Estimation and Inference for Counterfactual Treatment Effects

Yu-Chin Hsu*

Institute of Economics
Academia Sinica
&
Department of Finance
National Central University
&
Department of Economics
National Chengchi University

Tsung-Chih Lai†

Department of Economics
Feng Chia University

Robert P. Lieli‡

Department of Economics
Central European University

This version: September 12, 2018

* ychsu@econ.sinica.edu.tw. † Corresponding author. tclai@fcu.edu.tw. ‡ lielir@ceu.edu.
Acknowledgements: This paper replaces an earlier version titled "Forecasting Treatment Effects." 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, 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 (MOST103-2628-H-001-001-MY4) and the Career Development Award of Academia Sinica, Taiwan.

Abstract

This paper proposes statistical methods to evaluate the quantile counterfactual treatment effect (QCTE) when the composition of the population targeted by a status quo program changes. QCTE enables us to carry out an ex-ante assessment of the distributional impacts of policy interventions or conduct a meta-analysis to investigate possible explanations for treatment effect heterogeneity. Assuming unconfoundedness and invariance of the conditional distributions of 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 then propose a multiplier bootstrap procedure to construct uniform confidence bands, and provide similar results for the counterfactually treated subpopulation and the average effects. As an application, we estimate the QCTE of the Job Corps training program in the U.S. under various scenarios. Our results suggest that strong economic performance indeed explains the earlier finding in the literature that the program is ineffective at low quantiles of the earnings distribution. However, no supportive evidence is found for the skill hypothesis.

JEL Classification: C14, C31, J30.

Keywords: Program evaluation, counterfactual analysis, multiplier bootstrap, Job Corps

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 a population from which data are actually drawn, more recent studies have focused on extrapolating these estimates to new environments that the program might be extended to. For example, Stuart, Cole, Bradshaw, and Leaf (2011), Andrews and Oster (2018), and Kline and Tamer (2018) address the generalizability of treatment effect estimates obtained from a randomized trial where the sample is not representative of the ultimate target population. Hotz, Imbens, and Mortimer (2005), Hartman, Grieve, Ramsahai, and Sekhon (2015), and Dehejia, Pop-Eleches, and Samii (2017) consider extrapolating the treatment effect estimates from one location to another. While these studies provide 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.

Example 1 (Program 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 policymaker plans to expand the program to a new location where only individual characteristics are currently observed, and the first goal of the policymaker may be to predict the program effect in the new location prior to actual implementation.

Example 2 (Policy Intervention). The policymaker plans to manipulate (the distribution of) individual characteristics in a population currently targeted by some training program. For example, direct subsidies can be provided to change individuals' pre-treatment level of income. The policymaker may want to first predict the resulting change in the program effect, especially for the low-income subpopulation.

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 us to carry out an ex-ante assessment of the distributional impacts of policy interventions or conduct a meta-analysis to investigate possible explanations for treatment effect heterogeneity. Our framework is fully nonparametric in that we only restrict QCTE via general

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

conditions on the distributions of the potential outcomes in the counterfactual environment. 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 quantile counterfactual treatment effect for the treated (QCTT) in this paper, but for exposition we 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 argued for on a case-by-case basis (see Section 6 for an illustration).

Given the assumptions described above, QCTE can be nonparametrically identified and estimated. In the first step, we use a Nadaraya-Watson estimator to construct the conditional cumulative distribution functions (c.d.f.'s) of the status quo potential outcomes for $X = x$ 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 these two c.d.f.'s to obtain the estimated quantile functions. Taking the difference at any given quantile index gives the estimated value of QCTE at that point. 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 QCTT and state similar results.

As a special case, one may also integrate with respect to the empirical distribution of the status quo covariates in the second step described above. We then obtain an estimator for the distribution functions of the status quo potential outcomes, as an alternative to the inverse probability weighted (IPW) estimator proposed by Donald and Hsu (2014). Our estimator is first-order asymptotically equivalent to the IPW estimator in this special case.

We illustrate the usefulness of the proposed methods by a Monte Carlo simulation and an empirical study of the heterogeneous impact of Job Corps, the largest and most costly labor market program in the United States, on the distribution of earnings. According to Eren and Ozbeklik (2014), the program has not proved effective for individuals toward the bottom of the earnings distribution. Two possible explanations are offered for this finding: (i) strong economic conditions during the evaluation phase in the late 1990s; (ii) the relatively low skill and education level of the low earner subgroup. We test the empirical relevance of these explanations based on their implications for counterfactual analysis. Specifically, we first use data from an earlier training program, the Job Training Partnership Act (JTPA), to estimate the counterfactual effect of Job Corps on the population targeted by JTPA. If the strong economic conditions hypothesis is true, then the program effect should be insignificant in this counterfactual scenario as well. In addition, we reverse the roles of the two programs and estimate the counterfactual effect of JTPA for Job Corps individuals. Under the same hypothesis, the Job Corps cohort would be expected to attain a significantly larger treatment effect, as the early 1990s, the evaluation period for JTPA, was a recessionary period in the United States. The empirical evidence is consistent with the predictions.

To assess the role of the skill hypothesis, we artificially give extra education to individuals who do not benefit from Job Corps, while holding the education of others constant. If this hypothesis is true, then we would expect to see a significantly increased program effect at the corresponding earnings quantiles. However, the empirical findings do not support this hypothesis, suggesting that strong economic conditions are the primary reason for the apparent ineffectiveness of Job Corps at lower earnings quantiles.

Our paper is related to a number of previous studies on extrapolating treatment effects to different populations or settings. In their survey article, Athey and Imbens (2017) highlight three directions of research in the external validity literature. The first direction concerns instrumental variable settings, where the aim is to extrapolate the local average treatment effect, the average treatment effect for compliers, to other subgroups or the entire population (Angrist and Fernandez-Val, 2013; Kowalski, 2016; Bertanha and Imbens, 2018). The second direction considers extrapolation in the context of regression discontinuity designs, where estimates are generally valid only for units with values of the forcing variable close to the cutoff point (Angrist and Rokkanen, 2015; Dong and Lewbel, 2015; Bertanha and Imbens, 2018). The third direction, which is most closely related to this paper, extrapolates treatment effects from location to location, under the assumption that the differences between locations are due to the different compositions of individual characteristics (Hotz, Imbens, and Mortimer, 2005; Imbens, 2010; Allcott, 2015). As mentioned above, our paper makes the following contributions to this framework: (i) We relax the requirement that the status quo data be drawn from a randomized experiment; instead, we allow for observational data as long as the treatment assignment is unconfounded. (ii) We extend the analysis beyond average effects and provide a unified nonparametric framework for estimating QCTE and conducting uniform inference.

This last strand of the external validity literature, including our paper, is also related to previous work on estimating counterfactual distributions. For example, Firpo, Fortin, and Lemieux (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, Fernandez-Val, and Melly (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, 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. 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.[2] Third, we apply a new monotonization method to ensure that the unconditional distribution function estimators obtained in step two above 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, Lieli, and Lai (2018), which simply replaces any downward step in the c.d.f. estimate with a constant piece and is very easy to implement.

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 that are crucial for the validity of the multiplier bootstrap in Section 4. We also discuss how to conduct uniform inference in Section 4. Section 5 presents the Monte Carlo simulation and Section 6 the empirical study. Section 7 extends the analysis to the average case and the treated subset of the counterfactual population. Section 8 concludes. All proofs are collected in Appendix A.

2 Model Framework and Identification

2.1 Model

We first introduce the Rubin causal model in the status quo population. Let $D \in \{0,1\}$ be a binary treatment assignment and Y_d be the corresponding potential outcomes for $d = 0, 1$; that is, Y_1 is the outcome if an individual is exogenously assigned to treatment ($D = 1$), and Y_0 is the outcome in the absence of treatment ($D = 0$). The actual observed outcome is $Y = DY_1 + (1 - D)Y_0$. Next, we observe a k-dimensional vector of pre-treatment covariates $X = (X_1, \ldots, X_k)$ in the status quo environment and a covariate vector $X^* = (X_1^*, \ldots, X_k^*)$ in the counterfactual environment which is of the same dimension as X. The relationship between

[2]The tradeoff is that the multiplier bootstrap requires consistent estimation of the covariance functions. See Section 4.1 for more details.

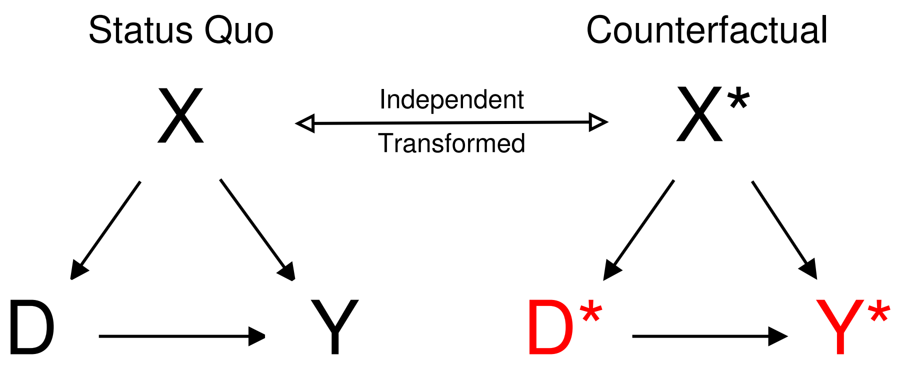


Figure 1. Model Framework

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 assignment, outcome, and potential outcomes in the counterfactual population are respectively denoted as D^*, Y^*, and Y_d^* for $d = 0, 1$ with $Y^* = D^*Y_1^* + (1 - D^*)Y_0^*$. Note that since the treatment is not yet implemented in the counterfactual population, neither D^* nor Y^* (and therefore both Y_0^* and Y_1^*) is observed in our model.

Figure 1 provides a graphical illustration of the model framework. The observed variables consisting of X, D, and Y in the status quo environment and X^* in the counterfactual environment are indicated in bold black. In contrast, the unobservables are shown in bold red.

2.2 Parameters of Interest

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 potential outcome distributions. Analogous to ATE and QTE, we define the average counterfactual treatment effect (ACTE) as the mean difference between Y_0^* and Y_1^*,

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

and the quantile counterfactual treatment effect (QCTE) as the difference between two quantile functions of Y_0^* and Y_1^* for some quantile index $\tau \in [0, 1]$,

$$\delta^*(\tau) = \mathbb{Q}_{Y_1^*}(\tau) - \mathbb{Q}_{Y_0^*}(\tau), \tag{2.2}$$

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

From a policymaker's standpoint, treatment effects for the treated subgroup are often more interesting than for the overall population. We then consider the average counterfactual treatment effect for the treated (ACTT) and the quantile counterfactual treatment effect for the treated (QCTT) as

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

where the expectation and quantile operators are taken with respect to the conditional distribution of Y_d^* given $D^*=1$. In the exposition we will focus primarily on QCTE and provide a shorter discussion of ACTT and QCTT in Section 7.

2.3 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 $p(x) = \mathbb{P}(D=1|X=x)$ 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 the support of X and redefining the status quo population as advocated by Crump, Hotz, Imbens, and Mitnik (2009). Note that Assumption 2.1 allows the use of observational (non-experimental) data in evaluating the status quo treatment. The second set of assumptions makes extrapolation of treatment effects possible.

Assumption 2.2 (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 2.2 appears frequently in the decomposition literature (Firpo, Fortin, and Lemieux, 2009; Fortin, Lemieux, and Firpo, 2011; Chernozhukov, Fernandez-Val, and Melly, 2013), and it stipulates that the difference between the status quo and counterfactual treatment effects arises solely from the different marginal distributions of X and X^*. This condition also comes from the "no macro-effects" assumption in Hotz, Imbens, and Mortimer (2005) and the policy invariance condition in Heckman and Vytlacil (2005, 2007) and Dong and Lewbel (2015). In addition, if one follows Rothe (2010) to specify a nonparametric and nonseparable 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 an error term representing unobserved heterogeneity, then a sufficient condition for Assumption 2.2(i) is the independence between ε_d and (X, X^*) as imposed in Rothe (2010):

$$\begin{aligned} F_{Y_d^*|X^*}(y|x) = \mathbb{P}\big(m_d(X^*, \varepsilon_d) \leq y \big| X^* = x\big) &= \mathbb{P}\big(m_d(x, \varepsilon_d) \leq y\big) \\ &= \mathbb{P}\big(m_d(X, \varepsilon_d) \leq y \big| X = x\big) = F_{Y_d|X}(y|x). \end{aligned}$$

Assumption 2.2(ii) is a support condition that is weaker than the complete overlap imposed in Hotz, Imbens, and Mortimer (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 2.2(ii) is violated, one could drop units in the counterfactual environment with covariates outside the common support and redefine ACTE and QCTE relative to the new support.

Lemma 1. *Suppose Assumptions 2.1 and 2.2 hold. QCTE is identified by*

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

To see Lemma 1, we first note under Assumption 2.2 that the distribution function $F_{Y_d^*}(y)$ 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. Identification results for ACTE and the treated cases can be found in Section 7.1.

3 Estimation and Asymptotic Properties

3.1 Estimation Procedure

Given the identification result in Lemma 1, we estimate QCTE in the following steps. First, using data from the status quo environment, we construct estimators for the conditional distribution functions $F_{Y_d|X}(y|x)$, $d = 0, 1$. Second, we average with respect to the empirical measure of X^* to estimate the unconditional distribution functions $F_{Y_d^*}(y)$. Third, we eliminate any non-monotonicity and invert the estimators to obtain estimates of the quantile functions $\mathbb{Q}_{Y_d^*}(\tau)$, $d = 0, 1$. Finally, one can estimate QCTE at any given quantile τ by taking the difference of these two estimates. To define the estimators more formally, we make the following assumption.

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* $0 < \lambda < \infty$*.*

Similarly to Rothe (2010), we use a kernel-based (Nadaraya-Watson) distribution function estimator in the first step:

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

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, and $\mathcal{K}_{x,h}(\cdot) = h^{-k}\, \mathcal{K}_x(\cdot/h)$ 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).[3] As the rate of convergence of the estimator will not be affected in either case, we will for simplicity assume that X is continuous.

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 $F_{Y_d^*}(y)$:

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

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

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

can be circumvented by either using the reweighting method in Rothe (2010), the rearranging method in Chernozhukov, Fernandez-Val, and Galichon (2009, 2010), or the monotonizing method in Hsu, Lieli, and Lai (2018), which is adopted in this paper. Specifically, define the functionals ϕ_1, ϕ_2, and ϕ so that for any function g with $\sup_{y\in\mathcal{Y}} g(y) > 0$,

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

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

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

The preliminary estimator $\widetilde{F}_{Y_d^*}$ is already a step function over $\mathcal{Y}$; the functional ϕ_2 simply rescales it to ensure that its maximum value is 1, and ϕ_1 eliminates any negative values or downward steps. (Downward steps are replaced by the value of the last upward step.) Thus, $\widehat{F}_{Y_d^*}(y)$ is a proper distribution function estimator. In Appendix B we provide an easy-to-implement procedure for the transformation in (3.3). We will also argue that $\widehat{F}_{Y_d^*}(y)$ and $\widetilde{F}_{Y_d^*}(y)$ are first-order asymptotically equivalent under the regularity conditions introduced below. This equivalence allows us to apply $\widehat{F}_{Y_d^*}(y)$ in practice, while concentrating on the limiting behavior of $\widetilde{F}_{Y_d^*}(y)$ in the theoretical derivation. Finally, our QCTE estimator is defined as $\widehat{\delta}^*(\tau) = \widehat{\mathbb{Q}}_{Y_1^*}(\tau) - \widehat{\mathbb{Q}}_{Y_0^*}(\tau)$, where

$$\widehat{\mathbb{Q}}_{Y_d^*}(\tau) = \inf\{y \in \mathcal{Y} : \widehat{F}_{Y_d^*}(y) \geq \tau\}. \tag{3.4}$$

3.2 Regularity Conditions

Before starting the asymptotic analysis of the proposed estimators, we gather all regularity conditions 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 $u^\gamma = \prod_{s=1}^k u_s^{\gamma_s}$. Furthermore, let r denote the order of the kernel function used in (3.1).

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 derivative is 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 derivative is 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* $\mathcal{K}_x$ *of order* r *satisfies:*

(i) $\int_{\mathcal{D}_x} \mathcal{K}_x(u)\, du = 1$.

(ii) $\int_{\mathcal{D}_x} u^\gamma \mathcal{K}_x(u)\, du = 0$ *for all* $|\gamma| = 1, \dots, r-1$.

(iii) $\int_{\mathcal{D}_x} |u^\gamma \mathcal{K}_x(u)|\, du < \infty$ *for* $|\gamma| = r$.

(iv) $\mathcal{K}_x(u) = 0$ *if* $|u| > 1$.

(v) $\mathcal{K}_x(u)$ *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) $n^{1/2}h^k/\log n \to \infty$.

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

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.[4] 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.

[4] If $\mathcal{Y} = \mathbb{R}$, one can still estimate QCTE at the parametric rate uniformly over compact subsets of the unit interval on which $f_{Y_d^*}(y)$ is bounded away from 0.

3.3 Asymptotic Properties

We now investigate the asymptotic properties of $\widehat{F}_{Y_d^*}(y)$ given the 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 $\boldsymbol{y} = (y_0, y_1)^T$, $\boldsymbol{F}(\boldsymbol{y}) = \big(F_{Y_0^*}(y_0), F_{Y_1^*}(y_1)\big)^T$, $\widehat{\boldsymbol{F}}(\boldsymbol{y}) = \big(\widehat{F}_{Y_0^*}(y_0), \widehat{F}_{Y_1^*}(y_1)\big)^T$, and $Z = (Y, D, X)$.

Lemma 2. *Suppose Assumptions 2.1, 2.2, and 3.1–3.6 hold. We then have:*

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

where $\mathcal{F}(\boldsymbol{y}) = \big(\mathcal{F}_0(y_0), \mathcal{F}_1(y_1)\big)^T$ *is a two-dimensional zero mean Gaussian process with covariance function* $\Psi^F(\boldsymbol{y}, \boldsymbol{y}') = \mathbb{E}\big[\varrho^F(\boldsymbol{y}, Z)\varrho^F(\boldsymbol{y}', Z)^T\big] + \mathbb{E}\big[\varphi^F(\boldsymbol{y}, X^*)\varphi^F(\boldsymbol{y}', X^*)^T\big]$*, and the convergence takes place in* $\ell^\infty(\mathcal{Y}) \times \ell^\infty(\mathcal{Y})$*, where* $\ell^\infty(\mathcal{Y})$ *is the space of bounded functions over* $\mathcal{Y}$*. Here,* $\varrho^F(\boldsymbol{y}, Z) = \big(\varrho_0^F(y_0, Z), \varrho_1^F(y_1, Z)\big)^T$ *and* $\varphi^F(\boldsymbol{y}, X^*) = \big(\varphi_0^F(y_0, X^*), \varphi_1^F(y_1, X^*)\big)^T$ *are defined as*

$$\begin{aligned} \varrho_d^F(y, Z) &= \frac{\mathbb{1}\{D = d\}\left[\mathbb{1}\{Y \le y\} - F_{Y_d|X}(y|X)\right]}{p(X)^d\left[1 - p(X)\right]^{1-d}} \frac{f_{X^*}(X)}{f_X(X)}, \\ \varphi_d^F(y, X^*) &= \sqrt{\lambda}\left[F_{Y_d|X}(y|X^*) - F_{Y_d^*}(y)\right]. \end{aligned} \tag{3.5}$$

The proof of Lemma 2 can be found in Appendix A; here, we give a brief outline of the argument. We first show that $\sqrt{n}\big(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\big)$ is asymptotically linear with the influence function representation:

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

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 2 holds by the functional central limit theorem for $\widetilde{\boldsymbol{F}} = \big(\widetilde{F}_{Y_0^*}, \widetilde{F}_{Y_1^*}\big)^T$ 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 $\sup_{y \in \mathcal{Y}}\big|\widehat{F}_{Y_d^*}(y) - \widetilde{F}_{Y_d^*}(y)\big| = o_p(n^{-1/2})$, which completes the proof.

There are several points on Lemma 2 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

$$\psi_d^F(y, Z) = \frac{\mathbb{1}\{D = d\}\left[\mathbb{1}\{Y \leq y\} - F_{Y_d|X}(y|X)\right]}{p(X)^d\left[1 - p(X)\right]^{1-d}} + F_{Y_d|X}(y|X) - F_{Y_d}(y),$$

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 that the quantile map is Hadamard differentiable, the asymptotic properties of the QCTE estimator can be obtained immediately from Lemma 2 by the functional delta method. We state the result in the following theorem.

Theorem 1. *Suppose Assumptions 2.1, 2.2, and 3.1–3.6 hold. We then have:*

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

where $\Delta(\tau)$ *is a Gaussian process with mean zero and covariance function* $\Psi(\tau) = \mathbb{E}\left[\varrho(\tau, Z)^2\right] + \mathbb{E}\left[\varphi(\tau, X^*)^2\right]$, *where*

$$\begin{aligned}
\varrho(\tau, Z) &= -\left[\frac{\varrho_1^F\left(\mathbb{Q}_{Y_1^*}(\tau), Z\right)}{f_{Y_1^*}\left(\mathbb{Q}_{Y_1^*}(\tau)\right)} - \frac{\varrho_0^F\left(\mathbb{Q}_{Y_0^*}(\tau), Z\right)}{f_{Y_0^*}\left(\mathbb{Q}_{Y_0^*}(\tau)\right)}\right], \\
\varphi(\tau, X^*) &= -\left[\frac{\varphi_1^F\left(\mathbb{Q}_{Y_1^*}(\tau), X^*\right)}{f_{Y_1^*}\left(\mathbb{Q}_{Y_1^*}(\tau)\right)} - \frac{\varphi_0^F\left(\mathbb{Q}_{Y_0^*}(\tau), X^*\right)}{f_{Y_0^*}\left(\mathbb{Q}_{Y_0^*}(\tau)\right)}\right],
\end{aligned} \tag{3.6}$$

where ϱ_d^F *and* φ_d^F *are given in* (3.5), *and the convergence takes place in* $\ell^\infty([0,1])$.

Theorem 1 allows for pointwise inference on QCTE. For example, suppose that we want to test whether the counterfactual treatment effect exists at the median:

$$H_0 : \delta^*(\tau) = 0 \quad \text{for } \tau = 0.5.$$

One can simply construct an ordinary t-statistic based on a consistent estimate of the asymptotic covariance function,

$$\widehat{\Psi}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\varrho}(\tau, Z_i)^2 + \frac{1}{n^*}\sum_{j=1}^{n^*}\widehat{\varphi}(\tau, X_j^*)^2, \tag{3.7}$$

where $\widehat{\varrho}$ and $\widehat{\varphi}$ are later provided in Section 4.1.

4 Inference over a Continuum of Quantile Indices

Many interesting hypotheses involve a continuum of quantiles, 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 hypotheses over a continuum of quantile indices:

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

for $0 \leq \tau_\ell < \tau_u \leq 1$.

In this section we propose the multiplier bootstrap to simulate critical values for testing the above null hypotheses or constructing one- and two-sided uniform confidence bands. The multiplier bootstrap method can be regarded as a more convenient alternative to the nonparametric bootstrap proposed by Rothe (2010). However, in our setting the choice of the multiplier hinges on the relationship between X and X^* in order to ensure that the simulated process preserves the same relationship. Such a problem does not appear in previous applications of the multiplier bootstrap technique (Barrett and Donald, 2003; Kline and Santos, 2012; Chernozhukov, Chetverikov, and Kato, 2013, 2016; Donald and Hsu, 2014; Hsu, 2016).

In Section 4.1, we describe the multiplier bootstrap procedure and show its validity. Section 4.2 discusses the hypothesis testing and uniform confidence bands. A step-by-step implementation is provided in Section 4.3.

4.1 Multiplier Bootstrap

To approximate the true limiting process $\Delta(\tau)$, 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 covariance kernel $\Psi(\tau)$. Monotonicity of the estimators for $F_{Y_d^*}(y)$ and $F_{Y_d|X}(y|x)$ is also necessary 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 arguments* y *and* x*.*

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

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

$$\begin{aligned}
\widehat{\varrho}(\tau, Z_i) &= -\left[\frac{\widehat{\varrho}_1^F\big(\widehat{\mathbb{Q}}_{Y_1^*}(\tau), Z_i\big)}{\widehat{f}_{Y_1^*}\big(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\big)} - \frac{\widehat{\varrho}_0^F\big(\widehat{\mathbb{Q}}_{Y_0^*}(\tau), Z_i\big)}{\widehat{f}_{Y_0^*}\big(\widehat{\mathbb{Q}}_{Y_0^*}(\tau)\big)}\right], \\
\widehat{\varphi}(\tau, X_j^*) &= -\left[\frac{\widehat{\varphi}_1^F\big(\widehat{\mathbb{Q}}_{Y_1^*}(\tau), X_j^*\big)}{\widehat{f}_{Y_1^*}\big(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\big)} - \frac{\widehat{\varphi}_0^F\big(\widehat{\mathbb{Q}}_{Y_0^*}(\tau), X_j^*\big)}{\widehat{f}_{Y_0^*}\big(\widehat{\mathbb{Q}}_{Y_0^*}(\tau)\big)}\right],
\end{aligned} \tag{4.2}$$

where $\widehat{\mathbb{Q}}_{Y_d^*}(\tau)$ is in (3.4) and

$$\begin{aligned}\widehat{\varrho}_d^F(y, Z_i) &= \frac{\mathbb{1}\{D_i = d\}\left[\mathbb{1}\{Y_i \le y\} - \widehat{F}_{Y_d|X}(y|X_i)\right]}{\widehat{p}(X_i)^d\left[1 - \widehat{p}(X_i)\right]^{1-d}} \frac{\widehat{f}_{X^*}(X_i)}{\widehat{f}_X(X_i)}, \\ \widehat{\varphi}_d^F(y, X_j^*) &= \sqrt{\widehat{\lambda}}\left[\widehat{F}_{Y_d|X}(y|X_j^*) - \widehat{F}_{Y_d^*}(y)\right],\end{aligned} \tag{4.3}$$

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 $\{(Z_i, X_j^*) : 1 \le i \le n, 1 \le j \le n^*, n, n^* \ge 1\}$. The simulated process for $\Delta(\tau)$ is then given by

$$\Delta^u(\tau) = \begin{cases} \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} U_i\left[\widehat{\varrho}(\tau, Z_i) + \widehat{\varphi}(\tau, X_i^*)\right] & \text{if} \quad X^* = \pi(X), \\ \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} U_i\widehat{\varrho}(\tau, Z_i) + \dfrac{1}{\sqrt{n^*}} \displaystyle\sum_{j=1}^{n^*} U_j^*\widehat{\varphi}(\tau, X_j^*) & \text{if} \quad X^* \perp\!\!\!\perp X. \end{cases} \tag{4.4}$$

As can be seen from (4.4), the choice of multiplier depends on the relationship between X and X^*: If $X^* = \pi(X)$, one can utilize a single multiplier U_i associated with $\widehat{\varrho}(\tau, Z_i) + \widehat{\varphi}(\tau, X_i^*)$ to simulate the process. If $X^* \perp\!\!\!\perp X$, we need to introduce two sets of independent multipliers U_i's and U_j^*'s to guarantee the independence between the simulated processes of $\widehat{\varrho}(\tau, Z_i)$ and $\widehat{\varphi}(\tau, X_j^*)$, since $\widehat{\varrho}(\tau, Z_i)$ and $\widehat{\varphi}(\tau, X_j^*)$ may not be independent in finite samples (but their convergents are) when X and X^* are independent.

The next theorem asserts the validity of the multiplier bootstrap and relies on the conditional multiplier central limit theorem. Assumption 4.1, including the monotonicity requirement, plays an important role in the proof of Theorem 2.

Theorem 2. *Suppose Assumptions 2.1, 2.2, 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 $\{X_j^* : 1, 2, \ldots\}$ with probability approaching one.

4.2 Hypothesis Testing and Uniform Confidence Bands

For the functional null hypotheses stated in (4.1), the one- and two-sided standardized Kolmogorov-Smirnov test statistics are given by

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

where $\widehat{\sigma}(\tau) = \widehat{\Psi}^{1/2}(\tau)$ and $\widehat{\Psi}(\tau)$ is defined in (3.7). Given critical values $\widehat{C}_\alpha^{\text{1-sided}}$ and $\widehat{C}_\alpha^{\text{2-sided}}$, which will be constructed later, the null hypotheses are rejected if the test statistics exceed the corresponding critical values. In other words, the decision rules are:

$$\text{Reject } H_0^{\text{1-sided}} \text{ if } \widehat{S}_n^{\text{1-sided}} > \widehat{C}_\alpha^{\text{1-sided}}. \qquad \text{Reject } H_0^{\text{2-sided}} \text{ if } \widehat{S}_n^{\text{2-sided}} > \widehat{C}_\alpha^{\text{2-sided}}.$$

Based on Theorem 2, we now elaborate how to simulate the critical values via multiplier bootstrap. For a nominal significance level α and for $\tau_\ell, \tau_u \in [0,1]$ with $\tau_\ell < \tau_u$, let $\widehat{C}_\alpha^{\text{1-sided}}$ and $\widehat{C}_\alpha^{\text{2-sided}}$ respectively denote the one- and two-sided critical values that satisfy

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

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

Here, $\widehat{C}_\alpha^{\text{1-sided}}$ and $\widehat{C}_\alpha^{\text{2-sided}}$ are, respectively, the $(1-\alpha)$th quantile of $\sup_{\tau \in [\tau_\ell, \tau_u]} \Delta^u(\tau)/\widehat{\sigma}(\tau)$ and $(1-\alpha)$th quantile of $\sup_{\tau \in [\tau_\ell, \tau_u]} \left|\Delta^u(\tau)\right|/\widehat{\sigma}(\tau)$. Once the critical values are constructed, we can also obtain one- and two-sided uniform confidence bands for QCTE over $[\tau_\ell, \tau_u]$. Specifically, the one-sided $(1-\alpha)$ uniform confidence band is given by

$$\left(\widehat{\delta}^*(\tau) - \widehat{C}_\alpha^{\text{1-sided}} \frac{\widehat{\sigma}(\tau)}{\sqrt{n}}, \quad +\infty \right), \tag{4.5}$$

and the two-sided $(1-\alpha)$ uniform confidence band is

$$\left(\widehat{\delta}^*(\tau) - \widehat{C}_\alpha^{\text{2-sided}} \frac{\widehat{\sigma}(\tau)}{\sqrt{n}}, \quad \widehat{\delta}^*(\tau) + \widehat{C}_\alpha^{\text{2-sided}} \frac{\widehat{\sigma}(\tau)}{\sqrt{n}} \right), \quad \tau \in [\tau_\ell, \tau_u]. \tag{4.6}$$

4.3 Implementation of Uniform Confidence Bands

We now provide a step-by-step implementation for constructing uniform confidence bands.

1. Suppose we have estimates $\widehat{\delta}^*(\tau)$ from Section 3.1 and $\widehat{\varrho}(\tau, Z_i)$ and $\widehat{\varphi}(\tau, X_j^*)$ (and therefore $\widehat{\sigma}(\tau)$) from Section 4.1 with $\tau \in \{\tau_\ell, \tau_\ell + 0.01, \ldots, \tau_u\}$.

2. Draw i.i.d. pseudo random variables $\{U_1, \ldots, U_n\}$ and $\{U_1^*, \ldots U_{n^*}^*\}$ with mean zero and unit variance B times for, say, $B = 1000$. For each repetition $b = 1, \ldots, B$, calculate the simulated process $\Delta_b^u(\tau)$ according to (4.4).

3. For the one-sided case, store the maximum value of $\Delta_b^u(\tau)/\widehat{\sigma}(\tau)$ over the grid of τ values set up in Step 1; that is, let $M_b = \max_\tau \Delta_b^u(\tau)/\widehat{\sigma}(\tau)$ for $b = 1, \ldots, B$.

4. Rank the M_b values in an ascending order so that $M_{(1)} \le \ldots \le M_{(B)}$. Next, define $M_{(\lfloor (1-\alpha)B \rfloor)}$ as the critical value $\widehat{C}_\alpha^{\text{1-sided}}$, where $\lfloor a \rfloor$ is the floor function returning the largest integer not greater than a. The one-sided $(1-\alpha)$ uniform confidence bands for $\{\delta^*(\tau) : \tau \in [\tau_\ell, \tau_u]\}$ are given by (4.5).

5. For the two-sided case, simply replace $\Delta_b^u(\tau)/\widehat{\sigma}(\tau)$ in Step 3 with $\left|\Delta_b^u(\tau)\right|/\widehat{\sigma}(\tau)$ and repeat

Step 4 for the 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 given by (4.6).

4.4 Uniformly Consistent and Monotone Estimators

In this section, we provide kernel-based estimators that satisfy Assumption 4.1 and can thus be used to construct the simulated process in (4.4). Regarding the monotonicity requirement in Assumption 4.1(ii), we use $\widehat{F}_{Y_d^*}(y)$ in (3.3) and let

$$\widehat{F}_{Y_d|X}(y|x) = \phi_1(\widetilde{F}_{Y_d|X})(y|x) \tag{4.7}$$

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 $\sup_{y\in\mathcal{Y},x\in\mathcal{X}} \left|\widehat{F}_{Y_d|X}(y|x) - F_{Y_d|X}(y|x)\right| = o_p(1)$ in Lemma 3 below.

The kernel estimators for $p(x)$, $f_X(x)$ and $f_{X^*}(x)$ are now given by

$$\begin{aligned} \widetilde{p}(x) &= \frac{\sum_{i=1}^{n} D_i\, \mathcal{K}_{x,h}(X_i - x)}{\sum_{i=1}^{n} \mathcal{K}_{x,h}(X_i - x)}, \\ \widetilde{f}_X(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{x,h}(X_i - x) \quad \text{and} \quad \widetilde{f}_{X^*}(x) = \frac{1}{n^*}\sum_{j=1}^{n^*} \mathcal{K}_{x,h}(X_j^* - x), \end{aligned} \tag{4.8}$$

where the uniform consistency of (4.8) is established in Härdle, Jansson, and Serfling (1998) for $\widetilde{p}(x)$ and Jones (1993) for $\widetilde{f}_X(x)$ and $\widetilde{f}_{X^*}(x)$. A minor disadvantage of applying a boundary kernel (even if second-order) is that the estimators in (4.8) are not necessarily positive. We tackle this issue by applying the trimming method in Donald, Hsu, and Lieli (2014a, 2014b) for $\widetilde{p}(x)$ and the method in Hsu, Lieli, and Lai (2018) for $\widetilde{f}_X(x)$ and $\widetilde{f}_{X^*}(x)$. For the former, let

$$\widehat{p}(x) = a_n\, \mathbb{1}\{\widetilde{p}(x) \le a_n\} + \widetilde{p}(x)\, \mathbb{1}\{a_n < \widetilde{p}(x) < 1 - a_n\} + (1 - a_n)\, \mathbb{1}\{\widetilde{p}(x) \ge 1 - a_n\}, \tag{4.9}$$

where $\{a_n \in (0, 1/2) : n \ge 1\}$ is a positive sequence converging to 0 and can be determined by Corollary 1 in Crump, Hotz, Imbens, and Mitnik (2009).[5] 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 density function estimators, we follow the trimming method in Hsu, Lieli, and Lai (2018) and let

$$\widehat{f}_X(x) = \max\{\widetilde{f}_X(x), b_n\}, \qquad \widehat{f}_{X^*}(x) = \max\{\widetilde{f}_{X^*}(x), b_n\}, \tag{4.10}$$

where $\{b_n : n \ge 1\}$ is a decreasing sequence of positive numbers converging to 0.[6]

[5] One can also discard the observations with $\widetilde{p}(x)$ outside the interval $[a_n, 1 - a_n]$ as in Crump, Hotz, Imbens, and Mitnik (2009).

[6] Despite not necessarily integrating to one for all n, the estimators in (4.10) are still uniformly consistent for

We lastly construct the estimator for $f_{Y_d^*}(y)$ similar to $\widetilde{F}_{Y_d^*}(y)$ in (3.2). In other words, we let

$$\widetilde{f}_{Y_d^*}(y) = \frac{1}{n^*}\sum_{j=1}^{n^*} \widetilde{f}_{Y_d|X}(y|X_j^*),$$

where

$$\widetilde{f}_{Y_d|X}(y|x) = \frac{\sum_{i=1}^{n} \mathcal{W}_{y,\eta}(Y_i - y)\, \mathbb{1}\{D_i = d\}\, \mathcal{K}_{x,h}(X_i - x)}{\sum_{i=1}^{n} \mathbb{1}\{D_i = d\}\, \mathcal{K}_{x,h}(X_i - x)}, \tag{4.11}$$

with $\mathcal{W}_{y,\eta}(\cdot) = \eta^{-1}\,\mathcal{W}_y(\cdot/\eta)$ a boundary kernel and $\eta = \eta_n$ the bandwidth in the y direction. As before, $\widetilde{f}_{Y_d^*}(y)$ could be negative in finite samples. We then employ the trimming method in (4.10) again and let

$$\widehat{f}_{Y_d^*}(y) = \max\{\widetilde{f}_{Y_d^*}(y), b_n\}. \tag{4.12}$$

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

Lemma 3. *Suppose Assumptions 2.1, 2.2, and 3.1–3.6 hold. Moreover, suppose Assumption 3.5 holds with $\mathcal{K}_x$ replaced by $\mathcal{W}_y$ and a_n, b_n, and $\eta = \eta_n \to 0$ as $n \to \infty$. The estimators in* (3.3), (4.7), (4.9), (4.10), *and* (4.12) *then satisfy Assumption 4.1.*

5 Simulation Study

We examine the finite sample properties of QCTE and QCTT estimators and the multiplier bootstrap via Monte Carlo simulations. Note that the asymptotic properties of the treated case are deferred to Section 7 for the ease of reading.

We consider the following data generating process. 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

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

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

We consider two counterfactual scenarios. The first one is the transformed case where $X^* = (X_1^*, X_2^*, X_3^*) = 0.75X$. The second corresponds to the independent case where each

$f_{X^*}(x)$ and $f_X(x)$ and hence meet Assumption 4.1(i).

element of X^* follows an i.i.d. standard exponential distribution truncated at 1.5. In each scenario, the counterfactual potential outcomes and treatment assignment are given by

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

We let status quo and counterfactual sample sizes n and n^* vary from 100, 200, and 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 for y. For the quantile indices, 100 equidistant grid points in $[0.1, 0.9]$ are considered.

In the estimation of QCTE and QCTT, we closely follow Rothe (2010) to construct higher-order boundary kernel $\mathcal{K}_x(u)$ that satisfies Assumption 3.5. More specifically, we let $\mathcal{K}_x(u) = \prod_{s=1}^k e_1^T S_{x_s}^{-1}(1, u_s, \ldots, u_s^p)^T \mathcal{K}(u_s)$, where $S_x = (\mu_{j+\ell,x})_{0\leq j,\ell\leq p}$ is a matrix of boundary kernel constants $\mu_{j,x} = \int_{\mathcal{D}_x} u^j \mathcal{K}(u)\, \mathrm{d}u$, $e_1 = (1, 0, \ldots, 0)^T$ is the unit vector, $p = r - k$ is the polynomial order, and $\mathcal{K}(u)$ is a standard univariate kernel (see Fan and Gijbels (1996) for more details). We let $\mathcal{K}(u)$ be the Epanechnikov kernel and choose $p = 1$ so that $\mathcal{K}_x(u)$ 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 $n_d = \sum_{i=1}^n \mathbb{1}\{D_i = d\}$ is the effective sample size. Furthermore, we use standard normal multipliers in both transformed and independent cases to simulate the critical values. The covariance functions are estimated using second-order boundary kernels for $\mathcal{K}_x(\cdot)$ and $\mathcal{W}_y(\cdot)$ with corresponding rule-of-thumb bandwidths $h_d = 2.12 s_X n_d^{-1/7}$ and $\eta = 2.34 s_Y n^{-1/5}$. The nominal coverage rate is 90%.

The simulation results are presented in Table I for the transformed case and Table II for the independent case. 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 Tables I and II, the finite sample performance of the proposed estimators is satisfactory in 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 convergence is indeed at the $\sqrt{n}$ rate. The empirical coverage rates of the uniform confidence bands are also close to 90% in both transformed and independent cases, thus validating the multiplier bootstrap method even when the sample size is relatively small.

Table I. Simulation Results (Transformed Case)

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 the critical values. The covariance functions are estimated using second-order boundary Epanechnikov kernels with rule-of-thumb bandwidths in both x and y directions.

Table II. Simulation Results (Independent Case)

n	n^*	QCTE			QCTT		
		IBias	RIMSE	Coverage Rate	IBias	RIMSE	Coverage 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 the critical values. The covariance functions are estimated using second-order boundary Epanechnikov kernels with rule-of-thumb bandwidths in both x and y directions.

6 Empirical Study

We apply our methods to investigate possible explanations for the heterogeneous effect of Job Corps on the earnings distribution. Job Corps is the largest and most comprehensive labor market program in the United States. It serves economically disadvantaged youths aged 16–24 through the delivery of academic, vocational, and social training, as well as residential living, health care, counseling, and job placement assistance. Established in 1964, it has trained approximately 60,000 youths annually through 120 centers nationwide. However, this comprehensiveness comes at a cost. In 2008, the total expenditure of Job Corps exceeded $1.58 billion and the cost per participant was around $26,000, roughly ten times the cost of a typical publicly-funded training program (Job Corps Annual Report, 2008). Despite its costliness, Job Corps is not uniformly effective. Specifically, Eren and Ozbeklik (2014) document a great deal of heterogeneity in the treatment effect and find that participants toward the bottom of the earnings/skill distribution benefit very little if at all. They propose two possible explanations for this finding.

The first explanation hinges on the U.S. economic boom in the late 1990s, the era of the Job Corps data used by Eren and Ozbeklik (2014). With inflation and unemployment rates both reaching their lowest since the 1960s, this economic boom generated large improvements for low-income individuals, including those who did not participate in any labor market program. These favorable conditions may have "washed out" the program effect. The second explanation states that Job Corps may only work well for participants with a relatively high initial skill or education level, but not for those at the bottom of the skill distribution.

We test these hypotheses by the following counterfactual exercises. First, under the strong economic conditions hypothesis, similar patterns should be observed for other U.S. training programs evaluated at the same time as Job Corps. Unfortunately, such information is not available, and we therefore propose an alternative approach based on an earlier training program, the Job Training Partnership Act (JTPA), with data collected in the early 1990s. Specifically, we estimate the counterfactual effect of Job Corps that would have prevailed had the JTPA individuals participated in the Job Corps program. If the strong economic conditions hypothesis is indeed true, then we still do not expect to see significant treatment effects at lower quantiles. Second, we reverse the roles of the two datasets and estimate the counterfactual effect of JTPA for Job Corps participants. As the JTPA outcomes stem from a recessionary period (the early 1990s), the same hypothesis implies that Job Corps individuals should also be more likely to benefit from receiving JTPA services.

Note that Assumption 2.2(i) is automatically satisfied in these exercises as we intend to reproduce the impact of Job Corps/JTPA for JTPA/Job Corps participants with all other things being equal. Moreover, the two sets of participants are plausibly treated as independent due to the different evaluation periods. In particular, by the end of the 1990s even the youngest JTPA applicants became ineligible for Job Corps due to the upper age limit in that program.

We finally examine the skill hypothesis by artificially increasing the level of education for those who do not benefit from Job Corps and then reestimate the program effect holding other factors constant. If the skill hypothesis is true, then one would expect to see an improved treatment effect estimate. This counterfactual exercise corresponds to the transformed case since the new education level is a deterministic transformation of the original one.

6.1 Data and Unconfoundedness Test

We use the same dataset as Eren and Ozbeklik (2014), which is extracted from the National Job Corps Study, an experimental evaluation of Job Corps undertaken between 1994 and 1996.[7] The design of this experiment is described in detail by Schochet, Burghardt, and McConnell (2008). We focus on the sample of youths who completed the 48-month interview. In this sample, a randomly assigned group of 6,828 individuals was offered training, while the other 4,485 individuals were excluded from receiving Job Corps services for 3 years. However, eligible members were allowed to refuse the offer and only 72% actually chose to receive Job Corps services. In other words, program participation was still subject to self-selection and hence likely correlated with potential outcomes in the absence of further controls. Below we will formally test the unconfoundedness assumption given a set of pre-treatment covariates and using randomized eligibility as an instrument.

The JTPA data come from the National JTPA Study, which is another randomized experiment conducted between 1987 and 1989; see Bloom et al. (1997) for a detailed description.[8] In the research sample, randomly selected applicants were offered job-related services (7,487 individuals) or were excluded from the program for 18 months (3,717 individuals). The participation rate of those offered services was even lower at about 64%. Again, self-selection is a concern, and we formally test the unconfoundedness assumption below.

Table III reports the summary statistics for both samples (we drop all observations with any missing values and treat males and females separately throughout). For Job Corps the outcome variable (Y) is average weekly earnings in the second year of the post-program period (1998). For JTPA, the recorded outcome is the sum of 30-month post-program earnings. We divide this variable by 130 weeks (to obtain average weekly earnings) and then convert to 1998 dollars using CPI-U. The treatment variable (D) is the actual program participation, and the baseline characteristics (X) include age, a dummy for minority status (black or hispanic), and a dummy for high-school education (including GED holders).

[7] Job Corps data are publicly available at `http://qed.econ.queensu.ca/jae/2014-v29.4/eren-ozbeklik/`.

[8] JTPA data are publicly available at `http://fmwww.bc.edu/repec/bocode/j/jtpa.dta`.

Table III. Summary Statistics

	Job Corps			JTPA		
	Part. [std. dev.]	No [std. dev.]	Diff. (t-stat.)	Part. [std. dev.]	No [std. dev.]	Diff. (t-stat.)
A. Males						
Number of obs.	2,711	3,466		2,005	2,772	
Outcome variable						
Average weekly earnings	241.88	223.85	18.03	217.85	176.18	41.67
	[213.97]	[203.34]	(3.36)	[202.46]	[193.21]	(7.16)
Characteristics						
Age	18.22	18.33	−0.11	32.80	32.94	−0.14
	[2.11]	[2.10]	(−2.03)	[9.66]	[9.27]	(−0.50)
Minority (black or hispanic)	0.70	0.70	0.00	0.36	0.34	0.02
	[0.46]	[0.46]	(−0.05)	[0.48]	[0.47]	(1.28)
Education (high school or GED)	0.20	0.21	−0.01	0.71	0.68	0.03
	[0.40]	[0.40]	(−0.90)	[0.45]	[0.47]	(2.46)
B. Females						
Number of obs.	2,065	2,517		2,557	3,175	
Outcome variable						
Average weekly earnings	170.32	165.02	5.30	143.74	122.28	21.46
	[168.64]	[169.59]	(1.06)	[136.94]	[132.94]	(5.97)
Characteristics						
Age	18.48	18.67	−0.19	32.99	33.41	−0.43
	[2.17]	[2.17]	(−2.99)	[9.63]	[9.78]	(−1.66)
Minority (black or hispanic)	0.81	0.78	0.02	0.38	0.38	0.00
	[0.39]	[0.41]	(2.05)	[0.49]	[0.49]	(−0.19)
Education (high school or GED)	0.26	0.31	−0.04	0.75	0.69	0.06
	[0.44]	[0.46]	(−3.24)	[0.43]	[0.46]	(5.07)

Note: The table reports means and standard deviations (in brackets) for the Job Corps and JTPA samples. The columns showing differences in means (by participation status) report the t-statistic (in parentheses) for the null hypothesis of equality in means. Average weekly earnings of the JTPA sample are calculated by dividing the sum of 30-month earnings by 130 weeks and adjusted to 1998 dollars using CPI-U.

Table IV. Unconfoundedness Test Results

		Specification			
	Obs.	Constant	Linear	Interaction	Quadratic
A. Job Corps					
Males	6,177	0.054	0.118	0.132	0.137
Females	4,582	0.040	0.140	0.163	0.166
B. JTPA					
Male	4,777	0.006	0.009	0.009	0.008
Female	5,732	0.497	0.675	0.618	0.627

Note: The table reports the p-values for the null hypothesis of unconfoundedness using the test proposed by Donald, Hsu, and Lieli (2014a) and randomized eligibility as an instrument. Different specifications indicate the inclusion of constant, linear, interaction, and quadratic terms of the covariates as power series in estimating the instrument propensity score by series logit estimation.

As shown in Table III, there are clear differences in individual characteristics by treatment status. For example, participants in both samples tend to be younger than non-participants, and JTPA participants are significantly more likely to have completed high school for both males and females. Table IV reports the p-values for the null hypothesis of unconfoundedness using the test proposed by Donald, Hsu, and Lieli (2014a) with randomized eligibility as a valid instrument.[9]

For Job Corps, unconfoundedness is rejected at 10% level only when the instrument propensity score is estimated by a constant (i.e., no covariates are used for conditioning). This means that Job Corps participation is not randomly assigned, which is of course not surprising. However, unconfoundedness can no longer be rejected given that the available covariates are adjusted for in various ways. For example, in the quadratic specification the p-values are 0.137 (for males) and 0.166 (for females). In this case the instrument propensity score is estimated by series logit with up to second powers of the terms in X included as controls (interactions too). Given the test results, we consider the Job Corps participation decision to be unconfounded conditional on the available covariates, for males and females alike.

The corresponding results for the JTPA sample are given in the lower panel of Table IV. Similarly to Donald, Hsu, and Lieli (2014a), the test does not reject unconfoundedness for females, but it does for males even when the covariates are included in a flexible way. Note, however, that despite the failure of the unconfoundedness for JTPA males, one can still estimate the QCTE of Job Corps for this group, as in this case the conditional distributions are estimated from Job Corps data only.

[9]Donald, Hsu, and Lieli (2014a) propose a Durbin-Wu-Hausman-type test for unconfoundedness by exploiting the fact that the local average treatment effect for the treated (LATT) will coincide with the average treatment effect for the treated (ATT) when the instrument satisfies one-sided noncompliance.

Before proceeding to the main results, it is necessary to mention that we treat the age variable as continuous and specify the bandwidth as $h_d = sn_d^{-1/3}$, where s is the standard deviation of age, and n_d is the sample size for $d = 0, 1$. Other specifications are similar to those in the simulation study. In addition, several sensitivity analyses are conducted in Section C for the robustness of our results.

6.2 Actual and Counterfactual Treatment Effect Estimates

We first present the QTE of Job Corps on the earnings distribution in Figure 2. These results extend the local quantile treatment effect (LQTE) estimates reported by Eren and Ozbeklik (2014) from the complier subpopulation to the whole population given unconfoundedness. The solid line is the QTE estimate with light and dark shaded areas representing 90% uniform and pointwise confidence bands, respectively, and the horizontal dashed line is the average effect (ATE).

Focusing on males (see Figure 2(a)), the zero QTE at lower quantiles reflects the fact that about 15% of individuals in the sample remain unemployed two years after the program. After that point, the estimate suddenly jumps to \$20 and is statistically significant around the 20th quantile, indicating a positive program effect on the extensive margin (i.e., on the probability of having positive earnings). Nevertheless, the intensive margin effect (the effect on earnings for those who already have positive earnings) is insignificant between the 20th and 40th quantiles according to the uniform confidence band. This finding, as argued earlier, may be attributed to strong economic conditions during the evaluation period, where the labor market favors less-skilled workers (the control group) more. In addition, Figure 2(b) shows a similar pattern for females. The extensive margin effect is again significantly positive at the 20th quantile and is equal to about \$12, while the intensive margin effect is insignificant above the third decile according to the uniform confidence band.

Figure 3 presents the counterfactual treatment effect estimates that are informative about the strong economic conditions hypothesis. In particular, Figure 3(a) presents the counterfactual effect of Job Corps for JTPA males. The solid line is the corresponding QCTE estimate, while, for comparison, the dashed lines represents the actual LQTE of the JTPA program and the dotted line the actual intention-to-treat (ITT) effect.[10] Focusing on the point estimates, it can be seen that QCTE fluctuates around zero and is smaller than LQTE and ITT for almost all quantiles, as predicted by the hypothesis. In other words, Job Corps seems to remain ineffective even for JTPA individuals, which is indirect evidence in support of the strong economic conditions explanation.[11]

The counterfactual effect of JTPA for Job Corps females, depicted in Figure 3(b), further reinforces the argument. The solid line is the corresponding QCTE estimate, while the dashed

[10] As the unconfoundedness assumption is rejected for JTPA males (see Table IV), we can only report LQTE and ITT instead of QTE for this sample. ITT can be regarded as the QTE of program eligibility on the earnings distribution. In addition, we set up a subsample of JTPA with age below 24 years old to meet Assumption 2.2(ii).

[11] We also perform the same analysis for females. The result is similar and is omitted due to space considerations.

line is the actual QTE of the Job Corps program for females.[12] As can be seen, QCTE is significant between the 30th and 50th quantiles and exceeds the actual QTE of Job Corps between the 40th and 60th quantiles. This finding lends further support to the strong economic conditions explanation by showing that Job Corps individuals would have benefited from participating in another training program implemented in a different period, especially during an economic downturn. This finding is also consistent with previous studies arguing for a direct relationship between program effectiveness and economic performance (Hotz, Imbens, Klerman, 2006; Lechner and Wunsch, 2009).

Figure 4 presents the results regarding the skill hypothesis. In this counterfactual exercise based on Job Corps, we increase the education level of males (females) with earnings between the 20th and 40th (30th and 50th) quantiles as if all of them had completed high school. Figure 4(a) shows the resulting QCTE for males and Figure 4(b) for females. The actual QTE estimate is shown in both cases as a dashed line, and it is clearly seen that the two estimates are not significantly different over the affected quantiles. This finding disagrees with the skill hypothesis, which predicts that a higher skill or education level should be accompanied by a significant increase in program performance. Nevertheless, it is also possible that the program's ineffectiveness is induced by low non-cognitive skills that cannot be measured by educational attainment (Heckman, Stixrud, and Urzua, 2006).

7 Theoretical Extensions: ACTE and the Counterfactual Treated Subpopulation

In addition to QCTE, here we extend the theoretical analysis to various counterfactual program effects which may also be appealing to practitioners and policymakers. We first discuss how to predict the mean program impact given counterfactual covariates in Section 7.1. The case for counterfactually treated subpopulation is then covered in Section 7.2.

7.1 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)**.**

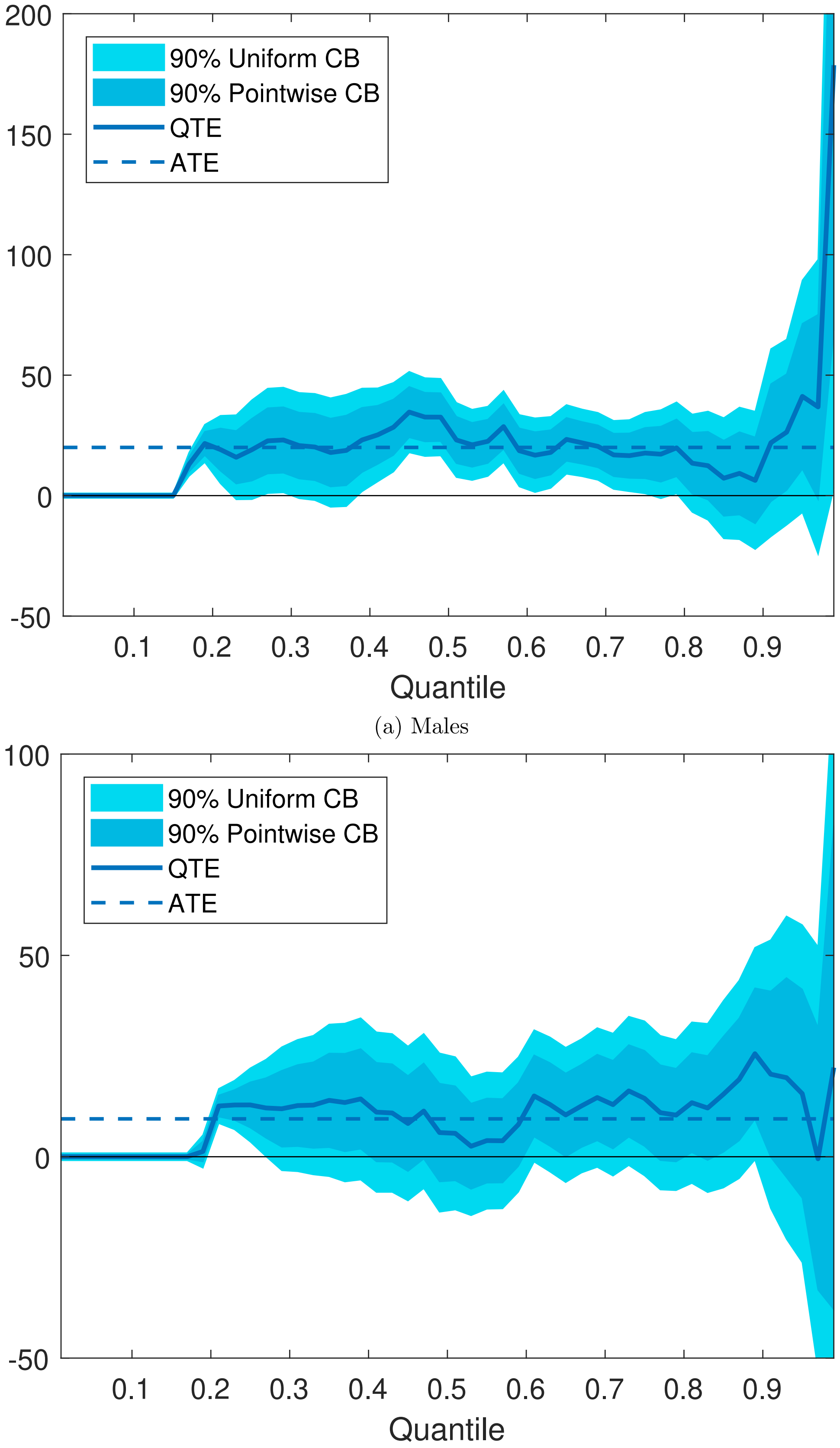
(i) $\mathbb{E}(Y_d|D,X) = \mathbb{E}(Y_d|X)$ *for* $d = 0, 1$.

(ii) $0 < p_\ell \leq p(X) \leq p_u < 1$.

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

(i) $\mathbb{E}(Y_d^*|X^* = x) = \mathbb{E}(Y_d|X = x)$ *for all* $x \in \mathcal{X}^*$, $d = 0, 1$.

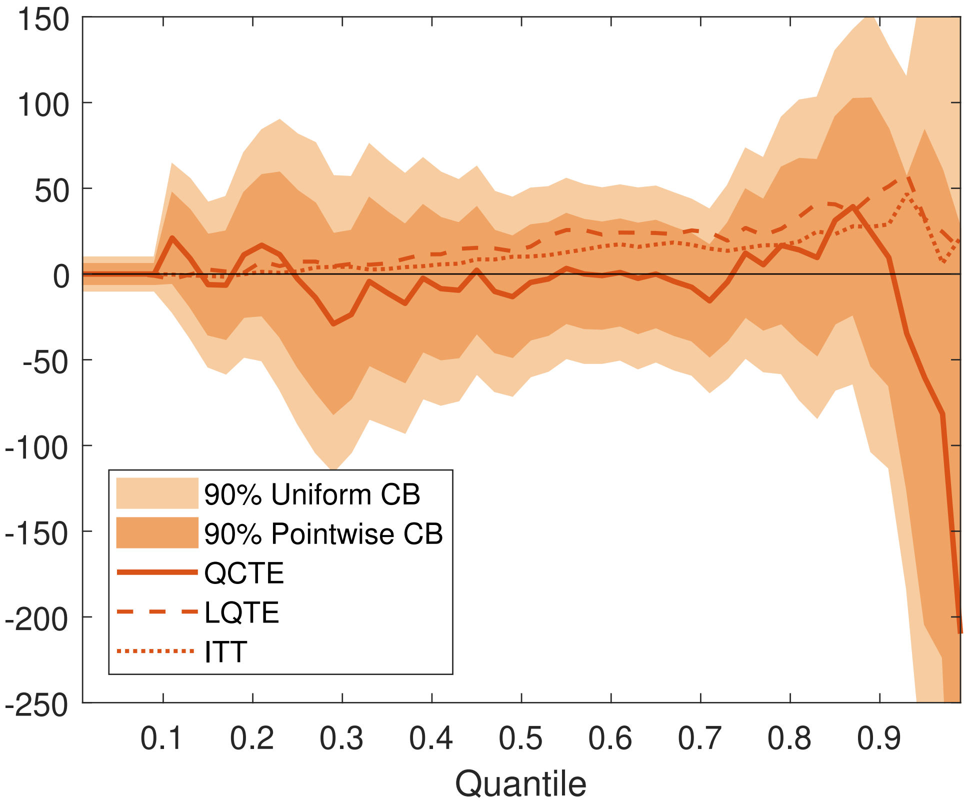
[12] However, the counterfactual effect of JTPA for Job Corps males cannot be consistently estimated since unconfoundedness does not hold for males in the JTPA sample.



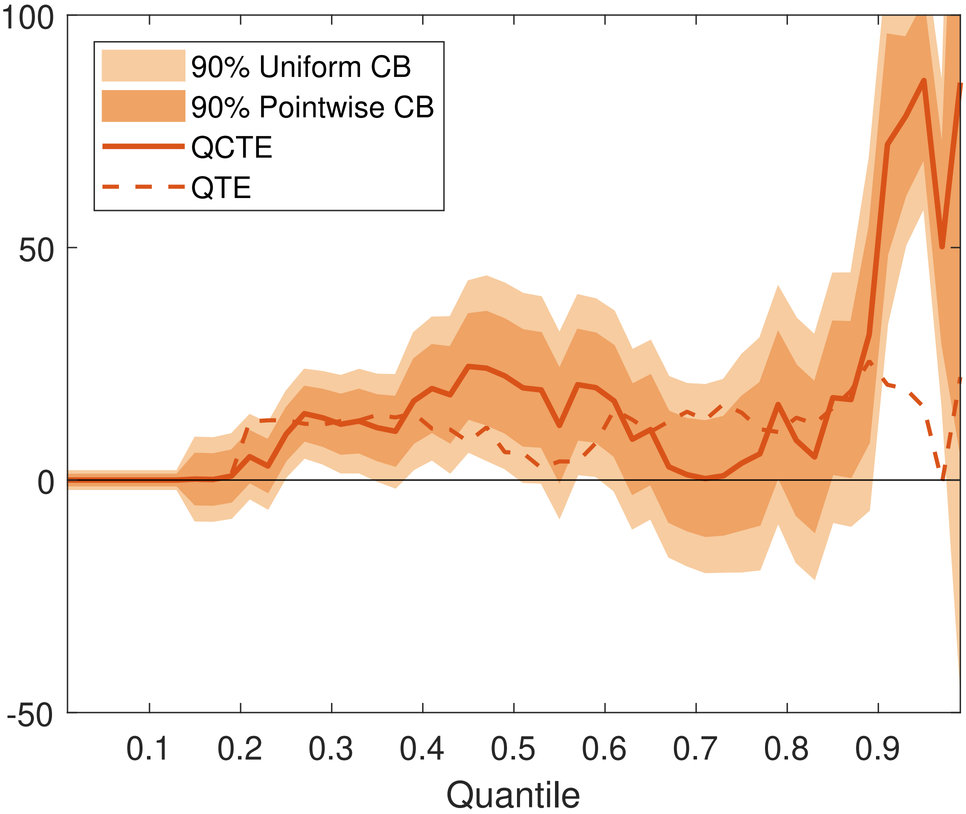
(a) Males

(b) Females

Figure 2. Actual Job Corps QTEs for males (panel (a)) and females (panel (b)), with 90% confidence bands. The horizontal dashed lines are the mean effects.

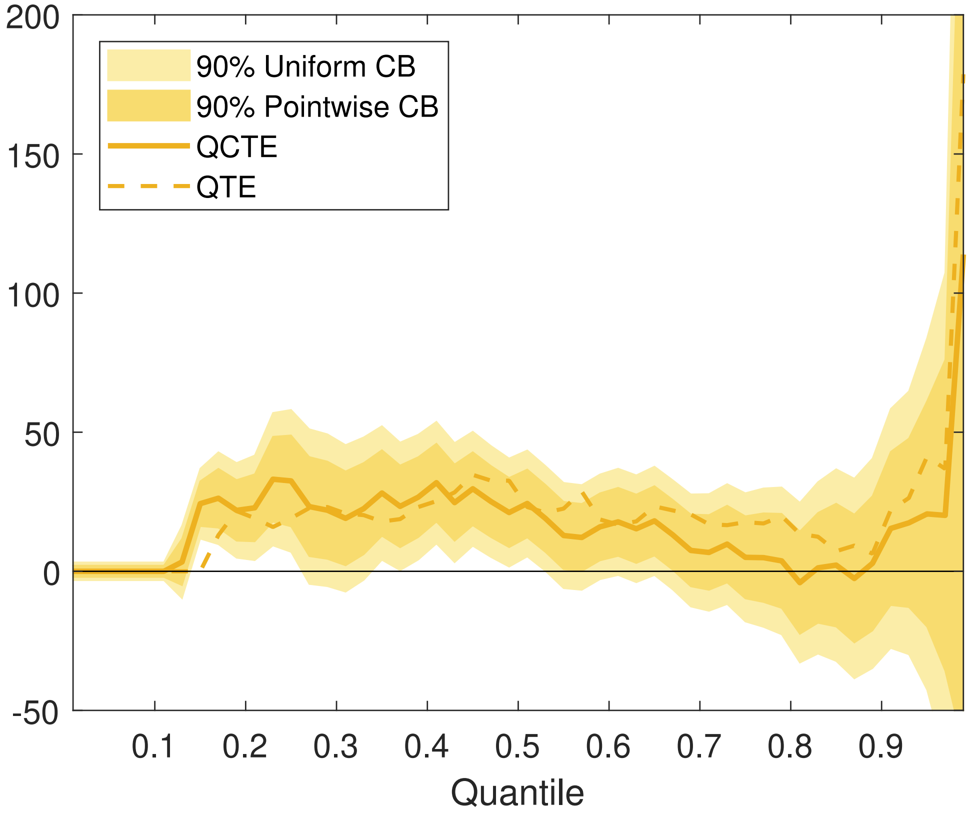


(a) The QCTE of Job Corps for JTPA males.

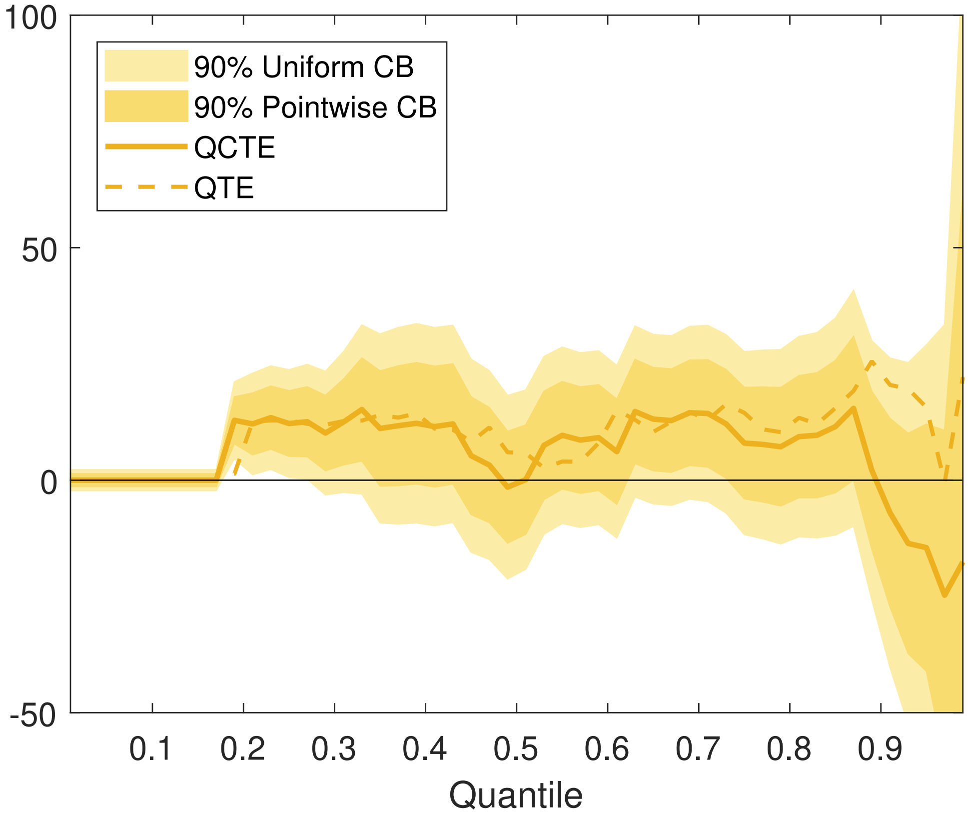


(b) The QCTE of JTPA for Job Corps females.

Figure 3. QCTEs with 90% confidence bands. The dashed and dotted lines in 3(a) are LQTE and ITT for JTPA males, respectively, and the dashed line in 3(b) is QTE for Job Corps females.



(a) Males.



(b) Females.

Figure 4. QCTEs for Job Corps males and females with increased education, with 90% confidence bands. The dashed lines are the original QTEs.

(ii) $\mathcal{X}^* \subseteq \mathcal{X}$.

Similar to Lemma 1, ACTE is identified under Assumptions 7.1 and 7.2 by

$$\delta^* = \mathbb{E}_{X^*}\left[\mathbb{E}(Y|D=1,X) - \mathbb{E}(Y|D=0,X)\right],$$

which can be estimated using the kernel-regression-based estimator proposed by Heckman, Ichimura, and Todd (1998).[13] In other words, we have:

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

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

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

We derive the asymptotic properties for $\widehat{\delta}^*$ under weaker regularity conditions stated below.

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

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

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

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

Corollary 1. *Suppose Assumptions 3.1, 3.2, 3.5, 3.6, and 7.1–7.4 hold. Then,*

$$\sqrt{n}\big(\widehat{\delta}^* - \delta^*\big) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}_{\delta^*}),$$

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

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

Several remarks on Corollary 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 $n^{-1/4}$

[13]One can also adopt Hahn's (1998) series approach for conditional mean $\breve{\mathbb{E}}(Y_d|X=x) = \{\breve{p}(x)^d[1-\breve{p}(x)]^{1-d}\}^{-1}\breve{\mathbb{E}}[\mathbb{1}\{D=d\}Y|X=x]$, where $\breve{p}(x)$ and $\breve{\mathbb{E}}[Y\,\mathbb{1}\{D=d\}|X=x]$ are both obtained by the series estimation. However, estimating $\mathbb{E}(Y_d|X=x)$ directly is sufficient and eliminates the need for estimating propensity score in this step.

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 $\mathbb{V}_{\delta^*}$ even when X and X^* are dependent. Fourth, compared to the semiparametric efficiency bound of the ATE estimator given in Hahn (1998):

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

it is true that $\mathbb{V}_{\delta^*}$ 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, Imbens, and Ridder (2003).

Since the asymptotic variance $\mathbb{V}_{\delta^*}$ can be consistently estimated by plugging $\widehat{p}(x)$, $\widehat{f}_X(x)$, and $\widehat{f}_{X^*}(x)$ in Section 4.4 into $\varrho_{\delta^*}(x)$, one can apply a standard normal approximation to test whether there is a counterfactual mean effect $H_0:\delta^*=0$. We omit the details for brevity. On the other hand, bootstrap methods may be valuable alternatives to conduct inference on ACTE.

7.2 The Treated Cases

We focus on ACTT and QCTT defined in (2.3) in this section. Note that since the counterfactual treatment assignment D^* is not observable in our framework, a different set of assumptions is introduced to identify the parameters of interest. Let $p^*(x)=\Pr(D^*=1|X^*=x)$ be the counterfactual propensity score for all $x\in\mathcal{X}^*$.

Assumption 7.5 (Unconfoundedness for the Untreated)**.**

(i) $Y_0 \perp\!\!\!\perp D|X$.

(ii) $p(X)>0$.

Assumption 7.6 (Invariance of Conditional Distributions for the Treated)**.**

(i) $F_{Y_d^*|X^*,D^*}(y|x,1)=F_{Y_d|X,D}(y|x,1)$ *for all* $x\in\mathcal{X}^*$, $d=0,1$.

(ii) $\mathcal{X}^*\subseteq\mathcal{X}$.

Assumption 7.7 (Invariance of Propensity Scores)**.** $p^*(x)=p(x)$ *for all* $x\in\mathcal{X}^*$.

It can easily be seen that Assumptions 7.5 and 7.6 are weaker than their counterparts Assumptions 2.1 and 2.2.[14] To identify treated parameters, however, we need to invoke Assumption 7.7 so that the counterfactual treatment assignment can be determined. It requires that the probabilities of receiving treatment must be the same for individuals who are observationally equivalent between counterfactual and status quo populations. Given these assumptions, ACTT and QCTT can be identified as follows.

[14]They can be further weakened for the ACTT case similar to Section 7.1.

Lemma 4. *Suppose Assumptions 7.5–7.7 hold. ACTT and QCTT are identified by*

$$\begin{aligned}
\delta_t^* =& \int_{\mathcal{X}} \frac{p(x)}{\mathbb{E}\left[p(X^*)\right]} \left\{\mathbb{E}(Y|X=x,D=1) - \mathbb{E}(Y|X=x,D=0)\right\} \, dF_{X^*}(x), \\
\delta_t^*(\tau) =& \inf\left\{y \in \mathcal{Y} : \int_{\mathcal{X}} \frac{p(x)}{\mathbb{E}\left[p(X^*)\right]} F_{Y|X,D}(y|x,1)\, dF_{X^*}(x) \geq \tau\right\} \\
& - \inf\left\{y \in \mathcal{Y} : \int_{\mathcal{X}} \frac{p(x)}{\mathbb{E}\left[p(X^*)\right]} F_{Y|X,D}(y|x,0)\, dF_{X^*}(x) \geq \tau\right\}.
\end{aligned}$$

According to Lemma 4, ACTT and QCTT estimators are given by

$$\begin{aligned}
\widehat{\delta}_t^* &= \sum_{j=1}^{n^*} \widehat{p}(X_j^*) \left[\widehat{\mathbb{E}}(Y_1|X=X_j^*) - \widehat{\mathbb{E}}(Y_0|X=X_j^*)\right] \Big/ \sum_{j=1}^{n^*} \widehat{p}(X_j^*), \\
\widehat{\delta}_t^*(\tau) &= \widehat{\mathbb{Q}}_{Y_1^*|D^*}(\tau|1) - \widehat{\mathbb{Q}}_{Y_0^*|D^*}(\tau|1),
\end{aligned}$$

where $\widehat{\mathbb{Q}}_{Y_d^*|D^*}(\tau|1) = \inf\{y \in \mathcal{Y} : \widehat{F}_{Y_d^*|D^*}(y|1) \geq \tau\}$ and

$$\widehat{F}_{Y_d^*|D^*}(y|1) = \sum_{j=1}^{n^*} \widehat{p}(X_j^*) \widehat{F}_{Y_d|X}(y|X_j^*) \Big/ \sum_{j=1}^{n^*} \widehat{p}(X_j^*).$$

Similarly, the asymptotic properties of the ACTT and QCTT estimators can be derived under a modified version of Assumption 3.3:

Assumption 7.8 (Distribution of Y_d^* for the Treated)**.**

(i) $F_{Y_d^*|D^*}(y|1)$ *has a compact support* $[y_{d\ell}^*, y_{du}^*] \subseteq \mathcal{Y}$.

(ii) $F_{Y_d^*|D^*}(y|1)$ *is continuous on* $\mathcal{Y}$.

(iii) $f_{Y_d^*|D^*}(y|1)$ *is bounded away from 0 and is two-times differentiable on* $\mathcal{Y}$.

Corollary 2. *Suppose Assumptions 3.1, 3.2, 3.4–3.6, and 7.5–7.8 hold. We then have:*

$$\sqrt{n}\big(\widehat{\delta}_t^* - \delta_t^*\big) \xrightarrow{d} \mathcal{N}\big(0, \mathbb{V}_{\delta_t^*}\big),$$

where $\mathbb{V}_{\delta_t^*} = \mathbb{E}\big[\varrho_{\delta_t^*}(Z)^2\big] + \mathbb{E}\big[\varphi_{\delta_t^*}(X^*)^2\big]$ *with*

$$\begin{aligned}
\varrho_{\delta_t^*}(Z) &= \frac{p(X)}{\mathbb{E}\left[p(X^*)\right]} \left\{\frac{D\big[Y - \mathbb{E}(Y_1|X)\big]}{p(X)} - \frac{(1-D)\big[Y - \mathbb{E}(Y_0|X)\big]}{1-p(X)}\right\} \frac{f_{X^*}(X)}{f_X(X)}, \\
\varphi_{\delta_t^*}(X^*) &= \sqrt{\lambda}\frac{p(X^*)}{\mathbb{E}\left[p(X^*)\right]} \big[\,\mathbb{E}(Y_1|X^*) - \mathbb{E}(Y_0|X^*) - \delta_t^*\big].
\end{aligned}$$

Moreover, we present:

$$\sqrt{n}\left(\widehat{\delta}_t^*(\cdot) - \delta_t^*(\cdot)\right) \Rightarrow \Delta_t(\cdot),$$

where $\Delta_t(\tau)$ is a Gaussian process with mean zero and covariance function $\Psi_t(\tau) = \mathbb{E}\big[\varrho_t(\tau, Z)^2\big] + \mathbb{E}\big[\varphi_t(\tau, X^*)^2\big]$ with

$$\varrho_t(\tau, Z) = -\left[\frac{\varrho^F_{1,t}\big(\mathbb{Q}_{Y_1^*|D^*}(\tau|1), Z\big)}{f_{Y_1^*|D^*}\big(\mathbb{Q}_{Y_1^*|D^*}(\tau|1)\big|1\big)} - \frac{\varrho^F_{0,t}\big(\mathbb{Q}_{Y_0^*|D^*}(\tau|1), Z\big)}{f_{Y_0^*|D^*}\big(\mathbb{Q}_{Y_0^*|D^*}(\tau|1)\big|1\big)}\right],$$
$$\varphi_t(\tau, X^*) = -\left[\frac{\varphi^F_{1,t}\big(\mathbb{Q}_{Y_1^*|D^*}(\tau|1), X^*\big)}{f_{Y_1^*|D^*}\big(\mathbb{Q}_{Y_1^*|D^*}(\tau|1)\big|1\big)} - \frac{\varphi^F_{0,t}\big(\mathbb{Q}_{Y_0^*|D^*}(\tau|1), X^*\big)}{f_{Y_0^*|D^*}\big(\mathbb{Q}_{Y_0^*|D^*}(\tau|1)\big|1\big)}\right],$$

where $\varrho^F_{d,t}(y, Z)$ and $\varphi^F_{d,t}(y, X^*)$ are given by

$$\varrho^F_{d,t}(y, Z) = \frac{p(X)}{\mathbb{E}\left[p(X^*)\right]} \frac{\mathbb{1}\{D = d\}\left[\mathbb{1}\{Y \leq y\} - F_{Y_d|X}(y|X)\right]}{p(X)^d\left[1 - p(X)\right]^{1-d}} \frac{f_{X^*}(X)}{f_X(X)},$$
$$\varphi^F_{d,t}(y, X^*) = \sqrt{\lambda}\frac{p(X^*)}{\mathbb{E}\left[p(X^*)\right]}\left[F_{Y_d|X}(y|X^*) - F_{Y_d^*}(y)\right],$$

and the convergence is in $\ell^\infty([0,1])$.

We also construct simulated process to approximate $\Delta_t(\cdot)$ similar to (4.4). That is:

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

where $\widehat{\varrho}_t$ and $\widehat{\varphi}_t$ can be estimated given $\widehat{f}_{Y_d^*|D^*}(y|1) = \max\{\widetilde{f}_{Y_d^*|D^*}(y|1), b_n\}$ with

$$\widetilde{f}_{Y_d^*|D^*}(y|1) = \sum_{j=1}^{n^*}\widehat{p}(X_j^*)\widetilde{f}_{Y_d|X}(y|X_j^*)\Big/\sum_{j=1}^{n^*}\widehat{p}(X_j^*),$$

where $\widehat{p}(x)$ and $\widetilde{f}_{Y_d|X}(y|x)$ are given in (4.9) and (4.11). One can show $\Delta^u_t(\cdot) \overset{p}{\Rightarrow} \Delta_t(\cdot)$ similar to Theorem 2. We omit the details for brevity.

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.

We then apply the proposed methods to study the heterogeneous impact of the Job Corps training program in the U.S. along the earnings distribution. Specifically, the literature has documented the ineffectiveness of the program for individuals with low earnings and put forth two explanations for this finding: (i) strong economic conditions during the evaluation period; (ii) insufficient skills to benefit from the program for individuals toward the bottom of the earnings distribution. We empirically examine these hypotheses and demonstrate three findings: (i) Job Corps would remain ineffective during the given evaluation period even if it were implemented for another population targeted by an earlier, more successful, program; (ii) the program effect for Job Corps individuals would become significantly larger had they participated in another training program in a different period of time; (iii) the effectiveness of Job Corps would not improve if individuals with low earnings were given extra education. Taken together, these findings suggest that only the strong economic conditions explanation contributes to the documented weak performance of Job Corps for individuals with low earnings, while no evidence can be found to support the skill hypothesis.

APPENDIX

A Proofs

Proof of Lemma 1:

By the law of iterated expectations, Assumption 2.2, Assumption 2.1(i), and $Y = Y_d$ for $D = d$,

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

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 2.2(ii), $F_{Y_d^*}(y)$ is identified. The quantile functions and the QCTE can be identified accordingly. $\square$

Proof of Lemma 2:

The proof consists of two parts. First, we show that $\sqrt{n}\big(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\big)$ is asymptotically linear in the following influence function representation:

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

Since $\varrho_d^F(y,\cdot)$ and $\varphi_d^F(y,\cdot)$ belong to Donsker classes for all $y \in \mathcal{Y}$ and the Cartesian product of two Donsker classes of functions is still a Donsker class as in van der Vaart (2000), Lemma 2 holds by the functional central limit theorem for $\widetilde{\boldsymbol{F}} = \big(\widetilde{F}_{Y_0^*}, \widetilde{F}_{Y_1^*}\big)^T$ in place of $\widehat{\boldsymbol{F}}$. Second, we complete the proof by establishing the first-order asymptotic equivalence between $\widehat{F}_{Y_d^*}(y)$ and $\widetilde{F}_{Y_d^*}(y)$. That is,

$$\sup_{y\in\mathcal{Y}} \big|\widehat{F}_{Y_d^*}(y) - \widetilde{F}_{Y_d^*}(y)\big| = o_p(n^{-1/2}). \qquad \text{(A.2)}$$

The derivation of (A.1) is similar to Theorem 1 of Rothe (2010). For simplicity, we assume $n^* = n$ so that $\lambda = 1$. Let P and P^* be the distribution function of X and X^*, respectively. Denote $\mathcal{G}_n = \sqrt{n}(\mathcal{P}_n - \mathcal{P})$, where $\mathcal{P}$ is the expectation under P and $\mathcal{P}_n$ is the empirical distribution under P such that for every measurable function $\phi : \mathcal{X} \to \mathbb{R}$, $\mathcal{P}\,\phi = \int \phi\,\mathrm{d}P$ and $\mathcal{P}_n\,\phi = n^{-1}\sum_{i=1}^{n}\phi(X_i)$. Define $\mathcal{G}_n^*$, $\mathcal{P}^*$ and $\mathcal{P}_n^*$ similarly under P^*.

To begin with, we rewrite $\sqrt{n}\big(\widehat{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\big)$ as

$$\sqrt{n}\left(\widehat{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\right) = \mathcal{G}_n^*\left(\widehat{F}_{Y_d|X}(y|x) - F_{Y_d|X}(y|x)\right) \tag{A.3}$$

$$+\sqrt{n}\,\mathcal{P}^*\left(\widehat{F}_{Y_d|X}(y|x) - F_{Y_d|X}(y|x)\right) \tag{A.4}$$

$$+\frac{1}{\sqrt{n}}\sum_{j=1}^{n^*}\big(F_{Y_d|X}(y|X_j^*) - F_{Y_d^*}(y)\big)\,. \tag{A.5}$$

It is true that (A.3) is $o_p(1)$ uniformly over $y \in \mathcal{Y}$ by Lemma 1 of Rothe (2010) and Lemma 19.24 of van der Vaart (2000). Next, we show that uniformly over $y \in \mathcal{Y}$,

$$(\text{A.4}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathbb{1}\{D_i = d\}\left[\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right]}{p(X_i)^d\left[1-p(X_i)\right]^{1-d}}\frac{f_{X^*}(X_i)}{f_X(X_i)} + o_p(1).$$

Define $g_d(y,x) = \mathbb{E}[\mathbb{1}\{Y \le y\}\,\mathbb{1}\{D = d\}|X = x]f_X(x)$, $g_d(x) = \mathbb{E}(\mathbb{1}\{D = d\}|X = x)f_X(x)$, $\widehat{g}_d(y,x) = n^{-1}\sum_{i=1}^n \mathbb{1}\{Y_i \le y\}\,\mathbb{1}\{D_i = d\}\,\mathcal{K}_{x,h}(X_i - x)$, and $\widehat{g}_d(x) = n^{-1}\sum_{i=1}^n \mathbb{1}\{D_i = d\}\,\mathcal{K}_{x,h}(X_i - x)$. Since $\widehat{F}_{Y_d|X}(y|x) = \widehat{g}_d(y,x)/\widehat{g}_d(x)$,

$$\mathcal{P}^*\left(\widehat{F}_{Y_d|X}(y|x) - F_{Y_d|X}(y|x)\right)$$

$$= \int \frac{1}{n}\sum_{i=1}^{n}\big(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\big)\,\frac{\mathbb{1}\{D_i = d\}\,\mathcal{K}_{x,h}(X_i - x)}{\widehat{g}_d(x)} f_{X^*}(x)\,\mathrm{d}x \tag{A.6}$$

$$+ \int \frac{1}{n}\sum_{i=1}^{n}\big(F_{Y_d|X}(y|X_i) - F_{Y_d|X}(y|x)\big)\,\frac{\mathbb{1}\{D_i = d\}\,\mathcal{K}_{x,h}(X_i - x)}{\widehat{g}_d(x)} f_{X^*}(x)\,\mathrm{d}x. \tag{A.7}$$

Apply a second-order Taylor expansion of $\widehat{g}_d(x)$ around $g_d(x)$ in (A.6),

$$(\text{A.6})$$

$$= \int \frac{1}{n}\sum_{i=1}^{n}\big(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\big)\,\frac{\mathbb{1}\{D_i = d\}\,\mathcal{K}_{x,h}(x - X_i)}{g_d(x)} f_{X^*}(x)\,\mathrm{d}x \tag{A.8}$$

$$- \int \frac{1}{n}\sum_{i=1}^{n}\big(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\big)\,\frac{\mathbb{1}\{D_i = d\}\,\mathcal{K}_{x,h}(x - X_i)}{g_d^2(x)}\big(\widehat{g}_d(x) - g_d(x)\big) f_{X^*}(x)\,\mathrm{d}x \tag{A.9}$$

$$+ o_p(n^{-1/2}), \tag{A.10}$$

where (A.10) is $o_p(n^{-1/2})$ uniformly in both x and y since $\|\widehat{g}_d(x) - g_d(x)\|_\infty = O_p\big((\log n/nh^k)^{1/2} + h^r\big) = o_p(n^{-1/4})$ by Assumption 3.6 and Lemma B.3 of Newey (1994), $|\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)| \le 1$, and the dominated convergence theorem.

For (A.8), we let $\zeta_d(x) = f_{X^*}(x)/g_d(x)$ which is r-times differentiable under Assumptions 3.2 and 3.4. Denote $\mathcal{K}_x^{(\gamma)}(u) = \partial^{|\gamma|}/(\partial^{\gamma_1}u_1,\ldots,\partial^{\gamma_k}u_k)\,\mathcal{K}_x(u)$ and $\zeta_d^{(\gamma)}(u) = \partial^{|\gamma|}/(\partial^{\gamma_1}u_1,\ldots,\partial^{\gamma_k}u_k)\zeta_d(u)$. By a standard change of variables $x = uh + X_i$ and a rth order Taylor expansion of $\mathcal{K}_{uh+X_i}(u)$ and $\zeta_d(uh + X_i)$

around $\mathcal{K}_{X_i}(u)$ and $\zeta_d(X_i)$, we then have uniformly over $y \in \mathcal{Y}$,

$$
\begin{aligned}
\text{(A.8)} &= \int \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right)\mathbb{1}\{D_i = d\}\,\mathcal{K}_{x,h}(x - X_i)\zeta_d(x)\,\mathrm{d}x \\
&= \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right)\mathbb{1}\{D_i = d\}\int \mathcal{K}_{uh+X_i}(u)\zeta_d(uh + X_i)\,\mathrm{d}u \\
&= \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right)\mathbb{1}\{D_i = d\}\int \Big(\mathcal{K}_{X_i}(u) + \cdots + (uh)^r\,\mathcal{K}_{\chi}^{(r)}(u)\Big) \\
&\qquad\qquad \times \Big(\zeta_d(X_i) + \cdots + (uh)^r\zeta_d^{(r)}(\chi)\Big)\,\mathrm{d}u \\
&= \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right)\left(\mathbb{1}\{D_i = d\}\zeta_d(X_i) + O_p(h^r)\right) \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{1}\{D_i = d\}}{p(X_i)^d[1 - p(X_i)]^{1-d}}\left(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right)\frac{f_{X^*}(X_i)}{f_X(X_i)} + o_p(n^{-1/2}),
\end{aligned}
$$

where χ is some value between $uh + X_i$ and X_i. The fourth equality follows from interchanging the differentiation and integration (which is true by the dominated convergence theorem) and Assumption 3.5. The last equality holds because $g_d(x) = p(x)^d[1-p(x)]^{1-d}f_X(x)$ and $O_p(h^r) = o_p(n^{-1/2})$ under Assumption 3.6.

Equation (A.9) can be derived in a similar manner. Let $\xi_d(x) = f_{X^*}(x)/g_d^2(x)$, which is r-times differentiable. By the definition of $\widehat{g}_d(x)$,

$$
\begin{aligned}
&\text{(A.9)} \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{1}\{D_i = d\}\left(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right) \\
&\qquad\qquad \times \int \left(\mathbb{1}\{D_j = d\}\,\mathcal{K}_{x,h}(X_j - x) - g_d(x)\right)\mathcal{K}_{x,h}(x - X_i)\xi_d(x)\,\mathrm{d}x \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{1}\{D_i = d\}\left(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right) \\
&\qquad \times \left\{\int \mathbb{1}\{D_j = d\}\,\mathcal{K}_{x,h}(X_j - x)\,\mathcal{K}_{x,h}(x - X_i)\xi_d(x)\,\mathrm{d}x - \int \mathcal{K}_{x,h}(x - X_i)\zeta_d(x)\,\mathrm{d}x\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{1}\{D_i = d\}\left(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right) \\
&\qquad \times \left\{\Big(\mathbb{1}\{D_j = d\}\,\mathcal{K}_{X_i,h}(X_j - X_i)\xi_d(X_i) + o_p(n^{-1/2})\Big) - \Big(\zeta_d(X_i) + o_p(n^{-1/2})\Big)\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{1}\{D_i = d\}\left(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right)\xi_d(X_i) \\
&\qquad\qquad \times \left\{\mathbb{1}\{D_j = d\}\,\mathcal{K}_{X_i,h}(X_j - X_i) - g(X_i)\right\} + o_p(n^{-1/2}) \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{1}\{D_i = d\}\left(\mathbb{1}\{Y_i \le y\} - F_{Y_d|X}(y|X_i)\right)\xi_d(X_i) \\
&\qquad \times \left\{\mathbb{1}\{D_j = d\}\,\mathcal{K}_{X_i,h}(X_j - X_i) - \mathbb{E}\left[\mathbb{1}\{D_j = d\}\,\mathcal{K}_{X_i,h}(X_j - X_i)\right]\right\} + o_p(n^{-1/2}),
\end{aligned}
\tag{A.11}
$$

where the last equality holds because $\mathbb{E}[\mathbb{1}\{D_j = d\}K_{x,h}(X_j - x)] - g(x) = O_p(h^r) = o_p(n^{-1/2})$ uniformly in x by Lemma B.2 of Newey (1994). As pointed out by Rothe (2010), the leading term in (A.11) is a degenerate second-order U-process. We therefore apply the uniform law of large numbers for U-processes

(Nolan and Pollard, 1987; Sherman, 1994) to show that (A.11) is $O_p(h^{-k}n^{-1}) + o_p(n^{-1/2}) = o_p(n^{-1/2})$ under Assumption 3.6.

Combining all the results obtained above, (A.6) is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\{D_i = d\}\left[\mathbb{1}\{Y_i \leq y\} - F_{Y_d|X}(y|X_i)\right]}{p(X_i)^d\left[1-p(X_i)\right]^{1-d}}\frac{f_{X^*}(X_i)}{f_X(X_i)} + o_p(n^{-1/2}).$$

One can also show that (A.7) is $o_p(n^{-1/2})$ through similar arguments. As a result, (A.4) is

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathbb{1}\{D_i = d\}\left[\mathbb{1}\{Y_i \leq y\} - F_{Y_d|X}(y|X_i)\right]}{p(X_i)^d[1-p(X_i)]^{1-d}}\frac{f_{X^*}(X_i)}{f_X(X_i)} + o_p(1),$$

and the asymptotic linear representation is as in (A.1). Next, since $\mathbb{1}\{Y \leq y\}$ is a type I function and the other functions in (A.1) are type II functions defined in Andrews (1994), $\varrho_d^F(y,\cdot)$ and $\varphi_d^F(y,\cdot)$ belong to some Donsker classes for all $y \in \mathcal{Y}$. By van der Vaart (2000, p.270) in which the Cartesian product of two Donsker classes of functions is still a Donsker class, Lemma 2 holds by the functional central limit theorem for $\widetilde{\boldsymbol{F}} = \big(\widetilde{F}_{Y_0^*}, \widetilde{F}_{Y_1^*}\big)^T$ in place of $\widehat{\boldsymbol{F}}$.

We now show the second part of the proof which claims that $\widehat{F}_{Y_d^*}(y)$ and $\widetilde{F}_{Y_d^*}(y)$ are asymptotically equivalent to the first-order approximation, or $\sup_{y\in\mathcal{Y}}\big|\widehat{F}_{Y_d^*}(y) - \widetilde{F}_{Y_d^*}(y)\big| = o_p(n^{-1/2})$ as in (A.2). For simplicity, assume that $\widetilde{F}_{Y_d^*}(0) \geq 0$ so that $\phi_1(\widetilde{F}_{Y_d^*})(y) = \sup_{y'\leq y}\widetilde{F}_{Y_d^*}(y')$ for all $y \in \mathcal{Y} = [0,\bar{y}]$.[15] From the first part of the proof it is true that $\sup_{y\in\mathcal{Y}}\big|\sqrt{n}\big(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\big)\big| = O_p(1)$, implying for any $\epsilon_1 > 0$ there exist $M > 0$ and large $N = N_M$ such that for all $n > N$,

$$\mathbb{P}\left(\sup_{y\in\mathcal{Y}}\left|\sqrt{n}\left(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\right)\right| \leq M\right) \geq 1-\epsilon_1. \tag{A.12}$$

Next, it can also be shown that $\sqrt{n}\big(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\big)$ is stochastic equicontinuous with respect to the pseudometric $\rho(y_1,y_2) = \big|F_{Y_d^*}(y_1) - F_{Y_d^*}(y_2)\big|^{1/2}$ for all $(y_1,y_2) \in \mathcal{Y}$ from Theorem 3.1 of Hsu, Lieli, and Lai (2018), meaning that for any $\epsilon_2 > 0$ and $\epsilon_3 > 0$, there exist $\delta > 0$ small enough and N_δ large enough that for all $n > N_\delta$,

$$\mathbb{P}\left(\sup_{\rho(y_1,y_2)\leq\delta}\left|\sqrt{n}\left(\widetilde{F}_{Y_d^*}(y_1) - F_{Y_d^*}(y_1)\right) - \sqrt{n}\left(\widetilde{F}_{Y_d^*}(y_2) - F_{Y_d^*}(y_2)\right)\right| \leq \epsilon_3\right) \geq 1-\epsilon_2. \tag{A.13}$$

If we pick a large N such that

$$2M/\sqrt{N} < \delta^2, \tag{A.14}$$

for $y_1 \leq y_2$ with $\rho(y_1,y_2) > \delta$ and for $n > N$ with $\sup_{y\in\mathcal{Y}}\big|\sqrt{n}(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y))\big| \leq M$,

$$\begin{aligned}
&\widetilde{F}_{Y_d^*}(y_1) - \widetilde{F}_{Y_d^*}(y_2)\\
=&\left(\widetilde{F}_{Y_d^*}(y_1) - F_{Y_d^*}(y_1)\right) - \left(\widetilde{F}_{Y_d^*}(y_2) - F_{Y_d^*}(y_2)\right) - \left(F_{Y_d^*}(y_2) - F_{Y_d^*}(y_1)\right)\\
\leq& 2M/\sqrt{n} - \delta^2 < 2M/\sqrt{N} - \delta^2 < 0,
\end{aligned}$$

where the inequality holds because $\sqrt{n}\big(\widetilde{F}_{Y_d^*}(y_1) - F_{Y_d^*}(y_1)\big) \leq M$, $\sqrt{n}\big(\widetilde{F}_{Y_d^*}(y_2) - F_{Y_d^*}(y_2)\big) \geq -M$, and

[15] If this is not the case, we can always redefine $\mathcal{Y} = [-\epsilon,\bar{y}]$ such that $\widetilde{F}_{Y_d^*}(-\epsilon) \geq 0$ for $\epsilon > 0$.

$F_{Y_d^*}(y_2) - F_{Y_d^*}(y_1) > \delta^2$ by the definition of $\rho(y_1, y_2)$. This implies that for $n > N$,

$$\phi_1(\widetilde{F}_{Y_d^*})(y) = \sup_{y' \le y} \widetilde{F}_{Y_d^*}(y') = \sup_{\{y' : y' \le y, \rho(y', y) \le \delta\}} \widetilde{F}_{Y_d^*}(y').$$

Consequently, for all $y \in \mathcal{Y}$ and for $n > N$ with $\sup_{y \in \mathcal{Y}} \left|\sqrt{n}(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y))\right| \le M$, we have:

$$\begin{aligned}
0 &\le \sqrt{n}\left(\phi_1(\widetilde{F}_{Y_d^*})(y) - \widetilde{F}_{Y_d^*}(y)\right) \\
&= \sqrt{n}\left(\phi_1(\widetilde{F}_{Y_d^*})(y) - F_{Y_d^*}(y) - \left(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\right)\right) \\
&= \sqrt{n}\left(\sup_{\{y' : y' \le y, \rho(y', y) \le \delta\}} \left(\widetilde{F}_{Y_d^*}(y') - F_{Y_d^*}(y) - \left(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\right)\right)\right) \\
&\le \sup_{\{y' : y' \le y, \rho(y', y) \le \delta\}} \sqrt{n}\left(\widetilde{F}_{Y_d^*}(y') - F_{Y_d^*}(y') - \left(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\right)\right) \\
&\le \sup_{\rho(y_1, y_2) \le \delta} \left|\sqrt{n}\left(\widetilde{F}_{Y_d^*}(y_1) - F_{Y_d^*}(y_1)\right) - \sqrt{n}\left(\widetilde{F}_{Y_d^*}(y_2) - F_{Y_d^*}(y_2)\right)\right|,
\end{aligned} \tag{A.15}$$

where the second inequality holds because $F_{Y_d^*}(y') \le F_{Y_d^*}(y)$ for $y' \le y$. Since it is also true that $\sup_{y \in \mathcal{Y}} \sqrt{n}\big(\phi_1(\widetilde{F}_{Y_d^*})(y) - 1\big) = \sup_{y \in \mathcal{Y}} \sqrt{n}\big(\widetilde{F}_{Y_d^*}(y) - 1\big) = o_p(1)$ by Theorem 3.1 of Hsu, Lieli, and Lai (2018), we have for all $y \in \mathcal{Y}$,

$$\begin{aligned}
&\sqrt{n}\left(\widehat{F}_{Y_d^*}(y) - \widetilde{F}_{Y_d^*}(y)\right) \\
=&\sqrt{n}\left(\frac{\phi_1(\widetilde{F}_{Y_d^*})(y)}{\sup_{y \in \mathcal{Y}} \phi_1(\widetilde{F}_{Y_d^*})(y)} - \widetilde{F}_{Y_d^*}(y)\right) \\
=&\sqrt{n}\left(\phi_1(\widetilde{F}_{Y_d^*})(y) - \widetilde{F}_{Y_d^*}(y)\right) - \phi_1(\widetilde{F}_{Y_d^*})(y)\sqrt{n}\left(\sup_{y \in \mathcal{Y}} \phi_1(\widetilde{F}_{Y_d^*})(y) - 1\right) + o_p(1) \\
=&\sqrt{n}\left(\phi_1(\widetilde{F}_{Y_d^*})(y) - \widetilde{F}_{Y_d^*}(y)\right) + o_p(1),
\end{aligned} \tag{A.16}$$

where the second equality follows from a mean-valued expansion of $\sup_{y \in \mathcal{Y}} \phi_1(\widetilde{F})_{Y_d^*}(y)$ around 1 and the last equality holds because $\phi_1(\widetilde{F}_{Y_d^*})(y) \xrightarrow{p} 1$ and $\sup_{y \in \mathcal{Y}} \sqrt{n}\big(\phi_1(\widetilde{F}_{Y_d^*})(y) - 1\big) = o_p(1)$. As a result, when conditions for (A.12), (A.13) and (A.14) hold, by (A.15) and (A.16),

$$\begin{aligned}
&\mathbb{P}\left(\sup_{y \in \mathcal{Y}} \left|\sqrt{n}\left(\widehat{F}_{Y_d^*}(y) - \widetilde{F}_{Y_d^*}(y)\right)\right| \le \epsilon_2\right) \\
\ge&\,\mathbb{P}\left(\sup_{\rho(y_1, y_2) \le \delta} \left|\sqrt{n}\left(\widetilde{F}_{Y_d^*}(y_1) - F_{Y_d^*}(y_1)\right) - \left(\widetilde{F}_{Y_d^*}(y_2) - F_{Y_d^*}(y_2)\right)\right| \le \epsilon_2\right) \ge 1 - \epsilon_3. \qquad \square
\end{aligned}$$

Proof of Theorem 1:

Given the quantile map is Hadamard differentiable, Theorem 1 follows immediately from Lemma 2 and the functional delta method. □

Proof of Theorem 2:

We consider only the independent case here since the proof for the transformed case is similar to that of Donald and Hsu (2014, Theorem 4.5). We first show that

$$\mathcal{F}_d^u(y) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \widehat{\varrho}_d^F(y, Z_i) + \frac{1}{\sqrt{n^*}} \sum_{j=1}^{n^*} U_j^* \widehat{\varphi}_d^F(y, X_j^*) \overset{p}{\Rightarrow} \mathcal{F}_d(y).$$

To see this, note that

$$\mathcal{F}_d^u(y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \varrho_d^F(y, Z_i) + \frac{1}{\sqrt{n^*}} \sum_{j=1}^{n^*} U_j^* \varphi_d^F(y, X_j^*) \tag{A.17}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \left[\widehat{\varrho}_d^F(y, Z_i) - \varrho_d^F(y, Z_i)\right] \tag{A.18}$$

$$+ \frac{1}{\sqrt{n^*}} \sum_{j=1}^{n^*} U_j^* \left[\widehat{\varphi}_d^F(y, X_j^*) - \varphi_d^F(y, X_j^*)\right]. \tag{A.19}$$

We now show (A.18) converges weakly to a zero process conditional on the sample path $\mathcal{Z} \equiv \{\omega \in Z_i : i = 1, 2, \dots\}$ with probability approaching one. That is,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \left[\widehat{\varrho}_d^F(y, Z_i) - \varrho_d^F(y, Z_i)\right] \overset{p}{\Rightarrow} 0. \tag{A.20}$$

Note that (A.20) is true if and only if for any subsequence k_n of n, there exists a further subsequence ℓ_n of k_n such that

$$\frac{1}{\sqrt{\ell_n}} \sum_{i=1}^{\ell_n} U_i \left[\widehat{\varrho}_{d,\ell_n}^F(y, Z_i) - \varrho_d^F(y, Z_i)\right] \overset{\text{a.s.}}{\Rightarrow} 0, \tag{A.21}$$

where $\widehat{\varrho}_{d,\ell_n}^F(y, z)$ denotes the estimator at ℓ_n. By Lemma 3, we have $\sup_{y\in\mathcal{Y}, z\in\mathcal{Z}} \left|\widehat{\varrho}_{d,\ell_n}^F(y,z) - \varrho_d^F(y,z)\right| \overset{\text{a.s.}}{\to} 0$ for any subsequence k_n of n and a further subsequence ℓ_n of k_n. We then define $\mathcal{Z}_{\ell_n} \equiv \big\{\omega \in \mathcal{Z} : \sup_{y\in\mathcal{Y}, z\in\mathcal{Z}} \left|\widehat{\varrho}_{d,\ell_n}^F(y,z)(\omega) - \varrho_d^F(y,z)\right| \to 0\big\}$, where $\widehat{\varrho}_{d,\ell_n}^F(y,z)(\omega)$ denotes the realization at ω and $\mathbb{P}(\mathcal{Z}_{\ell_n}) = 1$. For any $\omega \in \mathcal{Z}_{\ell_n}$, define

$$t_{\ell_n,i}(U_i, y|\omega) = \frac{U_i}{\sqrt{\ell_n}} \left[\widehat{\varrho}_{d,\ell_n}^F(y, Z_i)(\omega) - \varrho_d^F(y, Z_i)\right].$$

Note that since we have conditioned $t_{\ell_n,i}(U_i, y|\omega)$ on the sample path ω, the randomness comes from the U_i's which is independent of the sample path ω.

Next, we claim that the triangular array $\left\{t_{\ell_n,i}(U_i, y|\omega), 1 \le i \le \ell_n, \ell_n \ge 1\right\}$ satisfies assumptions (i)–(v) of Theorem 10.6 in Pollard (1990). If this is the case, we can then apply the functional central limit theorem to show:

$$\frac{1}{\sqrt{\ell_n}} \sum_{i=1}^{\ell_n} U_i \left[\widehat{\varrho}_{d,\ell_n}^F(y, Z_i)(\omega) - \varrho_d^F(y, Z_i)\right] \Rightarrow 0,$$

meaning that (A.21) and (A.20) would follow accordingly. By Theorem 3.1 in Hsu (2016), it is sufficient to check that $\widehat{\varrho}_{d,\ell_n}^F(y, Z_i)$ satisfies (i)–(iii) of Assumption 3.2 in Hsu (2016):

(i) $\left\{\widehat{\varrho}_{d,\ell_n}^F(y, Z_i) : y \in \mathcal{Y}, 1 \le i \le \ell_n, \ell_n \ge 1\right\}$ is manageable with respect to the envelope function $\left\{\widehat{\Omega}_{\ell_n}(Z_i) : 1 \le i \le \ell_n, \ell_n \ge 1\right\}$ in the sense of Definition 7.9 of Pollard (1990).

(ii) $\sup_{y_1,y_2\in\mathcal{Y}} \left|\ell_n^{-1} \sum_{i=1}^{\ell_n} \widehat{\varrho}_{d,\ell_n}^F(y_1, Z_i)\widehat{\varrho}_{d,\ell_n}^F(y_2, Z_i) - \mathbb{E}\left[\varrho_d^F(y_1, Z)\varrho_d^F(y_2, Z)\right]\right| \overset{p}{\to} 0.$

(iii) There exists $\delta > 0$ such that

$$\frac{1}{\ell_n} \sum_{i=1}^{\ell_n} \widehat{\Omega}_{\ell_n}^2(Z_i) - \frac{1}{\ell_n} \sum_{i=1}^{\ell_n} \Omega_{\ell_n}^2(Z_i) \overset{p}{\to} 0, \qquad \frac{1}{\ell_n} \sum_{i=1}^{\ell_n} \widehat{\Omega}_{\ell_n}^{2+\delta}(Z_i) - \frac{1}{\ell_n} \sum_{i=1}^{\ell_n} \Omega_{\ell_n}^{2+\delta}(Z_i) \overset{p}{\to} 0.$$

To check (i), recall that

$$\widehat{\varrho}^F_{d,\ell_n}(y, Z_i) = \frac{\mathbb{1}\{D_i = d\}\left[\mathbb{1}\{Y_i \leq y\} - \widehat{F}_{Y_d|X,\ell_n}(y|X_i)\right]}{\widehat{p}_{\ell_n}(X_i)^d\left[1 - \widehat{p}_{\ell_n}(X_i)\right]^{1-d}} \frac{\widehat{f}_{X^*,\ell_n}(X_i)}{\widehat{f}_{X,\ell_n}(X_i)},$$

where the subscript ℓ_n indicates estimators at ℓ_n. Since $\mathbb{1}\{Y_i \leq y\}$ for all $y \in \mathcal{Y}$ forms a Vapnik-Chervonenkis class of functions, it is manageable with respect to the envelope function of 1's. In addition, due to monotonicity $\widehat{F}_{Y_d|X,\ell_n}(y|x)$ satisfies Pollard's entropy condition as in (4.2) of Andrews (1994) with the envelope function being $M_{\ell_n} \geq 1$. Next, by construction, $a_{\ell_n} = \inf_{x\in\mathcal{X}} \widehat{p}_{\ell_n}(x) = \inf_{x\in\mathcal{X}} 1 - \widehat{p}_{\ell_n}(x)$ and $b_{\ell_n} = \inf_{x\in\mathcal{X}} \widehat{f}_{X,\ell_n}(x)$. Since $\widehat{f}_{X^*,\ell_n}(x)$ is uniformly bounded by, say B_{ℓ_n}, it belongs to a type II class of functions with the envelope function being B_{ℓ_n}. Taken all together, $\widehat{\varrho}^F_{d,\ell_n}(y, Z_i)$ is manageable with respect to a constant envelope function $\widehat{\Omega}_{\ell_n} = a_{\ell_n} b_{\ell_n} B_{\ell_n}(1 + M_{\ell_n}) > 0$, and hence (i) is satisfied.

To check (ii) and (iii), note that the functions involved in $\widehat{\varrho}^F_d(y, z)$ are uniformly consistent over $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$ by Lemma 3. It is thus easy to see (ii) and (iii) follow accordingly. In other words, the triangular array $\left\{t_{\ell_n,i}(U_i, y|\omega)\right\}$ for all $\omega \in \mathcal{Z}$ satisfies all requirements in Theorem 10.6 of Pollard (1990), meaning that conditional on the sample path ω and given the randomness coming from the U_i's,

$$\frac{1}{\sqrt{\ell_n}} \sum_{i=1}^{\ell_n} U_i \left[\widehat{\varrho}^F_{d,\ell_n}(y, X_i)(\omega) - \varrho^F_d(y, X_i)\right] \Rightarrow 0.$$

By a similar argument, it can be shown that (A.19) also converges weakly to a zero process conditional on the sample path $\{\omega \in X^*_j : j = 1, 2, \ldots\}$. Finally, by Corollary 2.9.3 in van der Vaart and Wellner (1996), it is true that (A.17) converges weakly to $\mathcal{F}_d(y)$ with probability approaching one.

We are now ready to show the conditional weak convergence of the simulated process for QCTE,

$$\Delta^u(\tau) = -\left[\frac{\mathcal{F}^u_1\left(\widehat{\mathbb{Q}}_{Y^*_1}(\tau)\right)}{\widehat{f}_{Y^*_1}\left(\widehat{\mathbb{Q}}_{Y^*_1}(\tau)\right)} - \frac{\mathcal{F}^u_0\left(\widehat{\mathbb{Q}}_{Y^*_0}(\tau)\right)}{\widehat{f}_{Y^*_0}\left(\widehat{\mathbb{Q}}_{Y^*_0}(\tau)\right)}\right] \overset{p}{\Rightarrow} \Delta(\tau). \tag{A.22}$$

Note that

$$\begin{aligned}
&\sup_{\tau\in[0,1]} \left|\widehat{f}_{Y^*_d}\left(\widehat{\mathbb{Q}}_{Y^*_d}(\tau)\right) - f_{Y^*_d}\left(\mathbb{Q}_{Y^*_d}(\tau)\right)\right| \\
\leq& \sup_{\tau\in[0,1]} \left|\widehat{f}_{Y^*_d}\left(\widehat{\mathbb{Q}}_{Y^*_d}(\tau)\right) - f_{Y^*_d}\left(\widehat{\mathbb{Q}}_{Y^*_d}(\tau)\right)\right| + \sup_{\tau\in[0,1]} \left|f_{Y^*_d}\left(\widehat{\mathbb{Q}}_{Y^*_d}(\tau)\right) - f_{Y^*_d}\left(\mathbb{Q}_{Y^*_d}(\tau)\right)\right| \\
\leq& \sup_{y\in\mathcal{Y}} \left|\widehat{f}_{Y^*_d}(y) - f_{Y^*_d}(y)\right| + C \sup_{\tau\in[0,1]} \left|\widehat{\mathbb{Q}}_{Y^*_d}(\tau) - \mathbb{Q}_{Y^*_d}(\tau)\right| = o_p(1),
\end{aligned}$$

for some constant C, and the second inequality comes from Assumption 3.3 that $f_{Y^*_d}(y)$ is two-times continuous differentiable on $\mathcal{Y}$. Moreover, it is true that $\sup_{\tau\in[0,1]} \left|\mathcal{F}^u_d\left(\widehat{\mathbb{Q}}_{Y^*_d}(\tau)\right) - \mathcal{F}^u_d\left(\mathbb{Q}_{Y^*_d}(\tau)\right)\right| = o_p(1)$ conditioning on the sample path with probability approaching one by the equicontinuity of $\mathcal{F}^u_d(y)$ and the uniform consistency of $\widehat{\mathbb{Q}}_{Y_d^*}(\tau)$. As a result, we have:

$$\begin{aligned}
&\sup_{\tau\in[0,1]}\left|\frac{\mathcal{F}_1^u\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)}{\widehat{f}_{Y_1^*}\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)}-\frac{\mathcal{F}_1^u\left(\mathbb{Q}_{Y_1^*}(\tau)\right)}{f_{Y_1^*}\left(\mathbb{Q}_{Y_1^*}(\tau)\right)}\right| \\
&\le \sup_{\tau\in[0,1]}\left|\frac{\mathcal{F}_1^u\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)}{\widehat{f}_{Y_1^*}\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)}-\frac{\mathcal{F}_1^u\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)}{f_{Y_1^*}\left(\mathbb{Q}_{Y_1^*}(\tau)\right)}\right|+\sup_{\tau\in[0,1]}\left|\frac{\mathcal{F}_1^u\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)}{f_{Y_1^*}\left(\mathbb{Q}_{Y_1^*}(\tau)\right)}-\frac{\mathcal{F}_1^u\left(\mathbb{Q}_{Y_1^*}(\tau)\right)}{f_{Y_1^*}\left(\mathbb{Q}_{Y_1^*}(\tau)\right)}\right| \\
&\le C\sup_{\tau\in[0,1]}\left|\widehat{f}_{Y_1^*}\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)-f_{Y_1^*}\left(\mathbb{Q}_{Y_1^*}(\tau)\right)\right|\sup_{\tau\in[0,1]}\left|\mathcal{F}_1^u\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)\right|+ \\
&\qquad\qquad C'\sup_{\tau\in[0,1]}\left|\mathcal{F}_1^u\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)-\mathcal{F}_1^u\left(\mathbb{Q}_{Y_1^*}(\tau)\right)\right| \\
&= C\cdot o_p(1)\cdot O_p(1)+C'\cdot o_p(1)=o_p(1).
\end{aligned}$$

The result regarding $\mathcal{F}_0^u\left(\widehat{\mathbb{Q}}_{Y_0^*}(\tau)\right)/\widehat{f}_{Y_0^*}\left(\widehat{\mathbb{Q}}_{Y_0^*}(\tau)\right)$ can be shown similarly and so is omitted. Finally, (A.22) holds because

$$\begin{aligned}
\Delta^u(\tau)=-&\left\{\left[\frac{\mathcal{F}_1^u\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)}{\widehat{f}_{Y_1^*}\left(\widehat{\mathbb{Q}}_{Y_1^*}(\tau)\right)}-\frac{\mathcal{F}_1^u\left(\mathbb{Q}_{Y_1^*}(\tau)\right)}{f_{Y_1^*}\left(\mathbb{Q}_{Y_1^*}(\tau)\right)}\right]+\left[\frac{\mathcal{F}_0^u\left(\widehat{\mathbb{Q}}_{Y_0^*}(\tau)\right)}{\widehat{f}_{Y_0^*}\left(\widehat{\mathbb{Q}}_{Y_0^*}(\tau)\right)}-\frac{\mathcal{F}_0^u\left(\mathbb{Q}_{Y_0^*}(\tau)\right)}{f_{Y_0^*}\left(\mathbb{Q}_{Y_0^*}(\tau)\right)}\right]\right. \\
&\left.+\left[\frac{\mathcal{F}_1^u\left(\mathbb{Q}_{Y_1^*}(\tau)\right)}{f_{Y_1^*}\left(\mathbb{Q}_{Y_1^*}(\tau)\right)}-\frac{\mathcal{F}_0^u\left(\mathbb{Q}_{Y_0^*}(\tau)\right)}{f_{Y_0^*}\left(\mathbb{Q}_{Y_0^*}(\tau)\right)}\right]\right\}\overset{p}{\Rightarrow}\Delta(\tau).
\end{aligned}$$

□

Proof of Lemma 3:

It suffices to check that

$$\begin{aligned}
&\sup_{y\in\mathcal{Y}}\left|\widehat{F}_{Y_d^*}(y)-F_{Y_d^*}(y)\right|+\sup_{y\in\mathcal{Y},x\in\mathcal{X}}\left|\widehat{F}_{Y_d|X}(y|x)-F_{Y_d|X}(y|x)\right|+\sup_{x\in\mathcal{X}}\left|\widehat{p}(x)-p(x)\right| \\
&+\sup_{x\in\mathcal{X}}\left|\widehat{f}_X(x)-f_X(x)\right|+\sup_{x\in\mathcal{X}}\left|\widehat{f}_{X^*}(x)-f_{X^*}(x)\right|+\sup_{y\in\mathcal{Y}}\left|\widehat{f}_{Y_d^*}(y)-f_{Y_d^*}(y)\right|=o_p(1).
\end{aligned} \tag{A.23}$$

From Lemma 2 it is true that $\sup_{y\in\mathcal{Y}}\left|\widehat{F}_{Y_d^*}(y)-F_{Y_d^*}(y)\right|=o_p(1)$. For the second and third terms in (A.23), the uniform consistency for $\widetilde{F}_{Y_d|X}(y|x)$ and $\widetilde{p}(x)$ has been established by Härdle, Jansson and Serfling (1988). We then follow Lemma 4.1 of Donald and Hsu (2014) to show that $\sup_{y\in\mathcal{Y},x\in\mathcal{X}}\left|\widehat{F}_{Y_d|X}(y|x)-F_{Y_d|X}(y|x)\right|=o_p(1)$. Suppose y' is the first point at which $\widetilde{F}_{Y_d|X}(y|x)$ jumps down, and then for $y'\le y<y'+\epsilon$, $\epsilon>0$, $\widehat{F}_{Y_d|X}(y|x)=\widetilde{F}_{Y_d|X}(y'-\epsilon|x)>\widetilde{F}_{Y_d|X}(y|x)$ and for $y'-\epsilon\le y<y'$, $\widehat{F}_{Y_d|X}(y|x)=\widetilde{F}_{Y_d|X}(y'-\epsilon|x)$. Next, for $y'\le y<y'+\epsilon$, if $\widehat{F}_{Y_d|X}(y|x)\le F_{Y_d|X}(y|x)$, we then have $F_{Y_d|X}(y|x)-\widetilde{F}_{Y_d|X}(y|x)>F_{Y_d|X}(y|x)-\widehat{F}_{Y_d|X}(y|x)>0$. On the other hand, if $\widehat{F}_{Y_d|X}(y|x)>F_{Y_d|X}(y|x)$, we then have $\widetilde{F}_{Y_d|X}(y'-\epsilon|x)-F_{Y_d|X}(y'-\epsilon|x)>\widehat{F}_{Y_d|X}(y|x)-F_{Y_d|X}(y|x)>0$ since $F_{Y_d|X}(y|x)$ is nondecreasing in y. These results imply that for $y'\le y<y'+\epsilon$,

$$\left|\widehat{F}_{Y_d|X}(y|x)-F_{Y_d|X}(y|x)\right|\le\max\left\{\left|\widetilde{F}_{Y_d|X}(y'-\epsilon|x)-F_{Y_d|X}(y'-\epsilon|x)\right|,\left|\widetilde{F}_{Y_d|X}(y|x)-F_{Y_d|X}(y|x)\right|\right\}.$$

Consequently, $\sup_{0\le y\le y'+\epsilon}\left|\widehat{F}_{Y_d|X}(y|x)-F_{Y_d|X}(y|x)\right|\le\sup_{0\le y\le y'+\epsilon}\left|\widetilde{F}_{Y_d|X}(y|x)-F_{Y_d|X}(y|x)\right|$ and we have $\sup_{y\in\mathcal{Y},x\in\mathcal{X}}\left|\widehat{F}_{Y_d|X}(y|x)-F_{Y_d|X}(y|x)\right|\le\sup_{y\in\mathcal{Y},x\in\mathcal{X}}\left|\widetilde{F}_{Y_d|X}(y|x)-F_{Y_d|X}(y|x)\right|=o_p(1)$ by induction. Since $|a_n|\le 1$ for all x, it is easy to see that $\sup_{x\in\mathcal{X}}\left|\widehat{p}(x)-p(x)\right|\le\sup_{x\in\mathcal{X}}\left|\widetilde{p}(x)-p(x)\right|=o_p(1)$.

For the fourth term in (A.23), we note that $\sup_{x\in\mathcal{X}}\left|\widetilde{f}_X(x)-f_X(x)\right|=o_p(1)$ is given by Jones (1993). Therefore, it is true that $\sup_{\{x:\widetilde{f}_X(x)\ge b_n\}}\left|\widehat{f}_X(x)-f_X(x)\right|=\sup_{\{x:\widetilde{f}_X(x)\ge b_n\}}\left|\widetilde{f}_X(x)-f_X(x)\right|\le\sup_{x\in\mathcal{X}}\left|\widetilde{f}_X(x)-f_X(x)\right|=o_p(1)$. Similar to Lemma 3.2 in Donald, Hsu, and Barrett (2012), for all x

such that $\widetilde{f}_X(x) < b_n$ we let x' satisfy $\widetilde{f}_X(x') = b_n$,

$$\begin{aligned}\left|\widehat{f}_X(x') - f_X(x)\right| &\leq \left|\widehat{f}_X(x') - f_X(x')\right| + |f_X(x') - f_X(x)| \\ &\leq |\widehat{f}_X(x') - f_X(x')| + M(x' - x) = o_p(1),\end{aligned}$$

where the second inequality holds by Assumption 3.2 for some $M > 0$. This implies the fact that $\sup_{\{x:\widetilde{f}_X(x)<b_n\}} \left|\widehat{f}_X(x) - f_X(x)\right| = o_p(1)$ and the uniform consistency of $\widehat{f}_X(x)$. For the other parts in (A.23), since $\sup_{y\in\mathcal{Y},x\in\mathcal{X}} \left|\widetilde{f}_{Y_d|X}(y|x) - f_{Y_d|X}(y|x)\right| = o_p(1)$ as shown by Hyndman, Bashtannyk and Grunwald (1996), the results regarding the fifth and the last terms follow similarly. $\square$

Proof of Corollary 1:

The proof is omitted since it is similar to the proof of Lemma 2 except replacing $\mathbb{1}\{Y_i \leq y\}$'s with Y_i's. $\square$

Proof of Lemma 4:

To see this, note that $\mathbb{Q}_{Y_d^*|D^*}(\tau|1) = \inf\left\{y \in \mathcal{Y} : F_{Y_d^*|D^*}(y|1) \geq \tau\right\}$ and

$$\begin{aligned}F_{Y_d^*|D^*}(y|1) &= \int_{\mathcal{X}^*} F_{Y_d^*|X^*,D^*}(y|x,1)\,\mathrm{d}F_{X^*|D^*}(x|1) = \int_{\mathcal{X}} F_{Y_d|X,D}(y|x,1) f_{X^*|D^*}(x|1)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} F_{Y|X,D}(y|x,d)\frac{p^*(x) f_{X^*}(x)}{\mathbb{P}(D^*=1)}\,\mathrm{d}x = \int_{\mathcal{X}} F_{Y|X,D}(y|x,d)\frac{p(x)}{\int_{\mathcal{X}} p(x) f_{X^*}(x)\,\mathrm{d}x}\,\mathrm{d}F_{X^*}(x) \\ &= \int_{\mathcal{X}} F_{Y|X,D}(y|x,d)\frac{p(x)}{\mathbb{E}\left[p(X^*)\right]}\,\mathrm{d}F_{X^*}(x),\end{aligned}$$

where the second equality follows from Assumption 7.6 and the third holds since $Y_1 = Y$ if $D = 1$, $F_{Y_0|X,D}(y|x,1) = F_{Y_0|X,D}(y|x,0) = F_{Y|X,D}(y|x,0)$ by Assumption 7.5(i) and by Bayes' theorem. The fourth equality is true given Assumption 7.7. Since we can observe Y, X, D, X^*, and $p(x) > 0$ for all $x \in \mathcal{X}$ by Assumption 7.5(ii), the last line is well-defined and identified. Thus, ACTT and QCTT can be identified as well. $\square$

Proof of Corollary 2:

The proof follows the same line of reasoning as in Lemma 2 and Theorem 1 and so is omitted. $\square$

B Implementation for the Monotonizing Method

This section shows the implementation of the monotonizing method in (3.3) which 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 $0 = Y_{(0)} < Y_{(1)} < \cdots < Y_{(n)} < Y_{(n+1)} = \bar{y}$. Let $\widetilde{M}_d = \sup_{y \in \mathcal{Y}} \widetilde{F}_{Y_d^*}(y)$. It is true that $\widetilde{M}_d \geq 1$ since $\widetilde{F}_{Y_d^*}(\bar{y}) = 1$. We then construct $\widehat{F}_{Y_d^*}(y)$ by induction:

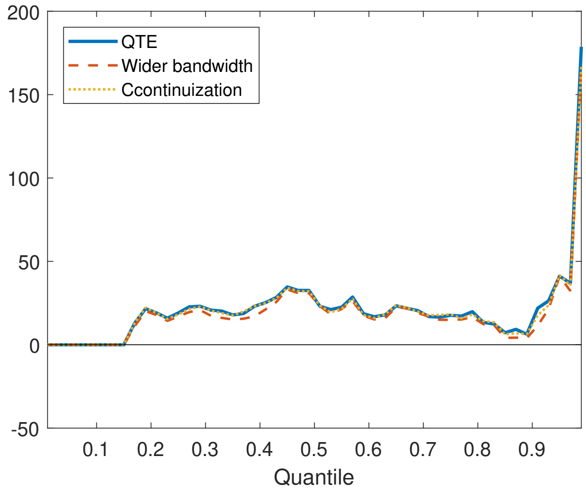
1. Define $\widehat{F}_{Y_d^*}(y) = 0$ for $Y_{(0)} \leq y < Y_{(1)}$.
2. Suppose $\widehat{F}_{Y_d^*}(y)$ is already defined for $Y_{(0)} \leq y < Y_{(i)}$, we then define $\widehat{F}_{Y_d^*}(y)$ for $Y_{(i)} \leq y < Y_{(i+1)}$ as

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

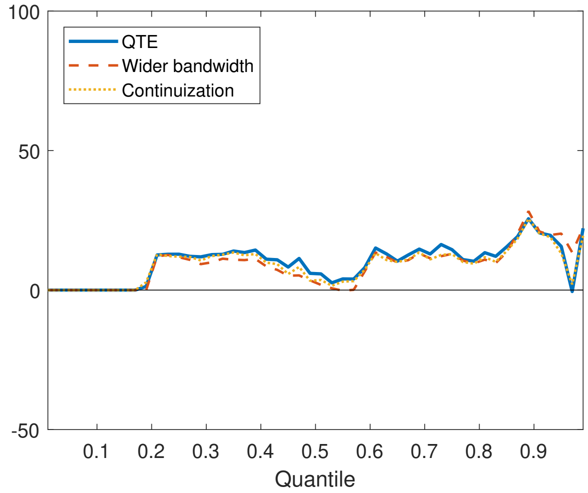
Continuing the same way, we can construct $\widehat{F}_{Y_d^*}(y)$ that is monotonically increasing and lies between the unit interval for all $y \in \mathcal{Y}$.

C Robustness Checks

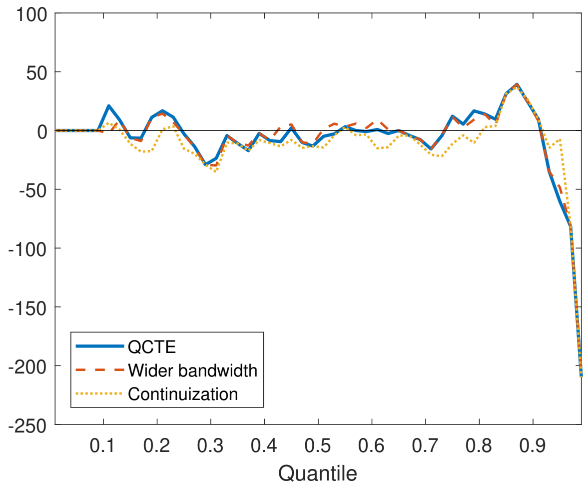
In this section, we undertake several sensitivity checks to examine the robustness of our empirical results. First, since the age variable is measured in years and the under-smoothed bandwidth $h_d < 1$ for $d = 0, 1$ in all cases, we then consider to adopt a wider bandwidth $1 < h_d < 2$ by specifying a larger bandwidth constant. Next, we "continuize" the age variable by adding a small random noise. Specifically, we add a uniformly distributed random number in the range [-0.5,0.5] to the integer-valued age to make it more continuous. The point estimates are presented in Figure 5 and are virtually identical to those reported in the main text. Thus, we conclude that these robustness checks leave our results intact.



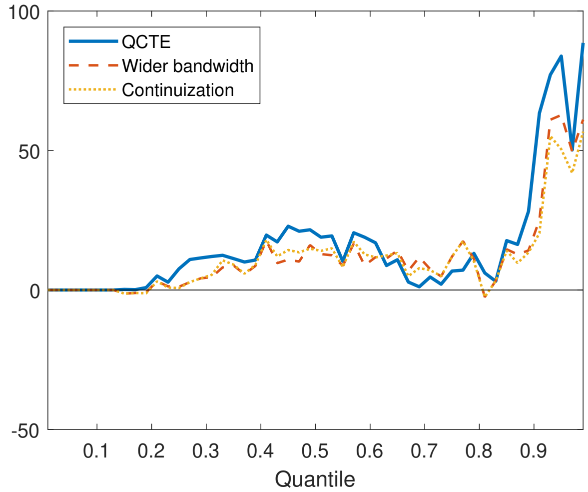
(a) Job Corps for males.



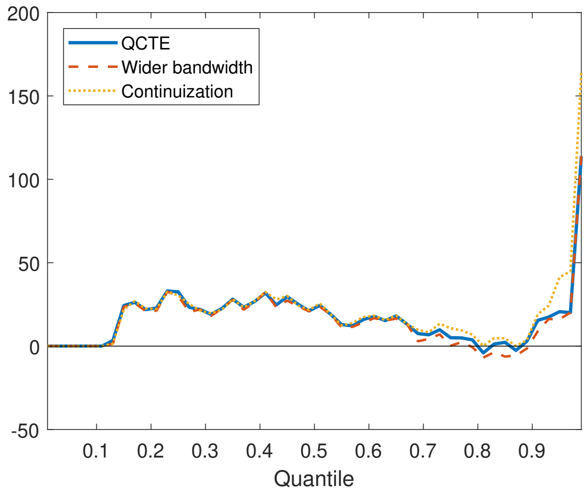
(b) Job Corps for females.



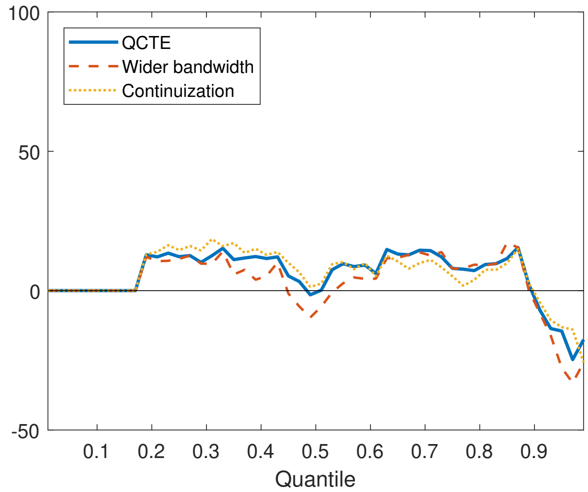
(c) Job Corps for JTPA males.



(d) JTPA for Job Corps females.



(e) Job Corps for males with increased education.



(f) Job Corps for females with increased education.

Figure 5. Robustness checks.

References

Allcott, H. (2015): "Site Selection Bias in Program Evaluation," *The Quarterly Journal of Economics*, **130**, 1117-1165.

Andrews, D. W. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, 4, 2247-2294.

Andrews, I. and E. Oster (2018): "Weighting for External Validity," NBER Working Paper 23826.

Angrist, J. D. and I. Fernandez-Val (2013): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," in D. Acemoglu, M. Arellano, and E. Dekel (Eds.), *Advances in Economics and Econometrics*, 401–434. Cambridge University Press.

Angrist, J. D. and M. Rokkanen (2015): "Wanna Get Away? RD Identification Away from the Cutoff," *Journal of the American Statistical Association*, **110**, 1331–1344.

Athey, S. and G. Imbens (2017): "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, **31**, 3–32.

Barrett, G. F. and S. G. Donald (2003): "Consistent Tests for Stochastic Dominance," *Econometrica*, **71**, 71–104.

Bertanha, M. and G. W. Imbens (2018): "External Validity in Fuzzy Regression Discontinuity Designs" NBER Working Paper 20773, New York: National Bureau of Economic Research.

Bloom, H. S., L. L. Orr, S. H. Bell, G. Cave, F. Doolittle, W. Lin, and J. M. Bos (1997): "The Benefits and Costs of JTPA Title II-A Programs: Key findings from the National Job Training Partnership Act study," *Journal of Human Resources*, 549–576.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors," *The Annals of Statistics*, **41**, 2786–2819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2016): "Empirical and Multiplier Bootstraps for Suprema of Empirical Processes of Increasing Complexity, and Related Gaussian Couplings," *Stochastic Processes and their Applications*.

Chernozhukov, V., I. Fernandez-Val and A. Galichon (2009): "Improving Point and Interval Estimators of Monotone Functions by Rearrangement," *Biometrika*, **96**, 559–575.

Chernozhukov, V., I. Fernandez-Val and A. Galichon (2010): "Quantile and Probability Curves without Crossing," *Econometrica*, **78**, 1093–1125.

Chernozhukov, V., I. Fernandez-Val, and B. Melly (2013): "Inference on Counterfactual Distributions," *Econometrica*, **81**, 2205–2268.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik, (2009): "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika* **96**, 187–199.

Dehejia, R., C. Pop-Eleches and C. Samii (2017): "From Local to Global: External Validity in a Fertility Natural Experiment," NBER Working Paper w21459. National Bureau of Economic Research.

Donald, S. G., and Y.-C. Hsu (2014): "Estimation and Inference for Distribution Functions and Quantile

Functions in Treatment Effect Models," *Journal of Econometrics*, **178**, 383–397.

Donald, S. G., Y.-C. Hsu, and G. F. Barrett (2012): "Incorporating Covariates in the Measurement of Welfare and Inequality: Methods and Applications," *The Econometrics Journal*, **15**, C1–C30.

Donald, S. G., Y.-C. Hsu, and R. P. Lieli (2014a): "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT," *Journal of Business and Economic Statistics*, **32**, 395–415.

Donald, S. G., Y.-C. Hsu, and R. P. Lieli (2014b): "Inverse Probability Weighted Estimation of Local Average Treatment Effects: A Higher Order MSE Expansion," *Statistics and Probability Letters*, **95**, 132–138.

Dong, Y. and A. Lewbel (2015): "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," *Review of Economics and Statistics*, **97**, 1081–1092.

Eren, O. and S. Ozbeklik (2014): "Who Benefits from Job Corps? A Distributional Analysis of an Active Labor Market Program," *Journal of Applied Econometrics*, **29**, 586–611.

Fan, J. and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications*, London; New York and Melbourne; Chapman and Hall.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, **75**, 259–276.

Firpo, S., N. M. Fortin and T. Lemieux (2009): "Unconditional Quantile Regressions," *Econometrica*, **77**, 953–973.

Fortin, N., T. Lemieux, and S. Firpo (2011): "Decomposition Methods in Economics," vol. 4, Part A of *Handbook of Labor Economics*, pp. 1 - 102. Elsevier.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, **66**, 315–331.

Härdle, W., P. Janssen and R. Serfling (1988): "Strong Uniform Consistency Rates for Estimators of Conditional Functionals," *The Annals of Statistics*, 1428–1449.

Hartman, E., R. Grieve, R. Ramsahai and J. S. Sekhon (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**, 757–778.

Heckman, J. J., H. Ichimura and P. Todd (1998): "Matching as an Econometric Evaluations Estimator," *Review of Economic Studies*, **65**, 261–294.

Heckman, J. J., J. Stixrud and S. Urzua (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, **24**, 411–482.

Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, **73**, 669–738.

Heckman, J. J. and E. Vytlacil (2007): "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast Their Effects in New Environments," in *Handbook of Econometrics*, Vol. 6B, ed. by J. Heckman and E. Leamer. Amsterdam: Elsevier, 4875–5144.

Hirano, K., G. Imbens and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, **71**, 1161–1189.

Hotz, V. J., G. Imbens and J. A. Klerman (2006): "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program," *Journal of Labor Economics*, **24**.

Hotz, V. J., G. Imbens and J. Mortimer (2005): "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations," *Journal of Econometrics*, **125**, 241–270.

Hsu, Y.-C. (2016): "Multiplier Bootstrap for Empirical Processes," Working Paper.

Hsu, Y.-C., R. P. Lieli and T.-C. Lai (2018): "Estimation and Inference for Distribution Functions and Quantile Functions in Endogenous Treatment Effect Models," Working Paper.

Hyndman, R. J., D. M. Bashtannyk and G. K. Grunwald (1996): "Estimating and Visualizing Conditional Densities," *Journal of Computational and Graphical Statistics*, **5**, 315-336.

Imbens, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, **48**, 399–423.

Imbens, G. W. and J. W. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, **47**, 5–86.

Jones, M. C. (1993): "Simple Boundary Correction for Kernel Density Estimation," *Statistics and Computing*, **3**, 135–146.

Kline, B. and E. Tamer (2018): "Identification of Treatment Effects with Selective Participation in a Randomized Trial," *The Econometrics Journal*, Forthcoming.

Kline, P. and A. Santos (2012): "A Score Based Approach to Wild Bootstrap Inference," *Journal of Econometric Methods*, **1**, 23–41.

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*. Springer Science and Business Media.

Kowalski, A. (2016): "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments," Working Paper.

Lechner, M. and C. Wunsch (2009): "Are Training Programs More Effective When Unemployment is High?," *Journal of Labor Economics*, **27**, 653–692.

Li, Q. and J. Racine (2008): "Nonparametric Estimation of Conditional CDF and Quantile Functions with Mixed Categorical and Continuous Data," *Journal of Business and Economic Statistics*, **26**, 423–434.

Newey, W. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, **10**, 233–253.

Nolan, D. and D. Pollard (1987): "U-Processes: Rates of Convergence," *Annals of Statistics*, **15**, 780-799.

Pollard, D. (1990): "Empirical Processes: Theory and Applications." CBMS Conference Series in Probability and Statistics, Vol. 2. Hayward, CA: Institute of Mathematical Statistics.

Rothe, C. (2010): “Nonparametric Estimation of Distributional Policy Effect,” *Journal of Econometrics*, **155**, 56–70.

Ruppert, D. and M. P. Wand (1994): “Multivariate Locally Weighted Least Squares Regression,” *The Annals of Statistics*, 1346–1370.

Schochet, P., J. Burghardt and S. McConnell (2008): “Does Job Corps Work? Impact Findings from the National Job Corps Study,” *The American Economic Review*, **98**, 1864–1886.

Sherman, R. (1994): “Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators,” *The Annals of Statistics,* **22**, 439-459.

Stuart, E. A., S. R. Cole, C. P. Bradshaw and P. J. Leaf (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**, 369–386.

van der Vaart, A. (2000): *Asymptotic Statistics.* Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Application to Statistics.* New York: Springer-Verlag.