

Relaxing monotonicity in the identification of local average treatment effects

Martin Huber and Giovanni Mellace

November 3, 2013

University of St. Gallen, Dept. of Economics

Abstract: In heterogeneous treatment effect models with endogeneity, the identification of the local average treatment effect (LATE) typically relies on an instrument that satisfies two conditions: (i) joint independence of the potential post-instrument variables and the instrument and (ii) monotonicity of the treatment in the instrument. We show that identification is still feasible when replacing monotonicity by a strictly weaker local monotonicity condition (given marginal potential outcomes). We demonstrate that the latter allows identifying the LATEs on the (i) compliers (whose treatment reacts to the instrument in the intended way), (ii) defiers (who react counter-intuitively), and (iii) both populations jointly. Furthermore, (i) and (iii) coincides with standard LATE if monotonicity holds. We also present an application to the quarter of birth instrument.

Keywords: instrumental variable, treatment effects, LATE, local monotonicity.

JEL classification: C14, C21, C26.

We have benefited from comments by Alberto Abadie, Joshua Angrist, Matias Cattaneo, Mario Fiorini, Markus Frölich, Stefan Hoderlein, Guido Imbens, Toru Kitagawa, Frank Kleibergen, Tobias Klein, Michael Lechner, Arthur Lewbel, Enno Mammen, Blaise Melly, Katrien Stevens, and seminar/conference participants at the University of St. Gallen (Oct 2011), Brown University (Nov 2011), the Austrian-Swiss labor meeting in Stuben (Jan 2012), Harvard (Feb 2012), MIT (Mar 2012), the Spring Meeting of Young Economists in Mannheim (Apr 2012), Boston College (Apr 2012), the North American Summer Meeting of the Econometric Society in Evanston (June-July 2012), the European Meeting of the Econometric Society in Malaga (Aug 2012), the European Association of Labour Economists meeting in Bonn (Sept 2012), the Italian Congress of Econometrics and Empirical Economics in Genoa (Jan 2013), and Rotterdam (Jun 2013). Martin Huber gratefully acknowledges financial support from the Swiss National Science Foundation grant PBSGP1_138770. Addresses for correspondence: Martin Huber (martin.huber@unisg.ch) and Giovanni Mellace (giovanni.mellace@unisg.ch), SEW, University of St. Gallen, Varnbühlstrasse 14, 9000 St. Gallen, Switzerland.

1 Introduction

In many economic evaluation problems, causal inference is complicated by endogeneity, implying that the explanatory or treatment variable of interest is correlated with unobserved factors that also affect the outcome. E.g., when estimating the returns to education, the schooling choice is plausibly influenced by unobserved ability (see for instance Card, 1999) which itself most likely has an impact on the earnings outcome. Due to the endogenous treatment selection (also known as selection on unobservables) the earnings effect of education is confounded with the unobserved terms. In the presence of endogeneity, identification often relies on an instrumental variable (IV) that generates exogenous variation in the treatment.

In heterogeneous treatment effect models with a binary treatment, an instrument is conventionally required to satisfy two assumptions. Firstly, it must be independent of the joint distribution of potential treatment states and potential outcomes, which excludes direct effects on the latter and implies that the instrument is (as good as) randomly assigned. Secondly, the treatment state has to vary with the instrument in a weakly monotonic manner. E.g., assignment should weakly increase actual program participation of all individuals in the population, i.e., *globally*. Under these assumptions, Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) show that the local average treatment effect (LATE) within the subpopulation of compliers, whose treatment states react to the instrument in the intended way, is identified. This is feasible because monotonicity rules out the existence of defiers, who react counter-intuitively to the instrument, e.g., by participating in a program if not being assigned to it and not participating under assignment. The non-existence of defiers implies the identification of the potential outcome distributions (including the means) of the compliers under treatment and non-treatment, see Imbens and Rubin (1997). The difference in mean potential outcomes is equivalent to the well known Wald formula that represents the LATE as ratio of the intention to treat effect and the share of compliers.

The contribution of this paper is to show that LATEs are identified (and under particular

assumptions $\sqrt{n}$ -consistently estimated) when relying on a condition that is strictly weaker than *global* monotonicity, while maintaining joint independence. We will refer to this condition as “*local* monotonicity” (LM). Crudely speaking and in contrast to (global) monotonicity, LM allows for the existence of both compliers and defiers, but requires that they do not occur at the same time at any support point of the outcome conditional on a particular treatment state. I.e., monotonicity is assumed to hold locally in subregions of the marginal potential outcome distributions (and may switch the sign across subregions), rather than over the entire support.

More specifically (and assuming a binary instrument), LM excludes the possibility that a subject is a defier if the joint density of her observed outcome and being treated conditional on receiving the instrument is larger than the respective joint density conditional on not receiving the instrument. In fact, under the independence of the instrument, this order of the joint densities is a sufficient condition for the existence of compliers (as outlined in Section 2). By ruling out defiers in such regions by LM, the potential outcomes of the compliers are locally identified. Conversely, in support regions in which the joint density of the outcome and the treatment when not receiving the instrument dominates the joint density when receiving the instrument, defiers necessarily exist by independence and LM rules out compliers to identify the potential outcomes of the defiers. Equivalent results hold for the joint densities under non-treatment. Therefore, we demonstrate that LM is sufficient for the identification of the marginal potential outcome distributions of the compliers and the defiers in either treatment state. Furthermore, it immediately follows that (global) monotonicity is a special case of LM, because the former requires that the joint densities are nested, see Kitagawa (2009).

As defiers are no longer assumed away, one improvement of generalizing monotonicity to LM is that we do not only identify the (i) LATE on the compliers, but also the LATEs (ii) on the defiers and (iii) on the joint population of compliers and defiers. Furthermore, the existence and proportion of defiers (and any other subpopulation) can be verified in the data to judge the relevance of (ii) and (iii). It will also be shown that (i) and (iii) coincide with the standard LATE with monotonicity if defiers do not exist. However, if the defiers’ proportion is larger than zero,

(i), (ii), and (iii) generally differ and standard LATE is inconsistent due to incorrectly invoking monotonicity. Finally, our discussion also reveals that our set of assumptions can be partially tested in the data. In fact, a necessary (but not sufficient) condition is the satisfaction of a particular scale constraint which has also been considered by Kitagawa (2009) and is based on the intuition that the proportion of any subpopulation must be equal across treatment states. If the latter is violated, point identification of LATEs is lost, while partial identification in the spirit of Manski (1990) might still be a worthwhile alternative, see for instance Huber and Mellace (2010).

Apart from the present work, several other studies have considered deviations from monotonicity and their implications for LATE identification. Small and Tan (2007) weaken (individual-level) monotonicity to stochastic monotonicity, requiring that conditional on unobservables which satisfy particular assumptions (e.g., conditional independence of the instrument), the share of compliers weakly dominates the share of defiers. Small and Tan (2007) show that the Wald estimator, albeit biased, retains some desirable properties (such as giving the correct sign of the effect) in the limit, but do not propose any method to point identify the LATE. Klein (2010) develops methods to assess the sensitivity of the LATE to random departures from monotonicity and to approximate the bias under particular conditions. In contrast, our framework allows for LATE identification under non-random violations, given that LM is satisfied. de Chaisemartin and D'Haultfoeuille (2012) represent monotonicity by a latent index model, see Vytlacil (2002), in which they relax the conventional rank invariance in the unobserved terms to rank similarity, see Chernozhukov and Hansen (2005). The unobserved terms affecting the treatment (e.g. taste) may therefore be a function of the instrument (which allows for the existence of defiers), but must have the same distribution with and without instrument conditional on the potential outcomes. Then, the probability limit of the Wald estimator identifies a causal effect on a specific mixture of subpopulations. In contrast, our assumptions allow identifying the LATE on subpopulations that are well defined in terms of their treatment response to the instrument.

Finally, de Chaisemartin (2012) relaxes monotonicity to two forms of stochastic monotonicity: The share of compliers weakly dominates the share of defiers (i) conditional on the (individual-level) treatment effect or (ii) conditional on the joint potential outcomes (which is closely related to Small and Tan (2007)). He shows that under (i), the probability limit of the Wald estimator point identifies the LATE within the remaining compliers after having discarded the share of compliers with the same size and average treatment effect as the defiers. The stronger (ii) also identifies local quantile treatment effects. As discussed in de Chaisemartin (2012), his approach can also be used to relax our LM assumption to allow for both compliers and defiers at particular outcome values. His weaker Assumption 2.9 identifies the LATEs within subgroups of compliers and defiers, while our stronger LM identifies the LATEs among all compliers and defiers.

As a practical illustration of our methods, we present an application to the 1980 U.S. census data on males born in 1940-49 considered in Angrist and Krueger (1991) to estimate the returns to education by using the birth quarter as instrument for education. Arguably, among students entering school in the same year, those who are born in an earlier quarter can drop out after less years of completed education at the age when compulsory schooling ends than those born later (in particular after the end of the academic year). This suggests that education is monotonically increasing in the quarter of birth. However, the postponement of school entry due to redshirting or unobserved school policies as discussed in Aliprantis (2012), Barua and Lang (2009), and Klein (2010) may reverse the relation of education and quarter of birth for some individuals and thus, violate monotonicity. We therefore relax global monotonicity and find statistically significant shares of both compliers and defiers.

The remainder of this paper is organized as follows. Section 2 discusses the assumptions and the identification of the LATEs on the compliers, defiers, and both populations jointly as well as the differences/connections to the standard LATE framework with monotonicity. Section 3 considers identification in the presence of bounded non-binary instruments. Section 4 proposes $\sqrt{n}$ -consistent and asymptotically normal estimators of the LATEs. Section 5 provides a brief simulation study. An empirical application to data from Angrist and Krueger (1991) is presented

in Section 6. Section 7 concludes.

2 Assumptions and identification

Suppose that we are interested in the average effect of a binary treatment $D \in \{1, 0\}$ (e.g., participation in a training program) on an outcome Y (e.g., labor market success such as earnings) evaluated at some point in time after the treatment. Under endogeneity, the effect of D is confounded with unobserved factors that affect both the treatment and the outcome. Identification of treatment effects generally requires an instrument, denoted by Z , that is correlated with the treatment but does not have a direct effect on the outcome (i.e., any impact other than through the treatment). In this section, we will consider the case of a binary instrument ($Z \in \{0, 1\}$) such as randomized treatment assignment, whereas Section 3 discusses the case of bounded non-binary instruments. Denote by $D(z)$ the potential treatment state for instrument $Z = z$, and by $Y(d)$ the potential outcome for treatment $D = d$ (see for instance Rubin, 1974, for a discussion of the potential outcome notation). For each subject, only one potential outcome is observed, because $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$.

Table 1: Types

Type T	D(1)	D(0)	Notion
a	1	1	Always takers
c	1	0	Compliers
d	0	1	Defiers
n	0	0	Never takers

As discussed in Angrist, Imbens, and Rubin (1996) and summarized in Table 1, the population can be categorized into four types (denoted by $T \in \{a, c, d, n\}$), depending on how the treatment state changes with the instrument. The compliers react on the instrument in the intended way by taking the treatment when $Z = 1$ and abstaining from it when $Z = 0$. For the remaining three types, $D(z) \neq z$ for either $Z = 1$, or $Z = 0$, or both: The always takers are always treated

irrespective of the instrument state, the never takers are never treated, and the defiers only take the treatment when $Z = 0$. It is obvious that we cannot directly observe the type any observation belongs to as either $D(1)$ or $D(0)$ remains unknown due to the fact that the actual treatment is $D = Z \cdot D(1) + (1 - Z) \cdot D(0)$. This implies that any observation i with a particular combination of the treatment and the instrument may belong to one of two types, see Table 2. Assuming an i.i.d. framework, we will show that the potential outcome distributions of the compliers and the defiers may nevertheless be identified under conditions that are weaker than the standard LATE assumptions of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996).

Table 2: Observed subgroups and types

Observed values of Z and D	Potential types T
$\{i : Z_i = 1, D_i = 1\}$	observation i belongs either to a or to c
$\{i : Z_i = 1, D_i = 0\}$	observation i belongs either to d or to n
$\{i : Z_i = 0, D_i = 1\}$	observation i belongs either to a or to d
$\{i : Z_i = 0, D_i = 0\}$	observation i belongs either to c or to n

To characterize the identification problem, we introduce further notation that heavily borrows from Kitagawa (2009) who, in contrast to this paper, considers partial identification of the average treatment effect (ATE) on the entire population. In a first step, we define shorthand expressions for the observed joint densities of the outcome and the treatment conditional on the instrument:

$$\begin{aligned}
p_1(y) &= f(y, D = 1 | Z = 1), & p_0(y) &= f(y, D = 0 | Z = 1), \\
q_1(y) &= f(y, D = 1 | Z = 0), & q_0(y) &= f(y, D = 0 | Z = 0).
\end{aligned} \tag{1}$$

I.e., $p_d(y)$ ($q_d(y)$) represents the joint density of $Y = y$ and $D = d$ given $Z = 1$ ($Z = 0$). Furthermore, denote by $\mathcal{Y}$ the support of Y and let $f(y(d))$ and $f(y(d), T = t)$ denote the density of the potential outcome and the joint density of the potential outcome and the type, respectively,

with $d \in \{0, 1\}$, $t \in \{a, c, d, n\}$, and $y \in \mathcal{Y}$. By Table 2, we have that for all $y \in \mathcal{Y}$,

$$p_1(y) = f(y(1), T = c|Z = 1) + f(y(1), T = a|Z = 1), \quad (2)$$

$$q_1(y) = f(y(1), T = d|Z = 0) + f(y(1), T = a|Z = 0), \quad (3)$$

$$p_0(y) = f(y(0), T = d|Z = 1) + f(y(0), T = n|Z = 1), \quad (4)$$

$$q_0(y) = f(y(0), T = c|Z = 0) + f(y(0), T = n|Z = 0), \quad (5)$$

$$f(y(1)|Z = 1) - p_1(y) = f(y(1), T = d|Z = 1) + f(y(1), T = n|Z = 1), \quad (6)$$

$$f(y(1)|Z = 0) - q_1(y) = f(y(1), T = c|Z = 0) + f(y(1), T = n|Z = 0), \quad (7)$$

$$f(y(0)|Z = 1) - p_0(y) = f(y(0), T = c|Z = 1) + f(y(0), T = a|Z = 1), \quad (8)$$

$$f(y(0)|Z = 0) - q_0(y) = f(y(0), T = d|Z = 0) + f(y(0), T = a|Z = 0). \quad (9)$$

Equations (2) to (5) make immediate use of the fact that any joint density $p_d(y)$, $q_d(y)$ observed in the data is constituted by the potential outcomes (given Z) of two different types. Equations (6) to (9) come from the law of total probability, implying that $f(y(d)|Z = z) = \sum_{t \in \{a, c, d, n\}} f(y(d), T = t|Z = z)$.

We now impose the first identifying assumption which invokes independence between Z and the joint distribution of the potential treatment states and outcomes, see Imbens and Angrist (1994):

Assumption 1:

$Z \perp (D(1), D(0), Y(1), Y(0))$ (joint independence),

where “ $\perp$ ” denotes independence. Assumption 1 is standard in the literature on the LATE and implies the randomization of the instrument (such that it is unrelated with factors affecting the treatment and/or outcome) and the exclusion of direct effects on the outcome. It follows that not only the potential outcomes, but also the types, which are defined by the potential treatment

states, are independent of the instrument. Therefore, equations (2) to (9) simplify to

$$p_1(y) = f(y(1), T = c) + f(y(1), T = a), \quad (10)$$

$$q_1(y) = f(y(1), T = d) + f(y(1), T = a), \quad (11)$$

$$p_0(y) = f(y(0), T = d) + f(y(0), T = n), \quad (12)$$

$$q_0(y) = f(y(0), T = c) + f(y(0), T = n), \quad (13)$$

$$f(y(1)) - p_1(y) = f(y(1), T = d) + f(y(1), T = n), \quad (14)$$

$$f(y(1)) - q_1(y) = f(y(1), T = c) + f(y(1), T = n), \quad (15)$$

$$f(y(0)) - p_0(y) = f(y(0), T = c) + f(y(0), T = a), \quad (16)$$

$$f(y(0)) - q_0(y) = f(y(0), T = d) + f(y(0), T = a), \quad (17)$$

see Kitagawa (2009) for a more detailed discussion.

To understand the implications of (10) to (17) for the identification of $f(y(d), T = t)$, consider, for instance, the always takers. (10) and (11) imply that $f(y(1), T = a)$ cannot be larger than $\min(p_1(y), q_1(y))$, under Assumption 1. Secondly, by (16) and (17), the upper bound of $f(y(0), T = a)$ is $f(y(0)) - \max(p_0(y), q_0(y))$ (which is, however, not observed because $f(y(0))$ is unknown). Similar results can be derived for all other types. E.g., $f(y(0), T = n)$ is bounded from above by $\min(p_0(y), q_0(y))$ and $f(y(1), T = n)$ by $f(y(1)) - \max(p_1(y), q_1(y))$. Furthermore, Assumption 1 also provides information on the local existence and relative importance of compliers and defiers. I.e., compliers necessarily exist locally if $p_1(y) > q_1(y)$ or $q_0(y) > p_0(y)$, respectively, because by (10) to (13) this means that compliers dominate defiers (whose proportion is at least zero): $f(y(1), T = c) > f(y(1), T = d) \geq 0$ or $f(y(0), T = c) > f(y(0), T = d) \geq 0$, respectively. In this case, $p_1(y) - q_1(y)$ or $q_0(y) - p_0(y)$, respectively, provide a lower bound on the compliers. Equivalently, $p_1(y) < q_1(y)$ or $q_0(y) < p_0(y)$ point to the local existence of defiers and their dominance over compliers, such that $q_1(y) - p_1(y)$ or $q_0(y) - p_0(y)$, respectively, bound their density of the defiers from below.

By (10) to (22), also any type proportion $\Pr(T = t)$ can be bounded. To this end, we define

the following integrals also used in Kitagawa (2009) to keep the notation tractable:

$$\delta_1 = \int_{\mathcal{Y}} \max(p_1(y), q_1(y)) dy, \quad (18)$$

$$\delta_0 = \int_{\mathcal{Y}} \max(p_0(y), q_0(y)) dy, \quad (19)$$

$$\lambda_1 = \int_{\mathcal{Y}} \min(p_1(y), q_1(y)) dy, \quad (20)$$

$$\lambda_0 = \int_{\mathcal{Y}} \min(p_0(y), q_0(y)) dy. \quad (21)$$

I.e., δ_d is the integrated density envelope of $(p_d(y), q_d(y))$, while λ_d is the inner integrated density envelope of $(p_d(y), q_d(y))$. Furthermore, note that

$$\int_{\mathcal{Y}} f(y(1), T = t) dy = \int_{\mathcal{Y}} f(y(0), T = t) dy = \Pr(T = t) \quad \forall t = \{a, c, d, n\}, \quad (22)$$

because very intuitively, the proportion of any type ($\Pr(T = t)$) is necessarily equal across the potential outcome distributions under treatment and non-treatment. This is what Kitagawa (2009) refers to as scale constraint.

Again, consider the always takers to investigate the identifying power of Assumption 1 and the scale constraint. As $f(y(1), T = a)$ and $f(y(0), T = a)$ are bounded by $\min(p_1(y), q_1(y))$ and $f(y(0)) - \max(p_0(y), q_0(y))$, respectively, it follows from (19), (20), and the scale constraint (22) that the proportion of always takers, $\Pr(T = a)$, is bounded from above by the minimum of λ_1 and $1 - \delta_0$ (with $\int_{\mathcal{Y}} f(y(1)) dy = 1$). Concerning the latter, note that (16) and (17) imply that $\max(p_0(y), q_0(y)) \leq f(y(0))$. Therefore, by integrating this expression we get $\delta_0 \leq 1$, otherwise Assumption 1 would be violated. Analogously, $\Pr(T = n)$ is bounded from above by the minimum of λ_0 and $1 - \delta_1$ (where $\delta_1 \leq 1$). Likewise, $\int_{\mathcal{Y}} p_1(y) dy - \lambda_1 = \Pr(D = 1 | Z = 1) - \lambda_1$ is a lower bound on the proportion of compliers under treatment, because the lower bound on $f(y(1), T = c)$ is $p_1(y) - q_1(y)$ if $p_1(y) > q_1(y)$ and zero if $p_1(y) \leq q_1(y)$. The respective bound under non-treatment is $\int_{\mathcal{Y}} q_0(y) dy - \lambda_0 = \Pr(D = 0 | Z = 0) - \lambda_0$, such that the maximum of the proportions under treatment and non-treatment provide a lower bound on $\Pr(T = c)$. Equivalently, the lower bound

on $\Pr(T = d)$ is obtained by the maximum of $\Pr(D = 1|Z = 0) - \lambda_1$ and $\Pr(D = 0|Z = 1) - \lambda_0$.

It is obvious that Assumption 1 only allows us to derive bounds on the densities and proportions of various types. To obtain point identification, we also impose a local monotonicity assumption (LM), which rules out that compliers and defiers exist at the same time for a given value of either potential outcome. Put differently, it is assumed that the marginal potential outcome distributions of compliers and defiers under treatment and non-treatment do not overlap, i.e. that compliers and defiers inhabit distinct regions of the support of $Y(1)$ and $Y(0)$, respectively.¹ However, in contrast to (global) monotonicity, neither of the two populations is ruled out completely.

Assumption 2:

Either $\Pr(D(1) \geq D(0)|Y(d) = y(d)) = 1$ or $\Pr(D(0) \geq D(1)|Y(d) = y(d)) = 1 \ \forall y(d)$ in the support of $Y(d)$ and $d \in \{0, 1\}$ (local monotonicity).

Assumption 2 implies that whenever compliers exist locally for some $Y(d) = y(d)$ ($\Pr(T = c|Y(d) = y(d)) > 0$), defiers are ruled out ($\Pr(T = d|Y(d) = y(d)) = 0$) and vice versa. To understand the logic of this restriction and its interaction with Assumption 1, several remarks are worth noting. Firstly, even though Assumption 2 does not specify the direction of LM for particular values of the observed outcome and the treatment, it must hold that $D(1) \geq D(0)$ whenever $p_1(y) > q_1(y)$ under treatment and $q_0(y) > p_0(y)$ under non-treatment, otherwise Assumption 1 is violated. I.e., Assumption 1 tells us which direction of LM is consistent with the data. Likewise, $D(0) \geq D(1)$ whenever $p_1(y) < q_1(y)$ under treatment and $q_0(y) < p_0(y)$ under non-treatment. By taking a look at (10),(11) and (12),(13), respectively, we also see that $D(1) = D(0)$ if $p_1(y) = q_1(y)$ under treatment and $q_0(y) = p_0(y)$ under non-treatment, such that only always takers or never takers, respectively, exist in this case. Therefore, Assumption 1 and

¹We thank Joshua Angrist and Toru Kitagawa for a fruitful discussion on the interpretation of LM.

2 together imply the following:

$$\Pr(D(1) \geq D(0)|p_1(y) \geq q_1(y)) = 1 \text{ and } \Pr(D(0) \geq D(1)|p_1(y) \leq q_1(y)) = 1 \quad \forall y \in \mathcal{Y},$$

$$\Pr(D(1) \geq D(0)|q_0(y) \geq p_0(y)) = 1 \text{ and } \Pr(D(0) \geq D(1)|q_0(y) \leq p_0(y)) = 1 \quad \forall y \in \mathcal{Y}.$$

This requires that, for instance, any treated complier ($D(1) > D(0)|p_1(y) > q_1(y)$, $y \in \mathcal{Y}$) would live in a region satisfying $q_0(y') > p_0(y')$, $y' \in \mathcal{Y}$, was she not treated (which, however, still allows for distinct support regions of complier outcomes across treatment states).

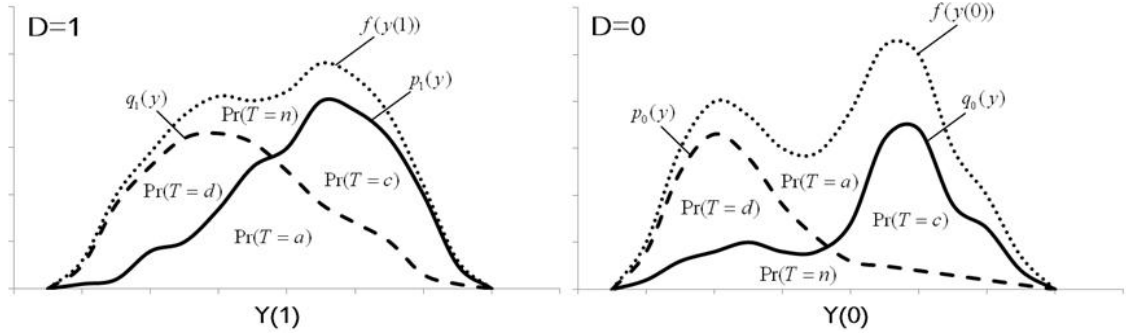
Secondly, imposing LM and ruling out defiers when $p_1(y) > q_1(y)$ or $q_0(y) > p_0(y)$ entails point identification of the compliers' density: $f(y(1), T = c) = p_1(y) - q_1(y)$ and $f(y(0), T = c) = q_0(y) - p_0(y)$, respectively, because in this case, $q_1(y)$, $p_0(y)$ only consist of always takers and never takers, respectively. I.e., the lower bounds on $f(y(1), T = c)$, $f(y(0), T = c)$ under Assumption 1 alone coincide with the point identified densities under Assumptions 1 and 2. Equivalent arguments hold for $p_1(y) < q_1(y)$ or $q_0(y) < p_0(y)$, which imply the local existence of defiers under Assumption 1 and the point identification of their densities under both assumptions, because $p_1(y)$, $q_0(y)$ now exclusively contain always takers and never takers, respectively. It therefore also follows that the share of compliers is point identified by $\Pr(D = 1|Z = 1) - \lambda_1 = \Pr(D = 0|Z = 0) - \lambda_0$, whereas the share of defiers is point identified by $\Pr(D = 1|Z = 0) - \lambda_1 = \Pr(D = 0|Z = 1) - \lambda_0$. Note that while $\Pr(D = 1|Z = 0) - \lambda_1 > 0$ and $\Pr(D = 0|Z = 1) - \lambda_0 > 0$ imply a deviation from (global) monotonicity conditional on the satisfaction of Assumption 1 as in our framework, they point to a violation of either monotonicity or Assumption 1 (or both) if not even the independence of the instrument can be ensured. This can be used to build tests for the joint satisfaction of Assumption 1 and monotonicity in the same way as for monotonicity alone given that Assumption 1 holds.²

Concerning the always takers, we have already discussed that $f(y(1), T = a) = q_1(y)$ if

²E.g., it is easy to see that testing the null hypothesis $H_0 : \Pr(D = 1|Z = 0) - \lambda_1 + \Pr(D = 0|Z = 1) - \lambda_0 = 0$ against the alternative $H_1 : \Pr(D = 1|Z = 0) - \lambda_1 + \Pr(D = 0|Z = 1) - \lambda_0 > 0$ jointly verifies Assumption 1 and monotonicity, with $\Pr(D = 1|Z = 0) - \lambda_1 + \Pr(D = 0|Z = 1) - \lambda_0$ being the sum of violations under treatment and non-treatment.

$p_1(y) > q_1(y)$ and $f(y(1), T = a) = p_1(y)$ if $p_1(y) < q_1(y)$ such that $\min(p_1(y), q_1(y))$ is the point identified density under treatment, while it was the upper bound under Assumption 1 alone. Likewise, $\min(p_0(y), q_0(y))$ point identifies the density of the never takers under non-treatment. It follows that $\lambda_1 = 1 - \delta_0$ and $1 - \delta_1 = \lambda_0$ are the respective proportions of the always takers and never takers. I.e., Assumptions 1 and 2 permit the identification of all type proportions and the density functions of the compliers and defiers under treatment and non-treatment. In contrast, the density function of the always takers is only identified under treatment (because $f(y(0))$ is unknown), that of the never takers only under non-treatment (because $f(y(1))$ is unknown). To visualize the results, Figure 1 displays the density functions $f(y(d)), p_d(y), q_d(y)$ as well as the locations and proportions of types defined by the intersections and differences of these densities for a hypothetical example under treatment and non-treatment.

Figure 1: Graphical illustration of the identification of type locations and proportions



Thirdly, the (global) monotonicity assumption of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) is a special case of LM. To see this, assume that defiers do not exist globally (i.e., assume positive monotonicity, a symmetric argument holds under negative monotonicity when compliers are ruled out) such that (11) and (12) reduce to $q_1(y) = f(y(1), T = a)$ and $p_0(y) = f(y(0), T = n)$, respectively, $\forall y \in \mathcal{Y}$. As discussed in Kitagawa (2009), this together with (10) to (13) implies the following nested density configuration: $p_1(y) \geq q_1(y)$ and $q_0(y) \geq p_0(y)$, $\forall y \in \mathcal{Y}$. Then, Assumption 2 simplifies to $\Pr(D(1) \geq D(0)) = 1$ which is (global) monotonicity.

We now formally discuss our identification results. The first proposition shows that under

Assumptions 1 and 2, the proportions of all types and the potential outcome distributions of the compliers, defiers, and always takers under treatment as well as those of the compliers, defiers, and never takers under non-treatment are identified.

Proposition 1. *Under Assumptions 1 and 2,*

1. $f(y(1), T = a) = \min(p_1(y), q_1(y))$,
2. $f(y(1), T = c) = p_1(y) - \min(p_1(y), q_1(y))$,
3. $f(y(1), T = d) = q_1(y) - \min(p_1(y), q_1(y))$,
4. $f(y(1), T = n) = f(y(1)) - \max(p_1(y), q_1(y))$
5. $f(y(0), T = n) = \min(p_0(y), q_0(y))$,
6. $f(y(0), T = c) = q_0(y) - \min(p_0(y), q_0(y))$,
7. $f(y(0), T = d) = p_0(y) - \min(p_0(y), q_0(y))$,
8. $f(y(0), T = a) = f(y(0)) - \max(p_0(y), q_0(y))$ for all $y \in \mathcal{Y}$.
9. *the type proportions are identified by*

$$\Pr(T = a) = \lambda_1, \Pr(T = c) = \Pr(D = 1|Z = 1) - \lambda_1 = \Pr(D = 0|Z = 0) - \lambda_0,$$

$$\Pr(T = d) = \Pr(D = 1|Z = 0) - \lambda_1 = \Pr(D = 0|Z = 1) - \lambda_0, \text{ and } \Pr(T = n) = \lambda_0.$$

Proof. See Appendix A. ■

Based on Proposition 1, the LATEs (i) on the compliers, (ii) on the defiers, and (iii) on the joint population of compliers and defiers are identified. The intuition is that since $f(y(d)|T = t) = \frac{f(y(d), T=t)}{\Pr(T=t)}$, we can use the results of Proposition 1 to identify the potential outcome distributions under treatment and non-treatment given $T \in \{c, d\}$ in order to identify the LATEs. Furthermore, if defiers do not exist, (i) and (iii) coincide with the standard LATE expression $\left(\frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} \right)$ under positive monotonicity, whereas (ii) and (iii) coincide with standard LATE under negative monotonicity if compliers do not exist. These results are formally stated in Proposition 2.

Proposition 2. *Under Assumptions 1 and 2,*

1.

$$\begin{aligned} E(Y(1) - Y(0)|T = c, d) &= \frac{\int_{\mathcal{Y}} y \cdot (\max(p_1(y), q_1(y)) - \min(p_1(y), q_1(y))) dy}{\Pr(D = 1|Z = 1) + \Pr(D = 1|Z = 0) - 2 \cdot \lambda_1} \\ &- \frac{\int_{\mathcal{Y}} y \cdot (\max(p_0(y), q_0(y)) - \min(p_0(y), q_0(y))) dy}{\Pr(D = 0|Z = 0) + \Pr(D = 0|Z = 1) - 2 \cdot \lambda_0}. \end{aligned} \quad (23)$$

2.

$$\begin{aligned} E(Y(1) - Y(0)|T = c) &= \frac{\int_{\mathcal{Y}} y \cdot (p_1(y) - \min(p_1(y), q_1(y))) dy}{\Pr(D = 1|Z = 1) - \lambda_1} \\ &- \frac{\int_{\mathcal{Y}} y \cdot (q_0(y) - \min(p_0(y), q_0(y))) dy}{\Pr(D = 0|Z = 0) - \lambda_0}. \end{aligned} \quad (24)$$

3.

$$\begin{aligned} E(Y(1) - Y(0)|T = d) &= \frac{\int_{\mathcal{Y}} y \cdot (q_1(y) - \min(p_1(y), q_1(y))) dy}{\Pr(D = 1|Z = 0) - \lambda_1} \\ &- \frac{\int_{\mathcal{Y}} y \cdot (p_0(y) - \min(p_0(y), q_0(y))) dy}{\Pr(D = 0|Z = 1) - \lambda_0}. \end{aligned} \quad (25)$$

4. If $\Pr(T = d) = 0$ and $\Pr(T = c) > 0$, (23) is equivalent to $E(Y(1) - Y(0)|T = c) = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)}$, whereas $E(Y(1) - Y(0)|T = d)$ is not defined.

5. If $\Pr(T = c) = 0$ and $\Pr(T = d) > 0$, (23) is equivalent to $E(Y(1) - Y(0)|T = d) = \frac{E(Y|Z=0) - E(Y|Z=1)}{E(D|Z=0) - E(D|Z=1)}$, whereas $E(Y(1) - Y(0)|T = c)$ is not defined.

Proof. See Appendix A. ■

Our discussion has shown that if the instrument satisfies Assumption 1, LATE identification does not necessarily rely on global monotonicity. In fact, the LATEs considered are equivalent to the LATE under monotonicity if the latter assumption is indeed satisfied, but can also be identified under the weaker LM, which is partially testable. Moreover, if Assumption 2 does not hold, neither does monotonicity, such that in terms of identification there appear to be no

gains when relying on standard LATE assumptions rather than the ones proposed in this section. However, albeit more general than monotonicity, LM may still be restrictive in applications, in particular with outcomes of limited support. E.g., for binary outcomes it requires that the potential outcomes of all compliers given a particular treatment state are either zero or one while all defier outcomes have the respective opposite value. Therefore, the plausibility of LM has to be critically judged in the empirical problem at hand.

To judge the implications of our assumptions in a structural model, consider the following two stage endogenous treatment selection model, with the first stage being characterized by a random coefficient model:³

$$\begin{aligned} Y_i &= \varphi(D_i, \epsilon_i), \\ D_i &= I\{\gamma_0 + \gamma_i Z_i + \nu_i > 0\}. \end{aligned} \tag{26}$$

i indexes a particular subject. $I\{\cdot\}$ is the indicator function which is equal to one if its argument holds true and zero otherwise. φ is a general function and ϵ_i, ν_i denote the unobservables in the outcome and treatment equation and may be arbitrarily correlated. γ_0, γ_i denote the constant term and the random coefficient on the instrument, respectively. Our assumptions require that whenever $p_1(Y_i) \geq q_1(Y_i)$ or $q_0(Y_i) \geq p_0(Y_i)$, respectively, γ_i is large enough to satisfy $D_i(1) = I(\gamma_0 + \gamma_i + \nu_i > 0) \geq D_i(0) = I(\gamma_0 + \nu_i > 0)$, which locally rules out defiers. A sufficient condition for this is $\gamma_i \geq 0$. For $p_1(Y_i) \leq q_1(Y_i)$ or $q_0(Y_i) \leq p_0(Y_i)$, respectively, it must hold that γ_i is small enough to satisfy $D_i(0) = I(\gamma_0 + \nu_i > 0) \geq D_i(1) = I(\gamma_0 + \gamma_i + \nu_i > 0)$. A sufficient condition for this is $\gamma_i \leq 0$. Note that global monotonicity would restrict γ_i in either one or the other way of any i , while Assumption 2 restricts γ_i only locally.

To give an idea about possible set ups in which LM holds while monotonicity does not, we provide two parametric examples that put further structure on the equations in (26). Firstly,

³We are indebted to Joshua Angrist for making valuable suggestions concerning potential models that fit our framework.

assume that the outcome equation is characterized by the following model:

$$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 D_i \epsilon_i + \epsilon_i, \quad (27)$$

where α_0 is the constant, α_1, α_2 are the coefficients on the treatment and its interaction (capturing individual effect heterogeneity), and ϵ_i is assumed to have finite first and second moments. In this case, $Y_i(0) = \alpha_0 + \epsilon_i$, $Y_i(1) = \alpha_0 + \alpha_1 + (1 + \alpha_2)\epsilon_i$ and the individual treatment effect is $\alpha_1 + \alpha_2\epsilon_i$. Moreover, assume that the coefficient on Z in the first stage is a deterministic function of ϵ_i :

$$\gamma_i = \beta_0 + \rho\epsilon_i, \quad (28)$$

where β_0 is a constant and ρ the coefficient on the unobserved term in the structural equation. For $\rho > 0$, it follows that $D_i(1) \geq D_i(0)$ for all $\epsilon_i \geq 0$ and $D_i(1) \leq D_i(0)$ for all $\epsilon_i \leq 0$, while the contrary holds for $\rho < 0$. As Y_i is a monotonic function of ϵ_i (unless α_2 is exactly -1 and $D_i = 1$), the outcomes of the compliers and defiers do not overlap conditional on the treatment state so that LM is satisfied.

Secondly, we consider an extension of our set up to a Roy (1951)-type model, which implies that the probability of treatment increases with the gains it creates. To this end, we maintain the previous outcome equation (27), but modify the first stage:

$$D_i = I\{Y_i(1) - Y_i(0) + \gamma_i Z_i + \nu_i > 0\} = I\{\beta_0 + \alpha_1 + (\alpha_2 + \rho Z_i)\epsilon_i + \nu_i > 0\}, \quad (29)$$

where the individual level treatment effect e.g., the returns to education or training, now influences the selection into treatment. In this case ν_i , if different from zero, may be interpreted as individual costs, disutility, or utility of the treatment not reflected by the treatment effect per se. The instrument Z exogenously shifts participation, but the direction depends again on ϵ_i as specified in (28). The expression left of the equality follows from substituting $Y_i(1) - Y_i(0)$ by $\alpha_1 + \alpha_2\epsilon_i$

and using (28). Again, this model implies a non-overlapping support of the potential outcomes of compliers and defiers due to γ_i being a deterministic function of ϵ_i and Y_i being monotonic in ϵ_i .

Restrictions of the kind used in our examples are not innocuous and raise the question whether there exist empirical problems in which they seem realistic. A potentially interesting application is the estimation of the returns to education based on quarter of birth instruments, see Angrist and Krueger (1991). As already discussed in Section 1, receiving a particular level of education (D) might not be monotonic in the quarter of birth (Z) due to postponing school entry (redshirting). Now assume that ϵ_i reflects the socio-economic background, see Bound, Jaeger, and Baker (1995) and Buckles and Hungerman (2013). Then, LM is satisfied if the wage (Y)⁴ is a positive function of socio-economic status and if it is the latter that also determines postponement. This is the case if parents with a high socio-economic status are more inclined to delay their children's school entry ($\rho < 0$), e.g. because they can more easily afford bearing child care costs for an extra year and/or behave more strategically in terms of schooling decisions compared to other groups. Empirical evidence pointing in this direction is provided in Bedard and Dhuey (2006), who report that children from the top quarter of the socioeconomic distribution are over-represented among those redshirting, and Aliprantis (2012), who finds that children delaying enrollment are disproportionately wealthy with better-educated parents and more books at home. Then, it is the defiers that are situated on the upper part of the wage distribution conditional on a particular level of education and compliers on the lower part.

However, assuming that redshirting is a deterministic function of ϵ_i is restrictive if other factors are likely to affect school entry postponement, too, and could thus lead to a failure of LM so that the marginal potential outcome distributions of compliers and defiers overlap. Fortunately, a violation of LM does not necessarily imply that our approach does not identify *any* causal effects at all. In fact, de Chaisemartin (2012) shows that our identification results

⁴As wage is non-negative, note that the models described before can easily be specified in a way that they generate a non-negative outcome, e.g., by letting α_0 in (27) be sufficiently large and letting ϵ_i be of bounded support such that $\alpha_0 + \alpha_1 D_i + \alpha_2 D_i \epsilon_i + \epsilon_i \geq 0$ always holds.

still apply to subgroups of compliers and defiers if LM is replaced by a weaker local *stochastic* monotonicity assumption which appears plausible in many empirical contexts:

Assumption 3:

$\forall y(1), y(0)$ in the support of $Y(1), Y(0)$:

$$\Pr(T = c|Y(1) = y(1)) \geq \Pr(T = d|Y(1) = y(1)) \text{ or } \Pr(T = c|Y(0) = y(0)) \geq \Pr(T = d|Y(0) = y(0))$$

implies that $\Pr(T = c|Y(1) = y(1), Y(0) = y(0)) \geq \Pr(T = d|Y(1) = y(1), Y(0) = y(0))$;

$$\Pr(T = c|Y(1) = y(1)) \leq \Pr(T = d|Y(1) = y(1)) \text{ or } \Pr(T = c|Y(0) = y(0)) \leq \Pr(T = d|Y(0) = y(0))$$

implies that $\Pr(T = c|Y(1) = y(1), Y(0) = y(0)) \leq \Pr(T = d|Y(1) = y(1), Y(0) = y(0))$

(local stochastic monotonicity).

In contrast to LM, Assumption 3, which is equivalent to Assumption 2.9 of de Chaisemartin (2012), allows for both compliers and defiers at any value of either marginal potential outcome distribution. It merely requires that if the share of one type weakly dominates the other conditional on *either* $Y(1)$ *or* $Y(0)$, it must also dominate conditional on both potential outcomes jointly, i.e. $Y(1)$ *and* $Y(0)$. Note that under Assumption 1, the data reveal the dominance conditional on one of the two potential outcomes: $p_1(y) \geq q_1(y)$ ($p_1(y) \leq q_1(y)$) implies that $\Pr(T = c|Y(1) = y) \geq \Pr(T = d|Y(1) = y)$ ($\Pr(T = c|Y(1) = y) \leq \Pr(T = d|Y(1) = y)$), while it follows from $q_0(y) \geq p_0(y)$ ($q_0(y) \leq p_0(y)$) that $\Pr(T = c|Y(0) = y) \geq \Pr(T = d|Y(0) = y)$ ($\Pr(T = c|Y(0) = y) \leq \Pr(T = d|Y(0) = y)$). When additionally invoking his local stochastic monotonicity, de Chaisemartin (2012) demonstrates (see his Theorem 2.5) that our identification results apply to subgroups of compliers and defiers. To be concise, the LATEs are identified for the fraction of compliers outnumbering the defiers whenever $\Pr(T = c|Y(1) = y(1), Y(0) = y(0)) \geq \Pr(T = d|Y(1) = y(1), Y(0) = y(0))$ and the fraction of defiers outnumbering the compliers whenever $\Pr(T = c|Y(1) = y(1), Y(0) = y(0)) \leq \Pr(T = d|Y(1) = y(1), Y(0) = y(0))$.

We conclude this section by discussing a testable necessary, albeit not sufficient condition for Assumptions 1 and 2, namely the satisfaction of the scale constraint (22). E.g., the proportion of always takers must be equal under treatment and non-treatment, i.e., $\lambda_1 = 1 - \delta_0$, and the same applies to the never takers, $\lambda_0 = 1 - \delta_1$, or any other subpopulation. Interestingly, one constraint

implies the other, which follows from Lemma A.2 in Kitagawa (2009):

$$\delta_1 + \delta_0 + \lambda_1 + \lambda_0 = 2,$$

because the four elements add up to the sum of the integrals of $p_1(y)$, $q_1(y)$, $p_0(y)$, and $q_0(y)$ (with $\int_{\mathcal{Y}} (p_1(y) + p_0(y))dy = \int_{\mathcal{Y}} (q_1(y) + q_0(y))dy = 1$). Rearranging the terms yields $(1 - \delta_0) - \lambda_1 = \lambda_0 - (1 - \delta_1)$. Therefore, it suffices to test one scale constraint, e.g., $\lambda_1 = 1 - \delta_0$, as its satisfaction also entails the validity of the remaining three. However, if $\lambda_1 \neq 1 - \delta_0$ (or equivalently $\lambda_0 \neq 1 - \delta_1$), point identification of LATEs is generally not feasible, at least without imposing further restrictive assumptions. It therefore appears advisable to test this implication in empirical applications and we do so in our estimations presented in Section 6. We nevertheless need to bear in mind that our assumptions may be violated even if the scale constraint is not rejected.

3 Non-binary instruments

This section discusses the identification of LATEs in the presence of a multi-valued discrete instrument with bounded support.⁵ Under (global) monotonicity, Frölich (2007) shows that if the support of Z is bounded so that $Z \in [z_{\min}, z_{\max}]$, where $z_{\min}$ and $z_{\max}$ are finite upper and lower bounds, it is possible to define and identify LATEs on the compliers with respect to any two distinct subsets of the support of Z . The proportion of compliers in general varies depending on the choice of subsets and is maximized when choosing the endpoints $z_{\min}$, $z_{\max}$. In our framework which allows for compliers and defiers, this result no longer holds in general without specifying LM more tightly. To see this, let z and $z' \in [z_{\min}, z_{\max}]$ denote two subsets such that $z \neq z'$. Define $\tilde{Z}$ as

$$\tilde{Z} = \begin{cases} 1 & \text{if } Z \in z \\ 0 & \text{if } Z \in z' \end{cases}. \quad (30)$$

⁵We thank Toru Kitagawa for very helpful comments concerning the case of non-binary instruments.

As an example, consider the case that the instrument can take three values, e.g. $Z \in \{0, 1, 2\}$, such that instead of Assumption 1 we invoke the following independence assumption:

Assumption 1a:

$$Z \perp (D(2), D(1), D(0), Y(1), Y(0)).$$

Without imposing any form of monotonicity, there now exist eight types according to $D(2), D(1), D(0)$, see Table 3. Positive monotonicity rules out types 3, 5, 6, and 7 so that only always takers (type 1), never takers (type 8) and compliers when switching the instrument from 0 to 1 (type 2) or from 1 to 2 (type 4) exist. In this framework, one could possibly think of five different definitions of z, z' : (i) $z = \{0\}, z' = \{1\}$, (ii) $z = \{1\}, z' = \{2\}$, (iii) $z = \{0\}, z' = \{2\}$, (iv) $z = \{0, 1\}, z' = \{2\}$, (v) $z = \{0\}, z' = \{1, 2\}$. (iii) maximizes the complier proportion, namely the joint proportion of types 2 and 4. This is the case because it may induce individuals to react on the treatment that are otherwise always or never takers when the instrument has less asymptotic power, i.e., operates over a smaller support, such as in (i), which only covers type 2, and in (ii), which covers type 4. In contrast, (iv) and (v) may be chosen to maximize finite sample power, because these setups encounter at least as many observations as (iii), at the cost of a weakly lower complier proportion.

Table 3: Types for $Z \in \{0, 1, 2\}$

Type T	D(2)	D(1)	D(0)
1	1	1	1
2	1	1	0
3	1	0	1
4	1	0	0
5	0	1	1
6	0	1	0
7	0	0	1
8	0	0	0

Identification becomes more complicated if we abandon (global) monotonicity. Without further restrictions, all eight types may exist, out of which two are pure compliers (types 2 and 4), two are pure defiers (types 5 and 7) and two even switch from compliance to defiance (type 6) or vice versa (type 3). Clearly, if LM is imposed w.r.t. $D(1), D(0)$ only, which allows identifying

LATEs within (i), or w.r.t. $D(2), D(1)$ only, which allows identifying LATEs within (ii), identification of LATEs in (iii) to (v) is generally not feasible. The reason is that the densities of compliers and defiers across (i) and (ii) may net each other out when coarsening the values of the instrument as in (iii) and (iv) or when considering endpoints only as in (v). I.e., some $y(1)$ and/or $y(0)$ might be inhabited by compliers in (i) and defiers in (ii) or vice versa such that any definition of z, z' not consisting of neighboring support points in Z does in general not identify LATEs. One possibility to establish identification is to assume that LM holds over all values in the support of the instrument.

Assumption 2a:

Either $\Pr(D(2) \geq D(1) \geq D(0)|y(d)) = 1$ or $\Pr(D(0) \geq D(1) \geq D(2)|y(d)) = 1$ at every value of the potential outcomes $y(d)$, $d = 0, 1$.

Assumption 2a rules out types 3 and 6 globally, implying that no individuals switch their treatment state in opposite directions for distinct pairs of instrument values. Furthermore, either defying types 5 and 7 or complying types 2 and 4 must not exist locally for any $y(d)$, meaning that over the entire range of instrument values, the support of defiers and compliers never overlaps. Under Assumptions 1a and 2a, the LATEs on types 2, 4, 5, and 7 are identified. I.e., (i) identifies the LATEs on types 2 and 7 and (ii) those on types 4 and 5. Analogously to the setup under global monotonicity, (iii) now maximizes both the proportions of compliers and defiers by identifying the LATEs on the types 2 and 4 jointly as well as on 5 and 7 jointly.

4 Estimation

This section discusses estimation based on the sample analogs of (23), (24), and (25). As discussed in Anderson, Linton, and Whang (2012), estimating parameters that involve integrals over the minimum (or maximum) of two densities, in our case of $p_d(y)$ and $q_d(y)$, is generally a non-standard problem. First of all, the limiting distribution of estimators containing such extrema can be non-normal if the two densities are equal over a non-zero mass set of points in the support

of y , referred to as contact set by the authors. Asymptotically, this is not an issue in our estimation problem, because as we argue in Section 2, $p_d(y) = q_d(y)$ immediately implies that both $f(y, T = c)$ and $f(y, T = d)$ are zero. Therefore, regions of y belonging to the contact set do not contribute to (23), (24), and (25). Nevertheless, even if $p_d(y)$ and $q_d(y)$ are not equal but sufficiently close to each other, the limiting distributions of our estimators can still be non-normal, see Anderson, Linton, and Whang (2012). For this reason, we will assume that $|p_d(y) - q_d(y)|$ is bounded away from zero with probability 1 whenever $p_d(y) \neq q_d(y)$, as formally stated in Proposition 3. This can be thought of as a local version of the first stage relevance assumption in the standard IV literature and implies that the instrument is locally not too weak.

We acknowledge that this assumption, which is solely required to ensure asymptotic normality of the estimators, may in principle be relaxed by adapting the methods proposed in Anderson, Linton, and Whang (2012) to our framework, which is left to future research. Note that by integration over y , the fact that $p_1(y) - q_1(y)$ ($q_1(y) - p_1(y)$) is bounded away from zero when $p_1(y) > q_1(y)$ ($q_1(y) > p_1(y)$) also implies that $\Pr(T = c)$ ($\Pr(T = d)$) is bounded away from zero (and an equivalent result follows under non-treatment). However, this does not necessarily imply that also their difference, $\Pr(T = c) - \Pr(T = d) = \Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)$, which corresponds to the probability limit of the denominator of the Wald estimator, is bounded away from zero. Therefore, the Wald estimator tends to infinity as complier and defier shares get arbitrarily close, while our methods are asymptotically unaffected.

If the outcome is discrete, all elements of the identification results outlined in Proposition 2 can be estimated at a parametric rate. The estimates of the densities (denoted by $\hat{p}$, $\hat{q}$) are then obtained using indicator functions for the values of Y : $\hat{p}_d(y) = \frac{1}{\sum Z_i} \sum (Z_i \cdot I\{Y_i = y, D_i = d\})$ and $\hat{q}_d(y) = \frac{1}{\sum (1 - Z_i)} \sum ((1 - Z_i) \cdot I\{Y_i = y, D_i = d\})$. The subsequent discussion focusses on the more complicated case of a continuous Y where the densities are estimated at a slower rate than $\sqrt{n}$. The LATE estimators can nevertheless be shown to be $\sqrt{n}$ -consistent and asymptotically normal under some regularity conditions. To this end, we characterize our estimation problem by a semiparametric two step GMM procedure and demonstrate that it belongs to the class of

MINPIN estimators, a general class of semiparametric two step M-estimators introduced in Andrews (1994a), see Appendix B. By applying his results it follows that the subsequent estimators of $E(Y(1) - Y(0)|T = c, d)$, $E(Y(1) - Y(0)|T = c)$, and $E(Y(1) - Y(0)|T = d)$ (denoted by $\hat{\mu}_{c,d}$, $\hat{\mu}_c$, and $\hat{\mu}_d$) have the desired properties given that the (plug-in) first step estimators $\hat{f}(Y_i|D = d, Z = z)$ satisfy particular conditions explained further below and in Appendix B.⁶

$$\begin{aligned}
\hat{\mu}_{c,d} &= \frac{\sum_{i=1}^n Y_i \cdot [I\{\hat{p}_1(Y_i) \geq \hat{q}_1(Y_i)\} \cdot (\hat{p}_1(Y_i) - \hat{q}_1(Y_i)) + I\{\hat{p}_1(Y_i) \leq \hat{q}_1(Y_i)\} \cdot (\hat{q}_1(Y_i) - \hat{p}_1(Y_i))]}{\hat{P}_{1|1} + \hat{P}_{1|0} - 2 \cdot \hat{\lambda}_1} \\
&- \frac{\sum_{i=1}^n Y_i \cdot [I\{\hat{p}_0(Y_i) \geq \hat{q}_0(Y_i)\} \cdot (\hat{p}_0(Y_i) - \hat{q}_0(Y_i)) + I\{\hat{p}_0(Y_i) \leq \hat{q}_0(Y_i)\} \cdot (\hat{q}_0(Y_i) - \hat{p}_0(Y_i))]}{\hat{P}_{0|0} + \hat{P}_{0|1} - 2 \cdot \hat{\lambda}_0}, \\
\hat{\mu}_c &= \frac{\sum_{i=1}^n Y_i \cdot I\{\hat{p}_1(Y_i) \geq \hat{q}_1(Y_i)\} \cdot (\hat{p}_1(Y_i) - \hat{q}_1(Y_i))}{\hat{P}_{1|1} - \hat{\lambda}_1} \\
&- \frac{\sum_{i=1}^n Y_i \cdot I\{\hat{p}_0(Y_i) \leq \hat{q}_0(Y_i)\} \cdot (\hat{q}_0(Y_i) - \hat{p}_0(Y_i))}{\hat{P}_{0|0} - \hat{\lambda}_0}, \\
\hat{\mu}_d &= \frac{\sum_{i=1}^n Y_i \cdot I\{\hat{p}_1(Y_i) \leq \hat{q}_1(Y_i)\} \cdot (\hat{q}_1(Y_i) - \hat{p}_1(Y_i))}{\hat{P}_{1|0} - \hat{\lambda}_1} \\
&- \frac{\sum_{i=1}^n Y_i \cdot I\{\hat{p}_0(Y_i) \geq \hat{q}_0(Y_i)\} \cdot (\hat{p}_0(Y_i) - \hat{q}_0(Y_i))}{\hat{P}_{0|1} - \hat{\lambda}_0},
\end{aligned}$$

where

$$\begin{aligned}
\hat{P}_{1|1} &= \frac{\sum_{i=1}^n D_i \cdot Z_i}{\sum_{i=1}^n Z_i}, \quad \hat{P}_{1|0} = \frac{\sum_{i=1}^n D_i \cdot (1 - Z_i)}{\sum_{i=1}^n (1 - Z_i)}, \quad \hat{P}_{0|1} = 1 - \hat{P}_{1|1}, \quad \hat{P}_{0|0} = 1 - \hat{P}_{1|0} \\
\hat{\lambda}_d &= \sum_{i=1}^n I\{\hat{p}_d(Y_i) \leq \hat{q}_d(Y_i)\} \cdot \hat{p}_d(Y_i) + I\{\hat{p}_d(Y_i) > \hat{q}_d(Y_i)\} \cdot \hat{q}_d(Y_i) \quad \text{for } d \in \{1, 0\} \\
\hat{p}_1(Y_i) &= \hat{P}_{1|1} \cdot \hat{f}(Y_i|D = 1, Z = 1), \quad \hat{q}_1(Y_i) = \hat{P}_{1|0} \cdot \hat{f}(Y_i|D = 1, Z = 0), \\
\hat{p}_0(Y_i) &= \hat{P}_{0|1} \cdot \hat{f}(Y_i|D = 0, Z = 1), \quad \hat{q}_0(Y_i) = \hat{P}_{0|0} \cdot \hat{f}(Y_i|D = 0, Z = 0), \\
\text{and } \hat{f}(Y_i|D = d, Z = z) &\text{ is a non-parametric preliminary estimator of } f(Y_i|D = d, Z = z).
\end{aligned}$$

Andrews (1994a) proposes a set of assumptions under which the class of MINPIN estimators is $\sqrt{n}$ -consistent and asymptotically normal. The assumptions include standard regularity con-

⁶Instead of evaluating the densities at the empirical data points, asymptotically equivalent estimators can be obtained by estimating the densities at an equidistant grid of values between the empirical lower and upper bound of the outcome support and taking the sample analogs of Proposition 2. The asymptotic properties can be derived in a similar manner and are not discussed here.

ditions (e.g., boundedness of the parameter space of the second step objects) and an orthogonality condition between first step objects (the densities) and second step parameters (the LATEs, λ_d , and $\Pr(D = d|Z = z)$). The latter ensures that the first step estimators of the densities $f(Y_i|D = d, Z = z)$ do not affect the asymptotic variances of the LATEs, which requires that they converge uniformly (rather than pointwise) sufficiently fast, i.e., at least at rate $n^{-\frac{1}{4}}$.⁷ As an example, Andrews (1995) discusses conditions under which nonparametric multidimensional kernel estimators satisfy this property in an estimation problem with a particular form of weak temporal dependence.

Here, the first step problem is econometrically less challenging because we only need to estimate one-dimensional densities in an i.d.d. framework. Uniform almost sure convergence (which implies convergence in probability) can be easily established if the support of Y is unbounded, see for instance Theorem 1.4 in Li and Racine (2007). If the support is bounded, the bias of (local constant) kernel density estimators is potentially large close to the boundaries and of lower order than in the interior. To obtain uniform convergence in this case, one may either use specific boundary kernels designed to overcome this problem, or a local linear density estimator (instead of local constant estimation) where the bias is of the same order at the boundaries as in the interior, see for instance Jones (1993), or adaptive bandwidth methods for boundaries as discussed in Dai and Sperlich (2010).⁸ Concerning the required rate of convergence of $n^{-\frac{1}{4}}$, the latter is easily obtained in the univariate case, where the fastest possible rate is $n^{-\frac{2}{5}}$.

A further important assumption in Andrews (1994a) is the smoothness of the expectations of the moment functions. Note that the presence of indicator functions (such as for example $I\{\hat{p}_1(Y_i) > \hat{q}_1(Y_i)\}$) in the LATE estimators does not allow imposing such smoothness conditions at the unit level as for example discussed in Newey (1994). However, as the estimators contain averages of these indicator functions, the smoothness condition of Andrews (1994a) is satisfied

⁷MINPIN estimators that do not satisfy the orthogonality condition may nevertheless be asymptotically normal. However, the variance term of the second step estimators will then also be a function of the first step estimation, see Andrews (1991).

⁸We refer to Cheng, Fan, and Marron (1997) and Dai and Sperlich (2010) for a comprehensive review of the literature on boundary corrections in kernel density estimation

in our case. This gives rise to a stochastic equicontinuity assumption on the empirical processes involved, which is a further requirement of asymptotic normality. Proposition 3 states that under a locally sufficiently strong instrument and Assumption E in Appendix B, which adapts the assumptions of Andrews (1994a) to our framework, the LATE estimators are $\sqrt{n}$ -consistent and asymptotically normal.

Proposition 3. *Assume that (i) for any $\eta \geq 1/2$ and for a non-negative function $g(\cdot)$ it holds that $\Pr(|p_d(y) - q_d(y)| > n^{-\eta} \cdot g(y)) = 1$ for all $y \in \mathcal{Y} : p_d(y) \neq q_d(y)$ and $d = 1, 0$, and (ii) Assumption E in Appendix B is satisfied. It follows that*

1. $\hat{\mu}_{c,d} \xrightarrow{P} E(Y(1) - Y(0)|T = c, d)$
2. $\hat{\mu}_c \xrightarrow{P} E(Y(1) - Y(0)|T = c)$
3. $\hat{\mu}_d \xrightarrow{P} E(Y(1) - Y(0)|T = d)$
4. $\hat{\lambda}_1 \xrightarrow{P} \Pr(T = a), \hat{\lambda}_0 \xrightarrow{P} \Pr(T = n)$
5. $\sqrt{n} \cdot (\hat{\mu}_{c,d} - E(Y(1) - Y(0)|T = c, d)) \xrightarrow{d} \mathcal{N}(0, V_{c,d})$
6. $\sqrt{n} \cdot (\hat{\mu}_c - E(Y(1) - Y(0)|T = c)) \xrightarrow{d} \mathcal{N}(0, V_c)$
7. $\sqrt{n} \cdot (\hat{\mu}_d - E(Y(1) - Y(0)|T = d)) \xrightarrow{d} \mathcal{N}(0, V_d)$
8. $\sqrt{n} \cdot (\hat{\lambda}_1 - \Pr(T = a)) \xrightarrow{d} \mathcal{N}(0, V_{\lambda_1}), \sqrt{n} \cdot (\hat{\lambda}_0 - \Pr(T = n)) \xrightarrow{d} \mathcal{N}(0, V_{\lambda_0})$

with $V_{c,d}$, V_c , V_d , V_{λ_1} , and V_{λ_0} as given in Appendix C.

Proof. See Appendix B. ■

Chen, Linton, and Keilegom (2003) generalize the results of Andrews (1994a) to a wider class of estimators (including the MINPIN class) and discuss the assumptions needed for the validity of bootstrap based inference. We show in Appendix B that $V_{c,d}$, V_c , V_d , and V_{λ_1} can be consistently estimated by bootstrapping under only slightly stronger assumptions than required for estimation per se. There, we also discuss in greater detail the assumptions of Andrews (1994a) and Chen, Linton, and Keilegom (2003) and the conditions under which they are satisfied in our setting.

5 Simulations

To investigate the finite sample properties of our estimator, we consider a brief simulation study in which the data are generated based on the following potential outcome distributions and shares of the various types:

$$\begin{aligned}
Y(1)|T = a &\sim \mathcal{U}(0, 4), \quad Y(0)|T = n \sim \mathcal{U}(0, 4), \\
Y(d)|T = c &\sim \mathcal{U}(1, 2) + \alpha d, \quad Y(d)|T = d \sim \mathcal{U}(2, 3) + \beta d, \quad d \in \{1, 0\}, \\
\Pr(T = a) &= 0.5 - 0.5 \cdot \Pr(T = c), \quad \Pr(T = n) = 0.5 - 0.5 \cdot \Pr(T = c) - \Pr(T = d), \\
Z &\sim \text{Bernoulli}(0.5).
\end{aligned}$$

$\mathcal{U}(A, B)$ is the continuous uniform distribution in the interval $[A, B]$. In our setup, the LATEs on the compliers and defiers are α and β , respectively, and the shares of always and never takers are defined as functions of the shares of compliers and defiers, which are to be defined. Note that the random assignment of Z implies the satisfaction of Assumption 1. Furthermore, Assumption 2 is satisfied because the supports of compliers' and defiers' potential outcomes do not overlap in either treatment state. Finally, it follows from the definition of types that $D = Z$ if $T = c$, $D = 1 - Z$ if $T = d$, $D = 1$ if $T = a$, and $D = 0$ if $T = n$.

We run 1000 simulations and consider two different scenarios. In the first one, we set the LATEs to $\alpha = 1$ and $\beta = -1$, the complier and defier shares to $\Pr(T = c) = 0.20$ and $\Pr(T = d) = 0.15$, respectively, and the sample sizes to $n = 1000, 4000, 16000$. This allows investigating how the performance of the estimator evolves as n increases when complier and defier shares are not too small. By our simulation design, $p_1(y) \geq q_1(y)$ ($q_0(y) \geq p_0(y)$) for all $y \in [0, 2)$, while $p_1(y) \leq q_1(y)$ ($q_0(y) \leq p_0(y)$) for all $y \in (0, 2]$. At $y = 2$, the respective functions switch discontinuously in terms of dominance, which jeopardizes density estimation in areas close to this point. It appears interesting to investigate whether the LATE estimators can perform satisfactorily even in the presence of such irregular behaviour of the density functions. In the second scenario, we aim at constructing a case that is comparable to the empirical application

presented in the next section, where the sample size is substantially larger, while the shares of compliers and defiers (under the satisfaction of Assumptions 1 and 2) are only around 1% or even less. We therefore set the simulation parameters close to the results obtained in the empirical application: $n = 250000$, $\Pr(T = c) = 0.012$, $\Pr(T = d) = 0.010$, $\alpha = 0.250$, $\beta = 0.300$.

Estimation is based on the sample analogs of Proposition 2. The estimates of $p_d(y)$, $q_d(y)$ are obtained by kernel density estimation (using the Gaussian kernel) within subgroups defined by the treatment and the instrument and evaluated on an equidistant grid of 1000 values between the simulated lower and upper bounds of the outcome support. The Silverman (1986) rule of thumb, denoted by b_s , is used to choose the bandwidths in each of the four subpopulations defined upon D and Z . As a second option, we also consider undersmoothing by defining the bandwidth as $b_s^{3/2}$ as also applied in Anderson, Linton, and Whang (2012).

The simulation results are reported in Table 4. Considering the first design, we see that the estimators of LATE_c , LATE_d , and $\text{LATE}_{c,d}$ based on b_s are generally somewhat biased under $n = 1000$. In contrast, undersmoothing ($b_s^{3/2}$) entails considerably lower biases which more than offset potential costs in terms of increased standard errors. Therefore, also mean squared errors (MSE) are substantially lower in the latter case. As the sample size increases, biases and variances go down under both bandwidth rules, but undersmoothing is always superior in terms of MSE and bias. Obviously, the undersmoothed estimators perform better in the area where compliers and defiers are ‘close’, namely at the discontinuity with $y = 2$, because the use of a smaller bandwidth entails comparably large bias reductions. Nevertheless, even estimation based on b_s appears satisfactory in larger samples despite the irregular behavior of the density functions considered. As expected, additional simulations (not presented here) also show that b_s becomes relatively more competitive compared to $b_s^{3/2}$ as the distance between the support regions of compliers and defiers increases (for instance when setting $\alpha = 2$, $\beta = -2$). Then, either p_d or q_d are discontinuous at a particular y , but not both at the same time. In contrast, the Wald estimator ($\text{LATE}(\text{mon})$) remains severely biased and unprecise in any setup even when $n = 16000$.

In the second scenario, the Silverman (1986) rule of thumb slightly dominates undersmoothing in terms bias and MSE, which may be due to the smaller magnitudes of discontinuities in p_d , q_d compared to the first scenario. In spite of the low complier and defier shares, the biases of the proposed estimators are not too severe thanks to the large sample size considered. Interestingly, the MSEs of the estimators using b_s are comparable to those first scenario with $n = 4000$ and considerably larger complier and defier proportions. This highlights the trade-off between the sample size and complier/defier shares in estimation for a given MSE. Again, LATE(mon) performs very poorly in terms of bias, standard error, and MSE, as it only exploits the difference in complier and defier shares as first stage (0.002), which creates a severe weak instrument problem.

Table 4: Simulations

	SCENARIO 1: $\Pr(T = c) = 0.2$, $\Pr(T = d) = 0.15$, $\alpha = 2$ and $\beta = -2$						
	<i>Silverman (1986) rule (b_s)</i>			<i>undersmoothing ($b_s^{3/2}$)</i>			LATE(mon)
	LATE _c	LATE _d	LATE _{c,d}	LATE _c	LATE _d	LATE _{c,d}	
n= 1000							
Bias	0.197	-0.274	0.033	0.036	-0.053	0.005	11.121
S.e.	0.131	0.162	0.139	0.144	0.183	0.117	79.284
MSE	0.056	0.101	0.020	0.022	0.036	0.014	6409.575
n= 4000							
Bias	0.143	-0.209	0.018	0.021	-0.036	0.000	6.920
S.e.	0.075	0.092	0.069	0.079	0.102	0.060	3.955
MSE	0.026	0.052	0.005	0.007	0.012	0.004	63.528
n= 16000							
Bias	0.106	-0.156	0.010	0.014	-0.022	-0.001	6.211
S.e.	0.038	0.046	0.031	0.039	0.052	0.028	1.231
MSE	0.013	0.026	0.001	0.002	0.003	0.001	40.092
	SCENARIO 2: $\Pr(T = c) = 0.012$, $\Pr(T = d) = 0.010$, $\beta = 0.250$ and $\alpha = 0.300$						
	<i>Silverman (1986) rule (b_s)</i>			<i>undersmoothing ($b_s^{3/2}$)</i>			LATE(mon)
	LATE _c	LATE _d	LATE _{c,d}	LATE _c	LATE _d	LATE _{c,d}	
n= 250000							
Bias	-0.057	-0.084	-0.070	-0.083	-0.108	-0.096	12.832
S.e.	0.170	0.216	0.094	0.169	0.207	0.093	529.225
MSE	0.032	0.054	0.014	0.036	0.055	0.018	280244.137

Note: Results are based on 1000 simulations. Bandwidths for kernel density estimation (based on the Gaussian kernel) are selected using the Silverman (1986) rule of thumb and the rule of thumb^{3/2} (undersmoothing), respectively. MSE: mean squared error. S.e.: standard error.

6 Empirical application

This section provides an application to 1980 U.S. census data analyzed by Angrist and Krueger (1991), which (among other cohorts) contain 486,926 males born in 1940-49. Angrist and Krueger (1991) assess the effect of education on wages by using the quarter of birth as instrument to control for potential endogeneity (e.g., due to unobserved ability) between the treatment and the outcome. The idea is that the quarter of birth instrument affects education through age-related schooling regulations. As documented in Angrist and Krueger (1992), state-specific rules require that a child must have attained the first grade admission age, which is six years in most cases, at a particular date during the year. Because schooling is compulsory until the age of 16 in most states, see Appendix 2 in Angrist and Krueger (1991), students who are born early in the year are in 10th grade when turning 16. As the school year usually starts in September and ends in July, these students have nine years of completed education if they decide to quit education as soon as possible. In contrast, students born after the end of the academic year but still entering school in the same year they turn six will have ten years of completed education at age 16. This suggests education to be monotonically increasing in the quarter of birth.

However, the quarter of birth instrument is far from being undisputed. E.g., Bound, Jaeger, and Baker (1995) challenge the validity of the exclusion restriction and present empirical results that point to systematic patterns in seasonality of birth (for instance w.r.t. performance in school, health, and family income) which may imply a direct association with the outcome. In line with these arguments, Buckles and Hungerman (2013) document large differences in maternal characteristics for births throughout the year (with winter births being more often realized by teenagers and unmarried women) based on U.S. birth certificate data and census data. For this reason, we will only consider quarters two and three in our analysis, i.e., the warmer seasons of the year. We acknowledge that this may not completely dissipate concerns about seasonality, but nevertheless assume that Assumption 1 is satisfied for the subsample born in the second or third quarters of the year.

Secondly, a crucial question for standard IV estimation is whether positive monotonicity holds for all individuals. This appears unlikely in the light of strategic school entry behavior as documented by Barua and Lang (2009), which may entail deviations from the schooling regulations. The authors present empirical evidence of redshirting based on 1960 U.S. census data, implying that many parents did not enroll their children at the earliest permissible entry age but postponed school entry. This occurred particularly often when born late in the year. Aliprantis (2012) provides further empirical support for redshirting based on the Early Childhood Longitudinal Study. Moreover, Klein (2010) acknowledges that postponement may also be induced by schools, which are generally not obliged to admit all children who turn six before the state-wide cutoff date. As discussed in Klein (2010), both redshirting and school policies may reverse the relation of education and the instrument for some individuals. Because children with postponement are close to seven when entering school and will just have started 10th grade when turning 16, some of them may decide to drop out immediately, with only nine years of completed education. In contrast, students born earlier will be at an advanced stage of the 10th grade when turning 16 and might therefore decide to complete the grade, thus having at least 10 years of completed education. For these individuals, compulsory schooling decreases in the quarter of birth and therefore violates monotonicity.

The implausibility of monotonicity motivates the use of our weaker LM, while the exclusion restriction will be maintained. As already mentioned, we confine our analysis to those males born in the second or third quarters (244,512 observations). The instrument Z is equal to zero if born in the second quarter and equal to one if born in the third quarter. Our treatment D is a binary indicator that is equal to zero if receiving high school education or less (i.e., up to 12 years of education) and one if obtaining at least some higher education (i.e., 13 years or more). I.e., we are interested in the returns to having at least some college education. According to our definition, roughly 48% (52%) of our sample receive lower (higher) education. The outcome variable Y is the log weekly wage.

As in Section 5, the densities $p_d(y)$, $q_d(y)$ are estimated by kernel density estimation within

subgroups defined by the treatment and the instrument and evaluated on an equidistant grid of 1000 values between the empirical lower and upper bounds of the outcome support. Again, the Silverman (1986) rule of thumb (b_s) and undersmoothing ($b_s^{3/2}$) are used for bandwidth choice. Concerning inference, we bootstrap the parameters of interest 1999 times to approximate their distributions and compute p-values by assessing the rank of the estimates in their respective re-centered bootstrap distributions. We use two-sided hypothesis tests, see for instance equation (6) of MacKinnon (2006), to obtain the p-values of the scale constraint and the LATE estimates and one-sided tests for the type proportions (the theoretical lower bound of which is zero). However, the bootstrap is inconsistent at boundaries of the parameter space, see Andrews (2000). In finite samples, we would therefore for instance expect it to be inadequate for the complier and defier shares if the latter are ‘too close’ to zero. For this reason, we also present p-values based on subsampling with replacement (also known as m out of n bootstrap), see Bickel, Götze, and van Zwet (1997), which remains pointwise consistent even at boundaries. The subsampling size is set to half of the sample size.

Table 5 presents the estimation results based on the different bandwidth rules (b_s , $b_s^{3/2}$) and inference approaches (bootstrapping, subsampling with replacement). The second column gives the estimates of $1 - \delta_0 - \lambda_1$, which tests the scale constraint and is necessarily zero if Assumptions 1 and 2 are satisfied. Under either bandwidth rule and inference method, the estimates are close to zero and insignificant (at the 10% level) so that our identifying assumptions are not rejected. The third and fourth columns contain the estimates of the complier and defier shares, respectively. Albeit very small (1% or even less), either proportion is significant at the 5% level in any estimation setup, which points to a violation of global monotonicity in the data, rendering the Wald estimator generally inconsistent. Furthermore, the difference between complier and defier shares is merely 0.2%. As this corresponds to the estimate of $\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)$, i.e. the denominator of the Wald estimator, the similar magnitudes of compliers and defiers entail a weak instrument problem for the latter. Indeed, a first stage OLS regression of D on a constant and Z yields a t-value of only 0.778 for the coefficient on Z , so that one would incorrectly conclude

that the share of compliers is not statistically different from zero when incorrectly assuming global monotonicity. Therefore, even in the special case that the LATEs on compliers and defiers are equal so that the Wald estimator is consistent despite ignoring defiers, see Angrist, Imbens, and Rubin (1996), it likely performs poorly under similar complier/defier shares due to an explosion of its variance. For this reason, our proposed methods not only come with potential gains in robustness, but also in efficiency, as compliers and defiers do not net each other out therein.

Comparing the various LATE estimates (columns 5-8) provides evidence in this direction. Due to its close-to-zero denominator, the precision of the Wald estimator (LATE(mon)) is extremely low and therefore no reliable conclusions on the returns to higher education can be drawn. In contrast, the LATE on the joint population of compliers and defiers, $\text{LATE}_{c,d}$, is highly significant even under subsampling and suggests that (at least some) higher education increases weekly wages by roughly 30 to 50%. Expectedly, the separately estimated effects on compliers and defiers (LATE_c , LATE_d) are somewhat less precise, but yet statistically significant at the 10% level in any setup. In conclusion, our estimates suggest that the wage effects of higher education are substantial and quite homogeneous across compliers and defiers.

Table 5: LATEs on log weekly wage among cohorts born in the 40s (n=244,512)

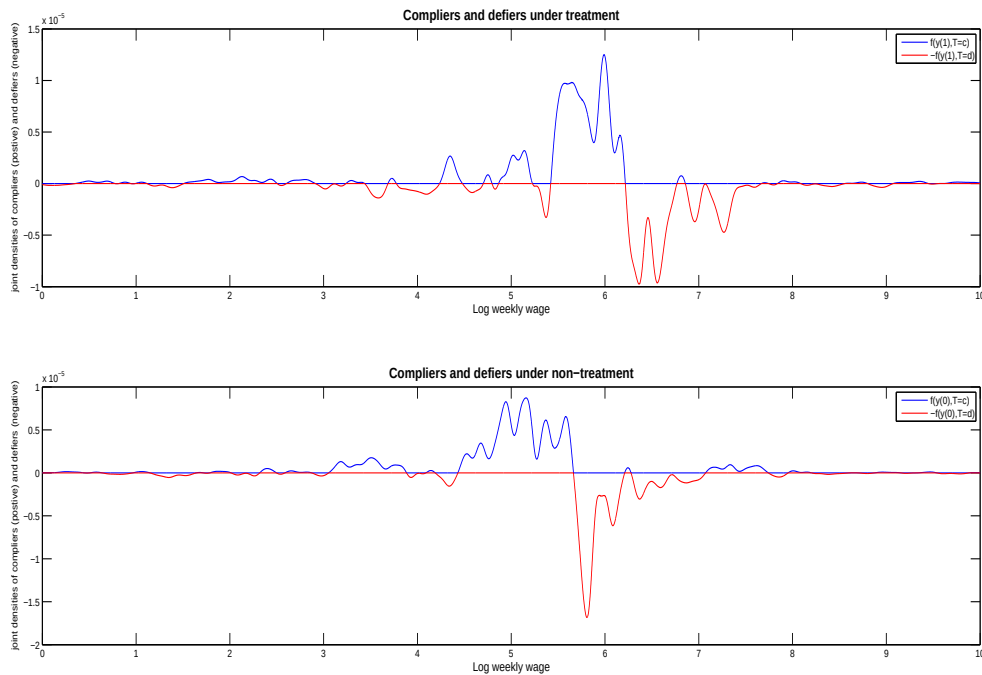
	$1 - \delta_0 - \lambda_1$	$\Pr(T = c)$	$\Pr(T = d)$	LATE_c	LATE_d	$\text{LATE}_{c,d}$	LATE(mon)
Estimate (b_s)	-0.001	0.008	0.006	0.445	0.435	0.445	0.184
P-value (bootstrap)	(0.644)	(0.001)	(0.001)	(0.001)	(0.035)	(0.000)	(0.906)
P-value (subsampling)	(0.737)	(0.010)	(0.031)	(0.012)	(0.073)	(0.000)	(0.915)
Estimate ($b_s^{3/2}$)	0.002	0.012	0.010	0.237	0.300	0.265	0.184
P-value (bootstrap)	(0.174)	(0.000)	(0.000)	(0.008)	(0.003)	(0.000)	(0.906)
P-value (subsampling)	(0.119)	(0.001)	(0.009)	(0.087)	(0.022)	(0.001)	(0.915)

Note: Bandwidths for kernel density estimation (based on the Gaussian kernel) are selected using the Silverman (1986) rule of thumb and the rule of thumb^{3/2} (undersmoothing), respectively. Bootstrap and subsampling p-values are based on resampling 1999 times. Subsampling draws 122,256 observations (half the sample size) with replacement from the original data.

Even though the results indicate that global monotonicity is violated so that our methods appear preferable to the Wald estimator, it has to be critically assessed whether LM is a plausible alternative. As already mentioned, the potential wages of compliers and defiers may overlap

despite the (asymptotic) non-rejection of the scale constraint. As a plausibility check, it therefore appears useful to also visually inspect the outcome support of both groups, given that there exists a reasonable theory concerning which outcome regions should be exclusively inhabited by either compliers or defiers. E.g., as discussed in Section 2, it may seem reasonable that individuals with high potential wages come from families with a high socio-economic status and that the latter also had a positive impact on the probability to postpone school entry, as evidenced in Bedard and Dhuey (2006) and Aliprantis (2012). This suggests that conditional on treatment, defiers are concentrated in the upper part of the wage distribution and compliers in the lower part.

Figure 2: Est. of $f(y(1), T = c)$, $f(y(1), T = d)$ and $f(y(0), T = c)$, $f(y(0), T = d)$



To verify our presumption, Figure 2 plots the estimates of the joint densities (using b_s) of the outcome values and being a complier or defier, respectively, under treatment ($f(y(1), T = c)$, $f(y(1), T = d)$) and non-treatment ($f(y(0), T = c)$, $f(y(0), T = d)$). Positive densities represent compliers, negative ones defiers. We see that our theory is roughly supported by the data. In either treatment state, defiers are concentrated in the upper part of the distribution, while

negative densities rare and small (in absolute terms) for lower values of log weekly wage, where complier densities are comparatively large. Still, LM may not be entirely convincing in the problem considered, because it could well be that socio-economic status is not the only factor determining redshirting behavior. Our estimates nevertheless bear a causal interpretation for subsets of the compliers and defiers if one is willing to replace LM by the weaker local stochastic monotonicity (in terms of joint potential outcomes) of de Chaisemartin (2012), see Assumption 3 in Section 2. In the latter case, the plots in Figure 2 suggest that the defiers generally outnumber the compliers among high values of $(Y(1), Y(0))$ while the opposite holds for low joint potential outcomes. It is then the respective ‘excess’ compliers/defiers (relative to the other group) for which the effects are estimated.

7 Conclusion

We have demonstrated that local average treatment effects (LATEs) are identified under strictly weaker conditions than the standard assumptions invoked in the literature, see Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). Under the joint independence of the instrument and the potential treatment states/outcomes, (global) monotonicity of the treatment in the instrument may be weakened to local monotonicity (LM). This brings the improvement that defiers need no longer be assumed away such that LATEs on the defiers as well as on the joint population of defiers and compliers are identified for the first time in addition to the effect on the compliers. Furthermore, our set of assumptions can be partly tested in the data. Even though improving on monotonicity, it nevertheless has to be acknowledged that LM is by no means innocuous, because it requires that the marginal potential outcome distributions of the latent defier and complier populations do not overlap in either treatment state. However, even if LM does not hold, our approach still identifies causal effects for subgroups of compliers and defiers, given that the weaker *stochastic* local monotonicity assumption suggested by de Chaisemartin (2012) holds, which appears plausible in many applications.

As an empirical illustration, we have applied our methods to U.S. census data previously analyzed by Angrist and Krueger (1991) to estimate the returns to higher education for males born in 1940-49 by using the birth quarter as instrument for education. We have documented the presence of both compliers and defiers and demonstrated that LATE estimation is not robust to ignoring defiers. In particular, the Wald estimator is very imprecise because compliers and defiers net each other out in its denominator and therefore create a weak instrument problem. In contrast, our methods assuming LM are much more efficient and predict similarly large returns to higher education for both compliers and defiers. Finally, we have also highlighted how a visual inspection of the estimated complier and defier distributions can help to assess the plausibility of LM.

A Proofs of the identification results

Proof of Proposition 1.

1. Invoking Assumption 1 and subtracting (11) from (10), we obtain

$$\begin{aligned}
 p_1(y) - q_1(y) &= f(y(1), T = c) - f(y(1), T = d) \\
 &= (\Pr(T = c|Y(1) = y) - \Pr(T = d|Y(1) = y)) \cdot f(y) \\
 &= (\Pr(D(1) > D(0)|Y(1) = y) - \Pr(D(1) < D(0)|Y(1) = y)) \cdot f(y), \quad (\text{A.1})
 \end{aligned}$$

where the last line follows from the fact that $\Pr(T = c) = \Pr(D(1) > D(0))$ and $\Pr(T = d) = \Pr(D(1) < D(0))$. Under Assumption 2, either $\Pr(D(1) \geq D(0)|Y(1) = y) = 1$ or $\Pr(D(1) \leq D(0)|Y(1) = y) = 1$, which implies that either $\Pr(D(1) < D(0)|Y(1) = y) = 0$ or $\Pr(D(1) > D(0)|Y(1) = y) = 0$. Therefore, it follows from (A.1) that $p_1(y) \geq q_1(y)$ if and only if $\Pr(D(1) < D(0)|Y(1) = y) = 0$ and $p_1(y) \leq q_1(y)$ if and only if $\Pr(D(1) > D(0)|Y(1) = y) = 0$.

Again considering (10) and (11), the equations can be rewritten as

$$\begin{aligned}
 p_1(y) &= \Pr(D(1) > D(0)|Y(1) = y) \cdot f(y) + f(y(1), T = a), \\
 q_1(y) &= \Pr(D(1) < D(0)|Y(1) = y) \cdot f(y) + f(y(1), T = a).
 \end{aligned}$$

Thus, $f(y(1), T = a) = q_1(y)$ if $p_1(y) \geq q_1(y)$, because then $\Pr(D(1) < D(0)|Y(1) = y) = 0$, and $f(y(1), T = a) = p_1(y)$ if $p_1(y) \leq q_1(y)$, because then $\Pr(D(1) > D(0)|Y(1) = y) = 0$. This implies $f(y(1), T = a) = \min(p_1(y), q_1(y))$.

2. This follows immediately by substituting $f(y(1), T = a) = \min(p_1(y), q_1(y))$ into (10).
3. This follows immediately by substituting $f(y(1), T = a) = \min(p_1(y), q_1(y))$ into (11).
4. Using a similar argument as in point 1, one can show that our assumptions together with (14) and (15) imply that $f(y(1), T = n) = f(y(1)) - p_1(y)$ if $p_1(y) \geq q_1(y)$ and $f(y(1), T = n) = f(y(1)) - q_1(y)$ if $p_1(y) \leq q_1(y)$, therefore $f(y(1), T = n) = f(y(1)) - \max(p_1(y), q_1(y))$.
5. Using a similar argument as in point 1, one can show that our assumptions together with (12) and (13) imply that $f(y(0), T = n) = q_0(y)$ if $q_0(y) \geq p_0(y)$ and $f(y(0), T = n) = p_0(y)$ if $q_0(y) \leq p_0(y)$, therefore $f(y(0), T = n) = \min(p_0(y), q_0(y))$.
6. This follows immediately by substituting $f(y(0), T = n) = \min(p_0(y), q_0(y))$ into (13).
7. This follows immediately by substituting $f(y(0), T = n) = \min(p_0(y), q_0(y))$ into (12).
8. Using a similar argument as in point 1, one can show that our assumptions together with (16) and (17) imply

that $f(y(0), T = a) = f(y(0)) - q_0(y)$ if $q_0(y) \geq p_0(y)$ and $f(y(0), T = a) = f(y(0)) - p_0(y)$ if $q_0(y) \leq p_0(y)$, therefore $f(y(0), T = a) = f(y(0)) - \max(p_0(y), q_0(y))$.

9. Integration of $f(y(1), T = a) = \min(p_1(y), q_1(y))$ immediately gives $\Pr(T = a) = \lambda_1$ and integration of $f(y(0), T = n) = \min(p_0(y), q_0(y))$ gives $\Pr(T = n) = \lambda_0$. Concerning the remaining type proportions note that the integrals over (10) and (11) give $\Pr(T = c) = \Pr(D = 1|Z = 1) - \Pr(T = a)$ and $\Pr(T = d) = \Pr(D = 1|Z = 0) - \Pr(T = a)$. Likewise, the integrals over (13) and (12) yield $\Pr(T = c) = \Pr(D = 0|Z = 0) - \Pr(T = n)$ and $\Pr(T = d) = \Pr(D = 0|Z = 1) - \Pr(T = n)$. It follows that $\Pr(T = c) = \Pr(D = 1|Z = 1) - \lambda_1 = \Pr(D = 0|Z = 0) - \lambda_0$ and $\Pr(T = d) = \Pr(D = 1|Z = 0) - \lambda_1 = \Pr(D = 0|Z = 1) - \lambda_0$.

■

Proof of Proposition 2.

1. It suffices to show that $f(y(1)|T = d, c) = \frac{\max(p_1(y), q_1(y)) - \min(p_1(y), q_1(y))}{\Pr(D=1|Z=1) + \Pr(D=1|Z=0) - 2 \cdot \lambda_1}$ since a symmetric argument can be used to demonstrate that $f(y(0)|T = d, c) = \frac{\max(p_0(y), q_0(y)) - \min(p_0(y), q_0(y))}{\Pr(D=0|Z=0) + \Pr(D=0|Z=1) - 2 \cdot \lambda_0}$. From Proposition 1, it follows that

$$\begin{aligned} f(y(1), T = c, d) &= f(y(1), T = c) + f(y(1), T = d) = p_1(y) + q_1(y) - 2 \cdot \min(p_1(y), q_1(y)) \\ &= \max(p_1(y), q_1(y)) - \min(p_1(y), q_1(y)). \end{aligned}$$

Therefore,

$$\begin{aligned} f(y(1)|T = c, d) &= \frac{f(y(1), T = c, d)}{\Pr(T = c, d)} = \frac{f(y(1), T = c, d)}{\pi_c + \pi_d} \\ &= \frac{\max(p_1(y), q_1(y)) - \min(p_1(y), q_1(y))}{\Pr(D = 1|Z = 1) + \Pr(D = 1|Z = 0) - 2 \cdot \lambda_1}, \end{aligned}$$

which ends this part of the proof.

2. Similarly to the proof of point 1 we will just show that $f(y(1)|T = c) = \frac{p_1(y) - \min(p_1(y), q_1(y))}{\Pr(D=1|Z=1) - \lambda_1}$. From Proposition 1 it follows that

$$f(y(1)|T = c, d) = \frac{f(y(1), T = c)}{\Pr(T = c)} = \frac{p_1(y) - \min(p_1(y), q_1(y))}{\Pr(D = 1|Z = 1) - \lambda_1},$$

which ends this part of the proof.

3. The proof of this point is symmetric to the one of point 2 and is therefore omitted.
4. Under positive monotonicity,

$$\begin{aligned} \max(p_1(y), q_1(y)) &= p_1(y), & \max(p_0(y), q_0(y)) &= q_0(y), & \min(p_1(y), q_1(y)) &= q_1(y), \\ \min(p_0(y), q_0(y)) &= p_0(y), & \lambda_1 &= \Pr(D = 1|Z = 0), & \lambda_0 &= \Pr(D = 0|Z = 1). \end{aligned}$$

Therefore, (23) simplifies to

$$\begin{aligned} E(Y(1) - Y(0)|T = c, d) &= \frac{\int_{\mathcal{Y}} y \cdot (p_1(y) - q_1(y))dy}{\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)} - \frac{\int_{\mathcal{Y}} y \cdot (q_0(y) - p_0(y))dy}{\Pr(D = 0|Z = 0) - \Pr(D = 0|Z = 1)} \\ &= \frac{\int_{\mathcal{Y}} y \cdot p_1(y)dy - \int_{\mathcal{Y}} y \cdot q_1(y)dy}{\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)} - \frac{\int_{\mathcal{Y}} y \cdot q_0(y)dy - \int_{\mathcal{Y}} y \cdot p_0(y)dy}{\Pr(D = 0|Z = 0) - \Pr(D = 0|Z = 1)}. \end{aligned}$$

Considering $\int_{\mathcal{Y}} y \cdot p_1(y)dy$, it is easy to see that

$$\begin{aligned} \int_{\mathcal{Y}} y \cdot p_1(y)dy &= \int_{\mathcal{Y}} y \cdot f(Y, D = 1|Z = 1)dy = \int_{\mathcal{Y}} y \cdot f(Y|Z = 1, D = 1) \cdot \Pr(D = 1|Z = 1)dy, \\ &= \Pr(D = 1|Z = 1) \cdot \int_{\mathcal{Y}} y \cdot f(Y|Z = 1, D = 1)dy = \Pr(D = 1|Z = 1) \cdot E(Y|Z = 1, D = 1). \end{aligned}$$

In a similar way it can be shown that

$$\begin{aligned} \int_{\mathcal{Y}} y \cdot q_1(y)dy &= \Pr(D = 1|Z = 0) \cdot E(Y|Z = 0, D = 1), \\ \int_{\mathcal{Y}} y \cdot q_0(y)dy &= \Pr(D = 0|Z = 0) \cdot E(Y|Z = 0, D = 0), \\ \int_{\mathcal{Y}} y \cdot p_0(y)dy &= \Pr(D = 0|Z = 1) \cdot E(Y|Z = 1, D = 0). \end{aligned}$$

Therefore,

$$\begin{aligned} E(Y(1) - Y(0)|T = c, d) &= \frac{\Pr(D = 1|Z = 1) \cdot E(Y|Z = 1, D = 1)}{\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)} + \frac{\Pr(D = 0|Z = 1) \cdot E(Y|Z = 1, D = 0)}{\Pr(D = 0|Z = 0) - \Pr(D = 0|Z = 1)} \\ &- \frac{\Pr(D = 1|Z = 0) \cdot E(Y|Z = 0, D = 1)}{\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)} - \frac{\Pr(D = 0|Z = 0) \cdot E(Y|Z = 0, D = 0)}{\Pr(D = 0|Z = 0) - \Pr(D = 0|Z = 1)} \\ &= \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)}, \end{aligned}$$

where we have made use of the fact that $\Pr(D = 0|Z = 0) - \Pr(D = 0|Z = 1) = 1 - \Pr(D = 1|Z = 0) - 1 + \Pr(D = 1|Z = 1) = \Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)$. It is easy to see that also (24) gives the same result:

$$\begin{aligned} E(Y(1) - Y(0)|T = c) &= \frac{\int_{\mathcal{Y}} y \cdot (p_1(y) - q_1(y))dy}{\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)} \\ &- \frac{\int_{\mathcal{Y}} y \cdot (q_0(y) - p_0(y))dy}{\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)}, \\ &= E(Y(1) - Y(0)|T = c, d). \end{aligned}$$

Finally, since the denominator of (25) is zero, this parameter is not defined.

5. The proof of this point is symmetric to the one of point 4 and is therefore omitted.

■

B Proof of the asymptotic properties of the estimators

In what follows we derive the asymptotic properties of $\hat{\mu}_c$, the estimator of $E(Y(1) - Y(0)|T = c)$, while those of $\hat{\mu}_{c,d}$ and $\hat{\mu}_d$ can be obtained in an equivalent way. Moreover, to simplify the notation and reduce the number of parameters in the subsequent discussion we use the fact that $\Pr(D = 0|Z = 0) - \lambda_0 = \Pr(D = 1|Z = 1) - \lambda_1$, implying that the identification result for the LATE on compliers can be expressed solely based on $\Pr(D = 1|Z = 1) - \lambda_1$:

$$E(Y(1) - Y(0)|T = c) = \frac{\int_{\mathcal{Y}} y \cdot (p_1(y) - \min(p_1(y), q_1(y))) dy}{\Pr(D = 1|Z = 1) - \lambda_1} - \frac{\int_{\mathcal{Y}} y \cdot (q_0(y) - \min(p_0(y), q_0(y))) dy}{\Pr(D = 1|Z = 1) - \lambda_1}.$$

We therefore consider the estimation of the latter expression, which is asymptotically equivalent to that proposed in Proposition 2. The proof is based on the fact that $\hat{\mu}_c$ and $\hat{\lambda}_d$ are the unique solutions of a two step semiparametric GMM optimization problem and belong to the class of MINPIN estimators as defined in Andrews (1994a). Consistency and asymptotic normality are shown by applying Theorem A-1 and 2 therein. We start by introducing some notation. Define $W = (Y, D, Z)$ to be the joint distribution of the variables. Denote by Θ the finite dimensional parameter set (we assume $\Theta \subset \mathbb{R}^4$) and by $\mathcal{T}$ the infinite dimensional parameter set of the first step. We assume $\mathcal{T}$ to be a pseudo-metric space with pseudo-metric ρ . The true values of the unknown parameters θ and τ are denoted by θ_0 and τ_0 , respectively. In estimation problem, the finite dimensional parameter vector is given by

$$\theta = \begin{pmatrix} \mu_c \\ \lambda_1 \\ P_{1|1} \\ P_{1|0} \end{pmatrix} = \begin{pmatrix} E(Y(1) - Y(0)|T = c) \\ \lambda_1 \\ \Pr(D = 1|Z = 1) \\ \Pr(D = 1|Z = 0) \end{pmatrix},$$

and the infinite dimensional parameter vector is

$$\tau(W_i) = \begin{pmatrix} \tau_1(W_i) \\ \tau_2(W_i) \\ \tau_3(W_i) \\ \tau_4(W_i) \end{pmatrix} = \begin{pmatrix} f(Y_i|D = 1, Z = 1) \\ f(Y_i|D = 1, Z = 0) \\ f(Y_i|D = 0, Z = 1) \\ f(Y_i|D = 0, Z = 0) \end{pmatrix}.$$

Let $\bar{m}_n(\theta, \tau) = \frac{\sum m(W_i, \theta, \tau(W_i))}{n}$ be a non-random measurable vector-valued function $\Theta \times \mathcal{T} \mapsto \mathbb{R}^4$, $\Theta \subset \mathbb{R}^4$, where

$$m(W_i, \theta, \tau(W_i)) = \begin{pmatrix} n \cdot \gamma_c(W_i) - \mu_c \cdot (P_{1|1} - \lambda_1) \\ n \cdot \gamma_{\lambda_1}(W_i) - \lambda_1 \\ (D_i - P_{1|1}) \cdot Z_i \\ (D_i - P_{1|0}) \cdot (1 - Z_i) \end{pmatrix},$$

with

$$\begin{aligned} \gamma_c(W_i) &= [Y_i \cdot I\{P_{1|1} \cdot \tau_1(W_i) \geq P_{1|0} \cdot \tau_2(Y_i)\} \cdot (P_{1|1} \cdot \tau_1(W_i) - P_{1|0} \cdot \tau_2(W_i))] \\ &- [Y_i \cdot I\{(1 - P_{1|0}) \cdot \tau_4(W_i) \geq (1 - P_{1|1}) \cdot \tau_3(W_i)\} \cdot ((1 - P_{1|0}) \cdot \tau_4(W_i) - (1 - P_{1|1}) \cdot \tau_3(W_i))], \end{aligned}$$

and

$$\gamma_{\lambda_1}(W_i) = I\{P_{1|0} \cdot \tau_2(W_i) \geq P_{1|1} \cdot \tau_1(Y_i)\} \cdot P_{1|1} \cdot \tau_1(W_i) + I\{P_{1|1} \cdot \tau_1(W_i) \geq P_{1|0} \cdot \tau_2(Y_i)\} \cdot P_{1|0} \cdot \tau_2(W_i).$$

Note that the first moment condition is the difference between n times the sample counterpart of the numerator of μ_c and its population equivalent ($\sum_{i=1}^n \mu_c \cdot (P_{1|1} - \lambda_1) = n \cdot \mu_c \cdot (P_{1|1} - \lambda_1)$).

Given a preliminary estimator $\hat{\tau}$, the estimator $\hat{\theta}$ solves the minimization problem

$$\min_{\theta \in \Theta} \bar{m}_n(\theta, \hat{\tau})' \bar{m}_n(\theta, \hat{\tau}).$$

To show that our estimator belongs to the class of MINPIN estimators, first consider the definition of a MINPIN estimator given in Andrews (1994a):

Definition 1. A sequence of MINPIN estimators $\{\hat{\theta}\}$ is any sequence of random variables that satisfies

$$d(\bar{m}_n(\hat{\theta}, \hat{\tau}), \hat{\kappa}) = \inf_{\theta \in \Theta} d(\bar{m}_n(\theta, \hat{\tau}), \hat{\kappa}) \quad w.p. \rightarrow 1,$$

where $\hat{\kappa}$ is similarly to τ a preliminary and possibly infinite dimensional estimator. Usually either $d(\cdot, \kappa) = \bar{m}' \kappa \bar{m} / 2$, where κ are weights, or κ does not exist as in our just identified case. Therefore, $\hat{\theta}$ is a MINPIN estimator with $d(m, \kappa) = \bar{m}' \bar{m} / 2$.

If we choose $\hat{\tau}$ such that Assumptions C and N of Andrews (1994a) are satisfied we can apply Theorem A-1 and Theorem 2 therein to show consistency and asymptotic normality of $\hat{\theta}$. For example, if the support of Y is not

bounded, one might want to estimate τ by kernel density estimation:

$$\hat{\tau}(W_i) = \left(\begin{array}{c} \frac{1}{l_1 \cdot n \cdot \sum_{i=1}^n \hat{\tau}_1(W_i)} \cdot \sum_{j=1}^n D_i \cdot Z_i \cdot K\left(\frac{Y_i - Y_j}{l_1}\right) \\ \frac{1}{l_2 \cdot n \cdot \sum_{i=1}^n \hat{\tau}_2(W_i)} \cdot \sum_{j=1}^n D_i \cdot (1 - Z_i) \cdot K\left(\frac{Y_i - Y_j}{l_2}\right) \\ \frac{1}{l_3 \cdot n \cdot \sum_{i=1}^n \hat{\tau}_3(W_i)} \cdot \sum_{j=1}^n (1 - D_i) \cdot Z_i \cdot K\left(\frac{Y_i - Y_j}{l_3}\right) \\ \frac{1}{l_4 \cdot n \cdot \sum_{i=1}^n \hat{\tau}_4(W_i)} \cdot \sum_{j=1}^n (1 - D_i) \cdot (1 - Z_i) \cdot K\left(\frac{Y_i - Y_j}{l_4}\right) \end{array} \right),$$

where $K(\cdot)$ is the kernel function (e.e., the Gaussian kernel) and l_1, l_2, l_3 and l_4 , are bandwidth parameters that are assumed to be optimally chosen. If the outcome is of bounded support, one may use boundary kernels, local linear density estimation, or adaptive bandwidth methods to overcome the poor properties of standard (local constant) kernel density estimation at the boundaries of the support of Y , see the discussion in Section 4).

We introduce some further notation required in our Assumption E below, which adapts Assumptions C and N of Andrews (1994a) to our framework. Let Θ_0 be a subset of Θ that contains a neighborhood around θ_0 and define

$$\begin{aligned} v_n(\tau) &= \sqrt{n} \cdot (\bar{m}_n(\theta_0, \tau) - E(\bar{m}_n(\theta_0, \tau))), \\ H &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} E(m_n(\theta_0, \tau_0)), \\ S &= \lim_{n \rightarrow \infty} Var(\sqrt{n} \cdot \bar{m}_n(\theta_0, \tau_0)). \end{aligned}$$

Assumption E:

1. $W_1 = (Y_1, D_1, Z_1), \dots, W_n = (Y_n, D_n, Z_n)$ is an i.i.d. sample from the joint distribution of (Y, D, Z) .
2. Θ is bounded, θ_0 lies in an interior of Θ and $E|Y|^{2+\eta} < \infty$ for some integer $\eta \geq 0$.
3. $E(m_n(\theta, \tau))$ is continuously differentiable in θ on Θ_0 and $\frac{\partial}{\partial \theta} E(m_n(\theta, \tau))$ satisfy weak law of large numbers over $\Theta \times \mathcal{T}$.
4. $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} E(m_n(\theta, \tau))$ and $\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n E(\bar{m}_n(\theta, \tau))}{n}$ exist uniformly over $(\theta, \tau) \in \Theta_0 \times \mathcal{T}$ and $\Theta \times \mathcal{T}$, respectively. The matrices S and H exist.
5. H is non singular and $\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n E(\bar{m}_n(\theta, \tau))}{n}$ is bounded away from zero for all θ outside any given neighborhood of θ_0 .
6. $f(Y|D=d, Z=z)$ is absolutely continuous with respect to Lebesgue measure for $d, z = 1, 0$.
7. $\Pr(\hat{\tau} \in \mathcal{T}) \rightarrow 1$ and $\hat{\tau} \xrightarrow{P} \tau$.
8. $\sqrt{n} \cdot E(m_n(\theta_0, \hat{\tau})) \xrightarrow{P} 0$.
9. $v_n(\tau_0) \xrightarrow{d} \mathcal{N}(0, S)$.
10. $v_n(\cdot)$ is stochastically equicontinuous at τ_0 .

Assumption E(1) can be relaxed to allow for some time dependence structure in the data. The first part of Assumption E(2) is standard and ensures that a sequence of consistent estimators of θ exists. The second part of Assumption E(2) is required to obtain uniform convergence of the first step estimators and to apply the weak law of large numbers and the central limit theorem. Assumptions E(3) to E(5) hold naturally under Assumptions E(1) and E(2) for H and S given below. Assumption E(6) is required for the first step estimation. Assumption E(7) is crucial and imposes uniform convergence of $\hat{\tau}$. When Y is bounded, uniform convergence can be obtained by using boundary kernels estimators (see Section 4). Assumption E(9) is satisfied by applying a standard central limit theorem. Assumption E(10) is a smoothness condition on the empirical process $v_n(\cdot)$ and is satisfied under Assumption E(1), E(2), and E(7) and the weak law of large numbers (see Andrews, 1994a and 1994b). To see this, consider the following pseudo-metric $\rho(\tau, \tau^*) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E \|\bar{m}_n(\theta_0, \tau) - \bar{m}_n(\theta_0, \tau^*)\|$. Under E(7), $\rho(\hat{\tau}, \tau_0) \xrightarrow{P} 0$ and under E(1) and the second part of E(2), $v_n(\hat{\tau}) - v_n(\tau_0) \xrightarrow{P} 0$ by the weak law of large numbers, which is the definition of stochastic equicontinuity given in Andrews (1994a). Finally, Assumption E(8) is an asymptotic orthogonality condition of θ and τ , which ensures that the estimation of τ does not affect the asymptotic variance of θ . As discussed in Andrews (1994a), E(8) holds under stochastic equicontinuity if $\sup_{y \in \mathcal{Y}} |\hat{\tau}(y) - \tau_0(y)| = o_p(n^{-\frac{1}{4}})$. Since the densities in τ are univariate, this rate of convergence can be easily obtained. Otherwise, Assumption E(8) could be replaced by $\sqrt{n} \cdot E(m_n(\theta_0, \hat{\tau})) \xrightarrow{d} \mathcal{N}(0, A)$ and in that case $\sqrt{n} \cdot (\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1}(S + A)(H^{-1})')$. I.e., asymptotic normality would still hold but the variance of $\hat{\theta}$ would be affected by the first step density estimation.

By applying Theorem A-1 and Theorem 2 of Andrews (1994a) under Assumption E, we have that

$$\hat{\theta} \xrightarrow{P} \theta_0 \quad \text{and} \quad \sqrt{n} \cdot (\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1}S(H^{-1})').$$

■

Let

$$\begin{aligned} \tilde{\mu}_c &= \mu_c \cdot \Pr(T = c) \\ \gamma_c(y) &= y \cdot (p_1(y) - \min(p_1(y), q_1(y))) - y \cdot (p_0(y) - \min(p_0(y), q_0(y))), \\ \gamma_{\lambda_1}(y) &= \min(p_1(y), q_1(y)), \\ h_{\mu_c, P_{1|1}} &= \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot p_1(y) dy - \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy - \mu_c, \\ h_{\mu_c, P_{1|0}} &= \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy - \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot p_1(y) dy - \mu_c. \end{aligned}$$

Then, H and S are given by

$$H = \begin{pmatrix} -\Pr(T=c) & \mu_c & h_{\mu_c, P_{1|1}} & h_{\mu_c, P_{1|0}} \\ 0 & -1 & \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy & \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \\ 0 & 0 & -E(Z) & 0 \\ 0 & 0 & 0 & -E(1-Z) \end{pmatrix},$$

and

$$S = \begin{pmatrix} S_1 & S_2 \\ S'_2 & S_3 \end{pmatrix},$$

where

$$S_1 = \begin{pmatrix} \int_{\mathcal{Y}} (n \cdot \gamma_c(y) - \tilde{\mu}_c)^2 dy & \int_{\mathcal{Y}} (n \cdot \gamma_{\lambda_1}(y) - \lambda_1) \cdot (n \cdot \gamma_{\lambda_1}(y) - \lambda_1) dy \\ \int_{\mathcal{Y}} (n \cdot \gamma_c(y) - \tilde{\mu}_c) \cdot (n \cdot \gamma_{\lambda_1}(y) - \lambda_1) dy & \int_{\mathcal{Y}} (n \cdot \gamma_{\lambda_1}(y) - \lambda_1)^2 dy \end{pmatrix},$$

$$S_2 = \begin{pmatrix} Cov((n \cdot \gamma_c(y) - \tilde{\mu}_c), (D - \Pr(D=1|Z=1)) \cdot Z) & Cov((n \cdot \gamma_c(y) - \tilde{\mu}_c), (D - \Pr(D=1|Z=0)) \cdot (1-Z)) \\ Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=1)) \cdot Z) & Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=0)) \cdot (1-Z)) \end{pmatrix},$$

$$S_3 = \begin{pmatrix} \Pr(D=1|Z=1) \cdot \Pr(D=0|Z=1) \cdot E(Z) & 0 \\ 0 & \Pr(D=1|Z=0) \cdot \Pr(D=0|Z=0) \cdot E(1-Z) \end{pmatrix}.$$

Finally, note that even though one can easily obtain consistent estimators of $\Omega = H^{-1}S(H^{-1})'$ (and therefore of the variances provided in the next section) by taking sample counterparts, it might be preferable to use the bootstrap instead. As it has already been pointed out in Section 4, Chen, Linton, and Keilegom (2003) provide conditions under which the variances of particular two step semiparametric estimators, including MINPIN estimators with i.i.d. observations, can be consistently estimated by bootstrapping, see Theorem B therein. It is easy to see that this is the case under a minor modification of our Assumption E. In particular one needs to replace “in probability” with “almost surely” in E(7) and use the strong rather than the weak law of large numbers. Moreover, E(6) to E(10) must hold in each bootstrap sample. Under these assumptions, $V_{c,d}$, V_c , V_d , and V_{λ_1} can be consistently estimated using the bootstrap.

C Variance formulae

Direct computation of the first and second elements of Ω gives

$$\begin{aligned}
V_c = & \frac{\int_{\mathcal{Y}} (n \cdot \gamma_c(y) - \tilde{\mu}_c)^2 dy + 2 \cdot \mu_c \cdot \int_{\mathcal{Y}} (n \cdot \gamma_c(y) - \tilde{\mu}_c)(n \cdot \gamma_{\lambda_1}(y) - \lambda_1) dy + \mu_c^2 \cdot \int_{\mathcal{Y}} (n \cdot \gamma_{\lambda_1}(y) - \lambda_1)^2 dy}{(\Pr(T=c))^2} \\
& + \frac{2 \cdot h_{\mu_c, P_{1|1}} \cdot Cov((n \cdot \gamma_c(y) - \tilde{\mu}_c), (D - \Pr(D=1|Z=1)) \cdot Z)}{(\Pr(T=c))^2 \cdot E(Z)} \\
& + \frac{2 \cdot \mu_c \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot Cov((n \cdot \gamma_c(y) - \tilde{\mu}_c), (D - \Pr(D=1|Z=1)) \cdot Z)}{(\Pr(T=c))^2 \cdot E(Z)} \\
& + \frac{2 \cdot \mu_c \cdot h_{\mu_c, P_{1|1}} \cdot Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=1)) \cdot Z)}{(\Pr(T=c))^2 \cdot E(Z)} \\
& - \frac{2 \cdot \mu_c^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=1)) \cdot Z)}{(\Pr(T=c))^2 \cdot E(Z)} \\
& + \frac{2 \cdot \mu_c \cdot h_{\mu_c, P_{1|1}} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot \Pr(D=1|Z=1) \cdot \Pr(D=0|Z=1) \cdot E(Z)}{(\Pr(T=c))^2 \cdot E(Z)} \\
& - \frac{2 \cdot \mu_c^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy^2 \cdot \Pr(D=1|Z=1) \cdot \Pr(D=0|Z=1) \cdot E(Z)}{(\Pr(T=c))^2 \cdot E(Z)} \\
& + \frac{h_{\mu_c, P_{1|1}}^2 \cdot \Pr(D=1|Z=1) \cdot \Pr(D=0|Z=1) \cdot E(Z)}{(\Pr(T=c))^2 \cdot E(Z)^2} \\
& + \frac{2 \cdot h_{\mu_c, P_{1|0}} \cdot Cov((n \cdot \gamma_c(y) - \tilde{\mu}_c), (D - \Pr(D=1|Z=0)) \cdot (1-Z))}{(\Pr(T=c))^2 \cdot E(1-Z)} \\
& + \frac{2 \cdot \mu_c \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot Cov((n \cdot \gamma_c(y) - \tilde{\mu}_c), (D - \Pr(D=1|Z=0)) \cdot (1-Z))}{(\Pr(T=c))^2 \cdot E(1-Z)} \\
& + \frac{2 \cdot \mu_c \cdot h_{\mu_c, P_{1|0}} \cdot Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=0)) \cdot (1-Z))}{(\Pr(T=c))^2 \cdot E(1-Z)} \\
& - \frac{2 \cdot \mu_c^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=0)) \cdot (1-Z))}{(\Pr(T=c))^2 \cdot E(1-Z)} \\
& + \frac{2 \cdot \mu_c \cdot h_{\mu_c, P_{1|0}} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot \Pr(D=1|Z=0) \cdot \Pr(D=0|Z=0) \cdot E(1-Z)}{(\Pr(T=c))^2 \cdot E(1-Z)} \\
& - \frac{2 \cdot \mu_c^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy^2 \cdot \Pr(D=1|Z=0) \cdot \Pr(D=0|Z=0) \cdot E(1-Z)}{(\Pr(T=c))^2 \cdot E(1-Z)} \\
& + \frac{h_{\mu_c, P_{1|0}}^2 \cdot \Pr(D=1|Z=0) \cdot \Pr(D=0|Z=0) \cdot E(1-Z)}{(\Pr(T=c))^2 \cdot E(1-Z)^2},
\end{aligned}$$

and

$$\begin{aligned}
V_{\lambda_1} = & \int_{\mathcal{Y}} (n \cdot \gamma_{\lambda_1}(y) - \lambda_1)^2 dy \\
& + \frac{E(D|Z=1, p_1(y) \leq q_1(y)) \cdot \Pr(D=1|Z=1) \cdot \Pr(D=0|Z=1) \cdot E(Z)}{E(Z)} \\
& - \frac{2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=1)) \cdot Z)}{E(Z)} \\
& + \frac{\int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot \Pr(D=1|Z=0) \cdot \Pr(D=0|Z=0) \cdot E(1-Z)}{E(1-Z)} \\
& - \frac{2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=0)) \cdot (1-Z))}{E(1-Z)}.
\end{aligned}$$

By replacing $m(W_i, \theta, \tau(W_i))$ with

$$m_d(W_i, \theta, \tau(W_i)) = \begin{pmatrix} n \cdot \gamma_c(W_i) - \mu_d \cdot (P_{1|1} - \lambda_1) \\ n \cdot \gamma_{\lambda_1}(W_i) - \lambda_1 \\ (D_i - P_{1|1}) \cdot Z_i \\ (D_i - P_{1|0}) \cdot (1 - Z_i) \end{pmatrix},$$

it can be shown that

$$\begin{aligned} V_d &= \frac{\int_{\mathcal{Y}} (n \cdot \gamma_d(y) - \tilde{\mu}_d)^2 dy + 2 \cdot \mu_d \cdot \int_{\mathcal{Y}} (n \cdot \gamma_d(y) - \tilde{\mu}_d)(n \cdot \gamma_{\lambda_1}(y) - \lambda_1) dy + \mu_d^2 \cdot \int_{\mathcal{Y}} (n \cdot \gamma_{\lambda_1}(y) - \lambda_1)^2 dy}{(\Pr(T = d))^2} \\ &+ \frac{2 \cdot h_{\mu_d, P_{1|1}} \cdot Cov((n \cdot \gamma_d(y) - \tilde{\mu}_d), (D - \Pr(D = 1|Z = 1)) \cdot Z)}{(\Pr(T = d))^2 \cdot E(Z)} \\ &+ \frac{2 \cdot \mu_d \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot Cov((n \cdot \gamma_d(y) - \tilde{\mu}_d), (D - \Pr(D = 1|Z = 1)) \cdot Z)}{(\Pr(T = d))^2 \cdot E(Z)} \\ &+ \frac{2 \cdot \mu_d \cdot h_{\mu_d, P_{1|1}} \cdot Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 1)) \cdot Z)}{(\Pr(T = d))^2 \cdot E(Z)} \\ &- \frac{2 \cdot \mu_d^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 1)) \cdot Z)}{(\Pr(T = d))^2 \cdot E(Z)} \\ &+ \frac{2 \cdot \mu_d \cdot h_{\mu_d, P_{1|1}} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = d))^2 \cdot E(Z)} \\ &- \frac{2 \cdot \mu_d^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy^2 \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = d))^2 \cdot E(Z)} \\ &+ \frac{h_{\mu_d, P_{1|1}}^2 \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = d))^2 \cdot E(Z)^2} \\ &+ \frac{2 \cdot h_{\mu_d, P_{1|0}} \cdot Cov((n \cdot \gamma_d(y) - \tilde{\mu}_d), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z))}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\ &+ \frac{2 \cdot \mu_d \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot Cov((n \cdot \gamma_d(y) - \tilde{\mu}_d), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z))}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\ &+ \frac{2 \cdot \mu_d \cdot h_{\mu_d, P_{1|0}} \cdot Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z))}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\ &- \frac{2 \cdot \mu_d^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot Cov((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z))}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\ &+ \frac{2 \cdot \mu_d \cdot h_{\mu_d, P_{1|0}} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot \Pr(D = 1|Z = 0) \cdot \Pr(D = 0|Z = 0) \cdot E(1 - Z)}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\ &- \frac{2 \cdot \mu_d^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy^2 \cdot \Pr(D = 1|Z = 0) \cdot \Pr(D = 0|Z = 0) \cdot E(1 - Z)}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\ &+ \frac{h_{\mu_d, P_{1|0}}^2 \cdot \Pr(D = 1|Z = 0) \cdot \Pr(D = 0|Z = 0) \cdot E(1 - Z)}{(\Pr(T = d))^2 \cdot E(1 - Z)^2}. \end{aligned}$$

Similarly by replacing $m(W_i, \theta, \tau(W_i))$ with

$$m_{c,d}(W_i, \theta, \tau(W_i)) = \begin{pmatrix} n \cdot \gamma_c(W_i) - \mu_{c,d} \cdot (P_{1|1} - \lambda_1) \\ n \cdot \gamma_{\lambda_1}(W_i) - \lambda_1 \\ (D_i - P_{1|1}) \cdot Z_i \\ (D_i - P_{1|0}) \cdot (1 - Z_i) \end{pmatrix},$$

it can be shown that

$$\begin{aligned} V_{c,d} &= \frac{\int_{\mathcal{Y}} (n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d})^2 dy + 2 \cdot \mu_{c,d} \cdot \int_{\mathcal{Y}} (n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d})(n \cdot \gamma_{\lambda_1}(y) - \lambda_1) dy + \mu_{c,d}^2 \cdot \int_{\mathcal{Y}} (n \cdot \gamma_{\lambda_1}(y) - \lambda_1)^2 dy}{(\Pr(T=c) + \Pr(T=d))^2} \\ &+ \frac{2 \cdot h_{\mu_{c,d}, P_{1|1}} \cdot \text{Cov}((n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d}), (D - \Pr(D=1|Z=1)) \cdot Z)}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(Z)} \\ &+ \frac{2 \cdot \mu_{c,d} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot \text{Cov}((n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d}), (D - \Pr(D=1|Z=1)) \cdot Z)}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(Z)} \\ &+ \frac{2 \cdot \mu_{c,d} \cdot h_{\mu_{c,d}, P_{1|1}} \cdot \text{Cov}((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=1)) \cdot Z)}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(Z)} \\ &- \frac{2 \cdot \mu_{c,d}^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot \text{Cov}((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=1)) \cdot Z)}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(Z)} \\ &+ \frac{2 \cdot \mu_{c,d} \cdot h_{\mu_{c,d}, P_{1|1}} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot \Pr(D=1|Z=1) \cdot \Pr(D=0|Z=1) \cdot E(Z)}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(Z)} \\ &- \frac{2 \cdot \mu_{c,d}^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy^2 \cdot \Pr(D=1|Z=1) \cdot \Pr(D=0|Z=1) \cdot E(Z)}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(Z)} \\ &+ \frac{h_{\mu_{c,d}, P_{1|1}}^2 \cdot \Pr(D=1|Z=1) \cdot \Pr(D=0|Z=1) \cdot E(Z)}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(Z)^2} \\ &+ \frac{2 \cdot h_{\mu_{c,d}, P_{1|0}} \cdot \text{Cov}((n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d}), (D - \Pr(D=1|Z=0)) \cdot (1-Z))}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(1-Z)} \\ &+ \frac{2 \cdot \mu_{c,d} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot \text{Cov}((n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d}), (D - \Pr(D=1|Z=0)) \cdot (1-Z))}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(1-Z)} \\ &+ \frac{2 \cdot \mu_{c,d} \cdot h_{\mu_{c,d}, P_{1|0}} \cdot \text{Cov}((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=0)) \cdot (1-Z))}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(1-Z)} \\ &- \frac{2 \cdot \mu_{c,d}^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot \text{Cov}((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=0)) \cdot (1-Z))}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(1-Z)} \\ &+ \frac{2 \cdot \mu_{c,d} \cdot h_{\mu_{c,d}, P_{1|0}} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot \Pr(D=1|Z=0) \cdot \Pr(D=0|Z=0) \cdot E(1-Z)}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(1-Z)} \\ &- \frac{2 \cdot \mu_{c,d}^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy^2 \cdot \Pr(D=1|Z=0) \cdot \Pr(D=0|Z=0) \cdot E(1-Z)}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(1-Z)} \\ &+ \frac{h_{\mu_{c,d}, P_{1|0}}^2 \cdot \Pr(D=1|Z=0) \cdot \Pr(D=0|Z=0) \cdot E(1-Z)}{(\Pr(T=c) + \Pr(T=d))^2 \cdot E(1-Z)^2}, \end{aligned}$$

where

$$\tilde{\mu}_{c,d} = \mu_{c,d} \cdot (\Pr(T = c) + \Pr(T = d)),$$

$$\tilde{\mu}_d = \mu_d \cdot \Pr(T = d),$$

$$\gamma_{c,d}(y) = y \cdot (\max(p_1(y), q_1(y)) - \min(p_1(y), q_1(y))) - y \cdot (\max(p_0(y), q_0(y)) - \min(p_0(y), q_0(y))),$$

$$\gamma_d(y) = y \cdot (q_1(y) - \min(p_1(y), q_1(y))) - y \cdot (q_0(y) - \min(p_0(y), q_0(y))),$$

$$\begin{aligned} h_{\mu_{c,d}, P_1|1} &= \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot p_1(y) dy - \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \\ &+ \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot q_1(y) dy - \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy - \mu_{c,d}, \end{aligned}$$

$$\begin{aligned} h_{\mu_{c,d}, P_1|0} &= \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy - \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot p_1(y) dy \\ &+ \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy - \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot q_1(y) dy - \mu_{c,d}, \end{aligned}$$

$$h_{\mu_d, P_1|1} = \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot q_1(y) dy - \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy - \mu_d,$$

$$h_{\mu_d, P_1|0} = \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy - \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot q_1(y) dy - \mu_d.$$

References

- ALIPRANTIS, D. (2012): “Redshirting, Compulsory Schooling Laws, and Educational Attainment,” *Journal of Educational and Behavioral Statistics*, 37, 316–338.
- ANDERSON, G., O. LINTON, AND Y.-J. WHANG (2012): “Nonparametric estimation and inference about the overlap of two distributions,” *Journal of Econometrics*, 171, 1–23.
- ANDREWS, D. W. (1991): “Asymptotics for Kernel-Based Non-Orthogonal Semiparametric Estimators,” *mimeo*.
- (1994a): “Asymptotics for semiparametric econometric models via stochastic equicontinuity,” *Econometrica*, 62(1), 43–72.
- (1994b): “Empirical process methods in econometrics,” in *Handbook of Econometrics*, ed. by R. F. Engle, and D. McFadden, pp. 2247–2294. Elsevier.
- (1995): “Nonparametric Kernel Estimation for Semiparametric Models,” *Econometric Theory*, 11, 560–586.
- ANDREWS, D. W. K. (2000): “Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space,” *Econometrica*, 68, 399–405.
- ANGRIST, J., G. IMBENS, AND D. RUBIN (1996): “Identification of Causal Effects using Instrumental Variables,” *Journal of American Statistical Association*, 91, 444–472 (with discussion).
- ANGRIST, J., AND A. KRUEGER (1991): “Does Compulsory School Attendance Affect Schooling and Earnings?,” *Quarterly Journal of Economics*, 106, 979–1014.
- (1992): “The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples,” *Journal of the American Statistical Association*, 87, 328–336.
- BARUA, R., AND K. LANG (2009): “School Entry, Educational Attainment, and Quarter of Birth: A Cautionary Tale of LATE,” *NBER Working Paper 15236*.
- BEDARD, K., AND E. DHUEY (2006): “The persistence of early childhood maturity: International evidence of long-run age effects,” *The Quarterly Journal of Economics*, 121, 1437–1472.
- BICKEL, P. J., F. GÖTZE, AND W. R. VAN ZWET (1997): “Resampling Fewer Than n Observations: Gains, Losses, and Remedies for Losses,” *Statistica Sinica*, 7, 1–31.
- BOUND, J., D. A. JAEGER, AND R. M. BAKER (1995): “Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak,” *Journal of the American Statistical Association*, 90, 443–450.
- BUCKLES, K. S., AND D. M. HUNGERMAN (2013): “Season of Birth and Later Outcomes: Old Questions, New Answers,” *forthcoming in the Review of Economics and Statistics*, 95, 711–724.
- CARD, D. (1999): “The Causal Effect of Education on Earnings,” in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1802–1863. North-Holland, Amsterdam.

- CHEN, X., O. LINTON, AND I. V. KEILEGOM (2003): “Estimation of Semiparametric Models When the Criterion Function Is Not Smooth,” *Econometrica*, 71, 1591–1608.
- CHENG, M.-Y., J. FAN, AND J. S. MARRON (1997): “On automatic boundary corrections,” *Annals of Statistics*, 25, 1691–1708.
- CHERNOZHUKOV, V., AND C. HANSEN (2005): “An IV Model of Quantile Treatment Effects,” *Econometrica*, 73, 245–261.
- DAI, J., AND S. SPERLICH (2010): “Simple and effective boundary correction for kernel densities and regression with an application to the world income and Engel curve estimation,” *Computational Statistics & Data Analysis*, 54, 2487 – 2497.
- DE CHAISEMARTIN, C. (2012): “All you need is LATE,” *mimeo, Paris School of Economics*.
- DE CHAISEMARTIN, C., AND X. D’HAULTFOEUILLE (2012): “LATE again, with defiers,” *mimeo, CREST*.
- FRÖLICH, M. (2007): “Nonparametric IV Estimation of Local Average Treatment Effects with Covariates,” *Journal of Econometrics*, 139, 35–75.
- HUBER, M., AND G. MELLACE (2010): “Sharp IV bounds on average treatment effects under endogeneity and noncompliance,” *University of St Gallen, Dept. of Economics Discussion Paper no. 2010-31*.
- IMBENS, G. W., AND J. ANGRIST (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–475.
- IMBENS, G. W., AND D. RUBIN (1997): “Estimating outcome distributions for compliers in instrumental variables models,” *Review of Economic Studies*, 64, 555–574.
- JONES, M. C. (1993): “Simple boundary correction for kernel density estimation,” *Statistics and Computing*, 3, 135–146.
- KITAGAWA, T. (2009): “Identification Region of the Potential Outcome Distribution under Instrument Independence,” *CeMMAP working paper 30/09*.
- KLEIN, T. J. (2010): “Heterogeneous treatment effects: Instrumental variables without monotonicity?,” *Journal of Econometrics*, 155, 99–116.
- LI, Q., AND J. S. RACINE (2007): *Nonparametric econometrics: theory and practice*. Princeton University Press, Princeton and Oxford.
- MACKINNON, J. G. (2006): “Bootstrap Methods in Econometrics,” *The Economic Record*, 82, S2–S18.
- MANSKI, C. F. (1990): “Nonparametric Bounds on Treatment Effects,” *American Economic Review, Papers and Proceedings*, 80, 319–323.
- NEWWEY, W. (1994): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62, 1349–1382.

- ROY, A. (1951): “Some Thoughts on the Distribution of Earnings,” *Oxford Economic Papers*, 3, 135–146.
- RUBIN, D. B. (1974): “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701.
- SILVERMAN, B. (1986): *Density estimation for statistics and data analysis*. Chapman and Hall, London.
- SMALL, D. S., AND Z. TAN (2007): “A Stochastic Monotonicity Assumption for the Instrumental Variables Method,” *Technical report, Department of Statistics, Wharton School, University of Pennsylvania*.
- VYTLACIL, E. (2002): “Independence, Monotonicity, and Latent Index Models: An Equivalence Result,” *Econometrica*, 70, 331–341.