Efficient Semiparametric Estimation of Quantile Treatment Effects*

Sergio Firpo
UC Berkeley - Department of Economics

This Draft: January, 2003

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 Fernando Ferreira, Carlos Flores, Gustavo Gonzaga, Jinyong Hahn, Michael Jansson, David Lee, Thierry Magnac, Rosa Matzkin, Dan McFadden, Joao de Mello, Deb Nolan, David Reinstein, Paul Ruud, Jeffrey Wooldridge and participants of the Econometrics Seminar at UC-Berkeley and of the Economics Seminar at PUC-Rio for comments. Financial support from CAPES - Brazil is acknowledged. All errors are mine. Electronic correspondence: *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 distribution of the treated and of 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 quantiles of the marginal distribution of the outcome. As is the case for any treatment effect parameter, identification restrictions are necessary for this parameter to be estimable. In this paper 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 are differences of minimizers of the sums 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 he 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. Barnow, Cain and Goldberger (1980), Heckman, Ichimura, Smith and Todd (1998) and Dehejia and Wahba (1999) are important examples. Further discussion 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 Dok-

sum (1974) and Lehmann (1974) implicitly argued that an observed individual would maintain his rank in the distribution regardless of his 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.[1]

[1]Note that there is no similar problem in estimation of the average treatment effect, as differences in means

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 $Pr[Y(j) \leq q] = \tau, \quad j = 0,1.$

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 $Pr[Y(j) \leq q|\, T = 1] = \tau, \quad j = 0,1.$

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.

Despite the relevance of QTE, the program evaluation literature on this topic is not as

always coincide with means of differences.

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 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 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:

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

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 functions, under the treatment and the control respectively, of vectors of unobservable attributes. 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 (ε_1, ε_0) on the potential outcomes:

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

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

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, he will weigh the gains and costs to him of both situations. Assume that an individual i predicts his expected earnings (given his vector X_i) and his costs for each of the alternatives. In other words, the individual i chooses the state that yields the largest expected utility:

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

where $u(\cdot)$ is utility function, $C_1(\cdot,\cdot)$ and $C_0(\cdot,\cdot)$ are some costs associated respectively with joining the training program and not joining it, and η_i is a vector of variables that is unobserved to the econometrician but not to the individual. Also, η_i is assumed to be independent of ($\varepsilon_{1,i}$,$\varepsilon_{0,i}$). The effect of η_i on the individual's utility will depend on whether or not he enters the job program. For example, η_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 $E[u(Y(1)) \,|\, X_i, \eta_i] - C_1(X_i, \eta_i) \geq E[u(Y(0)) \,|\, X_i, \eta_i] - C_0(X_i, \eta_i)$. That is:[2]

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

Note how this model fits into the Roy model (1951) of income distribution.[3] 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

[2] The indicator function $\mathbb{1}\{A\}$ is equal to one if A is true and zero otherwise.

[3] 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 η_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:

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

Equation (6) is the the unconfoundedness assumption labeled by Rubin (1977) . 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.[4]

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 $Pr[Y_i(0) \leq q_{0,\tau}]$.[5] Therefore, calculations of the difference $q_{1,\tau} - q_{0,\tau}$ for all τ in the interval [0,1] yield the distribution of the treatment effects.

[4] 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.

[5] 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 on his earnings only. 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.[6] 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:

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

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:

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

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

[6] 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:

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

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

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

Consider the case where for each $\tau \in \{\frac{1}{P}, \frac{2}{P}, ..., 1\}$, $a_{1,\tau} = a_{0,\tau} = a_\tau$. 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:

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

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 τ such that $a_\tau \neq 0$.[7]

A particular case of Equation (12) would be when $a_\tau = 0$ for all τ but for one τ'. 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.

[7] 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 quartiles (.75 and .25), 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 Δ_τ 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 (1983), 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 using instrumental variables (the *selection on unobservables* approach), and calculating bounds for the parameter of interest, as proposed by Manski (1997).[8] Part (ii) states

[8] For review and comparison of approaches see, for instance, Angrist and Krueger (1999) and Heckman, LaLonde and Smith (2000).

that for almost all values of X both treatment assignment levels have a positive probability of occurring.

Now consider that each one of the four types of quantiles defined previously, $q_{1,\tau}$, $q_{0,\tau}$, $q_{1,\tau|T=1}$, and $q_{0,\tau|T=1}$ do exist and are uniquely determined or, in other words, the distribution functions of the potential outcomes are continuous and not flat at the τ-percentile. These conditions appear in the following asumption:

ASSUMPTION 2 ***(Existence and Uniqueness of Quantiles):*** *For $j = 0, 1$, Y(j) is a continuous random variable with support in $\mathbb{R}$ such that for $\tau \in [0,1]$:*
(i) ***Existence****: $Q_{\tau,j} = \{q \in \mathbb{R} \mid \tau = Pr[Y(j) \leq q]\}$ and $Q_{\tau,j|T=1} = \{q \in \mathbb{R} \mid \tau = Pr[Y(j) \leq q | T = 1]\}$ are non-empty.*
*(ii)****Uniqueness****: Let $F_j(q) = Pr[Y(j) \leq q]$ and $F_{j|T=1}(q) = Pr[Y(j) \leq q | T = 1]$.*
Then $\frac{\partial F_j(q)}{\partial q}\Big|_{q=q_{j,\tau}} = f_j(q_{j,\tau}) > 0$ and $\frac{\partial F_{j|T=1}(q)}{\partial q}\Big|_{q=q_{j,\tau|T=1}} = f_{j|T=1}(q_{j,\tau|T=1}) > 0$

Under Assumptions 1 and 2 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 Assumptions 1 and 2, the following equalities hold:*

$\underline{q_{1,\tau}}$:

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

Proof: See Appendix I

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).[9] In fact, the lemma shows that there are 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 $x \in \mathcal{X}$ for the treated subset ($T = 1$) yields the desired result: $q_{1,\tau}$ will be the quantity that sets the expected value equal to τ.

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 τ.

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), be-

[9] An analogous result for $q_{0,\tau}$ would follow from the same lines of Lemma 1. For example, for the group C we would have $\tau = E\left[\frac{(1-T)\mathbb{1}\{Y \leq q_{0,\tau}\}}{1-p(X)}\right]$.

cause the quantile q enters as an argument, complicating computation.[10] 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 Assumptions 1 and 2, the following two sets of equalities hold:*

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

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

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

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

Proof: See Appendix I

[10] 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.

In the proof in the Appendix I, 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 Assumptions 1 and 2, the quantile treatment effect, Δ_τ, 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 Δ_τ 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 $q_{0,\tau|T=1}$, Δ_τ and $\Delta_{\tau|T=1}$are also functionals of the joint distribution of (Y,T,X). Therefore, Δ_τ and $\Delta_{\tau|T=1}$, are identified from data on (Y,T,X). □

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 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 a 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 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 ζ from some general statistical model. Say that this model contains a submodel that can be parameterized by a finite-dimensional parameter θ. 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 θ can be written as $E[\psi s'_\theta]$, where ψ is the influence function of ζ and s_θ is the score of that submodel. The semiparametric efficiency bound V_ζ will be equal to $E[\psi'_\theta \psi_\theta]$, where ψ_θ 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, β, and the average treatment effect on the treated, γ. For the quantile treatment effects setting, I also compute bounds for two parameters, namely, Δ_τ and $\Delta_{\tau|T=1}$. With Assumptions 1 and 2, the semiparametric efficiency bounds for Δ_τ and $\Delta_{\tau|T=1}$ can be calculated:

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

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

and

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

where for $j = 0, 1$:

$$g_{j,\Delta_\tau}(Y) = -\left(\frac{\mathbb{1}\{Y \leq q_{j,\tau}\} - \tau}{f_j(q_{j,\tau})}\right) \quad (15)$$

and

$$g_{j,\Delta_{\tau|T=1}}(Y) = -\left(\frac{\mathbb{1}\{Y \leq q_{j,\tau|T=1}\} - \tau}{f_{j|T=1}(q_{j,\tau|T=1})}\right) \quad (16)$$

Proof: See Appendix I

Note that the bounds V_{Δ_τ} and $V_{\Delta_{\tau|T=1}}$ are similar to the bounds computed by Hahn (1998) for the mean case. For β and γ the bounds, as computed by Hahn (1998), are respectively:[11]

$$V_\beta = E\left[\frac{V[Y|X,T=1]}{p(X)} + \frac{V[Y|X,T=0]}{1-p(X)} + ((E[Y|X,T=1]-\beta_1)-(E[Y|X,T=0]-\beta_0))^2\right]$$

and

$$V_\gamma = E\left[\frac{p(X)V[Y|X,T=1]}{p^2} + \frac{p(X)^2V[Y|X,T=0]}{p^2(1-p(X))} + \frac{p(X)\left((E[Y|X,T=1]-\gamma_1)-(E[Y|X,T=0]-\gamma_0)\right)^2}{p^2}\right].$$

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

[11] Using Hahn's notation, let $\beta_j = E[Y(j)]$ and $\gamma_j = E[Y(j)\,|\,T=1]$ for j=0, 1. Thus, $\beta = \beta_1 - \beta_0$ and $\gamma = \gamma_1 - \gamma_0$.

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 written as differences in expectations of random variables (implicitly for the QTE case) over the same density. This can be seen in the following equations:[12]

$$\Delta_\tau = \text{arg zero}_q E[\mathbb{1}\{Y(1) \leq q\} - \tau] - \text{arg zero}_q E[\mathbb{1}\{Y(0) \leq q\} - \tau] \tag{17}$$

and

$$\beta = E[Y(1)] - E[Y(0)] \tag{18}$$

Note that what ultimately determines the difference in the bounds is the distinction between the random variables $g_{j,\Delta_\tau}(Y(j))$ and $Y(j)$, respectively the influence functions of $q_{j,\tau}$ and of β_j when Y is independent of X.

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 further 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 estima-

[12] Note that this argument could be very well be applied to the comparison between the "on the treated" parameters, $\Delta_{\tau|T=1}$ and γ.

tion. In this section I use the *sample analogy principle*[13] to motivate the appropriateness of the usage of estimators of Δ_τ and $\Delta_{\tau|T=1}$ that are in fact solutions to minimization problems. Restricting then attention to one of the estimators, I present its large sample properties and also show that the asymptotic variance of the proposed estimator achieves the semiparametric efficiency bound.

5.1 MINIMIZATION 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 three different estimators for both the Δ_τ and $\Delta_{\tau|T=1}$ parameters. The estimators 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 $\hat{p}(x)$. In order to simplify the following argument, let me focus only on the three estimators of Δ_τ, which will be, for $E \in \{A,B,C\}$:

$$\hat{\Delta}_\tau^E = \hat{q}_{1,\tau}^E - \hat{q}_{0,\tau}^E \tag{19}$$

where for $j = 0,\ 1$:

$$\hat{q}_{j,\tau}^E = \arg\min_q \sum_{i=1}^N \hat{\omega}_{j,i}^E \rho_\tau(Y_i - q) \tag{20}$$

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

$$\rho_\tau(Y_i - q) = (Y_i - q)(\tau - \mathbb{1}\{Y_i - q \leq 0\})$$

[13] See for instance, Manski (1988)

The previous definitions of the estimators rely on the fact that sample quantiles can be found by minimizing a sum of check functions.[14] In our particular case, we have a weighted sum of check functions, which reflects the fact that as we do not observe the two potential outcomes for the same unit, a rescaling over the observed units is necessary. Also note that for the definition to be complete, I need to determine what the weights $\hat{\omega}^{E}_{j,i}$ are.

Once again for simplification, let us focus on the estimation technique C and concentrate on the sample quantile of the $Y(1)$'s distribution, $\hat{q}^{C}_{1,\tau}$. This sample quantile is defined as the minimizer of a weighted sum, where the weight of each unit is given by:

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

To get some intuition on why $\hat{q}^{C}_{1,\tau}$ is actually consistent for $q_{1,\tau}$, notice that an approximate first derivative of Equation (20) using the weight defined in Equation (21) and evaluated at $\hat{q}^{C}_{1,\tau}$ is equal to:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{T_i(\mathbb{1}\{Y_i \leq \hat{q}^{C}_{1,\tau}\} - \tau)}{\hat{p}(X_i)} \tag{22}$$

As $\hat{q}^{C}_{1,\tau}$ is the minimizer of the convex function expressed in Equation (20) using the weight defined in Equation (21), Equation (22) will converge in probability to zero as N increases. Therefore, Equation (22) is the sample analog of the identifying expression (Q^{C}_{1}) in Lemma 1.

Note that this intuition works also for the other two estimators of $q_{1,\tau}$: $\hat{q}^{A}_{1,\tau}$ and $\hat{q}^{B}_{1,\tau}$. For a more detailed discussion on how to find weights for the cases A and B, see the Appendix II.

The same line of reasoning could have been applied to estimation of $\Delta_{\tau|T=1}$. Each estimator will be defined as the difference between the solutions of two minimizations of sums of weighted check functions. For $E \in \{A,B,C\}$:

[14] See, for instance, Koenker and Bassett (1978).

$$\hat{\Delta}^{E}_{\tau|T=1} = \arg\min_{q} \sum_{i=1}^{N} \hat{\omega}^{E}_{1,i|T=1} \rho_\tau(Y_i - q) - \arg\min_{q} \sum_{i=1}^{N} \hat{\omega}^{E}_{0,i|T=1} \rho_\tau(Y_i - q) \tag{23}$$

In particular, for the estimation procedure indexed by C, the weights are equal to:

$$\hat{\omega}^{C}_{1,i|T=1} = \frac{T_i}{\sum_{l=1}^{N} T_l} \qquad \textit{and} \qquad \hat{\omega}^{C}_{0,i|T=1} = \frac{\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}(1-T_i)}{\sum_{l=1}^{N} T_l} \tag{24}$$

This result will be used later in the paper. Before that, however, let us turn our attention on the computation of the weights used in this subsection. In particular, let us concentrate on the calculation of $\hat{\omega}^{C}_{1,i}$.

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

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

$$G_N(q, \hat{p}) = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i}{\hat{p}(X_i)} (Y_i - q)(\tau - \mathbb{1}\{Y_i \leq q\}) \tag{25}$$

Equation (25) is a weighted sum of check functions. Following Koenker and Bassett (1978), I find sample quantiles as minimizers of sums of check functions. However, I have a weighted sum of check function, as the weights are the way used here to correct for the selection.

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.[15] These functions were chosen to be polynomials of x and the coefficients corresponding to those functions were estimated by maximum likelihood.

[15] The log odds ratio of $p(x)$ is equal to $ln(p(x)) - ln(1-p(x))$.

Start by defining $H_K(x) = [H_{K,j}(x)] \quad (j = 1,...,K)$, a vector of length K of polynomial functions of $x \in \mathcal{X}$ satisfying the following properties:

(*i*) $H_K : \mathcal{X} \to \mathbb{R}^K$;

(*ii*) (Constant included) $H_{K,1}(x) = 1$

If we want $H_K(x)$ 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 N and grows without bounds as N grows without bounds, that is, $K = K(N) \to \infty$ as $N \to \infty$.

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

$$\hat{p}(x) = L(H_K(x)'\hat{\pi}) \tag{26}$$

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

and

$$\hat{\pi} = \arg\max_{\pi} \frac{1}{N} \sum_{i=1}^{N} \left\{ T_i \, ln(L(H_K(X_i)'\pi)) + (1 - T_i) \, ln(1 - L(H_K(X_i)'\pi)) \right\} \tag{27}$$

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

5.3 Large Sample Properties

In this subsection I will prove that $\hat{q}^C_{1,\tau}$ 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 function of $q_{1,\tau}$.[16]

This subsection is divided into several parts, each one corresponds to a step in the proof:

[16] Thus, as $\Delta_\tau = q_{1,\tau} - q_{0,\tau}$ and as it can be shown by analogy that $\hat{q}^C_{0,\tau}$ equally satisfies the properties (*i*), (*ii*) and a properly modified version of (*iii*), the efficient influence function of Δ_τ will be equal to the difference between the efficient influence function of $q_{1,\tau}$ and $q_{0,\tau}$.

1. I state the assumptions and the results derived in Hirano, Imbens and Ridder (2002) for the asymptotics properties of the non-parametric estimation of the p-score in the first step by means of a power series approximation.

2. I use a tranformation from q to $t = q - q_{1,\tau}$ and define $Q_N(t, \hat{p})$, which is minimized by $\hat{t} = \hat{q}^C_{1,\tau} - q_{1,\tau}$.

3. I use another transformation, $u = \sqrt{N}t$, and show that $N\,(Q_N(u/\sqrt{N}, \hat{p}) - \tilde{Q}_N(u/\sqrt{N}))$ is $o_p(1)$ for fixed u, where $\tilde{Q}_N(u/\sqrt{N})$, which does not depend on $\hat{p}(x)$, is a quadratic random function.

4. I show that $\tilde{u}$, the argument that minimizes the random quadratic $N\,\tilde{Q}_N(u/\sqrt{N})$, is: (i) $O_p(1)$; and (ii) $\tilde{u} \xrightarrow{D} N(0, V_1)$, where V_1 is the semiparametric efficiency bound of $q_{1,\tau}$.

5. I show that the term $\hat{u} = \sqrt{N}\hat{t}$ is just $o_p(1)$ from $\tilde{u}$, or written in terms of q, that $\hat{q}^C_{1,\tau}$ is asymptotically equivalent to $\tilde{q}_{1,\tau} = \tilde{u}/\sqrt{N} + q_{1,\tau}$, which establishes the desired result.

5.3.1 Asymptotic Properties 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:

Assumption 3 ***(First Step):***

(i) $\mathcal{X}$ is a compact subset of $\mathbb{R}^r$;

(ii) the density of X, $f(x)$, satisfies $0 < \inf_{x\in\mathcal{X}} f(x) \leq \sup_{x\in\mathcal{X}} f(x) < \infty$

(iii) $p(x)$ is s-times continuously differentiable, where $s \geq 7r$ and r is the dimension of X;

(iv) the order of $H_K(x)$, K, is of the form $K = CN^{\alpha}$ where C is a constant and $\alpha \in (\frac{1}{4(\frac{s}{r}-1)}, \frac{1}{9})$

Newey (1995, 1997) has established that for orthogonal polynomials $H_K(x)$ and compact $\mathcal{X}$:

$$\zeta(K) = \sup_{x \in \mathcal{X}} ||H_K(x)|| \leq CK \tag{28}$$

where C is a generic constant. Note then that because of part (*iv*) of Assumption 3 ζ will be a function of N since K is assumed to be a function of N.

With part (*ii*) of Assumption 1 (Common Support) and 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) $\sup_{x \in \mathcal{X}} |p(x) - p_K(x)| \leq C\zeta(K)K^{-s/r} \leq C\zeta^{1-s/r} \leq CN^{(1-s/r)\alpha} = o(1)$; where:

$$p_K(x) = L(H_K(x)'\pi_K) \tag{29}$$

and:

$$\pi_K = \arg\max_{\pi} E\left\{ p(X)\, ln(L(H_K(X)'\pi)) + (1 - p(X))\, ln(1 - L(H_K(X)'\pi)) \right\}; \tag{30}$$

(II) $E||\hat{\pi} - \pi_K||^2 \leq C\frac{\zeta(K)}{N} \leq CN^{\alpha-1} = o(1)$;

(III) There is $\delta > 0$: $\lim_{N \to \infty} Pr[\delta < \inf_{X \in \mathcal{X}} \hat{p}(X) \leq \sup_{X \in \mathcal{X}} \hat{p}(X) < 1 - \delta] = 1$.

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

Note the importance of result (III) in simplifying the whole process of estimating $q_{1,\tau}$ by $\hat{q}^C_{1,\tau}$. 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.

5.3.2 CHANGE OF VARIABLES: t AND Q_N

First notice that:

$$
\begin{aligned}
\hat{q}^{C}_{1,\tau} &= \arg\min_{q} \frac{1}{N}\sum_{i=1}^{N} \frac{T_i}{\hat{p}(X_i)}(Y_i - q)(\tau - \mathbb{1}\{Y_i \leq q\}) \\
&= \arg\min_{q} \frac{1}{N}\sum_{i=1}^{N} \frac{T_i}{\hat{p}(X_i)}\Big[(Y_i - q)(\tau - \mathbb{1}\{Y_i \leq q\}) - (Y_i - q_{1,\tau})(\tau - \mathbb{1}\{Y_i \leq q_{1,\tau}\})\Big] \\
&= \arg\min_{q} \frac{1}{N}\sum_{i=1}^{N} \frac{T_i}{\hat{p}(X_i)}\Big[(\mathbb{1}\{Y_i \leq q_{1,\tau}\} - \tau)(q - q_{1,\tau}) + (Y_i - q)(\mathbb{1}\{Y_i \leq q_{1,\tau}\} - \mathbb{1}\{Y_i \leq q\})\Big]
\end{aligned} \tag{31}
$$

Now, define:

$$t = q - q_{1,\tau} \tag{32}$$

$$\hat{t} = \hat{q}^{C}_{1,\tau} - q_{1,\tau} \tag{33}$$

$$D(Y_i) = \mathbb{1}\{Y_i \leq q_{1,\tau}\} - \tau \tag{34}$$

$$R(Y_i, t) = (Y_i - (q_{1,\tau} + t))(\mathbb{1}\{Y_i \leq q_{1,\tau}\} - \mathbb{1}\{Y_i \leq q_{1,\tau} + t\}) \tag{35}$$

$$A(Y_i, t) = D(Y_i)\, t + R(Y_i, t) \tag{36}$$

$$Q_N(t, \hat{p}) = \frac{1}{N}\sum_{i=1}^{N} \frac{T_i}{\hat{p}(X_i)} A(Y_i, t) \tag{37}$$

A comment about some of the quantities above: The variable $D(Y_i)$ is the approximate first derivative of the check function $\rho(Y_i - q)$ with respect to q. It is approximate in the sense that $\rho(Y_i - q)$ is not differentiable for all q, as it involves indicator functions of whether q is less than or equal to some values in the data. $R(Y_i, q - q_{1,\tau})$ can be interpreted as the remainder term from a linear expansion about $q_{1,\tau}$ that uses $D(Y_i)$ as an approximated derivative.

Next, note that as $\hat{t} = \hat{q}^{C}_{1,\tau} - q_{1,\tau}$, then by Equation (31) it is also equal:

$$\hat{t} = \arg\min_t \frac{1}{N}\sum_{i=1}^{N} \frac{T_i}{\hat{p}(X_i)} \left[(\mathbb{1}\{Y_i \leq q_{1,\tau}\} - \tau)\, t + (Y_i - (q_{1,\tau} + t))(\mathbb{1}\{Y_i \leq q_{1,\tau}\} - \mathbb{1}\{Y_i \leq q_{1,\tau} + t\}) \right] \tag{38}$$

$$= \arg\min_t \frac{1}{N}\sum_{i=1}^{N} \frac{T_i}{\hat{p}(X_i)} \left[D(Y_i)\, t + R(Y_i, t) \right] \tag{39}$$

$$= \arg\min_t \frac{1}{N}\sum_{i=1}^{N} \frac{T_i}{\hat{p}(X_i)} A(Y_i, t) \tag{40}$$

$$= \arg\min_t Q_N(t, \hat{p}) \tag{41}$$

5.3.3 A Quadratic Approximation to the Objective Function

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

$$\tilde{Q}_N(t) = \frac{1}{N}\sum_{i=1}^{N} t\left(\frac{T_i D(Y_i)}{p(X_i)} - E[D(Y)\,|X_i, T = 1] \frac{T_i - p(X_i)}{p(X_i)} \right) + \frac{f_1(q_{1,\tau})}{2} t^2 \tag{42}$$

Now, define:

$$\varepsilon_N(t) = Q_N(t, \hat{p}) - \tilde{Q}_N(t) \tag{43}$$

and

$$u = \sqrt{N} t \tag{44}$$

The next lemma shows that $N\varepsilon_N(u/\sqrt{N})$ goes to zero in probability for each u, which means that the objective function is asymptotically equivalent to a quadratic random function. Before stating the lemma, let me first assume that the next regularity condition holds:

Assumption 4 ***(Lipschitz condition):*** *For $j = 0,\ 1$ and every t, the conditional density of $Y(j)$ given $X = x$, $f_j(\cdot\,|\,x)$, satisfies the following inequality, where $E[M(X)] < \infty$, and $\lambda > 0$:*

$$\left| f_j(q_{j,\tau}+t\,|\,x) - f_j(q_{j,\tau}\,|\,x) \right| \leq M(x)\,|t|^{\lambda} \tag{45}$$

LEMMA 4 ***(Bounding the differences in the Objective Functions):*** *Under Assumptions 1, 2, 3 and 4, for each u:*

$$N\varepsilon_N(u/\sqrt{N}) \xrightarrow{P} 0 \tag{46}$$

Proof: See Appendix I

5.3.4 ASYMPTOTIC PROPERTIES OF $\tilde{u}$

We have used Assumption 2 previously both for identification of quantiles of the potential outcomes and for an appropriate definition of the efficiency bounds. The same assumption is plays another role in this subsection; it guarantees that $\tilde{u}$, the argument that minimizes $N\tilde{Q}_N(u)$ is unique. From Equation (42) we have:

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

Then under Assumption 2, $N\tilde{Q}_N(u/\sqrt{N})$ has a unique minimum at:

$$\tilde{u} = \arg\min_u \sum_{i=1}^{N} \frac{u}{\sqrt{N}} \left(\frac{T_i D(Y_i)}{p(X_i)} - E[D(Y)\,|X_i, T=1] \frac{T_i - p(X_i)}{p(X_i)} \right) + \frac{u^2}{2} f_1(q_{1,\tau}) \tag{47}$$

$$= -\frac{1}{\sqrt{N}\, f_1(q_{1,\tau})} \sum_{i=1}^{N} \left(\frac{T_i D(Y_i)}{p(X_i)} - E[D(Y)\,|X_i, T=1] \frac{T_i - p(X_i)}{p(X_i)} \right) \tag{48}$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{T_i (g_{1,\Delta_\tau}(Y_i) - E[g_{1,\Delta_\tau}(Y)\,|X_i, T=1])}{p(X_i)} + E[g_{1,\Delta_\tau}(Y)\,|X_i, T=1] \right) \tag{49}$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_{1,i} \tag{50}$$

where the function g_{1,Δ_τ} was defined by Equation (15) and:

$$\psi_{1,i} = \frac{T_i (g_{1,\Delta_\tau}(Y_i) - E[g_{1,\Delta_\tau}(Y)\,|X_i, T=1])}{p(X_i)} + E[g_{1,\Delta_\tau}(Y)\,|X_i, T=1] \tag{51}$$

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

LEMMA 5 ***(Asymptotic Properties of $\tilde{u}$):*** *Let $\tilde{u} = arg\min_u N\tilde{Q}_N(u/\sqrt{N})$. Then, under Assumptions 1, 2 and 3:*

(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 I

5.3.5 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 5 holds. To get results about $\hat{u}$ and consequently about $\hat{q}^C_{1,\tau}$ 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 my case:

LEMMA 6 : ***(Nearness of Argmins (Hjort and Pollard (1993))*** *Under Assumptions 1, 2, 3 and 4 we have the following probabilistic bound on how far $\hat{u}$ can be from $\tilde{u}$: For each $\delta > 0$:*

$$Pr[|\hat{u} - \tilde{u}| \geq \delta] \leq Pr\left[\sup_{|u-\tilde{u}|\leq\delta} |N\varepsilon_N(u/\sqrt{N})| \geq \frac{1}{4} f_1(q_{1,\tau})\delta^2\right] \tag{52}$$

Moreover :

$$Pr\left[\sup_{|u-\tilde{u}|\leq\delta} |N\varepsilon_N(u/\sqrt{N})| \geq \frac{1}{4} f_1(q_{1,\tau})\delta^2\right] = o(1) \tag{53}$$

Proof: See Appendix I

Stating the final results:

THEOREM 2 : ***(Asymptotic Properties of ($\hat{q}^C_{1,\tau}$)*** *Let $\hat{q}^C_{1,\tau} = arg\min_q \frac{1}{N}\sum_{i=1}^N \frac{T_i}{\hat{p}(X_i)}(Y_i - q)(\tau - \mathbb{1}\{Y_i \leq q\})$ where $\hat{p}(x)$ is computed as described in subsection 5.2. Under Assumptions 1, 2, 3 and 4:*

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

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

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

(iii)$V_1 = E[\psi_1^2] = E\left[\frac{V[g_{1,\Delta_\tau}(Y)|X,T=1]}{p(X)} + E^2[g_{1,\Delta_\tau}(Y)|X,T=1]\right]$

Proof: Defining $\tilde{q}_{1,\tau} = \tilde{u}/\sqrt{N} + q_{1,\tau}$, by Lemma 6 we have:

$$\begin{aligned}\sqrt{N}|\hat{q}^C_{1,\tau} - \tilde{q}_{1,\tau}| &= |\sqrt{N}(\hat{q}^C_{1,\tau} - q_{1,\tau}) - \sqrt{N}(\tilde{q}^C_{1,\tau} - q_{1,\tau})| \\ &\leq |\hat{u} - \tilde{u}| \\ &= o_p(1)\end{aligned} \tag{54}$$

That is, $\hat{q}^C_{1,\tau}$ is asymptotically equivalent to $\tilde{q}_{1,\tau}$ and Theorem 2 follows immediately by Lemma 5. □

The same result obtained for $q_{1,\tau}$ could have been obtained analogously for $q_{0,\tau}$. In particular, with the same set of assumptions used in Theorem 2, it is possible to derive an asymptotic linear influence function for $\hat{q}^C_{0,\tau}$, ψ_0, which is analogous to ψ_1. In fact, $\psi_{0,i} = \frac{1-T_i}{1-p(X_i)}(g_{0,\Delta_\tau}(Y_i) - E[g_{0,\Delta_\tau}(Y)\,|\,X_i, T=0]) + E[g_{0,\Delta_\tau}(Y)\,|\,X_i, T=0]$.

A consequence of Theorem 2 is that $\hat{\Delta}^C_\tau$, which is equal to the difference between $\hat{q}^C_{1,\tau}$ and $\hat{q}^C_{0,\tau}$: (i) will also be consistent, (ii) will have an asymptotically linear influence function and, (iii) will be asymptotically normal:

THEOREM 3 *: **(Asymptotic Properties of $\hat{\Delta}^C_\tau$):** Under Assumptions 1, 2, 3 and 4:*

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

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

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

where $\psi_i = \psi_{1,i} - \psi_{0,i}$ *and*

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

Proof: Omitted.

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

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 those used for the overall quantile treatment effect, Δ_τ, to get results similar to 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 well described in LaLonde (1986). The program was designed as an experiment as applicants were randomly assigned into treatment. The treatment was work experience in a wide range of possible activities, like learning to operating a restaurant or a child care center, for a period not exceeding twelve months. Eligible participants were targeted from recipients of AFDC, former addicts, former offenders and young school dropouts. The NSW data set 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 this experimental data set as a benchmark for comparisons with the case 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). I use only a subsample from the PSID, which corresponds the subsample termed "PSID-1" by Dehejia and Wahba (1999). Summary statistics for the two data sets are presented in Table 1. As this table reveals, the non-experimental control group is essentially different from the treated group, which leads us to turn the attention to the parameters of the treatment effect on the treated. In what follows here, the outcome variable is earnings in 1978.[17] As in Dehejia and Wahba (1999) I consider

[17] Earnings are measured in 1982 US Dollars.

male workers only.

LaLonde finds that non-experimental control samples are poor substitutes for experimental data. Some reasons for those findings are described in the survey paper by Heckman, LaLonde and Smith (2000) and explored subsequently by Smith and Todd (2001). Three reasons that do not depend on the estimating procedures but on the data quality of the non-experimental data set are the following. A first reason relies on the fact that the non-experimental data set is not of the same type of the NSW, which implies that same variables are obtained from distinct questions and questionnaires. A second reason is the fact that comparisons groups obtained from surveys that do not cover only the original local labor market where the program took place should not be used to assess the impact of the program on that specific labor market. A third reason is that both data sets must have a sufficient number of relevant variables that explain the participation decision, which might not necessarily be the case for the NSW data set.

The choice of the estimation procedure also contributed for LaLonde's findings on the performance of comparisons using non-experimental samples. Dehejia and Wahba (1999) used the same data set as LaLonde (1986) and reached a different conclusion than LaLonde did. A first reason for the difference in conclusions come from the choice of which pre-program variables to include.[18] Another important difference relies on 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.

In Dehejia and Wahba's 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

[18] Dehejia and Wahba included information on previous two years earnings, what reduced the treated sample in about 40%. See Dehejia and Wahba (1999) and Smith and Todd (2001).

for the selection problem. One of these methods, reweighing using the estimated p-score, uses exactly the weights described by Equation (24), $\hat{\omega}^C_{1,i|T=1}$ and $\hat{\omega}^C_{0,i|T=1}$.

One data set Dehejia and Wahba use is a subset of 185 treated units and 2490 control observations from the PSID.[19] 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 $\sum_{i=1}^{n}(\hat{\omega}^C_{1,i|T=1} - \hat{\omega}^C_{0,i|T=1})Y_i$. For these specific treatment and control groups they find an average treatment effect on the treated of US$ 1129.[20] This is lower than the unadjusted experimental treatment effect of $1749, but larger than the initial numbers LaLonde computed using the non-experimental data.[21]

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 2 and in Figures 1 to 5.[22] I find that using experimental controls, treatment effects tend to be 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. Despite the fact that the counterfactual c.d.f. of the control group introduces a heterogeneity in effects not seen by using the experimental control, the difference between the two lies around zero, as it is shown by Figure 5.

An important feature of the estimated counterfactual distribution is that there are some

[19] As mentioned earlier 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.

[20] I replicated their calculations using the same p-score specification and got a slightly different number, $1120.

[21] The unadjusted for covariates treatment effect was computed using the experimental control sample of size 260. It is a simple difference in means between treated and control groups.

[22] In Table 2 and in Figures 3 and 5, the standard errors were computed by 100 bootstrap replications.

discrete jumps in the c.d.f., as some points had probability mass. A closer look at the data reveals that these points correspond to the observations from the non-experimental control sample that have the largest values of the estimated propensity-score and, therefore, the highest weight values. These "leverage points" are important in the sense that their large weights compensate the greater number of comparable treated individuals. For example, there is only one individual in the non-experimental control group that reported earnings of $2305 but his estimated p-score is larger than .98, what leads to a weight (.41) more than 600 times larger than the average weight (.0006) of the control group.

To assess the importance of the weights in the described method of finding the counterfactual distributions, consider a very simple example, in which we have just two data points, the first one with an outcome equals to 1 and a weight equals to 10; and a second point with 10 as outcome and 1 as weight. For this simple example, the function to be minimized is $10(1-q)(\tau - \mathbb{1}\{q \geq 1\}) + (10-q)(\tau - \mathbb{1}\{q \geq 10\})$. We can check that the value of the function at $q = 1$ is smaller than its value at $q = 10$ for all $\tau < 90/99$. This example, although very simplified, helps explaining the presence of jumps in the estimated counterfactual distribution.

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 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. Ashefelter and D. Card, 1277-1366, New York, Elsevier Science B.V.

AMEMIYA, T., (1982), "Two-Stage Least Absolute Deviations Estimators," *Econometrica*, 50, 689-711.

BARNOW, B., G. CAIN, AND A. GOLDBERGER, (1980), "Issues in the Analysis of the Selectivity Bias," *Evaluation Studies*, 5, 42-59.

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. Ashefelter 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*.

IMBENS, G., AND D. RUBIN, (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economic Studies*, October, 555-574.

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.

NEWEY, W., (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99-135.

NEWEY, W., (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349-1382.

NEWEY, W., (1995), "Convergence Rates for Series Estimators," in *Advances in Econometrics and Qualitatice Economics: Essays in Honor of C.R. Rao*, G. Maddal, P.C. Phillips, and T.N. Srinivasan, eds., Cambridge US, Basil-Blackwell.

NEWEY, 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 I

Proof of Lemma 1:

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

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

The first equality follows from the definition of $q_{1,\tau}$ and from Assumption 2. 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 $E[Z \,|\, X] = p(X)E[Z \,|\, X, T=1] + (1-p(X))E[Z \,|\, X, T=1]$, 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. □

Proof of Lemma 2:

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

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

The first equality follows from the definition of $q_{1,\tau|T=1}$ and from Assumption 2. 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 $E[Z \,|\, X] = p(X)E[Z \,|\, X, T=1] + (1-p(X))E[Z \,|\, X, T=1]$, where Z is some random variable. Finally, the last equality is a backward application of the law of iterated expectations.

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

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

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 $T = 0$. Eigth and ninth equalities hold by the same reasons the last two equalities for the $q_{1,\tau|T=1}$ case hold. □

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 σ-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:

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

and

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

where $f_1(y\,|\,x) = \int f(y,z\,|\,x)dz$ and $f_0(y\,|\,x) = \int f(z,y\,|\,x)dz$.

Working with the density of observed data, consider the regular parametric submodel indexed by θ, 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 $\phi(y,t,x) = \phi(y,t,x\,|\,\theta_0)$.

The score of a parametric submodel indexed by θ is given by:

$$s(y,t,x\,|\,\theta) = t\,s_1(y\,|\,x;\theta) + (1-t)s_0(y\,|\,x;\theta) + \frac{t-p(x\,|\,\theta)}{p(x\,|\,\theta)(1-p(x\,|\,\theta))}p'(x\,|\,\theta) + s_x(x\,|\,\theta)$$

where, for $j = 0, 1$:

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

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

and

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

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

In order to find the efficient influence functions of the parameters of interest, $\Delta_\tau(\theta)$ and $\Delta_{\tau|T=1}(\theta)$, I need first to define the tangent space of this statistical model. This will be the set $\mathcal{S}$ of all possible score functions, and it is defined as:

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

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

Next I show that both $\Delta_\tau(\theta)$ and $\Delta_{\tau|T=1}(\theta)$ are pathwise differentiable, that is, I show that for each one the derivative with respect to θ evaluated at θ_0 is equal to the expectation of the product of the score $s(Y,T,X)$ and the respective influence functions $\psi_{\Delta_\tau}(Y,T,X)$ and $\psi_{\Delta_{\tau|T=1}}(Y,T,X)$ 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. If an influence function belongs to the set $\mathcal{S}$, then its projection onto $\mathcal{S}$ is the original influence function itself. Therefore, the goal is to find an influence function which already belongs to the set of scores. A function that is in the set of the scores must be written as:

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

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

Starting with $q_{1,\tau}$, the first part of the parameter Δ_τ. For the parametric submodel indexed by θ, we have for all θ:

$$0 = \iint (\mathbb{1}\{y \leq q_{1,\tau}(\theta)\} - \tau) f_1(y \mid x; \theta) f(x \mid \theta) dy dx \tag{55}$$

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

$$\begin{aligned} 0 = & f_1(q_{1,\tau}) \frac{\partial q_{1,\tau}(\theta_0)}{\partial \theta} + \iint (\mathbb{1}\{y \leq q_{1,\tau}\} - \tau) s_1(y \mid x) f_1(y \mid x) f(x) dy dx \\ & + \iint (\mathbb{1}\{y \leq q_{1,\tau}\} - \tau) s_x(x) f_1(y \mid x) f(x) dy dx \end{aligned} \tag{56}$$

Note that:

$$\iint s_1(y \mid x) f_1(y \mid x) f(x) dy dx = 0 \tag{57}$$

$$\iint s_x(x) f_1(y \mid x) f(x) dy dx = \int s_x(x) f(x) dx = 0 \tag{58}$$

Hence the derivative of $q_{1,\tau}(\theta)$ evaluated at θ_0 is equal to:

$$\frac{\partial q_{1,\tau}(\theta_0)}{\partial \theta} = -\frac{\iint \mathbb{1}\{y \leq q_{1,\tau}\} s_1(y \mid x) f_1(y \mid x) f(x) dy dx}{f_1(q_{1,\tau})} - \frac{\iint \mathbb{1}\{y \leq q_{1,\tau}\} s_x(x) f_1(y \mid x) f(x) dy dx}{f_1(q_{1,\tau})} \tag{59}$$

After similar calculations for $q_{0,\tau}$, we can express the derivative of $\Delta_\tau(\theta)$ evaluated at θ_0 as:

$$\begin{aligned} \frac{\partial \Delta_\tau(\theta_0)}{\partial \theta} = & -\frac{\iint \mathbb{1}\{y \leq q_{1,\tau}\} s_1(y \mid x) f_1(y \mid x) f(x) dy dx}{f_1(q_{1,\tau})} + \frac{\iint \mathbb{1}\{y \leq q_{0,\tau}\} s_0(y \mid x) f_0(y \mid x) f(x) dy dx}{f_0(q_{0,\tau})} \\ & - \frac{\iint \mathbb{1}\{y \leq q_{1,\tau}\} s_x(x) f_1(y \mid x) f(x) dy dx}{f_1(q_{1,\tau})} + \frac{\iint \mathbb{1}\{y \leq q_{0,\tau}\} s_x(x) f_0(y \mid x) f(x) dy dx}{f_0(q_{0,\tau})} \end{aligned} \tag{60}$$

The next goal is to find a function of (Y, T, X) such that the expectation of the product of that function times the score is equal to Equation (60). A solution to this problem is the following:

$$\psi_{\Delta_\tau}(Y,T,X) = \frac{T(g_{1,\Delta_\tau}(Y) - E[g_{1,\Delta_\tau}(Y)\,|\,X,T=1])}{p(X)} - \frac{(1-T)(g_{0,\Delta_\tau}(Y) - E[g_{0,\Delta_\tau}(Y)\,|\,X,T=0])}{1-p(X)} + E[g_{1,\Delta_\tau}(Y)\,|\,X,T=1] - E[g_{0,\Delta_\tau}(Y)\,|\,X,T=0] \tag{61}$$

where the function g_{j,Δ_τ} was defined in Equation (15).

Note however that this influence function belongs to the set of the scores. In order to check that, we need only to verify that the following three equalities hold:

$$E\left[\frac{g_{1,\Delta_\tau}(Y) - E[g_{1,\Delta_\tau}(Y)\,|\,X,T=1]}{p(X)}\,\middle|\,X,T=1\right] = 0 \tag{62}$$

$$E\left[\frac{g_{0,\Delta_\tau}(Y) - E[g_{0,\Delta_\tau}(Y)\,|\,X,T=0]}{1-p(X)}\,\middle|\,X,T=0\right] = 0 \tag{63}$$

$$E\Big[E[g_{1,\Delta_\tau}(Y)\,|\,X,T=1] - E[g_{0,\Delta_\tau}(Y)\,|\,X,T=0]\Big] = 0 \tag{64}$$

Equations (62) and (63) hold by inspection. By the definition of g_{j,Δ_τ}, $E[g_{j,\Delta_\tau}(Y)\,|\,X,T=j] = 0$, so Equation (64) also holds. Hence, ψ_{Δ_τ} is the efficient influence function and has expected value equal to zero, since it is in the set of scores. Thus its variance is equal to $E[\psi^2_{\Delta_\tau}(Y,T,X)]$, which is the semiparametric efficiency bound for Δ_τ, V_{Δ_τ}.

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

$$0 = \iint \frac{p(x\,|\,\theta)}{\int p(x\,|\,\theta) f(x\,|\,\theta)dx}(\mathbb{1}\{y \le q_{1,\tau|T=1}(\theta)\} - \tau) f_1(y\,|\,x;\theta) f(x\,|\,\theta)dydx \tag{65}$$

Again I normalize: $q_{1,\tau|T=1} = q_{1,\tau|T=1}(\theta_0)$. The derivative evaluated at θ_0 is equal to:

$$\begin{aligned}\frac{\partial q_{1,\tau|T=1}(\theta_0)}{\partial\theta} = -\frac{1}{f_{1|T=1}(q_{1,\tau|T=1})}\Bigg(&\iint \mathbb{1}\{y \le q_{1,\tau|T=1}\} p(x) s_1(y\,|\,x) f_1(y\,|\,x) f(x)dydx \\ &+ \int (E[\mathbb{1}\{y \le q_{1,\tau|T=1}\}\,|\,X=x] - \tau) p'(x) f_1(y\,|\,x) f(x)dydx \\ &+ \int (E[\mathbb{1}\{y \le q_{1,\tau|T=1}\}\,|\,X=x] - \tau) p(x) s_x(x) f_1(y\,|\,x) f(x)dydx\Bigg)\end{aligned} \tag{66}$$

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 θ_0 as being:

$$\begin{aligned}\frac{\partial\Delta_{\tau|T=1}(\theta_0)}{\partial\theta} = &-\frac{\iint \mathbb{1}\{y\leq q_{1,\tau|T=1}\}p(x)s_1(y|x)f_1(y|x)f(x)dydx}{f_{1|T=1}(q_{1,\tau|T=1})}\\ &+\frac{\iint \mathbb{1}\{y\leq q_{0,\tau|T=1}\}p(x)s_0(y|x)f_0(y|x)f(x)dydx}{f_{0|T=1}(q_{0,\tau|T=1})}\\ &-\frac{\int(E[\mathbb{1}\{y\leq q_{1,\tau|T=1}\}|X=x]-\tau)p'(x)f_1(y|x)f(x)dydx}{f_{1|T=1}(q_{1,\tau|T=1})}\\ &+\frac{\int(E[\mathbb{1}\{y\leq q_{0,\tau|T=1}\}|X=x]-\tau)p'(x)f_0(y|x)f(x)dydx}{f_{0|T=1}(q_{0,\tau|T=1})}\\ &-\frac{\int(E[\mathbb{1}\{y\leq q_{1,\tau|T=1}\}|X=x]-\tau)p(x)s_x(x)f_1(y|x)f(x)dydx}{f_{1|T=1}(q_{1,\tau|T=1})}\\ &+\frac{\int(E[\mathbb{1}\{y\leq q_{0,\tau|T=1}\}|X=x]-\tau)p(x)s_x(x)f_0(y|x)f(x)dydx}{f_{0|T=1}(q_{0,\tau|T=1})}\end{aligned} \tag{67}$$

The efficient influence function for this case is equal:

$$\begin{aligned}\psi_{\Delta_{\tau|T=1}}(Y,T,X) = &\frac{T(g_{1,\Delta_{\tau|T=1}}(Y)-E[g_{1,\Delta_{\tau|T=1}}(Y)|X,T=1])}{p}\\ &-\frac{(1-T)p(X)(g_{0,\Delta_{\tau|T=1}}(Y)-E[g_{0,\Delta_{\tau|T=1}}(Y)|X,T=0])}{p(1-p(X))}\\ &+\frac{(T-p(X))}{p}(E[g_{1,\Delta_{\tau|T=1}}(Y)|X,T=1]-E[g_{0,\Delta_{\tau|T=1}}(Y)|X,T=0])\\ &+\frac{p(X)}{p}(E[g_{1,\Delta_{\tau|T=1}}(Y)|X,T=1]-E[g_{0,\Delta_{\tau|T=1}}(Y)|X,T=0])\end{aligned} \tag{68}$$

where the function $g_{j,\Delta_{\tau|T=1}}$ was defined by Equation (16).

As this influence function is in the set of scores, its expected value is zero and its variance is equal to $E[\psi^2_{\Delta_{\tau|T=1}}(Y,T,X)]$, 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 first to decompose $\varepsilon_N(t) = Q_N(t,\hat{p}) - \tilde{Q}_N(t)$ into three parts:

$$\varepsilon_N(t) = \left(\tilde{R}_N(t) - E[\tilde{R}_N(t)]\right) + \varepsilon_{1,N}(t) + o(t^2) \tag{69}$$

We saw that $A(Y,t)$ can be decomposed into two parts, $D(Y)\,t$ and $R(Y,t)$. Notice however, that we will be interested here in an approximation of $\frac{1}{N}\sum_{i=1}^{N}\frac{T_i D(Y_i)\,t}{\hat{p}(X_i)}$ by an expression that does not depend on $\hat{p}(X)$. An approximation that is analogous to first part of the sum of Equation (42) but that uses $R(Y,t)$ in the place of $D(Y)\,t$ is defined by $\tilde{R}_N(t)$ and written as:

$$\tilde{R}_N(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{T_i R(Y_i,t)}{p(X_i)} - E[R(Y,t)\,|X_i, T=1]\,\frac{T_i - p(X_i)}{p(X_i)}\right] \tag{70}$$

To proceed, I first show that Equation (69) actually holds. Then I show that N times each one of the three parts of Equation (69) evaluated at $u/\sqrt{N}$ will converge in probability to zero for each u.

I start by summing and subtracting several terms from $Q_N(t,\hat{p})$.

$$Q_N(t,\hat{p}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{T_i A(Y_i,t)}{\hat{p}(X_i)} - \frac{T_i A(Y_i,t)}{p(X_i)} + \frac{T_i A(Y_i,t)}{p^2(X_i)}(\hat{p}(X_i)-p(X_i))\right) \tag{71}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\left(\frac{T_i A(Y_i,t)}{p^2(X_i)}(\hat{p}(X_i)-p(X_i))\right) + E\left[\frac{E[A(Y,t)\,|\,X,T=1]}{p(X)}(\hat{p}(X)-p(X))\right] \tag{72}$$

$$-E\left[\frac{E[A(Y,t)\,|\,X,T=1]}{p(X)}(\hat{p}(X)-p(X))\right] - \frac{1}{N}\sum_{i=1}^{N}\tilde{\delta}(X_i,t)\frac{T_i - p_K(X_i)}{\sqrt{p_K(X_i)(1-p_K(X_i))}} \tag{73}$$

$$+\frac{1}{N}\sum_{i=1}^{N}(\tilde{\delta}(X_i,t)-\delta_K(X_i,t))\frac{T_i - p_K(X_i)}{\sqrt{p_K(X_i)(1-p_K(X_i))}} \tag{74}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\delta_K(X_i,t)\frac{T_i - p_K(X_i)}{\sqrt{p_K(X_i)(1-p_K(X_i))}} - \frac{1}{N}\sum_{i=1}^{N}\delta(X_i,t)\frac{T_i - p(X_i)}{\sqrt{p(X_i)(1-p(X_i))}} \tag{75}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\left(\frac{T_i A_i(t)}{p(X_i)} - E[A(t)\,|X_i,T=1]\frac{T_i - p(X_i)}{p(X_i)}\right) \tag{76}$$

where:

$$\tilde{\delta}(X_i,t) = -E\left[\frac{E[A(Y,t)\,|\,X,T=1]}{p(X)}L'(H_K(X)'\tilde{\pi})H_K(X)'\right]\tilde{\Sigma}^{-1}\sqrt{L'(H_K(X_i)'\pi_K)}H_K(X_i) \tag{77}$$

$$\delta_K(X_i,t) = -E\left[\frac{E[A(Y,t)\,|\,X,T=1]}{p(X)}L'(H_K(X)'\pi_K)H_K(X)'\right]\Sigma_K^{-1}\sqrt{L'(H_K(X_i)'\pi_K)}H_K(X_i) \tag{78}$$

$$\delta(X_i,t) = -E[A(Y,t)\,|X_i,T=1]\frac{\sqrt{p(X_i)(1-p(X_i))}}{p(X_i)} \tag{79}$$

$$\tilde{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} H_K(X_i)H_K(X_i)'L'(H_K(X_i)'\tilde{\pi}) \tag{80}$$

$$\Sigma = E[H_K(X)H_K(X)'L'(H_K(X)'\pi_K)] \tag{81}$$

Thus, by Equations (36) and (42):

$$Q_N(t,\hat{p}) = \frac{t}{N}\sum_{i=1}^{N}\left(\frac{T_i D(Y_i)}{p(X_i)} - E[D(Y)\,|X_i, T=1]\frac{T_i - p(X_i)}{p(X_i)}\right)$$

$$+\tilde{R}_N(t) - E[\tilde{R}_N(t)] + E[\tilde{R}_N(t)] + \varepsilon_{1,N}(t) \tag{82}$$

$$\tag{83}$$

where $\varepsilon_{1,N}(t)$ is equal to the sum of Equations (71) to (75).

In order to decompose $Q_N(t,\hat{p})$ into the sum of $\tilde{Q}_N(t)$ and $\varepsilon_N(t)$, from Equation (82) I show that $E[\tilde{R}_N(t)] = E[R(Y(1),t)] = \frac{t^2}{2} f_1(q_{1,\tau}) + o(t^2)$. I will do more than that. In fact, let me compute the first two conditional moments of $A(Y(1),t)$ given X and its first two unconditional moments.

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

$$E[D(Y(1))\,|X] = E[\mathbb{1}\{Y(1) \le q_{1,\tau}\} - \tau\,|X] \tag{84}$$

and

$$\begin{aligned} E[D(Y(1))] =& E[E[\mathbb{1}\{Y(1) \leq q_{1,\tau}\} - \tau \,|\, X]] \\ =& E[\mathbb{1}\{Y(1) \leq q_{1,\tau}\} - \tau] \\ =& 0 \end{aligned} \tag{85}$$

In order to compute $E[R(Y(1),t) \,|\, X = x]$, I will need to do integration by parts and use the Mean Value Theorem:

$$\begin{aligned} E[R(Y(1),t) \,|\, X = x] =& E[(Y(1) - (q_{1,\tau} + t))(\mathbb{1}\{Y(1) \leq q_{1,\tau}\} - \mathbb{1}\{Y(1) \leq q_{1,\tau} + t\}) \,|\, X = x] \\ =& \int_{q_{1,\tau}+t}^{q_{1,\tau}} (y - (q_{1,\tau} + t)) f_1(y \,|\, x) dy \\ =& (y - (q_{1,\tau} + t)) F_1(y \,|\, x) \Big|_{q_{1,\tau}+t}^{q_{1,\tau}} + \int_{q_{1,\tau}}^{q_{1,\tau}+t} F_1(y \,|\, x) dy \\ =& - F_1(q_{1,\tau} \,|\, x) t + F_1(q_{1,\tau} \,|\, x) t + \frac{1}{2} f_1(q_{1,\tau} + t^*(x,t) \,|\, x) t^2 \\ =& \frac{1}{2} f_1(q_{1,\tau} + t^*(x,t) \,|\, x) t^2 \end{aligned} \tag{86}$$

where $t^*(x,t)$ is some real number between 0 and t.

Under Assumption 4, the unconditional expectation of $R(Y(1),t)$ can be found by noticing the following:[23]

$$\begin{aligned} \left| E[R(Y(1),t)] - \frac{1}{2} f_1(q_{1,\tau}) t^2 \right| \leq& E\left[\left| E[R(Y(1),t) \,|\, X] - \frac{1}{2} f_1(q_{1,\tau} \,|\, X) t^2 \right| \right] \\ =& E\left[\left| \frac{1}{2} f_1(q_{1,\tau} + t^*(X,t) \,|\, X) t^2 - \frac{1}{2} f_1(q_{1,\tau} \,|\, X) t^2 \right| \right] \\ \leq& \frac{t^2}{2} E[M(X)] |t|^{\lambda} \\ =& o(t^2) \end{aligned} \tag{87}$$

[23] 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 $o(t^2) \to 0$ as $t \to 0$. 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 $t^2 a_N$.

Thus:

$$E[R(Y(1),t)] = \frac{1}{2} f_1(q_{1,\tau}) t^2 + o(t^2) \tag{88}$$

Finally, we have:

$$E[A(Y(1),t) \,|\, X = x] = t\, E[\mathbb{1}\{Y(1) \leq q_{1,\tau}\} - \tau \,|\, X = x] + \frac{t^2}{2} f_1(q_{1,\tau} + t^*(x,t) \,|\, x) \tag{89}$$

and

$$E[A(Y(1),t)] = \frac{1}{2} f_1(q_{1,\tau}) t^2 + o(t^2) \tag{90}$$

Now I compute the conditional second moment of $A(Y(1),t)$:

$$\begin{aligned} E[A^2(Y(1),t) \,|\, X = x] =& t^2 E[D^2(Y(1)) \,|\, X = x] + 2t\, E[D(Y(1)), R(Y(1),t) \,|\, X = x] \\ &+ E[R^2(Y(1),t) \,|\, X = x] \end{aligned} \tag{91}$$

where:

$$E[D^2(Y(1)) \,|\, X = x] = \int (\mathbb{1}\{y \leq q_{1,\tau}\} - \tau)^2 f_1(y \,|\, x) dy \tag{92}$$

The conditional expectation $E[R^2(Y(1),t) \,|\, X = x]$ is computed using similar steps to those used for the computation of the first conditional moment of $R(Y(1),t)$:

$$E[R^2(Y(1),t) \,|\, X = x] = \int_{q_{1,\tau}+t}^{q_{1,\tau}} (y - (q_{1,\tau} + t))^2 (\mathbb{1}\{y \leq q_{1,\tau}\} - \mathbb{1}\{y \leq q_{1,\tau} + t\}) f_1(y \,|\, x) dy \tag{93}$$

Consider the case in which $t > 0$:[24]

[24] The $t < 0$ case yields the same result times (-1).

$$
\begin{aligned}
E[R^2(Y(1),t)\,|\,X=x] &= \int_{q_{1,\tau}}^{q_{1,\tau}+t} (y-(q_{1,\tau}+t))^2 f_1(y\,|\,x)dy \\
&= (y-(q_{1,\tau}+t))^2 F_1(y\,|\,x)\Big|_{q_{1,\tau}}^{q_{1,\tau}+t} - 2\int_{q_{1,\tau}}^{q_{1,\tau}+t} (y-(q_{1,\tau}+t))F_1(y\,|\,x)dy \\
&= -F_1(q_{1,\tau}\,|\,x)t^2 + 2\left(\frac{1}{2}F_1(q_{1,\tau}\,|\,x)t^2 + \frac{1}{6}f_1(q_{1,\tau}+t^{**}(t,x)\,|\,x)t^3\right) \\
&= \frac{1}{3}f_1(q_{1,\tau}+t^{**}(t,x)\,|\,x)t^3
\end{aligned}
\tag{94}
$$

where where $t^{**}(x,t)$ is some real number between 0 and t.

The cross-term $E[D(Y(1))R(Y(1),t)\,|\,X=x]$ equals to:

$$
E[D(Y(1))R(Y(1),t)\,|\,X=x] = \int_{q_{1,\tau}+t}^{q_{1,\tau}} (y-(q_{1,\tau}+t))(\mathbb{1}\{y\le q_{1,\tau}\}-\tau)f_1(y\,|\,x)dy \tag{95}
$$

For $t>0$:

$$
\begin{aligned}
E[D(Y(1))R(Y(1),t)\,|\,X=x] &= \tau\int_{q_{1,\tau}}^{q_{1,\tau}+t} (y-(q_{1,\tau}+t))f_1(y\,|\,x)dy \\
&= -\tau E[R(Y(1),t)\,|\,X=x]
\end{aligned}
\tag{96}
$$

while for $t<0$:

$$
E[D(Y(1))R(Y(1),t)\,|\,X=x] = (1-\tau)E[R(Y(1),t)\,|\,X=x] \tag{97}
$$

Therefore:

$$
E[D(Y(1))R(Y(1),t)\,|\,X=x] = \frac{1}{2}f_1(q_{1,\tau}+t^*(x,t)\,|\,x)t^2(\mathbb{1}\{t<0\}-\tau) \tag{98}
$$

Calculations similar to those used to find Equation (88), which are based on Assumption 4 guarantee that:

$$E[R^2(Y(1),t)] = E[E[R^2(Y(1),t)\,|\,X]] = \frac{t^3}{3} f_1(q_{1,\tau}) + o(t^3) = O(t^3) \tag{99}$$

$$E[D(Y(1))R(Y(1),t)] = E[E[D(Y(1))R(Y(1),t)\,|\,X]] = O(t^2) \tag{100}$$

Also, we know that

$$E[D(Y(1))^2] = \tau(1-\tau) \tag{101}$$

Therefore,

$$E[A^2(Y(1),t)\,|\,X] = E[D^2(Y(1))\,|\,X]\,t^2 + 2tE[D(Y(1))R(Y(1),t)\,|\,X] + E[R^2(Y(1),t)\,|\,X] \tag{102}$$

and

$$E[A^2(Y(1),t)] = E[D^2(Y(1))]\,t^2 + 2tE[D(Y(1))R(Y(1),t)] + E[R^2(Y(1),t)] \tag{103}$$

$$= \tau(1-\tau)t^2 + O(t^2) + O(t^3) = O(t^2) \tag{104}$$

Finally, note that

$$\begin{aligned}
E[\tilde{R}_N(t)] &= E\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{T_i R(Y_i,t)}{p(X_i)} - E[R(Y,t)\,|X_i, T=1]\frac{T_i - p(X_i)}{p(X_i)}\right)\right] \\
&= E\left[\frac{T\,R(Y,t)}{p(X)} - E[R(Y,t)\,|X,T=1]\frac{T-p(X)}{p(X)}\right] \\
&= E\left[\frac{E[T\,R(Y,t)\,|\,X,T=1]p(X)}{p(X)}\right. \\
&\quad \left. - E[R(Y,t)\,|X,T=1]\left(\frac{(1-p(X))p(X)}{p(X)} - \frac{(p(X)(1-p(X))}{p(X)}\right)\right] \\
&= E[E[R(Y(1),t)\,|\,X]] \\
&= E[R(Y(1),t)] \\
&= \frac{1}{2} f_1(q_{1,\tau})t^2 + o(t^2)
\end{aligned} \tag{105}$$

Hence Equation (69) holds by Equations (42), (82) and (105). Therefore $N\varepsilon_N(u/\sqrt{N})$ is a sum of three components:

$$N\varepsilon_N(u/\sqrt{N}) = N(\tilde{R}_N(u/\sqrt{N}) - E[\tilde{R}_N(u/\sqrt{N})]) + N\varepsilon_{1,N}(u/\sqrt{N}) + N\,o(u^2/N)$$

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

Start with the last term, $N\,o(u^2/N)$. This is goes to zero for each u by definition.

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

$$\begin{aligned}
E[\tilde{R}_N^2(t)] &= \frac{1}{N}E\left[\left(\frac{T\,R(Y,t)}{p(X)} - E[R(Y,t)\,|\,X,T=1]\frac{T-p(X_i)}{p(X)}\right)^2\right] \\
&= \frac{1}{N}E\left[\frac{T^2R^2(Y,t)}{p^2(X)} + E^2[R(Y,t)\,|\,X,T=1]\left(\frac{T-p(X)}{p(X)}\right)^2\right. \\
&\quad \left. - 2\left(\frac{T\,R(Y,t)(T-p(X))E[R(Y,t)\,|\,X,T=1]}{p^2(X)}\right)\right] \\
&= \frac{1}{N}E\left[\frac{E[R^2(Y(1),t)\,|\,X]}{p(X)} + E^2[R(Y(1),t)\,|\,X]\left(\frac{1-p(X)}{p(X)}\right)\right. \\
&\quad \left. - 2E^2[R(Y(1),t)\,|\,X]\left(\frac{1-p(X)}{p(X)}\right)\right] \\
&= \frac{1}{N}E\left[\frac{V[R(Y(1),t)\,|\,X]}{p(X)} + E^2[R(Y(1),t)\,|\,X]\right] \\
&= \frac{1}{N}E\left[\frac{E[R^2(Y(1),t)\,|\,X]}{p(X)} - \frac{1-p(X)}{p(X)}E^2[R(Y(1),t)\,|\,X]\right] \qquad (106)
\end{aligned}$$

Hence, the expected $\tilde{R}_N(t)$ squared is equal to the sum of two terms, $\frac{1}{N}E\left[\frac{E[R^2(Y(1),t)\,|\,X]}{p(X)}\right]$ and $\frac{1}{N}E\left[\frac{p(X)-1}{p(X)}E^2[R(Y(1),t)\,|\,X]\right]$. The first one is equal to:

$$\begin{aligned}
\frac{1}{N}E\left[\frac{E[R^2(Y(1),t)\,|\,X]}{p(X)}\right] &= \frac{1}{N}E\left[\frac{1}{p(X)}\left(\frac{1}{3}f_1(q_{1,\tau}+t^{**}(t,X)\,|\,X)t^3\right)\right] \\
&\leq \frac{1}{cN}O(t^3) \qquad (107) \\
&= O(t^3/N)
\end{aligned}$$

whereas the second term is bounded by:

$$\begin{aligned}
&\frac{1-c}{cN}\left|E[E^2[R(Y(1),t)\,|\,X]]-\frac{1}{4}E[f_1^2(q_{1,\tau}\,|\,X)]t^4\right| \\
&\quad\le \frac{C}{N}E\left[\left|\frac{1}{4}E[f_1^2(q_{1,\tau}+t^*(X,t)\,|\,X)]t^4-\frac{1}{4}E[f_1^2(q_{1,\tau}\,|\,X)]t^4\right|\right] \\
&\quad= \frac{C}{N}\frac{t^4}{4}E[|\,f_1^2(q_{1,\tau}+t^*(X,t)\,|\,X)-f_1^2(q_{1,\tau}\,|\,X)|] \\
&\quad\le \frac{Ct^4}{4N}E[|\,f_1(q_{1,\tau}+t^*(X,t)\,|\,X)-f_1(q_{1,\tau}\,|\,X)|\,|2\,f_1(q_{1,\tau}\,|\,X)+f_1(q_{1,\tau}+t^*(X,t)\,|\,X)-f_1(q_{1,\tau}\,|\,X)|] \\
&\quad\le \frac{Ct^4}{4N}(E[M^2(X)]|t|^{2\lambda}+2f_1(q_{1,\tau})E[M(X)]|t|^{\lambda} \\
&\quad= o(t^4/N) \qquad (108)
\end{aligned}$$

for some positive constants c and C. Thus, $\frac{1}{N}E[E^2[R(Y(1),t)\,|\,X]]=O(t^4/N)$, and finally:

$$E[\tilde{R}_N^2(t)]=O(t^3/N) \qquad (109)$$

Therefore for each u:

$$\begin{aligned}
Var\left(N(\tilde{R}_N(u/\sqrt{N})-E[\tilde{R}_N(u/\sqrt{N})])\right) &= N^2\,O(|u|^3/N^{5/2}) \\
&= O(|u|^3/\sqrt{N}) \\
&= O(|u|^3 o(1)) \\
&= o(|u|^3) \qquad (110)
\end{aligned}$$

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

The missing part to prove Lemma 4 is to prove that for each u, $N\varepsilon_{1,N}(u/\sqrt{N})$ goes to zero in probability.

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 $N\varepsilon_{1,N}(u/\sqrt{N})$.

The main difference is that their terms do not depend on u, as instead of $A(Y, u/\sqrt{N})$ they have $Y/\sqrt{N}$, where $E[Y^2]$ is assumed to be finite. However, it is possible to bound $N\varepsilon_{1,N}(u/\sqrt{N})$ using exactly the same arguments they used, being just aware that we will have an extra term which will reflect the dependence on u.

I will show how the analogy between $N\varepsilon_{1,N}(u/\sqrt{N})$ and the remainder term in Hirano, Imbens and Ridder can be drawn. Consider for instance N times the absolute value of Equation (71) evaluated at $u/\sqrt{N}$:

$$\left| \sum_{i=1}^{N} \frac{T_i A(Y_i, u/\sqrt{N})}{\hat{p}(X_i)} - \frac{T_i A(Y_i, u/\sqrt{N})}{p(X_i)} + \frac{T_i A(Y_i, u/\sqrt{N})}{p^2(X_i)} (\hat{p}(X_i) - p(X_i)) \right| \tag{111}$$

$$\leq \sum_{i=1}^{N} \left| \frac{T_i A(Y_i, u/\sqrt{N})}{p^2(X_i)\hat{p}(X_i)} (\hat{p}(X_i) - p(X_i))^2 \right| \tag{112}$$

$$= \sum_{i=1}^{N} \left| \frac{T_i A(Y_i, u/\sqrt{N})}{p^2(X_i)\hat{p}(X_i)} (\hat{p}(X_i) - p_K(X_i))^2 \right| \tag{113}$$

$$+ \sum_{i=1}^{N} \left| \frac{T_i A(Y_i, u/\sqrt{N})}{p^2(X_i)\hat{p}(X_i)} (p_K(X_i) - p(X_i))^2 \right| \tag{114}$$

$$+ 2 \sum_{i=1}^{N} \left| \frac{T_i A(Y_i, u/\sqrt{N})}{p^2(X_i)\hat{p}(X_i)} (\hat{p}(X_i) - p_K(X_i))(p_K(X_i) - p(X_i)) \right| \tag{115}$$

$$\tag{116}$$

Let me start working with Equation (113):

$$
\begin{aligned}
\sum_{i=1}^{N}\left|\frac{T_i A(Y_i,u/\sqrt{N})}{p^2(X_i)\hat{p}(X_i)}(\hat{p}(X_i)-p_K(X_i))^2\right| &= \sum_{i=1}^{N}\left|\frac{T_i A(Y_i,u/\sqrt{N})}{p^2(X_i)\hat{p}(X_i)}\left(L'(H_K(X_i)'\tilde{\pi})H_K(X_i)'(\hat{\pi}-\pi_K)\right)^2\right| \\
&\leq (\inf_{x\in\mathcal{X}} p(x))^{-2}(\inf_{x\in\mathcal{X}}\hat{p}(x))^{-1}\sum_{i=1}^{N}\left|T_i A(Y_i,u/\sqrt{N})\left(L'(H_K(X_i)'\tilde{\pi})H_K(X_i)'(\hat{\pi}-\pi_K)\right)^2\right| \\
&\leq (\inf_{x\in\mathcal{X}} p(x))^{-2}(\inf_{x\in\mathcal{X}}\hat{p}(x))^{-1}\frac{1}{16}\zeta^2(N)\,||\hat{\pi}-\pi_K||^2\sum_{i=1}^{N}|A(Y_i(1),u/\sqrt{N})| \\
&= O(1)\,O_p(1)\,O(K^2)\,O_p(K/N)\,O_p(|u|) \\
&= O_p\left(\frac{K^3|u|}{N}\right) \\
&= O_p\left(N^{3\alpha-1}|u|\right) \\
&= O_p\left(o(1)\,|u|\right) \\
&= o_p(|u|) \qquad (117)
\end{aligned}
$$

In the first line of the above expression I used the Mean Value Theorem.[25]. In the third line I used a property of the logistic function and Newey's result presented in Equation (28). In the fourth line I used the common support assumption, results (II) and (III) of Lemma 3 and the Markov inequality with the previous result on the order of $E[A^2(Y(1),t)]$. Finally, in the sixth line I used Assumption 3.

The same logic could have been applied to Equations (114) and (115) yielding respectively:

$$
\begin{aligned}
\sum_{i=1}^{N}\left|\frac{T_i A(Y_i,u/\sqrt{N})}{p^2(X_i)\hat{p}(X_i)}(p_K(X_i)-p(X_i))^2\right| &= O(1)\,O_p(1)\,O_p(|u|)\,O(K^{2-2\frac{s}{r}}) \\
&= O_p(|u|\,N^{(2-2\frac{s}{r})\alpha}) \\
&= o_p(|u|) \qquad (118)
\end{aligned}
$$

and

[25]Note that for $\tilde{\pi}\in[\hat{\pi},\pi_K]$, $L'(H_K(x)'\tilde{\pi})>0$ where $L'(z)=\frac{dL(z)}{dz}=L(z)(1-L(z))$, yielding then that $\sup_z L'(z)=1/4$. Also note that $L''(z)=L'(z)(1-2L(z))$.

$$
\begin{aligned}
2\sum_{i=1}^{N}\left|\frac{T_i A(Y_i,u/\sqrt{N})}{p^2(X_i)\hat{p}(X_i)}(\hat{p}(X_i)-p_K(X_i))(p_K(X_i)-p(X_i))\right| & \\
&= O(1)\,O_p(1)\,O_p(|u|)\,O(K^{1-\frac{s}{r}})\,O_p(K^{\frac{3}{2}}/\sqrt{N}) \\
&= O_p\left(|u|\,\frac{K^{(5/2-s/r)\alpha}}{\sqrt{N}}\right) \\
&= o_p(|u|) \qquad (119)
\end{aligned}
$$

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 $|u|$. However, for a fixed u, the rate of convergence remains the same one they computed, $o_p(1)$.

Computation of bounds for Equations (72)-(75) follows again the same lines as in Hirano, Imbens and Ridder (2002). Therefore, and for reasons of space, a detailed proof that shows that each one of those equations times N, evaluated at $u/\sqrt{N}$, is $o_p(1)$ for fixed u is omitted. Note only however, that in the process of finding bounds for all of those four equations, we will face expressions depending either on $\sum_{i=1}^{N}|T_i A(Y_i,u/\sqrt{N})|$ or on $E[A(Y(1),u/\sqrt{N})\,|\,X]$. For the former I have already computed a probabilistic bound. But the latter, by the Markov inequality, is a random variable such that:

$$
\begin{aligned}
\left|\sum_{i=1}^{N}E[A(Y(1),u/\sqrt{N})\,|\,X_i]\right| &= O_p\left(\sqrt{NE[E^2[A(Y(1),u/\sqrt{N})\,|\,X]]}\right) \\
&= O_p\left(\sqrt{N(\frac{u^2}{N}O(1)+\frac{u^4}{N^2}O\left(\frac{u^4}{N^2}\right)+\frac{u^3}{N^{3/2}}O(1)O(u^2/N)}\right) \\
&= O_p(|u|+u^4N^{-7/2}+|u|^{7/2}N^{-2}) \\
&= O_p(|u|) \qquad (120)
\end{aligned}
$$

Hence, as claimed earlier, the proof that $N\varepsilon_{1,N}(u/\sqrt{N}) = o_p(|u|)$, will follow the same steps as in the proof by Hirano, Imbens and Ridder (2002). This happens because when they

have $\sum_{i=1}^{N}|Y_i(1)/\sqrt{N}| = O_p(1)$, I have $\sum_{i=1}^{N}|A(Y_i(1),u/\sqrt{N})| = O_p(|u|)$; and when they have $\left|\sum_{i=1}^{N}E[Y(1)/\sqrt{N}\,|\,X_i]\right| = O_p(1)$, I have $\left|\sum_{i=1}^{N}E[A(Y(1),u/\sqrt{N})\,|\,X_i]\right| = O_p(|u|)$. Thus, the only difference from their approach to mine is the $|u|$ term.

Finally, we conclude that for each fixed u, $N\varepsilon_N(u/\sqrt{N})$ goes to zero in probability.□

Proof of Lemma 5:

From Equation (50), 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 the Markov inequality:

$$Pr\left[\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_{1,i}\right| > M\right] < \frac{E[\psi_{1,i}^2]}{M^2} \tag{121}$$

Choosing M to satisfy $\frac{E[\psi_{1,i}^2]}{M^2} < \delta$, where δ is a small enough positive constant, there will exist a sample size N_δ such that for all $N > N_\delta$, Equation (121) will be satisfied.

Result (*ii*) follows by a Central Limit Theorem; while (*iii*) follows by noting that ψ_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 $q_{1,\tau}$.[26] □

Proof of Lemma 6:

First notice that $G_N(q,\hat{p}) = \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\hat{p}(X_i)}\rho_\tau(Y_i-q)$ is convex in q with probability approaching one, as it is a sum of zeros and convex functions in q. As a result, the transformed objective function, $Q_N(t,\hat{p})$ will be convex in t and the following random function must be convex in u:

$$\begin{aligned} B_N(u,\hat{p}) &= N\,Q_N(u/\sqrt{N},\hat{p}) - \sum_{i=1}^{N}\frac{u}{\sqrt{N}}\left(\frac{T_iD_i}{p(X_i)} - E[D\,|X_i,T=1]\frac{T_i-p(X_i)}{p(X_i)}\right) \\ &= \frac{1}{2}f_1(q_{1,\tau})u^2 + N\varepsilon_N(u/\sqrt{N}) \end{aligned} \tag{122}$$

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

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

[26] See the proof of the semiparametric efficiency bound in this appendix.

$$(1-\frac{\delta}{a})B_N(\tilde{u},\,\hat{p})+\frac{\delta}{a}B_N(u,\,\hat{p})\geq B_N(\tilde{u}+\delta,\,\hat{p}) \tag{123}$$

By Equation (122), this can be rewritten as:

$$\begin{aligned}\frac{\delta}{a}(B_N(u,\,\hat{p})-B_N(\tilde{u},\,\hat{p}))\geq & B(\tilde{u}+\delta)+N\varepsilon_N(\tilde{u}/\sqrt{N}+\delta)-\left(B(\tilde{u})+N\varepsilon_N(\tilde{u}/\sqrt{N})\right)\\ \geq & -2\sup_{|u-\tilde{u}|\leq\delta}|N\varepsilon_N(u/\sqrt{N})|\;+\inf_{|u-\tilde{u}|=\delta}|B(u)-B(\tilde{u})|\end{aligned} \tag{124}$$

Now, note that

$$\inf_{|u-\tilde{u}|=\delta}|B(u)-B(\tilde{u})|=\frac{1}{2}f_1(q_{1,\tau})\delta^2 \tag{125}$$

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

$$-2\sup_{|u-\tilde{u}|\leq\delta}|N\varepsilon_N(u/\sqrt{N})|\;+\frac{1}{2}f_1(q_{1,\tau})\delta^2>0 \tag{126}$$

then $\hat{u}$, the minimizer of $NQ_N(u/\sqrt{N},\hat{p})$, will be inside the δ-interval around $\tilde{u}$. Hence, I need to show that with probability approaching one, Equation (126) holds.

By the Hjort and Pollard's (1993) version of the Convexity Lemma, $sup_{u\in\mathcal{K}}|N\varepsilon_N(u/\sqrt{N})|=o_p(1)$ for each compact subset $\mathcal{K}$ of $\mathbb{R}$. Define:

$$\mathcal{K}_\delta=\{u\in\mathbb{R};|u-\tilde{u}|\leq\delta\} \tag{127}$$

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

$$\sup_{u\in\mathcal{K}_\delta}|N\varepsilon_N(u/\sqrt{N})|=o_p(1) \tag{128}$$

Thus, for each $\delta > 0$:

$$Pr[\sup_{u \in \mathcal{K}_\delta} |N\varepsilon_N(u/\sqrt{N})| \geq \frac{1}{4} f_1(q_{1,\tau})\delta^2] = o(1) \tag{129}$$

Hence with probability approaching one, for each $\delta > 0$, Equation (126) holds, which means that $\hat{u}$, the minimizer of $NQ_N(u/\sqrt{N}, \hat{p})$, will be inside the δ-interval around $\tilde{u}$ with probability approaching one:

$$|\hat{u} - \tilde{u}| = o_p(1) \tag{130}$$

□

APPENDIX II

For the set A, it is necessary to estimate in the first step the conditional expectation $m_1^A(x\,|\,q) = E[\mathbb{1}\{Y \leq q\} - \tau\,|\,X = x, T = 1]$ by $\hat{m}_1^A(x\,|\,q) = \hat{E}[\mathbb{1}\{Y \leq q\} - \tau\,|\,X = x, T = 1]$. This estimation problem can be written as:

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

where $\hat{\nu}_i^A$ 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:

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

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:

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

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

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

And:

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

Thus, finally:

$$\hat{q}_{1,\tau}^A=\arg\min_q\sum_{i=1}^{N}\hat{\omega}_{1,i}^A\rho_\tau(Y_i-q) \tag{131}$$

In the identification condition given by *B*, the weight associated with $\hat{q}_{1,\tau}^B$, $\hat{\omega}_{1,i}^B$, is equal to $\frac{1}{N}\sum_{j=1}^{N}\hat{\nu}_i^B(X_j)$. For the example where the conditional expectation is estimated by a kernel K with bandwidth h:

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

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.

TABLE 1: LaLonde/Dehejia and Wahba Data Set - Summary Statistics

Summary Statistics

Treatment Group
(Sample size = 185)

	Earnings (1978)	Age	Education	Dropout	Black	Hispanic	Married	Earnings (1974)	Earnings (1975)	Unemployed (1974)	Unemployed (1975)
Mean	6349.1	25.8	10.3	71%	84%	6%	19%	2095.6	1532.1	71%	60%
	(7867.4)	(7.2)	(2.0)	-	-	-	-	(4886.6)	(3219.3)	-	-
Min	0	17	4	-	-	-	-	0	0	-	-
Max	60307.9	48	16	-	-	-	-	35040.1	25142.2	-	-

Experimental Control Group
(Sample size = 260)

	Earnings (1978)	Age	Education	Dropout	Black	Hispanic	Married	Earnings (1974)	Earnings (1975)	Unemployed (1974)	Unemployed (1975)
Mean	4554.8	25.1	10.1	83%	83%	11%	15%	2107.0	1266.9	75%	68%
	(5483.8)	(7.1)	(1.6)	-	-	-	-	(5687.9)	(3103.0)	-	-
Min	0	17	3	-	-	-	-	0	0	-	-
Max	39483.5	55	14	-	-	-	-	39570.7	23032	-	-

Non-Experimental Control Group
(Sample size = 2490)

	Earnings (1978)	Age	Education	Dropout	Black	Hispanic	Married	Earnings (1974)	Earnings (1975)	Unemployed (1974)	Unemployed (1975)
Mean	21553.9	34.9	12.1	31%	25%	3%	87%	19428.8	19063.3	9%	10%
	(15555.3)	(10.4)	(3.1)	-	-	-	-	(13406.9)	(13596.9)	-	-
Min	0	18	0	-	-	-	-	0	0	-	-
Max	121174	55	17	-	-	-	-	137149	156653	-	-

TABLE 2: LaLonde/Dehejia and Wahba Data Set - Quantiles of Potential Earnings

QTE and Quantiles of Potentials Earnings (1978)

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\Delta_{\tau \mid T=1}$	0	0	711	21	1927	3879	4517	6027	5503
	(*126*)	(*538*)	(*1052*)	(*1357*)	(*1132*)	(*1275*)	(*1461*)	(*1853*)	(*3398*)
$\Delta_{\tau,exp}$	0	0	930	1163	1081	1446	1797	2246	2919
$\hat{q}^{C}_{1,\tau \mid T=1}$	0	0	930	2326	4232	6184	8174	10756	14582
$\hat{q}^{C}_{0,\tau \mid T=1}$	0	0	219	2305	2305	2305	3657	4729	9079
Quantiles of Non-Experimental Control Group ($\hat{q}_0$)	0	8866	13299	17733	20688	24315	27347	31623	38421
Quantiles of Experimental Control Group ($\hat{q}_{0,exp}$)	0	0	0	1163	3151	4738	6377	8510	11663
$\hat{q}^{C}_{0,\tau \mid T=1} - \hat{q}_{0,exp}$	0	0	219	1142	-846	-2433	-2720	-3781	2584
	(*126*)	(*515*)	(*893*)	(*1312*)	(*1289*)	(*1173*)	(*1263*)	(*1835*)	(*3207*)

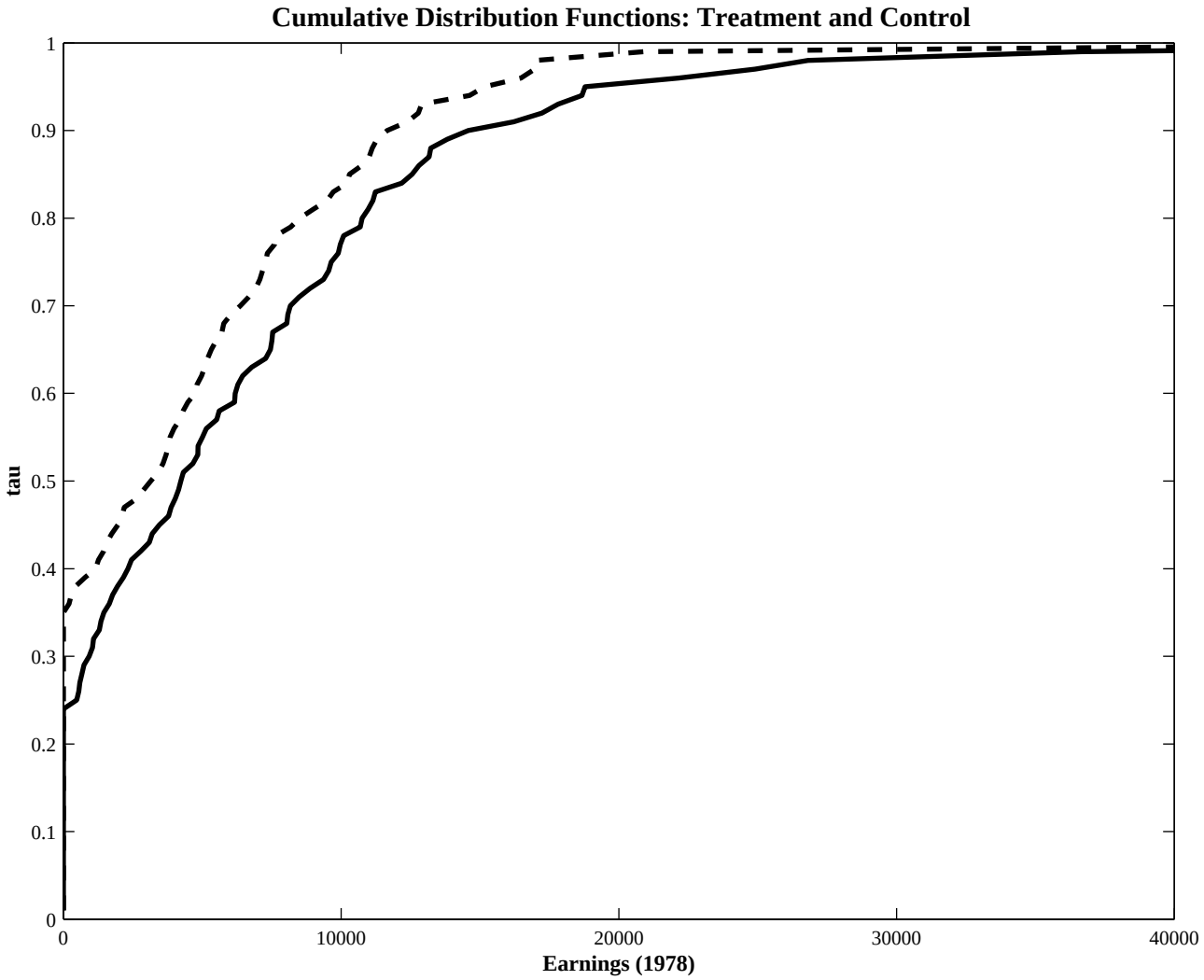

FIGURE 1: LaLonde/Dehejia and Wahba Experimental Data Set (Treatment: solid line; Control: dashed line)

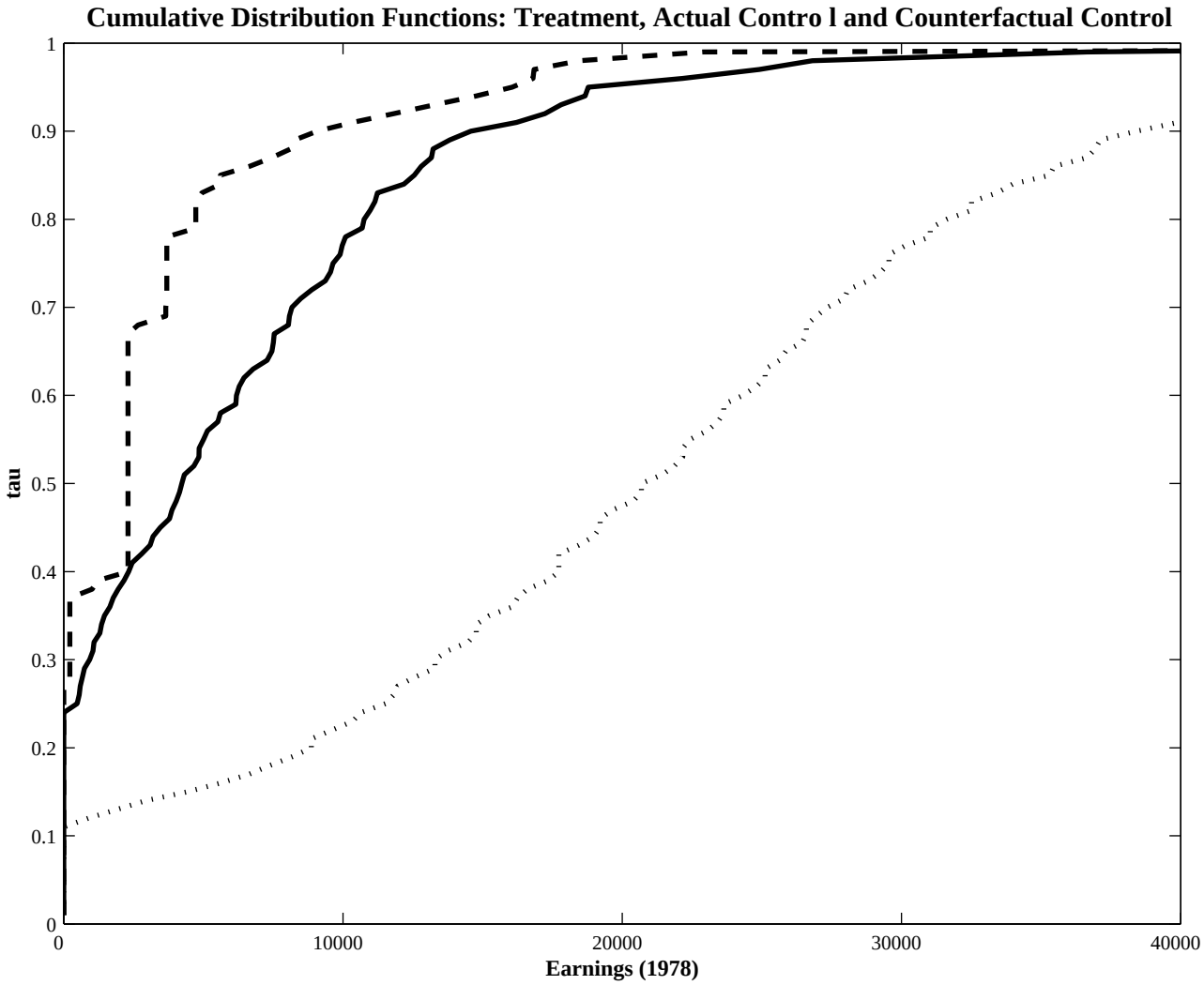

FIGURE 2: LaLonde/Dehejia and Wahba Non-Experimental Data Set (Treatment: solid line; Counterfactual Control: dashed line; Actual Control: dotted line)

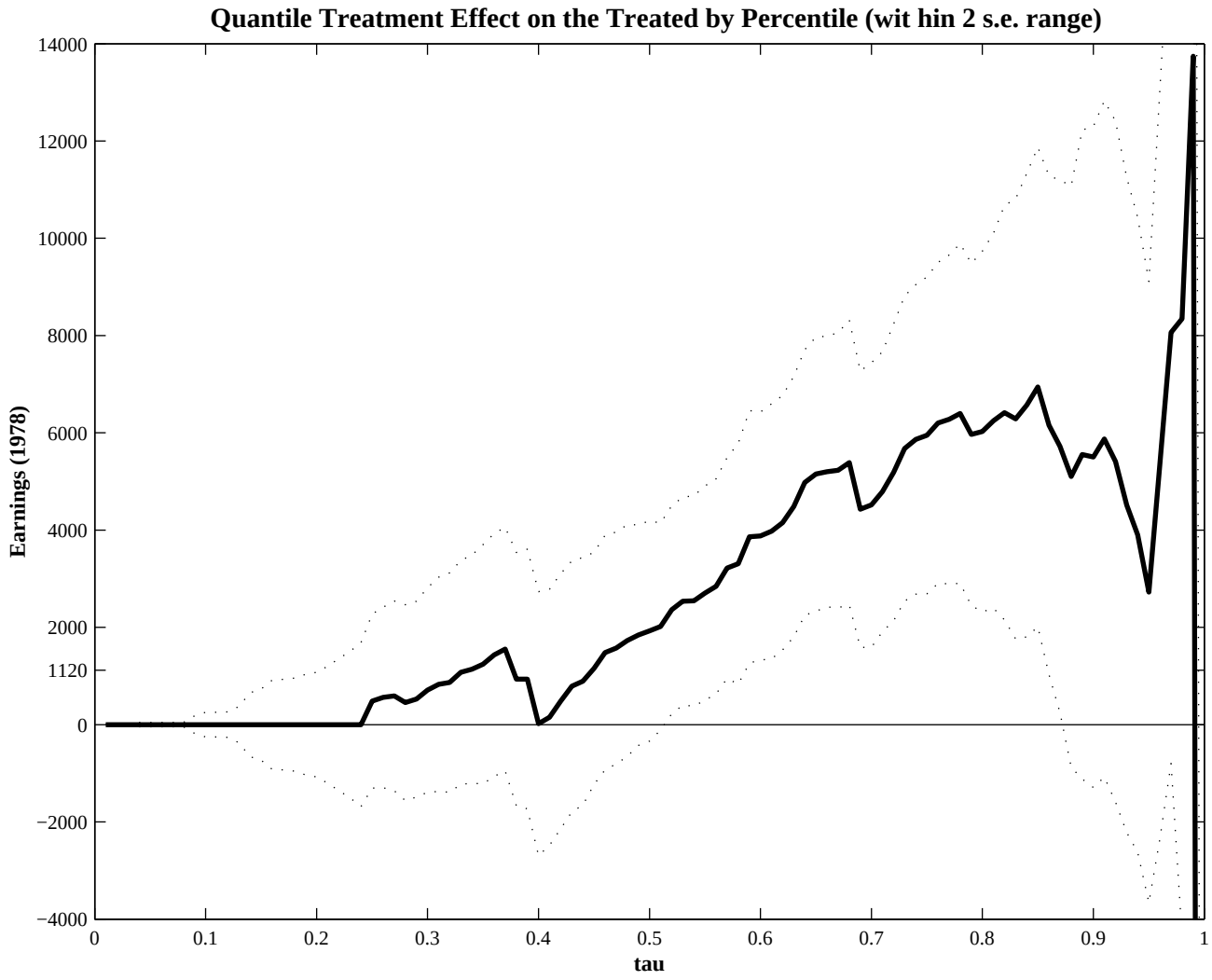

FIGURE 3: LaLonde/Dehejia and Wahba Non-Experimental Data Set

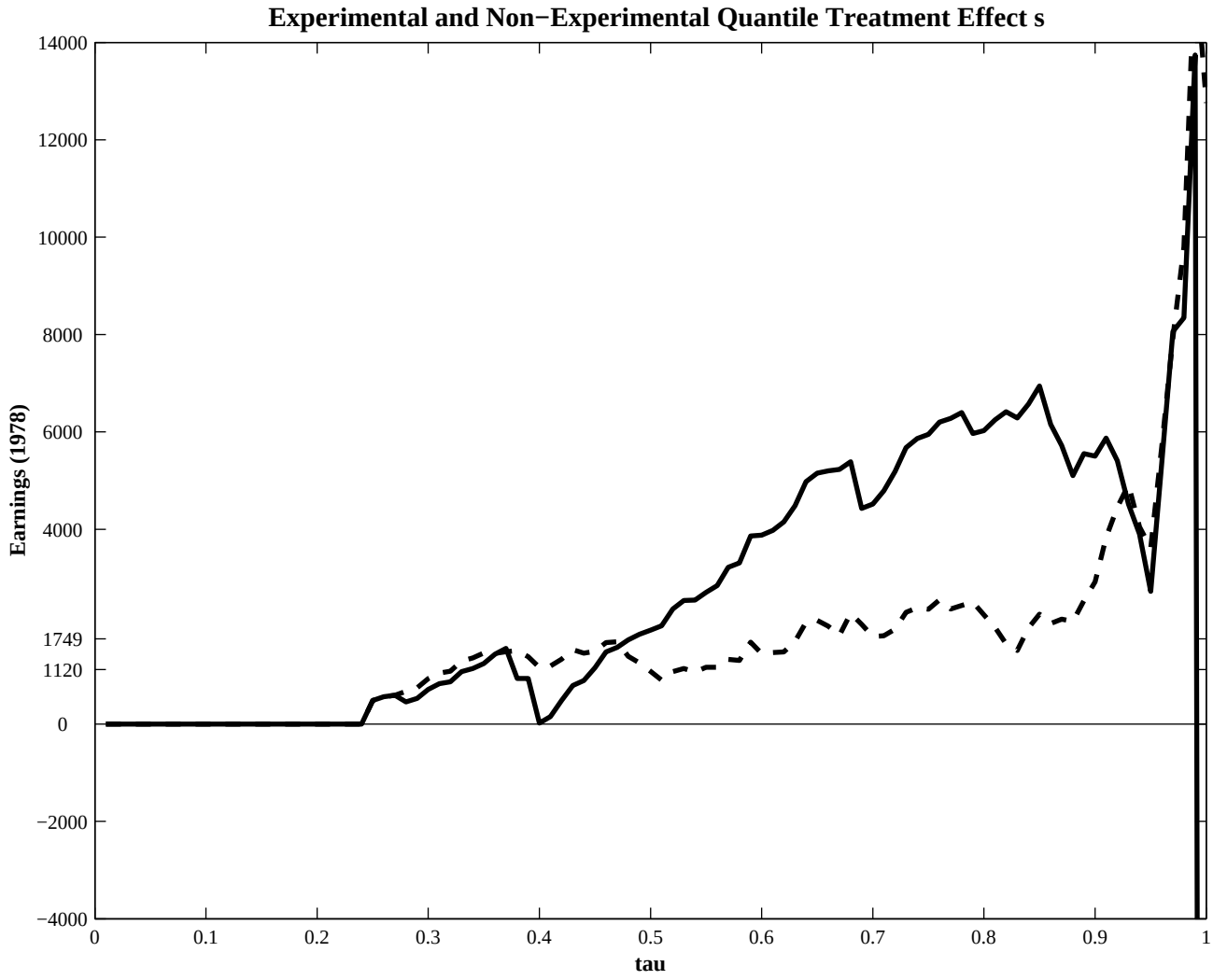

FIGURE 4: LaLonde/Dehejia and Wahba Data Set (Non-experimental QTE: solid line; Experimental QTE: dashed line)

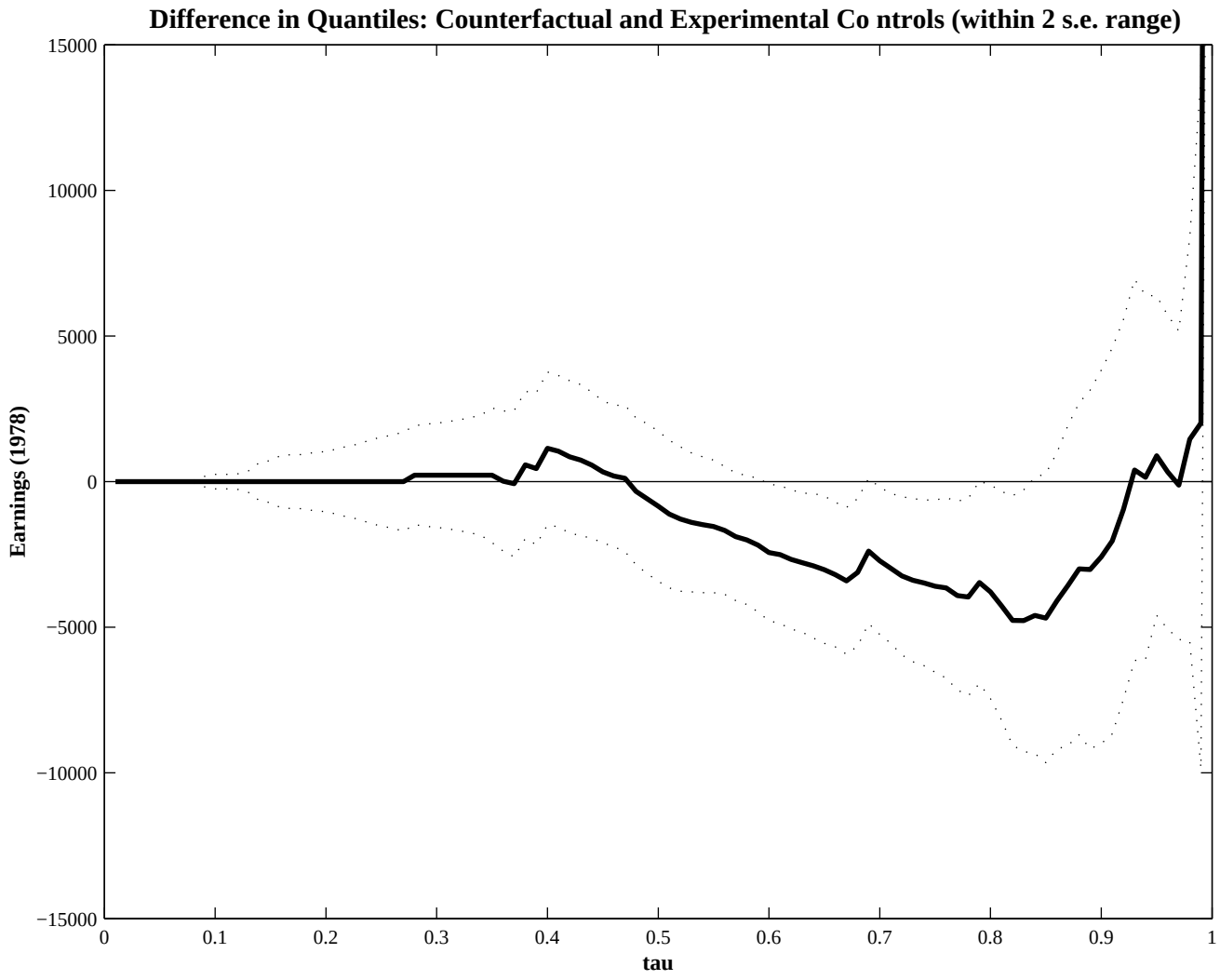

FIGURE 5: LaLonde/Dehejia and Wahba Data Set Non-experimental and Experimental Controls