Frölich, Markus; Melly, Blaise

Working Paper

Unconditional quantile treatment effects under endogeneity

IZA Discussion Papers, No. 3288

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Frölich, Markus; Melly, Blaise (2008) : Unconditional quantile treatment effects under endogeneity, IZA Discussion Papers, No. 3288, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/35065

DISCUSSION PAPER SERIES

IZA DP No. 3288

Unconditional Quantile Treatment Effects under Endogeneity

Markus Frölich
Blaise Melly

January 2008

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Unconditional Quantile Treatment Effects under Endogeneity

Markus Frölich
University of Mannheim, University of St. Gallen, IFAU and IZA

Blaise Melly
MIT and University of St. Gallen

Discussion Paper No. 3288
January 2008

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3288
January 2008

ABSTRACT

Unconditional Quantile Treatment Effects under Endogeneity*

This paper develops IV estimators for *unconditional* quantile treatment effects (QTE) when the treatment selection is *endogenous*. In contrast to conditional QTE, i.e. the effects conditional on a large number of covariates X, the unconditional QTE summarize the effects of a treatment for the entire population. They are usually of most interest in policy evaluations because the results can easily be conveyed and summarized. Last but not least, unconditional QTE can be estimated at $\sqrt{n}$ rate without any parametric assumption, which is obviously impossible for conditional QTE (unless all X are discrete). In this paper we extend the identification of unconditional QTE to endogenous treatments. Identification is based on a monotonicity assumption in the treatment choice equation and is achieved without any functional form restriction. Several types of estimators are proposed: regression, propensity score and weighting estimators. Root n consistency, asymptotic normality and attainment of the semiparametric efficiency bound are shown for our weighting estimator, which is extremely simple to implement. We also show that including covariates in the estimation is not only necessary for consistency when the instrumental variable is itself confounded but also for efficiency when the instrument is valid unconditionally. Monte Carlo simulations and two empirical applications illustrate the use of the proposed estimators.

JEL Classification: C13, C14, C21

Keywords: quantile treatment effects, nonparametric regression, instrumental variables

Corresponding author:

Markus Frölich
Universität Mannheim
Abteilung Volkswirtschaftslehre
L7, 3-5
68131 Mannheim
Germany
E-mail: froelich@uni-mannheim.de

* We have benefited from comments by Alberto Abadie, Joshua Angrist, Guido Imbens, Michael Lechner and seminar participants at the University of St. Gallen (March 15, 2007), the IZA Workshop "Heterogeneity in Micro Econometric Models" (June 8, 2007), Harvard (November 5, 2007), Uppsala (November 14, 2007), MIT (December 6, 2007). We thank David Card for providing us with the data used in section 5.

1 Introduction

In many research areas it is of first order importance to assess the distributional effects of policy variables. For instance, policy makers will evaluate differently two training programs having the same average effect but whose effects are concentrated in the lower end of the distribution for the first one and on the upper end for the second one. The ability of quantile treatment effects (QTE) to characterize the heterogeneous impacts of variables on different points of an outcome distribution makes them appealing in many economic applications. This has motivated the recent surge of interest in their identification and estimation using different sets of assumptions, particularly in the applied policy evaluation literature.

In this paper, we develop an instrumental variable (IV) model of unconditional QTE in the presence of endogeneity and obtain conditions for identification of the QTE without functional form assumptions. We are interested in the *unconditional* QTE, i.e. the difference between the unconditional quantiles of the treated outcome and the unconditional quantiles of the non-treated outcome for the population of interest. In contrast to conditional QTE, i.e. the effects conditional on a large number of covariates X, the unconditional QTE summarize the effects of a treatment for the entire population and are usually of most interest in policy evaluations. The results can easily be conveyed and summarized since the unconditional quantile function is a one dimensional function, whereas the conditional quantile functions are multidimensional functions (of the quantile on one side and of each of the covariates on the other side). In addition, and this is at least as important, unconditional QTE can be estimated at $\sqrt{n}$ rate without any parametric assumption, which is obviously impossible for conditional QTE (unless all X are discrete). Hence, we can estimate unconditional QTE more precisely than conditional QTE.

We allow for an endogenous binary treatment and we show how to use an IV to identify QTE. Our approach to IV is based on the framework developed by Imbens and Angrist (1994), but also permits non-binary instrumental variables. We identify the effects for the subpopulation that reacts on changes of the value of the instrument (compliers) based on a monotonicity assumption in the treatment choice equation.[1]

We assume that the instrument is independent of the outcome variable only conditionally on X. In many applications, the instrument is not randomly assigned and may itself be confounded.

[1] Note that in applications where two-sided noncompliance is impossible and the instrument has been randomized we identify the average treatment effect on the treated (ATET).

For example, college proximity may be used as an instrument to identify the returns to schooling, noting that living close to a college during childhood may induce some children to go to college but is unlikely to affect their wages many years later (Card, 1995). Nevertheless, since parents' residence is not randomly allocated, it is likely to be correlated with parent's profession, family income and wealth, which may affect the wage prospects of their children. In this case, distance to college may be a proper IV only after conditioning on some covariates. In addition, including covariates can be helpful to intercept all mediating causal paths between the instrument and the outcome variable. A crucial assumption is that the instrument has no direct impact on the outcome Y, other than via the treatment D. However, if there is another causal link between the instrument Z and the outcome Y and if this link runs via a mediating variable on the causal pathway, by conditioning on this variable we can still obtain identification. Naturally, since conditional independence over the whole support of X implies unconditional independence, our results also cover the case where the instrument is valid unconditionally, e.g. a randomized assignment to a training program or the Vietnam conscription lottery. We will show later that even in this configuration using covariates is useful since it increases the precision of the estimates.

In this model, we show that unconditional QTE for the compliers are identified. Several constructive identification results lead naturally to several types of fully nonparametric estimators: regression (or matching) on the covariates, regression on the propensity score, and weighting estimators. Additionally, a projection of the weights onto an appropriate space justifies using a weighted version of the traditional quantile regression algorithm proposed by Koenker and Bassett (1978).

We give then conditions under which the proposed weighting estimator is $\sqrt{n}$ consistent, asymptotically normally distributed and efficient. In order to show this last property we first derive the semiparametric efficiency bound for unconditional QTE in our model. Two further results complete the picture: Knowledge of the probability with which the instrument has been assigned does not change the value of the efficiency bound, but increasing the number of covariates does reduce the efficiency bound when the additional covariates are not required for consistency. Thus, incorporating the information contained in the covariates may be needed for consistency when the instrument is itself confounded or for efficiency when the value of the instrumental variable is assigned completely at random. These results can be combined by including some covariates to obtain consistency and additionally others for efficiency reasons. It is also worthwhile

to mention that these covariates are permitted to be endogenous.

Finally, Monte Carlo simulations and two empirical applications illustrate the use of the estimators. In the first application, we estimate unconditional QTE with the dataset used by Abadie, Angrist, and Imbens (2002, AAI in the following). In this case, the instrument has been completely randomized. As expected, incorporating covariates does not significantly change the results but reduces (moderately) their variance. As a pedagogical exercise we then manipulate the data such that the instrument becomes independent of the outcome only conditionally on a covariate. This allows us to measure the bias arising when the instrument is not valid independently of the covariates. In the second application, we estimate the return to college in the USA, an important issue in the debate on inequality, using college proximity as instrument with data from Card (1995).[2] In this case, controlling for additional covariates is critical for the validity of the instruments.

The estimators proposed in this paper are highly relevant for applied researchers. Our model corresponds to a situation often encountered in practice as illustrated by both applications. In addition to deriving the theoretical properties of these estimators, we also provide user-friendly computer programs that implement the estimators and provide analytical standard errors. Computer codes in the programming language R (free software available at www.r-project.org) and in Stata are available from the authors and should simplify the use of the results derived in this paper.

Of course, we are not the first to consider the estimation of QTE. This topic has been an active area of research during the last three decades. Koenker and Bassett (1978) proposed and derived the statistical properties of a parametric (linear) estimator for conditional quantile models. Due to its ability to capture heterogeneous effects, its theoretical properties have been studied extensively and it has been used in many empirical studies; see, for example, Powell (1986), Guntenbrunner and Jurečková (1992), Buchinsky (1994), Koenker and Xiao (2002), Angrist, Chernozhukov, and Fernández-Val (2006). Chaudhuri (1991) analyzed nonparametric estimation of conditional QTE.

All these estimators assume that the treatment selection is exogenous.[3] However, in observational studies, the variables of interest (e.g., education, prices) are often endogenous, making conventional quantile regression inconsistent and hence inappropriate for recovering the causal

[2] The returns to education have received a lot of attention with recent research interests aiming also particularly at higher education (see e.g. Black and Smith (2004 and 2005) about the returns to college).

[3] Also called "selection on observables", "conditional independence" or "unconfoundedness".

effects of these variables on the quantiles of economic outcomes. Therefore, Abadie, Angrist, and Imbens (2002) and Chernozhukov and Hansen (2005, 2006, 2007) have proposed linear instrumental variable quantile regression estimators. Chernozhukov, Imbens, and Newey (2007) and Horowitz and Lee (2007) have considered nonparametric IV estimation of conditional quantile functions. In a serie of papers, Chesher (2003, 2005, 2007) also examines nonparametric identification of conditional effects.[4] Hoderlein and Mammen (2007) consider marginal effects in nonseparable models. In our paper, on the other hand, we develop a fully nonparametric approach to identification and estimation of *unconditional* QTE.

The literature discussed in the preceding two paragraphs deals with the estimation of *conditional* QTE. When we estimate conditional QTE, either we must make a strong parametric assumption, or the estimates are not $\sqrt{n}$ consistent. The change of estimand, from the conditional to the unconditional effects, enables to take the best of both worlds: absence of functional form assumptions and $\sqrt{n}$ consistency.[5] Recently, Firpo (2007), Frölich (2007b) and Melly (2006) have examined the nonparametric estimation of unconditional QTE, under a selection on observables assumption. We contribute to this literature by allowing for endogenous treatment choice. As a matter of fact, our estimators simplify to those estimators when the instrument and the treatment are identical (exogenous treatment). On the other hand, our weighting estimator simplifies to the AAI estimator when there is no covariate such that the conditional QTE are also the unconditional ones. We show, however, that their approach is not generally applicable for the estimation of unconditional QTE. Thus, the model and estimators looked at in this paper both substantively complement and differ from the existing literature.

The paper is organized as follows. Section 2 presents the model. Section 3 gives the identification results and suggests natural estimators. The asymptotic properties of the model are derived in Section 4. In particular we derive the semiparametric efficiency bound and show that our weighting estimator is consistent, asymptotically normally distributed and efficient. Section 5 presents the results of the simulations and of two applications and Section 6 concludes.

[4] Imbens and Newey (2003) consider the case of a *continuous* treatment with identification based on a control function approach.

[5] At the same time, these are of most policy interest. Of course, conditional QTE are also interesting since they allow to analyze the heterogeneity (with respect to X) of the treatment effect. But for policy guidance, it would nevertheless be most helpful to examine QTE conditional on only a very small subset of discrete covariates, e.g. young men, older women etc.

2 Notation and Framework

We consider the effect of a *binary* treatment variable D on a continuous outcome variable Y. Let Y_i^1 and Y_i^0 be the potential outcomes of individual i. Hence, Y_i^1 would be realized if individual i were to receive treatment 1 and Y_i^0 would be realized otherwise. Most interest has focused on the estimation of average treatment effects

$$E[Y^1 - Y^0]$$

or average treatment effects on the treated

$$E[Y^1 - Y^0|D = 1].$$

Instead of considering only average effects, it is often of considerable interest to compare the distributional effects of the treatment as well. A standard example may be the impact of a program on income inequality. Another example which has received considerable public interest is *educational equality*, where many societies would prefer to provide every child with a fair chance into adult live. Here, Y is a measure of cognitive ability (e.g. obtained from Math and language tests) and D may be the introduction of computers in classroom (teaching). In this paper, we will identify and estimate the entire distribution functions of Y^1 and Y^0. Since *quantile treatment effects* (QTE) are an intuitive way to summarize the distributional impact of a treatment, we especially focus our attention on them:

$$\Delta^\tau = Q^\tau_{Y^1} - Q^\tau_{Y^0},$$

where $Q^\tau_{Y^1}$ is the τ quantile of Y^1. In the earnings example, $\Delta^{0.9}$ would be the impact of D on the high income part of the distribution. In fact, our results are not limited to the estimation of QTE. Since we identify the entire processes $Q^\tau_{Y^1}$ and $Q^\tau_{Y^0}$ for $\tau \in (0, 1)$, it would be straightforward to derive estimates and inference e.g. for the treatment effect on *inequality measures* such as the interquantile spread. A typical inequality measure is the inter-decile ratio that can be defined as

$$\frac{Q^{0.9}_{Y^1}}{Q^{0.1}_{Y^1}} - \frac{Q^{0.9}_{Y^0}}{Q^{0.1}_{Y^0}}$$

or as

$$\frac{Q^{0.9}_{Y^1}}{Q^{0.1}_{Y^1}} \frac{Q^{0.1}_{Y^0}}{Q^{0.9}_{Y^0}}.$$

Our main focus is on *unconditional* treatment effects, i.e. the effects of D in the population at large. We might also be interested in the effects in subpopulations defined by some, usually broadly defined, set A, e.g. women below the age of 25, which we define as

$$\Delta_A^\tau = Q_{Y^1|A}^\tau - Q_{Y^0|A}^\tau$$

where $Q_{Y^1|A}^\tau$ is the quantile in the subpopulation A. Notice that this focus differs from the existing literature on IV quantile regression, which focuses on *conditional* treatment effects, i.e. conditional on a set of variables X. We call our effects unconditional in the sense that A usually contains a very broadly defined set, while X usually consists of a large set of covariates often including continuous variables as well. More precisely, we will consider Δ_A^τ as an unconditional effect if the set A has positive probability mass, e.g. the subpopulation of women. In other words, A is not permitted to contain any continuous regressors.[6] Conditional and unconditional effects are interesting in their own rights. Whereas conditional effects may be more interesting in economic analysis of effects heterogeneity, for public policy unconditional effects will usually be more relevant. The reason for this is not only that policy and the public need more aggregated results for decision making, but also that unconditional effects can be estimated without parametric assumptions more precisely than conditional effects. We can achieve $\sqrt{n}$-consistency for *unconditional* QTE, whereas nonparametric estimation of *conditional* QTE will always be estimated at a lower rate (unless all X are discrete).

There are only relatively few contributions that examine explicitly unconditional distributional impacts of treatment. Firpo (2007), Frölich (2007b) and Melly (2006) consider estimation of treatment effects, when D is exogenous conditional on X. The usual concern with estimating treatment effects is endogeneity of D and we will rely on exclusion restrictions for the instrumental variables Z. We consider a setup related to the recent literature on nonparametric identification of nonseparable models:

$$\begin{aligned} Y_i &= \varphi(D_i, X_i, U_i) \\ D_i &= \zeta(Z_i, X_i, V_i), \end{aligned} \tag{1}$$

where U and V are possibly related unobservables and X are additional covariates, which are permitted to be correlated with U and/or V. We assume that, after having included X in the

[6] In our analysis, we will also need in a first step to condition on a large set of regressors X to make the instrumental variables conditions hold, but then average over the support of X to obtain unconditional effects.

model, Z is excluded from the function φ. The corresponding potential outcomes are

$$Y_i^d = \varphi(d, X_i, U_i)$$
$$D_i^z = \zeta(z, X_i, V_i).$$

In contrast to Chernozhukov and Hansen (2005), Chernozhukov, Imbens, and Newey (2007) and Chesher (2007), we impose triangularity, i.e. assume that Y does not enter in ζ, but do not need to assume any kind of monotonicity or rank invariance for φ.[7] We do impose, on the other hand, that the function ζ is (weakly) monotonous in its first argument, i.e. assume that an exogenous increase in Z_i can never decrease the value of D_i. This is the monotonicity assumption of Imbens and Angrist (1994). This assumption may be more plausible than monotonicity in φ in some applications, whereas in other applications it may be less appealing.[8]

Our method is well suited for binary D. Imbens and Newey (2003) and Chesher (2003) analyzed identification for continuous D and Chesher (2005) examined interval identification with discrete D. Heckman and Vytlacil (2005) analyzed (marginal) average treatment effects for continuous Z and focused on treatment effects conditional on X, whereas we aim for unconditional effects. In future work we will extend our results to discrete D.

We will focus our attention on the subgroup of *compliers*, which we define as all individuals who are responsive to a change in Z within the support of Z. Note that in applications where the instrument has been randomized and two-sided noncompliance is impossible, the compliers are the treated. In this case, which corresponds to the application of section 5.1 using the data of AAI, we actually obtain the QTE on the treated. Generally, we cannot identify the effect of D on Y for individuals for whom D_i^z does not vary with z in the support of Z.[9] If the instruments Z are sufficiently powerful to move everyone from $D_i = 0$ to $D_i = 1$, this will lead to the average treatment effect (ATE) in the entire population. In most applications, however, the instruments available are not so powerful and it is interesting in this case to consider effects in the *largest subpopulation* for which the effect is identified. In addition, if Y is bounded, we can derive bounds on the overall treatment effects because the size of the subpopulation of compliers is identified as

[7] Chernozhukov and Hansen (2005), Chernozhukov, Imbens, and Newey (2007) and Chesher (2007) assume that φ is monotonous in its third argument.

[8] In future work we are going to examine the estimation of unconditional QTE using the monotonicity assumption in the outcome equation and the combination of both assumptions.

[9] These are the always-participants or never-participants in the language of Imbens and Angrist (1994).

well. Therefore, we focus on the QTE for the compliers:

$$\Delta_c^\tau = Q_{Y^1|c}^\tau - Q_{Y^0|c}^\tau$$

where $Q_{Y^1|c}^\tau = \inf_q \Pr\left(Y^1 \leq q | \mathcal{T} = c\right) \geq \tau$, where $T_i = c$ means that individual i is a complier, as defined below.

If Z consists of a single *binary* variable and if it has a (weakly) monotonous impact on D, the largest subpopulation affected by moving the instrument will consist of the individuals for whom $D_i^1 > D_i^0$. More generally, the largest subpopulation affected would be obtained by moving Z from the smallest point of its support to its largest point. If there is only a single instrument Z with support $\mathcal{Z} = [z_{\min}, z_{\max}]$, this corresponds to hypothetically moving Z_i from $z_{\min}$ to $z_{\max}$ for every individual. If Z contains several instrumental variables, the largest subpopulation affected would be obtained by moving the instruments from z_1^* to z_2^* where

$$(z_1^*, z_2^*) = \arg\max_{z_1, z_2 \in \mathcal{Z}} \left| \int \left(E\left[D|X, Z = z_2\right] - E\left[D|X, Z = z_1\right] \right) dF_X \right|,$$

where the integral expression measures the size of this subpopulation, as further discussed below. With monotonicity z_1^* and z_2^* will be at the boundary of the support of $\mathcal{Z}$.[10]

In the following we will assume throughout that z_1^* and z_2^* are known (and not estimated) and that $\Pr(Z = z_1^*) > 0$ and $\Pr(Z = z_2^*) > 0$. This rules out continuous instruments, unless they have masspoints at z_1^* and z_2^*. Note that our identification results would also hold for continuous instruments, but $\sqrt{n}$ consistent estimation would not be possible anymore. We will develop estimators for those situations in future work.

To simplify the notation we will use the values 0 and 1 subsequently instead of $z_{\min}$ to $z_{\max}$ or z_1^* to z_2^*, respectively. Furthermore, we will in the following only refer to the effectively used sample $\{i : Z_i \in \{0, 1\}\}$ or in other words assume that $\Pr(Z = z_1^*) + \Pr(Z = z_2^*) = 1$. This is appropriate for our applications where the single instruments Z are binary. In other applications, where $\Pr(Z = z_1^*) + \Pr(Z = z_2^*) < 1$, our results apply with reference to the subsample $\{i : Z_i \in \{z_1^*, z_2^*\}\}$.[11]

[10] This may not be the case, if the impact of Z is monotonous only given X, such that the relationship determining D may be decreasing in z for some x and increasing for other x. Then, in principle, an even larger affected subpopulation could be defined by examining different values of z_1^* and z_2^* for every value of x.

[11] Consider $\Pr(Z = z_1^*) + \Pr(Z = z_2^*) = r < 1$ with $plim \frac{n}{N} = r$ where N is the total sample size and $n =$

By considering only the endpoints of the support of Z, recoding Z as 0 and 1, and with D being a binary treatment variable, we can partition the population into four groups defined as $\mathcal{T}_i = a$ if $D_i^1 = D_i^0 = 1$ (always treated), $\mathcal{T}_i = n$ if $D_i^1 = D_i^0 = 0$ (never treated), $\mathcal{T}_i = c$ if $D_i^1 > D_i^0$ (compliers), $\mathcal{T}_i = d$ if $D_i^1 < D_i^0$ (defiers). We assume that

Assumption 1:

i) Existence of compliers:	$\Pr(\mathcal{T} = c) = P_c > 0$
ii) Monotonicity:	$\Pr(\mathcal{T} = d) = 0$
iii) Independent instrument:	$(Y^d, \mathcal{T}) \perp\!\!\!\perp Z \mid X$
iv) Common support:	$0 < p(X) < 1 \quad a.s.$

where $p(x) = \Pr(Z = 1|X = x)$.

A comment on notation: We will often refer to $p(x)$ as the "propensity score", where one should note that it refers to the instrument Z and not, as usual, to the treatment D. The first assumption requires that at least some individuals react to movements in the instrument. The strength of the instrument can be measured by P_c, which is the probability mass of the compliants. The second assumption is often referred to as monotonicity. It requires that D_i^z either weakly increases with z for all individuals (or decreases for all individuals). The third assumption is the main instrumental variable assumption. It implicitly requires an exclusion restriction (=triangularity) and an unconfounded instrument restriction. In other words, Z_i should not affect the potential outcomes of individual i directly and those individuals for whom $Z = z$ is observed should not differ in their relevant unobserved characteristics from individuals with $Z \neq z$. Unless the instrument has been randomly assigned, this last restriction is often very unlikely to hold. However, *conditional* on a large set of covariates X, this assumption is often more plausible.[12] Note further that we do *not* need X to be exogenous. X can be related to U and V in (1) in any way. This may be important in many applications where X often contains lagged (dependent) variables that may well be related to unobserved ability U.[13]

$\sum_{i=1}^{N} 1(Z_i \in \{z_1^*, z_2^*\})$ the number of observations at the endpoints of the support of Z. When calculating the variance approximation for a particular application, the sample size n should be used. If r is much smaller than 1, there could be finite-sample precision gains by smoothing over Z. We leave this for future research.

[12] In our application, the distance to college instrument is clearly not randomly assigned and individuals with $Z = 1$ are certainly different from those with $Z = 0$. After conditioning on a number of X variables, particularly family background variables that capture the endogenous location choice, the assumption becomes more plausible.

[13] See also Frölich (2006).

The fourth assumption requires that the support of X is identical in the $Z = 0$ and the $Z = 1$ subpopulation. This assumption is needed since we first condition on X to make the instrumental variables assumption valid but then integrate out to obtain the unconditional treatment effects. An alternative set of assumptions, which leads to the same estimators later, replaces monotonicity (Assumption 1ii) with assuming that the average treatment effect is identical for compliers and defiers, conditional on X.

Finally, we also need to assume that the quantiles are unique and well-defined:

Assumption 2:

The random variables $Y^1|c$ and $Y^0|c$ are continuous with positive density in a neighborhood of $Q^{\tau}_{Y^1|c}$ and $Q^{\tau}_{Y^0|c}$, respectively.

3 Identification Results and Estimators

3.1 Identification

Theorem 1 stated below demonstrates that the unconditional QTE for the compliers are non-parametrically identified. Detailed proofs of this and all other theorems and lemmas are in the appendix. We now convey the intuition for the results. If Assumption 1 was valid without conditioning on X, the distribution function of Y^1 for the complier sub-population would be identified by

$$\frac{E\left[1\left(Y \leq u\right) D|Z=1\right] - E\left[1\left(Y \leq u\right) D|Z=0\right]}{E\left[D|Z=1\right] - E\left[D|Z=0\right]}.$$

This unconditional distribution function could then be inverted to obtain the unconditional quantile function. Since a similar result applies to the distribution of Y^0, identification of the QTE would directly follow from this simple result. However, conditioning on the X is necessary. Note that although it is unknown which observations are the compliers, the size of the complier sub-population with characteristics x is identified as

$$\Pr(\mathcal{T} = c|X = x) = E\left[D|X = x, Z = 1\right] - E\left[D|X = x, Z = 0\right],$$

and that, for all x with $\Pr(\mathcal{T} = c|X = x) > 0$, the conditional distribution function of Y^1 for the compliers (we get a similar result for Y^0) is identified as

$$F_{Y^1|X=x,\mathcal{T}=c}(u) = \frac{E\left[1\left(Y \leq u\right) D|X = x, Z = 1\right] - E\left[1\left(Y \leq u\right) D|X = x, Z = 0\right]}{\Pr(\mathcal{T} = c|X = x)}. \qquad (2)$$

We can, thus, identify the treatment effects for the compliers with characteristics $X = x$. More interesting, however, would be an estimate of the distribution for the subpopulation of all compliers, which is the largest population for which the effect is identified. The simple integration $\int F_{Y^1|X,\mathcal{T}=c}(u)dF_X$ of the conditional distribution using the observable distribution of X does not provide the solution to this problem. Moreover, the finite sample properties of such an estimator should be poor when $\Pr(\mathcal{T} = c|X = x)$ is small for certain values of x. If we want to obtain the unconditional distribution for the compliers, we need to weight the conditional distribution by the density of X for the compliers, $dF_{X|\mathcal{T}=c}$. This distribution is not observed but, by Bayes' law, $dF_{X|\mathcal{T}=c} = \frac{\Pr(\mathcal{T}=c|X=x)}{\Pr(\mathcal{T}=c)}dF_X$. Therefore, $F_{Y^1|\mathcal{T}=c}(u) = \int F_{Y^1|X,\mathcal{T}=c}(u)\frac{\Pr(\mathcal{T}=c|X=x)}{\Pr(\mathcal{T}=c)}dF_X$. Using (2), we obtain one of the results of Theorem 1.

Theorem 1 (Identification: Matching on X) *Under Assumption 1, the potential outcome distributions for the compliers are nonparametrically identified as*

$$F_{Y^1|c}(u) = \frac{\int \left(E\left[1\left(Y \leq u\right)D|X, Z = 1\right] - E\left[1\left(Y \leq u\right)D|X, Z = 0\right]\right)dF_X}{\int \left(E\left[D|X, Z = 1\right] - E\left[D|X, Z = 0\right]\right)dF_X}, \tag{3}$$
$$F_{Y^0|c}(u) = \frac{\int \left(E\left[1\left(Y \leq u\right)(D-1)|X, Z = 1\right] - E\left[1\left(Y \leq u\right)(D-1)|X, Z = 0\right]\right)dF_X}{\int \left(E\left[D|X, Z = 1\right] - E\left[D|X, Z = 0\right]\right)dF_X}.$$

which gives the QTE as the difference between the quantiles:

$$Q^{\tau}_{Y^1|c} = F^{-1}_{Y^1|c}(\tau) \qquad Q^{\tau}_{Y^0|c} = F^{-1}_{Y^0|c}(\tau).$$

Straightforward nonparametric estimators exist for all elements appearing in Theorem 1. $E\left[D|X = x, Z = z\right]$ can be estimated for instance by a local logit estimator or by a logistic series approximation. $E\left[1\left(Y \leq u\right)D|X, Z = 1\right]$ could also be estimated by a different local logit procedure for each u. An alternative that may be fruitful when we want to estimate the whole distribution (or at least for a large number of u) consists in estimating the conditional quantile function by local quantile regression and then to invert this function. Instead of using kernel weights, nearest neighbors estimators may also be used to estimate all conditional functions appearing in Theorem 1.

The estimators based directly on Theorem 1 can be qualified as regression (or matching) estimators because they correspond to a function of several nonparametric regressions on X. In the exogenous treatment evaluation literature, two alternative approaches are widely used: regression (or matching) on the propensity score and weighting estimators based on the propensity

score. The following Lemma shows that matching on the propensity score can also be used in our case

Lemma 2 (Propensity score matching) *Let $P = p(X)$ and dF_P be the distribution of P. Under Assumption 1 it follows that:*

$$F_{Y^1|c}(u) = \frac{\int \left(E\left[1\left(Y \leq u\right) D|P, Z=1\right] - E\left[1\left(Y \leq u\right) D|P, Z=0\right]\right) dF_P}{\int \left(E\left[D|P, Z=1\right] - E\left[D|P, Z=0\right]\right) dF_P}, \tag{4}$$
$$F_{Y^0|c}(u) = \frac{\int \left(E\left[1\left(Y \leq u\right) (D-1)|P, Z=1\right] - E\left[1\left(Y \leq u\right) (D-1)|X, Z=0\right]\right) dF_P}{\int \left(E\left[D|P, Z=1\right] - E\left[D|P, Z=0\right]\right) dF_P}.$$

If the propensity score is known or if a parametric functional form can be assumed for it, then matching on the propensity score has the advantage that it does not require high-dimensional nonparametric regressions. If the propensity score must be estimated, then high-dimensional nonparametric functions must be estimated anyway and it is an empirical question to find which one is better suited for a particular dataset.

Starting from (3) one can also derive an expression for the unconditional distribution functions by appropriate weighting of the observations. Note that

$$E\left[E\left[1\left(Y \leq u\right) D|X, Z=1\right]\right] = E\left[E\left[\frac{1\left(Y \leq u\right) DZ}{p(X)}|X\right]\right] = E\left[\frac{1\left(Y \leq u\right) DZ}{p(X)}\right]$$

and

$$E\left[E\left[1\left(Y \leq u\right) D|X, Z=0\right]\right] = E\left[E\left[\frac{1\left(Y \leq u\right) D(1-Z)}{1-p(X)}|X\right]\right] = E\left[\frac{1\left(Y \leq u\right) D(1-Z)}{1-p(X)}\right].$$

Moreover,

$$\begin{aligned} P_c &= E\left[E\left[D|X, Z=1\right] - E\left[D|X, Z=0\right]\right] \\ &= E\left[\frac{E\left[DZ|X\right]}{p(X)} - \frac{E\left[D\left(1-Z\right)|X\right]}{1-p(X)}\right] \\ &= E\left[D \cdot \frac{Z-p(X)}{p(X)\left(1-p(X)\right)}\right]. \end{aligned}$$

Thus, we can show that the expressions in (3) have the following equivalent representation:

Lemma 3 *Under Assumption 1, the potential outcome distributions are identified as*

$$F_{Y^1|c}(u) = \frac{E\left[1\left(Y < u\right) DW\right]}{E\left[DW\right]} \tag{5}$$
$$F_{Y^0|c}(u) = \frac{E\left[1\left(Y < u\right)\left(1-D\right) W\right]}{E\left[DW\right]},$$

where[14]

$$W = \frac{Z - p(X)}{p(X)(1 - p(X))}(2D - 1). \tag{6}$$

Hence, one could estimate the QTE by the difference

$$q_1 - q_0$$

of the solutions of the two moment conditions

$$\begin{aligned} E[1(Y < q_1)DW] &= \tau E[DW] \\ E[1(Y < q_0) \cdot (1 - D)W] &= \tau E[DW] \end{aligned} \tag{7}$$

or equivalently

$$\begin{aligned} E[\{1(Y < q_1) - \tau\}WD] &= 0 \\ E[\{1(Y < q_0) - \tau\}W(1 - D)] &= 0, \end{aligned} \tag{8}$$

because $E[W] = 2P_c$ and $E[DW] = P_c$. We could thus estimate q_0 and q_1 by these weighted univariate quantiles in the $D = 0$ and $D = 1$ populations. Alternatively, we could estimate the treatment effect directly by a weighted quantile regression:

Lemma 4 *Under Assumption 1, the solution of the following optimization problem*

$$(\alpha, \beta) = \arg\min_{a,b} E[\rho_\tau(Y - a - bD) \cdot W], \tag{9}$$

where $\rho_\tau(u) = u \cdot \{\tau - 1(u < 0)\}$*, is equivalent to the solutions to the moment conditions (8) in that the solution for* a *corresponds to* $Q^\tau_{Y^0|c}$ *and the solution for* b *corresponds to* $\Delta^\tau_c = Q^\tau_{Y^1|c} - Q^\tau_{Y^0|c}$.

Note that the sample objective function to (9) is typically non-convex since W_i is negative for $Z_i \neq D_i$. This complicates the optimization problem a little because local optima could exist. AAI notice a similar problem in their approach but our problem is less serious here because we

[14] As discussed below, these weights are different from the weights used by AAI. However, they appear to be the same as the weights κ_0 and κ_1 suggested in Theorem 1 of Abadie (2003). In that paper, he is interested in the conditional mean function. Therefore, as in AAI, he does not use κ_0 and κ_1 for estimation but only the combination κ, which we call W_{AII} below.

need to estimate only a *scalar* in the $D = 1$ population and another one in the $D = 0$ population. In other words, we can write (9) equivalently as

$$(Q^{\tau}_{Y^1|c}, Q^{\tau}_{Y^0|c}) = \left(\arg\min_{q_1} E\left[\rho_\tau(Y - q_1) \cdot W | D = 1\right], \quad \arg\min_{q_0} E\left[\rho_\tau(Y - q_0) \cdot W | D = 0\right] \right), \quad (10)$$

which are two separate *one-dimensional* estimation problems in the $D = 1$ and $D = 0$ populations such that we can easily use grid-search methods supported by visual inspection of the objective function for local minima.

Although the negativity of some of the weights W is not a serious problem, we consider two alternatives to it. The first alternative relates back to Theorem 1 in that we could estimate the cdf via (3), (4) or (5) instead of the quantiles via (9), particularly if one is interested in the entire distribution instead of only the effect at one single quantile, e.g. the median.

Alternatively, we could apply an iterated expectations argument to (9) to obtain

$$\begin{aligned}
(\alpha, \beta) &= \arg\min_{a,b} E\left[\rho_\tau(Y - a - bD) \cdot W\right] \\
&= \arg\min_{a,b} E\left[\rho_\tau(Y - a - bD) \cdot E\left[W | Y, D\right]\right] \\
&= \arg\min_{a,b} E\left[\rho_\tau(Y - a - bD) \cdot W^+\right]
\end{aligned}$$

where

$$W^+ = E\left[W \,|Y, D\right] = E\left[\frac{Z - p\left(X\right)}{p(X)\left(1 - p(X)\right)} \,|Y, D\right](2D - 1). \quad (11)$$

These new weights W^+ are always nonnegative as shown below. Hence, they can be used to develop an estimator with a linear programming representation. The sample objective function to (9) with W^+ instead of W is globally convex since it is the sum of convex functions, and the global optimum can be obtained in a finite number of iterations. However, we would need to estimate (11) first. Note that AAI suggest a similar projection approach, but their weights are conditional on Y, D *and* X. Hence, nonparametric estimation of their weights is more difficult and computationally demanding, whereas estimation of (11) requires only *univariate* nonparametric regression separately for the $D = 0$ and $D = 1$ populations.

We show now that these weights W^+ are always non-negative. If $D = 1$ the weights W^+

would be negative if $E\left[Z - p\left(X\right)|Y, D=1\right] < 0$. However,

$$\begin{aligned} E\left[Z - p\left(X\right)|Y, D=1\right] &= E\left[E\left[Z|X, Y^1, D=1\right] - p\left(X\right)|Y^1, D=1\right] \\ &= E\left[\frac{\Pr\left(Z=1, D=1|X, Y^1\right)}{\Pr\left(D=1|X, Y^1\right)} - p\left(X\right)|Y^1, D=1\right] \\ &= E\left[\frac{\Pr\left(Z=1, D=1|X, Y^1\right) - p\left(X\right)\cdot\Pr\left(D=1|X, Y^1\right)}{\Pr\left(D=1|X, Y^1\right)}|Y^1, D=1\right] \geq 0 \end{aligned}$$

because of

$$\begin{aligned} & \Pr\left(Z=1, D=1|X, Y^1\right) - p\left(X\right)\cdot\Pr\left(D=1|X, Y^1\right) \\ = \; & \Pr\left(Z=1, D=1|X, Y^1\right) - p\left(X\right)\cdot\left\{\Pr\left(D=1, Z=1|X, Y^1\right) + \Pr\left(D=1, Z=0|X, Y^1\right)\right\} \\ = \; & (1-p(X))\cdot\Pr\left(Z=1, D=1|X, Y^1\right) - p\left(X\right)\cdot\Pr\left(D=1, Z=0|X, Y^1\right) \\ = \; & (1-p(X))\cdot\Pr\left(Z=1, \mathcal{T}\in\{a,c\}|X, Y^1\right) - p\left(X\right)\cdot\Pr\left(Z=0, \mathcal{T}=a|X, Y^1\right) \\ = \; & (1-p(X))\cdot\Pr\left(Z=1|X, Y^1\right)\cdot\Pr\left(\mathcal{T}\in\{a,c\}|X, Y^1\right) - p\left(X\right)\cdot\Pr\left(Z=0|X, Y^1\right)\cdot\Pr\left(\mathcal{T}=a|X, Y^1\right) \\ = \; & (1-p(X))\cdot p(X)\cdot\Pr\left(\mathcal{T}\in\{a,c\}|X, Y^1\right) - p\left(X\right)\cdot(1-p(X))\cdot\Pr\left(\mathcal{T}=a|X, Y^1\right) \\ = \; & (1-p(X))\cdot p(X)\cdot\left\{\Pr\left(\mathcal{T}\in\{a,c\}|X, Y^1\right) - \Pr\left(\mathcal{T}=a|X, Y^1\right)\right\} \geq 0 \end{aligned}$$

because $\mathcal{T} \perp\!\!\!\perp Z|X, Y^d$ and $Y^d \perp\!\!\!\perp Z|X$ by Assumption (1iii).

On the other hand, if $D=0$ the weights W^+ would be negative if $E\left[Z - p\left(X\right)|Y, D=0\right] > 0$. By analogous derivations as above one can show that

$$E\left[Z - p\left(X\right)|Y, D=0\right] \leq 0.$$

Therefore, the weights W^+ are always non-negative.

3.2 Relationship to the Existing Literature

These results bear some resemblance with AAI, who suggested estimating a weighted linear quantile regression

$$\underset{\alpha,\beta}{\arg\min} E\left[\rho_\tau(Y - \alpha D - \beta' X)\cdot\left(1 - \frac{D(1-Z)}{1-p(X)} - \frac{(1-D)Z}{p(X)}\right)\right]. \tag{12}$$

However, both the model and the estimand are different. AAI impose a linear parametric specification, whereas our approach is entirely nonparametric. Furthermore, they identify the conditional treatment effects, i.e. conditional on X, whereas we are interested in the *unconditional* treatment effects.

Note that the approach of AAI generally *cannot* be used for estimating unconditional treatment effects since the weights in AAI are not appropriate for that case. In other words, one might be thinking to run a weighted quantile regression of Y on a constant and D by using equation (12) and replacing X by a constant in the first term. For that purpose, however, the weights of AAI are *not correct* as shown in the following proposition.

Proposition 5 *The solution of*

$$\arg\min_{\alpha,\beta} E\left[\rho_\tau(Y-\alpha-\beta D)\cdot\left(1-\frac{D(1-Z)}{1-p(X)}-\frac{(1-D)Z}{p(X)}\right)\right] \tag{13}$$

for β gives the difference between the τ quantiles of the treated compliers *and* non-treated compliers*, respectively:*

$$\beta = F^{-1}_{Y^1|c,D=1}(\tau) - F^{-1}_{Y^0|c,D=0}(\tau)$$

where

$$F_{Y^1|c,D=1}(u) = \Pr(Y^1 \leq u | D=1, T=c)$$
$$F_{Y^0|c,D=0}(u) = \Pr(Y^0 \leq u | D=0, T=c).$$

This difference is not very meaningful as one compares the Y^1 outcomes among the treated with the Y^0 outcomes among the non-treated. Therefore, in the general case the weights of AAI are only useful to estimate conditional quantile effects. Hence, if one is interested in nonparametric estimation of the unconditional QTE, one should use the weights in (9) but not those in (13). Note, however, that their weights can be used in a special case: when the instrumental variable is independent of X such that we can write $p(X) = p$.

To show this equivalence, we first define

$$W_{AAI} = 1 - \frac{D(1-Z)}{1-p} - \frac{(1-D)Z}{p}.$$

The following relation between the weights W, defined in 6, and W_{AAI} can be shown when $p(x) = p$ is a constant

$$W_{AAI} = (Dp + (1-D)(1-p))\, W.$$

This implies that, conditionally on D, W is a multiple of W_{AAI}. As shown in (10), the unconditional quantiles for the compliers can be estimated by univariate weighted quantiles separately in the $D=0$ and the $D=1$ population. Since multiplying with a positive constant does not

change the result of the minimization, this completes the proof of the equivalence of the weights if $p(x) = p$ is a constant:

$$Q^{\tau}_{Y^0|c} = \arg\min_{q_0} E\left[\rho_\tau(Y - q_0) \cdot W | D = 0\right] = \arg\min_{q_0} E\left[\rho_\tau(Y - q_0) \cdot W_{AAI} | D = 0\right]$$

$$Q^{\tau}_{Y^1|c} = \arg\min_{q_1} E\left[\rho_\tau(Y - q_1) \cdot W | D = 1\right] = \arg\min_{a} E\left[\rho_\tau(Y - q_1) \cdot W_{AAI} | D = 1\right].$$

This equivalence result holds only if $p(X)$ is constant, which implies that the instrument is valid without conditioning on any X. This is often the case when the instrument is randomly assigned, unless the randomization probabilities vary between strata, e.g. by gender or nationality or income groups. Therefore, AAI could have estimated the unconditional QTE with the weights used to estimate conditional QTE. However, they have changed the estimand by including additional covariates in the regression. In this sense, our weights W can be considered as a generalization of the weights W_{AAI} for the case when Z is not independent of X. In that case, the weights W_{AAI} do not work for the unconditional QTE whereas the weights W do.

On the other hand, if one were interested in estimating conditional QTE using a parametric specification, the weights W we propose in (6) could also be used. Hence, although not developed for this case, our weights W can be used in (12). More precisely

Proposition 6 *If one assumes a linear model for the conditional quantile for the compliers*

$$F_Y^{-1}(\tau | X, D, T = c) = X'\beta_0^{\tau} + \alpha_0^{\tau} D,$$

a weighted quantile regression with weights W would identify α_0^{τ} and β_0^{τ}.[15]

Hence, both types of weights, i.e. those of AAI and those in (6), would identify the conditional quantile treatment effects, but it is not clear which one will be more efficient. For the compliers, W varies with x whereas the weights in AAI are identical to one. In any case, both types of weights would be generally inefficient since they do not incorporate the conditional density function of the error term at the τ quantile. Hence, if one was mainly interested in estimating *conditional* QTE with a parametric specification, more efficient estimators could be developed.

Consider now another special case of our model: the case where treatment D is exogenous conditional on X. If D is exogenous, then we can use D as its own instrument and set $Z = D$

[15] Instead of W one could also use $E[W|Y, X, D]$, which are always nonnegative, but usually not the weights $W^+ = E[W|Y, D]$ as conditioning on X is necessary here.

and our representation in (3) simplifies to

$$F_{Y^1}(u) = \int E\left[1\left(Y \leq u\right)|X, D = 1\right] dF_X.$$

When the conditional distribution is estimated by local regression, we obtain the estimator proposed in Frölich (2007b); when it is estimated by non-parametric quantile regression, this is the estimator proposed by Melly (2006); when it is estimated by parametric methods we obtain the estimators proposed by Machado and Mata (2005), Gosling, Machin, and Meghir (2000) and, more generally, Chernozhukov, Fernández-Val, and Melly (2007). Furthermore, in this exogenous case, our weights simplify to

$$W = \frac{D}{p(x)} + \frac{1-D}{1-p(x)}.$$

These are exactly the weights proposed by Firpo (2007) and it shows that our methods generalize all existing approaches to estimating QTE under exogeneity. The weights of Firpo (2007) are always positive such that there is no need for estimating positive weights by projection. Finally, note that the whole population complies when we assume exogeneity such that we obtained the QTE for the population.

4 Asymptotic Properties

In the previous section the identification of unconditional QTE under endogeneity has been considered and several natural estimators have been suggested. In this section, we first analyze the asymptotic properties of one of the proposed estimators for the τ quantile treatment effect, Δ_c^τ.[16] We derive then the semiparametric efficiency bound and show that our estimator is indeed efficient. We finally show that the efficiency bound does not change if we know the propensity score $p(x)$ but does decrease if we include additional covariates when these covariates are not needed for consistency.

From Lemma 4, a natural estimator of $\Delta_c^\tau = Q_{Y^1|c}^\tau - Q_{Y^0|c}^\tau$ is given by

$$(\hat{Q}_{Y^0|c}^\tau, \hat{\Delta}_c^\tau) = \arg\min_{a,b} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - a - bD_i)\hat{W}_i \tag{14}$$

[16] We consider the weighting estimator, which is the simplest one to implement since it requires only one non-parametric regression.

or numerically equivalently via:

$$\hat{Q}^{\tau}_{Y^1|c} = \arg\min_{q_1} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - q_1) D_i \hat{W}_i$$
$$\hat{Q}^{\tau}_{Y^0|c} = \arg\min_{q_0} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - q_0) (1 - D_i) \hat{W}_i.$$

For this we need a first step estimator of the weights $\hat{W}_i$, which depend on a nonparametric estimate of $p(x)$. For concreteness, we develop the asymptotic distribution for $\hat{p}(x)$ being estimated by *local linear regression* in Theorem 7 and by *local logit regression* in Theorem 8. Alternative nonparametric estimators could be used as well, but local linear regression has several appealing properties. It has better boundary properties than Nadaraya-Watson regression and is easier to implement than local quadratic or local cubic regression, particularly when $\dim(X)$ is large.[17][18]

The local linear regression estimator of $p(x_0)$ at a location x_0 is defined as the value of a that solves the weighted least squares regression

$$\min_{a,b} \sum_{j=1}^{n} \left(Z_j - a - b'\left(X_j - x_0\right)\right)^2 K_j$$

where K_j is the product kernel:

$$K_j = K_h(X_j - x_0) = \frac{1}{h^L} \prod_{l=1}^{L} \kappa\left(\frac{X_{jl} - x_l}{h}\right),$$

where X_{jl} is the l-th element of X_j and x_l is the l-th element of x_0. Further, κ is a univariate kernel function of order λ, which is assumed to be integrating to one. The following kernel constants will be used later: $\mu_t = \int u^t \kappa(u) du$ and $\bar{\mu}_t = \int u^t \kappa^2(u) du$. The kernel function being of order λ means that $\mu_0 \neq 1$ and $\mu_t = 0$ for $0 < t < \lambda$ and $\mu_\lambda \neq 0$.

Assumption 3 gives regularity conditions under which the estimator is asymptotically normal and efficient:

[17] An alternative is series regression. The use of series methods as e.g. in Hirano, Imbens, and Ridder (2003) or Firpo (2007), however, seems to require much stronger smoothness assumptions. E.g. Firpo (2007) requires more than *seven* times $\dim(X)$ continuous derivatives of the propensity score. If X contains say 10 variables, more than 70 derivatives are needed. Although we need smoothness of several functions, we never require such a large amount of smoothness. In addition, when $\dim(X)$ is large, collinearity problems can make the implementation of series regression difficult in small samples.

[18] Since our estimator includes Firpo (2007) as a special case, for $Z = D$, the proofs below also complement his article when local linear or local logit estimation is used instead of series regression as in his article.

Assumption 3:

i) The data $\{(Y_i, D_i, Z_i, X_i)\}$ are iid from $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathcal{X}$ with $\mathcal{X} \subset \mathbb{R}^L$ being a compact set.

ii) $p(x)$ is bounded away from 0 and 1 over $\mathcal{X}$.

iii) Smoothness:

- $p(x)$ is 2 times continuously differentiable with second derivative Hölder continuous,
- $f(x)$ is $\lambda - 1$ times continuously differentiable with (λ-1)-th derivative Hölder continuous,
- $F_{Y|d,z,x}(y)$ is continuously differentiable with respect to y.

iv) Uniform consistency: The estimator $\hat{p}(x)$ satisfies

$$\sup_{x \in \mathcal{X}} |\hat{p}(x) - p(x)| \xrightarrow{p} 0.$$

v) The univariate Kernel function κ is compactly supported, bounded, Lipschitz and of order λ. We also assume that $\int \kappa(u)du = 1$.

vi) The bandwidth satisfies $nh^L / \ln n \to \infty$ and $nh^{2\lambda} \to 0$.

Since the estimated weights $\hat{W}$ imply a weighting by the inverses of $\hat{p}(x)$ and $1-\hat{p}(x)$, we need $\hat{p}(x)$ to be bounded away from zero and one. This is implied by Assumption (3ii) and (3iv). In Assumption (3iv) we simply assume $\hat{p}(x)$ to be uniformly consistent since there are many different sets of assumptions under which local linear estimation can be shown to be uniformly consistent. Some assumptions may be more appropriate in certain settings, other more in others, see e.g. Fan (1993), Masry (1996) or Gozalo and Linton (2000). For example, if we use a conventional second order kernel ($\lambda = 2$), the results of Gozalo and Linton (2000) apply to the local linear and also to the local logit estimator examined later. Their Theorem 1(ii) requires $f(x)$ to be bounded away from zero and further that $f(x)$ and $p(x)$ are continuous.[19]

Assumption (3v) and (3vi) are needed to reduce the bias term to a sufficiently small order. Together they require that $\lambda > L/2$. Hence, if X contains 4 or more continuous regressors, higher order kernels are required. With 3 or less continuous regressors, conventional kernels can be used.

To control the bias, for practical purposes we propose to use local linear regression with higher-order kernels, where we suggest using a product kernel. Instead of using higher order kernels, one could alternatively use local higher order polynomial regression instead of local linear regression.

[19] Apply their Theorem 1(ii) with $s = r = 0$. They also require the existence of $E[Z^2] < \infty$ and $Var(Z|X = x)$ to be a continuous function of x, which are trivially satisfied since Z is Bernoulli.

However, when the number of regressors in X is large, this could be inconvenient to implement in practice since a large number of interaction and higher order terms would be required, which could give rise to problems of local multicollinearity in small samples and/or for small bandwidth values. On the other hand, higher order kernels are very convenient to implement in practice when a product kernel is used. In addition, they conveniently permit to smooth over continuous and discrete regressors as suggested by Racine and Li (2004). Although the asymptotic theory is not affected by discrete regressors and the common solution is to conduct separate regressions within each cell spanned by the discrete regressors, smoothing over discrete regressors can increase precision in finite samples.

We could permit for a more general kernel function with multiple bandwidths as e.g. in Ruppert and Wand (1994) at the expense of a more complex notation. In practice, it appears to be common to rotate the data beforehand such that the covariance matrix is the identity matrix and to use a common bandwidth, instead of estimating a different bandwidth value for each X variable.

The following theorem gives the asymptotic distribution of $\hat{\Delta}_c^\tau$:

Theorem 7 (Asymptotic distribution) *Under Assumptions 1 to 3, the estimator (14) is $\sqrt{n}$ consistent, asymptotically normal and efficient:*

$$\sqrt{n}\left(\hat{\Delta}_c^\tau - \Delta_c^\tau\right) \xrightarrow{d} N\left(0, \mathcal{V}\right),$$

where

$$\begin{aligned}
\mathcal{V} = {} & \frac{1}{P_c^2 f^2_{Y^1|c}(Q^\tau_{Y^1|c})} E\left[\frac{\pi(X,1)}{p(X)} F_{Y|D=1,Z=1,X}(Q^\tau_{Y^1|c})\left(1 - F_{Y|D=1,Z=1,X}(Q^\tau_{Y^1|c})\right)\right] \\
& + \frac{1}{P_c^2 f^2_{Y^1|c}(Q^\tau_{Y^1|c})} E\left[\frac{\pi(X,0)}{1-p(x)} F_{Y|D=1,Z=0,X}(Q^\tau_{Y^1|c})\left(1 - F_{Y|D=1,Z=0,X}(Q^\tau_{Y^1|c})\right)\right] \\
& + \frac{1}{P_c^2 f^2_{Y^0|c}(Q^\tau_{Y^0|c})} E\left[\frac{1-\pi(X,1)}{p(X)} F_{Y|D=0,Z=1,X}(Q^\tau_{Y^0|c})\left(1 - F_{Y|D=0,Z=1,X}(Q^\tau_{Y^0|c})\right)\right] \\
& + \frac{1}{P_c^2 f^2_{Y^0|c}(Q^\tau_{Y^0|c})} E\left[\frac{1-\pi(X,0)}{1-p(X)} F_{Y|D=0,Z=0,X}(Q^\tau_{Y^0|c})\left(1 - F_{Y|D=0,Z=0,X}(Q^\tau_{Y^0|c})\right)\right] \\
& + E\Big[\frac{\pi(X,1)\vartheta^2_{11}(X) + (1-\pi(X,1))\,\vartheta^2_{01}(X)}{p(X)} + \frac{\pi(X,0)\vartheta^2_{10}(X) + (1-\pi(X,0))\,\vartheta^2_{00}(X)}{1-p(X)}\Big] \\
& - E\Big[p(X)(1-p(X))\left\{\frac{\pi(X,1)\vartheta_{11}(X) + (1-\pi(X,1))\,\vartheta_{01}(X)}{p(X)} + \frac{\pi(X,0)\vartheta_{10}(X) + (1-\pi(X,0))\,\vartheta_{00}(X)}{1-p(X)}\right\}^2\Big],
\end{aligned}$$

where $\vartheta_{dz}(x) = \frac{\tau - F_{Y|D=d,Z=z,X}(Q^{\tau}_{Y^d|_c})}{P_c \cdot f_{Y^d|_c}(Q^{\tau}_{Y^d|_c})}$ *and* $\pi(x,z) = \Pr(D = 1|X = x, Z = z)$ *and* $P_c = \int (\pi(x,1) - \pi(x,0))\, dF_X$ *is the fraction of compliers and*

$$f_{Y^1|_c}(u) = \left\{ \int \left(f_{Y|X,D=1,Z=1}(u)\pi(x,1) - f_{Y|X,D=1,Z=0}(u)\pi(x,0) \right) dF_X \right\} / P_c$$
$$f_{Y^0|_c}(u) = -\left\{ \int \left(f_{Y|X,D=0,Z=1}(u)\,(1 - \pi(x,1)) - f_{Y|X,D=0,Z=0}(u)\,(1 - \pi(x,0)) \right) dF_X \right\} / P_c.$$

(The proof of efficiency follows later.)

The variance contributions stem from two parts: First the weighting by W if the weights were known and second from the fact that the weights were estimated. The variance contribution to the estimation of $Q^{\tau}_{Y^1|_c}$ due to the weighting follows from the term $\left(\frac{ZD}{p(X)} - \frac{(1-Z)D}{1-p(X)}\right) \frac{\tau - 1(y \leq Q^{\tau}_{Y^1|_c})}{P_c \cdot f_{Y^1|_c}(Q^{\tau}_{Y^1|_c})}$ and the variance due to the nonparametric estimation of the weights is $-\left(\frac{\pi(x,1)(Z-p(x))}{p(x)}\right)\vartheta_{11}(X) - \left(\frac{\pi(x,0)(Z-p(x))}{1-p(x)}\right)\vartheta_{10}(X)$. The terms for the estimation of $Q^{\tau}_{Y^0|_c}$ are derived analogously, and the above variance expression for the QTE follows.

As it is often the case for semiparametric estimators, the first order asymptotic theory does not depend on the bandwidth value anymore, which has the unpleasant implication that it is not helpful for selecting bandwidth values. In principle, we could also extend the proof to derive a second order approximation to the mean squared error where the second order terms would depend on the bandwidth values. From the derivations in the proof, however, it appears that the second order terms would be very complex since they depend on higher order derivatives of several types of functions. Hence, although feasible to derive, it appears that the second order approximation would be of little practical value. Too many nuisance functions would have to be estimated and plugged in and the estimated bandwidth could be very noisy in finite sample. It thus appears that alternative approaches to bandwidth selection should be considered as e.g. the empirical bias bandwidth selector of Ruppert (1997) and further developed in Flossmann (2007). We leave this for future research.

Since we derive only first order asymptotics we can treat every point X_i where $p(X_i)$ is to be estimated as an interior point. In contrast to Nadaraya-Watson regression, the variance and bias of local linear regression are of the same order in the interior as at the boundary. The magnitude of the bias, however, is generally larger at the boundary. The impact of the boundary of X on the MSE of $\hat{\Delta}^{\tau}_c$ vanishes with h and would only show up if we were to derive the second

order asymptotics. This would be a further complication to the formula for the second order asymptotics as discussed above.

As an alternative to local linear regression, we also consider *local logit* regression in the following. Local linear regression has received extensive praise for its favorable properties, but does not ensure that $\hat{p}(x) \in (0,1)$. We could simply cap the estimates of $\hat{p}(x)$ above one or below zero by setting them to some value below one or above zero, but local logit regression performs often better than local linear regression in finite samples for binary dependent variables.

Define the log likelihood function for local logit regression at a location x_0 as

$$\ln L_n(x_0, a, b) = \frac{1}{n}\sum_{j=1}^{n} \left\{ Z_j \ln \Lambda\left(a + b'\left(X_j - x_0\right)\right) + \left(1 - Z_j\right) \ln\left(1 - \Lambda\left(a + b'\left(X_j - x_0\right)\right)\right)\right\} K_j$$

where $\Lambda(x) = \frac{1}{1+e^{-x}}$. Let $\hat{a}$ and $\hat{b}$ be the maximizers of $\ln L_n(x_0, a, b)$ and a_0 and b_0 be the values that maximize the expected value of the likelihood function $E\left[\ln \mathcal{L}_n(x_0, a, b)\right]$. Note that we are interested only in $\hat{a}$ and include $\hat{b}$ only to appeal to the well known properties that local likelihood or local estimating equations perform better if more than a constant term is included in the local approximation (see e.g. Fan and Gijbels (1996) and Carroll, Ruppert, and Welsh (1998)). We estimate $p(x_0)$ by $\hat{p}(x_0) = \Lambda(\hat{a}(x_0))$. As with local linear regression, we need uniform consistency of $\hat{p}(x_0)$. As we did with the local linear estimator, we simply assume uniform consistency here as uniform consistency can be achieved under different sets of regularity conditions, see e.g. Gozalo and Linton (2000) or Nielsen (2005). The following Assumption 4 and Theorem 8 are very similar to Assumption 3 and Theorem 7 stated above.

Assumption 4:

i) The data $\{(Y_i, D_i, Z_i, X_i)\}$ are iid from $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathcal{X}$ with $\mathcal{X} \subset \mathbb{R}^L$ being a compact set.

ii) $p(x)$ is bounded away from 0 and 1 over $\mathcal{X}$.

iii) Smoothness:

- $p(x)$ is λ times continuously differentiable with λ-th derivative Hölder continuous,
- $f(x)$ is $\lambda - 1$ times continuously differentiable with (λ-1)-th derivative Hölder continuous,
- $F_{Y|d,z,x}(y)$ is continuously differentiable with respect to y.

iv) Uniform consistency: The local logit estimator satisfies

$$\sup_{x \in \mathcal{X}} |\hat{p}(x) - p(x)| \xrightarrow{p} 0$$

v) The univariate Kernel function κ is compactly supported, bounded, Lipschitz and of order λ. We also assume that $\int \kappa(u)du = 1$.
vi) The bandwidth satisfies $nh^L/\ln n \to \infty$ and $nh^{2\lambda} \to 0$.

Theorem 8 (Local logit) *Under Assumptions 1, 2 and 4, the estimator (14), with $\hat{p}(x)$ estimated by local logit, is $\sqrt{n}$ consistent, asymptotically normal and efficient:*

$$\sqrt{n}\left(\hat{\Delta}_c^\tau - \Delta_c^\tau\right) \xrightarrow{d} N\left(0, \mathcal{V}\right).$$

Thus, both estimators have the same asymptotic distribution. An estimator for $\mathcal{V}$, the normalized asymptotic variance of $\hat{\Delta}_c^\tau$ is now suggested.[20] Even if the formula looks very complicated and difficult to estimate, straightforward estimators exist for each of its elements taken separately. In contrary to the asymptotic variance of the quantile regression estimator, for instance, there is no need to estimate conditional densities, which are typically difficult to estimate. On the other hand, we need to estimate the univariate densities of the counterfactual distributions, $f_{Y^1|c}(Q^\tau_{Y^0|c})$ and $f_{Y^1|c}(Q^\tau_{Y^0|c})$. As suggested by Firpo (2007), we can estimate such a density by a reweighted kernel estimator, using the weights already used to estimate the QTE. Firpo (2007) gives regularity conditions under which this estimator is consistent.

Consistent estimators for P_c, $Q^\tau_{Y^1|c}$ and $p(x)$ were given in the proof of Theorems 7 and 8. $\pi(x,z)$ can be estimated using a similar strategy and similar regularity conditions to that used to estimate $p(x)$. Finally, methods to estimate the conditional distribution $F_{Y|D,Z,X}$ are suggested, for instance, in Hall, Wolff, and Yao (1999). We use their local logit estimator in the estimation. The estimator of the variance obtained by inserting all these estimators in the asymptotic formula is consistent by the continuous mapping theorem.

We now show that $\hat{\Delta}_c^\tau$ is indeed efficient in the class of regular semiparametric estimators. In order to show that we derive the semiparametric efficiency bound[21] for Δ_c^τ in Theorem 9.

Theorem 9 (Efficiency bound) *Under Assumptions 1 and 2, the efficiency bound for $\Delta_c^\tau = Q^\tau_{Y^1|c} - Q^\tau_{Y^0|c}$ is $\mathcal{V}$. Therefore, $\hat{\Delta}_c^\tau$ attains the semiparametric efficiency bound.*

[20] This analytical estimator of the variance has also been implemented in Stata. The codes can be obtained from the authors.

[21] Frölich (2007a) derives a similar efficiency bound for average treatment effects and Hong and Nekipelov (2007) derives the bound for general separable models.

Note that this bound simplifies to the bound obtained by Firpo (2007) if we set $D = Z$. The following lemma shows that the efficiency bound is the same when the function $p(x)$ is known. In the leading example of experimental design with imperfect compliance, where Z is randomization into treatment and D is actual treatment receipt, the probability $\Pr(Z = 1|X)$ is often under the control of the institution conducting the experiment and thereby known. This probability might be constant or depend on X in a known way. In the application of AAI, for instance, it is known that the probability of being randomized is 2/3.

When the treatment is exogenous, Hahn (1998) shows that knowledge of the propensity does not help to estimate the treatment effects on the whole population but does help to estimate effects on the treated population. As discussed in Hirano, Imbens, and Ridder (2003) and Frölich (2004), the reason is that the propensity score is not helpful for estimating the conditional mean of the dependent variable. On the other hand, knowledge of the propensity score allows using the whole population to estimate the distribution of the covariates for the treated. The following lemma shows that our case is similar in this respect to the estimation of the ATE for the whole population. Intuitively, we need to integrate the conditional distribution over the whole population, even if the resulting treatment effects are defined for the compliers.

Lemma 10 (Knowledge of propensity score) *Under Assumptions 1 and 2, the efficiency bound for Δ_c^τ is $\mathcal{V}$ irrespective of whether the function $p(x)$ is known or unknown.*

Note that knowledge of $\Pr(D = 1|X, Z)$ would change the variance bound. However, we cannot find a plausible example where this probability would be known and simultaneously D would be endogenous.

Whereas knowledge of $p(x)$ does not affect the variance bound, including more variables in X, however, can reduce it. So far, in all our discussions, the reason for accounting for the X variables was to make Assumption 1 more plausible. As the following theorem shows, however, the X variables can also help to increase *efficiency*. We can therefore include some X variables to make the instrumental variables assumption more plausible and additional X variables to reduce the asymptotic variance. Consider two regressors sets X_1 and X_2 with $X_1 \subset X_2$. X_1 may be the empty set, which is the case when the instrumental variable is randomly assigned as e.g. in AAI. Suppose that *both* regressor sets satisfy Assumption 1. In other words, controlling for X_1 is sufficient to obtain consistent estimates, but X_2 may help to reduce variance.

Generally speaking, the *additional* variables in X_2 could be causal predictors of the instrument Z, and/or of the treatment variable D and/or of the outcome Y. If these variables were predictors of D or Y *and* also of the instrument Z, Assumption 1 would generally not be satisfied without controlling for these confounding variables. However, if the additional variables are only predictors of the instrument Z or if they are only predictors of D and/or Y, controlling for these variables is not necessary for consistency. Both estimators, one using X_1 and the other using X_2, would be consistent. Including additional variables that affect only Z but neither D nor Y usually leads to an efficiency loss. On the other hand, including variables affecting D or Y but not Z leads to efficiency gains. Hence, regressors that turn out as insignificant in a regression of Z on control variables could nevertheless be retained as regressors for efficiency reasons. In the particular case were Z is completely randomized, as e.g. in AAI, one could nevertheless gain efficiency by incorporating control variables in the estimation process.

We suppose for the following theorem that

$$\Pr\left(Z=1|X_1,X_2\right)=\Pr\left(Z=1|X_1\right). \tag{15}$$

Hence, the additional regressors in X_2 that are not included in X_1 do not affect the instrument but may be predictors of the potential outcomes Y and/or the treatment variable D. As mentioned above, without this assumption the estimator using X_1 would generally be inconsistent such that a comparison of variances would be of little interest.

Theorem 11 (Variance reduction) *Let X_1 and X_2 with $X_1 \subset X_2$ be two regressor sets that both satisfy Assumptions 1 and 2 as well as equation (15). Let $\mathcal{V}_1$ be the semiparametric variance bound when using regresor set X_1 and $\mathcal{V}_2$ be the semiparametric variance bound when using regresor set X_2, both referring to the same quantile τ of the QTE. Then*

$$\mathcal{V}_1 \geq \mathcal{V}_2.$$

Except for very special circumstances, the inequality would usually be strict.

5 Applications

5.1 Simulated datasets

We examine now the small sample performance of the proposed estimators on simulated datasets. In order to reveal systematic differences between the behaviors of the estimators, we use 29

different data generating processes (DGP). 200 replications of 400 observations are drawn for each of the DGP. The covariates X consist of three continuous regressors and the unconditional probabilities $\Pr(D=1)$ and $\Pr(Z=1)$ are set to 0.5. All other functions (Y as a function of X and D, D as a function of X and Z, Z as a function of X) vary from one DGP to the other. Therefore, we can examine how the ranking of the performance of the estimators changes when we change some of the parameters of the DGP. While the detailed results are available from the authors, we give here five general lessons we have gained from this exercise.

First, a regression on the nonparametrically estimated propensity score performs almost always worse than a direct regression on the covariates. This is not especially surprising since we know from the literature on ATE that the regression on the propensity score estimator is not efficient for this parameter. A similar result probably applies to the estimation of QTE. Moreover, this estimator is computationally the most demanding one since three nonparametric regressions are required: estimation of the propensity score, and regression of Y and D on the propensity score. For all these reasons, we would not recommend to use this estimator.

Secondly, among the weighting estimators, we find only moderate differences between using W and W^+ as weights and no clear ranking emerges between these two estimators. When we plot the results for a given sample, the estimates using W^+ appear to be smoother (as a function of the quantile at which the QTE is estimated). The projection of the weights on Y looks like an additional smoothing procedure in the Y dimension. Although it may produce more pleasing pictures, this does not necessarily deliver a better performance.

Third, as expected, the local linear estimator does not fit well a binary dependent variable when $\Pr(Z=1|X)$ has a large support even if we censor the estimated probabilities. In this case, the local logit estimator performs much better. However, a simple truncation of the weights (with recentering of the weights such that the estimated distribution functions are still well-defined distribution functions) is sufficient to obtain satisfactory results (as good as using local logit). Since fitting local linear regression is faster than fitting local logit models, this could be a good alternative at least for exploratory analysis and bandwidth selection.

Fourth, the simulations showed several factors determining whether the regression or the weighting estimator performs better. No difference can be found between these two approaches when all regressions are correctly parametrically specified. The finite-sample performance of the regression estimator deteriorates when the outcome equation becomes non-linear while

the performance of the weighting estimator deteriorates when the propensity score becomes non-linear. This is a direct consequence of the different functions estimated by each of these estimators. Finally, the performance of the weighting estimator deteriorates relatively more when the propensity score has much mass near 0 or 1. Note that each of these quantities (linearity of the functions and distribution of the propensity score) could be estimated to determine which estimator should be used for a given application. Alternatively, it should be possible to define a double robust estimator combining both approaches in an optimal way.

The last conclusion we draw from the Monte Carlo simulations concerns the choice of the bandwidths. As a first approximation cross-validation seems to be acceptable even when we know that it is not consistent. However, there is a potential for a non-negligible improvement. We have naively implemented under-smoothing by dividing the cross-validated bandwidths by 2 and 4, respectively. This rudimentary method delivers gains of 20% to 40% in MSE, showing that further research in this direction is worthwhile.

5.2 JTPA training programs

The impact of training programs on the earnings of participants is of great interest to economists, but its estimation is difficult because of the self-selection of the treatment status. A randomized training experiment conducted under the Job Training Partnership Act (JTPA) provides exogenous variation for addressing this issue. In this experiment, people were randomly offered training but they were able to refuse to participate. Therefore, the treatment was self-selected and potentially correlated with the potential outcome, the 30-month earnings data, but the treatment assignment provides a credible instrument for the treatment participants. In this sub-section we use the men subsample of the data of AAI to illustrate the estimation of unconditional QTE of JTPA training programs.

Since the instrument has been completely randomized, we do not need to condition on any covariates to ensure the validity of the exclusion restriction. We can therefore use the estimator proposed by AAI to estimate the unconditional QTE if we do not include the regressors X in the weighted quantile regression. However, Theorem 11 implies that we can improve efficiency by incorporating the information contained in the observed characteristics of the trainees. Therefore, we expect both estimators to be consistent but the second one to be more precise, although the efficiency gains might be small.

The AAI estimator is implemented exactly as in the original paper with the exception that we regress Y only on a constant and D.[22] In order to simplify the comparison, we apply the same estimation strategy as AAI but replace their weights by W^+ defined in (11). We estimate $p(X)$ by a linear regression of Z on X while the projection on Y and D is estimated using the same series approximation as in AAI. Figure 1 shows that the point estimates are, as expected, very similar and the difference between both estimators is never significant. Figure 2 plots the relative efficiency gains of the suggested estimator. We have bootstrapped the results 1000 times and then compared the standard errors of the estimated QTE and the length of a 95% percentile confidence interval. Values below 1 indicate a reduction in standard deviation or a shortening of the confidence interval. An efficiency gain of about 5% in the center of the distribution confirms the asymptotic results but also shows that the gains are modest in this application.

Naturally, the main motivation for our estimator does not consist in this gain but in the possibility of allowing Z to be a proper instrumental variable only after conditioning on some covariates X. In order to simulate this case, we now manipulate the value of Z such that Z becomes correlated with the potential outcome if we do not control for the covariates but is still a valid instrument after controlling for X. To keep this manipulation simple, we set the value of the corrupted instrument to 1 with probability 0.5 for married men having $Z = 0$. Then, we use the same procedure as before but with the corrupted Z as instrument. To eliminate the additional random component added by this manipulation we repeat it 1000 times and present the mean estimates in Figure 3.

The bias arising from not controlling for the covariates is very clear. Independently of their treatment status married men tend to have higher earnings than non-married men. Since the instrument is positively correlated with the marriage dummy, a positive bias can be observed at all quantiles. On the other hand, when we control for the covariates, the bias disappears and we get almost the same result as that obtained from the original data. This artificial exercise illustrates what can happen in many applications, as we will see now in the next sub-section.

5.3 Returns to college

The returns to education have received a lot of attention with recent research interests aiming also particularly at higher education (see e.g. Black and Smith (2004 and 2005) about the

[22] We can replicate their results if we add X to the regressors.

returns to college). In this section we apply the new quantile estimators to estimate the *returns to college* using *college proximity* as an instrument. The data is taken from Card (1995), who found that the 2SLS estimates of the returns to schooling were about 13% and thus twice as large as the corresponding OLS estimates. Here, we focus particularly on the treatment effect of having attended college. The data stems from the National Longitudinal Survey of Young Men (NLSYM), which began in 1966 with 5525 men between 14 to 24 years old. The sampling frame of the NLSYM oversampled neighborhoods with a large fraction of non-white residents.

We follow Card (1995) in that we examine wages in the year 1976 to mitigate the influence of attrition. About 20% of the sample attrited in the first three years of the survey and the total attrition rate was about 29% in 1976. Total attrition increased further to 35% until the 1981 wave. In 1976 the respondents were between 24 to 34 years old such that most of them should have completed college at that time. Eighty-three percent of men interviewed in 1976 report a valid wage observation. As pointed out in Card (1995), the characteristics of this working subsample are relatively similar to the original sample, the subsample of 1976 interviewees and the subsample with 1976 wages. Most noticeable is a smaller fraction of blacks. Therefore, we work with the same sample as in the original paper. Descriptive statistics can be found in the original paper.

The variable of interest Y is the log hourly wage in 1976[23], measured in dollars per hour. The binary indicator D having *attended college* is also taken from the 1976 wave. We define D as one if years of education is $\geq$12. About 50% have attended college, while the other 50% did not. The instrument is an indicator for the presence of an accredited 4-year college in the local labor market. Almost 68% of the observations were living in such neighborhoods in 1966. Our vector of covariates includes potential experience, race, and regions of residence taken from the 1966 and 1976 waves. We also include variables capturing the endogenous location choice of the parents: measures of parental education, interactions of mother's and father's education classes, and indicators for family structures at age 14.

As a first step to estimate the QTE of college attendance on wages, we examine the relationship between the instrumental variable Z and other background characteristics X that are likely to have a strong influence on earnings in 1976. Table 1 shows a probit regression of Z on X and

[23]The role of a monotone transformation of the dependent variable is completely transparent in the quantile setting, where $F^{-1}_{h(Y)}(\tau) = h\left(F^{-1}_{Y}(\tau)\right)$. We choose to use the log transformation in order to simplify the comparison with the existing literature.

Figure 4 shows the kernel density estimates of the distribution of $\Pr(Z = 1|X)$ in the $Z = 1$ and $Z = 0$ subpopulations. This figure shows that those individuals living near to a college ($Z = 1$) and those with $Z = 0$ do indeed seem to differ with respect to their family characteristics X. On the other hand, there does not seem to be a problem with respect to the common support since the support of $p(x)$ is rather similar in these two subpopulations.

For illustrative purposes we used a parametric estimate of $p(x)$ for Table 1 and Figure 4. For the following estimates of the QTE we use nonparametric regression. Our vector of covariates contains 3 continuous variables, 1 unordered variable and 8 indicator variables for 3010 observations. Given this large number of discrete variables we follow the suggestion by Racine and Li (2004) to also smooth over the dummy variables to improve precision in small samples. A product Epanechnikov kernel is used for the continuous and ordered variables whereas the dummy variables enter multiplicatively with a weight of one if the variable is identical to the evaluation point and a weight smaller than one otherwise. All smoothing parameters have been chosen by cross-validation.[24]

Figure 5 compares the returns to college assuming exogeneity with the returns obtained using college proximity as an instrument. The unconditional QTE assuming exogeneity are quite stable along the distribution and amount to about 20%. The instrumental variable strategy produces higher estimates of the return to college, especially on the lower part of the distribution. Note however that we must be cautious since the standard errors, not plotted to avoid overloading the figure, are quite high such that, as in Card (1995), we cannot reject that college is exogenous because both confidence intervals overlaps at all quantiles.

Figure 6 shows how important it is to incorporate covariates in the estimation. If we assume that the IV is valid without conditioning on X, then the estimated QTE attain incredibly high values (about 100%). Maybe contrarily to what could have been expected, Figure 6 also shows that controlling for family background is not determinant since the estimates do not really change when we exclude them from the conditioning set. The regional variables are much more important in this respect and the estimated returns to college are not credible when we exclude them.

[24] This is *not* a consistent way to estimate the optimal bandwidths. In the absence of a consistent method, we consider cross-validation as a reasonable second-best solution and leave this problem for further research.

6 Conclusions

This paper considers nonparametric identification and efficient estimation of unconditional quantile treatment effects under endogeneity. Unconditional QTE summarize the distributional effects for the whole population. They are easy to convey and can be estimated precisely even without parametric assumptions. We allow for endogeneity because the variables of interest are often self-selected in observational studies, making conventional methods inconsistent for the causal effects of these variables. In the spirit of Imbens and Angrist (1994), identification is based on the presence of an instrument satisfying a monotonicity assumption in the treatment choice equation. We only require the instruments to satisfy an exclusion restriction conditionally on the covariates, i.e. we allow the instrument to be itself unconditionally confounded. Incorporating covariates in the estimation makes therefore the IV assumption more credible in many applications.

We show that the unconditional QTE are nonparametrically identified in this setting. Natural estimators arise from the constructive identification results. We emphasize interesting links with the existing literature. On one side, we show that our estimators generalize all existing estimators for exogenous QTE. On the other hand, our weighting estimator simplifies to the Abadie, Angrist and Imbens (2002) estimator when there are no covariates. We show root n consistency and asymptotic normality of a weighting estimator, which is easy to implement. The derivation of the semiparametric efficiency bound for this model allows us to show efficiency for the suggested estimator. Interestingly, we also find that the semiparametric efficiency bound does not decrease when the propensity score is known, which is often the case when the instrument is under the control of the institution managing the experiment. On the other hand, including additional variables not needed for consistency decreases the asymptotic variance. Therefore, we may include some covariates for consistency and other covariates for efficiency.

The proposed estimators are easy to implement[25] and have numerous potential applications, as illustrated by the two empirical examples we presented. We first used our estimators to evaluate the effects of JTPA training programs using the variation induced by a randomized experiment. We can quantify the precision gained by including covariates in the estimation since the instrument has been completely randomized and is therefore valid unconditionally. In a second application, we estimated the returns to college using college proximity as an instrument. Controlling for covariates is critical in this case because the location choice made by the parents

[25] Moreover, user-friendly codes written in Stata and R can be obtained from the authors.

can be unconditionally correlated with unobserved characteristics. If we do not incorporate covariates, the returns to college attain incredibly high values when education is instrumentalized. Once we include covariates, the returns to college become plausible, and they are higher than under the assumption of exogeneity, as often found in the literature.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press, Cambridge, Mass.

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

Begun, J., W. Hall, W. Huang, and J. Wellner (1983): "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *Annals of Statistics*, 11, 432–452.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. John Hopkins University Press, Baltimore.

Black, D., and J. Smith (2004): "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching," *Journal of Econometrics*, 121, 99–124.

——— (2005): "Estimating the returns to college quality with multiple proxies for quality," *Journal of Labor Economics*, 24, 701–728.

Buchinsky, M. (1994): "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," *Econometrica*, 62, 405–458.

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. Christofides, E. Grant, and R. Swidinsky, pp. 201–222. University of Toronto Press, Toronto.

Carroll, R., D. Ruppert, and A. Welsh (1998): "Local Estimating Equations," *Journal of American Statistical Association*, 93, 214–227.

Chaudhuri, P. (1991): "Global nonparametric estimation of conditional quantile functions and their derivatives," *Journal of Multivariate Analysis*, 39, 246–269.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2007): "Inference on Counterfactual Distributions," mimeo.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

——— (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132, 491–525.

——— (2007): "Instrumental quantile regression: A robust inference approach," *Journal of Econometrics*, forthcoming.

Chernozhukov, V., G. Imbens, and W. Newey (2007): "Instrumental variable estimation of nonseparable models," *Journal of Econometrics*, 139, 4–14.

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

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

——— (2007): "Endogeneity and discrete outcomes," mimeo.

Fan, J. (1993): "Local Linear Regression Smoothers and their Minimax Efficiency," *Annals of Statistics*, 21, 196–216.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modeling and its Applications*. Chapman and Hall, London.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75, 259–276.

Flossmann, A. (2007): "Empirical Bias Bandwidth Choice for Local Polynomial Matching Estimators," Working paper University of Konstanz.

Frölich, M. (2004): "A note on the role of the propensity score for estimating average treatment effects," *Econometric Reviews*, 23, 167–174.

——— (2006): "A Note on Parametric and Nonparametric Regression in the Presence of Endogenous Control Variables," *IZA Discussion Paper*, 2126.

——— (2007a): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

——— (2007b): "Propensity score matching without conditional independence assumption - with an application to the gender wage gap in the UK," *Econometrics Journal*.

Gosling, A., S. Machin, and C. Meghir (2000): "The Changing Distribution of Male Wages in the U.K.," *Review of Economics Studies*, 67, 635–666.

Gozalo, P., and O. Linton (2000): "Local Nonlinear Least Squares: Using parametric information in nonparametric regression," *Journal of Econometrics*, 99, 63–106.

Guntenbrunner, C., and J. Jurečková (1992): "Regression Quantile and Regression Rank Score Process in the Linear Model and Derived Statistics," *Annals of Statistics*, 20, 305–330.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Hall, P., R. C. L. Wolff, and Q. Yao (1999): "Methods for estimating a conditional distribution function," *Journal of American Statistical Association*, 94(445), 154–163.

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

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Hjort, N., and D. Pollard (1993): "Asymptotics for minimisers of convex processes," Statistical Research Report, Department of Mathematics, University of Oslo.

Hoderlein, S., and E. Mammen (2007): "Identification of marginal effects in nonseparable models without monotonicity," *Econometrica*, 75, 1513–1518.

Hong, H., and D. Nekipelov (2007): "Semiparametric Efficiency in Nonlinear LATE Models," mimeo.

Horowitz, J., and S. Lee (2007): "Nonparametric instrumental variables estimation of a quantile regression model," *Econometrica*, 75, 1191–1208.

Imbens, G., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Imbens, G., and W. Newey (2003): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," presented at the EC2 conference London December 2003.

Koenker, R. (2005): *Quantile Regression*. Cambridge University Press, Cambridge.

Koenker, R., and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R., and Z. Xiao (2002): "Inference on the Quantile Regression Process," *Econometrica*, 70, 1583–1612.

Koshevnik, Y., and B. Levit (1976): "On a Non-parametric Analogue of the Information Matrix," *Theory of Probability and Applications*, 21, 738–753.

Machado, J., and J. Mata (2005): "Counterfactual decomposition of changes in wage distributions using quantile regression," *Journal of Applied Econometrics*, 20, 445–465.

Masry, E. (1996): "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis*, 17, 571–599.

Melly, B. (2006): "Estimation of counterfactual distribution using quantile regression," mimeo.

Newey, W. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Nielsen, S. (2005): "Local linear estimating equations: uniform consistency and rate of convergence," *Nonparametric Statistics*, 17, 493–511.

Pfanzagl, J., and W. Wefelmeyer (1982): *Contributions to a General Asymptotic Statistical Theory*. Springer Verlag, Heidelberg.

Powell, J. (1986): "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155.

Powell, J., J. Stock, and T. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

Racine, J., and Q. Li (2004): "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," *Journal of Econometrics*, 119, 99–130.

Ruppert, D. (1997): "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of American Statistical Association*, 92, 1049–1062.

Ruppert, D., and M. Wand (1994): "Multivariate Locally Weighted Least Squares Regression," *Annals of Statistics*, 22, 1346–1370.

Serfling, R. (1980): *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Stein, C. (1956): "Efficient Nonparametric Testing and Estimation," in *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press, Berkeley.

A Appendix (Proof of theorems)

A.1 Proof of Theorem (1):

We consider the derivation of

$$F_{Y^0|c}(u) = \frac{\int \left(E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=1\right] - E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=0\right]\right) dF_X}{\int \left(E\left[D|X, Z=1\right] - E\left[D|X, Z=0\right]\right) dF_X}. \tag{16}$$

(The results for $F_{Y^1|c}$ are analogous and are omitted.)

Consider first the expression

$$E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=1\right] - E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=0\right]$$

which by the law of total probability can be partitioned into the four subpopulations:

$$\begin{aligned}
&= E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=1, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a|X, Z=1\right) \\
&\qquad - E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=0, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a|X, Z=0\right) \\
&+ E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=1, \mathcal{T}=n\right] \Pr\left(\mathcal{T}=n|X, Z=1\right) \\
&\qquad - E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=0, \mathcal{T}=n\right] \Pr\left(\mathcal{T}=n|X, Z=0\right) \\
&+ E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=1, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c|X, Z=1\right) \\
&\qquad - E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=0, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c|X, Z=0\right).
\end{aligned}$$

Noting that the value of Z and $\mathcal{T}$ together determine the value of D and using that $\mathcal{T} \perp\!\!\!\perp Z|X$ from assumption 1, we obtain

$$\begin{aligned}
&= 0 \cdot \Pr\left(\mathcal{T}=a|X\right) \\
&- \left\{E\left[1\left(Y^0 \leq u\right)|X, Z=1, \mathcal{T}=n\right] - E\left[1\left(Y^0 \leq u\right)|X, Z=0, \mathcal{T}=n\right]\right\} \Pr\left(\mathcal{T}=n|X\right) \\
&+ \left\{0 + E\left[1\left(Y^0 \leq u\right)|X, Z=0, \mathcal{T}=c\right]\right\} \Pr\left(\mathcal{T}=c|X\right).
\end{aligned}$$

Now we use $Y^d \perp\!\!\!\perp Z|X, \mathcal{T}$ by assumption 1 to obtain

$$\begin{aligned}
&= - \left\{E\left[1\left(Y^0 \leq u\right)|X, \mathcal{T}=n\right] - E\left[1\left(Y^0 \leq u\right)|X, \mathcal{T}=n\right]\right\} \Pr\left(\mathcal{T}=n|X\right) \\
&+ \left\{E\left[1\left(Y^0 \leq u\right)|X, \mathcal{T}=c\right]\right\} \Pr\left(\mathcal{T}=c|X\right) \\
&= E\left[1\left(Y^0 \leq u\right)|X, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c|X\right).
\end{aligned}$$

Now we insert this result into the numerator of (16) to obtain

$$\begin{aligned}
&\int \left(E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=1\right] - E\left[1\left(Y \leq u\right) \cdot (D-1)|X, Z=0\right]\right) dF_X \\
&= \int E\left[1\left(Y^0 \leq u\right)|X, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c|X\right) dF_X \\
&= \int E\left[1\left(Y^0 \leq u\right)|X, \mathcal{T}=c\right] dF_{X|c} \cdot P_c \\
&= E\left[1\left(Y^0 \leq u\right)|\mathcal{T}=c\right] \cdot P_c
\end{aligned}$$

where the second last equality made use of Bayes' theorem:

$$dF_{X|c} \cdot P_c = \Pr\left(\mathcal{T}=c|X\right) \cdot dF_X.$$

Now consider the denominator of (16) and proceed as before. First notice that conditional on X

$$E\left[D|X, Z=1\right] - E\left[D|X, Z=0\right]$$

$$\begin{aligned}
&= E\left[D|X, Z=1, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a|X, Z=1\right) - E\left[D|X, Z=0, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a|X, Z=0\right) \\
&+ E\left[D|X, Z=1, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c|X, Z=1\right) - E\left[D|X, Z=0, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c|X, Z=0\right)
\end{aligned}$$

using that $\mathcal{T} \perp\!\!\!\perp Z|X$ from assumption 1, we obtain

$$\begin{aligned}
&= \Pr\left(\mathcal{T}=a|X\right) - \Pr\left(\mathcal{T}=a|X\right) + \Pr\left(\mathcal{T}=c|X\right) \\
&= \Pr\left(\mathcal{T}=c|X\right).
\end{aligned}$$

Inserting this into the denominator of (16) and again making use of Bayes' theorem

$$\int \Pr\left(\mathcal{T}=c|X\right) dF_X = \int dF_{X|c} \cdot P_c = P_c.$$

Putting these results together we obtain for the right hand side of (16)

$$\frac{E\left[1\left(Y^0 \leq u\right)|\mathcal{T}=c\right] \cdot P_c}{P_c} = \Pr\left(Y^0 \leq u|\mathcal{T}=c\right) = F_{Y^0|c}(u).$$

A.2 Proof of Lemma (2):

The following two variants of using iterated expectations show the equality for a typical component of the estimator

$$E\left[\frac{1\left(Y \leq u\right) DZ}{p(X)}\right] = \int E\left[\frac{1\left(Y \leq u\right) DZ}{p(X)}|X\right] dF_X = \int E\left[1\left(Y \leq u\right) D|X, Z=1\right] dF_X$$

$$E\left[\frac{1\left(Y\leq u\right)DZ}{p(X)}\right]=E\left[E\left[\frac{1\left(Y\leq u\right)DZ}{p(X)}|p(X)=\rho\right]\right]=E\left[E\left[\rho\frac{1\left(Y\leq u\right)D}{p(X)}|p(X)=\rho,Z=1\right]\right]$$
$$=E\left[E\left[1\left(Y\leq u\right)D|p(X)=\rho,Z=1\right]\right].$$

For the corresponding components in the $Z=0$ population, Z is replaced by $1-Z$ and $p(X)$ is replaced by $1-p(X)$ in the previous derivations.

A.3 Proof of Lemma (3):

Note that by iterated expectations

$$E\left[\frac{1\left(Y\leq u\right)DZ}{p(X)}\right]=\int E\left[\frac{1\left(Y\leq u\right)DZ}{p(X)}|X\right]dF_X=\int E\left[1\left(Y\leq u\right)D|X,Z=1\right]dF_X$$

and

$$E\left[\frac{1\left(Y\leq u\right)D(1-Z)}{1-p(X)}\right]=\int E\left[\frac{1\left(Y\leq u\right)D(1-Z)}{1-p(X)}|X\right]dF_X=\int E\left[1\left(Y\leq u\right)D|X,Z=0\right]dF_X$$

Hence, the equation (3) can be written as

$$F_{Y^1|c}(u)=\frac{\int\left(E\left[1\left(Y\leq u\right)D\left(\frac{Z-p(X)}{p(X)(1-p(X))}\right)\right]\right)dF_X}{P_c}$$

and analogously for $F_{Y^0|c}(u)$.

A.4 Proof of Lemma (4):

If the objective function has a unique interior solution, it follows that

$$\arg\min_{a,b}E\left[W\cdot\rho_\tau(Y-a-bD)\right] \tag{17}$$
$$=\arg\underset{a,b}{zeroE}\left[W\cdot\left\{\tau-1\left(Y<a+bD\right)\right\}\cdot\begin{pmatrix}1\\D\end{pmatrix}\right]. \tag{18}$$

This implies the moment conditions:

$$E\left[\frac{Z-p(X)}{p(X)\left(1-p(X)\right)}\left(2D-1\right)\left\{\tau-1\left(Y<a+bD\right)\right\}\right]=0$$
$$E\left[\frac{Z-p(X)}{p(X)\left(1-p(X)\right)}\left(2D-1\right)\left\{\tau-1\left(Y<a+bD\right)\right\}\cdot D\right]=0.$$

Multiplying the first moment condition with $(D+(1-D))$ inside the expectation operator and inserting the second moment condition gives:

$$E\left[\frac{Z-p(X)}{p(X)\left(1-p(X)\right)}\left(2D-1\right)\left\{\tau-1\left(Y<a+bD\right)\right\}\cdot(1-D)\right]=0$$
$$E\left[\frac{Z-p(X)}{p(X)\left(1-p(X)\right)}\left(2D-1\right)\left\{\tau-1\left(Y<a+bD\right)\right\}\cdot D\right]=0$$

which is equivalent to

$$E\left[\frac{Z-p(X)}{p(X)\left(1-p(X)\right)}\left\{\tau-1\left(Y<a\right)\right\}\cdot\left(1-D\right)\right]=0$$
$$E\left[\frac{Z-p(X)}{p(X)\left(1-p(X)\right)}\left\{\tau-1\left(Y<a+b\right)\right\}\cdot D\right]=0.$$

Renaming a with q_0 and $a+b$ with q_1 and subtracting the term $E\left[\tau\frac{Z-p(X)}{p(X)(1-p(X))}\right]$, which is zero, from the first moment condition gives

$$E\left[\frac{Z-p(X)}{p(X)\left(1-p(X)\right)}\left\{1\left(Y<q_0\right)\left(D-1\right)-\tau D)\right\}\right]=0$$
$$E\left[\frac{Z-p(X)}{p(X)\left(1-p(X)\right)}\left\{\tau-1\left(Y<q_1\right)\right\}\cdot D\right]=0,$$

which are identical to (7).

A.5 Proof of Proposition (5)

Replicating the previous proofs in reverse order, one can first show that the first order conditions to

$$\arg\min_{a,b}E\left[\left(1-\frac{D(1-Z)}{1-p(X)}-\frac{(1-D)Z}{p(X)}\right)\rho_\tau(Y-\alpha-\beta D)\right]$$

are:

$$E\left[D\left(\frac{Z-p(X)}{1-p(X)}\right)1\left(Y<\alpha+\beta\right)\right]=\tau\cdot P(T=c,D=1) \tag{19}$$
$$E\left[(D-1)\left(\frac{Z-p(X)}{p(X)}\right)1\left(Y<\alpha\right)\right]=\tau\cdot P(T=c,D=0).$$

where $\Pr(T=c,D=1)=\Pr(D=1|T=c)\Pr(T=c)=E\left[D\frac{Z-p(X)}{1-p(X)}\right]$ is the fraction of 'treated compliers' and $\Pr(T=c,D=0)=\Pr(D=0|T=c)\Pr(T=c)=E\left[(D-1)\frac{Z-p(X)}{p(X)}\right]$ is the fraction of 'non-treated compliers'. (Since the proof is very similar to the previous one it is ommitted.)

Define the distributions of the potential outcomes for *treated compliers* and *non-treated compliers* as

$$F_{Y^1|c,D=1}(u)=\Pr(Y^1\leq u|D=1,T=c)$$
$$F_{Y^0|c,D=0}(u)=\Pr(Y^0\leq u|D=0,T=c).$$

Analogously to the previous proofs one can show that these distributions are identified as

$$F_{Y^1|c,D=1}(u)=\frac{E\left[1\left(Y<u\right)\cdot D\cdot\frac{Z-p(X)}{1-p(X)}\right]}{\Pr(T=c,D=1)}$$
$$F_{Y^0|c,D=0}(u)=\frac{E\left[1\left(Y<u\right)\cdot(D-1)\cdot\frac{Z-p(X)}{p(X)}\right]}{\Pr(T=c,D=0)}.$$

Hence, $\alpha+\beta$ and α in (19) define the quantiles in the sense that $F_{Y^1|c,D=1}(\alpha_0+\beta_0)=\tau=F_{Y^0|c,D=0}(\alpha_0)$. This implies then that $F^{-1}_{Y^1|c,D=1}(\tau)=\alpha_0+\beta_0$ and $F^{-1}_{Y^0|c,D=0}(\tau)=\alpha_0$ and that

$$\beta_0=F^{-1}_{Y^1|c,D=1}(\tau)-F^{-1}_{Y^0|c,D=0}(\tau).$$

A.6 Proof of Proposition (6):

We show that $E\left[\rho_\tau(Y - X'b - aD) \cdot W\right]$ has expectation zero in the subpopulation of always- and never-participants, for every value of a and b. Note first that

$$\begin{aligned}
E\left[\rho_\tau(Y - X'\beta - \alpha D)W \,|\mathcal{T} = a\right] &= E\left[\rho_\tau(Y - X'\beta - \alpha D)\frac{Z - p\left(X\right)}{p(X)\left(1 - p(X)\right)}\left(2D - 1\right)|\mathcal{T} = a\right] \\
&= E\left[E\left[\rho_\tau(Y^1 - X'\beta - \alpha)\frac{Z - p\left(X\right)}{p(X)\left(1 - p(X)\right)}|X, Z, \mathcal{T} = a\right]|\mathcal{T} = a\right] \\
&= E\left[E\left[\rho_\tau(Y^1 - X'\beta - \alpha)|X, \mathcal{T} = a\right]\frac{E\left[Z|X, T = a\right] - p\left(X\right)}{p(X)\left(1 - p(X)\right)}|\mathcal{T} = a\right] \\
&= E\left[E\left[\rho_\tau(Y^1 - X'\beta - \alpha)|X, \mathcal{T} = a\right]\frac{E\left[Z|X\right] - p\left(X\right)}{p(X)\left(1 - p(X)\right)}|\mathcal{T} = a\right] \\
&= 0.
\end{aligned}$$

where $Y^d \perp\!\!\!\perp Z|X, \mathcal{T}$ and $\mathcal{T} \perp\!\!\!\perp Z|X$ has been used which follows from Assumption 1. The same result holds for the never-taker. Therefore,

$$\begin{aligned}
(\alpha, \beta) &= \arg\min_{a,b} E\left[\rho_\tau(Y - X'b - aD) \cdot W \,|\mathcal{T} = c\right] \\
&= \arg\min_{a,b} E\left[\left(\frac{D}{p(X)} + \frac{1 - D}{1 - p(X)}\right) \cdot \rho_\tau(Y - X'\beta - \alpha D) \;\; |\mathcal{T} = c\right],
\end{aligned}$$

which is the objective function of a weighted linear quantile regression for compliers. Note that all weights are strictly positive and finite because we assume that $0 < p(X) < 1$. Therefore, standard quantile regression results (see for instance Koenker (2005) Theorem 4.1 and 5.1) imply that this function is minimized at α_0^τ and β_0^τ as long as $E\left[\left(\frac{D}{p(X)} + \frac{1-D}{1-p(X)}\right)(X', D)'\,(X', D)\right]$ is positive definite, which has been assumed.

A.7 Proof of Theorem (7)

$\hat{\Delta}_c^\tau$ is the value of b that solves

$$\arg\min_{a,b} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - a - bD_i) \cdot \hat{W}_i. \tag{20}$$

Since D_i takes only two different values it is more convenient in the following to use the numerically equivalent representation:

$$\hat{\Delta}_c^\tau = \hat{q}_1 - \hat{q}_0 \qquad \text{where } (\hat{q}_1, \hat{q}_0) = \arg\min_{q_1, q_0} \left(\frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - q_1) \cdot D_i \hat{W}_i + \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - q_0) \cdot (1 - D_i)\,\hat{W}_i\right). \tag{21}$$

Here q_1 delivers an estimate of $Q^\tau_{Y^1|c}$ and q_0 an estimate of $Q^\tau_{Y^0|c}$. We will derive the joint asymptotic distribution of $\hat{q}_1$ and $\hat{q}_0$ which provides the distribution of $\hat{\Delta}_c^\tau$.

Define the objective function

$$G_n(q_1, q_0, \hat{W}) = \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i \left(\rho_\tau(Y_i - q_1)D_i + (1 - D_i)\,\rho_\tau(Y_i - q_0)\right) \\ - \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i \left(\rho_\tau(Y_i - Q^\tau_{Y^1|c})D_i + (1 - D_i)\,\rho_\tau(Y_i - Q^\tau_{Y^0|c})\right),$$

where the first term is identical to (21) and the second term does neither depend on q_1 nor q_0. Hence, the minimizers of G_n and of (21) are identical:

$$(\hat{q}_1, \hat{q}_0) = \arg\min_{q_1, q_0} G_n(q_1, q_0, \hat{W}). \tag{22}$$

The function G_n will be helpful later to derive the properties of $\hat{q}_1$ and $\hat{q}_0$.

A.7.1 Analysis of the approximate gradient

As a preliminary step we examine the properties of the approximate gradient. Instead of minimizing the objective function $G_n(q_1, q_0, \hat{W})$ we could alternatively consider the estimator which sets the moment function to zero

$$\hat{q}_1 = \arg zero \frac{1}{n}\sum_{i=1}^{n}(1(Y_i < q_1) - \tau)\cdot \hat{W}_i D_i,$$

and analogously for q_0. This is the approximate gradient of the objective function. We will first inspect the properties of this estimator and thereafter examine the estimator based on (22). These preliminary derivations are only casual to obtain a potential candidate for the influence function representation. Define the objective function $\Upsilon_n(q, W) = \frac{1}{n}\sum (1(Y_i < q) - \tau)\cdot W_i D_i$ where $\Upsilon_n(\hat{q}_1, \hat{W}) = 0$ and use a Taylor series expansion to obtain

$$0 = \Upsilon_n(\hat{q}_1, \hat{W}) = \Upsilon_n(Q^\tau_{Y^1|c}, \hat{W}) + (\hat{q}_1 - Q^\tau_{Y^1|c})\cdot \frac{\partial \Upsilon_n(q, \hat{W})}{\partial q}_{|q=Q^\tau_{Y^1|c}} + O_p\left((\hat{q}_1 - Q^\tau_{Y^1|c})^2\right).$$

The derivative is not everywhere defined but almost surely. One could thus approximate $\hat{q}_1$ as

$$\begin{aligned}\sqrt{n}(\hat{q}_1 - Q^\tau_{Y^1|c}) &= -\sqrt{n}\left[\frac{\partial \Upsilon_n(Q^\tau_{Y^1|c}, \hat{W})}{\partial q} + O_p\left(\hat{q}_1 - Q^\tau_{Y^1|c}\right)\right]^{-1}\Upsilon_n(Q^\tau_{Y^1|c}, \hat{W}) \\ &= -\sqrt{n}\frac{\Upsilon_n(Q^\tau_{Y^1|c}, \hat{W})}{P_c \cdot f_{Y^1|c}(Q^\tau_{Y^1|c})}\cdot(1 + o_p(1)) \end{aligned} \tag{23}$$

where we used that under certain regularity conditions $\partial \Upsilon_n(Q^\tau_{Y^1|c}, \hat{W})/\partial q$ converges to $\partial E[\Upsilon_n(Q^\tau_{Y^1|c}, W)]/\partial q = P_c \cdot f_{Y^1|c}(Q^\tau_{Y^1|c})$ where the last expression follows from Theorem 1 where we derived that $f_{Y^1|c}(u) = \left\{\int \left(f_{Y|X,D=1,Z=1}(u)\pi(x,1) - f_{Y|X,D=1,Z=0}(u)\pi(x,0)\right) dF_X\right\}/P_c$.

We will see below that the stochastic properties of the estimator are largely driven by the term $\Upsilon_n(Q^\tau_{Y^1|c}, \hat{W})$, which is the approximate gradient of the objective function G_n for $\hat{q}_1$ at $Q^\tau_{Y^1|c}$. (The terms

for $\hat{q}_0$ are analogous and are omitted here.) To lighten the notation we use $p_i = p(X_i)$ and $\hat{p}_i = \hat{p}(X_i)$ and $\pi_i(z) = \pi(X_i, z)$.

$$\begin{aligned}\Upsilon_n(Q^\tau_{Y^1|c}, \hat{W}) &= \frac{1}{n}\sum_{i=1}^n \hat{W}_i D_i \left(1\left(Y_i < Q^\tau_{Y^1|c}\right) - \tau\right) \\ &= \frac{1}{n}\sum_{i=1}^n \left(\frac{Z_i D_i}{\hat{p}_i} - \frac{(1-Z_i)D_i}{1-\hat{p}_i}\right)\left(1\left(Y_i < Q^\tau_{Y^1|c}\right) - \tau\right)\end{aligned}$$

where we used that $\hat{W} = \frac{ZD}{\hat{p}(X)} - \frac{(1-Z)D}{1-\hat{p}(X)} - \frac{Z(1-D)}{\hat{p}(X)} + \frac{(1-Z)(1-D)}{1-\hat{p}(X)}$. Now we use that $\frac{\hat{p}_i - p_i}{p_i^2}\left(1 - \frac{\hat{p}_i - p_i}{\hat{p}_i}\right) = \frac{1}{p_i} - \frac{1}{\hat{p}_i}$ to obtain

$$\begin{aligned}&= \frac{1}{n}\sum_{i=1}^n \left(\frac{Z_i D_i}{p_i} - \frac{(1-Z_i)D_i}{1-p_i}\right)\left(1\left(Y_i < Q^\tau_{Y^1|c}\right) - \tau\right) \\ &\quad - \frac{1}{n}\sum_{i=1}^n \frac{Z_i D_i}{p_i^2}\left(1\left(Y_i < Q^\tau_{Y^1|c}\right) - \tau\right)(\hat{p}_i - p_i)\left(1 - \frac{\hat{p}_i - p_i}{\hat{p}_i}\right) \\ &\quad - \frac{1}{n}\sum_{i=1}^n \frac{(1-Z_i)D_i}{(1-p_i)^2}\left(1\left(Y_i < Q^\tau_{Y^1|c}\right) - \tau\right)(\hat{p}_i - p_i)\left(1 + \frac{\hat{p}_i - p_i}{1-\hat{p}_i}\right).\end{aligned}$$

The first term captures the variance contribution due to the weighting if the weights were known. The second and third term capture the variance due to estimating the weights. From (45) and (46) and analogous derivations for the third term we obtain

$$\begin{aligned}&= \frac{1}{n}\sum_{i=1}^n \left(\frac{Z_i D_i}{p_i} - \frac{(1-Z_i)D_i}{1-p_i}\right)\left(1\left(Y_i < Q^\tau_{Y^1|c}\right) - \tau\right) \\ &- \frac{1}{n}\sum_{i=1}^n \frac{\pi(X_i,1)\vartheta_{11}(X_i)\,(Z_i - p_i)}{p\,(X_i)}\,(1 + o_p(1)) + o_p\left(\frac{1}{\sqrt{n}}\right) \\ &- \frac{1}{n}\sum_{i=1}^n \frac{\pi(X_i,0)\vartheta_{10}(X_i)\,(Z_i - p_i)}{1 - p\,(X_i)}\,(1 + o_p(1)) + o_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

and after some derivations

$$= \frac{1}{n}\sum_{i=1}^n \psi_1(Y_i, D_i, Z_i, X_i)\,(1 + o_p(1)) + o_p\left(\frac{1}{\sqrt{n}}\right)$$

where

$$\begin{aligned}\psi_1(Y_i, D_i, Z_i, X_i) &= \frac{Z_i D_i}{p_i}\chi_{11}(Y_i, X_i) - \frac{(1-Z_i)D_i}{1-p_i}\chi_{10}(Y_i, X_i) \\ &\quad + \frac{Z_i D_i - \pi(X_i,1)\,(Z_i - p_i)}{p\,(X_i)}\vartheta_{11}(X_i) - \frac{\pi(X_i,0)\,(Z_i - p_i) + (1-Z_i)D_i}{1 - p\,(X_i)}\vartheta_{10}(X_i) \quad (24)\end{aligned}$$

and

$$\chi_{dz}(y, x) = \left(1(y \le Q^\tau_{Y^d|c}) - \tau\right) - \vartheta_{dz}(x) \quad (25)$$

and

$$\vartheta_{dz}(x) = E\left[1(Y \le Q^\tau_{Y^d|c}) - \tau | D = d, Z = z, X = x\right]. \quad (26)$$

(Notice that χ_{dz} and ϑ_{dz} are defined differently here than in (68) and (69) in that they are not divided by P_c times the density at the quantile.) By inserting this into (23) and using a CLT for iid data we obtain that to first order

$$\sqrt{n}(\hat{q}_1 - Q^\tau_{Y^1|c}) \xrightarrow{d} N\left(0, Var\left(\frac{\psi_1(Y_i, D_i, Z_i, X_i)}{P_c \cdot f_{Y^1|c}(Q^\tau_{Y^1|c})}\right)\right).$$

Analogously, we can derive

$$\begin{aligned}\psi_0(Y_i, D_i, Z_i, X_i) = &-\frac{Z_i(1-D_i)}{p_i}\chi_{01}(Y_i, X_i) + \frac{(1-Z_i)(1-D_i)}{1-p_i}\chi_{00}(Y_i, X_i) \\ &- \frac{Z_i(1-D_i) - (1-\pi(X_i,1))(Z_i - p_i)}{p(X_i)}\vartheta_{01}(X_i) + \frac{(1-\pi(X_i,0))(Z_i-p_i) + (1-Z_i)(1-D_i)}{1-p(X_i)}\vartheta_{00}(X_i)\end{aligned} \tag{27}$$

and

$$\sqrt{n}(\hat{q}_0 - Q^\tau_{Y^0|c}) \xrightarrow{d} N\left(0, Var\left(\frac{\psi_0(Y_i, D_i, Z_i, X_i)}{P_c \cdot f_{Y^0|c}(Q^\tau_{Y^1|c})}\right)\right).$$

Finally it follows that

$$\sqrt{n}(\hat{\Delta}^\tau_c - \Delta^\tau_c) \xrightarrow{d} N\left(0, Var\left(\frac{\psi_1(Y_i, D_i, Z_i, X_i)}{P_c \cdot f_{Y^1|c}(Q^\tau_{Y^1|c})} - \frac{\psi_0(Y_i, D_i, Z_i, X_i)}{P_c \cdot f_{Y^0|c}(Q^\tau_{Y^1|c})}\right)\right).$$

This asymptotic variance is identical to the variance bound (71).

A.7.2 Analysis of the non-differentiable objective function

With the preliminaries of the previous subsection, we now embark on proving the theorem. Since G_n is not differentiable everywhere, although almost everywhere, we extend our previous heuristic proof by examining the approximate derivative and showing that the approximation error vanishes. This proof is somewhat similar to Firpo (2007). Define the statistic

$$\begin{aligned}\tilde{G}_n(q_1, q_0) = &\left(q_1 - Q^\tau_{Y^1|c}\right)\frac{1}{n}\sum_{i=1}^n \psi_1(Y_i, D_i, Z_i, X_i) + \left(q_0 - Q^\tau_{Y^0|c}\right)\frac{1}{n}\sum_{i=1}^n \psi_0(Y_i, D_i, Z_i, X_i) \\ &+ \left(q_1 - Q^\tau_{Y^1|c}\right)^2 \frac{f_{Y^1|c}(Q^\tau_{Y^1|c}) \cdot P_c}{2} + \left(q_0 - Q^\tau_{Y^0|c}\right)^2 \frac{f_{Y^0|c}(Q^\tau_{Y^0|c}) \cdot P_c}{2}\end{aligned} \tag{28}$$

where ψ_1 and ψ_0 are defined in (24) and (27).

We will show in the following, first that the difference between the two objective functions $G_n(q_1, q_0, \hat{W})$ and $\tilde{G}_n(q_1, q_0)$, which are both convex functions in q_1 and q_0, vanishes for any q_1 and q_0 as $n \to \infty$. Second, we derive the asymptotic distribution of the minimizers of $\tilde{G}_n(q_1, q_0)$, which will give results identical to those of the previous subsection. Finally, we show that the minimizers of $G_n(q_1, q_0, \hat{W})$ and $\tilde{G}_n(q_1, q_0)$ get close to each other such that their first order asymptotic distribution is the same.

A.7.3 Closeness of the two objective functions $G_n(q_1, q_0, \hat{W})$ and $\tilde{G}_n(q_1, q_0)$:

First, we need to show that the two objective functions $G_n(q_1, q_0, \hat{W})$ and $\tilde{G}_n(q_1, q_0)$ are getting close as $n \to \infty$. For this it is helpful to add and subtract the terms

$$\frac{1}{n}\sum_{i=1}^{n}\hat{W}_i D_i \left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right)\left(q_1 - Q^{\tau}_{Y^1|c}\right) + \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i (1-D_i)\left(1\left(Y_i < Q^{\tau}_{Y^0|c}\right) - \tau\right)\left(q_0 - Q^{\tau}_{Y^0|c}\right),$$

which is the linear term of an approximate Taylor expansion, to $G_n(q_1, q_0, \hat{W})$ to obtain

$$\left\| G_n(q_1, q_0, \hat{W}) - \tilde{G}_n(q_1, q_0) \right\|$$

$$\begin{aligned}
= \; & \Big|\Big| \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i D_i \left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right)\left(q_1 - Q^{\tau}_{Y^1|c}\right) - \left(q_1 - Q^{\tau}_{Y^1|c}\right)\frac{1}{n}\sum_{i=1}^{n}\psi_1(Y_i, D_i, Z_i, X_i) \\
& + \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i (1-D_i)\left(1\left(Y_i < Q^{\tau}_{Y^0|c}\right) - \tau\right)\left(q_0 - Q^{\tau}_{Y^0|c}\right) - \left(q_0 - Q^{\tau}_{Y^0|c}\right)\frac{1}{n}\sum_{i=1}^{n}\psi_0(Y_i, D_i, Z_i, X_i) \\
& + \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i D_i \left(\left(\tau - 1\left(Y_i < Q^{\tau}_{Y^1|c}\right)\right)\left(q_1 - Q^{\tau}_{Y^1|c}\right) + \rho_\tau(Y_i - q_1) - \rho_\tau(Y_i - Q^{\tau}_{Y^1|c})\right) \\
& \qquad - \left(q_1 - Q^{\tau}_{Y^1|c}\right)^2 \frac{f_{Y^1|c}(Q^{\tau}_{Y^1|c}) \cdot P_c}{2} \\
& + \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i (1-D_i)\left(\left(\tau - 1\left(Y_i < Q^{\tau}_{Y^0|c}\right)\right)\left(q_0 - Q^{\tau}_{Y^0|c}\right) + \rho_\tau(Y_i - q_0) - \rho_\tau(Y_i - Q^{\tau}_{Y^0|c})\right) \\
& \qquad - \left(q_0 - Q^{\tau}_{Y^0|c}\right)^2 \frac{f_{Y^0|c}(Q^{\tau}_{Y^0|c}) \cdot P_c}{2} \Big|\Big| \\
& \qquad \leq \|A_1\| + \|A_2\| + \|A_3\| + \|A_4\| = o_p\left(\frac{1}{n}\right). \qquad (29)
\end{aligned}$$

It remains to be shown that all these terms are $o_p(\frac{1}{n})$. Similarly to Hirano, Imbens, and Ridder (2003) and Firpo (2007), we consider a situation where n increases but $\sqrt{n}(q_1 - Q^{\tau}_{Y^1|c})$ and $\sqrt{n}(q_0 - Q^{\tau}_{Y^0|c})$ remain fixed. We will use throughout that $\hat{W} = \frac{ZD}{\hat{p}(X)} - \frac{(1-Z)D}{1-\hat{p}(X)} - \frac{Z(1-D)}{\hat{p}(X)} + \frac{(1-Z)(1-D)}{1-\hat{p}(X)}$ and $\frac{\hat{p}_i - p_i}{p_i^2}\left(1 - \frac{\hat{p}_i - p_i}{\hat{p}_i}\right) = \frac{1}{p_i} - \frac{1}{\hat{p}_i}$.

Term $\|A_1\| =$

$$\left|q_1 - Q^{\tau}_{Y^1|c}\right| \cdot \Big|\Big| \frac{1}{n}\sum_{i=1}^{n}\left(\hat{W}_i D_i \left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right) - \psi_1(Y_i, D_i, Z_i, X_i)\right)\Big|\Big|$$

$$\begin{aligned}
= \left|q_1 - Q^{\tau}_{Y^1|c}\right| \cdot \Big|\Big| \frac{1}{n}\sum_{i=1}^{n}\Big[& ZD\left(\frac{1}{p_i} - \frac{\hat{p}_i - p_i}{p_i^2}\left(1 - \frac{\hat{p}_i - p_i}{\hat{p}_i}\right)\right)\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right) \\
& - (1-Z)D\left(\frac{1}{1-p_i} + \frac{\hat{p}_i - p_i}{(1-p_i)^2}\left(1 + \frac{\hat{p}_i - p_i}{1-\hat{p}_i}\right)\right)\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right) \\
& - \psi_1(Y_i, D_i, Z_i, X_i)\Big]\Big|\Big|
\end{aligned}$$

and after inserting (24)

$$= \left|q_1 - Q^{\tau}_{Y^1|c}\right| \cdot \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{\pi(X_i,1)\,(Z_i - p_i)}{p\,(X_i)}\vartheta_{11}(X_i) - ZD\frac{\hat{p}_i - p_i}{p_i^2}\left(1 - \frac{\hat{p}_i - p_i}{\hat{p}_i}\right)\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right)$$
$$+ \frac{\pi(X_i,0)\,(Z_i - p_i)}{1 - p\,(X_i)}\vartheta_{10}(X_i) - (1-Z)D\frac{\hat{p}_i - p_i}{(1-p_i)^2}\left(1 + \frac{\hat{p}_i - p_i}{1-\hat{p}_i}\right)\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right)\Big]\Big|\Big|$$

using (45) and (46) and the analogous expressions for $\vartheta_{10}(X_i)$ we obtain

$$= \left|q_1 - Q^{\tau}_{Y^1|c}\right| \cdot \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\Big[\left(\frac{\pi(X_i,1)\vartheta_{11}(X_i)}{p\,(X_i)} + \frac{\pi(X_i,0)\vartheta_{10}(X_i)}{1-p\,(X_i)}\right)(Z_i - p_i)\,o_p(1) + o_p\left(\frac{1}{\sqrt{n}}\right)\Big]\Big|\Big|$$

$$= \sqrt{n}\left|q_1 - Q^{\tau}_{Y^1|c}\right| \cdot \Big|\Big|\frac{1}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n}\Big[\left(\frac{\pi(X_i,1)\vartheta_{11}(X_i)}{p\,(X_i)} + \frac{\pi(X_i,0)\vartheta_{10}(X_i)}{1-p\,(X_i)}\right)(Z_i - p_i)\,o_p(1) + o_p\left(\frac{1}{\sqrt{n}}\right)\Big]\Big|\Big|$$

$$= o_p\left(\frac{1}{n}\right)$$

where we used that $\frac{1}{n}\sum \mathcal{A}_i = O_p(E\left[\mathcal{A}_i\right] + \sqrt{Var\left(\mathcal{A}_i\right)/n})$.

The derivations for the **Term** $\|A_2\|$ are analogous and are omitted.

Next, we examine **Term** $\|A_3\|$. (The derivations for **Term** $\|A_4\|$ are analogous to these and are omitted.)

Term $\|A_3\| =$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_iD_i}{\hat{p}_i} - \frac{(1-Z_i)D_i}{1-\hat{p}_i}\right)(Y_i - q_1)\left\{1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1)\right\} - \left(q_1 - Q^{\tau}_{Y^1|c}\right)^2\frac{f_{Y^1|c}(Q^{\tau}_{Y^1|c})\cdot P_c}{2}\right\|$$
$$= \left\|\frac{1}{n}\sum_{i=1}^{n}D_i\,(1-Z_i)\,\frac{p_i - \hat{p}_i}{(1-\hat{p}_i)\,(1-p_i)}\,(Y_i - q_1)\left\{1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1)\right\}\right\|$$
$$+ \left\|\frac{1}{n}\sum_{i=1}^{n} - D_iZ_i\frac{\hat{p}_i - p_i}{\hat{p}_ip_i}\,(Y_i - q_1)\left\{1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1)\right\}\right\|$$
$$+ \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{D_iZ_i}{p_i}\,(Y_i - q_1)\left\{1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1)\right\} - E\left[\frac{D_iZ_i}{p_i}\,(Y_i - q_1)\left\{1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1)\right\}\right]\right\|$$
$$+ \Big|\Big|\frac{1}{n}\sum_{i=1}^{n} - D_i\frac{1-Z_i}{1-p_i}\,(Y_i - q_1)\left\{1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1)\right\}$$
$$+ E\left[D_i\frac{1-Z_i}{1-p_i}\,(Y_i - q_1)\left\{1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1)\right\}\right]\Big|\Big|$$
$$+ \left\|\frac{1}{n}\sum_{i=1}^{n}E\left[DW\cdot(Y - q_1)\left\{1\left(Y < Q^{\tau}_{Y^1|c}\right) - 1(Y < q_1)\right\}\right] - \left(q_1 - Q^{\tau}_{Y^1|c}\right)^2\frac{f_{Y^1|c}(Q^{\tau}_{Y^1|c})\cdot P_c}{2}\right\|.$$

Now we consider each of the five terms separately and show that they are all $o_p(n^{-1})$ where we consider a situation where n increases but $\sqrt{n}(q_1 - Q^{\tau}_{Y^1|c})$ and $\sqrt{n}(q_0 - Q^{\tau}_{Y^0|c})$ are fixed.

Now consider the second term of $\|A_3\|$. (Also note that the first term is analogous and the corresponding

derivations are omitted here.)

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} - D_i Z_i \frac{\hat{p}_i - p_i}{\hat{p}_i p_i} \left(Y_i - q_1\right) \left\{ 1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1) \right\} \right\| \\
&\leq \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Z_i}{p_i} \left(Y_i - q_1\right) \left\{ 1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1) \right\} \right\| \sup_{x\in\mathcal{X}} \left(\hat{p}(x) - p(x)\right) \cdot \left(\inf_{x\in\mathcal{X}} \hat{p}(x) \right)^{-1} \\
&= \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Z_i}{p_i} \left(Y_i - q_1\right) \left\{ 1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1) \right\} \right\| \cdot o_p(1) \cdot O_p(1)
\end{aligned}$$

because $p(x)$ is assumed to be bounded away from zero and one and because $\hat{p}(x)$ is uniformly consistent. Now we use (32) and (34) together with $\frac{1}{n}\sum \mathcal{A}_i = O_p(E\left[\mathcal{A}_i\right] + \sqrt{Var\left(\mathcal{A}_i\right)/n})$ to obtain

$$= \left\| O_p(n^{-1} + \sqrt{n^{-\frac{3}{2}-1}}) \right\| \cdot o_p(1) \cdot O_p(1) = o_p\left(\frac{1}{n}\right).$$

The derivations for the first term of $\|A_3\|$ are analogous and are omitted.

Third term of $\|A_3\|$:

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Z_i}{p(X_i)} \left(Y_i - q_1\right) \left\{ 1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1) \right\} - E\left[\frac{D_i Z_i}{p(X_i)} \left(Y_i - q_1\right) \left\{ 1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1) \right\} \right] \right\| \\
&= \left\| O_p(\sqrt{n^{-\frac{3}{2}-1}}) \right\| = o_p\left(\frac{1}{n}\right)
\end{aligned}$$

where we used (34) together with $\frac{1}{n}\sum\left(\mathcal{A}_i - E\left[\mathcal{A}_i\right]\right) = O_p(\sqrt{Var\left(\mathcal{A}_i\right)/n})$.

The derivations for the fourth term of $\|A_3\|$ are analogous and are omitted.

Fifth term of $\|A_3\|$: Here we use (33) to obtain

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} E\left[DW \cdot \left(Y - q_1\right) \left\{ 1\left(Y < Q^{\tau}_{Y^1|c}\right) - 1(Y < q_1) \right\} \right] - \frac{1}{2}\left(q_1 - Q^{\tau}_{Y^1|c}\right)^2 f_{Y^1|c}(Q^{\tau}_{Y^1|c}) \cdot P_c \right\| \\
&= \Bigg\| \frac{1}{n}\sum_{i=1}^{n} E\left[DW \left\{ \frac{f_{Y|X,Z,D}(Q^{\tau}_{Y^1|c})}{2}\left(q_1 - Q^{\tau}_{Y^1|c}\right)^2 + \frac{O\left(\left(\sqrt{n}\left(q_1 - Q^{\tau}_{Y^1|c}\right)\right)^3\right)}{n^{\frac{3}{2}}} \right\} \right] \\
&- \frac{\left(q_1 - Q^{\tau}_{Y^1|c}\right)^2}{2} f_{Y^1|c}(Q^{\tau}_{Y^1|c}) \cdot P_c \Bigg\| \\
&= \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{O\left(\left(\sqrt{n}\left(q_1 - Q^{\tau}_{Y^1|c}\right)\right)^3\right)}{n^{\frac{3}{2}}} \right\| = O_p\left(\frac{1}{n^{\frac{3}{2}}}\right) = o_p\left(\frac{1}{n}\right)
\end{aligned}$$

where we made use of

$$P_c \cdot f_{Y^1|c}(Q^{\tau}_{Y^1|c}) = E\left[DW \cdot f_{Y|D,X,Z}(Q^{\tau}_{Y^1|c}) \right] \qquad \text{and} \qquad P_c \cdot f_{Y^0|c}(Q^{\tau}_{Y^0|c}) = E\left[(1-D)W \cdot f_{Y|D,X,Z}(Q^{\tau}_{Y^0|c}) \right]$$

which followed from (63) and (64).

Some intermediate results that have been used in the derivations above. Consider the term:

$$
\begin{aligned}
&E\left[(Y_i - q_1)\left\{1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1)\right\}|X, D, Z\right] \\
&= \int_{q_1}^{Q^{\tau}_{Y^1|c}} (y - q_1) \cdot f_{Y|X,D,Z}(y)dy = \int_{q_1}^{Q^{\tau}_{Y^1|c}} \frac{\partial\left\{(y - q_1) \cdot F_{Y|X,D,Z}(y)\right\} - F_{Y|X,D,Z}(y)}{\partial y} dy \\
&= \left(Q^{\tau}_{Y^1|c} - q_1\right) \cdot F_{Y|X,D,Z}(Q^{\tau}_{Y^1|c}) - (q_1 - q_1) \cdot F_{Y|X,D,Z}(q_1) - \int_{q_1}^{Q^{\tau}_{Y^1|c}} F_{Y|X,D,Z}(y)dy \\
&= \left(Q^{\tau}_{Y^1|c} - q_1\right) \cdot F_{Y|X,D,Z}(Q^{\tau}_{Y^1|c}) - F_{Y|X,D,Z}(Q^{\tau}_{Y^1|c})\left(Q^{\tau}_{Y^1|c} - q_1\right) \\
&\qquad - f_{Y|X,D,Z}(Q^{\tau}_{Y^1|c})\left\{\frac{1}{2}\left(Q^{\tau}_{Y^1|c} - Q^{\tau}_{Y^1|c}\right)^2 - \frac{1}{2}\left(q_1 - Q^{\tau}_{Y^1|c}\right)^2\right\} \\
&\qquad - f'_{Y|X,D,Z}(Q^{\tau}_{Y^1|c})\left\{\frac{1}{6}\left(Q^{\tau}_{Y^1|c} - Q^{\tau}_{Y^1|c}\right)^3 - \frac{1}{6}\left(q_1 - Q^{\tau}_{Y^1|c}\right)^3\right\} - O\left(\left(q_1 - Q^{\tau}_{Y^1|c}\right)^4\right) \\
&= \frac{1}{2} f_{Y|X,D,Z}(Q^{\tau}_{Y^1|c})\left(q_1 - Q^{\tau}_{Y^1|c}\right)^2 + O\left(\left(q_1 - Q^{\tau}_{Y^1|c}\right)^3\right) \\
&= \frac{\frac{1}{2} f_{Y|X,D,Z}(Q^{\tau}_{Y^1|c})}{n}\left(\sqrt{n}\left(q_1 - Q^{\tau}_{Y^1|c}\right)\right)^2 + \frac{1}{n^{\frac{3}{2}}} O\left(\left(\sqrt{n}\left(q_1 - Q^{\tau}_{Y^1|c}\right)\right)^3\right) \qquad (30)
\end{aligned}
$$

where we used the expansion $F_{Y|X,D,Z}(y) = F_{Y|X,D,Z}(Q^{\tau}_{Y^1|c}) + (y - Q^{\tau}_{Y^1|c}) \cdot f_{Y|X,D,Z}(Q^{\tau}_{Y^1|c}) + \frac{1}{2}(y - Q^{\tau}_{Y^1|c})^2 \cdot f'_{Y|X,D,Z}(Q^{\tau}_{Y^1|c}) + O((y - Q^{\tau}_{Y^1|c})^3)$.

Also consider

$$
\begin{aligned}
&E\left[(Y_i - q_1)^2\left\{1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - 1(Y_i < q_1)\right\}|X, Z, D\right] \\
&= \int_{q_1}^{Q^{\tau}_{Y^1|c}} (y - q_1)^2 \cdot f_{Y|X,D,Z}(y)dy = \int_{q_1}^{Q^{\tau}_{Y^1|c}} \frac{\partial\left\{(y - q_1)^2 \cdot F_{Y|X,D,Z}(y)\right\} - 2(y - q_1) F_{Y|X,D,Z}(y)}{\partial y} dy \\
&= \left(Q^{\tau}_{Y^1|c} - q_1\right)^2 \cdot F_{Y|X,D,Z}(Q^{\tau}_{Y^1|c}) - (q_1 - q_1)^2 \cdot F_{Y|X,D,Z}(q_1) - \int_{q_1}^{Q^{\tau}_{Y^1|c}} 2(y - q_1) F_{Y|X,D,Z}(y)dy
\end{aligned}
$$

Now again expanding $F_{Y|X,D,Z}$

$$
\begin{aligned}
&= \left(Q^{\tau}_{Y^1|c} - q_1\right)^2 \cdot F_{Y|X,D,Z}(Q^{\tau}_{Y^1|c}) - 2F_{Y|X,D,Z}(Q^{\tau}_{Y^1|c}) \int_{q_1}^{Q^{\tau}_{Y^1|c}} (y - q_1)\, dy \\
&- 2f_{Y|X,D,Z}(Q^{\tau}_{Y^1|c}) \int_{q_1}^{Q^{\tau}_{Y^1|c}} (y - q_1)(y - Q^{\tau}_{Y^1|c})dy - f'_{Y|X,D,Z}(Q^{\tau}_{Y^1|c}) \int_{q_1}^{Q^{\tau}_{Y^1|c}} (y - q_1)\left((y - Q^{\tau}_{Y^1|c})^2 + O(y - Q^{\tau}_{Y^1|c})^3\right) dy \\
&= -2f_{Y|X,D,Z}(Q^{\tau}_{Y^1|c})\frac{1}{6}(q_1 - Q^{\tau}_{Y^1|c})^3 - f'_{Y|X,D,Z}(Q^{\tau}_{Y^1|c})\frac{1}{12}(q_1 - Q^{\tau}_{Y^1|c})^4 \\
&= -\frac{f_{Y|X,Z,D}(Q^{\tau}_{Y^1|c})}{3 \cdot n^{\frac{3}{2}}}\left(\sqrt{n}\left(q_1 - Q^{\tau}_{Y^1|c}\right)\right)^3 + O\left(\frac{1}{n^2}\right). \qquad (31)
\end{aligned}
$$

Furthermore

$$E\left[\left\{(Y_i-q_1)\left\{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-1(Y_i<q_1)\right\}\right\}^2|X,D,Z\right]$$
$$=1\left(Q^{\tau}_{Y^1|c}>q_1\right)\cdot E\left[(Y_i-q_1)^2\left\{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-1(Y_i<q_1)\right\}|X,D,Z\right]$$
$$-1\left(Q^{\tau}_{Y^1|c}<q_1\right)\cdot E\left[(Y_i-q_1)^2\left\{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-1(Y_i<q_1)\right\}|X,D,Z\right]$$
$$=O\left(\frac{1}{\sqrt{n}^3}\right),$$

which follows from (31).

Combining these intermediaries we obtain by using (30)

$$E\left[\frac{D_iZ_i}{p(X_i)}(Y_i-q_1)\left\{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-1(Y_i<q_1)\right\}\right]$$
$$=E\left[\pi(X,1)\cdot E\left[(Y-q_1)\left\{1\left(Y<Q^{\tau}_{Y^1|c}\right)-1(Y<q_1)\right\}|X,Z=1,D=1\right]\right]$$
$$=E\left[\pi(X,1)\cdot\left\{\frac{1}{2}f_{Y|X,Z=1,D=1}(Q^{\tau}_{Y^1|c})\left(q_1-Q^{\tau}_{Y^1|c}\right)^2+\frac{1}{n^{\frac{3}{2}}}O\left(\left(\sqrt{n}\left(q_1-Q^{\tau}_{Y^1|c}\right)\right)^3\right)\right\}\right]$$
$$=O\left(\frac{1}{n}\right) \quad (32)$$

and

$$E\left[D_iW_i(Y_i-q_1)\left\{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-1(Y_i<q_1)\right\}\right]$$
$$=E\left[D_iW_i\cdot\left\{\frac{1}{2}f_{Y|X,Z,D}(Q^{\tau}_{Y^1|c})\left(q_1-Q^{\tau}_{Y^1|c}\right)^2+\frac{1}{n^{\frac{3}{2}}}O\left(\left(\sqrt{n}\left(q_1-Q^{\tau}_{Y^1|c}\right)\right)^3\right)\right\}\right]. \quad (33)$$

We also need the variance expression

$$Var\left(\frac{D_iZ_i}{p_i}(Y_i-q_1)\left\{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-1(Y_i<q_1)\right\}\right)$$
$$=E\left[\left(\frac{D_iZ_i}{p_i}(Y_i-q_1)\left\{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-1(Y_i<q_1)\right\}\right)^2\right]-O\left(\frac{1}{n^2}\right)$$

by (32). Further

$$=E\left[\frac{\pi(X,1)}{p(X)}\cdot E\left[\left((Y-q_1)\left\{1\left(Y<Q^{\tau}_{Y^1|c}\right)-1(Y<q_1)\right\}\right)^2|X,Z=1,D=1\right]\right]-O\left(\frac{1}{n^2}\right)$$
$$=O\left(\frac{1}{\sqrt{n}^3}\right) \quad (34)$$

by (31).

A.7.4 Minimizer of $\tilde{G}_n(q_1, q_0)$

As a corollary to the final proof, we now examine the properties of the minimizers of the function $\tilde{G}_n(q_1, q_0)$, given in (28), and show that the minimizers of $\tilde{G}_n$ converge to an asymptotically normal distribution. Since $\tilde{G}_n(q_1, q_0)$ is differentiable and quadratic in q_1 and in q_0 its minimizers are defined by the first order condition:

$$0 = \frac{1}{n}\sum_{i=1}^{n}\psi_1(Y, D, Z, X) + \left(\tilde{q}_1 - Q^{\tau}_{Y^1|c}\right) f_{Y^1|c}(Q^{\tau}_{Y^1|c}) \cdot P_c \tag{35}$$

$$0 = \frac{1}{n}\sum_{i=1}^{n}\psi_0(Y, D, Z, X) + \left(\tilde{q}_0 - Q^{\tau}_{Y^0|c}\right) f_{Y^0|c}(Q^{\tau}_{Y^0|c}) \cdot P_c \tag{36}$$

$$\sqrt{n}\left(\tilde{q}_1 - Q^{\tau}_{Y^1|c}\right) = -\frac{1}{f_{Y^1|c}(Q^{\tau}_{Y^1|c}) \cdot P_c}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1(Y, D, Z, X) \tag{37}$$

$$\sqrt{n}\left(\tilde{q}_0 - Q^{\tau}_{Y^0|c}\right) = -\frac{1}{f_{Y^0|c}(Q^{\tau}_{Y^0|c}) \cdot P_c}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_0(Y, D, Z, X). \tag{38}$$

We thus obtain

$$\sqrt{n}\left(\tilde{q}_1 - Q^{\tau}_{Y^1|c}\right) \xrightarrow{d} N\left(0, Var\left(\frac{\psi_1(Y, D, Z, X)}{f_{Y^1|c}(Q^{\tau}_{Y^1|c}) \cdot P_c}\right)\right) \tag{39}$$

$$\sqrt{n}\left(\tilde{q}_0 - Q^{\tau}_{Y^0|c}\right) \xrightarrow{d} N\left(0, Var\left(\frac{\psi_0(Y, D, Z, X)}{f_{Y^0|c}(Q^{\tau}_{Y^0|c}) \cdot P_c}\right)\right) \tag{40}$$

$$\sqrt{n}\left(\tilde{q}_1 - \tilde{q}_0 - \Delta^{\tau}_c\right) \xrightarrow{d} N\left(0, Var\left(\frac{\psi_1(Y, D, Z, X)}{f_{Y^1|c}(Q^{\tau}_{Y^1|c}) \cdot P_c} - \frac{\psi_0(Y, D, Z, X)}{f_{Y^0|c}(Q^{\tau}_{Y^0|c}) \cdot P_c}\right)\right), \tag{41}$$

by a CLT for iid data.

A.7.5 Properties of $\hat{q}_1$ and $\hat{q}_0$

In the previous subsection we defined the statistic $\tilde{G}_n(q_1, q_0)$, which has unique minimizers $\tilde{q}_1$ and $\tilde{q}_0$, and derived their properties. In contrast, the minimizers $\hat{q}_1$ and $\hat{q}_0$ of the objective function G_n may not be unique as G_n may have "flat regions". Both functions, however, are convex and various approaches can be used to show that $\hat{q}_1$ and $\tilde{q}_1$ are close in the sense that $\hat{q}_1 = \tilde{q}_1 + o_p(\frac{1}{\sqrt{n}})$. Hence, the first order asymptotic distribution is identical for $\hat{q}_1$ and $\tilde{q}_1$. The same applies for $\hat{q}_0$ and $\tilde{q}_0$ and thus to $\hat{\Delta}^{\tau}_c$ and $\tilde{\Delta}^{\tau}_c$. In the following we make use of results in Hjort and Pollard (1993).

To simplify the notation in the following, we focus on q_1 and ignore q_0 since $\tilde{G}_n$ and G_n are both additively separable in q_1 and q_0. The two simplified statistics are thus

$$\tilde{G}_n(q_1) = \left(q_1 - Q^{\tau}_{Y^1|c}\right)\frac{1}{n}\sum_{i=1}^{n}\psi_1(Y_i, D_i, Z_i, X_i) + \left(q_1 - Q^{\tau}_{Y^1|c}\right)^2 \frac{f_{Y^1|c}(Q^{\tau}_{Y^1|c}) \cdot P_c}{2}$$

and

$$G_n(q_1, \hat{W}) = \frac{1}{n}\sum_{i=1}^{n} \hat{W}_i D_i \rho_\tau(Y_i - q_1) - \frac{1}{n}\sum_{i=1}^{n} \hat{W}_i D_i \rho_\tau(Y_i - Q^\tau_{Y^1|c}).$$

Lemma 2 of Hjort and Pollard (1993) states that for each $\delta > 0$ and for G_n and $\tilde{G}_n$ both convex functions in q_1 and with $\tilde{q}_1$ being the unique minimizer of $\tilde{G}_n$

$$\Pr\left(|\hat{q}_1 - \tilde{q}_1| \geq \delta\right) \leq \Pr\left(2 \sup_{|s-\tilde{q}_1|\leq\delta} \left|G_n(s, \hat{W}) - \tilde{G}_n(s)\right| \geq \inf_{|s-\tilde{q}_1|=\delta} \tilde{G}_n(s) - \tilde{G}_n(\tilde{q}_1)\right).$$

After some calculations and making use of (37) we obtain $\tilde{G}_n(\tilde{q}_1 - \delta) - \tilde{G}_n(\tilde{q}_1) = \delta^2 \cdot \frac{f_{Y^1|c}(Q^\tau_{Y^1|c}) \cdot P_c}{2}$ and the same result for $\tilde{G}_n(\tilde{q}_1 + \delta) - \tilde{G}_n(\tilde{q}_1)$. Further choosing $\delta = \varepsilon/\sqrt{n}$ we obtain:

$$\Pr\left(\sqrt{n}\,|\hat{q}_1 - \tilde{q}_1| \geq \varepsilon\right) \leq \Pr\left(2 \sup_{|s-\tilde{q}_1|\leq\frac{\varepsilon}{\sqrt{n}}} \left|G_n(s, \hat{W}) - \tilde{G}_n(s)\right| \geq \frac{\varepsilon^2}{n} \cdot \frac{f_{Y^1|c}(Q^\tau_{Y^1|c}) \cdot P_c}{2}\right). \tag{42}$$

Now we make use of the previous result in (29) that $|G_n(q_1, \hat{W}) - \tilde{G}_n(q_1)| = o_p(\frac{1}{n})$ for any value of q_1. By lemma 1 of Hjort and Pollard (1993) this also implies $\sup_{q_1 \in \mathcal{S}} |G_n(q_1, \hat{W}) - \tilde{G}_n(q_1)| = o_p(\frac{1}{n})$ for any compact set $\mathcal{S}$. Because the rightmost expression in (42) is always positive and of order $O_p(\frac{1}{n})$ whereas the expression $\sup \left|G_n(s, \hat{W}) - \tilde{G}_n(s)\right|$ is $o_p(\frac{1}{n})$, this now implies that

$$\Pr\left(\sqrt{n}\,|\hat{q}_1 - \tilde{q}_1| \geq \varepsilon\right) \xrightarrow{p} 0, \qquad \forall \varepsilon > 0$$

which completes the proof.

A.8 Proof of Theorem (8)

The proof of the theorem is essentially identical to the proof for local linear regression. We only have to show that (43) equals (45) when we estimate $\hat{p}_i$ by local logit regression. All these derivations are given in Appendix C.

B Properties of the local linear regression estimator

In this subsection several preliminaries for the proof of the previous theorem are derived. These results are stated in recursive order. First, the results most pertinent to the proofs are given followed by derivations for these intermediate results etc.

B.1 Properties of some V-statistics

Now we analyze the term

$$\frac{1}{n}\sum_{i=1}^{n} \frac{Z_i D_i}{p_i^2} \left(1\left(Y_i < Q^\tau_{Y^1|c}\right) - \tau\right)(\hat{p}_i - p_i) \tag{43}$$

with $L = \dim(X)$, when $\hat{p}$ is estimated by local linear regression, with some properties given in the subsequent section. Making use of expression (47) and defining

$$\varsigma_{ij} = \frac{Z_i D_i}{p_i^2}\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right) \cdot e_1' \left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1} \mathbb{X}_{j,i}K_{j,i} \times \left((Z_j - p_j) + (X_j - X_i)' \frac{\partial^2 p(X_i)}{\partial x \partial x'}(X_j - X_i) + O\left(h^3\right)\right)$$

we obtain that (43) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i D_i}{p_i^2}\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right)(\hat{p}_i - p_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varsigma_{ij} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\varsigma_{ij} + \varsigma_{ji}}{2}.$$

The latter term is a nondegenerate symmetric von Mises statistic. The von Mises statistic is asymptotically equivalent to the corresponding U-statistic, and its projection is

$$= \frac{2}{n}\sum_{i=1}^{n}\left(E\left[\frac{\varsigma_{ij} + \varsigma_{ji}}{2}|X_i, Z_i, D_i, Y_i\right] - E\left[\frac{\varsigma_{ij} + \varsigma_{ji}}{2}\right]\right) + E\left[\frac{\varsigma_{ij} + \varsigma_{ji}}{2}\right] + o_p\left(\frac{1}{\sqrt{n}}\right) \tag{44}$$

under the condition that $E\left[\left(\frac{\varsigma_{ij}+\varsigma_{ji}}{2}\right)^2\right] \leq o(n)$, see Serfling (1980, p.190) and Powell, Stock, and Stoker (1989). To verify this condition note that $E[\varsigma_{ij}^2] \leq o(n)$ by (53) and that $E|\varsigma_{ij}\varsigma_{ji}| \leq \sqrt{E[\varsigma_{ij}^2] \cdot E[\varsigma_{ji}^2]}$ by Hölder's inequality. From (50) and (52) we obtain that $E[\varsigma_{ij} + \varsigma_{ji}|X_i, Z_i, D_i, Y_i] = \left(\frac{\pi(X_i,1)}{p(X_i)}\vartheta_{11}(X_i) + O(h)\right) \cdot (Z_i - p_i) + O(h^{\lambda})$ and $E[\varsigma_{ij} + \varsigma_{ji}] = O(h^{\lambda})$. This gives

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\pi(X_i,1)}{p(X_i)}\vartheta_{11}(X_i) + O_p(h)\right) \cdot (Z_i - p_i) + O_p(h^{\lambda}) + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Under the condition that $nh^{2\lambda} \to 0$ the bias term is $o_p(n^{-\frac{1}{2}})$ and the asymptotic distribution is determined by the first term

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(X_i,1)\vartheta_{11}(X_i)(Z_i - p_i)}{p(X_i)}(1 + O_p(h)) + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{45}$$

Next, consider the term

$$\begin{aligned} & \left|\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i D_i}{p_i^2}\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right)(\hat{p}_i - p_i)\left(1 - \frac{\hat{p}_i - p_i}{\hat{p}_i}\right)\right| \\ \leq & \left|\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i D_i}{p_i^2}\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right)(\hat{p}_i - p_i)\right| \sup_{x\in\mathcal{X}}\left|\left(1 + \frac{p_i - \hat{p}_i}{\hat{p}_i}\right)\right| \\ = & \left|\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i D_i}{p_i^2}\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right) - \tau\right)(\hat{p}_i - p_i)\right| \cdot (1 + o_p(1)) \end{aligned} \tag{46}$$

since $\hat{p}(x)$ is uniformly consistent and $p(x)$ is bounded away from zero over the support of X. Hence, this term is of the same order as (43) and its asymptotic properties are determined by (45).

By analogous derivations we also obtain for

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)D_i}{p_i^2}\left(1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-\tau\right)(\hat{p}_i-p_i)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(X_i,0)\vartheta_{10}(X_i)\,(Z_i-p_i)}{1-p\,(X_i)}\,(1+O_p(h))+o_p\left(\frac{1}{\sqrt{n}}\right).$$

B.2 Local linear regression

Consider estimation of $p(x_0)$ at a location x_0. Define the regressor matrices $\mathbb{X}_j=\left(1,\left(\frac{X_j-x_0}{h}\right)'\right)'$ and $\mathbb{X}=(\mathbb{X}_1,\mathbb{X}_2,...,\mathbb{X}_n)'$ and $\mathbb{K}=diag(K_1,K_2,...,K_n)$. Since $p(x_0)$ is estimated by a weighted least squares regression, we can write the solution as

$$\hat{p}(x_0)=e_1'\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1}\sum_{j=1}^{n}\mathbb{X}_jK_jZ_j=e_1'\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1}\sum_{j=1}^{n}\mathbb{X}_jK_j\left(Z_j-p_j+p_j\right)$$

where e_1 is a column vector of zeros with first element being one and $p_j=p(X_j)$. A series expansion gives

$$=e_1'\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1}\sum_{j=1}^{n}\mathbb{X}_jK_j\left(Z_j-p_j\right)$$
$$+e_1'\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1}\sum_{j=1}^{n}\mathbb{X}_jK_j\left(p\left(x_0\right)+\left(X_j-x_0\right)'\frac{\partial p\left(x_0\right)}{\partial x}+\left(X_j-x_0\right)'\frac{1}{2}\frac{\partial^2 p\left(x_0\right)}{\partial x\partial x'}\left(X_j-x_0\right)+R_j\right)$$

where $\frac{\partial p(x_0)}{\partial x}$ is the $L\times 1$ vector of first derivatives and $\frac{\partial^2 p(x_0)}{\partial x\partial x'}$ the $L\times L$ matrix of second derivatives and R_j is the remainder term of all third order derivatives multiplied with the respective third order interaction terms of X_j-x_0. Since K_j has bounded support, the remainder term premultiplied with K_j is of order $O(K_j\cdot h^3)$. We thus obtain after some derivations that

$$=e_1'\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1}\sum_{j=1}^{n}\mathbb{X}_jK_j\left(Z_j-p_j\right)+p\left(x_0\right)+e_1'\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1}\sum_{j=1}^{n}\mathbb{X}_jK_j\left(\left(X_j-x_0\right)'\frac{1}{2}\frac{\partial^2 p\left(x_0\right)}{\partial x\partial x'}\left(X_j-x_0\right)+O\left(h^3\right)\right).$$

Now we replace x_0 with X_i to obtain the expression when estimating at a location X_i

$$\hat{p}(X_i)-p(X_i)=e_1'\left(\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\sum_{j=1}^{n}\mathbb{X}_{j,i}K_{j,i}\left(\left(Z_j-p_j\right)+\left(X_j-X_i\right)'\frac{1}{2}\frac{\partial^2 p\left(X_i\right)}{\partial x\partial x'}\left(X_j-X_i\right)+O\left(h^3\right)\right),\tag{47}$$

where $\mathbb{X}_{j,i}=\left(1,\left(\frac{X_j-X_i}{h}\right)'\right)'$ and $K_{j,i}=\prod_{l=1}^{L}\frac{1}{h}\kappa\left(\frac{X_{jl}-X_{il}}{h}\right)$ and $\mathbb{X}_i=(\mathbb{X}_{1,i},\mathbb{X}_{2,i},...,\mathbb{X}_{n,i})'$ and $\mathbb{K}_i=diag(K_{1,i},K_{2,i},...,K_{n,i})$.

B.3 Denominator of the local linear estimator

Under the assumption that $nh^L \to \infty$ and $h \to 0$ one can show that for a kernel of order λ:

$$\frac{1}{n}\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{X}_j\mathbb{X}_j'\prod_{l=1}^{L}\frac{1}{h}\kappa\left(\frac{X_{jl}-x_l}{h}\right)$$

$$= \begin{bmatrix} f(x_0)+O(h^\lambda) & h^{\lambda-1}\frac{\mu_\lambda}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(x_0)}{\partial x_1^{\lambda-1}}+O(h^\lambda) & \cdots & \cdots \\ h^{\lambda-1}\frac{\mu_\lambda}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(x_0)}{\partial x_1^{\lambda-1}}+O(h^\lambda) & h^{\lambda-2}\frac{\mu_\lambda}{(\lambda-2)!}\frac{\partial^{\lambda-2}f(x_0)}{\partial x_1^{\lambda-2}}+h^{\lambda-1}\frac{\mu_{\lambda+1}}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(x_0)}{\partial x_1^{\lambda-1}}+O(h^\lambda) & O(h^{2\lambda-2}) & \cdots \\ \vdots & O(h^{2\lambda-2}) & \ddots & O(h^{2\lambda-2}) \\ \vdots & \vdots & O(h^{2\lambda-2}) & \ddots \end{bmatrix} \tag{48}$$

This can be shown element-wise via mean square convergence. Only the derivations for the $(2,2)$ element are shown here, with the derivations for the other elements being analogous. Consider the $(2,2)$ element of $\frac{1}{n}\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)$ and denote it by ξ

$$\xi = \frac{1}{nh^L}\sum_{j=1}^{n}\left(\frac{X_{j1}-x_1}{h}\right)^2\prod_{l=1}^{L}\kappa\left(\frac{X_{jl}-x_l}{h}\right)$$

which has the expected value:

$$E[\xi] = \frac{1}{h^L}\int\cdots\int\left(\frac{X_{j1}-x_1}{h}\right)^2\prod_{l=1}^{L}\kappa\left(\frac{X_{jl}-x_l}{h}\right)f(X_j)dX_j.$$

With a change in variables: $u_l = \frac{X_{jl}-x_l}{h}$ and $u = (u_1,...,u_L)'$ and a Taylor series expansion and noting that κ is a kernel of order λ we obtain

$$\begin{aligned} &= \int\cdots\int u_1^2\prod_{l=1}^{L}\kappa\left(u_l\right)f(x_0+uh)du \\ &= \int\cdots\int u_1^2\prod_{l=1}^{L}\kappa\left(u_l\right)\left(\frac{u_1^{\lambda-2}h^{\lambda-2}}{(\lambda-2)!}\frac{\partial^{\lambda-2}f(x_0)}{\partial u_1^{\lambda-2}}+\frac{u_1^{\lambda-1}h^{\lambda-1}}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(x_0)}{\partial u_1^{\lambda-1}}+O(h^\lambda)\right)du \\ &= h^{\lambda-2}\frac{\mu_\lambda}{(\lambda-2)!}\frac{\partial^{\lambda-2}f(x_0)}{\partial x_1^{\lambda-2}}+h^{\lambda-1}\frac{\mu_{\lambda+1}}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(x_0)}{\partial x_1^{\lambda-1}}+O(h^\lambda) \end{aligned}$$

by bounded convergence.

To show convergence in mean square, it also needs to be shown that $Var\left(\xi\right)$ converges to zero

$$\begin{aligned} Var\left(\xi\right) &= \frac{1}{n^2h^{2L}}\sum_{j=1}^{n}Var\left(\left(\frac{X_{j1}-x_1}{h}\right)^2\prod_{l=1}^{L}\kappa\left(\frac{X_{jl}-x_l}{h}\right)\right) \\ &= \frac{1}{nh^{2L}}E\left[\left(\left(\frac{X_{j1}-x_1}{h}\right)^2\prod_{l=1}^{L}\kappa\left(\frac{X_{jl}-x_l}{h}\right)\right)^2\right]-\frac{1}{nh^{2L}}\left(E\left[\left(\frac{X_{j1}-x_1}{h}\right)^2\prod_{l=1}^{L}\kappa\left(\frac{X_{jl}-x_l}{h}\right)\right]\right)^2 \\ &= \frac{1}{nh^{2L}}\int h^Lu^4\prod_{l=1}^{L}\kappa^2\left(u_l\right)f(x_0+uh)du-\frac{1}{nh^{2L}}\left(h^L\int\cdots\int u^2\prod_{l=1}^{L}\kappa\left(u_l\right)f(x_0+uh)du\right)^2 \\ &= O\left(\frac{1}{nh^L}\right)-O\left(\frac{h^{2\lambda-4}}{n}\right), \end{aligned}$$

by bounded convergence and Taylor series expansion. As it has been assumed that $nh^L \to \infty$, the variance of ξ converges to zero. Hence, mean square convergence has been shown, which implies convergence in probability by Chebyshev's inequality.

From (48) one can derive after some tedious calculations that

$$e_1'\left(\frac{1}{n}\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1} = \frac{1}{f(x_0)+O(h)}\begin{pmatrix} 1+h\frac{\mu_{\lambda+1}}{\mu_\lambda}\frac{(\lambda-2)!}{(\lambda-1)!}\sum_{l=1}^{L}\frac{\partial^{\lambda-1}f(x_0)}{\partial x_l^{\lambda-1}}/\frac{\partial^{\lambda-2}f(x_0)}{\partial x_l^{\lambda-2}} \\ -h\left(\frac{\partial^{\lambda-1}f(x_0)}{\partial x_1^{\lambda-1}}/\frac{\partial^{\lambda-2}f(x_0)}{\partial x_1^{\lambda-2}}\right)\frac{(\lambda-2)!}{(\lambda-1)!} \\ \vdots \\ -h\left(\frac{\partial^{\lambda-1}f(x_0)}{\partial x_L^{\lambda-1}}/\frac{\partial^{\lambda-2}f(x_0)}{\partial x_L^{\lambda-2}}\right)\frac{(\lambda-2)!}{(\lambda-1)!} \end{pmatrix}' + O(h^2)$$

$$= \frac{1}{f(x_0)}\begin{pmatrix} 1+O(h) \\ -h\left(\frac{\partial^{\lambda-1}f(x_0)}{\partial x_1^{\lambda-1}}/\frac{\partial^{\lambda-2}f(x_0)}{\partial x_1^{\lambda-2}}\right)\frac{(\lambda-2)!'}{(\lambda-1)!}+O(h^2) \\ \vdots \\ -h\left(\frac{\partial^{\lambda-1}f(x_0)}{\partial x_L^{\lambda-1}}/\frac{\partial^{\lambda-2}f(x_0)}{\partial x_L^{\lambda-2}}\right)\frac{(\lambda-2)!'}{(\lambda-1)!}+O(h^2) \end{pmatrix}' \quad (49)$$

B.4 Further properties of local linear regression

For deriving the asymptotic distribution of the QTE estimator the expressions appearing in equation (44) are needed, in particular $E[\varsigma_{ij}|X_i,Z_i,D_i,Y_i]$ and $E[\varsigma_{ji}|X_i,Z_i,D_i,Y_i]$. These are derived below. To simplify notation, we frequently write $\frac{\partial^{\lambda-1}f(X_i)/\partial x^{\lambda-1}}{\partial^{\lambda-2}f(X_i)/\partial x^{\lambda-2}}$ as a shorthand notation for the column vector $\left(\left(\frac{\partial^{\lambda-1}f(x_0)}{\partial x_1^{\lambda-1}}/\frac{\partial^{\lambda-2}f(x_0)}{\partial x_1^{\lambda-2}}\right),...,\left(\frac{\partial^{\lambda-1}f(x_0)}{\partial x_L^{\lambda-1}}/\frac{\partial^{\lambda-2}f(x_0)}{\partial x_L^{\lambda-2}}\right)\right)'$.

Derive first $E[\varsigma_{ij}|X_i,Z_i,D_i,Y_i]$ which is:

$$E\Big[\frac{Z_iD_i}{p_i^2}\left(1\left(Y_i<Q^\tau_{Y^1|c}\right)-\tau\right)e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{j,i}$$
$$\times\left((Z_j-p_j)+(X_j-X_i)'\frac{1}{2}\frac{\partial^2 p(X_i)}{\partial x\partial x'}(X_j-X_i)+O\left(h^3\right)\right)\ |X_i,Z_i,D_i,Y_i\Big]$$

$$= \frac{Z_iD_i}{p_i^2}\left(1\left(Y_i<Q^\tau_{Y^1|c}\right)-\tau\right)\cdot h^\lambda\frac{\mu_\lambda}{f(X_i)}\frac{1}{2}\sum_{l=1}^{L}\frac{\partial^2 p(X_i)}{\partial x_l^2}$$
$$\times\left(\frac{\partial^{\lambda-2}f(X_i)/\partial x_l^{\lambda-2}}{(\lambda-2)!}-\frac{(\lambda-2)!}{(\lambda-1)!\,(\lambda-3)!}\frac{\partial^{\lambda-1}f(X_i)/\partial x_l^{\lambda-1}}{\partial^{\lambda-2}f(X_i)/\partial x_l^{\lambda-2}}\frac{\partial^{\lambda-3}f(X_i)}{\partial x_l^{\lambda-3}}\right)=O(h^\lambda) \quad (50)$$

by (51).

In the last expression, the following term has been used, where we make use of (49):

$$
\begin{aligned}
& E\Big[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{j,i}\left((Z_j-p_j)+(X_j-X_i)'\,\frac{1}{2}\frac{\partial^2 p\,(X_i)}{\partial x\partial x'}\,(X_j-X_i)+O\left(h^3\right)\right)\ |X_i,Z_i,D_i,Y_i\Big] \\
= \ & E\Big[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{j,i}\left((X_j-X_i)'\,\frac{1}{2}\frac{\partial^2 p\,(X_i)}{\partial x\partial x'}\,(X_j-X_i)+O\left(h^3\right)\right)\ |X_i\Big] \\
= \ & E\Big[\left(1+O(h)-h\frac{(\lambda-2)!}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(X_i)/\partial x^{\lambda-1}{}'}{\partial^{\lambda-2}f(X_i)/\partial x^{\lambda-2}}\,\frac{X_j-X_i}{h}\,(1+O(h))\right)\frac{K_{j,i}}{f(X_i)} \\
& \quad \times h^2\left(\frac{X_j-X_i}{h}'\,\frac{1}{2}\frac{\partial^2 p\,(X_i)}{\partial x\partial x'}\frac{X_j-X_i}{h}+O\left(h\right)\right)\ |X_i\Big] \\
= \ & \int\left(1+O(h)-h\frac{(\lambda-2)!}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(X_i)/\partial x^{\lambda-1}{}'}{\partial^{\lambda-2}f(X_i)/\partial x^{\lambda-2}}\,\frac{X_j-X_i}{h}\,(1+O(h))\right)\frac{K_{j,i}}{f(X_i)} \\
& \quad \times h^2\left(\frac{X_j-X_i}{h}'\,\frac{1}{2}\frac{\partial^2 p\,(X_i)}{\partial x\partial x'}\frac{X_j-X_i}{h}+O\left(h\right)\right)\ f(X_j)dX_j \\
= \ & \frac{h^2}{f(X_i)}\int\left(1+O(h)-h\frac{(\lambda-2)!}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(X_i)/\partial x^{\lambda-1}{}'}{\partial^{\lambda-2}f(X_i)/\partial x^{\lambda-2}}\,u\,(1+O(h))\right) \\
& \quad \times\prod_{l=1}^{L}\kappa\left(u_l\right)\left(u'\frac{1}{2}\frac{\partial^2 p\,(X_i)}{\partial x\partial x'}u+O\left(h\right)\right)\ f(X_i+uh)du
\end{aligned}
$$

where $u=\frac{X_j-X_i}{h}$. For $\lambda>2$ and using a Taylor series expansion we obtain:

$$
= \frac{h^\lambda\mu_\lambda}{f(X_i)}\frac{1}{2}\sum_{l=1}^{L}\frac{\partial^2 p\,(X_i)}{\partial x_l^2}\left(\frac{\partial^{\lambda-2}f(X_i)/\partial x_l^{\lambda-2}}{(\lambda-2)!}-\frac{(\lambda-2)!}{(\lambda-1)!\,(\lambda-3)!}\frac{\partial^{\lambda-1}f(X_i)}{\partial x_l^{\lambda-1}}\frac{\partial^{\lambda-3}f(X_i)}{\partial x_l^{\lambda-3}}/\frac{\partial^{\lambda-2}f(X_i)}{\partial x_l^{\lambda-2}}\right) \tag{51}
$$

and for $\lambda=2$ this term would be

$$
= \frac{h^\lambda\mu_\lambda}{2}\sum_{l=1}^{L}\frac{\partial^2 p\,(X_i)}{\partial x_l^2}.
$$

In both cases this term is of order $O(h^\lambda)$.

Now we derive $E[\varsigma_{ji}|X_i,Z_i,D_i,Y_i]$ which is:

$$
\begin{aligned}
& E\Big[\frac{Z_jD_j}{p_j^2}\left(1\left(Y_j<Q^\tau_{Y^1|c}\right)-\tau\right)e_1'\left(\frac{1}{n}\mathbb{X}_j'\mathbb{K}_j\mathbb{X}_j\right)^{-1}\mathbb{X}_{i,j}K_{i,j} \\
& \quad \times\left((Z_i-p_i)+(X_i-X_j)'\,\frac{1}{2}\frac{\partial^2 p\,(X_j)}{\partial x\partial x'}\,(X_i-X_j)+O\left(h^3\right)\right)\ |X_i,Z_i,D_i,Y_i\Big] \\
= \ & E\Big[E\left[\frac{Z_jD_j}{p_j^2}\left(1\left(Y_j<Q^\tau_{Y^1|c}\right)-\tau\right)|Z_i,X_1...X_n\right]e_1'\left(\frac{1}{n}\mathbb{X}_j'\mathbb{K}_j\mathbb{X}_j\right)^{-1}\mathbb{X}_{i,j}K_{i,j} \\
& \quad \times\left((Z_i-p_i)+(X_i-X_j)'\,\frac{1}{2}\frac{\partial^2 p\,(X_j)}{\partial x\partial x'}\,(X_i-X_j)+O\left(h^3\right)\right)\ |X_i,Z_i\Big] \\
= \ & E\Big[\frac{\pi(X_j,1)}{p_j}\vartheta_{11}(X_j)e_1'\left(\frac{1}{n}\mathbb{X}_j'\mathbb{K}_j\mathbb{X}_j\right)^{-1}\mathbb{X}_{i,j}K_{i,j}\left((Z_i-p_i)+(X_i-X_j)'\,\frac{1}{2}\frac{\partial^2 p\,(X_j)}{\partial x\partial x'}\,(X_i-X_j)+O\left(h^3\right)\right)\ |X_i,Z_i\Big]
\end{aligned}
$$

where ϑ_{11} was defined in (26). Now we enter (49) to obtain

$$\begin{aligned}
&= E\Big[\frac{\pi(X_j,1)}{p_j}\vartheta_{11}(X_j)\left(1+O(h)-h\frac{(\lambda-2)!}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(X_j)/\partial x^{\lambda-1}{}'}{\partial^{\lambda-2}f(X_j)/\partial x^{\lambda-2}}\frac{X_i-X_j}{h}(1+O(h))\right)\frac{K_{i,j}}{f(X_j)}\\
&\qquad\times\left((Z_i-p_i)+(X_i-X_j)'\frac{1}{2}\frac{\partial^2 p(X_j)}{\partial x\partial x'}(X_i-X_j)+O\left(h^3\right)\right)\ |X_i,Z_i\Big]\\
&= \int\frac{\pi(X_i-vh,1)}{p(X_i-vh)}\vartheta_{11}(X_i-vh)\left(1+O(h)-h\frac{(\lambda-2)!}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(X_i-vh)/\partial x^{\lambda-1}{}'}{\partial^{\lambda-2}f(X_i-vh)/\partial x^{\lambda-2}}v(1+O(h))\right)\\
&\qquad\times\left((Z_i-p_i)+\frac{h^2}{2}v'\frac{\partial^2 p(X_i-vh)}{\partial x\partial x'}v+O\left(h^3\right)\right)\prod_{l=1}^{L}\kappa(v_l)\,dv
\end{aligned}$$

where $v=\frac{X_i-X_j}{h}$ and by bounded convergence. With $\partial^2 p(x)/\partial x\partial x'$ Hölder continuous, the term Z_i-p_i clearly dominates the last expression and we obtain by bounded convergence

$$=\left(\frac{\pi(X_i,1)}{p(X_i)}\vartheta_{11}(X_i)+O(h)\right)\cdot(Z_i-p_i). \tag{52}$$

For an application of the projection theorem in (44) we need to show that $E[\varsigma_{ij}^2]\leq o(n)$. Therefore, consider the term

$$\begin{aligned}
&E\Big[\Big[\frac{Z_iD_i}{p_i^2}\left(1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-\tau\right)e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{j,i}\\
&\qquad\times\left((Z_j-p_j)+\frac{1}{2}(X_j-X_i)'\frac{\partial^2 p(X_i)}{\partial x\partial x'}(X_j-X_i)+O\left(h^3\right)\right)\Big]^2\Big]\\
=\ &E\Big[\frac{Z_iD_i}{p_i^4}\left(1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-\tau\right)^2\Big[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{j,i}\\
&\qquad\times\left((Z_j-p_j)+\frac{1}{2}(X_j-X_i)'\frac{\partial^2 p(X_i)}{\partial x\partial x'}(X_j-X_i)+O\left(h^3\right)\right)\Big]^2\Big]\\
=\ &E\Big[\frac{\pi(X_i,1)}{p^3(X_i)}E\left[\left(1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-\tau\right)^2|X_1,...,X_n,Z_i=1,D_i=1\right]\\
&\qquad\times E\Big[\left(e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{j,i}\left((Z_j-p_j)+(X_j-X_i)'\frac{1}{2}\frac{\partial^2 p(X_i)}{\partial x\partial x'}(X_j-X_i)+O\left(h^3\right)\right)\right)^2|X_i\Big]\Big]\\
=\ &o(n)
\end{aligned}\tag{53}$$

because of (54) and since $p(x)$ is bounded away from zero over its support as has been assumed. Here we

have used that

$$
\begin{aligned}
& E\left[\left[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{j,i}\left(\left(Z_j-p_j\right)+\left(X_j-X_i\right)'\frac{1}{2}\frac{\partial^2 p\left(X_i\right)}{\partial x\partial x'}\left(X_j-X_i\right)+O\left(h^3\right)\right)\right]^2|X_i\right] \\
= & E\left[\left(e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{j,i}\right)^2\left(p_j\left(1-p_j\right)+h^4\left(\frac{X_j-X_i}{h}'\frac{1}{2}\frac{\partial^2 p\left(X_i\right)}{\partial x\partial x'}\frac{X_j-X_i}{h}+O\left(h\right)\right)^2\right)|X_i\right] \\
= & E\left[\left(1+O(h)-h\frac{\partial^{\lambda-1}f(X_i)/\partial x^{\lambda-1}{}'}{\partial^{\lambda-2}f(X_i)/\partial x^{\lambda-2}}\frac{X_j-X_i}{h}\frac{(\lambda-2)!}{(\lambda-1)!}\right)^2\frac{K_{j,i}^2}{f^2(X_i)}\right. \\
& \qquad \left.\times\left(p_j\left(1-p_j\right)+h^4\left(\left(\frac{X_j-X_i}{h}\right)'\frac{1}{2}\frac{\partial^2 p\left(X_i\right)}{\partial x\partial x'}\left(\frac{X_j-X_i}{h}\right)+O\left(h\right)\right)^2\right)|X_i\right] \\
= & \int\frac{1}{h^{2L}}\left(1+O(h)-h\frac{\partial^{\lambda-1}f(X_i)/\partial x^{\lambda-1}{}'}{\partial^{\lambda-2}f(X_i)/\partial x^{\lambda-2}}u\frac{(\lambda-2)!}{(\lambda-1)!}\right)^2\frac{1}{f^2(X_i)}\prod_{l=1}^{L}\kappa^2\left(u_l\right) \\
& \qquad \times\left(p\left(X_i+uh\right)\left(1-p\left(X_i+uh\right)\right)+h^4\left(u'\frac{1}{2}\frac{\partial^2 p\left(X_i\right)}{\partial x\partial x'}u+O\left(h\right)\right)^2\right)f(X_i+uh)h^L du
\end{aligned}
$$

where $u=\frac{X_j-X_i}{h}$ and by bounded convergence

$$
=\frac{1}{h^L}\frac{p\left(X_i\right)\left(1-p\left(X_i\right)\right)}{f(X_i)}\bar{\mu}_0^L\left(1+O(h)\right)=O\left(\frac{1}{h^L}\right)=o(n) \tag{54}
$$

as it has been assumed that $nh^L\to\infty$.

C Properties of local logit

Define the log likelihood function for local logit regression at a location x_0 as

$$
\ln\mathcal{L}_n(x_0,a,b)=\frac{1}{n}\sum_{j=1}^{n}\left\{Z_j\ln\Lambda\left(a+b'\left(X_j-x_0\right)\right)+\left(1-Z_j\right)\ln\left(1-\Lambda\left(a+b'\left(X_j-x_0\right)\right)\right)\right\}\cdot K_j
$$

where $\Lambda(x)=\frac{1}{1+e^{-x}}$. We will denote derivativesof $\Lambda(x)$ by $\Lambda'(x)$, $\Lambda''(x)$, $\Lambda^{(3)}(x)$ etc. and also note that $\Lambda'(x)=\Lambda(x)\cdot(1-\Lambda(x))$. Let $\hat{a}$ and $\hat{b}$ be the maximizers of $\ln L_n(x_0,a,b)$ and a_0 and b_0 be the values that maximize the expected value of the likelihood function $E\left[\ln\mathcal{L}_n(x_0,a,b)\right]$. We will sometimes write $\hat{a}(x_0)$, $\hat{b}(x_0)$, $a_0(x_0)$ and $b_0(x_0)$ to make it explicit that these coefficients are different for different values of x_0. At other times we supress the dependence to ease notation and to focus attention on the properties at a particular x_0.

We estimate $p(x_0)$ by $\hat{p}(x_0)=\Lambda(\hat{a}(x_0))$. In the following we will also show that $\Lambda(a_0(x_0))$ is identical to $p(x_0)$ up to an $O(h^\lambda)$ term.

To derive this, note that since the likelihood function is globally convex, the maximizers are obtained by setting the first order conditions to zero. The values of $a_0(x_0)$ and $b_0(x_0)$ are thus implicitly defined by

the moment conditions

$$E\left[\left(Z_j-\Lambda\left(a_0+b_0'\left(X_j-x_0\right)\right)\right)\begin{pmatrix}1\\X_j-x_0\end{pmatrix}K_j\right]=0$$

$$=E\left[\left(p_j-\Lambda\left(a_0+b_0'\left(X_j-x_0\right)\right)\right)\begin{pmatrix}1\\X_j-x_0\end{pmatrix}K_j\right]=0. \tag{55}$$

Now examine only the first moment condition to obtain

$$\begin{aligned}0 &= \int\left(p(X_j)-\Lambda\left(a_0+b_0'\left(X_j-x_0\right)\right)\right)\cdot K_j\cdot f(X_j)dX_j\\ &= \int\left(p(x_0+uh)-\Lambda\left(a_0+b_0'uh\right)\right)\prod_{l=1}^{L}\kappa\left(u_l\right)f(x_0+uh)du\end{aligned}$$

where $u=\frac{X_j-x_0}{h}$. Now assuming that p is λ times differentiable and noting that the kernel is of order λ we obtain by Taylor expansion that

$$\left(p(x_0)-\Lambda\left(a_0\right)\right)f(x_0)+O(h^\lambda)=0$$

hence

$$p(x_0)=\Lambda\left(a_0\right)+O(h^\lambda).$$

Combining this with the previous results we thus have obtained an expression for $\hat{p}(x_0)-p(x_0)$

$$\hat{p}(x_0)-p(x_0)=\Lambda(\hat{a}(x_0))-\Lambda(a_0(x_0))+O(h^\lambda)$$

and by Taylor expansion of $\Lambda(\hat{a})$ which converges to $\Lambda(a_0)$

$$\hat{p}(x_0)-p(x_0)=(\hat{a}(x_0)-a_0(x_0))\cdot\Lambda'(a_0(x_0))\cdot(1+o_p(1))+O(h^\lambda).$$

(One could also explicitly consider the second order term, but for sake of brevity we omit this here.)

By entering (57) and (59) we obtain

$$\begin{aligned}&=\Lambda'(a_0(x_0))\cdot e_1'\left(\frac{1}{n}\sum\left\{\Lambda'(\boldsymbol{\beta}_0'\mathbb{X}_j)+\Lambda''(\boldsymbol{\beta}_0'\mathbb{X}_j)\mathbb{X}_j(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'+O_p(||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||^2\right\}\mathbb{X}_j\mathbb{X}_j'K_j\right)^{-1}\\ &\qquad\times\frac{1}{n}\sum\left(Z_j-p_j+p_j-\Lambda(a_0+b_0'\left(X_j-x_0\right))\right)K_j\mathbb{X}_j\cdot(1+o_p(1))+O(h^\lambda)\end{aligned}$$

where we defined $\boldsymbol{\beta}=(a,hb')'$ and $\mathbb{X}_j=\left(1,\left(\frac{X_j-x_0}{h}\right)'\right)'$ to obtain

$$\begin{aligned}&=\frac{1}{f(x_0)}\begin{pmatrix}1\\-h\frac{(\lambda-2)!}{(\lambda-1)!}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(x_0)\right)}{\partial x_1^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(x_0)\right)}{\partial x_1^{\lambda-2}}\right)\\ \vdots\\ -h\frac{(\lambda-2)!}{(\lambda-1)!}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(x_0)\right)}{\partial x_L^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(x_0)\right)}{\partial x_L^{\lambda-2}}\right)\end{pmatrix}'\\ &\qquad\times\frac{1}{n}\sum\left(Z_j-p_j+p_j-\Lambda(a_0+b_0'\left(X_j-x_0\right))\right)K_j\mathbb{X}_j\cdot(1+o_p(1))+O(h^\lambda),\end{aligned}$$

where $\partial^{\lambda}\left(\Lambda' f(x_0)\right)/\partial x_1^{\lambda}$ is defined in (58)

Now we can embark to show that (43) equals (45) when we estimate $\hat{p}_i$ by local logit regression. Analogously to Section B.1 we can write (43) as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i D_i}{p_i^2}\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right)-\tau\right)(\hat{p}_i - p_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varsigma_{ij}$$

where

$$\varsigma_{ij} = \frac{Z_i D_i}{p^2(X_i)}\left(1\left(Y_i < Q^{\tau}_{Y^1|c}\right)-\tau\right)\frac{1}{f(X_i)}\begin{pmatrix} 1 \\ -h\frac{(\lambda-2)!}{(\lambda-1)!}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_1^{\lambda-1}}\Big/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_1^{\lambda-2}}\right) \\ \vdots \\ -h\frac{(\lambda-2)!}{(\lambda-1)!}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_L^{\lambda-1}}\Big/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_L^{\lambda-2}}\right) \end{pmatrix}'$$
$$\times K_{j,i}\mathbb{X}_{j,i}\cdot\left(Z_j - p_j + p_j - \Lambda(a_0(X_i) + b_0'(X_i)(X_j - X_i))\right)\cdot(1+o_p(1)) + O(h^{\lambda}).$$

By (60) we obtain that $E\left[\varsigma_{ij}|X_i, Z_i, D_i, Y_i\right] = O_p(h^{\lambda})$ and by (61) that $E\left[\varsigma_{ji}|X_i, Z_i, D_i, Y_i\right]$

$$= \frac{\pi(X_i,1)\vartheta_{11}(X_i)}{p(X_i)}\left(Z_i - p(X_i)\right)(1+o_p(1)) + O_p(h^{\lambda}). \tag{56}$$

With these two results, and noting that $E[\varsigma_{ij}^2] \leq o(n)$ by (62), we obtain essentially the same expression as in (45), and we can apply the same derivations as in the proof of Theorem 9.

C.1 Further properties of local logit

C.1.1 Properties of $\hat{a}$

Now we need to examine $\hat{a}$ in more detail. Define first $\boldsymbol{\beta} = (a, hb')'$ and $\mathbb{X}_j = \left(1, \left(\frac{X_j - x_0}{h}\right)'\right)'$. The first order condition of the estimator is given by

$$0 = \frac{1}{n}\sum\left(Z_j - \Lambda(\hat{\boldsymbol{\beta}}'\mathbb{X}_j)\right)K_j\mathbb{X}_j'$$

$$= \frac{1}{n}\sum\left(Z_j - \Lambda(\boldsymbol{\beta}_0'\mathbb{X}_j) - \Lambda'(\boldsymbol{\beta}_0'\mathbb{X}_j)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\mathbb{X}_j - \Lambda''(\boldsymbol{\beta}_0'\mathbb{X}_j)\cdot(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\mathbb{X}_j\mathbb{X}_j'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) - O_p(||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||^3)\right)K_j\mathbb{X}_j'$$

by Taylor expansion. Further

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 = \left(\frac{1}{n}\sum\left\{\Lambda'(\boldsymbol{\beta}_0'\mathbb{X}_j) + \Lambda''(\boldsymbol{\beta}_0'\mathbb{X}_j)\mathbb{X}_j(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)' + O_p\left(||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||^2\right)\right\}\mathbb{X}_j\mathbb{X}_j'K_j\right)^{-1}\frac{1}{n}\sum\left(Z_j - \Lambda(\boldsymbol{\beta}_0'\mathbb{X}_j)\right)K_j\mathbb{X}_j.$$

As we are only interested in $\hat{a}$ and not in $\hat{b}$ we write

$$\hat{a} - a_0 = e_1'\left(\frac{1}{n}\sum\left\{\Lambda'(\boldsymbol{\beta}_0'\mathbb{X}_j) + \Lambda''(\boldsymbol{\beta}_0'\mathbb{X}_j)\mathbb{X}_j(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)' + O_p\left(||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||^2\right)\right\}\mathbb{X}_j\mathbb{X}_j'K_j\right)^{-1}$$
$$\times\frac{1}{n}\sum\left(Z_j - \Lambda(a_0 + b_0'(X_j - x_0))\right)K_j\mathbb{X}_j. \tag{57}$$

C.1.2 Denominator for local logit

We start with an approximation to the term

$$\frac{1}{n}\sum\left\{\Lambda'(\boldsymbol{\beta}_0'\mathbb{X}_j)+\Lambda''(\boldsymbol{\beta}_0'\mathbb{X}_j)\mathbb{X}_j(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'+O_p\left(||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||^2\right)\right\}\mathbb{X}_j\mathbb{X}_j'K_j.$$

Under the assumption that $nh^L\to\infty$ and $h\to 0$, which implies consistency of $\hat{a}$ and $\hat{b}$, one can show that for a kernel of order λ

$$=\begin{bmatrix} f(x_0)\Lambda'(a_0) & h^{\lambda-1}\frac{\mu_\lambda}{(\lambda-1)!}\frac{\partial^{\lambda-1}\left(\Lambda' f(x_0)\right)}{\partial x_1^{\lambda-1}} & \cdots & \cdots \\ h^{\lambda-1}\frac{\mu_\lambda}{(\lambda-1)!}\frac{\partial^{\lambda-1}\left(\Lambda' f(x_0)\right)}{\partial x_1^{\lambda-1}} & h^{\lambda-2}\frac{\mu_\lambda}{(\lambda-2)!}\frac{\partial^{\lambda-2}\left(\Lambda' f(x_0)\right)}{\partial x_1^{\lambda-2}} & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ \vdots & 0 & 0 & \ddots \end{bmatrix}(1+o_p(1))$$

where $\partial^\lambda\left(\Lambda' f(x_0)\right)/\partial x_l^\lambda$ is a shortcut notation for all the cross derivatives of Λ' and $f(x_0)$

$$\frac{\partial^\lambda\left(\Lambda' f(x_0)\right)}{\partial x_l^\lambda}\equiv\sum_{r=0}^{\lambda}\Lambda^{(r+1)}(a_0(x_0))\cdot\frac{\partial^{\lambda-r}f(x_0)}{\partial x_l^{\lambda-r}}. \tag{58}$$

The derivations are similar to those of Section B.3 and are omitted here. An additional complication compared to Section B.3 are the second order terms, which however are all of lower order when $(\hat{a}-a_0)$ and $(\hat{b}-b_0)$ are $o_p(1)$.

Similarly to Section B.3 we can now derive

$$e_1'\left(\frac{1}{n}\sum\left\{\Lambda'(\boldsymbol{\beta}_0'\mathbb{X}_j)+\Lambda''(\boldsymbol{\beta}_0'\mathbb{X}_j)\mathbb{X}_j(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'+O_p\left(||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||^2\right)\right\}\mathbb{X}_j\mathbb{X}_j'K_j\right)^{-1}$$

$$=\frac{1}{f(x_0)\Lambda'(a_0(x_0))}\begin{pmatrix} 1 \\ -h\frac{(\lambda-2)!}{(\lambda-1)!}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(x_0)\right)}{\partial x_1^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(x_0)\right)}{\partial x_1^{\lambda-2}}\right) \\ \vdots \\ -h\frac{(\lambda-2)!}{(\lambda-1)!}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(x_0)\right)}{\partial x_L^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(x_0)\right)}{\partial x_L^{\lambda-2}}\right) \end{pmatrix}'(1+o_p(1)) \tag{59}$$

C.1.3 Preliminaries for the U-statistics projection theorem

As a preliminary for the application of the projection theorem we calculate first $E\left[\varsigma_{ij}|X_i,Z_i,D_i,Y_i\right]=$

$$\int \frac{Z_iD_i}{p^2(X_i)}\left(1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-\tau\right)\frac{1}{f(X_i)}\begin{pmatrix}1\\ -h\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_1^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_1^{\lambda-2}}\right)\frac{(\lambda-2)!}{(\lambda-1)!}\\ \vdots\\ -h\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_L^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_L^{\lambda-2}}\right)\frac{(\lambda-2)!}{(\lambda-1)!}\end{pmatrix}'$$
$$\times K_{j,i}\mathbb{X}_{j,i}\cdot\left(Z_j-p(X_j)+p(X_j)-\Lambda(a_0(X_i)+b_0'(X_i)\left(X_j-X_i\right))\right)\cdot\left(1+o_p(1)\right)+O_p(h^{\lambda})$$
$$\times dF(X_j,Z_j)$$

$$=\frac{Z_iD_i}{p^2(X_i)}\frac{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-\tau}{f(X_i)}\int\begin{pmatrix}1\\ -h\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_1^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_1^{\lambda-2}}\right)\frac{(\lambda-2)!}{(\lambda-1)!}\\ \vdots\\ -h\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_L^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_L^{\lambda-2}}\right)\frac{(\lambda-2)!}{(\lambda-1)!}\end{pmatrix}'$$
$$\times\mathbb{X}_{j,i}K_{j,i}\cdot\left(p(X_j)-\Lambda(a_0(X_i)+b_0'(X_i)\left(X_j-X_i\right))\right)\cdot\left(1+o_p(1)\right)dF(X_j)+O(h^{\lambda})$$

$$=\frac{Z_iD_i}{p^2(X_i)}\frac{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-\tau}{f(X_i)}$$
$$\times\int\left\{1-h\frac{(\lambda-2)!}{(\lambda-1)!}\sum_{l=1}^{L}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_l^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_l^{\lambda-2}}\right)\left(\frac{X_{jl}-X_{il}}{h}\right)\right\}$$
$$\times K_{j,i}\cdot\left(p(X_j)-\Lambda(a_0(X_i)+b_0'(X_i)\left(X_j-X_i\right))\right)\cdot\left(1+o_p(1)\right)dF(X_j)+O_p(h^{\lambda})$$

$$=\frac{Z_iD_i}{p^2(X_i)}\frac{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-\tau}{f(X_i)}\times\int\left\{1-h\frac{(\lambda-2)!}{(\lambda-1)!}\sum_{l=1}^{L}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_l^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_l^{\lambda-2}}\right)u_l\right\}$$
$$\cdot\left(p(X_i+uh)-\Lambda(a_0(X_i)+uhb_0'(X_i))\right)\cdot f(X_i+uh)\left(1+o_p(1)\right)\prod_{l=1}^{L}\kappa\left(u_l\right)du+O_p(h^{\lambda})$$

where $u=\frac{X_j-X_i}{h}$ and by Taylor series expansion we obtain, using that $p(X_i)=\Lambda(a_0(X_i))+O_p(h^{\lambda})$ as has beeen derived above

$$=\frac{Z_iD_i}{p^2(X_i)}\frac{1\left(Y_i<Q^{\tau}_{Y^1|c}\right)-\tau}{f(X_i)}O_p(h^{\lambda})=O_p(h^{\lambda}).\tag{60}$$

Now we calculate $E\left[\varsigma_{ji}|X_i,Z_i,D_i,Y_i\right]=$

$$\int \frac{Z_jD_j}{p^2(X_j)}\frac{1\left(Y_j<Q^{\tau}_{Y^1|c}\right)-\tau}{f(X_j)}\begin{pmatrix}1\\ -h\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_j)\right)}{\partial x_1^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_j)\right)}{\partial x_1^{\lambda-2}}\right)\frac{(\lambda-2)!}{(\lambda-1)!}\\ \vdots\\ -h\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_j)\right)}{\partial x_L^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_j)\right)}{\partial x_L^{\lambda-2}}\right)\frac{(\lambda-2)!}{(\lambda-1)!}\end{pmatrix}'$$
$$\times K_{i,j}\mathbb{X}_{i,j}\cdot\left(Z_i-p_i+p_i-\Lambda(a_0(X_j)+b_0'(X_j)\left(X_i-X_j\right))\right)\cdot\left(1+o_p(1)\right)+O_p(h^{\lambda})$$
$$\times dF(Y_j,D_j,X_j,Z_j)$$

and by conditioning on all $X_1,...,X_n$ via iterated expectations we obtain

$$=\int\frac{\pi(X_j,1)\vartheta_{11}(X_j)}{p(X_j)f(X_j)}\times\left\{1-h\frac{(\lambda-2)!}{(\lambda-1)!}\sum_{l=1}^{L}\frac{\partial^{\lambda-1}\left(\Lambda' f(X_j)\right)/\partial x_l^{\lambda-1}}{\partial^{\lambda-2}\left(\Lambda' f(X_j)\right)/\partial x_l^{\lambda-2}}\left(\frac{X_{il}-X_{jl}}{h}\right)\right\}$$
$$\times K_{i,j}\cdot\left(Z_i-p_i+p_i-\Lambda(a_0(X_j)+b_0'(X_j)\left(X_i-X_j\right))\right)\cdot\left(1+o_p(1)\right)+O_p(h^{\lambda})$$
$$\times dF(X_j)$$

$$=\int\frac{\pi(X_i-vh,1)\vartheta_{11}\left(X_i-vh\right)}{p\left(X_i-vh\right)}\times\left\{1-h\frac{(\lambda-2)!}{(\lambda-1)!}\sum_{l=1}^{L}v_l\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i-vh)\right)/\partial x_l^{\lambda-1}}{\partial^{\lambda-2}\left(\Lambda' f(X_i-vh)\right)/\partial x_l^{\lambda-2}}\right\}$$
$$\times\left(Z_i-p(X_i)+p(X_i)-\Lambda(a_0\left(X_i-vh\right)+vhb_0'(X_i-vh))\right)\cdot\left(1+o_p(1)\right)+O_p(h^{\lambda})$$
$$\times\prod_{l=1}^{L}\kappa\left(v_l\right)dv$$

where $v=\frac{X_i-X_j}{h}$. Note that the term Z_i-p_i clearly dominates this expression because $p(X_i)-\Lambda(a_0(X_i))=O_p(h^{\lambda})$ and all other terms are multiplied by h or powers of it. We thus obtain

$$=\frac{\pi(X_i,1)\vartheta_{11}\left(X_i\right)}{p\left(X_i\right)}\left(Z_i-p(X_i)\right)\left(1+o_p(1)\right)+O_p(h^{\lambda}).\tag{61}$$

As a last element for applying the U-statistics projection we need to show that $E[\varsigma_{ij}^2]\leq o(n)$. The key element is to show that the following term is $o_p(n)$

$$E\Bigg[\Bigg[\frac{1}{f(X_i)}\begin{pmatrix}1\\ -h\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_1^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_1^{\lambda-2}}\right)\frac{(\lambda-2)!}{(\lambda-1)!}\\ \vdots\\ -h\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_L^{\lambda-1}}/\frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_L^{\lambda-2}}\right)\frac{(\lambda-2)!}{(\lambda-1)!}\end{pmatrix}'\cdot\mathbb{X}_{j,i}K_{j,i}$$
$$\times\left(Z_j-p(X_j)+p(X_j)-\Lambda(a_0(X_i)+b_0'(X_i)\left(X_j-X_i\right))\right)\cdot\left(1+o_p(1)\right)+O_p(h^{\lambda})\Bigg]^2\ |X_i\Bigg]$$

$$= E\Bigg[\Bigg\{1 - h\frac{(\lambda-2)!}{(\lambda-1)!}\sum_{l=1}^{L}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_l^{\lambda-1}} / \frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_l^{\lambda-2}}\right)\left(\frac{X_{jl}-X_{il}}{h}\right)\Bigg\}^2 \frac{K_{j,i}^2}{f^2(X_i)}$$
$$\times\left[\left(Z_j - p(X_j)\right)^2 + \left(p(X_j) - \Lambda(a_0(X_i) + b_0'(X_i)\left(X_j - X_i\right))\right)^2\right]\cdot\left(1+o_p(1)\right)\ \Big|X_i\Bigg]$$

$$= E\Bigg[\Bigg\{1 - h\frac{(\lambda-2)!}{(\lambda-1)!}\sum_{l=1}^{L}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_l^{\lambda-1}} / \frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_l^{\lambda-2}}\right)\left(\frac{X_{jl}-X_{il}}{h}\right)\Bigg\}^2 \frac{K_{j,i}^2}{f^2(X_i)}$$
$$\times\left[p(X_j)\left(1-p(X_j)\right) + \left(p(X_j) - \Lambda(a_0(X_i) + b_0'(X_i)\left(X_j - X_i\right))\right)^2\right]\cdot\left(1+o_p(1)\right)\ \Big|X_i\Bigg]$$

$$= \frac{1}{h^L}\frac{1}{f^2(X_i)}\int\left\{1 - h\frac{(\lambda-2)!}{(\lambda-1)!}\sum_{l=1}^{L}\left(\frac{\partial^{\lambda-1}\left(\Lambda' f(X_i)\right)}{\partial x_l^{\lambda-1}} / \frac{\partial^{\lambda-2}\left(\Lambda' f(X_i)\right)}{\partial x_l^{\lambda-2}}\right)u_l\right\}^2$$
$$\times\left[p(X_i+uh)\left(1-p(X_i+uh)\right) + \left(p(X_i+uh) - \Lambda(a_0(X_i) + uhb_0'(X_i))\right)^2\right]\cdot\ f(X_i+uh)$$
$$\times\prod_{l=1}^{L}\kappa^2\left(v_l\right)dv\cdot\left(1+o_p(1)\right)$$

where $u = \frac{X_j - X_i}{h}$ and we obtain

$$= \frac{\bar{\mu}_0^L}{h^L}\frac{p(X_i)\left(1-p(X_i)\right)}{f(X_i)}\cdot\left(1+o_p(1)\right) = o_p(n) \tag{62}$$

as it has been assumed that $nh^L \to \infty$.

D Additional lemmas used for the proofs

Here we restate the first two lemmas of Hjort and Pollard (1993) for the convenience of the reader.

Lemma 1 of Hjort and Pollard (1993): **From pointwise to uniform**. Suppose $A_n(s)$ is a sequence of convex random functions defined on an open convex set $\mathcal{S}$ in $\mathbb{R}^p$, which converges in probability to some $A(s)$, for each s. Then $\sup_{s\in K}|A_n(s) - A(s)|$ goes to zero in probability, for each compact subset K of $\mathcal{S}$.

Lemma 2 of Hjort and Pollard (1993): **Nearness of argmins**. Suppose $A_n(s)$ is convex as in Lemma 1 and is approximated by $B_n(s)$. Let α_n be the argmin of A_n, and assume that B_n has a unique argmin β_n. Then there is a probabilistic bound on how far α_n can be from β_n. For each $\delta > 0$,

$$\Pr\left(|\alpha_n - \beta_n| \geq \delta\right)$$
$$\leq \Pr\left\{2\cdot\sup_{|s-\beta_n|\leq\delta}|A_n(s) - B_n(s)|\ \geq\ \inf_{|s-\beta_n|=\delta}B_n(s) - B_n(\beta_n)\right\}.$$

E Proofs for the efficiency bounds:

E.1 Proof of Theorem (9):

Semiparametric efficiency bounds were introduced by Stein (1956) and developed by Koshevnik and Levit (1976), Pfanzagl and Wefelmeyer (1982), Begun, Hall, Huang, and Wellner (1983) and Bickel, Klaassen, Ritov, and Wellner (1993). See also the survey of Newey (1990) or Newey (1994).

We need to derive the efficiency bound for

$$\Delta_c^\tau = Q^\tau_{Y^1|c} - Q^\tau_{Y^0|c}.$$

For this it will be helpful to have an expression for $f_{Y^1|c}$ and $f_{Y^0|c}$. From Theorem (1) it follows that

$$f_{Y^1|c}(u) = \left\{ \int \left(f_{Y|X,D=1,Z=1}(u)\pi(x,1) - f_{Y|X,D=1,Z=0}(u)\pi(x,0) \right) dF_X \right\} / P_c \tag{63}$$

$$f_{Y^0|c}(u) = -\left\{ \int \left(f_{Y|X,D=0,Z=1}(u)\left(1-\pi(x,1)\right) - f_{Y|X,D=0,Z=0}(u)\left(1-\pi(x,0)\right) \right) dF_X \right\} / P_c \tag{64}$$

where $\pi(x,z) = \Pr(D=1|X=x,Z=z)$ and $P_c = \int (\pi(x,1)-\pi(x,0))\, dF_X$ is the fraction of compliers.

By **Assumption 2** the quantiles $Q^\tau_{Y^1|c}$ and $Q^\tau_{Y^0|c}$ are unique and defined as

$$0 = E\left[1(Y^1 \le Q^\tau_{Y^1|c}) - \tau | T = c\right] = \int \left(1(u \le Q^\tau_{Y^1|c}) - \tau\right) \cdot f_{Y^1|c}(u) du \tag{65}$$

$$0 = E\left[1(Y^0 \le Q^\tau_{Y^0|c}) - \tau | T = c\right] = \int \left(1(u \le Q^\tau_{Y^0|c}) - \tau\right) \cdot f_{Y^0|c}(u) du$$

where $f_{Y^d|c}$ are given above. We thus have expressed the quantiles in terms of the densities of the observed variables.

The joint density of the observed variables (Y, D, Z, X) with D and Z binary can be written as

$$f(y,d,z,x) = f(y|d,z,x)\, f(d|z,x)\, f(z|x)\, f(x)$$

$$= f(y|d,z,x) \left\{ \pi(x,z)^d \cdot (1-\pi(x,z))^{1-d} \right\} \left\{ p(x)^z \cdot (1-p(x))^{1-z} \right\} f(x).$$

Consider a regular parametric submodel indexed by θ with θ_0 corresponding to the true model: $f(y,d,z,x;\theta_0) = f(y,d,z,x)$. The density $f(y,d,z,x;\theta)$ can be written as

$$f(y,d,z,x;\theta) = f^{11}(y|x;\theta)^{dz} \cdot f^{10z}(y|x;\theta)^{d(1-z)} \cdot f^{01}(y|x;\theta)^{(1-d)z} \cdot f^{00}(y|x;\theta)^{(1-d)(1-z)}$$
$$\left\{ \pi(x,z;\theta)^d \cdot (1-\pi(x,z;\theta))^{1-d} \right\} \left\{ p(x;\theta)^z \cdot (1-p(x;\theta))^{1-z} \right\} f(x;\theta),$$

where $f^{dz}(y|x;\theta) = f(y|d,z,x;\theta)$.

We will assume throughout that all terms of the previous equation admit an interchange of the order of integration and differentiation, such that

$$\int \frac{\partial f(y,d,z,x;\theta)}{\partial \theta} dydddzdx = \frac{\partial}{\partial \theta} \int f(y,d,z,x;\theta)\, dydddzdx = 0.$$

Sufficient conditions for permitting interchanging differentiation and integration are, for example, given by Theorem 1.3.2 of Amemiya (1985).

The corresponding score of $f(y,d,z,x;\theta)$ is

$$\begin{aligned} s(y,d,z,x;\theta) &= \frac{\partial \ln f(y,d,z,x;\theta)}{\partial\theta} \\ &= dz\breve{f}^{11}(y|x;\theta) + d(1-z)\breve{f}^{10}(y|x;\theta) + (1-d)z\breve{f}^{01}(y|x;\theta) + (1-d)(1-z)\breve{f}^{00}(y|x;\theta) \\ &+ \frac{d-\pi(x,z;\theta)}{1-\pi(x,z;\theta)}\breve{\pi}(x,z,\theta) + \frac{z-p(x;\theta)}{1-p(x;\theta)}\breve{p}(x,\theta) + \breve{f}(x;\theta), \end{aligned}$$

where the subscript $\breve{f}$ defines a derivative of the log, i.e. $\breve{f}(x;\theta) = \partial \ln f(x;\theta)/\partial\theta$.

At the true value θ_0 the expectation of the score is zero. The tangent space of the model is the set of functions that are mean zero and satisfy the additive structure of the score:

$$\Im = \left\{ \begin{array}{c} dzs^{11}(y|x) + d(1-z)s^{10}(y|x) + (1-d)zs^{01}(y|x) + (1-d)(1-z)s^{00}(y|x) \\ +(d-\pi(x,z))\cdot s_\pi(x,z) + (z-p(x))\cdot s_p(x) + s_x(x) \end{array} \right\} \tag{66}$$

for any functions $s^{11}, s^{10}, s^{01}, s^{00}, s_x$ satisfying the mean-zero property: $E\left[s^{dz}|D=d,Z=z,X\right] = 0 = E[s_x(x)]$ and $s_\pi(x,z)$ and $s_p(x)$ being square-integrable measurable functions.

The *semiparametric variance bound* of Δ_c^τ is the variance of the projection on $\Im$ of a function $\psi(Y,D,Z,X)$ (with $E[\psi]=0$ and $E[\|\psi(\cdot)\|^2]<\infty$) that satisfies for all regular parametric submodels

$$\frac{\partial\Delta_c^\tau(F_\theta)}{\partial\theta}_{|\theta=\theta_0} = E[\psi(Y,D,Z,X)\cdot s(Y,D,Z,X)]_{|\theta=\theta_0} \tag{67}$$

If ψ itself already lies in the tangent space, the variance bound is given by $E\left[\psi^2\right]$.

As a first step to calculating the variance bound, we need to derive

$$\frac{\partial\Delta_c^\tau(\theta)}{\partial\theta} = \frac{\partial Q^\tau_{Y^1|c}(\theta)}{\partial\theta} - \frac{\partial Q^\tau_{Y^0|c}(\theta)}{\partial\theta}.$$

The identity (65) holds for all submodels θ such that we obtain

$$\frac{\partial}{\partial\theta}\int \left(1(u\le Q^\tau_{Y^1|c}(\theta)) - \tau\right)\cdot f_{Y^1|c}(u;\theta)du = 0$$

$$= (1-\tau)\frac{\partial}{\partial\theta}\int_{-\infty}^{Q^\tau_{Y^1|c}(\theta)} f_{Y^1|c}(u;\theta)du - \tau\frac{\partial}{\partial\theta}\int_{Q^\tau_{Y^1|c}(\theta)}^{\infty} f_{Y^1|c}(u;\theta)du$$

$$= f_{Y^1|c}(Q^\tau_{Y^1|c}(\theta);\theta)\cdot\frac{\partial Q^\tau_{Y^1|c}(\theta)}{\partial\theta} + \int\left(1(u\le Q^\tau_{Y^1|c}(\theta)) - \tau\right)\frac{\partial}{\partial\theta}f_{Y^1|c}(u;\theta)du = 0.$$

by Leibniz's rule of differentiation. We thus obtain that the derivative evaluated at the true θ_0 is

$$\frac{\partial\Delta_c^\tau(\theta)}{\partial\theta}_{|\theta=\theta_0} = \frac{\int\left(\tau - 1(u\le Q^\tau_{Y^1|c})\right)\frac{\partial}{\partial\theta}f_{Y^1|c}(u;\theta_0)du}{f_{Y^1|c}(Q^\tau_{Y^1|c})} - \frac{\int\left(\tau - 1(u\le Q^\tau_{Y^0|c})\right)\frac{\partial}{\partial\theta}f_{Y^0|c}(u;\theta_0)du}{f_{Y^0|c}(Q^\tau_{Y^0|c})}$$

where

$$\frac{\partial}{\partial\theta}f_{Y^1|c}(u;\theta_0) = \frac{1}{P_c}\frac{\partial}{\partial\theta}\left\{\int\left(f_{Y|X,D=1,Z=1}(u)\pi\left(x,1\right) - f_{Y|X,D=1,Z=0}(u)\pi\left(x,0\right)\right)f(x)dx\right\} \\ - f_{Y^1|c}(u;\theta_0)\frac{\partial \ln P_c(\theta_0)}{\partial\theta}$$

such that

$$\frac{\partial\Delta_c^\tau(\theta)}{\partial\theta}_{|\theta=\theta_0} \\ = \frac{\int\left(\tau - 1(u\leq Q^\tau_{Y^1|c})\right)\frac{\partial}{\partial\theta}\left\{\int\left(f_{Y|X,D=1,Z=1}(u)\pi\left(x,1\right) - f_{Y|X,D=1,Z=0}(u)\pi\left(x,0\right)\right)f(x)dx\right\}du}{P_c\cdot f_{Y^1|c}(Q^\tau_{Y^1|c})} \\ + \frac{\int\left(\tau - 1(u\leq Q^\tau_{Y^0|c})\right)\frac{\partial}{\partial\theta}\left\{\int\left(f_{Y|X,D=0,Z=1}(u)\left(1-\pi(x,1)\right) - f_{Y|X,D=0,Z=0}(u)\left(1-\pi(x,0)\right)\right)f(x)dx\right\}du}{P_c\cdot f_{Y^0|c}(Q^\tau_{Y^0|c})}$$

Define

$$\chi_{dz}(y,x) = \frac{\tau - 1(y\leq Q^\tau_{Y^d|c})}{P_c\cdot f_{Y^d|c}(Q^\tau_{Y^d|c})} - \vartheta_{dz}(x) \tag{68}$$

and

$$\vartheta_{dz}(x) = \frac{E\left[\tau - 1(Y\leq Q^\tau_{Y^d|c})|D=d,Z=z,X=x\right]}{P_c\cdot f_{Y^d|c}(Q^\tau_{Y^d|c})} = \frac{\tau - F_{Y|D=d,Z=z,X}(Q^\tau_{Y^d|c})}{P_c\cdot f_{Y^d|c}(Q^\tau_{Y^d|c})} \tag{69}$$

and choose $\psi(Y,D,Z,X)$ as

$$\psi(Y,D,Z,X) = \frac{ZD}{p(X)}\chi_{11}(Y,X) - \frac{(1-Z)D}{1-p(X)}\chi_{10}(Y,X) + \frac{Z(1-D)}{p(X)}\chi_{01}(Y,X) - \frac{(1-Z)(1-D)}{1-p(X)}\chi_{00}(Y,X) \\ + \left(\frac{DZ - \pi(x,1)\left(Z-p(x)\right)}{p(x)}\right)\vartheta_{11}(x) - \left(\frac{D(1-Z) + \pi(x,0)\left(Z-p(x)\right)}{1-p(x)}\right)\vartheta_{10}(x) \\ + \left(\frac{(1-D)Z - (1-\pi(x,1))\left(Z-p(x)\right)}{p(x)}\right)\vartheta_{01}(x) - \left(\frac{(1-D)(1-Z) + (1-\pi(x,0))\left(Z-p(x)\right)}{1-p(x)}\right)\vartheta_{00}(x), \tag{70}$$

which, after some tedious calculations, can be shown to satisfy (67). For this it is helpful to note that in every parametric submodel

$$E\left[\pi(X,1)\cdot\vartheta_{11}(X) - \pi(X,0)\cdot\vartheta_{10}(X)\right] = 0$$
$$E\left[(1-\pi(X,1))\cdot\vartheta_{01}(X) - (1-\pi(X,0))\cdot\vartheta_{00}(X)\right] = 0,$$

which can be derived from the calculations in the proof of Theorem (1).

Since ψ is mean zero and lies in the tangent set (66), the variance bound is

$$E\left[\psi(Y,D,Z,X)^2\right] \tag{71}$$

$$
\begin{aligned}
&= E\left[\left(\frac{ZD}{p(X)}\chi_{11}(Y,X) - \frac{(1-Z)D}{1-p(X)}\chi_{10}(Y,X) + \frac{Z(1-D)}{p(X)}\chi_{01}(Y,X) - \frac{(1-Z)(1-D)}{1-p(X)}\chi_{00}(Y,X)\right)^2\right] \\
&+ E\Big[\Big[\left(\frac{DZ - \pi(x,1)\left(Z-p(x)\right)}{p(x)}\right)\vartheta_{11}(x) - \left(\frac{D(1-Z) + \pi(x,0)\left(Z-p(x)\right)}{1-p(x)}\right)\vartheta_{10}(x) \\
&+ \left(\frac{(1-D)Z - (1-\pi(x,1))\left(Z-p(x)\right)}{p(x)}\right)\vartheta_{01}(x) - \left(\frac{(1-D)(1-Z) + (1-\pi(x,0))\left(Z-p(x)\right)}{1-p(x)}\right)\vartheta_{00}(x)\Big]^2\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[\left(\frac{ZD}{p(X)}\chi_{11}(Y,X)\right)^2\right] + E\left[\left(\frac{(1-Z)D}{1-p(X)}\chi_{10}(Y,X)\right)^2\right] \\
&+ E\left[\left(\frac{Z(1-D)}{p(X)}\chi_{01}(Y,X)\right)^2\right] + E\left[\left(\frac{(1-Z)(1-D)}{1-p(X)}\chi_{00}(Y,X)\right)^2\right] \\
&+ E\Big[\frac{\pi(X,1)}{p(X)}\vartheta_{11}^2(X) + \frac{1-\pi(X,1)}{p(X)}\vartheta_{01}^2(X) + \frac{\pi(X,0)}{1-p(X)}\vartheta_{10}^2(X) + \frac{1-\pi(X,0)}{1-p(X)}\vartheta_{00}^2(X) \\
&- p(X)(1-p(X))\left\{\frac{\pi(X,1)}{p(X)}\vartheta_{11}(X) + \frac{1-\pi(X,1)}{p(X)}\vartheta_{01}(X) + \frac{\pi(X,0)}{1-p(X)}\vartheta_{10}(X) + \frac{1-\pi(X,0)}{1-p(X)}\vartheta_{00}(X)\right\}^2\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{P_c^2 f^2_{Y^1|c}(Q^\tau_{Y^1|c})}E\left[\frac{\pi(X,1)}{p(X)}F_{Y|D=1,Z=1,X}(Q^\tau_{Y^1|c})\left(1 - F_{Y|D=1,Z=1,X}(Q^\tau_{Y^1|c})\right)\right] \\
&+ \frac{1}{P_c^2 f^2_{Y^1|c}(Q^\tau_{Y^1|c})}E\left[\frac{\pi(X,0)}{1-p(x)}F_{Y|D=1,Z=0,X}(Q^\tau_{Y^1|c})\left(1 - F_{Y|D=1,Z=0,X}(Q^\tau_{Y^1|c})\right)\right] \\
&+ \frac{1}{P_c^2 f^2_{Y^0|c}(Q^\tau_{Y^0|c})}E\left[\frac{1-\pi(X,1)}{p(X)}F_{Y|D=0,Z=1,X}(Q^\tau_{Y^0|c})\left(1 - F_{Y|D=0,Z=1,X}(Q^\tau_{Y^0|c})\right)\right] \\
&+ \frac{1}{P_c^2 f^2_{Y^0|c}(Q^\tau_{Y^0|c})}E\left[\frac{1-\pi(X,0)}{1-p(X)}F_{Y|D=0,Z=0,X}(Q^\tau_{Y^0|c})\left(1 - F_{Y|D=0,Z=0,X}(Q^\tau_{Y^0|c})\right)\right] \\
&+ E\Big[\frac{\pi(X,1)\vartheta_{11}^2(X) + (1-\pi(X,1))\,\vartheta_{01}^2(X)}{p(X)} + \frac{\pi(X,0)\vartheta_{10}^2(X) + (1-\pi(X,0))\,\vartheta_{00}^2(X)}{1-p(X)} \\
&- p(X)(1-p(X))\left\{\frac{\pi(X,1)\vartheta_{11}(X) + (1-\pi(X,1))\,\vartheta_{01}(X)}{p(X)} + \frac{\pi(X,0)\vartheta_{10}(X) + (1-\pi(X,0))\,\vartheta_{00}(X)}{1-p(X)}\right\}^2\Big]
\end{aligned}
$$

because

$$
E\left[\left(\frac{DZ}{p(X)}\chi_{11}(Y,X)\right)^2\right] = E\left[E\left[\frac{DZ}{p^2(X)}\chi_{11}^2(Y,X)|X\right]\right] = E\left[E\left[\frac{\pi(X,1)p(X)}{p^2(X)}E\left[\chi_{11}^2(Y,X)|D=Z=1,X\right]|X\right]\right]
$$

and

$$
E\left[\chi_{11}^2(Y,X)|D=Z=1,X\right] = \frac{F_{Y|D=1,Z=1,X}(Q^\tau_{Y^1|c})\left(1 - F_{Y|D=1,Z=1,X}(Q^\tau_{Y^1|c})\right)}{P_c^2 f^2_{Y^1|c}(Q^\tau_{Y^1|c})}
$$

and analogously for the other terms.

E.2 Proof of Lemma (10):

Now we derive the semiparametric efficiency bound when the function $p(x)$ is known. As in the previous theorem, we consider a regular parametric submodel indexed by θ with θ_0 corresponding to the true model:

$f(y,d,z,x;\theta_0) = f(y,d,z,x)$. The difference to the preceding theorem is that $p(x)$ is known and needs not to be estimated. Therefore the propensity score enters as $p(x) \equiv p(x;\theta_0)$ instead of $p(x;\theta)$ in the following density expression. The density $f(y,d,z,x;\theta)$ can be written as

$$f(y,d,z,x;\theta) = f^{11}(y|x;\theta)^{dz} \cdot f^{10z}(y|x;\theta)^{d(1-z)} \cdot f^{01}(y|x;\theta)^{(1-d)z} \cdot f^{00}(y|x;\theta)^{(1-d)(1-z)}$$
$$\left\{\pi(x,z;\theta)^d \cdot (1-\pi(x,z;\theta))^{1-d}\right\}\left\{p(x)^z \cdot (1-p(x))^{1-z}\right\} f(x;\theta),$$

where $f^{dz}(y|x;\theta) = f(y|d,z,x;\theta)$.

The corresponding score of $f(y,d,z,x;\theta)$ is thus identical to the preceeding theorem with the exception that the derivative of $p(x)$ with respect to θ is zero:

$$\begin{aligned} s(y,d,z,x;\theta) &= \frac{\partial \ln f(y,d,z,x;\theta)}{\partial\theta} \\ &= dz\breve{f}^{11}(y|x;\theta) + d(1-z)\breve{f}^{10}(y|x;\theta) + (1-d)z\breve{f}^{01}(y|x;\theta) + (1-d)(1-z)\breve{f}^{00}(y|x;\theta) \\ &+ \frac{d-\pi(x,z;\theta)}{1-\pi(x,z;\theta)}\breve{\pi}(x,z,\theta) + \breve{f}(x;\theta), \end{aligned}$$

where the subscript $\breve{f}$ defines a derivative of the log, i.e. $\breve{f}(x;\theta) = \partial \ln f(x;\theta)/\partial\theta$.

At the true value θ_0 the expectation of the score is zero. The tangent space of the model is the set of functions that are mean zero and satisfy the additive structure of the score:

$$\Im = \left\{ \begin{array}{c} dzs^{11}(y|x) + d(1-z)s^{10}(y|x) + (1-d)zs^{01}(y|x) + (1-d)(1-z)s^{00}(y|x) \\ + (d-\pi(x,z)) \cdot s_\pi(x,z) + s_x(x) \end{array} \right\} \tag{72}$$

for any functions $s^{11}, s^{10}, s^{01}, s^{00}, s_x$ satisfying the mean-zero property: $E\left[s^{dz}|D=d,Z=z,X\right] = 0 = E\left[s_x(x)\right]$ and $s_\pi(x,z)$ being a square-integrable measurable function.

Repeating the calculations of the previous theorem one obtains that the expression for $\frac{\partial \Delta_c^\tau(\theta)}{\partial\theta}|_{\theta=\theta_0}$ is identical to that obtained in the previous proof and thus not affected by knowledge of the propensity score. Now we define $\psi(Y,D,Z,X)$ as in (70) and note that ψ also lies in the tangent set (72). To see that ψ lies in the tangent set (72) it may be helpful to re-write ψ as

$$\begin{aligned} \psi(Y,D,Z,X) &= \frac{ZD}{p(X)}\chi_{11}(Y,X) - \frac{(1-Z)D}{1-p(X)}\chi_{10}(Y,X) + \frac{Z(1-D)}{p(X)}\chi_{01}(Y,X) - \frac{(1-Z)(1-D)}{1-p(X)}\chi_{00}(Y,X) \\ &+ (D-\pi(X,1)) \cdot Z\frac{\vartheta_{11}(X)-\vartheta_{01}(X)}{p(X)} \\ &+ (D-\pi(X,0)) \cdot (1-Z)\frac{\vartheta_{00}(X)-\vartheta_{10}(X)}{1-p(X)} \\ &+ \left[\pi(X,1)\vartheta_{11}(X) - \pi(X,0)\vartheta_{10}(X)\right] \\ &+ \left[(1-\pi(X,1))\vartheta_{01}(X) - (1-\pi(X,0))\,\vartheta_{00}(X)\right]. \end{aligned}$$

Note that the last two terms only depend on X and are mean zero by (8). The terms $\chi_{dz}(Y,X)$ are also mean zero conditional on D and Z. Finally, $Z\frac{\vartheta_{11}(X)-\vartheta_{01}(X)}{p(X)}$ and $(1-Z)\frac{\vartheta_{00}(X)-\vartheta_{10}(X)}{1-p(X)}$ are square-integrable because $p(x)$ is bounded away from zero and one and ϑ_{dz} is bounded away from infinity by Assumption 2.

After some tedious calculations, where we also make use of (8), we also obtain that

$$\frac{\partial \Delta_c^\tau(F_\theta)}{\partial \theta}_{|\theta=\theta_0} = E\left[\psi(Y,D,Z,X)\cdot s(Y,D,Z,X)\right]_{|\theta=\theta_0}$$

which shows that the semiparametric efficiency bound is not affected by knowledge of $p(x)$.

E.3 Proof of Theorem (11):

We consider two regressors sets X_1 and X_2 with $X_1 \subset X_2$. This contains the case where X_1 is the empty set. Suppose that both regressor sets satisfy the Assumption 1 and suppose further that

$$\Pr\left(Z=1|X_1,X_2\right) = \Pr\left(Z=1|X_1\right). \tag{73}$$

Hence, the additional regressors in X_2 that are not included in X_1 do not affect the instrument but may be predictors of the potential outcomes. (If the additional regressors in X_2 would have a causal effect on Z and on the potential outcomes, the regressor set X_1 would generally not be satisfying Assumption 1. In this case the estimator with X_1 only would not be consistent for the QTE.) Let $\mathcal{V}_2$ be the semiparametric variance bound when using regresor set X_2 and $\mathcal{V}_1$ be the semiparametric variance bound when using regresor set X_1, both refering to the same quantile τ of the QTE. We show in the following that

$$\mathcal{V}_1 \geq \mathcal{V}_2,$$

i.e. adding more regressors generally reduces the variance of the QTE.

We start from (71) which gave the semiparametric variance bound as:

$$\begin{aligned}
\mathcal{V}_1 = {} & E\left[\left(\frac{ZD}{p(X_1)}\chi_{11}(Y,X_1)\right)^2\right] + E\left[\left(\frac{(1-Z)D}{1-p(X_1)}\chi_{10}(Y,X_1)\right)^2\right] \\
& + E\left[\left(\frac{Z(1-D)}{p(X_1)}\chi_{01}(Y,X_1)\right)^2\right] + E\left[\left(\frac{(1-Z)(1-D)}{1-p(X_1)}\chi_{00}(Y,X_1)\right)^2\right] \\
& + E\Big[\frac{\pi(X_1,1)\vartheta_{11}^2(X_1)+(1-\pi(X_1,1))\,\vartheta_{01}^2(X_1)}{p(X_1)} + \frac{\pi(X_1,0)\vartheta_{10}^2(X_1)+(1-\pi(X_1,0))\,\vartheta_{00}^2(X_1)}{1-p(X_1)}\Big] \\
& - E\Big[p(X_1)(1-p(X_1))\Big[\frac{\pi(X_1,1)\vartheta_{11}(X_1)+(1-\pi(X_1,1))\,\vartheta_{01}(X_1)}{p(X_1)} \\
& \qquad + \frac{\pi(X_1,0)\vartheta_{10}(X_1)+(1-\pi(X_1,0))\,\vartheta_{00}(X_1)}{1-p(X_1)}\Big]^2\Big]
\end{aligned} \tag{74}$$

where $\chi_{dz}(y,x)$ and $\vartheta_{dz}(x)$ as defined in (68) and (69). The expressions for $\mathcal{V}_2$ are analogous.

As a preliminary calculation consider the first term in (74)

$$
\begin{aligned}
&E\left[\left(\frac{ZD}{p(X_1)}\chi_{11}(Y,X_1)\right)^2\right] = E\left[\frac{ZD}{p(X_1)^2}\left(\frac{\tau - 1(Y \leq Q^{\tau}_{Y^d|c})}{P_c \cdot f_{Y^d|c}(Q^{\tau}_{Y^d|c})} - \vartheta_{11}(X_1)\right)^2\right] \\
&= E\left[\frac{ZD}{p(X_1)^2}\left(\frac{\tau - 1(Y \leq Q^{\tau}_{Y^d|c})}{P_c \cdot f_{Y^d|c}(Q^{\tau}_{Y^d|c})} - \vartheta_{11}(X_2) + \vartheta_{11}(X_2) - \vartheta_{11}(X_1)\right)^2\right] \\
&= E\left[\frac{ZD}{p(X_1)^2}\left(\chi_{11}(Y,X_2) + \vartheta_{11}(X_2) - \vartheta_{11}(X_1)\right)^2\right] \\
&= E\left[\frac{ZD}{p(X_1)^2}\left(\chi_{11}(Y,X_2)\right)^2\right] + E\left[\frac{ZD}{p(X_1)^2}\left(\vartheta_{11}(X_2) - \vartheta_{11}(X_1)\right)^2\right] \\
&+ 2E\left[\frac{ZD}{p(X_1)^2}\chi_{11}(Y,X_2)\left(\vartheta_{11}(X_2) - \vartheta_{11}(X_1)\right)\right]
\end{aligned}
$$

where the last term is zero by using iterated expectations and conditioning on X_2. Using (73) we obtain

$$
= E\left[\frac{ZD}{p(X_2)^2}\left(\chi_{11}(Y,X_2)\right)^2\right] + E\left[\frac{ZD}{p(X_1)^2}\left(\vartheta_{11}(X_2) - \vartheta_{11}(X_1)\right)^2\right]. \tag{75}
$$

The derivations for the second, third and fourth term in (74) are analogous.

We will also use that

$$
E\left[\pi(X_2,1)\cdot\vartheta_{11}(X_2)|X_1\right] = \pi(X_1,1)\cdot\vartheta_{11}(X_1) \tag{76}
$$

or

$$
E\left[\pi(X_2,1)\vartheta_{11}(X_2) - \pi(X_1,1)\vartheta_{11}(X_1)|X_1\right] = 0
$$

because

$$
\begin{aligned}
&= \frac{1}{P_c \cdot f_{Y^d|c}(Q^{\tau}_{Y^d|c})} \cdot E\left[\pi(X_2,1)E\left[\left(\tau - 1(Y \leq Q^{\tau}_{Y^1|c})\right)|X_2, D = Z = 1\right]|X_1\right] \\
&\qquad - \frac{1}{P_c \cdot f_{Y^d|c}(Q^{\tau}_{Y^d|c})} \cdot E\left[\pi(X_1,1)E\left[\left(\tau - 1(Y \leq Q^{\tau}_{Y^1|c})\right)|X_1, D = Z = 1\right]|X_1\right] \\
&= \frac{1}{P_c \cdot f_{Y^d|c}(Q^{\tau}_{Y^d|c})} \cdot E\left[E\left[\frac{ZD}{p(X_1)}\left(\tau - 1(Y \leq Q^{\tau}_{Y^1|c})\right)|X_2\right] - E\left[\frac{ZD}{p(X_1)}\left(\tau - 1(Y \leq Q^{\tau}_{Y^1|c})\right)|X_1\right]|X_1\right] \\
&= \frac{1}{P_c \cdot f_{Y^d|c}(Q^{\tau}_{Y^d|c})} \cdot E\left[\frac{ZD}{p(X_1)}\left(\tau - 1(Y \leq Q^{\tau}_{Y^1|c})\right) - \frac{ZD}{p(X_1)}\left(\tau - 1(Y \leq Q^{\tau}_{Y^1|c})\right)\ |X_1\right] = 0,
\end{aligned}
$$

and analogously for the other terms $\pi(X_2,0)\cdot\vartheta_{10}(X_2)$ and $(1-\pi(X_2,1))\cdot\vartheta_{01}(X_2)$ and $(1-\pi(X_2,0))\cdot\vartheta_{00}(X_2)$.

Using (75) we obtain

$$
\mathcal{V}_1 - \mathcal{V}_2 =
$$

$$
\begin{aligned}
&E\left[\frac{ZD}{p(X_2)^2}\left(\chi_{11}(Y,X_2)\right)^2\right] + E\left[\frac{ZD}{p(X_1)^2}\left(\vartheta_{11}(X_2)-\vartheta_{11}(X_1)\right)^2\right] \\
&+ E\left[\frac{(1-Z)D}{\left(1-p(X_2)\right)^2}\left(\chi_{10}(Y,X_2)\right)^2\right] + E\left[\frac{(1-Z)D}{\left(1-p(X_1)\right)^2}\left(\vartheta_{10}(X_2)-\vartheta_{10}(X_1)\right)^2\right] \\
&+ E\left[\frac{Z(1-D)}{p(X_2)^2}\left(\chi_{01}(Y,X_2)\right)^2\right] + E\left[\frac{Z(1-D)}{p(X_1)^2}\left(\vartheta_{01}(X_2)-\vartheta_{01}(X_1)\right)^2\right] \\
&+ E\left[\frac{(1-Z)(1-D)}{\left(1-p(X_2)\right)^2}\left(\chi_{00}(Y,X_2)\right)^2\right] + E\left[\frac{(1-Z)(1-D)}{\left(1-p(X_1)\right)^2}\left(\vartheta_{00}(X_2)-\vartheta_{00}(X_1)\right)^2\right] \\
&+ E\Big[\frac{\pi(X_1,1)\vartheta_{11}^2(X_1)+(1-\pi(X_1,1))\,\vartheta_{01}^2(X_1)}{p(X_1)} + \frac{\pi(X_1,0)\vartheta_{10}^2(X_1)+(1-\pi(X_1,0))\,\vartheta_{00}^2(X_1)}{1-p(X_1)}\Big] \\
&- E\Big[p(X_1)(1-p(X_1))\Big[\frac{\pi(X_1,1)\vartheta_{11}(X_1)+(1-\pi(X_1,1))\,\vartheta_{01}(X_1)}{p(X_1)} \\
&\qquad + \frac{\pi(X_1,0)\vartheta_{10}(X_1)+(1-\pi(X_1,0))\,\vartheta_{00}(X_1)}{1-p(X_1)}\Big]^2\Big] \\
&- E\left[\left(\frac{ZD}{p(X_2)}\chi_{11}(Y,X_2)\right)^2\right] - E\left[\left(\frac{(1-Z)D}{1-p(X_2)}\chi_{10}(Y,X_2)\right)^2\right] \\
&- E\left[\left(\frac{Z(1-D)}{p(X_2)}\chi_{01}(Y,X_2)\right)^2\right] - E\left[\left(\frac{(1-Z)(1-D)}{1-p(X_2)}\chi_{00}(Y,X_2)\right)^2\right] \\
&- E\Big[\frac{\pi(X_2,1)\vartheta_{11}^2(X_2)+(1-\pi(X_2,1))\,\vartheta_{01}^2(X_2)}{p(X_2)} + \frac{\pi(X_2,0)\vartheta_{10}^2(X_2)+(1-\pi(X_2,0))\,\vartheta_{00}^2(X_2)}{1-p(X_2)}\Big] \\
&+ E\Big[p(X_2)(1-p(X_2))\Big[\frac{\pi(X_2,1)\vartheta_{11}(X_2)+(1-\pi(X_2,1))\,\vartheta_{01}(X_2)}{p(X_2)} \\
&\qquad + \frac{\pi(X_2,0)\vartheta_{10}(X_2)+(1-\pi(X_2,0))\,\vartheta_{00}(X_2)}{1-p(X_2)}\Big]^2\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[\frac{ZD}{p(X_1)^2}\left(\vartheta_{11}(X_2)-\vartheta_{11}(X_1)\right)^2\right] + E\left[\frac{(1-Z)D}{\left(1-p(X_1)\right)^2}\left(\vartheta_{10}(X_2)-\vartheta_{10}(X_1)\right)^2\right] \\
&+ E\left[\frac{Z(1-D)}{p(X_1)^2}\left(\vartheta_{01}(X_2)-\vartheta_{01}(X_1)\right)^2\right] + E\left[\frac{(1-Z)(1-D)}{\left(1-p(X_1)\right)^2}\left(\vartheta_{00}(X_2)-\vartheta_{00}(X_1)\right)^2\right] \\
&+ E\Big[\frac{\pi(X_1,1)\vartheta_{11}^2(X_1)+(1-\pi(X_1,1))\,\vartheta_{01}^2(X_1)}{p(X_1)} + \frac{\pi(X_1,0)\vartheta_{10}^2(X_1)+(1-\pi(X_1,0))\,\vartheta_{00}^2(X_1)}{1-p(X_1)}\Big] \\
&- E\Big[p(X_1)(1-p(X_1))\Big[\frac{\pi(X_1,1)\vartheta_{11}(X_1)+(1-\pi(X_1,1))\,\vartheta_{01}(X_1)}{p(X_1)} \\
&\qquad + \frac{\pi(X_1,0)\vartheta_{10}(X_1)+(1-\pi(X_1,0))\,\vartheta_{00}(X_1)}{1-p(X_1)}\Big]^2\Big] \\
&- E\Big[\frac{\pi(X_2,1)\vartheta_{11}^2(X_2)+(1-\pi(X_2,1))\,\vartheta_{01}^2(X_2)}{p(X_2)} + \frac{\pi(X_2,0)\vartheta_{10}^2(X_2)+(1-\pi(X_2,0))\,\vartheta_{00}^2(X_2)}{1-p(X_2)}\Big] \\
&+ E\Big[p(X_2)(1-p(X_2))\Big[\frac{\pi(X_2,1)\vartheta_{11}(X_2)+(1-\pi(X_2,1))\,\vartheta_{01}(X_2)}{p(X_2)} \\
&\qquad + \frac{\pi(X_2,0)\vartheta_{10}(X_2)+(1-\pi(X_2,0))\,\vartheta_{00}(X_2)}{1-p(X_2)}\Big]^2\Big]
\end{aligned}
$$

noting that $E[\frac{ZD}{p(X_1)^2}\left(\vartheta_{11}(X_2)-\vartheta_{11}(X_1)\right)^2+\frac{\pi(X_1,1)\vartheta_{11}^2(X_1)}{p(X_1)}-\frac{\pi(X_2,1)\vartheta_{11}^2(X_2)}{p(X_2)}|X_1]=0$ by (76) and analogously for the other terms we obtain

$$
\begin{aligned}
&= E\Big[p(X_2)(1-p(X_2))\Big[\frac{\pi(X_2,1)\vartheta_{11}(X_2)+(1-\pi(X_2,1))\,\vartheta_{01}(X_2)}{p(X_2)} \\
&\qquad + \frac{\pi(X_2,0)\vartheta_{10}(X_2)+(1-\pi(X_2,0))\,\vartheta_{00}(X_2)}{1-p(X_2)}\Big]^2\Big] \\
&- E\Big[p(X_1)(1-p(X_1))\Big[\frac{\pi(X_1,1)\vartheta_{11}(X_1)+(1-\pi(X_1,1))\,\vartheta_{01}(X_1)}{p(X_1)} \\
&\qquad + \frac{\pi(X_1,0)\vartheta_{10}(X_1)+(1-\pi(X_1,0))\,\vartheta_{00}(X_1)}{1-p(X_1)}\Big]^2\Big]
\end{aligned}
$$

again making use of (76) it follows

$$
\begin{aligned}
= E\Big[p(X_1)(1-p(X_1))\cdot Var\Big[&\frac{\pi(X_2,1)\vartheta_{11}(X_2)+(1-\pi(X_2,1))\,\vartheta_{01}(X_2)}{p(X_2)} \\
&+ \frac{\pi(X_2,0)\vartheta_{10}(X_2)+(1-\pi(X_2,0))\,\vartheta_{00}(X_2)}{1-p(X_2)}\;|X_1\Big]\Big]
\end{aligned}
$$

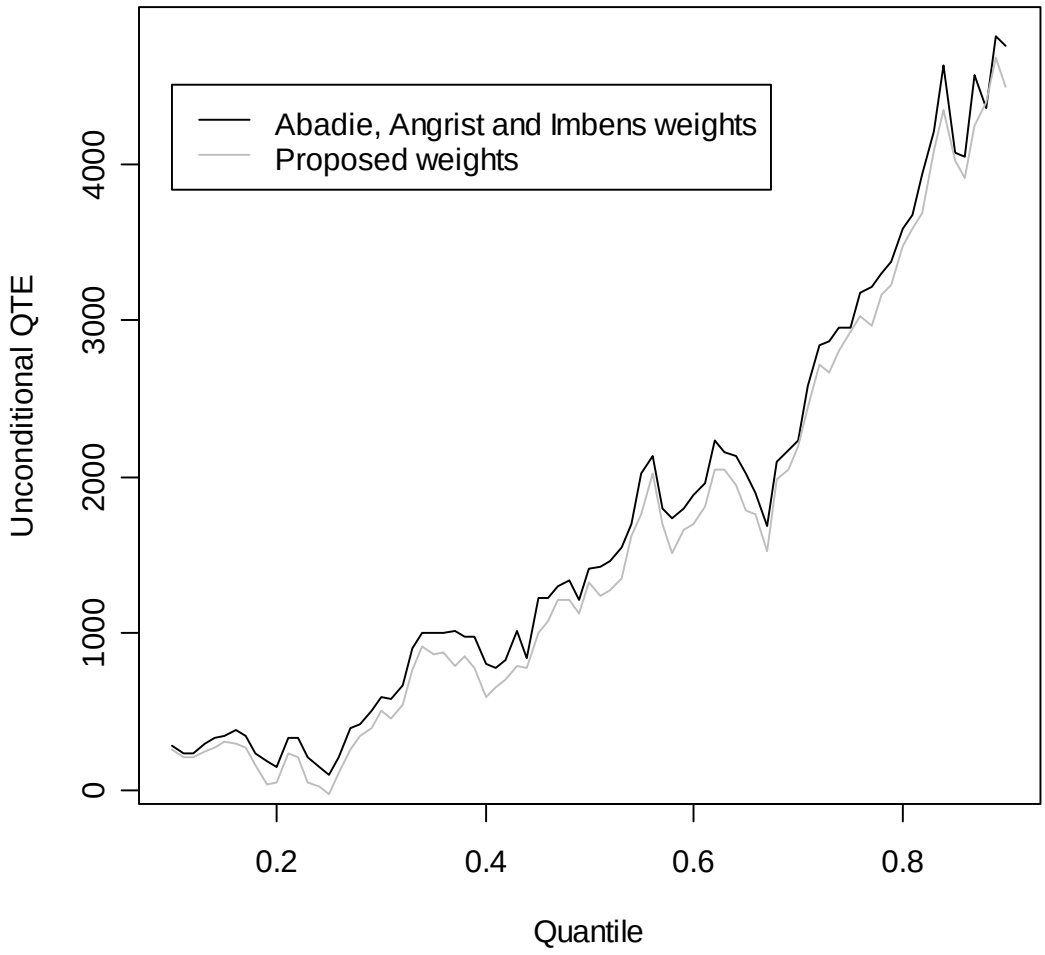
which is always non-negative.

Table 1: Probit regression of the instrument Z on X

	Coef.	Std.err	z-value
Experience	-0.004	0.026	-0.14
Experience squared	-1e-4	0.001	-0.10
Black	0.180*	0.075	2.40
SMSA 1976	0.264*	0.072	3.65
Dummy for lived in south in 1976	-0.081	0.098	-0.83
SMSA 1966	1.005*	0.070	14.30
Regional dummies for 1966			
Region 2	0.027	0.162	0.17
Region 3	-0.483*	0.153	-3.15
Region 4	-0.386*	0.169	-2.28
Region 5	-0.607*	0.174	-3.49
Region 6	-0.991*	0.184	-5.38
Region 7	-0.864*	0.184	-4.68
Region 8	-0.690*	0.204	-3.39
Region 9	-0.267	0.171	-1.56
Father's education	0.021	0.017	1.23
Mother's education	-0.017	0.016	-1.06
Father's education missing	-0.245	0.204	-1.20
Mother's education missing	-0.027	0.135	-0.20
Interactions of parental education:			
Mom and dad both > 12 yrs ed	0.002	0.305	0.01
Mom & dad >=12 and not both exactly 12	-0.307	0.276	-1.11
Mom=dad=12	-0.194	0.253	-0.77
Mom >=12 and dad missing	-0.177	0.175	1.01
Father >=12 and mom not in f1-f4	-0.208	0.250	-0.83
Mom>=12 and dad nonmissing	-0.221	0.239	-0.93
Mom and dad both >=9	-0.199	0.256	-0.78
Mom and dad both nonmissing	-0.109	0.220	-0.50
Living with mother and father at age 14	-0.084	0.102	-0.83
Living only with mother at age 14	0.034	0.139	0.25
Constant	0.460	0.345	1.33

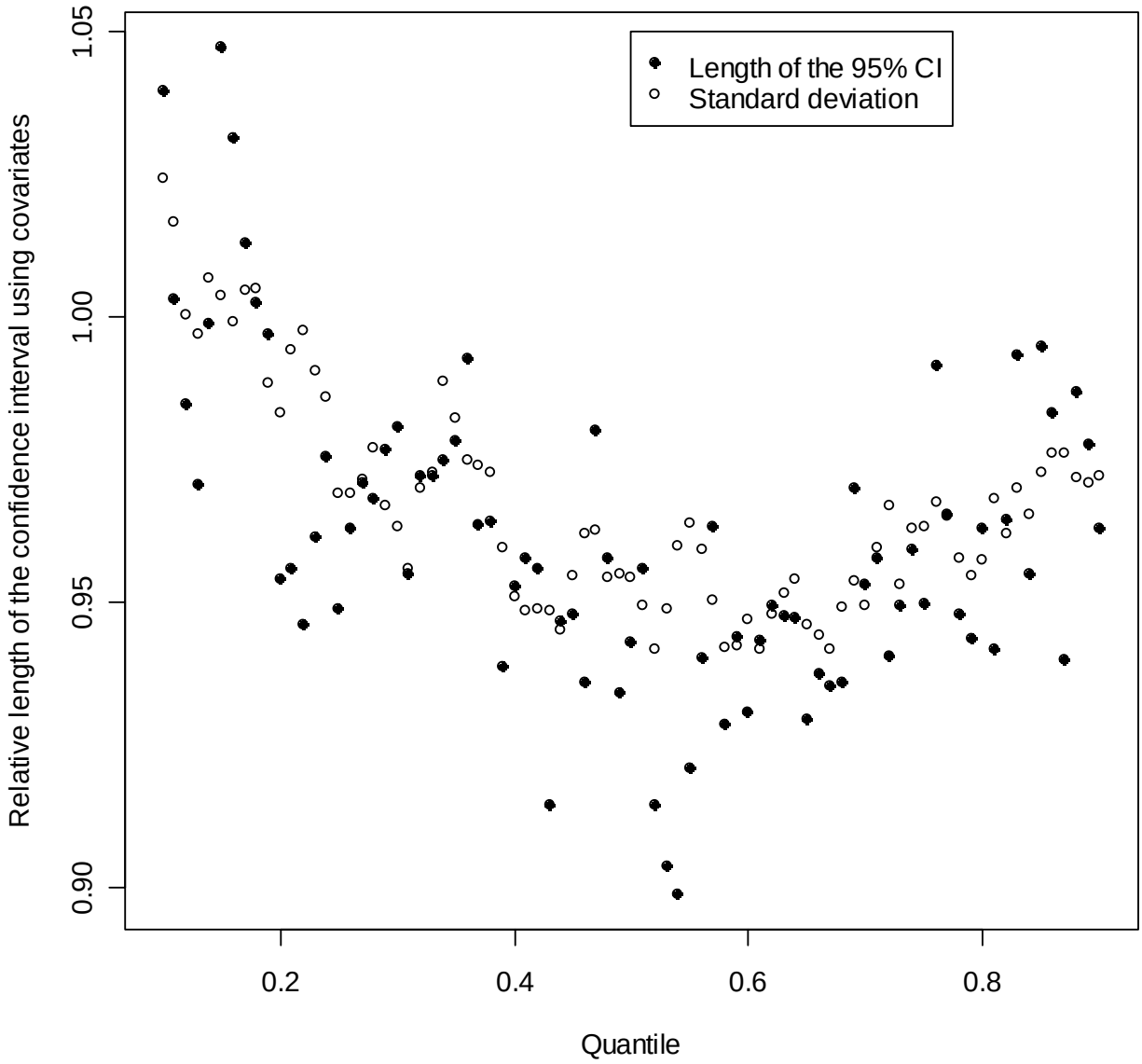
Note: Same specification as in Card (1995). Binary dependent variable: proximity of an accredited 4-year college in 1966. A * indicates that the coefficient is different from 0 at the 5% significance level. Potential experience in 1976 is constructed as age minus years of education minus 6.

Figure 1: Unconditional QTE of the JTPA training program



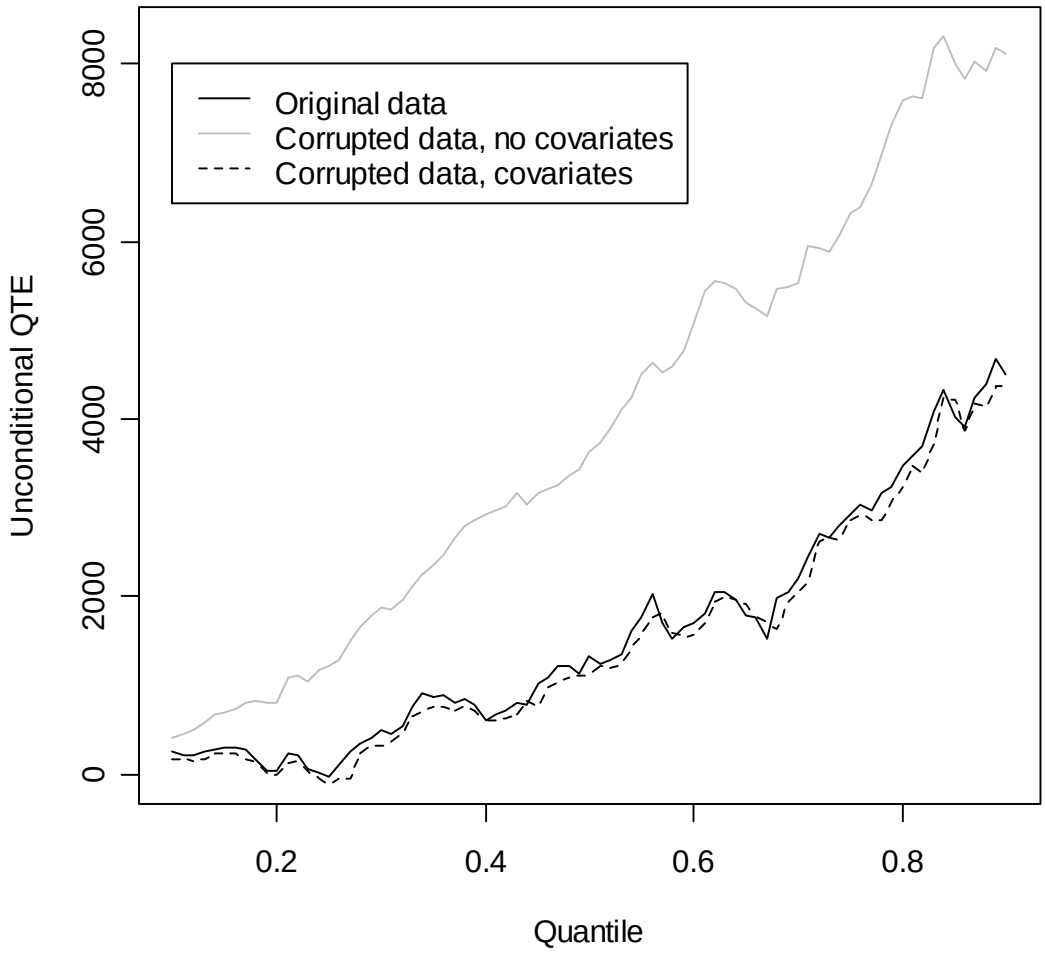
Note: Same data (only men) and same specification as in Abadie, Angrist and Imbens (2002).

Figure 2: Precision gain obtained by incorporating covariates



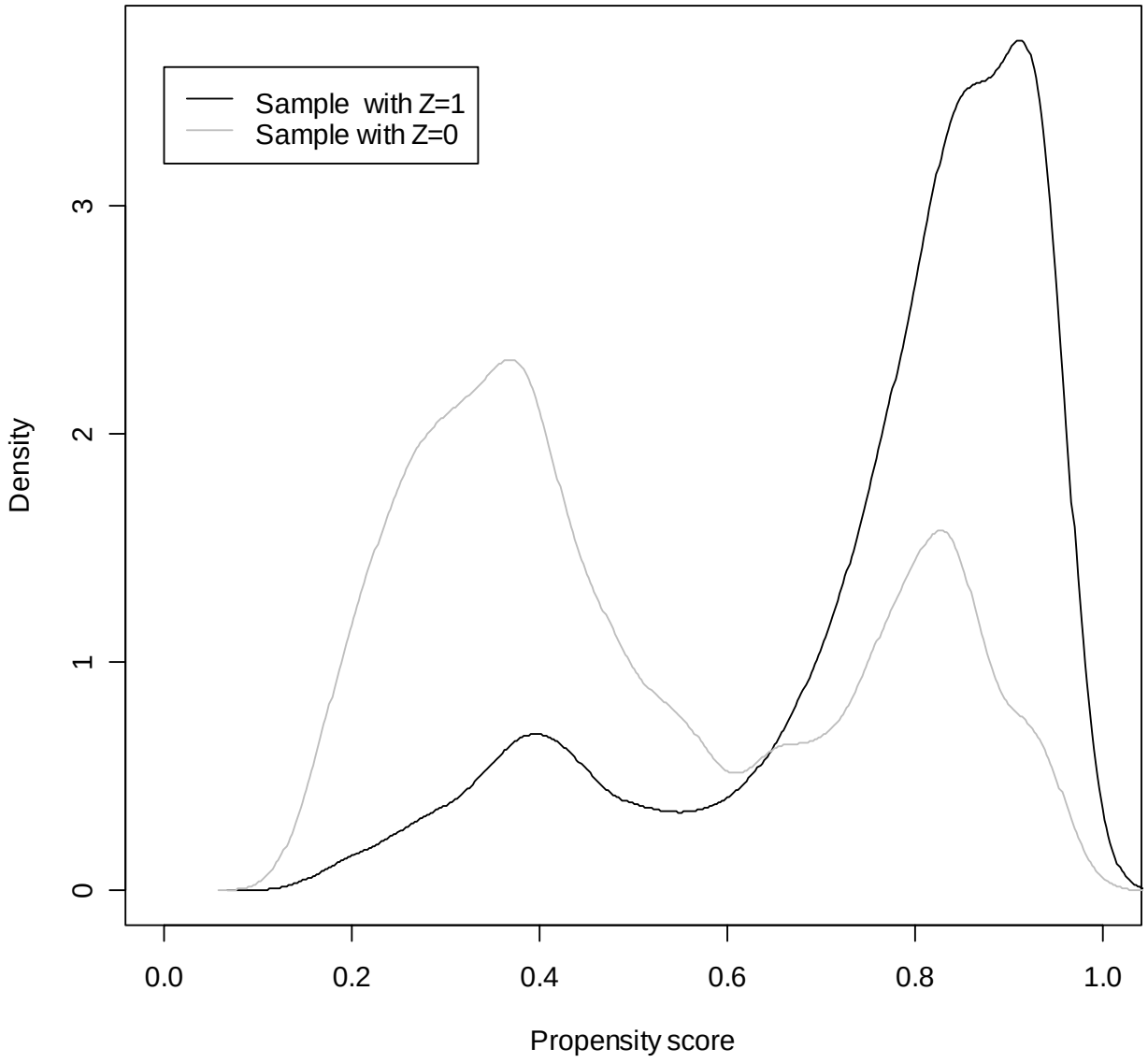
Note: The results plotted in Figure 1 have been bootstrapped 1000 times. We have calculated 95% percentile confidence intervals both with and without covariates. The bold points indicate the length of the confidence intervals incorporating covariates relatively to the length of the confidence interval without covariates. The empty circles indicate the same relative lengths of the standard deviation. Values below 1 indicate a reduction in standard deviation or a shortening of the confidence interval.

Figure 3: Bias arising from the manipulation of the instrument



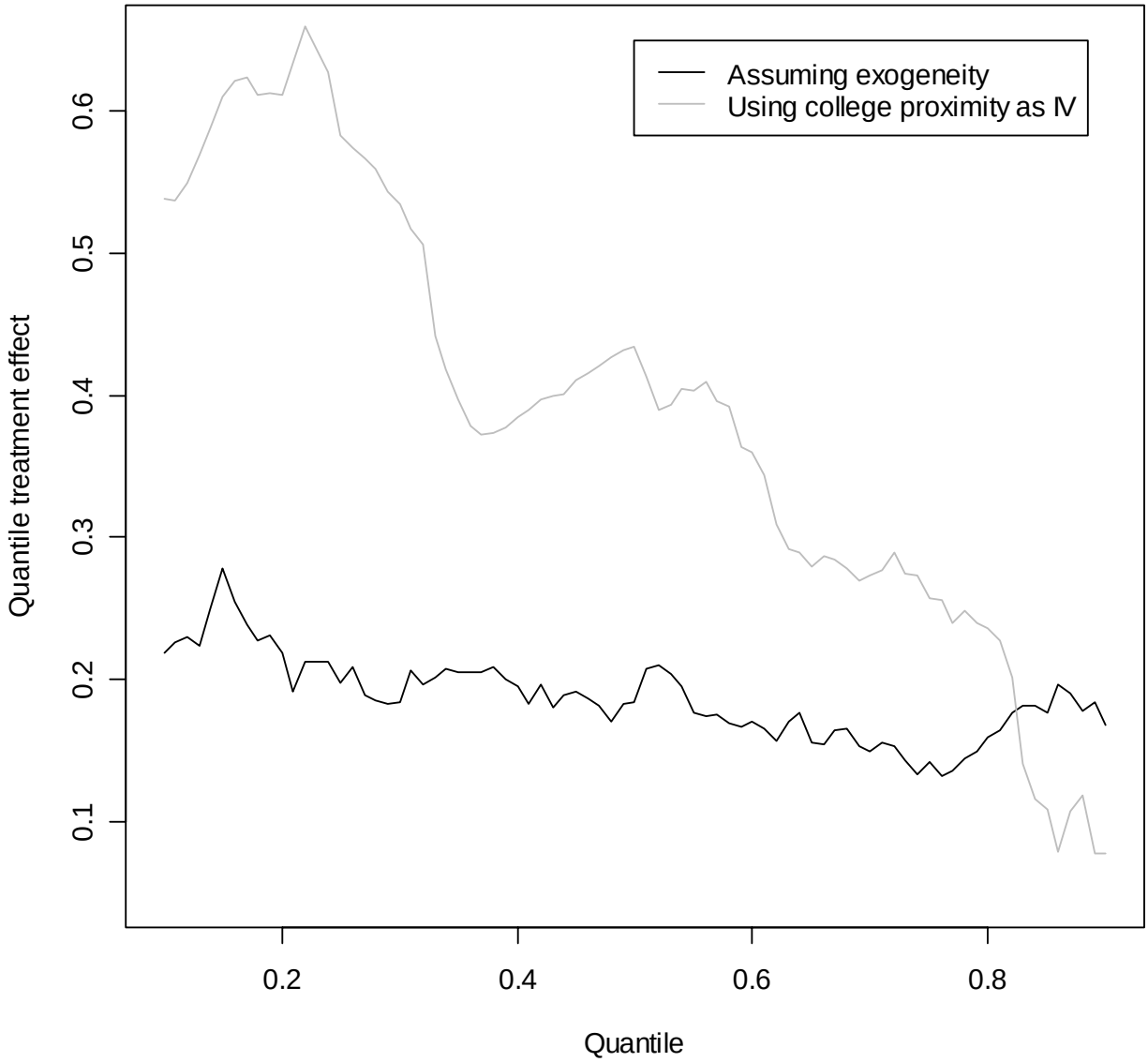
Note: The corrupted instrument $\tilde{Z}$ is set to 1 with probability 0.5 if Z=0 and the individual is married. In all other cases, $\tilde{Z} = Z$. We eliminate the random component of this data manipulation by averaging the results over 1000 corrupted samples.

Figure 4: Distribution of P(Z=1|X) in the Z=0 and Z=1 subpopulation



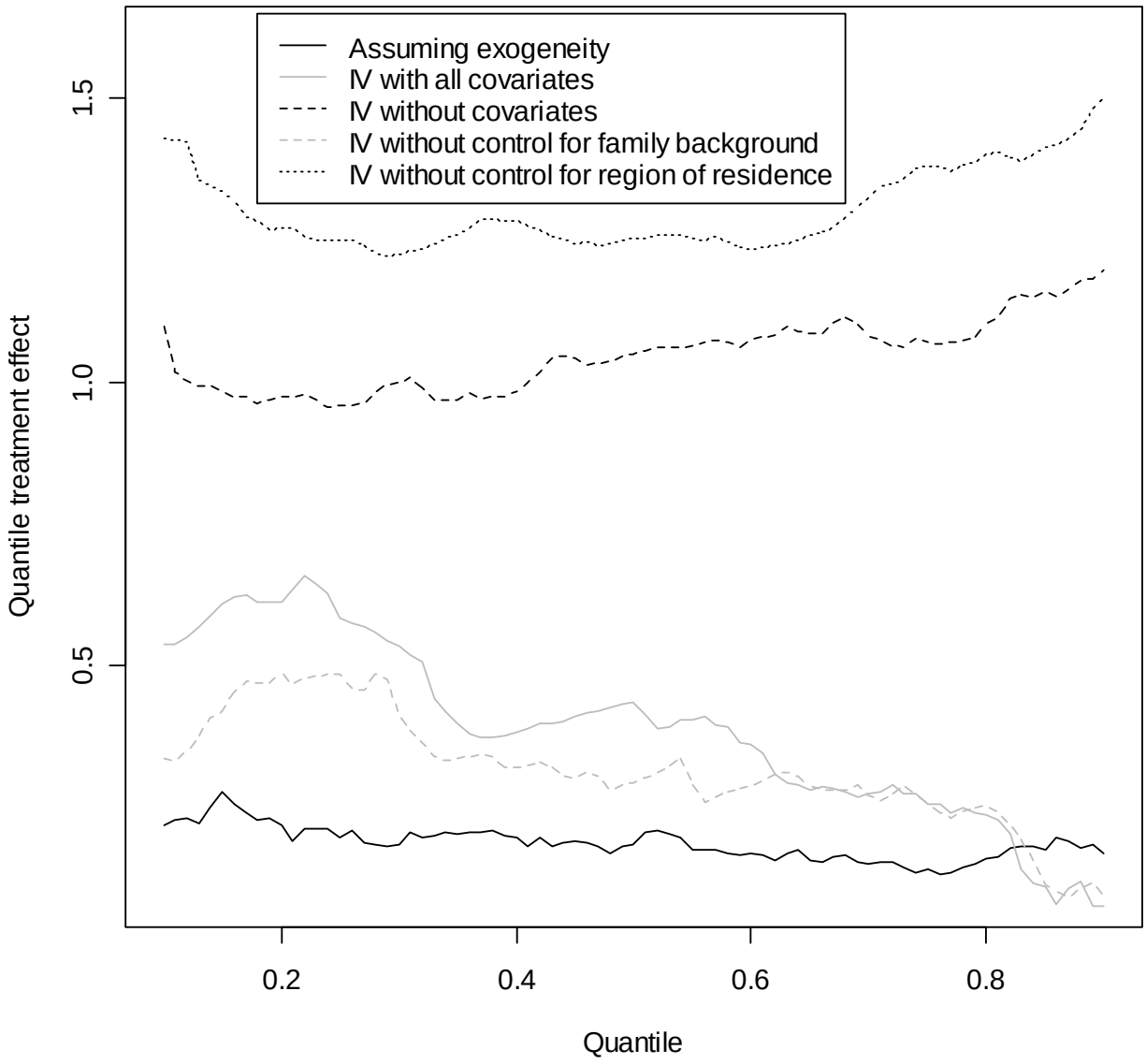
Note: Kernel density estimates of the propensity score $p(x)$ estimated by a parametric probit for the two sub-samples defined by the value of *Z*.

Figure 5: Nonparametric estimators of the QTE of college attendance



Note: Unconditional QTE of having a college degree on the log hourly wage. The weighting estimator defined in equation (14) with $p(x)$ estimated by local logit regression has been used. When we assume exogeneity, college attendance is used as its own IV. The bandwidths have been chosen by cross-validation. The covariates included are the same as those used by Card (1995) and in our Table 1 (naturally without experience squared and the interaction terms because our model is non-parametric).

Figure 6: QTE of college on earnings: the role of the covariates



Note: The results assuming exogeneity and using all covariates are the same as those in Figure 5. The other estimators differ only in the covariates used to estimate the weights. Only a constant has been used to calculate the "IV without covariates" results. All covariates except parental education and the family structure at age 14 have been used to calculate the "IV without control for family background" results. All covariates except the region of residence in 1966 and 1976 have been used to calculate the "IV without control for region of residence" results.