Testing for Unobserved Heterogeneous Treatment Effects with Observational Data

Yu-Chin Hsu
Institute of Economics, Academia Sinica;
Department of Finance, National Central University;
Department of Economics, National Chengchi University

Ta-Cheng Huang
Global Asia Institute, National University of Singapore

Haiqing Xu
Department of Economics, University of Texas at Austin

November 9, 2020

Abstract

Unobserved heterogeneous treatment effects have been emphasized in the recent policy evaluation literature (see e.g., Wager and Athey, 2018). This paper proposes a nonparametric test for unobserved heterogeneous treatment effects in a treatment effect model with a binary treatment assignment, allowing for individuals' self–selection to the treatment. Under the standard *local average treatment effects* assumptions, i.e., the no defiers condition, we derive testable model restrictions for the hypothesis of unobserved heterogeneous treatment effects. Also, we show that if the treatment outcomes satisfy a monotonicity assumption, these model restrictions are also sufficient. Then, we propose a modified Kolmogorov-Smirnov-type test which is consistent and simple to implement. Monte Carlo simulations show that our test performs well in finite samples. For illustration, we apply our test to study heterogeneous treatment effects of the Job Training Partnership Act on earnings and the impacts of fertility on family income, where the null hypothesis of homogeneous treatment effects gets rejected in the first case but fails to be rejected in the second application.

Keywords: endogenous treatment assignment, local average treatment effects, nonseparable model, unobserved heterogeneous treatment effects

1 Introduction

Heterogeneous treatment effects due to unobserved latent variables have been emphasized in the policy evaluation literature. See e.g. Imbens and Angrist (1994), Heckman, Smith, and Clements (1997), Heckman and Vytlacil (2001, 2005), Abadie, Angrist, and Imbens (2002), Abadie (2002, 2003), Blundell and Powell (2003) , Matzkin (2003), Chesher (2003, 2005), Chernozhukov and Hansen (2005), Florens, Heckman, Meghir, and Vytlacil (2008), Imbens and Newey (2009), Frölich and Melly (2013), D'Haultfœuille and Février (2015), Torgovitsky (2015), and Wager and Athey (2018), among many others. In the empirical study of treatment effects using observational data, the interpretation of the widely used instrumental variable (IV) estimation relies on the key hypothesis that after controlling for covariates, treatment effects are homogeneous across individuals. In the presence of unobserved heterogeneous treatment effects, the standard IV approach estimates the *local average treatment effects* (LATE), rather than the *average treatment effects* (ATE); see Imbens and Angrist (1994) and Imbens (2010).

In this paper, we develop a nonparametric test for the (unobserved) heterogeneous treatment effects. We use a nonparametric and nonseparable model, i.e., the error terms are not additively separable from the treatment indicator, to model heterogeneous treatment effects. Together with the endogeneity issue, introduced due to individuals' self-selection to treatment in observational data, it is well known in the literature that identification and estimation of nonseparable models are challenging. On the other hand, the homogeneous treatment effects assumption substantially simplifies econometrics analysis of treatment effects, since it implies the ATE is the same as the LATE, after controlling for observed heterogeneity (i.e., covariates), if the former is the main object of research interest. For instance, Angrist and Krueger (1991) use a two–stage least squares approach to estimate treatment effects of compulsory schooling on earnings. Therefore, by testing for heterogeneous treatment effects, our method assesses whether the complicated nonparametric and nonseparable treatment effect models is more appropriate (than e.g. the two–stage least squares approach) for a program evaluation assignment.

Though important, there are only a handful of papers on testing for such unobserved heterogeneity.[1] In the context of ideal social experiment data, i.e., lack of endogeneity,

[1] There exists a substantial literature for testing observed heterogeneity, i.e., whether (conditional) average

Heckman, Smith, and Clements (1997) develop a lower bound for the variance of heterogeneous treatment effects, thereby providing a test for whether or not the data are consistent with the identical treatment effects model. Moreover, Hoderlein and Mammen (2009) discuss specification tests for endogeneity as well as unobserved heterogeneity in nonseparable triangular models. Recently, Lu and White (2014) and Su, Tu, and Ullah (2015) establish nonparametric tests for unobserved heterogeneous treatment effects under the unconfoundedness assumption. In particular, Lu and White (2014) test unobserved heterogeneity in treatments effects via testing an equivalent independence condition on observables. Another closely related paper is Heckman, Schmierer, and Urzua (2010) who test the absence of self-selection on the gain to treatment in the generalized Roy model framework, allowing for (unobserved) heterogeneous treatment effects. Furthermore, our paper is also related to Heckman and Vytlacil (2005), who suggest an approach to test heterogeneity of the *marginal treatment effects* (MTE). Our test object of interest, however, focuses on whether there exists individual-level unobserved heterogeneity in treatment effects, rather than group-level (defined by a margin) unobserved heterogeneity, i.e., whether MTE varies across margins.

Motivated by Lu and White (2014), we show that in the presence of endogeneity, model restrictions arising from the homogeneous treatment effects hypothesis can also be characterized by a set of independence conditions that involve the LATE estimation. These testable implications are related to important literature on testing whether the distribution of potential outcomes is affected by the treatment. In the LATE framework, Abadie (2002) considers the null hypothesis of the equality between distribution functions of treatment and control groups, and also first-order and second-order stochastic dominance of the two distribution functions. Later, Lee and Whang (2009) and Chang, Lee, and Whang (2015) generalize Abadie (2002)'s test for distributional effects by allowing for treatment effects heterogeneity via covariates.[2] Our test problem differs from that literature in that we investigate whether the distribution function of the treatment group is an (unknown) constant

treatment effects vary across different subpopulations defined by observed covariates. For example, see Heckman, Smith, and Clements (1997), Crump, Hotz, Imbens, and Mitnik (2008), Chang, Lee, and Whang (2015), Abrevaya, Hsu, and Lieli (2015), Athey and Imbens (2016), Hsu (2017), and Lee, Okui, and Whang (2017), among many others.

[2]See also e.g., Jun, Lee, and Shin (2016) and Hsu (2017) for further extensions, and references therein.

shift from the control group's distribution. Specifically, the equality hypothesis on the two distribution functions is a special case of our test implication.

Nonparametric tests for conditional independence restrictions have been well studied in different contexts. See, e.g., Andrews (1997), Dauxois and Nkiet (1998), Su and White (2007, 2008, 2014), Huang (2010), Bouezmarni and Taamouti (2014), Hoderlein and White (2012), Linton and Gozalo (2014), and Huang, Sun, and White (2016), among many others. When one considers testing independence restrictions of variables that are nonparametrically constructed, however, a key technical issue arises in the case, in particular, where the nonparametric components are functions of continuous covariates (see, e.g., Lu and White, 2014). Motivated by Stinchcombe and White (1998), we modify the classic Kolmogorov–Smirnov tests by using the primitive function of CDF's. Such a modification is novel and plays a key role in our approach. Moreover, we establish the asymptotic properties of the proposed tests under the null and alternative hypotheses.

The remainder of this paper is organized as follows. In Section 2, we introduce the model and derive testable model restrictions. Section 3 discusses our test statistics and their asymptotic results. We distinguish whether the covariates include continuous variables. In Section 4, we conduct Monte Carlo experiments to study the finite-sample performance of the proposed test. Section 5 illustrates our testing approach by two empirical applications. All proofs are collected in the Appendix.

2 Model and Testable Restrictions

We consider the following nonseparable treatment effect model:

$$Y = g(D, X, \epsilon) \tag{1}$$

where $Y \in \mathrm{R}$ is the outcome variable, $D \in \{0, 1\}$ denotes the treatment status, $X \in \mathrm{R}^{d_X}$ is a vector of covariates, ϵ is an unobserved random disturbance of general form (e.g., without invoking any restriction on the dimensionality of ϵ), and g is an unknown but smooth function defined on $\{0, 1\} \times \mathcal{S}_{X\epsilon}$.[3] In particular, the treatment variable D is allowed to be correlated with ϵ to allow for selection to the treatment; see, e.g., Heckman, Smith, and Clements (1997). To deal with endogeneity, we introduce a binary instrumental variable

[3] For a generic random vector A, we use $\mathcal{S}_A$ to denote the support of A.

$Z \in \{0, 1\}$. Throughout the paper, we use upper case letters to denote random variables, and their corresponding lower case letters to stand for realizations of random variables.

As is motivated in the seminal paper by Matzkin (2003), the non-additivity of the structural relationship g in ϵ captures the idea of unobserved heterogeneous treatment effects in that the individual treatment effect, $g(1, X, \epsilon) - g(0, X, \epsilon)$, would depend on the unobserved individual heterogeneity ϵ, even after controlling for covariates X. Therefore, we have the following proposition.

Proposition 1 *Suppose* (1) *holds, then the homogeneous treatment effects hypothesis, i.e., for some measurable function* $\delta(\cdot) : \mathcal{S}_X \mapsto \mathrm{R}$*,*

$$\mathcal{H}_0 : \ g(1, X, \cdot) - g(0, X, \cdot) = \delta(X) \tag{2}$$

holds if and only if the structural relationship g *is additively separable in* ϵ *(w.r.t.* D*), i.e.,*

$$g(D, X, \epsilon) = m(D, X) + \nu(X, \epsilon), \tag{3}$$

where $m : \mathcal{S}_{DX} \mapsto \mathrm{R}$ *and* $\nu : \mathcal{S}_{X\epsilon} \mapsto \mathrm{R}$*.*

Proposition 1 directly follows Lu and White (2014). Note that if (3) holds, $\delta(x) = m(1, x) - m(0, x)$ in (2), which is the homogenous individual treatment effects across individuals with covariates $X = x$.

A key insight from Lu and White (2014) is that they further show the equivalence between the additive separability hypothesis and a conditional independence restriction on observables. In the presence of treatment endogeneity, we derive a similar result. For each $x \in \mathcal{S}_X$ and $z \in \{0, 1\}$, let $p(x, z) = \mathrm{P}(D = 1|X = x, Z = z)$ be the *propensity score.*

Assumption 1 *Suppose* $Z \perp\!\!\!\perp \epsilon | X$*. For all* $x \in \mathcal{S}_X$*,* $\mathbf{P}(Z = 1|X = x)$ *is bounded away from zero and one, and* $p(x, 0) \neq p(x, 1)$*.*

Assumption 1 is standard in the literature and requires the instrumental variable Z to be (conditionally) exogenous and relevant. See, e.g., Imbens and Angrist (1994) and Chernozhukov and Hansen (2005). Throughout, we maintain Assumption 1.

Moreover, let $\mu(x, z) = \mathbf{E}(Y|X = x, Z = z)$. Under $\mathcal{H}_0$ and Assumption 1, we have

$$\begin{aligned} \mu(x, z) &= \mathbf{E}\left[g(0, X, \epsilon) + \delta(X) \times D | X = x, Z = z\right] \\ &= \mathbf{E}\left[g(0, X, \epsilon)|X = x\right] + \delta(x) p(x, z), \quad \text{for} \quad z = 0, 1. \end{aligned}$$

In the above system of equations, we treat $\mathbf{E}[g(0,X,\epsilon)|X=x]$ and $\delta(x)$ as two unknowns. Solving the equations, we then identify LATE $\delta(x)$ as follows:

$$\delta(x)=\frac{\mu(x,1)-\mu(x,0)}{p(x,1)-p(x,0)}=\frac{\mathrm{Cov}(Y,Z|X=x)}{\mathrm{Cov}(D,Z|X=x)}. \tag{4}$$

See Imbens and Angrist (1994) for the LATE interpretation of (4). Note that $\delta(x)$ is well defined given $p(x,0)\neq p(x,1)$ under Assumption 1, and identified as well directly from the data regardless of the monotonicity of the selection.

Let $W\equiv Y+(1-D)\times\delta(X)$. Under the null hypothesis $\mathcal{H}_0$, we have

$$W=g(D,X,\epsilon)+(1-D)\times[g(1,X,\epsilon)-g(0,X,\epsilon)]=g(1,X,\epsilon)$$

which implies that W is conditionally independent of Z given X under Assumption 1. Therefore, we obtain the following lemma.

Lemma 1 *Suppose* (1) *and Assumption 1 hold. Then, $\mathcal{H}_0$ implies that $W\perp\!\!\!\perp Z|X$. On the other hand, if $W\perp\!\!\!\perp Z|X$, then the observed data can be rationalized by a structure that satisfies $\mathcal{H}_0$.*

Lemma 1 shows that the conditional independence condition is all the testable restrictions of $\mathcal{H}_0$. Regarding the first part of Lemma 1, intuitively, if treatment effects are homogeneous, we can estimate them by the IV method, and further construct potential outcomes that are independent of the instrumental variable.[4] Note that the conditional independence condition in Lemma 1 can be equivalently rewritten as

$$\begin{aligned}&\frac{\mathbf{P}(Y\leq y;D=1|X,Z=1)-\mathbf{P}(Y\leq y;D=1|X,Z=0)}{p(X,1)-p(X,0)}\\&\qquad=\frac{\mathbf{P}(Y\leq y-\delta(X);D=0|X,Z=0)-\mathbf{P}(Y\leq y-\delta(X);D=0|X,Z=1)}{p(X,1)-p(X,0)},\end{aligned}$$

provided $p(X,1)\neq p(X,0)$ almost surely. Under the additional monotonicity condition on the selection, both sides in the above equation can be interpreted as the conditional distribution of "potential outcomes" in Imbens and Rubin (1997).

According to Lemma 1, rejecting $W\perp\!\!\!\perp Z|X$ suggests unobserved heterogeneous treatment effects. It is also worth pointing out, however, that the structures of homogeneous

[4]Note that one could also define $W^a=Y-D\times\delta(X)$, which is equal to $g(0,X,\epsilon)$ under Assumption 1.

treatment effects can be fully distinguished from those of heterogeneous treatment effects by the testable condition in Lemma 1 under additional assumptions. These additional assumptions have been widely used for obtaining identification of quantile treatment effects, and LATE in the IV literature (see, e.g., Imbens and Angrist, 1994; Chernozhukov and Hansen, 2005).

Assumption 2 (Single–index error term) *There exists a measurable function* $\tilde{g} : \mathcal{S}_{DX} \times \mathrm{R} \mapsto \mathrm{R}$ *and* $\nu : \mathcal{S}_{X\epsilon} \mapsto \mathrm{R}$ *such that*

$$g(D, X, \epsilon) = \tilde{g}(D, X, \nu(X, \epsilon)).$$

Moreover, $\tilde{g}(d, x, \cdot)$ *is strictly increasing in the scalar–valued index* ν*.*

Assumption 2 imposes the monotonicity of the single–index error term, of which various simplified assumptions have also been made in the literature for identification and estimation of nonseparable functions. For instance, among many others, Matzkin (2003) and Chesher (2003) assume that the structural function g is strictly increasing in the scalar–valued error term ϵ. Note that Assumption 2 holds under the null hypothesis $\mathcal{H}_0$, represented in terms of (3). Hence, Assumption 2 narrows down the space of alternatives such that the model restrictions derived in Lemma 1 are also sufficient to distinguish the null and alternative hypotheses.

Assumption 3 (Monotone selection) *The selection to the treatment is given by*

$$D = \mathbb{1}\left[\theta(X, Z) - \eta \geq 0\right], \tag{5}$$

where θ *is an unknown function, and* $\eta \in \mathrm{R}$ *is an error term satisfying* $Z \perp\!\!\!\perp (\epsilon, \eta)|X$*.*

Imbens and Angrist (1994) first introduce the monotone selection assumption, which is essentially the "no defier" condition. Moreover, Vytlacil (2002) shows that such a monotonicity condition is observationally equivalent to the weak monotonicity of (5) in the error term η. In the recent literature, Kitagawa (2015), Huber and Mellace (2015), and Mourifié and Wan (2017) provide formal tests for this condition together with the IV validity.

Fix $X = x$, let $\mathcal{C}_x \equiv \{\eta \in \mathrm{R} : \min\{\theta(x, 0), \theta(x, 1)\} < \eta \leq \max\{\theta(x, 0), \theta(x, 1)\}\}$. Note that $\mathcal{C}_x$ is called as the "complier group" if $p(x, 0) < p(x, 1)$ (see Imbens and Angrist, 1994, for the concept of the "complier group.")

Assumption 4 *The support of $g(d, x, \epsilon)$ given $X = x$ and the complier group $\mathcal{C}_x$ equals to the support of $g(d, x, \epsilon)$ given $X = x$, i.e., $\mathcal{S}_{g(d,x,\epsilon)|X=x,\ \eta\in\mathcal{C}_x} = \mathcal{S}_{g(d,x,\epsilon)|X=x}$.*

Assumption 4 is a support condition, first introduced by Vuong and Xu (2017) as the effectiveness of the IV. It implies that $\mathcal{S}_{g(d,x,\epsilon)|X=x,\ \eta\in\mathcal{C}_x} = \mathcal{S}_{Y|D=d,X=x}$.[5] Note that the distribution of $g(d, x, \epsilon)$ given $X = x$ and $\eta \in \mathcal{C}_x$ can be identified; see, e.g., Imbens and Rubin (1997). Thus, Assumption 4 is testable. Specifically, for all $t \in \mathrm{R}$,

$$\mathbf{P}[g(d, x, \epsilon) \leq t|X = x, \eta \in \mathcal{C}_x] = \frac{\mathbf{P}(Y \leq t, D = d|X = x, Z = 1) - \mathbf{P}(Y \leq t, D = d|X = x, Z = 0)}{\mathbf{P}(D = d|X = x, Z = 1) - \mathbf{P}(D = d|X = x, Z = 0)},$$

from which we can identify the support $\mathcal{S}_{g(d,x,\epsilon)|X=x,\eta\in\mathcal{C}_x}$.

Assumption 4 allows one to use the data to address questions involving counterfactuals of outcomes of the "always–takers" and the "never–takers" groups. It is possible to provide sufficient primitive conditions for Assumption 4. For instance, if one assumes $\mathcal{S}_{\epsilon|X=x,\ \eta\in\mathcal{C}_x} = \mathcal{S}_{\epsilon|X=x}$, or even a stronger condition that (ϵ, η) has rectangular support conditional on $X = x$, then Assumption 4 holds. It is also worth noting that without Assumption 4, we can test the null hypothesis with respect to the subset $\epsilon \in \mathcal{S}_{\epsilon|X=x,\ \eta\in\mathcal{C}_x}$.

Theorem 1 *Let (1) and Assumptions 1-4 hold. Then $\mathcal{H}_0$ holds if and only if $W \perp\!\!\!\perp Z|X$.*

In Theorem 1, the only if part simply follows Lemma 1. For the if part, Assumptions 2–4 ensure point identification of individual treatment effects, i.e. $g(1, X, \epsilon) - g(0, X, \epsilon)$, as long as one of the potential outcomes $\{g(d, X, \epsilon) : d = 0, 1\}$ is obvserved; see Vuong and Xu (2017). It should also be noted that Theorem 1 is related to Lu and White (2014), who show that $\mathcal{H}_0$ holds if and only if $Y - \mathrm{E}(Y|D, X) \perp\!\!\!\perp D|X$ under the unconfoundedness condition (i.e., $D \perp\!\!\!\perp \epsilon|X$) and Assumption 2.

From now on, we maintain Assumptions 1-4. By Theorem 1, testing the null hypothesis $\mathcal{H}_0$ is equivalent to testing the conditional independence condition $W \perp\!\!\!\perp Z|X$.

3 Consistent tests

Based on Theorem 1, we now propose tests for unobserved treatment effect heterogeneity via testing the conditional independence restriction. Because Z is binary, the conditional

[5] To see this, note that $\mathcal{S}_{g(d,x,\epsilon)|X=x,\ \eta\in\mathcal{C}_x} \subseteq \mathcal{S}_{g(d,x,\epsilon)|D=d,X=x} \subseteq \mathcal{S}_{g(d,x,\epsilon)|X=x}$.

independence restriction in Theorem 1 is equivalent to

$$F_{W|XZ}(\cdot|x,0) = F_{W|XZ}(\cdot|x,1),\ \forall\ x \in \mathcal{S}_X.$$

Note that the variable W needs to be nonparametrically constructed from the data. In the following discussion, we distinguish the cases where the covariates X are continuous random variables because the continuous–covariates case is more difficult to deal with due to the nonparametric function $\delta(\cdot)$ in the construction of W. For expositional simplicity, we assume $X \in \mathrm{R}$ in the following discussion. It is straightforward to generalize our results to vector-valued covariates.

Notation. For a generic random vector $(A_1, A_2, A_3, \cdots)$, let $\mathbb{1}_{A_1A_2A_3\cdots}(a_1, a_2, a_3, \cdots) \equiv \mathbb{1}(A_1 = a_1, A_2 = a_2, A_3 = a_3 \cdots)$, and $\mathbb{1}^*_{A_1A_2A_3\cdots}(a_1, a_2, a_3, \cdots) \equiv \mathbb{1}(A_1 \le a_1, A_2 = a_2, A_3 = a_3, \cdots)$. Moreover, let K and h be a kernel function and a smoothing bandwidth, respectively, and $K_{A,h}(a) \equiv K((A-a)/h)/h$ for a random variable A.

3.1 Case 1: Discrete covariates

We first discuss the case where X takes only a finite number of values. Let $\{(Y_i, D_i, X_i, Z_i)' : i \le n\}$ be a random sample of $(Y, D, X, Z)'$. By Theorem 1, we test $\mathcal{H}_0$ via the following model restrictions:

$$F_{W|XZ}(\cdot\ |x,0) = F_{W|XZ}(\cdot\ |x,1),\ \forall\ x \in \mathcal{S}_X,$$

where $W = Y + (1-D) \times \delta(X)$ is generated from the observables.

We estimate $\delta(X_i)$ as follows

$$\hat{\delta}(X_i) = \frac{\sum_{j\neq i} Y_j Z_j \mathbb{1}_{X_j}(X_i) \times \sum_{j\neq i} \mathbb{1}_{X_j}(X_i) - \sum_{j\neq i} Y_j \mathbb{1}_{X_j}(X_i) \times \sum_{j\neq i} Z_j \mathbb{1}_{X_j}(X_i)}{\sum_{j\neq i} D_j Z_j \mathbb{1}_{X_j}(X_i) \times \sum_{j\neq i} \mathbb{1}_{X_j}(X_i) - \sum_{j\neq i} D_j \mathbb{1}_{X_j}(X_i) \times \sum_{j\neq i} Z_j \mathbb{1}_{X_j}(X_i)},$$

and further let $\widehat{W}_i = Y_i + (1-D_i) \times \hat{\delta}(X_i)$. We now define our test statistic:

$$\widehat{\mathcal{T}}_n = \sup_{(w,x)\in\mathcal{S}_{WX}} \sqrt{n}\ \left|\widehat{F}_{\widehat{W}|XZ}(w|x,0) - \widehat{F}_{\widehat{W}|XZ}(w|x,1)\right|,$$

where $\widehat{F}_{\widehat{W}|XZ}(w|x,z) = \frac{\sum_{i=1}^n 1^*_{\widehat{W}_i}(w) 1_{X_iZ_i}(x,z)}{\sum_{i=1}^n 1_{X_iZ_i}(x,z)}$.

Next, we establish the asymptotic properties of the test statistic $\widehat{\mathcal{T}}_n$. Let

$$f_{WD|XZ}(w,d|x,z) \equiv f_{W|DXZ}(w|d,x,z) \times \mathrm{P}(D = d|X = x, Z = z),$$

and

$$\kappa(w,x) \equiv -\frac{f_{WD|XZ}(w,0|x,1) - f_{WD|XZ}(w,0|x,0)}{p(x,1)-p(x,0)}.$$

Note that, under Assumptions 1 and 3, $\kappa(w,x) \geq 0$ since it becomes the conditional density of $g(0,x,\epsilon)$ given the complier group and $X = x$. Moreover, let

$$\psi_{wx} \equiv [\mathbb{1}(W \leq w) - F_{W|X}(w|x)] \times \Big[\frac{\mathbb{1}_{XZ}(x,1)}{\mathbf{P}(X=x,Z=1)} - \frac{\mathbb{1}_{XZ}(x,0)}{\mathbf{P}(X=x,Z=0)}\Big]; \tag{6}$$

$$\phi_{wx} \equiv \kappa(w,x)[W - \mathbf{E}(W|X=x)] \times \Big[\frac{\mathbb{1}_{XZ}(x,1)}{\mathbf{P}(X=x,Z=1)} - \frac{\mathbb{1}_{XZ}(x,0)}{\mathbf{P}(X=x,Z=0)}\Big]. \tag{7}$$

By the definition, ψ_{wx} and ϕ_{wx} are stochastic processes indexed by (w,x).

Assumption 5 *Let X be a discrete random variable with a finite support. Moreover, the probability distribution of Y given (D,X,Z) admits a uniformly continuous density function $f_{Y|DXZ}$ and* $\mathrm{E}(Y^2) < \infty$.

Theorem 2 *Suppose Assumptions 1–5 hold. Then, under $\mathcal{H}_0$,*

$$\widehat{\mathcal{T}}_n \xrightarrow{d} \sup_{(w,x)\in\mathcal{S}_{WX}} |\mathcal{Z}(w,x)|,$$

where $\mathcal{Z}(\cdot,x)$ is a mean–zero Gaussian process with a covariance kernel:

$$Cov\left[\mathcal{Z}(w,x),\mathcal{Z}(w',x)\right] = \mathrm{E}\left[(\psi_{wx}+\phi_{wx})(\psi_{w'x}+\phi_{w'x})\right],\ \forall w,w' \in \mathrm{R}.$$

Moreover, under $\mathcal{H}_1$, we have

$$n^{-\frac{1}{2}}\widehat{\mathcal{T}}_n \xrightarrow{p} \sup_{(w,x)\in\mathcal{S}_{WX}} \left|F_{W|XZ}(w|x,0) - F_{W|XZ}(w|x,1)\right|.$$

In the covariance kernel $\mathrm{Cov}\left[\mathcal{Z}(w,x),\mathcal{Z}(w',x)\right]$, ϕ_{wx} and $\phi_{w'x}$ appear due to the estimation of $\delta(x)$. By Theorem 2, our test is one–sided: reject $\mathcal{H}_0$ at significance level α if and only if $\widehat{\mathcal{T}}_n > c_\alpha$, where c_α is the $(1-\alpha)$-th quantile of $\sup_{(w,x)\in\mathcal{S}_{WX}} |\mathcal{Z}(w,x)|$.

Since the asymptotic distribution of $\sup_{(w,x)\in\mathcal{S}_{WX}} |\mathcal{Z}(w,x)|$ is complicated, we apply the multiplier bootstrap method to approximate the entire process for the critical value. See, e.g., van der Vaart and Wellner (1996), Delgado and Manteiga (2001), Barrett and Donald (2003), and Donald and Hsu (2014). Specifically, we simulate a sequence of i.i.d. pseudo–random variables $\{U_i : i = 1,\cdots,n\}$, which is independent of the random sample

$\{(Y_i, X_i, D_i, Z_i) : i = 1, \cdots, n\}$, with $\mathrm{E}[U] = 0$, $\mathrm{E}[U^2] = 1$, and $\mathrm{E}[U^4] < +\infty$. Then, we obtain the following simulated empirical process:

$$\widehat{\mathcal{Z}}^u(w, x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \times (\hat{\psi}_{wx,i} + \hat{\phi}_{wx,i}),$$

where $\hat{\psi}_{wx,i} + \hat{\phi}_{wx,i}$ is the estimated influence function. Namely,

$$\hat{\psi}_{wx,i} = \left[\mathbb{1}(\widehat{W}_i \le w) - \frac{\sum_{j=1}^{n} \mathbb{1}^*_{\widehat{W}_j X_j}(w, x)}{\sum_{j=1}^{n} \mathbb{1}_{X_j}(x)}\right] \times \left[\frac{\mathbb{1}_{X_i Z_i}(x, 0)}{\widehat{\mathbf{P}}(X = x, Z = 0)} - \frac{\mathbb{1}_{X_i Z_i}(z, 1)}{\widehat{\mathbf{P}}(X = x, Z = 1)}\right];$$

$$\hat{\phi}_{wx,i} = \hat{\kappa}(w, x) \left[\widehat{W}_i - \frac{\sum_{j=1}^{n} \widehat{W}_j \mathbb{1}_{X_j}(x)}{\sum_{j=1}^{n} \mathbb{1}_{X_j}(x)}\right] \times \left[\frac{\mathbb{1}_{X_i Z_i}(x, 0)}{\widehat{\mathbf{P}}(X = x, Z = 0)} - \frac{\mathbb{1}_{X_i Z_i}(x, 1)}{\widehat{\mathbf{P}}(X = x, Z = 1)}\right],$$

where $\widehat{\mathbf{P}}(X = x, Z = z) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{X_i Z_i}(x, z)$, and $\hat{\kappa}(w, x) = -\frac{\hat{f}_{WD|XZ}(w,0|x,1) - \hat{f}_{WD|XZ}(w,0|x,0)}{\hat{p}(x,1) - \hat{p}(x,0)}$. In the definition of $\hat{\kappa}(w, x)$, $\hat{f}_{WD|XZ}(w, 0|x, z) = \frac{\sum_{j=1}^{n} \mathbb{1}_{D_j X_j Z_j}(0,x,z) \times K_{\widehat{W}_j,h}(w)}{\sum_{j=1}^{n} \mathbb{1}_{X_j Z_j}(x,z)}$, where $\hat{p}(x, z) = \frac{\sum_{j=1}^{n} \mathbb{1}_{D_j X_j Z_j}(1,x,z)}{\sum_{j=1}^{n} \mathbb{1}_{X_j Z_j}(x,z)}$. For a given significant level α, the critical value $\hat{c}_n(\alpha)$ is obtained as the $(1-\alpha)$–quantile of the simulated distribution of $\sup_{w \in \mathrm{R},\ x \in \mathcal{S}_X} \left|\widehat{\mathcal{Z}}^u(w, x)\right|$. The validity of the multiplier bootstrap process in that $\widehat{\mathcal{Z}}^u(w, x) \Rightarrow \mathcal{Z}(w, x)$ conditional on the sample path with probability approaching one can be shown under suitable conditions, see, e.g., Donald and Hsu (2014). Then the validity of the multiplier bootstrap critical value $\hat{c}_n(\alpha)$ can be established by continuous mapping theorem, see, e.g., van der Vaart and Wellner (2007). We omit the proof for brevity.

3.2 Case 2: Continuous Covariates

We now consider the case where $X \in \mathrm{R}$ is continuously distributed. To extend the empirical process argument used in the proof of Theorem 2 to this case, we propose a modified Kolmogorov–Smirnov test statistic. Such a modification allows the generated variable W to be constructed from the unknown function $\delta(\cdot)$ as an infinite–dimensional parameter.

Let $\lambda(t) = -t \times 1(t \le 0)$ and $\Pi(w|x, z) = \mathrm{E}[\lambda(W - w)|X = x, Z = z]$. Note that $\lambda(\cdot)$ is continuous and has a directional derivative. By definition, $\Pi(\cdot|x, z)$ is the primitive function of the $F_{W|XZ}(\cdot|x, z)$, i.e.,

$$\frac{\partial}{\partial w}\Pi(w|x, z) = F_{W|XZ}(w|x, z).$$

Clearly, $F_{W|XZ}(\cdot|x,0) = F_{W|XZ}(\cdot|x,1)$ holds if and only if $\Pi(\cdot|x,0) = \Pi(\cdot|x,1)$. Thus, the model restriction $W \perp\!\!\!\perp Z|X$ can be equivalently characterized as follows

$$\Pi(w|x,0) = \Pi(w|x,1), \quad \forall(w,x) \in \mathcal{S}_{WX}.$$

In terms of probability distribution of W given $(X,Z) = (x,z)$, both $F_{W|XZ}(\cdot|x,z)$ and $\Pi(\cdot|x,z)$ contain the exact same amount of information. The latter, however, allows us to derive a test statistics and establish its limiting distribution using a pseudo sample of W. When covariates X is continuously distributed, the constructed pseudo sample $\{\hat{W}_i : i \leq n\}$ involves the nonparametric component $\hat{\delta}(X_i)$. We exploit the smoothness of $\lambda(\cdot)$ and show that this first-stage estimation error can be further aggregated out at the $\sqrt{N}$–rate in our test statistics defined on $\{\hat{W}_i : i \leq n\}$. It should be noted that when covariates X is discrete as discussed in the last subsection, or a mixture of both continuous and discrete components, we can also apply a similar testing procedure via testing $\Pi(\cdot|x,0) = \Pi(\cdot|x,1)$. Moreover, we assume that $\mathcal{S}_W$ is bounded for expositional simplicity.

We denote $f_{XZ}(x,z) \equiv f_{X|Z}(x|z) \times \mathbf{P}(Z=z)$. For $z \in \{0,1\}$, let $z' = 1-z$ and

$$G(w,x,z) = \mathbf{E}\left[\lambda(W-w)\mathbb{1}^*_{XZ}(x,z)f_{XZ}(X,z')\right].$$

Motivated by Stinchcombe and White (1998), we rewrite the above conditional expectation restrictions by the following unconditional ones:

$$G(w,x,0) = G(w,x,1), \quad \forall(w,x) \in \mathcal{S}_{WX}. \tag{8}$$

To see the equivalence, first note that

$$G(w,x,z) = \mathbf{E}\left[\lambda(W-w)\mathbb{1}(X \leq x)f_{X|Z}(X|z')|Z=z\right]\mathbf{P}(Z=0)\mathbf{P}(Z=1).$$

Moreover, by the law of iterated expectation,

$$\frac{\partial}{\partial x}\mathbf{E}[\lambda(W-w)\mathbb{1}(X \leq x)f_{X|Z}(X|z')|Z=z] = \Pi(w|x,z)f_{X|Z}(x|0)f_{X|Z}(x|1).$$

Therefore, we obtain the conditional expectation restrictions as the derivatives of (8). Note that the estimation of $G(w,x,z)$ avoids any denominator issues, which thereafter simplifies our asymptotic analysis.

We estimate $\delta(X_i)$ by

$$\hat{\delta}(X_i) = \frac{\sum_{j\neq i} Y_j Z_j K_{X_j,h}(X_i) \times \sum_{j\neq i} K_{X_j,h}(X_i) - \sum_{j\neq i} Y_j K_{X_j,h}(X_i) \times \sum_{j\neq i} Z_j K_{X_j,h}(X_i)}{\sum_{j\neq i} D_j Z_j K_{X_j,h}(X_i) \times \sum_{j\neq i} K_{X_j,h}(X_i) - \sum_{j\neq i} D_j K_{X_j,h}(X_i) \times \sum_{j\neq i} Z_j K_{X_j,h}(X_i)}.$$

Moreover, let

$$\hat{f}_{XZ}(X_i, z) = \frac{1}{n} \sum_{j\neq i} K_{X_j,h}(X_i) \mathbb{1}_{Z_j}(z);$$

$$\widehat{G}(w, x, z) = \frac{1}{n} \sum_{i=1}^{n} \lambda(\widehat{W}_i - w) \mathbb{1}^*_{X_i Z_i}(x, z) \hat{f}_{XZ}(X_i, z').$$

Thus, we define our test statistic as follows:

$$\widehat{\mathcal{T}}_n^c = \sup_{(w,x)\in\mathcal{S}_{WX}} \sqrt{n} \left| \widehat{G}(w, x, 0) - \widehat{G}(w, x, 1) \right|.$$

In the above definition, the support $\mathcal{S}_{WX}$ is assumed to be known for simplicity. In practice, this assumption can be relaxed by using a consistent set estimator $\hat{\mathcal{S}}_{WX}$ of $\mathcal{S}_{WX}$.

We show that the proposed test statistic $\widehat{\mathcal{T}}_n^c$ converges in distribution at the regular parametric rate under the null. The key step of our proof is to show that

$$\sup_{(w,x)\in\mathcal{S}_{WX}} \left| \widehat{G}(w, x, z) - \widetilde{G}(w, x, z) \right| = o_p\left(n^{-1/2}\right). \tag{9}$$

where $\widetilde{G}(w, x, z) = \frac{1}{n} \sum_{i=1}^{n} (w - \widehat{W}_i) \mathbb{1}(W_i \leq w) \times \mathbb{1}^*_{X_i Z_i}(x, z) \hat{f}_{XZ}(X_i, z')$. The above result requires that the nonparametric elements in the estimation of $\hat{\delta}(\cdot)$ should converge to the corresponding true values uniformly at a rate faster than $n^{-1/4}$. With (9), we obtain the limiting distribution of $\hat{G}$ via $\tilde{G}$. Note that the latter is continuous and differentiable in the first-stage kernel estimator $\hat{\delta}(\cdot)$ and can be represented as a V-process; See e.g. Powell, Stock, and Stoker (1989).

Assumption 6. *The support* $\mathcal{S}_{WX} \subseteq \mathrm{R}^2$ *is compact. For* $z = 0, 1$, $\inf_{(x,z)\in\mathcal{S}_{XZ}} f_{X|Z}(x|z) > 0$, $\sup_{(x,z)\in\mathcal{S}_{XZ}} f_{X|Z}(x|z) \leq \overline{f} < +\infty$, *and* $\inf_{x\in\mathcal{S}_X} |p(x, 1) - p(x, 0)| > 0$.

Assumption 7. *For* $z \in \{0, 1\}$, $f_{X|Z}(x|z)$, $p(x, z)$ *and* $\mu(x, z)$ *are continuous in* x.

Assumption 8. *For some* $\iota > \frac{1}{4}$, *we have* $h \to 0$ *and* $n^{\iota}/\sqrt{nh} \to 0$ *as* $n \to \infty$. *Moreover,* *the first-stage estimators satisfy the condition that*

$$\sup_{(x,z)\in\mathcal{S}_{XZ}}\left|\mathbf{E}\Big[\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{Z_j}(z)K_{X_j,h}(x)\Big]-f_{XZ}(x,z)\right|=O_p(n^{-\iota}),$$

$$\sup_{(x,z)\in\mathcal{S}_{XZ}}\left|\mathbf{E}\Big[\frac{1}{n}\sum_{j=1}^{n}D_j\mathbb{1}_{Z_j}(z)K_{X_j,h}(x)\Big]-p(x,z)f_{XZ}(x,z)\right|=O_p(n^{-\iota}),$$

$$\sup_{(x,z)\in\mathcal{S}_{XZ}}\left|\mathbf{E}\Big[\frac{1}{n}\sum_{j=1}^{n}Y_j\mathbb{1}_{Z_j}(z)K_{X_j,h}(x)\Big]-\mathbf{E}(Y|X=x,Z=z)f_{XZ}(x,z)\right|=O_p(n^{-\iota}).$$

Assumptions 6 and 7 are standard in the nonparametric estimation literature. Assumption 8 is a high-level condition that requires the nonparametric estimation bias to diminish uniformly at a rate faster than $n^{1/4}$. Such a condition on the bias term can be satisfied under additional primitive conditions on the kernel function and the bandwidth respectively, $K(\cdot)$ and h, as well as the smoothness of the underlying structural functions. See, e.g., Pagan and Ullah (1999).

Lemma 2 *Suppose Assumptions 6-8 hold. Then,* (9) *holds for* $z=0,1$.

By Lemma 2, it suffices to establish the limiting distribution of $\widetilde{G}(w,x,1)-\widetilde{G}(w,x,0)$ for the asymptotic properties of our test statistics. Note that in the definition of $\widetilde{G}(w,x,z)$, there is no nonparametric elements estimated in the indicator function.

To establish asymptotic properties for inference, we make the following assumption.

Assumption 9. *Both* $\sup_{x\in\mathcal{S}_X}\left|\mathrm{E}[\hat{\delta}(x)]-\delta(x)\right|$ *and* $\sup_{(x,z)\in\mathcal{S}_{XZ}}\left|\mathrm{E}[\hat{f}_{XZ}(x,z)]-f_{XZ}(x,z)\right|$ *are* $o_p(n^{-\frac{1}{2}})$.

Assumption 9 strengthens Assumption 8 by requiring the bias term in the first–stage nonparametric estimation to be smaller than $o_p(n^{-1/2})$, which can be established by using higher–order kernels (see, e.g., Powell, Stock, and Stoker, 1989).

Let $F^*_{WD|XZ}(w,d|x,z)\equiv F_{W|DXZ}(w|d,x,z)\times\mathbf{P}(D=d|X=x,Z=z)$ and

$$\kappa^c(w,x)=-\frac{F^*_{WD|XZ}(w,0|x,1)-F^*_{WD|XZ}(w,0|x,0)}{p(x,1)-p(x,0)}.$$

Moreover, we define two random processes indexed by (w, x) as follows:

$$\psi^c_{wx} = \Big\{\lambda(w - W) - \mathbf{E}[\lambda(w - W)|X]\Big\}\Big[\frac{\mathbb{1}^*_{XZ}(x, 1)}{f_{XZ}(X, 1)} - \frac{\mathbb{1}^*_{XZ}(x, 0)}{f_{XZ}(X, 0)}\Big] f_{XZ}(X, 0) f_{XZ}(X, 1);$$
$$\phi^c_{wx} = \kappa^c(w, X)[W - \mathbf{E}(W|X)]\Big[\frac{\mathbb{1}^*_{XZ}(x, 1)}{f_{XZ}(X, 1)} - \frac{\mathbb{1}^*_{XZ}(x, 0)}{f_{XZ}(X, 0)}\Big] f_{XZ}(X, 0) f_{XZ}(X, 1).$$

By definition, we have $\mathbf{E}(\psi^c_{wx}|X, Z) = \mathbf{E}(\phi^c_{wx}|X, Z) = 0$ under $\mathcal{H}_0$.

Theorem 3 *Suppose Assumptions 6–9 hold. Then, under $\mathcal{H}_0$,*

$$\widehat{\mathcal{T}}^c_n \xrightarrow{d} \sup_{(w,x)\in\mathcal{S}_{WX}} |\mathcal{Z}^c(w, x)|$$

where $\mathcal{Z}^c(\cdot, \cdot)$ is a mean–zero Gaussian process with the following covariance kernel

$$Cov\left[\mathcal{Z}^c(w, x), \mathcal{Z}^c(w', x')\right] = \mathrm{E}\left[(\psi^c_{wx} + \phi^c_{wx})(\psi^c_{w'x'} + \phi^c_{w'x'})\right], \ \forall (w, x), (w', x') \in \mathcal{S}_{WX}.$$

Moreover, under $\mathcal{H}_1$, we have

$$n^{-\frac{1}{2}}\widehat{\mathcal{T}}^c_n \xrightarrow{p} \sup_{(w,x)\in\mathcal{S}_{WX}} |G(w, x, 0) - G(w, x, 1)|.$$

Similar to the discrete–covariates case, we reject $\mathcal{H}_0$ at significance level α when $\widehat{\mathcal{T}}^c_n > c_\alpha$. The critical value c_α is obtained via the multiplier bootstrap method.

4 Monte Carlo Simulations

In this section, we investigate the finite sample performance of our tests with a simulation study. The data are simulated as follows:

$$Y = D + X + [\gamma + (1 - \gamma)D] \times \epsilon;$$
$$D = \mathbb{1}\left[\Phi(\eta) \leq 0.5 \times Z\right],$$

where (ϵ, η) conforms to a joint normal distribution with zero mean, unit variance and correlation coefficient $\rho = 0.7$, and $Z \sim Bernoulli(p)$ with $p = 0.25, 0.5$ and 0.75 respectively. Note that we also try different values for the correlation coefficient, and all the results are qualitatively similar.[6] For simplicity, X, Z and (ϵ, η) are mutually independent. Moreover, X is uniformly distributed on $\{1, 2, 3, 4\}$ and on $[0, 1]$, respectively, in the discrete covariates

[6] Additional Monte Carlo simulation results are available upon request.

and the continuous covariates case. Furthermore, parameter $\gamma \in [0,1]$ describes the degree of unobserved heterogeneous treatment effects in our specification. In particular, $\mathcal{H}_0$ holds if and only if $\gamma = 1$. Intuitively, the smaller γ is, the more power we expect from our tests. To investigate size and power of our tests, we choose $\gamma \in \{1, 0.75, 0.5\}$.

We consider sample size $n = 1000, 2000, 4000$, a nominal level of $\alpha = 5\%$, and $2,000$ Monte Carlo repetitions. To compute the suprema of the simulated stochastic processes, we use $n/10$ grids on the support of $[\min_{i=1}^n(\widehat{W}_i), \max_{i=1}^n(\widehat{W}_i)]$. Moreover, we use 500 multiplier bootstrap samples to simulate the p-values. Regarding the estimation of $\kappa(w, x)$, we choose the second–order Gaussian kernel function with the bandwidth, $h_n = c \cdot \mathrm{std}(\widehat{W}) \cdot n^{-1/5}$, and we set $c \in \{0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3\}$ to study the sensitivity of the test to the bandwidth.

Table 1 reports rejection probabilities of our simulations in the discrete-covariates case under the null hypothesis (i.e., $\gamma = 1$) and alternative hypotheses (i.e., $\gamma = 0.75, 0.5$). From Panel A, the level of our test is fairly well behaved: It gets closer to the nominal level as the sample size increases, and the rejection probabilities are not sensitive to the constant c for the bandwidth choice. Panels B and C show that the power of the test is reasonable. In particular, when γ is closer to 1, it is more difficult to detect such a "local" alternative. Therefore, we obtain relatively low power even when the sample size reaches $n = 2000$ in Panel B. For relatively "small" sample size, e.g., $n = 1000$, our results show that our test performs better with a larger bandwidth choice. Moreover, when p (i.e., the probability of $Z = 1$) is 0.5, all the results for size and power dominate the other two cases with $p = 0.25, 0.75$, which is expected by our asymptotic theory.

Next, we evaluate the performance of our tests in the case where the covariate X is continuous. To compute the suprema, we calculate the test statistic by using $n/20$ grid points in the support $[\min_{i=1}^n(\widehat{W}_i), \max_{i=1}^n(\widehat{W}_i)]$, as well as in the support $[\min_{i=1}^n(X_i), \max_{i=1}^n(X_i)]$. Table 2 reports the size and power properties of our test, which are qualitatively similar to those in the discrete-covariates case.[7]

[7] In an online supplemental appendix, we also conduct several other DGPs that are empirically relevant. In all cases, the size and power are qualitatively similar to the results we present in this paper.

5 Empirical Applications

5.1 Effect of Job Training Program on Earnings

We now apply our tests to study the effects of a job training program on earnings, i.e., the *National Job Training Partnership Act* (JTPA), commissioned by the Department of Labor of the U.S. This program funded training from 1983 to the late 1990's to increase employment and earnings for participants. The major component of JTPA aims to support training for the economically disadvantaged. The effects of JTPA training programs on earnings have also been studied by e.g., Heckman, Smith, and Clements (1997) and Abadie, Angrist, and Imbens (2002) under a general framework allowing for unobserved heterogeneous treatment effects. The data are publicly available at `https://upjohn.org/node/952`.

Our sample consists of 11,204 observations from the JTPA, a survey dataset from over $20,000$ adults and out-of-school youths who applied for JTPA in 16 local areas across the country between 1987 and 1989.[8] Each participant was randomly assigned to either a program group or a control group (1 out of 3 on average). Members of the program group were eligible to participate in JTPA services, including classroom training, on-the-job training or job search assistance, and other services, while members of control group were not eligible for JTPA services for 18 months. Following the literature (see, e.g., Bloom, Orr, Bell, Cave, Doolittle, Lin, and Bos, 1997), we use the program eligibility as an instrumental variable for the endogenous individual participation decision.

The outcome variable is individual earnings, measured by the sum of earnings in the 30-month period following the offer. The observed covariates include a set of dummies for race, high-school graduate, and marriage, whether the applicant worked at least 12 weeks in the 12 months preceding random assignment, and also 5 age-group dummies (22-24, 25-29, 30-35, 36-44, and 45-54), among others. Descriptive statistics can be found in Table 3. For simplicity, we group all applicants into 3 age categories (22-29, 30-35, and 36 and above), and pool all non-White applicants as minority applicants.

To implement the test, we use the second–order Gaussian kernel and set the smoothing parameter to $1.06 \times \mathrm{Std}(\widehat{W}) \times n^{-1/4}$ when we estimate $\kappa(x, \tau)$. For the critical value, we use $10,000$ multiplier bootstrap samples and search for the suprema by using $5,000$ grid

[8] JTPA services are provided at 649 sites, which might not be randomly chosen. For a given site, the applicants were randomly selected for the JTPA dataset.

points. The p-value of our test is 0.1204. Therefore, the null hypothesis (i.e., no unobserved heterogenous treatment effects) cannot be rejected at the 10% significance level. Our results are robust to the size of bootstrap samples, the number of grid points, and the choices of bandwidth. Note that our results are consistent with Abadie, Angrist, and Imbens (2002), who estimate quantile treatment effects under a linear specification. In particular, one cannot reject the null hypothesis that quantile treatment effects are invariant across different quantile levels.[9]

5.2 The Impact of Fertility on Family Income

The second empirical illustration considers the heterogeneous impacts of children on parents' labor supply and income. Recently, Frölich and Melly (2013) studied the heterogeneous effects of fertility on family income within the general LATE framework. To deal with the endogeneity of fertility decisions, Rosenzweig and Wolpin (1980), Angrist and Evans (1998), Bronars and Grogger (1994) and Jacobsen, Pearce, and Rosenbloom (1999), among many others, suggest using twin births as an instrumental variable.

Our data use the 1% and 5% Census Public Use Micro Sample (PUMS) from 1990 and 2000 censuses, consisting of 602,767 and 573,437 observations, respectively. The data are publicly available at `https://www.census.gov/main/www/pums.html`. Similar to Frölich and Melly (2013), our sample is restricted to 21- to 35-year-old married mothers with at least one child since we use twin birth as an instrument for fertility. The outcome variable of interest is the family's annual labor income.[10] The treatment variable is a dummy variable that takes the value 1 to indicate when a mother has two or more children. The instrumental variable is also a dummy variable and it equals 1 if the first birth is a twin. The covariates include mother's and father's age, race, educational level, and working status. Some covariates, i.e., age, years in education, and working hours per week, are treated as continuous variables. Summary statistics can be found in Table 4.

Similar to the previous empirical illustration, we use the second–order Gaussian kernel

[9] In Abadie, Angrist, and Imbens (2002), their empirical results (i.e., Table III) do not provide complete information on the variance–covariance matrix for quantile treatment effects at multiple quantile levels. So we cannot use their empirical results to conduct a formal test for our purpose.

[10] It includes wages, salary, armed forces pay, commissions, tips, piece-rate payments, cash bonuses earned before deductions were made for taxes, bonds, pensions, union dues, etc. See Frölich and Melly (2013) for more details.

with various bandwidth choices for a robustness check. For the critical value, we use $5,000$ bootstrapped samples and search for the suprema by using $1,000$ grids for each of the supports of both W and X's. The bandwidths are selected in the same manner as those in the JTPA case. The p-values of our tests are 0.0031 and 0.0004 for the 1990 and 2000 Censuses, respectively. These results suggest that the null hypothesis, i.e., homogeneous treatment effects, should be rejected at all usual significance levels.

5.3 Extensions

When Z takes multiple values rather than being binary, one could extend our approach of testing for unobserved heterogeneous treatment effects. Namely, let $W \equiv Y + (1-D)\times\delta(X)$ where $\delta(x) = \frac{\text{Cov}(Y,Z|X=x)}{\text{Cov}(D,Z|X=x)}$. Then we test $\mathcal{H}_0$ by testing $W \perp\!\!\!\perp Z|X$. When Z takes more than binary values in its support, this model restriction can be equivalently written as

$$F_{W|XZ}(\cdot\ |x,z) = F_{W|X}(\cdot\ |x),\ \forall\ (x,z) \in \mathcal{S}_{XZ},$$

or

$$\Pi(\cdot|x,z) = \text{E}[\lambda(W-\cdot)|X=x],\quad \forall(x,z) \in \mathcal{S}_{WXZ},$$

depending on covariates X or instruments Z contain any continuously distributed components or not.

Such a test, however, does not exploit model restrictions arising from multiplicity of Z. For instance, suppose w.l.o.g. $\mathcal{S}_Z = \{0,1,2\}$. Under $\mathcal{H}_0$ and Assumption 1, we have

$$\frac{\mathbf{E}(Y|X=x,Z=z) - \mathbf{E}(Y|X=x)}{p(x,z) - \mathbf{E}(D|X=x)} = \delta(x),\ \forall x.$$

As a matter of fact, our test does not exploit such a model restriction.

Our analysis naturally extends the case where the treatment variable D takes multiple values. For illustration purpose, suppose $\mathcal{S}_D = \{0,1,2\}$. Under the homogeneous treatment effects hypothesis, denote $\delta_1(x) \equiv g(1,x,\cdot) - g(0,x,\cdot)$ and $\delta_2(x) \equiv g(2,x,\cdot) - g(0,x,\cdot)$. For $d=1,2$, let $p_d(X,Z) = \mathbf{P}(D=d|X,Z)$ and $W_d \equiv \delta_d(X) + Y - \sum_{d'=1}^{2} \mathbb{1}(D=d') \times \delta_{d'}(X)$. Note that under $\mathcal{H}_0$, i.e. $g(d,x,\cdot) - g(0,x,\cdot) = \delta_d(x)$, we have

$$W_d = g(d,X,\epsilon).$$

By a similar argument, we test for unobserved heterogeneous treatment effects by testing $W_d \perp\!\!\!\perp Z|X$ for $d = 1, 2$. To complete our analysis, it suffices to establish the identification of $\delta_d(x)$. Note that under $\mathcal{H}_0$ and Assumption 1, we have

$$\mathbf{E}(Y|X = x, Z = z) = \mathbf{E}\left[g(0, X, \epsilon)|X = x\right] + \delta_1(x) \times p_1(x, z) + \delta_2(x) \times p_2(x, z), \quad \forall z.$$

Therefore, δ_d is identified if $\{(p_1(x, z), p_2(x, z))' : z \in \mathcal{S}_{Z|X=x}\}$ has the full rank. Note that such a rank condition requires $\mathcal{S}_{Z|X=x}$ contains at least three values.

A Appendix: Proofs

A.1 Proof of Proposition 1

Proof: For the "if" part, under (3), we have

$$g(1,x,\epsilon)-g(0,x,\epsilon)=m(1,x)-m(0,x)\equiv\delta(x),\quad \forall x\in\mathcal{S}_X.$$

For the "only if" part, (2) implies

$$g(d,x,\epsilon)=d\times[g(1,x,\epsilon)-g(0,x,\epsilon)]+g(0,x,\epsilon)=d\times\delta(x)+g(0,x,\epsilon).$$

Therefore, (3) holds in the sense $m(d,x)=d\times\delta(x)$ and $\nu(x,\epsilon)=g(0,x,\epsilon)$.

A.2 Proof of Lemma 1

Proof: The first part of Lemma 1 is straightforward given the discussion before Lemma 1. We now show the second part. It suffices to construct a structure that can rationalize the data and also satisfy (1), Assumption 1 and $\mathcal{H}_0$. Given the observed data, denoted by F^*_{YDXZ}, we now construct a data generating structure for it. In the following proof, we use $Q^*_{W|X}$ to denote the quantile function of W given X, obtained from F^*_{YDXZ}. Similarly, we define $\delta^*(x)$ and $p^*(x,z)$. To begin with our construction, let $\epsilon\sim U[0,1]$, $X\sim F^*_X$ and $Z\sim F^*_Z$. Moreover, let X, ϵ and Z be mutually independent for $F_{XZ\epsilon}$. To complete our construction, it suffices to define the probability distribution $\mathbf{P}(D=1|X,Z,\epsilon)$ and the function g for Y. Let

$$g(d,x,\tau)=Q^*_{W|X}(\tau|x)-(1-d)\times\delta^*(x)$$

and $Y=g(D,X,\epsilon)$. Regarding $\mathbf{P}(D=1|X,Z,\epsilon)$, given that we have constructed $F_{\epsilon|XZ}$, we can equivalently define the joint distribution of (D,ϵ) given X and Z. Let further

$$\mathbf{P}(D=1;\epsilon\le\tau|X=x,Z=z)=p^*(x,z)\times F^*_{Y|DXZ}(Q^*_{W|X}(\tau|x)|1,x,z),\quad \forall\ x,z.$$

By construction, Assumption 1 and $\mathcal{H}_0$ are satisfied. Thus, it suffices to show the observational equivalence. First, let $\tau=1$ in the construction of $P(D=1;\epsilon\le\tau|X,Z)$, then it follows that $\mathbf{P}(D=1|X=x,Z=z)=p^*(x,z)$. Moreover, note that

$$\begin{aligned}\mathbf{P}(Y\le y|D=1,X=x,Z=z)&=\frac{\mathbf{P}(g(1,x,\epsilon)\le y;D=1|X=x,Z=z)}{\mathbf{P}(D=1|X=x,Z=z)}\\&=\frac{\mathbf{P}(Q^*_{W|X}(\epsilon|x)\le y;D=1|X=x,Z=z)}{\mathbf{P}(D=1|X=x,Z=z)}=F^*_{Y|DXZ}(y|1,x,z).\end{aligned}$$

A.3 Proof of Theorem 1

Proof: Because Proposition 1 provides the "only if" part, then it suffices to show the "if" part. Suppose $W\perp\!\!\!\perp Z|X$. By the definition of W, we have: for any $y\in\mathrm{R}$,

$$\begin{aligned}&\mathbf{P}(Y\le y,D=1|X,Z=1)+\mathbf{P}(Y+\delta(X)\le y,D=0|X,Z=1)\\&\qquad=\mathbf{P}(Y\le y,D=1|X,Z=0)+\mathbf{P}(Y+\delta(X)\le y,D=0|X,Z=0).\end{aligned}$$

It follows that

$$\begin{aligned}&\mathbf{P}(Y\le y,D=1|X,Z=1)-\mathbf{P}(Y\le y,D=1|X,Z=0)\\&\qquad=\mathbf{P}(Y\le y-\delta(X),D=0|X,Z=0)-\mathbf{P}(Y\le y-\delta(X),D=0|X,Z=1).\end{aligned}\tag{10}$$

Denote $V\equiv\nu(X,\epsilon)$ and

$$\begin{aligned}\Delta_0(\tau,x)&\equiv\mathbf{P}(V\le\tau,D=0|X=x,Z=1)-\mathbf{P}(V\le\tau,D=0|X=x,Z=0);\\ \Delta_1(\tau,x)&\equiv\mathbf{P}(V\le\tau,D=1|X=x,Z=0)-\mathbf{P}(V\le\tau,D=1|X=x,Z=1).\end{aligned}$$

By Assumptions 1 and 3, we have

$$\Delta_0(\tau,x)=\mathbf{P}(V\le\tau,\eta\in\mathcal{C}_x|X=x)=\Delta_1(\tau,x)$$

which is strictly monotone in $\tau\in\mathcal{S}_{V|X=x,\ \eta\in\mathcal{C}_x}$. Moreover, there is $\mathcal{S}_{V|X=x,\ \eta\in\mathcal{C}_x}=\mathcal{S}_{V|X=x}$ under Assumptions 2 and 4.

Therefore, we have

$$\begin{aligned}&\mathbf{P}(Y\le y,D=1|X=x,Z=0)-\mathbf{P}(Y\le y,D=1|X=x,Z=1)\\&=\Delta_1(\tilde{g}^{-1}(1,x,y),x)\\&=\Delta_0(\tilde{g}^{-1}(1,x,y),x)\\&=\mathbf{P}(Y\le\tilde{g}(0,x,\tilde{g}^{-1}(1,x,y)),D=0|X=x,Z=1)\\&\quad-\mathbf{P}(Y\le\tilde{g}(0,x,\tilde{g}^{-1}(1,x,y)),D=0|X=x,Z=0),\end{aligned}$$

where $\tilde{g}^{-1}(1,x,\cdot)$ is the inverse function of $\tilde{g}(1,x,\cdot)$ and $\tilde{g}$ is a monotone function introduced in Assumption 2. Note that both sides are strictly monotone in $y\in\mathcal{S}_{\tilde{g}(1,X,V)|X=x}$ since $\Delta_d(\cdot,x)$ is strictly monotone on $\mathcal{S}_{V|X=x}$ under Assumption 4.

Combine the above result with (10), then we have

$$\tilde{g}(0,x,\tilde{g}^{-1}(1,x,y))=y-\delta(x),\quad\forall x\in\mathcal{S}_X,\ y\in\mathcal{S}_{\tilde{g}(1,x,V)|X=x}.$$

Let $y=\tilde{g}(1,x,\tau)$ for some $\tau\in\mathcal{S}_{V|X=x}$. Then the above equation becomes

$$\tilde{g}(0,x,\tau)=\tilde{g}(1,x,\tau)-\delta(x).$$

A.4 Proof of Theorem 2

Proof: Let $\mathbb{1}^*_{WXZ}(w,x,z)=\mathbb{1}(W\le w)\times\mathbb{1}_{XZ}(x,z)$ and $\mathbb{1}^*_{\widehat{W}XZ}(w,x,z)=\mathbb{1}(\widehat{W}\le w)\times\mathbb{1}_{XZ}(x,z)$. Let further $\mathbb{1}^*_{W(\tilde{\delta})XZ}(w,x,z)=\mathbb{1}(W(\tilde{\delta})\le w)\times\mathbb{1}_{XZ}(x,z)$, where $W(\tilde{\delta})=Y+(1-D)\tilde{\delta}(X)$, be a function indexed by $\tilde{\delta}(\cdot)\in\mathrm{R}^{\mathcal{S}_X}$. By the definition, $\mathbb{1}^*_{W(\delta)XZ}(w,x,z)=\mathbb{1}^*_{WXZ}(w,x,z)$ and $\mathbb{1}^*_{W(\hat{\delta})XZ}(w,x,z)=\mathbb{1}^*_{\widehat{W}XZ}(w,x,z)$.

We first derive the asymptotic of $\sqrt{n}[\widehat{F}_{W|XZ}(w|x,z)-F_{W|XZ}(w|x,z)]$. By the definition,

$$F_{W|XZ}(w|x,z)=\frac{\mathbf{E}[\mathbb{1}^*_{WXZ}(w,x,z)]}{\mathbf{E}[\mathbb{1}_{XZ}(x,z)]}\ \text{ and }\ \widehat{F}_{W|XZ}(w|x,z)=\frac{\mathbf{E}_n[\mathbb{1}^*_{\widehat{W}XZ}(w,x,z)]}{\mathbf{E}_n[\mathbb{1}_{XZ}(x,z)]}.$$

In the expectation $\mathbf{E}[\mathbb{1}^*_{W(\hat{\delta})XZ}(\cdot,x,z)]$ discussed below, we treat $\hat{\delta}$ as an index rather than a random object. Note that

$$\begin{aligned}
&\mathbf{E}_n[\mathbb{1}^*_{\widehat{W}XZ}(\cdot,x,z)] = \mathbf{E}_n[\mathbb{1}^*_{WXZ}(\cdot,x,z)] - \mathbf{E}[\mathbb{1}^*_{WXZ}(\cdot,x,z)] + \mathbf{E}[\mathbb{1}^*_{W(\hat{\delta})XZ}(\cdot,x,z)] \\
&\quad + \Big\{\mathbf{E}_n[\mathbb{1}^*_{W(\hat{\delta})XZ}(\cdot,x,z)] - \mathbf{E}[\mathbb{1}^*_{W(\hat{\delta})XZ}(\cdot,x,z)] - \mathbf{E}_n[\mathbb{1}^*_{W(\delta)XZ}(\cdot,x,z)] + \mathbf{E}[\mathbb{1}^*_{W(\delta)XZ}(\cdot,x,z)]\Big\} \\
&\qquad = \mathbf{E}_n[\mathbb{1}^*_{WXZ}(\cdot,x,z)] - \mathbf{E}[\mathbb{1}^*_{WXZ}(\cdot,x,z)] + \mathbf{E}[\mathbb{1}^*_{W(\hat{\delta})XZ}(\cdot,x,z)] + o_p(n^{-1/2}),
\end{aligned}$$

where the last step follows from the fact that $\sqrt{n}\big(\mathbf{E}_n[\mathbb{1}^*_{W(\delta)XZ}(\cdot,x,z)] + \mathbf{E}[\mathbb{1}^*_{W(\delta)XZ}(\cdot,x,z)]\big)$ is stochastically equicontinuous by the empirical process theory (see, e.g., van der Vaart and Wellner, 2007). By Taylor expansion,

$$\sqrt{n}\ \Big\{\mathbf{E}[\mathbb{1}^*_{W(\hat{\delta})XZ}(\cdot,x,z)] - F_{W|XZ}(w|x,z)\Big\} = \frac{\partial \mathbf{E}[\mathbb{1}^*_{W(\delta)XZ}(w,x,z)]}{\partial \delta} \times \sqrt{n}(\hat{\delta}-\delta) + o_p(1).$$

Note that $\frac{\partial \mathbf{E}[\mathbb{1}^*_{W(\delta)XZ}(w,x,z)]}{\partial \delta(x')} = 0$ for all $x' \neq x$ and $\frac{\partial \mathbf{E}[\mathbb{1}^*_{W(\delta)XZ}(w,x,z)]}{\partial \delta(x)} = -f_{W|DXZ}(w|0,x,z) \times \mathbf{P}(D = 0, X = x, Z = z)$. Therefore, we have

$$\begin{aligned}
&\sqrt{n}\ \Big\{\mathbf{E}[\mathbb{1}^*_{W(\hat{\delta})XZ}(\cdot,x,z)] - F_{W|XZ}(w|x,z)\Big\} \\
&\quad + \sqrt{n}\left\{\mathbf{E}_n[\mathbb{1}^*_{WXZ}(\cdot,x,z)] - \mathbf{E}[\mathbb{1}^*_{WXZ}(\cdot,x,z)]\right\} - f_{WDXZ}(w,0,x,z) \times \sqrt{n}[\hat{\delta}(x) - \delta(x)] + o_p(1).
\end{aligned}$$

Moreover, $\mathbf{E}_n[\mathbb{1}_{XZ}(x,z)] = \mathbf{P}(X = x, Z = z) + O_p(n^{-1/2})$ under the central limit theorem. Thus, by Slutsky's theorem, we have

$$\begin{aligned}
&\sqrt{n}\left[\widehat{F}_{W|XZ}(w|x,1) - \widehat{F}_{W|XZ}(w|x,0)\right] - \sqrt{n}\left[F_{W|XZ}(w|x,1) - F_{W|XZ}(w|x,0)\right] \\
&= \frac{\sqrt{n}\big\{\mathbf{E}_n[\mathbb{1}^*_{WXZ}(w,x,1)] - \mathbf{E}[\mathbb{1}^*_{WXZ}(w,x,1)]\big\} - f_{WDXZ}(w,0,x,1) \times \sqrt{n}[\hat{\delta}(x) - \delta(x)]}{\mathrm{P}(X = x, Z = 1)} \\
&\quad - \frac{\sqrt{n}\big\{\mathbf{E}_n[\mathbb{1}^*_{WXZ}(w,x,0)] - \mathbf{E}[\mathbb{1}^*_{WXZ}(w,x,0)]\big\} - f_{WDXZ}(w,0,x,0) \times \sqrt{n}[\hat{\delta}(x) - \delta(x)]}{\mathbf{P}(X = x, Z = 0)} \\
&\quad + \frac{\sqrt{n}\mathbf{P}(W \leq w, X = x, Z = 1)}{\mathbf{E}_n\mathbb{1}_{XZ}(x,1)} - \frac{\sqrt{n}\mathbf{P}(W \leq w, X = x, Z = 0)}{\mathbf{E}_n\mathbb{1}_{XZ}(x,0)} + o_p(1).
\end{aligned}$$

By applying Taylor expansion, we have

$$\begin{aligned}
&\frac{\sqrt{n}\mathbf{P}(W \leq w, X = x, Z = z)}{\mathbf{E}_n\mathbb{1}_{XZ}(x,z)} - \sqrt{n}\ F_{W|XZ}(w|x,z) \\
&\qquad = -F_{W|XZ}(w|x,z) \times \frac{\sqrt{n}\big[\mathbf{E}_n\mathbb{1}_{XZ}(x,z) - \mathbf{P}(X = x, Z = z)\big]}{\mathbf{P}(X = x, Z = z)} + o_p(1).
\end{aligned}$$

Moreover, applying Lemma 3, we have

$$
\begin{aligned}
& \sqrt{n}\left[\widehat{F}_{W|XZ}(w|x,1)-\widehat{F}_{W|XZ}(w|x,0)\right]-\sqrt{n}\left[F_{W|XZ}(w|x,1)-F_{W|XZ}(w|x,0)\right] \\
= \;& \sqrt{n}\mathbf{E}_n\left\{[\mathbb{1}(W\leq w)-F_{W|XZ}(w|x,1)]\times\frac{\mathbb{1}_{XZ}(x,1)}{\mathbf{P}(X=x,Z=1)}\right\} \\
- \;& \sqrt{n}\mathbf{E}_n\left\{[\mathbb{1}(W\leq w)-F_{W|XZ}(w|x,0)]\times\frac{\mathbb{1}_{XZ}(x,0)}{\mathbf{P}(X=x,Z=0)}\right\} \\
+ \;& \kappa(w,x)\times\sqrt{n}\mathbf{E}_n\left\{\left[W-\mathbf{E}(W|X=x,Z=0)\right]\times\frac{\mathbb{1}_{XZ}(x,1)}{\mathbf{P}(X=x,Z=1)}\right\} \\
- \;& \kappa(w,x)\times\sqrt{n}\mathbf{E}_n\left\{\left[W-\mathbf{E}(W|X=x,Z=1)\right]\times\frac{\mathbb{1}_{XZ}(x,0)}{\mathbf{P}(X=x,Z=0)}\right\}+o_p(1).
\end{aligned}
$$

Under the null hypothesis, there is

$$
\begin{aligned}
& \sqrt{n}\left[\widehat{F}_{W|XZ}(w|x,1)-\widehat{F}_{W|XZ}(w|x,0)\right] \\
= \;& \sqrt{n}\mathbf{E}_n\left\{[\mathbb{1}(W\leq w)-F_{W|X}(w|x)]\times\left[\frac{\mathbb{1}_{XZ}(x,1)}{\mathbf{P}(X=x,Z=1)}-\frac{\mathbb{1}_{XZ}(x,0)}{\mathbf{P}(X=x,Z=0)}\right]\right\} \\
+ \;& \kappa(w,x)\times\sqrt{n}\mathbf{E}_n\left\{\left[W-\mathbf{E}(W|X=x)\right]\times\left[\frac{\mathbb{1}_{XZ}(x,1)}{\mathbf{P}(X=x,Z=1)}-\frac{\mathbb{1}_{XZ}(x,0)}{\mathbf{P}(X=x,Z=0)}\right]\right\}+o_p(1) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi_{wx,i}+\phi_{wx,i})+o_p(1)
\end{aligned}
$$

where $\psi_{wx,i}$ and $\phi_{wx,i}$ are defined by (6) and (7). Following the empirical process theory (see, e.g., Kim and Pollard, 1990, Theorem 2.7) and the Continuous Mapping theorem (see, e.g., van der Vaart and Wellner, 2007), we have $\widehat{\mathcal{T}}_n \xrightarrow{d} \sup_{w\in\mathrm{R};\; x\in\mathcal{S}_X}\;|\mathcal{Z}(w,x)|$.

A.5 Proof of Lemma 2

Proof: Fix $X=x$ and w.l.o.g., let $z=1$. Note that

$$
\begin{aligned}
& \widehat{G}(w,x,1)-\widetilde{G}(w,x,1) \\
= \;& \mathbf{E}_n\left\{\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(w-\widehat{W})\left[\mathbb{1}(\widehat{W}\leq w)-\mathbb{1}(W\leq w)\right]\right\} \\
= \;& \mathbf{E}_n\left\{\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(w-\widehat{W})\left[\mathbb{1}(\widehat{W}\leq w)-\mathbb{1}(W\leq w)\right]\times\mathbb{1}(|W-w|\leq n^{-r})\right\} \\
+ \;& \mathbf{E}_n\left\{\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(w-\widehat{W})\left[\mathbb{1}(\widehat{W}\leq w)-\mathbb{1}(W\leq w)\right]\times\mathbb{1}(|W-w|> n^{-r})\right\} \\
\equiv \;& \mathrm{T}_1+\mathrm{T}_2
\end{aligned}
$$

where $r\in(\frac{1}{4},\iota)$. It suffices to show both T_1 and T_2 are $o_p(n^{-\frac{1}{2}})$.

First, note that

$$
\begin{aligned}
\mathrm{T}_1 = \;& \mathbf{E}_n\left\{\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(w-W)\left[\mathbb{1}(\widehat{W}\leq w)-\mathbb{1}(W\leq w)\right]\times\mathbb{1}(|W-w|\leq n^{-r})\right\} \\
& +\mathbf{E}_n\left\{\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(W-\widehat{W})\left[\mathbb{1}(\widehat{W}\leq w)-\mathbb{1}(W\leq w)\right]\times\mathbb{1}(|W-w|\leq n^{-r})\right\}.
\end{aligned}
$$

Because

$$\mathbf{E}\left|\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(w-W)\left[\mathbb{1}(\widehat{W}\le w)-\mathbb{1}(W\le w)\right]\times\mathbb{1}(|W-w|\le n^{-r})\right|$$
$$\le \mathbf{E}\left|\hat{f}_{XZ}(X_1,0)\times(w-W)\times\mathbb{1}(|W-w|\le n^{-r})\right| = O(1)\times O(n^{-2r}) = o(n^{-\frac{1}{2}}),$$

where the last step holds because $r>\frac{1}{4}$. Moreover,

$$\mathbf{E}\left|\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(W-\widehat{W})\left[\mathbb{1}(\widehat{W}\le w)-\mathbb{1}(W\le w)\right]\times\mathbb{1}(|W-w|\le n^{-r})\right|$$
$$\le \mathbf{E}\left|\hat{f}_{XZ}(X_1,0)\times(W-\widehat{W})\times\mathbb{1}(|W-w|\le n^{-r})\right| = O(1)\times O(n^{-\iota})\times O(n^{-r}) = o(n^{-\frac{1}{2}}).$$

Then, we have $\mathrm{T}_1 = o_p(n^{-\frac{1}{2}})$.

For term T_2, note that

$$\mathbf{E}|\mathrm{T}_2| \le \frac{\overline{K}}{h}\times\sqrt{\mathbf{E}[(w-\widehat{W})^2]}\times\sqrt{\mathbf{P}\left(|\widehat{W}-W|>n^{-r}\right)}$$
$$\le \frac{\overline{K}}{h}\times\sqrt{\mathbf{E}[\widehat{W}^2]-2w\cdot\mathbf{E}[\widehat{W}]+w^2}\times\sqrt{\mathbf{P}\left[|\hat{\delta}(X)-\delta(X)|>n^{-r}\right]},$$

where $\overline{K}$ is the upper bound of $K(\cdot)$. Because W is a bounded random variable and w belongs to a compact set, then $\sqrt{\mathbf{E}\widehat{W}^2-2w\cdot\mathbf{E}[\widehat{W}]+w^2}=O(1)$. Moreover, by Lemma 4, $\mathbf{E}|\mathrm{T}_2|\le o(n^{-k})$ for any $k>0$. Hence, $\mathrm{T}_2=o_p(n^{-\frac{1}{2}})$.

A.6 Proof of Theorem 3

Proof: By Lemma 2, we have

$$\widehat{\mathcal{T}}^c_n = \sqrt{n}\left|\widetilde{G}(w,x,1)-\widetilde{G}(w,x,0)\right| + o_p(1).$$

Note that

$$\widetilde{G}(w,x,z) = \mathrm{U}_1(w,x,z)+\mathrm{U}_2(w,x,z)+o_p(n^{-1/2})$$

where

$$\mathrm{U}_1(w,x,z) \equiv \frac{1}{n}\sum_{i=1}^n \mathbb{1}^*_{W_i}(w)\times\mathbb{1}^*_{X_iZ_i}(x,z)\times\hat{f}_{XZ}(X_i,z')\times(W_i-\widehat{W}_i);$$
$$\mathrm{U}_2(w,x,z) \equiv \frac{1}{n}\sum_{i=1}^n \mathbb{1}^*_{W_i}(w)\times\mathbb{1}^*_{X_iZ_i}(x,z)\times\hat{f}_{XZ}(X_i,z')\times(w-W_i).$$

Therefore,

$$\begin{aligned}
&\sqrt{n}\left[\widetilde{G}(w,x,1)-\widetilde{G}(w,x,0)\right]\\
&\quad=\sqrt{n}\left\{\mathrm{U}_1(w,x,1)-\mathrm{U}_1(w,x,0)-[\mathbf{E}\mathrm{U}_1(w,x,1)-\mathbf{E}\mathrm{U}_1(w,x,0)]\right\}\\
&\qquad+\sqrt{n}\left\{\mathrm{U}_2(w,x,1)-\mathrm{U}_2(w,x,0)-[\mathbf{E}\mathrm{U}_2(w,x,1)-\mathbf{E}\mathrm{U}_2(w,x,0)]\right\}\\
&\qquad+\sqrt{n}\left[\mathbf{E}\mathrm{U}_1(w,x,1)-\mathbf{E}\mathrm{U}_1(w,x,0)\right]+\sqrt{n}\left[\mathbf{E}\mathrm{U}_2(w,x,1)-\mathbf{E}\mathrm{U}_2(w,x,0)\right].
\end{aligned}$$

We first look at those U_2 terms. By definition,

$$\begin{aligned}\mathrm{U}_2(w,x,z) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\{\mathbb{1}^*_{X_iZ_i}(x,z)\lambda(W_i-w)\times K_{X_j,h}(X_i)\mathbb{1}(Z_j=z')\}\\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\zeta_{n,ij}(w,x,z)\end{aligned}$$

where $\zeta_{n,ij}(w,x,z)=\mathbb{1}^*_{X_iZ_i}(x,z)\times\lambda(W_i-w)\times K_{X_j,h}(X_i)\times\mathbb{1}(Z_j=z')$.

Let $\zeta^*_{n,ij}(w,x,z)=\frac{1}{2}\left[\zeta_{n,ij}(w,x,z)+\zeta_{n,ji}(w,x,z)\right]$. Then, $\zeta^*_{n,ij}$ is symmetric in indices i and j. Therefore,

$$\mathrm{U}_2(w,x,z)=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\zeta^*_{n,ij}(w,x,z),$$

which is a $\mathcal{U}$-process indexed by (w,x,z_ℓ). By Nolan and Pollard (1988, Theorem 5) and Powell, Stock, and Stoker (1989, Lemma 3.1),

$$\begin{aligned}&\mathrm{U}_2(w,x,z)-\mathbf{E}\mathrm{U}_2(w,x,z)\\ =\;&\frac{2}{n}\sum_{i=1}^{n}\left\{\mathbf{E}[\zeta^*_{n,ij}(w,x,z)|Y_i,D_i,X_i,Z_i]-\mathrm{E}[\zeta^*_{n,ij}(w,x,z)]\right\}+o_p(n^{-1/2}).\end{aligned}$$

where the $o_p(n^{-1/2})$ applies uniformly over (w,x). Note that

$$\begin{aligned}&\mathbf{E}[\zeta^*_{n,ij}(w,x,z)|Y_i,D_i,X_i,Z_i]\\ &\qquad=\frac{1}{2}\Big\{\mathbb{1}^*_{XZ}(x,z)f_{XZ}(X,z')\lambda(W-w)+\mathbb{1}^*_{XZ}(x,z')f_{XZ}(X,z)\Pi(w|X,z)\Big\}+o_p(1).\end{aligned}$$

Next, we derive $\mathbf{E}[\zeta^*_{n,ij}(w,x,z)]$. Let $u_1(w,x,z)=\mathbf{E}[\mathbb{1}^*_{XZ}(x,z)f_{XZ}(X,z')\lambda(W-w)]$ and $u_2(w,x,z)=\mathbf{E}[\mathbb{1}^*_{XZ}(x,z')f_{XZ}(X,z)\Pi(w|X,z)]$. Note that under $\mathcal{H}_0$

$$u_1(w,x,z)=u_2(w,x,z)=\int\mathbb{1}(X\leq x)\Pi(w|X)f_{X|Z}(X|1)f_{X|Z}(X|0)dX\times\mathbf{P}(Z=1)\mathbf{P}(Z=0),$$

invariant with z. Therefore, $\mathbf{E}[\zeta^*_{n,ij}(w,x,z)]=\frac{1}{2}[u_1(w,x,z)+u_2(w,x,z)]$ is also invariant with z. Let $u^e(w,x)=\mathbf{E}[\zeta^*_{n,ij}(w,x,z)]$. Moreover, by Powell, Stock, and Stoker (1989, Theorem 3.1),

$$\begin{aligned}&\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\left\{\mathbf{E}[\zeta^*_{n,ij}(w,x,z)|Y_i,D_i,X_i]-\mathbf{E}[\zeta^*_{n,ij}(w,x,z)]\right\}\\ &\qquad=\mathbf{E}_n\left\{\mathbb{1}^*_{XZ}(x,z)f_{XZ}(X,z')\lambda(W-w)-u^e(w,x)\right\}\\ &\qquad\qquad+\mathbf{E}_n\left\{\mathbb{1}^*_{XZ}(x,z')f_{XZ}(X,z)\Pi(w|X,z)-u^e(w,x)\right\}+o_p(n^{-\frac{1}{2}}),\end{aligned}$$

where the $o_p(n^{-1/2})$ holds uniformly over (w,x). Moreover, under $\mathcal{H}_0$, there is $\Pi(w|X,z)=\mathbf{E}(\lambda(W-w)|X)$. Thus,

$$\begin{aligned}&\mathrm{U}_2(w,x,1)-\mathrm{U}_2(w,x,0)-[\mathbf{E}\mathrm{U}_2(w,x,1)-\mathbf{E}\mathrm{U}_2(w,x,0)]\\ =\;&\mathbf{E}_n\left\{\Big[\frac{\mathbb{1}^*_{XZ}(x,1)}{f_{XZ}(X,1)}-\frac{\mathbb{1}^*_{XZ}(x,0)}{f_{XZ}(X,0)}\Big]f_{XZ}(X,0)f_{XZ}(X,1)\big[\lambda(W-w)-\mathbf{E}(\lambda(W-w)|X)\big]\right\}+o_p(n^{-\frac{1}{2}}).\end{aligned}$$

We now turn to $\mathrm{U}_1(w,x,z)$. Note that

$$\mathrm{U}_1(w,x,z) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{1}^*_{W_iX_iZ_i}(w,x,z)f_{XZ}(X_i,z')(1-D_i)[\hat{\delta}(X_i)-\delta(X_i)]\right\} + o_p(n^{-\frac{1}{2}}),$$

provided that $\mathbf{E}\left|\left[\hat{f}_{XZ}(X_i,z')-f_{XZ}(X_i,z')\right]\times\left[\hat{\delta}(X_i)-\delta(X_i)\right]\right| = o_p(n^{-\frac{1}{2}})$ holds. By a similar decomposition argument on $\hat{\delta}(X)-\delta(X)$ in Lemma 4, we have

$$\mathrm{U}_1(w,x,z) = -\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\xi_{n,ij}(w,x,z) + o_p(n^{-1/2})$$

where $\xi_{n,ij}(w,x,z) = \mathbb{1}^*_{W_iX_iZ_i}(w,x,z)f_{XZ}(X_i,z')(1-D_i)\frac{[W_j-\mathbf{E}(W_j|X_i)]K_{X_j,h}(X_i)}{p(X_i,1)-p(X_i,0)}\left[\frac{\mathbb{1}(Z_j=1)}{f_{XZ}(X_i,1)}-\frac{\mathbb{1}(Z_j=0)}{f_{XZ}(X_i,0)}\right]$. Moreover, let $\xi^*_{n,ij}(w,x,z)=\frac{1}{2}[\xi_{n,ij}(w,x,z)+\xi_{n,ji}(w,x,z)]$. By a similar argument for U_2,

$$\begin{aligned}&\mathrm{U}_1(w,x,z)-\mathbf{E}\mathrm{U}_1(w,x,z)\\ &\qquad = -\frac{2}{n}\sum_{i=1}^{n}\left\{\mathbf{E}[\xi^*_{n,ij}(w,x,z)|Y_i,D_i,X_i,Z_i]-\mathbf{E}[\xi^*_{n,ij}(w,x,z)]\right\}+o_p(n^{-1/2}).\end{aligned}$$

Note that $\mathbf{E}[\xi_{n,ij}(w,x,z)|Y_i,D_i,X_i,Z_i]=0$ and

$$\begin{aligned}&\mathbf{E}[\xi_{n,ji}(w,x,z)|Y_i,D_i,X_i,Z_i] = \mathbf{E}\left\{\mathbf{E}[\xi_{n,ji}(w,x,z)|X_j,Z_j,Y_i,D_i,X_i,Z_i]\middle|Y_i,D_i,X_i,Z_i\right\}\\
&= \mathbf{E}\Bigg\{\mathbb{1}^*_{X_jZ_j}(x,z)f_{XZ}(X_j,z')\mathbf{P}(W\le w;D=0|X_j,Z_j)[W_i-\mathbf{E}(W|X_j)]\\
&\quad\times\frac{K_{X_j,h}(X_i)}{p(X_j,1)-p(X_j,0)}\left[\frac{\mathbb{1}(Z_i=1)}{f_{XZ}(X_j,1)}-\frac{\mathbb{1}(Z_i=0)}{f_{XZ}(X_j,0)}\right]\Bigg|Y_i,D_i,X_i,Z_i\Bigg\}\\
&= F^*_{WD|XZ}(w,0|X_i,z)[W_i-\mathbf{E}(W|X_i)]\frac{f_{XZ}(X_i,0)f_{XZ}(X_i,1)}{p(X_i,1)-p(X_i,0)}\left[\frac{\mathbb{1}^*_{X_i,Z_i}(x,1)}{f_{XZ}(X_i,1)}-\frac{\mathbb{1}^*_{X_i,Z_i}(x,0)}{f_{XZ}(X_i,0)}\right]+o_p(1)\end{aligned}$$

where the last step comes from Bochner's Lemma (see, e.g., Rudin, 1962) and uses the fact the integrant equals zero if $Z_j=z'$.

Thus, we have

$$\begin{aligned}&\mathrm{U}_1(w,x,z)-\mathbf{E}\mathrm{U}_1(w,x,z)\\
&= -\mathbf{E}_n\left\{[W-\mathbf{E}(W|X)]\frac{F^*_{WD|XZ}(w,0|X,z)}{p(X,1)-p(X,0)}\left[\frac{\mathbb{1}^*_{XZ}(x,1)}{f_{XZ}(X,1)}-\frac{\mathbb{1}^*_{XZ}(x,0)}{f_{XZ}(X,0)}\right]f_{XZ}(X,1)f_{XZ}(X,0)\right\}+o_p(n^{-\frac{1}{2}}),\end{aligned}$$

where the $o_p(n^{-1/2})$ holds uniformly over (w,x). It follows that

$$\mathrm{U}_1(w,x,1)-\mathbf{E}\mathrm{U}_1(w,x,1)-[\mathrm{U}_1(w,x,0)-\mathbf{E}\mathrm{U}_1(w,x,0)] = \mathbf{E}_n\phi^c_{wx}+o_p(n^{-\frac{1}{2}}).$$

By Assumption 9, we have $\mathbf{E}\mathrm{U}_1(w,x;z)=o_p(n^{-\frac{1}{2}})$. Therefore, under $\mathcal{H}_0$,

$$\begin{aligned}&\sqrt{n}\left[\widetilde{G}(w,x,1)-\widetilde{G}(w,x,0)\right]\\
=\;&\sqrt{n}\left\{\mathrm{U}_1(w,x,1)-\mathrm{U}_1(w,x,0)-[\mathbf{E}\mathrm{U}_1(w,x,1)-\mathbf{E}\mathrm{U}_1(w,x,0)]\right\}\\
+\;&\sqrt{n}\left\{\mathrm{U}_2(w,x,1)-\mathrm{U}_2(w,x,0)-[\mathbf{E}\mathrm{U}_2(w,x,1)-\mathbf{E}\mathrm{U}_2(w,x,0)]\right\}+o_p(1)\\
=\;&\sqrt{n}\times\mathbf{E}_n(\psi^c_{wx}+\phi^c_{wx})+o_p(1),\end{aligned}$$

which converges to a zero-mean Gaussian process with the given covariance kernel.

B Technical Lemmas

Let $\Delta p(x) \equiv p(x,1) - p(x,0)$, which is strictly positive by Assumption 1.

Lemma 3 *Suppose Assumptions 1 and 5 hold. Then, we have*

$$\sqrt{n}[\hat{\delta}(x) - \delta(x)] = \frac{1}{\Delta p(x)} \times \sqrt{n}\mathbf{E}_n \left\{ \left[W - \mathbf{E}(W|X=x, Z=0)\right] \times \frac{\mathbb{1}_{XZ}(x,1)}{\mathbf{P}(X=x, Z=1)} \right\}$$
$$- \frac{1}{\Delta p(x)} \times \sqrt{n}\mathbf{E}_n \left\{ \left[W - \mathbf{E}(W|X=x, Z=1)\right] \times \frac{\mathbb{1}_{XZ}(x,0)}{\mathbf{P}(X=x, Z=0)} \right\} + o_p(1). \quad (11)$$

Proof of Lemma 3: Fix $X = x$. For expositional simplicity, we suppress x in the following proof. Moreover, let $\mathrm{A}_n(z) = \mathbf{E}_n[Y\mathbb{1}_{XZ}(x,z)]$, $\mathrm{B}_n(z) = \mathbf{E}_n[D\mathbb{1}_{XZ}(x,z)]$, $\mathrm{C}_n(z) = \mathbf{E}_n\mathbb{1}_{XZ}(x,z)$, $\mathrm{A}(z) = \mathbf{E}[Y\mathbb{1}_{XZ}(x,z)]$, $\mathrm{B}(z) = \mathbf{E}[D1_{XZ}(x,z)]$ and $\mathrm{C}(z) = \mathbf{E}\mathbb{1}_{XZ}(x,z) = \mathbf{P}(X=x, Z=z)$. By definition, note that

$$\hat{\delta} = \frac{\mathrm{A}_n(1)\mathrm{C}_n(0) - \mathrm{A}_n(0)\mathrm{C}_n(1)}{\mathrm{B}_n(1)\mathrm{C}_n(0) - \mathrm{B}_n(0)\mathrm{C}_n(1)} \quad \text{and} \quad \delta = \frac{\mathrm{A}(1)\mathrm{C}(0) - \mathrm{A}(0)\mathrm{C}(1)}{\mathrm{B}(1)\mathrm{C}(0) - \mathrm{B}(0)\mathrm{C}(1)}.$$

It follows that

$$\hat{\delta} - \delta = \frac{\mathrm{A}_n(1)\mathrm{C}_n(0) - \mathrm{A}_n(0)\mathrm{C}_n(1) - [\mathrm{A}(1)\mathrm{C}(0) - \mathrm{A}(0)\mathrm{C}(1)]}{\mathrm{B}_n(1)\mathrm{C}_n(0) - \mathrm{B}_n(0)\mathrm{C}_n(1)}$$
$$+ \left\{ \frac{\mathrm{A}(1)\mathrm{C}(0) - \mathrm{A}(0)\mathrm{C}(1)}{\mathrm{B}_n(1)\mathrm{C}_n(0) - \mathrm{B}_n(0)\mathrm{C}_n(1)} - \frac{\mathrm{A}(1)\mathrm{C}(0) - \mathrm{A}(0)\mathrm{C}(1)}{\mathrm{B}(1)\mathrm{C}(0) - \mathrm{B}(0)\mathrm{C}(1)} \right\} \equiv \mathrm{I} + \mathrm{II}.$$

We first look at the term I. By the Central Limit Theorem and Assumption 5, we have $\mathrm{A}_n(z) = \mathrm{A}(z) + O_p(n^{-1/2})$, $\mathrm{B}_n(z) = \mathrm{B}(z) + O_p(n^{-1/2})$ and $\mathrm{C}_n(z) = \mathrm{C}(z) + O_p(n^{-1/2})$. Therefore,

$$\begin{aligned}
\mathrm{I} &= \frac{[\mathrm{A}_n(1) - \mathrm{A}(1)]\,\mathrm{C}(0) + \mathrm{A}(1)\,[\mathrm{C}_n(0) - \mathrm{C}(0)]}{\mathrm{B}(1)\mathrm{C}(0) - \mathrm{B}(0)\mathrm{C}(1)} \\
&\quad - \frac{[\mathrm{A}_n(0) - \mathrm{A}(0)]\,\mathrm{C}(1) + \mathrm{A}(0)\,[\mathrm{C}_n(1) - \mathrm{C}(1)]}{\mathrm{B}(1)\mathrm{C}(0) - \mathrm{B}(0)\mathrm{C}(1)} + o_p(n^{-1/2}) \\
&= \frac{\mathrm{A}_n(1)\mathrm{C}(0) - \mathrm{A}(0)\mathrm{C}_n(1) - \mathrm{A}_n(0)\mathrm{C}(1) + \mathrm{A}(1)\mathrm{C}_n(0)}{\mathrm{B}(1)\mathrm{C}(0) - \mathrm{B}(0)\mathrm{C}(1)} \\
&\quad + \frac{2\,[\mathrm{A}(0)\mathrm{C}(1) - \mathrm{A}(1)\mathrm{C}(0)]}{\mathrm{B}(1)\mathrm{C}(0) - \mathrm{B}(0)\mathrm{C}(1)} + o_p(n^{-1/2}).
\end{aligned}$$

Specifically, we have

$$\begin{aligned}
\mathrm{I} &= \mathbf{E}_n\left\{\left[Y-\mathbf{E}(Y|X=x,Z=0)\right]\times \mathbb{1}_{XZ}(x,1)\right\}\times\frac{\mathbf{P}(X=x,Z=0)}{\mathrm{B}(1)\mathrm{C}(0)-\mathrm{B}(0)\mathrm{C}(1)} \\
&- \mathbf{E}_n\left\{\left[Y-\mathbf{E}(Y|X=x,Z=1)\right]\times \mathbb{1}_{XZ}(x,0)\right\}\times\frac{\mathbf{P}(X=x,Z=1)}{\mathrm{B}(1)\mathrm{C}(0)-\mathrm{B}(0)\mathrm{C}(1)} \\
&+ \frac{2\left[\mathrm{A}(0)\mathrm{C}(1)-\mathrm{A}(1)\mathrm{C}(0)\right]}{\mathrm{B}(1)\mathrm{C}(0)-\mathrm{B}(0)\mathrm{C}(1)}+o_p(n^{-1/2}) \\
&= \frac{1}{\Delta p(x)}\times\mathbf{E}_n\left\{\left[Y-\mathbf{E}(Y|X=x,Z=0)\right]\times\frac{\mathbb{1}_{XZ}(x,1)}{\mathbf{P}(X=x,Z=1)}\right\} \\
&- \frac{1}{\Delta p(x)}\times\mathbf{E}_n\left\{\left[Y-\mathbf{E}(Y|X=x,Z=1)\right]\times\frac{\mathbb{1}_{XZ}(x,0)}{\mathbf{P}(X=x,Z=0)}\right\}-2\delta(x)+o_p(n^{-1/2}).
\end{aligned}$$

For the term II, by a similar argument we have

$$\begin{aligned}
\mathrm{II} = \frac{-\delta(x)}{\Delta p(x)}&\times\mathbf{E}_n\left\{\left[D-p(x,0)\right]\times\frac{\mathbb{1}_{XZ}(x,1)}{\mathbf{P}(X=x,Z=1)}\right\} \\
&+\frac{\delta(x)}{\Delta p(x)}\times\mathbf{E}_n\left\{\left[D-p(x,1)\right]\times\frac{\mathbb{1}_{XZ}(x,0)}{\mathrm{P}(X=x,Z=0)}\right\}+2\delta(x)+o_p(n^{-1/2}).
\end{aligned}$$

By definition of W, we have $W-\mathrm{E}(W|X=x,Z=z)=Y-\mathrm{E}(Y|X=x,Z=z)-[D-p(x,z)]\times\delta(x)$. Summing up I and II, we obtain (11).

Lemma 4 *Suppose Assumptions 6–8 hold. Then for any $k>0$ and $r\in(\frac{1}{4},\iota)$,*

$$\sup_{x\in\mathcal{S}_X} n^k\times\mathbf{P}\left[|\hat{\delta}(x)-\delta(x)|>n^{-r}\right]\to 0.$$

Proof of Lemma 4: First, by a similar decomposition of $\hat{\delta}(x)-\delta(x)$ as that in the proof of Lemma 3, it suffices to show

$$\begin{aligned}
&\sup_x n^k\times\mathbf{P}\left\{|a_n(x,z)-a(x,z)|>\lambda_a\times n^{-r}\right\}\to 0; \\
&\sup_x n^k\times\mathbf{P}\left\{|b_n(x,z)-b(x,z)|>\lambda_b\times n^{-r}\right\}\to 0; \\
&\sup_x n^k\times\mathbf{P}\left\{|q_n(x,z)-q(x,z)|>\lambda_q\times n^{-r}\right\}\to 0,
\end{aligned}$$

where λ_a, λ_b and λ_q are strictly positive constants, and

$$\begin{aligned}
a_n(x,z) &= \frac{1}{n}\sum_{j=1}^n Y_j K_{X_j,h}(x)\mathbb{1}(Z_j=z), \quad a(x,z)=\mathbf{E}(Y|X=x,Z=z)\times q(x,z); \\
b_n(x,z) &= \frac{1}{n}\sum_{j=1}^n D_j K_{X_j,h}(x)\mathbb{1}(Z_j=z), \quad b(x,z)=\mathbf{E}(D|X=x,Z=z)\times q(x,z); \\
q_n(x,z) &= \frac{1}{n}\sum_{j=1}^n K_{X_j,h}(x)\mathbb{1}(Z_j=z).
\end{aligned}$$

For expositional simplicity, we only show the first result. It is straightforward that the rest follow a similar argument.

Let $T_{nxzj} = Y_j K(\frac{X_j - x}{h}) 1(Z_j = z)$ and $\tau_{nxz} = h \times [\lambda_a n^{-r} - |\mathbf{E} a_n(x,z) - a(x,z)|]$. Note that

$$\begin{aligned}
& \mathbf{P}\left[|a_n(x,z) - a(x,z)| > \lambda_a \times n^{-r}\right] \\
\leq \; & \mathbf{P}\left[|a_n(x,z) - \mathbf{E} a_n(x,z)| + |\mathbf{E} a_n(x,z) - a(x,z)| > \lambda_a \times n^{-r}\right] \\
= \; & \mathbf{P}\left\{ \frac{1}{n} \left| \sum_{j=1}^{n} (T_{nxzj} - \mathbf{E} T_{nxzj}) \right| > \tau_{nxz} \right\}.
\end{aligned}$$

Moreover, by Bernstein's tail inequality,

$$\mathbf{P}\left\{ \frac{1}{n} \left| \sum_{i=1}^{n} (T_{xzj} - \mathbf{E} T_{xzj}) \right| > \tau_{nxz} \right\} \leq 2\mathbf{E}\left(- \frac{n \times \tau_{nxz}^2}{2\mathrm{Var}\,(T_{nxzj}) + \frac{2}{3}\overline{K} \times \tau_{nxz}} \right).$$

where $\overline{K}$ is the upper bound of kernel K.

By Assumption 8, $|\mathbf{E} a_n(x,z) - a(x,z)| = O(n^{-\iota}) = o(n^{-r})$. Then, for sufficiently large n, there is $0.5\lambda_a n^{-r} h \leq \tau_n(x,z) \leq \lambda_a n^{-r} h$. Moreover,

$$\mathrm{Var}\,(T_{nxzj}) \leq \mathbf{E} T_{nxzj}^2 \leq \mathbf{E}\big[\mathbf{E}(Y^2|X) K^2(\frac{X-x}{h})\big] \leq Ch,$$

where $C = \sup_x \mathbf{E}[Y^2|X=x] \times \sup_x f_X(x) \times \overline{K} \times \int |K(u)| du < \infty$. It follows that

$$\mathbf{P}\left\{ \frac{1}{n} \left| \sum_{\ell=1}^{n} (T_{xzj} - \mathbf{E} T_{xzj}) \right| > \tau_{nxz} \right\} \leq 2\mathbf{E}\left(- \frac{\frac{\lambda_a}{4} n h n^{-2r}}{2C + \frac{2}{3}\overline{K}\lambda_a n^{-r}} \right).$$

For sufficiently large n, we have $\frac{2}{3}\overline{K}\lambda_a n^{-r} \leq 1$. Therefore, for sufficiently large n,

$$\mathbf{P}\left\{ \frac{1}{n} \Big| \sum_{\ell=1}^{n} (T_{xzj} - \mathbf{E} T_{xzj}) \Big| > \tau_{nxz} \right\} \leq 2\mathbf{E}\left(- \frac{n^{2\iota - 2r}}{2C+1} \right) = o(n^{-k})$$

where the inequality comes from Assumption 8. Note that the upper bound does not depend on x or z. Therefore,

$$\sup_{x,z} \mathbf{P}\left[|a_n(x,z) - a(x,z)| > \lambda_a \times n^{-r}\right] = o(n^{-k}).$$

References

Abadie, A. (2002): "Bootstrap tests for distributional treatment effects in instrumental variable models," *Journal of the American Statistical Association*, 97(457), 284–292.

——— (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 113(2), 231–263.

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70(1), 91–117.

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015): "Estimating conditional average treatment effects," *Journal of Business & Economic Statistics*, 33(4), 485–505.

Andrews, D. W. (1997): "A conditional Kolmogorov test," *Econometrica*, 65(5), 1097–1128.

Angrist, J. D., and W. N. Evans (1998): "Children and their parents' labor supply: Evidence from exogenous variation in family size," *American Economic Review*, 88(3), 450–477.

Angrist, J. D., and A. B. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

Athey, S., and G. Imbens (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360.

Barrett, G. F., and S. G. Donald (2003): "Consistent tests for stochastic dominance," *Econometrica*, 71(1), 71–104.

Bloom, H. S., L. L. Orr, S. H. Bell, G. Cave, F. Doolittle, W. Lin, and J. M. Bos (1997): "The benefits and costs of JTPA Title II-A programs: Key findings from the National Job Training Partnership Act study," *The Journal of Human Resources*, 32(3), 549–576.

Blundell, R., and J. L. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," *Econometric Society Monographs*, 36, 312–357.

Bouezmarni, T., and A. Taamouti (2014): "Nonparametric tests for conditional independence using conditional distributions," *Journal of Nonparametric Statistics*, 26(4), 697–719.

Bronars, S. G., and J. Grogger (1994): "The economic consequences of unwed motherhood: Using twin births as a natural experiment," *American Economic Review*, 32(5), 1141–1156.

Chang, M., S. Lee, and Y.-J. Whang (2015): "Nonparametric tests of conditional treatment effects with an application to single-sex schooling on academic achievements," *The Econometrics Journal*, 18(3), 307–346.

Chernozhukov, V., and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73(1), 245–261.

Chesher, A. (2003): "Identification in nonseparable models," *Econometrica*, 71(5), 1405–1441.

——— (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73(5), 1525–1550.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2008): "Nonparametric tests for treatment effect heterogeneity," *The Review of Economics and Statistics*, 90(3), 389–405.

Dauxois, J., and G. M. Nkiet (1998): "Nonlinear canonical analysis and independence tests," *The Annals of Statistics*, 26(4), 1254–1278.

Delgado, M. A., and W. G. Manteiga (2001): "Significance testing in nonparametric regression based on the bootstrap," *The Annals of Statistics*, 29(5), 1469–1507.

D'Haultfœuille, X., and P. Février (2015): "Identification of nonseparable triangular models with discrete instruments," *Econometrica*, 83(3), 1199–1210.

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(3), 383–397.

Florens, J.-P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," *Econometrica*, 76(5), 1191–1206.

Frölich, M., and B. Melly (2013): "Unconditional quantile treatment effects under endogeneity," *Journal of Business & Economic Statistics*, 31(3), 346–357.

Heckman, J. J., D. Schmierer, and S. Urzua (2010): "Testing the correlated random coefficient model," *Journal of Econometrics*, 158(2), 177–203.

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts," *The Review of Economic Studies*, 64(4), 487–535.

Heckman, J. J., and E. Vytlacil (2001): "Policy-relevant treatment effects," *American Economic Review*, 91(2), 107–111.

——— (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73(3), 669–738.

Hoderlein, S., and E. Mammen (2009): "Identification and Estimation of Marginal Effects in Nonseparable, Nonmomotonic Models," *The Econometrics Journal*, 12(1), 1–25.

Hoderlein, S., and H. White (2012): "Nonparametric identification in nonseparable panel data models with generalized fixed effects," *Journal of Econometrics*, 168(2), 300–314.

Hsu, Y.-C. (2017): "Consistent tests for conditional treatment effects," *The Econometrics Journal*, 20(1), 1–22.

Huang, M., Y. Sun, and H. White (2016): “A flexible nonparametric test for conditional independence,” *Econometric Theory*, 32(6), 1434–1482.

Huang, T.-M. (2010): “Testing conditional independence using maximal nonlinear conditional correlation,” *The Annals of Statistics*, 38(4), 2047–2091.

Huber, M., and G. Mellace (2015): “Testing instrument validity for LATE identification based on inequality moment constraints,” *The Review of Economics and Statistics*, 97(2), 398–411.

Imbens, G. W. (2010): “Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009),” *Journal of Economic Literature*, 48(2), 399–423.

Imbens, G. W., and J. D. Angrist (1994): “Identification and estimation of local average treatment effects,” *Econometrica*, 62(2), 467–475.

Imbens, G. W., and W. K. Newey (2009): “Identification and estimation of triangular simultaneous equations models without additivity,” *Econometrica*, 77(5), 1481–1512.

Imbens, G. W., and D. B. Rubin (1997): “Estimating distributions for outcome compliers models in instrumental variables,” *The Review of Economic Studies*, 64(4), 555–574.

Jacobsen, J. P., J. W. Pearce, and J. L. Rosenbloom (1999): “The effects of childbearing on married women’s labor supply and earnings: Using twin births as a natural experiment,” *The Journal of Human Resources*, 34(3), 449–474.

Jun, S. J., Y. Lee, and Y. Shin (2016): “Treatment effects with unobserved heterogeneity: A set identification approach,” *Journal of Business & Economic Statistics*, 34(2), 302–311.

Kim, J., and D. Pollard (1990): “Cube root asymptotics,” *The Annals of Statistics*, 18(1), 191–219.

Kitagawa, T. (2015): “A test for instrument validity,” *Econometrica*, 83(5), 2043–2063.

Lee, S., R. Okui, and Y.-J. Whang (2017): “Doubly robust uniform confidence band for the conditional average treatment effect function,” *Journal of Applied Econometrics*, 32(7), 1207–1225.

Lee, S., and Y.-J. Whang (2009): “Nonparametric tests of conditional treatment effects,” *Working Paper*.

Linton, O., and P. Gozalo (2014): “Testing conditional independence restrictions,” *Econometric Reviews*, 33(5-6), 523–552.

Lu, X., and H. White (2014): “Testing for separability in structural equations,” *Journal of Econometrics*, 182(1), 14–26.

Matzkin, R. L. (2003): “Nonparametric estimation of nonadditive random functions,” *Econometrica*, 71(5), 1339–1375.

Mourifié, I., and Y. Wan (2017): "Testing local average treatment effect assumptions," *The Review of Economics and Statistics*, 99(2), 305–313.

Nolan, D., and D. Pollard (1988): "Functional limit theorems for U-processes," *The Annals of Probability*, 16(3), 1291–1298.

Pagan, A., and A. Ullah (1999): *Nonparametric Econometrics*. Cambridge University Press.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica*, 57(6), 1403–1430.

Rosenzweig, M. R., and K. I. Wolpin (1980): "Testing the quantity-quality fertility model: The use of twins as a natural experiment," *Econometrica*, 48(1), 227–240.

Rudin, W. (1962): *Fourier Analysis on Groups*. Interscience.

Stinchcombe, M. B., and H. White (1998): "Consistent specification testing with nuisance parameters present only under the alternative," *Econometric Theory*, 14(03), 295–325.

Su, L., Y. Tu, and A. Ullah (2015): "Testing additive separability of error term in nonparametric structural models," *Econometric Reviews*, 34(6-10), 1057–1088.

Su, L., and H. White (2007): "A consistent characteristic function-based test for conditional independence," *Journal of Econometrics*, 141(2), 807–834.

——— (2008): "A nonparametric Hellinger metric test for conditional independence," *Econometric Theory*, 24(04), 829–864.

——— (2014): "Testing conditional independence via empirical likelihood," *Journal of Econometrics*, 182(1), 27–44.

Torgovitsky, A. (2015): "Identification of nonseparable models using instruments with small support," *Econometrica*, 83(3), 1185–1197.

van der Vaart, A. W., and J. A. Wellner (1996): "Weak Convergence," in *Weak Convergence and Empirical Processes*, pp. 16–28. Springer.

van der Vaart, A. W., and J. A. W. Wellner (2007): *Empirical processes indexed by estimated functions*vol. Volume 55 of *Lecture Notes–Monograph Series*, pp. 234–252. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Vuong, Q., and H. Xu (2017): "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity," *Quantitative Economics*, 8(2), 589–610.

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70(1), 331–341.

Wager, S., and S. Athey (2018): "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.

Tables

Table 1: Rejection probabilities ($\alpha = 5\%$) in the discrete–covariates case.

p	n	$c = 0.7$	$c = 0.8$	$c = 0.9$	$c = 1.0$	$c = 1.1$	$c = 1.2$	$c = 1.3$
Panel A: rejection probabilities at null hypothesis with $\gamma = 1$								
0.25	1000	0.0025	0.0045	0.0080	0.0105	0.0140	0.0190	0.0250
	2000	0.0130	0.0160	0.0230	0.0275	0.0330	0.0345	0.0415
	4000	0.0265	0.0320	0.0415	0.0460	0.0490	0.0530	0.0575
0.5	1000	0.0090	0.0120	0.0160	0.0235	0.0300	0.0395	0.0460
	2000	0.0250	0.0300	0.0340	0.0410	0.0415	0.0445	0.0490
	4000	0.0350	0.0430	0.0500	0.0525	0.0565	0.0610	0.0625
0.75	1000	0.0040	0.0075	0.0135	0.0180	0.0270	0.0335	0.0390
	2000	0.0140	0.0210	0.0245	0.0285	0.0360	0.0415	0.0480
	4000	0.0230	0.0280	0.0340	0.0390	0.0455	0.0505	0.0570
Panel B: rejection probabilities at alternative hypothesis with $\gamma = 0.75$								
0.25	1000	0.0125	0.0205	0.0340	0.0490	0.0605	0.0745	0.0885
	2000	0.0810	0.1065	0.1370	0.1610	0.1805	0.1985	0.2120
	4000	0.2610	0.2930	0.3160	0.3385	0.3600	0.3780	0.3935
0.5	1000	0.0390	0.0585	0.0775	0.1005	0.1185	0.1340	0.1405
	2000	0.1590	0.1920	0.2205	0.2485	0.2675	0.2830	0.2970
	4000	0.4360	0.4705	0.4945	0.5240	0.5395	0.5510	0.5730
0.75	1000	0.0230	0.0395	0.0540	0.0700	0.0855	0.1010	0.1100
	2000	0.0970	0.1260	0.1525	0.1710	0.1880	0.2050	0.2175
	4000	0.3035	0.3300	0.3565	0.3775	0.3955	0.4120	0.4245
Panel C: rejection probabilities at alternative hypothesis with $\gamma = 0.50$								
0.25	1000	0.1975	0.2760	0.3515	0.4145	0.4490	0.4790	0.5030
	2000	0.7335	0.8010	0.8445	0.8705	0.8870	0.8985	0.9045
	4000	0.9985	0.9990	0.9990	0.9990	0.9990	0.9990	0.9990
0.5	1000	0.5215	0.5915	0.6445	0.6860	0.7065	0.7155	0.7255
	2000	0.9630	0.9715	0.9750	0.9780	0.9825	0.9820	0.9835
	4000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.75	1000	0.3600	0.4330	0.4815	0.5135	0.5295	0.5370	0.5370
	2000	0.8645	0.8915	0.9070	0.9180	0.9220	0.9260	0.9265
	4000	0.9990	0.9990	0.9990	0.9995	0.9990	0.9990	0.9990

Table 2: Rejection probabilities ($\alpha = 5\%$) in the continuous–covariates case.

p	n	$c = 0.7$	$c = 0.8$	$c = 0.9$	$c = 1.0$	$c = 1.1$	$c = 1.2$	$c = 1.3$
Panel A: rejection probabilities at null hypothesis with $\gamma = 1$								
0.25	1000	0.0695	0.0630	0.0595	0.0575	0.0525	0.0540	0.0565
	2000	0.0620	0.0560	0.0555	0.0590	0.0570	0.0580	0.0590
	4000	0.0690	0.0690	0.0630	0.0650	0.0620	0.0570	0.0565
0.5	1000	0.0510	0.0520	0.0520	0.0505	0.0515	0.0520	0.0525
	2000	0.0590	0.0605	0.0575	0.0600	0.0630	0.0635	0.0650
	4000	0.0670	0.0620	0.0630	0.0620	0.0630	0.0555	0.0585
0.75	1000	0.0495	0.0485	0.0485	0.0480	0.0480	0.0470	0.0490
	2000	0.0450	0.0450	0.0490	0.0480	0.0470	0.0485	0.0455
	4000	0.0540	0.0560	0.0540	0.0510	0.0520	0.0515	0.0535
Panel B: rejection probabilities at alternative hypothesis with $\gamma = 0.75$								
0.25	1000	0.0805	0.0760	0.0730	0.0675	0.0635	0.0585	0.0585
	2000	0.1820	0.1570	0.1405	0.1210	0.1065	0.0920	0.0890
	4000	0.5730	0.5110	0.4550	0.4010	0.3560	0.3035	0.2655
0.5	1000	0.0960	0.0935	0.0890	0.0775	0.0720	0.0705	0.0690
	2000	0.3020	0.2700	0.2285	0.2000	0.1695	0.1490	0.1340
	4000	0.8160	0.7630	0.7170	0.6520	0.5940	0.5285	0.4805
0.75	1000	0.0585	0.0605	0.0580	0.0575	0.0560	0.0540	0.0520
	2000	0.1535	0.1400	0.1230	0.1080	0.0910	0.0780	0.0690
	4000	0.5450	0.4840	0.4300	0.3730	0.3220	0.2770	0.2410
Panel C: rejection probabilities at alternative hypothesis with $\gamma = 0.50$								
0.25	1000	0.6950	0.6620	0.6295	0.5940	0.5470	0.5200	0.4765
	2000	0.9925	0.9895	0.9850	0.9805	0.9720	0.9630	0.9525
	4000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.5	1000	0.9205	0.8970	0.8700	0.8370	0.8015	0.7560	0.7160
	2000	1.0000	1.0000	1.0000	0.9990	0.9985	0.9975	0.9970
	4000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.75	1000	0.7150	0.6665	0.6155	0.5685	0.5135	0.4500	0.4135
	2000	0.9990	0.9970	0.9935	0.9845	0.9705	0.9565	0.9370
	4000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3: Descriptive Statistics for the National JTPA Study

	All	$Z=1$ (eligible)	$Z=0$ (not eligible)
Men			
Number of observations	$5,102$	$3,399$	$1,703$
Training ($D=1$)	41.87%	62.28%	1.12%
High school or GED	69.32%	69.26%	69.43%
Married	35.26%	36.01%	33.75%
Minorities	38.38%	38.69%	37.76%
Work less than 13 weeks in the past year	40.02%	40.28%	39.05%
30 months earnings	$19,147$	$19,520$	$18,404$
Women			
Number of observations	$6,102$	$4,088$	$2,014$
Training ($D=1$)	44.61%	65.73%	1.74%
High school or GED	72.06%	72.85%	70.45%
Married	21.93%	22.48%	20.82%
Minorities	40.41%	40.58%	51.86%
Work less than 13 weeks in the past year	51.79%	51.75%	51.86%
30 months earnings	$13,029$	$13,439$	$12,197$

Note: Means are reported in this table for the National JTPA study 30-month earnings sample.

Table 4: Descriptive Statistics for the 1999 and 2000 Censuses

	1990			2000		
	All	$Z=1$ (twin birth)	$Z=0$ (no twin birth)	All	$Z=1$ (twin birth)	$Z=0$ (no twin birth)
Observations	602,767	6,524	596,243	573,437	8,569	564,868
Number of children	1.9276	2.5318	1.9209	1.8833	2.5196	1.8734
At least two children ($D=1$)	0.6500	1.0000	0.6461	0.6163	1.0000	0.6104
Mother						
Age in years	29.7894	29.9530	29.7876	30.0562	30.3943	30.0510
Years of education	12.9196	12.9623	12.9191	13.1131	13.2615	13.1108
Black	0.0637	0.0757	0.0636	0.0724	0.0816	0.07228
Asian	0.0326	0.0321	0.0326	0.0447	0.0335	0.0448
Other races	0.0537	0.0592	0.0536	0.0912	0.0806	0.0914
Currently at work	0.5781	0.5444	0.5785	0.5629	0.5132	0.5637
Usual hours per work	24.5660	23.3537	24.5795	25.1400	23.0491	25.1723
Wage or salary income last year	8942	8593	8946	14200	13757	14206
Father						
Age in years	32.5358	32.7534	32.5333	32.9291	33.3102	32.9232
Years of education	13.0436	13.0748	13.0432	13.0331	13.1806	13.0308
Black	0.0671	0.0796	0.0670	0.0800	0.0945	0.0798
Asian	0.0291	0.0263	0.0292	0.0402	0.0318	0.0403
Other races	0.0488	0.0529	0.0488	0.0919	0.0802	0.0921
Currently at work	0.8973	0.8922	0.8974	0.8512	0.8584	0.8511
Usual hours per work	42.7636	42.7704	42.7635	43.8805	43.8789	43.8805
Wage or salary income last year	27020	28039	27010	38041	41584	37987
Parents						
Wages or salary income last year	35,963	36,632	35,956	52,241	55,342	52,193

Note: Data from the 1% and 5% PUMS in 1990 and 2000. Own calculations using the PUMS sample weights. The sample consists of married mothers between 21 and 35 years of age with at least one child.