Counterfactual Treatment Effects: Estimation and Inference

Yu-Chin Hsu[a,b,c], Tsung-Chih Lai[d,*], and Robert P. Lieli[e]

[a]Institute of Economics, Academia Sinica, Taiwan
[b]Department of Finance, National Central University, Taiwan
[c]Department of Economics, National Chengchi University, Taiwan
[d]Department of Economics, Feng Chia University, Taiwan
[e]Department of Economics and Business, Central European University, Budapest

February 9, 2020

Abstract

This paper proposes statistical methods to evaluate the quantile counterfactual treatment effect (QCTE) if one were to change the composition of the population targeted by a status quo program. QCTE enables a researcher to carry out an *ex-ante* assessment of the distributional impact of certain policy interventions or to investigate the possible explanations for treatment effect heterogeneity. Assuming unconfoundedness and invariance of the conditional distributions of the potential outcomes, QCTE is identified and can be nonparametrically estimated by a kernel-based method. Viewed as a random function over the continuum of quantile indices, the estimator converges weakly to a zero mean Gaussian process at the parametric rate. We propose a multiplier bootstrap procedure to construct uniform confidence bands, and provide similar results for average effects and for the counterfactually treated subpopulation. We also present Monte Carlo simulations and two counterfactual exercises that provide insight into the heterogeneous earnings effects of the Job Corps training program in the United States.

Keywords: program evaluation, counterfactual analysis, multiplier bootstrap, Job Corps
JEL Classification: C14, C31, J30

1 Introduction

The program evaluation literature has shifted a fair amount of attention from internal validity to questions related to external validity over the past decade. In addition to credibly identifying and estimating treatment effects for the population from which data are actually drawn, recent studies have also focused on extrapolating these estimates to new environments that the program might be extended to. For example, Stuart et al. (2011), Kline and Tamer (2018), and Andrews and Oster (2019) address the generalizability of treatment effect estimates obtained from a randomized trial where the sample is not representative of the ultimate target population. Hotz et al. (2005), Allcott (2015), Hartman et al. (2015), and Dehejia et al. (2019) consider extrapolating treatment effect estimates from one location to another. While these studies provide

[*] Corresponding author. Department of Economics, Feng Chia University, 100 Wenhwa Road, Seatwen, Taichung, 40724 Taiwan. E-mail: tclai@fcu.edu.tw

sufficient conditions for nonparametric identification of extrapolated effects, a flexible model-free approach for estimation and inference is still lacking in this growing literature, even in the simplest case where program participation is randomly assigned.

This paper looks to fill this gap by developing statistical methods to extrapolate the treatment effects estimated for a *status quo* population to a *counterfactual* population.[1] Our approach is not limited to extrapolation from a randomized experiment, nor do we confine our analysis to predicting only the average treatment effect. To fix ideas, we present two examples that help clarify the counterfactual scenarios addressed in the paper.[2]

Example 1 (Independent Policy Implementation). Consider a job training program for a given population in a given location. Information is available about individual earnings, participation status (which may not be randomly assigned), and other observed characteristics. A policy maker is planning to expand the program to a new location where only individual characteristics are currently observed. The first task is to predict the program effect in the new location prior to actual implementation.

Example 2 (Dependent Policy Implementation). The policy maker is planning to manipulate (the distribution of) individual characteristics in a population currently targeted by some training program. For example, direct subsidies can be provided to individuals below an income threshold to change their pre-treatment level of income. Prior to deciding on actual implementation, the policy maker wants to predict the resulting change in the program effect, especially for the poorest prospective participants.

To describe the effect of extending or modifying a status quo treatment, we introduce a parameter called the quantile counterfactual treatment effect (QCTE), which we view as an unknown real-valued function defined over all possible quantile indices in the unit interval. QCTE enables one to carry out an ex-ante assessment of the distributional impacts of the policy interventions outlined above or to investigate possible explanations for treatment effect heterogeneity. Our framework is fully nonparametric in that we only restrict QCTE via general conditions on the distributions of observables and potential outcomes. Thus, we allow for heterogeneous effects across quantiles and capture any distributional impact the modified program might have. We also discuss the average counterfactual treatment effect (ACTE) and the average and quantile counterfactual treatment effects for the treated (ACTT and QCTT) in this paper, but the exposition will focus on QCTE.

To identify QCTE, we start by assuming that the status quo treatment assignment mechanism satisfies unconfoundedness, i.e., any systematic relationship between the potential outcomes and the treatment assignment can be accounted for by a vector X of observed covariates. In addition, we assume that the conditional distributions of the status quo and counterfactual potential outcomes are identical given $X = x$. This implies, for example, that for any individual with $X = x$, the expected treatment effect is the same regardless of whether the individual is drawn from the status quo or counterfactual population. In other words, we attribute any difference between the status quo and counterfactual treatment effects to the difference in the distribution of X across the two populations rather than the treatment operating in a fundamentally different way. This external validity assumption, while widely used in the literature, is admittedly strong and needs to be evaluated on a case-by-case basis. Alternatively, one could interpret this assumption as part of a thought experiment: how would a quantile or average treatment effect change if one were to hold constant the conditional distribution of the potential outcomes given X, but implemented changes in the marginal

[1] We refer to the two populations as "status quo" and "counterfactual" rather than "reference" and "target" to emphasize that the extrapolated population may be contrary to fact. See Example 2 below.

[2] For formal definitions, please see Assumption 2.2.

distribution of X? Such counterfactuals, similar in spirit to the decomposition by DiNardo et al. (1996), can provide insight into the source of treatment effect heterogeneity; see our application in Section 6.

Given the assumptions described above, QCTE can be nonparametrically identified and estimated in the following steps. First, we use a Nadaraya-Watson estimator to construct the conditional distribution functions of the status quo potential outcomes given $X = x$ using observations from the status quo population. Second, we integrate out x using the empirical distribution of the X-observations drawn from the *counterfactual* population and thus obtain estimates of the (unconditional) distribution functions of the counterfactual potential outcomes. Finally, after ensuring monotonicity, we invert the estimated distribution functions for the treated and control cases to obtain the corresponding quantile functions. Taking the difference at any given quantile index gives the estimated value of QCTE at that point.[3] Moreover, we show that this QCTE estimator, viewed as a random function over the continuum of quantile indices, converges weakly to a zero mean Gaussian process at the parametric rate. Exploiting this result, we propose a multiplier bootstrap procedure to construct uniform confidence bands for QCTE. We also propose estimation and inference methods for ACTE and state similar results.

We illustrate the use of the proposed methods by Monte Carlo simulations and two counterfactual exercises highlighting the heterogeneous impact of Job Corps, the largest and most comprehensive labor market program in the United States. In the first exercise, we take the earnings structure for males as given and replace the distribution of individual characteristics among males with that of females. We then estimate QCTE and repeat the procedure with the role of the two genders exchanged. Comparing QCTE with the actual QTE sheds light on the possible mechanisms behind the heterogeneity of the program effect by gender. In the second exercise, we hypothetically increase the education level of individuals with the most disadvantaged background characteristics. This exercise allows us to examine whether Job Corps works primarily for those who already have reasonably good labor market opportunities.

Our paper is related to a number of previous studies on extrapolating treatment effect estimates to other populations; see Athey and Imbens (2017) for a review. A large subset of this literature is concerned with instrumental variable settings, where the aim is to extrapolate the local average treatment effect, the average treatment effect for the complier subpopulation, to other subpopulations or the entire population of interest (Angrist, 2004; Angrist and Fernandez-Val, 2013; Kowalski, 2016; Brinch et al., 2017; Bertanha and Imbens, 2019). Of these papers, Angrist and Fernandez-Val (2013) is the most closely related to ours in terms of identification conditions. They assume that the complier populations associated with different instruments are systematically different from each other only in terms of a set of observed covariates. This makes it possible to extrapolate the local average treatment effect from one instrument to another or to the treated/overall population. Similarly, as explained above, our extrapolation scheme also relies on the premise that the difference between the status quo and the counterfactual environment is limited to the marginal distribution of the covariates X.

There is a more direct connection to previous work on estimating counterfactual distributions. For example, Firpo et al. (2009) use a recentered influence function regression approach to estimate the impact of a marginal increase in the covariates on the unconditional distribution of the outcome. Rothe (2010) and Chernozhukov et al. (2013) consider situations where the covariates are either drawn from a completely new distribution or are transformed. While our estimation procedure builds on Rothe (2010) to a large degree,

[3] As a special case, one may integrate with respect to the empirical distribution of the *status quo* covariates in the second step described above to obtain an estimator for the distribution functions of the status quo potential outcomes. This estimator is an alternative to the inverse probability weighted (IPW) estimator proposed by Donald and Hsu (2014) and is later shown to be first-order asymptotically equivalent to the IPW estimator.

there are several technical distinctions, which we highlight as follows.

First, we generalize the asymptotic analysis from a purely predictive setting to treatment effect models. This has non-trivial technical consequences, for example, the estimation error in the first stage will involve the propensity score. Furthermore, we also conduct inference for the treated subset of the counterfactual population, which is of course not defined in Rothe's (2010) simpler setup. Second, we use the multiplier bootstrap instead of the nonparametric one to simulate the asymptotic distribution of our estimators and conduct uniform inference. The main reason is computational convenience—the nonparametric bootstrap is potentially very time-consuming given that the entire nonparametric estimation procedure needs to be replicated for each new draw. In contrast, the computational burden of the multiplier bootstrap is reduced substantially as the resampling procedure is simultaneously simulated.[4] Third, we apply a new monotonization method to ensure that the unconditional distribution function estimators obtained in step two of the procedure are weakly increasing before we invert them. Non-monotonicity can arise, because of the use of a higher-order boundary kernel, which assigns negative weights to some observations. Rothe (2010) deals with this problem via a reweighting procedure, while we use the method proposed by Hsu et al. (2019), which simply replaces any downward step in the distribution function estimate with a constant piece and is very easy to implement.

In comparison with Chernozhukov et al. (2013), our fully nonparametric estimation procedure makes it unnecessary to specify parametric models for the conditional distribution or quantile functions of the potential outcomes. If these models are misspecified, the semiparametric approach by *ibid.* will generally result in inconsistent estimates. It is also worth noting that despite being fully nonparametric, the convergence rate of our estimator is independent of the dimension of X. Nevertheless, in our framework the counterfactual potential outcomes are not observed (even partially), so we can only consider the composition effect (a ceteris paribus change in the distribution of covariates) but not the structure effect (a ceteris paribus change in the conditional distribution).

Finally, the paper also has links to the literature on double randomization designs (Imai et al., 2013; Wunsch and Strobl, 2018), concerned with experimental identification of treatment effects mediated by specific variables. Indeed, it would be possible in our framework to re-interpret X as a set of (manipulable) mediators. If, in the simplest case, X and the treatment are both randomly assigned, the unconfoundedness assumption is satisfied and, furthermore, the distribution of the potential outcomes is independent of X. Our method can then be used to predict the treatment effect from a repeated experiment performed on the same population, where the mediator values are drawn from another distribution, while the treatment need not actually be repeated.

The remainder of this paper is organized as follows. Section 2 introduces the model framework, the parameters of interest, and the identification strategy. Section 3 covers the estimation procedure and the asymptotic properties under regularity conditions. In Section 4 we establish the validity of the multiplier bootstrap and provide uniform inference methods. Section 5 presents Monte Carlo simulations and Section 6 the empirical applications. Section 7 extends the analysis to the average case along with an overidentification test. Finally, Section 8 concludes. An on-line supplementary appendix that includes proofs of asymptotic properties of the estimators, and results on average and quantile counterfactual treatment effects for the treated is provided.[5]

[4] The tradeoff is that the multiplier bootstrap requires consistent estimation of the variance functions. We provide such estimators in Section 4.3.

[5] Supplemental material is available at `https://www.dropbox.com/s/2f29pdvdotct3go/Counterfactual_supplement.pdf`.

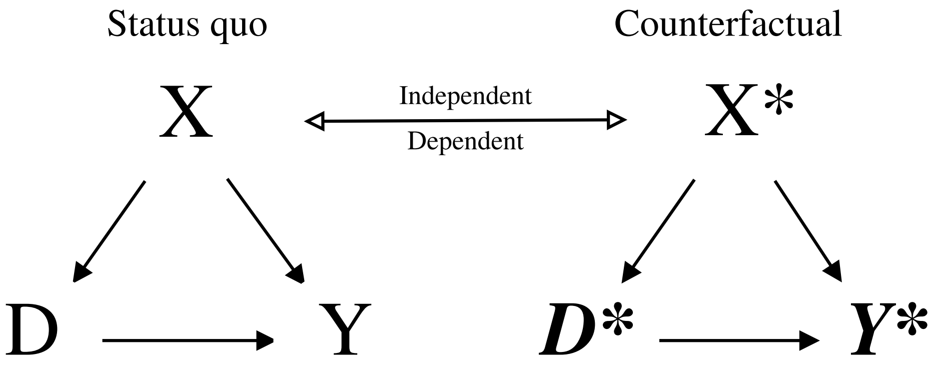


Figure 1. Model framework.

2 Model Framework and Identification

2.1 Model and Parameters of Interest

Following the Rubin causal model, we let $D \in \{0, 1\}$ be a binary treatment indicator and Y_d be the potential outcomes for $d = 0, 1$ in the status quo environment. That is, Y_1 is the outcome if an individual is exogenously assigned to the treatment ($D = 1$) and Y_0 is the outcome in the absence of treatment ($D = 0$). The actual observed outcome is $Y = DY_1 + (1 - D)Y_0$. We also observe a k-dimensional vector of pretreatment covariates $X = (X_1, \ldots, X_k)$ in the status quo environment and another covariate vector $X^* = (X_1^*, \ldots, X_k^*)$ in the counterfactual environment which is of the same dimension as X. The relationship between X and X^* depends on the context and we consider two cases in this paper: (i) X and X^* are statistically independent, and (ii) X^* is a deterministic transformation of X, i.e., $X^* = \pi(X)$ for some known function π. The corresponding treatment indicator, outcome, and potential outcomes in the counterfactual environment are denoted as D^*, Y^*, and Y_d^* for $d = 0, 1$, respectively, with $Y^* = D^*Y_1^* + (1 - D^*)Y_0^*$. Note that since the treatment has not been implemented in the counterfactual environment yet, neither D^* nor Y^* (and therefore Y_0^* and Y_1^*) is observed in our model. Figure 1 provides a graphical illustration of the model framework. The observed variables X, D, and Y in the status quo environment and X^* in the counterfactual environment are indicated in normal font. In contrast, the unobservables D^* and Y^* are shown in bold italics.

The average treatment effect (ATE) and the quantile treatment effect (QTE) are two commonly used parameters for evaluating the overall impact of a treatment or program. It is actually more precise to think of QTE as a family of parameters corresponding to various quantiles of interest. Thus, QTE is well suited for assessing treatment effect heterogeneity along the outcome distribution. Analogous to ATE and QTE, we define the average counterfactual treatment effect (ACTE) as the mean difference between Y_1^* and Y_0^*,

$$\delta^* = \mathrm{E}(Y_1^*) - \mathrm{E}(Y_0^*), \tag{2.1}$$

and the quantile counterfactual treatment effect (QCTE) as the difference between two quantile functions of Y_1^* and Y_0^*,

$$\delta^*(\tau) = \mathrm{Q}_{Y_1^*}(\tau) - \mathrm{Q}_{Y_0^*}(\tau), \quad \tau \in [0, 1], \tag{2.2}$$

where $\mathrm{Q}_{Y_d^*}(\tau) = \inf\{y \in \mathcal{Y} : F_{Y_d^*}(y) \geq \tau\}$ is the quantile function of Y_d^* with $\mathcal{Y}$ the support of Y and $F_{Y_d^*}(y)$ the distribution function of Y_d^*. Note that in developing the asymptotic theory for our estimator, we will treat QCTE as a function-valued parameter defined over $\tau \in [0, 1]$.

The policy maker sometimes may be interested in the treatment effects for the treated subgroup. In these cases, we can also consider the average counterfactual treatment effect for the treated (ACTT) and the quantile counterfactual treatment effect for the treated (QCTT) defined as

$$\delta_t^* = \mathrm{E}(Y_1^*|D^* = 1) - \mathrm{E}(Y_0^*|D^* = 1) \quad \text{and} \quad \delta_t^*(\tau) = \mathrm{Q}_{Y_1^*|D^*}(\tau|1) - \mathrm{Q}_{Y_0^*|D^*}(\tau|1), \tag{2.3}$$

where the expectation and quantile operators are taken with respect to the conditional distribution of Y_d^* given $D^* = 1$. However, since the analysis is similar to the overall cases and the quantile case is much more involved and challenging, we will focus primarily on QCTE in the exposition and relegate the discussion of ACTE to Section 7 and that of ACTT and QCTT to the supplementary appendix.

2.2 Identification

What makes the identification of counterfactual parameters challenging is that none of Y_0^*, Y_1^* and D^* are observed. We therefore need to employ rather strong identification assumptions that are nevertheless standard in the literature. Let $\mathcal{X}$ and $\mathcal{X}^*$ be the support of X and X^*, respectively, and let $p(x) = \mathrm{P}(D = 1|X = x)$ denote the propensity score for $x \in \mathcal{X}$. The first assumption ensures the internal validity of status quo estimates.

Assumption 2.1 (Unconfoundedness)**.**

(i) D is conditionally independent of (Y_0, Y_1) given X.

(ii) $p(x)$ is bounded away from 0 and 1 for all $x \in \mathcal{X}$.

Assumption 2.1(i) is also known as ignorability, selection on observables, or conditional independence. This assumption requires that conditional on X, there are no other unobserved confounders systematically associated with both the treatment assignment and the potential outcomes. The second part of Assumption 2.1, usually referred to as the overlap condition, requires the support of X to be the same across the treated and untreated subpopulations. If this condition is not met initially, one solution would be trimming and redefining $\mathcal{X}$ (see Busso et al. (2009) and Lechner and Strittmatter (2019) for the discussion of this issue.) Note that Assumption 2.1 allows the use of observational data in evaluating the status quo treatment. The next two assumptions make the extrapolation of the treatment effects possible.

Assumption 2.2. Assume that X^* satisfies that one of the following two conditions:

(i) (Independent Policy Implementation). X^* is independent of (Y_0, Y_1, D, X).

(ii) (Dependent Policy Implementation). X^* is a deterministic transformation of X, i.e., there exists a known function $\pi : \mathbb{R}^d \to \mathbb{R}^d$ such that $X^* = \pi(X)$.

Assumption 2.2 requires that either X^* is drawn from a different population than (Y_0, Y_1, D, X) or X^* is a deterministic transformation of X. Following Rothe (2010), we call the former case an independent policy implementation and the latter a dependent policy implementation. Furthermore, under Assumption 2.2, we have that the potential outcome distributions are independent of X^* conditional on X, i.e., X^* is unconfounded.[6] Assumption 2.2 generally deserves justification in a given application but it could also be interpreted as representing the types of *thought experiments* whose outcomes we are capable of predicting.

Assumption 2.3 (Invariance of Conditional Distributions)**.**

(i) The distribution of Y_d^* conditional on X^* is identical to the distribution of Y_d conditional on X for $d = 0, 1$. In other words, $F_{Y_d^*|X^*}(y|x) = F_{Y_d|X}(y|x)$ for all $x \in \mathcal{X}^*$.

(ii) $\mathcal{X}^*$ is a subset of $\mathcal{X}$.

The first part of Assumption 2.3 stipulates that the difference between the status quo and counterfactual treatment effects arises solely from the different marginal distributions of X and X^*. This assumption appears frequently in the decomposition literature (Firpo et al., 2009; Fortin et al., 2011; Chernozhukov et al. 2013), and is in the same spirit as the unconfounded location assumption in Hotz et al. (2005), the policy invariance condition in Heckman and Vytlacil (2005, 2007), and the conditional effect ignorability in Angrist and Fernandez-Val (2013). In addition, if one follows Rothe (2010) to specify a nonparametric structural model of the status quo and counterfactual potential outcomes as

$$Y_d = m_d(X, \varepsilon_d) \quad \text{and} \quad Y_d^* = m_d(X^*, \varepsilon_d), \quad d = 0, 1,$$

where m_d is unknown and ε_d represents unobserved characteristics, then a sufficient condition for Assumption 2.3(i) is the independence between ε_d and (X, X^*) as imposed in Rothe (2010). In other words, we follow Rothe (2010) to assume that both environments are in the same conditional equilibrium and consider counterfactual changes in the covariate distribution instead of changes in the unobserved characteristics.

Assumption 2.3(ii) is a support condition that is weaker than the complete overlap imposed in Hotz et al. (2005). This assumption is invoked so that our model need not be tied to any specific functional form. Nonetheless, the cost is that the possibility of extrapolating beyond the status quo support is ruled out. If Assumption 2.3(ii) is violated, one could drop units in the counterfactual environment with covariates outside the common support and redefine QCTE relative to the new support.

Lemma 2.1. Suppose Assumptions 2.1–2.3 hold. QCTE is identified by

$$\delta^*(\tau) = \inf_{y \in \mathcal{Y}} \left\{ \int_{\mathcal{X}} F_{Y|D,X}(y|1, x)\, \mathrm{d}F_{X^*}(x) \geq \tau \right\} - \inf_{y \in \mathcal{Y}} \left\{ \int_{\mathcal{X}} F_{Y|D,X}(y|0, x)\, \mathrm{d}F_{X^*}(x) \geq \tau \right\}.$$

To see Lemma 2.1, we first note thatunder Assumption 2.3, the distribution function is given by $F_{Y_d^*}(y) = \int_{\mathcal{X}} F_{Y_d|X}(y|x)\, \mathrm{d}F_{X^*}(x)$. As D is independent of Y_d conditional on X by Assumption 2.1, $F_{Y_d|X}(y|x)$ is identified by $F_{Y_d|X}(y|x) = F_{Y_d|D,X}(y|d, x) = F_{Y|D,X}(y|d, x)$ where the last equality holds because of $Y = Y_d$ for $D = d$. Once the distribution function is identified, the quantile functions and QCTE are identified as well. A more formal argument is provided in Appendix A.

[6] In case (i), (Y_0, Y_1, X) is jointly independent of X^*, which of course implies that (Y_0, Y_1) is independent of X^* conditional on X. In case (ii), X^* is a function of X, i.e., X^* is a constant given the value of X. Then X^* is independent of any other random variable given the value of X.

3 Estimation and Asymptotic Properties

3.1 Estimation

Given the identification result in Lemma 2.1, QCTE can be estimated by the following steps. First, construct estimators for the conditional distribution functions $F_{Y_d|X}(y|x)$ for $d = 0, 1$ using data from the status quo environment. Second, average with respect to the empirical measure of X^* for the unconditional distribution functions $F_{Y_d^*}(y)$. Third, eliminate any non-monotonicity and then take inversion to obtain estimates of the quantile functions $\mathrm{Q}_{Y_d^*}(\tau)$. Finally, define the QCTE estimator as the difference between the estimates of $\mathrm{Q}_{Y_1^*}(\tau)$ and $\mathrm{Q}_{Y_0^*}(\tau)$.

To be specific, for the first-step conditional distribution function estimation, we follow Rothe (2010) to use a kernel-based Nadaraya-Watson estimator:

$$\widetilde{F}_{Y_d|X}(y|x) = \frac{\sum_{i=1}^n 1\{Y_i \le y\}\, 1\{D_i = d\} K_{x,h}(X_i - x)}{\sum_{i=1}^n 1\{D_i = d\} K_{x,h}(X_i - x)}, \tag{3.1}$$

where $1\{\cdot\}$ denotes the indicator function and $K_{x,h}(\cdot) = h^{-k} K_x(\cdot/h)$ is a higher-order boundary kernel whose shape adapts when x is near the boundary of $\mathcal{X}$ with $h = h_n$ the bandwidth. Here, we implicitly assume that the underlying covariates are continuous. If X has both continuous and discrete components, one can either adjust the kernel for frequency (sample splitting) or employ the smoothing method advocated by Li and Racine (2008).[7] As the rate of convergence of the estimator will not be affected in either case, we will for simplicity assume that X is continuous throughout the paper.

In the second step, we evaluate $\widetilde{F}_{Y_d|X}(y|x)$ at the sample observations X_j^* from the counterfactual environment and take the sample average to estimate $F_{Y_d^*}(y)$. That is,

$$\widetilde{F}_{Y_d^*}(y) = \frac{1}{n^*} \sum_{j=1}^{n^*} \widetilde{F}_{Y_d|X}(y|X_j^*). \tag{3.2}$$

However, a practical issue is that $\widetilde{F}_{Y_d^*}(y)$ may be non-monotonic or lie outside the unit interval in finite samples due to negative weights introduced by the higher-order boundary kernel. This problem can be circumvented by either using the reweighting method in Rothe (2010), the rearranging method in Chernozhukov et al. (2009, 2010), or the monotonization method in Hsu et al. (2019) which is adopted in this paper. Specifically, define the functionals ϕ_1, ϕ_2, and ϕ so that for any function g with $\sup_{y \in \mathcal{Y}} g(y) > 0$,

$$\phi_1(g)(y) = \max\left\{0, \sup_{y' \le y} g(y')\right\}, \qquad \phi_2(g)(y) = \frac{g(y)}{\sup_{y' \in \mathcal{Y}} g(y')}, \qquad \phi = \phi_1 \circ \phi_2.$$

The properly monotonized version of (3.2) is then defined as

$$\widehat{F}_{Y_d^*}(y) = \phi(\widetilde{F}_{Y_d^*})(y). \tag{3.3}$$

Since the preliminary estimator $\widetilde{F}_{Y_d^*}$ is already a step function over $\mathcal{Y}$, the functional ϕ_1 simply replaces any downward steps by the value of the last upward step and then eliminates any negative values by setting them to 0. The functional ϕ_2 is used to rescale $\widetilde{F}_{Y_d^*}$ so that its maximum value is always 1. Thus, the

[7] For example, if $X = (X_1, X_2)$ with X_1 discrete and X_2 continuous, the frequency-based kernel is defined as $K_h(X - x) = 1\{X_1 = x_1\} K((X_2 - x_2)/h)/h$. One can also smooth the discrete variable by replacing $1\{X_1 = x_1\}$ with $1\{X_1 = x_1\} + \eta\, 1\{X_1 \ne x_1\}$, where $\eta \in (0, 1)$ and $\eta = \eta_n \to 0$ as $n \to \infty$.

monotonized estimator $\widehat{F}_{Y_d^*}(y)$ is a proper distribution function. Later we will show that $\widehat{F}_{Y_d^*}(y)$ and $\widetilde{F}_{Y_d^*}(y)$ are first-order asymptotically equivalent under the regularity conditions given below. This equivalence allows us to apply $\widehat{F}_{Y_d^*}(y)$ in practice (an easy-to-implement procedure is provided in Appendix B) and focus on the limiting behavior of $\widetilde{F}_{Y_d^*}(y)$ in the theoretical derivation. Finally, since the distribution function estimator is now invertible, our QCTE estimator is defined as

$$\widehat{\delta}^*(\tau) = \widehat{\mathrm{Q}}_{Y_1^*}(\tau) - \widehat{\mathrm{Q}}_{Y_0^*}(\tau), \tag{3.4}$$

where

$$\widehat{\mathrm{Q}}_{Y_d^*}(\tau) = \inf\{y \in \mathcal{Y} : \widehat{F}_{Y_d^*}(y) \geq \tau\}. \tag{3.5}$$

3.2 Regularity Conditions

We gather all regularity conditions for asymptotic analysis in this section. Similar conditions can be found in Rothe (2010). For a k-dimensional vector u and a k-dimensional vector of non-negative integers γ, let $|u| = \sum_{s=1}^{k} u_s$ and $u^\gamma = \prod_{s=1}^{k} u_s^{\gamma_s}$. Let r denote the order of the kernel function used in (3.1).

Assumption 3.1 (Sampling process)**.**

(i) $\{(Y_i, D_i, X_i)\}_{i=1}^n$ is a random sample from the joint distribution of (Y, D, X) and $\{X_j^*\}_{j=1}^{n^*}$ is a random sample from the distribution of X^*.

(ii) $\lim_{n,n^*\to\infty} n/n^* = \lambda$, where $0 < \lambda < \infty$.

Assumption 3.2 (Distributions of X and X^*)**.**

(i) $\mathcal{X}$ and $\mathcal{X}^*$ are Cartesian products of compact intervals. In other words, $\mathcal{X} = \prod_{s=1}^{k}[x_{\ell s}, x_{us}] \equiv [x_\ell, x_u]$, and $\mathcal{X}^* = \prod_{s=1}^{k}[x_{\ell s}^*, x_{us}^*] \equiv [x_\ell^*, x_u^*] \subseteq \mathcal{X}$.

(ii) The density functions $f_X(x)$ and $f_{X^*}(x)$ are bounded away from 0 on $\mathcal{X}$ and $\mathcal{X}^*$, respectively.

(iii) $f_X(x)$ and $f_{X^*}(x)$ are r-times differentiable on the interior of $\mathcal{X}$ and $\mathcal{X}^*$, respectively, and the derivatives are uniformly continuous and bounded.

Assumption 3.3 (Distribution of Y_d^*)**.**

(i) Y_d^* has a compact support $[y_{d\ell}^*, y_{du}^*] \subseteq \mathcal{Y}$. Without loss of generality, assume that $\mathcal{Y} \equiv [0, \bar{y}]$ with $\bar{y} < \infty$.

(ii) $F_{Y_d^*}(y)$ is continuous on $\mathcal{Y}$.

(iii) The density function $f_{Y_d^*}(y)$ is bounded away from 0 and is two-times differentiable on $\mathcal{Y}$.

Assumption 3.4 (Conditional Probability and Distribution)**.**

(i) $p(x)$ is r-times differentiable on the interior of $\mathcal{X}$ and the derivatives are uniformly continuous and bounded.

(ii) $F_{Y_d|X}(y|x)$ is r-times differentiable with respect to x on the interior of $\mathcal{X}$ and the derivatives are uniformly continuous and bounded.

Assumption 3.5 (Higher-Order Boundary Kernel)**.** Let $\mathcal{D}_x = \{u \in [-1,1] : x_\ell \leq x + hu \leq x_u\}$. The kernel function K_x of order r satisfies:

(i) $\int_{\mathcal{D}_x} K_x(u)\,\mathrm{d}u = 1$.

(ii) $\int_{\mathcal{D}_x} u^\gamma K_x(u)\,\mathrm{d}u = 0$ for all $|\gamma| = 1, \ldots, r-1$.

(iii) $\int_{\mathcal{D}_x} |u^\gamma K_x(u)|\,\mathrm{d}u < \infty$ for $|\gamma| = r$.

(iv) $K_x(u) = 0$ if $|u| > 1$.

(v) $K_x(u)$ is r-times differentiable with respect to both u and x, and the derivatives are uniformly continuous and bounded.

Assumption 3.6 (Bandwidth)**.** As $n \to \infty$, the bandwidth $h = h_n$ satisfies:

(i) $h \to 0$.

(ii) $n^{1/2}h^k \big/ \log n \to \infty$.

(iii) $n^{1/2}h^r \to 0$.

Assumption 3.1 impose conditions on the sampling process. Assumption 3.2 requires the distribution of the covariates to be continuous and sufficiently smooth. To estimate QCTE as a function of $\tau \in [0,1]$ at the parametric rate, $f_{Y_d^*}(y)$ needs to be bounded away from 0. This, of course, entails compact support.[8] Similarly to Assumption 3.2, Assumption 3.4 requires the smoothness of the propensity score as well as the conditional distribution function. Assumption 3.5 prescribes the use of a higher-order boundary kernel, which reduces first-stage estimation bias in both interior and boundary regions (see Ruppert and Wand, 1994). Assumption 3.6 determines the rate of convergence of the bandwidth toward 0. As mentioned in Rothe (2010), if h is of the form $h = cn^{-\theta}$ for some constants $c > 0$ and $\theta > 0$, then θ must lie in the interval $(1/2r, 1/2k)$ for $r > k$, meaning that the order of the kernel must exceed the dimension of X.

3.3 Asymptotic Properties

We now investigate the asymptotic properties of $\widehat{F}_{Y_d^*}(y)$ under regularity conditions. The asymptotics of $\widehat{F}_{Y_d^*}(y)$ are governed by the asymptotics of $\widetilde{F}_{Y_d^*}(y)$ which can be derived using arguments similar to Rothe (2010). Let $\boldsymbol{y} = (y_0, y_1)^T$, $\boldsymbol{F}(\boldsymbol{y}) = (F_{Y_0^*}(y_0), F_{Y_1^*}(y_1))^T$, $\widehat{\boldsymbol{F}}(\boldsymbol{y}) = (\widehat{F}_{Y_0^*}(y_0), \widehat{F}_{Y_1^*}(y_1))^T$, and $Z = (Y, D, X)$. Note that we drop asterisks on $\boldsymbol{F}$ and $\widehat{\boldsymbol{F}}$ for notational simplicity.

Lemma 3.1. Suppose Assumptions 2.1–2.3 and 3.1–3.6 hold. We then have:

$$\sqrt{n}\Big(\widehat{\boldsymbol{F}}(\cdot) - \boldsymbol{F}(\cdot)\Big) \Rightarrow \mathcal{F}(\cdot),$$

where $\mathcal{F}(\boldsymbol{y}) = (\mathcal{F}_0(y_0), \mathcal{F}_1(y_1))^T$ is a two-dimensional zero mean Gaussian process with covariance function $\Psi^F(\boldsymbol{y}, \boldsymbol{y}') = \mathrm{E}[\varrho^F(\boldsymbol{y}, Z)\varrho^F(\boldsymbol{y}', Z)^T] + \mathrm{E}[\varphi^F(\boldsymbol{y}, X^*)\varphi^F(\boldsymbol{y}', X^*)^T]$, and the convergence takes place in $\ell^\infty(\mathcal{Y}) \times$

[8] If $\mathcal{Y} = \mathbb{R}$, one can still estimate QCTE at the parametric rate uniformly over some compact subset of the unit interval. That is, the convergence holds in the space $\ell^\infty([\epsilon, 1-\epsilon])$ provided that the density function is strictly positive on the interval between the ϵ-th quantile and the (1-ϵ)-th quantile of $Y_d^*(y)$. Hence, by focusing on the "interior" quantiles, our results extend easily to cases where there are mass points at the boundary points of $\mathcal{Y}$. See our empirical examples in Section 6.

$\ell^\infty(\mathcal{Y})$, where $\ell^\infty(\mathcal{Y})$ is the space of bounded functions over $\mathcal{Y}$. Here, $\varrho^F(\boldsymbol{y}, Z) = (\varrho_0^F(y_0, Z), \varrho_1^F(y_1, Z))^T$ and $\varphi^F(\boldsymbol{y}, X^*) = (\varphi_0^F(y_0, X^*), \varphi_1^F(y_1, X^*))^T$ are defined as

$$\begin{aligned}\varrho_d^F(y, Z) &= \frac{1\{D = d\}\big[1\{Y \le y\} - F_{Y_d|X}(y|X)\big]}{p(X)^d[1 - p(X)]^{1-d}} \frac{f_{X^*}(X)}{f_X(X)}, \\ \varphi_d^F(y, X^*) &= \sqrt{\lambda}\big[F_{Y_d|X}(y|X^*) - F_{Y_d^*}(y)\big].\end{aligned} \tag{3.6}$$

The proof of Lemma 3.1 can be found in the supplementary material. Here, we give a brief outline of the argument. We first show that $\sqrt{n}(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y))$ is asymptotically linear with the influence function representation:

$$\begin{aligned}\sqrt{n}\Big(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\Big) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1\{D_i = d\}\big[1\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\big]}{p(X_i)^d[1 - p(X_i)]^{1-d}} \frac{f_{X^*}(X_i)}{f_X(X_i)} \\ &\quad + \frac{1}{\sqrt{n^*}} \sum_{j=1}^{n^*} \sqrt{\lambda}\big[F_{Y_d|X}(y|X_j^*) - F_{Y_d^*}(y)\big] + o_p(1) \\ &\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varrho_d^F(y, Z_i) + \frac{1}{\sqrt{n^*}} \sum_{j=1}^{n^*} \varphi_d^F(y, X_j^*) + o_p(1).\end{aligned}$$

As the functions $\varrho_d^F(y, \cdot)$, $y \in \mathcal{Y}$, and $\varphi_d^F(y, \cdot)$, $y \in \mathcal{Y}$, belong to Donsker classes and the Cartesian product of two Donsker classes is still a Donsker class (van der Vaart, 2000), Lemma 3.1 holds by the functional central limit theorem for $\widetilde{\boldsymbol{F}} = (\widetilde{F}_{Y_0^*}, \widetilde{F}_{Y_1^*})^T$ in place of $\widehat{\boldsymbol{F}}$. Finally, we show that $\widehat{F}_{Y_d^*}(y)$ and $\widetilde{F}_{Y_d^*}(y)$ are first-order asymptotic equivalent in that $\sup_{y\in\mathcal{Y}} |\widehat{F}_{Y_d^*}(y) - \widetilde{F}_{Y_d^*}(y)| = o_p(n^{-1/2})$, which completes the proof.

Several remarks on Lemma 3.1 are worth noting. First, the estimator avoids the curse of dimensionality in that it converges to a Gaussian process at the parametric rate despite the nonparametric estimation in the first stage. Second, there is no cross-product term in the expression for the asymptotic covariance function $\Psi^F(\boldsymbol{y}, \boldsymbol{y}')$ as $\varrho_d^F(y, Z)$ and $\varphi_d^F(y, X^*)$ are always uncorrelated regardless of the relationship between X and X^*. Third, ϱ_d^F accounts for the estimation error resulting from the first-stage estimation of $F_{Y_d|X}$. If the conditional distribution were known and need not be estimated, then φ_d^F alone would be the influence function of $\widehat{F}_{Y_d^*}$. Fourth, if we let $X^* = X$ and $\lambda = 1$, then the sum of $\varrho_d^F(y, Z)$ and $\varphi_d^F(y, X^*)$ would become

$$\frac{1\{D = d\}\big[1\{Y \le y\} - F_{Y_d|X}(y|X)\big]}{p(X)^d[1 - p(X)]^{1-d}} + F_{Y_d|X}(y|X) - F_{Y_d}(y),$$

which corresponds to the influence function of the IPW estimator proposed by Donald and Hsu (2014). In other words, our kernel-based imputation estimator is asymptotically equivalent to the IPW estimator in the status quo case, as mentioned earlier.

Given the quantile map is Hadamard differentiable, the asymptotic properties of the QCTE estimator can be obtained immediately from Lemma 3.1 by the functional delta method. We state the result in the following theorem.

Theorem 3.1. Suppose Assumptions 2.1–2.3 and 3.1–3.6 hold. We then have:

$$\sqrt{n}\Big(\widehat{\delta}^*(\cdot) - \delta^*(\cdot)\Big) \Rightarrow \Delta(\cdot),$$

where $\Delta(\tau)$ is a Gaussian process with mean zero and covariance function $\Psi(\tau_1, \tau_2) = \mathrm{E}[\psi(\tau_1)\psi(\tau_2)]$, where the variance function $\psi(\tau) = \mathrm{E}\big[\varrho(\tau, Z)^2\big] + \mathrm{E}\big[\varphi(\tau, X^*)^2\big]$ with

$$\begin{aligned} \varrho(\tau, Z) &= -\left[\frac{\varrho_1^F\Big(\mathrm{Q}_{Y_1^*}(\tau), Z\Big)}{f_{Y_1^*}\Big(\mathrm{Q}_{Y_1^*}(\tau)\Big)} - \frac{\varrho_0^F\Big(\mathrm{Q}_{Y_0^*}(\tau), Z\Big)}{f_{Y_0^*}\Big(\mathrm{Q}_{Y_0^*}(\tau)\Big)} \right], \\ \varphi(\tau, X^*) &= -\left[\frac{\varphi_1^F\Big(\mathrm{Q}_{Y_1^*}(\tau), X^*\Big)}{f_{Y_1^*}\Big(\mathrm{Q}_{Y_1^*}(\tau)\Big)} - \frac{\varphi_0^F\Big(\mathrm{Q}_{Y_0^*}(\tau), X^*\Big)}{f_{Y_0^*}\Big(\mathrm{Q}_{Y_0^*}(\tau)\Big)} \right], \end{aligned} \tag{3.7}$$

where ϱ_d^F and φ_d^F are given in (3.6), and the convergence takes place in $\ell^\infty([0,1])$.

Theorem 3.1 allows for pointwise inference on QCTE. For example, suppose we want to test whether the counterfactual treatment effect exists at the median, one can simply construct an ordinary t-statistic $\widehat{\delta}^*(\tau)/(\widehat{\psi}(\tau)/\sqrt{n})$ for $\tau = 0.5$ given a consistent estimate of the asymptotic variance function,

$$\widehat{\psi}(\tau) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varrho}(\tau, Z_i)^2 + \frac{1}{n^*}\sum_{j=1}^{n^*} \widehat{\varphi}(\tau, X_j^*)^2, \tag{3.8}$$

where $\widehat{\varrho}$ and $\widehat{\varphi}$ are later provided in Section 4.3.

4 Multiplier Bootstrap and Uniform Inference

Although Theorem 3.1 allows us to conduct pointwise inference on QCTE, there are many interesting hypotheses involving a continuum of quantile indices such as whether the counterfactual treatment has *any* effect along the outcome distribution. More generally, researchers may be interested in testing one-sided or two-sided functional hypotheses as follows:

$$H_0^{\text{1-sided}} : \delta^*(\tau) \leq 0 \text{ for } \tau \in [\tau_\ell, \tau_u], \qquad H_0^{\text{2-sided}} : \delta^*(\tau) = 0 \text{ for } \tau \in [\tau_\ell, \tau_u], \tag{4.1}$$

where $0 \leq \tau_\ell < \tau_u \leq 1$. In this paper, we propose a multiplier bootstrap procedure to simulate critical values for testing the above null hypotheses or constructing one-sided and two-sided uniform confidence bands. The multiplier bootstrap can be regarded as a more convenient alternative to the empirical bootstrap proposed by Rothe (2010) and Chernozhukov et al. (2013). However, in our setting the choice of the multiplier hinges on the relationship between X and X^* to preserve the same relationship in the simulated processes. Such a problem does not appear in previous applications of the multiplier bootstrap technique (see, e.g., Horowitz (2019) for a recent review).

In Section 4.1, we describe the multiplier bootstrap procedure and show its validity. Section 4.2 discusses uniform inference methods including hypothesis testing and uniform confidence bands. A step-by-step implementation of the uniform confidence bands is also carried out in Section 4.2. Finally, Section 4.3 provides a set of uniformly consistent nonparametric estimators that are useful in generating simulated processes.

4.1 Multiplier Bootstrap

To approximate the true limiting process, one must show the estimation errors associated with the simulated process are asymptotically negligible. This requires a uniformly consistent estimation of the functions involved in the variance function. Moreover, monotonicity of the estimators for $F_{Y_d^*}(y)$ and $F_{Y_d|X}(y|x)$ is also needed for manageability of the simulated processes. The following assumption formally states the availability of such estimators.

Assumption 4.1 (Uniform Consistency and Monotonicity)**.**

(i) $\widehat{F}_{Y_d^*}(y)$, $\widehat{F}_{Y_d|X}(y|x)$, $\widehat{p}(x)$, $\widehat{f}_X(x)$, $\widehat{f}_{X^*}(x)$, and $\widehat{f}_{Y_d^*}(y)$ are uniformly consistent in both y and x.

(ii) $\widehat{F}_{Y_d^*}(y)$ and $\widehat{F}_{Y_d|X}(y|x)$ are monotone in y for all $x \in \mathcal{X}$.

Suppose Assumption 4.1 holds, we can estimate $\varrho(\tau, Z)$ and $\varphi(\tau, X^*)$ by

$$\widehat{\varrho}(\tau, Z_i) = -\left[\frac{\widehat{\varrho}_1^F\left(\widehat{\mathrm{Q}}_{Y_1^*}(\tau), Z_i\right)}{\widehat{f}_{Y_1^*}\left(\widehat{\mathrm{Q}}_{Y_1^*}(\tau)\right)} - \frac{\widehat{\varrho}_0^F\left(\widehat{\mathrm{Q}}_{Y_0^*}(\tau), Z_i\right)}{\widehat{f}_{Y_0^*}\left(\widehat{\mathrm{Q}}_{Y_0^*}(\tau)\right)}\right],$$

$$\widehat{\varphi}(\tau, X_j^*) = -\left[\frac{\widehat{\varphi}_1^F\left(\widehat{\mathrm{Q}}_{Y_1^*}(\tau), X_j^*\right)}{\widehat{f}_{Y_1^*}\left(\widehat{\mathrm{Q}}_{Y_1^*}(\tau)\right)} - \frac{\widehat{\varphi}_0^F\left(\widehat{\mathrm{Q}}_{Y_0^*}(\tau), X_j^*\right)}{\widehat{f}_{Y_0^*}\left(\widehat{\mathrm{Q}}_{Y_0^*}(\tau)\right)}\right],$$

where $\widehat{\mathrm{Q}}_{Y_d^*}(\tau)$ is defined in (3.5) and for $d = 0, 1$,

$$\widehat{\varrho}_d^F(y, Z_i) = \frac{1\{D_i = d\}\Big[1\{Y_i \leq y\} - \widehat{F}_{Y_d|X}(y|X_i)\Big]}{\widehat{p}(X_i)^d[1 - \widehat{p}(X_i)]^{1-d}} \frac{\widehat{f}_{X^*}(X_i)}{\widehat{f}_X(X_i)},$$

$$\widehat{\varphi}_d^F(y, X_j^*) = \sqrt{\widehat{\lambda}}\Big[\widehat{F}_{Y_d|X}(y|X_j^*) - \widehat{F}_{Y_d^*}(y)\Big],$$

with $\widehat{\lambda} = n/n^*$. Let $\{U_1, \ldots, U_n\}$ and $\{U_1^*, \ldots U_{n^*}^*\}$ be i.i.d. pseudo-random variables with mean zero and variance one that are independent of each other and the whole sample process $\{(Z_i, X_j^*) : 1 \leq i \leq n, 1 \leq j \leq n^*, n, n^* \geq 1\}$. The simulated process for $\Delta(\tau)$ is then given by

$$\Delta^u(\tau) = \begin{cases} \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} U_i[\widehat{\varrho}(\tau, Z_i) + \widehat{\varphi}(\tau, X_i^*)] & \text{if } X^* = \pi(X), \\ \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} U_i \widehat{\varrho}(\tau, Z_i) + \dfrac{1}{\sqrt{n^*}} \displaystyle\sum_{j=1}^{n^*} U_j^* \widehat{\varphi}(\tau, X_j^*) & \text{if } X^* \perp\!\!\!\perp X. \end{cases} \tag{4.2}$$

As can be seen from (4.2), the choice of multiplier depends on the relationship between X and X^*. If X^* is a deterministic transformation of X, one can simply utilize a single multiplier U_i associated with $\widehat{\varrho}(\tau, Z_i) + \widehat{\varphi}(\tau, X_i^*)$ to simulate the limiting process. If X^* and X are independent, on the other hand, one needs to introduce two mutually independent multipliers U_i and U_j^* to ensure the independence between the processes generated by $U_i\widehat{\varrho}(\tau, Z_i)$ and $U_j^*\widehat{\varphi}(\tau, X_j^*)$. The fact that we need to impose an additional multiplier is because $\widehat{\varrho}(\tau, Z_i)$ and $\widehat{\varphi}(\tau, X_j^*)$ are only asymptotically independent, i.e., they may not be independent in finite samples (but their probability limits are) even if X^* is independent of X.

The next theorem asserts the validity of the multiplier bootstrap procedure based on conditional multiplier central limit theorem. Note that Assumption 4.1 plays an important role in the proof of Theorem 4.1.

Theorem 4.1. Suppose Assumptions 2.1–2.3, 3.1–3.6 and 4.1 hold. We then have:

$$\Delta^u(\cdot) \overset{p}{\Rightarrow} \Delta(\cdot),$$

conditional on the sample paths $\{Z_i : i = 1, 2, \ldots\}$ and $\{X_j^* : j = 1, 2, \ldots\}$ with probability approaching one.

4.2 Uniform Inference

Based on Theorem 4.1, we briefly discuss uniform inference methods on QCTE via multiplier bootstrap. To test functional hypotheses stated in (4.1), it is common to apply the Kolmogorov-Smirnov test, where the one-sided and two-sided standardized test statistics are given, respectively, by

$$\widehat{KS}_n^{\text{1-sided}} = \sup_{\tau\in[\tau_\ell,\tau_u]} \frac{\sqrt{n}\widehat{\delta}^*(\tau)}{\widehat{\sigma}(\tau)}, \qquad \widehat{KS}_n^{\text{2-sided}} = \sup_{\tau\in[\tau_\ell,\tau_u]} \frac{\left|\sqrt{n}\widehat{\delta}^*(\tau)\right|}{\widehat{\sigma}(\tau)},$$

with $\widehat{\sigma}(\tau) \equiv \widehat{\psi}^{1/2}(\tau)$ and $\widehat{\psi}(\tau)$ defined in (3.8). For any nominal significance level α, the null hypotheses are rejected if the test statistics exceed the corresponding critical values $\widehat{C}_\alpha^{\text{1-sided}}$ and $\widehat{C}_\alpha^{\text{2-sided}}$,

$$\widehat{C}_\alpha^{\text{1-sided}} = \inf_{a\in\mathbb{R}}\left\{\mathrm{P}\left(\sup_{\tau\in[\tau_\ell,\tau_u]} \frac{\Delta^u(\tau)}{\widehat{\sigma}(\tau)} \le a\right) \ge 1-\alpha\right\},$$
$$\widehat{C}_\alpha^{\text{2-sided}} = \inf_{a\in\mathbb{R}}\left\{\mathrm{P}\left(\sup_{\tau\in[\tau_\ell,\tau_u]} \frac{|\Delta^u(\tau)|}{\widehat{\sigma}(\tau)} \le a\right) \ge 1-\alpha\right\},$$

where $\Delta^u(\tau)$ is the simulated process in (4.2). Apart from hypothesis testing, one can also construct uniform confidence bands based on the simulated critical values $\widehat{C}_\alpha^{\text{1-sided}}$ and $\widehat{C}_\alpha^{\text{2-sided}}$. That is, the one-sided $(1-\alpha)$ uniform confidence bands for QCTE over $[\tau_\ell,\tau_u]$ are

$$\left(\widehat{\delta}^*(\tau) - \widehat{C}_\alpha^{\text{1-sided}}\frac{\widehat{\sigma}(\tau)}{\sqrt{n}}, \quad +\infty\right) \quad \text{and} \quad \left(-\infty, \quad \widehat{\delta}^*(\tau) + \widehat{C}_\alpha^{\text{1-sided}}\frac{\widehat{\sigma}(\tau)}{\sqrt{n}}\right), \tag{4.3}$$

whereas the two-sided $(1-\alpha)$ uniform confidence band over $[\tau_\ell,\tau_u]$ is given by

$$\left(\widehat{\delta}^*(\tau) - \widehat{C}_\alpha^{\text{2-sided}}\frac{\widehat{\sigma}(\tau)}{\sqrt{n}}, \quad \widehat{\delta}^*(\tau) + \widehat{C}_\alpha^{\text{2-sided}}\frac{\widehat{\sigma}(\tau)}{\sqrt{n}}\right). \tag{4.4}$$

In practice, the simulated one-sided and two-sided critical values can be easily obtained by noting that $\widehat{C}_\alpha^{\text{1-sided}}$ and $\widehat{C}_\alpha^{\text{2-sided}}$ are actually the $(1-\alpha)$th quantile of $\sup_{\tau\in[\tau_\ell,\tau_u]}\Delta^u(\tau)/\widehat{\sigma}(\tau)$ and the $(1-\alpha)$th quantile of $\sup_{\tau\in[\tau_\ell,\tau_u]}|\Delta^u(\tau)|/\widehat{\sigma}(\tau)$, respectively. Therefore, one can simulate the limiting processes many times via multiplier bootstrap to approximate the quantiles as well as critical values. To be more specific, a step-by-step implementation of uniform confidence bands is provided as follows.

1. Given a set of prespecified quantile indices $\tau \in \{\tau_\ell,\dots,\tau_u\}$, obtain QCTE estimates $\widehat{\delta}^*(\tau)$ according to (3.4) and $\widehat{\sigma}(\tau) = \widehat{\psi}^{1/2}(\tau)$ where $\widehat{\psi}(\tau)$ is defined in (3.8).

2. For each bootstrap replication $b = 1,\dots,B$, say $B = 1000$, draw i.i.d. pseudo random variables $\{U_1,\dots,U_n\}$ and $\{U_1^*,\dots U_{n^*}^*\}$ with mean zero and unit variance, and then calculate the simulated process $\Delta_b^u(\tau)$ based on (4.2).

3. For the one-sided case, store the maximum value of $\Delta_b^u(\tau)/\widehat{\sigma}(\tau)$ over all grid points of τ in each bootstrap replication. That is, let $M_b \equiv \max_{\tau\in[\tau_\ell,\tau_u]}\Delta_b^u(\tau)/\widehat{\sigma}(\tau)$ for $b = 1,\dots,B$.

4. Rank $\{M_b : b = 1,\dots,B\}$ in an ascending order such that $M_{(1)} \le \dots \le M_{(B)}$. Next, define $M_{(\lfloor(1-\alpha)B\rfloor)}$ as critical value $\widehat{C}_\alpha^{\text{1-sided}}$, where $\lfloor c\rfloor$ is the floor function returning the largest integer not greater than c. The one-sided $(1-\alpha)$ uniform confidence bands for $\{\delta^*(\tau) : \tau \in [\tau_\ell,\tau_u]\}$ are given by (4.3).

5. For the two-sided case, simply replace $\Delta_b^u(\tau)/\widehat{\sigma}(\tau)$ in Step 3 with $|\Delta_b^u(\tau)|/\widehat{\sigma}(\tau)$ and repeat Step 4 to obtain critical value $\widehat{C}_\alpha^{\text{2-sided}}$. The two-sided $(1-\alpha)$ uniform confidence band for $\{\delta^*(\tau) : \tau \in [\tau_\ell, \tau_u]\}$ is then given by (4.4).

4.3 Uniformly Consistent and Monotone Estimators for Assumption 4.1

Here we provide kernel-based nonparametric estimators that satisfy Assumption 4.1 and can thus be used to construct simulated processes according to (4.2). Regarding the monotonicity requirement in Assumption 4.1(ii), we use $\widehat{F}_{Y_d^*}(y)$ in (3.3) and let

$$\widehat{F}_{Y_d|X}(y|x) = \phi_1(\widetilde{F}_{Y_d|X})(y|x), \tag{4.5}$$

where $\widetilde{F}_{Y_d|X}(y|x)$ is defined in (3.1). It is easy to see that $\widehat{F}_{Y_d}(y)$ and $\widehat{F}_{Y_d|X}(y|x)$ satisfy Assumption 4.1(ii). To meet Assumption 4.1(i), we will show $\sup_{y\in\mathcal{Y},x\in\mathcal{X}} |\widehat{F}_{Y_d|X}(y|x) - F_{Y_d|X}(y|x)| = o_p(1)$ in Lemma 4.1 below.

Next, the kernel-based estimators for $p(x)$, $f_X(x)$ and $f_{X^*}(x)$ are given by

$$\widetilde{p}(x) = \frac{\sum_{i=1}^n D_i K_{x,h}(X_i - x)}{\sum_{i=1}^n K_{x,h}(X_i - x)}, \quad \widetilde{f}_X(x) = \frac{1}{n}\sum_{i=1}^n K_{x,h}(X_i - x), \quad \widetilde{f}_{X^*}(x) = \frac{1}{n^*}\sum_{j=1}^{n^*} K_{x,h}(X_j^* - x). \tag{4.6}$$

Uniform consistency of the above estimators are already established in, e.g., Härdle et al. (1998) and Jones (1993). However, a minor disadvantage of applying a boundary kernel (even if second-order) is that the estimators in (4.6) are not always positive. We then tackle this issue by applying the trimming method in Donald et al. (2014) for $\widetilde{p}(x)$ and the method in Hsu et al. (2019) for $\widetilde{f}_X(x)$ and $\widetilde{f}_{X^*}(x)$. For the former, let

$$\widehat{p}(x) = a_n\, 1\{\widetilde{p}(x) \le a_n\} + \widetilde{p}(x)\, 1\{a_n < \widetilde{p}(x) < 1 - a_n\} + (1 - a_n)\, 1\{\widetilde{p}(x) \ge 1 - a_n\}, \tag{4.7}$$

where $\{a_n \in (0, 1/2) : n \ge 1\}$ is a positive sequence converging to 0 and can be determined by criteria outlined in Busso et al. (2009) and Lechner and Strittmatter (2019). It is straightforward to see that $\widehat{p}(x)$ is a proper propensity score estimator in that the estimate is bounded away from 0 and 1. For the latter density function estimators, we follow the trimming method in Hsu et al. (2019) and let

$$\widehat{f}_X(x) = \max\{\widetilde{f}_X(x), b_n\}, \qquad \widehat{f}_{X^*}(x) = \max\{\widetilde{f}_{X^*}(x), b_n\}, \tag{4.8}$$

where $\{b_n : n \ge 1\}$ is a decreasing sequence of positive numbers converging to 0. Despite not necessarily integrating to one for all n, the estimators in (4.8) are still uniformly consistent for $f_{X^*}(x)$ and $f_X(x)$ and hence meet the requirements in Assumption 4.1(i).

Finally, the estimator for $f_{Y_d^*}(y)$ can be constructed in the same manner as $\widetilde{F}_{Y_d^*}(y)$ in (3.2). That is, let

$$\widetilde{f}_{Y_d^*}(y) = \frac{1}{n^*}\sum_{j=1}^{n^*} \widetilde{f}_{Y_d|X}(y|X_j^*), \qquad \widetilde{f}_{Y_d|X}(y|x) = \frac{\sum_{i=1}^n W_{y,\eta}(Y_i - y)\, 1\{D_i = d\} K_{x,h}(X_i - x)}{\sum_{i=1}^n 1\{D_i = d\} K_{x,h}(X_i - x)},$$

where $W_{y,\eta}(\cdot) = \eta^{-1} W_y(\cdot/\eta)$ with W_y a boundary kernel (possibly different from K_x) and $\eta = \eta_n$ the bandwidth in the y direction. As we argue above, $\widetilde{f}_{Y_d^*}(y)$ could be negative in finite samples. We then employ the trimming method in (4.8) again to define

$$\widehat{f}_{Y_d^*}(y) = \max\{\widetilde{f}_{Y_d^*}(y), b_n\}. \tag{4.9}$$

The next lemma summarizes the discussion and formally states that all the proposed estimators meet the requirements of the multiplier bootstrap method.

Lemma 4.1. Suppose Assumptions 2.1, 2.3 and 3.1–3.6 hold. Moreover, suppose Assumption 3.5 holds with K_x replaced by W_y, and a_n, b_n, and $\eta = \eta_n \to 0$ as $n \to \infty$. The estimators in (3.3), (4.5) and (4.7)–(4.9) then satisfy Assumption 4.1.

5 Simulation Study

In this section, we examine the finite sample properties of the QCTE and QCTT estimators as well as the multiplier bootstrap procedure via Monte Carlo simulations.[9] The data generating process is as follows. Let $X = (X_1, X_2, X_3)$ be a three-dimensional random vector with each element following a standard exponential distribution truncated at 2. Let $Y = DY_1 + (1-D)Y_0$ with

$$D = 1\{(X_1 + X_2)/2 > \varepsilon_D\}, \qquad Y_1 = 4 + X_2 - 2X_3 + \varepsilon_1, \qquad Y_0 = 3 - \sqrt{X_2 + X_3} \cdot \varepsilon_0,$$

where ε_D, ε_1, and ε_0 are independently drawn from a standard exponential distribution truncated at 1. Note that in our design only X_2 appears in both the outcome and selection equations, indicating that D is unconfounded of (Y_1, Y_0) given X.

We consider two counterfactual scenarios. The first one corresponds to the dependent policy implementation where $X^* = (X_1^*, X_2^*, X_3^*) = 0.75X$. The second one corresponds to the independent policy implementation where each element of X^* follows an i.i.d. standard exponential distribution truncated at 1.5. In each scenario, the counterfactual treatment assignment and potential outcomes are given by

$$D^* = 1\{(X_1^* + X_2^*)/2 > \varepsilon_D\}, \qquad Y_1^* = 4 + X_2^* - 2X_3^* + \varepsilon_1, \qquad Y_0^* = 3 - \sqrt{X_2^* + X_3^*} \cdot \varepsilon_0.$$

We vary the status quo and counterfactual sample sizes n and n^* from 100, 200 to 400 with $n \geq n^*$. The numbers of Monte Carlo replications and bootstrap samples are both set to 1000. We use all of the different values of Y_i as the grid points to estimate the distribution functions. For the quantile functions and QCTE and QCTT, 100 equidistant grid points in $[0.1, 0.9]$ are considered.

To construct a higher-order boundary kernel $K_x(u)$ that satisfies Assumption 3.5, we closely follow Rothe (2010), who utilizes the equivalent kernels for the local polynomial estimation (see Fan and Gijbels (1996) for more details). Specifically, let $K_x(u) = \prod_{s=1}^k e_1^T S_{x_s}^{-1}(1, u_s, \ldots, u_s^p)^T K(u_s)$ where $e_1 = (1, 0, \ldots, 0)^T$ is the unit vector, $S_x = (\mu_{j+\ell,x})_{0\leq j,\ell\leq p}$ is a matrix of boundary kernel constants $\mu_{j,x} = \int_{\mathcal{D}_x} u^j K(u)\,\mathrm{d}u$, $p = r - k$ is the polynomial order, and $K(u)$ is a standard univariate kernel. We let $K(u)$ be the Epanechnikov kernel and choose $p = 1$ so that $K_x(u)$ is a fourth-order boundary kernel. The bandwidth for $d = 0, 1$ is given by $h_d = 3.2 s_X n_d^{-1/7}$, where 3.2 is the rule-of-thumb bandwidth constant for fourth-order Epanechnikov kernel with three-dimensional vector X, s_X is the sample standard deviation of X, and $n_d = \sum_{i=1}^n 1\{D_i = d\}$ is the effective sample size. Note that the power of n_d must fall into the interval $(-1/2k, -1/2r) = (-1/6, -1/8)$ to satisfy Assumption 3.6. For the variance function estimation, we use second-order boundary kernels for $K_x(\cdot)$

[9] The theoretical analysis of the QCTT estimator is deferred to the supplementary appendix due to space constraints.

Table 1. Simulation results (dependent policy implementation).

n	n^*	QCTE			QCTT		
		IBias	RIMSE	Cov. Rate	IBias	RIMSE	Cov. Rate
100	100	0.125	0.178	0.827	0.131	0.189	0.851
200	100	0.096	0.136	0.814	0.105	0.149	0.831
200	200	0.090	0.128	0.869	0.100	0.141	0.881
400	100	0.078	0.112	0.821	0.086	0.121	0.855
400	200	0.070	0.101	0.863	0.077	0.110	0.871
400	400	0.064	0.092	0.939	0.071	0.100	0.925

Reported are Monte Carlo estimates among 1000 replications. In each replication, QCTE and QCTT are estimated over 100 equidistant grid points in $[0.1, 0.9]$ using fourth-order boundary Epanechnikov kernel with bandwidth $h_d = 3.2 s_X n_d^{-1/7}$, where s_X is the sample standard deviation and n_d is the effective sample size. The nominal coverage rate of the two-sided uniform confidence bands is 90%. We use 1000 bootstrap samples with standard normal multipliers to simulate critical values. The variance functions are estimated using second-order boundary Epanechnikov kernels with rule-of-thumb bandwidths in both x and y directions. See text for more details.

Table 2. Simulation results (independent policy implementation).

n	n^*	QCTE			QCTT		
		IBias	RIMSE	Cov. Rate	IBias	RIMSE	Cov. Rate
100	100	0.130	0.182	0.852	0.131	0.184	0.867
200	100	0.096	0.135	0.824	0.094	0.134	0.870
200	200	0.091	0.128	0.897	0.092	0.131	0.898
400	100	0.072	0.101	0.820	0.071	0.101	0.891
400	200	0.067	0.094	0.858	0.067	0.095	0.882
400	400	0.063	0.089	0.899	0.064	0.090	0.902

Reported are Monte Carlo estimates among 1000 replications. In each replication, QCTE and QCTT are estimated over 100 equidistant grid points in $[0.1, 0.9]$ using fourth-order boundary Epanechnikov kernel with bandwidth $h_d = 3.2 s_X n_d^{-1/7}$, where s_X is the sample standard deviation and n_d is the effective sample size. The nominal coverage rate of the two-sided uniform confidence bands is 90%. We use 1000 bootstrap samples with standard normal multipliers to simulate critical values. The variance functions are estimated using second-order boundary Epanechnikov kernels with rule-of-thumb bandwidths in both x and y directions. See text for more details.

and $W_y(\cdot)$ with corresponding rule-of-thumb bandwidths $h_d = 2.12 s_X n_d^{-1/7}$ and $\eta = 2.34 s_Y n^{-1/5}$. We also use standard normal multipliers in both dependent and independent cases to simulate critical values. The nominal coverage rate is 90%.

Tables 1 and 2 present the simulation results for the dependent and independent policy implementations, respectively. In each case, we report the Monte Carlo estimates of the integrated bias (IBias), the root integrated mean squared error (RIMSE), and the coverage rate of the two-sided uniform confidence band. As can be seen from the tables, the finite sample performance of the proposed estimators is satisfactory in the sense that the estimates of IBias vanish as the sample sizes grow. In addition, the RIMSE estimates decrease by almost half as n and n^* increase from 100 to 400, suggesting that the rate of convergence is indeed at the $\sqrt{n}$ rate. The empirical coverage rates of the uniform confidence bands are also close to 90% in both cases. This result validates the multiplier bootstrap method even when the sample size is relatively small.

6 Empirical Study

In this section, we illustrate the application of our methods through two counterfactual exercises concerned with the heterogeneous earnings effects of Job Corps, the largest and most comprehensive training program in the U.S. Established in 1964, Job Corps serves approximately 60,000 disadvantaged youths aged 16–24 each year by providing academic, vocational, and social training, as well as residential living, health care, counseling, and job placement assistance. It is well documented that Job Corps is more effective for males than for females in improving participants' post-program earnings. For example, Schochet et al. (2008) show that males benefit more from Job Corps than females on average in a large-scale experimental evaluation. Eren and Ozbeklik (2014) also report more positive quantile earnings effects for males than for females. To understand the potential mechanism behind these findings, Strittmatter (2019) decomposes these program effects into (i) a structural gender earnings inequality component arising from different labor market opportunities by gender and (ii) differences in Job Corps trainability due to different characteristics by gender. However, his analysis focuses on the intention to treat effect, i.e., the effect of eligibility to participate in the Job Corps program. This may, of course, differ from the effect of actual program participation. Moreover, *ibid.* uses pointwise, rather than uniform, inference procedures in his analysis, so some of his conclusions may be affected by multiple testing issues.

In our first counterfactual exercise, we use our methods to estimate the counterfactual effect of Job Corps that would have prevailed for females had they faced the male earnings structure, and vice versa. The counterfactual and status quo estimates allow us to shed light on potential mechanisms underlying the program effect heterogeneity by gender. For example, to isolate the contribution of structural gender earnings inequality, one may compare the QCTE for females with the male earnings structure with the actual QTE for females. This exercise corresponds to the independent policy implementation scheme since individuals are randomly sampled from an experimental evaluation of Job Corps (we will briefly describe the experiment below).

In the second exercise, we hypothetically increase the education level of a more disadvantaged subgroup, and then reestimate the program effect with other things being equal. This exercise is designed to examine the skill hypothesis proposed by Frumento et al. (2012) and Eren and Ozbeklik (2014), who speculate that the Job Corps may only work well for participants with better labor market opportunities (i.e., higher education), but not for those at the bottom of the skill distribution. This exercise corresponds to the dependent policy implementation scheme since the new level of education is a deterministic transformation of the original one.

It is worth noting that both exercises meet Assumption 2.3(i) by virtue of the questions being asked. For example, what we are interested in in the first exercise is the program effect for males/females that would prevail under the earnings structure of the opposite sex, i.e., *assuming* that condition 2.3(i) holds. The causal interpretation of the resulting QCTE then comes from unconfoundedness. The same consideration applies to the second exercise. A clear limitation of the counterfactual analysis is that it fails to incorporate spillover or general equilibrium effects, but this problem is beyond the scope of the current paper.

6.1 Data and Unconfoundedness Test

Similarly to Eren and Ozbeklik (2014), our data are extracted from the National Job Corps Study (NJCS), an experimental evaluation of Job Corps undertaken between 1994 and 1996.[10] A detailed description of the experiment design can be found in Schochet et al. (2008). We focus on the sample of youths who completed

[10] The data are publicly available at http://qed.econ.queensu.ca/jae/2014-v29.4/eren-ozbeklik/.

the 48-month interview. In this sample, a randomly chosen group of 6,828 eligible applicants was offered training, while the other 4,485 applicants were excluded from receiving Job Corps services for three years. However, eligible members were allowed to refuse the offer and only 72% actually chose to receive Job Corps services. Thus, actual program participation has a self-selection component and is likely to correlate with potential outcomes without further controls.

ú8Table 3 reports the means and standard deviations of the observed variables by gender and by participation status. The outcome variable (Y) is the earnings per week in year four after randomization. The treatment variable (D) is the actual program participation. The baseline characteristics (X) include age, a race dummy (1 if nonwhite), and a dummy for high school education (including GED holders). Note that we drop all observations with any missing values and exclude those with a high school diploma or GED at age 16. As shown in the far-right column of Table 3, there are significant differences in characteristics between male and female applicants, where females tend to be older, less white, and more educated than males. In addition, we observe significant differences in characteristics by participation status for females, but not for males. These results differ from the comparisons in Eren and Ozbeklik (2014) and Strittmatter (2019) since they report the differences in characteristics by random eligibility rather than by actual participation.

To address question whether the observed covariates are sufficient to control for self-selection into the unconfoundedness test proposed by Donald et al. (2014). The instrument is the randomly assigned eligibility indicator in the NJCS.[11] In Table 4, we present the p-values for the null hypothesis of unconfoundedness for male and female subpopulations under various specifications of the propensity scores.

As shown by the results, the selection-on-observables assumption is rejected at the 10% level only when the propensity scores are estimated by a constant (i.e., no covariates are used for conditioning). This case of the test corresponds to the null of completely random participation. Given that the baseline covariates (age, race, education) are adjusted for in various ways, the p-values are all above 0.1, albeit the difference is not large, especially for males. Nevertheless, we conclude that there is no strong evidence against random participation for either gender conditional on age, race and education.

Regarding further implementation details, we treat the age variable as continuous and use the second-order boundary Epanchnikov kernel, which can be constructed as in the simulation study. The corresponding bandwidth is $h_d = 2.34 s_X n_d^{-1/3}$ for $d = 0, 1$. We focus on quantiles that are beyond the proportion of the mass point at zero, i.e., individuals with zero earnings. Specifically, we consider 50 equidistant quantile indices in the interval $[0.18, 0.99]$ for males and $[0.21, 0.99]$ for females, since 1063/6166 (17.24%) males and 943/4579 (20.59%) females have zero earnings in year 4 after random assignment, i.e., the second year of the post-program period. The rest of the setup is similar to that used in the simulation study. The robustness checks provided in the supplementary material suggest that the results are not sensitive to these settings.

6.2 Empirical Results

In Figure 2 we first present the actual QTEs of Job Corps participation by gender. These results complement the local quantile treatment effects reported in Eren and Ozbeklik (2014) and the ITT effects reported in Strittmatter (2019). The solid lines in Figures 2(a) and 2(b) are the QTE estimates for males and females, respectively, with light and dark shaded areas representing 90% uniform and pointwise confidence bands.

[11] Donald et al. (2014) show that, given the binary instrument satisfying one-sided non-compliance, the local average treatment effect for the treated and the average treatment effect for the treated would be identical under the null hypothesis of unconfoundedness. They then propose a Durbin-Wu-Hausman-type test based on this observation. The one-sided noncompliance condition in our case requires that non-eligible applicants must not enroll in Job Corps, which is satisfied by the experimental design of NJCS.

Table 3. Summary statistics.

	Male			Female			Gender
	Part. (std. dev.)	No (std. dev.)	Diff. [t-stat.]	Part. (std. dev.)	No (std. dev.)	Diff. [t-stat.]	Diff. [t-stat.]
Earnings per week	241.73	223.91	17.81	170.32	165.16	5.16	64.23
in year 4	(214.07)	(203.41)	[3.31]	(168.64)	(169.63)	[1.03]	[17.62]
Age	18.23	18.33	−0.11	18.48	18.67	−0.20	−0.30
	(2.11)	(2.10)	[−1.95]	(2.17)	(2.17)	[−3.04]	[−7.12]
Race	0.70	0.70	0.00	0.81	0.78	0.02	−0.10
(nonwhite=1)	(0.46)	(0.46)	[−0.08]	(0.39)	(0.41)	[2.00]	[−11.48]
Education	0.19	0.20	−0.01	0.26	0.31	−0.04	−0.09
(HS or GED=1)	(0.40)	(0.40)	[−1.07]	(0.44)	(0.46)	[3.18]	[−10.54]
Sample size	2,703	3,463		2,065	2,514		

The table reports means and standard deviations (in parentheses) for NJCS individuals between 16 and 24 years old, excluding those with a high school diploma or GED at age 16. The columns showing differences in means by participation status and by gender report t-statistics (in brackets) for the null hypotheses of equality in means, respectively.

Table 4. Unconfoundedness test results.

		Specification			
	Sample size	Constant	Linear	Interaction	Quadratic
		$X = \{\text{age, race, education}\}$			
Male	6,166	0.051	0.110	0.118	0.126
Female	4,579	0.041	0.141	0.163	0.163

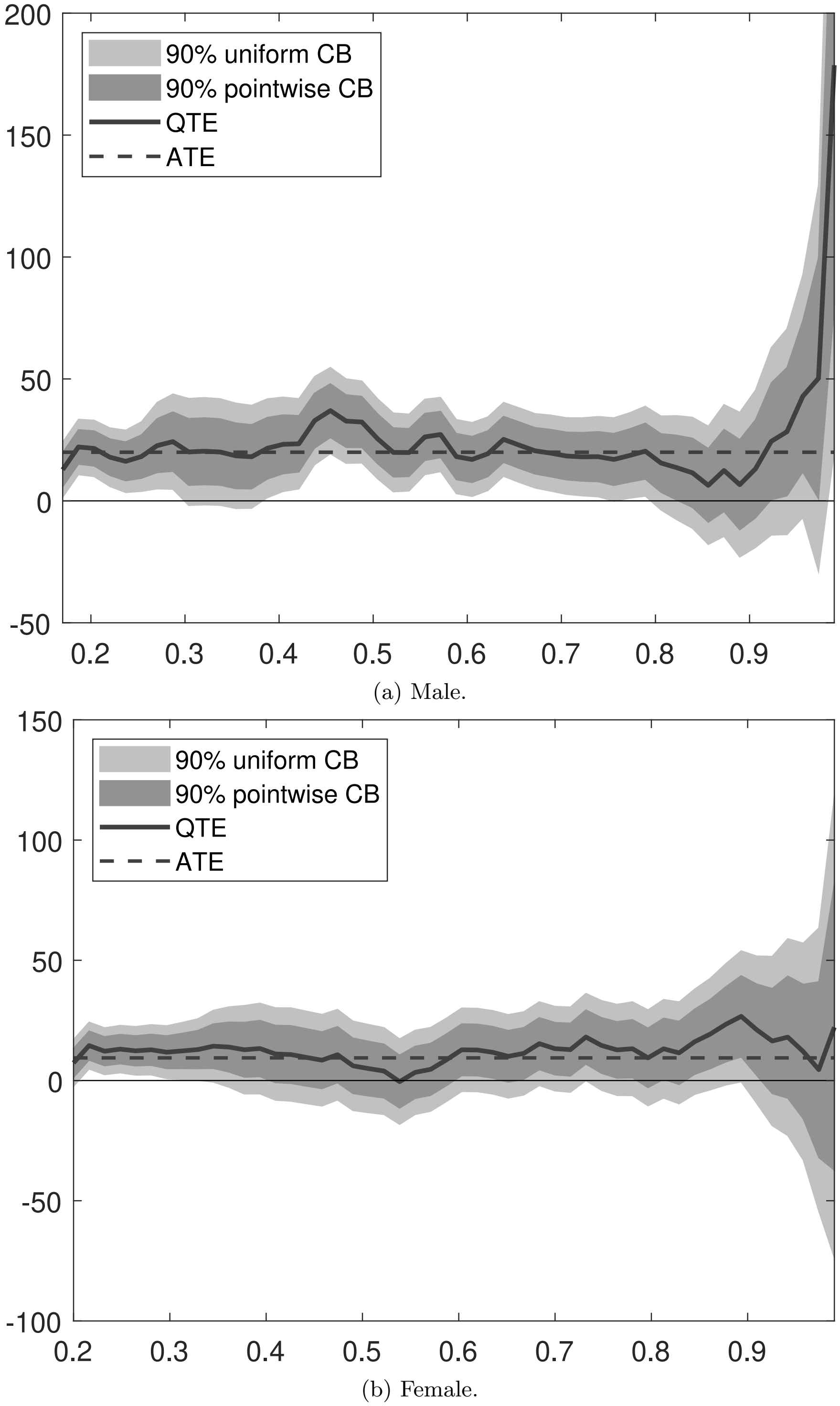
The table reports p-values for the null hypothesis of unconfoundedness using the test proposed by Donald et al. (2014) and random eligibility of NJCS as the instrument. Different specifications indicate the inclusion of constant, linear, interaction, and quadratic terms of covariates as power series in estimating the instrument propensity score. See Donald et al. (2014) for more details.

The horizontal dashed lines are the corresponding ATEs. The unit on the vertical axis is dollars per week.

Focusing on the male case in Figure 2(a), we see that the QTE between the 2nd and 3rd earning deciles is statistically significant and is around $20 (per week). Proceeding to higher quantiles, the program effect loses statistical significance between the 3rd and 4th deciles according to the uniform confidence band, but is again significantly positive between the 40th and 75th percentiles. Finally, the point estimate escalates beyond the 9th decile, though it again loses significance. As for the female case, Figure 2(b) shows a more uniform QTE pattern along the entire earnings distribution. Specifically, the QTE estimate fluctuates around the ATE estimate, $10, and the effect is only significant at the bottom of the earnings distribution.

Figure 3 presents the QCTEs of Job Corps on the earnings distribution that would prevail for a given gender if the wage structure were switched to that of the opposite sex. To compare the counterfactual effects with the status quo more easily, we also depict the actual QTEs for males and females as a dashed and dotted line, respectively. In Figure 3(a), it is clearly seen that the QCTE for females with the male earnings structure resembles the actual QTE for males across the earnings distribution. Accordingly, the QCTE is substantially higher than the actual QTE for females. Similar patterns can be observed in Figure 3(b), where we depict the QCTE for males with the female earnings structure. Again, this counterfactual is closer to the actual female QTE than the male QTE. Taken together, our finding lends further support to the argument in Strittmatter (2019), who claims that the effect heterogeneity by gender arises mainly from structural gender earnings inequality rather than from gender differences in Job Corps trainability.

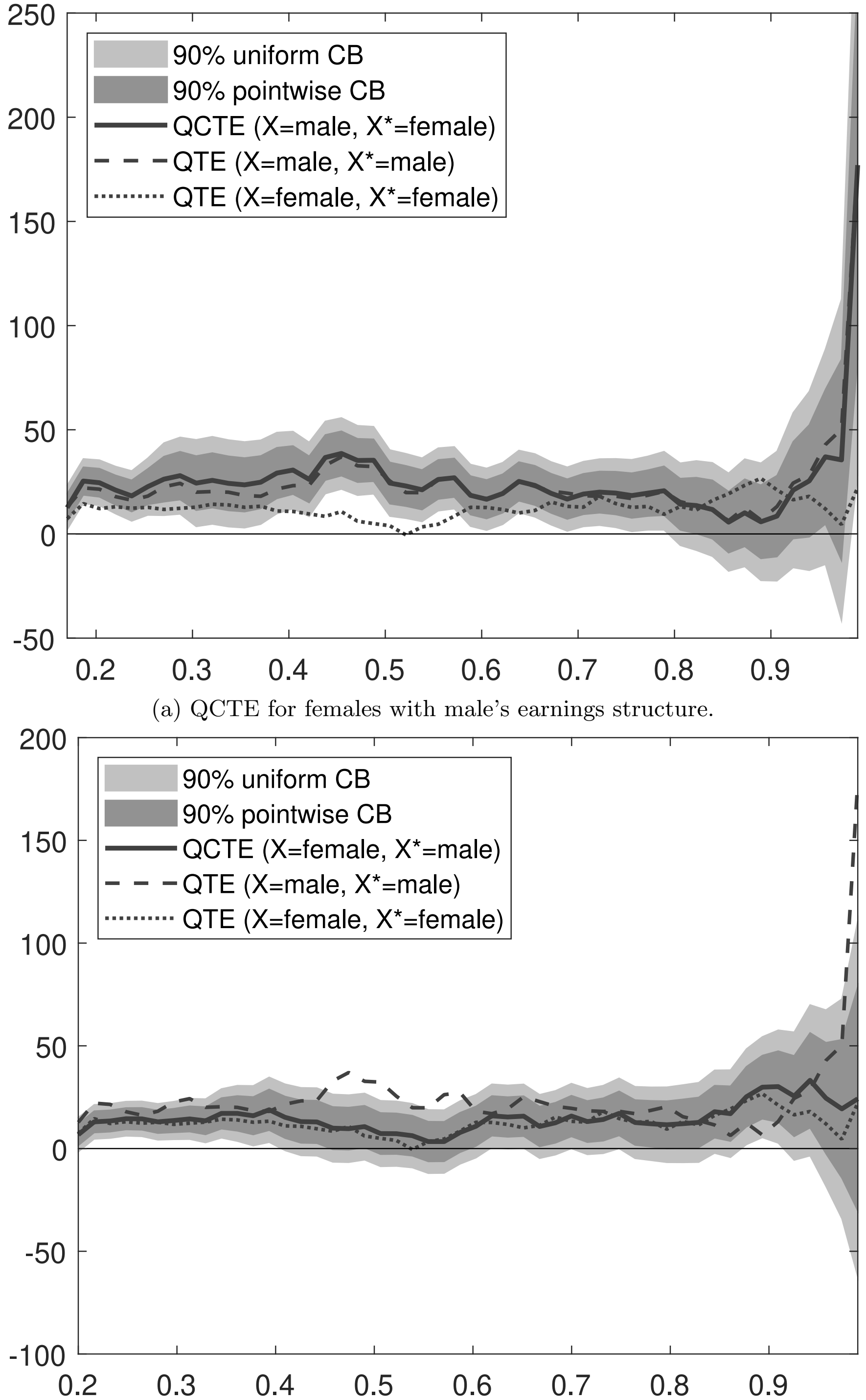
Lastly, Figure 4 presents the results from the second counterfactual exercise in which we artificially give high school education to all nonwhite males and females with ages between 17 and 19 years. Figure 4(a) shows the resulting QCTE for males and Figure 4(b) for females. The original QTEs are plotted as dashed lines in both cases for reference. Surprisingly, the QCTEs with extra education are uniformly below the original QTEs for both genders. In particular, the QCTEs are not only insignificant but mostly negative beyond the 7th decile for both males and females. This finding is contrary to the speculated skill hypothesis which predicts that a higher skill or higher education level should be accompanied by a significant increase in the Job Corps program effectiveness. Instead, it is more in line with the arguments of Card et al. (2018), stating that individuals with weak labor market attachment are likely to experience large effects from work first programs such as Job Corps.



(a) Male.

(b) Female.

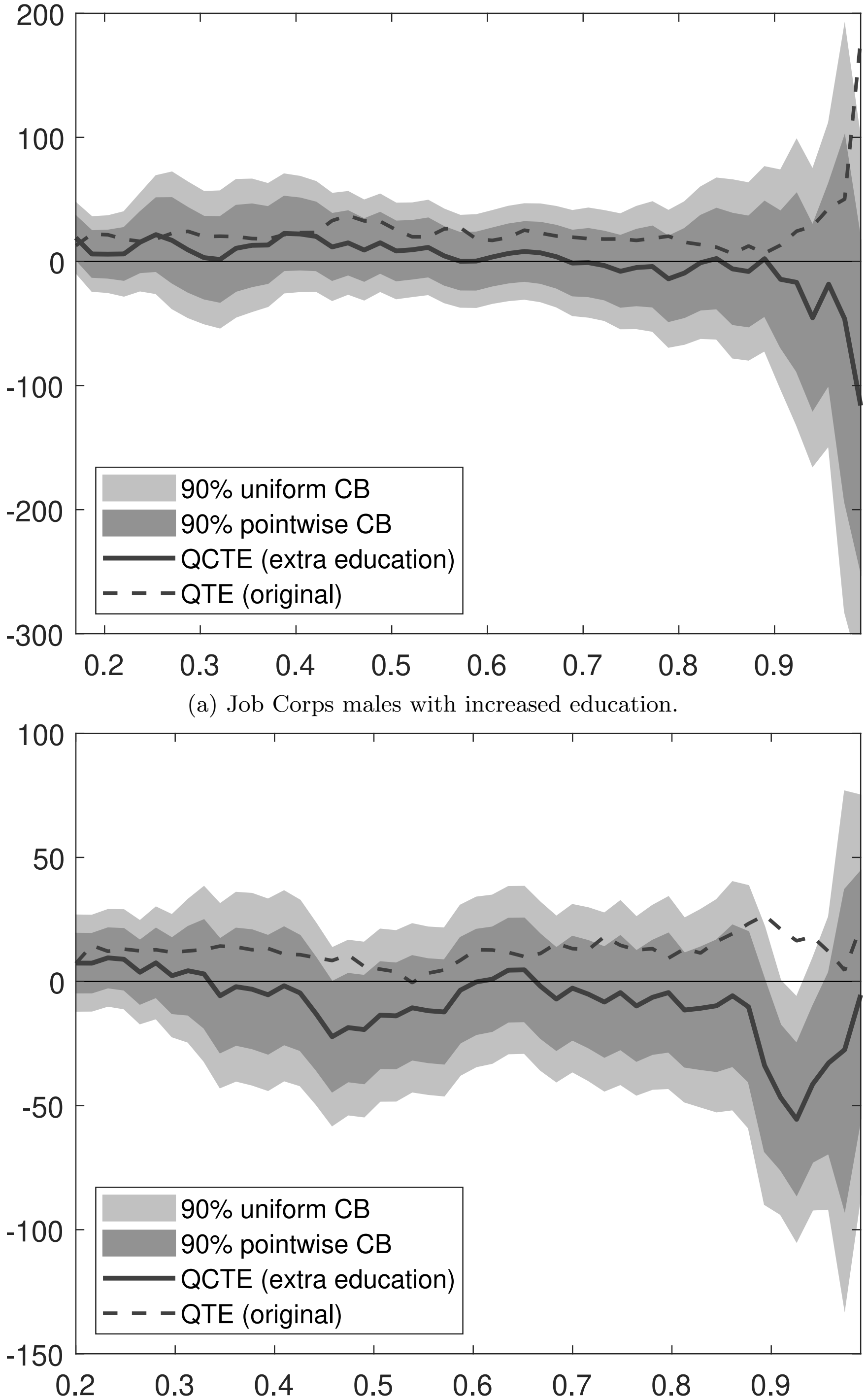
Figure 2. QTEs of Job Corps on earnings by gender, with 90% uniform and pointwise confidence bands. The horizontal dashed lines are the corresponding ATEs.



(a) QCTE for females with male's earnings structure.

(b) QCTE for males with female's earnings structure.

Figure 3. QCTEs of Job Corps on earnings that would have prevailed for females and males if the wage structure were switched to that of the opposite sex, with 90% uniform and pointwise confidence bands. The dashed and dotted lines are QTEs of Job Corps on earnings for males and for females, respectively.



(a) Job Corps males with increased education.

(b) Job Corps females with increased education.

Figure 4. QCTEs of Job Corps on earnings with increased education by gender, with 90% uniform and pointwise confidence bands. The dashed lines are the original QTEs by gender.

7 ACTE and Overidentification Test

7.1 ACTE

Similar to QCTE, here we provide the theoretical analysis of the predicted mean program impact given counterfactual covariates, namely the ACTE. Since ACTE defined in (2.1) only depends on the means but not the entire distributions of Y_0^* and Y_1^*, the identification assumptions can be weakened as follows.

Assumption 7.1 (Mean Unconfoundedness)**.**

(i) D is conditionally mean independent of (Y_0, Y_1) given X, i.e., $\mathrm{E}(Y_d|D, X) = \mathrm{E}(Y_d|X)$ for $d = 0, 1$.

(ii) $p(x)$ is bounded away from 0 and 1 for all $x \in \mathcal{X}$.

Assumption 7.2 (Invariance of Conditional Means)**.**

(i) The conditional mean of Y_d^* given X^* is identical to that of Y_d given X for $d = 0, 1$. In other words, $\mathrm{E}(Y_d^*|X^* = x) = \mathrm{E}(Y_d|X = x)$ for all $x \in \mathcal{X}^*$.

(ii) $\mathcal{X}^*$ is a subset of $\mathcal{X}$.

Under Assumptions 7.1 and 7.2, ACTE is identified by $\delta^* = \mathrm{E}_{X^*}[\mathrm{E}(Y|D = 1, X) - \mathrm{E}(Y|D = 0, X)]$ and can be estimated either by the kernel-based estimator proposed in Heckman et al. (1998) or by the series-based estimator proposed in Hahn (1998). We adopt the former approach such that

$$\widehat{\delta}^* = \frac{1}{n^*}\sum_{j=1}^{n^*}\Big[\widehat{\mathrm{E}}(Y_1|X = X_j^*) - \widehat{\mathrm{E}}(Y_0|X = X_j^*)\Big],$$

where $\widehat{\mathrm{E}}(Y_d|X = x)$ is the Nadaraya-Watson estimator,

$$\widehat{\mathrm{E}}(Y_d|X = x) = \frac{\sum_{i=1}^{n} Y_i\, 1\{D_i = d\}K_{x,h}(X_i - x)}{\sum_{i=1}^{n} 1\{D_i = d\}K_{x,h}(X_i - x)}.$$

The asymptotic properties for $\widehat{\delta}^*$ can be derived under weaker regularity conditions stated below.

Assumption 7.3 (Moment of Y_d)**.** $\mathrm{E}(|Y_d|^r) < \infty$.

Assumption 7.4 (Conditional Probability and Moment)**.**

(i) $p(x)$ is r-times differentiable on the interior of $\mathcal{X}$ and the derivatives are uniformly continuous and bounded.

(ii) $\mathrm{E}(Y_d|X = x)$ is r-times differentiable with respect to x on the interior of $\mathcal{X}$ and the derivatives are uniformly continuous and bounded.

Corollary 7.1. *Suppose Assumptions 3.1, 3.2, 3.5, 3.6 and 7.1–7.4 hold. Then,*

$$\sqrt{n}\Big(\widehat{\delta}^* - \delta^*\Big) \xrightarrow{d} \mathcal{N}\big(0, \sigma_{\delta^*}^2\big),$$

where $\sigma_{\delta^*}^2 = \mathrm{E}\big[\varrho_{\delta^*}(Z)^2\big] + \mathrm{E}\big[\varphi_{\delta^*}(X^*)^2\big]$ *with*

$$\varrho_{\delta^*}(Z) = \left\{\frac{D[Y - \mathrm{E}(Y_1|X)]}{p(X)} - \frac{(1 - D)[Y - \mathrm{E}(Y_0|X)]}{1 - p(X)}\right\}\frac{f_{X^*}(X)}{f_X(X)},$$
$$\varphi_{\delta^*}(X^*) = \sqrt{\lambda}[\mathrm{E}(Y_1|X^*) - \mathrm{E}(Y_0|X^*) - \delta^*].$$

Several remarks on Corollary 7.1 are made below. First, the first-stage estimation error except for the leading term $\varrho_{\delta^*}(Z)$ is small enough in that it vanishes at a rate faster than $n^{-1/4}$ and can be neglected in the final estimation. Second, $\varphi_{\delta^*}(X^*)$ accounts for the uncertainty in replacing the expectation with a sample average, and it would also represent the influence function of $\widehat{\delta}^*$ if the conditional mean were known. Third, it can be verified that $\varrho_{\delta^*}(Z)$ and $\varphi_{\delta^*}(X^*)$ are uncorrelated, resulting in no covariance term in the asymptotic variance $\sigma^2_{\delta^*}$ even when X and X^* are dependent. Fourth, compared to the semiparametric efficiency bound of the ATE estimator given in Hahn (1998):

$$\mathrm{E}\left\{\frac{\mathrm{Var}(Y_1|X)}{p(X)}+\frac{\mathrm{Var}(Y_0|X)}{1-p(X)}+[\mathrm{E}(Y_1-Y_0|X)-\mathrm{E}(Y_1-Y_0)]^2\right\},$$

it is true that $\sigma^2_{\delta^*}$ attains this bound if $X^*=X$ and $\lambda=1$. Put differently, when applying to the ATE case, our kernel-based estimator is as efficient as the series estimator in Hahn (1998) and the IPW estimator in Hirano et al. (2003).

Since the asymptotic variance $\sigma^2_{\delta^*}$ can be consistently estimated by plugging $\widehat{p}(x)$, $\widehat{f}_X(x)$, and $\widehat{f}_{X^*}(x)$ in Section 4.3 into $\varrho_{\delta^*}(x)$, one can apply a standard t-test to test whether there is a counterfactual mean effect $H_0:\delta^*=0$. More interestingly, in the next section we will introduce an over-identification test based on a similar testing procedure.

7.2 Over-identification Test

In this section, we propose a simple t-test to see if our method works properly in some cases where Y_0^* or Y_1^* is always observed.[12] Without loss of generality, we assume that Y_0^* is always observed. The basic idea of our test is that when Y_0^* is always observed, then we will have two estimators for $\mu_0^*=\mathrm{E}(Y_0^*)$. The first one could be the sample average of the observed Y_0^*'s and the second estimator can be obtained from Section 7.1. More specifically, the two estimators are respectively given by

$$\widetilde{\mu}_0^*=\frac{1}{n^*}\sum_{j=1}^{n^*}Y^*_{0,j},\qquad\qquad \widehat{\mu}_0^*=\frac{1}{n^*}\sum_{j=1}^{n^*}\widehat{\mathrm{E}}(Y_0|X=X_j^*).$$

Suppose that our method works properly, then both $\widetilde{\mu}_0^*$ and $\widehat{\mu}_0^*$ should have the same probability limit. If not, then these two estimators in general will have different probability limits. We can therefore test the validity of our method by examining whether $\widetilde{\mu}_0^*$ and $\widehat{\mu}_0^*$ have the same limits.

Under the same regularity conditions in Section 7.1, we have:

$$\sqrt{n}(\widetilde{\mu}_0^*-\widehat{\mu}_0^*)\xrightarrow{d}\mathcal{N}(0,\sigma^2)$$

where $\sigma^2=\mathrm{E}[\varrho(Z)^2]+\mathrm{E}[\varphi(X^*)^2]$ with

$$\varrho(Z)=-\left\{\frac{(1-D)[Y-\mathrm{E}(Y_0|X)]}{1-p(X)}\right\}\frac{f_{X^*}(X)}{f_X(X)},\quad \varphi(X^*)=\sqrt{\lambda}[Y_0^*-\mathrm{E}(Y_0|X^*)].$$

Thus, given a consistent estimator $\widehat{\sigma}^2$ for σ^2, one can use a simple two-sided t-test with the test statistic $\sqrt{n}(\widetilde{\mu}_0^*-\widehat{\mu}_0^*)/\widehat{\sigma}$. We conclude this section by noting that our method may not work properly when $|\sqrt{n}(\widetilde{\mu}_0^*-\widehat{\mu}_0^*)/\widehat{\sigma}|$ is too large.

[12] We thank a referee for suggesting this idea.

8 Conclusion

This paper proposes a unified nonparametric approach to the estimation and inference for the quantile treatment effect in a counterfactual environment. In particular, we extrapolate the changes in the effect of a status quo treatment under the assumption that the treatment is implemented in a population with a different distribution of observed covariates. Thus, instead of speculating about the new treatment effect, a researcher or policy maker can formally estimate it before actual implementation. While the analysis hinges on strong identifying conditions (unconfoundedness and invariance of conditional distributions), these assumptions can still be reasonable in some applications and, at the very least, make the extrapolation process transparent.

We derive the asymptotic properties of the proposed kernel-based estimator and provide a multiplier bootstrap procedure suitable for conducting uniform inference on the quantile counterfactual treatment effect over a continuum of quantile indices. We state similar results for the average counterfactual treatment effect and the counterfactually treated subpopulation. In our assessment, the multiplier bootstrap is more convenient to implement in this setting than a standard nonparametric bootstrap procedure.

As an empirical illustration, we apply the methods to two counterfactual exercises regarding the heterogeneous impact of Job Corps with respect to gender and education. By exchanging the individual characteristics of the two genders while keeping the earnings structure fixed, our first finding reinforces the argument in Strittmatter (2019) that the heterogeneous Job Corps effect by gender can be attributed to structural gender earnings inequality rather than gender differences in Job Corps trainability. By giving extra education to a more disadvantaged subgroup, our second finding indicates that the efficacy of Job Corps does not necessarily improve with better labor market opportunities. While this result does not support the skill hypothesis raised by Frumento et al. (2012) and Eren and Ozbeklik (2014), it is consistent with the finding of Card et al. (2018) arguing that job search assistance programs appear to be relatively more successful for disadvantaged participants.

Appendix A Proof of Lemma 2.1:

By the law of iterated expectations, Assumption 2.3, Assumption 2.1(i), and $Y = Y_d$ for $D = d$, we have

$$\begin{aligned} F_{Y_d^*}(y) &= \int_{\mathcal{X}^*} F_{Y_d^*|X^*}(y|x)\,\mathrm{d}F_{X^*}(x) = \int_{\mathcal{X}} F_{Y_d|X}(y|x)\,\mathrm{d}F_{X^*}(x) \\ &= \int_{\mathcal{X}} F_{Y_d|D,X}(y|d,x)\,\mathrm{d}F_{X^*}(x) = \int_{\mathcal{X}} F_{Y|D,X}(y|d,x)\,\mathrm{d}F_{X^*}(x), \end{aligned}$$

where $F_{Y|D,X}(y|d,x)$ is well defined for all d and x under Assumption 2.1(ii). Since X^* is defined on the same sample space as X that takes values inside $\mathcal{X}$ with probability 1 by Assumption 2.3(ii), $F_{Y_d^*}(y)$ is identified. Accordingly, the quantile functions and the QCTE can be identified as well. □

Appendix B Implementation of the Monotonization Method

This section describes the implementation of the monotonization method in (3.3) that can be easily computed. First, without loss of generality assume that there are no ties between Y_i's. Since $\widetilde{F}_{Y_d^*}(y)$ is a step function with jumps at the Y_i's, let $Y_{(i)}$ denote the ith smallest element among the Y_i's and add $Y_{(0)} = 0$ and $Y_{(n+1)} = \bar{y}$. In other words, we have $0 = Y_{(0)} < Y_{(1)} < \cdots < Y_{(n)} < Y_{(n+1)} = \bar{y}$. Denote $\overline{F}_d \equiv \sup_{y\in\mathcal{Y}} \widetilde{F}_{Y_d^*}(y)$ for $d = 0, 1$. Note that $\overline{F}_d \geq 1$ since $\widetilde{F}_{Y_d^*}(\bar{y}) = 1$. We then construct $\widehat{F}_{Y_d^*}(y)$ by induction:

1. Define $\widehat{F}_{Y_d^*}(y) = 0$ for $Y_{(0)} \leq y < Y_{(1)}$.
2. Suppose $\widehat{F}_{Y_d^*}(y)$ is already defined for $Y_{(0)} \leq y < Y_{(i)}$, we then define $\widehat{F}_{Y_d^*}(y)$ for $Y_{(i)} \leq y < Y_{(i+1)}$ as

$$\widehat{F}_{Y_d^*}(y) = \widehat{F}_{Y_d^*}(Y_{(i-1)})\,1\left\{\frac{\widetilde{F}_{Y_d^*}(Y_{(i)})}{\overline{F}_d} \leq \widehat{F}_{Y_d^*}(Y_{(i-1)})\right\} + \frac{\widetilde{F}_{Y_d^*}(Y_{(i)})}{\overline{F}_d}\,1\left\{\frac{\widetilde{F}_{Y_d^*}(Y_{(i)})}{\overline{F}_d} > \widehat{F}_{Y_d^*}(Y_{(i-1)})\right\}.$$

Continuing this way, we can construct $\widehat{F}_{Y_d^*}(y)$ that is monotonically increasing and lies between the unit interval for all $y \in \mathcal{Y}$.

Acknowledgement

This paper replaces earlier versions titled “Forecasting Treatment Effects” and “Estimating Counterfactual Treatment Effects to Assess External Validity.” We thank Irene Botosaru, Le-Yu Chen, Xu Cheng, Yingying Dong, Hiroaki Kaido, Toru Kitagawa, Chung-Ming Kuan, Ying-Ying Lee, Tatsushi Oka, Yuya Sasaki, Zhentao Shi, Ji-Liang Shiu, the editor Jianqing Fan, two anonymous referees, and conference participants of SETA 2015, AMES 2016 and TER 2016 for helpful suggestions and comments. All errors are our responsibility. Yu-Chin Hsu gratefully acknowledges research support from the Ministry of Science and Technology of Taiwan (103-2628-H-001-001-MY4 and 107-2410-H-001-034-MY3) and the Career Development Award of Academia Sinica, Taiwan. Tsung-Chih Lai gratefully acknowledges research support from the Ministry of Science and Technology of Taiwan (106-2410-H-035-003).

References

[1] Allcott, H. (2015). "Site selection bias in program evaluation." *The Quarterly Journal of Economics*, 130(3), 1117-1165.

[2] Andrews, I., and Oster, E. (2019). "A simple approximation for evaluating external validity bias." *Economics Letters*, 178, 58-62.

[3] Angrist, J. D. (2004). "Treatment effect heterogeneity in theory and practice." *The economic journal*, 114(494), C52-C83.

[4] Angrist, J. D., and Fernandez-Val I. (2013). "ExtrapoLATE-ing: External validity and overidentification in the LATE framework." In *Advances in Economics and Econometrics: Tenth World Congress, Vol. III: Econometrics*, ed. D. Acemoglu, M. Arellano, and E. Dekel, 401-434. Cambridge University Press.

[5] Athey, S., and Imbens, G. W. (2017). "The state of applied econometrics: Causality and policy evaluation." *Journal of Economic Perspectives*, 31(2), 3-32.

[6] Bertanha, M., and Imbens, G. W. (2019). "External validity in fuzzy regression discontinuity designs." *Journal of Business and Economic Statistics*, 1-39.

[7] Brinch, C. N., Mogstad, M., and Wiswall, M. (2017). "Beyond LATE with a discrete instrument." *Journal of Political Economy*, 125(4), 985-1039.

[8] Busso, M., DiNardo, J., and McCrary, J. (2009). "Finite sample properties of semiparametric estimators of average treatment effects." *Journal of Business and Economic Statistics*, forthcoming.

[9] Card, D., Kluve, J., and Weber, A. (2018). "What works? A meta analysis of recent active labor market program evaluations." *Journal of the European Economic Association*, 16(3), 894-931.

[10] Chernozhukov, V., Fernandez-Val, I., and Galichon, A. (2009). "Improving point and interval estimators of monotone functions by rearrangement." *Biometrika*, 96(3), 559-575.

[11] Chernozhukov, V., Fernandez-Val, I., and Galichon, A. (2010). "Quantile and probability curves without crossing." *Econometrica*, 78(3), 1093-1125.

[12] Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2013). "Inference on counterfactual distributions." *Econometrica*, 81(6), 2205-2268.

[13] Dehejia, R., Pop-Eleches, C., and Samii, C. (2019). "From local to global: External validity in a fertility natural experiment." *Journal of Business and Economic Statistics*, 1-27.

[14] DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). "Labor market institutions and the distribution of wages, 1973-1992: a semiparametric approach." *Econometrica*, 64(5), 1001-1044.

[15] Donald, S. G., and Hsu, Y. C. (2014). "Estimation and inference for distribution functions and quantile functions in treatment effect models." *Journal of Econometrics*, 178, 383-397.

[16] Donald, S. G., Hsu, Y. C., and Lieli, R. P. (2014). "Testing the unconfoundedness assumption via inverse probability weighted estimators of (L) ATT." *Journal of Business and Economic Statistics*, 32(3), 395-415.

[17] Eren, O., and Ozbeklik, S. (2014). "Who benefits from job corps? a distributional analysis of an active labor market program." *Journal of Applied Econometrics*, 29(4), 586-611.

[18] Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66* (Vol. 66). CRC Press.

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

[20] Firpo, S., Fortin, N. M., and Lemieux, T. (2009). "Unconditional quantile regressions." *Econometrica*, 77(3), 953-973.

[21] Fortin, N., Lemieux, T., and Firpo, S. (2011). *Decomposition methods in economics. In Handbook of labor economics* (Vol. 4, pp. 1-102). Elsevier.

[22] Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2012). "Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data." *Journal of the American Statistical Association*, 107(498), 450-466.

[23] Hahn, J. (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects." *Econometrica*, 315-331.

[24] Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2015). "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3), 757-778.

[25] Heckman, J. J., Ichimura, H., and Todd, P. (1998). "Matching as an econometric evaluation estimator." *The review of economic studies*, 65(2), 261-294.

[26] Heckman, J. J., and Vytlacil, E. (2005). "Structural equations, treatment effects, and econometric policy evaluation." *Econometrica*, 73(3), 669-738.

[27] Heckman, J. J., and Vytlacil, E. J. (2007). "Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments." *Handbook of econometrics*, 6, 4875-5143.

[28] Hirano, K., Imbens, G. W., and Ridder, G. (2003). "Efficient estimation of average treatment effects using the estimated propensity score." *Econometrica*, 71(4), 1161-1189.

[29] Horowitz, J. L. (2019). "Bootstrap methods in econometrics." *Annual Review of Economics*, 11, 193-224.

[30] Hotz, V. J., Imbens, G. W., and Mortimer, J. H. (2005). "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics*, 125(1-2), 241-270.

[31] Hsu, Y.-C., Lieli, R. P., and Lai, T.-C. (2019). "Estimation and inference for distribution functions and quantile functions in endogenous treatment effect models." Working Paper.

[32] Imai, K., Tingley, D., and Yamamoto, T. (2013). "Experimental designs for identifying causal mechanisms." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 5-51.

[33] Imbens, G. W. (2010). "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic literature*, 48(2), 399-423.

[34] Imbens, G. W., and Wooldridge, J. M. (2009). "Recent developments in the econometrics of program evaluation." *Journal of economic literature*, 47(1), 5-86.

[35] Kline, B., and Tamer, E. (2018). "Identification of treatment effects with selective participation in a randomized trial." *The Econometrics Journal*, 21(3), 332-353.

[36] Kowalski, A. E. (2016). "Doing more when you're running LATE: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments" (No. w22363). National Bureau of Economic Research.

[37] Lechner, M., and Strittmatter, A. (2019). "Practical procedures to deal with common support problems in matching estimation." *Econometric Reviews*, 38(2), 193-207.

[38] Li, Q., and Racine, J. S. (2008). "Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data." *Journal of Business and Economic Statistics*, 26(4), 423-434.

[39] Rothe, C. (2010). "Nonparametric estimation of distributional policy effects." *Journal of Econometrics*, 155(1), 56-70.

[40] Ruppert, D., and Wand, M. P. (1994). "Multivariate locally weighted least squares regression." *The annals of statistics*, 1346-1370.

[41] Schochet, P. Z., Burghardt, J., and McConnell, S. (2008). "Does job corps work? Impact findings from the national job corps study." *American Economic Review*, 98(5), 1864-86.

[42] Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). "The use of propensity scores to assess the generalizability of results from randomized trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2), 369-386.

[43] Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

[44] Wunsch, C., and Strobl, R. (2018). "Identification of Causal Mechanisms Based on Between-Subject Double Randomization Designs."