Treatment Effects Estimators Under Misspecification[*]

Ping Yu[†]
The University of Hong Kong

First Version: January 2016
This Version: May 2016

Abstract

We conduct misspecification analysis for four treatment effects estimators, IVE, IV-QRE, LSE and QRE, in the framework of HV (2005). We first derive the pseudo-true values which shed more light on why the IVE and IV-QRE are consistent when the essential heterogeneity is absent and why the LSE and QRE are consistent when the selection effect is further excluded. For example, when the essential heterogeneity is absent, the IVE is consistent because of an offsetting property, while the IV-QRE is consistent due to a counterfactual-quantiles matching property inherited from rank similarity. We then check two responses to model misspecification. First, we conduct a local sensitivity analysis which provides a measure of sensitivity when the key identification assumption is locally perturbed along a prespecified direction. Second, we summarize the sharp bounds for the ATE in the literature and develop corresponding bounds for the QTE. We illustrate our analyses by studying the effect of veteran status on earnings in Angrist (1990).

Keywords: misspecification, pseudo-true value, local sensitivity analysis, path derivative, rank similarity, counterfactual-quantiles matching, partial identification, MTE, LATE, IVE, IV-QRE, LSE, QRE
JEL-Classification: C21, C26

[*] I want to thank Yu-Chin Hsu, Jack Porter, Keunkwan Ryu, Xiaojun Song, Yundong Tu and seminar participants at PKU-GSM for helpful comments. The usual disclaimer applies.

[†] School of Economics and Finance, The University of Hong Kong, Pokfulam Road, Hong Kong; email: pingyu@hku.hk.

"*Essentially, all models are wrong, but some are useful.*"

Box and Draper (1987), p. 424

1 Introduction

Any econometric model is based on econometricians' past experiences (and/or past literature which is based on further past experiences and literature), i.e., any model is the result of an interaction of subjectivity and objectivity rather than a genuine reflection of objectivity. On the other hand, econometric models are built by human beings and so a good econometric model should be helpful to aid understanding of the reality for human beings. Modeling means structures/assumptions beyond simple data description; to be more straight, modeling is a kind of belief although we try to make our belief close to the reality.[1] Certainly, belief is not reality, so any model is potentially wrong. However, to delve into the essence of a problem, modeling is sometimes inevitable. More structures are imposed in a model, potentially farther from the reality is the model.

Econometricians keep an eye on model misspecification for a long time. To adapt to the needs of applications, they use a range of models, from fully nonparametric to fully parametric. The recent trend is to impose less structures (i.e., more nonparametric) such that the misspecification problem can be alleviated or the model is more robust to misspecification. Based on my personal understanding, there are four responses to model misspecification. First, use a more structured model while admit that the model is possibly misspecified and check what the estimator would converge to under misspecification. The probability limit of the estimator is often termed as the pseudo-true value, so we label this response as *pseudo-true analysis*. This tradition dates back (in econometrics) at least to Halbert White's pioneering work in the 1980s. For example, White (1980, 1981) study the misspecification problem in the OLS estimation given that the conditional mean may not take the linear form of the covariates; White (1982) examines the consequences and detection of model misspecification when using maximum likelihood techniques for estimation and inference. Later, in the GMM framework, Hall and Inoue (2003) study the consequences of misspecification and develop the asymptotics for the pseudo–true parameters; in quantile regression, Angrist et al. (2006) study the estimation and inference in a misspecified model. Second, take a nonparametric setup but find some "smart" sufficient conditions for point identification; see, e.g., Matzkin (1994, 2007) for a summary of literature and Matzkin (2013) and Lewbel (2016) for an introduction.[2] Generally speaking, the "key" identification conditions must be carefully examined. In a nonparametric setup, such an examination seems rare (maybe because such conditions are already hard to find and also because the setup is already nonparametric), while in a more structured setup, there are usually two ways to check the validity or robustness of these conditions. The first way is to check whether these conditions are consistent with the data, so-called *misspecification testing* or *goodness-of-fit testing*. The literature dates back at least to Ramsey (1969, 1970)'s RESET in econometrics and to Karl Pearson (1900)'s chi-square test in statistics. The literature on this topic is vast, see Holly (1987), Godfrey (1988) and White (1987, 1994) for a summary of early literature and Gonzáez-Manteiga and Crujeiras (2013) for a summary of recent developments. The second way is to check the sensitivity of the identified parameter to these conditions. Such a checking seems routine for applied econometricians, see, e.g., Mroz (1987) for an early example. The treatment effect literature of sensitivity analysis dates back at least to Cornfield et al. (1959); see Glynn et al. (1986) for the Bayesian method and Rosenbaum (2002) for a variety of sensitivity analyses in the statistical literature; see Vijverberg (1993) for sensitivity analysis in a

[1] From the philosophical perspective, there is nothing called "reality" or "objectivity". Anything must involve subjectivity; otherwise, how can we understand it?

[2] Early literature of identification concentrates on identification of simultaneous equations system; see, e.g., Hausman (1983) and Hsiao (1983) for a summary of literature.

Roy model and Athey and Imbens (2015) and references therein for recent developments in the econometric literature. Third, allow the model to be partially identified, i.e., the identified objects are sets rather than points.[3] This part of literature builds on insights in early literature by Peterson (1976) and Holland (1986a) and was developed by Charles Manski and his co-authors in the 1990s and 2000s; see Manski (1995, 2003, 2007) for a summary and Tamer (2010) for an introduction to the literature. Fourth, select the true model or average a sequence of models; see Claeskens and Hjort (2008) for an introduction. Some of the four responses are interwined, e.g., the goodness-of-fit testing is closely related to model selection.

This paper tries to examine the four responses to model misspecification in the treatment effects evaluation. It is well known that one main objective of microeconometrics is to explore causal relationships between a response variable and some covariates. As a result, treatment effects evaluation is a main task for micro-econometricians. Usually, causal inference takes the potential outcomes approach rather than regards the treatment status as a usual covariate. This approach goes back at least to Neyman (1923, 1935) and Fisher (1918, 1925, 1935) on agricultural experiment, but the modern version is usually attributed to Rubin (1974, 1978), so this framework is often termed as the *Neyman-Fisher-Rubin causal model*. The literature on this topic is tremendous. Roughly speaking, one strand of literature emphasizes "effects of causes" and is more reduced-formed, and another strand emphasizes "causes of effects" and is more structural. The former is the tradition of statistics and is borrowed to econometrics, see Holland (1986b) for a historical view and Imbens and Wooldridge (2009) and Imbens and Rubin (2015) for a summary of literature on this tradition. The latter is developed by James Heckman and his co-authors in the 1980s and 1990s; see Heckman and Vytlacil (2007a, b) for a summary of relevant literature on this tradition. The distinction between these two traditions makes their interpretation of treatment effects estimates quite different; nevertheless, the econometric techniques they use are often overlapped or intertwined. Careful assessment of these two traditions is out of the scope of this paper; rather, we confine ourselves to the mathematical side of these two traditions, which seems to have more in common.

We will examine four treatment effects estimators in the framework of Heckman and Vytlacil (2005) (HV hereafter) which allows for both the selection effect (i.e., the untreated outcome may depend on the treatment status, which is the usual endogeneity problem) and the essential heterogeneity (i.e., the treatment status may depend on idiosyncratic gains, which is new for treatment effects evaluation). This framework is equivalent to the framework used in the first tradition (e.g., the setup of Imbens and Angrsit (1994) (IA hereafter)) in some sense; see Yu (2015b) for a bare comparison of these two frameworks. The four treatment effects estimators we will examine are the instrumental variable estimator (IVE), the instrumental variable quantile regression estimator (IV-QRE), the least squares estimator (LSE) and the quantile regression estimator (QRE). The IVE and LSE attempt to estimate the average treatment effect (ATE) and the IV-QRE and QRE attempt to estimate the quantile treatment effect (QTE). The IVE in this general framework is examined first by IA where the instruments are discrete. When the instrument is binary, the IVE is estimating the treatment effect for a subpopulation called compliers, so-called the *local average treatment effect* (LATE); when the instrument is multi-valued, it is estimating an average of LATEs. HV's framework allows for more general instruments so that the average treatment effect for a marginal population, so called the *marginal treatment effect* (MTE), can be identified, and the LATE can be expressed as a weighted average of MTEs.[4] Only if the essential heterogeneity is absent (or the endogeneity is present only in the form of selection effect), the IVE is estimating the ATE; otherwise, the ATE can only be partially identified. Excluding the essential heterogeneity is quite useful in practice, e.g., Angrsit and Fernández-Val (2013) essentially impose this

[3] A weaker version of partial identification is weak identification where the parameter is only "locally" unidentified, e.g., the weak instruments problem.

[4] This does not mean IA did not explore the general instruments case, see, e.g., Angrist et al. (2000) for the continuous instruments scenario, where they also show that the MTE is a limit form of the LATE.

restriction to extrapolate the LATE estimation. The IV-QRE is put forward in Chernozhukov and Hansen (2005) (CH hereafter). After imposing a key rank invariance or rank similarity assumption, they show that the QTE can be identified by the IV-QRE. We examine this key assumption carefully in this paper and show that this assumption essentially excludes the essential heterogeneity in the context of QTE. In other words, the IV-QRE is the quantile counterpart of IVE. Finally, the LSE can identify the ATE and the QRE can identify the QTE only if neither the selection effect nor the essential heterogeneity can be present. They are useful in practice because valid instruments are hard to find in applications while they can achieve identification without instruments when the required conditions are satisfied.

For each of the four estimators, we presume they are estimating the "global" treatment effects (i.e., the ATE or the QTE), so the model is misspecified. We examine the four responses to model misspecification in this specific context. First, we derive the pseudo-true value of each estimator, which allows us to shed more light on why each estimator is consistent to the parameter of interest when the required restrictions discussed above are imposed. It turns out that the IVE can identify the two counterfactual means in an intriguing way. It extrapolates the mean of an identifiable subpopulation to the unidentifiable subpopulation, which may induce bias, and meanwhile uses an weighted average of the two counterfactual means for the identifiable subpopulation to replace the genuine mean, which may also induce bias, but these two biases luckily offset each other when the essential heterogeneity is excluded so that a consistent estimator is achieved. The IV-QRE also extrapolates to achieves consistency but in a different way. It turns out that under rank similarity, the two marginal counterfactual cumulative distribution functions (cdfs), i.e., the cdfs for a marginal population, and the two population counterfactual cdfs satisfy a very lucky relationship which we label as *counterfactual-quantiles matching*; this relationship guarantees the consistency of the IV-QRE in estimating the two population counterfactual cdfs. For both estimators, we concentrate on using the propensity score as the instrument, but we also point out the difference when a general instrument is used. As to the LSE and QRE, we decompose the bias under misspecification into two components, one from the selection effect and the other from the essential heterogeneity, which makes it clear why only if both sources of endogeneity are excluded can consistency be achieved. This decomposition also makes it clear that the bias from the selection effect spreads over all subpopulations, no matter identifiable or not, which explains why the IVE and IV-QRE need extrapolation to achieve consistency. Second, we conduct a different kind of sensitivity analysis from those mentioned above, called the *local sensitivity analysis* (LSA), to assess fragility of the estimators to their corresponding key identification assumptions. This kind of analysis is meaningful since specifying a model is usually based on past experiences and the specification should be close to reality (otherwise, why use a very wrong model?); in other words, possible misspecifications are only local. Our sensitivity analysis is in the spirit of Carneiro et al. (2010). Specifically, we perturb the identified system a little bit (or locally) in a prespecified direction and check how sensitive the identified parameter is to such a local perturbation; in other words, a path derivative is developed. Such a path derivative is standard in semiparametric efficiency analysis, so the LSA can be treated as an extension of the path derivative there to a more general context. Most importantly, all components in the path derivative (excluding the direction which needs to be prespecified) can be estimated from the data, so the LSA is very practical and is suggested to become a routine analysis whenever a key identification assumption is imposed. Third, we conduct some partial identification analysis when the general framework is maintained. As mentioned, the ATE and QTE cannot be point identified now. Heckman and Vytlacil (2001b) develop sharp bounds for the ATE, and we conduct a parallel analysis and develop sharp bounds for the QTE. These bounds can be used to assess the validity of the identification assumptions.

By passing, we mention that we do not discuss the misspecification testing in this paper because there have already been some related developments in the literature. For example, Angrsit and Fernández-Val (2013) check the presence of essential heterogeneity by the popular J-test, Heckman et al. (2010) and Heckman and Schmierer (2010) develop a few conditional moment tests for the same purpose, and the famous Hausman (1978)'s test can be used to check the presence of both the selection effect and the essential heterogeneity. We do not discuss model selection or model averaging either. Recall that model selection or model averaging can be applied only if the data contain some but imprecise information on the true model. However, in the framework of HV, the data do not contain any information on the the untreated outcome of always-takers or treated outcome of never-takers, so it is impossible to trade off bias and variance of global treatment effects estimators. Nevertheless, there is some literature on covariates selection in treatment effects evaluation under unconfoundedness, see Kitagawa and Muris (2016) and references therein.

This paper is organized as follows. In Section 2, we state the assumptions which will be maintained throughout the paper and also specify a running example which will be examined for each estimator. Sections 3-6 will conduct the pseudo-true analysis, the local sensitivity analysis and the partial identification analysis for each of the four estimators, respectively. Section 7 will use an empirical example from Angrist (1990) to illustrate the analyses in Sections 3-6, and Section 8 concludes. To save space, we relegate some discussions to two supplementary materials S.1 and S.2. S.1 contains the proofs that are not given in the main text, and S.2 contains points that we do not want to expand in the main text.

2 Maintained Assumptions and A Running Example

We first state the nonlinear and nonseparable outcome model as in HV,

$$\begin{aligned} Y_1 &= \mu_1(X, U_1), \\ Y_0 &= \mu_0(X, U_0), \end{aligned} \tag{1}$$

where Y_1 and Y_0 are the potential outcome for the treated and control group, respectively, X includes all relevant covariates, and U_1 and U_0 are random errors. The participation decision

$$D = 1(\mu_D(X, Z) - V \geq 0), \tag{2}$$

where Z includes the instruments (usually some policy variables) for the choice process, V is a scalar random error in the participation decision, and $1(\cdot)$ is the indicator function. Both X and Z appearing as the arguments of μ_D does not lose generality since $\mu_D(X, Z)$ may not depend on all elements of X. By transforming $\mu_D(X, Z)$ and V by the conditional cdf $F_{V|X,Z}$, we can rewrite

$$D = 1(p(X, Z) - U_D \geq 0), \tag{3}$$

where $U_D|X, Z \sim U(0, 1)$, the uniform distribution on $[0, 1]$, and $p(X, Z)$ is the propensity score. As shown in Vytlacil (2006), there is a larger class of latent index models that will have a representation of this form. The equations in (1) imply that Z is excluded from the outcome equations; this is the usual *exclusion assumption*. Equation (3) implies the *monotonicity assumption* of IA, i.e., when Z changes from z to z', the direction of D change is the same for all individuals. Actually, as shown in Vytlacil (2002), they are equivalent in some sense; see Yu (2015b) for more discussions on the implication of (3). In the setup (1) and (3), we define the MTE as $MTE(u_D) \equiv E[Y_1 - Y_0|U_D = u_D]$, $u_D \in [0, 1]$.

For the ATE, we maintain the following assumptions.

Assumption A:

(A-1) $p(X, Z)$ is a nondegenerate random variable conditional on X.

(A-2) The random vectors (U_1, U_D) and (U_0, U_D) are independent of Z conditional on X. In Dawid (1979)'s notation, $(U_1, U_D) \perp Z|X$ and $(U_0, U_D) \perp Z|X$.
(A-3) The distribution of U_D is absolutely continuous with respect to Lebesgue measure.
(A-4) $E\left|Y_d\right| < \infty$, $d = 0, 1$.
(A-5) $1 > P(D = 1|X) > 0$.
(A-6) $X_1 = X_0$ almost everywhere, where X_d denotes a value of X if D is set to d.

These assumptions are repetition of those in HV and are prevalent in the literature of heterogeneous treatment effects. A necessary condition for (A-1) is that Z contains an extra random variable beyond X. (A-2) is the *ignorability assumption*; it allows for both the selection effect ($U_0 \not\perp D|X$) and the essential heterogeneity ($(U_1 - U_0) \not\perp D|X$). (A-3)-(A-6) are regularity assumptions, e.g., (A-6) is the "no feedback" condition which excludes the effect of D on X so conditioning on X does not mask the effects of D. Assumptions (A-1)-(A-5), combined with the setup (1) and (2), impose testable restrictions on the distribution of (Y, D, Z, X); see HV (p. 678) for the index sufficiency restriction and the monotonicity restriction. As discussed in Yu (2015b), the framework of HV above and that of IA are roughly equivalent. As emphasized in AIR (1996), the key assumptions in IA's framework are the ignorability assumption and the monotonicity assumption. Of course, the exclusion assumption is also important in some applications. See Kitagawa (2015) and the references therein for tests of joint validity of these three assumptions.

To facilitate our analysis, we impose the following assumption on the propensity score $p(X, Z)$.

Assumption P: $\text{supp}(p(X,Z)|X = x) = \left[\underline{p}_x, \overline{p}_x\right] \subset [0,1]$ for each $x \in \text{supp}(X)$, where $\text{supp}(\cdot)$ means the support of a random variable. In other words, the conditional density function of $p(X, Z)$ given $X = x$, say $f_{P(X,Z)|X}(p|x)$, is positive for $p \in \left(\underline{p}_x, \overline{p}_x\right)$.

This assumption tries to reflect the reality of the support of $p(X, Z)$. Usually, the support of $p(X, Z)$ is different for different x values and is a subset of $[0, 1]$ staying in the middle of $[0, 1]$. We can relax this assumption by allowing the support of $p(X, Z)$ to be segments of intervals without difficulty, but we find Assumption P will provide clean results without losing the essence of our problem so we maintain it throughout the paper. Note also that we implicitly assume Z includes some continuous components; otherwise, $\text{supp}(p(X,Z)|X = x)$ cannot be an interval. We focus on this general case and treat the case with only discrete instruments as a special case.

For the QTE, we do not require (A-4). Nevertheless, we replace (A-4) by some regularity conditions on the distribution of Y_d given $U_D = u_D$, $d = 0, 1$. We label the combined assumptions as

Assumption Q:

(A-1), (A-2), (A-3), (A-5) and (A-6) plus
(Q-4) $\text{supp}(Y_d|X = x, U_D = u_D) = \mathcal{Y}_{xd}$, the conditional density $f_{Y_d|X,U_D}(y_d|x, u_D)$ is continuous in (y_d, u_D) for almost every $x \in\text{supp}(X)$, $y_d \in \mathcal{Y}_{xd}$, $d = 0, 1$, and $u_D \in [0, 1]$, where $\mathcal{Y}_{x0} = \left[\underline{y}_{x0}, \overline{y}_{x0}\right]$ and $\mathcal{Y}_{x1} = \left[\underline{y}_{x1}, \overline{y}_{x1}\right]$ are compact and need not be the same.

Assumption (Q-4) implies that the support of Y_0 and Y_1 may be different as typical in applications. Note also that the support of Y_d is compact, does not depend on U_D (although may depend on X), and $f_{Y_d|X,U_D}(y_d|x, u_D) > 0$ for $y_d \in \left(\underline{y}_{xd}, \overline{y}_{xd}\right)$ and $u_D \in [0, 1]$. This implies that $\text{supp}(Y_d|X = x) = \mathcal{Y}_{xd}$, $f_{Y_d|X}(y_d|x)$ is continuous in y_d for $y_d \in \mathcal{Y}_{xd}$, and $f_{Y_d|X}(y_d|x) > 0$ for $y_d \in \left(\underline{y}_{xd}, \overline{y}_{xd}\right)$. These regularities can be relaxed in an obvious way without affecting our general results, but to make the discussion clean and avoid technical complications, we maintain assumption (Q-4) throughout the paper.

We now specify a running example that will be used for illustration in all four estimators. First, let $Y_1 = 2U$, $Y_0 = U$ and $D = 1(Z - V \geq 0)$, where

$$\begin{pmatrix} U \\ V \\ Z \end{pmatrix} \sim N(\mathbf{0}, \mathbf{\Sigma}) \text{ with } \mathbf{\Sigma} = \begin{pmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

We label this case as the selection-effect-only case. For the endogeneity-free case, change the correlation between U and V to 0. When there are both the selection effect and essential heterogeneity, let $Y_1 = V + 2U, Y_0 = 2V + U$, and all other specifications are the same as in the seletion-effect-only case. Note that $p(Z) = E[D|Z] = \int_{-\infty}^{Z} \phi(v)dv = \Phi(Z) \sim U[0,1]$, where we use $\Phi(\cdot)$ and $\phi(\cdot)$ to represent the cdf and the probability density function (pdf) of a standard normal distribution.

We add more comments on our running example. The label "selection-effect-only" is applied only to the QTE scenario. As will be explained in Section 4.1, the essential heterogeneity in the QTE case is defined as $F_{U_1|U_D}(u|u_D) = F_{U_0|U_D}(u|u_D)$ for $u \in [0,1]$ and $u_D \in [0,1]$, which is different from the definition in the ATE case - $E[U_1|U_D = u_D] = E[U_0|U_D = u_D]$ for $u_D \in [0,1]$. As shown in S.2.5, these two definitions do not imply each other. Our running example satisfies $F_{U_1|U_D}(u|u_D) = F_{U_0|U_D}(u|u_D)$ but not $E[U_1|U_D = u_D] = E[U_0|U_D = u_D]$. Nevertheless, the examples in the ATE sections (i.e., Section 3 and 5) do not involve whether $E[U_1 - U_0|U_D = u_D] = 0$ but are only related to $p(Z)$.

A word on notation: first, we depress the conditioning on $X = x$ throughout the paper to simplify notations. $\mathcal{C}(\mathcal{I})$ is the space of continuous functions on a compact set $\mathcal{I}$. d is always used for indicating the two treatment statuses, so is not written out explicitly as "$d = 0, 1$" throughout the paper. $\mu_d \equiv E[Y_d], \Delta = \mu_1 - \mu_0$, $F_d(y_d) \equiv P(Y_d \leq y_d)$ and $\Delta(\tau) = Q_1(\tau) - Q_0(\tau) \equiv F_1^{-1}(\tau) - F_0^{-1}(\tau)$, where $F_d^{-1}(\tau) = \inf\{y_d | F_d(y_d) \geq \tau\}$. We use the superscript star to indicate pseudo-true values for the IVE and IV-QRE and use the upper bar to indicate pseudo-true values for the LSE and QRE. For example, μ_d^* is the probability limit of the IVE for μ_d, and $\overline{\Delta}(\tau)$ is the probability limit of the QRE for $\Delta(\tau)$.

3 IVE

In this section, we first derive the pseudo-true value of IVE when the propensity score is used as the instrument[5] and explain why the IVE is consistent when the essential heterogeneity is absent; we then conduct the local sensitivity analysis and a parallel analysis when a general instrument is used, and conclude by summarizing the partial identification results in Heckman and Vytlacil (2001b). To start, recall that the moment conditions used by the IVE are

$$E\left[\begin{pmatrix} 1 \\ J(Z) \end{pmatrix}(Y - \mu_0^* - D \cdot \Delta^*)\right] = 0,$$

where $J(Z)$ is a scalar function of Z. Note also that for evaluating the ATE, we can use the additively separable outcome model, $Y_d = \mu_d(X) + U_d$, without loss of generality since we can redefine U_d as $Y_d - E[Y_d|X]$. In this setup, the selection effect means $E[U_0|U_D] \neq 0$ and the essential heterogeneity means $E[U_1 - U_0|U_D] \neq 0$, where X is depressed.

[5] As to why use $p(Z)$ as an instrument, note from Chamberlain (1987) that in the model $Y = \mu_0^* + D \cdot \Delta^* + U$ with $E[U|Z] = 0$, the optimal instruments are $E[(1, D)|Z] = (1, p(Z))$.

3.1 Pseudo-true Values

The following theorem states the pseudo-true values of IVE under misspecification when the propensity score is used as the instrument.

Theorem 1 *Under Assumptions A and P, when* $J(Z) = p(Z)$*,*

$$\begin{aligned}
\mu_1^* &= \int_0^{\underline{p}} E\left[Y_1|U_D = u_D\right] du_D + \int_{\underline{p}}^{\overline{p}} \left\{E\left[Y_1|U_D = u_D\right] h_1(u_D) + E\left[Y_0|U_D = u_D\right](1 - h_1(u_D))\right\} du_D \\
&\quad + \int_{\overline{p}}^{1} E\left[Y_0|U_D = u_D\right] du_D, \\
\mu_0^* &= \int_{\overline{p}}^{1} E\left[Y_0|U_D = u_D\right] du_D + \int_{\underline{p}}^{\overline{p}} \left\{E\left[Y_1|U_D = u_D\right] h_0(u_D) + E\left[Y_0|U_D = u_D\right](1 - h_0(u_D))\right\} du_D \\
&\quad + \int_{0}^{\underline{p}} E\left[Y_1|U_D = u_D\right] du_D,
\end{aligned}$$

and

$$\Delta^* = \mu_1^* - \mu_0^* = \int_{\underline{p}}^{\overline{p}} MTE(u_D)h(u_D)du_D = \int_0^1 MTE(u_D)h(u_D)du_D$$

where

$$\begin{aligned}
h_1(u_D) &= \frac{1}{Var\left(p(Z)\right)} \int_{u_D}^{1} \left[E\left[p(Z)^2\right] - pE\left[p(Z)\right] + \left(p - E\left[p(Z)\right]\right)\right] dF_{p(Z)}(p), \\
h_0(u_D) &= \frac{1}{Var\left(p(Z)\right)} \int_{u_D}^{1} \left[E\left[p(Z)^2\right] - pE\left[p(Z)\right]\right] dF_{p(Z)}(p).
\end{aligned}$$

and

$$h(u_D) = h_1(u_D) - h_0(u_D) = \frac{1}{Var\left(p(Z)\right)} \int_{u_D}^{1} \left[p - E\left[p(Z)\right]\right] dF_{p(Z)}(p).$$

Note that

$$\mu_1 = \int_0^1 E\left[Y_1|U_D = u_D\right] du_D = \int_0^{\underline{p}} E\left[Y_1|U_D = u_D\right] du_D + \int_{\underline{p}}^{\overline{p}} E\left[Y_1|U_D = u_D\right] du_D + \int_{\overline{p}}^{1} E\left[Y_1|U_D = u_D\right] du_D, \tag{4}$$

so in μ_1^*, we use a weighted average of $E\left[Y_1|U_D = u_D\right]$ and $E\left[Y_0|U_D = u_D\right]$ to approximate $E\left[Y_1|U_D = u_D\right]$ for $u_D \in [\underline{p}, \overline{p}]$ and use $E\left[Y_0|U_D = u_D\right]$ to approximate $E\left[Y_1|U_D = u_D\right]$ for $u_D \in [\overline{p}, 1]$. Similarly, in μ_0^*, we use a (different) weighted average of $E\left[Y_1|U_D = u_D\right]$ and $E\left[Y_0|U_D = u_D\right]$ to approximate $E\left[Y_0|U_D = u_D\right]$ for $u_D \in [\underline{p}, \overline{p}]$ and use $E\left[Y_1|U_D = u_D\right]$ to approximate $E\left[Y_0|U_D = u_D\right]$ for $u_D \in [0, \underline{p}]$. The approximation error depends on how close $E\left[Y_0|U_D = u_D\right]$ $\left(E\left[Y_1|U_D = u_D\right]\right)$ is to $E\left[Y_1|U_D = u_D\right]$ $\left(E\left[Y_1|U_D = u_D\right]\right)$ for $u_D \in [\underline{p}, 1]$ $(u_D \in [0, \overline{p}])$.

Since we can *only* identify $E\left[Y_1|U_D = u_D\right]$ and $E\left[Y_0|U_D = u_D\right]$ for $u_D \in [\underline{p}, \overline{p}]$ by

$$E\left[Y_1|U_D = u_D\right] = \left.\frac{dE\left[YD|p(Z) = p\right]}{dp}\right|_{p=u_D}, E\left[Y_0|U_D = u_D\right] = -\left.\frac{dE\left[Y(1-D)|p(Z) = p\right]}{dp}\right|_{p=u_D},$$

the true values μ_1 and μ_0 are not identifiable.[6] However, since

$$\int_0^{\underline{p}} E\left[Y_1|U_D=u_D\right]du_D = \underline{p}E\left[Y|D=1,p(Z)=\underline{p}\right], \int_{\overline{p}}^1 E\left[Y_0|U_D=u_D\right]du_D = (1-\overline{p})E\left[Y|D=0,p(Z)=\overline{p}\right],$$

μ_1^* and μ_0^* use the identifiable to replace the unidentifiable and thus are estimable,[7] where $E\left[Y|D=1,p(Z)=\underline{p}\right]$ is the mean of Y_1 for always takers and $E\left[Y|D=0,p(Z)=\overline{p}\right]$ is the mean of Y_0 for never-takers, and both are estimable.

From the proof of Theorem 1,

$$\begin{aligned}
\mu_1^* &= E\left[\frac{p(Z)}{Var\left(p(Z)\right)}\left\{\left(E\left[p(Z)^2-E\left[p(Z)\right]\right]\right)+p(Z)\left(1-E\left[p(Z)\right]\right)\right\}E\left[Y|D=1,p(Z)\right]\right] \\
&+E\left[\frac{1-p(Z)}{Var\left(p(Z)\right)}\left\{\left(E\left[p(Z)^2-E\left[p(Z)\right]\right]\right)+p(Z)\left(1-E\left[p(Z)\right]\right)\right\}E\left[Y|D=0,p(Z)\right]\right], \\
\mu_0^* &= E\left[\frac{p(Z)}{Var\left(p(Z)\right)}\left\{E\left[p(Z)^2\right]-p(Z)E\left[p(Z)\right]\right\}E\left[Y|D=1,p(Z)\right]\right] \\
&+E\left[\frac{1-p(Z)}{Var\left(p(Z)\right)}\left\{E\left[p(Z)^2\right]-p(Z)E\left[p(Z)\right]\right\}E\left[Y|D=0,p(Z)\right]\right].
\end{aligned}$$

In other words, μ_1^* is an expected weighted average of $E\left[Y|D=1,p(Z)\right]$ and $E\left[Y|D=0,p(Z)\right]$, while μ_0^* is a different expected weighted average of these two conditional expectations, and the two conditional expectations and the weights are all estimable. Note that given a $p(Z)$ value, the weights in μ_d^* for $E\left[Y|D=1,p(Z)\right]$ and $E\left[Y|D=0,p(Z)\right]$ are positive but their sum need not be 1; nevertheless, the expectation of this sum is 1. In other words, μ_d^* is a weighted average of $E\left[Y|D=1,p(Z)=p\right]$ and $E\left[Y|D=0,p(Z)=p\right]$ over all possible p values.

As to h_d and h, first note that they are only related to $p(Z)$ but do not involve Y_d. It is not hard to check $h_d(u_D)=1$ for $u_D\in[0,\underline{p}]$ and $h_d(u_D)=0$ for $u_D\in[\overline{p},1]$, which implies $h(u_D)=0$ for $u_D\in[0,\underline{p}]\cup[\overline{p},1]$. This also implies that the three terms of μ_d^* in Theorem 1 can be combined:

$$\begin{array}{c}
\mu_1^* = \int_0^1\left\{E\left[Y_1|U_D=u_D\right]h_1(u_D)+E\left[Y_0|U_D=u_D\right]\left(1-h_1(u_D)\right)\right\}du_D \\
\mu_0^* = \int_0^1\left\{E\left[Y_1|U_D=u_D\right]h_0(u_D)+E\left[Y_0|U_D=u_D\right]\left(1-h_0(u_D)\right)\right\}du_D.
\end{array} \tag{5}$$

Note also that $h(u_D)$ is the same as $h^{\mathrm{IV}}_{p(Z)}(u_D)$ in HV, so we reproduce the weighted average expression of Δ^* as in HV by analyzing μ_1^* and μ_0^* separately. The new thing here is the new weights h_1 and h_0, whose properties are studied in the next proposition. For completeness, we also state the properties of h as studied in HV.

Proposition 1 *h_1, h_0 and h satisfy the following properties:*

(i) $\int_{\underline{p}}^{\overline{p}} h_1(u_D)du_D = 1-\underline{p}$, $\int_0^1 h_1(u_D)du_D = 1$;

(ii) $\int_{\underline{p}}^{\overline{p}} h_0(u_D)du_D = -\underline{p}$, $\int_0^1 h_0(u_D)du_D = 0$;

(iii) $\int_{\underline{p}}^{\overline{p}} h(u_D)du_D = \int_0^1 h(u_D)du_D = 1$;

[6] This implies $MTE(u_D)=E\left[Y_1|U_D=u_D\right]-E\left[Y_0|U_D=u_D\right]=\left.\frac{dE[Y|p(Z)=p]}{dp}\right|_{p=u_D}$, $u_D\in[\underline{p},\overline{p}]$, as in Heckman and Vytlacil (1999, 2001a).

[7] In S.2.1, we provide alternative formulas for μ_d^* to show that it is estimable.

(iv) *when* $u_D < \frac{E[p(Z)]-E\left[p(Z)^2\right]}{1-E[p(Z)]}$*,* $h_1(u_D)$ *is strictly increasing, and when* $u_D > \frac{E[p(Z)]-E\left[p(Z)^2\right]}{1-E[p(Z)]}$*,* $h_1(u_D)$ *is strictly decreasing;*

(v) *when* $u_D < \frac{E\left[p(Z)^2\right]}{E[p(Z)]}$*,* $h_0(u_D)$ *is strictly decreasing, and when* $u_D > \frac{E\left[p(Z)^2\right]}{Ep(Z)}$*,* $h_0(u_D)$ *is strictly increasing;*

(vi) *when* $u_D < E\left[p(Z)\right]$*,* $h\left(u_D\right)$ *is strictly increasing, and when* $u_D > E\left[p(Z)\right]$*,* $h(u_D)$ *is strictly decreasing.*

From (iv), there is a point u_{1D}^* such that when $u_D \in (\underline{p}, u_{1D}^*)$, μ_1^* overweights $E\left[Y_1|U_D = u_D\right]$, and when $u_D \in (u_{1D}^*, \overline{p}]$, μ_1^* underweights $E\left[Y_1|U_D = u_D\right]$. Over $u_D \in [\underline{p}, \overline{p})$, $E\left[Y_0|U_D = u_D\right]$ is underweighted and the weight can even be negative when $u_D \in (\underline{p}, u_{1D}^*)$. From (v), μ_0^* always underweights $E\left[Y_1|U_D = u_D\right]$ for $u_D \in (\underline{p}, \overline{p}]$. Because $h_0(\overline{p}) = 0$ and h_0 is strictly increasing at the left of $\overline{p}$, there is a point u_{0D}^* such that when $u_D \in (u_{0D}^*, \overline{p})$, the weight is negative, so $E\left[Y_0|U_D = u_D\right]$ is overweighted over these u_D's. From (vi), h is always positive on $(\underline{p}, \overline{p})$.

In summary, h_1 is a proper weighting scheme on $[0,1]$ since $h_1(u_D) \geq 0$ and $\int_0^1 h_1(u_D)du_D = 1$, but is not a proper weighting scheme on $[\underline{p}, \overline{p}]$ since $\int_{\underline{p}}^{\overline{p}} h_1(u_D)du_D = 1 - \underline{p} < 1$; h_0 is not a proper weighting scheme on either $[0,1]$ or $[\underline{p}, \overline{p}]$ since it can be negative and $\int_{\underline{p}}^{\overline{p}} h_0(u_D)du_D = -\underline{p} < 1$, $\int_0^1 h_0(u_D)du_D = 0 < 1$; while h is a a proper weighting scheme on both $[0,1]$ and $[\underline{p}, \overline{p}]$ since $h(u_D) \geq 0$ and $\int_0^1 h(u_D)du_D = \int_{\underline{p}}^{\overline{p}} h(u_D)du_D = 1$. Nevertheless, from (5), since $\int_0^1 \left[h_1(u_D) + (1 - h_1(u_D))\right] du_D = 1$ and $\int_0^1 \left[h_0(u_D) + (1 - h_0(u_D))\right] du_D = 1$, both μ_1^* and μ_0^* are weighted averages of $E\left[Y_1|U_D = u_D\right]$ and $E\left[Y_0|U_D = u_D\right]$ over $u_D \in [0,1]$. Note also that the weights in μ_1^* on $E\left[Y_1|U_D = u_D\right]$ over $u_D \in [0, \overline{p}]$ satisfy $\int_0^{\overline{p}} h_1(u_D)du_D = 1$ and the weights on $E\left[Y_0|U_D = u_D\right]$ over $u_D \in [\underline{p}, 1]$ satisfy $\int_{\underline{p}}^1 \left[1 - h_1(u_D)\right] du_D = 0$. Similarly, the weights in μ_0^* on $E\left[Y_0|U_D = u_D\right]$ over $u_D \in [\underline{p}, 1]$ satisfy $\int_{\underline{p}}^1 \left[1 - h_0(u_D)\right] du_D = 1$ and the weights on $E\left[Y_1|U_D = u_D\right]$ over $u_D \in [0, \overline{p}]$ satisfy $\int_0^{\overline{p}} h_0(u_D)du_D = 0$. This implies that in the unconfounded case, where $E\left[Y_d|U_D = u_D\right] = \mu_d$, $\mu_d^* = \mu_d$; that is, in the unconfounded case, μ_d can be identified with or without instruments.

The following example intuitively illustrates the shape of h_1, h_0 and h.

Example 1 *Consider the setup of treatment status* D *in the running example. In this example,* $p(Z) \sim U[0,1]$*, so* $\underline{p} = 0$ *and* $\overline{p} = 1$*.*

$$h_1(u_D) = (1 + 3u_D)(1 - u_D),\ h_0(u_D) = (1 - 3u_D)(1 - u_D) \text{ and } h(u_D) = 6u_D(1 - u_D)$$

are shown in Figure 1. From Figure 1, $h_1(u_D)$ *is increasing on* $[0, 1/3]$ *and decreasing on* $[1/3, 1]$*,* $h_0(u_D)$ *is decreasing on* $[0, 2/3]$ *and increasing on* $[2/3, 1]$*, and* $h(u_D)$ *is increasing on* $[0, 1/2]$ *and decreasing on* $[1/2, 1]$*.* μ_1^* *overestimates* $E\left[Y_1|U_D = u_D\right]$ *for* $u_D \in (0, 2/3)$ *and underestimates* $E\left[Y_1|U_D = u_D\right]$ *for* $u_D \in (2/3, 1]$*, while* μ_0^* *overestimates* $E\left[Y_0|U_D = u_D\right]$ *for* $u_D \in (1/3, 1)$ *and underestimates* $E\left[Y_0|U_D = u_D\right]$ *for* $u_D \in [0, 1/3)$*. The area between* $h_1(u_D)$ *and the horizontal axis is* 1*, while the area between* $h_0(u_D)$ *and the horizontal axis is* 0*. Since* h_1 *is always above* h_0 *for* $u_D \in (0,1)$*,* $h(u_D)$ *is positive on* $u_D \in (0,1)$ *and the area between it and the horizontal axis is* 1*.*

Without unconfoundedness, the following proposition shows that it is still possible to have $\mu_d^* = \mu_d$ as long as there is no essential heterogeneity,

Proposition 2 *If* $E[U_1 - U_0|U_D = u_D] = 0$ *for all* $u_D \in [\underline{p}, 1]$*,* $\mu_1^* = \mu_1$*; if* $E[U_1 - U_0|U_D = u_D] = 0$ *for all* $u_D \in [0, \overline{p}]$*,* $\mu_0^* = \mu_0$*. If* $E[U_1 - U_0|U_D = u_D] = 0$ *for all* $u_D \in [\underline{p}, \overline{p}]$*,* $\Delta^* = \Delta$*.*

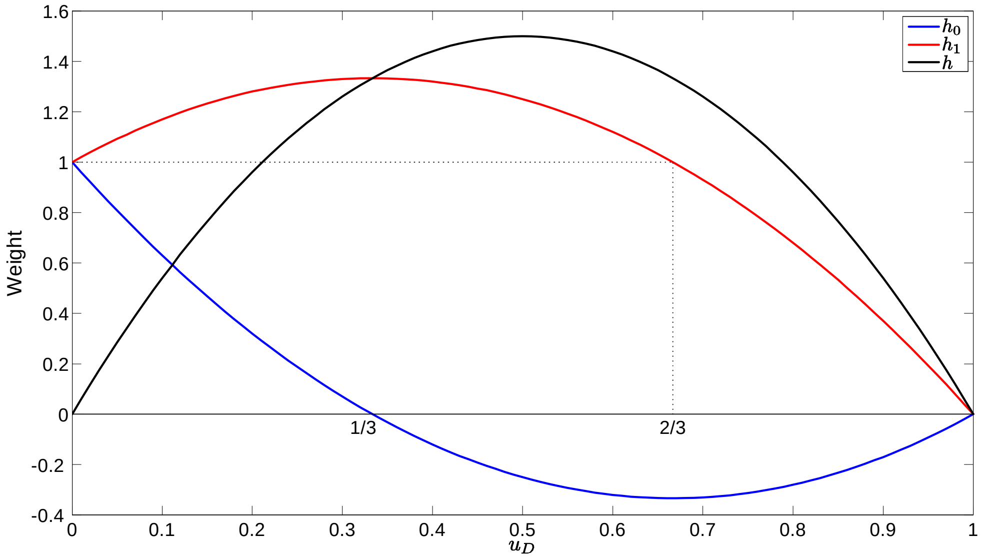

Figure 1: h_1, h_0 and $h = h_1 - h_0$

Proof. Since when $E[U_1 - U_0|U_D = u_D] = 0$, μ_1 takes the form of (4) with $E[Y_1|U_D = u_D] = E[Y_0|U_D = u_D] + \Delta$,

$$\mu_1^* - \mu_1 = \Delta \left[\int_{\underline{p}}^{\overline{p}} h_1(u_D) du_D - \int_{\underline{p}}^{1} du_D \right] = \Delta \left[(1 - \underline{p}) - (1 - \underline{p}) \right] = 0$$

from Proposition 1(i). Similarly,

$$\mu_0^* - \mu_0 = \Delta \left[\int_{0}^{\underline{p}} du_D + \int_{\underline{p}}^{\overline{p}} h_0(u_D) du_D \right] = \Delta \left[\underline{p} - \underline{p} \right] = 0$$

from Proposition 1(ii). Finally,

$$\Delta^* - \Delta = \int_{\underline{p}}^{\overline{p}} E[U_1 - U_0|U_D = u_D] h(u_D) du_D = 0.$$

■

Suppose $\mu_1 > \mu_0 > 0$. μ_1^* overestimates $E[Y_1|U_D = u_D]$ over $u_D \in (\underline{p}, u_{1D}^*)$ but underestimates $E[Y_1|U_D = u_D]$ over $u_D \in (u_{1D}^*, 1]$. These two forces offset each other and we get a consistent estimator. Similarly, μ_0^* overestimates $E[Y_0|U_D = u_D]$ over $u_D \in [0, u_{0D}^*]$ but underestimates $E[Y_0|U_D = u_D]$ over $u_D \in [u_{0D}^*, \overline{p}]$ such that a consistent estimator is achieved. Here, u_{1D}^* and u_{0D}^* are defined after Proposition 1. From the above proof, we can see that if the pseudo-true values of an estimator can be expressed in the weighted average form as in Theorem 1, then as long as the weights satisfy Proposition 1(i) and (ii) (especially, $\int_{\underline{p}}^{\overline{p}} h_1(u_D) du_D = 1 - \underline{p}$ and $\int_{\underline{p}}^{\overline{p}} h_0(u_D) du_D = -\underline{p}$), the results of Proposition 2 can be applied.

3.2 Local Sensitivity Analysis

Take $E[U_1 - U_0|U_D = u_D]$ as a function in $\mathcal{C}[0,1]$, the space of continuous functions on $[0,1]$. Assume $E[U_1 - U_0|U_D = u_D]$ stays in a parametrized space

$$\mathcal{G}^*_\alpha = \left\{ G_\alpha(\cdot) : \alpha \in M, 0 \in M, G_0(u_D) = 0, u_D \in [0,1], \text{ and } \int_0^1 G_\alpha(u_D) du_D = 0 \right\};$$

then μ^*_d and Δ^* depend on α, where $\int_0^1 G_\alpha(u_D) du_D = 0$ because we normalize $E[U_1] = E[U_0] = 0$. We can now study the sensitivity of μ^*_d and Δ^* to the local deviation of $E[U_1 - U_0|U_D = u_D]$ from zero along the path $\mathcal{G}^*_\alpha$. Specifically, give a direction of deviation, say $g(\cdot)$, let $\lim_{\alpha \to 0} \frac{G_\alpha(\cdot)}{\alpha} \to g(\cdot)$; then our targets are

$$D^*_d(g) \equiv \lim_{\alpha \to 0} \frac{\mu^*_d(\alpha) - \mu_d}{\alpha} \text{ and } D^*_\Delta(g) \equiv \lim_{\alpha \to 0} \frac{\Delta^*(\alpha) - \Delta}{\alpha}$$

or path derivatives along the path $\mathcal{G}^*_\alpha$, where we use $\mu^*_d(\alpha)$ and $\Delta^*(\alpha)$ to indicate the dependence of μ^*_d and Δ^* on α, and note that $\mu^*_d(0) = \mu_d$ and $\Delta^*(0) = \Delta$. We impose the following conditions to guarantee the limits in D^*_d and D^*_Δ exist.

Assumption LA: $\sup_{u_D \in [0,1], \alpha \in \mathcal{N}} \left| \frac{G_\alpha(u_D)}{\alpha} \right| < \infty$, where $\mathcal{N}$ is a neighborhood of 0.

Proposition 3 *Under Assumptions A, P and LA,*

$$\begin{aligned} D^*_1(g) &= -\int_{\underline{p}}^{\overline{p}} g(u_D)(1 - h_1(u_D)) du_D - \int_{\overline{p}}^1 g(u_D) du_D = \int_0^1 g(u_D) h_1(u_D) du_D, \\ D^*_0(g) &= \int_0^{\underline{p}} g(u_D) du_D + \int_{\underline{p}}^{\overline{p}} g(u_D) h_0(u_D) du_D = \int_0^1 g(u_D) h_0(u_D) du_D, \end{aligned}$$

and

$$D^*_\Delta(g) = \int_{\underline{p}}^{\overline{p}} g(u_D) h(u_D) du_D = \int_0^1 g(u_D) h(u_D) du_D = D^*_1(g) - D^*_0(g).$$

$D^*_1(g)$ depends on $g(u_D)$ for $u_D \in [\underline{p}, 1]$, and $D^*_0(g)$ depends on $g(u_D)$ for $u_D \in [0, \overline{p}]$. This is because only on these areas of u_D, μ^*_1 and μ^*_0 have mis-estimated $E[Y_1|U_D = u_D]$ and $E[Y_0|U_D = u_D]$, respectively. Since h_d and h are estimable, $D^*_d(g)$ and $D^*_\Delta(g)$ are estimable.

The following example numerically illustrates $D^*_d(g)$ and $D^*_\Delta(g)$ for a specific g.

Example 2 *Consider the setup in the running example. Suppose $g(u_D) = 1 - 2u_D$; i.e., the individual with higher propensity to participate has a larger idiosyncratic gain. Then*

$$\begin{aligned} D^*_1(g) &= \int_0^1 (1 - 2u_D)(1 + 3u_D)(1 - u_D) du_D = \frac{1}{6}, \\ D^*_0(g) &= \int_0^1 (1 - 2u_D)(1 - 3u_D)(1 - u_D) du_D = \frac{1}{6} \end{aligned}$$

and

$$D^*_\Delta(g) = \int_0^1 6u_D(1 - u_D)(1 - 2u_D) du_D = 0.$$

In this simple case, although there is essential heterogeneity locally, Δ can be consistently estimated. This is because both μ^*_1 and μ^*_0 overestimate their corresponding true value while the biases offset each other. Of course, it is easy to construct examples where Δ cannot be consistently estimated.

The local sensitivity analysis above is closely related to semiparametric efficiency bound calculation in the literature; see, e.g., Newey (1990) and Bickel et al. (1998) for an introduction. For a real parameter β, we need to check how it changes as the response to a local variation of the parametric submodel f_θ to calculate its semiparametric efficiency bound, where if we use $\beta(\theta)$ to denote the dependency of β on θ, then $\beta(\theta_0)$ equals the true value. The local sensitivity analysis has two differences. First, we do not need to parametrize the density (or distribution) function, but parametrize any characteristic of the model, e.g., $E[U_1 - U_0|U_D = u_D]$ in this example. Second, the parameter of interest need not be real-valued, but can belong to any normed space. Although for this example, μ_d^* and Δ^* are indeed real-valued, in Section 4.3 below, the parameters of interest belong to $\mathcal{C}([0,1])$. Finally, we mention that if the parameter of interest is real-valued and its path derivative is a continuous linear functional from the Hilbert space that $g(\cdot)$ stays to $\mathbb{R}$, then by the Riesz representation theorem, its path derivative can be represented as an inner product, where we still use $g(\cdot)$ to denote the path derivative of the parametrized characteristic of the model. Conversely, if the path derivative of the interested parameter can be represented as an inner product, then it is a continuous linear functional. For example, $D_1^*(g) = \int_0^1 s(u_D)g(u_D)du_D$ is an inner product on $L^2([0,1])$, where $s(u_D) = -1(\underline{p} \le u_D \le \overline{p})(1 - h_1(u_D)) - 1(\overline{p} \le u_D \le 1)$, so it is a continuous linear functional. Sometimes, we may restrict the space $g(\cdot)$ staying to be a subspace of a Hilbert space, e.g., in this example, we have restricted $g(\cdot) \in \mathcal{C}_0([0,1])$, where $\mathcal{C}_0([0,1]) \equiv \left\{g(\cdot) : g \in \mathcal{C}([0,1]) \text{ and } \int_0^1 g(u_D)du_D = 0\right\}$.

In practice, it is quite often that there are a few possible directions of deviation and there is some prior information on the distribution of these deviations. Rigorously, suppose $g \in \mathcal{G}$ and there is a prior distribution $\mathcal{F}(\cdot)$ on $\mathcal{G}$, where $\mathcal{G}$ can be parametrized or fully nonparametric.[8] Then we can summarize such prior information by

$$D_d^*(\mathcal{G}) \equiv \int D_1^*(g)\, d\mathcal{F}(g) \text{ and } D_\Delta^*(\mathcal{G}) \equiv \int D_\Delta^*(g)\, d\mathcal{F}(g).$$

This is somewhat like *Bayesian model averaging*. This technique can be applied to any local sensitivity anslysis so we will not repeat it in the future.

3.3 Using General Instruments

When a general instrument $J(Z)$ is used in the IVE, the pseudo-true values are stated in the following proposition. To distinguish from the pseudo-true values as $p(Z)$ is used as the instrument, we add the subscript J to all objects.

Proposition 4 *Under Assumptions A and P,*

$$\begin{aligned}
\mu_{J1}^* &= \int_0^{\underline{p}} E[Y_1|U_D = u_D]\, du_D + \int_{\underline{p}}^{\overline{p}} \{E[Y_1|U_D = u_D]\, h_{J1}(u_D) + E[Y_0|U_D = u_D](1 - h_{J1}(u_D))\}\, du_D \\
&\quad + \int_{\overline{p}}^1 E[Y_0|U_D = u_D]\, du_D, \\
\mu_{J0}^* &= \int_{\overline{p}}^1 E[Y_0|U_D = u_D]\, du_D + \int_{\underline{p}}^{\overline{p}} \{E[Y_1|U_D = u_D]\, h_{J0}(u_D) + E[Y_0|U_D = u_D](1 - h_{J0}(u_D))\}\, du_D \\
&\quad + \int_0^{\underline{p}} E[Y_1|U_D = u_D]\, du_D,
\end{aligned}$$

[8] Note that $\mathcal{C}_0([0,1])$ is separable, so any $g \in \mathcal{G}$ can be approximated arbitrarily well by countably many basis functions, e.g., classical orthogonal polynomials, Fourier series, neural nets, or wavelets. Since it is easy to evaluate $D_1^*(\cdot), D_0^*(\cdot)$ and $D_\Delta^*(\cdot)$ at these basis functions, $D_1^*(g), D_0^*(g)$ and $D_\Delta^*(g)$ for a general $g \in \mathcal{G}$ can be easily approximated by linear combinations of $D_1^*(\cdot), D_0^*(\cdot)$ and $D_\Delta^*(\cdot)$ values at these basis functions.

and

$$\Delta_J^* = \mu_{J1}^* - \mu_{J0}^* = \int_{\underline{p}}^{\overline{p}} MTE(u_D) h_J(u_D) du_D = \int_0^1 MTE(u_D) h_J(u_D) du_D$$

where

$$\begin{aligned} & h_{J1}(u_D) \\ = \; & \frac{1}{Cov\left(J(Z), p(Z)\right)} \int_{u_D}^1 \left[E\left[p(Z)J(Z)\right] - E\left[J(Z)|p(Z)=p\right] E\left[p(Z)\right] + \left(E\left[J(Z)|p(Z)=p\right] - E\left[J(Z)\right]\right) \right] dF_{p(Z)}(p), \\ & h_{J0}(u_D) \\ = \; & \frac{1}{Cov\left(J(Z), p(Z)\right)} \int_{u_D}^1 \left[E\left[p(Z)J(Z)\right] - E\left[J(Z)|p(Z)=p\right] E\left[p(Z)\right] \right] dF_{p(Z)}(p) \end{aligned}$$

and

$$h_J(u_D) = h_{J1}(u_D) - h_{J0}(u_D) = \frac{1}{Cov\left(J(Z), p(Z)\right)} \int_{u_D}^1 \left[E\left[J(Z)|p(Z)=p\right] - E\left[J(Z)\right] \right] dF_{p(Z)}(p).$$

The structures of μ_{Jd}^* and Δ_J^* are similar to those of μ_d^* and Δ^*, and the only difference is the weights, so we concentrate on h_{Jd} and h_J in the following discussion. Similar to h_d, it is not hard to check $h_{Jd}(u_D) = 1$ for $u_D \in [0, \underline{p}]$ and $h_{Jd}(u_D) = 0$ for $u_D \in [\overline{p}, 1]$, and $h_J(u_D) = 0$ for $u_D \in [0, \underline{p}] \cup [\overline{p}, 1]$. $h_J(u_D)$ is exactly $h_{\mathrm{IV}}(u_D|J)$ in HV, so we reproduce the weighted average expression of Δ_J^* by analyzing μ_{J1}^* and μ_{J0}^* separately.

We can still prove the first two properties in Proposition 1 for h_{J1} and h_{J0}, so the results in Proposition 2 still hold for μ_{J1}^* and μ_{J0}^*. However, the last three properties of Proposition 1 cannot be proved generally. Nevertheless, if $E\left[J(Z)|p(Z)=p\right]$ is a strictly increasing or decreasing function of p, we can still prove similar properties for h_{Jd} and h_J. To avoid repetition, we summarize the properties of h_{Jd} and h_J in S.2.3. When $E\left[J(Z)|p(Z)=p\right]$ is not monotone in p, $h_J(u_D)$ need not be positive as emphasized in HV; see Figure 4 of HV for the possibility that $h_J\left(u_D\right)$ is negative at some area of u_D. From the alternative expression of h_J,

$$h_J\left(u_D\right) = \frac{E\left[J(Z) - E\left[J(Z)\right] | p(Z) \geq u_D\right]}{Cov\left(J(Z), p(Z)\right)} P\left(p(Z) \geq u_D\right) = \frac{Cov\left(J(Z), 1\left(p(Z) \geq u_D\right)\right)}{Cov\left(J(Z), 1\left(p(Z) \geq U_D\right)\right)},$$

we can see if $E\left[J(Z)|p(Z) \geq p\right]$ is weakly monotone in p, then $h_J\left(u_D\right)$ is nonnegative for any u_D. This condition is weaker than the weak monotonicity of $E\left[J(Z)|p(Z)=p\right]$.

We re-emphasize a point made by HV here. For different $J(Z)$, μ_{Jd}^* and Δ_J^* are different; in other words, the estimands depend on the instruments employed (even if the set of Z is fixed) unless the essential heterogeneity is excluded; see Heckman (2010) and footnote 6 of Carneiro et al. (2011) for an intuitive explanation.

We next discuss the discrete instrument case. A discrete instrument can appear when Z is discrete and $p(Z)$ is used as the instrument or Z is continuous but the employed instrument $J(Z)$ is discrete (e.g., $J(Z)$ is an indicator function for different ranges of $p(Z)$). We consider only the former case since the latter can be similarly discussed. As in Section 4.3 of HV, suppose $p(Z)$ can take only K values, $\{p_1, \cdots, p_K\}$, with $0 < p_1 < p_2 < \cdots < p_K < 1$. In this case, $h_1(u_D) = h_1(p_{k+1})$ for $u_D \in (p_k, p_{k+1}]$, $k = 1, \cdots, K-1$, and similarly for $h_0(u_D)$ and $h(u_D)$. As a result,

$$\begin{aligned}
\mu_1^* &= \sum_{k=0}^{K} \int_{p_k}^{p_{k+1}} \left\{ E\left[Y_1|U_D=u_D\right] h_1(p_{k+1}) + E\left[Y_0|U_D=u_D\right](1-h_1(p_{k+1})) \right\} du_D \\
&= \sum_{k=0}^{K} (p_{k+1}-p_k)\left[h_1(p_{k+1})\mu_1^k + (1-h_1(p_{k+1}))\mu_0^k\right] \\
&= \sum_{k=0}^{K} (p_{k+1}-p_k)\left\{ E\left[Y_0|U_D=\widetilde{u}_{Dk}\right] + \int_{p_k}^{p_{k+1}} MTE(u_D)\frac{h_1(p_{k+1})}{p_{k+1}-p_k} du_D \right\},
\end{aligned}$$

where

$$\mu_d^k = \frac{1}{p_{k+1}-p_k}\int_{p_k}^{p_{k+1}} E\left[Y_d|U_D=u_D\right] du_D = \frac{E\left[Y\cdot 1\left(D=d\right)|p(Z)=p_{k+1}\right] - E\left[Y\cdot 1\left(D=d\right)|p(Z)=p_k\right]}{P\left(D=d|p(Z)=p_{k+1}\right) - P\left(D=d|p(Z)=p_k\right)},$$

$k=0,1,\cdots,K$, are identifiable,[9] μ_d^0 is the mean of Y_d for always-takers $\mathcal{A}\equiv\{U_D\le p_1\}$, μ_d^K is the mean of Y_d for never-takers $\mathcal{N}\equiv\{U_D>p_K\}$, μ_d^k, $k=1,\cdots,K-1$, is the mean of Y_d for compliers $\mathcal{C}_k\equiv\{p_k<U_D\le p_{k+1}\}$ who are induced to switch from $D=0$ to $D=1$ as $p(Z)$ changes from p_k to p_{k+1}, $\widetilde{u}_{Dk}$ is defined by $E\left[Y_0|U_D=\widetilde{u}_{Dk}\right]=\mu_0^k$, $p_0\equiv 0$, $p_{K+1}\equiv 1$, $h_1(p_1)=1$, $h_1(p_{K+1})=0$, and

$$h_1(p_{k+1}) = \frac{Cov\left(p(Z),p(Z)1\left(p(Z)\ge p_{k+1}\right)\right) - Cov\left(p(Z)^2,1\left(p(Z)\ge p_{k+1}\right)\right) + Cov\left(p(Z),1\left(p(Z)\ge p_{k+1}\right)\right)}{Var\left(p(Z)\right)},$$

$k=1,\cdots,K-1$, are identifiable. Similarly,

$$\begin{aligned}
\mu_0^* &= \sum_{k=0}^{K} \int_{p_k}^{p_{k+1}} \left\{ E\left[Y_1|U_D=u_D\right] h_0(p_{k+1}) + E\left[Y_0|U_D=u_D\right](1-h_0(p_{k+1})) \right\} du_D \\
&= \sum_{k=0}^{K} (p_{k+1}-p_k)\left[h_0(p_{k+1})\mu_1^k + (1-h_0(p_{k+1}))\mu_0^k\right] \\
&= \sum_{k=0}^{K} (p_{k+1}-p_k)\left\{ E\left[Y_0|U_D=\widetilde{u}_{Dk}\right] + \int_{p_k}^{p_{k+1}} MTE(u_D)\frac{h_0(p_k)}{p_{k+1}-p_k} du_D \right\},
\end{aligned}$$

where the $\widetilde{u}_{Dk}$'s are the same as in μ_1^*, $h_0(p_1)=1$, $h_0(p_{K+1})=0$, and

$$h_0(p_{k+1}) = \frac{Cov\left(p(Z),p(Z)1\left(p(Z)\ge p_{k+1}\right)\right) - Cov\left(p(Z)^2,1\left(p(Z)\ge p_{k+1}\right)\right)}{Var\left(p(Z)\right)}, k=1,\cdots,K-1,$$

are identifiable. Therefore,

$$\Delta^* = \sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)h(p_{k+1})\int_{p_k}^{p_{k+1}} MTE(u_D)\frac{1}{p_{k+1}-p_k}du_D = \sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right)h(p_{k+1})LATE\left(p_k,p_{k+1}\right)$$

as shown in HV, where

$$LATE\left(p_k,p_{k+1}\right) \equiv \frac{1}{p_{k+1}-p_k}\int_{p_k}^{p_{k+1}} MTE(u_D)du_D = \mu_1^k - \mu_0^k = \frac{E\left[Y|p(Z)=p_{k+1}\right] - E\left[Y|p(Z)=p_k\right]}{p_{k+1}-p_k}$$

[9] In μ_d^0 and μ_d^K, $E\left[Y\cdot 1\left(D=1\right)|p(Z)=0\right] = E\left[Y\cdot 1\left(D=0\right)|p(Z)=1\right] = P\left(D=1|p(Z)=0\right) = P\left(D=0|p(Z)=1\right) = 0$.

is the average treatment effect for compliers $\mathcal{C}_k$,

$$h(p_{k+1}) = h_1(p_{k+1}) - h_0(p_{k+1}) = \frac{Cov\left(p(Z), 1\left(p(Z) \geq p_{k+1}\right)\right)}{Var\left(p(Z)\right)}, k = 1, \cdots, K-1, \tag{6}$$

and the summation is from $k = 1$ to $k = K-1$ because $h_1(p_1) = h_0(p_1) = 1$ and $h_1(p_{K+1}) = h_0(p_{K+1}) = 0$. Finally, we explain why $\sum_{k=1}^{K-1}\left(p_{k+1} - p_k\right) h(p_{k+1}) = 1$. Note that $p(Z) = \sum_{k=0}^{K}\left(p_{k+1} - p_k\right) 1(p(Z) \geq p_{k+1}) = p_1 + \sum_{k=1}^{K-1}\left(p_{k+1} - p_k\right) 1(p(Z) \geq p_{k+1}) + (1 - p_K)$, so

$$Var\left(p(Z)\right) = \sum_{k=1}^{K-1}\left(p_{k+1} - p_k\right) Cov\left(p(Z), 1\left(p(Z) \geq p_{k+1}\right)\right),$$

which implies $\sum_{k=1}^{K-1}\left(p_{k+1} - p_k\right) h(p_{k+1}) = 1$.

A special case is $K = 2$ as discussed in IA, where Z can take only two values. In this case,

$$h_1(p_2) = \frac{1 - p_1}{p_2 - p_1}, h_0(p_2) = -\frac{p_1}{p_2 - p_1} \text{ and } h(p_2) = \frac{1}{p_2 - p_1},$$

so

$$\begin{aligned}
\mu_1^* &= p_1\mu_{1|\mathcal{A}} + (p_2 - p_1)\left[\frac{1 - p_1}{p_2 - p_1}\mu_{1|\mathcal{C}} - \frac{1 - p_2}{p_2 - p_1}\mu_{0|\mathcal{C}}\right] + (1 - p_2)\mu_{0|\mathcal{N}} \\
&= p_1\mu_{1|\mathcal{A}} + (1 - p_1)\mu_{1|\mathcal{C}} + (1 - p_2)\left(\mu_{0|\mathcal{N}} - \mu_{0|\mathcal{C}}\right), \\
\mu_0^* &= (1 - p_2)\mu_{0|\mathcal{N}} + (p_2 - p_1)\left[\frac{p_2}{p_2 - p_1}\mu_{0|\mathcal{C}} - \frac{p_1}{p_2 - p_1}\mu_{1|\mathcal{C}}\right] + p_1\mu_{1|\mathcal{A}} \\
&= (1 - p_2)\mu_{0|\mathcal{N}} + p_2\mu_{0|\mathcal{C}} + p_1\left(\mu_{1|\mathcal{A}} - \mu_{1|\mathcal{C}}\right), \\
\Delta^* &= \mu_{1|\mathcal{C}} - \mu_{0|\mathcal{C}} \equiv LATE,
\end{aligned}$$

where Δ^* is exactly the LATE in IA, $p_1, p_2 - p_1$ and $1 - p_2$ are the probabilities of always-takers, compliers and never-takers, and $\mu_{1|\mathcal{A}}$, $\mu_{1|\mathcal{C}}$, $\mu_{0|\mathcal{C}}$ and $\mu_{0|\mathcal{N}}$ are the means for always-takers $\mathcal{A} \equiv \{U_D \leq p_1\}$, compliers $\mathcal{C} \equiv \{p_1 < U_D \leq p_2\}$ and never-takers $\mathcal{N} \equiv \{U_D > p_2\}$ in the two counterfactual statuses. Obviously, μ_d^* is a weighted average of these four means, but μ_1^* overweights $\mu_{1|\mathcal{C}}$ and negatively weights $\mu_{0|\mathcal{C}}$ while μ_0^* overweights $\mu_{0|\mathcal{C}}$ and negatively weights $\mu_{1|\mathcal{C}}$. Overall, if $\mu_{1|\mathcal{C}} - \mu_{0|\mathcal{C}} = \mu_{1|\mathcal{N}} - \mu_{0|\mathcal{N}}$, then $\mu_1^* = p_1\mu_{1|\mathcal{A}} + (1 - p_1)\mu_{1|\mathcal{C}} + (1 - p_2)\left(\mu_{1|\mathcal{N}} - \mu_{1|\mathcal{C}}\right) = p_1\mu_{1|\mathcal{A}} + (p_2 - p_1)\mu_{1|\mathcal{C}} + (1 - p_2)\mu_{1|\mathcal{N}} = \mu_1$. Similarly, if $\mu_{1|\mathcal{C}} - \mu_{0|\mathcal{C}} = \mu_{1|\mathcal{A}} - \mu_{0|\mathcal{A}}$, then $\mu_0^* = \mu_0$. Finally, if $\mu_{1|\mathcal{C}} - \mu_{0|\mathcal{C}} = \mu_{1|\mathcal{N}} - \mu_{0|\mathcal{N}} = \mu_{1|\mathcal{A}} - \mu_{0|\mathcal{A}}$, which is the conditonal constant effects assumption (Restriction 2) of Angrist (2004) or the conditional effect ignorability assumption (Assumption 3) of Angrsit and Fernández-Val (2013), then $\Delta^* = \Delta$.

Finally, when a general instrument $J(Z)$ is used, $D_d^*(g)$ and $D_\Delta^*(g)$ in Section 3.2 take similar forms except replacing $h_d(u_D)$ and $h(u_D)$ by $h_{Jd}(u_D)$ and $h_J(u_D)$, so similar analysis as above can be applied. For example, when $p(Z)$ is discrete,

$$\begin{aligned}
D_1^*(g) &= -\sum_{k=1}^{K-1}\left(p_{k+1} - p_k\right)\left(1 - h_1(p_{k+1})\right) g_k - (1 - p_K) g_K, \\
D_0^*(g) &= p_1 g_0 + \sum_{k=1}^{K-1}\left(p_{k+1} - p_k\right) h_0(p_{k+1}) g_k,
\end{aligned}$$

and $D^*_\Delta(g) = D^*_1(g) - D^*_0(g)$, where

$$g_k = \frac{1}{p_{k+1} - p_k} \int_{p_k}^{p_{k+1}} g(u_D) du_D, k = 0, 1, \cdots, K. \tag{7}$$

Of course, we can specify g_k directly, $k = 0, 1, \cdots, K$, rather than $g(\cdot)$ in practice.

3.4 Partial Identification Analysis

We summarize the sharp bounds for μ_d and Δ developed in Heckman and Vytlacil (2001b) in this subsection. To avoid trivial bounds for μ_d and Δ, we maintain the bounds of Y_d as in assumption (Q-4). Specifically, we impose a weaker version of assumption (Q-4):

Assumption (Q-4′): $P(Y_d \in \mathcal{Y}_{xd} | X = x) = 1$ for almost every $x \in$supp(X).

Note that assumption (A-2) implies that supp$(Y_d|X = x)$ does not depend on Z, so the conditioning on Z in (Q-4′) is not necessary. As usual, we depress the conditioning on $X = x$ to simplify notations. It turns out that

$$\begin{aligned}
\underline{I}_1 &\equiv \overline{p}E[Y|p(Z) = \overline{p}, D = 1] + (1-\overline{p})\underline{y}_1 \leq \mu_1 \leq \overline{p}E[Y|p(Z) = \overline{p}, D = 1] + (1-\overline{p})\overline{y}_1 \equiv \overline{I}_1, \\
\underline{I}_0 &\equiv \left(1-\underline{p}\right) E\left[Y|p(Z) = \underline{p}, D = 0\right] + \underline{py}_0 \leq \mu_0 \leq \left(1-\underline{p}\right) E\left[Y|p(Z) = \underline{p}, D = 0\right] + \underline{p}\overline{y}_0 \equiv \overline{I}_0, \\
\underline{I}_\Delta &\equiv \underline{I}_1 - \overline{I}_0 \leq \Delta \leq \overline{I}_1 - \underline{I}_0 \equiv \overline{I}_\Delta,
\end{aligned}$$

where $\underline{p}$ and $\overline{p}$ are defined in Assumption P. Note that $\left[\underline{I}_\Delta, \overline{I}_\Delta\right]$ does not necessarily cover 0. Since the width of the bounds for Δ is $\overline{I}_1 - \underline{I}_1 + \overline{I}_0 - \underline{I}_0 = (1-\overline{p})\left(\overline{y}_1 - \underline{y}_1\right) + \underline{p}\left(\overline{y}_0 - \underline{y}_0\right)$, $\underline{p} = 0$ and $\overline{p} = 1$ are necessary and sufficient for point identification of Δ under Assumption A and (Q-4′). To evaluate the bounds for μ_1, we need $\overline{p}$, $\underline{y}_1$, $\overline{y}_1$ and $E[Y|p(Z) = \overline{p}, D = 1]$; to evaluate the bounds for μ_0, we need $\underline{p}$, $\underline{y}_0$, $\overline{y}_0$ and $E\left[Y|p(Z) = \underline{p}, D = 0\right]$; to evaluate the bounds for Δ, we need all of them. Anyway, we need only evaluate two conditional means to construct these bounds.

4 IV-QRE

The structure of this section is the same as that of Section 3 except that we explain at the beginning why the IV-QRE, as an estimator of QTE, implicitly excludes the essential heterogeneity to guarantee its consistency.

4.1 Understanding the IV-QRE

CH express

$$Y_d = q(d, X, U_d) \text{ with } U_d|X \sim U(0,1)$$

by the Skorohod representation, where $q(d, x, \tau)$ is the quantile function of Y_d conditional on $X = x$. This representation is a special case of the setup (1) and is essential in developing their identification results. CH impose the following assumptions on the model:

A1. Potential Outcomes: Conditional on $X = x$, for each d, $Y_d = q(d, x, U_d)$, where $q(d, x, \tau)$ is strictly increasing in τ and $U_d \sim U(0,1)$.[10]
A2. Independence: Conditional on $X = x$, $\{U_d\}$ are independent of Z.

[10] Note that U_d here should be understood as the conditional rank given $X = x$.

A3. Selection: $D \equiv \delta(Z, X, V)$ for some unknown function δ and random vector V.
A4. Rank Invariance (RI) or Rank Similarity (RS): Conditional on $X = x$, $Z = z$, (a) $\{U_d\}$ are equal to each other; or, more generally, (b) $\{U_d\}$ are identically distributed, conditional on V.
A5. Observed Variables: Observed variables consist of $Y \equiv q(D, X, U_D)$, D, X, and Z.[11]

The frameworks of CH and HV do not include each other. Roughly speaking, HV's setup is more general in the outcome equations while CH's setup is more general in the selection equation. For example, by the Skorohod representation, HV's framework represents Y_d as

$$Y_d = q(d, X, V, U_d) \text{ with } U_d|\,(X, V) \sim U(0, 1),$$

in other words, there are two random errors in each counterfactual outcome, while Y_d in CH's framework has only one random error; see Yu (2014) for more discussions on these two representations of Y_d. Under rank invariance, $Y = q(D, X, U)$ in CH's framework while $Y = q(D, X, V, U)$ in HV's framework, where $U = U_1 = U_0$. On the other hand, Assumption A3 does not impose any restriction on D, e.g., there is no restriction on the dimension of V, while HV assume D takes the additive latent index form (2) with only one random error. Under the general setup of outcome equations, the monotonicity assumption, which is implied by the indexed choice model, is hard to relax; see Section 6 of HV for more discussions. Given the indexed structure of D, it is without loss of generality to assume $Z \perp V|X$; see the discussions in Section 2. Because we maintain the framework of HV, we strengthen the assumption on D to (3) but maintain the restrictive assumption on Y_d by CH. The relaxing on D and strengthening on Y_d by CH allow for a weaker assumption on the relationship between Z and the error terms. For example, HV require joint independence $(U_1, U_0, U_D) \perp Z|X$, while CH require only $(U_1, U_0) \perp Z|X$ and Z can even be dependent of V. As emphasized in Yu (2015b), the joint independence between Z and (U_d, U_D) (in contrast to $U_d \perp Z|X$ and $U_D \perp Z|X$) is very important for all developments in HV. Finally, note that in HV's framework, V in A3 is equal to U_D and $(U_d, U_D) \perp Z|X$ implies $F_{U_d|X,Z,U_D}(u|x, z, u_D) = F_{U_d|X,U_D}(u|x, u_D)$,[12] so the rank similarity assumption can be restated as

Assumption RS: $F_{U_1|U_D}(u|u_D) = F_{U_0|U_D}(u|u_D)$ for $u \in [0, 1]$ and $u_D \in [0, 1]$, where the conditioning on $X = x$ is depressed.

From this assumption, we denote $F_{U_1|U_D}(u|u_D) = F_{U_0|U_D}(u|u_D)$ as $F_{U|U_D}(u|u_D)$. See Yu (2015a) for testing rank invariance and Dong and Shen (2015) and Yu (2016b) for testing rank similarity in various contexts of treatment effects evaluation.

We now explain why Assumption RS essentially excludes the essential heterogeneity in the context of QTE. Intuitively, $E[U_1 - U_0|U_D] = 0$ reduces two random errors to one in the context of ATE, $F_{U_1|U_D}(u|u_D) = F_{U_0|U_D}(u|u_D)$ is doing the same thing in the context of QTE. $F_{U_1|U_D}(u|u_D) = F_{U_0|U_D}(u|u_D)$ implies that the τth conditional quantiles of U_1 and U_0 given $U_D = u_D$ are the same, which seems a natural counterpart of $E[U_1|U_D] = E[U_0|U_D]$. In the unconfounded case, $F_{U_1|U_D}(\tau|u_D) = \tau = F_{U_0|U_D}(\tau|u_D)$. Under Assumption RS, $F_{U_1|U_D}(\tau|u_D)$ need not be τ but equals $F_{U_0|U_D}(\tau|u_D)$. This is just like the assumption $E[U_1 - U_0|U_D] = 0$: although $E[U_1|U_D = u_D]$ need not be zero, $E[U_1|U_D = u_D] = E[U_0|U_D = u_D]$. On the other hand, Assumption RS does not imply $E[U_1 - U_0|U_D] = 0$. To see why, recall that the definitions of U_1 and U_0 in the ATE case are different from those in the QTE case. The U_d in the former case equals $q_d(U_d) - \int_0^1 q_d(u)du$ in the latter case, where $q_d(u) = q(d, u)$ and the conditioning on $X = x$ is depressed. So $E[U_1 - U_0|U_D = u_D]$ in the

[11] Note that their $U_D = DU_1 + (1 - D)U_0$ is different from our U_D.

[12] Intuitively, if $Z \perp (U_1, U_0)$, but $Z \not\perp U_D$, then Z can affect U_d through its correlation with U_D (which must be correlated with U_d).

former case equals $\int [q_1(u) - q_0(u)]\, dF_{U|U_D}(u|u_D) - \int [q_1(u) - q_0(u)]\, du$ in the latter case under Assumption RS, and need not be zero.[13]

Assumptions A1-A5 imply

$$P\left(Y \leq q(D)|Z\right) = \tau, \tag{8}$$

where $q(d) = q(d, \tau)$. Under a key full rank condition, CH show that there is a unqiue $q = (q(0), q(1))$ satisfying (8) when Z is binary.[14] Before examining this full rank condition, we first check the set $\mathcal{L}$ of (y_0, y_1) defined in their (2.12). First, for $(y_0, y_1) \in \mathcal{L}$, $z = 0, 1$,

$$\begin{aligned}
&P\left(Y < (1-D)y_0 + Dy_1|Z = z\right) \\
&= P\left(Y < (1-D)y_0 + Dy_1|Z = z, D = 0\right) P(D = 0|Z = z) \\
&\quad + P\left(Y < (1-D)y_0 + Dy_1|Z = z, D = 1\right) P(D = 1|Z = z) \\
&= P\left(Y_0 < y_0|Z = z, D = 0\right)(1 - p(z)) + P\left(Y_1 < y_1|Z = z, D = 1\right) p(z) \\
&= \int_{p(z)}^{1} F_{Y_0|U_D}(y_0|u_D)du_D + \int_0^{p(z)} F_{Y_1|U_D}(y_1|u_D)du_D \in [\tau - \delta, \tau + \delta]
\end{aligned} \tag{9}$$

for some $\delta > 0$, where the second to last equality indicates that $P\left(Y \leq q(D)|Z\right)$ is the distribution of a mixed population from observable treated and untreated invidivuals given $Z = z$. Second, for $z \in \{0, 1\}$ with $P(D = d|Z = z) > 0$ and $(y_0, y_1) \in \mathcal{L}$,

$$f_{Y|Z,D}(y_d|z, d) \geq \underline{f} \tag{10}$$

for some $\underline{f} > 0$. Since $F_{Y|Z,D}(y_1|z, 1) = \frac{1}{p(z)} \int_0^{p(z)} F_{Y_1|U_D}(y_1|u_D)du_D$ and $F_{Y|Z,D}(y_0|z, 0) = \frac{1}{1-p(z)} \int_{p(z)}^{1} F_{Y_0|U_D}(y_0|u_D)du_D$,

$$f_{Y|Z,D}(y_1|z, 1) = \frac{1}{p(z)} \int_0^{p(z)} f_{Y_1|U_D}(y_1|u_D)du_D,\ f_{Y|Z,D}(y_0|z, 0) = \frac{1}{1 - p(z)} \int_{p(z)}^{1} f_{Y_0|U_D}(y_0|u_D)du_D,$$

where note that $p(z) = P(D = 1|Z = z) > 0$ and $1 - p(z) = P(D = 0|Z = z) > 0$ are assumed. As a result, (10) is implied by our assumption (Q-4) for any $y_d \in \left(\underline{y}_d, \overline{y}_d\right)$. It seems that the main restriction on $\mathcal{L}$ is (9); (10) is only a regularity condition. The full rank condition is to assume that

$$\Pi'(y_0, y_1) = \begin{pmatrix} f_{Y,D|Z}(y_0, 0|0) & f_{Y,D|Z}(y_1, 1|0) \\ f_{Y,D|Z}(y_0, 0|1) & f_{Y,D|Z}(y_1, 1|1) \end{pmatrix} = \begin{pmatrix} \int_{p(0)}^{1} f_{Y_0|U_D}(y_0|u_D)du_D & \int_0^{p(0)} f_{Y_1|U_D}(y_1|u_D)du_D \\ \int_{p(1)}^{1} f_{Y_0|U_D}(y_0|u_D)du_D & \int_0^{p(1)} f_{Y_1|U_D}(y_1|u_D)du_D \end{pmatrix} \tag{11}$$

is full rank for any $(y_0, y_1) \in \mathcal{L}$, where $\Pi'(y_0, y_1)$ is the Jacobian of the moment equations (8) with respect to (y_0, y_1). It is not hard to see that assumption (Q-4) implies that $\Pi'(y_0, y_1)$ is continuous for $(y_0, y_1) \in \mathcal{L}$, so the regularity condition in CH's Theorem 2 is satisfied. Note that the full rankness of $\Pi'(y_0, y_1)$ is equivalent to a monotone likelihood ratio condition

$$\frac{f_{Y,D|Z}(y_1, 1|1)}{f_{Y,D|Z}(y_0, 0|1)} > \frac{f_{Y,D|Z}(y_1, 1|0)}{f_{Y,D|Z}(y_0, 0|0)}. \tag{12}$$

The following proposition shows that this condition is equivalent to the positivity of the probability of compliers, $p(1) - p(0)$, after regularizing $p(1) \geq p(0)$. Such a condition seems easier to check than (12) since we do not need to check the joint conditional density of Y and D.

[13] As to the definition of selection effect, in the ATE case, it is $E\left[U_0|U_D = u_D\right] \neq 0$ for $u_D \in [0, 1]$, and in the QTE case, it is $F_{U_0|U_D}(u_0|u_D) \neq u_0$ for $u_0 \in [0, 1]$ and $u_D \in [0, 1]$. Obviously, $E\left[U_0|U_D = u_D\right] = \int q_0(u)dF_{U_0|U_D}(u|u_D) - \int q_0(u)du = 0$ if $F_{U_0|U_D}(u_0|u_D) = u_0$, so the selection effect in the ATE case is stronger than the counterpart in the QTE case.

[14] When Z is not binary, we just pick two values that Z can take to identify $(q(0), q(1))$. More general identification assumptions are further discussed in Chernozhukov and Hansen (2013).

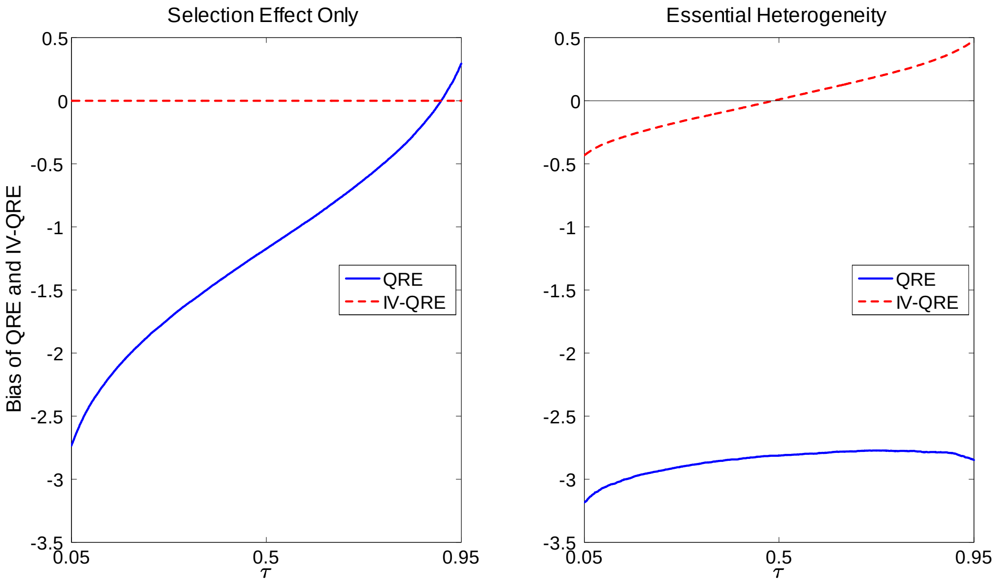

Figure 2: Asymptotic Biases of the QRE and IV-QRE When There is Only Selection Effect and When There is ALSO the Essential Heterogeneity

Proposition 5 *If Assumption (Q-4) holds and* $p(1) \geq p(0)$*, (12) holds if and only if* $p(1) > p(0)$*.*

Proof. From (11), (12) is equivalent to

$$\frac{\int_0^{p(1)} f_{Y_1|U_D}(y_1|u_D)du_D}{\int_{p(1)}^1 f_{Y_0|U_D}(y_0|u_D)du_D} > \frac{\int_0^{p(0)} f_{Y_1|U_D}(y_1|u_D)du_D}{\int_{p(0)}^1 f_{Y_0|U_D}(y_0|u_D)du_D}. \qquad (13)$$

Necessity: If $p(1) = p(0)$, then the left hand side (LHS) is equal to the right hand side (RHS). Sufficiency: When $p(1) > p(0)$, Assumption (Q-4) implies that $\int_{p(0)}^{p(1)} f_{Y_1|U_D}(y_1|u_D)du_D > 0$ and $\int_{p(0)}^{p(1)} f_{Y_0|U_D}(y_0|u_D)du_D > 0$ for $(y_0, y_1) \in \mathcal{L}$, which implies that the numerator of the LHS is larger and the denominator is smaller than the counterparts of the RHS. So (12) is equivalent to $p(1) > p(0)$, i.e., Z has a nontrivial impact on D (rather than (Y, D) as in (12)). ■

We close this subsection by a simple example to illustrate the bias of the QRE when the selection effect exists and the bias of the IV-QRE when the essential heterogeneity exists.

Example 3 *Consider the setup in the running example. The left panel of Figure 2 shows that although the IV-QRE is consistent to estimate the QTE, the QRE is inconsistent when the selection effect exists. The right panel shows that when there is also the essential heterogeneity, the IV-QRE is not consistent to estimate the QTE either although compared with the QRE, it corrects part of the endogeneity bias. In summary, the QRE is inconsistent as long as* D *is endogenous, and the IV-QRE is consistent only when the endogeneity comes solely from the selection effect.*

4.2 Pseudo-true Values

First, similar to the IVE, suppose the moment conditions used by the IV-QRE are

$$E\left[\left(\begin{array}{c}1\\ J(Z)\end{array}\right)\left(\tau-1\left(Y\leq Q_0^*(\tau)+D\cdot\Delta^*(\tau)\right)\right)\right]=0,$$

where $J(Z)$ is a general scalar instrument, and note that we use $Q_d(\cdot)$ instead of $q(d,\cdot)$ as in CH to denote the true quantile functions.

Theorem 2 *Under Assumptions P and Q, when $J(Z)=p(Z)$,*

$$\begin{aligned}F_1^*(y_1) &= \int_0^{\underline{p}} F_{Y_1|U_D}(y_1|u_D)du_D+\int_{\underline{p}}^{\overline{p}}\left[F_{Y_1|U_D}(y_1|u_D)(1-F_{p(Z)}(u_D))+F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)|u_D)F_{p(Z)}(u_D)\right]du_D\\ &\quad+\int_{\overline{p}}^1 F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)|u_D)du_D,\end{aligned}$$

and

$$\begin{aligned}F_0^*(y_0) &= \int_{\overline{p}}^1 F_{Y_0|U_D}(y_0|u_D)du_D+\int_{\underline{p}}^{\overline{p}}\left[F_{Y_1|U_D}(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)|u_D)(1-F_{p(Z)}(u_D))+F_{Y_0|U_D}(y_0|u_D)F_{p(Z)}(u_D)\right]du_D\\ &\quad+\int_0^{\underline{p}} F_{Y_1|U_D}(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)|u_D)du_D,\end{aligned}$$

where

$$\widetilde{F}_1\left(y_1\right)=\int_{\underline{p}}^{\overline{p}}F_{Y_1|U_D}(y_1|u_D)h\left(u_D\right)du_D \text{ and } \widetilde{F}_0\left(y_0\right)=\int_{\underline{p}}^{\overline{p}}F_{Y_0|U_D}(y_0|u_D)h\left(u_D\right)du_D,$$

and $h\left(u_D\right)$ is defined in Theorem 1.

As in the μ_d^* case, the three terms of $F_d^*(y_d)$ can be combined since $F_{p(Z)}(u_D)=0$ for $u_D\in[0,\underline{p}]$ and $F_{p(Z)}(u_D)=1$ for $u_D\in[\overline{p},1]$. The weight $h\left(u_D\right)$ in $\widetilde{F}_d$ is exactly the weight in the IVE case. Note that since $h\left(u_D\right)$ is nonnegative with $\int h\left(u_D\right)du_D=1$ and $h\left(0\right)=h\left(1\right)=0$, both $\widetilde{F}_1\left(\cdot\right)$ and $\widetilde{F}_0\left(\cdot\right)$ are proper cdfs such that $\widetilde{F}_d^{-1}(\cdot)$ in $F_d^*(y_d)$ is well defined. They are counterfactual cdfs for a mixed population with the weight $h(u_D)$ on the subpopulation $U_D=u_D$.

Since

$$F_d(y_d)=\int F_{Y_d|U_D}(y_d|u_D)du_D=\int_0^{\underline{p}}F_{Y_d|U_D}(y_d|u_D)du_D+\int_{\underline{p}}^{\overline{p}}F_{Y_d|U_D}(y_d|u_D)du_D+\int_{\overline{p}}^1F_{Y_d|U_D}(y_d|u_D)du_D,$$

and we can *only* identify $F_{Y_1|U_D}(y_1|u_D)$ and $F_{Y_0|U_D}(y_0|u_D)$ for $u_D\in\left[\underline{p},\overline{p}\right]$ by (see, e.g., Yu (2014))

$$F_{Y_1|U_D}(y_1|u_D)=\left.\frac{dE\left[1\left(Y\leq y_d\right)D|p(Z)=p\right]}{dp}\right|_{p=u_D},F_{Y_0|U_D}(y_0|u_D)=-\left.\frac{dE\left[1\left(Y\leq y_d\right)\left(1-D\right)|p(Z)=p\right]}{dp}\right|_{p=u_D},$$

the true cdfs $F_1(y_1)$ and $F_0(y_0)$ are not identifiable. As an alternative identification scheme, the IV-QRE uses the weighted average of $F_{Y_1|U_D}(y_1|u_D)$ and $F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)|u_D)$ $(F_{Y_1|U_D}(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)|u_D)$ and $F_{Y_0|U_D}(y_0|u_D))$ to approximate $F_{Y_1|U_D}(y_1|u_D)$ $(F_{Y_0|U_D}(y_0|u_D))$ for $u_D\in\left[\underline{p},\overline{p}\right]$ and uses $F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)|u_D)$ $(F_{Y_1|U_D}(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)|u_D))$ to approximate $F_{Y_1|U_D}(y_1|u_D)$ $(F_{Y_0|U_D}(y_0|u_D))$ for $u_D\in[\overline{p},1]$ $(u_D\in\left[0,\underline{p}\right])$. This is somewhat similar to the identification scheme of μ_d^*. Note that

$$\int_0^{\underline{p}} F_{Y_1|U_D}(y_1|u_D)du_D = \underline{p}F_{Y|D=1,p(Z)=\underline{p}}(y_1), \int_{\overline{p}}^1 F_{Y_0|U_D}(y_0|u_D)du_D = (1-\overline{p})F_{Y|D=0,p(Z)=\overline{p}}(y_0),$$

and $\widetilde{F}_d\left(y_d\right)$ are all identifiable from the data, where $F_{Y|D=1,p(Z)=\underline{p}}(y_1)$ is the cdf of Y_1 for always takers and $F_{Y|D=0,p(Z)=\overline{p}}(y_0)$ is the cdf of Y_0 for never-takers, and both are identifiable, so the IV-QRE uses the identifiable to approximate the unidentifiable.[15] The approximation error depends on how close $F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)|u_D)$ ($F_{Y_1|U_D}(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)|u_D)$) is to $F_{Y_1|U_D}(y_1|u_D)$ ($F_{Y_0|U_D}(y_0|u_D)$) for $u_D \in [\underline{p}, 1]$ ($u_D \in [0, \overline{p}]$). In the unconfounded case, $F_{Y_d|U_D}(y_d|u_D) = F_d(y_d)$, so $\widetilde{F}_d\left(y_d\right) = F_d\left(y_d\right)$ and thus

$$\begin{aligned} F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)|u_D) &= F_0(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)) = F_0(F_0^{-1}F_1\left(y_1\right)) = F_1\left(y_1\right) = F_{Y_1|U_D}(y_1|u_D), \\ F_{Y_1|U_D}(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)|u_D) &= F_1(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)) = F_1(F_1^{-1}F_0\left(y_0\right)) = F_0\left(y_0\right) = F_{Y_0|U_D}(y_0|u_D), \end{aligned}$$

for any $u_D \in [0,1]$ and the IV-QRE of $Q_d(\cdot)$ is consistent.[16]

From the proof of Theorem 2,

$$\begin{aligned} F_1^*\left(y\right) &= E\left[p(Z)\cdot F_{Y|D=1,p(Z)}\left(y\right) + (1-p(Z))\cdot F_{Y|D=0,p(Z)}\left(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y\right)\right)\right], \\ F_0^*\left(y\right) &= E\left[p(Z)\cdot F_{Y|D=1,p(Z)}\left(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y\right)\right) + (1-p(Z))\cdot F_{Y|D=0,p(Z)}\left(y\right)\right]. \end{aligned}$$

In other words, $F_1^*(\cdot)$ is an expected weighted average of $F_{Y|D=1,p(Z)}\left(\cdot\right)$ and $F_{Y|D=0,p(Z)}\left(\widetilde{F}_0^{-1}\widetilde{F}_1\left(\cdot\right)\right)$, while $F_0^*\left(\cdot\right)$ is an expected weighted average of $F_{Y|D=1,p(Z)}\left(\widetilde{F}_1^{-1}\widetilde{F}_0\left(\cdot\right)\right)$ and $F_{Y|D=0,p(Z)}\left(\widetilde{F}_0^{-1}\widetilde{F}_1\left(\cdot\right)\right)$, and all these components and the weights are estimable.

In the unconfounded case, $F_{Y_d|U_D}(y_d|u_D) = F_d(y_d)$, and thus the IV-QRE of $Q_d(\cdot)$ is consistent. The amazing result of CH is that in the absence of unconfoundedness, when the ranks of Y_1 and Y_0 for individuals with $U_D = u_D$ are preserved, they can still prove $F_d^*(\cdot) = F_d(\cdot)$. One may suspect that this is because

$$\begin{aligned} 0 \neq & \int_{\underline{p}}^{\overline{p}} \left[F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)|u_D) - F_{Y_1|U_D}(y_1|u_D)\right] F_{p(Z)}(u_D)du_D \\ = & \int_{\overline{p}}^1 \left[F_{Y_1|U_D}(y_1|u_D) - F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)|u_D)\right] du_D \neq 0, \end{aligned}$$

and

$$\begin{aligned} 0 \neq & \int_{\underline{p}}^{\overline{p}} \left[F_{Y_1|U_D}(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)|u_D) - F_{Y_0|U_D}(y_0|u_D)\right] (1-F_{p(Z)}(u_D))du_D \\ = & \int_0^{\underline{p}} \left[F_{Y_0|U_D}(y_0|u_D) - F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_0\right)|u_D)\right] du_D \neq 0 \end{aligned}$$

as in the IVE case, i.e., the biases in estimating $F_{Y_1|U_D}(y_1|u_D)$ ($F_{Y_0|U_D}(y_0|u_D)$) over $u_D \in [\underline{p}, \overline{p}]$ and over $u_D \in [\overline{p}, 1]$ ($[0, \underline{p}]$) offset each other. The following proposition shows that under rank similarity, this is not the case.

[15] In S.2.2, we provide alternative formulas for $F_d^*(y_d)$ to show that it is estimable.

[16] For point identification of $(Q_1(\tau), Q_0(\tau))$ in the unconfounded case, note that since $f_{Y_1|U_D}(y_1|u_D)$ and $f_{Y_0|U_D}(y_0|u_D)$ does not depend on u_D, (13) reduces to $p(1)/(1-p(1)) > p(0)/(1-p(0))$, which is equivalent to $p(1) > p(0)$ as in Proposition 5.

Proposition 6 *Under Assumption RS,* $F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)|u_D) = F_{Y_1|U_D}(y_1|u_D)$*, and* $F_{Y_1|U_D}(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)|u_D) = F_{Y_0|U_D}(y_0|u_D)$ *for any* $u_D \in [0,1]$ *although* $F_{Y_d|U_D}(y_d|u_D)$ *need not equal* $F_{Y_d}(y_d)$ *as in the unconfounded case.*

Proof. We only prove the first statement since the second statement can be similarly proved. Recall that under Assumption RS, the common conditional cdf of U_1 and U_0 given $U_D = u_D$ is denoted as $F_{U|U_D}(\cdot|u_D)$. The key implication of Assumption RS is $\widetilde{F}_1\left(\cdot\right) = F_U^h(F_1\left(\cdot\right))$ and $\widetilde{F}_0\left(\cdot\right) = F_U^h(F_0\left(\cdot\right))$, where the same mixture cdf $F_U^h \equiv \int F_{U|U_D}(\cdot|u_D)h\left(u_D\right)du_D$ appears in both $\widetilde{F}_1$ and $\widetilde{F}_0$. To see why, first note that

$$F_{Y_1|U_D}(y_1|u_D) = P\left(Y_1 \leq y_1|U_D = u_D\right) = P\left(U_1 \leq F_1\left(y_1\right)|U_D = u_D\right) = F_{U_1|U_D}(F_1\left(y_1\right)|u_D),$$

where the second equality is from Assumption A1, so

$$\widetilde{F}_1\left(y_1\right) = \int F_{Y_1|U_D}(y_1|u_D)h\left(u_D\right)du_D = \int F_{U|U_D}(F_1\left(y_1\right)|u_D)h\left(u_D\right)du_D = F_U^h(F_1\left(y_1\right)),$$

and similarly, $\widetilde{F}_0\left(y_0\right) = F_U^h(F_0\left(y_0\right))$. Since $F_U^h(\cdot)$ need not be an identity function, $\widetilde{F}_d\left(\cdot\right)$ need not equal $F_d\left(\cdot\right)$ as in the unconfounded case.

Now,

$$\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right) = F_0^{-1}\left(F_U^h\right)^{-1}F_U^h(F_1\left(y_1\right)) = F_0^{-1}(F_1\left(y_1\right)).$$

On the other hand, since $F_{Y_d|U_D}(y_d|u_D) = F_{U|U_D}(F_d\left(y_d\right)|u_D)$,

$$F_{Y_0|U_D}^{-1}\left(F_{Y_1|U_D}(y_1|u_D)|u_D\right) = F_0^{-1}\left(F_{U|U_D=u_D}\right)^{-1}F_{U|U_D=u_D}(F_1\left(y_1\right)) = F_0^{-1}(F_1\left(y_1\right)).$$

This finishes the proof. ■

The above proof shows that if the pseudo-true values of an estimator take the form in Theorem 2, then as long as the weight $h(\cdot)$ satisfies Proposition 1(iii) (especially, $h(u_D) \geq 0$ for $u_D \in [\underline{p}, \overline{p}]$ and $\int_{\underline{p}}^{\overline{p}} h(u_D)du_D = 1$), the results of Proposition 6 hold. For future reference, we label the property in Proposition 6 as *counterfactual-quantiles matching*. This following example numerically illustrates this property.

Example 4 *Consider the running example with only the selection effect. In this simple example,*

$$\begin{aligned}
F_{Y_1|U_D}(y_1|u_D) &= P\left(2U \leq y_1|V = \Phi^{-1}(u_D)\right) = \Phi\left(\frac{y_1/2 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right), \\
F_{Y_0|U_D}(y_0|u_D) &= P\left(U \leq y_0|V = \Phi^{-1}(u_D)\right) = \Phi\left(\frac{y_0 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right), \\
\widetilde{F}_1(y_1) &= \int_0^1 F_{Y_1|U_D}(y_1|u_D)h\left(u_D\right)du_D = \int_0^1 \Phi\left(\frac{y_1/2 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)6u_D(1-u_D)du_D, \\
\widetilde{F}_0(y_0) &= \int_0^1 F_{Y_0|U_D}(y_0|u_D)h\left(u_D\right)du_D = \int_0^1 \Phi\left(\frac{y_0 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)6u_D(1-u_D)du_D;
\end{aligned}$$

also, we need only consider the middle term of $F_1^*(y_1)$ *and* $F_0^*(y_0)$*, i.e.,*

$$\begin{aligned}
F_1^*(y_1) &= \int_0^1 \left[\Phi\left(\frac{y_1/2 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)(1-u_D) + \Phi\left(\frac{\widetilde{F}_0^{-1}\left(\widetilde{F}_1\left(y_1\right)\right) - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)u_D\right]du_D, \\
F_0^*(y_0) &= \int_0^1 \left[\Phi\left(\frac{\widetilde{F}_1^{-1}\left(\widetilde{F}_0\left(y_0\right)\right)/2 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)(1-u_D) + \Phi\left(\frac{y_0 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)u_D\right]du_D.
\end{aligned}$$

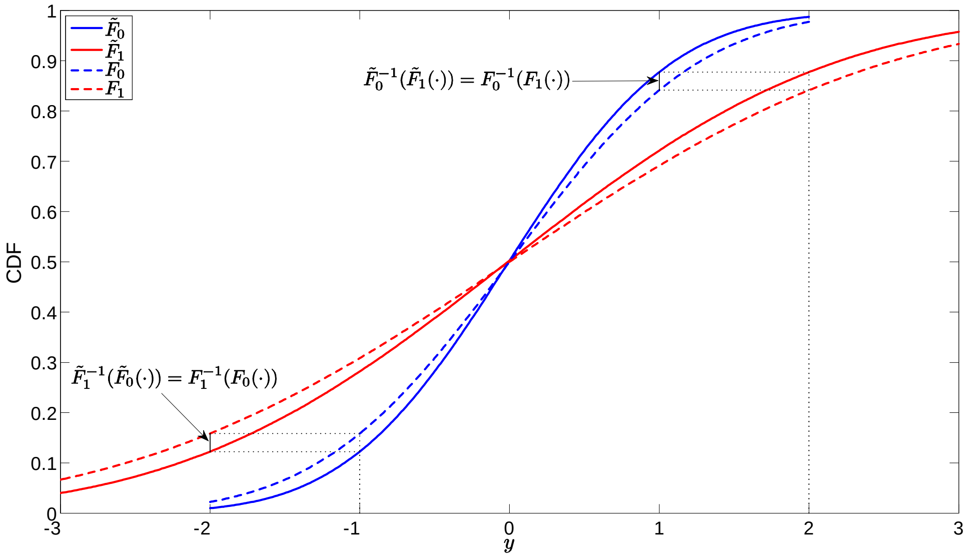

Figure 3: $\widetilde{F}_1(\cdot) \neq F_1(\cdot)$ and $\widetilde{F}_0(\cdot) \neq F_0(\cdot)$, but $\widetilde{F}_0^{-1}\left(\widetilde{F}_1(\cdot)\right) = F_0\left(F_1(\cdot)\right)$ and $\widetilde{F}_1^{-1}\left(\widetilde{F}_0(\cdot)\right) = F_1\left(F_0(\cdot)\right)$

The true cdfs are

$$\begin{aligned} F_1(y_1) &= \Phi\left(\frac{y_1}{2}\right) = \int_0^1 \Phi\left(\frac{y_1/2 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right) du_D, \\ F_0(y_0) &= \Phi\left(y_0\right) = \int_0^1 \Phi\left(\frac{y_0 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right) du_D, \end{aligned}$$

so the difference between $F_1(y_1)$ and $F_1^*(y_1)$ ($F_0(y_0)$ and $F_0^*(y_0)$) depends on how different between $\widetilde{F}_0^{-1}\left(\widetilde{F}_1\left(y_1\right)\right)$ and $F_0^{-1}\left(F_1\left(y_1\right)\right) = y_1/2$ ($\widetilde{F}_1^{-1}\left(\widetilde{F}_0\left(y_0\right)\right)$ and $F_1^{-1}\left(F_0\left(y_0\right)\right) = 2y_0$). Although different from the unconfounded case, $\widetilde{F}_1(y_1) \neq F_1(y_1)$ and $\widetilde{F}_0(y_0) \neq F_0(y_0)$, $\widetilde{F}_0^{-1}\left(\widetilde{F}_1\left(y_1\right)\right) = F_0^{-1}\left(F_1\left(y_1\right)\right)$ and $\widetilde{F}_1^{-1}\left(\widetilde{F}_0\left(y_0\right)\right) = F_1^{-1}\left(F_0\left(y_0\right)\right)$ as shown in Figure 3.

On the contrary, the following example shows that with essential heterogeneity, $\widetilde{F}_0^{-1}\left(\widetilde{F}_1\left(y_1\right)\right) \neq F_0^{-1}\left(F_1\left(y_1\right)\right)$ and $\widetilde{F}_1^{-1}\left(\widetilde{F}_0\left(y_0\right)\right) \neq F_1^{-1}\left(F_0\left(y_0\right)\right)$.

Example 5 *Consider the running example with both the selection effect and the essential heterogeneity. In this case,*

$$\begin{aligned} F_{Y_1|U_D}(y_1|u_D) &= P\left(V + 2U \leq y_1 | V = \Phi^{-1}(u_D)\right) = \Phi\left(\frac{y_1/2 - 1.2\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right), \\ F_{Y_0|U_D}(y_0|u_D) &= P\left(2V + U \leq y_0 | V = \Phi^{-1}(u_D)\right) = \Phi\left(\frac{y_0 - 2.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right), \\ \widetilde{F}_1(y_1) &= \int_0^1 F_{Y_1|U_D}(y_1|u_D) h\left(u_D\right) du_D = \int_0^1 \Phi\left(\frac{y_1/2 - 1.2\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right) 6u_D(1-u_D) du_D, \\ \widetilde{F}_0(y_0) &= \int_0^1 F_{Y_0|U_D}(y_0|u_D) h\left(u_D\right) du_D = \int_0^1 \Phi\left(\frac{y_0 - 2.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right) 6u_D(1-u_D) du_D, \end{aligned}$$

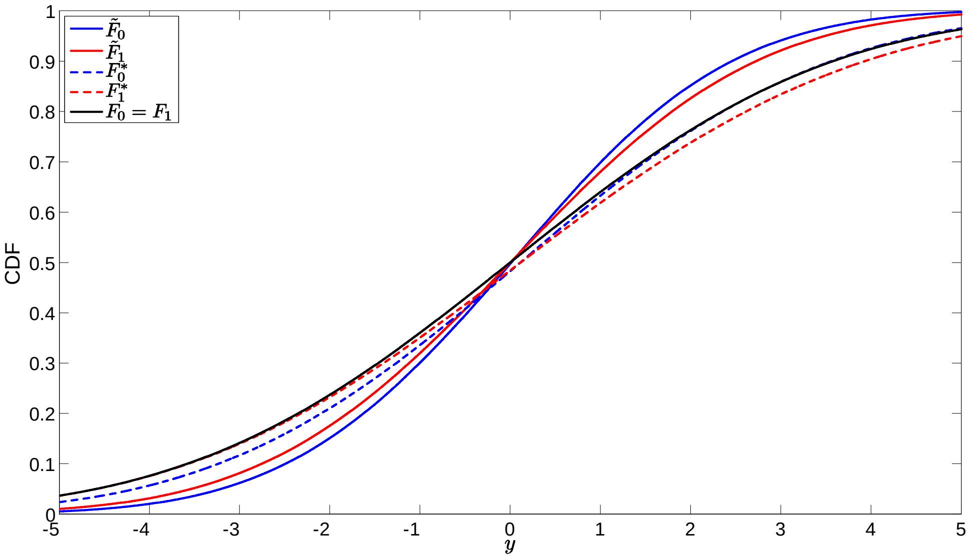

Figure 4: $\widetilde{F}_0^{-1}\left(\widetilde{F}_1\left(y_1\right)\right) \neq F_0^{-1}\left(F_1\left(y_1\right)\right)$, $\widetilde{F}_1^{-1}\left(\widetilde{F}_0\left(y_0\right)\right) \neq F_1^{-1}\left(F_0\left(y_0\right)\right)$, $F_1^*(\cdot) \neq F_1(\cdot) = F_0(\cdot) \neq F_0^*(\cdot)$

and

$$\begin{aligned} F_1^*(y_1) &= \int_0^1 \left[\Phi\left(\frac{y_1/2 - 1.2\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)(1-u_D) + \Phi\left(\frac{\widetilde{F}_0^{-1}\left(\widetilde{F}_1\left(y_1\right)\right) - 2.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)u_D\right]du_D, \\ F_0^*(y_0) &= \int_0^1 \left[\Phi\left(\frac{\widetilde{F}_1^{-1}\left(\widetilde{F}_0\left(y_0\right)\right)/2 - 1.2\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)(1-u_D) + \Phi\left(\frac{y_0 - 2.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)u_D\right]du_D. \end{aligned}$$

The true cdfs are

$$\begin{aligned} F_1(y_1) &= \Phi\left(\frac{y_1}{\sqrt{7.8}}\right) = \int_0^1 \Phi\left(\frac{y_1/2 - 1.2\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)du_D, \\ F_0(y_0) &= \Phi\left(\frac{y_0}{\sqrt{7.8}}\right) = \int_0^1 \Phi\left(\frac{y_0 - 2.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right)du_D. \end{aligned}$$

Figure 4 shows that not only $\widetilde{F}_1(y_1) \neq F_1(y_1)$ *and* $\widetilde{F}_0(y_0) \neq F_0(y_0)$*, but* $\widetilde{F}_0^{-1}\left(\widetilde{F}_1\left(y_1\right)\right) \neq F_0^{-1}\left(F_1\left(y_1\right)\right) = y_1$ *and* $\widetilde{F}_1^{-1}\left(\widetilde{F}_0\left(y_0\right)\right) \neq F_1^{-1}\left(F_0\left(y_0\right)\right) = y_0$*. Figure 4 also shows* $F_1^*(y_1) \neq F_1(y_1)$ *and* $F_0^*(y_0) \neq F_0(y_0)$*, which explains why the IV-QRE of the QTE is not consistent in Figure 2.*

Finally, given $F_d^*(\cdot)$, we can express $\Delta^*(\tau)$ in the following corollary.

Corollary 1 *Under Assumptions P and Q, when* $J(Z) = p(Z)$*,*

$$\Delta^*(\tau) = \widetilde{\Delta}\left(\widetilde{F}_0\left(Q_0^*(\tau)\right)\right) = \widetilde{\Delta}\left(\widetilde{F}_1\left(Q_1^*(\tau)\right)\right),$$

where $\widetilde{\Delta}\left(\cdot\right) = \widetilde{F}_1^{-1}\left(\cdot\right) - \widetilde{F}_0^{-1}\left(\cdot\right)$*.*

Proof. From the proof of Theorem 2, $\widetilde{F}_1\left(Q_1^*(\tau)\right)=\widetilde{F}_0\left(Q_0^*(\tau)\right)$, so

$$\begin{aligned}\Delta^*(\tau) &= \widetilde{F}_1^{-1}\widetilde{F}_0\left(Q_0^*(\tau)\right)-Q_0^*(\tau)=\widetilde{F}_1^{-1}\widetilde{F}_0\left(Q_0^*(\tau)\right)-\widetilde{F}_0^{-1}\widetilde{F}_0\left(Q_0^*(\tau)\right)\\ &= \widetilde{\Delta}\left(\widetilde{F}_0\left(Q_0^*(\tau)\right)\right)=\widetilde{\Delta}\left(\widetilde{F}_1\left(Q_1^*(\tau)\right)\right).\end{aligned}$$

■

This corollary states that the pseudo-QTE is the QTE for a mixed population at transformed quantile levels. This interesting result is the main result of Wüthrich (2015) as stated in his Theorem 2 and 7. This corollary generalizes his result to the continuous and/or multiple instruments case where the definition of the mixed population is different.

We close this subsection by two further comments. First, given F_d^*, we can estimate μ_d by $\int y_d dF_d^*(y_d)$, but this estimator need not equal μ_d^* even if Assumption RS holds; S.2.5 provide more discussions and some numerical illustration. Second, although $\widetilde{F}_d$ is monotone, its finite sample analog need not be, but we can rearrange the original estimator before inversion as suggested by Chernozhukov et al. (2010). An interesting problem is to compare the estimator of $\Delta^*(\tau)$ which rearranges the inverse quantile regression estimator of Chernozhukov and Hansen (2006) and our estimator of $\Delta^*(\tau)$ which rearranges $\widetilde{F}_d$ in the expression of F_d^*, but a detailed comparison is beyond the scope of this paper.

4.3 Local Sensitivity Analysis

From Proposition 6, the key assumption to guarantee the consistency of IV-QRE is

$$F_{Y_1|U_D}(y_1|u_D)=F_{U|U_D}(F_1(y_1)|u_D) \text{ and } F_{Y_0|U_D}(y_0|u_D)=F_{U|U_D}(F_0(y_0)|u_D),$$

for the same function $F_{U|U_D}(\cdot|u_D)$ for any $u_D\in\left[\underline{p},\overline{p}\right]$. Generally,

$$F_{Y_1|U_D}(y_1|u_D)=F_{U_1|U_D}(F_1(y_1)|u_D) \text{ and } F_{Y_0|U_D}(y_0|u_D)=F_{U_0|U_D}(F_0(y_0)|u_D),$$

where $F_{U_1|U_D}(\cdot|\cdot)$ and $F_{U_0|U_D}(\cdot|\cdot)$ may not be the same,[17] so we assume $F_{U_1|U_D}(\cdot|\cdot)-F_{U_0|U_D}(\cdot|\cdot)$ falls in the parametrized space

$$\mathcal{F}_\alpha^*=\left\{F^\alpha(\cdot|\cdot):\alpha\in M,0\in M,F^\alpha(0|u_D)=F^\alpha(1|u_D)=0,\int_0^1F^\alpha(u|u_D)du_D=0,F^0(u|u_D)=0,u,u_D\in[0,1]\right\},$$

where $F^\alpha(0|u_D)=F^\alpha(1|u_D)=0$ because $F_{U_1|U_D}(0|u_D)=F_{U_0|U_D}(0|u_D)=0$ and $F_{U_1|U_D}(1|u_D)=F_{U_0|U_D}(1|u_D)=1$, and $\int_0^1F^\alpha(u|u_D)du_D=0$ because $\int_0^1F_{U_1|U_D}(u|u_D)du_D=\int_0^1F_{U_0|U_D}(u|u_D)du_D=u$. If we use $g(\cdot|\cdot)$ for shorthand of $\frac{\partial}{\partial\alpha}F^\alpha(\cdot|\cdot)\big|_{\alpha=0}$, then $g(0|u_D)=g(1|u_D)=0$ and $\int_0^1g(u|u_D)du_D=0$. Because $\int_0^1g(u|u_D)du_D=0$, $g(u|u_D)$ cannot be the same for all $u_D\in[0,1]$ unless $g(u|u_D)=0$. We assume $g(\cdot|\cdot)\in\mathcal{C}\left([0,1]^2\right)$; otherwise, there is a point mass shift in u for some u_D, or there is a sharp change in the shape of $F_{U_d|U_D}(u|u_D)$ for two close u_D values.

Our targets are

$$D_{F_d}^*(g)(y_d)\equiv\lim_{\alpha\to0}\frac{F_d^*(y_d;\alpha)-F_d(y_d)}{\alpha} \text{ and } D_\Delta^*(g)(\tau)\equiv\lim_{\alpha\to0}\frac{\Delta^*(\tau;\alpha)-\Delta(\tau)}{\alpha}$$

[17] Note that $F_{U_1|U_D}(\cdot|\cdot)=F_{U_0|U_D}(\cdot|\cdot)$ means that the two copula functions $C_d(u_d,u_D)=P(U_d\le u_d,U_D\le u_D)$, $d=0,1$, are completely the same. Deviation from $F_{U_1|U_D}(\cdot|\cdot)=F_{U_0|U_D}(\cdot|\cdot)$ means difference between $C_1(u_1,u_D)$ and $C_0(u_0,u_D)$. Note further that since U_d and U_D are not independent, $F_{U_d|U_D}(\cdot|u_D)$ is not the same for all $u_D\in[0,1]$.

or path derivatives of $F_d^*(y_d)$ and $\Delta^*(\tau)$ along the path $\mathcal{F}_\alpha^*$, where we use $F_d^*(y_d;\alpha)$ and $\Delta^*(\tau;\alpha)$ to indicate the dependence of F_d^* and Δ^* on α, assume $\lim_{\alpha\to 0}\frac{F^\alpha(\cdot|\cdot)}{\alpha}\to g(\cdot|\cdot)$, and note that $F_d^*(y_d;0)=F_d(y_d)$ and $\Delta^*(\tau;0)=\Delta(\tau)$. We impose the following conditions to guarantee the limits in $D_{F_d}^*$ and D_Δ^* exist.

Assumption LF: $\sup_{u_D\in[0,1],u\in[0,1],\alpha\in\mathcal{N}}\left|\frac{F^\alpha(u|u_D)}{\alpha}\right|<\infty$, where $\mathcal{N}$ is a neighborhood of 0.

Proposition 7 *Under Assumptions P, Q and LF,*

$$
\begin{aligned}
D_{F_1}^*(g)(y_1) &= \int_{\underline{p}}^{\overline{p}} D_{F_{Y_1|U_D}}(g)(y_1|u_D)\,F_{p(Z)}(u_D)du_D+\int_{\overline{p}}^{1} D_{F_{Y_1|U_D}}(g)(y_1|u_D)\,du_D\\
&= \int_0^1 D_{F_{Y_1|U_D}}(g)(y_1|u_D)\,F_{p(Z)}(u_D)du_D,\\
D_{F_0}^*(g)(y_0) &= \int_0^{\underline{p}} D_{F_{Y_0|U_D}}(g)(y_0|u_D)\,du_D+\int_{\underline{p}}^{\overline{p}} D_{F_{Y_0|U_D}}(g)(y_0|u_D)\,(1-F_{p(Z)}(u_D))du_D\\
&= \int_0^1 D_{F_{Y_0|U_D}}(g)(y_0|u_D)\,(1-F_{p(Z)}(u_D))du_D,
\end{aligned}
$$

and

$$
D_\Delta^*(g)(\tau)=\frac{D_{F_0}^*(g)\left(F_0^{-1}(\tau)\right)}{f_0\left(F_0^{-1}(\tau)\right)}-\frac{D_{F_1}^*(g)\left(F_1^{-1}(\tau)\right)}{f_1\left(F_1^{-1}(\tau)\right)},
$$

where

$$
\begin{aligned}
D_{F_{Y_1|U_D}}(g)(y_1|u_D) &= \frac{f_{Y_0|U_D}\left(F_0^{-1}F_1(y_1)\,|u_D\right)}{\widetilde{f}_0\left(F_0^{-1}F_1(y_1)\right)}\int_{\underline{p}}^{\overline{p}} g(F_1(y_1)\,|u_D)h(u_D)\,du_D-g(F_1(y_1)\,|u_D),\\
D_{F_{Y_0|U_D}}(g)(y_0|u_D) &= g(F_0(y_0)\,|u_D)-\frac{f_{Y_1|U_D}\left(F_1^{-1}F_0(y_0)\,|u_D\right)}{\widetilde{f}_1\left(F_1^{-1}F_0(y_0)\right)}\int_{\underline{p}}^{\overline{p}} g(F_0(y_0)\,|u_D)h(u_D)\,du_D
\end{aligned}
$$

$f_{Y_d|U_D}$ *is the pdf of* $F_{Y_d|U_D}$, $\widetilde{f}_d(\cdot)=\int_{\underline{p}}^{\overline{p}} f_{Y_d|U_D}(\cdot|u_D)h(u_D)\,du_D$ *is the pdf of* $\widetilde{F}_d$, *and* f_d *is the pdf of* F_d.

If $g(\cdot|\cdot)\in\mathcal{C}\left([0,1]^2\right)$, then $D_{F_d}^*(g)(y_d)$ is a continuous linear mapping from $\mathcal{C}_0\left([0,1]^2\right)$, a subspace of $\mathcal{C}\left([0,1]^2\right)$, to $\mathcal{C}(\mathcal{Y}_d)$, and $D_\Delta^*(g)(\tau)$ is a continuous linear mapping from $\mathcal{C}_0\left([0,1]^2\right)$ to $\mathcal{C}([0,1])$, where $\mathcal{C}_0\left([0,1]^2\right)\equiv\left\{g(\cdot|\cdot):g\in\mathcal{C}\left([0,1]^2\right),g(0|u_D)=g(1|u_D)=0\text{ and }\int_0^1 g(u|u_D)du_D=0\text{ for }u,u_D\in[0,1]\right\}$.

As $D_1^*(g)$ in Proposition 3, $D_{F_1}^*(g)(y_1)$ depends on $g(F_1(y_1)\,|u_D)$ only for $u_D\in[\underline{p},1]$; as $D_0^*(g)$, $D_{F_0}^*(g)(y_0)$ depends on $g(F_0(y_0)\,|u_D)$ only for $u_D\in[0,\overline{p}]$. This is understandable since only on these areas of u_D, F_1^* and F_0^* have mis-estimated $F_{Y_1|U_D}(y_1|u_D)$ and $F_{Y_0|U_D}(y_0|u_D)$, respectively.

$D_{F_{Y_1|U_D}}(y_1|u_D)$ is the effect of mis-estimating $F_{Y_1|U_D}(y_1|u_D)$ as $F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1(y_1)\,|u_D)$ and $D_{F_{Y_0|U_D}}(y_0|u_D)$ is the effect of mis-estimating $F_{Y_0|U_D}(y_0|u_D)$ as $F_{Y_1|U_D}(\widetilde{F}_1^{-1}\widetilde{F}_0(y_0)\,|u_D)$. The terms in $D_{F_{Y_1|U_D}}(y_1|u_D)$ and $D_{F_{Y_0|U_D}}(y_0|u_D)$ can be well interpreted. For example, the first term of $D_{F_{Y_1|U_D}}(y_1|u_D)$ is from the misspecification of $F_{U_1|U_D}$ as $F_{U_0|U_D}$ in $\widetilde{F}_0$ of $F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1(y_1)\,|u_D)$, and the second term is from the misspecification of $F_{U_1|U_D}$ as $F_{U_0|U_D}$ in $F_{Y_0|U_D}$ of $F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1(y_1)\,|u_D)$. Note that

$$
\begin{aligned}
D_{F_1}^*(g)(y_1) \;=\; & \frac{\int_{\underline{p}}^{\overline{p}} f_{Y_0|U_D}\left(F_0^{-1}F_1(y_1)\,|u_D\right)F_{p(Z)}(u_D)du_D+\int_{\overline{p}}^{1} f_{Y_0|U_D}\left(F_0^{-1}F_1(y_1)\,|u_D\right)du_D}{\int_{\underline{p}}^{\overline{p}} f_{Y_0|U_D}(F_0^{-1}F_1(y_1)\,|u_D)h(u_D)\,du_D}\\
&\cdot\int_{\underline{p}}^{\overline{p}} g(F_1(y_1)\,|u_D)h(u_D)\,du_D-\left[\int_{\underline{p}}^{\overline{p}} g(F_1(y_1)\,|u_D)F_{p(Z)}(u_D)du_D+\int_{\overline{p}}^{1} g(F_1(y_1)\,|u_D)du_D\right],
\end{aligned}
$$

where all terms (except the terms involving g which need to be specified) are estimable. For example, although $f_{Y_0|U_D}(\cdot|u_D)$ for $u_D \in (\overline{p}, 1]$ is not estimable, $\int_{\overline{p}}^{1} f_{Y_0|U_D}(\cdot|u_D)\, du_D$ can be estimated by the derivative of $\int_{\overline{p}}^{1} F_{Y_0|U_D}(\cdot|u_D) du_D$ which is equal to $(1-\overline{p})F_{Y|D=0,p(Z)=\overline{p}}(\cdot)$ and is estimable. Other terms such as $F_d(\cdot)$, $F_{p(Z)}(\cdot)$ and $f_{Y_0|U_D}(\cdot|u_D)$ for $u_D \in [\underline{p}, \overline{p}]$ are all estimable. Similarly,

$$\begin{aligned} D^*_{F_0}(g)(y_0) &= -\frac{\int_0^{\underline{p}} f_{Y_1|U_D}\left(F_1^{-1}F_0(y_0)|u_D\right) du_D + \int_{\underline{p}}^{\overline{p}} f_{Y_1|U_D}\left(F_1^{-1}F_0(y_0)|u_D\right)(1-F_{p(Z)}(u_D))du_D}{\int_{\underline{p}}^{\overline{p}} f_{Y_1|U_D}(F_1^{-1}F_0(y_0)|u_D)h(u_D)\, du_D} \\ &\quad \cdot \int_{\underline{p}}^{\overline{p}} g(F_0(y_0)|u_D)h(u_D)\, du_D + \left[\int_0^{\underline{p}} g(F_0(y_0)|u_D)du_D + \int_{\underline{p}}^{\overline{p}} g(F_0(y_0)|u_D)(1-F_{p(Z)}(u_D))du_D\right], \end{aligned}$$

where although $f_{Y_1|U_D}(\cdot|u_D)$ for $u_D \in [0, \underline{p})$ is not estimable, $\int_0^{\underline{p}} f_{Y_1|U_D}(\cdot|u_D)\, du_D$ can be estimated by the derivative of $\int_0^{\underline{p}} F_{Y_1|U_D}(\cdot|u_D)\, du_D$ which is equal to $\underline{p}F_{Y|D=1,p(Z)=\underline{p}}(\cdot)$ and is estimable. Finally, $D^*_{\Delta}(g)(\tau) = \frac{D^*_{F_0}(g)\left(F_0^{-1}(\tau)\right)}{f_0\left(F_0^{-1}(\tau)\right)} - \frac{D^*_{F_1}(g)\left(F_1^{-1}(\tau)\right)}{f_1\left(F_1^{-1}(\tau)\right)}$ involves further of $f_d\left(F_d^{-1}(\tau)\right)$, while $f_d\left(F_d^{-1}(\tau)\right)$ can be estimated by $\left.\frac{dF_d(y_d)}{dy_d}\right|_{y_d=F_d^{-1}(\tau)}$ or $\frac{1}{f_d\left(F_d^{-1}(\tau)\right)}$ can be estimated by $\frac{Q_d(\tau)}{d\tau}$, the sparsity function for $Q_d(\cdot)$.

In summary, $D^*_{F_d}(g)(\cdot)$ and $D^*_{\Delta}(g)(\cdot)$ are estimable for a given $g(\cdot|\cdot)$; only $f_{Y_0|U_D}(\cdot|\cdot)$ is involved in $D^*_{F_1}(g)$ and only $f_{Y_1|U_D}(\cdot|\cdot)$ is involved in $D^*_{F_0}(g)$, which is an indication of misspecification. Given $D^*_{F_1}(g)(y_1)$ and $D^*_{F_0}(g)(y_0)$, we can study the path derivative of any Hadamard differentiable functional of F_1 and F_0, e.g., Lorenz curves and Gini coefficients. Since such kind of calculation is standard, the details are omitted.

Example 6 *Consider the setup in the running example with only the selection effect. Suppose $g(u|u_D) = 6u(1-u)(2u_D - 1)$. It implies that the ranks of individuals who want most to participate move up, and visa versa. Since only $F_d(y_d)$ is involved in $D^*_{F_d}(g)(y_d)$, we plot $D^*_{F_d}(g)$ as a function of $\tau_d \equiv F_d(y_d)$ instead of y_d and denote it as $D^*_{F_d}(g)(\tau_d)$. To provide more intuition on the form of $D^*_{F_d}(g)(\tau_d)$ and $D^*_{\Delta}(g)(\tau)$, we report their formulas for this example below.*

$$\begin{aligned} D^*_{F_1}(g)(\tau_1) &= \frac{\int_0^1 f_{Y_0|U_D}\left(F_0^{-1}(\tau_1)|u_D\right)F_{p(Z)}(u_D)du_D}{\int_0^1 f_{Y_0|U_D}(F_0^{-1}(\tau_1)|u_D)h(u_D)\, du_D}\int_0^1 g(\tau_1|u_D)h(u_D)\, du_D - \int_0^1 g(\tau_1|u_D)F_{p(Z)}(u_D)du_D, \\ D^*_{F_0}(g)(\tau_0) &= -\frac{\int_0^1 f_{Y_1|U_D}\left(F_1^{-1}(\tau_0)|u_D\right)(1-F_{p(Z)}(u_D))du_D}{\int_0^1 f_{Y_1|U_D}(F_1^{-1}(\tau_0)|u_D)h(u_D)\, du_D}\int_0^1 g(\tau_0|u_D)h(u_D)\, du_D \\ &\quad + \int_0^1 g(\tau_0|u_D)(1-F_{p(Z)}(u_D))du_D, \end{aligned}$$

and

$$D^*_{\Delta}(g)(\tau) = \frac{D^*_{F_0}(g)(\tau)}{f_0\left(F_0^{-1}(\tau)\right)} - \frac{D^*_{F_1}(g)(\tau)}{f_1\left(F_1^{-1}(\tau)\right)}$$

where

$$\begin{aligned} f_{Y_1|U_D}(y_1|u_D) &= \frac{1}{2}\phi\left(\frac{y_1/2 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right), \\ f_{Y_0|U_D}(y_0|u_D) &= \phi\left(\frac{y_0 - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right), \\ F_0^{-1}(\tau) &= \Phi^{-1}(\tau), F_1^{-1}(\tau) = 2\Phi^{-1}(\tau), \\ f_0(y_0) &= \phi(y_0), f_1(y_1) = \frac{1}{2}\phi\left(\frac{y_1}{2}\right), \\ h(u_D) &= 6u_D(1-u_D), F_{p(Z)}(u_D) = u_D. \end{aligned}$$

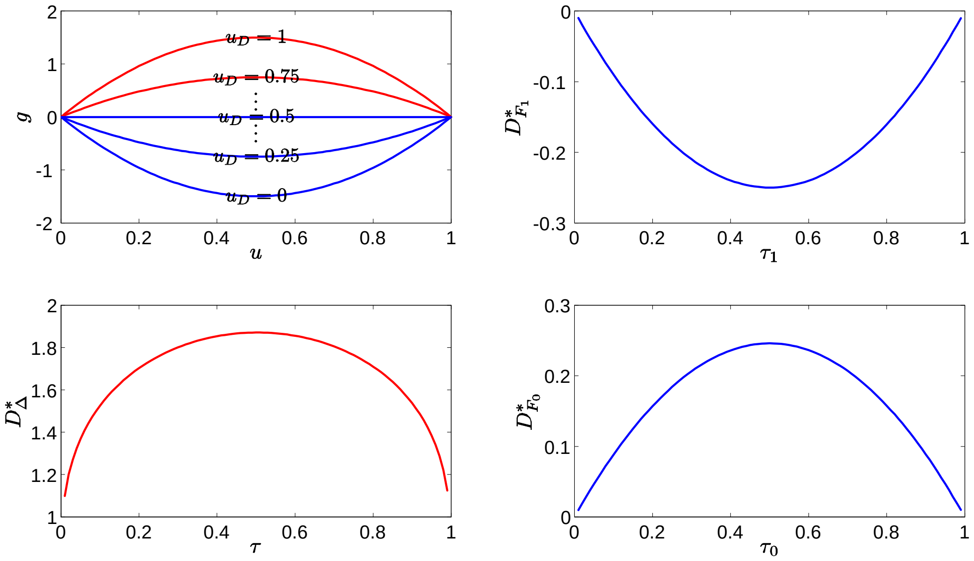

Figure 5: $g(u|u_D)$, $D^*_{F_1}(g)(\tau_1)$, $D^*_{F_0}(g)(\tau_0)$ and $D^*_{\Delta}(g)(\tau)$

Figure 5 shows $D^*_{F_d}(g)(\tau_d)$ and $D^*_{\Delta}(g)(\tau)$. For this local misspecification, $F_1^*(\cdot)$ tends to underestimate $F_1(\cdot)$ (especially in the middle quantiles), while $F_0^*(\cdot)$ tends to overestimate $F_0(\cdot)$ (especially in the middle quantiles). In other words, F_1^* first-order stochastically dominates F_1, while F_0^* is first-order stochastically dominated by F_0. This results in an overestimation of $\Delta(\cdot)$ (especially in the middle quantiles).

4.4 Using General Instruments

Now, consider a general instrument $J(Z)$. As in Section 3.3, we add a subscript J to distinguish from the case with $p(Z)$ as the instrument.

Proposition 8 *Under Assumptions P and Q, $F^*_{J1}(y_1)$ and $F^*_{J0}(y_0)$ take the same form as in Theorem 2 except replacing $\widetilde{F}_1(y_1)$ and $\widetilde{F}_0(y_0)$ by*

$$\begin{aligned}\widetilde{F}_{J1}(y_1) &= \int F_{Y_1|U_D}(y_1|u_D)h_J(u_D)\,du_D,\\ \widetilde{F}_{J0}(y_0) &= \int F_{Y_0|U_D}(y_1|u_D)h_J(u_D)\,du_D,\end{aligned}$$

respectively, where $h_J(u_D)$ is defined in Proposition 4.

Similar to the IVE case, the estimands depend on the instruments employed. Moreover, a general instrument $J(Z)$ may make the intuitive interpretation of $F^*_d(y_d)$ break down. This is because $h_J(u_D)$ may not be nonnegative for all u_D when a general instrument $J(Z)$ is used (although $\int h_J(u_D)\,du_D = 1$ and $h_J(0) = h_J(1) = 0$ still hold) such that $\widetilde{F}_{Jd}(\cdot)$ may not be a genuine cdf anymore, i.e., $\widetilde{F}^{-1}_{Jd}(\cdot)$ in $F^*_{Jd}(y_1)$ may not be well defined. As a result, the monotonicity assumption in Assumption 7 of Wütherich (2015), where Z may be a continuous scalar and $J(Z) = Z$, need not hold unless $p(Z)$ is monotonically increasing in Z. In the discrete Z case, we can always reorder $p(Z)$ to make it monotonically increasing. This is why $\widetilde{F}_1(\cdot)$ and $\widetilde{F}_0(\cdot)$ are genuine cdfs in his Theorem 6.

As in Section 3.3, we discuss the discrete instrument case here. Besides considering the case with discrete Z and using $p(Z)$ as the instrument, we also consider the case in Wütherich (2015) where Z is a discrete scalar and $J(Z) = Z$. We first discuss $\widetilde{F}_d(\cdot)$. Assume $p(Z)$ takes K values as in Section 3.3; then

$$\widetilde{F}_d\left(y_d\right) = \sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right) h(p_{k+1}) \int_{p_k}^{p_{k+1}} F_{Y_d|U_D}(y_d|u_D)\frac{1}{p_{k+1}-p_k} du_D = \sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right) h(p_{k+1}) F_d^k(y_d),$$

where $h(p_{k+1})$ is defined in (6), and

$$\begin{aligned} F_d^k(y_d) &= \frac{1}{p_{k+1}-p_k}\int_{p_k}^{p_{k+1}} F_{Y_d|U_D}(y_d|u_D)du_D \\ &= \frac{E\left[1\left(Y \leq y_d\right)\cdot 1\left(D=d\right)|p(Z)=p_{k+1}\right] - E\left[1\left(Y \leq y_d\right)\cdot 1\left(D=d\right)|p(Z)=p_k\right]}{P\left(D=d|p(Z)=p_{k+1}\right) - P\left(D=d|p(Z)=p_k\right)}, \end{aligned}$$

$k = 1, \cdots, K-1$, is the cdf of Y_d for compliers $\mathcal{C}_k$ and is identifiable. If Z takes K values, $\{z_1, \cdots, z_K\}$, with $-\infty < z_1 < \cdots < z_K < \infty$ such that $p(Z)$ is a strictly monotone function of Z,[18] then as $J(Z) = Z$,

$$\widetilde{F}_{Jd}\left(y_d\right) = \sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right) h_J(z_{k+1}) F_d^k(y_1),$$

where

$$h_J(z_{k+1}) = \frac{Cov\left(Z, 1(Z \geq z_{k+1})\right)}{\sum_{k=1}^{K-1}\left(p_{k+1}-p_k\right) Cov\left(Z, 1(Z \geq z_{k+1})\right)}$$

is equivalent to the weight in Theorem 6 of Wütherich (2015). Given $\widetilde{F}_d\left(y_d\right)$, we can see

$$F_1^*(y_1) = \sum_{k=0}^{K}\left(p_{k+1}-p_k\right)\left\{F_1^k(y_1)\left[1-F_{p(Z)}(p_k)\right] + F_0^k(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right))F_{p(Z)}(p_k)\right\},$$

and

$$F_0^*(y_0) = \sum_{k=0}^{K}\left(p_{k+1}-p_k\right)\left\{F_1^k(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right))\left[1-F_{p(Z)}(p_k)\right] + F_0^k(y_0)F_{p(Z)}(p_k)\right\},$$

which are equivalent to the formulas in Theorem 6 of Wütherich (2015) when $\widetilde{F}_d\left(y_d\right)$ is replaced by $\widetilde{F}_{Jd}\left(y_d\right)$, where $F_1^0(y_1) = E\left[1\left(Y \leq y_1\right)D|p(Z)=p_1\right]/p_1$ is the cdf of Y_1 for always-takers, $F_0^K(y_0) = E\left[1\left(Y \leq y_0\right)\left(1-D\right)|p(Z)=p_K\right]/\left(1-p_K\right)$ is the cdf of Y_0 for never-takers, and $F_d^k(\cdot)$, $k = 1, \cdots, K-1$, is defined above. When $K = 2$, these formulas can be much simplified. In this case,

$$\widetilde{F}_d\left(y_d\right) = F_{Y_d|\mathcal{C}}(y_d),$$

so

$$\begin{aligned} F_1^*(y_1) &= p_1 F_{Y_1|\mathcal{A}}(y_1) + (p_2-p_1)\left[F_{Y_1|\mathcal{C}}(y_1)(1-q_1) + F_{Y_0|\mathcal{C}}(F_{Y_0|\mathcal{C}}^{-1}F_{Y_1|\mathcal{C}}\left(y_1\right))q_1\right] + (1-p_2)F_{Y_0|\mathcal{N}}(F_{Y_0|\mathcal{C}}^{-1}F_{Y_1|\mathcal{C}}\left(y_1\right)) \\ &= p_1 F_{Y_1|\mathcal{A}}(y_1) + (p_2-p_1)F_{Y_1|\mathcal{C}}(y_1) + (1-p_2)F_{Y_0|\mathcal{N}}(F_{Y_0|\mathcal{C}}^{-1}F_{Y_1|\mathcal{C}}\left(y_1\right)), \\ F_0^*(y_0) &= \left(1-p_2\right)F_{Y_0|\mathcal{N}}(y_0) + (p_2-p_1)\left[F_{Y_1|\mathcal{C}}(F_{Y_1|\mathcal{C}}^{-1}F_{Y_0|\mathcal{C}}\left(y_0\right))(1-q_1) + F_{Y_0|\mathcal{C}}(y_0)q_1\right] + p_1 F_{Y_1|\mathcal{A}}(F_{Y_1|\mathcal{C}}^{-1}F_{Y_0|\mathcal{C}}\left(y_0\right)) \\ &= \left(1-p_2\right)F_{Y_0|\mathcal{N}}(y_0) + (p_2-p_1)F_{Y_0|\mathcal{C}}(y_0) + p_1 F_{Y_1|\mathcal{A}}(F_{Y_1|\mathcal{C}}^{-1}F_{Y_0|\mathcal{C}}\left(y_0\right)), \end{aligned}$$

[18] Note that different Z values may generate the same $p(Z)$ value, in which case, we combine the corresponding Z values.

as claimed in Theorem 1 of Wütherich (2015), where $q_1 = P(Z = z_1)$, $F_{Y_1|\mathcal{A}}(\cdot), F_{Y_1|\mathcal{C}}(\cdot), F_{Y_0|\mathcal{C}}(\cdot)$ and $F_{Y_0|\mathcal{N}}(\cdot)$ are the cdfs for always-takers, compliers and never-takers in the two counterfactual statuses and can be identified as in Abadie (2002) and Imbens and Rubin (1997).

Finally, when a general instrument $J(Z)$ is used, $D^*_{F_d}(g)(\cdot)$ and $D^*_{\Delta}(g)(\cdot)$ in Section 4.3 take similar forms except replacing $h(u_D)$ by $h_J(u_D)$ at all appearances of $h(u_D)$, so similar analysis as above can be applied. For example, when $p(Z)$ is discrete,

$$\begin{aligned} D^*_{F_1}(g)(\tau_1) &= \frac{\sum_{k=1}^{K-1}(p_{k+1}-p_k)\,F_{p(Z)}(p_k)f_0^k\left(F_0^{-1}(\tau_1)\right)+(1-p_K)f_0^K\left(F_0^{-1}(\tau_1)\right)}{\sum_{k=1}^{K-1}(p_{k+1}-p_k)\,h(p_{k+1})f_0^k(F_0^{-1}(\tau_1))} \\ &\quad\cdot\sum_{k=1}^{K-1}(p_{k+1}-p_k)\,h(p_{k+1})g_k(\tau_1) \\ &\quad-\left[\sum_{k=1}^{K-1}(p_{k+1}-p_k)\,F_{p(Z)}(p_k)g_k(\tau_1)+(1-p_K)g_K(\tau_1)\right], \\ D^*_{F_0}(g)(\tau_0) &= -\frac{p_1f_1^0\left(F_1^{-1}F_0(y_0)\right)+\sum_{k=1}^{K-1}(p_{k+1}-p_k)\left(1-F_{p(Z)}(p_k)\right)f_1^k\left(F_1^{-1}(\tau_0)\right)}{\sum_{k=1}^{K-1}(p_{k+1}-p_k)\,h(p_{k+1})f_1^k\left(F_1^{-1}(\tau_0)\right)} \\ &\quad\cdot\sum_{k=1}^{K-1}(p_{k+1}-p_k)\,h(p_{k+1})g_k(\tau_0) \\ &\quad+\left[p_1g_0(\tau_0)+\sum_{k=1}^{K-1}(p_{k+1}-p_k)\left(1-F_{p(Z)}(p_k)\right)g_k(\tau_0)\right], \end{aligned}$$

and $D^*_{\Delta}(g)(\tau) = \frac{D^*_{F_0}(g)(\tau)}{f_0\left(F_0^{-1}(\tau)\right)} - \frac{D^*_{F_1}(g)(\tau)}{f_1\left(F_1^{-1}(\tau)\right)}$, where $\tau_d = F_d(y_d)$ as in Example 6,

$$f_d^k(y_d) = \frac{1}{p_{k+1}-p_k}\int_{p_k}^{p_{k+1}} f_{Y_d|U_D}(y_d|u_D)\,du_D, k = 1, \cdots, K-1,$$

can be estimated by the derivative of $F_d^k(y_d)$,

$$f_0^K(y_0) = \frac{1}{1-p_K}\int_{p_K}^{1} f_{Y_0|U_D}(y_0|u_D)\,du_D, f_1^0(y_1) = \frac{1}{p_1}\int_0^{p_1} f_{Y_1|U_D}(y_1|u_D)\,du_D$$

can be estimated by the derivative of

$$\begin{aligned} F_0^K(y_0) &= \frac{1}{1-p_K}\int_{p_K}^{1} F_{Y_0|U_D}(y_0|u_D)\,du_D = \frac{E\left[1(Y\le y_0)(1-D)\,|p(Z)=p_K\right]}{1-p_K} \\ &= P\left(Y\le y_0|p(Z)=p_K, D=0\right), \\ F_1^0(y_1) &= \frac{1}{p_1}\int_0^{p_1} F_{Y_1|U_D}(y_1|u_D)\,du_D = \frac{E\left[1(Y\le y_1)\,D|p(Z)=p_1\right]}{p_1} \\ &= P\left(Y\le y_1|p(Z)=p_1, D=1\right), \end{aligned}$$

respectively, and

$$g_k(u) = \frac{1}{p_{k+1}-p_k}\int_{p_k}^{p_{k+1}} g(u|u_D)du_D, k = 0, 1, \cdots, K. \tag{14}$$

4.5 Partial Identification Analysis

We now develop sharp bounds for $Q_d(\tau)$ and $\Delta(\tau)$. Our analyses are parallel to those in Heckman and Vytlacil (2001b).

Theorem 3 *Suppose Assumptions P and Q with (Q-4) replaced by (Q-4′) hold.*

(i) *$Q_1(\tau)$ and $Q_0(\tau)$, $\tau \in (0,1)$, have sharp bounds*

$$\begin{aligned} \underline{I}_1(\tau) &\leq Q_1(\tau) \leq \overline{I}_1(\tau), \\ \underline{I}_0(\tau) &\leq Q_0(\tau) \leq \overline{I}_0(\tau), \end{aligned}$$

where

$$\begin{aligned} \underline{I}_1(\tau) &= \begin{cases} Q_{Y|p(Z),D}\left(1-\frac{1-\tau}{\overline{p}}\middle|\,\overline{p},1\right), & \text{if } \overline{p} > 1-\tau, \\ \underline{y}_1, & \text{otherwise,} \end{cases}, \\ \overline{I}_1(\tau) &= \begin{cases} Q_{Y|p(Z),D}\left(\frac{\tau}{\overline{p}}\middle|\,\overline{p},1\right), & \text{if } \overline{p} \geq \tau, \\ \overline{y}_1, & \text{otherwise,} \end{cases} \end{aligned}$$

and

$$\begin{aligned} \underline{I}_0(\tau) &= \begin{cases} Q_{Y|p(Z),D}\left(1-\frac{1-\tau}{1-\underline{p}}\middle|\,\underline{p},0\right), & \text{if } \underline{p} < \tau, \\ \underline{y}_0, & \text{otherwise,} \end{cases} \\ \overline{I}_0(\tau) &= \begin{cases} Q_{Y|p(Z),D}\left(\frac{\tau}{1-\underline{p}}\middle|\,\underline{p},0\right), & \text{if } \underline{p} \leq 1-\tau, \\ \overline{y}_0, & \text{otherwise.} \end{cases} \end{aligned}$$

This implies $\Delta(\tau)$ has sharp bounds,

$$\underline{I}_\Delta(\tau) \equiv \underline{I}_1(\tau) - \overline{I}_0(\tau) \leq \Delta(\tau) \leq \overline{I}_1(\tau) - \underline{I}_0(\tau) \equiv \overline{I}_\Delta(\tau).$$

(ii) *$\overline{p} = 1$ ($\underline{p} = 0$) is sufficient for point identification of $Q_1(\tau)$ ($Q_0(\tau)$), and $\overline{p} = 1$ and $\underline{p} = 0$ are sufficient for point identification of $\Delta(\tau)$ for any fixed $\tau \in (0,1)$. When assumption (Q-4′) is replaced by (Q-4), these sufficient conditions are also necessary.*

Example 7 *We use a specific example to provide some intuition for why $\underline{I}_1(\tau) \leq Q_1(\tau) \leq \overline{I}_1(\tau)$;*[19] *similar intuition can be applied to the bounds for $Q_0(\tau)$. From the proof of the theorem,*

$$P\left(Y \leq y|p(Z) = \overline{p}, D = 1\right)\overline{p} \leq P(Y_1 \leq y) \leq P\left(Y \leq y|p(Z) = \overline{p}, D = 1\right)\overline{p} + (1-\overline{p}).$$

Suppose $(Y_1, V) \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$; then

$$P\left(Y \leq y|p(Z) = \overline{p}, D = 1\right)\overline{p} = \int_0^{\overline{p}} \Phi\left(\frac{y - \rho\Phi^{-1}(u_D)}{\sqrt{1-\rho^2}}\right) du_D.$$

Figure 6 shows the bounds for $P(Y_1 \leq y)$ when $\overline{p} = 0.8$ and $\rho = 0.5$. Inverting the bounds for $P(Y_1 \leq y)$, we can get the bounds for $Q_1(\tau)$. When $\tau \leq 1-\overline{p}$, $\underline{I}_1(\tau) = \underline{y}_1$; when $\tau > \overline{p}$, $\overline{I}_1(\tau) = \overline{y}_1$. Only if $\tau \in (1-\overline{p}, \overline{p})$, both bounds are nontrivial. This is not always possible; only if $\overline{p} > \max(\tau, 1-\tau) \geq 1/2$ ($\underline{p} < \min(\tau, 1-\tau)$), neither the left nor the right bound for $Q_1(\tau)$ ($Q_0(\tau)$) is trivial. Pushing $\tau \to 0$ or 1, we can see that there are nontrivial bounds for $Q_1(\tau)$ ($Q_0(\tau)$) for all τ if and only if $\overline{p} = 1$ ($\underline{p} = 0$), but from Theorem 3(ii), these nontrivial bounds coincide. In the figure, $\overline{I}_1(\tau_2) = 0$, so $\left[\underline{I}_1(\tau), \overline{I}_1(\tau)\right]$ need not cover 0 for a given τ. It is

[19] The setup of this example is different from that of the running example because $\underline{p} = 0$ and $\overline{p} = 1$ there such that point identification can be achieved.

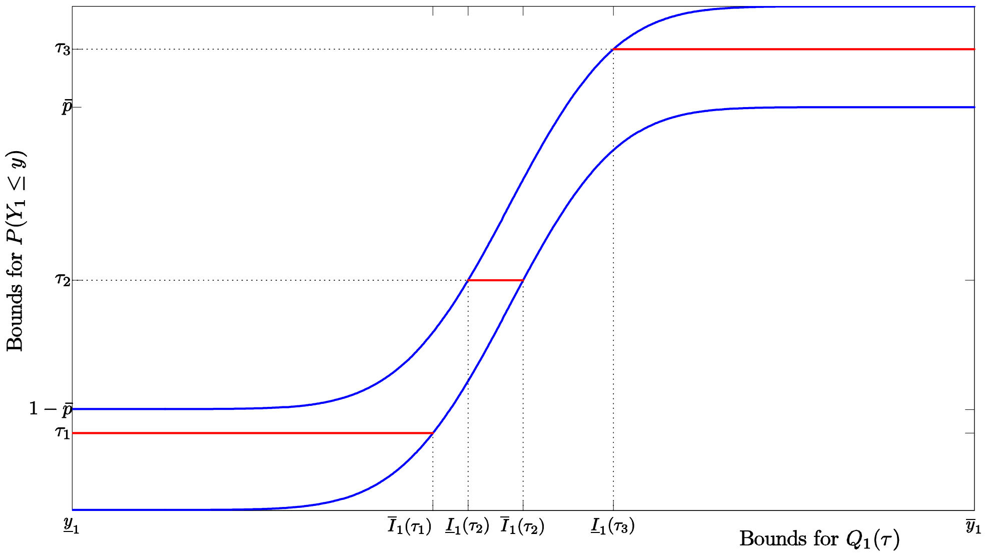

Figure 6: Intuition for $\underline{I}_1(\tau) \leq Q_1(\tau) \leq \overline{I}_1(\tau)$: $\overline{p} = 0.8$, $\rho = 0.5$, $\tau_1 = 0.15$, $\tau_2 = 0.46$ and $\tau_3 = 0.91$

also not hard to see that neither $\left[\underline{I}_0(\tau), \overline{I}_0(\tau)\right]$ nor $\left[\underline{I}_\Delta(\tau), \overline{I}_\Delta(\tau)\right]$ need cover 0 for a given τ.

We first provide some comments on the bounds in Theorem 3(i). Note that $\underline{I}_d(\tau)$ and $\overline{I}_d(\tau)$ are increasing functions of τ; hence the bound for $Q_d(\tau)$ shifts to the right as τ increases. Also observe that

$$1 - \frac{1-\tau}{\overline{p}} \leq \tau \leq \frac{\tau}{\overline{p}} \text{ and } 1 - \frac{1-\tau}{1-\underline{p}} \leq \tau \leq \frac{\tau}{\underline{p}};$$

hence $Q_{Y|p(Z),D}\left(\tau|\overline{p}, 1\right)$ and $Q_{Y|p(Z),D}\left(\tau|\underline{p}, 0\right)$ lie within the bound for $Q_1(\tau)$ and $Q_0(\tau)$, respectively. This implies that $F_1(\cdot) = F_{Y|p(Z),D}(\cdot|\overline{p}, 1)$ and $F_0(\cdot) = F_{Y|p(Z),D}(\cdot|\underline{p}, 0)$ are not rejectable in the absence of other information. Evaluating the bounds for $Q_1(\tau)$ requires knowledge of $\overline{p}$, $\underline{y}_1$, $\overline{y}_1$ and $Q_{Y|p(Z),D}\left(\cdot|\,\overline{p}, 1\right)$ at $1 - \frac{1-\tau}{\overline{p}}$ and $\frac{\tau}{\overline{p}}$; evaluating the bounds for $Q_0(\tau)$ requires knowledge of $\underline{p}$, $\underline{y}_0$, $\overline{y}_0$ and $Q_{Y|p(Z),D}\left(\cdot|\,\underline{p}, 0\right)$ at $1 - \frac{1-\tau}{1-\underline{p}}$ and $\frac{\tau}{1-\underline{p}}$; evaluating the bounds for $\Delta(\tau)$ requires knowledge of all of them; nevertheless, only four conditional quantiles are needed for all these evaluations. Estimators of these objects can be constructed in an obvious way, so are omitted here.

We next turn to the sufficient and necessary conditions for point identification in Theorem 3(ii). Comparing with the identification result in Proposition 5, we require not only $p(1) > p(0)$ but further that $p(1) = 1$ and $p(0) = 0$ for point identifying $Q_1(\tau)$ and $Q_0(\tau)$. This is of course due to the general assumption on the outcome equations as discussed in Section 4.1. Our results are parallel to those of Heckman and Vytlacil (2001b) but with subtle differences. Although $\underline{p} = 0$ and $\overline{p} = 1$ are sufficient for point identification of $\Delta(\tau)$, they are not necessary when Y_d is not continuously distributed. Here, note that since assumption (Q-4$'$) requires only $P\left(Y_d \in \mathcal{Y}_{xd}|X = x\right) = 1$ not $\mathrm{supp}(Y_d|X = x) = \mathcal{Y}_{xd}$, the bounds for $\Delta(\tau)$ in the theorem can be applied to cases with discrete, continuous or mixed response variables. In S.2.4, we provide an example to illustrate that $\underline{p} = 0$ and $\overline{p} = 1$ are not necessary for point identification of $\Delta(\tau)$ when Y_d is binary. Note further that we require only that $Y|\left(p(Z) = \overline{p}, D = 1\right)$ and $Y|\left(p(Z) = \underline{p}, D = 0\right)$ are continuously distributed with a positive density on $(\underline{y}_1, \overline{y}_1)$ and $(\underline{y}_0, \overline{y}_0)$ respectively to make $\underline{p} = 0$ and $\overline{p} = 1$ necessary for point identification of $\Delta(\tau)$, but Assumption (Q-4) implies this condition.

5 LSE

The LSE estimates the ATE by the coefficient of D in the linear regression of Y on $(1, D)$. The following theorem states the resulting pseudo-true values.

Theorem 4 *Under Assumptions A and P,*

$$\begin{aligned}\overline{\mu}_1 &= \int_0^{\underline{p}} \frac{E\left[Y_1|U_D = u_D\right]}{E\left[p(Z)\right]} du_D + \int_{\underline{p}}^{\overline{p}} E\left[Y_1|U_D = u_D\right] \frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} du_D, \\ \overline{\mu}_0 &= \int_{\underline{p}}^{\overline{p}} E\left[Y_0|U_D = u_D\right] \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]} du_D + \int_{\overline{p}}^{1} \frac{E\left[Y_0|U_D = u_D\right]}{E\left[1 - p(Z)\right]} du_D\end{aligned}$$

and

$$\overline{\Delta} = \overline{\mu}_1 - \overline{\mu}_0 = \int_0^1 MTE(u_D)\omega(u_D)du_D,$$

where

$$\omega(u_D) = \begin{cases} 1 + \frac{E[U_1|U_D=u_D]\omega_1(u_D) - E[U_0|U_D=u_D]\omega_0(u_D)}{MTE(u_D)} & \textit{if } MTE(u_D) \neq 0 \\ 0 \textit{ otherwise,} & \end{cases}$$

with

$$\omega_1(u_D) = \frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} \textit{ and } \omega_0(u_D) = \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]}.$$

Since $E\left[p(Z)\right] = \int_0^1 \left(1 - F_{p(Z)}(p)\right) dp = \underline{p} + \int_{\underline{p}}^{\overline{p}} \left(1 - F_{p(Z)}(p)\right) dp$, $\overline{\mu}_1$ is a weighted average of $E\left[Y_1|U_D = u_D\right]$ among $u_D \in [0, \overline{p}]$, where the weight on $[0, \underline{p}]$ is a constant $\frac{1}{E[p(Z)]}$ and the weight on $[\underline{p}, \overline{p}]$ is $\frac{1-F_{p(Z)}(u_D)}{E[p(Z)]}$. Similarly, since $1 - E\left[p(Z)\right] = 1 - \overline{p} + \int_{\underline{p}}^{\overline{p}} F_{p(Z)}(u_D)$, $\overline{\mu}_0$ is a weighted average of $E\left[Y_0|U_D = u_D\right]$ among $u_D \in \left[\underline{p}, 1\right]$, where the weight on $[\overline{p}, 1]$ is a constant $\frac{1}{1-E[p(Z)]}$ and the weight on $[\underline{p}, \overline{p}]$ is $\frac{F_{p(Z)}(u_D)}{1-E[p(Z)]}$. Actually, since $F_{p(Z)}(u_D) = 0$ for $u_D \in [0, \underline{p}]$ and $F_{p(Z)}(u_D) = 1$ for $u_D \in [\overline{p}, 1]$, $\overline{\mu}_d$ can be re-written as

$$\overline{\mu}_1 = \int_0^1 E\left[Y_1|U_D = u_D\right] \frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} du_D \text{ and } \overline{\mu}_0 = \int_0^1 E\left[Y_0|U_D = u_D\right] \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]} du_D.$$

When $E\left[Y_1|U_D = u_D\right] = \mu_1$, $\overline{\mu}_1 = \mu_1$, and when $E\left[Y_0|U_D = u_D\right] = \mu_0$, $\overline{\mu}_0 = \mu_0$. Different from μ_1^*, $\overline{\mu}_1$ does not extrapolate $E\left[Y_1|U_D = u_D\right]$ to $(\overline{p}, 1]$, rather, it averages $E\left[Y_1|U_D = u_D\right]$ over u_D's that are guaranteed to be treated given the data. Similarly, $\overline{\mu}_0$ does not extrapolate $E\left[Y_0|U_D = u_D\right]$ to $[0, \underline{p})$ but averages $E\left[Y_0|U_D = u_D\right]$ over u_D's that are guaranteed to be untreated given the data. Actually, it is not hard to see that $\overline{\mu}_1 = E\left[\frac{D}{E[D]}Y\right] = E\left[\frac{D}{E[D]}Y_1\right]$ uses only treated data and $\overline{\mu}_0 = E\left[\frac{1-D}{1-E[D]}Y\right] = E\left[\frac{1-D}{1-E[D]}Y_0\right]$ uses only untreated data. As mentioned before, $\displaystyle\int_0^{\underline{p}} E\left[Y_1|U_D = u_D\right] du_D = \underline{p}E\left[Y|D = 1, p(Z) = \underline{p}\right]$ and $\int_{\overline{p}}^1 E\left[Y_0|U_D = u_D\right] du_D = (1 - \overline{p})E\left[Y|D = 1, p(Z) = \overline{p}\right]$ are identifiable, so $\overline{\mu}_1$ and $\overline{\mu}_0$ are indeed identifiable.

From the formula of $\overline{\mu}_1$ and $\overline{\mu}_0$, we can see

$$\begin{aligned}
\overline{\Delta} - \Delta &= \int_0^{\underline{p}} \frac{E\left[U_1 - U_0 | U_D = u_D\right]}{E\left[p(Z)\right]} du_D + \int_{\underline{p}}^{\overline{p}} E\left[U_1 - U_0 | U_D = u_D\right] \frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} du_D \text{ (I)} \\
&+ \int_0^{\underline{p}} \frac{E\left[U_0 | U_D = u_D\right]}{E\left[p(Z)\right]} du_D + \int_{\underline{p}}^{\overline{p}} E\left[U_0 | U_D = u_D\right] \left(\frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} - \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]} \right) du_D \\
&- \int_{\overline{p}}^1 \frac{E\left[U_0 | U_D = u_D\right]}{E\left[1 - p(Z)\right]} du_D \text{ (II)}.
\end{aligned}$$

As a result, when $E\left[U_1 - U_0 | U_D = u_D\right] = 0$ over $u_D \in [0, \overline{p}]$, the (I) part of bias will be gone. When $E\left[U_0 | U_D = u_D\right] = 0$ over $u_D \in [0, 1]$, the (II) part of bias will be gone. The (I) part of bias comes solely from $\overline{\mu}_1$, while the (II) part of bias comes from both $\overline{\mu}_1$ and $\overline{\mu}_0$. To understand such a result, note that $Y = Y_0 + (Y_1 - Y_0)D = \mu_0 + \Delta \cdot D + U_0 + (U_1 - U_0)D$. To consistently estimate Δ, we need $E\left[U_1 - U_0 | D = 1\right] = 0$ and $E\left[U_0 | D\right] = 0$.[20] Violation of the former generates the (I) part of bias and violation of the latter generates the (II) part of bias. The IVE eliminates the (II) part instead of the first part of bias, so it must extend the estimation of μ_1 to $(\overline{p}, 1]$ and the estimation of μ_0 to $[0, \underline{p})$ to cure this bias. Note further that the weight $\omega(u_D)$ is the same as in HV. $\omega(u_D)$ need not be positive, and $\int_0^1 \omega(u_D) du_D$ need not be one, so even if $E\left[U_1 - U_0 | U_D = u_D\right] = 0$, $\overline{\Delta} = \Delta \int_0^1 \omega(u_D) du_D$ need not be Δ.

To provide more intuition on the discussion above, we consider the discrete instrument case here. We will use the same setup and notations as in Section 3.3. Now,

$$\begin{aligned}
\overline{\mu}_1 &= \sum_{k=0}^{K-1} \frac{1 - F_{p(Z)}(p_k)}{E\left[p(Z)\right]} \left(p_{k+1} - p_k\right) \int_{p_k}^{p_{k+1}} \frac{E\left[Y_1 | U_D = u_D\right]}{p_{k+1} - p_k} du_D = \sum_{k=0}^{K-1} \frac{1 - F_{p(Z)}(p_k)}{E\left[p(Z)\right]} \left(p_{k+1} - p_k\right) \mu_1^k, \\
\overline{\mu}_0 &= \sum_{k=1}^{K} \frac{F_{p(Z)}(p_k)}{E\left[1 - p(Z)\right]} \left(p_{k+1} - p_k\right) \int_{p_k}^{p_{k+1}} \frac{E\left[Y_0 | U_D = u_D\right]}{p_{k+1} - p_k} du_D = \sum_{k=1}^{K} \frac{F_{p(Z)}(p_k)}{E\left[1 - p(Z)\right]} \left(p_{k+1} - p_k\right) \mu_0^k, \\
\overline{\Delta} &= \frac{p_1}{E\left[p(Z)\right]} \mu_1^0 + \sum_{k=1}^{K-1} \left(p_{k+1} - p_k\right) \left[\frac{1 - F_{p(Z)}(p_k)}{E\left[p(Z)\right]} \mu_1^k - \frac{F_{p(Z)}(p_k)}{E\left[1 - p(Z)\right]} \mu_0^k \right] - \frac{1 - p_K}{E\left[1 - p(Z)\right]} \mu_0^K.
\end{aligned}$$

When $K = 2$,

$$\begin{aligned}
\overline{\mu}_1 &= \frac{p_1}{E\left[p(Z)\right]} \mu_{1|\mathcal{A}} + \frac{1 - F_{p(Z)}(p_1)}{E\left[p(Z)\right]} \left(p_2 - p_1\right) \mu_{1|\mathcal{C}}, \\
\overline{\mu}_0 &= \frac{F_{p(Z)}(p_1)}{E\left[1 - p(Z)\right]} \left(p_2 - p_1\right) \mu_{0|\mathcal{C}} + \frac{1 - p_2}{E\left[1 - p(Z)\right]} \mu_{0|\mathcal{N}}, \\
\overline{\Delta} &= \frac{p_1}{E\left[p(Z)\right]} \mu_{1|\mathcal{A}} + \left(p_2 - p_1\right) \left[\frac{1 - F_{p(Z)}(p_1)}{E\left[p(Z)\right]} \mu_{1|\mathcal{C}} - \frac{F_{p(Z)}(p_1)}{E\left[1 - p(Z)\right]} \mu_{0|\mathcal{C}} \right] - \frac{1 - p_2}{E\left[1 - p(Z)\right]} \mu_{0|\mathcal{N}},
\end{aligned}$$

where $\overline{\mu}_1$ is a weighted average of $\mu_{1|\mathcal{A}}$ and $\mu_{1|\mathcal{C}}$, and $\overline{\mu}_0$ is a weighted average of $\mu_{0|\mathcal{C}}$ and $\mu_{0|\mathcal{N}}$. Indeed, there is no extrapolation in $\overline{\mu}_d$, and $\overline{\Delta}$ is different from the LATE. When $\mu_{1|\mathcal{A}} = \mu_{1|\mathcal{C}}$, $\overline{\mu}_1 = \mu_1$; when $\mu_{0|\mathcal{C}} = \mu_{0|\mathcal{N}}$, $\overline{\mu}_0 = \mu_0$; when $\mu_{1|\mathcal{A}} = \mu_{1|\mathcal{C}}$ and $\mu_{0|\mathcal{C}} = \mu_{0|\mathcal{N}}$, $\overline{\Delta} = \Delta$.

The bounds in Section 3.4 can still be used to assess the validity of LSE. We next conduct the local sensitivity analysis on $\overline{\mu}_d$ and $\overline{\Delta}$. Assume $(E[U_1 - U_0 | U_D = u_D], E\left[U_0 | U_D = u_D\right])$ stays in a parametrized

[20] Strictly speaking, we only require $E\left[U_0 + (U_1 - U_0)D | D = 1\right] = E\left[U_0 + (U_1 - U_0)D | D = 0\right]$, i.e., $E[U_1 | D = 1] = E\left[U_0 | D = 0\right]$. However, $E[U_1 | D = 1] - E\left[U_0 | D = 0\right] = \left(E\left[U_1 | D = 1\right] - E\left[U_0 | D = 1\right]\right) + \left(E\left[U_0 | D = 1\right] - E\left[U_0 | D = 0\right]\right)$, so $E\left[U_1 - U_0 | D = 1\right] = 0$ and $E\left[U_0 | D\right] = 0$ are sufficient for consistency of the LSE.

space

$$\begin{aligned}\overline{\mathcal{G}}_\alpha &= \Big\{ \left(G_\alpha(\cdot), G^0_\alpha(\cdot)\right) : \alpha \in M, 0 \in M, \left(G_0(u_D), G^0_0(u_D)\right) = 0, u_D \in [0,1], \\ &\qquad \int_0^1 G_\alpha(u_D)du_D = 0, \text{ and } \int_0^1 G^0_\alpha(u_D)du_D = 0 \Big\}.\end{aligned}$$

Our targets are

$$\overline{D}_d\left(g, g^0\right) \equiv \lim_{\alpha\to 0} \frac{\overline{\mu}_d(\alpha) - \mu_d}{\alpha} \text{ and } \overline{D}_\Delta\left(g, g^0\right) = \lim_{\alpha\to 0} \frac{\overline{\Delta}(\alpha) - \Delta}{\alpha},$$

where we use $\overline{\mu}_d(\alpha)$ and $\overline{\Delta}(\alpha)$ to indicate the dependence of $\overline{\mu}_d$ and $\overline{\Delta}$ on α with $\overline{\mu}_d(0) = \mu_d$ and $\overline{\Delta}(0) = \Delta$, and assume $\lim_{\alpha\to 0} \frac{G_\alpha(\cdot)}{\alpha} \to g(\cdot)$ and $\lim_{\alpha\to 0} \frac{G^0_\alpha(\cdot)}{\alpha} \to g^0(\cdot)$. We impose the following conditions to guarantee the limits in $\overline{D}_d$ and $\overline{D}_\Delta$ exist.

Assumption LA$'$: $\sup_{u_D\in[0,1],\alpha\in\mathcal{N}} \left|\frac{G_\alpha(u_D)}{\alpha}\right| < \infty$, and $\sup_{u_D\in[0,1],\alpha\in\mathcal{N}} \left|\frac{G^0_\alpha(u_D)}{\alpha}\right| < \infty$ where $\mathcal{N}$ is a neighborhood of 0.

Proposition 9 *Under Assumptions A, P and LA$'$,*

$$\begin{aligned}\overline{D}_1\left(g, g^0\right) &= \int_0^{\underline{p}} \frac{g(u_D) + g^0(u_D)}{E\left[p(Z)\right]} du_D + \int_{\underline{p}}^{\overline{p}} \frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} \left[g(u_D) + g^0(u_D)\right] du_D \\ &= \int_0^1 \frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} \left[g(u_D) + g^0(u_D)\right] du_D, \\ \overline{D}_0\left(g, g^0\right) &= \int_{\underline{p}}^{\overline{p}} g^0(u_D) \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]} du_D + \int_{\overline{p}}^1 \frac{g^0(u_D)}{E\left[1 - p(Z)\right]} du_D \\ &= \int_0^1 g^0(u_D) \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]} du_D,\end{aligned}$$

and

$$\begin{aligned}\overline{D}_\Delta\left(g, g^0\right) &= \int_0^{\underline{p}} \frac{g(u_D)}{E\left[p(Z)\right]} du_D + \int_{\underline{p}}^{\overline{p}} \frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} g(u_D) du_D \\ &\quad + \int_0^{\underline{p}} \frac{g^0(u_D)}{E\left[p(Z)\right]} du_D + \int_{\underline{p}}^{\overline{p}} \left(\frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} - \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]}\right) g^0(u_D) du_D - \int_{\overline{p}}^1 \frac{g^0(u_D)}{E\left[1 - p(Z)\right]} du_D \\ &= \int_0^1 \frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} g(u_D) du_D + \int_0^1 \left(\frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} - \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]}\right) g^0(u_D) du_D \\ &= \overline{D}_1\left(g, g^0\right) - \overline{D}_0\left(g, g^0\right).\end{aligned}$$

The proof is similar to that of Proposition 3, so omitted. Note that $\overline{D}_0\left(g, g^0\right)$ only involves g^0, while $\overline{D}_1\left(g, g^0\right)$ and $\overline{D}_\Delta\left(g, g^0\right)$ involve both g and g^0. In other words, only the misspecification of $E\left[U_0|U_D = u_D\right]$ will affect the consistency of $\overline{\mu}_0$, while the misspecification of both $E\left[U_0|U_D = u_D\right]$ and $E\left[U_1 - U_0|U_D = u_D\right]$ affect the consistency of $\overline{\mu}_1$. From this proposition, $\overline{D}_d\left(g, g^0\right)$ and $\overline{D}_\Delta\left(g, g^0\right)$ are continuous linear functionals on $\mathcal{C}_0\left([0,1]\right)^2$, where $\mathcal{C}_0\left([0,1]\right)$ is defined in Section 3.2.

The following example numerically illustrates these path derivatives.

Example 8 *Consider the setup in the running example with both the selection effect and the essential heterogeneity. Suppose $g(u_D) = 1 - 2u_D$ as in Example 2 and $g^0(u_D) = u_D - 0.5$, i.e., the economic status of* *the individual with higher propensity to participate is relatively low when the treatment is absent. Then*

$$\begin{aligned}
\overline{D}_1\left(g,g^0\right) &= \int_0^1 2\left(1-u_D\right)\left(0.5-u_D\right)du_D = \frac{1}{6}, \\
\overline{D}_0\left(g,g^0\right) &= \int_0^1 2u_D(u_D-0.5)du_D = \frac{1}{6}, \\
\overline{D}_\Delta\left(g,g^0\right) &= \overline{D}_1\left(g,g^0\right) - \overline{D}_0\left(g,g^0\right) = 0.
\end{aligned}$$

In this local misspecification, both $\overline{\mu}_1$ and $\overline{\mu}_0$ overestimate their corresponding true value, but the biases offset each other such that Δ can be consistently estimated. Of course, it is easy to construct examples where Δ cannot be consistently estimated.

Finally, when $p(Z)$ is discrete,

$$\begin{aligned}
\overline{D}_1\left(g,g^0\right) &= \sum\nolimits_{k=0}^{K-1}\left(p_{k+1}-p_k\right)\left(g_k+g_k^0\right)\frac{1-F_{p(Z)}(p_k)}{E\left[p(Z)\right]}, \\
\overline{D}_0\left(g,g^0\right) &= \sum\nolimits_{k=1}^{K}\left(p_{k+1}-p_k\right)g_k^0\frac{F_{p(Z)}(p_k)}{1-E\left[p(Z)\right]},
\end{aligned}$$

and $\overline{D}_\Delta\left(g,g^0\right) = \overline{D}_1\left(g,g^0\right) - \overline{D}_0\left(g,g^0\right)$, where g_k, $k=0,1,\cdots,K-1$, is defined in (7), and

$$g_k^0 = \frac{1}{p_{k+1}-p_k}\int_{p_k}^{p_{k+1}} g^0(u_D)du_D, k=0,1,\cdots,K.$$

6 QRE

The QRE estimates the QTE by the coefficient of D in the quantile regression of Y on $(1,D)$. The following theorem states the resulting pseudo-true value.

Theorem 5 *Under Assumptions P and Q,*

$$\begin{aligned}
\overline{F}_1(y_1) &= \int_0^{\underline{p}}\frac{F_{Y_1|U_D}(y_1|u_D)}{E\left[p(Z)\right]}du_D + \int_{\underline{p}}^{\overline{p}} F_{Y_1|U_D}(y_1|u_D)\frac{1-F_{p(Z)}(u_D)}{E\left[p(Z)\right]}du_D, \\
\overline{F}_0(y_0) &= \int_{\underline{p}}^{\overline{p}} F_{Y_0|U_D}(y_0|u_D)\frac{F_{p(Z)}(u_D)}{E\left[1-p(Z)\right]}du_D + \int_{\overline{p}}^1\frac{F_{Y_0|U_D}(y_0|u_D)}{E\left[1-p(Z)\right]}du_D,
\end{aligned}$$

and

$$\overline{\Delta}\left(\tau\right) = \overline{F}_1^{-1}\left(\tau\right) - \overline{F}_0^{-1}\left(\tau\right).$$

The formula for $\overline{F}_d$ is similar to $\overline{\mu}_d$ except that $E\left[Y_d|U_D=u_D\right]$ is replaced by $F_{Y_d|U_D}(y_d|u_D)$, so some comments on $\overline{\mu}_d$ can be applied to $\overline{F}_d$, e.g., there is no extrapolation, $\overline{F}_d$ is estimable and the two terms in $\overline{F}_d$ can be combined. On the other hand, different from F_1^*, $\overline{F}_1$ involves only $F_{Y_1|U_D}$ but not $F_{Y_0|U_D}$; similarly, $\overline{F}_0$ involves only $F_{Y_0|U_D}$ but not $F_{Y_1|U_D}$. This makes the expression for $\overline{F}_d$ and $\overline{\Delta}\left(\tau\right)$ much neater. From the last section, $\int_0^{\underline{p}}\frac{1}{E[p(Z)]}du_D + \int_{\underline{p}}^{\overline{p}}\frac{1-F_{p(Z)}(u_D)}{E[p(Z)]}du_D = 1$ and $\frac{1-F_{p(Z)}(u_D)}{E[p(Z)]} \geq 0$, so $\overline{F}_1(y_1)$ is a genuine cdf. Similarly, $\overline{F}_0(y_0)$ is a genuine cdf. This makes the inverting operation in $\overline{\Delta}\left(\tau\right)$ valid. This also implies that if $F_{Y_d|U_D}(y_d|u_D) = F_{Y_d}(y_d)$ as in the unconfounded case, $\overline{F}_d = F_d$, and thus $\overline{\Delta}\left(\tau\right) = \Delta\left(\tau\right)$. Note also that different from F_d^* and μ_d^*, $\overline{\mu}_d = \int y_d d\overline{F}_d(y_d)$ regardless of whether there is endogeneity.

We can also decompose the bias in $\overline{F}_d$ into the part due to the selection effect and the part due to the essential heterogeneity. The selection effect means $F_{U_0|U_D}(u_0|u_D) \neq u_0$ (or $F_{Y_0|U_D}(y_0|u_D) \neq F_{Y_0}(y_0)$) and the essential heterogeneity means $F_{U_1|U_D} \neq F_{U_0|U_D}$ (or $F^{-1}_{Y_0|U_D}\left(F_{Y_1|U_D}(y_1|u_D)|u_D\right) \neq F_0^{-1}(F_1(y_1))$). Obviously, $\overline{F}_0$ only suffers from the selection effect, while $\overline{F}_1$ suffers from both. To see that, note that

$$\overline{F}_0(y_0) - F_0(y_0) = \int_{\underline{p}}^{\overline{p}} \left[F_{U_0|U_D}(F_0(y_0)|u_D) - F_0(y_0)\right] \frac{F_{p(Z)}(u_D)}{E[1-p(Z)]} du_D + \int_{\overline{p}}^{1} \frac{\left[F_{U_0|U_D}(F_0(y_0)|u_D) - F_0(y_0)\right]}{E[1-p(Z)]} du_D$$

and

$$\begin{aligned}
\overline{F}_1(y_1) - F_1(y_1) &= \int_0^{\underline{p}} \frac{F_{U_1|U_D}(F_1(y_1)|u_D) - F_1(y_1)}{E[p(Z)]} du_D + \int_{\underline{p}}^{\overline{p}} \left[F_{U_1|U_D}(F_1(y_1)|u_D) - F_1(y_1)\right] \frac{1 - F_{p(Z)}(u_D)}{E[p(Z)]} du_D \\
&= \int_0^{\underline{p}} \frac{F_{U_1|U_D}(F_1(y_1)|u_D) - F_{U_0|U_D}(F_1(y_1)|u_D) + F_{U_0|U_D}(F_1(y_1)|u_D) - F_1(y_1)}{E[p(Z)]} du_D \\
&\quad + \int_{\underline{p}}^{\overline{p}} \left[F_{U_1|U_D}(F_1(y_1)|u_D) - F_{U_0|U_D}(F_1(y_1)|u_D) + F_{U_0|U_D}(F_1(y_1)|u_D) - F_1(y_1)\right] \frac{1 - F_{p(Z)}(u_D)}{E[p(Z)]} du_D,
\end{aligned}$$

where $F_{U_1|U_D}(F_1(y_1)|u_D) - F_{U_0|U_D}(F_1(y_1)|u_D) \neq 0$ is due to the essential heterogeneity and $F_{U_0|U_D}(F_d(y_d)|u_D) - F_d(y_d) \neq 0$ is due to the selection effect.

As in the last section, we consider the discrete instrument case to aid intuition. Here, we employ the notations in Section 4.4.

$$\begin{aligned}
\overline{F}_1(y_1) &= \sum_{k=0}^{K-1} \frac{1 - F_{p(Z)}(p_k)}{E[p(Z)]} (p_{k+1} - p_k) \int_{p_k}^{p_{k+1}} \frac{F_{Y_1|U_D}(y_1|u_D)}{p_{k+1} - p_k} du_D = \sum_{k=0}^{K-1} \frac{1 - F_{p(Z)}(p_k)}{E[p(Z)]} (p_{k+1} - p_k) F_1^k(y_1), \\
\overline{F}_0(y_0) &= \sum_{k=1}^{K} \frac{F_{p(Z)}(p_k)}{E[1-p(Z)]} (p_{k+1} - p_k) \int_{p_k}^{p_{k+1}} \frac{F_{Y_0|U_D}(y_0|u_D)}{p_{k+1} - p_k} du_D = \sum_{k=1}^{K} \frac{F_{p(Z)}(p_k)}{E[1-p(Z)]} (p_{k+1} - p_k) F_0^k(y_0)
\end{aligned}$$

take similar forms as $\overline{\mu}_d$. When $K = 2$,

$$\begin{aligned}
\overline{F}_1(y_1) &= \frac{p_1}{E[p(Z)]} F_{1|\mathcal{A}}(y_1) + \frac{1 - F_{p(Z)}(p_1)}{E[p(Z)]} (p_2 - p_1) F_{1|\mathcal{C}}(y_1), \\
\overline{F}_0(y_0) &= \frac{F_{p(Z)}(p_1)}{E[1-p(Z)]} (p_2 - p_1) F_{0|\mathcal{C}}(y_0) + \frac{1 - p_2}{E[1-p(Z)]} F_{0|\mathcal{N}}(y_0),
\end{aligned}$$

so $\overline{F}_1(\cdot)$ is a weighted average of $F_{1|\mathcal{A}}(\cdot)$ and $F_{1|\mathcal{C}}(\cdot)$, and $\overline{F}_0$ is a weighted average of $F_{0|\mathcal{C}}(\cdot)$ and $F_{0|\mathcal{N}}(\cdot)$. Like $\overline{\mu}_d$, there is no extrapolation in $\overline{F}_d(\cdot)$, and different from $F_d^*(\cdot)$, $\overline{F}_d(\cdot)$ involves only two rather than three types of participants.

The bounds developed in Section 4.5 can still be used to assess the validity of QRE. We next conduct the local sensitivity analysis on $\overline{F}_d$ and $\overline{\Delta}(\tau)$. Assume $\left(F_{U_1|U_D}(\cdot|\cdot) - F_{U_0|U_D}(\cdot|\cdot), F_{U_0|U_D}(\cdot|\cdot)\right)$ stays in a parametrized space

$$\begin{aligned}
\overline{\mathcal{F}}_\alpha = \Big\{ &(F^\alpha(\cdot|\cdot), F_0^\alpha(\cdot|\cdot)) : \alpha \in M, 0 \in M, F^\alpha(0|u_D) = F^\alpha(1|u_D) = 0, \textstyle\int_0^1 F^\alpha(u|u_D) du_D = 0, \\
&F_0^\alpha(0|u_D) = 0, F_0^\alpha(1|u_D) = 1, F_0^\alpha(u|u_D) \text{ is nondecreasing in } u, \textstyle\int_0^1 F_0^\alpha(u|u_D) du_D = u, \\
&F^0(u|u_D) = 0, F_0^0(u|u_D) = u, u, u_D \in [0,1] \Big\}.
\end{aligned}$$

Our targets are

$$\overline{D}_{F_d}\left(g,g^0\right)(y_d) \equiv \lim_{\alpha\to 0}\frac{\overline{F}_d(y_d;\alpha)-F_d(y_d)}{\alpha} \text{ and } \overline{D}_\Delta\left(g,g^0\right)(\tau) = \lim_{\alpha\to 0}\frac{\overline{\Delta}(\tau;\alpha)-\Delta(\tau)}{\alpha},$$

where we use $\overline{F}_d(y_d;\alpha)$ and $\overline{\Delta}(\tau;\alpha)$ to indicate the dependence of $\overline{F}_d(y_d)$ and $\overline{\Delta}(\tau)$ on α with $\overline{F}_d(y_d;0) = F_d(y_d)$ and $\overline{\Delta}(\tau;0) = \Delta(\tau)$, and assume $\lim_{\alpha\to 0}\frac{F^\alpha(u|u_D)}{\alpha} \to g(\cdot|\cdot) \in \mathcal{C}\left([0,1]^2\right)$ and $\lim_{\alpha\to 0}\frac{F_0^\alpha(u|u_D)-u}{\alpha} \to g^0(u|u_D) \in \mathcal{C}\left([0,1]^2\right)$. We impose the following conditions to guarantee the limits in $D^*_{\overline{F}_d}$ and $D^*_{\overline{\Delta}}$ exist.

Assumption LF′: $\sup_{u_D\in[0,1],u\in[0,1],\alpha\in\mathcal{N}}\left|\frac{F^\alpha(u|u_D)}{\alpha}\right| < \infty$ and $\sup_{u_D\in[0,1],u\in[0,1],\alpha\in\mathcal{N}}\left|\frac{F_0^\alpha(u|u_D)-u}{\alpha}\right| < \infty$, where $\mathcal{N}$ is a neighborhood of 0.

Proposition 10 *Under Assumptions P, Q and LF′,*

$$\begin{aligned}
\overline{D}_{F_1}\left(g,g^0\right)(y_1) &= \int_0^{\underline{p}}\frac{g(F_1(y_1)|u_D)+g^0(F_1(y_1)|u_D)}{E\left[p(Z)\right]}du_D + \int_{\underline{p}}^{\overline{p}}\frac{1-F_{p(Z)}(u_D)}{E\left[p(Z)\right]}\left[g(F_1(y_1)|u_D)+g^0(F_1(y_1)|u_D)\right]du_D \\
&= \int_0^1\frac{1-F_{p(Z)}(u_D)}{E\left[p(Z)\right]}\left[g(F_1(y_1)|u_D)+g^0(F_1(y_1)|u_D)\right]du_D, \\
\overline{D}_{F_0}\left(g,g^0\right)(y_0) &= \int_{\underline{p}}^{\overline{p}}\frac{F_{p(Z)}(u_D)}{E\left[1-p(Z)\right]}g^0(F_0(y_0)|u_D)du_D + \int_{\overline{p}}^1\frac{g^0(F_0(y_0)|u_D)}{E\left[1-p(Z)\right]}du_D \\
&= \int_0^1\frac{F_{p(Z)}(u_D)}{E\left[1-p(Z)\right]}g^0(F_0(y_0)|u_D)du_D,
\end{aligned}$$

and

$$\overline{D}_\Delta\left(g,g^0\right)(\tau) = \frac{\overline{D}_{F_0}\left(g,g^0\right)\left(F_0^{-1}(\tau)\right)}{f_0\left(F_0^{-1}(\tau)\right)} - \frac{\overline{D}_{F_1}\left(g,g^0\right)\left(F_1^{-1}(\tau)\right)}{f_1\left(F_1^{-1}(\tau)\right)}.$$

The proof for $\overline{D}_{F_d}$ is similar to that in Proposition 3 and the proof for $\overline{D}_\Delta(\cdot)$ is similar to that for $D^*_\Delta(\cdot)$, so omitted. Similar to $\overline{D}_0\left(g,g^0\right)$, $\overline{D}_{F_0}\left(g,g^0\right)(y_0)$ only involves g^0, and similar to $\overline{D}_1\left(g,g^0\right)$ and $\overline{D}_\Delta\left(g,g^0\right)$, $\overline{D}_{F_1}\left(g,g^0\right)(y_1)$ and $\overline{D}_\Delta\left(g,g^0\right)(\tau)$ involve both g and g^0. From this proposition, $\overline{D}_{F_d}\left(g,g^0\right)(y_d)$ is a continuous linear mapping from $\mathcal{C}_0\left([0,1]^2\right)^2$ to $\mathcal{C}(\mathcal{Y}_d)$, and $D_\Delta(g)(\tau)$ is a continuous linear mapping from $\mathcal{C}_0\left([0,1]^2\right)^2$ to $\mathcal{C}([0,1])$, where $\mathcal{C}_0\left([0,1]^2\right)$ is defined in Section 4.3.

The following example numerically illustrates these path derivatives.

Example 9 *Consider the setup in the running example with both the selection effect and the essential heterogeneity. Suppose $g(u|u_D) = 6u(1-u)(2u_D-1)$ as in Example 6 and $g^0(u|u_D) = 3u(1-u)(1-2u_D)$. g^0 implies that the ranks of individuals who want most to participate move down in their untreated state (but not as much as in g), and visa versa.*

Since only $F_d(y_d)$ is involved in $\overline{D}_{F_d}\left(g,g^0\right)(y_d)$, as in $D^*_{\overline{F}_d}$, we plot $\overline{D}_{F_d}\left(g,g^0\right)$ as a function of $\tau_d \equiv F_d(y_d)$ instead of y_d and denote it as $\overline{D}_{F_d}\left(g,g^0\right)(\tau_d)$. To provide more intuition on the form of $\overline{D}_{F_d}\left(g,g^0\right)(\tau_d)$ and $\overline{D}_\Delta\left(g,g^0\right)(\tau)$, we report their formulas for this example below.

$$\begin{aligned}
\overline{D}_{F_1}\left(g,g^0\right)(\tau_1) &= \int_0^1 2(1-u_D)\left[g(\tau_1|u_D)+g^0(\tau_1|u_D)\right]du_D, \\
\overline{D}_{F_0}\left(g,g^0\right)(\tau_0) &= \int_0^1 2u_D\cdot g^0(\tau_0|u_D)du_D,
\end{aligned}$$

and

$$\overline{D}_\Delta\left(g,g^0\right)(\tau) = \frac{\overline{D}_{F_0}\left(g,g^0\right)(\tau)}{f_0\left(F_0^{-1}(\tau)\right)} - \frac{\overline{D}_{F_1}\left(g^0\right)(\tau)}{f_1\left(F_1^{-1}(\tau)\right)}$$

where

$$\begin{aligned} F_0^{-1}(\tau) &= F_1^{-1}(\tau) = \sqrt{7.8}\Phi^{-1}(\tau), \\ f_0(y) &= f_1(y) = \frac{1}{\sqrt{7.8}}\phi\left(\frac{y}{\sqrt{7.8}}\right). \end{aligned}$$

Figure 7 shows $\overline{D}_{F_d}\left(g, g^0\right)(\tau_d)$ *and* $\overline{D}_\Delta\left(g, g^0\right)(\tau)$. *For this local misspecification,* $\overline{D}_{F_1} = \overline{D}_{F_0} < 0$, *so* $\overline{F}_d(\cdot)$ *tends to underestimate* $F_d(\cdot)$ *(especially in the middle quantiles), but the biases offset each other such that a consistent estimation of* $\Delta(\cdot)$ *is achieved. Of course, it is easy to construct examples where* Δ *cannot be consistently estimated.*

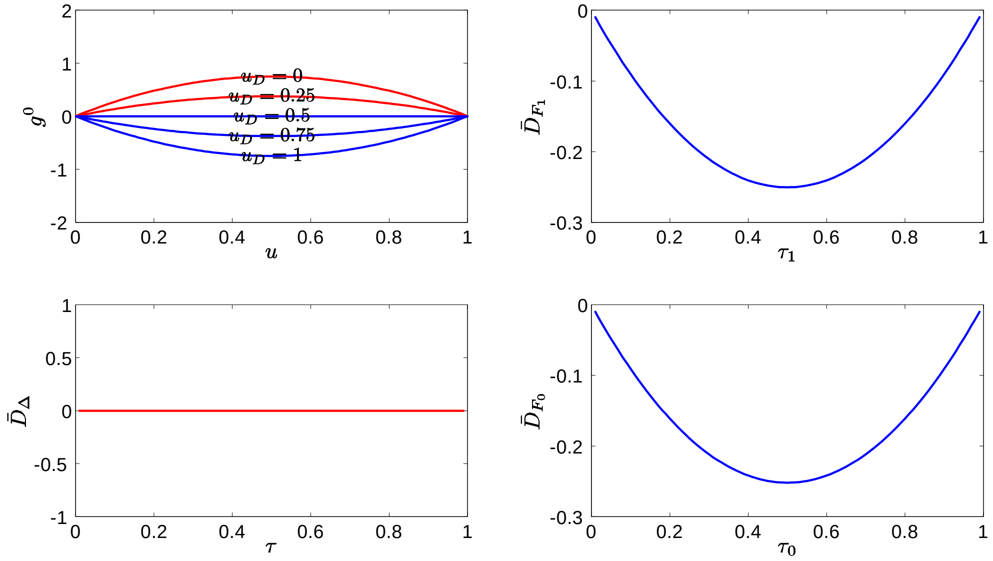

Figure 7: $g^0(u|u_D)$, $\overline{D}_{F_1}\left(g, g^0\right)(\tau_1)$, $\overline{D}_{F_0}\left(g, g^0\right)(\tau_0)$ and $\overline{D}_\Delta\left(g, g^0\right)(\tau)$

Finally, when $p(Z)$ is discrete,

$$\begin{aligned} \overline{D}_{F_1}\left(g, g^0\right)(\tau_1) &= \sum\nolimits_{k=0}^{K-1}(p_{k+1} - p_k)\left(g_k(\tau_1) + g_k^0(\tau_1)\right)\frac{1 - F_{p(Z)}(p_k)}{E\left[p(Z)\right]}, \\ \overline{D}_{F_0}\left(g, g^0\right)(\tau_0) &= \sum\nolimits_{k=1}^{K}(p_{k+1} - p_k)\, g_k^0(\tau_0)\frac{F_{p(Z)}(p_k)}{1 - E\left[p(Z)\right]}, \end{aligned}$$

and $\overline{D}_\Delta\left(g, g^0\right)(\tau) = \frac{\overline{D}_{F_0}\left(g, g^0\right)(\tau)}{f_0\left(F_0^{-1}(\tau)\right)} - \frac{\overline{D}_{F_1}\left(g^0\right)(\tau)}{f_1\left(F_1^{-1}(\tau)\right)}$, where $g_k(u)$, $k = 0, 1, \cdots, K-1$, is defined in (14), and

$$g_k^0(u) = \frac{1}{p_{k+1} - p_k}\int_{p_k}^{p_{k+1}} g^0(u|u_D)du_D, k = 0, 1, \cdots, K.$$

7 Empirical Illustration

In this section, we use the data from Angrist (1990) to illustrate the implementation of our analyses. For this application, Y is the annual earning, D is the Vietnam veteran status and Z is the U.S. draft lottery.

The data set contains information about 11637 white men, born in 1950-1953, from the March Current Population Surveys of 1979 and 1981-1985; among which, 2461 are Vietnam veterans and 3234 are eligible for U.S. military service. See Abadie (2002) for additional information on the data and the construction of the variables.

Using notations in Sections 3.3, 3.4, 4.4 and 4.5, $K = 2, p_1 = \underline{p} = 0.179$ and $p_2 = \overline{p} = 0.295$, so there are $p_1 = 17.9\%$ always-takers, $1 - p_2 = 70.5\%$ never-takers, and $p_2 - p_1 = 11.6\%$ compliers. Also,

$$\begin{aligned}
\mu_{1|\mathcal{A}} &= \frac{E[YD|Z=0]}{p_1} = 11301.54, \\
\mu_{1|\mathcal{C}} &= \frac{E[YD|Z=1] - E[YD|Z=0]}{p_2 - p_1} = 11637.21, \\
\mu_{0|\mathcal{C}} &= \frac{E[Y(1-D)|Z=1] - E[Y(1-D)|Z=0]}{p_1 - p_2} = 12915.00, \\
\mu_{0|\mathcal{N}} &= \frac{E[Y(1-D)|Z=1]}{1 - p_2} = 11462.32,
\end{aligned}$$

so by the results in Section 3.3,

$$\mu_1^* = 10552.86,\ \mu_0^* = 11830.64 \text{ and } \Delta^* = LATE = -1277.78.$$

As expected, $\Delta^* < 0$; i.e., participating the Vietnam war is harmful to subsequent earnings. It is interesting to observe that $\mu_{1|\mathcal{A}} < \mu_{1|\mathcal{C}}$ and $\mu_{0|\mathcal{C}} > \mu_{0|\mathcal{N}}$; i.e., the average earning of always-taker veterans is lower than that of complier veterans, and the average earning of never-taker non-veterans is lower than that of complier non-veterans. As to the shape of h_d and h, note that $h_1(p_2) = 7.10, h_0(p_2) = -1.55$ and $h(p_2) = 8.65$, so there is indeed overweighting of $E[Y_1|U_D = u_D]$ in μ_1^* and $E[Y_0|U_D = u_D]$ in μ_0^* for $u_D \in (p_1, p_2]$.

We next conduct the local sensitivity analysis and the partial identification analysis for the IVE. For the former, suppose $g(u_D) = (1 - 2u_D)\overline{Y}$, where $\overline{Y} = 11560.42$ is the sample mean of Y. Here, we add $\overline{Y}$ to the specification of $g(u_D)$ in Example 2 to accomodate the scale of Y. For this g,

$$\begin{aligned}
g_0 &= \tfrac{1}{p_1} \int_0^{p_1} g(u_D) du_D = 0.15\overline{Y}, \\
g_1 &= \tfrac{1}{p_2 - p_1} \int_{p_1}^{p_2} g(u_D) du_D = 0.53\overline{Y}, \\
g_2 &= \tfrac{1}{1 - p_2} \int_{p_2}^{1} g(u_D) du_D = -0.29\overline{Y},
\end{aligned} \tag{15}$$

so

$$\begin{aligned}
D_1^*(g) &= -(p_2 - p_1)(1 - h_1(p_2)) g_1 - (1 - p_2) g_2 = 0.58\overline{Y} = 6688.54, \\
D_0^*(g) &= p_1 g_0 + (p_2 - p_1) h_0(p_2) g_1 = -0.068\overline{Y} = -784.71, \\
D_\Delta^*(g) &= D_1^*(g) - D_0^*(g) = 0.65\overline{Y} = 7473.25,
\end{aligned}$$

i.e., for this local misspecification, we tend to overestimate μ_1 and Δ but underestimate μ_0. As to the partial identification analysis, it turns out that

$$\begin{aligned}
\mu_1 &\in [\underline{I}_1, \overline{I}_1] = [3372.67, 39098.01], \\
\mu_0 &\in [\underline{I}_0, \overline{I}_0] = [9574.66, 20828.66], \\
\Delta &\in [\underline{I}_\Delta, \overline{I}_\Delta] = [-17455.99, 29523.34].
\end{aligned}$$

All these bounds cover the corresponding pseudo-true values, and $[\underline{I}_\Delta, \overline{I}_\Delta]$ covers 0, i.e., the bounds for Δ

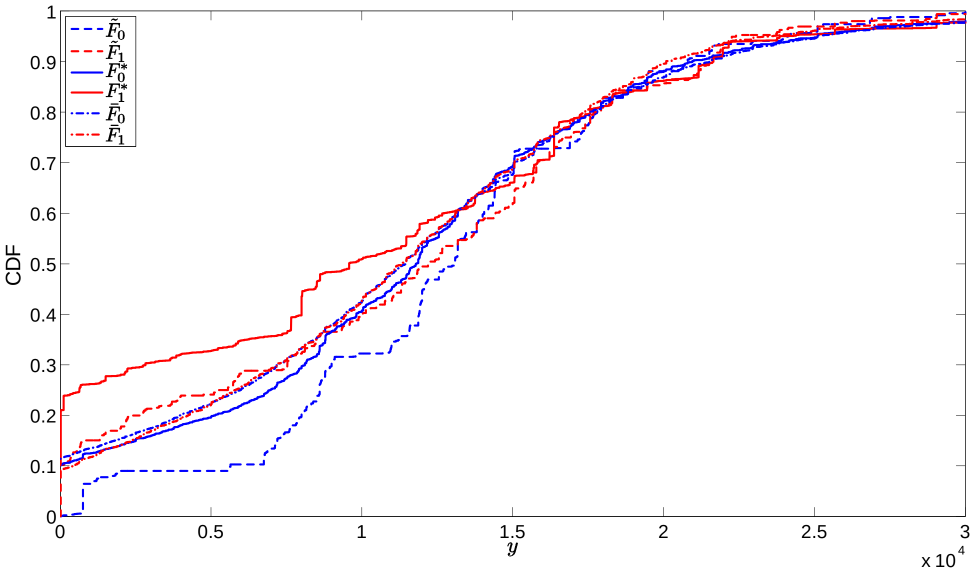

Figure 8: $\widetilde{F}_d(\cdot)$, $F_d^*(\cdot)$ and $\overline{F}_d(\cdot)$

cannot exclude null ATE.

We now turn to the analysis for the IV-QTE. From Abadie (2002) and Imbens and Rubin (1997),

$$\begin{aligned}
F_{Y_1|\mathcal{A}} &= \frac{E\left[1\left(Y \leq y_1\right)D|Z=0\right]}{p_1} = P\left(Y \leq y_1|Z=0, D=1\right), \\
F_{Y_1|\mathcal{C}} &= \frac{E\left[1\left(Y \leq y_1\right)D|Z=1\right] - E\left[1\left(Y \leq y_1\right)D|Z=0\right]}{p_2 - p_1}, \\
F_{Y_0|\mathcal{C}} &= \frac{E\left[1\left(Y \leq y_0\right)\left(1-D\right)|Z=1\right] - E\left[1\left(Y \leq y_0\right)\left(1-D\right)|Z=0\right]}{p_1 - p_2}, \\
F_{Y_0|\mathcal{N}} &= \frac{E\left[1\left(Y \leq y_0\right)\left(1-D\right)|Z=1\right]}{1-p_2} = P\left(Y \leq y_0|Z=1, D=0\right).
\end{aligned}$$

Figure 8 shows $\widetilde{F}_d(\cdot)(= F_{Y_d|\mathcal{C}}(\cdot))$ and $F_d^*(\cdot)$ whose formulas are stated in Section 4.4. Recall that $\mu_{d|\mathcal{C}} = \int y_d dF_{Y_d|\mathcal{C}}(y_d)$ but $\mu_d^* \neq \int y_d dF_d^*(y_d)$. Note also that from the tests in Section 4.1 of Kitagawa (2015), the inversion of $F_{Y_d|\mathcal{C}}(\cdot)$ in $F_d^*(\cdot)$ is justified. In finite samples, we rearrange $F_{Y_d|\mathcal{C}}(\cdot)$ before inversion to guarantee the monotonicity of $F_d^*(\cdot)$; note that $F_{Y_1|\mathcal{A}}(\cdot)$ and $F_{Y_0|\mathcal{N}}(\cdot)$ are automatically monotone in finite samples. These functions are similar to those reported in Section 5.3 of Wütherich (2015) and Section 3.1 of Chernozhukov et al. (2010). Figure 9 shows the sharp bounds for $Q_d(\tau)$ and $\Delta(\tau)$, $\tau \in [0.12, 0.95]$. Here, we restrict $\tau > 0.12$ because there are about 9% $Y_1 = 0$ and 11% $Y_0 = 0$ in the sample. From these bounds, we can see that for $\tau \in [0.12, 0.95]$, (i) $Q_d^*(\tau) \in [\underline{I}_d(\tau), \overline{I}_d(\tau)]$; (ii) $0 \in \left[\underline{I}_\Delta(\tau), \overline{I}_\Delta(\tau)\right]$ and $\Delta^*(\tau) \in \left[\underline{I}_\Delta(\tau), \overline{I}_\Delta(\tau)\right]$;[21] (iii) the width of the bounds depends on τ.

For the local sensitivity analysis, suppose $g(u|u_D) = 6u(1-u)(2u_D - 1)$ as in Example 6. Different from

[21] $\Delta^*(\tau)$ is generally negative especially for $\tau < 0.5$. This matches the result that $\Delta^* < 0$ although $\Delta^* \neq \int \Delta^*(\tau) d\tau$.

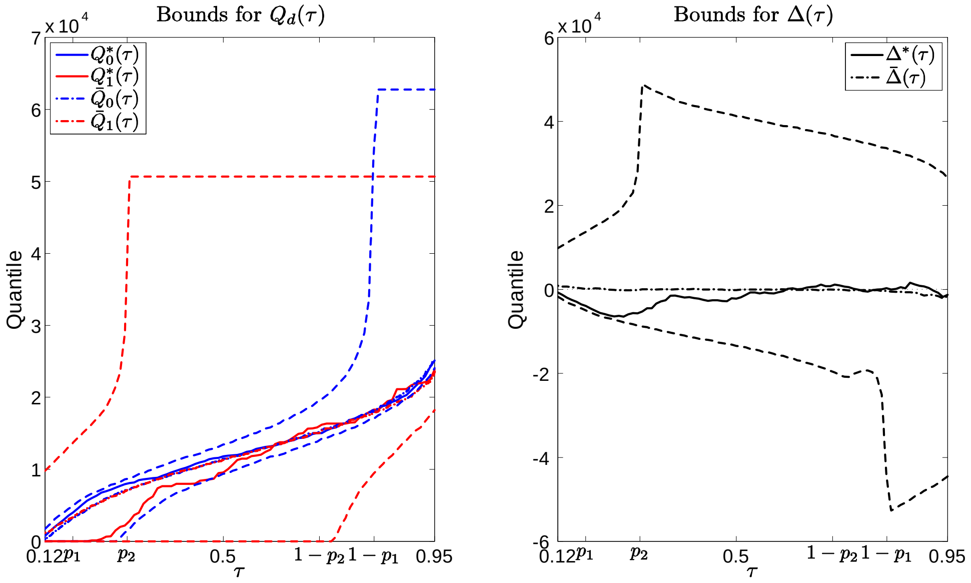

Figure 9: Bounds for $Q_d(\tau)$ and $\Delta(\tau)$, $\tau \in [0.12, 0.95]$

$g(u_D)$ in the IVE case, $g(u|u_D)$ is unit-free since both u and u_D are in $[0,1]$. For this g,

$$\begin{aligned} g_0(u) &= \tfrac{1}{p_1}\int_0^{p_1} g(u|u_D)du_D = -6g_0u(1-u), \\ g_1(u) &= \tfrac{1}{p_2-p_1}\int_{p_1}^{p_2} g(u|u_D)du_D = -6g_1u(1-u), \\ g_2(u) &= \tfrac{1}{1-p_2}\int_{p_2}^{1} g(u|u_D)du_D = -6g_2u(1-u), \end{aligned} \tag{16}$$

where g_k, $k=0,1,2$, are defined in the local sensitivity analysis for the IVE. As in Example 6, we report $D_{F_d}^*(g)$ as a function of $\tau_d \equiv F_d(y_d)$. Specifically,

$$\begin{aligned} D_{F_1}^*(g)(\tau_1) &= \frac{(p_2-p_1)\,q_1 f_0^1\left(F_0^{-1}(\tau_1)\right)+(1-p_2)f_0^2\left(F_0^{-1}(\tau_1)\right)}{f_0^1(F_0^{-1}(\tau_1))}g_1(\tau_1) \\ &\quad -\left[(p_2-p_1)\,q_1g_1(\tau_1)+(1-p_2)g_2(\tau_1)\right] \\ D_{F_0}^*(g)(\tau_0) &= -\frac{p_1f_1^0\left(F_1^{-1}(\tau_0)\right)+(p_2-p_1)(1-q_1)f_1^1\left(F_1^{-1}(\tau_0)\right)}{f_1^1\left(F_1^{-1}(\tau_0)\right)}g_1(\tau_0) \\ &\quad +\left[p_1g_0(\tau_0)+(p_2-p_1)(1-q_1)\,g_1(\tau_0)\right], \end{aligned}$$

and

$$D_\Delta^*(g)(\tau) = D_{F_0}^*(g)(\tau)\cdot s_0(\tau) - D_{F_1}^*(g)(\tau)\cdot s_1(\tau),$$

where $q_1 = P(Z=0)$, and $s_d(\tau) = \frac{1}{f_d\left(F_d^{-1}(\tau)\right)}$ is the sparsity which includes the information about the scale of $D_{\Delta}^{*}(g)(\tau)$. $f_0^1(y_0)$ and $f_1^1(y_1)$ are estimated by the sample analog of

$$\begin{aligned}
& \frac{E\left[L_b(Y-y_0)(1-D)|Z=1\right]-E\left[L_b(Y-y_0)(1-D)|Z=0\right]}{p_1-p_2} \\
= & \frac{E\left[L_b(Y-y_0)|Z=1,D=0\right](1-p_2)-E\left[L_b(Y-y_0)|Z=0,D=0\right](1-p_1)}{p_1-p_2} \\
\approx & \frac{f_{10}(y_0)(1-p_2)-f_{00}(y_0)(1-p_1)}{p_1-p_2}
\end{aligned}$$

and

$$\begin{aligned}
& \frac{E\left[L_b(Y-y_1)D|Z=1\right]-E\left[L_b(Y-y_1)D|Z=0\right]}{p_2-p_1} \\
= & \frac{E\left[L_b(Y-y_1)|Z=1,D=1\right]p_2-E\left[L_b(Y-y_1)|Z=0,D=1\right]p_1}{p_2-p_1} \\
\approx & \frac{f_{11}(y_1)p_2-f_{01}(y_1)p_1}{p_2-p_1},
\end{aligned}$$

respectively, and $f_0^2(y_0)$ and $f_1^0(y_1)$ are estimated by the sample analog of

$$\frac{E\left[L_b(Y-y_0)\left(1-D\right)|Z=1\right]}{1-p_2}=E\left[L_b(Y-y_0)|Z=1,D=0\right]\approx f_{10}(y_0)$$

and

$$\frac{E\left[L_b(Y-y_1)D|Z=0\right]}{p_1}=E\left[L_b(Y-y_1)|Z=0,D=1\right]\approx f_{01}(y_1),$$

respectively, where $L_b(\cdot)=\frac{1}{b}L\left(\frac{\cdot}{b}\right)$ for a kernel function $L(\cdot)$ and a bandwidth b, and $f_{zd}(\cdot)$ is the conditional density of Y given $Z=z$ and $D=d$. We use the standard normal kernel and select b based on Botev et al. (2010). $s_d(\tau)$ is estimated based on the difference quotient, as discussed by Siddiqui (1960) and Hendricks and Koenker (1992):

$$\frac{Q_d^*(\tau+b)-Q_d^*(\tau-b)}{2b},$$

where $Q_d^*(\cdot)=F_d^{*-1}(\cdot)$ is consistent to $Q_d(\cdot)$ under Assumption RS, $b=n_d^{-1/5}\left[4.5\phi^4\left(\Phi^{-1}(\tau)\right)/\left(\left(2\Phi^{-1}(\tau)\right)^2+1\right)\right]^{1/5}$ depends on d and τ and is based on Bofinger (1975) with $n_d=\sum_{i=1}^{n}1(D_i=d)$.[22] Figure 10 shows $D_{F_d}^{*}(g)(\tau_d)$ and $D_{\Delta}^{*}(g)(\tau)$ for $\tau_d,\tau\in[0.12,0.95]$. From this figure, we can see a few facts for this local misspecification: (i) $F_0^*(\cdot)$ tends to overestimate $F_0(\cdot)$ but the magnitude is quite small; (ii) $F_1^*(\cdot)$ tends to underestimate $F_1(\cdot)$ especially at lower quantiles; (iii) $\Delta^*(\cdot)$ tends to overestimate $\Delta(\cdot)$ especially when the quantile index is not small, where both the large values of $\Delta(\cdot)$ and the wiggles of $\Delta(\cdot)$ are inherited from $s_d(\cdot)$. These results match those in the IVE case: $D_0^*(g)<0$, $D_1^*(g)>0$ and $D_{\Delta}^*(g)>0$.

We next consider the LSE. From $\overline{\mu}_1=\frac{E[DY]}{E[D]}$ and $\overline{\mu}_0=\frac{E[(1-D)Y]}{1-E[D]}$, we have $\overline{\mu}_1=11352.55\in[\underline{I}_1,\overline{I}_1]$, $\overline{\mu}_0=11616.17\in[\underline{I}_0,\overline{I}_0]$ and $\overline{\Delta}=-263.62\in[\underline{I}_{\Delta},\overline{I}_{\Delta}]$, where the bounds are stated in the analysis for the IVE, $|\overline{\mu}_1|>|\mu_1^*|$, $|\overline{\mu}_0|<|\mu_0^*|$ and $\left|\overline{\Delta}\right|<|\Delta^*|$. For the local sensitivity analysis, suppose $g(u_D)=(1-2u_D)\overline{Y}$ and $g^0(u_D)=(u_D-0.5)\overline{Y}$ as in Example 8 but add $\overline{Y}$ as in the local sensitivity analysis for the IVE. For

[22] An alternative bandwidth in Hall and Sheather (1988) is designed for confidence interval construction for quantiles, rather than optimizing MSE-performance for the sparsity estimated itself as in Bofinger (1975), so is not suitable here.

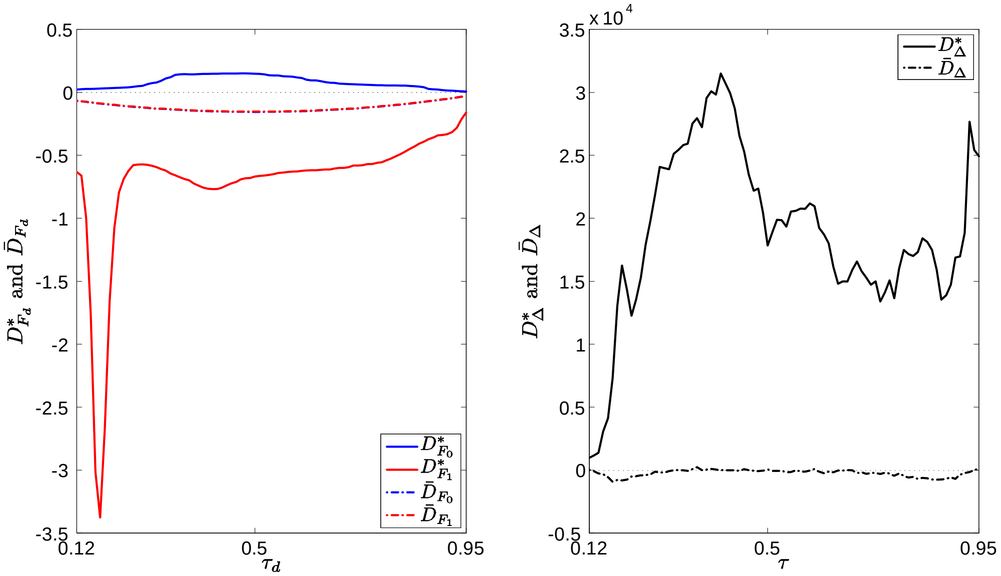

Figure 10: $D^*_{F_d}\left(g\right)\left(\tau_d\right)$, $\overline{D}_{F_d}\left(g\right)\left(\tau_d\right)$, $D^*_\Delta\left(g\right)\left(\tau\right)$ and $\overline{D}_\Delta\left(g\right)\left(\tau\right)$ for $\tau_d, \tau \in [0.12, 0.95]$

this specification of local deviation, g_k, $k = 0, 1, 2$, are defined in (15), and

$$\begin{aligned} g_0^0 &= \tfrac{1}{p_1}\int_0^{p_1} g^0(u_D)du_D = -0.5g_0\overline{Y}, \\ g_1^0 &= \tfrac{1}{p_2-p_1}\int_{p_1}^{p_2} g^0(u_D)du_D = -0.5g_1\overline{Y}, \\ g_2^0 &= \tfrac{1}{1-p_2}\int_{p_2}^{1} g^0(u_D)du_D = -0.5g_2\overline{Y}. \end{aligned}$$

So

$$\begin{aligned} \overline{D}_1\left(g, g^0\right) &= \frac{p_1}{E\left[D\right]}\left(g_0 + g_0^0\right) + \frac{\left(p_2 - p_1\right)\left(1 - q_1\right)}{E\left[D\right]}\left(g_1 + g_1^0\right) = 0.102\overline{Y} = 1183.20, \\ \overline{D}_0\left(g, g^0\right) &= \frac{\left(p_2 - p_1\right)q_1}{1 - E\left[D\right]}g_1^0 + \frac{1 - p_2}{1 - E\left[D\right]}g_2^0 = 0.104\overline{Y} = 1202.73, \\ \overline{D}_\Delta\left(g, g^0\right) &= \overline{D}_1\left(g, g^0\right) - \overline{D}_0\left(g, g^0\right) = -0.00169\overline{Y} = -19.52. \end{aligned}$$

We finally discuss the QRE. Note that $\overline{F}_1(y_1) = \frac{E[D\cdot 1(Y\leq y_1)]}{E[D]}$ and $\overline{F}_0\left(y_0\right) = \frac{E[(1-D)\cdot 1(Y\leq y_0)]}{1-E[D]}$; we plot $\overline{F}_d(y_d)$ also in Figure 8 for comparison, where $\overline{F}_d(y_d)$ is automatically monotone. Different from F_1^* and F_0^*, $\overline{F}_1$ and $\overline{F}_0$ are much closer to each other. $\overline{Q}_d(\tau)$ and $\overline{\Delta}\left(\tau\right)$ are plotted in Figure 9 for comparison; it turns out that $\overline{Q}_d(\tau) \in [\underline{I}_d(\tau), \overline{I}_d(\tau)]$, $\overline{\Delta}(\tau) \in \left[\underline{I}_\Delta(\tau), \overline{I}_\Delta(\tau)\right]$ and $\overline{\Delta}\left(\tau\right)$ is much smaller than $\Delta^*\left(\tau\right)$ in absolute value for $\tau \in [0.12, 0.95]$. For the local sensitivity analysis, suppose $g(u|u_D) = 6u(1-u)(2u_D - 1)$ and $g^0(u|u_D) = 3u(1-u)(1-2u_D)$ as in Example 9. For this specification of local deviation, $g_k(u)$, $k = 0, 1, 2$, are defined in (16), and

$$\begin{aligned} g_0^0(u) &= \tfrac{1}{p_1}\int_0^{p_1} g^0(u|u_D)du_D = -0.5g_0(u), \\ g_1^0(u) &= \tfrac{1}{p_2-p_1}\int_{p_1}^{p_2} g^0(u|u_D)du_D = -0.5g_1(u), \\ g_2^0(u) &= \tfrac{1}{1-p_2}\int_{p_2}^{1} g^0(u|u_D)du_D = -0.5g_2(u). \end{aligned}$$

So

$$
\begin{aligned}
\overline{D}_{F_1}\left(g,g^0\right)(\tau_1) &= \frac{p_1}{E\left[D\right]}\left(g_0(\tau_1)+g_0^0(\tau_1)\right)+\frac{(p_2-p_1)\left(1-q_1\right)}{E\left[D\right]}\left(g_1(\tau_1)+g_1^0(\tau_1)\right)\\
\overline{D}_{F_0}\left(g,g^0\right)(\tau_0) &= \frac{(p_2-p_1)\,q_1}{1-E\left[D\right]}g_1^0(\tau_0)+\frac{1-p_2}{1-E\left[D\right]}g_2^0(\tau_0),
\end{aligned}
$$

and

$$\overline{D}_{\Delta}\left(g,g^0\right)(\tau)=\overline{D}_{F_0}\left(g,g^0\right)(\tau)\cdot s_0\left(\tau\right)-\overline{D}_{F_1}\left(g^0\right)(\tau)\cdot s_1\left(\tau\right),$$

which are also plotted in Figure 10 for comparison, where $s_d\left(\tau\right)$ is similarly estimated as in the local sensitivity analysis for the IV-QRE but replacing $Q_d^*\left(\cdot\right)$ by $\overline{Q}_d\left(\cdot\right)=\overline{F}_d^{-1}\left(\cdot\right)$. From Figure 10, we can see a few facts for this local misspecification: (i) $\overline{D}_{F_1}\approx\overline{D}_{F_0}$ at all quantiles, which is inherited from $\overline{F}_1\approx\overline{F}_0$; (ii) $\overline{F}_d\left(\cdot\right)$ tends to underestimate $F_d\left(\cdot\right)$ but the magnitude is quite small; (iii) $\overline{\Delta}\left(\cdot\right)$ tends to underestimate $\Delta\left(\cdot\right)$ especially at extreme quantiles, but the magnitude is relatively small compared to $\Delta^*\left(\cdot\right)$.

We conclude this section by one more comment on the analyses above. Neither of the four estimators contradicts the partial identification analysis, so it seems that the sharp bounds are too wide to be informative in this application. On the other hand, the local sensitivity analysis seems more informative. For example, although the consistency of the LSE and QRE requires stronger assumption than the IVE and IV-QRE, they are more robust in the sense that they are less sensitive to the local perturbation from the correctly-specified models.

8 Conclusion

In this paper, we analyze the IVE, IV-QRE, LSE and QRE, in the framework of HV, i.e., we analyze four treatment effects estimators under misspecification. We derive the pseudo-true values, conduct the local sensitivity analysis, and develop sharp bounds for the QTE which are parallel to the bounds in Heckman and Vytlacil (2001b).

We concentrate on the identification issue in this paper. This can be justified by Varian (2014), "In this period of "big data," it seems strange to focus on sampling uncertainty, which tends to be small with large datasets, while completely ignoring model uncertainty, which may be quite large." Nevertheless, for the size of most treatment datasets, inferences may still be important and we plan to investigate this issue in another occasion. Actually, because all the four estimators are defined by moment conditions, their asymptotic distributions can be developed in a standard way. On the other hand, because all our analyses are conditional on X, we may be interested in pseudo-parameters averaging over the distribution of X. This is somewhat like the average derivative estimator in Stoker (1986) and PSS (1989) and the techniques there may be employed. Another limitation of this paper is to maintain the monotonicity assumption throughout our analysis. This assumption promises a neat analysis although the consistency of the IVE and IV-QRE does not require such a restriction (as long as the essential heterogeneity is excluded); we will relax this assumption and conduct a parallel analysis in a companion paper Yu (2016a).

References

Abadie, A., 2002, Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models, *Journal of the American Statistical Association*, 97, 284-292.

Angrist, J.D., 1990, Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records, *American Economic Review*, 80, 313-336.

Angrist, J.D., 2004, Treatment Effect Heterogeneity in Theory and Practice, *Economic Journal*, 114, C52-C83.

Angrist, J.D., V. Chernozhukov and I. Fernández-Val, 2006, Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure, *Econometrica*, 74, 539-563.

Angrist, J.D. and I. Fernández-Val, 2013, ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework, *Advances in Economics and Econometrics*, Tenth World Congress, Vol III, D. Acemoglu, M. Arellano and E. Dekel, eds., Cambridge: Cambridge University Press.

Angrist, J.D., K. Graddy and G.W. Imbens, 2000, The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish, *Review of Economic Studies*, 67, 499-527.

Angrist, J.D., G.W. Imbens and D.B. Rubin (AIR), 1996, Identification of Causal Effects Using Instrumental Variables (with discussion), *Journal of the American Statistical Association*, 91, 444-472.

Athey, S. and G. Imbens, 2015, A Measure of Robustness to Misspecification, *American Economic Review: Papers and Proceedings*, 105, 476-480.

Bickel, P.J., C.A.J. Klassen, Y. Ritov and J.A. Wellner, 1998, *Efficient and Adaptive Estimation for Semiparametric Models*, New York: Springer.

Bofinger, E., 1975, Estimation of a Density Function Using Order Statistics, *Australian Journal of Statistics*, 17, 1-7.

Botev, Z.I., J.F. Grotowski and D.P. Kroese, 2010, Kernel Density Estimation via Diffusion, *The Annals of Statistics*, 38, 2916–2957.

Box, G.E.P. and N.R. Draper, 1987, *Empirical Model Building and Response Surfaces*, New York: John Wiley & Sons.

Carneiro, P., J.J. Heckman and E. Vytlacil, 2010, Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin, *Econometrica*, 78, 377-394.

Carneiro, P., J.J. Heckman and E. Vytlacil, 2011, Estimating Marginal Returns to Education, *American Economic Review*, 101, 2754-2781.

Chamberlain, G., 1987, Asymptotic Efficiency in Estimation with Conditional Moment Restrictions, *Journal of Econometrics*, 34, 305-334.

Chernozhukov, V. and C. Hansen, 2005, An IV Model of Quantile Treatment Effects, *Econometrica*, 73, 245-261.

Chernozhukov, V. and C. Hansen, 2006, Instrumental Quantile Regression Inference for Structural and Treatment Effect Models, *Journal of Econometrics*, 132, 491-525.

Chernozhukov, V. and C. Hansen, 2013, Quantile Models with Endogeneity, *Annual Review of Economics*, 5, 57-81.

Chernozhukov, V., I. Fernández-Val and A. Galichon, 2010, Quantile and Probability Curves without Crossing, *Econometrica*, 78, 1093-1125.

Claeskens, G. and N.L. Hjort, 2008, *Model Selection and Model Averaging*, New York: Cambridge University Press.

Cornfield, J., Haenszei, W., Hammond, E., Lilienfeld, A., Shimkin, M., and E. Wynder, 1959, Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions, *Journal of the National Cancer Institute*, 22, 173-203.

Dawid, A.P., 1979, Conditional Independence in Statistical Theory, *Journal of the Royal Statistical Society, Series B*, 41, 1-31.

Dong, Y.-Y. and S. Shen, 2015, Testing for Rank Invariance or Similarity in Program Evaluation: The Effect of Training on Earnings Revisited, mimeo, UC-Irvine.

Fisher, R.A., 1918, The Causes of Human Variability, *Eugenics Review*, 10, 213-220.

Fisher, R.A., 1925, *Statistical Methods for Research Workers*, 1st ed., Edinburgh: Oliver and Boyd.

Fisher, R.A., 1935, *The Design of Experiments*, 1st ed., London: Oliver and Boyd.

Glynn, R.J., Laird, N.M. and D.B. Rubin, 1986, Selection Modeling versus Mixture Modeling with Nonignorable Nonresponse, *Drawing Inferences from Self-Selected Samples*, Howard Wainer, eds., New York: Springer-Verlag.

Godfrey, L.G., 1988, *Misspecification Tests in Econometrics: the Lagrange Multiplier Principle and Other Approaches*, New York: Cambridge University Press.

Gonzáez-Manteiga, W. and R.M. Crujeiras, 2013, An Updated Review of Goodness-of-Fit Tests for Regression Models (with discussion), Test, 22, 361-447.

Hall, A.R. and A. Inoue, 2003, The Large Sample Behavior of the Generalized Method of Moments Estimator in Misspecified Models, *Journal of Econometrics*, 114, 361-394.

Hall, P. and S. Sheather, 1988, On the Distribution of a Studentized Quantile, *Journal of the Royal Statistical Society, Series B*, 50, 381-391.

Hausman, J.A., 1978, Specification Testing in Econometrics, *Econometrica*, 46, 1251-1271.

Hausman, J.A., 1983, Specification and Estimation of Simultaneous Equation Models, *Handbook of Econometrics*, Vol. 1, Z. Griliches and M.D. Intriligator, eds., Amsterdam: Elsevier Science B.V., Ch. 7, 391-448.

Heckman, J.J., 2010, Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy, *Journal of Economic Literature*, 48, 356-398.

Heckman, J.J. and D. Schmierer, 2010, Tests of Hypotheses Arising in the Correlated Random Coefficient Model, *Economic Modelling*, 27, 1355-1367.

Heckman, J.J., D. Schmierer and S. Urzua, 2010, Testing the Correlated Random Coefficient Model, *Journal of Econometrics*, 158, 177-203.

Heckman, J.J. and E.J. Vytlacil, 1999, Local Instrumental Variables and Latent Variable Models for Identifying the Bounding Treatment Effects, *Proceedings of the National Academy of Sciences*, 96, 4730-4734.

Heckman, J.J. and E.J. Vytlacil, 2001a, Local Instrumental Variables, *Nonlinear Statistical Modeling*: *Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, C. Hsiao, K. Morimune and J.L. Powell, eds., New York: Cambridge University Press, 1-46.

Heckman, J.J. and E.J. Vytlacil, 2001b, Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect, *Econometric Evaluation of Labour Market Policies*, M. Lechner and M. Pfeiffer, eds., Heidelberg: Physica-Verlag, 1-15.

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

Heckman, J.J. and E.J. Vytlacil, 2007a, Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation, *Handbook of Econometrics*, Vol. 6B, J.J. Heckman and E.E. Leamer, eds., Amsterdam: Elsevier Science B.V., Ch. 70, 4779-4874.

Heckman, J.J. and E.J. Vytlacil, 2007b, Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast Their Effects in New Environments, *Handbook of Econometrics*, Vol. 6B, J.J. Heckman and E.E. Leamer, eds., Amsterdam: Elsevier Science B.V., Ch. 71, 4875-5143.

Hendricks, B. and R. Koenker, 1992, Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity, *Journal of the American Statistical Association*, 87, 58-68.

Holland, P.W., 1986a, A Comment on Remarks by Rubin and Hartigan, *Drawing Inferences from Self-Selected Samples,* Howard Wainer, eds., New York: Springer-Verlag.

Holland, P.W., 1986b, Statistics and Causal Inference (with discussion), *Journal of the American Statistical Association*, 81, 945–970.

Holly, A., 1987, Specification Tests: An Overview, *Advances in Econometrics*, 5th World Congress, Vol I, T.F. Bewley, eds., New York: Cambridge University Press, Ch. 2, 59-97.

Hsiao, C., 1983, Identification, *Handbook of Econometrics*, Vol. 1, Z. Griliches and M.D. Intriligator, eds., Amsterdam: Elsevier Science B.V., Ch. 4, 223-283.

Imbens, G.W. and J.D. Angrist (IA), 1994, Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62, 467-475.

Imbens, G.W., Rubin, D.B., 1997, Estimating Outcome Distributions for Compliers in Instrumental Variables Models, *Review of Economic Studies*, 64, 555-574.

Imbens, G.W. and D.B. Rubin, 2015, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge: Cambridge University Press.

Imbens, G.W. and J.M. Wooldridge, 2009, Recent Developments in the Econometrics of Program Evaluation, *Journal of Economic Literature*, 47, 5-86.

Kitagawa, T., 2015, A Test for Instrument Validity, *Econometrica*, 83, 2043-2063.

Kitagawa, T. and C. Muris, 2016, Model Averaging in Semiparametric Estimation of Treatment Effects, *Journal of Econometrics*, 193, 271-289.

Lewbel, A., 2016, The Identification Zoo - Meanings of Identification in Econometrics, mimeo, BC.

Manski, C., 1995, *Identification Problems in the Social Sciences*, Cambridge, Mass.: Harvard University Press.

Manski, C., 2003, *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Manski, C., 2007, *Identification for Prediction and Decision*, Cambridge, Mass.: Harvard University Press.

Matzkin, R.L., 1994, Restrictions of Economic Theory in Nonparametric Methods, *Handbook of Econometrics*, Vol. 4, R.F. Eagle and D.L. McFadden, eds., Amsterdam: Elsevier Science B.V., Ch. 42, 2523-2558.

Matzkin, R.L., 2007, Nonparametric Identification, *Handbook of Econometrics*, Vol. 6B, J.J. Heckman and E.E. Leamer, eds., Amsterdam: Elsevier Science B.V., Ch. 73, 5307-5368.

Matzkin, R.L., 2013, Nonparametric Identification in Structural Economic Models, *Annual Review of Economics*, 5, 457-486.

Mroz, T.A., 1987, The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, *Econometrica*, 55, 765–799.

Newey, W.K., 1990, Semiparametric Efficiency Bounds, *Journal of Applied Econometrics*, 5, 99-135.

Neyman, J., 1923, On the Application of Probability Theory to Agricultural Experiments, Essays on Principles, Section 9, *Roczniki Nauk Rolniczych Tom X*, 1-51 (Annals of Agricultural Sciences), reprinted in *Statistical Science*, 5, 465–472.

Neyman, J., 1935, Statistical Problems in Agricultural Experiments, *Supplement to the Journal of the Royal Statistical Society*, 2, 107-180.

Pearson, K., 1900, On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that It can be reasonably supposed to have arisen from Random Sampling, *Philosophical Magazine Series 5*, 50, 157–175.

Peterson, A.V., 1976, Bounds for a Joint Distribution Function with Fixed Sub-Distribution Functions: Application to Competing Risks, *Proceedings of the National Academy of Sciences of the United States of America*, 73, 11-13.

Powell, J.L., J.H. Stock and T.M. Stocker (PSS), 1989, Semiparametric Estimation of Index Coefficients, *Econometrica*, 57, 1403-1430.

Ramsey, J.B., 1969, Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis, *Journal of the Royal Statistical Society*, *Series B*, 31, 350-371.

Ramsey, J.B., 1970, Models, Specification Error, and Inference: A Discussion of Some Problems in Econometric Methodology, *Bulletin of the Oxford Institute of Economics and Statistics*, 32, 301-318.

Rosenbaum, P.R., 2002, Sensitivity to Hidden Bias, *Observational Studies*, Leipzig: Springer-Verlag, Ch. 4, 105-170.

Rubin, D.B., 1974, Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies, *Journal of Educational Psychology*, 66, 688-701.

Rubin, D.B., 1978, Bayesian Inference for Causal Effects: The Role of Randomization, *Annals of Statistics*, 6, 34–58.

Siddiqui, M., 1960, Distribution of Quantiles from a Bivariate Population, *Journal of Research of the National Bureau of Standards*, 64B, 145-150.

Stoker, T.M., 1986, Consistent Estimation of Scaled Coefficients, *Econometrica*, 1461-1481.

Tamer, E., 2010, Partial Identification in Econometrics, *Annual Review of Economics*, 2, 167-195.

Varian, H.R., 2014, Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives*, 28, 3-27.

Vijverberg, W.P.M., 1993, Measuring the Unidentified Parameter of the Extended Roy Model of Selectivity, *Journal of Econometrics*, 57, 69-89.

Vytlacil, E.J., 2002, Independence, Monotonicity, and Latent Index Models: An Equivalence Result, *Econometrica*, 70, 331-341.

Vytlacil, E.J., 2006, A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results, *Oxford Bulletin of Economics and Statistics*, 68, 515-518.

White, H., 1980, Using Least Squares to Approximate Unknown Regression Functions, *International Economic Review*, 21, 149-170.

White, H., 1981, Consequences and Detection of Misspecified Nonlinear Regression Models, *Journal of the American Statistical Association*, 76, 419-433.

White, H., 1982, Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, 50, 1-25.

White, H., 1987, Specification Testing in Dynamic Models, *Advances in Econometrics*, 5th World Congress, Vol I, T.F. Bewley, eds., New York: Cambridge University Press, Ch. 1, 1-58.

White, H., 1994, *Estimation, Inference and Specification Analysis*, New York: Cambridge University Press.

Wüthrich, K., 2015, A Comparison of Two Quantile Models with Endogeneity, mimeo, Bern.

Yu, P., 2014, Marginal Quantile Treatment Effect, mimeo, HKU.

Yu, P., 2015a, Testing Rank Preservation Under Unconfoundedness, mimeo, HKU.

Yu, P., 2015b, On the Interpretation of LATE and MTE, mimeo, HKU.

Yu, P., 2016a, Treatment Effects Estimators Under Misspecification II: Without the Monotonicity Assumption, work in progress, HKU.

Yu, P., 2016b, Testing Rank Similarity With and Without Covariates, work in progress, HKU.

Supplementary Material S.1

Proof of Theorem 1. The moment conditions for the IVE are

$$E\left[\begin{pmatrix} 1 \\ p(Z) \end{pmatrix}(Y-\mu_0^*-D\cdot\Delta^*)\right]=0,$$

which implies

$$\begin{aligned}
&\begin{pmatrix} \int\left[p\left(\mu_0^*+\Delta^*\right)+(1-p)\mu_0^*\right]dF_{p(Z)}(p) \\ \int\left[p^2\left(\mu_0^*+\Delta^*\right)+p(1-p)\mu_0^*\right]dF_{p(Z)}(p) \end{pmatrix} \\
=\ &\begin{pmatrix} \int\left[\int_0^p E\left[Y_1|U_D=u_D\right]du_D+\int_p^1 E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \\ \int p\left[\int_0^p E\left[Y_1|U_D=u_D\right]du_D+\int_p^1 E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \end{pmatrix},
\end{aligned}$$

or rewritten as

$$\begin{aligned}
&\begin{pmatrix} E\left[p(Z)\right] & 1-E\left[p(Z)\right] \\ E\left[p(Z)^2\right] & E\left[p(Z)\right]-E\left[p(Z)^2\right] \end{pmatrix}\begin{pmatrix} \mu_1^* \\ \mu_0^* \end{pmatrix} \\
=\ &\begin{pmatrix} \int\left[\int_0^p E\left[Y_1|U_D=u_D\right]du_D+\int_p^1 E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \\ \int p\left[\int_0^p E\left[Y_1|U_D=u_D\right]du_D+\int_p^1 E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \end{pmatrix},
\end{aligned}$$

where $\mu_1^*=\mu_0^*+\Delta^*$. It follows that

$$\begin{aligned}
\begin{pmatrix} \mu_1^* \\ \mu_0^* \end{pmatrix} &= \frac{1}{Var\left(p(Z)\right)}\begin{pmatrix} E\left[p(Z)^2\right]-E\left[p(Z)\right] & 1-E\left[p(Z)\right] \\ E\left[p(Z)^2\right] & -E\left[p(Z)\right] \end{pmatrix} \\
&\quad \begin{pmatrix} \int\left[\int_0^p E\left[Y_1|U_D=u_D\right]du_D+\int_p^1 E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \\ \int p\left[\int_0^p E\left[Y_1|U_D=u_D\right]du_D+\int_p^1 E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \end{pmatrix} \\
&= \frac{1}{Var\left(p(Z)\right)}\begin{pmatrix} \int_{\underline{p}}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]+\left(p-E\left[p(Z)\right]\right)\right] \\ \qquad\cdot\left[\int_0^p E\left[Y_1|U_D=u_D\right]du_D+\int_p^1 E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \\ \int_{\underline{p}}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]\right] \\ \qquad\cdot\left[\int_0^p E\left[Y_1|U_D=u_D\right]du_D+\int_p^1 E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \end{pmatrix},
\end{aligned}$$

so by Fubini's theorem,

$$\mu_1^*=\int_0^{\underline{p}} E\left[Y_1|U_D=u_D\right]\frac{1}{Var\left(p(Z)\right)}\int_{\underline{p}}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]+\left(p-E\left[p(Z)\right]\right)\right]dF_{p(Z)}(p)du_D$$

$$
\begin{aligned}
&+\int_{\underline{p}}^{\overline{p}} E\left[Y_1|U_D=u_D\right]\frac{1}{Var\left(p(Z)\right)}\int_{u_D}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]+\left(p-E\left[p(Z)\right]\right)\right]dF_{p(Z)}(p)du_D \\
&+\int_{\underline{p}}^{\overline{p}} E\left[Y_0|U_D=u_D\right]\frac{1}{Var\left(p(Z)\right)}\int_{\underline{p}}^{u_D}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]+\left(p-E\left[p(Z)\right]\right)\right]dF_{p(Z)}(p)du_D \\
&+\int_{\overline{p}}^{1} E\left[Y_0|U_D=u_D\right]\frac{1}{Var\left(p(Z)\right)}\int_{\underline{p}}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]+\left(p-E\left[p(Z)\right]\right)\right]dF_{p(Z)}(p)du_D \\
= \;& \int_0^{\underline{p}} E\left[Y_1|U_D=u_D\right]du_D+\int_{\underline{p}}^{\overline{p}}\left\{E\left[Y_1|U_D=u_D\right]h_1(u_D)+E\left[Y_0|U_D=u_D\right]\left(1-h_1(u_D)\right)\right\}du_D \\
&+\int_{\overline{p}}^{1} E\left[Y_0|U_D=u_D\right]du_D,
\end{aligned}
$$

and

$$
\begin{aligned}
\mu_0^* \;=\;& \int_{\overline{p}}^{1} E\left[Y_0|U_D=u_D\right]du_D+\int_{\underline{p}}^{\overline{p}}\left\{E\left[Y_1|U_D=u_D\right]h_0(u_D)+E\left[Y_0|U_D=u_D\right]\left(1-h_0(u_D)\right)\right\}du_D \\
&+\int_0^{\underline{p}} E\left[Y_1|U_D=u_D\right]du_D,
\end{aligned}
$$

where

$$
h_1(u_D)=\frac{1}{Var\left(p(Z)\right)}\int_{u_D}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]+\left(p-E\left[p(Z)\right]\right)\right]dF_{p(Z)}(p)
$$

and

$$
h_0(u_D)=\frac{1}{Var\left(p(Z)\right)}\int_{u_D}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]\right]dF_{p(Z)}(p).
$$

This means

$$
\Delta^*=\mu_1^*-\mu_0^*=\int_{\underline{p}}^{\overline{p}} MTE(u_D)h(u_D)du_D=\int_0^1 MTE(u_D)h(u_D)du_D,
$$

where

$$
h(u_D)=h_1(u_D)-h_0(u_D)=\frac{1}{Var\left(p(Z)\right)}\int_{u_D}^{\overline{p}}\left(p-E\left[p(Z)\right]\right)dF_{p(Z)}(p)
$$

satisfies $h(u_D)=0$ for $u_D\in[0,\underline{p}]\cup[\overline{p},1]$. ■

Proof of Proposition 1. First,

$$
\begin{aligned}
\int_{\underline{p}}^{\overline{p}} h_1(u_D)du_D \;=\;& \int_{\underline{p}}^{\overline{p}}\frac{1}{Var\left(p(Z)\right)}\int_{u_D}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]+\left(p-E\left[p(Z)\right]\right)\right]dF_{p(Z)}(p)du_D \qquad (17)\\
=\;& \frac{1}{Var\left(p(Z)\right)}\int_{\underline{p}}^{\overline{p}}\int_{\underline{p}}^{p}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]+\left(p-E\left[p(Z)\right]\right)\right]du_D dF_{p(Z)}(p) \\
=\;& \frac{1}{Var\left(p(Z)\right)}\int_{\underline{p}}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]+\left(p-E\left[p(Z)\right]\right)\right]\int_{\underline{p}}^{p}du_D dF_{p(Z)}(p) \\
=\;& \frac{1}{Var\left(p(Z)\right)}\int_{\underline{p}}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]+\left(p-E\left[p(Z)\right]\right)\right]\left(p-\underline{p}\right)dF_{p(Z)}(p) \\
=\;& 1-\underline{p},
\end{aligned}
$$

where the second equality is from Fubini's theorem. Similarly,

$$\begin{aligned}\int_{\underline{p}}^{\overline{p}} h_0(u_D)du_D &= \frac{1}{Var\left(p(Z)\right)}\int_{\underline{p}}^{\overline{p}}\int_{u_D}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]\right]dF_{p(Z)}(p)du_D \qquad (18)\\ &= \frac{1}{Var\left(p(Z)\right)}\int_{\underline{p}}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]\right]\int_{\underline{p}}^{p}du_D dF_{p(Z)}(p)\\ &= \frac{1}{Var\left(p(Z)\right)}\int_{\underline{p}}^{\overline{p}}\left[E\left[p(Z)^2\right]-pE\left[p(Z)\right]\right]\left(p-\underline{p}\right)dF_{p(Z)}(p)\\ &= -\underline{p}.\end{aligned}$$

So

$$\begin{aligned}\int_0^1 h_1(u_D)du_D &= \int_0^{\underline{p}} h_1(u_D)du_D + \int_{\underline{p}}^{\overline{p}} h_1(u_D)du_D + \int_{\overline{p}}^1 h_1(u_D)du_D\\ &= \underline{p}+1-\underline{p}+0=1,\end{aligned}$$

and

$$\begin{aligned}\int_0^1 h_0(u_D)du_D &= \int_0^{\underline{p}} h_0(u_D)du_D + \int_{\underline{p}}^{\overline{p}} h_0(u_D)du_D + \int_{\overline{p}}^1 h_0(u_D)du_D\\ &= \underline{p}-\underline{p}+0=0,\end{aligned}$$

which implies

$$\int_{\underline{p}}^{\overline{p}} h(u_D)du_D = \int_0^1 h_1(u_D)du_D - \int_0^1 h_0(u_D)du_D = 1,$$

as shown in HV.

Since

$$h_1'(u_D) = \left[E\left[p(Z)\right]-E\left[p(Z)^2\right]-\left(1-E\left[p(Z)\right]\right)u_D\right]\frac{f_{p(Z)}(u_D)}{Var\left(p(Z)\right)},$$

when $u_D < \frac{E[p(Z)]-E\left[p(Z)^2\right]}{1-E[p(Z)]}$, $h_1'(u_D) > 0$ and $u_D > \frac{E[p(Z)]-E\left[p(Z)^2\right]}{1-E[p(Z)]}$, $h'(u_D) < 0$. Similarly,

$$h_0'(u_D) = \left[E\left[p(Z)\right]u_D - E\left[p(Z)^2\right]\right]\frac{f_{p(Z)}(u_D)}{Var\left(p(Z)\right)},$$

so when $u_D < \frac{E\left[p(Z)^2\right]}{E[p(Z)]}$, h_0 is decreasing and when $u_D > \frac{E\left[p(Z)^2\right]}{Ep(Z)}$, h_0 is increasing. Finally,

$$h'\left(u_D\right) = \left(E\left[p(Z)\right]-u_D\right)\frac{f_{p(Z)}(u_D)}{Var\left(p(Z)\right)},$$

so $h\left(u_D\right)$ is increasing when $u_D < E\left[p(Z)\right]$ and decreasing when $u_D > E\left[p(Z)\right]$. ■

Proof of Proposition 3. From Theorem 1, when $E[U_1-U_0|U_D=u_D]=G_\alpha(u_D)$, $E\left[Y_1-Y_0|U_D=u_D\right]=$

$\Delta + G_\alpha(u_D)$, so

$$\begin{aligned}
\mu_1^*(\alpha) &= \mu_1 - \int_{\underline{p}}^{\overline{p}} (\Delta + G_\alpha(u_D))(1 - h_1(u_D))\, du_D - \int_{\overline{p}}^{1} (\Delta + G_\alpha(u_D))\, du_D \\
&= \mu_1 - \int_{\underline{p}}^{\overline{p}} G_\alpha(u_D)(1 - h_1(u_D))\, du_D - \int_{\overline{p}}^{1} G_\alpha(u_D) du_D, \\
\mu_0^*(\alpha) &= \mu_0 + \int_{0}^{\underline{p}} (\Delta + G_\alpha(u_D))\, du_D + \int_{\underline{p}}^{\overline{p}} (\Delta + G_\alpha(u_D))\, h_0(u_D) du_D \\
&= \mu_0 + \int_{0}^{\underline{p}} G_\alpha(u_D) du_D + \int_{\underline{p}}^{\overline{p}} G_\alpha(u_D) h_0(u_D) du_D, \\
\Delta^*(\alpha) &= \Delta + \int_0^1 G_\alpha(u_D) h(u_D) du_D = \Delta + \int_{\underline{p}}^{\overline{p}} G_\alpha(u_D) h(u_D) du_D.
\end{aligned}$$

The first equalities in the proposition follow straightforwardly from Assumption LA by exchanging the integration and taking limit. The second equality for $D_1^*(g)$ is from the normalization $\int_0^1 g(u_D) du_D = 0$ and $h_1(u_D) = 0$ for $u_D \in [\underline{p}, 1]$, and the second equalities for $D_0^*(g)$ and $D_\Delta^*(g)$ are from $h_0(u_D) = 0$ for $u_D \in [\underline{p}, 1]$ and $h(u_D) = 0$ for $u_D \in [0, \underline{p}] \cup [\overline{p}, 1]$ respectively. The third equality for $D_\Delta^*(g)$ is from $h(u_D) = h_1(u_D) - h_0(u_D)$. ■

Proof of Proposition 4. The moment conditions for the IVE when $J(Z)$ is used as the instrument are

$$E\left[\begin{pmatrix} 1 \\ J(Z) \end{pmatrix} (Y - \mu_{J0}^* - D \cdot \Delta_J^*)\right] = 0,$$

which implies

$$\begin{aligned}
&\begin{pmatrix} \int \left[p\left(\mu_{J0}^* + \Delta_J^*\right) + (1-p)\mu_{J0}^*\right] dF_{p(Z)}(p) \\ \int \left[pj\left(\mu_{J0}^* + \Delta_J^*\right) + j(1-p)\mu_{J0}^*\right] dF_{p(Z),J(Z)}(p,j) \end{pmatrix} \\
= &\begin{pmatrix} \int \left[\int_0^p E\left[Y_1|U_D = u_D\right] du_D + \int_p^1 E\left[Y_0|U_D = u_D\right] du_D\right] dF_{p(Z)}(p) \\ \int j \left[\int_0^p E\left[Y_1|U_D = u_D\right] du_D + \int_p^1 E\left[Y_0|U_D = u_D\right] du_D\right] dF_{p(Z),J(Z)}(p,j) \end{pmatrix},
\end{aligned}$$

or rewritten as

$$\begin{aligned}
&\begin{pmatrix} E\left[p(Z)\right] & 1 - E\left[p(Z)\right] \\ E\left[p(Z)J(Z)\right] & E\left[J(Z)\right] - E\left[p(Z)J(Z)\right] \end{pmatrix} \begin{pmatrix} \mu_{J1}^* \\ \mu_{J0}^* \end{pmatrix} \\
= &\begin{pmatrix} \int \left[\int_0^p E\left[Y_1|U_D = u_D\right] du_D + \int_p^1 E\left[Y_0|U_D = u_D\right] du_D\right] dF_{p(Z)}(p) \\ \int j \left[\int_0^p E\left[Y_1|U_D = u_D\right] du_D + \int_p^1 E\left[Y_0|U_D = u_D\right] du_D\right] dF_{p(Z),J(Z)}(p,j) \end{pmatrix},
\end{aligned}$$

where where $\mu_{J1}^{*}=\mu_{J0}^{*}+\Delta_{J}^{*}$. It follows that

$$\begin{aligned}
\begin{pmatrix} \mu_{J1}^{*} \\ \mu_{J0}^{*} \end{pmatrix} &= \frac{1}{Cov\left(J(Z),p(Z)\right)}\begin{pmatrix} E\left[p(Z)J(Z)\right]-E\left[J(Z)\right] & 1-E\left[p(Z)\right] \\ E\left[p(Z)J(Z)\right] & -E\left[p(Z)\right] \end{pmatrix} \\
&\quad \begin{pmatrix} \int\left[\int_{0}^{p}E\left[Y_1|U_D=u_D\right]du_D+\int_{p}^{1}E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \\ \int j\left[\int_{0}^{p}E\left[Y_1|U_D=u_D\right]du_D+\int_{p}^{1}E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z),J(Z)}(p,j) \end{pmatrix} \\
&= \frac{1}{Cov\left(J(Z),p(Z)\right)}\begin{pmatrix} \int_{\underline{p}}^{\overline{p}}\left[E\left[p(Z)J(Z)\right]-E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right]+\left(E\left[J(Z)|p(Z)=p\right]-E\left[J(Z)\right]\right)\right] \\ \quad\cdot\left[\int_{0}^{p}E\left[Y_1|U_D=u_D\right]du_D+\int_{p}^{1}E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \\ \int_{\underline{p}}^{\overline{p}}\left[E\left[p(Z)J(Z)\right]-E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right]\right] \\ \quad\cdot\left[\int_{0}^{p}E\left[Y_1|U_D=u_D\right]du_D+\int_{p}^{1}E\left[Y_0|U_D=u_D\right]du_D\right]dF_{p(Z)}(p) \end{pmatrix},
\end{aligned}$$

so by Fubini's theorem,

$$\begin{aligned}
\mu_{J1}^{*} &= \int_{0}^{\underline{p}}E\left[Y_1|U_D=u_D\right]\frac{1}{Cov\left(J(Z),p(Z)\right)}\int_{\underline{p}}^{\overline{p}}\begin{bmatrix} E\left[p(Z)J(Z)\right]-E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right] \\ +\left(E\left[J(Z)|p(Z)=p\right]-E\left[J(Z)\right]\right) \end{bmatrix}dF_{p(Z)}(p)du_D \\
&\quad +\int_{\underline{p}}^{\overline{p}}E\left[Y_1|U_D=u_D\right]\frac{1}{Cov\left(J(Z),p(Z)\right)}\int_{u_D}^{\overline{p}}\begin{bmatrix} E\left[p(Z)J(Z)\right]-E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right] \\ +\left(E\left[J(Z)|p(Z)=p\right]-E\left[J(Z)\right]\right) \end{bmatrix}dF_{p(Z)}(p)du_D \\
&\quad +\int_{\underline{p}}^{\overline{p}}E\left[Y_0|U_D=u_D\right]\frac{1}{Cov\left(J(Z),p(Z)\right)}\int_{\underline{p}}^{u_D}\begin{bmatrix} E\left[p(Z)J(Z)\right]-E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right] \\ +\left(E\left[J(Z)|p(Z)=p\right]-E\left[J(Z)\right]\right) \end{bmatrix}dF_{p(Z)}(p)du_D \\
&\quad +\int_{\overline{p}}^{1}E\left[Y_0|U_D=u_D\right]\frac{1}{Cov\left(J(Z),p(Z)\right)}\int_{\underline{p}}^{\overline{p}}\begin{bmatrix} E\left[p(Z)J(Z)\right]-E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right] \\ +\left(E\left[J(Z)|p(Z)=p\right]-E\left[J(Z)\right]\right) \end{bmatrix}dF_{p(Z)}(p)du_D \\
&= \int_{0}^{\underline{p}}E\left[Y_1|U_D=u_D\right]du_D+\int_{\underline{p}}^{\overline{p}}\left\{E\left[Y_1|U_D=u_D\right]h_{J1}(u_D)+E\left[Y_0|U_D=u_D\right]\left(1-h_{J1}(u_D)\right)\right\}du_D \\
&\quad +\int_{\overline{p}}^{1}E\left[Y_0|U_D=u_D\right]du_D,
\end{aligned}$$

and

$$\begin{aligned}
\mu_{J0}^{*} &= \int_{\overline{p}}^{1}E\left[Y_0|U_D=u_D\right]du_D+\int_{\underline{p}}^{\overline{p}}\left\{E\left[Y_1|U_D=u_D\right]h_{J0}(u_D)+E\left[Y_0|U_D=u_D\right]\left(1-h_{J0}(u_D)\right)\right\}du_D \\
&\quad +\int_{0}^{\underline{p}}E\left[Y_1|U_D=u_D\right]du_D,
\end{aligned}$$

where

$$h_{J1}(u_D)=\frac{1}{Cov\left(J(Z),p(Z)\right)}\int_{u_D}^{\overline{p}}\left[E\left[p(Z)J(Z)\right]-E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right]+\left(E\left[J(Z)|p(Z)=p\right]-E\left[J(Z)\right]\right)\right]dF_{p(Z)}(p)$$

and

$$h_{J0}(u_D)=\frac{1}{Cov\left(J(Z),p(Z)\right)}\int_{u_D}^{\overline{p}}\left[E\left[p(Z)J(Z)\right]-E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right]\right]dF_{p(Z)}(p).$$

This means

$$\Delta_J^* = \mu_{J1}^* - \mu_{J0}^* = \int_{\underline{p}}^{\overline{p}} MTE(u_D)h_J(u_D)du_D = \int_0^1 MTE(u_D)h_J(u_D)du_D,$$

where

$$h_J(u_D) = h_{J1}(u_D) - h_{J0}(u_D) = \frac{1}{Cov\left(J(Z), p(Z)\right)}\int_{u_D}^{\overline{p}} \left(E\left[J(Z)|p(Z)=p\right] - E\left[J(Z)\right]\right) dF_{p(Z)}(p)$$

satisfies $h_J(u_D) = 0$ for $u_D \in [0, \underline{p}] \cup [\overline{p}, 1]$. ■

Proof of Theorem 2. When use $p(Z)$ as the instrument, the moment conditions are

$$E\left[\begin{pmatrix} 1 \\ p(Z) \end{pmatrix}(\tau - 1(Y \le Q_0^*(\tau) + D\cdot\Delta^*(\tau)))\right] = 0.$$

Using a similar analysis as in the proof of Theorem 1, we have

$$\begin{pmatrix} \int \left[\int_0^p F_{Y_1|U_D}(Q_1^*(\tau)|u_D)du_D + \int_p^1 F_{Y_0|U_D}(Q_0^*(\tau)|u_D)du_D\right] dF_{p(Z)}(p) \\ \int p\left[\int_0^p F_{Y_1|U_D}(Q_1^*(\tau)|u_D)du_D + \int_p^1 F_{Y_0|U_D}(Q_0^*(\tau)|u_D)du_D\right] dF_{p(Z)}(p) \end{pmatrix} = \begin{pmatrix} \tau \\ \tau E[p(Z)] \end{pmatrix}$$

or

$$\begin{pmatrix} \int \left[\int_0^p F_{Y_1|U_D}(Q_1^*(\tau)|u_D)du_D + \int_p^1 F_{Y_0|U_D}(Q_0^*(\tau)|u_D)du_D\right] dF_{p(Z)}(p) \\ \int \frac{p}{E[p(Z)]}\left[\int_0^p F_{Y_1|U_D}(Q_1^*(\tau)|u_D)du_D + \int_p^1 F_{Y_0|U_D}(Q_0^*(\tau)|u_D)du_D\right] dF_{p(Z)}(p) \end{pmatrix} = \begin{pmatrix} \tau \\ \tau \end{pmatrix}.$$

where $Q_1^*(\tau) = Q_0^*(\tau) + \Delta^*(\tau)$. This implies

$$\begin{aligned} & \int p\left(\frac{p}{E\left[p(Z)\right]} - 1\right)\left[\frac{1}{p}\int_0^p F_{Y_1|U_D}(Q_1^*(\tau)|u_D)du_D\right] dF_{p(Z)}(p) \\ = & \int (1-p)\left(1 - \frac{p}{E\left[p(Z)\right]}\right)\left[\frac{1}{1-p}\int_p^1 F_{Y_0|U_D}(Q_0^*(\tau)|u_D)du_D\right] dF_{p(Z)}(p). \end{aligned}$$

Deviding both sides by $\int p\left(\frac{p}{E[p(Z)]} - 1\right) dF_{p(Z)}(p) = \frac{Var(p(Z))}{E[p(Z)]} > 0$, we have

$$\begin{aligned} \widetilde{F}_1\left(Q_1^*(\tau)\right) & \equiv \int p\frac{p - E\left[p(Z)\right]}{Var(p(Z))}\left[\frac{1}{p}\int_0^p F_{Y_1|U_D}(Q_1^*(\tau)|u_D)du_D\right] dF_{p(Z)}(p) \\ & = \int (1-p)\left(\frac{E\left[p(Z)\right] - p}{Var(p(Z))}\right)\left[\frac{1}{1-p}\int_p^1 F_{Y_0|U_D}(Q_0^*(\tau)|u_D)du_D\right] dF_{p(Z)}(p) \equiv \widetilde{F}_0\left(Q_0^*(\tau)\right), \end{aligned}$$

where $\int p\frac{p-E[p(Z)]}{Var(p(Z))} dF_{p(Z)}(p) = \int (1-p)\left(\frac{E[p(Z)]-p}{Var(p(Z))}\right) dF_{p(Z)}(p) = 1$,

$$\begin{aligned} & \int p\frac{p - E\left[p(Z)\right]}{Var(p(Z))}\left[\frac{1}{p}\int_0^p F_{Y_1|U_D}(Q_1^*(\tau)|u_D)du_D\right] dF_{p(Z)}(p) \\ = & \int F_{Y_1|U_D}(Q_1^*(\tau)|u_D)\int_{u_D}^1 \frac{p - E\left[p(Z)\right]}{Var(p(Z))} dF_{p(Z)}(p)du_D = \int F_{Y_1|U_D}(Q_1^*(\tau)|u_D)h_{p(Z)}^1\left(u_D\right)du_D, \end{aligned}$$

and

$$
\begin{aligned}
&\int (1-p)\left(\frac{E\left[p(Z)\right]-p}{Var(p(Z))}\right)\left[\frac{1}{1-p}\int_p^1 F_{Y_0|U_D}(Q_0^*(\tau)|u_D)du_D\right]dF_{p(Z)}(p) \\
=\ &\int F_{Y_0|U_D}(Q_0^*(\tau)|u_D)\int_0^{u_D}\frac{E\left[p(Z)\right]-p}{Var(p(Z))}dF_{p(Z)}(p)du_D = \int F_{Y_0|U_D}(Q_0^*(\tau)|u_D)h_{p(Z)}^0\left(u_D\right)du_D
\end{aligned}
$$

with

$$
h_{p(Z)}^1\left(u_D\right) = \int_{u_D}^1 \frac{p-E\left[p(Z)\right]}{Var(p(Z))}dF_{p(Z)}(p) \text{ and } h_{p(Z)}^0\left(u_D\right) = \int_0^{u_D}\frac{E\left[p(Z)\right]-p}{Var(p(Z))}dF_{p(Z)}(p).
$$

Note that

$$
h_{p(Z)}^1\left(u_D\right) - h_{p(Z)}^0\left(u_D\right) = 0,
$$

and they are exactly the same as $h(u_D)$ in Theorem 1.

From $\widetilde{F}_1\left(Q_1^*(\tau)\right) = \widetilde{F}_0\left(Q_0^*(\tau)\right)$, $Q_0^*(\tau) = \widetilde{F}_0^{-1}\widetilde{F}_1\left(Q_1^*(\tau)\right)$. Substituting in $\int\left[\int_0^p F_{Y_1|U_D}(Q_1^*(\tau)|u_D)du_D + \int_p^1 F_{Y_0|U_D}(Q_0^*(\tau)|u_D)du_D\right]dF_{p(Z)}(p) = \tau$, we have

$$
\begin{aligned}
F_1^*\left(Q_1^*(\tau)\right) &\equiv \int\left[\int_0^p F_{Y_1|U_D}(Q_1^*(\tau)|u_D)du_D + \int_p^1 F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(Q_1^*(\tau)\right)|u_D)du_D\right]dF_{p(Z)}(p) \\
&= \int_0^{\underline{p}} F_{Y_1|U_D}(Q_1^*(\tau)|u_D)du_D + \int_{\underline{p}}^{\overline{p}}\left[\begin{array}{c} F_{Y_1|U_D}(Q_1^*(\tau)|u_D)(1-F_{p(Z)}(u_D)) \\ +F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(Q_1^*(\tau)\right)|u_D)F_{p(Z)}(u_D) \end{array}\right]du_D \\
&\quad + \int_{\overline{p}}^1 F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(Q_1^*(\tau)\right)|u_D)du_D \\
&= \tau,
\end{aligned}
$$

where the second equality is from Fubini's theorem. Since this is valid for any τ, the IV-QRE is estimating $F_1^*(y_1)$ by the formula stated in the theorem. Similarly, we can prove the result for $F_0^*(y_0)$. ■

Proof of Proposition 7. First consider F_1^*. The key is to study the path derivative of $F_{Y_0|U_D}(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)|u_D)$ along the path $\mathcal{F}_\alpha^*$. It is better to use $F_{U_1|U_D}$ as a benchmark and express $F_{U_0|U_D}^\alpha = F_{U_1|U_D} - F^\alpha$. Specifically,

$$
\begin{aligned}
&\lim_{\alpha\to 0}\frac{F_{Y_0|U_D}^\alpha(\widetilde{F}_0^{\alpha-1}\widetilde{F}_1\left(y_1\right)|u_D) - F_{Y_1|U_D}(y_1|u_D)}{\alpha} \\
=\ &\lim_{\alpha\to 0}\frac{F_{U_0|U_D}^\alpha\left(F_0\left(\widetilde{F}_0^{\alpha-1}\widetilde{F}_1\left(y_1\right)\right)|u_D\right) - F_{U_1|U_D}\left(F_1(y_1)|u_D\right)}{\alpha} \\
=\ &\lim_{\alpha\to 0}\frac{\begin{array}{c} F_{U_0|U_D}^\alpha\left(F_0\left(\widetilde{F}_0^{\alpha-1}\widetilde{F}_1\left(y_1\right)\right)|u_D\right) - F_{U_1|U_D}\left(F_0\left(\widetilde{F}_0^{\alpha-1}\widetilde{F}_1\left(y_1\right)\right)|u_D\right) \\ +F_{U_1|U_D}\left(F_0\left(\widetilde{F}_0^{\alpha-1}\widetilde{F}_1\left(y_1\right)\right)|u_D\right) - F_{U_1|U_D}\left(F_0\left(\widetilde{F}_0^{-1}\widetilde{F}_1\left(y_1\right)\right)|u_D\right) \end{array}}{\alpha} \\
=\ &-g(F_1\left(y_1\right)|u_D) + f_{U_1|U_D}\left(F_1\left(y_1\right)|u_D\right)\frac{f_0\left(F_0^{-1}F_1\left(y_1\right)\right)}{\widetilde{f}_0\left(F_0^{-1}F_1\left(y_1\right)\right)}\int_{\underline{p}}^{\overline{p}} g(F_1\left(y_1\right)|u_D)h\left(u_D\right)du_D \\
=\ &\frac{f_{Y_0|U_D}\left(F_0^{-1}\left(F_1\left(y_1\right)\right)|u_D\right)}{\widetilde{f}_0\left(F_0^{-1}F_1\left(y_1\right)\right)}\int_{\underline{p}}^{\overline{p}} g(F_1\left(y_1\right)|u_D)h\left(u_D\right)du_D - g(F_1\left(y_1\right)|u_D)
\end{aligned}
$$

where $F_{Y_0|U_D}^\alpha(\cdot|u_D) = F_{U_0|U_D}^\alpha\left(F_0(\cdot)|u_D\right)$, $\widetilde{F}_0^\alpha\left(\cdot\right) = \int_{\underline{p}}^{\overline{p}} F_{Y_0|U_D}^\alpha(\cdot|u_D)h\left(u_D\right)du_D$, $f_{U_1|U_D}\left(\cdot|u_D\right)$ is the pdf of $F_{U_1|U_D}\left(\cdot|u_D\right)$, $f_{Y_1|U_D}\left(\cdot|u_D\right) = f_{U_1|U_D}\left(F_1\left(\cdot\right)|u_D\right)\cdot f_1(\cdot)$ is the pdf of $F_{Y_1|U_D}\left(\cdot|u_D\right)$, $f_d(\cdot) = \int_0^1 f_{Y_d|U_D}\left(\cdot|u_D\right)du_D$

is the pdf of $F_d(\cdot)$, $\widetilde{f}_0\left(\cdot\right) = \int_{\underline{p}}^{\overline{p}} f_{Y_0|U_D}(\cdot|u_D)h\left(u_D\right)du_D$ is the pdf of $\widetilde{F}_0(\cdot)$, the second to last equality is from Assumption LF, and the last equality is from

$$f_{U_1|U_D}\left(F_1\left(y_1\right)|u_D\right) = f_{U_0|U_D}\left(F_1\left(y_1\right)|u_D\right) = f_{U_0|U_D}\left(F_0(F_0^{-1}\left(F_1\left(y_1\right)\right))|u_D\right) = \frac{f_{Y_0|U_D}\left(F_0^{-1}\left(F_1\left(y_1\right)\right)|u_D\right)}{f_0\left(F_0^{-1}\left(F_1\left(y_1\right)\right)\right)}$$

as $\alpha = 0$. As a result,

$$\begin{aligned}
& \lim_{\alpha\to 0}\frac{F_1^*(y_1;\alpha) - F_1(y_1)}{\alpha} \\
= & \int_{\underline{p}}^{\overline{p}}\left[\frac{f_{Y_0|U_D}\left(F_0^{-1}\left(F_1\left(y_1\right)\right)|u_D\right)}{\widetilde{f}_0\left(F_0^{-1}F_1\left(y_1\right)\right)}\int_{\underline{p}}^{\overline{p}} g(F_1\left(y_1\right)|u_D)h\left(u_D\right)du_D - g(F_1\left(y_1\right)|u_D)\right]F_{p(Z)}(u_D)du_D \\
& + \int_{\overline{p}}^{1}\left[\frac{f_{Y_0|U_D}\left(F_0^{-1}\left(F_1\left(y_1\right)\right)|u_D\right)}{\widetilde{f}_0\left(F_0^{-1}F_1\left(y_1\right)\right)}\int_{\underline{p}}^{\overline{p}} g(F_1\left(y_1\right)|u_D)h\left(u_D\right)du_D - g(F_1\left(y_1\right)|u_D)\right]du_D
\end{aligned}$$

which is the formula in the proposition.

Next consider F_0^*. The key is to study the path derivative of $F_{Y_1|U_D}(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)|u_D)$ along the path $\mathcal{F}_\alpha^*$. It is better to use $F_{U_0|U_D}$ as a benchmark this time and express $F_{U_1|U_D}^\alpha = F_{U_0|U_D} + F^\alpha$. Specifically,

$$\begin{aligned}
& \lim_{\alpha\to 0}\frac{F_{Y_1|U_D}^\alpha(\widetilde{F}_1^{\alpha-1}\widetilde{F}_0\left(y_0\right)|u_D) - F_{Y_0|U_D}(y_0|u_D)}{\alpha} \\
= & \lim_{\alpha\to 0}\frac{F_{U_1|U_D}^\alpha(F_1\left(\widetilde{F}_1^{\alpha-1}\widetilde{F}_0\left(y_0\right)\right)|u_D) - F_{Y_0|U_D}(y_0|u_D)}{\alpha} \\
= & \lim_{\alpha\to 0}\frac{\begin{array}{c}F_{U_1|U_D}^\alpha(F_1\left(\widetilde{F}_1^{\alpha-1}\widetilde{F}_0\left(y_0\right)\right)|u_D) - F_{U_0|U_D}(F_1\left(\widetilde{F}_1^{\alpha-1}\widetilde{F}_0\left(y_0\right)\right)|u_D) \\ +F_{U_0|U_D}(F_1\left(\widetilde{F}_1^{\alpha-1}\widetilde{F}_0\left(y_0\right)\right)|u_D) - F_{U_0|U_D}(F_1\left(\widetilde{F}_1^{-1}\widetilde{F}_0\left(y_0\right)\right)|u_D)\end{array}}{\alpha} \\
= & g(F_0\left(y_0\right)|u_D) - f_{U_0|U_D}\left(F_0\left(y_0\right)|u_D\right)\frac{f_1\left(F_1^{-1}F_0\left(y_0\right)\right)}{\widetilde{f}_1\left(F_1^{-1}F_0\left(y_0\right)\right)}\int_{\underline{p}}^{\overline{p}} g(F_0\left(y_0\right)|u_D)h\left(u_D\right)du_D \\
= & g(F_0\left(y_0\right)|u_D) - \frac{f_{Y_1|U_D}\left(F_1^{-1}\left(F_0\left(y_0\right)\right)|u_D\right)}{\widetilde{f}_1\left(F_1^{-1}F_0\left(y_0\right)\right)}\int_{\underline{p}}^{\overline{p}} g(F_0\left(y_0\right)|u_D)h\left(u_D\right)du_D,
\end{aligned}$$

where $F_{Y_1|U_D}^\alpha(\cdot|u_D) = F_{U_1|U_D}^\alpha\left(F_1(\cdot)|u_D\right)$, $\widetilde{F}_1^\alpha\left(\cdot\right) = \int_{\underline{p}}^{\overline{p}} F_{Y_1|U_D}^\alpha(\cdot|u_D)h\left(u_D\right)du_D$, $f_{U_0|U_D}\left(\cdot|u_D\right)$ is the pdf of $F_{U_0|U_D}\left(\cdot|u_D\right)$, $f_{Y_0|U_D}\left(\cdot|u_D\right) = f_{U_0|U_D}\left(F_0\left(\cdot\right)|u_D\right)f_0(\cdot)$ is the pdf of $F_{Y_0|U_D}\left(\cdot|u_D\right)$, $f_d(\cdot) = \int_0^1 f_{Y_d|U_D}\left(\cdot|u_D\right)du_D$ is the pdf of $F_d(\cdot)$, $\widetilde{f}_1\left(\cdot\right) = \int_{\underline{p}}^{\overline{p}} f_{Y_1|U_D}(\cdot|u_D)h\left(u_D\right)du_D$ is the pdf of $\widetilde{F}_1(\cdot)$, and the last equality is from

$$f_{U_0|U_D}\left(F_0\left(y_0\right)|u_D\right) = f_{U_1|U_D}\left(F_0\left(y_0\right)|u_D\right) = f_{U_1|U_D}\left(F_1(F_1^{-1}\left(F_0\left(y_0\right)\right))|u_D\right) = \frac{f_{Y_1|U_D}\left(F_1^{-1}\left(F_0\left(y_0\right)\right)|u_D\right)}{f_1\left(F_1^{-1}\left(F_0\left(y_0\right)\right)\right)}$$

as $\alpha = 0$. As a result,

$$\begin{aligned}
&\lim_{\alpha\to 0}\frac{F_0^*(y_0;\alpha)-F_0(y_0)}{\alpha}\\
=\;& \int_0^{\underline{p}}\left[g(F_0\left(y_0\right)|u_D)-\frac{f_{Y_1|U_D}\left(F_1^{-1}\left(F_0\left(y_0\right)\right)|u_D\right)}{\widetilde{f}_1\left(F_1^{-1}F_0\left(y_0\right)\right)}\int_{\underline{p}}^{\overline{p}}g(F_0\left(y_0\right)|u_D)h\left(u_D\right)du_D\right]du_D\\
&+\int_{\underline{p}}^{\overline{p}}\left[g(F_0\left(y_0\right)|u_D)-\frac{f_{Y_1|U_D}\left(F_1^{-1}\left(F_0\left(y_0\right)\right)|u_D\right)}{\widetilde{f}_1\left(F_1^{-1}F_0\left(y_0\right)\right)}\int_{\underline{p}}^{\overline{p}}g(F_0\left(y_0\right)|u_D)h\left(u_D\right)du_D\right](1-F_{p(Z)}(u_D))du_D,
\end{aligned}$$

which is the formula in the proposition.

Finally consider $\Delta^*\left(\tau\right)$. Since $\Delta^*\left(\tau;\alpha\right) = F_1^{*-1}\left(\tau;\alpha\right) - F_0^{*-1}\left(\tau;\alpha\right)$, the result in the proposition is straightforward by the relationship between the derivatives of the cdf and the quantile. ■

Proof of Proposition 8. The moment conditions for the IV-QRE when $J(Z)$ is used as the instrument are

$$E\left[\left(\begin{array}{c}1\\ J(Z)\end{array}\right)(\tau-1(Y\leq Q_{J0}^*(\tau)+D\cdot\Delta_J^*(\tau)))\right]=0,$$

which implies

$$\left(\begin{array}{c}\int\left[\int_0^p F_{Y_1|U_D}(Q_{J1}^*(\tau)|u_D)du_D+\int_p^1 F_{Y_0|U_D}(Q_{J0}^*(\tau)|u_D)du_D\right]dF_{p(Z)}(p)\\ \int\frac{j}{E[J(Z)]}\left[\int_0^p F_{Y_1|U_D}(Q_{J1}^*(\tau)|u_D)du_D+\int_p^1 F_{Y_0|U_D}(Q_{J0}^*(\tau)|u_D)du_D\right]dF_{p(Z),J(Z)}(p,j)\end{array}\right)=\left(\begin{array}{c}\tau\\ \tau\end{array}\right).$$

So

$$\begin{aligned}
&\int p\left(\frac{j}{E\left[J(Z)\right]}-1\right)\left[\frac{1}{p}\int_0^p F_{Y_1|U_D}(Q_{J1}^*(\tau)|u_D)du_D\right]dF_{p(Z),J(Z)}(p,j)\\
=\;&\int(1-p)\left(1-\frac{j}{E\left[J(Z)\right]}\right)\left[\frac{1}{1-p}\int_p^1 F_{Y_0|U_D}(Q_{J0}^*(\tau)|u_D)du_D\right]dF_{p(Z),J(Z)}(p,j).
\end{aligned}$$

Deviding both sides by $\int p\left(\frac{j}{E[J(Z)]}-1\right)dF_{p(Z),J(Z)}(p,j)=\frac{Cov(p(Z),J(Z))}{E[J(Z)]}$, we have

$$\begin{aligned}
\widetilde{F}_{J1}\left(Q_{J1}^*(\tau)\right)\;\equiv\;&\int p\frac{j-E\left[J(Z)\right]}{Cov\left(J(Z),p(Z)\right)}\left[\frac{1}{p}\int_0^p F_{Y_1|U_D}(Q_{J1}^*(\tau)|u_D)du_D\right]dF_{p(Z),J(Z)}(p,j)\\
=\;&\int(1-p)\left(\frac{E\left[J(Z)\right]-j}{Cov\left(J(Z),p(Z)\right)}\right)\left[\frac{1}{1-p}\int_p^1 F_{Y_0|U_D}(Q_{J0}^*(\tau)|u_D)du_D\right]dF_{p(Z),J(Z)}(p,j)\equiv\widetilde{F}_{J0}\left(Q_{J0}^*(\tau)\right),
\end{aligned}$$

By similar arguments as in the proof of Theorem 2, we can show the results in the proposition with

$$\begin{aligned}
\widetilde{F}_{J1}\left(y_1\right)\;\equiv\;&\int\frac{j-E\left[J(Z)\right]}{Cov\left(J(Z),p(Z)\right)}\left[\int_0^p F_{Y_1|U_D}(y_1|u_D)du_D\right]dF_{p(Z),J(Z)}(p,j)\\
=\;&\int F_{Y_1|U_D}(y_1|u_D)h_J\left(u_D\right)du_D,\\
\widetilde{F}_{J0}\left(y_0\right)\;\equiv\;&\int\left(\frac{E\left[J(Z)\right]-j}{Cov\left(J(Z),p(Z)\right)}\right)\left[\int_p^1 F_{Y_0|U_D}(Q_{J0}^*(\tau)|u_D)du_D\right]dF_{p(Z),J(Z)}(p,j)\\
=\;&\int F_{Y_0|U_D}(y_1|u_D)h_J\left(u_D\right)du_D,
\end{aligned}$$

where

$$\begin{aligned} h_J(u_D) &= \int_{-\infty}^{\infty} \frac{j - E[J(Z)]}{Cov(J(Z), p(Z))} \int_{u_D}^{1} dF_{p(Z)|J(Z)}(p|j) dF_{J(Z)}(j) \\ &= \int_{u_D}^{1} \int_{-\infty}^{\infty} \frac{j - E[J(Z)]}{Cov(J(Z), p(Z))} dF_{J(Z)|p(Z)}(j|p) dF_{p(Z)}(p) \\ &= \int_{u_D}^{1} \frac{E[J(Z)|p(Z) = p] - E[J(Z)]}{Cov(J(Z), p(Z))} dF_{p(Z)}(p). \end{aligned}$$

■

Proof of Theorem 3. From the discussion in Section 4.2, $\int_0^{\overline{p}} F_{Y_1|U_D}(y|u_D)du_D$ and $\int_{\underline{p}}^1 F_{Y_0|U_D}(y|u_D)du_D$ can be identified, but the distribution of (D, Y, Z) contains no information on $\int_{\overline{p}}^1 F_{Y_1|U_D}(y|u_D)du_D$ and $\int_0^{\underline{p}} F_{Y_0|U_D}(y|u_D)du_D$. Nevertheless, note that

$$\begin{aligned} P(Y \leq y|p(Z) = \overline{p}, D = 1)\overline{p} \leq P(Y_1 \leq y) = \int_0^{\overline{p}} F_{Y_1|U_D}(y|u_D)du_D + \int_{\overline{p}}^1 F_{Y_1|U_D}(y|u_D)du_D \\ \leq P(Y \leq y|p(Z) = \overline{p}, D = 1)\overline{p} + (1 - \overline{p}), \end{aligned} \tag{19}$$

and

$$\begin{aligned} P\left(Y \leq y|p(Z) = \underline{p}, D = 0\right)\left(1 - \underline{p}\right) \leq P(Y_0 \leq y) = \int_{\underline{p}}^1 F_{Y_1|U_D}(y|u_D)du_D + \int_0^{\underline{p}} F_{Y_0|U_D}(y|u_D)du_D \\ \leq P\left(Y \leq y|p(Z) = \underline{p}, D = 0\right)\left(1 - \underline{p}\right) + \underline{p}. \end{aligned}$$

We concentrate on bounding $Q_1(\tau)$ since the results for $Q_0(\tau)$ can be similarly derived. We divide the proof into five steps. The first four steps are similar to the proof of Proposition 2 in Manski (1994).

Step 1: $\overline{I}_1(\tau)$ is an upper bound for $Q_1(\tau)$.

By (19),

$$P(Y \leq y|p(Z) = \overline{p}, D = 1)\overline{p} \geq \tau \Longrightarrow P(Y_1 \leq y) \geq \tau. \tag{20}$$

The premise of (20) is empty if that $\overline{p} < \tau$. Suppose that $\overline{p} \geq \tau$. Then the definition of $\overline{I}_1(\tau)$ states that

$$\overline{I}_1(\tau) \equiv \min\left\{t : P(Y \leq t|p(Z) = \overline{p}, D = 1) \geq \frac{\tau}{\overline{p}}\right\}.$$

It follows that $P(Y_1 \leq \overline{I}_1(\tau)) \geq \tau$. Hence $Q_1(\tau) \leq \overline{I}_1(\tau)$.

Step 2: $\underline{I}_1(\tau)$ is a lower bound for $Q_1(\tau)$.

By (19),

$$P(Y \leq y|p(Z) = \overline{p}, D = 1)\overline{p} + (1 - \overline{p}) < \tau \Longrightarrow P(Y_1 \leq y) < \tau,$$

which can be rewritten as

$$P(Y \leq y|p(Z) = \overline{p}, D = 1) < 1 - \frac{1 - \tau}{\overline{p}} \Longrightarrow P(Y_1 \leq y) < \tau. \tag{21}$$

The premise of (21) is empty if $\overline{p} \leq 1-\tau$. Suppose that $\overline{p} > 1-\tau$. Then the definition of $\underline{I}_1(\tau)$ states that

$$\underline{I}_1(\tau) \equiv \min\left\{t : P\left(Y \leq t|p(Z) = \overline{p}, D = 1\right) \geq 1 - \frac{1-\tau}{\overline{p}}\right\}.$$

It follows that, for all $\eta > 0$, $P\left(Y_1 \leq \underline{I}_1(\tau) - \eta\right) < \tau$. Hence $Q_1(\tau) \geq \underline{I}_1(\tau)$.

Step 3: $\overline{I}_1(\tau)$ is the least upper bound for $Q_1(\tau)$.

First let $\overline{p} \geq \tau$. For any $\lambda > 0$,

$$P\left(Y_1 \leq \overline{I}_1(\tau) - \lambda\right) = P\left(Y \leq \overline{I}_1(\tau) - \lambda|p(Z) = \overline{p}, D = 1\right)\overline{p} + \int_{\overline{p}}^{1} F_{Y_1|U_D}(\overline{I}_1(\tau) - \lambda|u_D)du_D.$$

Suppose $F_{Y_1|U_D}(\overline{I}_1(\tau) - \lambda|u_D) = 0$ for $u_D \in (\overline{p}, 1]$, as is possible in the absence of other information. Then the definition of $\overline{I}_1(\tau)$ implies that

$$P\left(Y_1 \leq \overline{I}_1(\tau) - \lambda\right) = P\left(Y \leq \overline{I}_1(\tau) - \lambda|p(Z) = \overline{p}, D = 1\right)\overline{p} < \tau.$$

Hence $Q_1(\tau) > \overline{I}_1(\tau) - \lambda$. Now let $\overline{p} < \tau$. For any $t < \overline{y}_1$,

$$P(Y_1 \leq t) = P\left(Y \leq t|p(Z) = \overline{p}, D = 1\right)\overline{p} + \int_{\overline{p}}^{1} F_{Y_1|U_D}(t|u_D)du_D. \tag{22}$$

Suppose that $F_{Y_1|U_D}(t|u_D) = 0$ for $u_D \in (\overline{p}, 1]$. Then

$$P(Y_1 \leq t) = P\left(Y \leq t|p(Z) = \overline{p}, D = 1\right)\overline{p} < \tau.$$

Hence $Q_1(\tau) > t$.

Step 4: $\underline{I}_1(\tau)$ is the greatest lower bound for $Q_1(\tau)$.

First let $\overline{p} > 1-\tau$. For any $\lambda > 0$,

$$P\left(Y_1 \leq \underline{I}_1(\tau) + \lambda\right) = P\left(Y \leq \underline{I}_1(\tau) + \lambda|p(Z) = \overline{p}, D = 1\right)\overline{p} + \int_{\overline{p}}^{1} F_{Y_1|U_D}(\underline{I}_1(\tau) + \lambda|u_D)du_D.$$

Suppose that $F_{Y_1|U_D}(\underline{I}_1(\tau) + \lambda|u_D) = 1$ for $u_D \in (\overline{p}, 1]$, as is possible in the absence of other information. Then the definition of $\underline{I}_1(\tau)$ implies that

$$P\left(Y_1 \leq \underline{I}_1(\tau) + \lambda\right) = P\left(Y \leq \underline{I}_1(\tau) + \lambda|p(Z) = \overline{p}, D = 1\right)\overline{p} + (1-\overline{p}) \geq \tau.$$

Hence $Q_1(\tau) \leq \underline{I}_1(\tau) + \lambda$. Now, let $\overline{p} \leq 1-\tau$. Let $t > \underline{y}_1$ and suppose that $F_{Y_1|U_D}(t|u_D) = 1$ for $u_D \in (\overline{p}, 1]$. Then by (22),

$$P(Y_1 \leq t) = P\left(Y \leq t|p(Z) = \overline{p}, D = 1\right)\overline{p} + (1-\overline{p}) \geq \tau.$$

Hence $Q_1(\tau) \leq t$.

Step 5: $\overline{I}_1(\tau) = \underline{I}_1(\tau)$ if $\overline{p} = 1$; $\overline{p} = 1$ if $\overline{I}_1(\tau) = \underline{I}_1(\tau)$ when $Y|\left(p(Z) = \overline{p}, D = 1\right)$ is continuously distributed with a positive density on $(\underline{y}_1, \overline{y}_1)$.

Suppose $\overline{p} = 1$; then $\underline{I}_1(\tau) = Q_{Y|p(Z),D}(\tau|1, 1) = \overline{I}_1(\tau)$.

Fix $\tau \in (0, 1/2]$. When $0 \leq \overline{p} < \tau$, $\underline{I}_1(\tau) = \underline{y}_1 < \overline{y}_1 = \overline{I}_1(\tau)$. When $1 > \overline{p} > 1-\tau$, $\underline{I}_1(\tau) = Q_{Y|p(Z),D}\left(1 - \frac{1-\tau}{\overline{p}}\middle|\overline{p}, 1\right) < Q_{Y|p(Z),D}\left(\frac{\tau}{\overline{p}}\middle|\overline{p}, 1\right) = \overline{I}_1(\tau)$. When $1-\tau \geq \overline{p} \geq \tau$, $\underline{I}_1(\tau) = \underline{y}_1 < Q_{Y|p(Z),D}\left(\frac{\tau}{\overline{p}}\middle|\overline{p}, 1\right)$. So when $\tau \leq 1/2$, $Q_1(\tau)$ cannot be point identified unless $\overline{p} = 1$. Similarly, when $1 > \tau > 1/2$, $Q_1(\tau)$ cannot be point identified unless $\overline{p} = 1$. ■

Proof of Theorem 4. The moment conditions for $(\overline{\mu}_1, \overline{\mu}_0)$ are

$$E\left[\begin{pmatrix} 1 \\ D \end{pmatrix}\left(Y - \overline{\mu}_0 - D \cdot \overline{\Delta}\right)\right] = 0,$$

where $\overline{\Delta} = \overline{\mu}_1 - \overline{\mu}_0$, i.e.,

$$\begin{aligned}
\begin{pmatrix} \overline{\mu}_1 \\ \overline{\mu}_0 \end{pmatrix} &= \begin{pmatrix} E\left[\frac{DY}{E[D]}\right] \\ E\left[\frac{(1-D)Y}{E[1-D]}\right] \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{E[D]}\int_{\underline{p}}^{\overline{p}}\left[\int_0^p E\left[Y_1|U_D = u_D\right]du_D\right]dF_{p(Z)}(p) \\ \frac{1}{E[1-D]}\int_{\underline{p}}^{\overline{p}}\left[\int_p^1 E\left[Y_0|U_D = u_D\right]du_D\right]dF_{p(Z)}(p) \end{pmatrix} \\
&= \begin{pmatrix} \int_{\underline{p}}^{\overline{p}} \frac{p}{E[D]}\left[\frac{1}{p}\int_0^p E\left[Y_1|U_D = u_D\right]du_D\right]dF_{p(Z)}(p) \\ \int_{\underline{p}}^{\overline{p}} \frac{1-p}{E[1-D]}\left[\frac{1}{1-p}\int_p^1 E\left[Y_0|U_D = u_D\right]du_D\right]dF_{p(Z)}(p) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{E[D]}\left[\int_0^{\underline{p}} E\left[Y_1|U_D = u_D\right]\int_{\underline{p}}^{\overline{p}} dF_{p(Z)}(p)du_D + \int_{\underline{p}}^{\overline{p}} E\left[Y_1|U_D = u_D\right]\int_{u_D}^{\overline{p}} dF_{p(Z)}(p)du_D\right] \\ \frac{1}{E[1-D]}\left[\int_{\underline{p}}^{\overline{p}} E\left[Y_0|U_D = u_D\right]\int_{\underline{p}}^{u_D} dF_{p(Z)}(p)du_D + \int_{\overline{p}}^1 E\left[Y_0|U_D = u_D\right]\int_{\underline{p}}^{\overline{p}} dF_{p(Z)}(p)du_D\right] \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{E[p(Z)]}\left[\int_0^{\underline{p}} E\left[Y_1|U_D = u_D\right]du_D + \int_{\underline{p}}^{\overline{p}} E\left[Y_1|U_D = u_D\right]\left(1 - F_{p(Z)}(u_D)\right)du_D\right] \\ \frac{1}{E[1-p(Z)]}\left[\int_{\underline{p}}^{\overline{p}} E\left[Y_0|U_D = u_D\right]F_{p(Z)}(u_D) + \int_{\overline{p}}^1 E\left[Y_0|U_D = u_D\right]du_D\right] \end{pmatrix}.
\end{aligned}$$

Given $\overline{\mu}_1$ and $\overline{\mu}_0$, $\overline{\Delta} = \overline{\mu}_1 - \overline{\mu}_0$.

To derive the weight for $\overline{\Delta}$ in HV, note that

$$\begin{aligned}
\overline{\Delta} - \Delta &= \int_0^1 E\left[U_1 - U_0|U_D = u_D\right]\frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]}du_D + \int_0^1 E\left[U_0|U_D = u_D\right]\left(\frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} - \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]}\right)du_D \\
&= \int_0^1 MTE(u_D)\frac{E\left[U_1|U_D = u_D\right]\omega_1(u_D) - E\left[U_0|U_D = u_D\right]\omega_0(u_D)}{MTE(u_D)}du_D.
\end{aligned}$$

where

$$\omega_1(u_D) = \frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]}, \omega_0(u_D) = \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]}.$$

Since $\Delta = \int_0^1 MTE(u_D)du_D$, $\overline{\Delta} = \int_0^1 MTE(u_D)\omega(u_D)du_D$ with $\omega(u_D) = 1 + \frac{E[U_1|U_D = u_D]\omega_1(u_D) - E[U_0|U_D = u_D]\omega_0(u_D)}{MTE(u_D)}$.
■

Proof of Theorem 5. The moment conditions to identify $\left(\overline{Q}_0(\tau), \overline{\Delta}(\tau)\right)$ are

$$E\left[\begin{pmatrix} 1 \\ D \end{pmatrix}\left(\tau - 1(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau))\right)\right] = 0,$$

i.e.,

$$\begin{pmatrix} \tau \\ \tau E[D] \end{pmatrix} = \begin{pmatrix} P\left(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau)\right) \\ E[D \cdot 1(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau))] \end{pmatrix}.$$

We calculate these moment conditions in our framework. First,

$$P\left(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau)\right) = \int P\left(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau) | p(Z) = p\right) dF_{p(Z)}(p),$$

where

$$\begin{aligned} & P\left(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau) | p(Z) = p\right) \\ &= P\left(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau) | p(Z) = p, D = 1\right) P(D = 1 | p(Z) = p) \\ &\quad + P\left(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau) | p(Z) = p, D = 0\right) P(D = 0 | p(Z) = p) \\ &= P\left(Y_1 \leq \overline{Q}_1(\tau) | U_D \leq p\right) p + P\left(Y_0 \leq \overline{Q}_0(\tau) | U_D > p\right)(1 - p) \\ &= \int_0^p F_{Y_1|U_D}(\overline{Q}_1(\tau)|u_D) du_D + \int_p^1 F_{Y_0|U_D}(\overline{Q}_0(\tau)|u_D) du_D \end{aligned}$$

with $\overline{Q}_1(\tau) = \overline{Q}_0(\tau) + \overline{\Delta}(\tau)$. Second,

$$E[D \cdot 1(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau))] = \int E[D \cdot 1(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau)) | p(Z) = p] dF_{p(Z)}(p),$$

where

$$\begin{aligned} & E[D \cdot 1(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau)) | p(Z) = p] \\ &= E[D \cdot 1(Y \leq \overline{Q}_0(\tau) + D \cdot \overline{\Delta}(\tau)) | p(Z) = p, D = 1] P(D = 1 | p(Z) = p) \\ &= P\left(Y_1 \leq \overline{Q}_1(\tau) | U_D \leq p\right) p \\ &= \int_0^p F_{Y_1|U_D}(\overline{Q}_1(\tau)|u_D) du_D \end{aligned}$$

In summary,

$$\begin{pmatrix} \tau \int (1 - p)\, dF_{p(Z)}(p) \\ \tau \int p dF_{p(Z)}(p) \end{pmatrix} = \begin{pmatrix} \int \left[\int_p^1 F_{Y_0|U_D}(\overline{Q}_0(\tau)|u_D) du_D\right] dF_{p(Z)}(p) \\ \int \left[\int_0^p F_{Y_1|U_D}(\overline{Q}_1(\tau)|u_D) du_D\right] dF_{p(Z)}(p) \end{pmatrix},$$

or

$$\begin{aligned} \overline{F}_0\left(\overline{Q}_0(\tau)\right) &\equiv \int_{\underline{p}}^{\overline{p}} \left[\frac{1 - p}{1 - E\left[p(Z)\right]} \frac{1}{1 - p} \int_p^1 F_{Y_0|U_D}(\overline{Q}_0(\tau)|u_D) du_D\right] dF_{p(Z)}(p) \\ &= \int_{\underline{p}}^{\overline{p}} F_{Y_0|U_D}(\overline{Q}_0(\tau)|u_D) \frac{F_{p(Z)}(u_D)}{E\left[1 - p(Z)\right]} du_D + \int_{\overline{p}}^1 \frac{F_{Y_0|U_D}(\overline{Q}_0(\tau)|u_D)}{E\left[1 - p(Z)\right]} du_D = \tau, \\ \overline{F}_1\left(\overline{Q}_1(\tau)\right) &\equiv \int_{\underline{p}}^{\overline{p}} \left[\frac{p}{E\left[p(Z)\right]} \frac{1}{p} \int_0^p F_{Y_1|U_D}(\overline{Q}_1(\tau)|u_D) du_D\right] dF_{p(Z)}(p) \\ &= \int_0^{\underline{p}} \frac{F_{Y_1|U_D}(\overline{Q}_1(\tau)|u_D)}{E\left[p(Z)\right]} du_D + \int_{\underline{p}}^{\overline{p}} F_{Y_1|U_D}(\overline{Q}_1(\tau)|u_D) \frac{1 - F_{p(Z)}(u_D)}{E\left[p(Z)\right]} du_D = \tau, \end{aligned}$$

where the second equalities use Fubini's theorem. Since the formula for $\overline{F}_d\left(\overline{Q}_d(\tau)\right)$ is valid for any τ, the QRE is estimating $\overline{F}_d(y_d)$ by the formula stated in the theorem. Given $\overline{F}_d(y_d)$, the formula for $\overline{\Delta}(\tau)$ is followed. ■

Supplementary Material S.2

S.2.1. Alternative Formulas for μ_d^*

Different parts of μ_d^* in Theorem 1 can be re-expressed as

$$\begin{aligned}\int_0^{\underline{p}} E\left[Y_1|U_D = u_D\right] du_D &= \underline{p}E\big[Y|D = 1, p(Z) = \underline{p}\big],\\ \int_{\overline{p}}^1 E\left[Y_0|U_D = u_D\right] du_D &= (1-\overline{p})E\left[Y|D = 0, p(Z) = \overline{p}\right],\end{aligned}$$

and

$$\begin{aligned}&\int_{\underline{p}}^{\overline{p}} E\left[Y_1|U_D = u_D\right] h_1(u_D)du_D\\ =& \int_{\underline{p}}^{\overline{p}} \frac{E\left[p(Z)^2\right] - pE\left[p(Z)\right] + (p - E\left[p(Z)\right])}{Var\left(p(Z)\right)}\big[pE\left[Y|D = 1, p(Z) = p\right] - \underline{p}E\big[Y|D = 1, p(Z) = \underline{p}\big]\big]\, dF_{p(Z)}(p)\\ &\int_{\underline{p}}^{\overline{p}} E\left[Y_0|U_D = u_D\right](1 - h_1(u_D))du_D\\ =& \int_{\underline{p}}^{\overline{p}} \frac{E\left[p(Z)^2\right] - pE\left[p(Z)\right] + (p - E\left[p(Z)\right])}{Var\left(p(Z)\right)}\left[(1-p)\,E\left[Y|D = 0, p(Z) = p\right] - (1-\overline{p})\,E\left[Y|D = 0, p(Z) = \overline{p}\right]\right] dF_{p(Z)}(p)\\ &\int_{\underline{p}}^{\overline{p}} E\left[Y_1|U_D = u_D\right] h_0(u_D)du_D\\ =& \int_{\underline{p}}^{\overline{p}} \frac{E\left[p(Z)^2\right] - pE\left[p(Z)\right]}{Var\left(p(Z)\right)}\big[pE\left[Y|D = 1, p(Z) = p\right] - \underline{p}E\big[Y|D = 1, p(Z) = \underline{p}\big]\big]\, dF_{p(Z)}(p)\\ &\int_{\underline{p}}^{\overline{p}} E\left[Y_0|U_D = u_D\right](1 - h_0(u_D))du_D\\ =& \int_{\underline{p}}^{\overline{p}} \frac{E\left[p(Z)^2\right] - pE\left[p(Z)\right]}{Var\left(p(Z)\right)}\left[(1-p)\,E\left[Y|D = 0, p(Z) = p\right] - (1-\overline{p})\,E\left[Y|D = 0, p(Z) = \overline{p}\right]\right] dF_{p(Z)}(p).\end{aligned}$$

Finally,

$$\begin{aligned}&\int_{\underline{p}}^{\overline{p}} MTE(u_D)h(u_D)du_D\\ =& \int_{\underline{p}}^{\overline{p}} E\left[Y_1|U_D = u_D\right] h(u_D)du_D - \int_{\underline{p}}^{\overline{p}} E\left[Y_0|U_D = u_D\right] h(u_D)du_D\\ =& \int_{\underline{p}}^{\overline{p}} \frac{p - E\left[D\right]}{Var(p(Z))}\big[pE\left[Y|D = 1, p(Z) = p\right] - \underline{p}E\big[Y|D = 1, p(Z) = \underline{p}\big]\big]\, dF_{p(Z)}(p)\\ &- \int_{\underline{p}}^{\overline{p}} \frac{p - E\left[D\right]}{Var(p(Z))}\left[(1-p)\,E\left[Y|D = 0, p(Z) = p\right] - (1-\overline{p})\,E\left[Y|D = 0, p(Z) = \overline{p}\right]\right] dF_{p(Z)}(p),\end{aligned}$$

where $pE\left[Y|D = 1, p(Z) = p\right]$ can be rewritten as $E\left[YD|p(Z) = p\right]$ and $(1-p)\,E\left[Y|D = 0, p(Z) = p\right]$ can be rewritten as $E\left[Y\left(1-D\right)|p(Z) = p\right]$.

S.2.2. Alternative Formulas for $F_d^*\left(y_d\right)$

Different parts of $F_d^*\left(y_d\right)$ in Theorem 2 can be re-expressed as

$$\begin{aligned}
\int_0^{\underline{p}} F_{Y_1|U_D}(y_1|u_D)du_D &= \underline{p}F_{Y|D=1,p(Z)=\underline{p}}(y_1), \\
\int_{\overline{p}}^1 F_{Y_0|U_D}(y_0|u_D)du_D &= (1-\overline{p})F_{Y|D=0,p(Z)=\overline{p}}(y_0),
\end{aligned}$$

and

$$\begin{aligned}
\widetilde{F}_1\left(y_1\right) &= \int_{\underline{p}}^{\overline{p}} \frac{p-E\left[D\right]}{Var(p(Z))}\left[pF_{Y|D=1,p(Z)=p}(y_1) - \underline{p}F_{Y|D=1,p(Z)=\underline{p}}(y_1)\right]dF_{p(Z)}(p), \\
\widetilde{F}_0\left(y_0\right) &= \int_{\underline{p}}^{\overline{p}} \left(\frac{E\left[D\right]-p}{Var(p(Z))}\right)\left[(1-p)F_{Y|D=0,p(Z)=p}\left(y_0\right) - (1-\overline{p})F_{Y|D=0,p(Z)=\overline{p}}(y_0)\right]dF_{p(Z)}(p), \\
& \int_{\underline{p}}^{\overline{p}} F_{Y_1|U_D}(y_1|u_D)(1-F_{p(Z)}(u_D))du_D \\
&= \int_{\underline{p}}^{\overline{p}} \left[pF_{Y|D=1,p(Z)=p}(y_1) - \underline{p}F_{Y|D=1,p(Z)=\underline{p}}(y_1)\right]dF_{p(Z)}(p), \\
& \int_{\underline{p}}^{\overline{p}} F_{Y_0|U_D}(y_1|u_D)F_{p(Z)}(u_D)du_D \\
&= \int_{\underline{p}}^{\overline{p}} \left[(1-p)F_{Y|D=0,p(Z)=p}\left(y_0\right) - (1-\overline{p})F_{Y|D=0,p(Z)=\overline{p}}(y_0)\right]dF_{p(Z)}(p),
\end{aligned}$$

where $pF_{Y|D=1,p(Z)=p}(y_1)$ can be rewritten as $E\left[1\left(Y \leq y_1\right)D|p(Z)=p\right]$ and $(1-p)F_{Y|D=0,p(Z)=p}\left(y_0\right)$ can be rewritten as $E\left[1\left(Y \leq y_0\right)\left(1-D\right)|p(Z)=p\right]$.

Note that the expressions for $\int_{\underline{p}}^{\overline{p}} F_{Y_1|U_D}(y_1|u_D)(1-F_{p(Z)}(u_D))du_D$ and $\int_{\underline{p}}^{\overline{p}} F_{Y_0|U_D}(y_1|u_D)F_{p(Z)}(u_D)du_D$ can be extended to the formulas for the LSE and QRE.

S.2.3. Properties of h_{Jd} and h_J

First note that h_{Jd} and h_J are only functions of $p(Z)$ and $J(Z)$ and do not involve Y_1 and Y_0. For completeness, we restate the first three properties of h_d and h for h_{Jd} and h_J.

Proposition 11 *h_{J1}, h_{J0} and h_J satisfy the following properties:*

(i) $\int_{\underline{p}}^{\overline{p}} h_{J1}(u_D)du_D = 1-\underline{p}$, $\int_0^1 h_{J1}(u_D)du_D = 1$;

(ii) $\int_{\underline{p}}^{\overline{p}} h_{J0}(u_D)du_D = -\underline{p}$, $\int_0^1 h_{J0}(u_D)du_D = 0$;

(iii) $\int_{\underline{p}}^{\overline{p}} h_J(u_D)du_D = \int_0^1 h_J(u_D)du_D = 1$;

(iv) *if $E\left[J(Z)|p(Z)=p\right]$ is a continuous and strictly increasing or decreasing function of p, then when $u_D < u_{D1}^*$, $h_{J1}(u_D)$ is strictly increasing, and when $u_D > u_{D1}^*$, $h_{J1}(u_D)$ is strictly decreasing, where $u_{D1}^* \in [\underline{p},\overline{p}]$ is defined by $E\left[J(Z)|p(Z)=u_{D1}^*\right] = \frac{E[J(Z)]-E[p(Z)J(Z)]}{1-E[p(Z)]}$;*

(v) *if $E\left[J(Z)|p(Z)=p\right]$ is a continuous and strictly increasing or decreasing function of p, when $u_D < u_{D0}^*$, $h_{J0}(u_D)$ is strictly decreasing, and when $u_D > u_{D0}^*$, $h_{J0}(u_D)$ is strictly increasing, where $u_{D0}^* \in [\underline{p}, \overline{p}]$ is defined by $E\left[J(Z)|p(Z)=u_{D0}^*\right] = \frac{E[p(Z)J(Z)]}{E[p(Z)]}$;*

(vi) *if $E\left[J(Z)|p(Z)=p\right]$ is a continuous and strictly increasing or decreasing function of p, when $u_D < u_D^*$, $h_J\left(u_D\right)$ is strictly increasing, and when $u_D > u_D^*$, $h_J(u_D)$ is strictly decreasing, where $u_D^* \in [\underline{p}, \overline{p}]$ is defined by $E\left[J(Z)|p(Z)=u_D^*\right] = E\left[J(Z)\right]$.*

Proof. First,

$$\begin{aligned}
\int_{\underline{p}}^{\overline{p}} h_{J1}(u_D)du_D &= \int_{\underline{p}}^{\overline{p}} \frac{1}{Cov\left(J(Z),p(Z)\right)} \int_{u_D}^{1} \left[\begin{array}{c} E\left[p(Z)J(Z)\right] - E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right] \\ +\left(E\left[J(Z)|p(Z)=p\right] - E\left[J(Z)\right]\right) \end{array} \right] dF_{p(Z)}(p)du_D \\
&= \frac{1}{Cov\left(J(Z),p(Z)\right)} \int_{\underline{p}}^{\overline{p}} \left[\begin{array}{c} E\left[p(Z)J(Z)\right] - E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right] \\ +\left(E\left[J(Z)|p(Z)=p\right] - E\left[J(Z)\right]\right) \end{array} \right] \int_{\underline{p}}^{p} du_D dF_{p(Z)}(p) \\
&= \frac{1}{Cov\left(J(Z),p(Z)\right)} \int_{\underline{p}}^{\overline{p}} \left[\begin{array}{c} E\left[p(Z)J(Z)\right] - E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right] \\ +\left(E\left[J(Z)|p(Z)=p\right] - E\left[J(Z)\right]\right) \end{array} \right] \left(p-\underline{p}\right) dF_{p(Z)}(p) \\
&= 1-\underline{p},
\end{aligned}$$

where the second equality is from Fubini's theorem. Similarly,

$$\begin{aligned}
\int_{\underline{p}}^{\overline{p}} h_{J0}(u_D)du_D &= \frac{1}{Cov\left(J(Z),p(Z)\right)} \int_{\underline{p}}^{\overline{p}} \int_{u_D}^{\overline{p}} \left[E\left[p(Z)J(Z)\right] - E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right]\right] dF_{p(Z)}(p)du_D \\
&= \frac{1}{Cov\left(J(Z),p(Z)\right)} \int_{\underline{p}}^{\overline{p}} \left[E\left[p(Z)J(Z)\right] - E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right]\right] \int_{\underline{p}}^{p} du_D dF_{p(Z)}(p) \\
&= \frac{1}{Cov\left(J(Z),p(Z)\right)} \int_{\underline{p}}^{\overline{p}} \left[E\left[p(Z)J(Z)\right] - E\left[J(Z)|p(Z)=p\right]E\left[p(Z)\right]\right] \left(p-\underline{p}\right) dF_{p(Z)}(p) \\
&= -\underline{p}.
\end{aligned}$$

So

$$\begin{aligned}
\int_0^1 h_{J1}(u_D)du_D &= \int_0^{\underline{p}} h_{J1}(u_D)du_D + \int_{\underline{p}}^{\overline{p}} h_{J1}(u_D)du_D + \int_{\overline{p}}^{1} h_{J1}(u_D)du_D \\
&= \underline{p} + 1 - \underline{p} + 0 = 1,
\end{aligned}$$

and

$$\begin{aligned}
\int_0^1 h_{J0}(u_D)du_D &= \int_0^{\underline{p}} h_{J0}(u_D)du_D + \int_{\underline{p}}^{\overline{p}} h_{J0}(u_D)du_D + \int_{\overline{p}}^{1} h_{J0}(u_D)du_D \\
&= \underline{p} - \underline{p} + 0 = 0,
\end{aligned}$$

which implies

$$\int_{\underline{p}}^{\overline{p}} h_J(u_D)du_D = \int_0^1 h_J(u_D)du_D = \int_0^1 h_{J1}(u_D)du_D - \int_0^1 h_{J0}(u_D)du_D = 1,$$

as shown in HV.

For the last three properties, note that

$$\begin{aligned} h'_{J1}(u_D) &= \left[E\left[J(Z)\right] - E\left[p(Z)J(Z)\right] - \left(1 - E\left[p(Z)\right]\right) E\left[J(Z)|p(Z) = u_D\right]\right] \frac{f_{p(Z)}(u_D)}{Cov\left(J(Z), p(Z)\right)}, \\ h'_{J0}(u_D) &= \left[E\left[p(Z)\right] E\left[J(Z)|p(Z) = u_D\right] - E\left[p(Z)J(Z)\right]\right] \frac{f_{p(Z)}(u_D)}{Cov\left(J(Z), p(Z)\right)}, \\ h'_J\left(u_D\right) &= \left[E\left[J(Z)\right] - E\left[J(Z)|p(Z) = u_D\right]\right] \frac{f_{p(Z)}(u_D)}{Cov\left(J(Z), p(Z)\right)}, \end{aligned}$$

so the shape of h_{Jd} and h_J depends on the property of $E\left[J(Z)|p(Z) = u_D\right]$. If $j(u_D) \equiv E\left[J(Z)|p(Z) = u_D\right]$ is strictly increasing,

$$Cov\left(J(Z), p(Z)\right) = \int_{\underline{p}}^{\overline{p}} p\left(E\left[J(Z)|p(Z) = p\right] - E\left[J(Z)\right]\right) dF_{p(Z)}(p) > 0,$$

and similarly, $Cov\left(J(Z), p(Z)\right) < 0$ if $E\left[J(Z)|p(Z) = u_D\right]$ is strictly decreasing.

Consider the strictly increasing case only since the other case can be similarly analyzed. For $h'_{J1}(u_D)$, note that

$$\frac{E\left[J(Z)\right] - E\left[p(Z)J(Z)\right]}{1 - E\left[p(Z)\right]} \in \left[\underline{j}, \overline{j}\right],$$

where $\underline{j} = \min_{u_D \in [\underline{p}, \overline{p}]} j(u_D)$, and $\overline{j} = \max_{u_D \in [\underline{p}, \overline{p}]} j(u_D)$. So there exists a $u^*_{D1} \in [\underline{p}, \overline{p}]$ such that

$$j(u^*_{D1}) = \frac{E\left[J(Z)\right] - E\left[p(Z)J(Z)\right]}{1 - E\left[p(Z)\right]},$$

and when $u_D < u^*_{D1}$, $h'_{J1}(u_D) > 0$ and $u_D > u^*_{D1}$, $h'_{J1}(u_D) < 0$. For $h'_{J0}(u_D)$, note that

$$\frac{E\left[p(Z)J(Z)\right]}{E\left[p(Z)\right]} \in \left[\underline{j}, \overline{j}\right],$$

so there exists a $u^*_{D0} \in [\underline{p}, \overline{p}]$ such that

$$j(u^*_{D0}) = \frac{E\left[p(Z)J(Z)\right]}{E\left[p(Z)\right]},$$

and when $u_D < u^*_{D0}$, $h'_{J0}(u_D) < 0$ and $u_D > u^*_{D0}$, $h'_{J0}(u_D) > 0$. For $h'_J\left(u_D\right)$, note that $E\left[J(Z)\right] \in \left[\underline{j}, \overline{j}\right]$, so there exists a $u^*_D \in [\underline{p}, \overline{p}]$ such that $j(u^*_D) = E\left[J(Z)\right]$ and when $u_D < u^*_D$, $h'_J(u_D) > 0$ and $u_D > u^*_D$, $h'_J(u_D) < 0$. ■

S.2.4. Unnecessity of $\underline{p} = 0$ and $\overline{p} = 1$ for Point Identification of $\Delta(\tau)$

The following example illustrates that $\underline{p} = 0$ and $\overline{p} = 1$ are not necessary for point identification of $\Delta(\tau)$ when Y_d is binary. A similar example where the distribution of Y_d is a mixture of continuous and discrete is available upon request.

Example 10 *Suppose* $Y_d \in \{0, 1\}$. $\overline{p}_1 \equiv P(Y = 0|p(Z) = \overline{p}, D = 1) \in (0, 1)$ *and* $\underline{p}_0 \equiv P(Y = 0|p(Z) =$

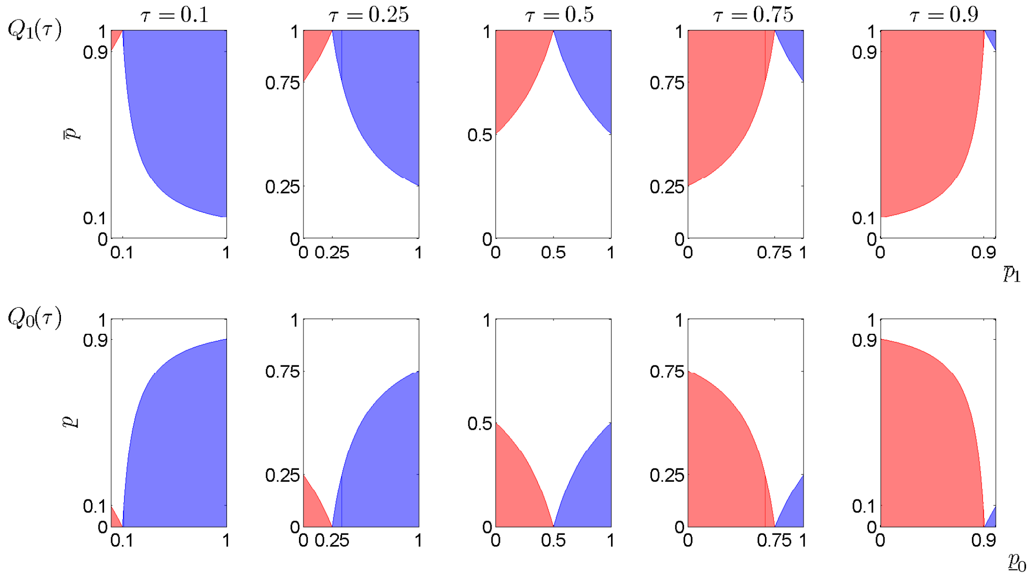

Figure 11: $\overline{p}$ ($\underline{p}$) and $\overline{p}_1$ ($\underline{p}_0$) Combination for Point Identification of $Q_1(\tau)$ ($Q_0(\tau)$): Red Area for $Q_d(\tau) = 1$ and Blue Area for $Q_d(\tau) = 0$

$\underline{p}, D = 0) \in (0,1)$. *First check the bounds for* $Q_1(\tau)$*:*

$$\underline{I}_1(\tau) = \begin{cases} 1 & \textit{if } \overline{p} > 1-\tau \textit{ and } 1 - \frac{1-\tau}{\overline{p}} > \overline{p}_1, \\ 0, & \textit{if } \overline{p} \le 1-\tau \textit{ or } [\overline{p} > 1-\tau \textit{ and } 1 - \frac{1-\tau}{\overline{p}} \le \overline{p}_1], \end{cases}$$
$$\overline{I}_1(\tau) = \begin{cases} 0, & \textit{if } \overline{p} \ge \tau \textit{ and } \frac{\tau}{\overline{p}} \le \overline{p}_1, \\ 1, & \textit{if } \overline{p} < \tau \textit{ or } [\overline{p} \ge \tau \textit{ and } \frac{\tau}{\overline{p}} > \overline{p}_1]. \end{cases}$$

When $\max\left\{1-\tau, \frac{\tau}{\overline{p}_1}\right\} < \overline{p} \le \frac{1-\tau}{1-\overline{p}_1}$ *or* $\frac{\tau}{\overline{p}_1} \le \overline{p} \le 1-\tau$, $\underline{I}_1(\tau) = \overline{I}_1(\tau) = 0$*; when* $\frac{1-\tau}{1-\overline{p}_1} < \overline{p} < \tau$ *or* $\max\left\{\frac{1-\tau}{1-\overline{p}_1}, \tau\right\} < \overline{p} < \frac{\tau}{\overline{p}_1}$, $\underline{I}_1(\tau) = \overline{I}_1(\tau) = 1$. *Similarly, when* $1 - \frac{1-\tau}{1-\underline{p}_0} \le \underline{p} < \min\left\{\tau, 1 - \frac{\tau}{\underline{p}_0}\right\}$ *or* $\tau \le \underline{p} \le 1 - \frac{\tau}{\underline{p}_0}$, $\underline{I}_0(\tau) = \overline{I}_0(\tau) = 0$*; when* $1-\tau < \underline{p} < 1 - \frac{1-\tau}{1-\underline{p}_0}$ *or* $1 - \frac{\tau}{\underline{p}_0} < \underline{p} < \min\left\{1-\tau, 1 - \frac{1-\tau}{1-\underline{p}_0}\right\}$, $\underline{I}_0(\tau) = \overline{I}_0(\tau) = 1$. *Figure 11 shows the combination of* $\overline{p}$ ($\underline{p}$) *and* $\overline{p}_1$ ($\underline{p}_0$) *for point identification of* $Q_1(\tau)$ ($Q_0(\tau)$) *at* $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$. *Obviously,* $\underline{p} = 0$ *and* $\overline{p} = 1$ *are not necessary for point identification of* $\Delta(\tau)$. *Only if* $\overline{p}_1 = \underline{p}_0 = \tau$, $\underline{p} = 0$ *and* $\overline{p} = 1$ *are necessary. Note also that* $\overline{p} \ge \min\{\tau, 1-\tau\}$ *and* $\overline{p} \le \max\{\tau, 1-\tau\}$ *for point identification of* $\Delta(\tau)$ *for any* $\overline{p}_1$, $\underline{p}_0 \in (0,1)$ *as predicted by Theorem ? of Yu (2016a).*

S.2.5. $\int y_d dF_d^*(y_d) \neq \mu_d^*$ Under Assumption RS

As shown in Section 4.1, $E[U_1 - U_0 | U_D = u_D]$ in the ATE case equals $\int [q_1(u) - q_0(u)] dF_{U|U_D}(u|u_D) - \int [q_1(u) - q_0(u)] du$ in the QTE case under Assumption RS, and need not be zero. Under Assumption RS, $F_d^*(y_d) = F_d(y_d)$ so that $\int y_d dF_d^*(y_d) = \int y_d dF_d(y_d) = \mu_d$ while μ_d^* need not equal μ_d since $E[U_1 - U_0 | U_D = u_D]$ need not be zero.

We use the running example to illustrate this result. In this example with only the selection effect,

$$\begin{aligned} q_1(u) &= 2\Phi^{-1}(u), q_0(u) = \Phi^{-1}(u), \\ F_{U_1|U_D}(u|u_D) &= F_{U_0|U_D}(u|u_D) = \Phi\left(\frac{\Phi^{-1}(u) - 0.7\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right), \end{aligned}$$

that is, Assumption RS holds and $\int y_d dF_d^*(y_d) = \mu_d$. On the other hand,

$$E[U_1 - U_0|U_D = u_D] = \int \left[q_1(u) - q_0(u)\right] dF_{U|U_D}(u|u_D) = 0.7\Phi^{-1}(u_D) \neq 0.$$

In other words, there is essential heterogeneity in the ATE case. It is not hard to check that

$$\begin{aligned} \mu_1^* &= \int \left\{E\left[Y_1|U_D = u_D\right] h_1(u_D) + E\left[Y_0|U_D = u_D\right](1 - h_1(u_D))\right\} du_D \\ &= \int \left\{1.4u_D(1+3u_D)(1-u_D) + 0.7u_D\left[1 - (1+3u_D)(1-u_D)\right]\right\} du_D \\ &\approx 0.64 \neq 0 = \mu_1 \end{aligned}$$

and

$$\begin{aligned} \mu_0^* &= \int \left\{E\left[Y_1|U_D = u_D\right] h_0(u_D) + E\left[Y_0|U_D = u_D\right](1 - h_0(u_D))\right\} du_D \\ &= \int \left\{1.4u_D(1-3u_D)(1-u_D) + 0.7u_D\left[1 - (1-3u_D)(1-u_D)\right]\right\} du_D \\ &\approx 0.29 \neq 0 = \mu_0. \end{aligned}$$

From this example, the IV-QRE can handle some cases that the IVE cannot handle. Of course, the IVE can also handle some cases that the IV-QRE cannot handle. For example, in the running example, let $Y_1 = V + 2U$ and $Y_0 = 2V + \frac{0.4}{0.7}U$; then

$$\begin{aligned} F_{U_1|U_D}(u|u_D) &= P\left(Y_1 \leq F_1^{-1}(u)|V = \Phi^{-1}(u_D)\right) = \Phi\left(\frac{\frac{\sqrt{7.8}}{2}\Phi^{-1}(u) - 1.2\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right), \\ F_{U_0|U_D}(u|u_D) &= P\left(Y_0 \leq F_0^{-1}(u)|V = \Phi^{-1}(u_D)\right) = \Phi\left(\frac{\frac{7\sqrt{a}}{4}\Phi^{-1}(u) - 4.2\Phi^{-1}(u_D)}{\sqrt{1-0.7^2}}\right). \end{aligned}$$

where $a = 28/5 + 16/49$. Obviously, $F_{U_1|U_D}(u|u_D) \neq F_{U_0|U_D}(u|u_D)$, so Assumption RS does not hold. However,

$$\begin{aligned} E\left[U_1|U_D = u_D\right] &= Cov\left(Y_1, V\right) u_D = 2.4u_D, \\ E\left[U_0|U_D = u_D\right] &= Cov\left(Y_0, V\right) u_D = 2.4u_D, \end{aligned}$$

i.e., $E\left[U_1 - U_0|U_D = u_D\right] = 0$. Certainly, it is not hard to find cases that both the IVE and IV-QRE can handle, e.g., in the running example, let $Y_1 = 1 + U$ and $Y_0 = U$.