Estimating Counterfactual Treatment Effects to Assess External Validity

Yu-Chin Hsu[*]

Institute of Economics
Academia Sinica

Tsung-Chih Lai[†]

Department of Economics
Feng Chia University

Robert P. Lieli[‡]

Department of Economics
Central European University

This version: July 25, 2017

[*] ychsu@econ.sinica.edu.tw. [†] Corresponding author. tclai@fcu.edu.tw. [‡] lielir@ceu.edu.

Acknowledgements: This paper replaces an earlier version that was titled "Forecasting Treatment Effects." We thank Irene Botosaru, Le-Yu Chen, 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 the research support from Ministry of Science and Technology of Taiwan (MOST103-2628-H-001-001-MY4) and Career Development Award of Academia Sinica, Taiwan.

Abstract

We propose statistical methods for assessing the external validity of treatment effect estimates obtained in a specific status-quo environment. In particular, we estimate *counterfactual* quantile treatment effects that would obtain if one were to change the composition of the population targeted by the status-quo treatment. Assuming unconfoundedness, and the invariance of the conditional distributions of the potential outcomes, the parameter of interest 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. Exploiting this result, we propose a multiplier bootstrap procedure to construct uniform confidence bands. We provide similar results for the counterfactually treated subpopulation and the average effect. As an application, we estimate the counterfactual quantile treatment effect of the Job Corps training program in the U.S. under various scenarios. The results suggest that strong economic conditions and the skill hypotheses both help explain the earlier finding in the literature that the program was ineffective at low quantiles of the earnings distribution.

JEL Classification: C13, C31, J24, J30.

Keywords: counterfactual analysis, external validity, program evaluation, multiplier bootstrap, Job Corps.

1 Introduction

The focus of the program evaluation literature has traditionally been on internal validity, i.e., on credible identification and estimation of treatment effect parameters for the population from which the data are actually drawn. Parameters such as the average treatment effect (ATE, Hahn, 1998; Heckman, Ichimura, and Todd, 1998; Hirano, Imbens, and Ridder, 2003) or a quantile treatment effect (QTE, Firpo 2007) characterize the causal impact of a treatment in the specific environment where it is implemented. Nevertheless, to inform policy decisions about potential extensions or modifications of a treatment, program evaluation studies must also possess some degree of external validity, i.e., they must be relevant for other populations the treatment might be extended to. While this is a well-known concern (see, e.g., Imbens 2010), discussions of external validity in empirical studies typically remain informal and speculative. It is only relatively recently that econometric research has started addressing questions of external validity in a more formal way (a short review is provided below). Our paper contributes to this growing literature by developing statistical methods for extrapolating the effect of a *status quo* treatment, whose impact can be credibly estimated under the unconfoundedness assumption, to a *counterfactual* environment.

The following two examples help to clarify the counterfactual scenarios we address. The first one concerns program implementation/extension (Hotz, Imbens, and Mortimer, 2005) and the second policy intervention (Rothe, 2010):

Example 1 (Program Implementation). Consider a job training program for a given population in a given region. Information is available about individual earnings, participation status, and other individual characteristics. A policymaker plans to expand the program to a new region where only individual characteristics are currently observed. The goal is to predict the program effect in the new region 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 goal is to predict the resulting change in the program effect, especially for the treated 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. Our framework is fully nonparametric in that we only restrict this function via general conditions on the counterfactual potential outcome distributions. Thus, we allow for heterogeneous effects across quantiles and capture any distributional impact the modified program might have. We also develop identification and estimation results for the average counterfactual treatment effect (ACTE) and the quantile counterfactual treatment effect for the treated (QCTT), but the exposition will focus on QCTE.

In order 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 captured 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 to the status-quo or the counterfactual population. In other words, we attribute any difference between the status quo and counterfactual treatment effect to the difference in the distribution of X across the two environments rather than the treatment operating in a fundamentally different way. This is admittedly a strong assumption, which needs to be argued for on a case by case basis.

Under these assumptions, QCTE is nonparametrically identified and can be estimated in a number of steps. We first use a Nadaraya-Watson estimator to construct the conditional distribution functions of the status quo potential outcomes for $X = x$ using the observations on X from the status quo environment. Second, we integrate out x using the empirical measure of the X-observations drawn from the *counterfactual* environment, and hence 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, and we propose a multiplier bootstrap procedure to construct uniform confidence bands for it. We also propose estimators for ACTE and QCTT, and state similar results.

As 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 of the distribution functions of the status quo potential outcomes, an alternative to the inverse probability weighted estimators proposed by Donald and Hsu (2014). Our estimator is first-order asymptotically equivalent.

We illustrate the proposed method by Monte Carlo simulations and an empirical application focusing on the heterogeneous earnings impact of Job Corps, the largest and most costly active labor market program in the United States. As reported by 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 (the National Job Corps Study was conducted from 1994 to 1998); (ii) the skill hypothesis which states that the program is effective only for individuals with sufficiently high levels of education. We test the empirical relevance of these explanations based on their implications for external validity. In particular, we first use data from the National Supported Work Demonstration (NSW), an earlier program in the mid-1970s, to estimate the counterfactual effect of Job Corps for the population targeted by the NSW. If the first hypothesis is true,

we should find a reduced or insignificant program effect for this population as well. In addition, we reverse the roles of the two programs and estimate the counterfactual program effect as if the Job Corps cohort were to participate in the NSW. Under the strong economic conditions hypothesis, we expect the Job Corps cohort would receive a significantly larger program effect in another time period. The empirical results are consistent with the predictions. Second, we take the individuals who do not benefit from the Job Corps program and artificially give them high school education. If the skill hypothesis is true, we would expect to see a significantly increased counterfactual program effect at the corresponding earnings quantiles. Again, the empirical results confirm this prediction; thus, the evidence suggests that both explanations contribute to the ineffectiveness documented by Eren and Ozbeklik (2014).

As mentioned above, this paper is not the first in the program evaluation literature to formally address problems related to external validity. Recent research, surveyed by Athey and Imbens (2016), points at least into three directions. The first is concerned with instrumental variables settings, where the aim is to extrapolate the local average treatment effect, the average treatment effect for individuals whose treatment status is affected by the instrument, to other subgroups or the entire population (Angrist and Fernandez-Val, 2013; Bertanha and Imbens, 2016; Kowalski, 2016). 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, 2016). Most closely related to this paper is the third direction, where the difference in treatment effects across the status quo and counterfactual environments is attributed to the different composition of individual characteristics (Hotz, Imbens, and Mortimer, 2005; Hotz, Imbens, and Klerman, 2006; Allcott, 2015). Our paper intends to make at least two contributions here: (i) We relax the requirement that the data on the status quo treatment should come from a randomized experiment; instead, we allow for observational data as long as treatment assignment is unconfounded. (ii) We extend the analysis beyond average effects and provide a unified nonparametric framework for estimating and conducting uniform inference on QCTE.

This last strand of the external validity literature, including our paper, is also closely 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. Chernozhukov, Fernandez-Val and Melly (2013), and Rothe (2010) consider situations where the covariates are either drawn from a completely new distribution or transformed from the original values. While the former authors propose a semiparametric method, our estimation procedure builds on Rothe's fully nonparametric approach. Nevertheless, there are several technical distinctions between this paper and Rothe (2010), 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 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. (The tradeoff is that the multiplier bootstrap requires consistent estimation of the covariance functions.) 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 can assign 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 (2016), which simply replaces any downward step in the c.d.f. estimate with a constant piece, and is very easy to implement.

The rest of the 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 which are crucial for the validity of the multiplier bootstrap discussed in Section 4. Here we also discuss how to conduct uniform inference for QCTE. Section 5 presents the simulation study and the empirical application. In Section 6 we extend the analysis to average effect case and the treated subset of the counterfactual population. Finally Section 7 concludes. All proofs are collected in Appendix A.

2 Model Framework and Identification

2.1 Model

Following the Rubin causal model, we let $D \in \{0, 1\}$ be a binary treatment indicator 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 actually observed outcome is then $Y = DY_1 + (1 - D)Y_0$. As in Heckman and Vytlacil (2005, 2007) and Fortin, Lemieux, and Firpo (2011), we consider a nonparametric and nonseparable structural model of the potential outcomes as

$$Y_d = m_d(X, \varepsilon_d), \quad d = 0, 1, \tag{2.1}$$

where $X = (X_1, \ldots, X_k)$ is a k-dimensional vector of pre-treatment covariates and ε_d is an error term representing unobserved heterogeneity. The function m_d, unknown to the econometrician, determines an individual's potential outcome Y_d given observed and unobserved characteristics X and ε_d. One can think of an individual's participation decision D being based on some estimate of Y_d, $d = 0, 1$ given X and partial or full knowledge of ϵ_d and m_d. We do not impose

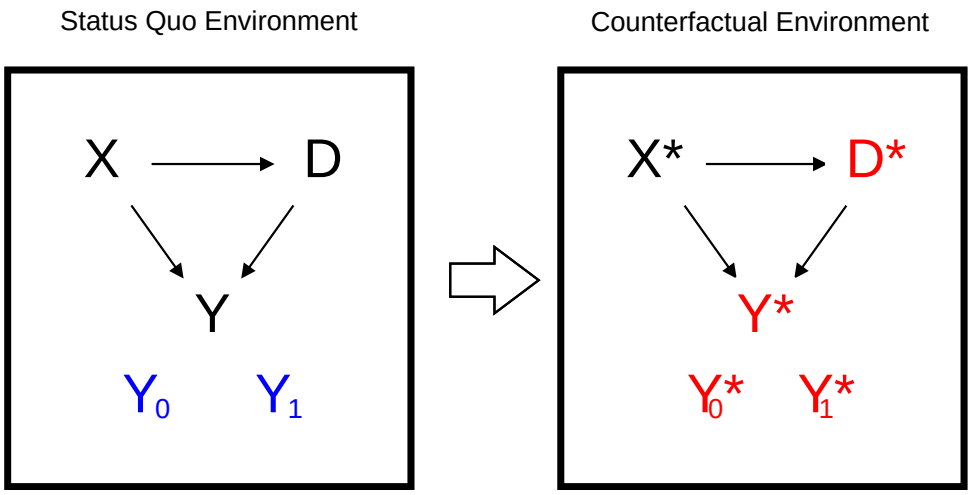


Figure 1: Model Framework

parametric restrictions on m_d or the distribution of ε_d. The dimension of ε_d is also unrestricted. We will hereafter refer to the model (2.1) as the "status quo".

We now introduce the *counterfactual potential outcomes* in the same structural sense:

$$Y_d^* = m_d(X^*, \varepsilon_d), \quad d = 0, 1, \tag{2.2}$$

where the asterisks in (2.2) indicate the counterfactual counterparts. Thus, the difference between the counterfactual and status quo potential outcomes is attributed entirely to the change in the observed covariates rather than more fundamental changes in the treatment mechanism. The counterfactual treatment status corresponding to X^* is denoted D^*. As pointed out by Rothe (2010), the relationship between X and X^* depends on the context and we consider two cases in this paper: (i) X and X^* are statistically independent and (ii) X^* is a deterministic transformation of X, that is, $X^* = \pi(X)$ for some known function π.

Figure 1 provides a graphical illustration of our model framework. The observed variables in the status quo environment are X, D and the potential outcome corresponding to D. In the counterfactual environment the only observed variable is X^*.

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^*,

$$\delta^* = \mathbb{E}(Y_1^*) - \mathbb{E}(Y_0^*), \tag{2.3}$$

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

$$\delta^*(\tau) = \mathbb{Q}_{Y_1^*}(\tau) - \mathbb{Q}_{Y_0^*}(\tau), \tag{2.4}$$

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

$$\delta_t^* = \mathbb{E}(Y_1^*|D^*=1) - \mathbb{E}(Y_0^*|D^*=1) \quad \text{and} \quad \delta_t^*(\tau) = \mathbb{Q}_{Y_1^*|D^*}(\tau|1) - \mathbb{Q}_{Y_0^*|D^*}(\tau|1), \tag{2.5}$$

where $D^* \in \{0,1\}$ indicates the unknown counterfactual treatment assignment, and the expectation and quantile operators are 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 6.

2.3 Identification

What makes the identification of counterfactual parameters particularly challenging is that Y_0^*, Y_1^* and D^* are all unobserved. We therefore need to employ rather strong identification assumptions which 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 conditional 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). In contrast to Hotz, Imbens, and Mortimer (2005), 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 $d = 0, 1$. That is, $F_{Y_d^*|X^*}(y|x) = F_{Y_d|X}(y|x)$ 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 the counterfactual treatment effects arises solely from the different marginal distributions of X and X^*.This condition is similar to 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). A sufficient condition for Assumption 2.2(i) is the independence between ϵ_d and X and ϵ_d and X^*, as imposed by 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.

Under Assumptions 2.1 and 2.2, the QCTE parameter is nonparametrically identified.

Lemma 1. *Suppose Assumptions 2.1 and 2.2 hold. The 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, first note that under Assumption 2.2, 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 $Y = Y_d$ 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 case can be found in Section 6.1.

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 treatment, 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:

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

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 continuous as well as discrete components, one can either adjust the kernel for frequency (sample splitting) or employ the smoothing method advocated by Li and Racine (2008).[1] 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 for $F_{Y_d^*}(y)$:

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

[1]For example, if $X = (X_1, X_2)$ with $X_1 \in \{0, 1\}$ and X_2 continuous, then 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 $\mathbb{1}\{X_1 = x_1\}$ above with $\mathbb{1}\{X_1 = x_1\} + \eta\,\mathbb{1}\{X_1 \ne x_1\}$, where $\eta \in (0, 1)$ and $\eta = \eta_n \to 0$ as $n \to \infty$.

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 the negative weights introduced by the higher-order boundary kernel. This problem can be circumvented by either 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 (2016), 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 implementing 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 asymptotic 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 derivations.

Finally, our QCTE estimator is defined as $\widehat{\delta}^*(\tau) = \widehat{\mathbb{Q}}_{Y_1^*}(\tau) - \widehat{\mathbb{Q}}_{Y_0^*}(\tau)$, where $\tau \in [0,1]$ is a quantile index and

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

3.2 Regularity Conditions

Before starting the asymptotic analysis of the proposed estimators, we gather all regularity conditions in this subsection. 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. That is, $\mathcal{X} = \prod_{s=1}^k [x_{\ell s}, x_{us}] \equiv [x_\ell, x_u]$ 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 2-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* $|\gamma| = 1, \ldots, r-1$*.*

(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.[2] 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 $\theta > 0$, 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.

[2] 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 $\widehat{F}_{Y_d^*}(y)$ is governed by the asymptotics of $\widetilde{F}_{Y_d^*}(y)$, which, in turn, 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, 3.1–3.6 hold. Then,*

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

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 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 $\lambda = 1$, 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 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, 3.1–3.6 hold. Then,*

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

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

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

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 there is a counterfactual treatment effect at the median, i.e.,

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

Given a consistent estimate of the asymptotic covariance function,

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

where $\widehat{\varrho}$ and $\widehat{\varphi}$ are provided in Section 4.1, one can simply construct an ordinary t-statistic and apply standard normal critical values.

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). 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. Next, in Section 4.2, we discuss hypothesis testing and provide step-by-step instructions for constructing uniform confidence bands.

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 uniformly consistent estimation of the functions involved in the covariance kernel $\Psi(\tau)$. The monotonicity of the estimators for $F_{Y_d^*}(y)$ and $F_{Y_d|X}(y|x)$ is also necessary for the 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 $\{U_1, \dots, U_n\}$ and $\{U_1^*, \dots U_{n^*}^*\}$ 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

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

As can be seen from (4.4) that 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^* is independent of X, we use two sets of independent multipliers U_i's and U_j^*'s to force independence between the simulated counterparts of $\widehat{\varrho}(\tau, Z_i)$ and $\widehat{\varphi}(\tau, X_j^*)$. Although $\varrho(\tau, Z)$ and $\varphi(\tau, X^*)$ are independent if $X \perp\!\!\!\perp X^*$, this does not guarantee that $\widehat{\varrho}(\tau, Z_i)$ and $\widehat{\varphi}(\tau, X_j^*)$ are also independent in finite samples. As a result, independence is enforced through the choice of independent multipliers $\{U_i\}$ and $\{U_j^*\}$.

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 requirements, 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. Then,*

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

conditional on the sample paths $\{Z_i : i = 1, 2, \dots\}$ *and* $\{X_j : 1, 2, \dots\}$ *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 is rejected if the test statistics exceed the

corresponding critical values. That is, 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 explain how to simulate the critical values via the 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 $\widehat{C}_\alpha^{\text{1-sided}}$ denote the one- and two-sided critical values that satisfy

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

That is, $\widehat{C}_\alpha^{\text{1-sided}}$ 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 lower and upper one-sided $(1-\alpha)$ uniform confidence bands are given by

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

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

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

A step-by-step procedure for constructing the uniform confidence bands is as follows:

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 $\Delta_b^u(\tau)$ as in (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 $M_{(1)} \le \ldots \le M_{(B)}$. Then 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).

4.3 Uniformly Consistent and Monotone Estimators

In this subsection 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). Moreover, to meet Assumption 4.1(i) we show that $\sup_{y\in\mathcal{Y},x\in\mathcal{X}} \left|\widehat{F}_{Y_d|X}(y|x) - F_{Y_d|X}(y|x)\right| = o_p(1)$ in Lemma 3 below.

Next, the kernel estimators for $p(x)$, $f_X(x)$ and $f_{X^*}(x)$ are 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) has been 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 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 (2016) for $\widetilde{f}_X(x)$ and $\widetilde{f}_{X^*}(x)$. For the former, we 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).[3] 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 (2016) 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.[4]

[3] 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).

[4] Despite not necessarily integrating to one for all n, the estimators in (4.10) are still uniformly consistent for $f_{X^*}(x)$ and $f_X(x)$ and hence meet Assumption 4.1(i).

Lastly, we construct the estimator for $f_{Y_d^*}(y)$ similar to $\widetilde{F}_{Y_d^*}(y)$ in (3.2). That is, 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, 3.1–3.6 hold. Also 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 $n \to \infty$. Then the estimators in* (3.3)*,* (4.7)*,* (4.9)*,* (4.10) *and* (4.12) *satisfy Assumption 4.1.*

5 Monte Carlo Simulation and Empirical Study

5.1 Simulation Study

We now examine the finite sample performance of the QCTE estimator and the multiplier bootstrap method via Monte Carlo simulations. Note that the simulation results for the QCTT are carried out here as well to verify the asymptotic properties of the treated cases which we defer to Section 6 for ease of reading.

Consider the following data generating process. Let $X = (X_1, X_2, X_3)$ be a three-dimensional random vector with each element obeying an i.i.d. standard exponential distribution truncated at 2. Let $Y = DY_1 + (1 - D)Y_0$ with $D = \mathbb{1}\{(X_1 + X_2)/2 > \varepsilon_D\}$ and

$$Y_1 = 4 + X_2 - 2X_3 + \varepsilon_{Y_1} \qquad\qquad Y_0 = 3 - \sqrt{X_2 + X_3}\varepsilon_{Y_0},$$

where ε_D, ε_{Y_1}, ε_{Y_0} are i.i.d. standard exponential distributed truncated at 1. Note that X_2 appears in both selection and outcome equations, indicating that D is conditional independent of (Y_1, Y_0) given X. We consider two counterfactual scenarios. The first one is the transformed case where $X^* = (X_1^*, X_2^*, X_3^*) = 0.5X$. The second corresponds to the independent case where

each element of X^* follows an i.i.d. standard exponential distribution truncated at 1. In each scenario, the counterfactual treatment assignment D^* and the counterfactual potential outcomes (Y_1^*, Y_0^*) are given by

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

We let status quo and counterfactual sample sizes n and n^* vary from 100, 200 and 400 with $n \geq n^*$. The number of Monte Carlo replications and the number of bootstrap repetitions are both set to be 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 $[0.2, 0.8]$ 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). In the implementation, 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 $h_d = 3\widehat{\sigma}_X n_d^{-1/7}$ where $\widehat{\sigma}_X$ is the sample standard deviation of X and $n_d = \sum_{i=1}^{n} \mathbb{1}\{D_i = d\}$ is the effective sample size. We use standard normal multipliers in both transformed and independent cases to simulate the critical values, in which the corresponding covariance functions of the QCTE/QCTT are estimated using second-order boundary kernels with rule-of-thumb bandwidths. The nominal coverage rate is set to 90%.

The simulation results are presented in Table 1 for the transformed case and Table 2 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 1 and 2, 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 increases from 100 to 400, suggesting 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, which validates the multiplier bootstrap method even when the sample size is relatively small.

Table 1: Simulation Results – Transformed Case

n	n^*	QCTE			QCTT		
		IBias	RIMSE	Coverage Rate	IBias	RIMSE	Coverage Rate
100	100	0.034	0.209	0.876	-0.051	0.212	0.895
200	100	0.029	0.152	0.883	-0.048	0.157	0.902
200	200	0.020	0.147	0.890	-0.047	0.152	0.890
400	100	0.026	0.112	0.871	-0.039	0.118	0.910
400	200	0.025	0.107	0.867	-0.032	0.109	0.905
400	400	0.028	0.108	0.860	-0.036	0.109	0.903

Reported are Monte Carlo estimates among 1000 replications. In each replication, the QCTE and QCTT are estimated over 100 equidistant grid points in $[0.2, 0.8]$ using fourth-order boundary Epanechnikov kernel with bandwidth $h_d = 3\widehat{\sigma}_X n_d^{-1/7}$, where $\widehat{\sigma}_X$ is the sample standard deviation and n_d is the effective sample size. For the coverage rates of the two-sided uniform confidence bands with nominal level of 90%, the simulated critical values are calculated with 1000 bootstrap repetitions using standard normal multipliers. The covariance functions are estimated using second-order boundary Epanechnikov kernels with rule-of-thumb bandwidths.

Table 2: Simulation Results – Independent Case

n	n^*	QCTE			QCTT		
		IBias	RIMSE	Coverage Rate	IBias	RIMSE	Coverage Rate
100	100	0.053	0.203	0.893	-0.018	0.201	0.917
200	100	0.041	0.143	0.932	-0.018	0.143	0.935
200	200	0.042	0.141	0.908	-0.020	0.139	0.938
400	100	0.044	0.110	0.904	-0.007	0.105	0.953
400	200	0.037	0.105	0.896	-0.011	0.104	0.945
400	400	0.040	0.103	0.904	-0.012	0.100	0.940

Reported are Monte Carlo estimates among 1000 replications. In each replication, the QCTE and QCTT are estimated over 100 equidistant grid points in $[0.2, 0.8]$ using fourth-order boundary Epanechnikov kernel with bandwidth $h_d = 3\widehat{\sigma}_X n_d^{-1/7}$, where $\widehat{\sigma}_X$ is the sample standard deviation and n_d is the effective sample size. For the coverage rates of the two-sided uniform confidence bands with nominal level of 90%, the simulated critical values are calculated with 1000 bootstrap repetitions using standard normal multipliers. The covariance functions are estimated using second-order boundary Epanechnikov kernels with rule-of-thumb bandwidths.

5.2 Empirical Application

This section demonstrates the usefulness of our proposed methods by investigating two possible explanations regarding the heterogeneous effects of Job Corps on earnings distribution. Job Corps is the largest and most comprehensive education and job training program in the United

States. Established in 1964, the purpose is to assist at-risk youths between the ages of 16 and 24. In 2008, Job Corps enrolled about 60,000 students at a total cost of more than $1.5 billion funded by the federal government. Eligible participants received academic, vocational, and social training services at over 110 centers nationwide for an average of 8–9 months.

To credibly estimate the impact of Job Corps on earnings distribution, we exploit data from the National Job Corps Study (NJCS), a randomized experiment conducted during 1994–1998. In NJCS research sample eligible applicants were randomly assigned to a control or treatment group, where the control members were barred from enrolling in Job Corps for 3 years, while the treatment group could enroll in Job Corps. The NJCS data is publicly available and has been extensively studied by Schochet, Burghardt, and McConnell (2008), Lee (2009), Flores, Flores-Lagunes, Gonzales, and Neumann (2012), Flores and Flores-Lagunes (2013) and Eren and Ozbeklik (2014) among others.[5,6] In particular, Eren and Ozbeklik (2014) evaluate the distributional impact of Job Corps and show a great deal of heterogeneity along the earnings distribution. Their findings suggest that Job Corps is not effective for males at the bottom of the distribution, and two plausible explanations are also given: (i) strong economic performance in the era of NJCS and (ii) relatively low skill endowment for some subgroups.

The first explanation hinges on the U.S. economic boom in the late 1990s. With inflation and unemployment rates both reaching their lowest in the second half of the 20th century, this economic boom generated large improvements for individuals at the bottom of the earnings distribution, especially for those who did not participate in any labor market program. As a result, the gains of Job Corps may be offset by the benefits from strong economic conditions for the lower quantiles. The second explanation hypothesizes that Job Corps only works well for those with relatively higher education. In other words, it may simply be the case that Job Corps is ineffective for those who are at the bottom of the skill distribution.

We propose to verify these two hypotheses based on their implications for external validity. First, suppose the strong economic conditions explanation is true, one should see a similar pattern in evaluating other U.S. training programs implemented in the same period (i.e., late 1990s). Unfortunately, information about such programs is sometimes not available. We therefore construct a contemporaneous counterfactual program based on past experience—the National Supported Work Demonstration (NSW). The NSW, targeting disadvantaged workers aged 17–55, was implemented during 1975–1979 and is also a benchmark dataset studied by LaLonde (1986), Dehejia and Wahba (1999, 2002), Smith and Todd (2005) and Firpo (2007) among others.[7] We incorporate the NSW data to estimate the counterfactual effect of Job Corps that would have prevailed had the members of the NSW participated in Job Corps. Note that the individual characteristics between Job Corps and NSW are likely to be independent in this case since the data are collected two decades apart.

However, even if the counterfactual program effect is found to follow a similar pattern as that

[5]See Schochet, Burghardt, and Glazerman (2001) for more details about Job Corps and NJCS.
[6]The NJCS data can be downloaded from `http://qed.econ.queensu.ca/jae/datasets/eren001/`.
[7]The NSW data can be downloaded from `http://users.nber.org/~rdehejia/data/nswdata2.html`.

of Job Corps, one cannot identify whether the similarity is attributed to the same evaluation period or other similar nature between the actual and counterfactual programs. We then tackle this issue by reversing the roles of Job Corps and NSW to estimate the counterfactual program effect as if the Job Corps cohort were to participate in the NSW which was implemented in the mid-1970s. If the Job Corps cohort would receive a significant positive effect in a different time period and the NSW cohort would receive nothing in the late 1990s, it is very likely that the ineffectiveness of Job Corps is due to the strong economic performance observed in the same era of NJCS.

The second explanation presumes that low skill or low education would lead to weak program performance. To examine this hypothesis, we propose to focus on individuals who do not benefit from Job Corps and artificially give them higher education. We then estimate the ceteris paribus program effect given new education level to see if there is a significantly increased effect as predicted by the hypothesis. We argue that this approach would be better than comparing subgroup effects by educational attainment in that a clear causal interpretation can be made given all other things being equal. It is also understood that this manipulation would correspond to the transformed case, where the new education level is a deterministic transformation of the original one.

Before proceeding any further, we first describe the summary statistics of the NJCS and NSW samples in Table 3. Note that we focus on male individuals with ages between 17 and 24 to meet Assumption 2.2(ii). Individual characteristics include age, race and education. The outcome variable is the average weekly earnings in year 4, which is the second year of the post-program period.[8] As shown by Table 3, there are barely no differences in characteristics between participants and non-participants in NJCS, whereas the NSW members tend to be older and mostly non-white with lower education. It is also worth noting that the NJCS sample is shown to be unconfounded by exploiting randomized program eligibility and the test proposed by Donald, Hsu, and Lieli (2014a). More specifically, we show in Table 4 that the unconfoundedness cannot be rejected given the characteristics mentioned above under various specifications. We take this as a justification even when a relatively small conditional set is considered. Finally, note that the NSW sample is also unconfounded given the fact that it is drawn from a field experiment where individuals were randomly assigned to participate in the program.

[8]We divide earnings in 1978 by 52 weeks for the average weekly earnings of NSW.

Table 3: Descriptive Statistics

Variable	Job Corps		NSW	
	Participation	No	Participation	No
Average Weekly Earnings	257.116	236.626	105.942	99.344
	(216.592)	(209.226)	(96.268)	(110.402)
Age	18.995	19.017	20.202	19.967
	(1.915)	(1.891)	(2.231)	(2.205)
Race (Non-White=1)	0.607	0.613	0.905	0.908
	(0.489)	(0.487)	(0.294)	(0.289)
Education (High School=1)	0.260	0.265	0.244	0.113
	(0.439)	(0.442)	(0.431)	(0.317)
Sample Size	1918	2589	168	240

Both samples include males aged 17–24. The average weekly earnings of the NSW sample is calculated by dividing earnings in 1978 by 52 weeks. Standard deviations are in parentheses.

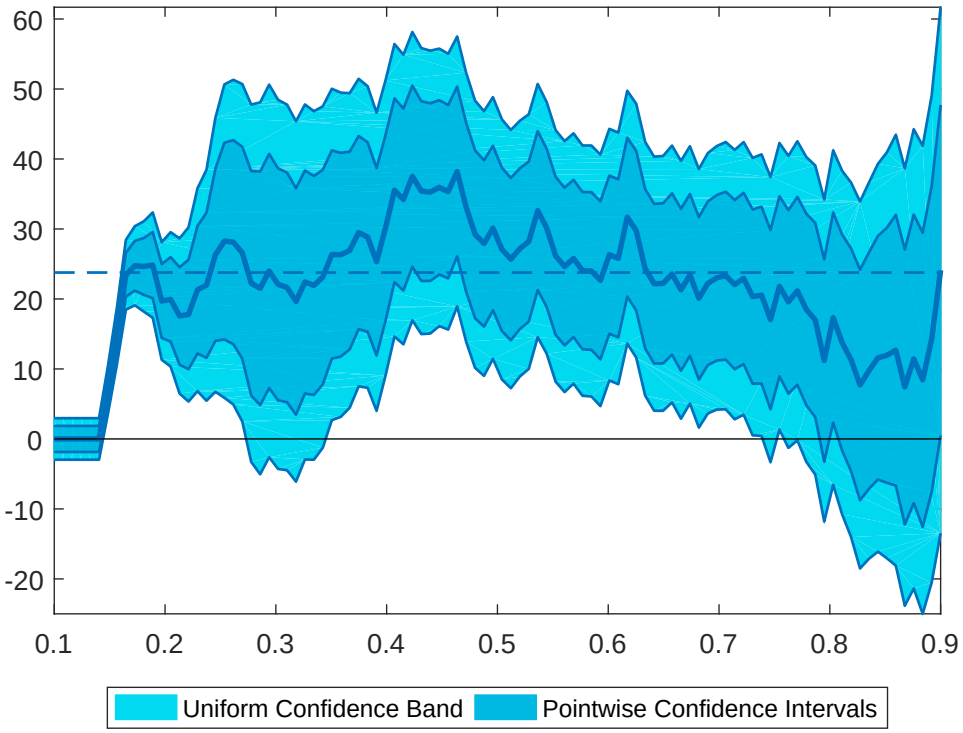
Table 4: Test for Unconfoundedness

Specification	Covariates	p-value
Linear	Age, Race, Edu	0.179
Interaction	Age, Race, Edu, Age×Race, Age×Edu, Race×Edu	0.186
Quadratic	Age, Race, Edu, Age×Race, Age×Edu, Race×Edu, Age^2	0.193

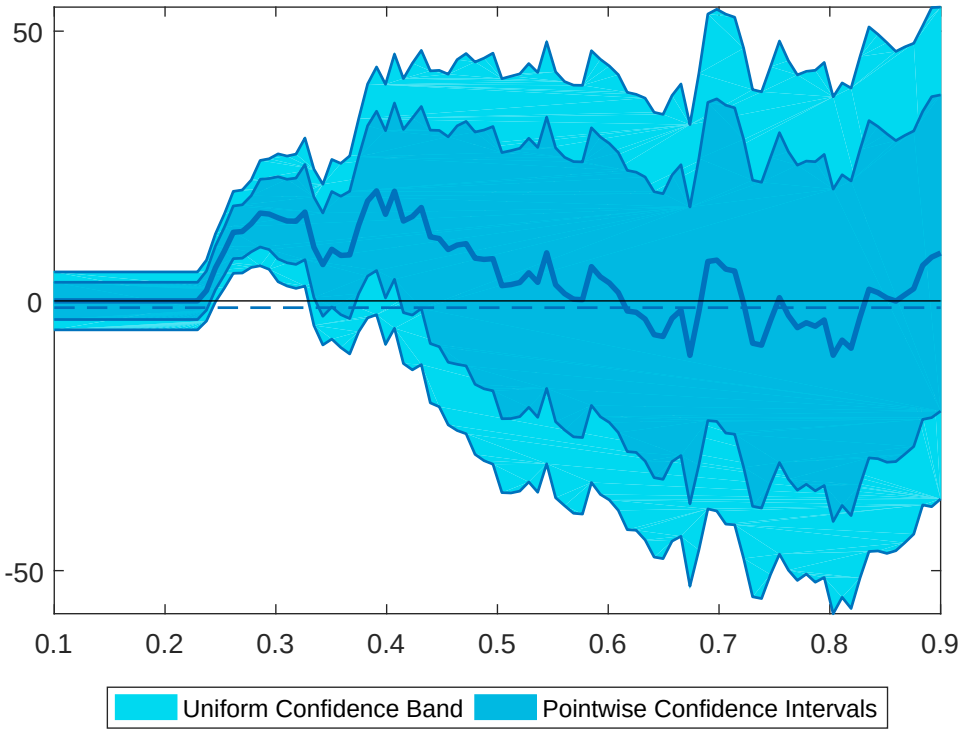
We use NJCS sample of males aged 17–24 and series logit estimation for the computation of the propensity scores.

Using the same specifications as for the simulation study, our main empirical findings are threefold. First, Figure 2 illustrates the actual program effects of Job Corps and NSW on earnings distributions. In the case of Job Corps, Figure 2(a) generalizes the local QTE in Eren and Ozbeklik (2014) to the overall QTE under unconfoundedness. The thick line represents the QTE estimates with light shaded area 90% uniform confidence band and dark shaded area 90% pointwise confidence intervals, as well as the horizontal dashed line the ATE estimate. Along the earnings distribution, it can be first seen that the flat zero QTEs (and the confidence band/intervals) for the lower tail reflect that some individuals remain unemployed in the post-Job Corps period.[9] After that, the program effect suddenly jumps to \$25 along with uniform and statistically significant QTEs between the 15th and 25th quantiles. However, Job Corps is not effective for males around 30th quantile according to the uniform confidence band. This is

[9]Roughly 14% (17%) of the participants (non-participants) report zero weekly earnings in year 4.

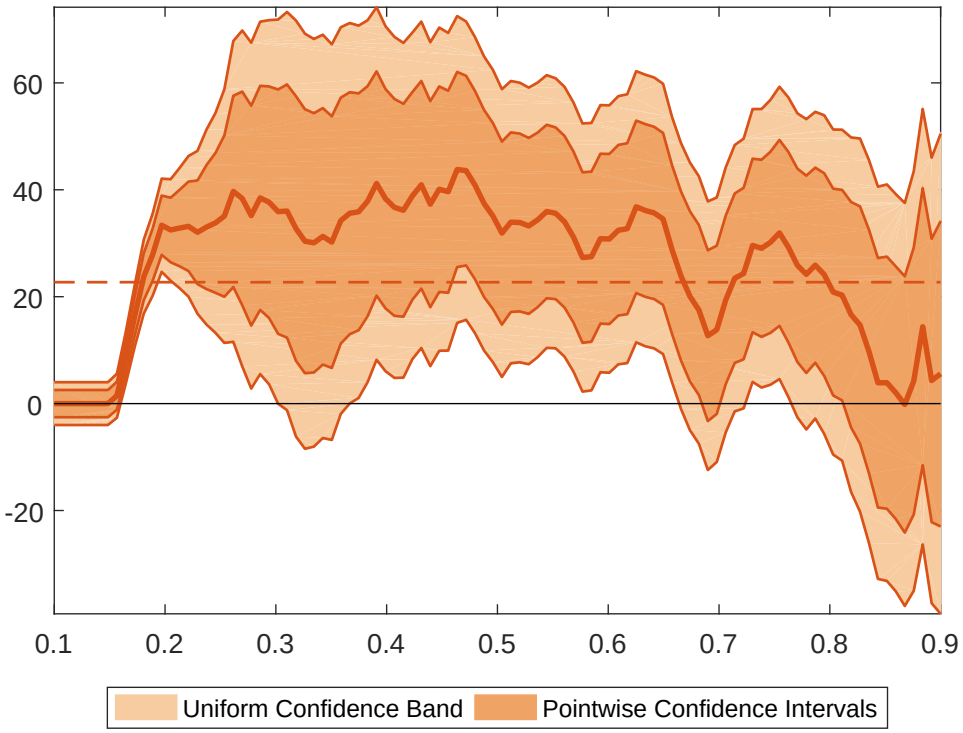


(a) Job Corps

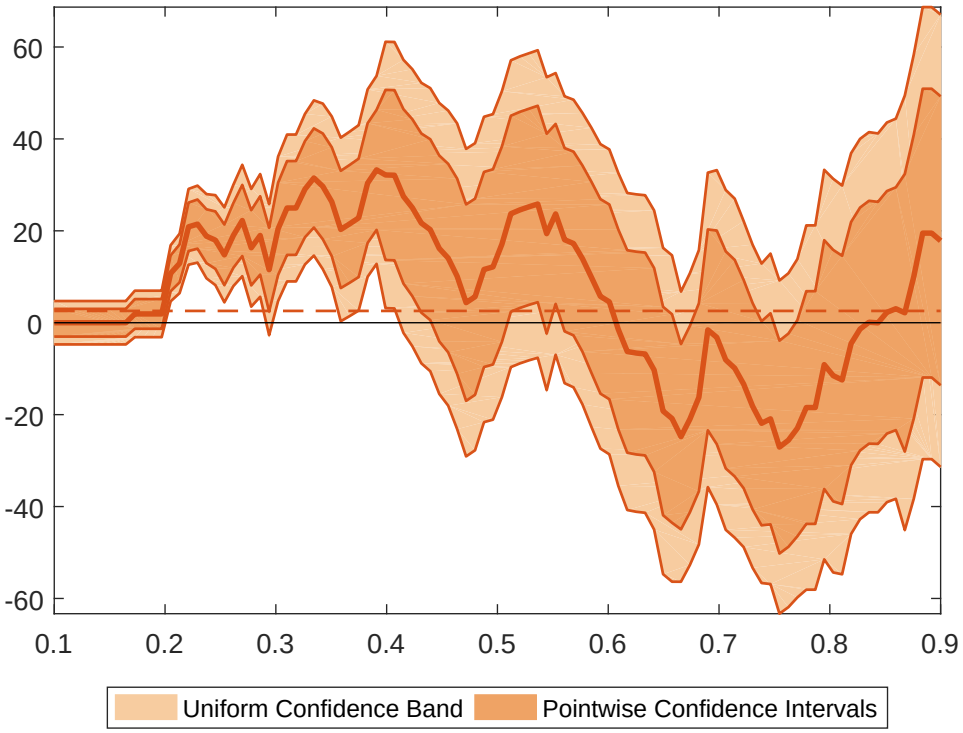


(b) NSW

Figure 2: Program Effects on Earnings Distributions



(a) Job Corps Program with NSW Cohort



(b) NSW Program with Job Corps Cohort

Figure 3: Counterfactual Program Effects (Strong Economic Conditions)

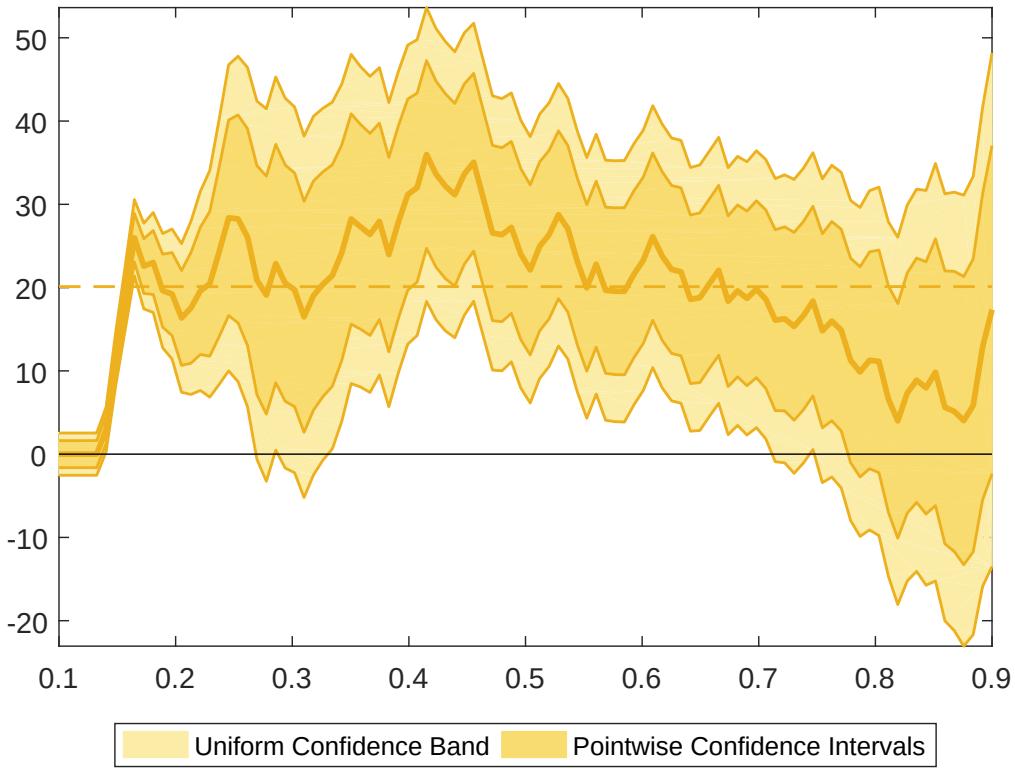


Figure 4: Counterfactual Program Effect (Skill Hypothesis)

consistent with the findings in Eren and Ozbeklik (2014) whereas the same conclusion cannot be reached from the pointwise confidence intervals. Finally, the QTEs are again significantly positive between the 35th and 75th quantiles, suggesting a great heterogeneity in the impact of Job Corps. In Figure 2(b) we depict the effects of NSW similarly using a sample with ages between 17 and 24. Interestingly, opposite pattern is observed in that the NSW is only effective for individuals around the third decile of the earnings distribution and ineffective elsewhere.

Next, we present the counterfactual program effects in Figure 3. Specifically, Figure 3(a) shows the effect of Job Corps that would have prevailed had the NSW cohort participated in Job Corps. It can be readily seen from Figures 2(a) and 3(a) that the ineffectiveness around the 30th quantile exists in both cases, yielding supportive evidence for the strong economic conditions hypothesis. In addition, we focus on the Job Corps cohort and depict the counterfactual program effect as if they were to participate in the NSW in Figure 3(b). A simple comparison between Figures 2(b) and 3(b) indicates that this counterfactual program performs better than the original NSW in that a significant impact around \$20 between the 20th and 40th quantiles can be found. This finding not only suggests the Job Corps cohort would have benefited from another job training program implemented in a different time period, but also ensures that the strong economic conditions indeed explains (part of) the ineffectiveness.

Finally, the result regarding the skill hypothesis is presented in Figure 4. In this counterfactual exercise, we manipulate education level of individuals with earnings between the 25th and 35th quantiles as if all of them were to graduate from high schools. To be more precise, we assign high school education to 255 non-graduates out of 450 individuals whose average weekly earnings are between \$71 and \$150. By giving extra education, we expect to see a significant program effect around the 30th quantile1, which is exactly the case in Figure 4. Put differently, our finding suggests that skill hypothesis could be one of the explanation of the weak performance of Job Corps, as proposed by Eren and Ozbeklik (2014).

6 Theoretical Extensions: The ACTE and the Counterfactual Treated Subpopulation

In addition to the QCTE, we extend the theoretical analysis to various counterfactual program effects here which may also be appealing to practitioners and policymakers. We first discuss how to predict mean program impact given counterfactual covariates in Section 6.1. The case for counterfactually treated subpopulation is then covered in Section 6.2.

6.1 The ACTE

Since the ACTE defined in (2.3) only depends on the means but not the entire distributions of Y_d^* and Y_1^*, the identification assumptions can be weakened as follows.

Assumption 6.1 (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 6.2 (Invariance of Conditional Means)**.**

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

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

Similar to Lemma 1, the ACTE is identified under Assumptions 6.1 and 6.2 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).[10] That is,

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

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

Assumption 6.4 (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 6.1–6.4 hold. Then,*

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

[10] One can also adopt Hahn's (1998) series approach for conditional mean $\breve{\mathbb{E}}(Y_d|X = x) = \{\breve{p}(x)^d[1 - \breve{p}(x)]^{1-d}\}^{-1}\breve{\mathbb{E}}[Y\,\mathbb{1}\{D = d\}|X = x]$ 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 of estimating propensity score in this step.

where $\mathbb{V}_{\delta^*} = \mathbb{E}\big[\varrho_{\delta^*}(Z)\big]^2 + \mathbb{E}\big[\varphi_{\delta^*}(X^*)\big]^2$ *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 $n^{-1/4}$ and can be neglected in the final estimation. Second, $\varphi_{\delta^*}(X^*)$ accounts for the uncertainty in replacing expectation with 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),

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

it is true that $\mathbb{V}_{\delta^*}$ attains this bound if $X^* = X$ and $\lambda = 1$. Put it 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 4.3 into $\varrho_{\delta^*}(x)$, one can apply 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 the ACTE.

6.2 The Treated Cases

We focus on ACTT and QCTT defined in (2.5) in this subsection. 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 6.5 (Unconfoundedness for the Untreated)**.**

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

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

Assumption 6.6 (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 6.7 (Invariance of Propensity Scores)**.** $p^*(x) = p(x)$ *for all* $x \in \mathcal{X}^*$.

It can easily be seen that Assumptions 6.5 and 6.6 are weaker than their counterparts Assumptions 2.1 and 2.2.[11] To identify treated parameters, however, we need to invoke Assumption 6.7 so that the counterfactual treatment assignment can be determined: It requires the counterfactual participation status must be consistent with the status quo one for individuals who are observationally equivalent between counterfactual and status quo environments. Given these assumptions, the ACTT and QCTT can be identified as follows.

Lemma 4. *Suppose Assumptions 6.5–6.7 hold. The ACTT and QCTT are identified by*

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

According to Lemma 4, the 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 6.8 (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 2-times differentiable on* $\mathcal{Y}$.

Corollary 2. *Suppose Assumptions 3.1, 3.2, 3.4–3.6 and 6.5–6.8 hold. Then,*

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

[11] They can be further weakened for the ACTT case similar to Section 6.1.

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

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

Moreover,

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

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

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

where $\varrho_{d,t}^F(y, Z)$ *and* $\varphi_{d,t}^F(y, X^*)$ *are given by*

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

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,

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

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 $\Delta_t^u(\cdot) \overset{p}{\Rightarrow} \Delta_t(\cdot)$ similar to Theorem 2. We omit the details for brevity.

7 Conclusion

This paper proposes a unified nonparametric approach to the identification and estimation of quantile treatment effects 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 external validity, a researcher or policy maker can formally estimate the new treatment effect 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 about the counterfactual quantile 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 apply the proposed methods to study the heterogeneous impact of the Job Corps training program in the U.S. along the earnings distribution. In particular, 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) individuals toward the bottom of the earnings distribution have insufficient skill to benefit from the program. We examine the implications of these hypotheses for the external validity of the original finding, and demonstrate that (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 cohort would be larger had they participated in another training program implemented in different period of time; (iii) the effectiveness of Job Corps would improve if individuals with low earnings were given extra education. Taken together, these findings suggest that both explanations contribute to the documented weak performance of Job Corps for individuals with low earnings.

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. □

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,

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

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

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

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

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

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 $\widehat{F}_{Y_d|X}(y|x) = \widehat{g}_d(y,x)/\widehat{g}_d(x)$,

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

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

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

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

(A.6)

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

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

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

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 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 have that uniformly over $y \in \mathcal{Y}$,

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

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.

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

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

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

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

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

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

and the asymptotic linear representation is as in (A.1). Next, since $\mathbb{1}\{Y \le y\}$ is a type I function and the rest 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 $y \in \mathcal{Y}$. Also by van der Vaart (2000, p.270) that the Cartesian product of two Donsker classes of functions is still a Donsker class, Lemma 2 holds by the functional central limit theorem for $\widetilde{\boldsymbol{F}} = \left(\widetilde{F}_{Y_0^*}, \widetilde{F}_{Y_1^*}\right)^T$ 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 $\sup_{y\in\mathcal{Y}}\left|\widehat{F}_{Y_d^*}(y) - \widetilde{F}_{Y_d^*}(y)\right| = o_p(n^{-1/2})$ as in (A.2). For simplicity assume that $\widetilde{F}_{Y_d^*}(0) \ge 0$ so that $\phi_1(\widetilde{F}_{Y_d^*})(y) = \sup_{y'\le y}\widetilde{F}_{Y_d^*}(y')$ for all $y \in \mathcal{Y} = [0,\bar{y}]$.[12] From the first part of the proof it is true that $\sup_{y\in\mathcal{Y}}\left|\sqrt{n}\big(\widetilde{F}_{Y_d^*}(y) - F_{Y_d^*}(y)\big)\right| = O_p(1)$, implying for any $\epsilon_1 > 0$ there exist $M > 0$ and large $N = N_M$ such that for all $n > N$,

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

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 $\rho(y_1,y_2) = \left|F_{Y_d^*}(y_1) - F_{Y_d^*}(y_2)\right|^{1/2}$ for all $(y_1,y_2) \in \mathcal{Y}$ from Theorem 3.1 of Hsu, Lieli, and Lai (2016), meaning that for any $\epsilon_2 > 0$ and $\epsilon_3 > 0$, there exists $\delta > 0$ small enough and N_δ large enough that for all $n > N_\delta$,

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

If we pick a large N such that

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

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

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

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

[12] If this is not the case, we can always redefine $\mathcal{Y} = [-\epsilon,\bar{y}]$ such that $\widetilde{F}_{Y_d^*}(-\epsilon) \ge 0$ 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 (2016), 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),

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

□

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 $\mathcal{Z}_{\ell_n} \equiv \left\{\omega \in \mathcal{Z} : \sup_{y\in\mathcal{Y}, z\in\mathcal{Z}} \left|\widehat{\varrho}_{d,\ell_n}^F(y, z)(\omega) - \varrho_d^F(y, z)\right| \to 0\right\}$, 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 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 the monotonicity $\widehat{F}_{Y_d|X,\ell_n}(y|x)$ satisfies Pollard's entropy condition as in (4.2) of Andrews (1994) with 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 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 sample path $\{\omega \in X^*_j : j = 1, 2, \dots\}$. 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, or

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

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,

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

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 so is omitted. Finally, (A.22) holds because

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

□

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, then for $y'\leq y<y'+\epsilon$, $\epsilon>0$, $\widehat{F}_{Y_d|X}(y|x)=\widetilde{F}_{Y_d|X}(y'-\epsilon|x)>\widetilde{F}_{Y_d|X}(y|x)$ and for $y'-\epsilon\leq y<y'$, $\widehat{F}_{Y_d|X}(y|x)=\widetilde{F}_{Y_d|X}(y'-\epsilon|x)$. Next, for $y'\leq y<y'+\epsilon$, if $\widehat{F}_{Y_d|X}(y|x)\leq F_{Y_d|X}(y|x)$, 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 $y'\leq y<y'+\epsilon$,

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

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

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 $\sup_{\{x:\widetilde{f}_X(x)\geq b_n\}}\left|\widehat{f}_X(x)-f_X(x)\right|=\sup_{\{x:\widetilde{f}_X(x)\geq b_n\}}\left|\widetilde{f}_X(x)-f_X(x)\right|\leq\sup_{x\in\mathcal{X}}\left|\widetilde{f}_X(x)-f_X(x)\right|=o_p(1)$. Similar to Lemma 3.2 in Donald, Hsu, and Barrett (2012), for all x

such that $\widetilde{f}_X(x) < b_n$ and 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 rest 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. □

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. □

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 6.6 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 6.5(i) and by Bayes' theorem. The fourth equality is true given Assumption 6.7. Since we can observe Y, X, D, X^* and $p(x) > 0$ for all $x \in \mathcal{X}$ by Assumption 6.5(ii), the last line is well-defined and identified. Thus, the ACTT and QCTT can be identified as well. □

Proof of Corollary 2:

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

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 $Y_{(n+1)} = \bar{y}$. That is, 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 $y \in \mathcal{Y}$.

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.

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 (2016): "The State of Applied Econometrics – Causality and Policy Evaluation," arXiv preprint arXiv:1607.00699.

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

Bertanha, M. and G. W. Imbens (2016): "External Validity in Fuzzy Regression Discontinuity Designs" NBER Working Paper 20773, New York: National Bureau of Economic Research.

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. H. and S. Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, **94**, 1053–1062.

Dehejia, R. H. and S. Wahba (2002): "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *Review of Economics and statistics*, **84**, 151–161.

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.

Flores C, A. Flores-Lagunes, A. Gonzales, and T. C. Neumann (2012): "Estimating the Effects of Length of Exposure to Instruction in a Training Program: The Case of Job Corps," *Review of Economics and Statistics*, **941**, 153–171.

Flores, C. A. and A. Flores-Lagunes (2013): "Partial Identification of Local Average Treatment Effects with an Invalid Instrument," *Journal of Business and Economic Statistics*, **31**, 534–545.

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

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

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. 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 (2016): "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, 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.

LaLonde R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, **76**, 604–620.

Lee D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, **76**, 1071-1102.

Li, Q. and J. Racine (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, Princeton, NJ.

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. Z., 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.

Smith, J. A. and P. E. Todd (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of econometrics*, **125**, 305–353.

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.