UC San Diego
Recent Work

Title

Identification and Estimation of Unconditional Policy Effects of an Endogenous Binary Treatment: an Unconditional MTE Approach

Permalink

Authors

Martínez-Iriarte, Julian
Sun, Yixiao

Publication Date

2021-07-22

Identification and Estimation of Unconditional Policy Effects of an Endogenous Binary Treatment: an Unconditional MTE Approach[*]

Julian Martinez-Iriarte[†]
Department of Economics
UC Santa Cruz

Yixiao Sun[‡]
Department of Economics
UC San Diego

February 1, 2022

Abstract

This paper studies identification and estimation of unconditional policy effects when treatment status is binary and endogenous. We introduce a new class of unconditional marginal treatment effects (MTE) based on the influence function of the functional underlying the policy target. We show that an unconditional policy effect can be represented as a weighted average of the newly defined unconditional MTEs over the individuals who are indifferent about their treatment status. We provide conditions for point identification of the unconditional policy effects. When a quantile is the functional of interest, we characterize the asymptotic bias of the unconditional quantile regression (UQR) estimator that ignores the endogeneity of the treatment and elaborate on the channels that the endogeneity can render the UQR estimator inconsistent. We show that, even if the treatment status is exogenous, the UQR estimator can still be inconsistent when there are common covariates affecting both the treatment status and the outcome variable. To overcome the inconsistency of the UQR estimator, we introduce the UNconditional Instrumental Quantile Estimator (UNIQUE) and establish its consistency and asymptotic distribution. In the empirical application, we estimate the effect of changing college enrollment status, induced by higher tuition subsidy, on the quantiles of the wage distribution.

Keywords: marginal treatment effect, marginal policy-relevant treatment effect, selection models, instrumental variables, unconditional policy effect, unconditional quantile regressions.
JEL: C14, C31, C36.

1 Introduction

An unconditional policy effect is the effect of a change in a target covariate on the (unconditional) distribution of an outcome variable of interest.[1] When the target covariate has a continuous distribution, we may be interested in shifting its location and evaluating the effect of such a shift on the distribution of the outcome variable. For example, we may consider increasing the number of years of education for *every* worker in order to improve the median of the wage distribution. When the change of the covariate distribution is small, such an effect may be referred to as the marginal unconditional policy effect.

In this paper, we consider a binary target covariate that indicates the treatment status. In this case, a location shift is not possible, and the only way to change its distribution is to change the proportion of treated individuals. We analyze the impact of such a marginal change on a general functional of the distribution of the outcome. For example, when the functional of interest is the mean of the outcome, this corresponds to the marginal policy-relevant treatment effect (MPRTE) of Carneiro, Heckman and Vytlacil (2010, 2011). For the case of quantiles, we obtain an unconditional quantile effect (UQE). Previously, Firpo, Fortin and Lemieux (2009) proposed using an unconditional quantile regression (UQR) to estimate the UQE.[2] However, we show that their identification strategy can break down under endogeneity. An extensive analysis of the resulting asymptotic bias of the UQR estimator is provided.

The first contribution of this paper is to introduce a new class of unconditional marginal treatment effects (MTEs) and show that the corresponding unconditional policy effect can be represented as a weighted average of these unconditional MTEs. The new MTEs are based on the influence function of the functional of the outcome distribution that we care about. The weights depend on the characteristics of the subpopulation who is indifferent about the treatment take-up. For example, the UQE is a weighted average of the unconditional MTEs built on the influence function of quantiles. This shows that the MPRTE and UQE belong to the same family of parameters. To the best of our knowledge, this was not previously recognized in either the literature on MTEs or the literature on unconditional policy effects.

To illustrate the usefulness of this general approach, we provide an extensive analysis of the unconditional quantile effects. This is empirically important since the UQR estimator proposed by Firpo, Fortin and Lemieux (2009) is consistent for the UQE only if a certain distributional invariance assumption holds. Such assumption is unlikely to hold when the treatment status is endogenous. We note that treatment endogeneity is the rule rather than the exception in economic applications.

The second contribution of this paper is to provide a closed-form expression for the asymptotic

---

[1]There are several groups of variables in this framework. The target covariate is the variable a policy maker aims to change. We often refer to the target covariate as the treatment or treatment variable. The outcome variable is the variable that a policy maker ultimately cares about. A policy maker hopes to change the target covariate in order to achieve a desired effect on the distribution of the outcome variable. Sometimes, a policy maker can not change the treatment variable directly and has to change it via intervening some other covariates, which may be referred to as the policy covariates. There may also be covariates that will not be intervened.

[2]Mukhin (2019) generalizes Firpo, Fortin and Lemieux (2009) to allow for non-marginal changes in continuous covariates. Further work in the case of continuous covariates include Sasaki, Ura and Zhang (2020) who allow for a high-dimensional setting, Inoue, Li and Xu (2021) who tackle the two-sample problem, and Martinez-Iriarte, Montes-Rojas and Sun (2022) who analyze location-scale and compensated shifts.

bias of the UQR estimator when endogeneity is neglected. The asymptotic bias can be traced back to two sources. First, the subpopulation of the individuals at the margin of indifference might have different characteristics than the whole population. We refer to this source of the bias as the *marginal heterogeneity bias.* Second, the treatment effect for a marginal individual might be different from an apparent effect obtained by comparing the treatment group with the control group. It is the marginal subpopulation, not the whole population or any other subpopulation, that contributes to the UQE. We refer to the second source of the bias as the *marginal selection bias.*

For the case with a binary treatment, the UQR of Firpo, Fortin and Lemieux (2009) does not account for the endogeneity and the difference between the marginal subpopulation and the whole population in terms of their covariate distributions and treatment effects. As a result, it can be plagued with biases from both sources. We show that in some cases, even if the treatment status is exogenous, the presence of other covariates[3] can render the UQR estimator inconsistent. For example, this can occur in situations where the treatment status is partly determined by the covariates that also affect the outcome variable, that is, when the selection equation and the outcome equation have covariates in common. In such a case with no endogeneity, there is no *marginal selection bias,* but the UQR estimator can still suffer from a severe *marginal heterogeneity bias.* This intriguing result cautions the use of the UQR without careful consideration.

One may ask if, as in a linear model, the asymptotic bias of the UQR estimator can be signed based on a correlation coefficient. The answer to this question is negative. The reason is that the asymptotic bias can be non-uniform across quantiles. We could have a positive bias for the 10th quantile while a negative bias for the 11th quantile. Strong assumptions have to be imposed on the data generating process to sign the bias *a priori*.

The third contribution of this paper is to show that if assumptions similar to instrument validity are imposed on the policy variables under intervention, the resulting UQE can be point identified using the local instrumental variable approach as in Carneiro and Lee (2009). Based on this, we introduce the UNconditional Instrumental QUantile Estimator (UNIQUE) and develop methods of statistical inference based on the UNIQUE when the binary treatment is endogeneous. We take a nonparametric approach but allow the propensity score function to be either parametric or non-parametric. We establish the asymptotic distribution of the UNIQUE. This is a formidable task, as the UNIQUE is a four-step estimator, and we have to pin down the estimation error from each step.

Related Literature. This paper is related to the literature on marginal treatment effects. Introduced by Bjorklund and Moffitt (1987), MTEs can be used as a building block for many different causal parameters of interest as shown in Heckman and Vytlacil (2001*a*). For an excellent review, the reader is referred to Mogstad and Torgovitsky (2018). As a main departure from this literature, this paper introduces unconditional MTEs targeted at studying unconditional policy effects. As mentioned above, if we focus on the mean, the unconditional marginal policy effect we study corresponds to the MPRTE of Carneiro, Heckman and Vytlacil (2010). In this case, the unconditional MTE is the same as the usual MTE. For quantiles, the unconditional marginal policy effect is studied by Firpo, Fortin and Lemieux (2009), but under a setting that does not allow for endogeneity.[4]

---

[3]UQR does not explicitly account for the presence of additional covariates. See the last paragraph of Section 3 in Firpo, Fortin and Lemieux (2009) or Corollary 3 in the working paper version Firpo, Fortin and Lemieux (2007).

[4]For the case of continuous endogenous covariates, Rothe (2010) shows that the control function approach of Imbens

The unconditional MTE is novel in this case, as well as in the case with a more general nonlinear functional of interest. In our setting that allows some covariates to enter both the outcome equation and the selection equation, conditioning on the propensity score is not enough, but we show that the propensity score plays a key role in averaging the unconditional MTEs to obtain the unconditional policy effect. This is in contrast with Zhou and Xie (2019) where the marginal treatment effect parameter is defined based on conditioning on the propensity score. Among other contributions, Mogstad, Torgovitsky and Walters (2020) provide an extensive list of MTE applications which are, in principle, applicable to the unconditional MTEs introduced in this paper.

Our general treatment of the problem using functionals is closely related to that of Rothe (2012). Rothe (2012) analyzes the effect of an arbitrary change in the distribution of a target covariate, either continuous or discrete, on some feature of the distribution of the outcome variable. By assuming a form of conditional independence, for the case of continuous target covariates, Rothe (2012) generalizes the approach of Firpo, Fortin and Lemieux (2009). However, for the case of a discrete treatment, instead of point identifying the effect as we do here, bounds are obtained by assuming that either the highest-ranked or lowest-ranked individuals enter the program under the new policy.

We are not the first to consider unconditional quantile regressions under endogeneity. Kasy (2016) focuses on ranking counterfactual policies and, for the case of discrete regressors, allows for endogeneity. However, one key difference from our approach is that the counterfactual policies analyzed in Kasy (2016) are randomly assigned conditional on a covariate vector. In our setting, selection into treatment follows a threshold-crossing model, where we use the exogenous variation of an instrument to obtain different counterfactual scenarios. Martinez-Iriarte (2020) introduces the quantile breakdown frontier in order to perform a sensitivity analysis to departures from the distributional invariance assumption employed by Firpo, Fortin and Lemieux (2009).

Estimation of the MTE and parameters derived from the MTE curve is discussed in Heckman, Urzua and Vytlacil (2006), Carneiro and Lee (2009), Carneiro, Heckman and Vytlacil (2010, 2011), and, more recently, Sasaki and Ura (2021). All of these studies make a linear-in-parameters assumption regarding the conditional means of the potential outcomes, which yields a tractable partially-linear model for the MTE curve. This strategy is not very helpful in our case because our newly defined unconditional MTE might involve a nonlinear function of the potential outcomes. For example, for the case of quantiles, there is an indicator function involved. Using our expression for the weights, we can write the UQE as a quotient of two average derivatives. One of them, however, involves as a regressor the estimated propensity score as in the setting of Hahn and Ridder (2013). We provide conditions, different from those in Hahn and Ridder (2013), under which the error from estimating the propensity score function, either parametrically or nonparametrically, does not affect the asymptotic variance of the UNIQUE. This may be of independent interest.

Outline. Section 2 introduces the new MTE curve and shows how it relates to the unconditional policy effect. Section 3 presents a model for studying the UQE under endogeneity. Section 4 considers intervening an instrumental variable in order to change the treatment status and establishes the identification of the corresponding UQE. Section 5 introduces and studies the UNIQUE under a

---

and Newey (2009) can be used to achieve identification.

parametric specification of the propensity score. Section 6 provides simulation evidence. In Section 7 we revisit the empirical application of Carneiro, Heckman and Vytlacil (2011) and focus on the unconditional quantile effect. Section 8 concludes. The Appendix contains all proofs as well as some additional explanations or derivations of some results in the main text and considers the UNIQUE under a nonparametric specification of the propensity score.

Notation. For any generic random variable W_1, we denote its CDF and pdf by $F_{W_1}(\cdot)$ and $f_{W_1}(\cdot)$, respectively. We denote its conditional CDF and pdf conditional on a second random variable W_2 by $F_{W_1|W_2}(\cdot|\cdot)$ and $f_{W_1|W_2}(\cdot|\cdot)$, respectively.

2 Unconditional Policy Effects under Endogeneity

2.1 Policy Intervention and Unconditional Policy Effects

We employ the potential outcomes framework. For each individual, there are two potential outcomes: $Y(0)$ and $Y(1)$, where $Y(0)$ is the outcome had she received no treatment and $Y(1)$ is the outcome had she received treatment. We assume that the potential outcomes are given by

$$Y(0) = r_0(X, U_0), \text{ and } Y(1) = r_1(X, U_1),$$

for a pair of unknown functions r_0 and r_1. The vector $X \in \mathbb{R}^{d_X}$ consists of observables and (U_0, U_1) consists of unobservables. Depending on the individual's actual choice of treatment, denoted by D, we observe either $Y(0)$ or $Y(1)$, but we can never observe both. The observed outcome is denoted by Y:

$$Y = (1-D)\,Y(0) + D Y(1) = (1-D)\,r_0(X, U_0) + D r_1(X, U_1). \tag{1}$$

Following Heckman and Vytlacil (1999, 2001a, 2005), we assume that selection into treatment is determined by a threshold-crossing equation

$$D = \mathbb{1}\{V \le \mu(W)\}, \tag{2}$$

where $W = (Z, X)$ and $Z \in \mathbb{R}^{d_Z}$ consists of covariates that do not affect the potential outcomes directly. In the above, the unknown function $\mu(W)$ can be regarded as the benefit from the treatment and V as the cost of the treatment. Individuals decide to take up the treatment if and only if its benefit outweighs its cost. Alternatively, we can think of $\mu(W)$ as the utility and V as the disutility from participating in the program. While we observe (D, W, Y), we observe neither $U := (U_0, U_1)'$ nor V. Also, we do not restrict the dependence among U, W, and V. Hence, they can be mutually dependent and D could be endogenous.

To change the treatment take-up rate, we manipulate Z. More specifically, we assume that Z is a vector of continuous random variables and consider a policy intervention that changes Z into $Z_\delta = \mathcal{G}(W, \delta)$ for a vector of smooth functions $\mathcal{G}(\cdot, \cdot) \in \mathbb{R}^{d_Z}$. We assume that $\mathcal{G}(W, 0) = Z$ so that the *status quo* policy corresponds to $\delta = 0$. For example, we can take $\mathcal{G}(W, \delta) = Z + \delta$. In this case, we change the value of Z by δ for everyone in the population. When Z has two components, say, $Z = (Z_1, Z_2)'$, we may consider intervening only one component. For example, $\mathcal{G}(W, \delta) =$

$(Z_1 + s(\delta), Z_2)'$ for some function $s(\delta)$ with $s(0) = 0$. More general interventions, such as $\mathcal{G}(W, \delta) = Z(1 + \delta), Z + \delta Z^2, Z + \delta ZX$ when $d_X = d_Z = 1$, are allowed. Even though Z can be correlated with other variables, observed or not, we assume that Z does not contain causal factors of other variables so that the induced change in Z does not cause other variables to change. However, we do not rule out that other variables may cause Z or there is a common factor that causes Z and other variables, leading to a nonzero correlation between Z and other variables.[5]

When we induce the covariate Z to change, the distribution of this covariate will change. However, we do not specify the new distribution *a priori*. Instead, we specify the policy rule that pins down how the value of the covariate will be changed for each individual in the population. Our intervention may then be regarded as a *value* intervention. This is in contrast to a *distribution* intervention that stipulates a new covariate distribution directly. An advantage of our policy rule is that it is directly implementable in practice while a hypothetical distribution intervention is not. The latter intervention may still have to be implemented via a value intervention, which is our focus here.

With the induced change in Z, the selection equation becomes

$$D_\delta = \mathbf{1}\{V \le \mu(Z_\delta, X)\} = \mathbf{1}\{V \le \mu(\mathcal{G}(W, \delta), X)\}. \tag{3}$$

The outcome equation, in turn, is now

$$Y_\delta = (1 - D_\delta)Y(0) + D_\delta Y(1) = (1 - D_\delta)r_0(X, U_0) + D_\delta r_1(X, U_1). \tag{4}$$

These two equations are the same as the *status quo* equations; the only exception is that Z has been replaced by Z_δ. We have maintained the structural forms of the outcome equation and the treatment selection equation. Importantly, we have also maintained the stochastic dependence among $U, W,$ and V, which is is manifested through the use of the same notation $U, W,$ and V in equations (3) and (4) as in equations (1) and (2). Our policy intervention has a *ceteris paribus* interpretation at the population level*: we apply the same form of intervention on Z for all individuals in the population but hold all else, including the causal mechanism and the stochastic dependence among the *status quo* variables, constant.

Note that the model described by (3) and (4) coincides with the model given in (1) and (2) if we set $\delta = 0$. For notational convenience, when $\delta = 0$, we drop the subscript, and we write Y and D for Y_0 and D_0, respectively. It may be useful to reiterate that, regardless of the value of δ, the pattern of stochastic dependence among $U, W,$ and V is the same. In particular, the conditional distribution of (U, V) given W is invariant to the value of δ.

Under the *status quo* policy regime, the propensity score is $P(w) := \Pr(D = 1 | W = w)$. In view of (2), we can represent it as

$$P(w) = \Pr(V \le \mu(W) | W = w) = F_{V|W}(\mu(w)|w).$$

---

[5]The case where the induced change in Z may cause other variables to change requires a model for such a mechanism. Our framework can, in principle, accommodate this if such a model is available. We leave the details for future studies.

If the conditional CDF $F_{V|W}(\cdot|w)$ is a strictly increasing function for all $w \in \mathcal{W}$, the support of W, we have

$$D = \mathbb{1}\left\{V \le \mu\left(W\right)\right\} = \mathbb{1}\left\{F_{V|W}(V|W) \le F_{V|W}(\mu\left(W\right)|W)\right\} = \mathbb{1}\left\{U_D \le P(W)\right\},$$

where $U_D := F_{V|W}(V|W)$ measures an individual's relative resistance to the treatment. It can be shown that U_D is uniform on $[0,1]$ and is independent of W.

Under the counterfactual policy regime, $W = (Z, X)$ becomes $W_\delta := (Z_\delta, X)$, and we have

$$D_\delta = \mathbb{1}\left\{V \le \mu\left(W_\delta\right)\right\} = \mathbb{1}\left\{F_{V|W}(V|W) \le F_{V|W}(\mu\left(W_\delta\right)|W)\right\} = \mathbb{1}\left\{U_D \le P_\delta(W)\right\}, \qquad (5)$$

where
$$P_\delta(W) := F_{V|W}(\mu\left(W_\delta\right)|W) = F_{V|W}(\mu\left(\mathcal{G}(W,\delta), X\right)|W)$$

is a function of W and δ. Note that U_D in (5) is still defined as $F_{V|W}(V|W)$, and so it does not change under the counterfactual policy regime. This is to say that relative to others, an individual's resistance to the treatment is maintained across the two policy regimes. In particular, U_D is still uniform on [0,1] and independent of W. The propensity score under the new policy regime is then equal to

$$P_\delta(w) = \Pr(D_\delta = 1|W = w) = \Pr\left[U_D \le P_\delta(w)|W = w\right] = \Pr\left[U_D \le P_\delta(w)\right] = P_\delta\left(w\right).$$

To maintain the generality of the policy change, we will not specify the policy function $\mathcal{G}(\cdot, \delta)$ for now, but we assume that the policy will increase the participation rate (in expectation) by δ. That is, under the intervention $\mathcal{G}(\cdot, \delta)$, $P_\delta(\cdot)$ satisfies

$$E\left[P_\delta(W)\right] = E\left[P(W)\right] + \delta,$$

and for $\delta = 0$, $P_0(w) = P(w)$ for all $w \in \mathcal{W}$, the support of W.

Let $\mathcal{F}^*$ be the space of finite signed measures ν on $\mathcal{Y} \subseteq \mathbb{R}$ with distribution function $F_\nu(y) = \nu(-\infty, y]$ for $y \in \mathcal{Y}$. We endow $\mathcal{F}^*$ with the usual supremum norm: for two distribution functions F_{ν_1} and F_{ν_2} associated with the respective signed measures ν_1 and ν_2 on $\mathcal{Y}$, we define $\|F_{\nu_1} - F_{\nu_2}\|_\infty :=$ $\sup_{y \in \mathcal{Y}} |F_{\nu_1}(y) - F_{\nu_2}(y)|$. With some abuse of notation, denote F_Y as the distribution of Y: $F_Y(y) = \nu_Y(-\infty, y]$ where ν_Y is the measure induced by the distribution of Y. Define F_{Y_δ} similarly. Clearly, both F_Y and F_{Y_δ} belong to $\mathcal{F}^*$. We consider a general functional $\rho : \mathcal{F}^* \to \mathbb{R}$ and study the general unconditional policy effect.

Definition 1. *General Unconditional Policy Effect*

The general unconditional policy effect for the functional ρ is defined as

$$\Pi_\rho := \lim_{\delta \to 0} \frac{\rho(F_{Y_\delta}) - \rho(F_Y)}{\delta},$$

whenever this limit exists.

The definition above is the same as that of the marginal partial distributional policy effect defined

in Rothe (2012). Examples of ρ are the mean as in the marginal policy-relevant treatment effect of Carneiro, Heckman and Vytlacil (2010, 2011), and the quantiles as in the unconditional quantile effects of Firpo, Fortin and Lemieux (2009). In order to ensure that Π_ρ exists, we will assume a certain smoothness of ρ and the closeness of F_{Y_δ} to F_Y.

We first consider a Hadamard differentiable functional ρ. For completeness, we provide the definition of Hadamard differentiability below.

Definition 2. $\rho : \mathcal{F}^* \to \mathbb{R}$ *is Hadamard differentiable at $F \in \mathcal{F}^*$ if there exists a linear and continuous functional $\dot{\rho}_F : \mathcal{F}^* \to \mathbb{R}$ such that for any $G \in \mathcal{F}^*$ and $G_\delta \in \mathcal{F}^*$ with*

$$\lim_{\delta \to 0} \|G_\delta - G\|_\infty = 0,$$

we have

$$\lim_{\delta \to 0} \frac{\rho(F + \delta G_\delta) - \rho(F)}{\delta} = \dot{\rho}_F(G).$$

To see how we can use the Hadamard differentiability to obtain Π_ρ, we write

$$\frac{\rho(F_{Y_\delta}) - \rho(F_Y)}{\delta} = \frac{\rho(F_Y + \delta G_\delta) - \rho(F_Y)}{\delta},$$

for

$$G_\delta = \frac{F_{Y_\delta} - F_Y}{\delta}. \tag{6}$$

As long as we can show that $\lim_{\delta \to 0} \|G_\delta - G\|_\infty = 0$ for some G, then, we obtain $\Pi_\rho = \dot{\rho}_{F_Y}(G)$.

Next, we provide sufficient conditions for $\lim_{\delta \to 0} \|G_\delta - G\|_\infty = 0$. We first make a support assumption.

Assumption 1. *Support Assumption For $d = 0, 1$, the support $\mathcal{Y}(d)$ of $Y(d)$ given (U_D, W) does not depend on (U_D, W).*

For some $\varepsilon > 0$, define $N_\varepsilon := \{\delta : |\delta| \leq \varepsilon\}$. For every $\delta \in N_\varepsilon$ and $y \in \mathcal{Y}$, we have

$$
\begin{aligned}
F_{Y_\delta}(y) &= (p + \delta) \Pr(Y(1) \leq y | D_\delta = 1) + (1 - p - \delta) \Pr(Y(0) \leq y | D_\delta = 0) \\
&= \int_{\mathcal{Y}(1)} \mathbf{1}\{\tilde{y} \leq y\} (p + \delta) f_{Y(1)|D_\delta}(\tilde{y}|1) d\tilde{y} \\
&\quad + \int_{\mathcal{Y}(0)} \mathbf{1}\{\tilde{y} \leq y\} (1 - p - \delta) f_{Y(0)|D_\delta}(\tilde{y}|0) d\tilde{y},
\end{aligned}
\tag{7}
$$

where we have used the support assumption so that the support of $Y(d)$ given $D_\delta = d$ is still $\mathcal{Y}(d)$.

The goal is to linearize $(p + \delta) f_{Y(1)|D_\delta}(\tilde{y}|1)$ and $(1 - p - \delta) f_{Y(0)|D_\delta}(\tilde{y}|0)$ around $\delta = 0$. To this end, we make some technical assumptions.

Assumption 2. *Regularity Conditions*

(a) *For $d = 0, 1$, the random variables $(Y(d), U_D, W)$ are absolutely continuous with joint density given by $f_{Y(d), U_D|W} f_W$.*

(b) *(i) For $d = 0, 1$, $u \mapsto f_{Y(d)|U_D,W}(y|u,w)$ is continuous for all $y \in \mathcal{Y}(d)$ and all $w \in \mathcal{W}$. (ii) For $d = 0, 1$, $\sup_{y \in \mathcal{Y}(d)} \sup_{w \in \mathcal{W}} \sup_{\delta \in N_\varepsilon} f_{Y(d)|U_D,W}(y|P_\delta(w), w) < \infty$.*

(c) *(i) $p = E[P(W)] \in (0,1)$. (ii) For all $w \in \mathcal{W}$, the map $\delta \mapsto P_\delta(w)$ is continuously differentiable on N_ε.*
(iii) $\sup_{w \in \mathcal{W}} \sup_{\delta \in N_\varepsilon} \left| \frac{\partial P_\delta(w)}{\partial \delta} \right| < \infty$.

Assumption 3. ***Domination Conditions*** *For $d = 0, 1$,*

$$\int_{\mathcal{Y}(d)} \sup_{\delta \in N_\varepsilon} f_{Y(d)|D_\delta}(y|d) dy < \infty \text{ and } \int_{\mathcal{Y}(d)} \sup_{\delta \in N_\varepsilon} \left| \frac{\partial f_{Y(d)|D_\delta}(y|d)}{\partial \delta} \right| dy < \infty.$$

In Assumption 2 "for all $w \in \mathcal{W}$" can be replaced by "for almost all $w \in \mathcal{W}$", and the supremum over $w \in \mathcal{W}$ can be replaced by the essential supremum over $w \in \mathcal{W}$.

Lemma 1. *Let Assumptions 1 and 2 hold. The map $\delta \mapsto (p + \delta) f_{Y(1)|D_\delta}(\tilde{y}|1)$ is continuously differentiable on N_ε for all $\tilde{y} \in \mathcal{Y}(1)$ and the map $\delta \mapsto (1 - p - \delta) f_{Y(0)|D_\delta}(\tilde{y}|0)$ is continuously differentiable on N_ε for all $\tilde{y} \in \mathcal{Y}(0)$.*

The results of Lemma 1 can be used to approximate $F_{Y_\delta}(y)$ in (7) by taking an expansion around $\delta = 0$.

Lemma 2. *Let Assumptions 1–3 hold. Then*

$$\begin{aligned} F_{Y_\delta}(y) &= F_Y(y) + \delta E\left[\left\{ F_{Y(1)|U_D,W}(y|P(W),W) - F_{Y(0)|U_D,W}(y|P(W),W) \right\} \dot{P}(W) \right] \\ &+ R_F(\delta; y), \end{aligned}$$

where $R_F(\delta; y) \to 0$ uniformly over $y \in \mathcal{Y} := \mathcal{Y}(0) \cup \mathcal{Y}(1)$ as $\delta \to 0$, and

$$\dot{P}(w) := \left. \frac{\partial P_\delta(w)}{\partial \delta} \right|_{\delta=0}.$$

Remark 1. *Lemma 2 provides a linear approximation to F_{Y_δ}, the CDF of the outcome variable under D_δ. Essentially, it says that the proportion of individuals with outcome below y under the new policy regime, that is, $F_{Y_\delta}(y)$, will be equal to the proportion of individuals with outcome below y under the existing policy regime, that is, $F_Y(y)$, plus an adjustment given by the marginal entrants. Consider $\delta > 0$ and $P_\delta(w) > P(w)$ for all $w \in \mathcal{W}$ as an example. In this case, because of the policy intervention, the individuals who are on the margin, namely those with $u_D = P(w)$, will switch their treatment status from 0 to 1. Such a switch contributes to $F_{Y_\delta}(y)$ by the amount $F_{Y(1)|U_D,W}(P(w),w) - F_{Y(0)|U_D,W}(P(w),w)$, averaged over the distribution of W for a certain subpopulation. We will show later that the subpopulation is exactly the group of individuals who are on the margin under the existing policy regime.*

Remark 2. *The linear approximation to F_{Y_δ} is uniform over $\mathcal{Y}$ as $\delta \to 0$. We need the uniform approximation because we consider a general Hadamard differentiable ρ. In the special case when ρ does not depend on the whole distribution, the uniformity of the approximation over the whole support $\mathcal{Y}$ may not be necessary. In the quantile case, we maintain the uniformity because we consider all quantile levels in $(0,1)$.*

Lemma 2 ensures that if we take

$$G(y) = E\left[\left\{F_{Y(1)|U_D,W}(y|P(W),W) - F_{Y(0)|U_D,W}(y|P(W),W)\right\}\dot{P}(W)\right], \tag{8}$$

then we have $\lim_{\delta\to 0}\|G_\delta - G\|_\infty = 0$ for G_δ defined in (6). Hence, under the Hadamard differentiability of ρ, we obtain

$$\Pi_\rho = \dot{\rho}_{F_Y}(G) = \int_{\mathcal{Y}}\psi(y,\rho,F_Y)dG(y), \tag{9}$$

where $\psi(y,\rho,F_Y)$ is the influence function of ρ at F_Y. It is defined as

$$\psi(y,\rho,F_Y) := \lim_{\epsilon\to 0+}\frac{\rho\left((1-\epsilon)F_Y + \epsilon\Delta_y\right) - \rho(F_Y)}{\epsilon},$$

where Δ_y is the probability measure that assigns mass 1 to the single point $\{y\}$. Plugging (8) into (9) yields the following theorem.

Theorem 1. *Let Assumptions 1–3 hold. Assume further that $\rho : \mathcal{F}^* \to \mathbb{R}$ is Hadamard differentiable. Then*

$$\Pi_\rho = \int_{\mathcal{Y}}\psi(y,\rho,F_Y)E\left[\left\{f_{Y(1)|U_D,W}(y|P(W),W) - f_{Y(0)|U_D,W}(y|P(W),W)\right\}\dot{P}(W)\right]dy.$$

Definition 3. *The unconditional marginal treatment effect for the ρ functional is*

$$\text{MTE}_\rho(u,w) := E\left[\psi(Y(1),\rho,F_Y) - \psi(Y(0),\rho,F_Y)|U_D = u, W = w\right]. \tag{10}$$

With the above definition, we can write Π_ρ as

$$\Pi_\rho = \int_{\mathcal{W}}\text{MTE}_\rho(P(w),w)\dot{P}(w)f_W(w)dw.$$

The general unconditional policy effect Π_ρ can be represented as a weighted average of $\text{MTE}_\rho(u,w)$ over the marginal subpopulation: for the individuals with $W = w$, only those for whom $u_D = P(w)$ will contribute to the unconditional effect. Among the group defined by $W = w$, there is a subgroup of individuals who are indifferent between participating and not participating: those for whom $u_D = P(w)$, that is, those for whom v satisfies $F_{V|W}(v|w) = P(w)$. A small incentive will induce a change in the treatment status for only this subgroup of individuals. It is the change in their treatment status, and hence the change in the composition of $Y(1)$ and $Y(0)$ in the observed outcome Y, that changes its unconditional characteristics such as the quantiles.

2.2 The Role of $\dot{P}(w)$

Theorem 1 shows that the unconditional effect depends also on $\dot{P}(w)$. Under Assumption 2(c), we have

$$\int_{\mathcal{W}}\frac{\partial P_\delta(w)}{\partial\delta}f_W(w)\,dw = \frac{\partial}{\partial\delta}\int_{\mathcal{W}}P_\delta(w)f_W(w)\,dw = \frac{\partial}{\partial\delta}(p+\delta) = 1$$

for any $\delta \in N_\varepsilon$, hence $\int_{\mathcal{W}}\dot{P}(w)f_W(w)\,dw = 1$. Thus the integrals in Π_ρ can be regarded as a weighted mean with the weight given by $\dot{P}(w)$.

It is important to note that $\dot{P}(w)$ depends on the form of the policy function $\mathcal{G}(\cdot, \delta)$, which determines how we change the propensity score and who the marginal entrants are. Different policy interventions result in different changes in the propensity score, which then lead to different sets of marginal entrants and different unconditional effects.

For intuition on this, consider the case where $\delta > 0$ and $P_\delta(w) \geq P(w)$ for all $w \in \mathcal{W}$. Then we have

$$
\begin{aligned}
\dot{P}(w) &= \lim_{\delta \to 0} \frac{\Pr(U_D \leq P_\delta(w)) - \Pr(U_D \leq P(w))}{\delta} \\
&= \lim_{\delta \to 0} \frac{\Pr(P(W) < U_D \leq P_\delta(W) \mid W = w)}{\delta}.
\end{aligned}
$$

Thus, $\dot{P}(w)$ measures the relative contribution to the overall improvement in the participation rate (i.e., δ) for the individuals with $W = w$. For each value of W, only individuals on the margin ("the marginal individuals") will change their treatment status and contribute to the overall improvement in the participation rate. The relative "thickness" of the margin depends on w and is measured by $\dot{P}(w)$.

We can use Figure 1 to convey the intuition behind $\dot{P}(w)$. The figure illustrates the marginal individuals under the existing and new policy regimes. Under the existing policy regime, the marginal individuals lie on the 45-degree line in the $(P(w), u_D)$-plane. For easy reference, we call it the marginal curve, which is the set of points $\{(P(w), u_D) : u_D = P(w)\}$. Under the new policy regime, the marginal curve is now $\{(P(w), u_D) : u_D = P_\delta(w)\}$. Note that we can rewrite $u_D = P_\delta(w)$ as $u_D = P(w) + [P_\delta(w) - P(w)]$. Thus the new marginal curve can be obtained by shifting every point on the original marginal curve up by $P_\delta(w) - P(w)$. The magnitude of the upward shift is approximately $\dot{P}(w)\delta$, which is, in general, different for different values of w. The integral of the difference of the two marginal curves (i.e., the area of the gray region) weighted by the marginal density $f_W(\cdot)$ of W is equal to δ.[6]

To understand the weight $f_W(w)\dot{P}(w)$ that appears in Theorem 1, let ϵ be a small positive number. Then, $f_W(w)\epsilon$ measures the proportion of individuals for whom W is in $[w - \epsilon/2, w + \epsilon/2]$. Note that for $W \in [w - \epsilon/2, w + \epsilon/2]$, the propensity scores under D and D_δ are approximately $P(w)$ and $P_\delta(w)$. The proportion of the individuals for whom $W \in [w - \epsilon/2, w + \epsilon/2]$ and who have switched their treatment status from 0 to 1 is then equal to

$$
f_W(w) \cdot [P_\delta(w) - P(w)] \cdot \epsilon.
$$

Scaling this by δ, which is the overall proportion of the individuals who have switched the treatment status, we obtain $f_W(w) \cdot [P_\delta(w) - P(w)] / \delta \cdot \epsilon$. Thus, we can regard $f_W(w) \cdot [P_\delta(w) - P(w)] / \delta$ as the density function of W among those who have switched the treatment status from 0 to 1 as a result of the policy intervention.

---

[6]Note that the two marginal curves coincide in the limit as $\delta \to 0$, and so in the limit we can define the marginal individuals as those with $U_D = P(W)$. In our discussions, "the marginal individuals" may refer to the group of individuals with $U_D = P(W)$ or the group of individuals with $U_D = P_\delta(W)$. Which group we refer to should be clear from the context.
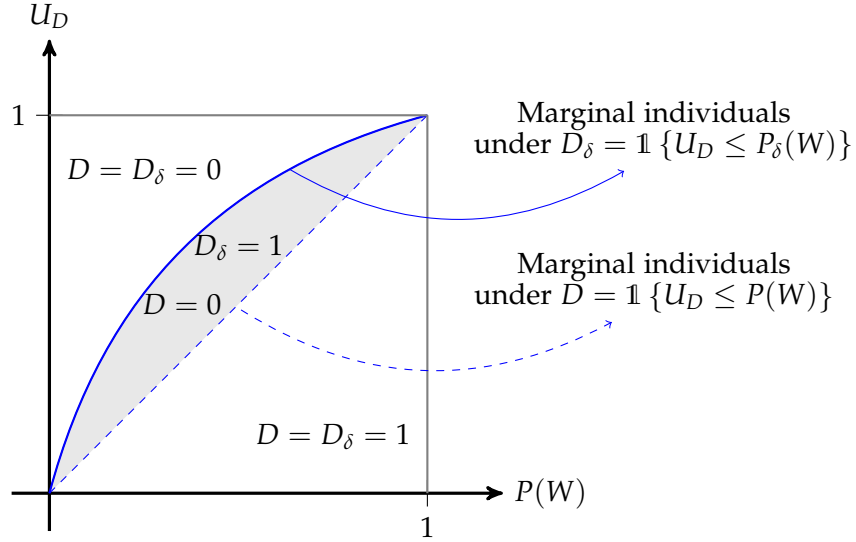
Figure 1: *Marginal individuals under different policies.*

Mathematically, we have

$$\frac{\Pr\left(W \in [w - \epsilon/2, w + \epsilon/2] \,|\, D = 0, D_\delta = 1\right)}{\epsilon}$$

$$= \frac{\Pr\left(W \in [w - \epsilon/2, w + \epsilon/2], D = 0, D_\delta = 1\right)}{\epsilon \cdot \Pr\left(D = 0, D_\delta = 1\right)}$$

$$= \frac{\Pr\left(W \in [w - \epsilon/2, w + \epsilon/2], U_D \in [P(W), P_\delta(W)]\right)}{\epsilon \cdot \delta}.$$

Taking the limit as $\epsilon \to 0$, we obtain

$$\lim_{\epsilon \to 0} \frac{\Pr\left(W \in [w - \epsilon/2, w + \epsilon/2] \,|\, D = 0, D_\delta = 1\right)}{\epsilon} = f_W(w) \frac{P_\delta(w) - P(w)}{\delta}.$$

Thus $f_W(w) \left[P_\delta(w) - P(w)\right] / \delta$ is the density of W among those who respond positively to the policy intervention, that is, those with $D = 0$ and $D_\delta = 1$. Graphically, $f_W(w) \left[P_\delta(w) - P(w)\right] / \delta$ is the conditional density of W conditional on $(P(W), U_D)$ being in the gray region in Figure 1.

Letting $\delta \to 0$, we obtain

$$\lim_{\delta \to 0} f_W(w) \frac{P_\delta(w) - P(w)}{\delta} = f_W(w) \dot{P}(w).$$

That is, $f_W(w) \dot{P}(w)$ is the limit of the density of W among those with $D = 0, D_\delta = 1$. We can therefore refer to $f_W(w) \dot{P}(w)$ as the density of the distribution of W over the marginal subpopulation that consists of all marginal individuals.

In view of the above interpretation of $f_W(w) \dot{P}(w)$, Theorem 1 shows that the unconditional effect is equal to the change in the influence functions for the marginal individuals, weighted by the density of the distribution of W over those marginal individuals.

Noting that $f_W(w)$ is the density of the distribution of W over the entire population, we can regard $\dot{P}(w)$ as the Radon-Nikodym (henceforth RN) derivative of the subpopulation distribution

11

with respect to the population distribution. Even if $\dot{P}(w)$ is not positive for all $w \in \mathcal{W}$, the RN interpretation is still valid. In this case, the distribution with density $f_W(w)\dot{P}(w)$ with respect to the Lebesgue measure is a signed measure.

2.3 $\dot{P}(w)$ and MPRTE

While Theorem 1 covers general functionals, it does not cover the mean functional $\rho(F_Y) = \int_{\mathcal{Y}} y dF_Y(y)$ unless $\mathcal{Y}$ is a bounded set. When $\mathcal{Y}$ is unbounded, the mean functional is not continuous on $(\mathcal{F}^*, \|\cdot\|_\infty)$ and hence is not Hadamard differentiable (see, for example, Exercise 7 in Chapter 20 in van der Vaart (2000)). In such a case, we opt for a direct approach by first showing that the remainder from Lemma 2 satisfies

$$\lim_{\delta \to 0} \frac{1}{\delta} \int_{\mathcal{Y}} y dR_F(\delta; y) = 0, \tag{11}$$

and then exploiting the linearity of the mean functional to obtain

$$\Pi_\rho = \int_{\mathcal{Y}} y E\left[\left\{f_{Y(1)|U_D,W}(y|P(W),W) - f_{Y(0)|U_D,W}(y|P(W),W)\right\}\dot{P}(W)\right] dy.$$

The result in (11) holds if the following stronger version of Assumption 3 holds:

Assumption 4. *Stronger Domination Conditions* For $d = 0, 1$,

$$\int_{\mathcal{Y}(d)} \sup_{\delta \in N_\varepsilon} y \cdot f_{Y(d)|D_\delta}(y|d) dy < \infty, \text{ and } \int_{\mathcal{Y}(d)} \sup_{\delta \in N_\varepsilon} \left| y \cdot \frac{\partial f_{Y(d)|D_\delta}(y|d)}{\partial \delta} \right| dy < \infty.$$

Corollary 1. *Let Assumptions 1, 2, and 4 hold. Then, for the mean functional, we have*

$$\Pi_\rho = E\left\{E\left[Y(1) - Y(0)|U_D = P(W), W\right]\dot{P}(W)\right\}. \tag{12}$$

Heckman and Vytlacil (2001b, 2005) consider the *policy-relevant treatment effect* defined as

$$\text{PRTE}_\delta = \frac{E(Y_\delta) - E(Y)}{E(D_\delta) - E(D)} = \frac{E(Y_\delta) - E(Y)}{\delta}. \tag{13}$$

Taking the limit $\delta \to 0$ yields the *marginal policy-relevant treatment effect* (MPRTE) of Carneiro, Heckman and Vytlacil (2010): MPRTE $= \lim_{\delta \to 0} \text{PRTE}_\delta$. Carneiro, Heckman and Vytlacil (2010, 2011) show that MPRTE can be represented in terms of the marginal treatment effect curve. Indeed, if we drop X for simplicity and assume that Z is independent of U conditional on U_D, the usual MTE curve is $\text{MTE}(u) := E[Y(1) - Y(0)|U_D = u]$. Then

$$\text{MPRTE} = -\int_0^1 \text{MTE}(u) \frac{\partial F_{\mathcal{P}_\delta}(u)}{\partial \delta}\bigg|_{\delta=0} du, \tag{14}$$

where $F_{\mathcal{P}_\delta}$ is the CDF of the random variable $\mathcal{P}_\delta := P_\delta(Z)$. In the Appendix we show that equations (12) and (14) are equivalent. The expression in (12) has the advantage of explicitly depending on

$\dot{P}(w)$ which, as we mentioned, has the interpretation as the density of the distribution of W over the marginal individuals.

This shows that our results include the existing results on MPRTE as special cases. We note that unlike the mean functional, which is linear, the quantile functional is not linear. In the latter case and for more general nonlinear functionals, we can not use the law of iterated expectations as in the linear case to represent an unconditional effect as the expectation of the corresponding conditional effect. Because of this fundamental difference, our results are not straightforward extensions of existing results in the literature.

2.4 Understanding the Unconditional MTE

To further understand the unconditional MTE, we provide another perspective here. This subsection is for heuristics only; rigorous developments have already been given in the previous subsections.

If $\dot{\rho}_F$ is the continuous derivative of ρ at F_Y, we can expect

$$\lim_{\delta \to 0} \frac{\rho(F_{Y_\delta}) - \rho(F_Y)}{\delta} = \dot{\rho}_{F_Y} \left(\lim_{\delta \to 0} \frac{F_{Y_\delta} - F_Y}{\delta} \right).$$

One way to represent $(F_{Y_\delta} - F_Y)/\delta$ is to replace $E(Y_\delta)$ and $E(Y)$ in (13) by $E[\mathbb{1}\{Y_\delta \leq y\}]$ and $E[\mathbb{1}\{Y \leq y\}]$, respectively. So, for a given $y \in \mathcal{Y}$,

$$\frac{F_{Y_\delta}(y) - F_Y(y)}{\delta} = \frac{E[\mathbb{1}\{Y_\delta \leq y\} - E[\mathbb{1}\{Y \leq y\}]}{\delta}. \tag{15}$$

In the spirit of Heckman and Vytlacil (2001b, 2005), we may then regard $(F_{Y_\delta}(y) - F_Y(y))/\delta$ as a policy-relevant treatment effect: it is the effect of the policy on the percentage of individuals whose value of Y is less than y. The effect is tied to a particular value y, and we obtain a continuum of policy-relevant effects indexed by $y \in \mathcal{Y}$ if y is allowed to vary over $\mathcal{Y}$. The limit

$$\lim_{\delta \to 0} \frac{F_{Y_\delta}(y) - F_Y(y)}{\delta}$$

can then be regarded as a continuum of marginal policy-relevant treatment effects indexed by $y \in \mathcal{Y}$. By the results of Carneiro and Lee (2009), for each $y \in \mathcal{Y}$, the marginal policy-relevant treatment effect can be represented as a weighted average of the following *policy-relevant "distributional"* MTE:

$$
\begin{aligned}
\mathrm{MTE}_d(u, w; y) &= E\left[\mathbb{1}\{Y(1) \leq y\} - \mathbb{1}\{Y(0) \leq y\} | U_D = u, W = w\right] \\
&= F_{Y(1)|U_D, W}(y|u, w) - F_{Y(0)|U_D, W}(y|u, w).
\end{aligned}
$$

The composition $\dot{\rho}_F \circ \mathrm{MTE}_d$, defined as $\int_{\mathcal{Y}} \psi(y, \rho, F_Y) \mathrm{MTE}_d(u, w; dy)$, is then

$$
\begin{aligned}
(\dot{\rho}_F \circ \mathrm{MTE}_d)(u, w) &= \int_{\mathcal{Y}} \psi(y, \rho, F_Y) \left[F_{Y(1)|U_D, W}(dy|u, w) - F_{Y(0)|U_D, W}(dy|u, w) \right] \\
&= \int_{\mathcal{Y}} \psi(y, \rho, F_Y) \left[f_{Y(1)|U_D, W}(y|u, w) - f_{Y(0)|U_D, W}(y|u, w) \right] dy \\
&= E\left[\psi(Y(1), \rho, F_Y) - \psi(Y(0), \rho, F_Y) | U_D = u, W = w\right].
\end{aligned}
$$

This shows that $\dot{\rho}_F \circ \mathrm{MTE}_d$ is exactly the unconditional MTE defined in (10). The unconditional MTE is therefore a composition of the underlying influence function with the *policy-relevant distributional MTE*.

3 Unconditional Quantile Regressions under Endogeneity

In the rest of the paper, we focus on the quantile functional: $\rho_\tau[F_Y] = F_Y^{-1}(\tau)$, and in this section, we apply our results to study the unconditional quantile regression of Firpo, Fortin and Lemieux (2009) in great detail. We provide an in-depth analysis of the sources of the asymptotic bias of the UQR estimator.

3.1 UQR with a Binary Regressor

First, we review the method of UQR proposed by Firpo, Fortin and Lemieux (2009), and show how the identification strategy breaks down when D is endogenous. Let y_τ be the τ-quantile of Y, and let $y_{\tau,\delta}$ be the τ-quantile of Y_δ. That is, That is, $\Pr(Y \le y_\tau) = \Pr(Y_\delta \le y_{\tau,\delta}) = \tau$. We are interested in the behavior of $y_{\tau,\delta}$ as $\delta \to 0$.

Definition 4. *Unconditional Quantile Effect*

The unconditional quantile effect (UQE) is defined as

$$\Pi_\tau := \lim_{\delta \to 0} \frac{y_{\tau,\delta} - y_\tau}{\delta}$$

whenever this limit exists.

By definition, we have

$$\Pr\left(Y_\delta \le y\right) = \Pr\left(D_\delta = 1\right) \Pr\left(Y_\delta \le y | D_\delta = 1\right) + \Pr\left(D_\delta = 0\right) \Pr\left(Y_\delta \le y | D_\delta = 0\right).$$

When W is not present, Corollary 3 in the working paper Firpo, Fortin and Lemieux (2007) makes the following assumption to achieve identification:

$$\Pr\left(Y_\delta \le y | D_\delta = d\right) = \Pr\left(Y \le y | D = d\right) \text{ for } d = 0 \text{ and } 1. \tag{16}$$

We refer to this as *distributional invariance*, and it readily identifies the counterfactual distribution:

$$\Pr\left(Y_\delta \le y\right) = \Pr\left(D_\delta = 1\right) \Pr\left(Y \le y | D = 1\right) + \Pr\left(D_\delta = 0\right) \Pr\left(Y \le y | D = 0\right).$$

Under some mild conditions, Firpo, Fortin and Lemieux (2007) obtain

$$\Pi_\tau = \frac{1}{f_Y(y_\tau)} \left[\Pr(Y > y_\tau | D = 1) - \Pr(Y > y_\tau | D = 0)\right]. \tag{17}$$

The distributional invariance assumption given in (16) is key for obtaining the above identification result. It requires that the conditional distribution of the outcome variable conditioning

14

on the treatment status remains the same across two policy regimes. If treatments are randomly assigned under both policy regimes (e.g., $D_\delta = \mathbb{1}\{U_D \leq P_\delta(W)\}$ and (U_D, W) is independent of (U_0, U_1)), then (U_0, U_1) is clearly independent of D_δ. In this case, both $\Pr(Y_\delta \leq y | D_\delta = d)$ and $\Pr(Y \leq y | D = d)$ are equal to $\Pr(Y(d) \leq y)$ and the distributional invariance assumption is satisfied. When we allow for D_δ to be correlated with U, however, distributional invariance does not hold in general. For example, when $d = 1$,

$$\Pr(Y_\delta \leq y | D_\delta = 1) = \Pr(Y(1) \leq y | D_\delta = 1) = \Pr(r_1(X, U_1) \leq y | U_D \leq P_\delta(W)),$$

and $\Pr(Y \leq y | D = 1) = \Pr(r_1(X, U_1) \leq y | U_D \leq P(W))$. These two conditional probabilities are different under the general dependence of (W, U, U_D).

3.2 Asymptotic Bias of the UQR Estimator

In order to obtain an expression for the asymptotic bias in UQR when distributional invariance does not hold, we work with the same potential outcomes and threshold crossing model as in equations (1)–(3). The following corollary follows directly from Theorem 1 by choosing ρ to be the quantile functional $\rho_\tau[F_Y] = F_Y^{-1}(\tau)$ with the influence function $\psi(y, \rho_\tau, F_Y) = 1\{\tau - \mathbb{1}\{y \leq \rho_\tau[F_Y]\}\}/f_Y(y_\tau)$.

Corollary 2. *Let Assumptions 1–3 hold. Assume further that $f_Y(y_\tau) > 0$. Then*

$$
\begin{aligned}
\Pi_\tau &= \lim_{\delta \to 0} \frac{y_{\tau,\delta} - y_\tau}{\delta} \\
&= \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\{Y(0) \leq y_\tau\} | U_D = P(w), W = w\right] \dot{P}(w) f_W(w) \, dw \\
&\quad - \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\{Y(1) \leq y_\tau\} | U_D = P(w), W = w\right] \dot{P}(w) f_W(w) \, dw.
\end{aligned}
\tag{18}
$$

The next corollary decomposes the unconditional quantile effect into an apparent component that neglects the adjustment given by $\dot{P}(w)$ and a bias component.[7]

Corollary 3. *Let the assumptions in Corollary 2 hold. Then*

$$\Pi_\tau = A_\tau - B_\tau,$$

where

$$
\begin{aligned}
A_\tau &= \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\{Y \leq y_\tau\} | D = 0, W = w\right] f_W(w) \, dw \\
&\quad - \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\{Y \leq y_\tau\} | D = 1, W = w\right] f_W(w) \, dw,
\end{aligned}
\tag{19}
$$

---

[7]We use "A" to denote the **A**pparent component and use "B" to denote the **B**ias component.

15

and $B_\tau = B_{1\tau} + B_{2\tau}$, for

$$B_{1\tau} = \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y|D,W}(y_\tau|1,w) - F_{Y|D,W}(y_\tau|0,w) \right] \dot{P}(w) f_W(w) \, dw$$
$$- \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y|D,W}(y_\tau|1,w) - F_{Y|D,W}(y_\tau|0,w) \right] f_W(w) \, dw$$

and

$$B_{2\tau} = \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y(0)|D,W}(y_\tau|0,w) - F_{Y(0)|U_D,W}(y_\tau|P(w),w) \right] \dot{P}(w) f_W(w) \, dw$$
$$- \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y(1)|D,W}(y_\tau|1,w) - F_{Y(1)|U_D,W}(y_\tau|P(w),w) \right] \dot{P}(w) f_W(w) \, dw.$$

To facilitate understanding of Corollary 3, we can define and organize the average influence functions (AIF) in a table:

	AIF for $Y(1)$ − AIF for $Y(0)$
U_D	$E_w \left[\psi_\tau(Y(1)) - \psi_\tau(Y(0)) \, | U_D = P(w) \right]$
D	$E_w \left[\psi_\tau(Y(1)) \, | D = 1 \right] - E_w \left[\psi_\tau(Y(0)) \, | D = 0 \right]$

where $\psi_\tau(\cdot)$ is short for $\psi(\cdot, \rho_\tau, F_Y)$, the influence function of the quantile functional. In the above, $E_w[\cdot]$ stands for the conditional mean operator given $W = w$. For example, $E_w \left[\psi_\tau(Y(0)) \, | D = 0 \right]$ stands for $E \left[\psi_\tau(Y(0)) \, | D = 0, W = w \right]$. Let

$$\psi_{\Delta, U_D}(w) := E_w \left[\psi_\tau(Y(1)) - \psi_\tau(Y(0)) \, | U_D = P(w) \right],$$
$$\psi_{\Delta, D}(w) := E_w \left[\psi_\tau(Y(1)) \, | D = 1 \right] - E_w \left[\psi_\tau(Y(0)) \, | D = 0 \right].$$

The unconditional quantile effect Π_τ is the average of the difference $\psi_{\Delta, U_D}(w)$ with respect to the distribution of W over the *marginal subpopulation*. The average *apparent* effect A_τ is the average of the difference $\psi_{\Delta, D}(w)$ with respect to the distribution of W over the whole population distribution. It is also equal to the limit of the UQR estimator of Firpo, Fortin and Lemieux (2009), where the endogeneity of the treatment selection is ignored.[8] Note that if our model contains no covariate W, then

$$A_\tau = \frac{1}{f_Y(y_\tau)} E \left[\mathbb{1}\{Y \le y_\tau\} \, | D = 0 \right] - \frac{1}{f_Y(y_\tau)} E \left[\mathbb{1}\{Y \le y_\tau\} \, | D = 1 \right]$$
$$= \frac{1}{f_Y(y_\tau)} \left[\Pr(Y > y_\tau | D = 1) - \Pr(Y > y_\tau | D = 0) \right]. \tag{20}$$

This is identical to the unconditional quantile effect given in (17).

---

[8]To see why this is the case, we note that, in its simplest form, the UQR involves regressing the "influence function" $\hat{f}_Y(\hat{y}_\tau)^{-1}(1 - \mathbb{1}\{Y_i \le \hat{y}_\tau\})$ on D_i and W_i by OLS and using the estimated coefficient on D_i as the estimator of the unconditional quantile effect. Here $\hat{f}_Y(y_\tau)$ is a consistent estimator of $f_Y(y_\tau)$ and $\hat{y}_\tau$ is a consistent estimator of y_τ. It is now easy to see that the UQR estimator converges in probability to A_τ if the conditional expectations in (19) are linear in W.

The discrepancy between Π_τ and A_τ gives rise to the asymptotic *bias* B_τ of the UQR estimator of Firpo, Fortin and Lemieux (2009):

$$
\begin{aligned}
B_\tau &= A_\tau - \Pi_\tau = E\left[\psi_{\Delta,D}\left(W\right)\right] - E\left[\psi_{\Delta,U_D}\left(W\right)\dot{P}\left(W\right)\right] \\
&= \underbrace{E\left\{\psi_{\Delta,D}\left(W\right)\left[1 - \dot{P}\left(W\right)\right]\right\}}_{B_{1\tau}} + \underbrace{E\left\{\left[\psi_{\Delta,D}\left(W\right) - \psi_{\Delta,U_D}\left(W\right)\right]\dot{P}\left(W\right)\right\}}_{B_{2\tau}}.
\end{aligned}
\tag{21}
$$

It is easy to see that $B_{1\tau}$ and $B_{2\tau}$ given above are identical to those given in Corollary 3.

The decomposition in Equation (21) traces the asymptotic bias back to two sources. The first one, $B_{1\tau}$, captures the heterogeneity of the averaged apparent effects averaged over two different subpopulations. For every w, $\psi_{\Delta,D}\left(w\right)$ is the average effect of D on $\left[\tau - \mathbf{1}\left\{Y \leq y_\tau\right\}\right]/f_Y\left(y_\tau\right)$ for the individuals with $W = w$. These effects are averaged over two different distributions of W: the distribution of W for the marginal subpopulation (i.e., $\dot{P}\left(w\right)f_W\left(w\right)$) and the distribution of W for the whole population (i.e., $f_W\left(w\right)$). $B_{1\tau}$ is equal to the difference of these two average effects. If the effect $\psi_{\Delta,D}\left(w\right)$ does not depend on w, then $B_{1\tau} = 0$. If $\dot{P}(w) = 1$, then the distribution of W over the whole population is the same as that over the subpopulation, and hence $B_{1\tau} = 0$ as well. For $B_{1\tau} \neq 0$, it is necessary that there is an effect heterogeneity (i.e., $\psi_{\Delta,D}\left(w\right)$ depends on w) and a distributional heterogeneity (i.e., $\dot{P}\left(w\right) \neq 1$ so that the distribution of W over the marginal subpopulation is different from that over the whole population). To highlight the necessary conditions for a nonzero $B_{1\tau}$, we refer to $B_{1\tau}$ as the *marginal heterogeneity bias.*

The second bias component, $B_{2\tau}$, embodies the second source of the bias and has a difference-in-differences interpretation. Each of $\psi_{\Delta,D}\left(\cdot\right)$ and $\psi_{\Delta,U_D}\left(\cdot\right)$ is the difference in the average influence functions associated with the counterfactual outcomes $Y\left(1\right)$ and $Y\left(0\right)$. However, $\psi_{\Delta,D}\left(\cdot\right)$ is the difference over the two subpopulations who actually choose $D = 1$ and $D = 0$, while $\psi_{\Delta,U_D}\left(\cdot\right)$ is the difference over the marginal subpopulation. So $\psi_{\Delta,D}\left(\cdot\right) - \psi_{\Delta,U_D}\left(\cdot\right)$ is a difference in differences. $B_{2\tau}$ is simply the average of this difference in differences with respect to the distribution of W over the marginal subpopulation. This term arises because the change in the distributions of Y for those with $D = 1$ and those with $D = 0$ is different from that for those whose U_D is just above $P\left(w\right)$ and those whose U_D is just below $P\left(w\right)$. Thus we can label $B_{2\tau}$ as a *marginal selection bias.*

If $\psi_{\Delta,D}\left(w\right) = \psi_{\Delta,U_D}\left(w\right)$ for almost all $w \in \mathcal{W}$, then $B_{2\tau} = 0$. The condition $\psi_{\Delta,D}\left(w\right) = \psi_{\Delta,U_D}\left(w\right)$ is the same as

$$
E_w\left[\psi_\tau\left(Y\left(1\right)\right) - \psi_\tau\left(Y\left(0\right)\right)|U_D = P\left(w\right)\right] = E_w\left[\psi_\tau\left(Y\left(1\right)\right)|D = 1\right] - E_w\left[\psi_\tau\left(Y\left(0\right)\right)|D = 0\right].
$$

Equivalently,

$$
\begin{aligned}
&E_w\left[\psi_\tau\left(Y\left(1\right)\right)|U_D = P\left(w\right)\right] - E_w\left[\psi_\tau\left(Y\left(1\right)\right)|D = 1\right] \\
&= E_w\left[\psi_\tau\left(Y\left(0\right)\right)|U_D = P\left(w\right)\right] - E_w\left[\psi_\tau\left(Y\left(0\right)\right)|D = 0\right].
\end{aligned}
$$

The condition resembles the parallel-paths assumption or the constant-bias assumption in a difference-in-differences analysis. If U_D is independent of $\left(U_0, U_1\right)$ given W, then this condition holds and $B_{2\tau} = 0$.

In general, when U_D is not independent of (U_0, U_1) given W, and W enters the selection equation, we have $B_{1\tau} \neq 0$ and $B_{2\tau} \neq 0$, hence $\Pi_\tau \neq A_\tau$. If $\dot{P}(w)$ is not identified, then $B_{1\tau}$ is not identified. In general, $B_{2\tau}$ is not identified without additional assumptions. Therefore, in the absence of additional assumptions, the asymptotic bias can not be eliminated, and Π_τ is not identified.

It is not surprising that in the presence of endogeneity, the UQR estimator of Firpo, Fortin and Lemieux (2009) is asymptotically biased. The virtue of Corollary 3 is that it provides a closed-form characterization and clear interpretations of the asymptotic bias. To the best of our knowledge, this bias formula is new in the literature. If point identification can not be achieved, then the bias formula can be used in a bound analysis or sensitivity analysis. From a broad perspective, the asymptotic bias B_τ is the unconditional quantile counterpart of the endogenous bias of the OLS estimator in a linear regression framework.[9]

Remark 3. *Consider a setting of full independence: $V \perp U_0 \perp U_1 \perp W$ (i.e., every subset of these variables is independent of the rest). In this case, $B_{2\tau} = 0$ and by equation (18), the UQE is*

$$
\begin{aligned}
\Pi_\tau &= \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\left\{Y(0) \leq y_\tau\right\} | W = w\right] \dot{P}(w) f_W(w)\, dw \\
&\quad - \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\left\{Y(1) \leq y_\tau\right\} | W = w\right] \dot{P}(w) f_W(w)\, dw.
\end{aligned}
$$

Following (19), the apparent effect is

$$
\begin{aligned}
A_\tau &= \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\left\{Y(0) \leq y_\tau\right\} | W = w\right] f_W(w)\, dw \\
&\quad - \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\left\{Y(1) \leq y_\tau\right\} | W = w\right] f_W(w)\, dw.
\end{aligned}
$$

In general, therefore, we will still have a marginal heterogeneity bias term given by $B_{1\tau}$ unless $\dot{P}(w) = 1$ or the difference $E\left[\mathbb{1}\left\{Y(0) \leq y_\tau\right\} | W = w\right] - E\left[\mathbb{1}\left\{Y(1) \leq y_\tau\right\} | W = w\right]$ does not depend on w. In general, both conditions fail if some covariate (i.e., X) in W enters both the outcome equations and the selection equation nontrivially. In this case, the unconditional quantile regression estimator of Firpo, Fortin and Lemieux (2009), which converges to A_τ, will be asymptotically biased even though there is no endogeneity.

In the presence of endogeneity, it is not easy to evaluate or even sign the asymptotic bias. In general, the joint distribution of (U, U_D) given the covariate W is needed for this purpose. This is not atypical. For a nonlinear estimator such as the unconditional quantile estimator, its asymptotic properties often depend on the full data generating process in a nontrivial way. This is in sharp contrast with a linear estimator such as the OLS in a linear regression model whose properties depend on only the first few moments of the data.

---

[9]The bias decomposition is not unique. Corollary 3 gives only one possibility. We can also write

$$
B_\tau = \underbrace{E\left\{\psi_{\Delta, U_D}(W)\left[1 - \dot{P}(W)\right]\right\}}_{\tilde{B}_{1\tau}} + \underbrace{E\left[\psi_{\Delta, D}(W) - \psi_{\Delta, U_D}(W)\right]}_{\tilde{B}_{2\tau}}.
$$

The interpretations of $\tilde{B}_{1\tau}$ and $\tilde{B}_{2\tau}$ are similar to those of $B_{1\tau}$ and $B_{2\tau}$ with obvious and minor modifications. In this case, it is even more revealing to call $\tilde{B}_{1\tau}$ the marginal heterogeneity bias, because a necessary condition for a nonzero $\tilde{B}_{1\tau}$ is that there is an effect heterogeneity among the marginal subpopulation.

Figure 2 shows the non-uniformity of the asymptotic bias, computed as $B_\tau := A_\tau - \Pi_\tau$. It is derived from a model (details can be found in Section B of the Appendix) where the endogeneity is controlled by a single correlation parameter ρ. The case $\rho = 0$ corresponds to the case of an exogenous treatment and illustrates the *marginal heterogeneity bias*.
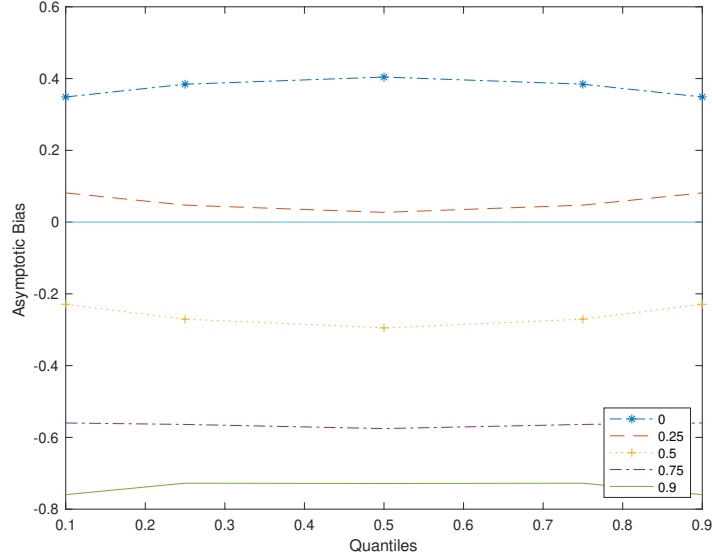

Figure 2: Asymptotic bias of the UQR estimator for different values of ρ.

4 Unconditional Quantile Effect under Instrumental Intervention

In the previous section, we have shown that the UQR estimator of Firpo, Fortin and Lemieux (2009) is asymptotically biased under endogeneity or marginal heterogeneity, and the UQE is in general unidentified. In this section, we impose additional assumptions on Z in order to achieve the point identification of the UQE. The solution we propose to tackle the endogeneity problem for the quantile functional can be easily generalized to deal with a general functional. To save space, we do not pursue this straightforward generalization.

4.1 Instrumental Intervention

We impose the following additional assumptions on Z, taken directly from Heckman and Vytlacil (1999, 2001a, 2005)

Assumption 5. *Relevance and Exogeneity*

(a) $\mu(Z, X)$ is a non-degenerate random variable conditional on X.

(b) (U_0, U_1, V) is independent of Z conditional on X.

Assumption 5(a) is a relevance assumption: for any given level of X, the variable Z can induce some variation in D. Assumption 5(b) is referred to as an exogeneity assumption. The two assumptions are essentially the conditions for a valid instrumental variable, hence we will refer to Z as the instrumental variable and refer to the intervention as the instrumental intervention.

Assumption 5(b) allows us to write

$$U_D = F_{V|Z,X}(V|Z,X) = F_{V|X}(V|X).$$

Assumption 5(b) then implies that (U_0, U_1, U_D) is independent of Z conditional on X. If conditioning on the covariate X that enters the structural functions $r_0(\cdot, U_0)$ and $r_1(\cdot, U_1)$ is not enough to ensure the independence of (U_0, U_1, U_D) from Z, we may augment the conditioning set to include additional control variables, say X_C. The control variables in X_C do not enter r_0, r_1, or μ. However, we can think that X_C enters $r_0, r_1,$ and μ *trivially* so that with some abuse of notation we can write

$$
\begin{aligned}
Y(0) &= r_0(X_S, X_C, U_0), \; Y(1) = r_1(X_S, X_C, U_1), \\
D &= \mathbb{1}\{V \leq \mu(Z, X_S, X_C)\},
\end{aligned}
$$

where X_S is the structural variable that enters r_0, r_1 and μ nontrivially and X_C is the control variable that enters r_0, r_1 and μ trivially (i.e., r_0, r_1 and μ are constant functions of X_C given other variables). Letting $X = (X_S, X_C)$ and $W = (Z, X_S, X_C)$, the above model then takes the same form as the model in (1) and (2). With this conceptual change, our results remain valid for the case with additional controls.

Using the influence function of the quantile functional, we define an unconditional marginal treatment effect for the τ-quantile, which will be a basic building block for the unconditional quantile effect.

Definition 5. *The unconditional marginal treatment effect for the τ-quantile is defined as*

$$
\begin{aligned}
\mathrm{MTE}_\tau(u, x) &:= E\left[\left.\frac{\tau - \mathbb{1}\{Y(1) \leq y_\tau\}}{f_Y(y_\tau)} - \frac{\tau - \mathbb{1}\{Y(0) \leq y_\tau\}}{f_Y(y_\tau)}\right| U_D = u, X = x\right] \\
&= \frac{1}{f_Y(y_\tau)} E\left[\mathbb{1}\{Y(0) \leq y_\tau\} - \mathbb{1}\{Y(1) \leq y_\tau\} \mid U_D = u, X = x\right].
\end{aligned}
$$

The MTE_τ is different from the quantile analogue of the marginal treatment effect of Carneiro and Lee (2009) and Yu (2014), which is defined as $F_{Y(1)|U_D,X}^{-1}(\tau|u,x) - F_{Y(0)|U_D,X}^{-1}(\tau|u,x)$. An unconditional quantile effect can not be represented as an integrated version of the latter.

The MTE_τ given in Definition 5 is a special case of the general unconditional MTE in (10) with $\psi(y, \rho, F_Y) = [\tau - \mathbb{1}\{y \leq y_\tau\}] / f_Y(y_\tau)$. Here, because of Assumption 5(b), conditioning on Z is not necessary and has been dropped.

Define $\Delta(y_\tau) := (\mathbb{1}\{Y(0) \leq y_\tau\} - \mathbb{1}\{Y(1) \leq y_\tau\}) / f_Y(y_\tau)$, which underlies the above definition

of $\text{MTE}_\tau(u,x)$. The random variable $\Delta(y_\tau)$ can take three values:

$$\Delta(y_\tau) = \begin{cases} f_Y(y_\tau)^{-1} & \text{if } Y(0) \le y_\tau \text{ and } Y(1) > y_\tau \\ 0 & \text{if } \left[Y(0) > y_\tau \text{ and } Y(1) > y_\tau\right] \text{ or } \left[Y(0) \le y_\tau \text{ and } Y(1) \le y_\tau\right] \\ -f_Y(y_\tau)^{-1} & \text{if } Y(0) > y_\tau \text{ and } Y(1) \le y_\tau \end{cases}$$

For a given individual, $\Delta(y_\tau) = 1/f_Y(y_\tau)$ when the treatment induces the individual to "cross" the τ-quantile y_τ of Y from below, and $\Delta(y_\tau) = -1/f_Y(y_\tau)$ when the treatment induces the individual to "cross" the τ-quantile y_τ of Y from above. In the first case, the individual benefits from the treatment while in the second case the treatment harms her. The intermediate case, $\Delta(y_\tau) = 0$, occurs when the treatment induces no quantile crossing of any type. Thus, the unconditional expected value $f_Y(y_\tau) \times E[\Delta(y_\tau)]$ equals the difference between the proportion of individuals who benefit from the treatment and the proportion of individuals who are harmed by it. For the UQE, whether the treatment is beneficial or harmful is measured in terms of quantile crossing. Among the individuals with characteristics $U_D = u$ and $X = x$, $\text{MTE}_\tau(u,x)$ is then equal to the rescaled (by $1/f_Y(y_\tau)$) difference between the proportion of individuals who benefit from the treatment and the proportion of individuals who are harmed by it. Thus, $\text{MTE}_\tau(u,x)$ is positive if more individuals increase their outcome above y_τ, and it is negative if more individuals decrease their outcome below y_τ.

To bring our setting closer to the case of the shift of a continuous covariate that Firpo, Fortin and Lemieux (2009) study, we consider the intervention

$$\mathcal{G}(W,\delta) = (Z_1 + g(W)s(\delta), Z_2, \cdots, Z_{d_Z}) \tag{22}$$

where $g(\cdot)$ is a measurable function and $s(\delta)$ is a smooth function satisfying $s(0) = 0$. That is, we intervene to change the first component of Z. Note that while $s(\delta)$ is the same for all individuals, $g(W)$ depends on the value of W and hence it is individual-specific. Thus, we allow the intervention to be heterogeneous.[10] In empirical applications, Z may consist of a few variables, and Z_1 is the target variable that we consider to change. The unconditional policy effect is specific to the variable Z_1 that we choose to intervene in order to improve the treatment adoption rate.

Lemma 3. *Assume that (i) (V,X) are absolutely continuous random variables with joint density $f_{V,X}(v,x)$ given by $f_{V|X}(v|x) f_X(x)$; (ii) $f_{V|X}(v|x)$ is continuous in v for almost all $x \in \mathcal{X}$; (iii) $\mu(z,x)$ is continuously differentiable in z_1 for almost all $w \in \mathcal{W}$; (iv) letting $\mu'_{z_1}(z,x)$ be the partial derivative of $\mu(z,x)$ with respect to z_1, we have*

$$E \sup_{\delta \in N_\varepsilon} \left| f_{V|X}(\mu(\mathcal{G}(W,\delta),X)|X) \mu'_{z_1}(\mathcal{G}(W,\delta),X) g(W) \right| < \infty;$$

---

[10]A simple homogenous intervention is obtained by setting $g(\cdot) \equiv 1$, in which case we have an additive shift in Z. An example of a heterogeneous intervention is given by $g(W) = Z$, in which case we obtain a multiplicative shift of order $1 + s(\delta)$. Both cases have been studied by Carneiro, Heckman and Vytlacil (2010).

(v) for each $\delta \in N_\varepsilon$,

$$E\left[f_{V|X}\left(\mu\left(\mathcal{G}\left(W,\delta\right),X\right)|X\right)\mu'_{z_1}\left(\mathcal{G}\left(W,\delta\right),X\right)g(W)\right] \neq 0;$$

(vi) $s(\delta)$ is a differential function in a neighborhood of zero and $s(0) = 0$.

Then

$$\left.\frac{\partial s(\delta)}{\partial\delta}\right|_{\delta=0} = \frac{1}{E\left[f_{V|W}\left(\mu\left(W\right)|W\right)\mu'_{z_1}\left(W\right)g(W)\right]},$$

$$\left.\frac{\partial P_\delta(z,x)}{\partial\delta}\right|_{\delta=0} = \frac{f_{V|W}\left(\mu(w)|w\right)\mu'_{z_1}\left(w\right)g(w)}{E\left[f_{V|W}\left(\mu\left(W\right)|W\right)\mu'_{z_1}\left(W\right)g(W)\right]}.$$

Corollary 4. *Let Assumptions 1–3 and 5, and the assumptions of Lemma 3 hold. Assume further that $f_Y(y_\tau) > 0$. Then, the unconditional quantile effect of the shift in Z given in (22) is*

$$\Pi_\tau = \int_{\mathcal{W}} \mathrm{MTE}_\tau\left(P\left(w\right),x\right)\dot{P}\left(w\right)f_W\left(w\right)dw,$$

where

$$\dot{P}(w) = \left.\frac{\partial P_\delta(w)}{\partial\delta}\right|_{\delta=0} = \frac{f_{V|X}\left(\mu(w)|x\right)\mu'_{z_1}\left(w\right)g(w)}{E\left[f_{V|X}\left(\mu\left(W\right)|X\right)\mu'_{z_1}\left(W\right)g\left(W\right)\right]}.$$

4.2 Identification of the UQE

To investigate the identifiability of Π_τ, we study the identifiability of the unconditional MTE_τ and the weight function or the RN derivative $\left.\frac{\partial P_\delta(w)}{\partial\delta}\right|_{\delta=0}$ separately. The proposition below shows that $\mathrm{MTE}_\tau(u,x)$ is identified for every $u = P(w)$ for some $w \in \mathcal{W}$.

Proposition 1. *Let Assumptions 2(a), 2(b), and 5(b) hold. Then, for every $u = P(w)$ with $w \in \mathcal{W}$, we have*

$$\mathrm{MTE}_\tau(u,x) = -\frac{1}{f_Y\left(y_\tau\right)}\frac{\partial E\left[\mathbb{1}\left\{Y \leq y_\tau\right\}|P(W) = u, X = x\right]}{\partial u}.$$

Proposition 1 can be proved using Theorem 1 in Carneiro and Lee (2009). In the Appendix we provide a self-contained proof that is directly connected to the idea of shifting the propensity score. For a more general functional ρ, we can show that

$$\mathrm{MTE}_\rho(u,x) = \frac{\partial E\left[\psi(Y,\rho,F_Y)|P(W) = u, X = x\right]}{\partial u}$$

for every u equal to $P(w)$ for some $w \in \mathcal{W}$.

Now we turn to the identification of the RN derivative $\dot{P}(w)$ given in Corollary 4. Under Assumption 5(b), the propensity score becomes

$$P(w) = \Pr\left(V \leq \mu\left(W\right)|W = w\right) = F_{V|Z,X}(\mu(z,x)|z,x) = F_{V|X}(\mu(z,x)|x).$$

Therefore,

$$\frac{\partial P(w)}{\partial z_1} = f_{V|X}(\mu(z,x)|x)\mu'_{z_1}(z,x) = f_{V|X}(\mu(w)|x)\mu'_{z_1}(w).$$

It is now clear that $\dot{P}(w)$ can be represented using $\frac{\partial P(w)}{\partial z_1}$ and $g(w)$. We formalize this in the following proposition.

Proposition 2. *Let Assumption 5 and the assumptions in Lemma 3 hold. Then*

$$\dot{P}(w) = \left.\frac{\partial P_\delta(w)}{\partial \delta}\right|_{\delta=0} = \frac{\frac{\partial P(w)}{\partial z_1}g(w)}{E\left[\frac{\partial P(W)}{\partial z_1}g(W)\right]}. \tag{23}$$

Since $g(w)$ is known and $\frac{\partial P(w)}{\partial z_1}$ is identified, $\dot{P}(w)$ is also identified. As in the case of MTE$_\tau$, Assumption 5(b) plays a key role in identifying $\dot{P}(w)$. Without the assumption that V is independent of Z conditional on X, we can have only that

$$\dot{P}(w) = \frac{f_{V|W}(\mu(w)|w)\mu'_{z_1}(w)g(w)}{E\left[f_{V|W}(\mu(W)|W)\mu'_{z_1}(W)g(W)\right]}$$

and

$$\frac{\partial P(w)}{\partial z_1} = f_{V|Z,X}(\mu(w)|w)\mu'_{z_1}(w) + \left.\frac{\partial F_{V|Z,X}(\mu(w)|\tilde{z},x)}{\partial \tilde{z}_1}\right|_{\tilde{z}=z}.$$

The presence of the second term in the above equation invalidates the identification result in (23).

Using Propositions 1 and 2, we can represent Π_τ as

$$\Pi_\tau = -\frac{1}{f_Y(y_\tau)}\int_{\mathcal{W}}\frac{\partial E\left[\mathbb{1}\left\{Y \leq y_\tau\right\}|P(W)=P(w),X=x\right]}{\partial P(w)}\frac{\frac{\partial P(w)}{\partial z_1}g(w)}{E\left[\frac{\partial P(W)}{\partial z_1}g(W)\right]}f_W(w)\,dw. \tag{24}$$

All objects in the above are point identified, hence Π_τ is point identified.

5 Unconditional Instrumental Quantile Estimation

This section is devoted to the estimation and inference of the UQE under the instrumental intervention in (22). We assume that the propensity score function is parametric, and we leave the case with a nonparametric propensity score to Section D of the Appendix. In order to simplify the notation, we set $g(\cdot) \equiv 1$ for the remainder of this paper.[11]

Letting

$$m_0(y_\tau, P(w), x) := E\left[\mathbb{1}\left\{Y \leq y_\tau\right\}|P(W)=P(w),X=x\right]$$

and using (24), we have

---

[11]The presence of $g(\cdot)$ amounts to a change of measure: from a measure with density $f_W(w)$ to a measure with density $g(w)f_W(w)$. When $g(\cdot)$ is not equal to a constant function, we only need to change the population expectation operator $E[h(W)]$ that involves the distribution of W into $E[h(W)g(W)]$ and the empirical average operator $\mathbb{P}_n[h(W)]$ into $\mathbb{P}_n[h(W)g(W)]$. All of our results will remain valid.

$$\Pi_\tau = -\frac{1}{f_Y(y_\tau)} E\left[\frac{\partial P(W)}{\partial z_1}\right]^{-1} E\left[\frac{\partial m_0(y_\tau, P(W), X)}{\partial z_1}\right].$$

Π_τ consists of two average derivatives and a density evaluated at a point, some of which depend on the unconditional τ-quantile y_τ. Altogether Π_τ depends on four unknown quantities. The method of unconditional instrumental quantile estimation involves first estimating the four quantities separately and then plugging these estimates into Π_τ to obtain the estimator $\hat{\Pi}_\tau$. See (32) for the formula of $\hat{\Pi}_\tau$.

We consider estimating the four quantities in the next few subsections. For a given sample $\{O_i = (Y_i, Z_i, X_i, D_i)\}_{i=1}^n$, we will use $\mathbb{P}_n$ to denote the empirical measure. The expectation of a function $\chi(O)$ with respect to $\mathbb{P}_n$ is then $\mathbb{P}_n\chi = n^{-1}\sum_{i=1}^n \chi(O_i)$.

5.1 Estimating the Quantile and Density

For a given τ, we estimate y_τ using the (generalized) inverse of the empirical distribution function of Y: $\hat{y}_\tau = \inf\{y : \mathbb{F}_n(y) \geq \tau\}$, where

$$\mathbb{F}_n(y) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_i \leq y\}.$$

The following asymptotic result can be found in Serfling (1980).[12]

Lemma 4. *If the density $f_Y(\cdot)$ of Y is positive and continuous at y_τ, then $\hat{y}_\tau - y_\tau = \mathbb{P}_n\psi_Q(y_\tau) + o_p(n^{-1/2})$ where*

$$\psi_Q(y_\tau) := \frac{\tau - \mathbb{1}\{Y \leq y_\tau\}}{f_Y(y_\tau)}.$$

We use a kernel density estimator to estimate $f_Y(y)$. We maintain the following assumptions on the kernel function and the bandwidth.

Assumption 6. *Kernel Assumption*

The kernel function $K(\cdot)$ satisfies (i) $\int_{-\infty}^\infty K(u)du = 1$, (ii) $\int_{-\infty}^\infty u^2 K(u)du < \infty$, and (iii) $K(u) = K(-u)$, and it is twice differentiable with Lipschitz continuous second-order derivative $K''(u)$ satisfying (i) $\int_{-\infty}^\infty K''(u)udu < \infty$ and (ii) there exist positive constants C_1 and C_2 such that $|K''(u_1) - K''(u_2)| \leq C_2|u_1 - u_2|^2$ for $|u_1 - u_2| \geq C_1$.

Assumption 7. *Rate Assumption $n \uparrow \infty$ and $h \downarrow 0$ such that $nh^3 \uparrow \infty$ but $nh^5 = O(1)$.*

The non-standard condition $nh^3 \uparrow \infty$ is due to the estimation of y_τ. Since we need to expand $\hat{f}_Y(\hat{y}_\tau) - \hat{f}_Y(y_\tau)$, which involves the derivative of $\hat{f}_Y(y)$, we have to impose a slower rate of decay for h to control the remainder. The details can be found in the proof of Lemma 5. We note, however, that $nh^3 \uparrow \infty$ implies the usual rate condition $nh \uparrow \infty$.

The estimator of $f_Y(y)$ is then given by

$$\hat{f}_Y(y) = \frac{1}{n}\sum_{i=1}^n K_h(Y_i - y),$$

---

[12]See Section 2.5.1. Actually, Serfling (1980) provides a better rate for the remainder.

where $K_h(u) := K(u/h)/h$.

Lemma 5. *Let Assumptions 6 and 7 hold. Then*

$$\hat{f}_Y(y) - f_Y(y) = \mathbb{P}_n \psi_{f_Y}(y) + B_{f_Y}(y) + o_p(h^2),$$

where

$$\psi_{f_Y}(y) := K_h(Y - y) - E[K_h(Y - y)] = O_p(n^{-1/2}h^{-1/2})$$

and

$$B_{f_Y}(y) = \frac{1}{2}h^2 f_Y''^2(y) \int_{-\infty}^{\infty} u^2 K(u) du.$$

Furthermore, for the quantile estimator $\hat{y}_\tau$ of y_τ that satisfies Lemma 4, we have

$$\hat{f}_Y(\hat{y}_\tau) - \hat{f}_Y(y_\tau) = f_Y(\hat{y}_\tau) - f_Y(y_\tau) + R_{f_Y} = f_Y'(y_\tau)\mathbb{P}_n \psi_Q(y_\tau) + R_{f_Y},$$

where

$$R_{f_Y} = o_p(n^{-1/2}h^{-1/2}).$$

In order to isolate the contributions of $\hat{f}$ and $\hat{y}_\tau$, we can use Lemma 5 to write

$$\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau) = \hat{f}_Y(y_\tau) - f_Y(y_\tau) + f_Y(\hat{y}_\tau) - f_Y(y_\tau) + R_{f_Y}. \tag{25}$$

The first pair of terms on the right-hand side of (25) represents the dominant term and reflects the uncertainty in the estimation of f_Y. The second pair of terms reflects the error from estimating y_τ. In order to ensure that $R_{f_Y} = o_p(n^{-1/2}h^{-1/2})$, we need $nh^3 \uparrow \infty$, as stated in Assumption 7.

5.2 Estimating the Average Derivatives

To estimate the two average derivatives, we make a parametric assumption on the propensity score, leaving the nonparametric specification to the appendix.

Assumption 8. *The propensity score $P(Z, X, \alpha_0)$ is known up to a finite-dimensional vector $\alpha_0 \in \mathbb{R}^{d_\alpha}$.*

Under Assumption 8, the parameter Π_τ can be written as

$$\Pi_\tau = -\frac{1}{f_Y(y_\tau)} \cdot \frac{1}{T_1} \cdot T_2, \tag{26}$$

where

$$T_1 = E\left[\frac{\partial P(Z, X, \alpha_0)}{\partial z_1}\right] \text{ and } T_2 = E\left[\frac{\partial m_0(y_\tau, P(Z, X, \alpha_0), X)}{\partial z_1}\right].$$

First, we estimate T_1, the mean of the derivative of the propensity score, by

$$T_{1n}(\hat{\alpha}) := \frac{1}{n}\sum_{i=1}^{n} \frac{\partial P(z, x, \hat{\alpha})}{\partial z_1}\bigg|_{(z,x)=(Z_i, X_i)}$$

25

where $\hat{\alpha}$ is an estimator of α_0. To save space, we slightly abuse notation and write

$$\frac{\partial P(Z, X, \hat{\alpha})}{\partial z_1} = \frac{\partial P(z, x, \hat{\alpha})}{\partial z_1}\Bigg|_{(z,x)=(Z,X)}.$$

We adopt this convention in the rest of the paper.

Lemma 6. *Suppose that*

(a) $\hat{\alpha}$ *admits the representation* $\hat{\alpha} - \alpha_0 = \mathbb{P}_n\psi_{\alpha_0} + o_p(n^{-1/2})$, *where* $\psi_{\alpha_0}(W_i)$ *is a mean-zero* $d_\alpha \times 1$ *random vector with* $E\|\psi_{\alpha_0}(W_i)\|^2 < \infty$, *and* $\|\cdot\|$ *denotes the Euclidean norm;*

(b) *the variance of* $\frac{\partial P(Z, X, \alpha_0)}{\partial z_1} := \frac{\partial P(z, x, \alpha_0)}{\partial z_1}\big|_{(z,x)=(Z,X)}$ *is finite;*

(c) *the* $d_\alpha \times 1$ *derivative vector* $\frac{\partial^2 P(z, x, \alpha)}{\partial \alpha \partial z_1}$ *exists for all* z *and* x *and for* α *in an open neighborhood around* α_0;

(d) *for* $\alpha \in \mathcal{A}_0$, *a neighborhood around* α_0, *the map* $\alpha \mapsto E\left[\frac{\partial^2 P(Z, X, \alpha)}{\partial \alpha \partial z_1}\right]$ *is continuous and a uniform law of large numbers holds:* $\sup_{\alpha \in \mathcal{A}_0} \left\|\mathbb{P}_n \frac{\partial^2 P(Z_i, X_i, \alpha)}{\partial \alpha \partial z_1} - E\left[\frac{\partial^2 P(Z, X, \alpha)}{\partial \alpha \partial z_1}\right]\right\| \xrightarrow{p} 0.$

Then, $T_{1n}(\hat{\alpha}) - T_1$ *has the following stochastic approximation*

$$T_{1n}(\hat{\alpha}) - T_1 = E\left[\frac{\partial^2 P(Z, X, \alpha_0)}{\partial z_1 \partial \alpha}\right]' \mathbb{P}_n\psi_{\alpha_0} + \mathbb{P}_n\psi_{\partial P} + o_p(n^{-1/2}),$$

where

$$\psi_{\partial P} := \frac{\partial P(Z, X, \alpha_0)}{\partial z_1} - E\left[\frac{\partial P(Z, X, \alpha_0)}{\partial z_1}\right].$$

We can rewrite the main result of Lemma 6 as

$$
\begin{aligned}
T_{1n}(\hat{\alpha}) - T_1 &= T_{1n}(\alpha_0) - T_1 + T_{1n}(\hat{\alpha}) - T_{1n}(\alpha_0) \\
&= T_{1n}(\alpha_0) - T_1 + E\left[\frac{\partial P(Z, X, \hat{\alpha})}{\partial z_1}\right] - E\left[\frac{\partial P(Z, X, \alpha_0)}{\partial z_1}\right] + o_p(n^{-1/2}).
\end{aligned}
\tag{27}
$$

Equation (27) has the same interpretation as equation (25). It consists of a pair of leading terms that ignores the estimation uncertainty in $\hat{\alpha}$ but accounts for the variability of the sample mean, and another pair that accounts for the uncertainty in $\hat{\alpha}$ but ignores the variability of the sample mean.

We estimate the second average derivative T_2 by

$$T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) := \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{m}(\hat{y}_\tau, P(Z_i, X_i, \hat{\alpha}), X_i)}{\partial z_1};\tag{28}$$

see (29) for an explicit construction. We can regard $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})$ as a four-step estimator. The first step estimates y_τ, the second step estimates α_0, the third step estimates the conditional expectation $m_0(y, P(Z, X, \alpha_0), X)$ using the generated regressor $P(Z, X, \hat{\alpha})$, and the fourth step averages the derivative (with respect to Z_1) over X and the generated regressor $P(Z, X, \hat{\alpha})$.

We use the series method to estimate m_0. To alleviate notation, define the vector $\tilde{w}(\alpha) := (P(z, x, \alpha), x)'$ and $\tilde{W}_i(\alpha) := (P(Z_i, X_i, \alpha), X_i)'$. We write $\tilde{w} = \tilde{w}(\alpha_0)$ and $\tilde{W}_i = \tilde{W}_i(\alpha_0)$ to suppress their dependence on the true parameter value α_0. Both $\tilde{w}(\alpha)$ and $\tilde{W}_i(\alpha)$ are in $\mathbb{R}^{d_X+1}$. Let $\phi^J(\tilde{w}(\alpha)) = (\phi_{1J}(\tilde{w}(\alpha)), \ldots, \phi_{JJ}(\tilde{w}(\alpha)))'$ be a vector of J basis functions of $\tilde{w}(\alpha)$ with finite second moments. Here, each $\phi_{jJ}(\cdot)$ is a differentiable basis function. Then, the series estimator of $m_0(y_\tau, \tilde{w}(\alpha))$ is $\hat{m}(\hat{y}_\tau, \tilde{w}(\hat{\alpha})) = \phi^J(\tilde{w}(\hat{\alpha}))'\hat{b}(\hat{\alpha}, \hat{y}_\tau)$, where $\hat{b}(\hat{\alpha}, \hat{y}_\tau)$ is:

$$\hat{b}(\hat{\alpha}, \hat{y}_\tau) = \left(\sum_{i=1}^n \phi^J(\tilde{W}_i(\hat{\alpha}))\phi^J(\tilde{W}_i(\hat{\alpha}))' \right)^{-1} \sum_{i=1}^n \phi^J(\tilde{W}_i(\hat{\alpha}))\mathbb{1}\{Y_i \leq \hat{y}_\tau\}.$$

The estimator of the average derivative T_2 is then

$$T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) = \frac{1}{n} \sum_{i=1}^n \frac{\partial \phi^J(\tilde{W}_i(\hat{\alpha}))}{\partial z_1}' \hat{b}(\hat{\alpha}, \hat{y}_\tau). \tag{29}$$

We use the path derivative approach of Newey (1994) to obtain a decomposition of $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) - T_2$, which is similar to that in Section 2.1 of Hahn and Ridder (2013). To describe the idea, let $\{F_\theta\}$ be a path of distributions indexed by $\theta \in \mathbb{R}$ such that F_{θ_0} is the true distribution of $O := (Y, Z, X, D)$. The parametric assumption on the propensity score need not be imposed on the path.[13] The score of the parametric submodel is $S(O) = \frac{\partial \log dF_\theta(O)}{\partial \theta}\big|_{\theta=\theta_0}$. For any θ, we define

$$T_{2,\theta} = E_\theta \left[\frac{\partial m_\theta(y_{\tau,\theta}, \tilde{W}(\alpha_\theta))}{\partial z_1} \right]$$

where m_θ, $y_{\tau,\theta}$, and α_θ are the probability limits of $\hat{m}$, $\hat{y}_\tau$, and $\hat{\alpha}$, respectively, when the distribution of O is F_θ. Note that when $\theta = \theta_0$, we have $T_{2,\theta_0} = T_2$. Suppose the set of scores $\{S(O)\}$ for all parametric submodels $\{F_\theta\}$ can approximate any zero-mean, finite-variance function of O in the mean square sense.[14] If the function $\theta \to T_{2,\theta}$ is differentiable at θ_0 and we can write

$$\frac{\partial T_{2,\theta}}{\partial \theta}\bigg|_{\theta=\theta_0} = E[\Gamma(O)S(O)] \tag{30}$$

for some mean-zero and finite second-moment function $\Gamma(\cdot)$ and any path F_θ, then, by Theorem 2.1 of Newey (1994), the asymptotic variance of $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})$ is $E[\Gamma(O)^2]$.

In the next lemma, we will show that $\theta \to T_{2,\theta}$ is differentiable at θ_0. Suppose, for the moment, this is the case. Then, by the chain rule, we can write

$$\frac{\partial T_{2,\theta}}{\partial \theta}\bigg|_{\theta=\theta_0} = \frac{\partial}{\partial \theta} E_\theta \left[\frac{\partial m_0(y_\tau, \tilde{W}(\alpha_0))}{\partial z_1} \right]\bigg|_{\theta=\theta_0} + \frac{\partial}{\partial \theta} E \left[\frac{\partial m_\theta(y_\tau, \tilde{W}(\alpha_0))}{\partial z_1} \right]\bigg|_{\theta=\theta_0}$$

$$+ \frac{\partial}{\partial \theta} E \left[\frac{\partial m_0(y_{\tau,\theta}, \tilde{W}(\alpha_0))}{\partial z_1} \right]\bigg|_{\theta=\theta_0} + \frac{\partial}{\partial \theta} E \left[\frac{\partial m_0(y_\tau, \tilde{W}(\alpha_\theta))}{\partial z_1} \right]\bigg|_{\theta=\theta_0}.$$

---

[13]As we show later, the error from estimating the propensity score does not affect the asymptotic variance of $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})$.

[14]This is the "generality" requirement of the family of distributions in Newey (1994).

To use Theorem 2.1 of Newey (1994), we need to write all these terms in an outer-product form, namely the form of the right-hand side of (30). To search for the required function $\Gamma(\cdot)$, we follow Newey (1994) and examine one component of $T_{2,\theta}$ at a time by treating the remaining components as known.

The next lemma provides the conditions under which we can ignore the error from estimating the propensity score in our asymptotic analysis. The following notation is used: $Z_{-1} = (Z_2, \ldots, Z_{d_Z})$, $W_{-1} = (W_2, \ldots, W_{d_W}) = (Z_{-1}, X)$ for $d_W = d_X + d_Z$, and $\mathcal{W}_{-1}$ is the support of W_{-1}.

Lemma 7. *Assume that*

(a) *(Z, X) is absolutely continuous with density $f_{ZX}(z, x)$ and*

   (i) *$f_{ZX}(z, x)$ is continuously differentiable with respect to z_1 in $\mathcal{Z} \times \mathcal{X}$,*

   (ii) *for each $w_{-1} \in \mathcal{W}_{-1}$, $f_{Z_1|W_{-1}}(z_1|w_{-1}) = 0$ for any z_1 on the boundary of $\mathcal{Z}_1(w_{-1})$, the support of Z_1 conditional on $W_{-1} = w_{-1}$.*

(b) *$m(y_\tau, \tilde{w})$ is continuously differentiable with respect to z_1 for all orders, and for a neighborhood Θ_0 of θ_0, the following holds:*

$$(i)\ E\left[\sup_{\theta \in \Theta_0}\left|\frac{\partial}{\partial \alpha_\theta}\frac{\partial m_0(y_\tau, \tilde{W}(\alpha_\theta))}{\partial z_1}\right|\right] < \infty, \quad (ii)\ E\sup_{\theta \in \Theta_0}\left|\frac{\partial}{\partial \alpha_\theta}E\left[\frac{\partial \log f_W(W)}{\partial z_1}\left|\tilde{W}(\alpha_\theta)\right.\right]\right| < \infty,$$

$$(iii)\ E\sup_{\theta \in \Theta_0}\left|\frac{\partial}{\partial \alpha_\theta}\left\{m_0(y_\tau, \tilde{W}(\alpha_\theta))E\left[\frac{\partial \log f_W(W)}{\partial z_1}\left|\tilde{W}(\alpha_\theta)\right.\right]\right\}\right| < \infty.$$

Then

$$\frac{\partial}{\partial \theta}E\left[\frac{\partial m_0(y_\tau, \tilde{W}(\alpha_\theta))}{\partial z_1}\right]\Bigg|_{\theta=\theta_0} = 0.$$

The next lemma establishes a stochastic approximation of $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) - T_2$ and provides the influence function as well. The assumptions of the lemma are adapted from Newey (1994). These assumptions are not necessarily the weakest possible.

Lemma 8. *Suppose that*

(a) (i) *the support $\mathcal{W}$ of W is $[w_{1l}, w_{1u}] \times \cdots \times [w_{d_Wl}, w_{d_Wu}]$; (ii) for $w = (w_1, \ldots, w_{d_W})$, $f_W(w)$ is bounded below by $C \times \Pi_{j=1}^{d_W}(w_j - w_{jl})^\kappa (w_{ju} - w_j)^\kappa$ for some $C > 0, \kappa > 0$; (iii) $\int_{\mathcal{W}} \sup_{\theta \in \Theta_0}\left|\frac{\partial}{\partial \theta} f_W(w; \theta)\right| dw < \infty$; (iv) $E\{[\frac{\partial}{\partial z_1} \log f_W(W; \theta)]^2\} < \infty$ for $\theta \in \Theta_0$;*

(b) *there is a constant C such that $\partial^\ell m_0(y_\tau, \tilde{w})/\partial z_1^\ell \leq C^\ell$ for all $\ell \in \mathbb{N}$;*

(c) *the number of series terms, J, satisfies $J(n) = O(n^\varrho)$ for some $\varrho > 0$, and $J^{7+2\kappa} = O(n)$;*

(d) *for each $w_{-1} \in \mathcal{W}_{-1}$, the map $(y, z_1) \mapsto m_0(y, \tilde{w})$ is differentiable, and*

$$E\left[\sup_{y \in \mathcal{Y}}\left|\frac{\partial^2 m_0(y, \tilde{W})}{\partial y \partial z_1}\right|\right] < \infty;$$

(e) the following stochastic equicontinuity conditions hold:

$$E\frac{\partial}{\partial z_1}\left[\hat{m}(y_\tau, \tilde{W}) - m_0(y_\tau, \tilde{W})\right] = \mathbb{P}_n\frac{\partial}{\partial z_1}\left[\hat{m}(y_\tau, \tilde{W}_i) - m_0(y_\tau, \tilde{W}_i)\right] + o_p(n^{-1/2}),$$

$$E\frac{\partial}{\partial z_1}\left[m_0(\hat{y}_\tau, \tilde{W}) - m_0(\hat{y}_\tau, \tilde{W})\right] = \mathbb{P}_n\frac{\partial}{\partial z_1}\left[m_0(\hat{y}_\tau, \tilde{W}_i) - m_0(y_\tau, \tilde{W}_i)\right] + o_p(n^{-1/2});$$

(f) the assumptions of Lemma 7 hold.

Then, we have the decomposition

$$
\begin{aligned}
T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) - T_2 &= T_{2n}(y_\tau, m_0, \alpha_0) - T_2 + T_{2n}(y_\tau, \hat{m}, \alpha_0) - T_{2n}(y_\tau, m_0, \alpha_0) \\
&\quad + T_{2n}(\hat{y}_\tau, m_0, \alpha_0) - T_{2n}(y_\tau, m_0, \alpha_0) + o_p(n^{-1/2}) \\
&= \mathbb{P}_n\psi_{\partial m_0} + \mathbb{P}_n\psi_{m_0} + \mathbb{P}_n\tilde{\psi}_Q(y_\tau) + o_p(n^{-1/2}),
\end{aligned}
\tag{31}
$$

where

$$\psi_{\partial m_0} := \frac{\partial m_0(y_\tau, \tilde{W})}{\partial z_1} - T_2,$$

$$\psi_{m_0} := -\left[\mathbf{1}\left\{Y \leq y_\tau\right\} - m_0(y_\tau, \tilde{W}\right] \times E\left[\frac{\partial \log f_W(W)}{\partial z_1}\middle| \tilde{W}\right],$$

and

$$\tilde{\psi}_Q(y_\tau) := E\left[\frac{\partial f_{Y|\tilde{W}}(y_\tau|\tilde{W})}{\partial z_1}\right]\left[\frac{\tau - \mathbf{1}\left\{Y \leq y_\tau\right\}}{f_Y(y_\tau)}\right].$$

Lemma 8 characterizes the contribution of each stage to the influence function of $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})$. The contribution from estimating m_0, given by $\mathbb{P}_n\psi_{m_0}$, corresponds to the one in Proposition 5 of Newey (1994) (p. 1362).

5.3 Estimating the UQE

With $\hat{y}_\tau, \hat{f}_Y, \hat{m}, \hat{\alpha}$ given in the previous subsections, we estimate the UQE by

$$\hat{\Pi}_\tau(\hat{y}_\tau, \hat{f}_Y, \hat{m}, \hat{\alpha}) = -\frac{1}{\hat{f}_Y(\hat{y}_\tau)}\frac{T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})}{T_{1n}(\hat{\alpha})}.\tag{32}$$

This is our unconditional instrumental quantile estimator (UNIQUE). With the asymptotic linear representations of all three components $\hat{f}_Y(\hat{y}_\tau)$, $T_{1n}(\hat{\alpha})$, and $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})$, we can obtain the asymptotic linear representation of $\hat{\Pi}_\tau(\hat{y}_\tau, \hat{f}_Y, \hat{m}, \hat{\alpha})$. The next theorem follows from combining Lemmas 5, 6, 7, and 8.

Theorem 2. *Under the assumptions of Lemmas 5, 6, 7, and 8, we have*

$$
\begin{aligned}
\hat{\Pi}_\tau - \Pi_\tau &= \frac{T_2}{f_Y(y_\tau)^2 T_1} \left[\mathbb{P}_n \psi_{f_Y}(y_\tau) + B_{f_Y}(y_\tau) \right] + \frac{T_2}{f_Y(y_\tau)^2 T_1} f_Y'(y_\tau) \mathbb{P}_n \psi_Q(y_\tau) \\
&+ \frac{T_2}{f_Y(y_\tau) T_1^2} \mathbb{P}_n \psi_{\partial P} + \frac{T_2}{f_Y(y_\tau) T_1^2} E \left[\frac{\partial^2 P(Z, X, \alpha_0)}{\partial z_1 \partial \alpha_0'} \right] \mathbb{P}_n \psi_{\alpha_0} \\
&- \frac{1}{f_Y(y_\tau) T_1} \mathbb{P}_n \psi_{\partial m_0} - \frac{1}{f_Y(y_\tau) T_1} \mathbb{P}_n \psi_{m_0} - \frac{1}{f_Y(y_\tau) T_1} \mathbb{P}_n \tilde{\psi}_Q(y_\tau) + R_\Pi,
\end{aligned}
\tag{33}
$$

where

$$
\begin{aligned}
R_\Pi &= O_p \left(|\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)|^2 \right) + O_p \left(n^{-1} \right) + O_p \left(n^{-1/2} |\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)| \right) \\
&+ O_p \left(|R_{f_Y}| \right) + o_p(n^{-1/2}) + o_p(h^2).
\end{aligned}
$$

Furthermore, under Assumption 7, $\sqrt{nh} R_\Pi = o_p(1)$.

Equation (33) consists of six influence functions and a bias term. The bias term $B_{f_Y}(y_\tau)$ arises from estimating the density and is of order $O(h^2)$. The six influence functions reflect the impact of each estimation stage. The rate of convergence of $\hat{\Pi}_\tau$ is slowed down through $\mathbb{P}_n \psi_{f_Y}(y_\tau)$, which is of order $O_p(n^{-1/2} h^{-1/2})$. We can summarize the results of Theorem 2 in a single equation:

$$
\hat{\Pi}_\tau - \Pi_\tau = \mathbb{P}_n \psi_{\Pi_\tau} + \tilde{B}_{f_Y}(y_\tau) + o_p(n^{-1/2} h^{-1/2}),
$$

where ψ_{Π_τ} collects all the influence functions in (33) except for the bias, and

$$
\tilde{B}_{f_Y}(y_\tau) := \frac{T_2}{f_Y(y_\tau)^2 T_1} B_{f_Y}(y_\tau).
$$

If $nh^5 \to 0$, then the bias term is $o(n^{-1/2} h^{-1/2})$. The following corollary provides the asymptotic distribution of $\hat{\Pi}_\tau$.

Corollary 5. *Under the assumptions of Theorem 2 and the assumption that $nh^5 \to 0$,*

$$
\sqrt{nh} \left(\hat{\Pi}_\tau - \Pi_\tau \right) = \sqrt{n} \mathbb{P}_n \sqrt{h} \psi_{\Pi_\tau} + o_p(1) \Rightarrow \mathcal{N}(0, V_\tau),
$$

where

$$
V_\tau = \lim_{h \downarrow 0} E \left[h \psi_{\Pi_\tau}^2 \right].
\tag{34}
$$

From the perspective of asymptotic theory, all of the following terms are all of order $O_p(h) = o_p(1)$ and hence can be ignored in large samples: $\sqrt{nh} \mathbb{P}_n \psi_Q(y_\tau)$, $\sqrt{nh} \mathbb{P}_n \psi_{\partial P}$, $\sqrt{nh} \mathbb{P}_n \psi_{\alpha_0}$, $\sqrt{nh} \mathbb{P}_n \psi_{\partial m_0}$, and $\sqrt{nh} \mathbb{P}_n \psi_{m_0}$. The asymptotic variance is then given by

$$
V_\tau = \frac{T_2^2}{f_Y(y_\tau)^4 T_1^2} \lim_{h \downarrow 0} E \left[h \psi_f^2(y_\tau, h) \right] = \frac{T_2^2}{f_Y(y_\tau)^3 T_1^2} \int_{-\infty}^{\infty} K^2(u) \, du.
$$

However, V_τ ignores all estimation uncertainties except that in $\hat{f}_Y(y_\tau)$, and we do not expect it to reflect the finite-sample variability of $\sqrt{nh}(\hat{\Pi}_\tau - \Pi_\tau)$ well. To improve the finite-sample perfor-

mances, we keep the dominating term from each source of estimation errors and employ a sample counterpart of $Eh\psi_{\Pi_\tau}^2$ to estimate V_τ. The details can be found in Section C of the Appendix.

5.4 Testing the Null of No Effect

We can use Corollary 5 for hypothesis testing on Π_τ. Since $\hat{\Pi}_\tau$ converges to Π_τ at a nonparametric rate, in general, the test will have power only against a local departure of a nonparametric rate. However, if we are interested in testing the null of a zero effect, that is, $H_0 : \Pi_\tau = 0$ vs. $H_1 : \Pi_\tau \neq 0$, we can detect a parametric rate of departure from the null. The reason is that, by (26), $\Pi_\tau = 0$ if and only if $T_2 = 0$, and T_2 can be estimated at the usual parametric rate. Hence, instead of testing $H_0 : \Pi_\tau = 0$ vs. $H_1 : \Pi_\tau \neq 0$, we can test the equivalent hypotheses $H_0 : T_2 = 0$ vs. $H_1 : T_2 \neq 0$.

Our test is based on the estimator $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})$ of T_2. In view of its influence function given in Lemma 8, we can estimate the asymptotic variance of $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})$ by

$$\hat{V}_2 = \frac{1}{n}\sum_{i=1}^n \left(\hat{\psi}_{\partial m,i} + \hat{\psi}_{m,i} + \hat{E}\left[\frac{\partial f_{Y|\tilde{W}(\alpha_0)}(\hat{y}_\tau|\tilde{W}(\hat{\alpha}))}{\partial z_1} \right] \hat{\psi}_{Q,i}(\hat{y}_\tau) \right)^2.$$

We can then form the test statistic:

$$T_{2n}^o := \sqrt{n}\frac{T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})}{\sqrt{\hat{V}_2}}. \tag{35}$$

By Lemma 8 and using standard arguments, we can show that $T_{2n}^o \Rightarrow \mathcal{N}(0,1)$. To save space, we omit the details here.

6 Simulation Evidence

For our simulation, we consider the following model: $Y(0) = U_0$, $Y(1) = \beta + U_1$, and $D = \mathbb{1}\{V \leq Z\}$ where $Z \in \mathbb{R}$. By Corollary 4, and assuming that $f_{U_0|V} = f_{U_1|V}$, we have[15]

$$\Pi_\tau = \frac{1}{f_Y(y_\tau)} \int_{\mathcal{Z}} \left[\int_{y_\tau-\beta}^{y_\tau} f_{U_0|V}(u|z)du \right] \tilde{f}_Z(z)dz, \tag{36}$$

where

$$\tilde{f}_Z(z) = \frac{f_V(z)f_Z(z)}{\int_{\mathcal{Z}} f_V(\check{z})f_Z(\check{z})d\check{z}}.$$

In order to compute Π_τ numerically, we assume that Z is standard normal and independent of (U_0, U_1, V), which is jointly normal with mean 0 and variance-covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}.$$

---

[15]The details can be found in Section B of the Appendix.

Here, ρ is the correlation between U_0 and V, and between U_1 and V. It is also the parameter that governs the endogeneity of D.

Estimation of Π_τ requires estimating y_τ, f_Y, T_1, and T_2. The quantiles are estimated in the usual way. To estimate f_Y, we use a Gaussian Kernel with bandwidth $h = 1.06 \times \hat{\sigma}_Y \times n^{-1/5}$, where $\hat{\sigma}_Y$ is the sample standard deviation of Y. This is Silverman's rule of thumb. To estimate T_1, we use a probit model. To estimate T_2, we run a cubic series regression.

6.1 Testing the Null Hypothesis of No Effect

If we set $\beta = 0$, and by (36), $\Pi_\tau = 0$ and so the null hypothesis of a zero effect holds. The test statistic T_{2n}^o is constructed following equation (35). Because the test statistic does not involve estimating the density (and also T_1), the test has nontrivial power again $1/\sqrt{n}$-departures (i.e., $\beta = c/\sqrt{n}$ for some $c \neq 0$) from the null.

To simulate the power function of the nominal 5% test, we consider a range of 25 values of β between -1 and 1. The endogeneity, governed by the parameter ρ, takes five values: 0, 0.25, 0.5, 0.75, and 0.9. We perform 1,000 simulations with 1,000 observations. For different values of τ, the power functions are shown below in Figure 3. The test has the desired level, except for the extreme quantile $\tau = 0.1$, where under high endogeneity, the rejection probability does not increase fast enough. Omitted is the power function for the median, $\tau = 0.5$, which is basically indistinguishable from that of $\tau = 0.4$. Furthermore, simulation results not reported here show that the power functions for $\tau = 0.6, 0.7, 0.8, 0.9$ are very similar to those of $\tau = 0.4, 0.3, 0.2, 0.1$, respectively.

6.2 Empirical Coverage of Confidence Intervals

In this subsection we investigate the empirical coverage of confidence intervals built using $\hat{V}_\tau$, the variance estimator given in (A.34). Since $\sqrt{nh}\left(\hat{\Pi}_\tau - \Pi_\tau\right) \approx \mathcal{N}(0, \hat{V}_\tau)$, a 95% confidence interval for Π_τ can be constructed using

$$\hat{\Pi}_\tau \pm 1.96 \times \frac{\sqrt{\hat{V}_\tau}}{\sqrt{nh}}.$$

We use a grid of β that takes values $-1, -0.5, -0.25, 0, 0.25, 0.5$, and 1. For the endogeneity parameter, ρ, we take the values $0, 0.25, 0.5, 0.75$, and 0.9. Finally, τ takes the values from 0.1 to 0.9 with an increment of 0.1. We note that, for values of $\beta \neq 0$, where the effect is not 0, we need to numerically compute the value of Π_τ. We perform 1,000 simulations with 1,000 observations each. The results are reported in the tables below for $\tau = 0.1$ and $\tau = 0.5$. It is clear that the confidence intervals have reasonable coverage accuracy in almost all cases.

7 Empirical Application

We estimate the unconditional quantile effect of expanding college enrollment on (log) wages. The outcome variable Y is the log wage, and the binary treatment is the college enrollment status. Thus $p = \Pr(D = 1)$ is the proportion of individuals who ever enrolled in a college. Arguably, the cost
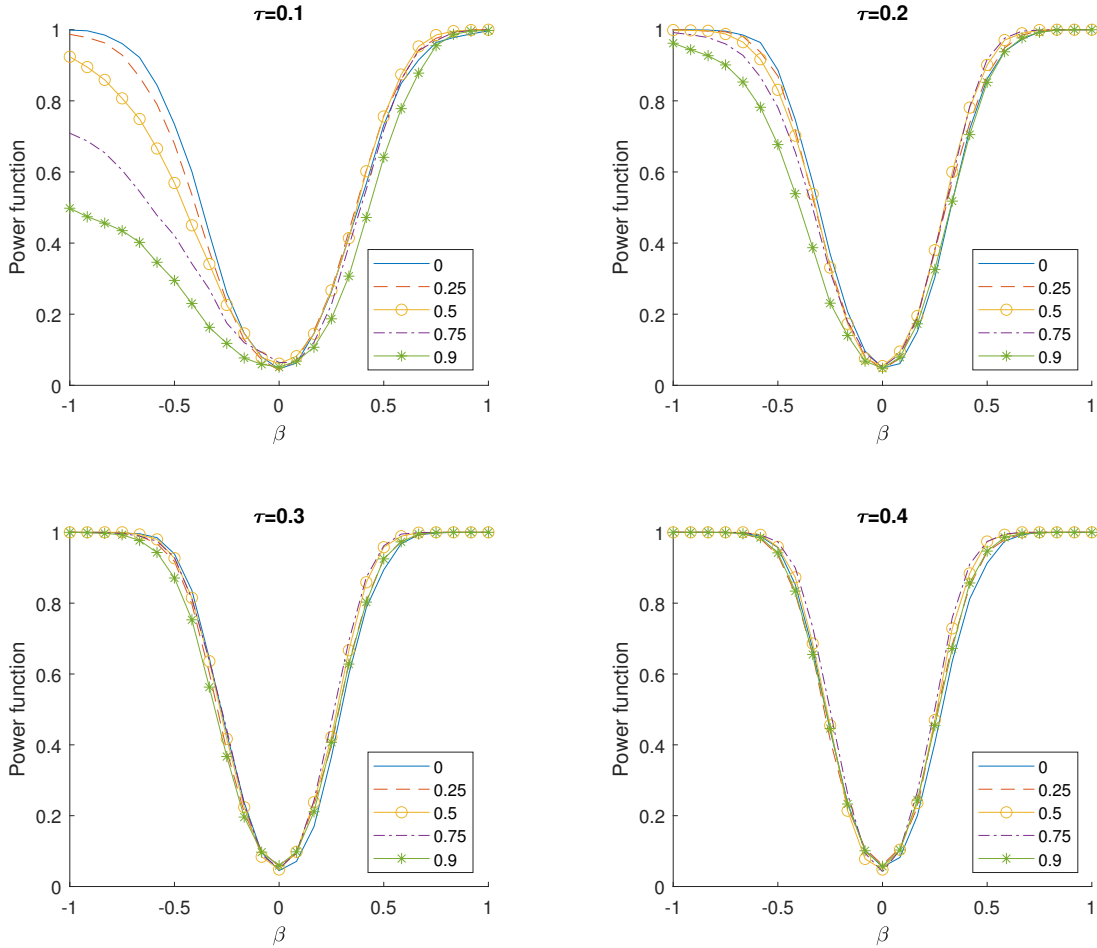
Figure 3: Power functions for different values of ρ and τ.

of tuition (Z_1) is an important factor that affects the college enrollment status but not the wage. In order to alter the proportion of enrolled individuals, we consider a policy that subsidizes tuition by a certain amount. The UQE is the effect of this policy on the different quantiles of the unconditional distribution of wages when the subsidy is small. This policy shifts Z_1, the tuition, to $Z_{1\delta} = Z_1 + s(\delta)$ for some $s(\delta)$, which is the same for all individuals, and induces the college enrollment to increase from p to $p + \delta$. Note that we do not need to specify $s(\delta)$ because we look at the limiting version as $\delta \to 0$. In practice, we may set $s(\delta)$ equal to a small percentage of the total tuition, say 1%.

We use the same data as in Carneiro, Heckman and Vytlacil (2010) and Carneiro, Heckman and Vytlacil (2011): a sample of white males from the 1979 National Longitudinal Survey of Youth (NLSY1979). The web appendix to Carneiro, Heckman and Vytlacil (2011) contains a detailed description of the variables. The outcome variable Y is the log wage in 1991. The treatment indicator D is equal to 1 if the individual ever enrolled in college by 1991, and 0 otherwise. The other covariates are AFQT score, mother's education, number of siblings, average log earnings 1979–2000 in the county of residence at age 17, average unemployment 1979–2000 in the state of residence at age

33

$\beta \backslash \rho$	0	0.25	0.5	0.75	0.9
-1	0.978	0.982	0.974	0.973	0.948
-0.5	0.955	0.958	0.969	0.977	0.947
-0.25	0.962	0.961	0.969	0.957	0.969
0	0.952	0.961	0.958	0.949	0.964
0.25	0.958	0.962	0.966	0.965	0.958
0.5	0.967	0.969	0.972	0.965	0.963
1	0.979	0.974	0.974	0.979	0.97

Table 1: Empirical coverage of 95% confidence intervals for $\tau = 0.1$.

$\beta \backslash \rho$	0	0.25	0.5	0.75	0.9
-1	0.94	0.911	0.916	0.91	0.692
-0.5	0.939	0.942	0.944	0.944	0.921
-0.25	0.955	0.941	0.943	0.946	0.955
0	0.941	0.957	0.959	0.945	0.953
0.25	0.949	0.942	0.942	0.967	0.954
0.5	0.946	0.942	0.941	0.957	0.951
1	0.905	0.934	0.922	0.921	0.955

Table 2: Empirical coverage of 95% confidence intervals for $\tau = 0.5$.

17, urban residence dummy at age 14, cohort dummies, years of experience in 1991, average local log earnings in 1991, and local unemployment in 1991. We collect these variables into a vector and denote it by X.

We assume that the following four variables (denoted by Z_1, Z_2, Z_3, Z_4) enter the selection equation but not the outcome equation: tuition at local public four-year colleges at age 17, presence of a four-year college in the county of residence at age 14, local earnings at age 17, and local unemployment at age 17. The total sample size is 1747, of which 882 individuals had never enrolled in a college ($D = 0$) by 1991, and 865 individuals had enrolled in a college by 1991 ($D = 1$). We compute the UQE of a marginal shift in the tuition at local public four-year colleges at age 17.

To estimate the propensity score, we use a parametric logistic specification. To estimate the conditional expectation function m_0, we use a series regression using both the estimated propensity score and the covariates X as the regressors. Due to the large number of variables involved, a penalization of $\lambda = 10^{-4}$ was imposed on the L_2-norm of the coefficients, excluding the constant term as in ridge regressions. We compute the UQE at the quantile level $\tau = 0.1, 0.15, \ldots, 0.9$. For each τ, we also construct the 95% (pointwise) confidence interval.

Figure 4 presents the results. The UQE ranges between 0.22 and 0.47 across the quantiles with an average of 0.37. When we estimate the unconditional mean effect, we obtain an estimate of 0.21, which is somewhat consistent with the quantile cases. We interpret these estimates in the following way: the effect of a δ (small) increase in college enrollment induced by an additive change in tuition increases (log) wages between $0.22 \times \delta$ and $0.47 \times \delta$ across quantiles. For example, for $\delta = 0.01$, we obtain an increase in the quantiles of the wage distribution between of 0.22% and 0.47%.
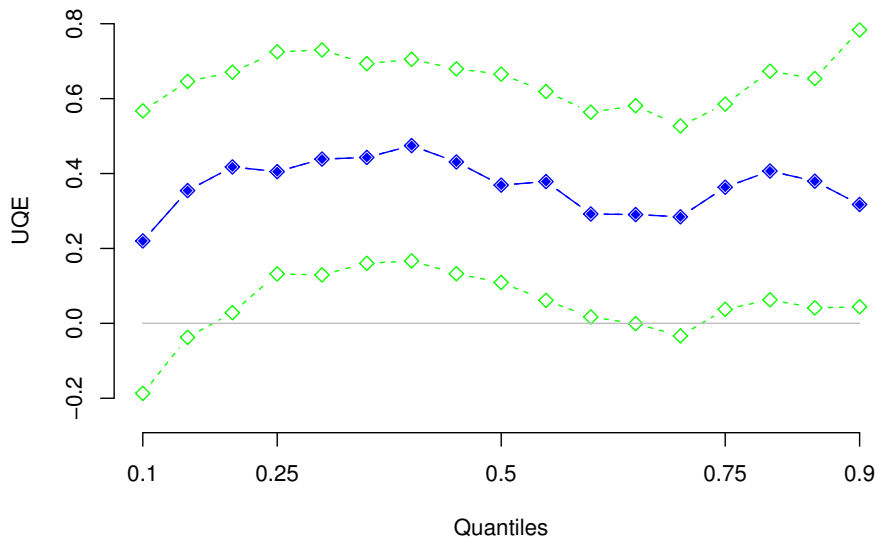
Figure 4: *Solid line: UQE; Dashed line: 95% confidence intervals.*

8   Conclusion

In this paper we study the unconditional policy effect with an endogenous binary treatment. Framing the selection equation as a threshold-crossing model allows us to introduce a novel class of unconditional marginal treatment effects and represent the unconditional effect as a weighted average of these unconditional marginal treatment effects. When the policy variable used to change the participation rate satisfies a conditional exogeneity condition, it is possible to recover the unconditional policy effect using the proposed UNIQUE method.

To illustrate the usefulness of the unconditional MTEs, we focus on the unconditional quantile effect. We find that the unconditional quantile regression estimator that neglects endogeneity can be severely biased. The bias may not be uniform across quantiles. Any attempt to sign the bias *a priori* requires very strong assumptions on the data generating process. More intriguingly, the unconditional quantile regression estimator can be inconsistent even if the treatment status is exogenously determined. This happens when the treatment selection is partly determined by some covariates that also influence the outcome variable.

We find that the unconditional quantile effect and the marginal policy-relevant treatment effect can be seen as part of the same family of policy effects. It is possible to view the latter as a robust version of the former. Both of them are examples of a general unconditional policy effect. To the best of our knowledge, this connection has not been established in either literature.

References

Bjorklund, Anders, and Robert Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in

Self-Selection." *The Review of Economics and Statistics*, 69(1): 42–49.

Carneiro, Pedro, and Sokbae Lee. 2009. "Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality." *Journal of Econometrics*, 149(2): 191–208.

Carneiro, Pedro, James J. Heckman, and Edward Vytlacil. 2010. "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin." *Econometrica*, 78(1): 377–394.

Carneiro, Pedro, James J. Heckman, and Edward Vytlacil. 2011. "Estimating Marginal Returns to Education." *American Economic Review*, 101(6): 2754–2781.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2007. "Unconditional Quantile Regressions." *NBER Technical Working Paper 339*.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2009. "Unconditional Quantile Regressions." *Econometrica*, 77(3): 953–973.

Hahn, Jinyong, and Geert Ridder. 2013. "Asymptotic Variance of Semiparametric Estimators with Generated Regressors." *Econometrica*, 81(1): 315–340.

Heckman, James J., and Edward Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences of the United States of America*, 96(8): 4730–4734.

Heckman, James J., and Edward Vytlacil. 2001*a*. "Local Instrumental Variables." In *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*. , ed. C. Hsiao, K. Morimune and J. Powell, 1–46. Cambridge, UK:Cambrigde University Press.

Heckman, James J., and Edward Vytlacil. 2001*b*. "Policy Relevant Treatment Effects." *American Economic Review*, 91(2): 107–111.

Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics*, 88(3): 389–432.

Ibragimov, I. A., and R. Z. Hasminskii. 1981. *Statistical Estimation: Asymptotic Theory.* Springer-Verlag.

Imbens, Guido W., and Whitney K. Newey. 2009. "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity." *Econometrica*, 77(5): 1481–1512.

Inoue, Atsushi, Tong Li, and Qi Xu. 2021. "Two Sample Unconditional Quantile Effect." ARXIV: https://arxiv.org/pdf/2105.09445.pdf.

Kasy, Maximilian. 2016. "Partial Identification, Distributional Preferences, and The Welfare Ranking of Policies." *The Review of Economics and Statistics*, 98(March): 111–131.

Martinez-Iriarte, Julian. 2020. "Sensitivity Analysis in Unconditional Quantile Effects." Working Paper.

Martinez-Iriarte, Julian, Gabriel Montes-Rojas, and Yixiao Sun. 2022. "Location-Scale and Compensated Effects in Unconditional Quantile Regressions." Working Paper.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters. 2020. "Policy Evaluation with Multiple Instrumental Variables." Working Paper.

Mogstad, Magne, and Alexander Torgovitsky. 2018. "Identification and Extrapolation of Causal Effects with Instrumental Variables." *Annual Review of Economics*, 10: 577–613.

Mukhin, Yaroslav. 2019. "On Counterfactual Analysis of Differentiable Functionals." Working Paper.

Newey, Whitney K. 1994. "The Asymptotic Variance of Semiparametric Estimators." *Econometrica*, 62(6): 1349–1382.

Pagan, Adrian, and Aman Ullah. 1999. *Nonparametric Econometrics.* Cambridge University Press.

Rothe, Christoph. 2010. "Identification of Unconditional Partial Effects in Nonseparable Models." *Economics Letters*, 109(3): 171–174.

Rothe, Christoph. 2012. "Partial Distributional Policy Effects." *Econometrica*, 80(5): 2269–2301.

Rudin, Walter. 1976. *Principles of Mathematical Analysis.* New York: McGraw-Hill.

Sasaki, Y., T. Ura, and Y. Zhang. 2020. "Unconditional Quantile Regression with High Dimensional Data." Working Paper.

Sasaki, Yuya, and Takuya Ura. 2021. "Estimation and Inference for Policy Relevant Treatment Effects." *Journal of Econometrics* (Forthcoming).

Serfling, Robert J. 1980. *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

van der Vaart, A. W. 2000. *Asymptotic Statistics.* Cambrigde University Press.

Yu, Ping. 2014. "Marginal Quantile Treatment Effect." Working Paper.

Zhou, Xiang, and Yu Xie. 2019. "Marginal Treatment Effects from a Propensity Score Perspective." *Journal of Political Economy*, 127(6): 3070–3084.

Appendix

A Additional Proofs

Proof of Lemma 1. Using the selection equation $D_\delta = \mathbb{1}\{U_D \le P_\delta(W)\}$, we have

$$
\begin{aligned}
F_{Y(1)|D_\delta}(\tilde{y}|1) &= \Pr\left(Y(1) \le \tilde{y}|D_\delta = 1\right) \\
&= \frac{1}{\Pr(D_\delta = 1)} \Pr\left(Y(1) \le \tilde{y}, D_\delta = 1\right) \\
&= \frac{1}{\Pr(D_\delta = 1)} \int_{\mathcal{W}} \Pr\left(Y(1) \le \tilde{y}, D_\delta = 1|W = w\right) f_W(w)\, dw \\
&= \frac{1}{\Pr(D_\delta = 1)} \int_{\mathcal{W}} \Pr\left(Y(1) \le \tilde{y}, U_D \le P_\delta(w)|W = w\right) f_W(w)\, dw.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
F_{Y(1)|D_\delta}(\tilde{y}|1) &= \frac{1}{p+\delta} \int_{\mathcal{W}} F_{Y(1), U_D|W}(\tilde{y}, P_\delta(w)|w) f_W(w)\, dw \\
&= \frac{1}{p+\delta} \int_{\mathcal{W}} \int_{-\infty}^{\tilde{y}} \int_{-\infty}^{P_\delta(w)} f_{Y(1), U_D|W}(y, u|w) f_W(w)\, du\, dy\, dw \\
&= \frac{1}{p+\delta} \int_{-\infty}^{\tilde{y}} \int_{\mathcal{W}} \left[\int_{-\infty}^{P_\delta(w)} f_{Y(1), U_D|W}(y, u|w)\, du\right] f_W(w)\, dw\, dy,
\end{aligned}
$$

where the order of integration can be switched because the integrands are non-negative. It then follows that

$$
\begin{aligned}
f_{Y(1)|D_\delta}(\tilde{y}|1) &= \frac{1}{p+\delta} \int_{\mathcal{W}} \left[\int_{-\infty}^{P_\delta(w)} f_{Y(1), U_D|W}(\tilde{y}, u|w)\, du\right] f_W(w)\, dw \\
&= \frac{1}{p+\delta} \int_{\mathcal{W}} \left[\int_{-\infty}^{P_\delta(w)} f_{Y(1)|U_D,W}(\tilde{y}|u, w)\, du\right] f_W(w)\, dw. \tag{A.1}
\end{aligned}
$$

So,

$$
(p+\delta)\, f_{Y(1)|D_\delta}(\tilde{y}|1) = \int_{\mathcal{W}} \left[\int_{-\infty}^{P_\delta(w)} f_{Y(1)|U_D,W}(\tilde{y}|u, w)\, du\right] f_W(w)\, dw. \tag{A.2}
$$

Under Assumptions 2(b) and 2(c), we can differentiate both sides of (A.2) with respect to δ under the integral sign to get

$$
\frac{\partial\left[(p+\delta)\, f_{Y(1)|D_\delta}(\tilde{y}|1)\right]}{\partial\delta} = \int_{\mathcal{W}} f_{Y(1)|U_D,W}(\tilde{y}|P_\delta(w), w) \frac{\partial P_\delta(w)}{\partial\delta} f_W(w)\, dw. \tag{A.3}
$$

Under Assumptions 2(b.i) and 2(c.ii), $f_{Y(1)|U_D,W}(\tilde{y}|P_\delta(w), w)\partial P_\delta(w)/\partial\delta$ is continuous in δ for each $\tilde{y} \in \mathcal{Y}(1)$ and $w \in \mathcal{W}$. In view of Assumptions 2(b.ii) and 2(c.iii), we can invoke the dominated convergence theorem to show that the map $\delta \mapsto \frac{\partial\left[(p+\delta) f_{Y(1)|D_\delta}(\tilde{y}|1)\right]}{\partial\delta}$ is continuous for each $\tilde{y} \in \mathcal{Y}(1)$.

For the case of $(1 - p - \delta) f_{Y(0)|D_\delta}(\tilde{y}|0)$, we have

$$\frac{\partial f_{Y(0)|D_\delta}(\tilde{y}|0)}{\partial \delta} = \frac{\partial}{\partial \delta} \frac{\partial F_{Y(0)|D_\delta}(\tilde{y}|0)}{\partial \tilde{y}}.$$

Using the selection equation, we can write $F_{Y(0)|D_\delta}(\tilde{y}|0)$ as

$$
\begin{aligned}
F_{Y(0)|D_\delta}(\tilde{y}|0) &= \Pr(Y(0) \le \tilde{y}|D_\delta = 0) \\
&= \Pr(Y(0) \le \tilde{y}|U_D > P_\delta(W))) \\
&= \frac{1}{\Pr[U_D > P_\delta(W)]} \Pr(Y(0) \le \tilde{y}, U_D > P_\delta(W)) \\
&= \frac{1}{1 - p - \delta} [\Pr(Y(0) \le \tilde{y}) - \Pr(Y(0) \le \tilde{y}, U_D \le P_\delta(W))] \\
&= \frac{1}{1 - p - \delta} \left[F_{Y(0)}(\tilde{y}) - \int_{\mathcal{W}} F_{Y(0),U_D|W}(\tilde{y}, P_\delta(w)|w) f_W(w) dw \right] \\
&= \frac{1}{1 - p - \delta} \left[F_{Y(0)}(\tilde{y}) - \int_{\mathcal{W}} \int_{-\infty}^{\tilde{y}} \int_{-\infty}^{P_\delta(w)} f_{Y(0),U_D|W}(y, u|w) f_W(w) du dy dw \right] \\
&= \frac{1}{1 - p - \delta} \left[F_{Y(0)}(\tilde{y}) - \int_{-\infty}^{\tilde{y}} \int_{\mathcal{W}} \int_{-\infty}^{P_\delta(w)} f_{Y(0),U_D|W}(y, u|w) f_W(w) du dw dy \right],
\end{aligned}
$$

where the orders of integrations can be switched because the integrands are non-negative. Therefore,

$$(1 - p - \delta) f_{Y(0)|D_\delta}(\tilde{y}|0) = f_{Y(0)}(\tilde{y}) - \int_{\mathcal{W}} \int_{-\infty}^{P_\delta(w)} f_{Y(0),U_D|W}(\tilde{y}, u|w) f_W(w) du dw. \tag{A.4}$$

Using Assumptions 2(b) and 2(c), we have

$$
\begin{aligned}
\frac{\partial \left[(1 - p - \delta) f_{Y(0)|D_\delta}(\tilde{y}|0) \right]}{\partial \delta} &= -\int_{\mathcal{W}} f_{Y(0),U_D|W}(\tilde{y}, P_\delta(w))|w) \frac{\partial P_\delta(w)}{\partial \delta} f_W(w) dw \\
&= -\int_{\mathcal{W}} f_{Y(0)|U_D,W}(\tilde{y}|P_\delta(w), w) \frac{\partial P_\delta(w)}{\partial \delta} f_W(w) dw. \tag{A.5}
\end{aligned}
$$

The continuity of $\delta \mapsto \frac{\partial \left[(1-p-\delta) f_{Y(0)|D_\delta}(\tilde{y}|0) \right]}{\partial \delta}$ follows from the same arguments for the continuity of $\delta \mapsto \frac{\partial \left[(p+\delta) f_{Y(1)|D_\delta}(\tilde{y}|1) \right]}{\partial \delta}$. ∎

Proof of Lemma 2. For any δ in N_ε, we have

$$F_{Y_\delta}(y) = \int_{\mathcal{Y}(1)} \mathbf{1}\{\tilde{y} \le y\}(p+\delta) f_{Y(1)|D_\delta}(\tilde{y}|1) d\tilde{y} + \int_{\mathcal{Y}(0)} \mathbf{1}\{\tilde{y} \le y\}(1-p-\delta) f_{Y(0)|D_\delta}(\tilde{y}|0) d\tilde{y}.$$

We proceed to take the first order Taylor expansion of $\delta \mapsto (p + \delta) f_{Y(1)|D_\delta}$ and $\delta \mapsto (1 - p -$

2

$\delta)f_{Y(0)|D_\delta}$ around $\delta = 0$, which is possible by Lemma 1. Using (A.3), we have

$$(p+\delta)f_{Y(1)|D_\delta}(\tilde{y}|1)$$
$$= pf_{Y(1)|D}(\tilde{y}|1) + \delta \cdot \left[\int_{\mathcal{W}} f_{Y(1)|U_D,W}(\tilde{y}|P(w),w)\dot{P}(w)f_W(w)dw\right] + R(\delta;\tilde{y},1), \quad \text{(A.6)}$$

where

$$R(\delta;\tilde{y},1) := \delta \cdot \left[\frac{\partial\left[(p+\delta)f_{Y(1)|D_\delta}(\tilde{y}|1)\right]}{\partial\delta}\Bigg|_{\delta=\tilde{\delta}_1} - \frac{\partial\left[(p+\delta)f_{Y(1)|D_\delta}(\tilde{y}|1)\right]}{\partial\delta}\Bigg|_{\delta=0}\right] \quad \text{(A.7)}$$

and $0 \leq \tilde{\delta}_1 \leq \delta$. The middle point $\tilde{\delta}_1$ depends on δ. For the case of $d = 0$, we have a similar expansion:

$$(1-p-\delta)f_{Y(0)|D_\delta}(\tilde{y}|0) = (1-p)f_{Y(0)|D}(\tilde{y}|0) \quad \text{(A.8)}$$
$$- \delta \cdot \left[\int_{\mathcal{W}} f_{Y(0)|U_D,W}(\tilde{y}|P(w),w)\dot{P}(w)f_W(w)dw\right] + R(\delta;\tilde{y},0),$$

where

$$R(\delta;\tilde{y},0) := \delta \cdot \left[\frac{\partial\left[(1-p-\delta)f_{Y(0)|D_\delta}(\tilde{y}|0)\right]}{\partial\delta}\Bigg|_{\delta=\tilde{\delta}_0} - \frac{\partial\left[(1-p-\delta)f_{Y(0)|D_\delta}(\tilde{y}|0)\right]}{\partial\delta}\Bigg|_{\delta=0}\right] \quad \text{(A.9)}$$

and $0 \leq \tilde{\delta}_0 \leq \delta$. The middle point $\tilde{\delta}_0$ depends on δ.

Hence

$$F_{Y_\delta}(y) = F_Y(y) + \delta\int_{\mathcal{Y}(1)}\int_{\mathcal{W}} \mathbb{1}\{\tilde{y} \leq y\} f_{Y(1)|U_D,W}(\tilde{y}|P(w),w)\dot{P}(w)f_W(w)dwd\tilde{y}$$
$$- \delta\int_{\mathcal{Y}(0)}\int_{\mathcal{W}} \mathbb{1}\{\tilde{y} \leq y\} f_{Y(0)|U_D,W}(\tilde{y}|P(w),w)\dot{P}(w)f_W(w)dwd\tilde{y} + R_F(\delta;y) \quad \text{(A.10)}$$

where the remainder $R_F(\delta;y)$ is

$$R_F(\delta;y) := \int_{\mathcal{Y}(1)} \mathbb{1}\{\tilde{y} \leq y\}R(\delta;\tilde{y},1)d\tilde{y} + \int_{\mathcal{Y}(0)} \mathbb{1}\{\tilde{y} \leq y\}R(\delta;\tilde{y},0)d\tilde{y}. \quad \text{(A.11)}$$

The next step is to show that the remainder in (A.11) is $o(|\delta|)$ uniformly over $y \in \mathcal{Y} = \mathcal{Y}(0) \cup \mathcal{Y}(1)$ as $\delta \to 0$, that is, $\lim_{\delta\to 0}\sup_{y\in\mathcal{Y}}\left|\frac{R_F(\delta;y)}{\delta}\right| = 0$. Using (A.7) and (A.9), we get

$$\sup_{y\in\mathcal{Y}}\left|\frac{R_F(\delta;y)}{\delta}\right| \leq \int_{\mathcal{Y}(1)}\left|\frac{\partial\left[(p+\delta)f_{Y(1)|D_\delta}(\tilde{y}|1)\right]}{\partial\delta}\Bigg|_{\delta=\tilde{\delta}_1} - \frac{\partial\left[(p+\delta)f_{Y(1)|D_\delta}(\tilde{y}|1)\right]}{\partial\delta}\Bigg|_{\delta=0}\right|d\tilde{y}$$

$$\int_{\mathcal{Y}(0)}\left|\frac{\partial\left[(1-p-\delta)f_{Y(0)|D_\delta}(\tilde{y}|0)\right]}{\partial\delta}\Bigg|_{\delta=\tilde{\delta}_0} - \frac{\partial\left[(1-p-\delta)f_{Y(0)|D_\delta}(\tilde{y}|0)\right]}{\partial\delta}\Bigg|_{\delta=0}\right|d\tilde{y}.$$

Assumption 3 allows us to take the limit $\delta \to 0$ under the integral signs. Also, by Lemma 1, both

3

$\frac{\partial\left[(p+\delta)f_{Y(1)|D_\delta}(\tilde{y}|1)\right]}{\partial\delta}$ and $\frac{\partial\left[(1-p-\delta)f_{Y(0)|D_\delta}(\tilde{y}|0)\right]}{\partial\delta}$ are continuous in δ. We get the desired result by noting that $\lim_{\delta\to0}\sup_{y\in\mathcal{Y}}\left|\frac{R_F(\delta;y)}{\delta}\right| = 0$. So, uniformly over $y\in\mathcal{Y}$ as $\delta\to0$

$$F_{Y_\delta}(y) = F_Y(y) + \delta\int_{\mathcal{Y}(1)}\int_{\mathcal{W}}\mathbb{1}\left\{\tilde{y}\le y\right\}f_{Y(1)|U_D,W}(\tilde{y}|P(w),w)\dot{P}(w)f_W(w)dwd\tilde{y}$$

$$- \delta\int_{\mathcal{Y}(0)}\int_{\mathcal{W}}\mathbb{1}\left\{\tilde{y}\le y\right\}f_{Y(0)|U_D,W}(\tilde{y}|P(w),w)\dot{P}(w)f_W(w)dwd\tilde{y} + o(|\delta|).$$

∎

Proof of Theorem 1. Most of the proof is in the main text immediately preceding the statement of the theorem. We only need to show that

$$G(y) = E\left[\left\{F_{Y(1)|U_D,W}(y|P(W),W) - F_{Y(0)|U_D,W}(y|P(W),W)\right\}\dot{P}(W)\right]$$

is differentiable. To this end, we consider the limit

$$\lim_{h\to0}\frac{1}{h}\left[G(y+h) - G(y)\right]$$

$$= \lim_{h\to0}E\left[\left\{\frac{F_{Y(1)|U_D,W}(y+h|P(W),W) - F_{Y(1)|U_D,W}(y|P(W),W)}{h}\right\}\dot{P}(W)\right]$$

$$- \lim_{h\to0}E\left[\left\{\frac{F_{Y(0)|U_D,W}(y+h|P(W),W) - F_{Y(0)|U_D,W}(y|P(W),W)}{h}\right\}\dot{P}(W)\right].$$

But

$$\sup_h\left|\frac{F_{Y(d)|U_D,W}(y+h|P(W),W) - F_{Y(d)|U_D,W}(y|P(W),W)}{h}\right|$$

$$\le \sup_{\tilde{y}\in\mathcal{Y}(d)}\left|f_{Y(d)|U_D,W}(\tilde{y}|P(W),W)\right|.$$

This and Assumptions 2(b.iii) and 2(c.iii) allow us to use the dominated convergence theorem to obtain

$$\lim_{h\to0}\frac{1}{h}\left[G(y+h) - G(y)\right] = E\left[\left\{f_{Y(1)|U_D,W}(y|P(W),W) - f_{Y(0)|U_D,W}(y|P(W),W)\right\}\dot{P}(W)\right]$$

and hence $G(y)$ is indeed differentiable with derivative given above. Therefore,

$$\int_{\mathcal{Y}}\psi(y,\rho,F_Y)dG(y) = \int_{\mathcal{Y}}\psi(y,\rho,F_Y)$$

$$\times E\left[\left\{f_{Y(1)|U_D,W}(y|P(W),W) - f_{Y(0)|U_D,W}(y|P(W),W)\right\}\dot{P}(W)\right]dy.$$

∎

Proof of Corollary 1. We can see that

$$
\int_{\mathcal{Y}} y E\left[\left\{f_{Y(1)|U_D,W}(y|P(W),W) - f_{Y(0)|U_D,W}(y|P(W),W)\right\} \dot{P}(W)\right] dy
$$
$$
= E\left\{E\left[Y(1) - Y(0) | U_D = P(W), W\right] \dot{P}(W)\right\}.
$$

Hence it suffices to show that $\lim_{\delta \to 0} \delta^{-1} \int_{\mathcal{Y}} y dR_F(\delta; y) = 0$. Under Assumption 4, we have

$$
\left| \frac{1}{\delta} \int_{\mathcal{Y}} y dR_F(\delta; y) \right| \leq p \int_{\mathcal{Y}(1)} |\tilde{y}| \cdot \left| \frac{\partial f_{Y(1)|D_\delta}(\tilde{y}|1)}{\partial \delta} \right|_{\delta=\tilde{\delta}_1} - \frac{\partial f_{Y(1)|D_\delta}(\tilde{y}|1)}{\partial \delta} \bigg|_{\delta=0} \right| d\tilde{y}
$$
$$
+ (1-p) \int_{\mathcal{Y}(0)} |\tilde{y}| \cdot \left| \frac{\partial f_{Y(0)|D_\delta}(\tilde{y}|0)}{\partial \delta} \right|_{\delta=\tilde{\delta}_0} - \frac{\partial f_{Y(0)|D_\delta}(\tilde{y}|0)}{\partial \delta} \bigg|_{\delta=0} \right| d\tilde{y}
$$
$$
+ \int_{\mathcal{Y}(1)} |\tilde{y}| \cdot \left| f_{Y(1)|D}(\tilde{y}|1) - f_{Y(1)|D_\delta}(\tilde{y}|1) \right| d\tilde{y} + \int_{\mathcal{Y}(0)} |\tilde{y}| \cdot \left| f_{Y(0)|D}(\tilde{y}|0) - f_{Y(0)|D_\delta}(\tilde{y}|0) \right| d\tilde{y}.
$$

As in the proof of Lemma 2, each term in the above upper bound converges to zero as $\delta \to 0$. Therefore,

$$
\lim_{\delta \to 0} \frac{1}{\delta} \int_{\mathcal{Y}} y dR_F(\delta; y) = 0,
$$

as desired. ∎

Derivation of Equation (14). Note that

$$
E(Y_\delta) = \int_0^1 E(Y_\delta|\mathcal{P}_\delta = t) f_{\mathcal{P}_\delta}(t) dt. \tag{A.12}
$$

Now, consider $E(Y_\delta|\mathcal{P}_\delta = t)$. Depending on the value of $\mathcal{P}_\delta$ relative to the index U_D, we observe the potential outcome $Y(0)$ or $Y(1)$. By the independence of $\mathcal{P}_\delta$ from U_D and the law of iterated expectations, we have

$$
E(Y_\delta|\mathcal{P}_\delta = t) = \int_0^1 E(Y_\delta|\mathcal{P}_\delta = t, U_D = u) du
$$
$$
= \int_0^t E(Y(1)|U_D = u) du + \int_t^1 E(Y(0)|U_D = u) du
$$
$$
= \int_0^1 \Big[\mathbb{1}\{0 \leq u \leq t\} E(Y(1)|U_D = u)
$$
$$
+ \mathbb{1}\{t \leq u \leq 1\} E(Y(0)|U_D = u) \Big] du. \tag{A.13}
$$

Plugging (A.13) back into (A.12), we get

$$
\begin{aligned}
& E(Y_\delta) \\
= {} & \int_0^1 \left\{ \int_0^1 \left[\mathbf{1}\{0 \le u \le t\} E(Y(1)|U_D = u) + \mathbf{1}\{t \le u \le 1\} E(Y(0)|U_D = u) \right] du \right\} f_{\mathcal{P}_\delta}(t) dt \\
= {} & \int_0^1 \left\{ \int_0^1 \left[\mathbf{1}\{u \le t \le 1\} E(Y(1)|U_D = u) + \mathbf{1}\{0 \le t \le u\} E(Y(0)|U_D = u) \right] f_{\mathcal{P}_\delta}(t) dt \right\} du \\
= {} & \int_0^1 \left[(1 - F_{\mathcal{P}_\delta}(u)) E(Y(1)|U_D = u) + F_{\mathcal{P}_\delta}(u) E(Y(0)|U_D = u) \right] du. \tag{A.14}
\end{aligned}
$$

Going back to (13) using (A.14), we get

$$
\begin{aligned}
\text{PRTE} = {} & \frac{1}{\delta} \int_0^1 \left[(1 - F_{\mathcal{P}_\delta}(u)) E(Y(1)|U_D = u) + F_{\mathcal{P}_\delta}(u) E(Y(0)|U_D = u) \right] du \\
& - \frac{1}{\delta} \int_0^1 \left[(1 - F_{\mathcal{P}}(u)) E(Y(1)|U_D = u) - F_{\mathcal{P}}(u) E(Y(0)|U_D = u) \right] du \\
= {} & \frac{1}{\delta} \int_0^1 \text{MTE}(u)(F_{\mathcal{P}}(u) - F_{\mathcal{P}_\delta}(u)) du. \tag{A.15}
\end{aligned}
$$

Taking the limit in (A.15) as $\delta \to 0$ yields

$$
\begin{aligned}
\text{MPRTE} = {} & \lim_{\delta \to 0} -\frac{1}{\delta} \int_0^1 \text{MTE}(u)(F_{\mathcal{P}_\delta}(u) - F_{\mathcal{P}}(u)) du \\
= {} & -\int_0^1 \text{MTE}(u) \frac{\partial F_{\mathcal{P}_\delta}(u)}{\partial \delta} \bigg|_{\delta=0} du.
\end{aligned}
$$

∎

Proof of Corollary 2. It follows from Theorem 1 by considering $\psi(y, \rho, F_Y) = \frac{\tau - \mathbf{1}\{y \le y_\tau\}}{f_Y(y_\tau)}$. ∎

Proof of Corollary 3. For each $d = 0$ and 1, we have

$$
\begin{aligned}
& \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbf{1}\{Y(d) \le y_\tau\} | U_D = P(w), W = w \right] \dot{P}(w) f_W(w) dw \\
= {} & \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbf{1}\{Y(d) \le y_\tau\} | D = d, W = w \right] \dot{P}(w) f_W(w) dw \\
& + \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbf{1}\{Y(d) \le y_\tau\} | U_D = P(w), W = w \right] \dot{P}(w) f_W(w) dw \\
& - \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbf{1}\{Y(d) \le y_\tau\} | D = d, W = w \right] \dot{P}(w) f_W(w) dw \\
= {} & \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbf{1}\{Y(d) \le y_\tau\} | D = d, W = w \right] f_W(w) dw \\
& - \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbf{1}\{Y(d) \le y_\tau\} | D = d, W = w \right] f_W(w) \left[1 - \dot{P}(w) \right] dw \\
& - \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y(d)|D,W}(y_\tau|d, w) - F_{Y(d)|U_D, W}(y_\tau|P(w), w) \right] \dot{P}(w) f_W(w) dw \\
:= {} & A_\tau(d) - B_{1\tau}(d) - B_{2\tau}(d)
\end{aligned}
$$

where

$$A_\tau(d) = \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\{Y(d) \le y_\tau\} | D = d, W = w\right] f_W(w)\, dw,$$

$$B_{1\tau}(d) = \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\{Y(d) \le y_\tau\} | D = d, W = w\right] f_W(w)\left[1 - \dot{P}(w)\right] dw,$$

$$B_{2\tau}(d) = \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y(d)|D,W}(y_\tau|d, w) - F_{Y(d)|U_D,W}(y_\tau|P(w), w)\right] \dot{P}(w) f_W(w)\, dw.$$

So

$$
\begin{aligned}
\Pi_\tau &= A_\tau(0) - B_{1\tau}(0) - B_{2\tau}(0) - [A_\tau(1) - B_{1\tau}(1) - B_{2\tau}(1)] \\
&= A_\tau - B_{1\tau} - B_{2\tau},
\end{aligned}
$$

where

$$
\begin{aligned}
A_\tau &= A_\tau(0) - A_\tau(1) \\
&= \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\{Y \le y_\tau\} | D = 0, W = w\right] f_W(w)\, dw \\
&\quad - \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} E\left[\mathbb{1}\{Y \le y_\tau\} | D = 1, W = w\right] f_W(w)\, dw,
\end{aligned}
$$

$$
\begin{aligned}
B_{1\tau} &= B_{1\tau}(0) - B_{1\tau}(1) \\
&= \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y(0)|D,W}(y_\tau|0, w) - F_{Y(1)|D,W}(y_\tau|1, w)\right]\left[1 - \dot{P}(w)\right] f_W(w)\, dw \\
&= \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y|D,W}(y_\tau|0, w) - F_{Y|D,W}(y_\tau|1, w)\right]\left[1 - \dot{P}(w)\right] f_W(w)\, dw,
\end{aligned}
$$

and

$$
\begin{aligned}
B_{2\tau} &= B_{2\tau}(0) - B_{2\tau}(1) \\
&= \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y(0)|D,W}(y_\tau|0, w) - F_{Y(0)|U_D,W}(y_\tau|P(w), w)\right] \dot{P}(w) f_W(w)\, dw \\
&\quad + \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y(1)|U_D,W}(y_\tau|P(w), w) - F_{Y(1)|D,W}(y_\tau|1, w)\right] \dot{P}(w) f_W(w)\, dw.
\end{aligned}
$$

∎

Proof of Lemma 3. For a given δ, $s(\delta)$ satisfies $\Pr(D_\delta = 1) = p + \delta$. But

$$\Pr(D_\delta = 1) = E\left[P_\delta(W)\right] = \int_{\mathcal{W}} F_{V|W}(\mu(\mathcal{G}^*(w, s(\delta)), x)|w) f_W(w) dw,$$

where $\mathcal{G}^*(W, s) = (Z_1 + g(W)s, Z_2, \cdots, Z_{d_Z})$. Note that $\mathcal{G}^*(W, s)$ takes the same form as $\mathcal{G}(W, \delta)$ but their second arguments are different. So

$$p + \delta = \int_{\mathcal{W}} F_{V|W}(\mu(\mathcal{G}^*(w, s(\delta)), x)|w) f_W(w) dw. \tag{A.16}$$

Note that $s(0) = 0$. We need to find the derivative of the implicit function $s(\delta)$ with respect to δ. Define

$$t(\delta, s) = p + \delta - \int_{\mathcal{W}} F_{V|W}(\mu(\mathcal{G}^*(w,s), x)|w) f_W(w) dw. \tag{A.17}$$

By Theorem 9.28 in Rudin (1976), we need to show that $t(\delta, s)$ is continuously differentiable for (δ, s) in a neighborhood of $(0,0)$. We do this by showing that the partial derivatives of (A.17) with respect to δ and s exist and are continuous (See Theorem 9.21 in Rudin (1976)).

For the partial derivative with respect to δ, we have $\partial t(\delta, s)/\partial \delta = 1$, which is obviously continuous in (δ, s). For the partial derivative with respect to s, we use Assumption (iii) in the lemma to obtain

$$\frac{\partial t(\delta, s)}{\partial s} = -\int_{\mathcal{W}} f_{V|W}(\mu(\mathcal{G}^*(w,s), x)|w) \mu'_{z_1}(\mathcal{G}^*(w,s), x) g(w) f_W(w) dw.$$

The function is trivially continuous in δ. In view of the continuity of $f_{V|W}(v|w)$ in v for almost all w, the dominated convergence theorem implies that $\partial t(\delta, s)/\partial s$ is also continuous in s. Therefore, we can apply the implicit function theorem to obtain $s'(\delta)$ in a neighborhood of $\delta = 0$. Taking the derivative of (A.16) with respect to δ, we get

$$\begin{aligned}
\frac{\partial s(\delta)}{\partial \delta} &= \frac{1}{\int_{\mathcal{W}} f_{V|W}(\mu(\mathcal{G}(w,\delta), x)|w) \mu'_{z_1}(\mathcal{G}(w,\delta), x) g(w) f_W(w) dw} \\
&= \frac{1}{E\left[f_{V|W}(\mu(\mathcal{G}(W,\delta), X)|W) \mu'_{z_1}(\mathcal{G}(W,\delta), X) g(W)\right]}.
\end{aligned}$$

Next, we have

$$P_\delta(z, x) = \Pr(D_\delta = 1|Z = z, X = x) = F_{V|W}(\mu(\mathcal{G}(w,\delta), x)|w).$$

So

$$\begin{aligned}
\frac{\partial P_\delta(z, x)}{\partial \delta} &= f_{V|W}(\mu(\mathcal{G}(w,\delta), x)|w) \mu'_{z_1}(z + g(w)s(\delta), x) g(w) \frac{\partial s(\delta)}{\partial \delta} \\
&= \frac{f_{V|W}(\mu(\mathcal{G}(w,\delta), x)|w) \mu'_{z_1}(\mathcal{G}(w,\delta), x) g(w)}{E\left[f_{V|W}(\mu(\mathcal{G}(W,\delta), X)|W) \mu'_{z_1}(\mathcal{G}(W,\delta), X) g(W)\right]}.
\end{aligned}$$

It then follows that

$$\left.\frac{\partial s(\delta)}{\partial \delta}\right|_{\delta=0} = \frac{1}{E\left[f_{V|W}(\mu(W)|W) \mu'_{z_1}(W) g(W)\right]},$$

and

$$\left.\frac{\partial P_\delta(z, x)}{\partial \delta}\right|_{\delta=0} = \frac{f_{V|W}(\mu(w)|w) \mu'_{z_1}(w) g(w)}{E\left[f_{V|W}(\mu(W)|W) \mu'_{z_1}(W) g(W)\right]}.$$

∎

Proof of Corollary 4. It follows from an application of Lemma 3 to Corollary 3. ∎

Proof of Proposition 1. Note that for any bounded function $\mathcal{L}(\cdot)$, we have

$$
\begin{aligned}
E\left[\mathcal{L}(Y)|P(W) = P(w), X = x\right] = {} & E\left[\mathcal{L}(Y(1))|D = 1, P(W) = P(w), X = x\right] \\
& \times \Pr(D = 1|P(W) = P(w), X = x) \\
& + E\left[\mathcal{L}(Y(0))|D = 0, P(W) = P(w), X = x\right] \\
& \times \Pr(D = 0|P(W) = P(w), X = x).
\end{aligned}
$$

But, using $D = \mathbb{1}\left\{U_D \leq P(W)\right\}$, we have

$$
\begin{aligned}
\Pr(D = 1|P(W) = P(w), X = x) & = \Pr(U_D \leq P(W)|P(W) = P(w), X = x) \\
& = \Pr(U_D \leq P(w)|P(W) = P(w), X = x) \\
& = P(w),
\end{aligned}
$$

because U_D is independent of W. So,

$$
\begin{aligned}
E\left[\mathcal{L}(Y)|P(W) = P(w), X = x\right] = {} & E\left[\mathcal{L}(Y(1))|D = 1, P(W) = P(w), X = x\right] P(w) \\
& + E\left[\mathcal{L}(Y(0))|D = 0, P(W) = P(w), X = x\right] (1 - P(w)). \\
= {} & E\left[\mathcal{L}(Y(1)))|U_D \leq P(w), P(W) = P(w), X = x\right] P(w) \\
& + E\left[\mathcal{L}(Y(0))|U_D > P(w), P(W) = P(w), X = x\right] (1 - P(w)) \\
= {} & E\left[\mathcal{L}(Y(1)))|U_D \leq P(w), X = x\right] P(w) \\
& + E\left[\mathcal{L}(Y(0))|U_D > P(w), X = x\right] (1 - P(w)),
\end{aligned}
$$

where the last line follows because $U = (U_0, U_1)$ is independent of Z given X and U_D.

Now

$$
\begin{aligned}
E\left[\mathcal{L}(Y(1)))|U_D \leq P(w), X = x\right] & = E\left\{E\left[\mathcal{L}(Y(1))|U_D, X = x\right]|U_D \leq P(w), X = x\right\} \\
& = E\left\{E\left[\mathcal{L}(Y(1))|U_D, X = x\right]|U_D \leq P(w)\right\} \\
& = \frac{\int_0^{P(w)} E\left[\mathcal{L}(Y(1))|U_D = u, X = x\right] du}{P(w)},
\end{aligned}
$$

where the first equality uses the law of iterated expectations, the second equality uses the independence of U_D from X, and the last equality uses $U_D \sim$ uniform on $[0,1]$. Similarly,

$$
E\left[\mathcal{L}(Y(0))|U_D > P(w), X = x\right] = \frac{\int_{P(w)}^1 E\left[\mathcal{L}(Y(1))|U_D = u, X = x\right] du}{1 - P(w)}.
$$

So we have

$$
\begin{aligned}
E\left[\mathcal{L}(Y)|P(W) = P(w), X = x\right] = {} & \int_0^{P(w)} E\left[\mathcal{L}(Y(1))|U_D = u, X = x\right] du \\
& + \int_{P(w)}^1 E\left[\mathcal{L}(Y(0))|U_D = u, X = x\right] du.
\end{aligned}
$$

9

By taking $\mathcal{L}(\cdot) = \mathbb{1}\{\cdot \leq y_\tau\}$, we have

$$
\begin{aligned}
E\left[\mathbb{1}\{Y \leq y_\tau\}\,|P(W) = P(w), X = x\right] &= \int_0^{P(w)} E\left[\mathbb{1}\{Y(1) \leq y_\tau\}\,|U_D = u, X = x\right] du \\
&+ \int_{P(w)}^1 E\left[\mathbb{1}\{Y(0) \leq y_\tau\}\,|U_D = u, X = x\right] du.
\end{aligned}
$$

Under Assumptions 2(a) and 2(b), we can invoke the fundamental theorem of calculus to obtain

$$
\begin{aligned}
\frac{\partial E\left[\mathbb{1}\{Y \leq y_\tau\}\,|P(W) = P(w), X = x\right]}{\partial P(w)} &= E\left[\mathbb{1}\{Y(1) \leq y_\tau\}\,|U_D = P(w), X = x\right] \\
&- E\left[\mathbb{1}\{Y(0) \leq y_\tau\}\,|U_D = P(w), X = x\right] \\
&= \mathrm{MTE}_\tau(P(w), x). \qquad\qquad\text{(A.18)}
\end{aligned}
$$

That is,

$$
\mathrm{MTE}_\tau(u, x) = \frac{\partial E\left[\mathbb{1}\{Y \leq y_\tau\}\,|P(W) = u, X = x\right]}{\partial u}
$$

for any u such that there is a $w \in \mathcal{W}$ satisfying $P(w) = u$. ∎

Proof of Lemma 5. We have

$$
\begin{aligned}
\hat{f}_Y(y) - f_Y(y) &= \frac{1}{nh}\sum_{i=1}^n K\left(\frac{Y_i - y}{h}\right) - f_Y(y) \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{h} K\left(\frac{Y_i - y}{h}\right) - E\hat{f}_Y(y) + E\hat{f}_Y(y) - f_Y(y) \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{h} K\left(\frac{Y_i - y}{h}\right) - E\hat{f}_Y(y) + B_{f_Y}(y) + o_p(h^2),
\end{aligned}
$$

where

$$
B_{f_Y}(y) = \frac{1}{2}h^2 f_Y''^2(y)\int_{-\infty}^{\infty} u^2 K(u)du.
$$

We write this concisely as

$$
\hat{f}_Y(y) - f_Y(y) = \frac{1}{n}\sum_{i=1}^n \psi_{f_Y, i}(y, h) + B_{f_Y}(y) + o_p(h^2),
$$

where

$$
\psi_{f_Y, i}(y, h) := \frac{1}{h} K\left(\frac{Y_i - y}{h}\right) - E\frac{1}{h}K\left(\frac{Y_i - y}{h}\right) = O_p(n^{-1/2}h^{-1/2}).
$$

This completes the proof of the first result.

We now prove the second result of the lemma. Since $K(u)$ is twice continuously differentiable, we use a Taylor expansion to obtain

$$
\hat{f}_Y(\hat{y}_\tau) - \hat{f}_Y(y_\tau) = \hat{f}_Y'(y_\tau)(\hat{y}_\tau - y_\tau) + \frac{1}{2}\hat{f}''(\tilde{y}_\tau)(\hat{y}_\tau - y_\tau)^2 \qquad\qquad\text{(A.19)}
$$

for some $\tilde{y}_\tau$ between $\hat{y}_\tau$ and y_τ. The first and second derivatives are

$$\hat{f}'_Y(y) = -\frac{1}{nh^2} \sum_{i=1}^n K'\left(\frac{Y_i - y}{h}\right), \quad \hat{f}''_Y(y) = \frac{1}{nh^3} \sum_{i=1}^n K''\left(\frac{Y_i - y}{h}\right).$$

To find the order of $\hat{f}''_Y(y)$, we calculate its mean and variance. We have

$$
\begin{aligned}
E\left[\hat{f}''_Y(y)\right] &= \frac{1}{nh^3} \sum_{i=1}^n K''\left(\frac{Y_i - y}{h}\right) = O\left(\frac{1}{h^2}\right), \\
var\left[\hat{f}''_Y(y)\right] &\leq \frac{n}{(nh^3)^2} E\left[K''\left(\frac{Y_i - y}{h}\right)\right]^2 = O\left(\frac{1}{nh^5}\right).
\end{aligned}
$$

Therefore, when $nh^3 \to \infty$,

$$\hat{f}''_Y(y) = O_p\left(h^{-2}\right),$$

for any y. That is, for any $\epsilon > 0$, there exists an $M > 0$ such that

$$\Pr\left(h^2 \left|\hat{f}''_Y(y_\tau)\right| > \frac{M}{2}\right) < \frac{\epsilon}{2}$$

when n is large enough.

Suppose we choose M so large that we also have

$$\Pr\left(\sqrt{n}\left|\tilde{y}_\tau - y_\tau\right| > M\right) < \frac{\epsilon}{2}$$

when n is large enough. Then, when n is large enough,

$$
\begin{aligned}
\Pr\left(h^2 \hat{f}''_Y(\tilde{y}_\tau) > M\right) &\leq \Pr\left(h^2 \left|\hat{f}''_Y(\tilde{y}_\tau) - \hat{f}''_Y(y_\tau)\right| > \frac{M}{2}\right) + \Pr\left(h^2 \left|\hat{f}''_Y(y_\tau)\right| > \frac{M}{2}\right) \\
&\leq \Pr\left(h^2 \left[\hat{f}''_Y(\tilde{y}_\tau) - \hat{f}''_Y(y_\tau)\right] > \frac{M}{2}\right) + \frac{\epsilon}{2} \\
&\leq \Pr\left(h^2 \left[\hat{f}''_Y(\tilde{y}_\tau) - \hat{f}''_Y(y_\tau)\right] > \frac{M}{2}, \sqrt{n}\left|\tilde{y}_\tau - y_\tau\right| < M\right) + \epsilon. \quad \text{(A.20)}
\end{aligned}
$$

We split $\sqrt{n}\left|\tilde{y}_\tau - y_\tau\right| < M$ into two cases. When $\sqrt{n}\left|\tilde{y}_\tau - y_\tau\right| < \sqrt{h}$, we have

$$
\begin{aligned}
h^2 \left|\hat{f}''_Y(\tilde{y}_\tau) - \hat{f}''_Y(y_\tau)\right| &\leq \frac{1}{nh} \sum_{i=1}^n \left|K''\left(\frac{Y_i - y_\tau}{h}\right) - K''\left(\frac{Y_i - \tilde{y}_\tau}{h}\right)\right| \\
&\leq \frac{1}{nh} \cdot n \cdot L_K \frac{1}{h}\left|y_\tau - \tilde{y}_\tau\right| \leq L_K \cdot \frac{1}{h^2}\frac{\sqrt{h}}{\sqrt{n}} = L_K \cdot \frac{1}{\sqrt{nh^3}}
\end{aligned}
$$

by the Lipschitz continuity of $K''(\cdot)$ with Lipschitz constant L_K. When $\sqrt{h} \leq \sqrt{n}\left|\tilde{y}_\tau - y_\tau\right| < M$, we have

$$\left|\frac{Y_i - y_\tau}{h} - \frac{Y_i - \tilde{y}_\tau}{h}\right| = \frac{\sqrt{n}\left|\tilde{y}_\tau - y_\tau\right|}{h} > \frac{1}{\sqrt{h}} \to \infty.$$

11

Using the second condition on $K''(\cdot)$, we have, for $\sqrt{h} \le \sqrt{n} |\tilde{y}_\tau - y_\tau| < M$,

$$\left| K''\left(\frac{Y_i - y_\tau}{h}\right) - K''\left(\frac{Y_i - \tilde{y}_\tau}{h}\right) \right| \le C_2 \frac{(\tilde{y}_\tau - y_\tau)^2}{h^2} \le C_2 \frac{M^2}{nh^2},$$

and

$$\begin{aligned}
h^2 \left| \hat{f}_Y''(\tilde{y}_\tau) - \hat{f}_Y''(y_\tau) \right| &\le \frac{1}{nh} \sum_{i=1}^{n} \left| K''\left(\frac{Y_i - y_\tau}{h}\right) - K''\left(\frac{Y_i - \tilde{y}_\tau}{h}\right) \right| \\
&\le C_2 \frac{M^2}{nh^3} = O_p\left(\frac{1}{\sqrt{nh^3}}\right).
\end{aligned}$$

Hence, in both cases, $h^2 \left| \hat{f}_Y''(\tilde{y}_\tau) - \hat{f}_Y''(y_\tau) \right| = O_p\left(n^{-1/2} h^{-3/2}\right)$. As a result,

$$\Pr\left(h^2 \left[\hat{f}_Y''(\tilde{y}_\tau) - \hat{f}_Y''(y_\tau) \right] > \frac{M}{2}, \sqrt{n} |\tilde{y}_\tau - y_\tau| < M\right) \to 0.$$

Combining this with (A.20), we obtain

$$h^2 \hat{f}_Y''(\tilde{y}_\tau) = O_p(1) \text{ and } \hat{f}_Y''(\tilde{y}_\tau)(\hat{y}_\tau - y_\tau)^2 = O_p\left(n^{-1} h^{-2}\right).$$

In view of (A.19), we then have

$$\hat{f}_Y(\hat{y}_\tau) - \hat{f}_Y(y_\tau) = \hat{f}_Y'(y_\tau)(\hat{y}_\tau - y_\tau) + O_p(n^{-1} h^{-2}).$$

Now, using Lemma 4, we can write

$$\begin{aligned}
\hat{f}_Y(\hat{y}_\tau) - \hat{f}_Y(y_\tau) &= f_Y'(y_\tau)(\hat{y}_\tau - y_\tau) + \left[\hat{f}_Y'(y_\tau) - f_Y'(y_\tau) \right] (\hat{y}_\tau - y_\tau) + O_p(n^{-1} h^{-2}) \\
&= f_Y'(y_\tau) \frac{1}{n} \sum_{i=1}^{n} \psi_{Q,i}(y_\tau) + R_{f_Y},
\end{aligned}$$

where

$$R_{f_Y} := \left[\hat{f}_Y'(y_\tau) - f_Y'(y_\tau) \right] [\hat{y}_\tau - y_\tau] + o_p(n^{-1/2}) + O_p(n^{-1} h^{-2})$$

and the $o_p(n^{-1/2})$ term is the error of the linear asymptotic representation of $\hat{y}_\tau - y_\tau$.

In order to obtain the order of R_{f_Y}, we use the following results:

$$\begin{aligned}
\hat{f}_Y'(y) &= f_Y'(y) + O_p\left(\frac{1}{\sqrt{nh^3}} + h^2\right), \\
\hat{y}_\tau &= y_\tau + O_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

The rate on the derivative of the density can be found on page 56 of Pagan and Ullah (1999).

Therefore,

$$
\begin{aligned}
R_{f_Y} &= o_p(n^{-1/2}) + O_p\left(\frac{1}{\sqrt{nh^3}} + h^2\right)O_p\left(\frac{1}{\sqrt{n}}\right) + O_p(n^{-1}h^{-2}). \\
&= o_p(n^{-1/2}) + O_p\left(n^{-1}h^{-3/2}\right) + O_p\left(n^{-1/2}h^2\right) + O_p(n^{-1}h^{-2}). \\
&= o_p(n^{-1/2}) + O_p\left(n^{-1}h^{-3/2}\right) + O_p(n^{-1}h^{-2}).
\end{aligned}
$$

because, since by Assumption 7, $h \downarrow 0$, so $O_p\left(n^{-1/2}h^2\right) = o_p(n^{-1/2})$. We need to show that $\sqrt{nh}R_{f_Y} = o_p(1)$. We do this term by term. First,

$$
\sqrt{nh} \times o_p(n^{-1/2}) = o_p(h^{1/2}) = o_p(1)
$$

because $h \downarrow 0$. Second,

$$
\sqrt{nh} \times O_p\left(n^{-1}h^{-3/2}\right) = O_p\left(n^{-1/2}h^{-1}\right) = o_p(1)
$$

as long as $nh^2 \uparrow \infty$, which is guaranteed by Assumption 7, since it is implied by $nh^3 \uparrow \infty$. Finally,

$$
\sqrt{nh} \times O_p(n^{-1}h^{-2}) = O_p(n^{-1/2}h^{-3/2}) = o_p(1)
$$

since by Assumption 7 $nh^3 \uparrow \infty$. Therefore, $\sqrt{nh}R_{f_Y} = o_p(1)$. ∎

Proof of Lemma 6. We have the following decomposition:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Z_i, X_i, \hat{\alpha})}{\partial z_1} - E\left[\frac{\partial P(Z, X, \alpha_0)}{\partial z_1}\right] &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Z_i, X_i, \hat{\alpha})}{\partial z_1} - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Z_i, X_i, \alpha_0)}{\partial z_1} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Z_i, X_i, \alpha_0)}{\partial z_1} - E\left[\frac{\partial P(Z, X, \alpha_0)}{\partial z_1}\right].
\end{aligned}
$$

Under Condition (b) of the lemma, we have

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Z_i, X_i, \alpha_0)}{\partial z_1} - E\left[\frac{\partial P(Z, X, \alpha_0)}{\partial z_1}\right] = \frac{1}{n}\sum_{i=1}^{n}\psi_{\partial P, i} = O_p(n^{-1/2}),
$$

where

$$
\psi_{\partial P, i} = \frac{\partial P(Z_i, X_i, \alpha_0)}{\partial z_1} - E\left[\frac{\partial P(Z, X, \alpha_0)}{\partial z_1}\right].
$$

For the first term, we have by applying the mean value theorem coordinate-wise

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Z_i, X_i, \hat{\alpha})}{\partial z_1} - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Z_i, X_i, \alpha_0)}{\partial z_1} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 P(Z_i, X_i, \tilde{\alpha})}{\partial\alpha\partial z_1}\right)'(\hat{\alpha} - \alpha_0),
$$

where $\tilde{\alpha}$ is a vector with (not necessarily equal) coordinates between α_0 and $\hat{\alpha}$. Under Conditions (c)

and (d) of the lemma, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 P(Z_i, X_i, \tilde{\alpha})}{\partial\alpha\partial z_1} \xrightarrow{p} E\left[\frac{\partial^2 P(Z, X, \alpha_0)}{\partial\alpha\partial z_1}\right]. \qquad (A.21)$$

Using (A.21) together with the linear representation of $\hat{\alpha}$ in Condition (a), we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Z_i, X_i, \hat{\alpha})}{\partial z_1} - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Z_i, X_i, \alpha_0)}{\partial z_1} &= \left\{ E\left[\frac{\partial^2 P(Z, X, \alpha_0)}{\partial\alpha\partial z_1}\right]' + o_p(1) \right\} \left\{ \mathbb{P}_n\psi_{\alpha_0} + o_p(n^{-1/2}) \right\} \\
&= E\left[\frac{\partial^2 P(Z, X, \alpha_0)}{\partial\alpha\partial z_1}\right]' \mathbb{P}_n\psi_{\alpha_0} + o_p(n^{-1/2}).
\end{aligned}$$

We can then write

$$\begin{aligned}
T_{1n}(\hat{\alpha}) - T_1 &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(Z_i, X_i, \hat{\alpha})}{\partial z_1} - E\left[\frac{\partial P(Z, X, \alpha_0)}{\partial z_1}\right] \\
&= E\left[\frac{\partial^2 P(Z, X, \alpha_0)}{\partial\alpha\partial z_1}\right]' \mathbb{P}_n\psi_{\alpha_0} + \mathbb{P}_n\psi_{\partial P} + o_p(n^{-1/2}).
\end{aligned}$$

∎

Proof of Lemma 7. Recall that

$$\begin{aligned}
m_0\left(y_\tau, \tilde{w}\left(\alpha_\theta\right)\right) &:= m_0\left(y_\tau, P(w, \alpha_\theta), x\right) \\
&= E\left[\mathbf{1}\left\{Y \leq y_\tau\right\} | P(W, \alpha_\theta) = P(w, \alpha_\theta), X = x\right].
\end{aligned}$$

In order to emphasize the dual roles of α_θ, we define

$$\tilde{m}_0(y_\tau, u, x; P\left(\cdot, \alpha_{\theta_2}\right)) = E\left[\mathbf{1}\left\{Y \leq y_\tau\right\} | P(W, \alpha_{\theta_2}) = u, X = x\right].$$

Since y_τ is fixed, we regard $\tilde{m}_0$ as a function of (u, x) that depends on the function $P\left(\cdot, \alpha_{\theta_2}\right)$. Then

$$\begin{aligned}
\tilde{m}_0(y_\tau, P(w, \alpha_{\theta_1}), x; P\left(\cdot, \alpha_{\theta_2}\right))|_{\theta_1=\theta_2=\theta} &= E\left[\mathbf{1}\left\{Y \leq y_\tau\right\} | P(W, \alpha_\theta) = P(w, \alpha_\theta), X = x\right] \\
&= m_0\left(y_\tau, P(w, \alpha_\theta), x\right).
\end{aligned}$$

As in Hahn and Ridder (2013), we employ $\tilde{m}_0\left(y_\tau, u, x; P\left(\cdot, \alpha_{\theta_2}\right)\right)$ as an expositional device only.

The functional of interest is

$$\begin{aligned}
\mathcal{H}[m_0] =: \quad \mathcal{H}[m_0\left(y_\tau, P(\cdot, \alpha_\theta), \cdot\right)] &= \int_{\mathcal{W}} \frac{\partial m_0\left(y_\tau, P(w, \alpha_\theta), x\right)}{\partial z_1} f_W(w)\, dw \\
&= \int_{\mathcal{W}} \frac{\partial \tilde{m}_0(y_\tau, P(w, \alpha_{\theta_1}), x; P\left(\cdot, \alpha_{\theta_2}\right))}{\partial z_1}\bigg|_{\theta_1=\theta_2=\theta} f_W(w)\, dw.
\end{aligned}$$

Under Condition (b.i) of the lemma, we can exchange $\frac{\partial}{\partial \alpha_\theta}$ with $\int_{\mathcal{W}}$ and obtain

$$
\begin{aligned}
\left.\frac{\partial}{\partial \alpha_\theta} \mathcal{H}[m_0]\right|_{\theta=\theta_0} &= \int_{\mathcal{W}} \frac{\partial}{\partial z_1} \left.\frac{\partial \tilde{m}_0(y_\tau, P(w, \alpha_{\theta_1}), x; P(\cdot, \alpha_{\theta_2}))}{\partial \alpha_{\theta_1}}\right|_{\theta_1=\theta_2=\theta_0} f_W(w)\, dw \\
&\quad + \int_{\mathcal{W}} \frac{\partial}{\partial z_1} \left.\frac{\partial \tilde{m}_0(y_\tau, P(w, \alpha_{\theta_1}), x; P(\cdot, \alpha_{\theta_2}))}{\partial \alpha_{\theta_2}}\right|_{\theta_1=\theta_2=\theta_0} f_W(w)\, dw \\
&= \int_{\mathcal{W}} \frac{\partial}{\partial z_1} \left[\tilde{m}'_{0,\alpha}(y_\tau, P(w, \alpha_{\theta_0}), x; P(\cdot, \alpha_{\theta_0}))\right] f_W(w)\, dw,
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{m}'_{0,\alpha}(y_\tau, P(w, \alpha_\theta), x; P(\cdot, \alpha_\theta)) &= \left.\frac{\partial \tilde{m}_0(y_\tau, P(w, \alpha_{\theta_1}), x; P(\cdot, \alpha_\theta))}{\partial \alpha_{\theta_1}}\right|_{\theta_1=\theta} \\
&\quad + \left.\frac{\partial \tilde{m}_0(y_\tau, P(w, \alpha_\theta), x; P(\cdot, \alpha_{\theta_2}))}{\partial \alpha_{\theta_2}}\right|_{\theta_2=\theta}.
\end{aligned}
$$

Under Condition (a) of the lemma, we can use integration by parts. Letting $z_{-1} = (z_2, \ldots, z_{d_Z})$ and $[z_{1l}(w_{-1}), z_{1u}(w_{-1})]$ be the support of Z_1 given $W_{-1} = w_{-1}$, we have

$$
\begin{aligned}
&\int_{\mathcal{W}} \frac{\partial}{\partial z_1} \left[\tilde{m}'_{0,\alpha}(y_\tau, P(w, \alpha_{\theta_0}), x; P(\cdot, \alpha_{\theta_0}))\right] f_W(w)\, dw \\
&= \int_{\mathcal{X}} \int_{z_{1l}(w_{-1})}^{z_{1u}(w_{-1})} \frac{\partial}{\partial z_1} \left[\tilde{m}'_{0,\alpha}(y_\tau, P(w, \alpha_{\theta_0}), x; P(\cdot, \alpha_{\theta_0}))\right] f_{Z_1|W_{-1}}(z_1|w_{-1})\, dz_1 f_{W_{-1}}(w_{-1})\, dw_{-1} \\
&= \int_{\mathcal{X}} \tilde{m}'_{0,\alpha}(y_\tau, P(w, \alpha_{\theta_0}), x; P(\cdot, \alpha_{\theta_0})) f_{Z_1|W_{-1}}(z_1|w_{-1})\Big|_{z_{1l}(w_{-1})}^{z_{1u}(w_{-1})} f_{W_{-1}}(w_{-1})\, dw_{-1} \\
&\quad - \int_{\mathcal{W}} \tilde{m}'_{0,\alpha}(y_\tau, P(w, \alpha_{\theta_0}), x; P(\cdot, \alpha_{\theta_0})) \frac{\partial \log f_{Z_1|W_{-1}}(z_1|w_{-1})}{\partial z_1} f_W(w)\, dw \\
&= -\int_{\mathcal{W}} \tilde{m}'_{0,\alpha}(y_\tau, P(w, \alpha_{\theta_0}), x; P(\cdot, \alpha_{\theta_0})) \frac{\partial \log f_W(w)}{\partial z_1} f_W(w)\, dw.
\end{aligned}
$$

Define

$$
v(u, x; P(\cdot, \alpha_\theta)) = E\left[\left.\frac{\partial \log f_W(W)}{\partial z_1}\right| P(W, \alpha_\theta) = u, X = x\right].
$$

By the law of iterated expectations, we have

$$
\begin{aligned}
&\int_{\mathcal{W}} \tilde{m}_0(y_\tau, P(w, \alpha_\theta), x; P(\cdot, \alpha_\theta)) v(P(w, \alpha_\theta), x; P(\cdot, \alpha_{\theta_0})) f_W(w)\, dw \\
&= E\left[\mathbf{1}\{Y \le y_\tau\} v(P(W, \alpha_\theta), X; P(\cdot, \alpha_{\theta_0}))\right]. \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad (A.22)
\end{aligned}
$$

This holds because

$$
\tilde{m}_0(y_\tau, P(W, \alpha_\theta), X; P(\cdot, \alpha_\theta)) = E\left[\mathbf{1}\{Y \le y_\tau\} | \mathcal{I}_\theta\right]
$$

where $\mathcal{I}_\theta$ is the sub-σ algebra generated by $(P(W, \alpha_\theta), X)$ and $v(P(W, \alpha_\theta), X; P(\cdot, \alpha_{\theta_0}))$ is a function of $(P(W, \alpha_\theta), X)$ and is thus $\mathcal{I}_\theta$-measurable.

Differentiating (A.22) with respect to α_θ and evaluating the resulting equation at $\theta = \theta_0$, we have

$$E\left[\frac{\partial \tilde{m}_0(y_\tau, P(W, \alpha_{\theta_1}), X; P(\cdot, \alpha_{\theta_0}))}{\partial \alpha_{\theta_1}}\bigg|_{\theta_1 = \theta_0} \nu\left(P(W, \alpha_{\theta_0}), X; P(\cdot, \alpha_{\theta_0})\right)\right]$$

$$+ \; E\left[\frac{\partial \tilde{m}_0(y_\tau, P(W, \alpha_{\theta_0}), X; P(\cdot, \alpha_{\theta_2}))}{\partial \alpha_{\theta_2}}\bigg|_{\theta_2 = \theta_0} \nu\left(P(W, \alpha_{\theta_0}), X; P(\cdot, \alpha_{\theta_0})\right)\right]$$

$$= \; E\left\{\left[\mathbf{1}\left\{Y \leq y_\tau\right\} - m\left(y_\tau, P(W, \alpha_{\theta_0}), X\right)\right] \frac{\partial \nu\left(P(W, \alpha_\theta), X; P(\cdot, \alpha_{\theta_0})\right)}{\partial \alpha_\theta}\bigg|_{\theta = \theta_0}\right\}, \qquad \text{(A.23)}$$

where we have used Conditions (b.ii) and (b.iii) of the lemma to exchange the differentiation with the expectation.

Using (A.23) and the law of iterated expectations, we have

$$\int_{\mathcal{W}} \frac{\partial}{\partial z_1}\left[\tilde{m}'_{0,\alpha}\left(y_\tau, P(w, \alpha_{\theta_0}), x; P\left(\cdot, \alpha_{\theta_0}\right)\right)\right] f_W(w)\, dw$$

$$= \; -\int_{\mathcal{W}} \tilde{m}'_{0,\alpha}\left(y_\tau, P(w, \alpha_{\theta_0}), x; P\left(\cdot, \alpha_{\theta_0}\right)\right) \frac{\partial \log f_W(w)}{\partial z_1} f_W(w)\, dw$$

$$= \; \int_{\mathcal{W}} \tilde{m}'_{0,\alpha}\left(y_\tau, P(w, \alpha_{\theta_0}), x; P\left(\cdot, \alpha_{\theta_0}\right)\right) \nu\left(P(w, \alpha_{\theta_0}), x; P(\cdot, \alpha_{\theta_0})\right) f_W(w)\, dw$$

$$= \; E\left\{\left[\mathbf{1}\left\{Y \leq y_\tau\right\} - m\left(y_\tau, P(W, \alpha_{\theta_0}), X\right)\right] \frac{\partial \nu\left(P(W, \alpha_\theta), X; P(\cdot, \alpha_{\theta_0})\right)}{\partial \alpha_\theta}\bigg|_{\theta = \theta_0}\right\}.$$

Note that $\frac{\partial \nu\left(P(W, \alpha_\theta), X; P(\cdot, \alpha_{\theta_0})\right)}{\partial \alpha_\theta}\bigg|_{\theta = \theta_0}$ is a measurable function of $P(W, \alpha_{\theta_0})$, X, and $\partial P(W, \alpha_{\theta_0})/\partial \alpha_{\theta_0}$. Also,

$$E\left\{\mathbf{1}\left\{Y \leq y_\tau\right\} | P(W, \alpha_{\theta_0}), \partial P(W, \alpha_{\theta_0})/\partial \alpha_{\theta_0}, X\right\}$$

$$= \; E\left\{\mathbf{1}\left\{Y \leq y_\tau\right\} | P(W, \alpha_{\theta_0}), X\right\} = m\left(y_\tau, P(W, \alpha_{\theta_0}), X\right)$$

because $\partial P(W, \alpha_{\theta_0})/\partial \alpha_{\theta_0}$ contains only information on Z and Z is independent of (U_D, U_1, U_2) given X. These imply that

$$\int_{\mathcal{W}} \frac{\partial}{\partial z_1}\left[\tilde{m}'_{0,\alpha}\left(y_\tau, P(w, \alpha_{\theta_0}), x; P\left(\cdot, \alpha_{\theta_0}\right)\right)\right] f_W(w)\, dw = 0.$$

Hence,

$$\frac{\partial}{\partial \theta} E\left[\frac{\partial m_0(y_\tau, P(Z, X, \alpha_\theta), X)}{\partial z_1}\right]\bigg|_{\theta = \theta_0} = 0.$$

∎

Proof of Lemma 8. First, we prove that the decomposition in (31) is valid. We start by showing that

$$T_{2,\theta} = E_\theta\left[\frac{\partial m_\theta(y_{\tau,\theta}, \tilde{W}(\alpha_\theta))}{\partial z_1}\right]$$

is differentiable at θ_0. For this, it suffices to show that each of the four derivatives below exists at

16

$\theta = \theta_0$:

$$\frac{\partial}{\partial \theta} E_\theta \left[\frac{\partial m_0(y_\tau, \tilde{W}(\alpha_0))}{\partial z_1} \right] ; \frac{\partial}{\partial \theta} E \left[\frac{\partial m_\theta(y_\tau, \tilde{W}(\alpha_0))}{\partial z_1} \right] ;$$
$$\frac{\partial}{\partial \theta} E \left[\frac{\partial m_0(y_{\tau,\theta}, \tilde{W}(\alpha_0))}{\partial z_1} \right] ; \frac{\partial}{\partial \theta} E \left[\frac{\partial m_0(y_\tau, \tilde{W}(\alpha_\theta))}{\partial z_1} \right] . \tag{A.24}$$

By Lemma 7, the last derivative exists and is equal to zero at $\theta = \theta_0$. We deal with the rest three derivatives in (A.24) one at a time. Consider the first derivative. Under Condition (a.iii) and Condition (b) with $\ell = 0$ of the lemma, we have

$$\int_{\mathcal{W}} \left[\frac{\partial m_0(y_\tau, \tilde{w}(\alpha_0))}{\partial z_1} \sup_{\theta \in \Theta_0} \left| \frac{\partial f_W(w; \theta)}{\partial \theta} \right| dw \right] < \infty.$$

So $E_\theta \left[\frac{\partial m_0(y_\tau, \tilde{W}(\alpha_0))}{\partial z_1} \right]$ is differentiable in θ and

$$\frac{\partial}{\partial \theta} E_\theta \left[\frac{\partial m_0(y_\tau, \tilde{W}(\alpha_0))}{\partial z_1} \right] \bigg|_{\theta=\theta_0} = E \left[\frac{\partial m_0(y_\tau, \tilde{W}(\alpha_0))}{\partial z_1} S(O) \right] .$$

Hence, the contribution associated with the first derivative is simply the influence function of $T_{2n}(y_\tau, m_0, \alpha_0) - T_2$.

Now, for the second derivative in (A.24), Theorem 7.2 in Newey (1994) shows that the assumptions of the lemma imply the following:

1. There is a function $\gamma_{m_0}(o)$ and a measure $\hat{F}_{m_0}$ such that $E[\gamma_{m_0}(O)] = 0$, $E[\gamma_{m_0}(O)^2] < \infty$, and for all $\hat{m}$ such that $\|\hat{m} - m_0\|_\infty$ is small enough,

$$E \left[\frac{\partial \hat{m}(y_\tau, P(Z, X, \alpha_0), X)}{\partial z_1} - \frac{\partial m_0(y_\tau, P(Z, X, \alpha_0), X)}{\partial z_1} \right] = \int \gamma_{m_0}(o) d\hat{F}_{m_0}(o).$$

2. The following approximation holds

$$\int \gamma_{m_0}(o) d\hat{F}_{m_0}(o) = \frac{1}{n} \sum_{i=1}^n \gamma_{m_0}(O_i) + o_p(n^{-1/2}).$$

It can be shown that $\gamma_{m_0}(\cdot)$ equals $\psi_{m_0}(\cdot)$ defined in the lemma (c.f., Proposition 5 of Newey (1994)).

For a parametric submodel F_θ, we then have, when θ is close enough to θ_0:

$$\frac{1}{\theta - \theta_0} E \left[\frac{\partial m_\theta(y_\tau, P(Z, X, \alpha_0), X)}{\partial z_1} - \frac{\partial m_0(y_\tau, P(Z, X, \alpha_0), X)}{\partial z_1} \right]$$
$$= \frac{1}{\theta - \theta_0} \int \gamma_{m_0}(o) d \left[F_\theta(o) - F_{\theta_0}(o) \right],$$

where we have used $E[\gamma_{m_0}(O)] = \int \gamma_{m_0}(o) dF_{\theta_0}(o) = 0$. Under Condition (a.iv), $\int \gamma_{m_0}(o)^2 dF_\theta(o) < \infty$ for $\theta \in \Theta_0$. Then, by Lemma 7.2 in Ibragimov and Hasminskii (1981), the second derivative in

(A.24) exists and satisfies

$$\frac{\partial}{\partial\theta}E\left[\frac{\partial m_\theta(y_\tau,P(Z,X,\alpha_0),X)}{\partial z_1}\right]\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\int \gamma_{m_0}(o)dF_\theta(o)\Big|_{\theta=\theta_0} = E[\gamma_{m_0}(O)S(O)].$$

This shows that $\gamma_{m_0}(o)$ is the influence function of $E\left[\frac{\partial\hat{m}(y_\tau,P(Z,X,\alpha_0),X)}{\partial z_1}\right]$. That is,

$$E\left[\frac{\partial\hat{m}(y_\tau,P(Z,X,\alpha_0),X)}{\partial z_1}\right] - E\left[\frac{\partial m_0(y_\tau,P(Z,X,\alpha_0),X)}{\partial z_1}\right] = \frac{1}{n}\sum_{i=1}^n \gamma_{m_0}(O_i) + o_p(n^{-1/2}).$$

This, combined with the stochastic equicontinuity assumption, implies that

$$T_{2n}(y_\tau,\hat{m},\alpha_0) - T_{2n}(y_\tau,m_0,\alpha_0) = \frac{1}{n}\sum_{i=1}^n \gamma_{m_0}(O_i) + o_p(n^{-1/2}).$$

Now, the dominating condition in Condition (d) ensures that the third derivative in (A.24) exists and

$$\frac{\partial}{\partial\theta}E\left[\frac{\partial m_0(y_{\tau,\theta},\tilde{W}(\alpha_0))}{\partial z_1}\right]\Big|_{\theta=\theta_0} = E\left[\frac{\partial^2 m_0(y_\tau,\tilde{W}(\alpha_0))}{\partial y_\tau \partial z_1}\right]\frac{\partial y_{\tau,\theta}}{\partial\theta}\Big|_{\theta=\theta_0}.$$

Given the approximation

$$\hat{y}_\tau - y_\tau = \mathbb{P}_n\psi_Q(y_\tau) + o_p(n^{-1/2}),$$

from Lemma 4, we have

$$\frac{\partial y_{\tau,\theta}}{\partial\theta}\Big|_{\theta=\theta_0} = E\left[\psi_Q(y_\tau)S(O)\right].$$

Hence,

$$\frac{\partial}{\partial\theta}E\left[\frac{\partial m_0(y_{\tau,\theta},\tilde{W}(\alpha_0))}{\partial z_1}\right]\Big|_{\theta=\theta_0} = E\left[\frac{\partial^2 m_0(y_\tau,\tilde{W}(\alpha_0))}{\partial y_\tau \partial z_1}\right]E\left[\psi_Q(y_\tau)S(O)\right].$$

This gives us the contribution from the estimation of y_τ. Alternatively, this expression gives us the influence function of

$$E\left[\frac{\partial m_0(\hat{y}_\tau,\tilde{W}(\alpha_0))}{\partial z_1}\right] - E\left[\frac{\partial m_0(y_\tau,\tilde{W}(\alpha_0))}{\partial z_1}\right],$$

because

$$E\left[\frac{\partial m_0(\hat{y}_\tau,\tilde{W}(\alpha_0))}{\partial z_1}\right] - E\left[\frac{\partial m_0(y_\tau,\tilde{W}(\alpha_0))}{\partial z_1}\right] = E\left[\frac{\partial^2 m_0(y_\tau,\tilde{W}(\alpha_0))}{\partial y_\tau \partial z_1}\right](\hat{y}_\tau - y_\tau) + o_p(n^{-1/2}).$$

Using the stochastic equicontinuity assumption, we then get that

$$
\begin{aligned}
T_{2n}(\hat{y}_\tau,m_0,\alpha_0) - T_{2n}(y_\tau,m_0,\alpha_0) &= E\left[\frac{\partial^2 m_0(y_\tau,\tilde{W}(\alpha_0))}{\partial y_\tau \partial z_1}\right](\hat{y}_\tau - y_\tau) + o_p(n^{-1/2})\\
&= E\left[\frac{\partial f_{Y|\tilde{W}(\alpha_0)}(y_\tau|\tilde{W}(\alpha_0))}{\partial z_1}\right]\mathbb{P}_n\psi_Q(y_\tau) + o_p(n^{-1/2})\\
&= \mathbb{P}_n\tilde{\psi}_Q(y_\tau) + o_p(n^{-1/2}).
\end{aligned}
$$

To sum up, we have shown that

$$
\begin{aligned}
T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) - T_2 &= [T_{2n}(y_\tau, m_0, \alpha_0) - T_2] + [T_{2n}(y_\tau, \hat{m}, \alpha_0) - T_{2n}(y_\tau, m_0, \alpha_0)] \\
&\quad + [T_{2n}(\hat{y}_\tau, m_0, \alpha_0) - T_{2n}(y_\tau, m_0, \alpha_0)] + o_p(n^{-1/2}) \\
&= \mathbb{P}_n \psi_{\partial m_0} + \mathbb{P}_n \psi_{m_0} + \mathbb{P}_n \tilde{\psi}_Q(y_\tau) + o_p(n^{-1/2}).
\end{aligned}
$$

∎

Proof of Theorem 2. Consider the following difference

$$
\begin{aligned}
\Pi_{\tau,z} - \hat{\Pi}_{\tau,z} &= \frac{1}{\hat{f}_Y(\hat{y}_\tau)} \frac{T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})}{T_{1n}(\hat{\alpha})} - \frac{1}{f_Y(y_\tau)} \frac{T_2}{T_1} \\
&= \frac{T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) f_Y(y_\tau) T_1 - \hat{f}_Y(\hat{y}_\tau) T_{1n}(\hat{\alpha}) T_2}{\hat{f}_Y(\hat{y}_\tau) T_{1n}(\hat{\alpha}) f_Y(y_\tau) T_1} \\
&= \frac{f_Y(y_\tau) T_1}{\hat{f}_Y(\hat{y}_\tau) T_{1n}(\hat{\alpha}) f_Y(y_\tau) T_1} [T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) - T_2] \\
&\quad - \frac{\hat{f}_Y(\hat{y}_\tau) T_{1n}(\hat{\alpha}) T_2 - f_Y(y_\tau) T_1 T_2}{\hat{f}_Y(\hat{y}_\tau) T_{1n}(\hat{\alpha}) f_Y(y_\tau) T_1} \\
&= \frac{f_Y(y_\tau) T_1}{\hat{f}_Y(\hat{y}_\tau) T_{1n}(\hat{\alpha}) f_Y(y_\tau) T_1} [T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) - T_2] \\
&\quad - \frac{T_{1n}(\hat{\alpha}) T_2}{\hat{f}_Y(\hat{y}_\tau) T_{1n}(\hat{\alpha}) f_Y(y_\tau) T_1} \left[\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau) \right] \\
&\quad - \frac{T_2 f_Y(y_\tau)}{\hat{f}_Y(\hat{y}_\tau) T_{1n}(\hat{\alpha}) f_Y(y_\tau) T_1} [T_{1n}(\hat{\alpha}) - T_1]. \tag{A.25}
\end{aligned}
$$

We can rearrange (A.25) as

$$
\begin{aligned}
\hat{\Pi}_{\tau,z} - \Pi_{\tau,z} &= \frac{T_2}{\hat{f}_Y(\hat{y}_\tau) f_Y(y_\tau) T_1} \left[\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau) \right] \\
&\quad + \frac{T_2}{\hat{f}_Y(\hat{y}_\tau) T_{1n}(\hat{\alpha}) T_1} [T_{1n}(\hat{\alpha}) - T_1] \\
&\quad - \frac{1}{\hat{f}_Y(\hat{y}_\tau) T_{1n}(\hat{\alpha})} [T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) - T_2]. \tag{A.26}
\end{aligned}
$$

By appropriately defining the remainders, we can express (A.26) as

$$
\begin{aligned}
\hat{\Pi}_{\tau,z} - \Pi_{\tau,z} &= \frac{T_2}{f_Y(y_\tau)^2 T_1} \left[\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau) \right] \\
&\quad + \frac{T_2}{f_Y(y_\tau) T_1^2} [T_{1n}(\hat{\alpha}) - T_1] \\
&\quad - \frac{1}{f_Y(y_\tau) T_1} [T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha}) - T_2] + R_1 + R_2 + R_3. \tag{A.27}
\end{aligned}
$$

The definitions of R_1, R_2 and R_3 can be found below. Now we are ready to separate the contribution

19

of each stage of the estimation. We shall use equations (25), (27), and (31).

$$
\begin{aligned}
\hat{\Pi}_{\tau,z} - \Pi_{\tau,z} &= \frac{T_2}{f_Y(y_\tau)^2 T_1}\left[\hat{f}_Y(y_\tau) - f_Y(y_\tau)\right] + \frac{T_2}{f_Y(y_\tau)^2 T_1}\left[f_Y(\hat{y}_\tau) - f_Y(y_\tau)\right] \\
&+ \frac{T_2}{f_Y(y_\tau)T_1^2}\left[T_{1n}(\alpha_0) - T_1\right] + \frac{T_2}{f_Y(y_\tau)T_1^2}\left\{E\left[\frac{\partial P(W,\hat{\alpha})}{\partial z_1}\right] - T_1\right\} \\
&- \frac{1}{f_Y(y_\tau)T_1}\left[T_{2n}(y_\tau,m_0,\alpha_0) - T_2\right] \\
&- \frac{1}{f_Y(y_\tau)T_1}\left[T_{2n}(y_\tau,\hat{m},\alpha_0) - T_{2n}(y_\tau,m_0,\alpha_0)\right] \\
&- \frac{1}{f_Y(y_\tau)T_1}\left[T_{2n}(\hat{y}_\tau,m_0,\alpha_0) - T_{2n}(y_\tau,m_0,\alpha_0)\right] \\
&+ R_1 + R_2 + R_3 + \frac{T_2}{f_Y(y_\tau)^2 T_1}R_{f_Y} + o_p(n^{-1/2}).
\end{aligned}
\tag{A.28}
$$

Finally, we establish the rate for the remainders R_1, R_2 and R_3 in (A.27). We deal with each component of the remainder separately. The first remainder is

$$
\begin{aligned}
R_1 &= \frac{T_2}{\hat{f}_Y(\hat{y}_\tau)f_Y(y_\tau)T_1}\left[\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)\right] - \frac{T_2}{f_Y(y_\tau)^2 T_1}\left[\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)\right] \\
&= \frac{T_2}{f_Y(y)T_1}\left[\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)\right]\left[\frac{1}{\hat{f}_Y(\hat{y}_\tau)} - \frac{1}{f_Y(y_\tau)}\right] \\
&= -\frac{T_2}{f_Y(y)^2 T_1 \hat{f}_Y(\hat{y}_\tau)}\left[\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)\right]^2 \\
&= O_p\left(|\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)|^2\right).
\end{aligned}
\tag{A.29}
$$

The second remainder is

$$
\begin{aligned}
R_2 &= \frac{T_2}{\hat{f}_Y(\hat{y}_\tau)T_{1n}(\hat{\alpha})T_1}\left[T_{1n}(\hat{\alpha}) - T_1\right] - \frac{T_2}{f_Y(y_\tau)T_1^2}\left[T_{1n}(\hat{\alpha}) - T_1\right] \\
&= \frac{T_2}{T_1}\left[T_{1n}(\hat{\alpha}) - T_1\right]\left[\frac{1}{\hat{f}_Y(\hat{y}_\tau)T_{1n}(\hat{\alpha})} - \frac{1}{f_Y(y_\tau)T_1}\right] \\
&= \frac{T_2}{T_1}\left[T_{1n}(\hat{\alpha}) - T_1\right]\left[\frac{f_Y(y_\tau)T_1 - \hat{f}_Y(\hat{y}_\tau)T_{1n}(\hat{\alpha})}{\hat{f}_Y(\hat{y}_\tau)T_{1n}(\hat{\alpha})f_Y(y_\tau)T_1}\right] \\
&= \frac{T_2}{T_1}\left[T_{1n}(\hat{\alpha}) - T_1\right]\left[\frac{f_Y(y_\tau)(T_1 - T_{1n}(\hat{\alpha})) - T_{1n}(\hat{\alpha})(\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau))}{\hat{f}_Y(\hat{y}_\tau)T_{1n}(\hat{\alpha})f_Y(y_\tau)T_1}\right] \\
&= O_p\left(|T_{1n}(\hat{\alpha}) - T_1|^2\right) + O_p\left(|T_{1n}(\hat{\alpha}) - T_1||\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)|\right) \\
&= O_p\left(n^{-1}\right) + O_p\left(n^{-1/2}|\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)|\right).
\end{aligned}
\tag{A.30}
$$

The third remainder is

$$
\begin{aligned}
R_3 &= \frac{1}{f_Y(y_\tau)T_1}\left[T_{2n}(\hat{y}_\tau,\hat{m},\hat{\alpha})-T_2\right] - \frac{1}{\hat{f}_Y(\hat{y}_\tau)T_{1n}(\hat{\alpha})}\left[T_{2n}(\hat{y}_\tau,\hat{m},\hat{\alpha})-T_2\right] \\
&= O_p\left(|T_{2n}(\hat{y}_\tau,\hat{m},\hat{\alpha})-T_2||T_{1n}(\hat{\alpha})-T_1|\right) \\
&\quad + O_p\left(|T_{2n}(\hat{y}_\tau,\hat{m},\hat{\alpha})-T_2||\hat{f}_Y(\hat{y}_\tau)-f_Y(y_\tau)|\right) \\
&= O_p\left(n^{-1}\right)+O_p\left(n^{-1/2}|\hat{f}_Y(\hat{y}_\tau)-f_Y(y_\tau)|\right) \tag{A.31}
\end{aligned}
$$

because it has the same denominator as R_2 in (A.30). Finally, we compute the rate for $\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)$. To do so, we use the results in Lemma 5. Equation (25) tells us

$$
\begin{aligned}
\hat{f}_Y(\hat{y}_\tau)-f_Y(y_\tau) &= \hat{f}_Y(y_\tau)-f_Y(y_\tau)+f_Y(\hat{y}_\tau)-f_Y(y_\tau)+R_{f_Y} \\
&= O_p(n^{-1/2}h^{-1/2})+O_p(h^2)+O_p(n^{-1/2})+O_p(|R_{f_Y}|) \\
&= O_p(n^{-1/2}h^{-1/2})+O_p(h^2). \tag{A.32}
\end{aligned}
$$

The remainder R_Π is defined as

$$
R_\Pi := R_1 + R_2 + R_3 + \frac{T_2}{f_Y(y_\tau)^2 T_1}R_{f_Y} + o_p(n^{-1/2}) + o_p(h^2).
$$

So,

$$
\begin{aligned}
R_\Pi &= O_p\left(|\hat{f}_Y(\hat{y}_\tau)-f_Y(y_\tau)|^2\right)+O_p\left(n^{-1}\right)+O_p\left(n^{-1/2}|\hat{f}_Y(\hat{y}_\tau)-f_Y(y_\tau)|\right) \\
&\quad + O_p\left(|R_{f_Y}|\right)+o_p(n^{-1/2})+o_p(h^2) \\
&= O_p\left(n^{-1}h^{-1}+n^{-1}h^{-1/2}\right)+o_p\left(n^{-1/2}h^{-1/2}+h^2\right) \\
&= o_p\left(n^{-1/2}h^{-1/2}\right), \tag{A.33}
\end{aligned}
$$

where we have used the rate conditions maintained in Assumption 7. $\blacksquare$

B Asymptotic Bias of the UQR Estimator with Exogenous Treatment

Consider the model

$$
\begin{aligned}
Y(0) &= q(X)+U_0, \\
Y(1) &= q(X)+\beta+U_1, \\
D &= \mathbb{1}\{V \le \mu(Z,X)\},
\end{aligned}
$$

where $(Z,X) \in \mathbb{R}^2$ is independent of (U_0,U_1,V). For simplicity, we assume that $\mu(Z,X)=Z+X$ and consider $Z_\delta = Z+s(\delta)$. The selection equation under the new or counterfactual policy regime is

$$
D_\delta = \mathbb{1}\{V \le Z+s(\delta)+X\}.
$$

where $s(\delta)$ satisfies $E[D_\delta] = p + \delta$. Let $W = (Z, X)$ and $w = (z, x)$. We follow Corollary 4 to obtain Π_τ. The propensity score under the new regime is

$$
\begin{aligned}
P_\delta(W) &= \Pr(D_\delta = 1|W) = F_{V|W}(Z + s(\delta) + X|W) \\
&= F_V(Z + s(\delta) + X)
\end{aligned}
$$

because V is independent of both Z and X. Now,

$$
\dot{P}(W) = \left.\frac{\partial P_\delta(W)}{\partial \delta}\right|_{\delta=0} = f_V(Z + X)\left.\frac{\partial s(\delta)}{\partial \delta}\right|_{\delta=0}.
$$

Since $E[D_\delta] = p + \delta$, we have

$$
\left.\frac{\partial E[D_\delta]}{\partial \delta}\right|_{\delta=0} = \left.\frac{\partial s(\delta)}{\partial \delta}\right|_{\delta=0} E[f_V(Z + X)] = 1,
$$

which implies that $\left.\frac{\partial s(\delta)}{\partial \delta}\right|_{\delta=0} = 1/E[f_V(Z + X)]$. Therefore, $\dot{P}(W) = f_V(Z + X)/E[f_V(Z + X)]$. Next,

$$
\begin{aligned}
&E\left[\mathbf{1}\left\{Y(0) \le y_\tau\right\}|V = \mu(w), W = w\right] \\
&= E\left[\mathbf{1}\left\{q(x) + U_0 \le y_\tau\right\}|V = \mu(w), W = w\right] \\
&= E\left[\mathbf{1}\left\{q(x) + U_0 \le y_\tau\right\}|V = \mu(w)\right] \\
&= F_{U_0|V}(y_\tau - q(x)|\mu(w)).
\end{aligned}
$$

Similarly,

$$
E\left[\mathbf{1}\left\{Y(1) \le y_\tau\right\}|V = \mu(w), W = w\right] = F_{U_1|V}(y_\tau - q(x) - \beta|\mu(w)).
$$

Using Corollary 4 and assuming that $f_{U_0|V} = f_{U_1|V}$, we have

$$
\Pi_\tau = \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[\int_{y_\tau-q(x)-\beta}^{y_\tau-q(x)} f_{U_0|V}(u|\mu(w))du\right] \dot{P}(w)f_W(w)dw.
$$

It follows from Corollary 3 that the apparent effect is

$$
\begin{aligned}
A_\tau &= \frac{1}{f_Y(y_\tau)} \int_{-\infty}^{\infty} E\left[\mathbf{1}\left\{Y \le y_\tau\right\}|D = 0, W = w\right] f_W(w)dw \\
&\quad - \frac{1}{f_Y(y_\tau)} \int_{-\infty}^{\infty} E\left[\mathbf{1}\left\{Y \le y_\tau\right\}|D = 1, W = w\right] f_W(w)dw.
\end{aligned}
$$

Using the explicit forms of the potential outcomes, we get

$$E\left[\mathbb{1}\left\{Y \le y_\tau\right\}|D = 0, W = w\right]$$
$$= E\left[\mathbb{1}\left\{Y(0) \le y_\tau\right\}|D = 0, W = w\right]$$
$$= E\left[\mathbb{1}\left\{q(X) + U_0 \le y_\tau\right\}|D = 0, W = w\right]$$
$$= E\left[\mathbf{1}\left\{q(X) + U_0 \le y_\tau\right\}|V > \mu(w), W = w\right]$$
$$= \Pr(U_0 \le y_\tau - q(x)|V > \mu(w))$$
$$= \frac{\int_{\mu(w)}^\infty \left[\int_{-\infty}^{y_\tau - q(x)} f_{U_0|V}(u|v)du\right] f_V(v)\, dv}{1 - F_V(\mu(w))}$$

and

$$E\left[\mathbb{1}\left\{Y \le y_\tau\right\}|D = 1, W = w\right]$$
$$= E\left[\mathbb{1}\left\{Y(1) \le y_\tau\right\}|D = 1, W = w\right]$$
$$= E\left[\mathbf{1}\left\{q(X) + \beta + U_1 \le y_\tau\right\}|D = 1, W = w\right]$$
$$= E\left[\mathbf{1}\left\{q(X) + \beta + U_1 \le y_\tau\right\}|V \le \mu(W), W = w\right]$$
$$= \Pr(U_1 \le y_\tau - q(x) - \beta|V \le \mu(w))$$
$$= \frac{\int_{-\infty}^{\mu(w)} \left[\int_{-\infty}^{y_\tau - q(x) - \beta} f_{U_1|V}(u|v)du\right] f_V(v)\, dv}{F_V(\mu(w))}.$$

Hence

$$A_\tau = \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \frac{\int_{\mu(w)}^\infty \left[\int_{-\infty}^{y_\tau - q(x)} f_{U_0|V}(u|v)du\right] f_V(v)\, dv}{1 - F_V(\mu(w))} f_W(w)dw$$
$$- \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \frac{\int_{-\infty}^{\mu(w)} \left[\int_{-\infty}^{y_\tau - q(x) - \beta} f_{U_1|V}(u|v)du\right] f_V(v)\, dv}{F_V(\mu(w))} f_W(w)dw.$$

Now, for the scaling factor $f_Y(\cdot)$, we have a mixture

$$f_Y(y_\tau) = f_{Y(1)|D}(y_\tau|1)\Pr(D = 1) + f_{Y(0)|D}(y_\tau|0)\Pr(D = 0).$$

The mixing weights are $\Pr(D = 1) = \Pr(V \le \mu(W))$ and $\Pr(D = 0) = \Pr(V > \mu(W))$. To obtain the mixing densities, we note that

$$F_{Y(0)|D}(y_\tau|0) = \Pr(Y(0) \le y_\tau|D = 0) = \frac{\Pr(Y(0) \le y_\tau, D = 0)}{\Pr(D = 0)}$$
$$= \frac{1}{\Pr(D = 0)} \Pr(q(X) + U_0 \le y_\tau, V > \mu(W))$$
$$= \frac{1}{\Pr(D = 0)} \int_{\mathcal{W}} \Pr(q(x) + U_0 \le y_\tau, V > \mu(w)) f_W(w)dw$$
$$= \frac{1}{\Pr(D = 0)} \int_{\mathcal{W}} \left[F_{U_0}(y_\tau - q(x)) - F_{U_0,V}(y_\tau - q(x), \mu(w))\right] f_W(w)dw.$$

Hence, the density $f_{Y(0)|D}(y_\tau|0)$ is

$$f_{Y(0)|D}(y_\tau|0) = \frac{1}{\Pr(D=0)} \int_W \left[f_{U_0}(y_\tau - q(x)) - \int_{-\infty}^{\mu(w)} f_{U_0,V}(y_\tau - q(x), \breve{w}) d\breve{w} \right] f_W(w) dw.$$

For the other case, we have

$$
\begin{aligned}
F_{Y(1)|D}(y_\tau|1) &= \Pr(Y(1) \le y_\tau | D = 1) = \frac{\Pr(Y(1) \le y_\tau, D = 1)}{\Pr(D = 1)} \\
&= \frac{1}{\Pr(D=1)} \Pr(q(X) + \beta + U_1 \le y_\tau, V \le \mu(W)) \\
&= \frac{1}{\Pr(D=1)} \int_W \Pr(q(x) + \beta + U_1 \le y_\tau, V \le \mu(w)) f_W(w) dw \\
&= \frac{1}{\Pr(D=1)} \int_W F_{U_1,V}(y_\tau - q(x) - \beta, \mu(w)) f_W(w) dw,
\end{aligned}
$$

and so the density $f_{Y(1)|D}(y_\tau|1)$ is

$$f_{Y(1)|D}(y_\tau|1) = \frac{1}{\Pr(D=1)} \int_W \left[\int_{-\infty}^{\mu(w)} f_{U_1,V}(y_\tau - q(x) - \beta, \breve{w}) d\breve{w} \right] f_W(w) dw.$$

The density $f_Y(y_\tau)$ is then

$$
\begin{aligned}
f_Y(y_\tau) &= \int_W \left[f_{U_0}(y_\tau - q(x)) - \int_{-\infty}^{\mu(w)} f_{U_0,V}(y_\tau - q(x), \breve{w}) d\breve{w} \right] f_W(w) dw \\
&\quad + \int_W \left[\int_{-\infty}^{\mu(w)} f_{U_1,V}(y_\tau - q(x) - \beta, \breve{w}) d\breve{w} \right] f_W(w) dw.
\end{aligned}
$$

To compute Π_τ and A_τ numerically, we set $\beta = 1$, $q(X) = X$, $\mu(Z, X) = Z + X$, and we assume that X and Z are independent standard normals and that (U_0, U_1, V) is normal with mean 0 and variance

$$\Sigma = \begin{bmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix},$$

where ρ is the correlation between U_0 and V, and between U_1 and V. Different values of ρ lead to different degrees of endogeneity. When $\rho = 0$, the treatment selection is exogenous.

Figure 2 plots the asymptotic bias, computed as $B_\tau := A_\tau - \Pi_\tau$, as a function of τ for $\rho = 0, .25, .5, .75, .9$. We can see that the asymptotic bias is not uniform across quantiles. Hence, any attempt to sign the bias based on the "sign" or degree of the endogeneity (i.e., the sign or magnitude of ρ) is futile.

When $\rho = 0$, we only have *marginal heterogeneity bias* because the *marginal selection bias* 0, as

$$
\begin{aligned}
F_{Y(0)|D,W}(y_\tau|0, w) &= F_{Y(0)|U_D,W}(y_\tau|P(w), w) = F_{Y(0)|W}(y_\tau|w), \\
F_{Y(1)|D,W}(y_\tau|1, w) &= F_{Y(1)|U_D,W}(y_\tau|P(w), w) = F_{Y(1)|W}(y_\tau|w).
\end{aligned}
$$

The *marginal heterogeneity bias* is

$$
\begin{aligned}
B_{1\tau} &= \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y|D,W}(y_\tau|1,w) - F_{Y|D,W}(y_\tau|0,w) \right] \dot{P}(w) f_W(w)\, dw \\
&\quad - \frac{1}{f_Y(y_\tau)} \int_{\mathcal{W}} \left[F_{Y|D,W}(y_\tau|1,w) - F_{Y|D,W}(y_\tau|0,w) \right] f_W(w)\, dw.
\end{aligned}
$$

In this example, $\dot{P}(w) \neq 1$ and

$$
F_{Y|D,W}(y_\tau|1,w) - F_{Y|D,W}(y_\tau|0,w) = \Pr(X + \beta + U_1 < y_\tau) - \Pr(X + U_0 < y_\tau) \neq 0
$$

for all $\beta \neq 0$. So the *marginal heterogeneity bias* $B_{1\tau}$ is not equal to 0 when $\beta \neq 0$.

C   Estimation of the Asymptotic Variance of the UNIQUE

The asymptotic variance V_τ in (34) can be estimated by the plug-in estimator

$$
\hat{V}_\tau = \frac{h}{n} \sum_{i=1}^n \hat{\psi}_{\Pi_\tau, i}^2, \tag{A.34}
$$

where, by Theorem 2,

$$
\begin{aligned}
\hat{\psi}_{\Pi_\tau, i} &= \frac{\hat{T}_{2n}}{\hat{f}_Y(\hat{y}_\tau)^2 \hat{T}_{1n}} \hat{\psi}_{f,i}(\hat{y}_\tau) + \frac{\hat{T}_{2n}}{\hat{f}_Y(\hat{y}_\tau)^2 \hat{T}_{1n}} \hat{f}_Y'(\hat{y}_\tau) \hat{\psi}_{Q,i}(\hat{y}_\tau) \\
&\quad + \frac{\hat{T}_{2n}}{\hat{f}_Y(\hat{y}_\tau) \hat{T}_{1n}^2} \hat{\psi}_{\partial P, i} + \frac{\hat{T}_{2n}}{\hat{f}_Y(\hat{y}_\tau) \hat{T}_{1n}^2} \left[\frac{1}{n} \sum_{j=1}^n \frac{\partial^2 P(W_j, \hat{\alpha})}{\partial z_1 \partial \hat{\alpha}'} \right] \hat{\psi}_{\alpha, i} \\
&\quad - \frac{1}{\hat{f}_Y(\hat{y}_\tau) \hat{T}_{1n}} \hat{\psi}_{\partial m, i} - \frac{1}{\hat{f}_Y(\hat{y}_\tau) \hat{T}_{1n}} \hat{\psi}_{m, i} - \frac{1}{\hat{f}_Y(\hat{y}_\tau) \hat{T}_{1n}} \hat{E} \left[\frac{\partial f_{Y|\tilde{W}(\alpha_0)}(\hat{y}_\tau|\tilde{W}(\hat{\alpha}))}{\partial z_1} \right] \hat{\psi}_{Q, i}(\hat{y}_\tau).
\end{aligned}
$$

In this equation, $\hat{T}_{2n} = T_{2n}(\hat{y}_\tau, \hat{m}, \hat{\alpha})$, $\hat{T}_{1n} = T_{1n}(\hat{\alpha})$,

$$
\begin{aligned}
\hat{\psi}_{f,i}(\hat{y}_\tau) &= K_h([Y_i - \hat{y}_\tau]) - \frac{1}{n} \sum_{j=1}^n K_h([Y_j - \hat{y}_\tau]) \\
\hat{\psi}_{Q,i}(\hat{y}_\tau) &= \frac{\tau - \mathbf{1}\{Y_i \leq \hat{y}_\tau\}}{\hat{f}_Y(\hat{y}_\tau)}, \\
\hat{\psi}_{\partial P,i} &= \frac{\partial P(W_i, \hat{\alpha})}{\partial z_1} - \frac{1}{n} \sum_{j=1}^n \frac{\partial P(W_j, \hat{\alpha})}{\partial z_1}, \\
\hat{\psi}_{\alpha,i} &= \left(-\frac{1}{n} \sum_{j=1}^n \frac{[P_\partial(W_j, \hat{\alpha})]^2 W_j W_j'}{P(W_j, \hat{\alpha})[1 - P(W_j, \hat{\alpha})]} \right)^{-1} \frac{P_\partial(W_i, \hat{\alpha}) W_i [D_i - P(W_i, \hat{\alpha})]}{P(W_i, \hat{\alpha})[1 - P(W_i, \hat{\alpha})]}, \\
\hat{\psi}_{\partial m,i} &= \frac{\partial \hat{m}(\hat{y}_\tau, \tilde{W}_i(\hat{\alpha}))}{\partial z_1} - \hat{T}_{2n} \\
\hat{\psi}_{m,i} &= -\left(\mathbf{1}\{Y_i \leq \hat{y}_\tau\} - m_0(\hat{y}_\tau, \tilde{W}_i(\hat{\alpha}))\right) \times \hat{E}\left[\frac{\partial \log f_W(W_i)}{\partial z_1} \bigg| \tilde{W}_i(\hat{\alpha}) \right],
\end{aligned}
$$

and

$$\hat{E}\left[\frac{\partial f_{Y|\tilde{W}(\alpha_0)}(\hat{y}_\tau|\tilde{W}(\hat{\alpha}))}{\partial z_1}\right] = \frac{1}{n}\sum_{i=1}^n \left.\frac{\partial \hat{f}_{Y|\tilde{W}(\hat{\alpha}),X}(\hat{y}_\tau|p,X_i)}{\partial p}\right|_{p=P(W_i,\hat{\alpha})} \left.\frac{\partial P(z,X_i,\hat{\alpha})}{\partial z_1}\right|_{z=Z_i}.$$

Most of these plug-in estimates are self-explanatory. For example, $\hat{\psi}_{\alpha,i}$ is the estimated influence function for the MLE when $P(W_i,\hat{\alpha}) = P(W_i\hat{\alpha})$ and $P_\partial(a) = \partial P(a)/\partial a$. If the propensity score function does not take a linear index form, then we need to make some adjustment to $\hat{\psi}_{\alpha,i}$. We only need to find the influence function for the MLE, which is an easy task, and then plug $\hat{\alpha}$ into the influence function.

The only remaining quantity that needs some explanation is $\hat{\psi}_{m,i}$, which involves a nonparametric regression of $\frac{\partial \log f_W(W_i)}{\partial z_1}$ on $\tilde{W}_i(\hat{\alpha}) := [P(W_i,\hat{\alpha}), X_i]$. We let

$$\hat{E}\left[\frac{\partial \log f_W(W_i)}{\partial z_1}\middle| \tilde{W}_i(\hat{\alpha})\right] = -\phi^J(\tilde{W}_i(\hat{\alpha}))' \left(\sum_{\ell=1}^n \phi^J(\tilde{W}_\ell(\hat{\alpha}))\phi^J(\tilde{W}_\ell(\hat{\alpha}))'\right)^{-1} \sum_{\ell=1}^n \frac{\partial \phi^J(\tilde{W}_\ell(\hat{\alpha}))}{\partial z_1}.$$

To see why this may be consistent for $E\left[\frac{\partial \log f(W_i)}{\partial z_1}\middle| \tilde{W}_i(\alpha_0)\right]$, we note that using integration by parts, the above is just a series approximation to $E\left[\frac{\partial \log f_W(W_i)}{\partial z_1}\middle| \tilde{W}_i(\alpha_0)\right]$.

The consistency of $\hat{V}_\tau$ can be established by using the uniform law of large numbers. The arguments are standard but tedious. We omit the details here.

D Unconditional Instrumental Quantile Estimation under Nonparametric Propensity Score

In this section, we drop Assumption 8 and estimate the propensity score non-parametrically using the series method. With respect to the results in Section 5, we only need to modify Lemma 6, since Lemma 8 shows that we do not need to account for the error from estimating the propensity score.

Let $\hat{P}(w)$ denote the nonparametric series estimator of $P(w)$. The estimator of $T_1 := E\left[\frac{\partial P(W)}{\partial z_1}\right]$ is now

$$T_{1n}(\hat{P}) := \frac{1}{n}\sum_{i=1}^n \left.\frac{\partial \hat{P}(w)}{\partial z_1}\right|_{w=W_i}.$$

The estimator of T_2 is the same as in (28) but with $P(W_i,\hat{\alpha})$ replaced by $\hat{P}(W_i)$:

$$T_{2n}(\hat{y}_\tau,\hat{m},\hat{P}) := \frac{1}{n}\sum_{i=1}^n \frac{\partial \hat{m}(\hat{y}_\tau,\hat{P}(W_i),X_i)}{\partial z_1}, \tag{A.35}$$

where, as in (28), $\hat{m}$ is the series estimator of m. The formula is the same as before, and we only need to replace $P(W_i,\hat{\alpha})$ by $\hat{P}(W_i)$. The nonparametric UNIQUE becomes

$$\begin{aligned}
\hat{\Pi}_\tau(\hat{y}_\tau,\hat{f}_Y,\hat{m},\hat{P}) :=& \quad -\frac{1}{\hat{f}_Y(\hat{y}_\tau)}\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \hat{P}(W_i)}{\partial z_1}\right]^{-1} \frac{1}{n}\sum_{i=1}^n \frac{\partial \hat{m}(\hat{y}_\tau,\hat{P}(W_i),X_i)}{\partial z_1} \\
=& \quad -\frac{1}{\hat{f}_Y(\hat{y}_\tau)}\frac{T_{2n}(\hat{y}_\tau,\hat{m},\hat{P})}{T_{1n}(\hat{P})}.
\end{aligned} \tag{A.36}$$

26

The following lemma follows directly from Theorem 7.2 of Newey (1994).

Lemma 9. *Let Assumption (a) of Lemma 7 and Assumptions (a) and (c) of Lemma 8 hold. Assume further that $P(z, x)$ is continuously differentiable with respect to z_1 for all orders, and that there is a constant C such that $\partial^\ell P(z, x)/\partial z_1^\ell \leq C^\ell$ for all $\ell \in \mathbb{N}$. Then*

$$T_{1n}(\hat{P}) - T_1 = \mathbb{P}_n \psi_{\partial P_s} + \mathbb{P}_n \psi_{P_s} + o_p(n^{-1/2}),$$

where we define

$$\psi_{\partial P_s} := \frac{\partial P(W)}{\partial z_1} - T_1$$

and

$$\psi_{P_s} := -(D - P(W)) \times \frac{\partial \log f_W(W)}{\partial z_1}.$$

Using a proof similar to that of Lemma 8, we can show that $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{P})$ and $T_{2n}(\hat{y}_\tau, \hat{m}, P)$ have the same influence function. That is, we have

$$T_{2n}(\hat{y}_\tau, \hat{m}, \hat{P}) - T_2 = \mathbb{P}_n \psi_{\partial m_0} + \mathbb{P}_n \psi_{m_0} + \mathbb{P}_n \tilde{\psi}_Q(y_\tau) + o_p(n^{-1/2}),$$

where

$$\psi_{\partial m_0} := \frac{\partial m_0(y_\tau, P(W), X)}{\partial z_1} - T_2,$$

$$\psi_{m_0} := -(\mathbf{1}\{Y \leq y_\tau\} - m_0(y_\tau, P(W), X)) \times E\left[\frac{\partial \log f_W(W)}{\partial z_1}\middle| P(W), X\right],$$

and

$$\tilde{\psi}_Q(y_\tau) = E\left[\frac{\partial f_{Y|P(W),X}(y_\tau|P(W), X)}{\partial z_1}\right] \psi_Q(y_\tau).$$

Given the asymptotic linear representations of $T_{1n}(\hat{P}) - T_1$ and $T_{2n}(\hat{y}_\tau, \hat{m}, \hat{P}) - T_2$, we can directly use Lemma 8, together with Lemma 5, to obtain an asymptotic linear representation of $\hat{\Pi}_\tau(\hat{y}_\tau, \hat{f}_Y, \hat{m}, \hat{P})$.

Theorem 3. *Under the assumptions of Lemmas 5, 8, and 9, we have*

$$
\begin{aligned}
\hat{\Pi}_\tau - \Pi_\tau &= \frac{T_2}{f_Y(y_\tau)^2 T_1}\left[\mathbb{P}_n \psi_{f_Y}(y_\tau) + B_{f_Y}(y_\tau)\right] + \frac{T_2}{f_Y(y_\tau)^2 T_1} f_Y'(y_\tau)\mathbb{P}_n \psi_Q(y_\tau)\\
&+ \frac{T_2}{f_Y(y_\tau)T_1^2}\mathbb{P}_n \psi_{\partial P_s} + \frac{T_2}{f_Y(y_\tau)T_1^2}\mathbb{P}_n \psi_{P_s}\\
&- \frac{1}{f_Y(y_\tau)T_1}\mathbb{P}_n \psi_{\partial m_0} - \frac{1}{f_Y(y_\tau)T_1}\mathbb{P}_n \psi_{m_0}\\
&- \frac{1}{f_Y(y_\tau)T_1}\mathbb{P}_n \tilde{\psi}_Q(y_\tau) + R_\Pi,
\end{aligned}
\tag{A.37}
$$

where

$$R_\Pi = O_p\left(|\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)|^2\right) + O_p\left(n^{-1}\right) + O_p\left(n^{-1/2}|\hat{f}_Y(\hat{y}_\tau) - f_Y(y_\tau)|\right)$$
$$+ \quad O_p\left(|R_{f_Y}|\right) + o_p(n^{-1/2}) + o_p(h^2). \tag{A.38}$$

Furthermore, under Assumption 7, $\sqrt{nh}R_\Pi = o_p(1)$.

We summarize the results of Theorem 3 in a single equation:

$$\hat{\Pi}_\tau - \Pi_\tau = \mathbb{P}_n\psi_{\Pi_\tau} + \tilde{B}_{f_Y}(y_\tau) + o_p(n^{-1/2}h^{-1/2}),$$

where ψ_{Π_τ} collects all the influence functions in (A.37) except for the bias, R_Π is absorbed in the $o_p(n^{-1/2}h^{-1/2})$ term, and

$$\tilde{B}_{f_Y}(y_\tau) := \frac{T_2}{f_Y(y_\tau)^2 T_1} B_{f_Y}(y_\tau).$$

The bias term is $o_p(n^{-1/2}h^{-1/2})$ by Assumption 7. The following corollary provides the asymptotic distribution of $\hat{\Pi}_\tau$.

Corollary 6. *Under the assumptions of Theorem 3,*

$$\sqrt{nh}\left(\hat{\Pi}_\tau - \Pi_\tau\right) = \sqrt{n}\mathbb{P}_n\sqrt{h}\psi_{\Pi_\tau} + o_p(1) \Rightarrow \mathcal{N}(0, V_\tau),$$

where

$$V_\tau = \lim_{h\downarrow 0} E\left[h\psi_{\Pi_\tau}^2\right].$$

The asymptotic variance takes the same form as the asymptotic variance in Corollary 5. Estimating the asymptotic variance and testing for a zero unconditional effect are entirely similar to the case with a parametric propensity score. We omit the details to avoid repetition and redundancy. From the perspective of implementation, there is no substantive difference between a parametric approach and a nonparametric approach to the propensity score estimation.