

# Efficient Semiparametric Estimation of Quantile Treatment Effects\*

Sergio Firpo

UC Berkeley - Department of Economics

This Draft: November 17, 2002

## Abstract

This paper presents calculations of semiparametric efficiency bounds for quantile treatment effects parameters when selection to treatment is based on observable characteristics. The paper also presents three estimation procedures for these parameters, all of which have two steps: a nonparametric estimation and a computation of the difference between the solutions of two distinct minimization problems. Root- $N$  consistency, asymptotic normality, and the achievement of the semiparametric efficiency bound is shown for one of the three estimators. In the final part of the paper, an empirical application to a job training program reveals the importance of heterogeneous treatment effects, showing that for this program the effects are concentrated in the upper quantiles of the earnings distribution.

**Keywords:** *Quantile Treatment Effects, Propensity Score, Semiparametric Efficiency Bounds, Efficient Estimation, Semiparametric Estimation*

---

\*Preliminary version: Comments are welcome. I am indebted to Guido Imbens and to Jim Powell for their advice, support and many suggestions. I am also grateful to Carlos Flores, Jinyong Hahn, David Lee, Rosa Matzkin, Dan McFadden, Deb Nolan, David Reinstein, Paul Ruud and Jeffrey Wooldridge for comments. Financial support from CAPES - Brazil is acknowledged. All errors are mine. Electronic correspondence: [firpo@econ.berkeley.edu](mailto:firpo@econ.berkeley.edu)

# 1 INTRODUCTION

## 1.1 THE PROBLEM

In program evaluation studies it is often important to learn not only about the average treatment effects, but about the distributional effects of a treatment. In particular, the policy-maker might be interested in the effect of the treatment on the dispersion of the outcome, or its effect on the lower tail of the outcome distribution.

One way of capturing this effect in a setting with binary treatment and scalar outcomes is by computing the quantiles of the treated and the control outcomes. Using quantiles, discretized versions of the distribution functions of treated and controls can be calculated. Also, quantiles are used in many inequality measurements as, for instance, quantile ratios, inter-quantile ranges, concentration functions, and the Gini coefficient. Finally, differences in quantiles are important as the effects of a treatment may be heterogeneous, varying along the outcome distribution.

The parameter of interest in this paper, labeled the quantile treatment effect, is the difference between the treated and the control groups in the marginal quantiles of the outcome. As is the case for any treatment effect parameter, identification restrictions are necessary for this parameter to be estimable. In this case the relevant restriction is the assumption that selection to treatment is based on observable variables.

It is common practice in calculations of average treatment effects to first compute a conditional average treatment effect, and then to integrate over the distribution of covariates to recover the unconditional average treatment effect. However, as the mean of the quantiles is not equal to the quantile of the mean, integrating a first-stage computation of the conditional quantiles (of the treated and the control outcomes) will not yield the marginal quantiles. Instead, this paper demonstrates how to use the identification assumption that selection to treatment is based on observable variables to calculate the marginal quantiles for the treated and for the control outcomes without computing the corresponding conditional quantiles.

Quantile treatment effects have been indirectly computed for the case in which selection into the treatment group is based on observable characteristics. DiNardo, Fortin and Lemieux (1996) have suggested a way of estimating counterfactual densities of control groups in a binary treatment/scalar outcome setting. Apparently however, no further development, refinement, or derivations of large sample properties of this procedure have been proposed in the literature.

I show in this paper how to estimate the quantile treatment effects in three different ways. All three proposed estimation techniques involve two steps. The first is nonparametric, and the estimators may differ by the number and type of estimated functionals. In the second step all estimators equivalently solve minimum distance problems and are minimizers of the sum of check functions. This second step is typical of quantile estimation. I then focus on a two-step estimation technique that involves estimation of only one function in the first step: the propensity score. I show that this estimator is root- $N$  consistent and asymptotically normal. I also calculate the semiparametric efficiency bound and show that the quantile treatment effects estimator achieves it. Finally, I provide an empirical application, to illustrate the techniques and show its practicality. The estimates suggest that for several quantiles the treatment effect is quite different from the mean treatment effect. Thus, the application demonstrates how the techniques developed in this paper can provide evidence of heterogeneity in the impact of a treatment.

## 1.2 QUANTILE TREATMENT EFFECTS

In a binary treatment/scalar outcome setting, one is often interested in learning the impact of the treatment on the outcome. We define the potential outcome of being treated,  $Y(1)$ , as the outcome that an individual would have experienced (or perhaps did experience) had she been exposed to the treatment. Analogously, we define the potential outcome of not being treated,  $Y(0)$ , as the (hypothetical or actual) outcome had the individual not been exposed to the treatment. For any given individual we observe only one potential outcome, the other one,

sometimes called the counterfactual outcome, constitutes missing data.

The fact that potential outcomes are partially unobservable leads us to the use of some identification restrictions, a requirement that is common to the identification of any treatment effects parameter. A typical strategy to deal with this problem is to assume that given a set of observed covariates, individuals are randomly assigned either to the treatment group or to the control group. That assumption was termed by Rubin (1977) the *unconfoundedness assumption* and it characterizes the *selection on observables* branch of the program evaluation literature. Heckman, Ichimura, Smith and Todd (1998) and Dehejia and Wahba (1999) are important examples. Further explanation of these identifying assumptions will be provided in later sections.

Several parameters can be defined in order to capture the effects of a treatment. In most cases, the focus is on the *average treatment effect* (ATE) defined as the difference in the means of the potential outcomes. One reason that many program evaluation studies focus on average treatment effects is that for the special case in which the treatment has a homogeneous effect, it is possible to interpret ATE as the effect of the treatment on a single observation. Note, however, that the average treatment effect does not depend on homogeneity assumptions to be well-defined.

Indeed, treatment effects may be heterogeneous, varying greatly along the outcome distribution. The presence of heterogeneity in treatment effects is very important when evaluating programs, as policy-makers are often interested in the distributional consequences of the treatment. This is true, for example, for a wide range of social programs such as welfare, unemployment insurance, subsidized job training, the minimum wage, agrarian reform, and micro-credit provision.

A parameter of interest in the presence of heterogeneous treatment effects is the *quantile treatment effect* (QTE). As originally defined by Lehmann (1974) and Doksum (1974), the QTE corresponds, for any fixed percentile, to the horizontal distance between two cumulative

distribution functions. In defining QTE as a treatment effect at the individual level, both Doksum (1974) and Lehmann (1974) implicitly argued that an observed individual would maintain her rank in the distribution regardless of her treatment status. This paper will refer to this type of assumption as a *rank invariance* assumption.

Rank invariance assumptions are strong assumptions as they require that the relative value (rank) of the potential outcome for a given individual would be the same under treatment as under non-treatment. There are two ways to deal with cases in which rank invariance is an unreasonable assumption. The first one is due to Heckman, Smith, and Clements (1997), who suggested computing bounds for the QTE, allowing for several possibilities of re-orderings of the ranks. According to them, the outcome for the same individual may differ from one distribution to another based on how observable and unobservable attributes impact each one of the potential outcomes. However, while the effect of observable characteristics can be measured, unobservable characteristics can interact with treatment status in many unknown ways, leaving open the possibility of a sharp reordering of ranks. Bounds for the QTE that capture these alternatives were proposed by Heckman, Smith and Clements (1997).

The second approach to dealing with failures of the rank invariance assumption argues that even without this assumption, one can still have a meaningful parameter for policy purposes. Consider the case in which all the policy-maker is interested is in learning about the marginal distributions of the potential outcomes. A good way to summarize interesting aspects of these distributions is by computing their quantiles. In this case, quantile treatment effects can be defined as simple differences between quantiles of the marginal distributions of potential outcomes. As an example, suppose that one is interested in the difference in medians between two distributions, and not in the effects of treatment on a typical individual. In such a setting it is not necessary to have any knowledge about the joint distribution of outcomes for the treated and control groups, so the rank invariance assumption could be dropped. Note, however, that if rank invariance holds, then the simple differences in quantiles turn out to be the quantiles of

the treatment.<sup>1</sup>

This definition of quantile treatment effects, together with the selection on observables approach, allows identification of various QTE parameters that differ by the subpopulation they refer to. Following the approach of Heckman and Robb (1986) and Hirano, Imbens and Ridder (2002), who suggest several parameters of interest for the mean case, two QTE parameters will be the object of study in this paper. They are labeled the *quantile treatment effect* and the *quantile treatment effect on the treated*, the former being the QTE parameter for the whole population under consideration and the latter the parameter for those individuals subject to treatment. Defining  $T$  as the indicator variable of treatment, these parameters can be expressed as:

**Quantile Treatment Effect:**  $\Delta_{\tau} = q_{1,\tau} - q_{0,\tau}$ ,

where  $q_{j,\tau}$  is such that  $Pr[Y(j) \leq q] = \tau$ ,  $j = 0, 1$ .

**Quantile Treatment Effect on the Treated:**  $\Delta_{\tau|T=1} = q_{1,\tau|T=1} - q_{0,\tau|T=1}$ ,

where  $q_{j,\tau|T=1}$  is such that  $Pr[Y(j) \leq q|T = 1] = \tau$ ,  $j = 0, 1$ .

The role that the observable covariates play in identification of both ATE and QTE is made clearer in the QTE case. This is because, as stated earlier, the computation of quantile treatment effects does not use the conditional quantiles. Computation of conditional quantiles is unnecessary since the quantiles of the marginal distributions of the potential outcomes are the object of interest and the mean of the quantile is not the quantile of the mean. Hence, for QTE, the covariates serve only to remove the selection bias.

Quantile treatment effects are also useful in describing the center of the distribution of the treatment. In particular the *median treatment effect* (MTE), the QTE for the fifty percentile, is a central measure of the treatment effect, like ATE. However, MTE has an additional and desirable feature not present in ATE: its corresponding estimator is robust to the presence of data outliers.

---

<sup>1</sup>Note that there is no similar problem in estimation of the average treatment effect, as differences in means always coincide with means of differences.

Despite the relevance of QTE, the program evaluation literature on this topic is not as vast as that of its main competitor, ATE. Traditionally, expectations have received more attention in the literature than quantiles. Pioneer papers on quantile estimation, such as those by Koenker and Bassett (1978) and, in an instrumental variables setting, by Amemiya (1982) and Powell (1983) have helped to bridge this gap. In the treatment effects literature, some recent contributions have also been made to the study of the distributional effects of the treatment. Among them, Abadie, Angrist and Imbens (2002), and Chernozhukov and Hansen (2001) have proposed instrumental variables versions of the QTE. Imbens and Rubin (1997) and Abadie (2002) proposed methods to estimate some distributional features for a subset of the treated units, again in an instrumental variables setting. Distributional effects have also been studied empirically, as in the papers of Freeman (1980), Card (1996), and DiNardo, Fortin and Lemieux (1996).

In this paper three different semiparametric ways of estimating each QTE parameter are presented. Each one corresponds to a particular way that the parameter can be identified from the observable data. These three ways will differ by the number and by the sort of functionals of the observed data involved in estimating the parameter. I focus my attention on the estimation technique that requires estimation of only the propensity score. This estimator is the QTE analogue of the ATE estimator proposed by Hirano, Imbens, and Ridder (2002), and involves reweighting observations by the inverse of the propensity score. The estimator will be equal to the difference between two quantiles, which can be expressed as the solution to minimization problems, where the minimand, a sum of check functions, is a convex empirical process. Using the empirical process literature consistency and asymptotic normality results are derived. As the estimator has asymptotic variance equal to the semiparametric efficiency bound (which I compute using the techniques suggested in both Newey (1990) and Bickel, Klassen, Ritov, and Wellner (1993)), this is an efficient estimator for the QTE parameters.

The remainder of this paper is divided as follows. The next section presents a simple

model of quantile treatment effects. In the third section I demonstrate how the identification assumptions allow expressing of the parameters of interest as functionals of the observed data. Semiparametric efficiency bounds for QTE parameters are presented in Section 4, while section 5 presents the three estimation techniques (mentioned above) and large sample properties. Section 6 presents an empirical application for the estimator. Section 7 concludes.

## 2 A SIMPLE MODEL OF QUANTILE TREATMENT EFFECTS

I start by assuming that there is an available random sample of  $N$  individuals (units). For each unit  $i$ , let  $X_i$  be a random vector of observed covariates with compact support  $\mathcal{X} \subset \mathbb{R}^r$ . Define  $Y_i(1)$  as the potential outcome for individual  $i$  under treatment, and  $Y_i(0)$  the potential outcome for the same individual without the treatment. Let the treatment assignment be defined as  $T_i$ , which equals one if individual  $i$  is exposed to treatment and equals zero otherwise. As we only observe each unit at one treatment status, we say that the unobserved outcome is the counterfactual outcome. Thus, the observed outcome can be expressed as:

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0), \quad \forall i \tag{1}$$

To motivate, consider  $Y_i$  as the observed earnings of individual  $i$  in a model of the impact of a job training program on worker earnings. In this example,  $T_i$  is the indicator for the receipt of training.

Potential outcomes depend on both observed and unobserved individual characteristics. For each individual  $i$ , let  $\varepsilon_{1,i}$  and  $\varepsilon_{0,i}$  be vectors of unobservable attributes under the treatment and the control respectively. In a job training program model for example, earnings of each individual are a function of their pre-program observable characteristics, such as past earnings, employment status, education, age, job experience, gender, and union status; they are also a function of unobservable attributes, such as ability, motivation and some possible idiosyncratic shock.

Specifying the impact of  $X$  and  $(\varepsilon_1, \varepsilon_0)$  on the potential outcomes:

$$Y_i(1) = G_1(X_i, \varepsilon_{1,i}) \quad (2)$$

$$Y_i(0) = G_0(X_i, \varepsilon_{0,i}) \quad (3)$$

I assume self-selection into treatment: individuals can decide whether or not to be treated. When an individual  $i$  faces the decision whether or not to join the job training program, she will weigh the gains and costs to her of both situations. Assume that an individual  $i$  predicts her expected earnings (given her vector  $X_i$ ) and her costs for each of the alternatives. In other words, the individual  $i$  chooses the state that yields the largest expected utility:

$$\max \{E[Y(1) | X_i] - C_1(X_i, \eta_i); E[Y(0) | X_i] - C_0(X_i, \eta_i)\} \quad (4)$$

where  $C_1(\cdot, \cdot)$  and  $C_0(\cdot, \cdot)$  are some costs associated respectively with joining the training program and not joining it, and  $\eta_i$  is a vector of variables that is unobserved to the econometrician but not to the individual. Also,  $\eta_i$  is assumed to be independent of  $(\varepsilon_{1,i}, \varepsilon_{0,i})$ . The effect of  $\eta_i$  on the individual's utility will depend on whether or not she enters the job program. For example,  $\eta_i$  might be a reservation wage that enters as an argument to a foregone earnings function. Individual  $i$  will then choose to take part in the program if  $E[Y(1) | X_i] - C_1(X_i, \eta_i) \geq E[Y(0) | X_i] - C_0(X_i, \eta_i)$ . That is:

$$T = \mathbb{1}\{E[Y(1) | X_i] - E[Y(0) | X_i] - (C_1(X_i, \eta_i) - C_0(X_i, \eta_i)) \geq 0\} \quad (5)$$

Note how this model fits into the Roy model (1951) of income distribution.<sup>2</sup> In the Roy model, an individual chooses the greater of the potential earnings given by two different occupations. Here, the choice is based on the individual's expected earnings and on some individual

---

<sup>2</sup>See also Heckman and Honore (1990).

cost. Thus, after controlling for  $X_i$ , the choice of getting treatment will be independent of the individual potential earnings, which depends only on  $X_i$  and  $(\varepsilon_{1,i}, \varepsilon_{0,i})$ . That will hold as long as  $\eta_i$  and  $(\varepsilon_{1,i}, \varepsilon_{0,i})$  are independent and the functional form of potential earnings is the one described in Equations (2) and (3). The independence result can be written as:

$$(Y_i(1), Y_i(0)) \text{ is jointly independent of } T_i \text{ given } X_i \quad \forall i \quad (6)$$

Equation (6) is the assumption termed by Rubin (1977) as the unconfoundedness assumption. This assumption was derived here as a result, but we needed to put some structure on the form of the potential outcomes and on the form of the decision rule. We also needed to put stochastic restrictions on the unobserved variables. Note however, that unless there is a gain in insight to writing the model with the structure presented in Equations (2)-(5), Equation (6) could actually have been our starting point.

I will maintain the structure of the above model for now. In this model, a rank invariance assumption can be obtained by imposing two additional requirements:

- (i)  $\forall x \in \mathcal{X}$ ,  $G_1(x, \cdot)$  and  $G_0(x, \cdot)$  are either (a) strictly increasing functions or (b) strictly decreasing functions;
- (ii)  $\forall i$ ,  $\varepsilon_{1,i}$  and  $\varepsilon_{0,i}$  are perfectly positively correlated.

These two assumptions ensure that people do not change their position in the earnings ranks in each one of the possible two states. These are strong assumptions, in particular part (ii). This is the case when skills that are useful in one regime may not be as useful in another regime.<sup>3</sup>

However, note that if these two extra requirements hold, then for every individual  $i$  such that  $Pr[Y_i(1) \leq q_{1,\tau}]$ , it must be the case that  $Pr[Y_i(0) \leq q_{0,\tau}]$ .<sup>4</sup> Therefore, calculations of the difference  $q_{1,\tau} - q_{0,\tau}$  for all  $\tau$  in the interval  $[0,1]$  yield the distribution of the treatment effects.

---

<sup>3</sup>In terms of the Roy model (1951), in a world with only two occupations, hunting and fishing, that assumption implies that the most able hunters are also the most able fishermen.

<sup>4</sup>The same would be true for the quantiles of the distribution of potential outcomes given  $T = 1$ , that is, if  $Pr[Y_i(1) \leq q_{1,\tau|T=1}]$ , then  $Pr[Y_i(0) \leq q_{0,\tau|T=1}]$ .

As rank invariance is in many cases a too strong assumption, I also motivate the interest in the differences in quantiles in a different way. Assume that there is a social welfare function,  $V$ , such that  $V$  depends on the individual utility functions. For simplicity, assume that each individual utility depends only on her earnings. Therefore, we can write  $V$  as a function of the earnings distribution of the whole population. In order to simplify the argument, imagine that there are two possible scenarios: we either treat everyone or treat no one.<sup>5</sup> Under the first scenario, the distribution of earnings is then equal to distribution of  $Y(1)$ , which has the cumulative distribution function  $F_1$ ; while in the second scenario, the earnings distribution equals that of  $Y(0)$ , whose cumulative distribution function is  $F_0$ . Ignoring social choice problems, assume that the policy-maker has to choose between these two distributions in order to maximize the social welfare function:

$$V^* = \max_{F_1, F_0} V(F) \quad (7)$$

In order to compare  $V(F_1)$  with  $V(F_0)$  the policy-maker will need to calculate approximate distributions of the potential earnings,  $F_1$  and  $F_0$ , and a good way to summarize a distribution is to compute its quantiles. If we compute a sufficient number of quantiles, we will end up having a discretized approximation of the distribution.

Consider then that each distribution is approximated by the calculation of a number  $P$  of quantiles. When  $P$  is equal to 100, we say that each quantile corresponds to a percentile. Doing that for both distributions, we have:

$$V_1 = V(q_{1, \frac{1}{P}}, q_{1, \frac{2}{P}}, \dots, q_{1, 1}) \quad (8)$$

$$V_0 = V(q_{0, \frac{1}{P}}, q_{0, \frac{2}{P}}, \dots, q_{0, 1}) \quad (9)$$

---

<sup>5</sup>Alternatives, as discussed in Manski (1997), include allowing individuals to choose their treatment status or assigning them to treatment based on observed characteristics.

The policy maker chooses between treatment and no treatment according to whether  $V_1$  is greater than  $V_0$ .

Say that both  $V_1$  and  $V_0$  are linear in the quantiles, that is, say that:

$$\begin{aligned} V_1 &= V(q_{1,\frac{1}{P}}, q_{1,\frac{2}{P}}, \dots, q_{1,1}) \\ &= \sum_{j=1}^P a_{1,\frac{j}{P}} q_{1,\frac{j}{P}} \end{aligned} \quad (10)$$

$$\begin{aligned} V_0 &= V(q_{0,\frac{1}{P}}, q_{0,\frac{2}{P}}, \dots, q_{0,1}) \\ &= \sum_{j=1}^P a_{0,\frac{j}{P}} q_{0,\frac{j}{P}} \end{aligned} \quad (11)$$

where  $a_{1,\frac{j}{P}}$  and  $a_{0,\frac{j}{P}}$ , ( $j = 1, \dots, P$ ) are parameters of the social welfare function.

Consider the case where for each  $\tau \in \{\frac{1}{P}, \frac{2}{P}, \dots, 1\}$ ,  $a_{1,\tau} = a_{0,\tau} = a_\tau$ . This is a fairly intuitive case: The weights on the social welfare function are the same whether or not the treatment is implemented. In this case, the decision to run a job training program would be consistent with the following inequality:

$$V_1 - V_0 = \sum_{j=1}^P a_{\frac{j}{P}} (q_{1,\frac{j}{P}} - q_{0,\frac{j}{P}}) \geq 0 \quad (12)$$

Equation (12) motivates the difference in quantiles as the main object of interest for the policy-maker. The decision to continue running the program depends crucially on the quantile treatment effects for all the quantiles of interest, that is, for all  $\tau$  such that  $a_\tau \neq 0$ .<sup>6</sup>

A particular case of Equation (12) would be when  $a_\tau = 0$  for all  $\tau$  but for one  $\tau'$ . This is the case, for example, when all the policy-maker is interested in is whether the training increases the earnings of those at the lower tail of the distribution.

---

<sup>6</sup>Note how this differs from the case in which the policy-maker wants to maximize the average outcome. In this case, the parameter of interest would simply be the average treatment effect.

Other types of social welfare functions would lead to the calculation of other treatment effect parameters. For example, say that  $V_1 = \frac{q_{1,0.25}}{q_{1,0.75}}$  and that  $V_0 = \frac{q_{0,0.25}}{q_{0,0.75}}$ . This is the case in which the policy-maker aims to run a job training program that decreases earnings inequality measured in a particular way. In this example, if  $V_1 - V_0 \geq 0$ , then the program reduces the gap between quantiles, that is, reduces earnings inequality.

### 3 IDENTIFICATION OF QUANTILE TREATMENT EFFECTS PARAMETERS

As potential outcomes are only partially observed, in order to identify from the observed data both  $\Delta_\tau$  and  $\Delta_{\tau|T=1}$ , the quantile treatment effects and the quantile treatment effects on the treated, we need an identification restriction. Instead of writing that restriction in terms of unobserved components (as in the previous section), I will start from a more general setting, in which we do not need to know the functional form of the potential outcomes. Let the propensity score,  $Pr[T = 1|X = x]$ , be written as  $p(x)$ , and its expectation,  $E[p(X)]$ , be written as  $p$ . Thus, the identification assumption used here, following Rosenbaum and Rubin, is:

**ASSUMPTION 1 (*Strong Ignorability - Rosenbaum and Rubin (1983)*):** For almost all values of  $X$ :

(i) **Unconfoundedness** :  $(Y(1), Y(0))$  is jointly independent from  $T$  given  $X$ ;

(ii) **Common Support**:  $c < p(x) < 1 - c$ , for some  $c > 0$

Although it is a strong assumption, many studies of the effect of treatments or programs make an assumption similar to that of part (i) of Assumption 1 as, for example, Heckman, Ichimura, Smith, and Todd (1998) and Dehejia and Wahba (1999). Alternatives to this assumption are the use of instrumental variables (the *selection on unobservables* approach), and calculation of bounds for the parameter of interest, as proposed by Manski (1997).<sup>7</sup> Part (ii)

---

<sup>7</sup>For review and comparison of approaches see, for instance, Angrist and Krueger (1999) and Heckman, Lalonde and Smith (2000).

states that for almost all values of  $X$  both treatment assignment levels have a positive probability of occurring. Under Assumption 1 both the overall quantile treatment effect and the quantile treatment effect on the treated become estimable from the data on  $(Y, T, X)$ . To show this, I first prove that the quantiles of the potential outcome distributions can be written as implicit functions of the observed data:

LEMMA 1 (*Identification of Quantiles*): Under Assumption 1, the following equalities hold:<sup>8</sup>

$q_{1,\tau}$ :

$$\begin{aligned}
 \tau &= \\
 (Q1_A) \quad &= E[\Pr\{Y \leq q_{1,\tau} | X, T = 1\}] \\
 (Q1_B) \quad &= E\left[\frac{E[T \mathbb{I}\{Y \leq q_{1,\tau}\} | X]}{p(X)}\right] \\
 (Q1_C) \quad &= E\left[\frac{T \mathbb{I}\{Y \leq q_{1,\tau}\}}{p(X)}\right]
 \end{aligned}$$

Lemma 1 shows that there are multiple ways of expressing each quantile of the potential outcome distributions in terms of the observed data  $(Y, T, X)$ .<sup>9</sup> In fact, the lemma shows that there at least three ways of identifying the quantiles using the observed data  $(Y, T, X)$ . These are divided into three groups denoted by  $A$ ,  $B$  and  $C$  (which are the indices for each expression in Lemma 1). Each group will differ according to the number and type of conditional expectations to be taken inside the expectation symbol.

In the first identification group, indexed by  $A$ , the computation of a conditional probability function in the first step is required. This function is the probability of  $Y$  being less than or equal to  $q$  given that  $X = x$  and  $T = 1$ . Taking the expectation over all  $x \in X$  for the treated subset ( $T = 1$ ) yields the desired result:  $q_{1,\tau}$  will be the quantity that sets the expected value equal to  $\tau$ .

---

<sup>8</sup>The indicator function  $\mathbb{I}\{A\}$  is equal to one if  $A$  is true and zero otherwise.

<sup>9</sup>An analogous result for  $q_{0,\tau}$  would follow directly from Lemma 1.

The equation indexed by  $B$  also requires computation of a conditional expectation in the first step. However, as this conditional expectation function is not restricted to the subset of treated units, one needs to divide by the probability of being treated given  $X = x$  (the “propensity score”). Notice then, that the first step involves two conditional expectations computations. This is the price paid for not restricting computation to the subset of treated units. Also, as in expression  $A$ , in expression  $B$   $q_{1,\tau}$  will be the quantity that sets the expected value of the ratio of conditional functions equal to  $\tau$ .

Finally, expression  $C$  is the simplest of the three. The first step requires computation of just one conditional expectation function, namely, the propensity score. Notice, that expression  $A$  also requires just one conditional expectation computation in the first step. The main difference lies in the role that the quantile  $q$  plays. In  $A$  one first has to compute  $\kappa_1(x; q) = E[\mathbb{1}\{Y \leq q\} | X = x, T = 1]$ . This function does not simply depend on  $(y, t, x)$ , because the quantile  $q$  enters as an argument, complicating computation.<sup>10</sup> This is different for expression  $C$ . In  $C$ , the p-score computation does not involve  $q$ ; in fact, it does not involve the random variable  $Y$  nor any functional of its distribution. Finally, to get  $q_{1,\tau}$ , one needs to proceed as in the other steps and compute an unconditional expectation.

As Lemma 1 does not directly yield a way to identify the quantiles of the potential outcomes for the actual treated units, it is necessary to postulate another set of results for that special case:

**LEMMA 2 (*Identification of Quantiles for the Treated*):** *Under Assumption 1, the following two sets of equalities hold:*

---

<sup>10</sup>However, as we will see in a later section, this does not have a real impact on the estimation procedure for  $q_{1,\tau}$  based on expression  $A$ . This is due to the fact that we are able to estimate a quantile by a minimization procedure that does not involve  $q$  in the first step.

$q_{1,\tau|T=1}$ :

$$\begin{aligned}
\tau &= \\
(QT1_A) \quad &= E \left[ \frac{p(X)Pr[Y \leq q_{1,\tau|T=1} | X, T = 1]}{p} \right] \\
(QT1_B) \quad &= E \left[ \frac{E[T \mathbb{I}\{Y \leq q_{1,\tau|T=1}\} | X]}{p} \right] \\
(QT1_C) \quad &= E \left[ \frac{T \mathbb{I}\{Y \leq q_{1,\tau|T=1}\}}{p} \right]
\end{aligned}$$

$q_{0,\tau|T=1}$ :

$$\begin{aligned}
\tau &= \\
(QT0_A) \quad &= E \left[ \frac{p(X)Pr[Y \leq q_{0,\tau|T=1} | X, T = 0]}{p} \right] \\
(QT0_B) \quad &= E \left[ \frac{p(X)}{(1-p(X))p} E[(1-T) \mathbb{I}\{Y \leq q_{0,\tau|T=1}\} | X] \right] \\
(QT0_C) \quad &= E \left[ \frac{p(X)}{(1-p(X))p} (1-T) \mathbb{I}\{Y \leq q_{0,\tau|T=1}\} \right]
\end{aligned}$$

**Proof:** See appendix

In the proof in the appendix, one can see that Assumption 1 plays no role in the identification of  $q_{1,\tau|T=1}$ . Heckman, Ichimura, and Todd (1997) have stressed such result when looking for identification conditions for the average treatment effects on the treated.

Identification of the quantile treatment effect parameters parameters is a straightforward consequence of Lemmas 1 and 2, as stated in the next corollary.

**COROLLARY 1 (*Identification of quantile treatment effect parameters*):** *Under Assumption 1, the quantile treatment effect,  $\Delta_\tau$ , and the quantile treatment effect on the treated,  $\Delta_{\tau|T=1}$ , are identified from data on  $(Y, T, X)$ .*

**Proof:** Note that from Lemmas 1 and 2 the four parameters  $q_{1,\tau}$ ,  $q_{0,\tau}$ ,  $q_{1,\tau|T=1}$ , and  $q_{0,\tau|T=1}$  are functionals of the joint distribution of  $(Y, T, X)$ . As  $\Delta_\tau$  equals the difference between  $q_{1,\tau}$  and  $q_{0,\tau}$ ; and  $\Delta_{\tau|T=1}$  equals the difference between  $q_{1,\tau|T=1}$  and  $q_{0,\tau|T=1}$ ,  $\Delta_\tau$  and  $\Delta_{\tau|T=1}$  are

also functionals of the joint distribution of  $(Y, T, X)$ . Therefore,  $\Delta_\tau$  and  $\Delta_{\tau|T=1}$ , are identified from data on  $(Y, T, X)$ .  $\square$

For  $\Delta_{\tau|T=1}$ , the method given by group *A* requires the computation of the p-score in addition to the computation of one conditional expectation given  $T = 1$  and  $X$  for  $(QT1_A)$ , and another conditional expectation given  $T = 0$  and  $X$  for  $(QT0_A)$ . The method in group *B* requires computing one conditional expectation given  $X$  for  $(QT1_B)$  and computing another conditional expectation as well as the p-score for  $(QT0_B)$ . Finally, for group *C* all that it is required is the p-score computation for  $(QT0_C)$ . Notice that the expectation of the p-score,  $p$ , is required for all three groups.

A comparison between Lemmas 1 and 2 reveals the presence of an interesting asymmetry in the former but not in the latter. Using procedures *B* and *C*, the computation of  $q_{1,\tau|T=1}$  requires a fewer first step calculations of conditional functions than the computation of  $q_{0,\tau|T=1}$ . This difference does not hold for  $q_{1,\tau}$  versus  $q_{0,\tau}$ , since the computation of these are symmetric and both computations involve the same number and sort of functionals.

From an estimation point of view the classification of these three groups of methods is relevant not only for the QTE, but for mean-based measures, such as the ATE, as well. Using sample analogues, Hahn (1998) has suggested estimation of the ATE based on an identifying approach similar to that described by *B*. Dehejia and Wahba (1999) proposed (among other techniques) estimating the average treatment effect on the treated by reweighing the control sample using the estimated p-score; this is analogous to the identification set *C*. Hirano, Imbens and Ridder (2002), going into more detail, have also focused on the estimation of ATE using the analogue of the set *C* for identification.

Estimation of the quantile treatment effect on the treatment based on the set *C* of identifying assumptions has been implicit in the applied literature. DiNardo, Fortin, and Lemieux (1996) proposed estimation of the counterfactual density of outcomes for the control group using a method similar to  $(QT0_C)$ . They argue in a footnote that, once the counterfactual density

is estimated, it is possible to recover the counterfactual quantiles and therefore the difference between the quantiles of the treated group and the counterfactual quantiles of the control group. However, as is made clear by expression  $(QT0_C)$ , there is no need to first compute densities if the ultimate goal is the estimation of quantiles.

In Section 5 of this paper I present the estimation counterparts of all three sets of equations for both the overall quantile treatment effect and the quantile treatment effect on the treated.

## 4 SEMIPARAMETRIC EFFICIENCY BOUNDS

As Lemmas 1 and 2 suggest, estimation of quantile treatment effects can be attempted using two-step procedure, where the first step is a non-parametric estimation of a conditional expectation function. The preliminary step must be non-parametric since the joint distribution of  $(Y(0), Y(1))$  is not parametrically specified. Semiparametric estimation for the ATE can be found in Hahn (1998), Heckman, Ichimura, Smith, and Todd (1998) and Hirano, Imbens and Ridder (2002).

A semiparametric analog of the Cramer-Rao lower bound was first introduced by Stein (1956) and further developed by Begun, Hall, Huang and Wellner (1983) and by Bickel, Klassen, Ritov, and Wellner (1993). The semiparametric efficiency bound concept was popularized in the econometric literature by a review article by Newey (1990). In general terms, the bound corresponds to the largest variance over all possible regular parametric specifications of the nonparametric component of the model. Such bound is indeed a (not necessarily achievable) lower bound for the asymptotic variance of distribution-free, root- $N$  consistent estimators.

More formally, consider a finite-dimensional parameter  $\zeta$  from some general statistical model. Say that this model contains a submodel that can be parameterized by a finite-dimensional parameter  $\theta$ . Thus, for this submodel we write  $\zeta(\theta)$ . If this parameter is *differentiable* in the sense described by Bickel, Klassen, Ritov, and Wellner (1993), then its derivative with respect

to  $\theta$  can be written as  $E[\psi s'_\theta]$ , where  $\psi$  is the influence function of  $\zeta$  and  $s_\theta$  is the score of that submodel. The semiparametric efficiency bound  $V_\zeta$  will be equal to  $E[\psi'_\theta \psi_\theta]$ , where  $\psi_\theta$  is equal to  $E[\psi s'_\theta](E[s_\theta s'_\theta])^{-1} s_\theta$ , the “projection” onto the space spanned by all scores.

Hahn (1998) uses the setup described above to compute the semiparametric efficiency bounds for both the average treatment effect,  $\beta$ , and the average treatment effect on the treated,  $\gamma$ . For the quantile treatment effects setting, I also compute bounds for two parameters, namely,  $\Delta_\tau$  and  $\Delta_{\tau|T=1}$ . For QTE parameters however, computation of such bounds requires an additional regularity condition, which is assumed here to hold:

**ASSUMPTION 2 (*Local Identification*):** For  $j = 0, 1$ , define  $F_j(q)$  and  $F_{j|T=1}(q)$  as being respectively  $Pr[Y(j) \leq q]$  and  $Pr[Y(j) \leq q|T = 1]$ . Then:

- (i)  $\left. \frac{\partial F_j(q)}{\partial q} \right|_{q=q_{j,\tau}} = f_j(q_{j,\tau}) > 0$ ; and
- (ii)  $\left. \frac{\partial F_{j|T=1}(q)}{\partial q} \right|_{q=q_{j,\tau|T=1}} = f_{j|T=1}(q_{j,\tau|T=1}) > 0$

Assumption 2 guarantees that the distribution functions of the potential outcomes are not flat at the  $\tau$ -percentile. This is equivalent to assuming the uniqueness of the respective quantiles.

With Assumptions 1 and 2, the semiparametric efficiency bounds for  $\Delta_\tau$  and  $\Delta_{\tau|T=1}$  can be calculated:

**THEOREM 1 : (*Bounds for  $\Delta_\tau$  and  $\Delta_{\tau|T=1}$* ):** Under Assumptions 1 and 2, the semiparametric efficiency bounds for  $\Delta_\tau$  and  $\Delta_{\tau|T=1}$  are respectively equal to:

$$V_{\Delta_\tau} = E \left[ \frac{E[g_{1,\Delta_\tau}^2(Y, X)|X, T = 1]}{p(X)} + \frac{E[g_{0,\Delta_\tau}^2(Y, X)|X, T = 0]}{1 - p(X)} + (h_{1,\Delta_\tau}(X) - h_{0,\Delta_\tau}(X))^2 \right] \quad (13)$$

and

$$V_{\Delta\tau|T=1} = E \left[ \frac{p(X)E[g_{1,\Delta\tau|T=1}^2(Y,X)|X,T=1]}{p^2} + \frac{p^2(X)E[g_{0,\Delta\tau|T=1}^2(Y,X)|X,T=0]}{p^2(1-p(X))} + \frac{p(X)(h_{1,\Delta\tau|T=1}(X) - h_{0,\Delta\tau|T=1}(X))^2}{p^2} \right] \quad (14)$$

where for  $j = 0, 1$ :

$$g_{j,\Delta\tau}(Y,X) = - \left( \frac{\mathbb{I}\{Y \leq q_{j,\tau}\} - \tau - E[\mathbb{I}\{Y(j) \leq q_{j,\tau}\} - \tau | X]}{f_j(q_{j,\tau})} \right), \quad (15)$$

$$h_{j,\Delta\tau}(Y,X) = - \left( \frac{E[\mathbb{I}\{Y(j) \leq q_{j,\tau}\} - \tau | X]}{f_j(q_{j,\tau})} \right), \quad (16)$$

$$g_{j,\Delta\tau|T=1}(Y,X) = - \left( \frac{\mathbb{I}\{Y \leq q_{j,\tau|T=1}\} - \tau - E[\mathbb{I}\{Y(j) \leq q_{j,\tau|T=1}\} - \tau | X, T=1]}{f_{j|T=1}(q_{j,\tau|T=1})} \right), \quad (17)$$

and

$$h_{j,\Delta\tau|T=1}(Y,X) = - \left( \frac{E[\mathbb{I}\{Y(j) \leq q_{j,\tau|T=1}\} - \tau | X, T=1]}{f_{j|T=1}(q_{j,\tau|T=1})} \right) \quad (18)$$

**Proof:** See appendix

Note that the bounds  $V_{\Delta\tau}$  and  $V_{\Delta\tau|T=1}$  are similar to the bounds computed by Hahn (1998) for the mean case. For  $\beta$  and  $\gamma$  the bounds, as computed by Hahn (1998), are respectively:

$$V_{\beta} = E \left[ \frac{V[Y|X,T=1]}{p(X)} + \frac{V[Y|X,T=0]}{1-p(X)} + (E[Y|X,T=1] - E[Y|X,T=0] - \beta)^2 \right]$$

and

$$V_\gamma = E \left[ \frac{p(X)V[Y|X, T=1]}{p^2} + \frac{p(X)^2V[Y|X, T=0]}{p^2(1-p(X))} + \frac{p(X)(E[Y|X, T=1] - E[Y|X, T=0] - \gamma)^2}{p^2} \right].$$

There are two reasons for the similarity between the semiparametric efficiency bounds of the QTE and the ATE parameters. First, both the QTE and the ATE are parameters from the same statistical model and, therefore, can be expressed as functionals of the same distribution of the data. But this is not enough for the similarity. In fact, the second reason is the important one: both the QTE and the ATE are *expectations* of different random variables but over the same density. Note also that the difference in the random variables is what determines the difference in the bounds.

The role of the propensity score in efficient estimation of ATE has received a great deal of attention in the recent literature. Examples include Heckman, Ichimura, Smith and Todd (1998), Hahn (1998) and Hirano, Imbens, and Ridder (2002). The latter provide intuition for Hahn's result that knowing the true propensity score does not lead to efficient estimation of the ATE. For the QTE parameters the same results apply since both cases share the same statistical model, and thus the propensity score plays the same role. Because of this similarity, this result will not be explored in this paper.

## 5 EFFICIENT ESTIMATION

Once we know which parameters we want to estimate and we know the minimum attainable asymptotic variance of any semiparametric estimator, we can propose candidates for estimation. In this section I use the *sample analogy principle*<sup>11</sup> to come up with estimators of  $\Delta_\tau$  and  $\Delta_{\tau|T=1}$ . I show that these estimators are in fact solutions to minimization problems, and this facilitates both the computation and the derivation of their large sample properties. Restricting then attention to one of the estimators, I present its large sample properties and also show

---

<sup>11</sup>See for instance, Manski (1988)

that the asymptotic variance of the proposed estimator achieves the semiparametric efficiency bound.

## 5.1 SAMPLE ANALOG APPROACH

According to Lemmas 1 and 2 there are at least three ways of identifying the quantiles of the potential outcome distribution. From the sets  $A$ ,  $B$  and  $C$  of identification expressions, it is possible to derive sample analogs that can be used to estimate both  $\Delta_\tau$  and  $\Delta_\tau|_{T=1}$ . The three sets will differ among themselves by the number and type of conditional expectations functions to be non-parametrically estimated in a first step. As a piece of notation, let the first step estimators of functionals of  $(Y, T, X)$  be denoted by a “hat” on it. For example, the nonparametric estimator of the p-score, will be  $\hat{p}(x)$ . For a random  $\varepsilon_N$  that converges to zero with probability one at an appropriate rate, let the three estimators of  $\Delta_\tau$  be:

$$\hat{\Delta}_\tau^A = \hat{q}_{1,\tau}^A - \hat{q}_{0,\tau}^A \quad (19a)$$

$$\hat{\Delta}_\tau^B = \hat{q}_{1,\tau}^B - \hat{q}_{0,\tau}^B \quad (19b)$$

$$\hat{\Delta}_\tau^C = \hat{q}_{1,\tau}^C - \hat{q}_{0,\tau}^C \quad (19c)$$

where:

$$\frac{1}{N} \sum_{i=1}^N \hat{E}[\mathbf{1}\{Y \leq \hat{q}_{1,\tau}^A\} - \tau | X_i, T = 1] \leq \varepsilon_N \quad \text{and} \quad \frac{1}{N} \sum_{i=1}^N \hat{E}[\mathbf{1}\{Y \leq \hat{q}_{0,\tau}^A\} - \tau | X_i, T = 0] \leq \varepsilon_N; \quad (20a)$$

$$\frac{1}{N} \sum_{i=1}^N \frac{\hat{E}[T(\mathbf{1}\{Y \leq \hat{q}_{1,\tau}^B\} - \tau) | X_i]}{\hat{p}(X_i)} \leq \varepsilon_N \quad \text{and} \quad \frac{1}{N} \sum_{i=1}^N \frac{\hat{E}[(1-T)(\mathbf{1}\{Y \leq \hat{q}_{0,\tau}^B\} - \tau) | X_i]}{1 - \hat{p}(X_i)} \leq \varepsilon_N; \quad (20b)$$

and

$$\frac{1}{N} \sum_{i=1}^N \frac{T_i(\mathbb{1}\{Y_i \leq \hat{q}_{1,\tau}^C\} - \tau)}{\hat{p}(X_i)} \leq \varepsilon_N \quad \text{and} \quad \frac{1}{N} \sum_{i=1}^N \frac{(1 - T_i)(\mathbb{1}\{Y_i \leq \hat{q}_{0,\tau}^C\} - \tau)}{1 - \hat{p}(X_i)} \leq \varepsilon_N. \quad (20c)$$

Let us concentrate on the estimation differences between the procedures  $A$ ,  $B$ , and  $C$  for  $q_{1,\tau}$  only, as estimation of  $q_{0,\tau}$  follows by analogy. First look at the difference between  $C$  and  $B$ : Procedure  $C$  is the simplest one, as the first step only requires estimation of the propensity score. In contrast, the first step for procedure  $B$  requires not only the p-score estimation but also the estimation of a more complicated function of the data,  $E[T(\mathbb{1}\{Y \leq q\} - \tau) | X = x]$ . Notice that this function depends not only on  $(Y, T, X)$ , but also on the quantile  $q$ , which enters as an argument. This would seem to apparently complicate actual estimation. However, as will be shown later in this section,  $\hat{q}_{1,\tau}^B$  can be found by an equivalent yet simpler two-step procedure. In that setup the estimated quantile minimizes a weighted sum in the second step, while in the first step the weights, which do not depend on  $q$ , are computed.

The main difference between procedures  $B$  and  $A$  is that in  $A$  just one conditional expectation function is computed. This is the same function estimated in the numerator of  $B$ , but estimation in  $A$  is restricted to data on the treated units, making information from the control group irrelevant. Also, as in  $B$ ,  $A$  can be expressed by a simpler but equivalent two-step procedure.

By an analogy the three estimators of  $\Delta_{\tau|T=1}$  could be found in the same way as for the overall case, and because of that, I do not present them here. Note, however that for the treated case, two differences are important: First, there is an asymmetry between the estimation of  $q_{1,\tau|T=1}$  and the estimation of  $q_{0,\tau|T=1}$ ; this results because we are estimating quantiles of the distribution of potential outcomes for the treated only. Second, the estimator of  $\hat{q}_{1,\tau|T=1}^C$  is a simple sample quantile, and there is no nonparametric first step associated with it.

## 5.2 MINIMIZATION APPROACH

Now abstract temporarily from the possible problem of having to estimate the p-score in a first step, as is the case, for instance, for the set  $C$  of estimators. It is straightforward to show that the previous estimators are indeed solutions to minimization problems, since sample quantiles can be found by minimizing a sum of check functions.<sup>12</sup>

In order to simplify the argument, let us first focus on the estimation of  $\Delta_\tau$ , starting with the simplest estimation technique  $C$ , and concentrating on the sample quantile of the  $Y(1)$  distribution,  $\hat{q}_{1,\tau}^C$ , which solves  $\frac{1}{N} \sum_{i=1}^N \frac{T_i(\mathbb{1}\{Y_i \leq \hat{q}_{1,\tau}^C\} - \tau)}{\hat{p}(X_i)} \leq \epsilon_N$ . This sample quantile can equivalently be written as the minimizer of a weighted sum, where the weight of each unit is given by:

$$\hat{\omega}_{1,i}^C = \frac{T_i}{N\hat{p}(X_i)} \quad (21)$$

The sample quantile is equal to:

$$\hat{q}_{1,\tau}^C = \arg \min_q \sum_{i=1}^N \hat{\omega}_{1,i}^C \rho_\tau(Y_i - q) \quad (22a)$$

where the check function  $\rho_\tau(\cdot)$  evaluated at  $Y_i - q$  is:

$$\rho_\tau(Y_i - q) = (Y_i - q)(\tau - \mathbb{1}\{Y_i \leq q\})$$

There are two advantages to writing  $\hat{\Delta}_\tau^C$  as the difference between two minimizers: First, in computational terms,  $\hat{\Delta}_\tau^C$  is the difference between the solutions of two simple linear programs. Second, the asymptotic properties of minimizers of convex random functions are well established in the statistical and econometric literature.

The equivalence result just described for the set  $C$  of estimators can be extended for the other two sets of estimators,  $A$  and  $B$ . In fact, all three estimators can be seen as solutions to weighted quantile problems, as shown in the next paragraphs.

---

<sup>12</sup>See, for instance, Koenker and Bassett (1978).

In the analogy approach, for the set  $A$ , it is necessary to estimate in the first step the conditional expectation  $m_1^A(x|q) = E[\mathbb{I}\{Y \leq q\} - \tau | X = x, T = 1]$  by  $\hat{m}_1^A(x|q) = \hat{E}[\mathbb{I}\{Y \leq q\} - \tau | X = x, T = 1]$ .

Note that this estimation problem can be written as:

$$\begin{aligned}\hat{m}_1^A(x|q) &= \hat{E}[\mathbb{I}\{Y \leq q\} - \tau | X = x, T = 1] \\ &= \hat{E}[T \mathbb{I}\{Y \leq q\} - \tau | X = x, T = 1] \\ &= \sum_{i=1}^N \hat{v}_i^A(x) (\mathbb{I}\{Y_i \leq q\} - \tau)\end{aligned}$$

where  $\hat{v}_i^A$  is a weight that is chosen according to the choice of non-parametric estimation technique. For example, suppose that for the non-parametric estimation we use a smoothing function  $K_h(\cdot)$ , which is equal to  $h^{-k}K(\cdot/h)$  and where  $K(\cdot)$  is a kernel function and  $h$  is a bandwidth. Then:

$$\hat{v}_i^A = \frac{K_h(X_i - x)T_i}{\sum_{l=1}^N K_h(X_l - x)T_l}$$

The unconditional expectation function,  $E[m_1^A(X|q)]$ , can be estimated by  $\frac{1}{N} \sum_{j=1}^N \hat{m}_1^A(X_j|q)$ .

But this expression can be rewritten as:

$$\begin{aligned}\frac{1}{N} \sum_{j=1}^N \hat{m}_1^A(X_j|q) &= \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^N \hat{v}_i^A(X_j) (\mathbb{I}\{Y_i \leq q\} - \tau) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \hat{v}_i^A(X_j) (\mathbb{I}\{Y_i \leq q\} - \tau)\end{aligned}$$

Now, define  $\hat{\omega}_{1,i}^A$  as being equal to  $\frac{1}{N} \sum_{j=1}^N \hat{v}_i^A(X_j)$ . Then:

$$\frac{1}{N} \sum_{j=1}^N \hat{m}_1^A(X_j|q) = \sum_{i=1}^N \hat{\omega}_{1,i}^A (\mathbb{I}\{Y_i \leq q\} - \tau)$$

And:

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^N \hat{m}_1^A(X_j | \hat{q}_{1,\tau}^A) &= \sum_{i=1}^N \hat{\omega}_{1,i}^A (\mathbb{I}\{Y_i \leq \hat{q}_{1,\tau}^A\} - \tau) \\ &= \frac{1}{N} \sum_{i=1}^N \hat{E}[\mathbb{I}\{Y \leq \hat{q}_{1,\tau}^A\} - \tau | X_i, T = 1] \leq \epsilon_N \end{aligned}$$

Thus, finally:

$$\hat{q}_{1,\tau}^A = \arg \min_q \sum_{i=1}^N \hat{\omega}_{1,i}^A \rho_\tau(Y_i - q) \quad (22b)$$

Analogously, it is possible to rewrite the first non-parametric step of the approach *B* and get  $\hat{q}_{1,\tau}^B$  as the minimizer of a sum of check functions:

$$\hat{q}_{1,\tau}^B = \arg \min_q \sum_{i=1}^N \hat{\omega}_{1,i}^B \rho_\tau(Y_i - q) \quad (22c)$$

where  $\hat{\omega}_{1,i}^B$  is equal to  $\frac{1}{N} \sum_{j=1}^N \hat{v}_i^B(X_j)$ . For the example where the conditional expectation is estimated by a kernel  $K$  with bandwidth  $h$ :

$$\hat{v}_i^B = \frac{K_h(X_i - x) T_i}{\sum_{l=1}^N K_h(X_l - x) \hat{p}(x)}$$

As an interesting by-product, note that if the kernel function and the bandwidth are exactly the same for the cases *A* and *B*, then the weights  $\hat{\omega}_{1,i}^A$  and  $\hat{\omega}_{1,i}^B$  must be equal. Also note that these weights,  $\hat{\omega}_{1,i}^A$  and  $\hat{\omega}_{1,i}^B$ , sum to 1 over  $i$ , regardless of whether they are estimated using kernel smoothing or using some other non-parametric estimation technique.

Completing the procedure of writing the Equations 19a to 19c as differences of minimizers, note that  $\hat{q}_{0,\tau}^A$ ,  $\hat{q}_{0,\tau}^B$ , and  $\hat{q}_{0,\tau}^C$  can be written as

$$\hat{q}_{0,\tau}^E = \arg \min_q \sum_{i=1}^N \hat{\omega}_{0,i}^E \rho_\tau(Y_i - q) \quad (23)$$

for some appropriate choice of  $\hat{\omega}_{0,i}^E$ , where  $E \in \{A, B, C\}$ . As a matter of fact,  $\hat{\omega}_{0,i}^C$  is equal to:

$$\hat{\omega}_{0,i}^C = \frac{1 - T_i}{N(1 - \hat{p}(X_i))} \quad (24)$$

The same line of reasoning can be applied to the three estimators of  $\Delta_{\tau|T=1}$ . I omit the formal proof that they are differences between minimizers. However, note that for the estimation procedure indexed by  $C$ , the sample quantiles can be written as:

$$\hat{q}_{j,\tau|T=1}^C = \arg \min_q \sum_{i=1}^N \hat{\omega}_{j,i|T=1}^C \rho_{\tau}(Y_i - q), \quad j = 0, 1. \quad (25)$$

where the weights are equal to:

$$\hat{\omega}_{1,i|T=1}^C = \frac{T_i}{\sum_{l=1}^N T_l} \quad \text{and} \quad \hat{\omega}_{0,i|T=1}^C = \frac{\frac{\hat{p}(X_i)}{1 - \hat{p}(X_i)}(1 - T_i)}{\sum_{l=1}^N T_l} \quad (26)$$

This result will be used later in the paper.

### 5.3 FEASIBLE ESTIMATION

For the remainder of the paper, I shall restrict the discussion to estimators that use the set  $C$  of identifying equations. As argued before, these are the simplest estimators.

I focus only on  $\hat{q}_{1,\tau}^C$  since extensions for  $\hat{q}_{0,\tau}^C$  and for  $\hat{\Delta}_{\tau|T=1}^C$  follow immediately.

The estimator  $\hat{q}_{1,\tau}^C$  is a two-step estimator. In the first step, we estimate the p-score non-parametrically. In the second stage, we minimize:

$$G_N(q, \hat{p}) = \frac{1}{N} \sum_{i=1}^N \frac{T_i}{\hat{p}(X_i)} (Y_i - q)(\tau - \mathbb{I}\{Y_i \leq q\}) \quad (27)$$

My specific methods were as follows: To estimate the p-score, I used a logistic power series approximation, i.e., the log odds ratio of the p-score was approximated by a series of

functions.<sup>13</sup> These functions were chosen to be polynomials of  $x$  and the coefficients corresponding to those functions were estimated by maximum likelihood.

Start by defining  $H(x) = [H_j(x)]$  ( $j = 1, \dots, K$ ), a vector of length  $K$  of polynomial functions of  $x$  satisfying the following properties:

- (i)  $H : \mathcal{X} \rightarrow \mathbb{R}^K$ ;
- (ii) (Orthogonal array of polynomials)  $E[H_j(X)H_l(X)] = \mathbb{1}\{j = l\}$ , ( $j, l = 1, \dots, K$ );
- (iii) (Constant included)  $H_1(x) = 1$ ,  $\forall x \in \mathcal{X}$  with probability one.

If we want  $H(x)$  to include polynomials of  $x$  up to the order  $n$ , then it is sufficient to choose  $K$  such that  $K \geq (n+1)^r$ . In what follows, I will assume that  $K$  is a function of the sample size and grows without bounds as  $N$  grows without bounds, that is,  $K = K(N) \rightarrow \infty$  as  $N \rightarrow \infty$ .

Next, the propensity score is estimated. Let  $\hat{p}(x)$  be:

$$\hat{p}(x) = L(H(x)' \hat{\pi}) \tag{28}$$

where  $L : \mathbb{R} \rightarrow \mathbb{R}$ ,  $L(z) = (1 + \exp(-z))^{-1}$

and

$$\hat{\pi} = \arg \max_{\pi} \frac{1}{n} \sum_{i=1}^N \left\{ T_i \ln(L(H(X_i)' \pi)) + (1 - T_i) \ln(1 - L(H(X_i)' \pi)) \right\} \tag{29}$$

Thus, after estimating the p-score, I can minimize  $G_N(q, \hat{p})$  with respect to  $q$ , obtaining  $\hat{q}_{1,\tau}^C$ .

## 5.4 LARGE SAMPLE PROPERTIES

In this subsection I will prove that  $\hat{q}_{1,\tau}^C$  is (i) root- $N$  consistent for  $q_{1,\tau}$ ; (ii) asymptotically normal; and (iii) has asymptotic variance equal to the expected square of the efficient influence

---

<sup>13</sup>The log odds ratio of  $p(x)$  is equal to  $\ln(p(x)) - \ln(1 - p(x))$ .

function of  $q_{1,\tau}$ .<sup>14</sup>

This subsection is divided into several parts:

1. I show that the non-parametric estimation of the p-score in the first step by means of a power series approximation yields a uniformly consistent estimator of  $p(x)$ . The regularity conditions for this result are stated and the result proved. Also it is shown that the estimator  $\hat{p}(x)$  is bounded in probability from 0 and 1.
2. I define  $Q_N(t, \hat{p})$ , where, for any  $q$ ,  $t = q - q_{1,\tau}$  and show that  $Q_N(t, \hat{p})$  is minimized by  $\hat{t} = \hat{q}_{1,\tau}^C - q_{1,\tau}$ .
3. I show how the objective function  $Q_N(t, \hat{p})$ , which is random and convex in  $t$ , can be written as the sum of a quadratic random function of  $t$ ,  $\tilde{Q}_N(t)$  and a remainder term.
4. I define  $u = \sqrt{N}t$  and  $\hat{u} = \sqrt{N}\hat{t}$  and show that for each  $u$ , under the regularity conditions that yielded the uniform consistency of  $\hat{p}(x)$ ,  $N(Q_N(u/\sqrt{N}, \hat{p}) - \tilde{Q}_N(u/\sqrt{N}))$  goes to zero in probability.
5. I show that under some additional mild requirements  $\tilde{u}$ , the argument that minimizes the random quadratic  $N\tilde{Q}_N(u/\sqrt{N})$ , is: (i)  $O_p(1)$ ; and (ii)  $\tilde{u} = N(0, V_1) + o_p(1)$ , where  $V_1$  is the semiparametric efficiency bound of  $q_{1,\tau}$ .
6. I show that  $\hat{u} - \tilde{u} = o_p(1)$ , or written in terms of  $q$ , that  $\hat{q}_{1,\tau}^C$  is asymptotically equivalent to  $\tilde{q}_{1,\tau} = \tilde{u}/\sqrt{N} + q_{1,\tau}$ , which establishes the desired result.

#### 5.4.1 UNIFORM CONSISTENCY OF THE FIRST STEP

The suggested approach to estimating the p-score guarantees, under certain regularity conditions, that  $\hat{p}(x)$ , the estimator of the p-score, is uniformly consistent for the true  $p(x)$ . To assure that this holds, I make the following assumptions:

---

<sup>14</sup>Thus, as  $\Delta_\tau = q_{1,\tau} - q_{0,\tau}$  and as it can be shown by analogy that  $\hat{q}_{0,\tau}^C$  equally satisfies the properties (i), (ii) and a properly modified version of (iii), the efficient influence function of  $\Delta_\tau$  will be equal to the difference between the efficient influence function of  $q_{1,\tau}$  and  $q_{0,\tau}$ .

ASSUMPTION 3 (*First Step*):

- (i)  $\mathcal{X}$  is a compact subset of  $\mathbb{R}^r$ ;
- (ii)  $p(x)$  is  $s$ -times continuously differentiable;
- (iii) (*Common Support*)  $0 < c < p(x) < 1 - c \quad \forall x \in \mathcal{X}$ ;
- (iv)  $f(x) > 0 \quad \forall x \in \mathcal{X}$ ;
- (v)  $\frac{\zeta^4(N)}{N} = o_p(1)$ , where  $\zeta(N) = \sup_{x \in \mathcal{X}} \|H(x)\|$ .

Newey (1995) has established that for orthogonal polynomials  $H(x)$  and compact  $\mathcal{X}$ :<sup>15</sup>

$$\sup_{x \in \mathcal{X}} \|H(x)\| = CK \quad (30)$$

where  $C$  is a generic constant. Note then that  $\zeta$  will be a function of  $N$  since  $K$  is assumed to be a function of  $N$ .

With Assumption 3 in hand we can invoke some of the results derived by Hirano, Imbens and Ridder (2002) in a format of a lemma:

LEMMA 3 (*First Step*): Under Assumptions 1 and 3 the following results hold:

(I)  $\sup_{x \in \mathcal{X}} |p(x) - p_K(x)| = C\zeta(N)K^{-s/r}(N) = C\zeta^{1-s/r}(N)$ ; where  $p_K(x) = L(H(x)'\pi_K)$  and:

$$\pi_K = \arg \max_{\pi} E \left\{ p(X) \ln(L(H(X)'\pi)) + (1 - p(X)) \ln(1 - L(H(X)'\pi)) \right\}; \quad (31)$$

(II)  $E \|\hat{\pi} - \pi_K\|^2 = C \frac{\zeta(N)}{N}$ ;

(III) (a)  $\lim_{N \rightarrow \infty} Pr[\inf_{X \in \mathcal{X}} \hat{p}(X) > \delta] = 1$ ; and (b)  $\lim_{N \rightarrow \infty} Pr[\sup_{X \in \mathcal{X}} \hat{p}(X) < 1 - \delta] = 1$ ;

(IV) For  $\tilde{\pi} \in [\hat{\pi}, \pi_K]$ ,  $L'(H(x)'\tilde{\pi}) > 0$  where  $L'(z) = \frac{dL(z)}{dz}$ .<sup>16</sup>

**Proof:** See Hirano, Imbens and Ridder (2002).

<sup>15</sup>See also Newey (1997).

<sup>16</sup>Note that  $L'(z) = L(z)(1 - L(z))$ , yielding then that  $\sup_z L'(z) = 1/4$ . Also note that  $L''(z) = L'(z)(1 - 2L(z))$ .

Results (I), (II) and (IV) plus an extra requirement on  $s$ , the differentiability order of  $p(x)$ , assure that  $\hat{p}(x)$  will converge to  $p(x)$  uniformly in probability:

$$\begin{aligned}
Pr \left[ \sup_{x \in \mathcal{X}} |\hat{p}(x) - p(x)| > \delta \right] &\leq Pr \left[ \sup_{x \in \mathcal{X}} |L(H(X)'\hat{\pi}) - L(H(x)'\pi_K)| > \delta \right] + Pr \left[ \sup_{x \in \mathcal{X}} |p_K(x) - p(x)| > \delta \right] \\
&\leq Pr \left[ \sup_{x \in \mathcal{X}} |L'(H(x)'\tilde{\pi})H(x)'(\hat{\pi} - \pi_K)| > \delta \right] + Pr \left[ \sup_{x \in \mathcal{X}} |p_K(x) - p(x)| > \delta \right] \\
&\leq \frac{C_1}{4} \zeta(N) \sqrt{\frac{\zeta(N)}{N}} + C_2 \zeta^{1-s/r}(N)
\end{aligned} \tag{32}$$

Now for  $s > r$  and by Assumption 3 (v), we have:

$$\lim_{N \rightarrow \infty} Pr \left[ \sup_{X \in \mathcal{X}} |\hat{p}(X) - p(X)| > \delta \right] = 0 \tag{33}$$

Note the importance of result (III) in simplifying the whole process of estimating  $q_{1,\tau}$  by  $\hat{q}_{1,\tau}^C$ . As  $\hat{p}(x)$  is bounded in probability from 0 and 1, there is no need to use a trimming function in order to avoid dividing a number by zero. This nice feature of the first step estimator comes from the logit approximation.<sup>17</sup>

#### 5.4.2 CHANGE OF VARIABLES: $t$ AND $Q_N$

Let  $q_{1,\tau}$  be in  $Q$ , a subset of  $\mathbb{R}$ . Then notice that:

$$\begin{aligned}
\hat{q}_{1,\tau}^C &= \arg \min_{q \in Q} \frac{1}{N} \sum_{i=1}^N \frac{T_i}{\hat{p}(X_i)} (Y_i - q)(\tau - \mathbb{1}\{Y_i \leq q\}) \\
&= \arg \min_{q \in Q} \frac{1}{N} \sum_{i=1}^N \frac{T_i}{\hat{p}(X_i)} \left[ (Y_i - q)(\tau - \mathbb{1}\{Y_i \leq q\}) - (Y_i - q_{1,\tau})(\tau - \mathbb{1}\{Y_i \leq q_{1,\tau}\}) \right] \\
&= \arg \min_{q \in Q} \frac{1}{N} \sum_{i=1}^N \frac{T_i}{\hat{p}(X_i)} \left[ (\mathbb{1}\{Y_i \leq q_{1,\tau}\} - \tau)(q - q_{1,\tau}) + (Y_i - q)(\mathbb{1}\{Y_i \leq q_{1,\tau}\} - \mathbb{1}\{Y_i \leq q\}) \right]
\end{aligned} \tag{34}$$

---

<sup>17</sup>As an example of a two-step estimator that requires this type of trimming, see for instance, Khan and Powell (2001).

Now, define:

$$t = q - q_{1,\tau}, \quad \forall q \in \mathcal{Q} \quad (35)$$

$$\mathcal{T} = \{t \in \mathbb{R} \mid t = q - q_{1,\tau}, q \in \mathcal{Q}\} \quad (36)$$

$$\hat{t} = \hat{q}_{1,\tau}^C - q_{1,\tau} \quad (37)$$

$$D_i = \mathbb{1}\{Y_i \leq q_{1,\tau}\} - \tau \quad (38)$$

$$R_i(t) = (Y_i - (q_{1,\tau} + t))(\mathbb{1}\{Y_i \leq q_{1,\tau}\} - \mathbb{1}\{Y_i \leq q_{1,\tau} + t\}) \quad (39)$$

$$A_i(t) = D_i t + R_i(t) \quad (40)$$

$$Q_N(t, \hat{p}) = \frac{1}{N} \sum_{i=1}^N \frac{T_i}{\hat{p}(X_i)} A_i(t) \quad (41)$$

A comment about some quantities above: The variable  $D$  is the approximate first derivative of  $G_N(q, \hat{p})$  with respect to  $q$ . It is approximate in the sense that  $G_N(q, \hat{p})$  is not differentiable for any  $q$ , as it involves indicator functions of whether  $q$  is less than or equal to some values in the data.  $R(t)$  can be interpreted as the remainder term from a linear expansion about zero that uses  $D$  as an approximated derivative.

Next, note that:

$$\hat{t} = \arg \min_{t \in \mathcal{T}} \frac{1}{N} \sum_{i=1}^N \frac{T_i}{\hat{p}(X_i)} \left[ (\mathbb{1}\{Y_i \leq q_{1,\tau}\} - \tau) t + (Y_i - (q_{1,\tau} + t)) (\mathbb{1}\{Y_i \leq q_{1,\tau}\} - \mathbb{1}\{Y_i \leq q_{1,\tau} + t\}) \right] \quad (42)$$

$$= \arg \min_{t \in \mathcal{T}} \frac{1}{N} \sum_{i=1}^N \frac{T_i}{\hat{p}(X_i)} [D_i t + R_i(t)] \quad (43)$$

$$= \arg \min_{t \in \mathcal{T}} \frac{1}{N} \sum_{i=1}^N \frac{T_i}{\hat{p}(X_i)} A_i(t) \quad (44)$$

$$= \arg \min_{t \in \mathcal{T}} Q_N(t, \hat{p}) \quad (45)$$

### 5.4.3 FIRST APPROXIMATION OF THE OBJECTIVE FUNCTION

I begin by defining some useful expressions: First, consider the random function  $\tilde{Q}_N(t)$ , which will be shown to be a quadratic approximation to  $Q_N(t, \hat{p})$ , but does not depend on the first step  $\hat{p}(X)$ :

$$\tilde{Q}_N(t) = \frac{1}{N} \sum_{i=1}^N t \left( \frac{T_i D_i}{p(X_i)} - E[D | X_i, T = 1] \frac{T_i - p(X_i)}{p(X_i)} \right) + \frac{f_1(q_{1,\tau})}{2} t^2 \quad (46)$$

We saw that  $A(t)$  can be decomposed into two parts,  $Dt$  and  $R(t)$ . Notice however, that we will be interested here in an approximation of  $\frac{1}{N} \sum_{i=1}^N \frac{T_i D_i t}{\hat{p}(X_i)}$  by an expression that does not depend on  $\hat{p}(X)$ . An approximation that is analogous to the one above that uses  $R(t)$  in the place of  $Dt$  will be shown to be bounded in probability. Let me define this approximation by  $\tilde{R}_N(t)$ :

$$\tilde{R}_N(t) = \frac{1}{N} \sum_{i=1}^N \frac{T_i R_i(t)}{p(X_i)} - E[R(t) | X_i, T = 1] \frac{T_i - p(X_i)}{p(X_i)} \quad (47)$$

Let me state the approximation result as a lemma:

LEMMA 4 (*First Approximation of the Objective Function*): Under Assumptions 1:

$$\begin{aligned} Q_N(t, \hat{p}) &= \tilde{Q}_N(t) + (\tilde{R}_N(t) - E[\tilde{R}_N(t) | T = 1]) + \varepsilon_{1,N}(t) + o(t^2) \\ &= \tilde{Q}_N(t) + \varepsilon_N(t) \end{aligned} \quad (48)$$

**Proof:** See appendix.

### 5.4.4 SECOND NORMALIZATION: $u = \sqrt{N}t$

Now, define:

$$u = \sqrt{N}t \quad (49)$$

Therefore, Equation (48) times  $N$  can be written as:

$$N Q_N(u/\sqrt{N}, \hat{p}) = N \tilde{Q}_N(u/\sqrt{N}) + N \varepsilon_N(u/\sqrt{N}) \quad (50)$$

The next lemma shows that for a fixed  $t$ ,  $N \varepsilon_N(u/\sqrt{N})$  goes to zero in probability for each  $u$ :

**LEMMA 5 (*Bounding the differences in the Objective Functions*):** *Under Assumptions 1 and 3, for a fixed  $t$  and each  $u$ :*

$$N \varepsilon_N(u/\sqrt{N}) \xrightarrow{P} 0 \quad (51)$$

**Proof:** See appendix

#### 5.4.5 ASYMPTOTIC PROPERTIES OF $\tilde{u}$

I assume from now on that the following regularity condition holds:

**ASSUMPTION 4 (*Interior*):** *The  $\tau$ -quantile of  $Y(1)$ 's marginal distribution,  $q_{1,\tau}$ , lies on the interior of  $Q$ , an open and convex subset of  $\mathbb{R}$ .*

Also, Let me now restate one of the necessary assumptions for the derivation of the semi-parametric efficiency bound of  $\Delta_\tau$ :

**ASSUMPTION 5 (*Local Identification of  $q_{1,\tau}$* ):** *The density of  $Y(1)$  exists and is bounded away from zero at  $q_{1,\tau}$ , that is,  $f_1(q_{1,\tau}) > 0$*

That assumption plays a different role here; it guarantees that  $\tilde{u}$ , the argument that minimizes  $N \tilde{Q}_N(u)$  is unique. Remember Equation (46):

$$\tilde{Q}_N(t) = \sum_{i=1}^N \frac{t}{\sqrt{N}} \left( \frac{T_i D_i}{p(X_i)} - E[D|X_i, T=1] \frac{T_i - p(X_i)}{p(X_i)} \right) + \frac{t^2}{2} f_1(q_{1,\tau})$$

Then under Assumption 5,  $N \tilde{Q}_N(u/\sqrt{N})$  has a unique minimum at:

$$\tilde{u} = \arg \min_u \sum_{i=1}^N \frac{u}{\sqrt{N}} \left( \frac{T_i D_i}{p(X_i)} - E[D|X_i, T=1] \frac{T_i - p(X_i)}{p(X_i)} \right) + \frac{u^2}{2} f_1(q_{1,\tau}) \quad (52)$$

$$= -\frac{1}{\sqrt{N} f_1(q_{1,\tau})} \sum_{i=1}^N \left( \frac{T_i D_i}{p(X_i)} - E[D|X_i, T=1] \frac{T_i - p(X_i)}{p(X_i)} \right) \quad (53)$$

$$= -\frac{1}{\sqrt{N} f_1(q_{1,\tau})} \sum_{i=1}^N \left( \frac{T_i(D_i - E[D|X_i, T=1])}{p(X_i)} + E[D|X_i, T=1] \right) \quad (54)$$

$$= -\frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{T_i(\mathbb{I}\{Y_i \leq q_{1,\tau}\} - \tau - E[\mathbb{I}\{Y(1) \leq q_{1,\tau}\} - \tau | X_i, T=1])}{p(X_i) f_1(q_{1,\tau})} \right)$$

$$- \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{E[\mathbb{I}\{Y(1) \leq q_{1,\tau}\} - \tau | X_i, T=1]}{f_1(q_{1,\tau})} \right) \quad (55)$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^N \left( \frac{T_i g_{1,\Delta\tau}(Y_i, X_i)}{p(X_i)} + h_{1,\Delta\tau}(X_i) \right) \quad (56)$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^N \Psi_{1,i} \quad (57)$$

where the functions  $g_{1,\Delta\tau}$  and  $h_{1,\Delta\tau}$  are those defined by Equations (15) and (16) and:

$$\Psi_{1,i} = \frac{T_i g_{1,\Delta\tau}(Y_i, X_i)}{p(X_i)} + h_{1,\Delta\tau}(X_i) \quad (58)$$

Let me now write the main result of this subsection as a lemma:

LEMMA 6 (*Asymptotic Properties of  $\tilde{u}$* ): Let  $\tilde{u} = \arg \min_u N \tilde{Q}_N(t)$ . Then, under Assumptions 1, 3, 4 and 5:

(i)  $\tilde{u} = O_p(1)$ ;

(ii)  $\tilde{u} \xrightarrow{D} N(0, E[\Psi_{1,i}^2])$ ;

(iii)  $E[\Psi_{1,i}^2] = V_1$ , the semiparametric efficiency bound for  $q_{1,\tau}$ .

**Proof:** See appendix

#### 5.4.6 NEARNESS OF ARGMINS

Defining  $\hat{u} = \sqrt{N}\hat{t}$ , I show the desired result that  $\hat{u} - \tilde{u} = o_p(1)$ , which will imply that  $\sqrt{N}(\hat{q}_{1,\tau}^C - q_{1,\tau})$  is (i)  $O_p(1)$ , (ii) and asymptotically normal (iii) and has an asymptotic variance that is equal to the semiparametric efficiency bound for  $q_{1,\tau}$ . Before I do that, let me state and prove an intermediate lemma.

We have already seen that Lemma 6 holds. To get results about  $\hat{u}$  and consequently about  $\hat{q}_{1,\tau}^C$  I will use a result in Hjort and Pollard (1993) on the nearness of minimizers of convex random functions.

I apply Hjort and Pollard's Lemma 2 directly to our case:

LEMMA 7 : (*Nearness of Argmins (Hjort and Pollard (1993))*) Under Assumptions 1, 3, 4 and 5, we have the following probabilistic bound on how far  $\hat{u}$  can be from  $\tilde{u}$ : For each  $\delta > 0$ :

$$Pr[|\hat{u} - \tilde{u}| \geq \delta] \leq Pr \left[ \sup_{|u - \tilde{u}| \leq \delta} |N \epsilon_N(u/\sqrt{N})| \geq \frac{1}{4} f_1(q_{1,\tau}) \delta^2 \right] \quad (59)$$

Moreover :

$$Pr \left[ \sup_{|u - \tilde{u}| \leq \delta} |N \epsilon_N(u/\sqrt{N})| \geq \frac{1}{4} f_1(q_{1,\tau}) \delta^2 \right] = o(1) \quad (60)$$

**Proof:** See appendix

Stating the final results:

**THEOREM 2 : (Asymptotic Properties of  $(\hat{q}_{1,\tau}^C)$ )** Let  $\hat{q}_{1,\tau}^C = \arg \min_{q \in Q} \frac{1}{N} \sum_{i=1}^N \frac{T_i}{\hat{p}(X_i)} (Y_i - q)(\tau - \mathbb{1}\{Y_i \leq q\})$  where  $\hat{p}(x)$  is computed as described in subsection 5.3. Under Assumptions 1, 3, 4 and 5:

$$(i) \sqrt{N}(\hat{q}_{1,\tau}^C - q_{1,\tau}) = O_p(1)$$

$$(ii) \sqrt{N}(\hat{q}_{1,\tau}^C - q_{1,\tau}) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \Psi_{1,i} + o_p(1)$$

$$\text{where } \frac{1}{\sqrt{N}} \sum_{i=1}^N \Psi_{1,i} \xrightarrow{D} N(0, V_1);$$

$$(iii) V_1 = E[\Psi_1^2] = E \left[ \frac{E[g_{1,\Delta\tau}^2(Y,X)|X,T=1]}{p(X)} + h_{1,\Delta\tau}^2 \right]$$

**Proof:** Defining  $\tilde{q}_{1,\tau} = \tilde{u}/\sqrt{N} + q_{1,\tau}$ , by Lemma 7 we have:

$$\begin{aligned} \sqrt{N}|\hat{q}_{1,\tau}^C - \tilde{q}_{1,\tau}| &= |\sqrt{N}(\hat{q}_{1,\tau}^C - q_{1,\tau}) - \sqrt{N}(\tilde{q}_{1,\tau} - q_{1,\tau})| \\ &\leq |\hat{u} - \tilde{u}| \\ &= o_p(1) \end{aligned} \tag{61}$$

That is,  $\hat{q}_{1,\tau}^C$  is asymptotically equivalent to  $\tilde{q}_{1,\tau}$  and Theorem 2 follows immediately by Lemma 6.  $\square$

The same result obtained for  $q_{1,\tau}$  could have been obtained analogously for  $q_{0,\tau}$ . In particular, with a set of assumptions analogous to those used in Theorem 2, it is possible to derive an asymptotic linear influence function for  $\hat{q}_{0,\tau}^C$ ,  $\Psi_0$ , which is analogous to  $\Psi_1$ . In fact,  $\Psi_{0,i} = \frac{1-T_i}{1-p(X_i)} g_{0,\Delta\tau}(Y_i, X_i) + h_{0,\Delta\tau}(Y_i, X_i)$ .

A consequence of Theorem 2 is that  $\hat{\Delta}_\tau^C$ , which is equal to the difference between  $\hat{q}_{1,\tau}^C$  and  $\hat{q}_{0,\tau}^C$ : (i) will also be consistent, (ii) will have an asymptotically linear influence function and, (iii) will be asymptotically normal. The extra conditions needed for this result are that Assumptions 4 and 5 have properly modified counterparts for the  $q_{0,\tau}$  case. I name these Assumptions 4' and 5' in stating the following theorem:

**THEOREM 3 : (Asymptotic Properties of  $\hat{\Delta}_\tau^C$ ):** Under Assumptions 1, 3, 4, 4', 5 and 5':

$$(i) \hat{\Delta}_\tau^C - \Delta_\tau \xrightarrow{P} 0$$

$$(ii) \sqrt{N}(\hat{\Delta}_\tau^C - \Delta_\tau) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \Psi_i + o_p(1)$$

$$(iii) \sqrt{N}(\hat{\Delta}_\tau^C - \Delta_\tau) \xrightarrow{D} N(0, V_\tau)$$

where  $\Psi_i = \Psi_{1,i} - \Psi_{0,i}$

$$\text{and } V_{\Delta_\tau} = E[\Psi^2] = E \left[ \frac{E[g_{1,\Delta_\tau}^2(Y,X)|X,T=1]}{p(X)} + \frac{E[g_{0,\Delta_\tau}^2(Y,X)|X,T=0]}{1-p(X)} + (h_{1,\Delta_\tau}(X) - h_{0,\Delta_\tau}(X))^2 \right]$$

**Proof:** Omitted.

Theorem 3 shows that besides  $\hat{\Delta}_\tau^C$  being root- $N$  consistent and asymptotically linear, it is efficient, as it achieves the semiparametric lower bound for  $\Delta_\tau$ .

Estimation of the quantile treatment effect on the treated,  $\Delta_\tau|_{T=1}$ , will yield a similar result, which could have been obtained using analogous steps to get the result of Theorem 3.

## 6 EMPIRICAL APPLICATION

In this section I consider one empirical application for the QTE estimators proposed in the previous sections. This application uses the job training program data set first analyzed by Lalonde (1986) and later by many others, including Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2001) and Abadie and Imbens (2002). The original data set from the ‘‘National Supported Work Program’’ (NSW) is an experimental data set where treatment was randomly assigned to the eligible participants. This data set is well described in Lalonde (1986) and basically consists of information on earnings and employment (outcome variables); whether treated or not; and background characteristics, such as education, ethnicity, age, and employment variables before treatment.

Lalonde uses the experimental data set as a benchmark for comparisons with his other analyses in which control samples come from non-experimental data sets, as for example, control samples based on Panel Study of Income Dynamics (PSID) and on Westat’s Matched Current Population Survey-Social Security Administration File (CPS-SSA). He finds that non-experimental control samples are poor substitutes for experimental data, as they are composed of individuals with observable characteristics quite distant from those of the individuals in the

treatment group.

Dehejia and Wahba (1999) use the same data set as Lalonde (1986) but reach a different conclusion. This discrepancy results mainly from the parametric nature of Lalonde’s analysis using non-experimental control groups. While Lalonde estimates parametric wage regressions for treatment and control groups which are intrinsically different from each other, Dehejia and Wahba use a more flexible methodology. Their methods involve considering differentially the control units based on some closeness measure of their observable characteristics to characteristics of the treatment group.

As the non-experimental control groups were essentially different from the treated group, Dehejia and Wahba were interested only in computing the average treatment effect for the treated. In their setting, the outcome variable was 1978 earnings. For their estimation of the ATE on the treated, they estimate the propensity score in a first step using logistic regressions and propose several ways of using it to control for the selection problem. One of these methods, reweighing using the estimated p-score, uses exactly the weights described by Equation (26),  $\hat{\omega}_{1,i|T=1}^C$  and  $\hat{\omega}_{0,i|T=1}^C$ .

One data set Dehejia and Wahba use is a subset of 185 treated units and 2490 control observations from the PSID.<sup>18</sup> Dehejia and Wahba estimate the p-score using logistic regression. The specification of the logit model is an issue in their paper, and it varies for each control sample, because they are trying to find a specification that best “balances” each covariate between treated and control groups. Next, they compute the average treatment effect on the treated, which is equal to  $\sum_{i=1}^n (\hat{\omega}_{1,i|T=1}^C - \hat{\omega}_{0,i|T=1}^C) Y_i$ . For these specific treatment and control groups they find an average treatment effect on the treated of US\$ 1129.<sup>19</sup> This is lower than the unadjusted experimental treatment effect of \$1749, but above the initial numbers Lalonde computed using the non-experimental data.<sup>20</sup>

---

<sup>18</sup>This corresponds to the control sample labelled by Lalonde (1986) and Dehejia and Wahba (1999) as PSID-1, as they constructed more than one control group based on PSID.

<sup>19</sup>I replicated their calculations using the same p-score specification and got a slightly different number, \$1120.

<sup>20</sup>The unadjusted for covariates treatment effect was computed using the experimental control sample of size

Using the same data, I analyze the treatment and control subsets to generate estimates of the quantile treatment effect on the treated for each percentile. I also perform an “experimental” QTE estimation, which is just the difference between the quantiles of the treated and the experimental controls, without any weighting. My results are presented in Table 1 and in Figures 1 to 4.<sup>21</sup> I find that using experimental controls, treatment effects tend to be much more homogenous than in the observational setting. With a non-experimental control sample, treatment effects seem to be above the median until almost the upper end of the distribution. At the extreme upper quantiles, the very high earnings of the control sample induce a negative effect. Another interesting result is that the value I find for the median treatment effect using the non-experimental data is \$1927, which is relatively close to the estimated experimental mean effect of \$1749.

## 7 CONCLUSION

In this paper I motivated interest in the quantile treatment effects by constructing a simple model where (i) the individual decision to be in the treatment group depends on a vector of observable covariates, and (ii) the policy-maker aims to learn features of the marginal distributions of potential outcomes.

This paper has also shown how to estimate the quantile treatment effects in three different ways, using a two-step procedure. The estimator that (in the first step) involves only estimation of the propensity score is shown to be root- $N$  consistent and asymptotically normal. I also calculated the semiparametric efficiency bound and proved that this quantile treatment effects estimator achieves it.

The empirical application was designed to show how to apply the estimator and how it differs from the usual average treatment effects estimator. In this particular example, estimation of the quantiles of the potential outcomes revealed the presence of heterogenous impacts of the

---

260. It is a simple difference in means between treated and control groups.

<sup>21</sup>In Table 1 and in Figure 3, the standard errors were computed by 100 bootstrap replications.

treatment. This heterogeneity could never be captured by the estimator of average treatment effects.

A natural extension to this paper would be the computation and estimation of inequality measures for the potential outcomes of being treated and not being treated. Several relevant inequality measures are of interest in the applied literature. The framework developed here could be extended to estimate and predict the response of such inequality measures to a treatment.

## REFERENCES

- ABADIE, A., (2002), "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of American Statistical Association*, 97, 284-292.
- ABADIE, A., J. ANGRIST, AND G. IMBENS, (2002), "Instrumental Variables Estimation of Quantile Treatment Effects," *Econometrica*, 70, 91-117.
- ABADIE, A. AND G. IMBENS, (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," preprint.
- ANGRIST, J., AND A. KRUEGER, (1999), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, 1277-1366, New York, Elsevier Science B.V.
- AMEMIYA, T., (1982), "Two-Stage Least Absolute Deviations Estimators," *Econometrica*, 50, 689-711.
- BEGUN, J., W. HALL, W. HUANG, AND J. WELLNER, (1983), "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *Annals of Statistics*, 11, 432-452.
- BICKEL, P., C. KLASSEN, Y. RITOV, AND J. WELLNER, (1993), *Efficient and Adaptive Estimation for Semiparametric Models*. New York, Springer-Verlag.
- CARD, D., (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 64, 957-979.
- CHERNOZHUKOV, V., AND C. HANSEN, (2001), "An IV Model of Quantile Treatment Effects," *MIT Department of Economics Working Paper*, No. 02-06.
- DEHEJIA, R., AND S. WAHBA, (1999), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053-1062.

- DINARDO, J., N. FORTIN, AND T. LEMIEUX, (1996), "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64, 1001-1044.
- DOKSUM, K., (1974), "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *Annals of Statistics*, 2, 267-277.
- FREEMAN, D., (1980), "Unionism and the Dispersion of Wages," *Industrial and Labor Relations Review*, 34, 3-23.
- HAHN, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315-331.
- HECKMAN, J., AND B. HONORE (1990) "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121-1149.
- HECKMAN, J., AND J. HOTZ, (1989), "Alternative Methods for Evaluating the Impact of Training Programs," (with discussion), *Journal of the American Statistical Association*.
- HECKMAN, J., H. ICHIMURA, AND P. TODD, (1997), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65(2), 261-294.
- HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD, (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017-1098.
- HECKMAN, J., R. LALONDE, AND J. SMITH, (2000), "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, 1277-1366., New York, Elsevier Science B.V.
- HECKMAN, J., R. ROBB, (1986), "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcome," in *Drawing Inference from Self-Selected Samples*, ed. by H. Wainer, pp. 63-107. New York, Springer-Verlag.

- HECKMAN, J., J. SMITH, AND N. CLEMENTS, (1997), "Making the Most out of Programme Evaluations and Social Experiments Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64(4), 487-535.
- HIRANO, K., G. IMBENS, AND G. RIDDER, (2002), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," forthcoming, *Econometrica*.
- HJORT, N., AND D. POLLARD, (1993), "Asymptotics for Minimizers of Convex Processes," preprint, [www.stat.yale.edu/Preprints/1993/93may-1.pdf](http://www.stat.yale.edu/Preprints/1993/93may-1.pdf).
- IMBENS, G., AND D. RUBIN, (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economic Studies*, October, 555-574.
- KHAN, S., AND J. POWELL, (2001) "Two Step Estimation of Semiparametric Censored Regression Models," *Journal of Econometrics*, 103, 73-110.
- KOENKER, R., AND G. BASSETT, (1978), "Regression Quantiles," *Econometrica*, 46, 33-50.
- LALONDE, R., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604-620.
- LEHMANN, E. (1974) *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco, Holden-Day.
- MANSKI, C. (1988), *Analog Estimation Methods in Econometrics*, New York, Chapman and Hall.
- MANSKI, C. (1997), "The Mixing Problem in Programme Evaluation," *Review of Economic Studies*, October, 537-554.
- NEWBY, W., (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99-135.

- NEWKEY, W., (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349-1382.
- NEWKEY, W., (1995), "Convergence Rates for Series Estimators," in *Advances in Econometrics and Qualitative Economics: Essays in Honor of C.R. Rao*, G. Maddal, P.C. Phillips, and T.N. Srinivasan, eds., Cambridge US, Basil-Blackwell.
- NEWKEY, W., (1997), "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147-168.
- POWELL, J., (1983), "The Asymptotic Normality of Two-Stage Least Absolute Deviations Estimators," *Econometrica*, 51, 1569-1576.
- RUBIN, D., (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2(1), 1-26.
- ROSENBAUM, P., AND D. RUBIN, (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.
- ROY, A., (1951) "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3(2), 135-146.
- SMITH, J. A. AND P. E. TODD, (2001), "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *American Economic Review*, Papers and Proceedings, 91, 112-118.
- STEIN, C., (1956), "Efficient Nonparametric Testing and Estimation," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1. Berkeley, University of California Press.

## APPENDIX

### Proof of Lemma 1:

Starting from the definition of the  $\tau$ -quantile of  $Y(1)$  I show how to express  $q_{1,\tau}$  in terms of the observed data  $(Y, T, X)$ :

$$\begin{aligned}
 \tau &= Pr[Y(1) \leq q_{1,\tau}] \\
 &= E[Pr[Y(1) \leq q_{1,\tau} | X]] \\
 &= E[Pr[Y(1) \leq q_{1,\tau} | X, T = 1]] \\
 (Q1_A) \quad &= E[Pr[Y \leq q_{1,\tau} | X, T = 1]] \\
 &= E[E[T \mathbb{1}\{Y \leq q_{1,\tau}\} | X, T = 1]] \\
 (Q1_B) \quad &= E \left[ \frac{E[T \mathbb{1}\{Y \leq q_{1,\tau}\} | X]}{p(X)} \right] \\
 (Q1_C) \quad &= E \left[ \frac{T \mathbb{1}\{Y \leq q_{1,\tau}\}}{p(X)} \right]
 \end{aligned}$$

The first equality follows from the definition of  $q_{1,\tau}$ . The second is an application of the law of iterated expectations. The third equality follows from the ignorability assumption (Assumption 1). The fourth results from the definition of  $Y$ ,  $Y = TY(1) + (1 - T)Y(0)$ . The fifth equality comes from  $E[\mathbb{1}\{A\}] = Pr[A]$  (where  $A$  is some event) and from the fact that the expectation is conditional on  $T = 1$ . The sixth is a consequence from  $E[Z | X] = p(X)E[Z | X, T = 1] + (1 - p(X))E[Z | X, T = 0]$ , where  $Z$  is some random variable. Finally, the last equality is a backward application of the law of iterated expectations.

An analogous result for  $q_{0,\tau}$  could have been derived following essentially the same steps as above.  $\square$

### Proof of Lemma 2:

$q_{1,\tau|T=1}$ :

$$\begin{aligned}
\tau &= \frac{\Pr[Y(1) \leq q_{1,\tau|T=1} | T = 1]}{\Pr[Y(1) \leq q_{1,\tau|T=1}, T = 1]} \\
&= \frac{p}{E \left[ \frac{\Pr[Y(1) \leq q_{1,\tau|T=1}, T = 1 | X]}{p} \right]} \\
&= E \left[ \frac{\Pr[Y \leq q_{1,\tau|T=1}, T = 1 | X]}{p} \right] \\
(QT1_A) \quad &= E \left[ \frac{p(X) \Pr[Y \leq q_{1,\tau|T=1} | X, T = 1]}{p} \right] \\
&= E \left[ \frac{p(X) E[T \mathbb{1}\{Y \leq q_{1,\tau|T=1}\} | X, T = 1]}{p} \right] \\
(QT1_B) \quad &= E \left[ \frac{E[T \mathbb{1}\{Y \leq q_{1,\tau|T=1}\} | X]}{p} \right] \\
(QT1_C) \quad &= E \left[ \frac{T \mathbb{1}\{Y \leq q_{1,\tau|T=1}\}}{p} \right]
\end{aligned}$$

The first equality follows from the definition of  $q_{1,\tau|T=1}$ . The second is an application of the Bayes' rule. The third equality follows from an application of the law of iterated expectations. The fourth results from  $Y = TY(1) + (1 - T)Y(0)$ . The fifth equality comes from another application of the Bayes' rule. Sixth equality is a consequence from the fact that the expectation is conditional on  $T = 1$ . Seventh uses the relation  $E[Z | X] = p(X)E[Z | X, T = 1] + (1 - p(X))E[Z | X, T = 0]$ , where  $Z$  is some random variable. Finally, the last equality is a backward application of the law of iterated expectations.

$q_{0,\tau|T=1}$  :

$$\begin{aligned}
\tau &= \frac{Pr[Y(0) \leq q_{0,\tau|T=1} | T = 1]}{Pr[Y(0) \leq q_{0,\tau|T=1}, T = 1]} \\
&= \frac{p}{E\left[\frac{Pr[Y(0) \leq q_{0,\tau|T=1}, T = 1 | X]}{p}\right]} \\
&= E\left[\frac{p(X)Pr[Y(0) \leq q_{0,\tau|T=1} | X, T = 1]}{p}\right] \\
&= E\left[\frac{p(X)Pr[Y(0) \leq q_{0,\tau|T=1} | X, T = 0]}{p}\right] \\
(QT0_A) \quad &= E\left[\frac{p(X)Pr[Y \leq q_{0,\tau|T=1} | X, T = 0]}{p}\right] \\
&= E\left[\frac{p(X)E[(1-T)\mathbb{I}\{Y \leq q_{0,\tau|T=1}\} | X, T = 0]}{p}\right] \\
(QT0_B) \quad &= E\left[\frac{p(X)}{(1-p(X))p}E[(1-T)\mathbb{I}\{Y \leq q_{0,\tau|T=1}\} | X]\right] \\
(QT0_C) \quad &= E\left[\frac{p(X)}{(1-p(X))p}(1-T)\mathbb{I}\{Y \leq q_{0,\tau|T=1}\}\right]
\end{aligned}$$

Equalities 1 to 3 hold by the same reasons equalities 1-3 hold for the  $q_{1,\tau|T=1}$  case. The fourth equality comes from an application of the Bayes' rule. The fifth equality follows from Assumption 1. Sixth results from  $Y = TY(1) + (1-T)Y(0)$ . Seventh equality is a consequence from the fact that the expectation is conditional on  $T = 0$ . Eighth and ninth equalities hold by the same reasons the last two equalities for the  $q_{1,\tau|T=1}$  case hold.  $\square$

**Proof of Theorem 1:**

This proof is an extension to the quantile case of the proofs by Hahn (1998) and by Hirano, Imbens and Ridder (2002) for the quantile case. Both references use the machinery presented by Bickel, Klassen, Ritov, and Wellner (1993), Newey (1990) and Newey (1994). Start defining the densities, with respect to some  $\sigma$ -finite measure, of  $(Y(1), Y(0), T, X)$  and of the observed data  $(Y, T, X)$ . Under Assumption 1, both densities represent the same statistical model and are, therefore, equivalent. These densities can be written as:

$$\phi(y(1), y(0), t, x) = f(y(1), y(0) | x) p(x)^t (1 - p(x))^{1-t} f(x).$$

and

$$\phi(y, t, x) = [f_1(y | x) p(x)]^t [f_0(y | x) (1 - p(x))]^{1-t} f(x),$$

where  $f_1(\cdot | x) = \int f(\cdot, y(0) | x) dy(0)$  and  $f_0(\cdot | x) = \int f(y(1), \cdot | x) dy(1)$ .

Working with the density of observed data, consider the regular parametric submodel indexed by  $\theta$ , a finite dimensional vector:

$$\phi(y, t, x | \theta) = [f_1(y | x; \theta) p(x | \theta)]^t [f_0(y | x; \theta) (1 - p(x | \theta))]^{1-t} f(x | \theta),$$

By a normalization argument, let  $\phi(y, t, x) = \phi(y, t, x | \theta)$  when  $\theta = \theta_0$ .

The score of a parametric submodel indexed by  $\theta$  is given by:

$$s(y, t, x | \theta) = t s_1(y | x; \theta) + (1 - t) s_0(y | x; \theta) + \frac{t - p(x | \theta)}{p(x | \theta) (1 - p(x | \theta))} \dot{p}(x | \theta) + s_x(x | \theta)$$

where, for  $j = 0, 1$ :

$$s_j(y | x; \theta) = \frac{\partial}{\partial \theta} \log f_j(y | x; \theta)$$

$$\dot{p}(x | \theta) = \frac{\partial}{\partial \theta} p(x | \theta)$$

and

$$s_x(x|\theta) = \frac{\partial}{\partial \theta} \log f(x|\theta).$$

Again I normalize:  $s(y, t, x) = s(y, t, x|\theta)$  when  $\theta = \theta_0$ .

The tangent space of this model is the set of functions  $\mathcal{S}$  defined as:

$$\mathcal{S} = \left\{ S : \mathbb{R} \times \{0, 1\} \times \mathcal{X} \rightarrow \mathbb{R} \mid \begin{array}{l} S(y, t, x) = t s_1(y|x) + (1-t) s_0(y|x) + a(x)(t - p(x)) + s_x(x); \quad \text{and} \\ E[s_j(Y, X) | X = x, T = j] = E[s_x(X)] = 0, \forall x \quad \text{and} \quad j = 0, 1 \end{array} \right\}$$

where  $a(x)$  is some square-integrable measurable function of  $x$ .

For the two parameters of interest in this paper,  $\Delta_\tau(\theta)$  and  $\Delta_{\tau|T=1}(\theta)$ , I need to first show that they are pathwise differentiable. In other words, I need to show that for each one their derivatives with respect to  $\theta$  evaluated at  $\theta_0$  are equal to the expectation of the product of the score  $s(Y, T, X)$  and their influence functions  $\psi_{\Delta_\tau}(Y, T, X)$  and  $\psi_{\Delta_{\tau|T=1}}(\theta)$  respectively.

After I show pathwise differentiability, I find the projection of the influence function on the set of scores. That projection is often called the efficient influence function. Notice that as expected with projections in general, the efficient influence function will be unique. Also, if an influence function belongs to the set  $\mathcal{S}$ , then its projection onto  $\mathcal{S}$  is the original influence function itself. Therefore, if I can find an influence function which already belongs to the set of scores, that will be the efficient one, and will be written as:

$$\psi = T c_1(Y, X) + (1 - T) c_0(Y, X) + a(X)(T - p(X)) + c_x(X),$$

where  $E[c_j(Y, X) | X = x, T = j] = E[c_x(X)] = 0, \forall x \quad \text{and} \quad j = 0, 1$ .

Start with  $\Delta_\tau$ . In particular, start with  $q_{1,\tau}$ . For the parametric submodel indexed by  $\theta$ , we have:

$$0 = \iint (\mathbb{I}\{y \leq q_{1,\tau}(\theta)\} - \tau) f_1(y|x; \theta) f(x|\theta) dy dx \quad (62)$$

Thus, normalizing  $q_{1,\tau} = q_{1,\tau}(\theta_0)$ , by an application of Leibniz's rule we have:

$$\begin{aligned} 0 = & f_1(q_{1,\tau}) \frac{\partial q_{1,\tau}(\theta_0)}{\partial \theta} + \iint (\mathbb{I}\{y \leq q_{1,\tau}\} - \tau) s_1(y|x) f_1(y|x) f(x) dy dx \\ & + \iint (\mathbb{I}\{y \leq q_{1,\tau}\} - \tau) s_x(x) f_1(y|x) f(x) dy dx \end{aligned} \quad (63)$$

But because:

$$\iint s_1(y|x) f_1(y|x) f(x) dy dx = 0 \quad (64)$$

$$\int s_x(x) f(x) dx = 0 \quad (65)$$

The derivative of  $q_{1,\tau}(\theta)$  evaluated at  $\theta_0$  is equal to:

$$\frac{\partial q_{1,\tau}(\theta_0)}{\partial \theta} = - \frac{\iint \mathbb{I}\{y \leq q_{1,\tau}\} s_1(y|x) f_1(y|x) f(x) dy dx}{f_1(q_{1,\tau})} - \frac{\iint \mathbb{I}\{y \leq q_{1,\tau}\} s_x(x) f_1(y|x) f(x) dy dx}{f_1(q_{1,\tau})} \quad (66)$$

As the same sort of calculations are true for  $q_{0,\tau}$ , we can express the derivative of  $\Delta_\tau(\theta)$  evaluated at  $\theta_0$  as being:

$$\begin{aligned} \frac{\partial \Delta_\tau(\theta_0)}{\partial \theta} = & - \frac{\iint \mathbb{I}\{y \leq q_{1,\tau}\} s_1(y|x) f_1(y|x) f(x) dy dx}{f_1(q_{1,\tau})} + \frac{\iint \mathbb{I}\{y \leq q_{0,\tau}\} s_0(y|x) f_0(y|x) f(x) dy dx}{f_0(q_{0,\tau})} \\ & - \frac{\iint \mathbb{I}\{y \leq q_{1,\tau}\} s_x(x) f_1(y|x) f(x) dy dx}{f_1(q_{1,\tau})} + \frac{\iint \mathbb{I}\{y \leq q_{0,\tau}\} s_x(x) f_0(y|x) f(x) dy dx}{f_0(q_{0,\tau})} \end{aligned} \quad (67)$$

For this particular case, a function that times the score has expected value is the following:

$$\Psi_{\Delta_\tau}(Y, T, X) = \frac{T g_{1,\Delta_\tau}(Y, X)}{p(X)} - \frac{(1-T) g_{0,\Delta_\tau}(Y, X)}{1-p(X)} + h_{1,\Delta_\tau}(Y, X) - h_{0,\Delta_\tau}(Y, X) \quad (68)$$

where the functions  $g_{j,\Delta_\tau}$  and  $h_{j,\Delta_\tau}$  are those defined by Equations (15) and (16).

Note however that this influence function belongs to the set of the scores. In order to check that, we need that:

$$E \left[ \frac{g_{1,\Delta_\tau}(Y, X)}{p(X)} \mid T = 1 \right] = 0 \quad (69)$$

$$E \left[ \frac{g_{0,\Delta_\tau}(Y, X)}{1 - p(X)} \mid T = 0 \right] = 0 \quad (70)$$

$$E[h_{1,\Delta_\tau}(Y, X) - h_{0,\Delta_\tau}(Y, X) \mid X] = 0 \quad (71)$$

By the definitions of  $g_{j,\Delta_\tau}$  and  $h_{j,\Delta_\tau}$ , one can check that indeed Equations (69)-(71) hold. Hence,  $\psi_{\Delta_\tau}$  is the efficient influence function has expected value equal to zero, since it is on the set of scores. Thus its variance is equal to  $E[\psi_{\Delta_\tau}^2(Y, T, X)]$ , which is the semiparametric efficiency bound for  $\Delta_\tau$ ,  $V_{\Delta_\tau}$ .

Now we do the same for  $\Delta_{\tau|T=1}$ . For a parametric submodel indexed by  $\theta$ , we have:

$$0 = \iint \frac{p(x|\theta)}{\int p(x|\theta)f(x|\theta)dx} (\mathbb{I}\{y \leq q_{1,\tau|T=1}(\theta)\} - \tau) f_1(y|x; \theta) f(x|\theta) dy dx \quad (72)$$

Again I normalize:  $q_{1,\tau|T=1} = q_{1,\tau|T=1}(\theta_0)$ . The derivative evaluated at  $\theta_0$  is equal to:

$$\begin{aligned} \frac{\partial q_{1,\tau|T=1}(\theta_0)}{\partial \theta} &= -\frac{1}{f_{1|T=1}(q_{1,\tau|T=1})} \left( \iint \mathbb{I}\{y \leq q_{1,\tau|T=1}\} p(x) s_1(y|x) f_1(y|x) f(x) dy dx \right. \\ &\quad + \int (E[\mathbb{I}\{y \leq q_{1,\tau|T=1}\} \mid X = x] - \tau) \dot{p}(x) f_1(y|x) f(x) dy dx \\ &\quad \left. + \int (E[\mathbb{I}\{y \leq q_{1,\tau|T=1}\} \mid X = x] - \tau) p(x) s_x(x) f_1(y|x) f(x) dy dx \right) \end{aligned} \quad (73)$$

As the same sort of calculations are true for  $q_{0,\tau|T=1}$ , we can express the derivative of  $\Delta_{\tau|T=1}(\theta)$  evaluated at  $\theta_0$  as being:

$$\begin{aligned}
\frac{\partial \Delta_{\tau|T=1}(\theta_0)}{\partial \theta} &= - \frac{\iint \mathbb{I}\{y \leq q_{1,\tau|T=1}\} p(x) s_1(y|x) f_1(y|x) f(x) dy dx}{f_{1|T=1}(q_{1,\tau|T=1})} \\
&+ \frac{\iint \mathbb{I}\{y \leq q_{0,\tau|T=1}\} p(x) s_0(y|x) f_0(y|x) f(x) dy dx}{f_{0|T=1}(q_{0,\tau|T=1})} \\
&- \frac{f(E[\mathbb{I}\{y \leq q_{1,\tau|T=1}\} | X = x] - \tau) \dot{p}(x) f_1(y|x) f(x) dy dx}{f_{1|T=1}(q_{1,\tau|T=1})} \\
&+ \frac{f(E[\mathbb{I}\{y \leq q_{0,\tau|T=1}\} | X = x] - \tau) \dot{p}(x) f_0(y|x) f(x) dy dx}{f_{0|T=1}(q_{0,\tau|T=1})} \\
&- \frac{f(E[\mathbb{I}\{y \leq q_{1,\tau|T=1}\} | X = x] - \tau) p(x) s_x(x) f_1(y|x) f(x) dy dx}{f_{1|T=1}(q_{1,\tau|T=1})} \\
&+ \frac{f(E[\mathbb{I}\{y \leq q_{0,\tau|T=1}\} | X = x] - \tau) p(x) s_x(x) f_0(y|x) f(x) dy dx}{f_{0|T=1}(q_{0,\tau|T=1})}
\end{aligned} \tag{74}$$

The efficient influence function for this case is equal:

$$\begin{aligned}
\Psi_{\Delta_{\tau|T=1}}(Y, T, X) &= \frac{T g_{1,\Delta_{\tau|T=1}}(Y, X)}{p} - \frac{(1-T) p(X) g_{0,\Delta_{\tau|T=1}}(Y, X)}{p(1-p(X))} \\
&+ (h_{1,\Delta_{\tau|T=1}}(X) - h_{0,\Delta_{\tau|T=1}}(X)) \frac{(T-p(X))}{p} \\
&+ (h_{1,\Delta_{\tau|T=1}}(X) - h_{0,\Delta_{\tau|T=1}}(X)) \frac{p(X)}{p}
\end{aligned} \tag{75}$$

where the functions  $g_{j,\Delta_{\tau|T=1}}$  and  $h_{j,\Delta_{\tau|T=1}}$  are those defined by Equations (17) and (18).

As this influence function is in the set of scores, its expected value is zero and its variance is equal to  $E[\Psi_{\Delta_{\tau|T=1}}^2(Y, T, X)]$ , which is the semiparametric efficiency bound for  $\Delta_{\tau|T=1}$ ,  $V_{\Delta_{\tau|T=1}}$ .

□

**Proof of Lemma 4:** In order to prove Lemma 4, I will need to sum and subtract several terms from the original definition of  $Q_N(t, \hat{p})$  in Equation 41 in such a way these terms can be absorbed into  $\varepsilon_N(t)$ . That is:

$$Q_N(t, \hat{p}) = \frac{1}{N} \sum_{i=1}^N \frac{T_i A_i(t)}{\hat{p}(X_i)} - \frac{T_i A_i(t)}{p(X_i)} + \frac{T_i A_i(t)}{p^2(X_i)} (\hat{p}(X_i) - p(X_i)) \quad (76)$$

$$- \frac{1}{N} \sum_{i=1}^N \frac{T_i A_i(t)}{p^2(X_i)} (\hat{p}(X_i) - p(X_i)) + E \left[ \frac{E[A(t) | X, T = 1]}{p(X)} (\hat{p}(X) - p(X)) \right] \quad (77)$$

$$E \left[ \frac{E[A(t) | X, T = 1]}{p(X)} (\hat{p}(X) - p(X)) \right] - \frac{1}{N} \sum_{i=1}^N \tilde{\delta}(X_i, t) \frac{T_i - p_K(X_i)}{\sqrt{p_K(X_i)(1 - p_K(X_i))}} \quad (78)$$

$$+ \frac{1}{N} \sum_{i=1}^N (\tilde{\delta}(X_i, t) - \delta_K(X_i, t)) \frac{T_i - p_K(X_i)}{\sqrt{p_K(X_i)(1 - p_K(X_i))}} \quad (79)$$

$$+ \frac{1}{N} \sum_{i=1}^N \delta_K(X_i, t) \frac{T_i - p_K(X_i)}{\sqrt{p_K(X_i)(1 - p_K(X_i))}} - \frac{1}{N} \sum_{i=1}^N \delta(X_i, t) \frac{T_i - p(X_i)}{\sqrt{p(X_i)(1 - p(X_i))}} \quad (80)$$

$$+ \frac{1}{N} \sum_{i=1}^N \frac{T_i A_i(t)}{p(X_i)} - E[A(t) | X_i, T = 1] \frac{T_i - p(X_i)}{p(X_i)} \quad (81)$$

where:

$$\tilde{\delta}(X_i, t) = -E \left[ \frac{E[A(t) | X, T = 1]}{p(X)} L'(H(X)' \tilde{\pi}) H(X)' \right] \tilde{\Sigma}^{-1} \sqrt{L'(H(X_i)' \pi_K) H(X_i)} \quad (82)$$

$$\delta_K(X_i, t) = -E \left[ \frac{E[A(t) | X, T = 1]}{p(X)} L'(H(X)' \pi_K) H(X)' \right] \Sigma_K^{-1} \sqrt{L'(H(X_i)' \pi_K) H(X_i)} \quad (83)$$

$$\delta(X_i, t) = -E[A(t) | X_i, T = 1] \frac{\sqrt{p(X_i)(1 - p(X_i))}}{p(X_i)} \quad (84)$$

$$\tilde{\Sigma} = \frac{1}{N} \sum_{i=1}^N H(X_i) H(X_i)' L'(H(X_i)' \tilde{\pi}) \quad (85)$$

$$\Sigma = E[H(X)H(X)'L'(H(X)'\pi_K)] \quad (86)$$

Thus, by Equations (40) and (46):

$$Q_N(t, \hat{p}) = \frac{t}{N} \sum_{i=1}^N \left( \frac{T_i D_i}{p(X_i)} - E[D|X_i, T=1] \frac{T_i - p(X_i)}{p(X_i)} \right) \\ + \tilde{R}_N(t) - E[\tilde{R}_N(t) | T=1] + E[\tilde{R}_N(t) | T=1] + \varepsilon_{1,N}(t) \quad (87)$$

$$(88)$$

where  $\varepsilon_{1,N}(t)$  is equal to the sum of Equations (76) to (80).

In order to decompose  $Q_N(t, \hat{p})$  into the sum of  $\tilde{Q}_N(t)$  and  $\varepsilon_N(t)$ , from Equation 87 I need to show that  $E[\tilde{R}_N(t) | T=1] = E[R(t) | T=1] = \frac{t^2}{2} f_1(q_{1,\tau}) + o(t^2)$ . I will do more than that. In fact, let me compute the first two conditional moments of  $A(t)$  given  $T=1$  and given  $X$  and  $T=1$ .

Starting with the first moments of  $A(t)$ ,  $E[A(t) | X, T=1] = E[D | X, T=1]t + E[R(t) | X, T=1]$  and  $E[A(t) | T=1] = E[D | T=1]t + E[R(t) | T=1]$ , where:

$$E[D | X, T=1] = E[\mathbb{1}\{Y(1) \leq q_{1,\tau}\} - \tau | X] \quad (89)$$

$$E[D | T=1] = E[E[\mathbb{1}\{Y(1) \leq q_{1,\tau}\} - \tau | X]] \\ = E[\mathbb{1}\{Y(1) \leq q_{1,\tau}\} - \tau] \\ = 0 \quad (90)$$

$$E[R(t) | X = x, T=1] = E[(Y(1) - (q_{1,\tau} + t))(\mathbb{1}\{Y(1) \leq q_{1,\tau}\} - \mathbb{1}\{Y(1) \leq q_{1,\tau} + t\}) | X = x] \\ = \int_{q_{1,\tau} + t}^{q_{1,\tau}} (y - (q_{1,\tau} + t)) f_1(y|x) dy \quad (91)$$

But integrating by parts and using a second order approximation argument:<sup>22</sup>

$$\begin{aligned}
E[R(t) | X = x, T = 1] &= q_{1,\tau} F_1(q_{1,\tau} | x) - (q_{1,\tau} + t) F_1(q_{1,\tau} + t | x) + \int_{q_{1,\tau}}^{q_{1,\tau} + t} F_1(y | x) dy \\
&\quad - (q_{1,\tau} + t) (F_1(q_{1,\tau} | x) - F_1(q_{1,\tau} + t | x)) \\
&= \frac{1}{2} f_1(q_{1,\tau} | x) t^2 + o(t^2)
\end{aligned} \tag{92}$$

$$\begin{aligned}
E[R(t) | T = 1] &= E[E[(Y(1) - (q_{1,\tau} + t))(\mathbb{1}\{Y(1) \leq q_{1,\tau}\} - \mathbb{1}\{Y(1) \leq q_{1,\tau} + t\}) | X]] \\
&= E\left[\frac{1}{2} f_1(q_{1,\tau} | x)\right] t^2 + o(t^2) \\
&= \frac{1}{2} f_1(q_{1,\tau}) t^2 + o(t^2)
\end{aligned} \tag{93}$$

Therefore, we have:

$$E[A(t) | X = x, T = 1] = t E[\mathbb{1}\{Y(1) \leq q_{1,\tau}\} - \tau | X = x] + \frac{1}{2} f_1(q_{1,\tau} | x) t^2 + o(t^2) \tag{94}$$

and

$$E[A(t) | T = 1] = \frac{1}{2} f_1(q_{1,\tau}) t^2 + o(t^2) \tag{95}$$

Now we compute:

$$E[A^2(t) | T = 1] = t^2 E[D^2 | T = 1] + 2t \text{Cov}[D, R(t) | T = 1] + E[R^2(t) | T = 1] \tag{96}$$

where:

---

<sup>22</sup>Let me be clear about the notation. There are two ways that the remainder terms of the above Taylor approximation go to zero. The first and natural one is to say that  $o(t^2) \rightarrow 0$  as  $t \rightarrow 0$ . But the remainder term might go to zero even for fixed  $t$ . This is the case when there is sequence  $a_N = o(1)$  and the remainder term is in fact equal to  $t^2 a_N$ .

$$\begin{aligned}
E[D^2 | T = 1] &= E[(\mathbb{1}\{Y\} \leq q_{1,\tau} - \tau)^2 | T = 1] \\
&= E[E[(\mathbb{1}\{Y(1)\} \leq q_{1,\tau} - \tau)^2 | X]] \\
&= E[(\mathbb{1}\{Y(1)\} \leq q_{1,\tau} - \tau)^2] = \tau(1 - \tau)
\end{aligned}$$

Again, integration by parts and second order Taylor approximation yield the final result for the second moment of  $R(t)$  and for its covariance with  $D$ :

$$E[R^2(t) | T = 1] = o(t^2) \quad (97)$$

$$\text{Cov}[D, R(t) | T = 1] = o(t^2) \quad (98)$$

Therefore,

$$E[A^2(t) | T = 1] = E[D^2 | T = 1]t^2 + 2t\text{Cov}[D, R(t) | T = 1] + E[R^2(t) | T = 1] \quad (99)$$

$$= \tau(1 - \tau)t^2 + o(t^2) \quad (100)$$

Instead of actually computing  $E[A^2(t) | X, T = 1]$  I will use a crude bound for it, which comes from the fact that if the unconditional expectation of  $A^2(t)$  is  $o(t^2)$ , then the conditional one cannot be of a smaller order. Thus  $E[A^2(t) | X, T = 1]$  will be  $o(t^2)$ . The same reason allows to write  $E[R^2(t) | X, T = 1]$  as being  $o(t^2)$ .

Finally, note that

$$\begin{aligned}
E[\tilde{R}_N(t) | T = 1] &= \frac{1}{N} \sum_{i=1}^N E \left[ \frac{T_i R_i(t)}{p(X_i)} - E[R(t) | X_i, T = 1] \frac{T_i - p(X_i)}{p(X_i)} \middle| T = 1 \right] \\
&= E \left[ \frac{R(t)}{p(X)} \middle| T = 1 \right] - E \left[ E[R(t) | X, T = 1] \frac{1 - p(X)}{p(X)} \middle| T = 1 \right] \\
&= E_{X|T=1} \left[ \frac{E[R(t) | X, T = 1]}{p(X)} \right] - E_{X|T=1} \left[ \frac{E[R(t) | X, T = 1] (1 - p(X))}{p(X)} \right] \\
&= E_{X|T=1} [E[R(t) | X, T = 1]] \\
&= E[R(t) | T = 1] \\
&= \frac{1}{2} f_1(q_{1,\tau}) t^2 + o(t^2) \tag{101}
\end{aligned}$$

Hence Equation (48) holds by Equations (46), (87) and (101), that is,  $Q_N(t, \hat{p}) = \tilde{Q}_N(t) + \varepsilon_N(t)$ .  $\square$

**Proof of Lemma 5:** Remember that  $N\varepsilon_N(u/\sqrt{N})$  is a sum of three components:

$$N\varepsilon_N(u/\sqrt{N}) = N(\tilde{R}_N(u/\sqrt{N}) - E[\tilde{R}_N(u/\sqrt{N}) | T = 1]) + N\varepsilon_{1,N}(u/\sqrt{N}) + No(u^2/N)$$

I now show that each one of these components goes to zero in probability for each  $u$ .

Start with the last term,  $No(u^2/N)$ . This goes to zero for each  $u$  by definition.

Now the first part of the sum:  $N(\tilde{R}_N(u/\sqrt{N}) - E[\tilde{R}_N(u/\sqrt{N}) | T = 1])$ . This is mean zero and its variance can be computed by first calculating  $E[\tilde{R}_N^2(t) | T = 1]$ :

$$\begin{aligned}
E[\tilde{R}_N^2(t) | T = 1] &= \frac{1}{N} E \left[ \left( \frac{T R(t)}{p(X)} - E[R(t) | X, T = 1] \frac{T - p(X_i)}{p(X)} \right)^2 \middle| T = 1 \right] \\
&= \frac{1}{N} E \left[ \left( \frac{T(R(t) - E[R(t) | X, T = 1])}{p(X)} + E[R(t) | X, T = 1] \right)^2 \middle| T = 1 \right] \\
&= \frac{1}{N} E \left[ \frac{(R(t) - E[R(t) | X, T = 1])^2}{p(X)^2} + 2 \frac{(R - E[R(t) | X, T = 1]) E[R(t) | X, T = 1]}{p(X)} \right. \\
&\quad \left. + E^2[R(t) | X, T = 1] \middle| T = 1 \right] \\
&= \frac{1}{N} E_{X|T=1} \left[ \frac{1}{p^2(X)} \left( E[R(t) | X, T = 1] - 2E^2[R(t) | X, T = 1] \right. \right. \\
&\quad \left. \left. + E^2[R(t) | X, T = 1] + E^2[R(t) | X, T = 1] p^2(X) \right) \right] \\
&= \frac{1}{N} E_{X|T=1} \left[ \frac{E[R^2(t) | X, T = 1]}{p^2(X)} - \frac{(1 - p^2(X)) E^2[R(t) | X, T = 1]}{p^2(X)} \right] \\
&= \frac{o(t^2)}{N} \tag{102}
\end{aligned}$$

Therefore for fixed  $t$  and each  $u$ :

$$\begin{aligned}
\text{Var} \left( N(\tilde{R}_N(u/\sqrt{N}) - E[\tilde{R}_N(u/\sqrt{N}) | T = 1]) \right) &= N^2 \frac{o(u^2/N)}{N} \\
&= N o(u^2/N) \rightarrow 0 \tag{103}
\end{aligned}$$

Then we can finally conclude that for each  $u$ ,  $N(\tilde{R}_N(u/\sqrt{N}) - E[\tilde{R}_N(u/\sqrt{N}) | T = 1])$  goes to zero in probability.

The missing part to prove Lemma 5 is to prove that for each  $u$ ,  $N \varepsilon_{1,N}(u/\sqrt{N})$  goes to zero in probability. In order to do that, it suffices to show that for fixed  $t$ , Equations (76) to (80) will be  $o_p(1/\sqrt{N})$ .

Hirano, Imbens and Ridder (2002) have computed their first step in the exact same way I do. Also, in their Theorem 1 they have a remainder term to bound very similar to  $\varepsilon_{1,N}(t)$ . The main difference is that their terms do not depend on  $t$ , as instead of  $A(t)/N$  they have  $Y/\sqrt{N}$ , where  $E[Y^2]$  is assumed to be finite.<sup>23</sup> However, it is possible to bound  $\varepsilon_{1,N}(t)$  using exactly the same arguments they used, being just aware that we will possibly have an extra term which will reflect both the dependence on  $t$  and the order of the approximation.<sup>24</sup>

I will show how the analogy between  $\varepsilon_{1,N}(t)$  and the remainder term in Hirano, Imbens and Ridder can be drawn. Consider for instance the absolute value of Equation (76):

$$\left| \frac{1}{N} \sum_{i=1}^N \frac{T_i A_i(t)}{\hat{p}(X_i)} - \frac{T_i A_i(t)}{p(X_i)} + \frac{T_i A_i(t)}{p^2(X_i)} (\hat{p}(X_i) - p(X_i)) \right| \leq \frac{1}{N} \sum_{i=1}^N \left| \frac{T_i A_i(t)}{p^2(X_i) \hat{p}(X_i)} (\hat{p}(X_i) - p(X_i))^2 \right| \quad (104)$$

$$= \frac{1}{N} \sum_{i=1}^N \left| \frac{T_i A_i(t)}{p^2(X_i) \hat{p}(X_i)} (\hat{p}(X_i) - p_K(X_i))^2 \right| \quad (105)$$

$$+ \frac{1}{N} \sum_{i=1}^N \left| \frac{T_i A_i(t)}{p^2(X_i) \hat{p}(X_i)} (p_K(X_i) - p(X_i))^2 \right| \quad (106)$$

$$+ \frac{2}{N} \sum_{i=1}^N \left| \frac{T_i A_i(t)}{p^2(X_i) \hat{p}(X_i)} (\hat{p}(X_i) - p_K(X_i)) (p_K(X_i) - p(X_i)) \right| \quad (107)$$

$$(108)$$

Let me start working with Equation (105) and use the Mean Value Theorem:

$$\frac{1}{N} \sum_{i=1}^N \left| \frac{T_i A_i(t)}{p^2(X_i) \hat{p}(X_i)} (\hat{p}(X_i) - p_K(X_i))^2 \right| = \frac{1}{N} \sum_{i=1}^N \left| \frac{T_i A_i(t)}{p^2(X_i) \hat{p}(X_i)} (L'(H(X_i)' \tilde{\pi}) H(X_i)' (\hat{\pi} - \pi))^2 \right| \quad (109)$$

$$\leq \frac{C}{N} \sum_{i=1}^N |T_i A_i(t)| (H(X_i)' (\hat{\pi} - \pi))^2 + o_p(1) \quad (110)$$

---

<sup>23</sup>In Hirano, Imbens and Ridder (2002), the requirement is in terms of the unconditional  $E(Y^2)$  since for their Theorem 1,  $Y$  was considered to be unobserved if  $T = 0$ . For both cases however, what matters are the moments of functions of  $Y(1)$ , which can be found in the setup of this paper, given unconfoundedness, by conditional moments given  $T = 1$ .

<sup>24</sup>This last difference comes from the fact that they scaled by  $\sqrt{N}$ , while I am using  $N$ .

Inequality (110) holds because of the common support assumption and results (III) and (IV) of Lemma 3. Now an application of the Cauchy-Schwarz inequality and by result (II) of the same Lemma 3, we have:

$$\frac{C}{N} \sum_{i=1}^N |T_i A_i(t)| (H(X_i)'(\hat{\pi} - \pi))^2 + o_p(1) \leq C \zeta^2(N) \|(\hat{\pi} - \pi)\|^2 \frac{1}{N} \sum_{i=1}^N |T_i A_i(t)| + o_p(1) \quad (111)$$

$$\leq C \frac{\zeta^3(N)}{N} \frac{1}{N} \sum_{i=1}^N |T_i A_i(t)| + o_p(1) \quad (112)$$

But note that because  $E[A(t) | T = 1] = \frac{1}{2} f_1(q_{1,\tau}) t^2 + o(t^2)$  and  $E[A^2(t) | T = 1] = o(t^2)$ , then

$$\frac{1}{N} \sum_{i=1}^N |T_i A_i(t)| = a(t) + o_p(1) \quad (113)$$

where  $a(t)$  equals  $E[|A(t)| | T = 1]$ , which is a deterministic function of  $t$ . As  $E[A(t) | T = 1]$  is  $o(t^2)$ ,  $a(t)$  will also be at least  $o(t^2)$ . Therefore:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \left| \frac{T_i A_i(t)}{p^2(X_i) \hat{p}(X_i)} (\hat{p}(X_i) - p_K(X_i))^2 \right| &\leq C \frac{\zeta^3(N)}{N} \frac{1}{N} \sum_{i=1}^N |T_i A_i(t)| + o_p(1) \\ &\leq C a(t) \frac{\zeta^3(N)}{N} + o_p(1) \end{aligned} \quad (114)$$

The same logic could have been applied to Equations (106) and (107) yielding respectively:

$$\frac{1}{N} \sum_{i=1}^N \left| \frac{T_i A_i(t)}{p^2(X_i) \hat{p}(X_i)} (p_K(X_i) - p(X_i))^2 \right| \leq C a(t) \zeta^2(N) K(N)^{-2\frac{\xi}{\tau}} + o_p(1) \quad (115)$$

and

$$\frac{2}{N} \sum_{i=1}^N \left| \frac{T_i A_i(t)}{p^2(X_i) \hat{p}(X_i)} (\hat{p}(X_i) - p_K(X_i)) (p_K(X_i) - p(X_i)) \right| \leq C a(t) \frac{\zeta^{\frac{5}{2}}(N) K(N)^{-\frac{\xi}{\tau}}}{\sqrt{N}} + o_p(1) \quad (116)$$

Now note that these bounds are similar to those computed by Hirano, Imbens and Ridder (2002) for the same sort of approximation. The only difference is that here we have the extra term  $a(t)/\sqrt{N}$ .

Computation of bounds for Equations (77)-(80) follows again the same lines as in Hirano, Imbens and Ridder (2002). In the process of finding bounds for all of those four equations, we will face expressions depending either on  $\frac{1}{N} \sum_{i=1}^N |T_i A_i(t)|$  or on  $E[A(t) | X, T = 1]$ . For the former I have already computed a probabilistic bound. But the latter we have seen to be a random variable such that:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N E[A(t) | X_i, T = 1] &= E[A(t) | T = 1] + o_p(1) \\ &= \frac{1}{2} f_1(q_1, \tau) t^2 + o(t^2) + o_p(1) \\ &= b(t) + o_p(1) \end{aligned} \tag{117}$$

Thus, for each one of Equations (76)-(80) I will have either  $a(t)/\sqrt{N}$  or  $b(t)/\sqrt{N}$  times the expressions for which Hirano, Imbens and Ridder (2002) have found probabilistic bounds. They show indeed that if  $K(N)$  is of the form  $N^\alpha$ ,  $p(x)$  is at least  $4r$ -times differentiable and  $\frac{1}{4(\frac{r}{2}-1)} < \alpha < \frac{1}{9}$  then their bounds are all  $o_p(1)$ . Therefore the sum of the bounds of the Equations (76)-(80) are  $C \frac{a(t)+b(t)}{\sqrt{N}}$  times  $o_p(1)$ , that is to say that for each fixed  $t$ , the sum will be  $o_p(1/\sqrt{N})$ .

Now we are able to verify that:

$$\begin{aligned} N \epsilon_{1,N}(u/\sqrt{N}) &= N o_p \left( \frac{a(u/\sqrt{N}) + b(u/\sqrt{N})}{\sqrt{N}} \right) \\ &= o_p(u^2/\sqrt{N}) \end{aligned} \tag{118}$$

We can finally conclude that for fixed  $t$  and for each  $u$ ,  $N \epsilon_N(u/\sqrt{N})$  goes to zero in probability.  $\square$

**Proof of Lemma 6:** From Equation (57), for result (i) I need to show that  $\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi_{1,i}$  is  $O_p(1)$ . This will follow by Chebyshev inequality:

$$Pr \left[ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi_{1,i} \right| > M \right] < \frac{E[\psi_{1,i}^2]}{M^2} \quad (119)$$

Choosing  $M$  to satisfy  $\frac{E[\psi_{1,i}^2]}{M^2} < \delta$ , where  $\delta$  is a small enough positive constant, there will exist a sample size  $N_\delta$  such that for all  $N > N_\delta$ , Equation (119) will be satisfied.

Result (ii) follows by a Central Limit Theorem; while (iii) follows by noting that  $\psi_1$  is the efficient influence function of  $q_{1,\tau}$ , and therefore, its expected square is  $E[\psi_1^2] = V_1$ , the semiparametric efficiency bound for  $q_{1,\tau}$ .<sup>25</sup>  $\square$

**Proof of Lemma 7:** By convexity of  $N Q_N(u/\sqrt{N}, \hat{p})$ , the following random function must also be convex in  $u$ :

$$\begin{aligned} B_N(u) &= N Q_N(u/\sqrt{N}, \hat{p}) - \sum_{i=1}^N \frac{u}{\sqrt{N}} \left( \frac{T_i D_i}{p(X_i)} - E[D | X_i, T = 1] \frac{T_i - p(X_i)}{p(X_i)} \right) \\ &= \frac{1}{2} f_1(q_{1,\tau}) u^2 + N \epsilon_N(u/\sqrt{N}) \end{aligned} \quad (120)$$

Let me call  $B(u)$  the quadratic  $\frac{1}{2} f_1(q_{1,\tau}) u^2$ .

Now, by convexity of  $B_N(u)$  for any  $u$  such that  $|u - \tilde{u}| = a > \delta$ :

$$\left(1 - \frac{\delta}{a}\right) B_N(\tilde{u}) + \frac{\delta}{a} B_N(u) \geq B_N(\tilde{u} + \delta) \quad (121)$$

By Equation (120), this can be rewritten as:

$$\begin{aligned} \frac{\delta}{a} (B_N(u) - B_N(\tilde{u})) &\geq B(\tilde{u} + \delta) + N \epsilon_N(\tilde{u}/\sqrt{N} + \delta) - \left( B(\tilde{u}) + N \epsilon_N(\tilde{u}/\sqrt{N}) \right) \\ &\geq -2 \sup_{|u - \tilde{u}| \leq \delta} |N \epsilon_N(u/\sqrt{N})| + \inf_{|u - \tilde{u}| = \delta} |B(u) - B(\tilde{u})| \end{aligned} \quad (122)$$

<sup>25</sup> See the proof of the semiparametric efficiency bound in the appendix.

Now, note that

$$\inf_{|u-\tilde{u}|=\delta} |B(u) - B(\tilde{u})| = \frac{1}{2} f_1(q_1, \tau) \delta^2 \quad (123)$$

Thus, for all  $u$  outside the  $\delta$ -interval around  $\tilde{u}$ , if:

$$-2 \sup_{|u-\tilde{u}| \leq \delta} |N \epsilon_N(u/\sqrt{N})| + \frac{1}{2} f_1(q_1, \tau) \delta^2 > 0 \quad (124)$$

then  $\hat{u}$ , the minimizer of  $N Q_N(u/\sqrt{N}, \hat{p})$ , will be inside the  $\delta$ -interval around  $\tilde{u}$ . Hence, I need to show that with probability approaching one, Equation (124) holds.

By the Convexity Lemma,  $\sup_{u \in \mathcal{K}} |N \epsilon_N(t)| = o_p(1)$  for each compact subset  $\mathcal{K}$  of  $\mathbb{R}$ . Define:

$$\mathcal{K}_\delta = \{u \in \mathbb{R}; |u - \tilde{u}| \leq \delta\} \quad (125)$$

Because  $\mathcal{K}_\delta$  is a bounded and closed subset of  $\mathbb{R}$ , it is compact. Therefore:

$$\sup_{u \in \mathcal{K}_\delta} |N \epsilon_N(u/\sqrt{N})| = o_p(1) \quad (126)$$

Thus, for each  $\delta > 0$ :

$$Pr[\sup_{u \in \mathcal{K}_\delta} |N \epsilon_N(u/\sqrt{N})| \geq \frac{1}{4} f_1(q_1, \tau) \delta^2] = o(1) \quad (127)$$

Hence with probability approaching one, for each  $\delta > 0$ , Equation (124) holds, which means that  $\hat{u}$ , the minimizer of  $N Q_N(u/\sqrt{N}, \hat{p})$ , will be inside the  $\delta$ -interval around  $\tilde{u}$  with probability approaching one:

$$|\hat{u} - \tilde{u}| = o_p(1) \quad (128)$$

□

TABLE 1: Lalonde/Dehejia and Wahba Data Set

**QTE and Quantiles of Potentials Outcomes (in 1978 US\$)**

$\tau$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\Delta_{\tau T=1}$	0 (126)	0 (538)	711 (1052)	21 (1357)	1927 (1132)	3879 (1275)	4517 (1461)	6027 (1853)	5503 (3398)
$\Delta_{\tau,exp}$	0	0	930	1163	1081	1446	1797	2246	2919
$\hat{q}_{1,\tau T=1}^C$	0	0	930	2326	4232	6184	8174	10756	14582
$\hat{q}_{0,\tau T=1}^C$	0	0	219	2305	2305	2305	3657	4729	9079
Quantiles of Non-Experimental Control Group	0	8866	13299	17733	20688	24315	27347	31623	38421
Quantiles of Experimental Control Group	0	0	0	1163	3151	4738	6377	8510	11663

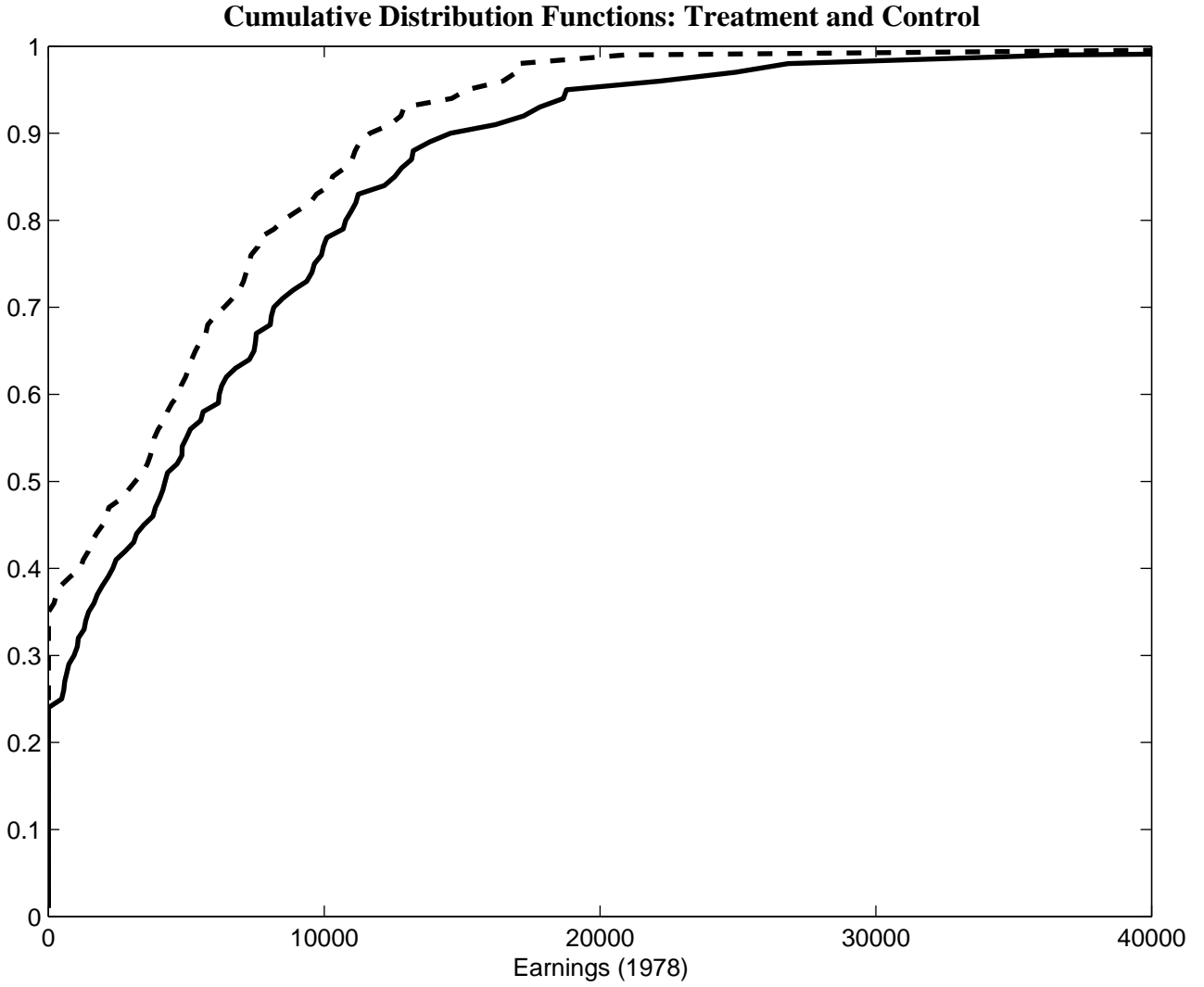


FIGURE 1: Lalonde/Dehejia and Wahba Experimental Data Set (Treatment: solid line; Control: dashed line)

**Cumulative Distribution Functions: Treatment, Actual Control and Counterfactual Control**

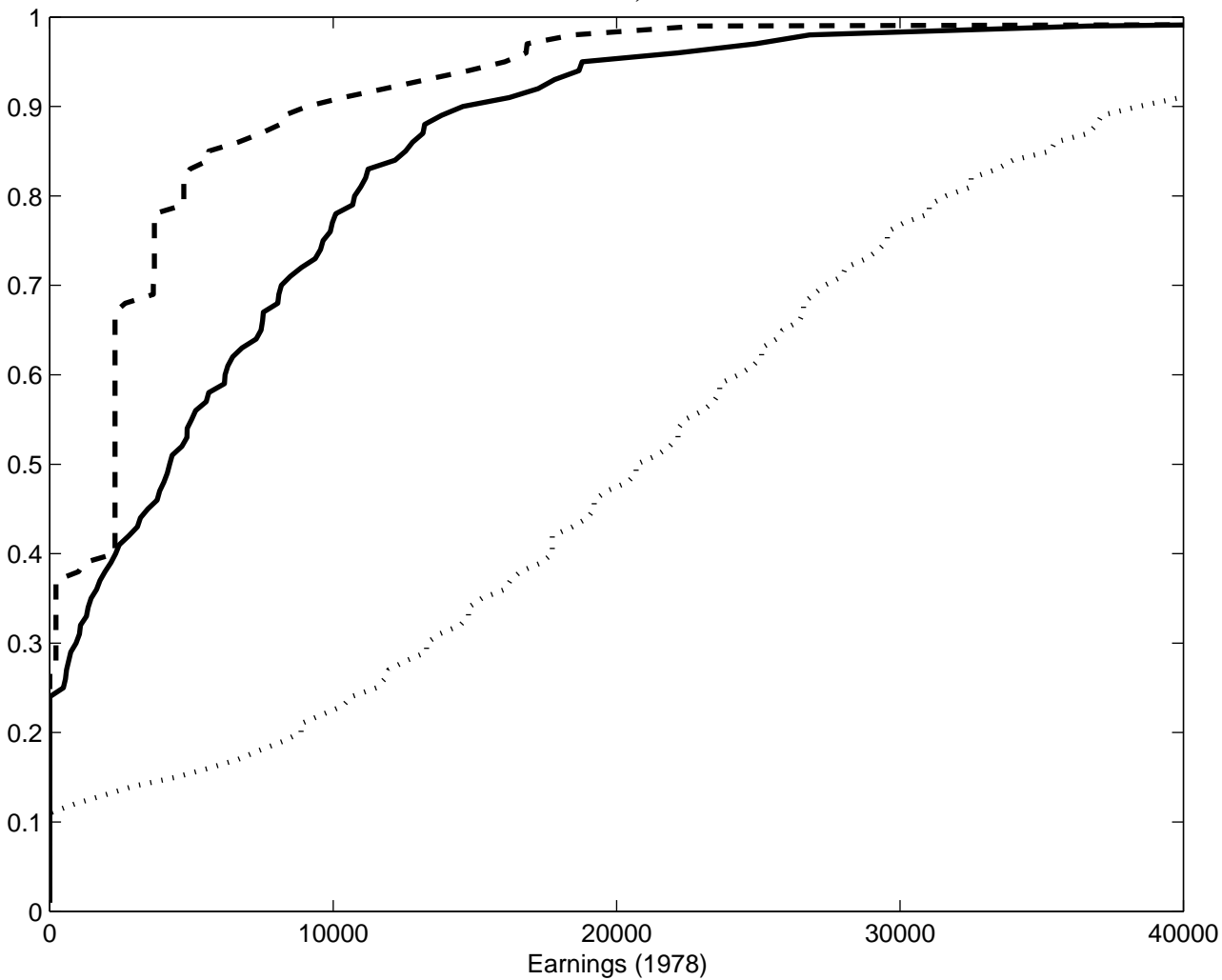


FIGURE 2: Lalonde/Dehejia and Wahba Non-Experimental Data Set (Treatment: solid line; Counterfactual Control: dashed line; Actual Control: dotted line)

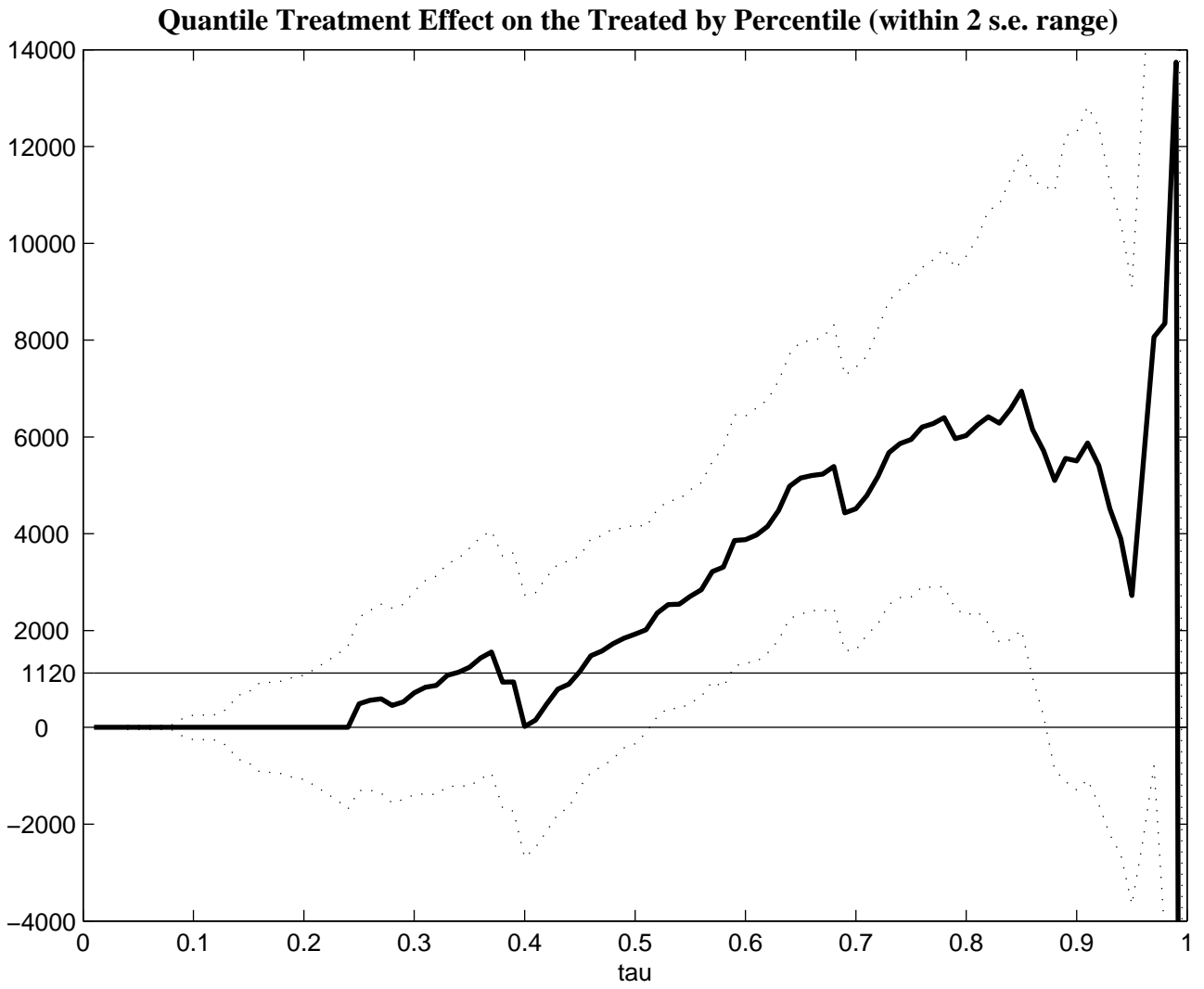


FIGURE 3: Lalonde/Dehejia and Wahba Non-Experimental Data Set

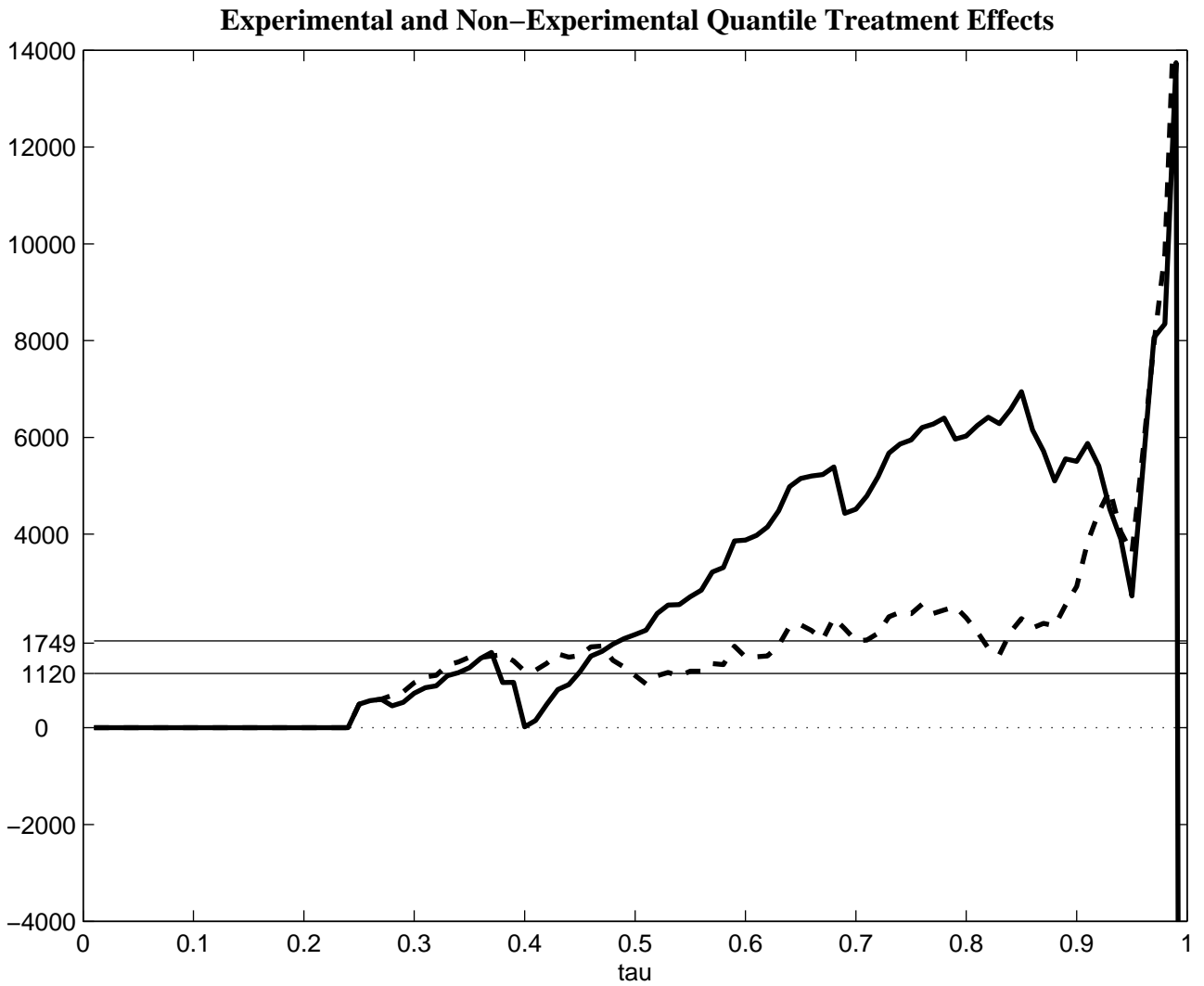


FIGURE 4: Lalonde/Dehejia and Wahba Data Set (Non-experimental QTE: solid line; Experimental QTE: dashed line)