pollard (2011). Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*.

Hu, F. and L.-X. Zhang (2020). On the theory of covariate-adaptive designs. *arXiv preprint arXiv:2004.02994*.

Hu, Y. and F. Hu (2012). Asymptotic properties of covariate-adaptive randomization. *The*

Annals of Statistics 40(3), 1794 – 1815.

Jiang, L., P. C. Phillips, Y. Tao, and Y. Zhang (2023). Regression-adjusted estimation of quantile treatment effects under covariate-adaptive randomizations. *Journal of Econometrics 234*(2), 758–776.

Knight, K. (1998). Limiting distributions for l1 regression estimators under general conditions. *Annals of statistics 26*(2), 755–770.

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Kremer, J. M., H. K. Genant, L. W. Moreland, A. S. Russell, P. Emery, C. Abud-Mendoza, J. Szechinski, T. Li, Z. Ge, J.-C. Becker, et al. (2006). Effects of abatacept in patients with methotrexate-resistant active rheumatoid arthritis: a randomized trial. *Annals of internal medicine 144*(12), 865–876.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604–620.

Lin, Y. and Z. Su (2012). Balancing continuous and categorical baseline covariates in sequential clinical trials using the area between empirical cumulative distribution functions. *Statistics in medicine 31*(18), 1961–1971.

Lin, Y., M. Zhu, and Z. Su (2015). The pursuit of balance: an overview of covariate-adaptive randomization techniques in clinical trials. *Contemporary clinical trials 45*, 21–25.

Liu, H., F. Tu, and W. Ma (2023). Lasso-adjusted treatment effect estimation under covariate-

adaptive randomization. *Biometrika 110*(2), 431–447.

Liu, X., F. Hu, and W. Ma (2025). The properties of covariate-adaptive randomization procedures with possibly unequal allocation ratio. *The Annals of Applied Statistics 19*(2), 907–925.

Liu, Y. and F. Hu (2023). The impacts of unobserved covariates on covariate-adaptive randomized experiments. *The Annals of Statistics 51*(5), 1895–1920.

Liu, Y., L. Xia, and F. Hu (2025). Testing heterogeneous treatment effect with quantile regression under covariate-adaptive randomization. *Journal of Econometrics 249*, 105808.

Liu, Y., Y. Zhou, P. Li, and F. Hu (2024). Cluster-adaptive network a/b testing: From randomization to estimation. *Journal of Machine Learning Research 25*(170), 1–48.

Ma, W., F. Hu, and L. Zhang (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association 110*(510), 669–680.

Ma, W., P. Li, L.-X. Zhang, and F. Hu (2024). A new and unified family of covariate adaptive randomization procedures and their properties. *Journal of the American Statistical Association 119*(545), 151–162.

Ma, W., Y. Qin, Y. Li, and F. Hu (2020). Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association 115*(531), 1488–1497.

Ma, W., F. Tu, and H. Liu (2022). Regression analysis for covariate-adaptive randomization: a robust and efficient inference perspective. *Statistics in Medicine 41*(29), 5645–5661.

Ma, Z. and F. Hu (2013). Balancing continuous covariates based on kernel densities. *Contemporary clinical trials 34*(2), 262–269.

Morgan, K. L. and D. B. Rubin (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics 40*(2), 1263 – 1282.

Pocock, S. J. and R. Simon (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 103–115.

Qin, Y., Y. Li, W. Ma, H. Yang, and F. Hu (2022). Adaptive randomization via mahalanobis distance. *Statistica Sinica 1*, 1–40.

Rosenberger, W. F. and J. M. Lachin (2015). *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

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

Smith, R. L. (1984). Sequential treatment allocation using biased coin designs. *Journal of the Royal Statistical Society Series B: Statistical Methodology 46*(3), 519–543.

Taves, D. R. (1974). Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics 15*(5), 443–453.

Taves, D. R. (2010). The use of minimization in clinical trials. *Contemporary clinical trials 31*(2), 180–184.

van der Heijde, D., L. Klareskog, V. Rodriguez-Valverde, C. Codreanu, H. Bolosiu, J. Melo-

Gomes, J. Tornero-Molina, J. Wajdula, R. Pedersen, S. Fatenejad, et al. (2006). Comparison of etanercept and methotrexate, alone and combined, in the treatment of rheumatoid arthritis: two-year clinical and radiographic results from the tempo study, a double-blind, randomized trial. *Arthritis & Rheumatism 54*(4), 1063–1074.

Ye, T., Y. Yi, and J. Shao (2022). Inference on the average treatment effect under minimization and other covariate-adaptive randomization methods. *Biometrika 109*(1), 33–47.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases 27*(7), 365–375.

Zhang, L.-X. (2019). Theory on covariate-adaptive randomized clinical trials: Efficiency, selection bias and randomization methods. *arXiv preprint arXiv:1912.03636*.

Zhang, L.-X. (2023). Asymptotic properties of multi-treatment covariate adaptive randomization procedures for balancing observed and unobserved covariates. *arXiv preprint arXiv:2305.13842*.

Zhang, Y. and X. Zheng (2020). Quantile treatment effects and bootstrap inference under covariate-adaptive randomization. *Quantitative Economics 11*(3), 957–982.

Zhou, Q., P. A. Ernst, K. L. Morgan, D. B. Rubin, and A. Zhang (2018). Sequential rerandomization. *Biometrika 105*(3), 745–752.

Yuhang Tao

School of Mathematical Sciences, Zhejiang University, Hangzhou, China

E-mail: taoyh2023@zju.edu.cn

Li-Xin Zhang

School of Statistics and Data Science, Zhejiang Gongshang University, Hangzhou, China

Center for Data Sceince, Zhejiang University, Hangzhou, China

ORCID ID: 0000-0002-0121-0187

E-mail: stazlx@mail.zjgsu.edu.cn