Centre de Recherche en Économie et Statistique

CREST

Série des Documents de Travail

n° 2012-12

LATE again, without Monotonicity

C. de CHAISEMARTIN[1]

June, 2012

1 Paris School of Economics and CREST (Laboratoire de Microéconométrie)
chaisemartin@pse.ens.fr

LATE again, without monotonicity.*

Clément de Chaisemartin[†]

June 2012

Abstract

Monotonicity is not necessary for the Wald ratio to identify a Local Average Treatment Effect. Under random assignment and exclusion restriction, if for every value of potential outcomes there are more compliers than defiers, the Wald ratio identifies the average treatment effect within a subpopulation of compliers. I use a simple Roy selection model to show that this "less defiers than compliers" condition is substantially weaker than monotonicity. It has two implications which are testable from the data, and it is closely related to those testable implications. Similarly, the local monotonicity condition in Huber & Mellace (2012) is not necessary for their identification results to hold and can also be replaced by a substantially weaker condition.

Keywords: local average treatment effect, instrumental variable, monotonicity, local monotonicity, defiers.
JEL Codes: C21, C26

*I am very grateful to Alberto Abadie, Josh Angrist, Xavier D'Haultfoeuille, Martin Huber, Tobias Klein and Kevin Lang for their helpful comments. I gratefully acknowledge financial support from the région Ile de France.

†Paris School of Economics and CREST, chaisemartin@pse.ens.fr.

1 Introduction

Since the seminal work of Imbens & Angrist (1994) and Angrist et al. (1996), the use of instruments for identifying causal effects has been thought of as depending on three crucial assumptions: random assignment, exclusion restriction and monotonicity. Random assignment states that the instrument is assigned to individuals independently of their potential outcomes and treatments. Exclusion restriction implies that the instrument should not have an effect on the outcome other than through its effect on treatment. Monotonicity means that the effect of the instrument on the treatment should go in the same direction for all observations in the sample. Many other treatment effects model also rely on similar monotonicity conditions. Important examples include the fuzzy regression discontinuity model in Hahn et al. (2001), or the linear IV-quantile regression model in Abadie et al. (2002).

Monotonicity may be problematic in some applications. Barua & Lang (2010) argue that quarter of birth, used as an instrument for school entry age, violates this monotonicity condition, because of heterogenous strategic behavior of parents when choosing the entry date at school of their child. Another potential example is the use of sibling-sex composition as an IV when studying the effect of childbearing on labor supply (see Angrist & Evans, 1998). In this paper, it appears that the share of parents who have a third child is 7 percentage points higher when their first two children have the same sex than when they have different sex, which implies that some parents have a preference for diversity. But the authors also find that the share of parents who have a third child is approximately 1.5 percentage points higher when their first two children are girls than when they are boys, which implies that some parents have a preference for boys. Such parents could be more likely to have a third child if their first two children are a boy and girl, and less likely to do so if their first two children are boys, which would violate monotonicity.

Monotonicity is also problematic in randomized experiments relying on an encouragement design protocol, in which the treatment group typically receives a financial incentive to get treated, or a flyer describing the benefits from treatment, or both. Indeed, in such protocols, monotonicity means that the incentive for treatment should have a positive impact on willingness to get treated for all observations. In practice, it might be the case that those incentives have the expected positive effect on a majority of observations, but still discourage some observations to get treated. For instance, the fact they are offered to get paid to receive the treatment might cast some doubt on its actual efficacy among some observations.

The aim of this note is to show that actually, the Wald ratio still identifies a local average treatment effect (LATE) under random assignment and exclusion restriction, if monotonicity is replaced by the substantially weaker condition that conditional on potential outcomes, there should be less defiers than compliers. Indeed, in such instances, there always exists a subpopulation of compliers of same size than defiers and with the same distribution of potential outcomes. Therefore, treatment effects among this subpopulation of compliers and among defiers cancel out in the numerator of the Wald ratio, so that this ratio identifies the average treatment effect among the remaining part of compliers.

Angrist et al. (1996) show that the Wald ratio also identifies an average treatment effect if one assumes that treatment effects are the same among defiers and compliers instead of assuming monotonicity. One can show that those two assumptions are actually "polar cases" of the Less Defiers than Compliers (LDTC) condition I substitute to monotonicity. Indeed, LDTC implies that the joint distribution of potential outcomes should not to be "too" different among defiers than among compliers. The extent to which those two distributions are allowed to differ depends on the ratio of the shares of compliers and defiers. When this ratio is large, meaning that there are almost no defiers and that monotonicity almost holds, those two joint distributions can be very different. On the contrary, when this ratio is close to 1, meaning that monotonicity is very strongly violated, those two distributions should be almost the same for LDTC to hold.

I also consider a simple Roy selection model to show that conditions on the structural parameters of the model which are sufficient for LDTC to hold are substantially weaker than conditions ensuring monotonicity. Then, I show that the LDTC condition is partly testable since it implies that the difference between two densities should be non-negative, as in Kitagawa (2008). Moreover, the LDTC condition is very "close" from its testable implications. Indeed, one can test from the data that conditional on one potential outcome there are less defiers than compliers, while the LDTC condition states that there should be less defiers than compliers conditional on both potential outcomes.

Importantly, it is sufficient to slightly strengthen the LDTC condition by assuming that there are less defiers than compliers conditional on potential outcomes and covariates, to obtain that the distribution of covariates among the same subpopulation of compliers is also identified from the data. This means that the population on which treatment effects are identified can also be characterized in terms of its observable characteristics, as it is the case under monotonicity. Finally, the same approach can be used to show that the local monotonicity condition put forward Huber & Mellace (2012) to relax monotonicity is not necessary for their identification results to hold and can be replaced by a substantially weaker "local stochastic monotonicity" condition.

The issue of whether monotonicity is necessary for identifying local average treatment effects has received attention recently. Small & Tan (2007) replace monotonicity by a stochastic monotonicity assumption. Under this assumption, the standard Wald parameter does not identify a causal effect anymore, but satisfies the no sign reversal property (namely, its sign would be positive if the treatment effect is positive with probability one). Klein (2010) considers “local” violations of monotonicity, and shows that the bias of the Wald parameter can be well approximated if such violations are small. As mentioned above, Huber & Mellace (2012) show that it is possible to identify average treatment effects on compliers, defiers or both under a local monotonicity condition. Finally, Chaisemartin & D'Haultfoeuille (2012) show that monotonicity is equivalent to a rank invariance assumption, and can be replaced by a weaker rank similarity assumption. Here, I show that it is possible to relax monotonicity substantially while keeping a causal interpretation of the Wald parameter.

2 The "Less Defiers than Compliers" (LDTC) assumption

For any random variables S and T let f_S and $f_{S|T}$ denote the density of S and the density of S conditional on T. Throughout the paper, I adopt the following convention to alleviate the notational burden: for any random variable S and any event A, let $f_{S,A}(s) = f_{S|A}(s)P(A)$, with the convention that $f_{S,A}(s) = 0$ if $P(A) = 0$. Let $\mathcal{S}(T)$ denote the support of T.

My framework is the same as the one of Angrist et al. (1996). Let Z be a binary instrument. Let $D(z)$ denote the potential treatment when $Z = z$. Let $Y(d, z)$ denote potential outcomes as functions of treatment and of the instrument. We only observe $D = D(Z)$ and $Y = Y(D, Z)$. Following Imbens & Angrist (1994), never takers (NT) are observations such that $D(0) = 0$ and $D(1) = 0$, always takers (AT) are observations such that $D(0) = 1$ and $D(1) = 1$, compliers (C) are observations such that $D(0) = 0$ and $D(1) = 1$, and defiers ($\overline{D}$) are observations such that $D(0) = 1$ and $D(1) = 0$. Let $\mathcal{T} = \{NT, AT, C, \overline{D}\}$ be the set of all possible types. Without loss of generality, assume that $P(D = 1|Z = 1) > P(D = 1|Z = 0)$ (if it is not the case, one can merely switch the words "defiers" and "compliers" in what follows).

The first assumption in Angrist et al. (1996) is that the instrument is exogenous: it is assumed to be independent of potential treatments and potential outcomes.

Assumption 2.1 *(Instrument exogeneity)*

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

They also assume that the instrument has an impact on the outcome only through its impact on treatment.

Assumption 2.2 *(Exclusion restriction)*
$Y(d,0) = Y(d,1) = Y(d)$ *for* $d \in \{0,1\}$.

Last, they suppose that the instrument has a monotonous effect on Z.

Assumption 2.3 *(Instrument monotonicity)*
Almost surely, $D(1) \geq D(0)$.

In what follows, Assumptions 2.1 and 2.2 are maintained, while Assumption 2.3 is replaced by the following Assumption:

Assumption 2.4 *(Less Defiers than Compliers: LDTC)*

$$f_{Y(0),Y(1),\overline{D}}(y_0, y_1) \leq f_{Y(0),Y(1),C}(y_0, y_1), \tag{2.1}$$

or equivalently:

$$P(\overline{D}|Y(0), Y(1)) \leq P(C|Y(0), Y(1)), \tag{2.2}$$

almost surely.

When monotonicity is verified, the left hand side of Equation (2.1) is equal to 0, so that Assumption 2.4 is automatically verified. Therefore, Assumption 2.4 is strictly weaker than monotonicity.

Angrist et al. (1996) show that the Wald ratio identifies an average treatment effect if there are no defiers, or if treatment effects are the same among defiers and compliers. Since $P(D = 1|Z = 1) > P(D = 1|Z = 0)$ by assumption, one can show that Assumption 2.1 implies $0 \leq P(\overline{D}) < P(C)$. Therefore, when $P(\overline{D})$ is different from 0, Equation (2.1) rewrites as follows:

$$\frac{f_{Y(0),Y(1)|\overline{D}}(y_0, y_1)}{f_{Y(0),Y(1)|C}(y_0, y_1)} \leq \frac{P(C)}{P(\overline{D})}. \tag{2.3}$$

This makes it clear that the two assumptions pointed out in Angrist et al. (1996) are polar cases of Assumption 2.4. Indeed, when $P(\overline{D})$ is close to 0, $\frac{P(C)}{P(\overline{D})}$ is large, and treatment effects can be very different among defiers and compliers. When $P(\overline{D})$ is close from $P(C)$, since the two densities in Equation (2.3) must be positive and integrate to 1, Assumption 2.4 implies that treatment effects must be extremely similar among compliers and defiers. But Equation (2.3) also shows that in between those two polar cases, there is a continuum of cases in which Assumption 2.4 is verified because some kind of "trade-off" between those two polar assumptions takes place. Indeed, Equation (2.3) shows that the more defiers there are, the less different compliers and defiers must be for Assumption 2.4 to hold, and conversely.

Note also that the extent to which defiers and compliers can differ depends on $\frac{P(C)}{P(\overline{D})}$. This quantity is not identified from the data, but under Assumption 2.1 one can show that

$$\frac{P(D=1|Z=1)}{P(D=1|Z=0)} \leq \frac{P(C)}{P(\overline{D})}.$$

Therefore, estimating this lower bound provides a simple way of assessing how different compliers and defiers can be under Assumption 2.4.

Then, I use a simple Roy selection model to further emphasize that the LDTC condition is substantially weaker than monotonicity. To fix ideas, consider the example of a randomized experiment relying on an encouragement design protocol, in which the treatment group receives a flyer describing the advantages it can expect from treatment, and a financial incentive to get treated. To simplify, assume that $Y(1)-Y(0)$ is strictly greater than 0 with probability 1. This is a restrictive assumption which is taken only to simplify the computations. The point I am trying to make here, which is that structural assumptions which ensure that LCTD holds are substantially weaker than conditions ensuring monotonicity, is still true without this restrictive assumption. Assume that selection into treatment is determined by the following Roy Model:

$$D = 1\{\alpha\,(Y(1)-Y(0)) + Z(\beta-\alpha)\,(Y(1)-Y(0)) \geq \varepsilon - \lambda Z\}. \tag{2.4}$$

α is a random coefficient equal to the ratio between the perceived and true benefits of treatment before reading the flyer. β represents the same ratio, but after reading the flyer. λ is a strictly positive constant which represents the financial incentive given to treatment group observations that undertake the treatment. ε is a strictly positive unobserved heterogeneity term which represents the cost of treatment. $Z=1$ when an observation receives the flyer describing the treatment and the financial incentive to get treated. In this model, when $Z=0$, an observation will undertake the treatment if and only if the perceived benefits of treatment before reading the flyer are greater than the cost of getting treated. When $Z=1$, it will get treated if and only if the perceived benefits after reading the flyer are greater than the difference between the cost of getting treated and the incentive to get treated.

Therefore, the instrument has two effects on selection into treatment: it reduces the cost via the financial incentive, but it also changes the perception of the benefits of treatment via the flyer. In particular, the flyer will reduce the left-hand-side of Equation (2.4) for observations with $\alpha > \beta$, which is the reason why there might be defiers in this model, despite the fact that Z reduces the right hand side for all observations.

If $\alpha \leq \beta$ almost surely, meaning that the flyer and the financial incentive never have a negative impact on how observations evaluate the benefits from treatment, monotonicity holds. On the contrary, one can show that

$$P(\alpha \leq \beta | Y(1), Y(0)) \geq P(\alpha > \beta | Y(1), Y(0)) \tag{2.5}$$

is sufficient for Assumption 2.4 to hold. This means that for each value of $Y(1)$ and $Y(0)$ there should be less observations whose perceived benefits of treatment decrease after reading the flyer and receiving the incentive, than observations whose perceived benefits increase. Therefore, the condition on α and β which ensure that Assumption 2.4 holds is substantially weaker than the condition under which monotonicity holds.

3 Testability

Assumption LDTC has implications which are testable from the data. Consider the following assumption.

Assumption 3.1
$f_{Y,D=d|Z=1-d}(y) \leq f_{Y,D=d|Z=d}(y)$ for $d \in \{0; 1\}$

Assumption 3.1 is the same condition as the one used in Kitagawa (2008) to test jointly Assumptions 2.1, 2.2 and 2.3.

Lemma 3.1 *Suppose Assumptions 2.1 and 2.2 hold and $P(D = 1|Z = 1) > P(D = 1|Z = 0)$. Then, Assumption 2.4 $\Rightarrow$ Assumption 3.1.*

Proof:

$$\begin{aligned} f_{Y,D=d|Z=d}(y) - f_{Y,D=d|Z=1-d}(y) &= f_{Y(d),D(d)=d}(y) - f_{Y(d),D(1-d)=d}(y) \\ &= f_{Y(d),D(d)=d,D(1-d)=1-d}(y) - f_{Y(d),D(1-d)=d,D(d)=1-d}(y) \\ &= f_{Y(d),C}(y) - f_{Y(d),\overline{D}}(y). \end{aligned} \tag{3.1}$$

The first equality follows from Assumptions 2.1 and 2.2, others follow from standard arguments.

Integrating Equation (2.1) over the support of $Y(1-d)$ implies that $f_{Y(d),\overline{D}}(y) \leq f_{Y(d),C}(y)$. Combining this with Equation (3.1) shows that Assumption 2.4 implies Assumption 3.1.

QED.

Lemma 3.1 shows that under Assumptions 2.1 and 2.2, Assumption 2.4 has two implications which are directly testable from the data, for instance through Kitagawa (2008) test. Under Assumptions 2.1 and 2.2, those two testable implications rewrite as

$$\begin{aligned} P(\overline{D}|Y(0)) &\leq P(C|Y(0)) \\ P(\overline{D}|Y(1)) &\leq P(C|Y(1)). \end{aligned}$$

This shows that under Assumptions 2.1 and 2.2, it is possible to test whether conditional on one potential outcome there are more compliers than defiers, while Assumption 2.4 states that there should be more compliers than defiers conditional on both potential outcomes. This means that Assumption 2.4 is very "close" from its testable implications.

Kitagawa (2008) shows that Assumption 3.1 is the weakest testable implication of Assumptions 2.1, 2.2 and 2.3: if it holds then it is possible to construct a DGP such that Assumptions 2.1, 2.2 and 2.3 are verified. Since Assumptions 2.1, 2.2 and 2.4 are weaker than Assumptions 2.1, 2.2 and 2.3, this implies both that Assumption 3.1 is the weakest testable implication of Assumptions 2.1, 2.2 and 2.4 as well, and that Assumptions 2.1, 2.2 and 2.4 and Assumptions 2.1, 2.2 and 2.3 are observationally equivalent.

4 Identification of a LATE under LDTC.

The following Theorem shows that under Assumptions 2.1 and 2.2, if Assumption 2.4 is verified, then for every true underlying data distribution there always exists a subpopulation of compliers such that the Wald ratio identifies the average treatment effect among this subpopulation. I also obtain a result on the marginal distributions of potential outcomes within this subpopulation that generalizes the result obtained in Imbens & Rubin (1997) under monotonicity.

Theorem 4.1 *Suppose Assumptions 2.1, 2.2 and 2.4 hold and* $P(D = 1|Z = 1) > P(D = 1|Z = 0)$*. Then, there exists a subpopulation of compliers* C^* *such that*

$$P(C^*) = P(D = 1|Z = 1) - P(D = 1|Z = 0),$$

and

$$f_{Y(d),C^*}(y) = f_{Y,D=d|Z=d}(y) - f_{Y,D=d|Z=1-d}(y).$$

Consequently,

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

Proof:
First, notice that

$$\begin{aligned}
P(D=1|Z=1)-P(D=1|Z=0) &= P(D(1)=1)-P(D(0)=1) \\
&= P(D(1)=1,D(0)=0)-P(D(0)=1,D(1)=0) \\
&= P(C)-P(\overline{D}). \qquad (4.1)
\end{aligned}$$

The first equality follows from Assumption 2.1, others follow from standard arguments.

Constructing C^* merely amounts to withdraw randomly a share

$$p(y_0,y_1)=\frac{f_{Y(0),Y(1),\overline{D}}(y_0,y_1)}{f_{Y(0),Y(1),C}(y_0,y_1)}$$

of compliers such that $Y(0)=y_0$ and $Y(1)=y_1$ for each (y_0,y_1) in $\mathcal{S}(Y(0)\times Y(1))$. Indeed, Assumption 2.4 ensures that $p(y_0,y_1)$ is smaller than 1.

Formally, let $(B(y_0,y_1))_{y_0,y_1}$ denote a family of independent Bernoulli variables defined over $\mathcal{S}(Y(0)\times Y(1))$, independent of $(Y(0),Y(1))$ and $(D(0),D(1))$, and such that $P(B(y_0,y_1)=1)=p(y_0,y_1)$. Let C^{**} denote the event $C\cap\{B\,(Y(0),Y(1))=1\}$ and $C^*=C\setminus C^{**}$. Note that in general, C^* is not uniquely defined, unless $p(y_0,y_1)=0$ or 1 for every (y_0,y_1).

$$\begin{aligned}
f_{Y(0),Y(1),C^{**}}(y_0,y_1) &= f_{Y(0),Y(1),C,B(Y(0),Y(1))=1}(y_0,y_1) \\
&= f_{Y(0),Y(1),C,B(y_0,y_1)=1}(y_0,y_1) \\
&= f_{Y(0),Y(1),C}(y_0,y_1)P(B(y_0,y_1)=1) \\
&= f_{Y(0),Y(1),C}(y_0,y_1)p(y_0,y_1) \\
&= f_{Y(0),Y(1),\overline{D}}(y_0,y_1).
\end{aligned}$$

The first equality arises from the definition of C^{**}, the second one is due to the fact that the event $Y(0)=y_0,Y(1)=y_1,B(Y(0),Y(1))=1$ is equal to $Y(0)=y_0,Y(1)=y_1,B(y_0,y_1)=1$. The third equality follows from independence.

Consequently, since $C^{**}\subseteq C$,

$$\begin{aligned}
f_{Y(0),Y(1),C^*}(y_0,y_1) &= f_{Y(0),Y(1),C}(y_0,y_1)-f_{Y(0),Y(1),C^{**}}(y_0,y_1) \\
&= f_{Y(0),Y(1),C}(y_0,y_1)-f_{Y(0),Y(1),\overline{D}}(y_0,y_1), \qquad (4.2)
\end{aligned}$$

which is positive under Assumption 2.4. Integrating Equation (4.2) over the support of $Y(1-d)$ implies that

$$f_{Y(d),C^*}(y_d)=f_{Y(d),C}(y_d)-f_{Y(d),\overline{D}}(y_d), \qquad (4.3)$$

Combining Equation (4.3) with Equation (4.1) implies that

$$P(C^*) = P(D = 1|Z = 1) - P(D = 1|Z = 0).$$

Equation (4.3) combined with Equation (3.1) implies that

$$f_{Y(d),C^*}(y) = f_{Y,D=d|Z=d}(y) - f_{Y,D=d|Z=1-d}(y),$$

which is positive under Assumption 2.4 as shown in Lemma 3.1. This proves the result.
QED.

To illustrate Theorem 4.1, consider the example in Figure 1. Let $Y(0)$ and $Y(1)$ be binary. Assume that $P(AT) = P(NT) = 0$, $P(C) = 0.75$ and $P(D) = 0.25$. Also, assume that the joint distributions of $Y(1)$ and $Y(0)$ among compliers and defiers are as presented in Figure 1. In each cell of the table, $P(A)$ should be understood as $P(A, Y(0) = y_0, Y(1) = y_1)$ for the values y_0 and y_1 corresponding to that cell, and for $A \in \{C, D, C^{**}\}$.

Y(0) \ Y(1)	0	1
0	P(C): 0.40 P(D): 0.05 P(C**): 0.05	P(C): 0.15 P(D): 0.05 P(C**): 0.05
1	P(C): 0.15 P(D): 0.10 P(C**): 0.10	P(C): 0.05 P(D): 0.05 P(C**): 0.05

Figure 1: Assumption 2.4 is sufficient.

The LDTC assumption is verified here since for every $(y_0, y_1) \in \{0; 1\}^2$,

$$P(\overline{D}, Y(0) = y_0, Y(1) = y_1) \leq P(C, Y(0) = y_0, Y(1) = y_1).$$

To construct C^{**}, one can merely set

$$P(C^{**}, Y(0) = y_0, Y(1) = y_1) = P(\overline{D}, Y(0) = y_0, Y(1) = y_1).$$

As a result, C^{**} has same size than defiers, and the same marginal distributions of potential outcomes. Moreover, C^{**} is a valid subpopulation of compliers since for every $(y_0, y_1) \in \{0; 1\}^2$,

$$0 \leq P(C^{**}, Y(0) = y_0, Y(1) = y_1) \leq P(C, Y(0) = y_0, Y(1) = y_1).$$

Therefore, in this simple example it is easy to see that the LDTC assumption is indeed sufficient to ensure that there exists subpopulations of compliers C^{**} of same size and with same marginal distributions of potential outcomes than defiers. Therefore, $C^* = C \setminus C^{**}$ is such that

$$P(C^*) = P(C) - P(\overline{D}) = P(D = 1|Z = 1) - P(D = 1|Z = 0),$$

and

$$f_{Y(d),C^*}(y) = f_{Y(d),C}(y) - f_{Y(d),\overline{D}}(y) = f_{Y,D=d|Z=d}(y) - f_{Y,D=d|Z=1-d}(y).$$

As a result, Theorem 4.1 holds.

The LDTC Assumption is sufficient for Theorem 4.1 to hold. However, it is not necessary, as shown in the counterexample in Figure 2.

Y(0) \ Y(1)	0	1
0	P(C): 0.40 P(D): 0.05 P(C**): 0.00	P(C): 0.15 P(D): 0.05 P(C**): 0.10
1	P(C): 0.15 P(D): 0.05 P(C**): 0.10	P(C): 0.05 P(D): 0.10 P(C**): 0.05

Figure 2: Assumption 2.4 is not necessary.

The LDTC assumption is not verified here since

$$P(\overline{D}, Y(0) = 1, Y(1) = 1) = 0.10 > P(C, Y(0) = 1, Y(1) = 1) = 0.05.$$

But C^{**} is a valid subpopulation of compliers since for every $(y_0, y_1) \in \{0; 1\}^2$,

$$0 \leq P(C^{**}, Y(0) = y_0, Y(1) = y_1) \leq P(C, Y(0) = y_0, Y(1) = y_1).$$

Moreover, the marginal distributions of $Y(1)$ and $Y(0)$ are the same in C^{**} than among defiers since the total of each line and each column are the same in the two populations.

In this second example, Assumption 3.1 is verified, since the total of each line and each column is greater for compliers than defiers. However, Figure 3 presents a last counterexample to show that Assumption 3.1 is not sufficient for Theorem 4.1 to hold. Indeed, the joint distributions of potential outcomes among compliers and defiers presented below are such that Assumption 3.1 is satisfied. Still, it is not possible to construct a subpopulation C^{**} of compliers with same marginal distributions of potential outcomes than defiers. Indeed, the maximum possible value of $P(C^{**}, Y(0) = 1, Y(1) = 1)$ is 0.05, because

$P(C, Y(0) = 1, Y(1) = 1) = 0.05$. As a result, to ensure that the marginal distributions are the same in C^{**} than among defiers, one must set

$$P(C^{**}, Y(0) = 1, Y(1) = 0) = P(C^{**}, Y(0) = 0, Y(1) = 1) = 0.15$$

and

$$P(C^{**}, Y(0) = 0, Y(1) = 0) = -0.10,$$

which is impossible.

Y(0) \ Y(1)	0	1
0	P(C): 0.40 P(D): 0.00 P(C**): -0.10	P(C): 0.15 P(D): 0.05 P(C**): 0.15
1	P(C): 0.15 P(D): 0.05 P(C**): 0.15	P(C): 0.05 P(D): 0.15 P(C**): 0.05

Figure 3: Assumption 3.1 is not sufficient.

Actually, it is possible to show that Assumption 3.1 is sufficient to ensure that there exists two subpopulations of compliers C_0^* and C_1^* such that

$$f_{Y(0),C_0^*}(y) = f_{Y(0),C}(y) - f_{Y(0),\overline{D}}(y) = f_{Y,D=0|Z=0}(y) - f_{Y,D=0|Z=1}(y)$$

and

$$f_{Y(1),C_1^*}(y) = f_{Y(1),C}(y) - f_{Y(1),\overline{D}}(y) = f_{Y,D=1|Z=1}(y) - f_{Y,D=1|Z=0}(y),$$

but the preceding counterexample shows that it is not sufficient to ensure that there exists a unique population C^* such that $f_{Y(0),C^*}(y) = f_{Y,D=0|Z=0}(y) - f_{Y,D=0|Z=1}(y)$ and $f_{Y(1),C^*}(y) = f_{Y,D=1|Z=1}(y) - f_{Y,D=1|Z=0}(y)$.

5 Characterizing C^*.

One key issue when estimating a LATE is to be able to describe the population of compliers to which it applies. Indeed, LATE are by definition local effects. Therefore, it is important to describe the population on which those effects are measured, so as to be able to assess whether the findings of the analysis are generalizable to other populations. In the standard model of Angrist et al. (1996), the distribution of any vector of covariates among compliers is identified from the data. Here, the LDTC assumption is not sufficient to ensure that the distribution of covariates in the subpopulation C^* is identified. However, a very weak strengthening of Assumption 2.4 is sufficient to have that the distribution of X within that population is identified.

Let X denote a vector of covariates. Consider the following assumption.

Assumption 5.1 *(LDTC-X)*

$$f_{Y(0),Y(1),X,\overline{D}}(y_0, y_1, x) \leq f_{Y(0),Y(1),X,C}(y_0, y_1, x), \tag{5.1}$$

or equivalently:

$$P(\overline{D}|Y(0), Y(1), X) \leq P(C|Y(0), Y(1), X), \tag{5.2}$$

almost surely.

It suffices to integrate (5.1) over the support of X to see that Assumption 5.1 is stronger than Assumption 2.4. But it also has more testable implications. Indeed, it is easy to show that under Assumptions 2.1 and 2.2 it implies that we should have

$$f_{Y,D=d,X=x|Z=1-d}(y) \leq f_{Y,D=d,X=x|Z=d}(y)$$

for $d \in \{0; 1\}$ and for any $x \in \mathcal{S}(X)$. The next Theorem shows that under Assumptions 2.1, 2.2 and 5.1, there exists a subpopulation of compliers C^* such that treatment effects and the distribution of X are both identified from the data within that population.

Theorem 5.1 *Suppose Assumptions 2.1, 2.2 and 5.1 hold and $P(D = 1|Z = 1) > P(D = 1|Z = 0)$. Then, there exists a subpopulation of compliers C^* such that*

$$P(C^*) = P(D = 1|Z = 1) - P(D = 1|Z = 0),$$

$$f_{Y(d),C^*}(y) = f_{Y,D=d|Z=d}(y) - f_{Y,D=d|Z=1-d}(y),$$

and

$$f_{X,C^*}(x) = f_{X,D=d|Z=d}(x) - f_{X,D=d|Z=1-d}(x).$$

Proof:
First, notice that

$$\begin{aligned} f_{X,D=d|Z=d}(x) - f_{X,D=d|Z=1-d}(x) &= f_{X,D(d)=d}(x) - f_{X,D(1-d)=d}(x) \\ &= f_{X,D(d)=d,D(1-d)=1-d}(x) - f_{X,D(1-d)=d,D(d)=1-d}(x) \\ &= f_{X,C}(x) - f_{X,\overline{D}}(x). \end{aligned} \tag{5.3}$$

The first equality follows from Assumption 2.1, others follow from standard arguments. Here, constructing C^* merely amounts to withdraw randomly a share

$$p(y_0, y_1, x) = \frac{f_{Y(0),Y(1),X,\overline{D}}(y_0, y_1, x)}{f_{Y(0),Y(1),X,C}(y_0, y_1, x)}$$

of compliers such that $Y(0) = y_0$, $Y(1) = y_1$, $X = x$ for each (y_0, y_1, x) in $\mathcal{S}(Y(0) \times Y(1) \times X)$. Indeed, Assumption 5.1 ensures that $p(y_0, y_1, x)$ is smaller than 1.

Formally, let $(B(y_0, y_1, x))_{y_0,y_1,x}$ denote a family of independent Bernoulli variables defined over $\mathcal{S}(Y(0) \times Y(1) \times X)$, independent of $(Y(0), Y(1))$, $(D(0), D(1))$ and X, and such that $P(B(y_0, y_1, x) = 1) = p(y_0, y_1, x)$. Let C^{**} denote the event $C \cap \{B(Y(0), Y(1), X) = 1\}$ and $C^* = C \setminus C^{**}$. Using the same steps as in the proof of Theorem 4.1, one can show that

$$f_{Y(0),Y(1),X,C^*}(y_0, y_1, x) = f_{Y(0),Y(1),X,C}(y_0, y_1, x) - f_{Y(0),Y(1),X,\overline{D}}(y_0, y_1, x), \tag{5.4}$$

which is positive under Assumption 5.1. Integrating Equation (5.4) over the support of $Y(1-d) \times X$ implies that Equation (4.3) holds here as well, which in turn implies that

$$P(C^*) = P(D = 1|Z = 1) - P(D = 1|Z = 0)$$

and

$$f_{Y(d),C^*}(y) = f_{Y,D=d|Z=d}(y) - f_{Y,D=d|Z=1-d}(y).$$

Similarly, integrating Equation (5.4) over the support of $Y(0) \times Y(1)$ implies that

$$f_{X,C^*}(x) = f_{X,C}(x) - f_{X,\overline{D}}(x). \tag{5.5}$$

Combining Equations (5.5) and (5.3) yields

$$f_{X,C^*}(x) = f_{X,D=d|Z=d}(x) - f_{X,D=d|Z=1-d}(x).$$

This completes the proof.
QED.

6 Identification of a LATE under Local Stochastic Monotonicity.

Using an approach similar to the one used to generalize the result in Angrist et al. (1996), it is also possible to generalize the results in Huber & Mellace (2012). Huber & Mellace (2012) show that when Assumption 3.1 is not verified in the data, which implies that neither Assumption 2.3 or 2.4 holds, it is possible to replace those assumptions by a local monotonicity condition under which treatment effects are identified among compliers and defiers. This Assumption writes as follows:

Assumption 6.1 *(Local monotonicity)*
Almost surely, $P(D(1) \geq D(0)|Y(0), Y(1)) = 1$ *or* $P(D(0) \geq D(1)|Y(0), Y(1)) = 1.$

It is weaker than monotonicity since it implies that there can be defiers in the population. But it means that among observations such that $Y(0) = y_0$ and $Y(1) = y_1$, there cannot be both defiers and compliers. Therefore, it implies that the distribution of potential outcomes among compliers and defiers must be extremely different, since their support cannot overlap. Consider now the following assumption:

Assumption 6.2 *(Local Stochastic Monotonicity)*
For some $d \in \{0; 1\}$, *for every* $y_d \in \mathcal{S}(Y(d))$,

$$\begin{aligned} f_{Y(d),\overline{D}}(y_d) \leq f_{Y(d),C}(y_d) \quad &\Rightarrow \quad f_{Y(0),Y(1),\overline{D}}(y_0, y_1) \leq f_{Y(0),Y(1),C}(y_0, y_1), \;\; \forall y_{1-d} \in \mathcal{S}(Y(1-d)) \\ f_{Y(d),\overline{D}}(y_d) \geq f_{Y(d),C}(y_d) \quad &\Rightarrow \quad f_{Y(0),Y(1),\overline{D}}(y_0, y_1) \geq f_{Y(0),Y(1),C}(y_0, y_1), \;\; \forall y_{1-d} \in \mathcal{S}(Y(1-d)), \end{aligned}$$

or equivalently,

$$\begin{aligned} P(\overline{D}|Y(d)) \leq P(C|Y(d)) \quad &\Rightarrow \quad P(\overline{D}|Y(0), Y(1)) \leq P(C|Y(0), Y(1)) \\ P(\overline{D}|Y(d)) \geq P(C|Y(d)) \quad &\Rightarrow \quad P(\overline{D}|Y(0), Y(1)) \geq P(C|Y(0), Y(1)) \end{aligned}$$

almost surely.

This condition is weaker than Assumption 6.1 since the support of $Y(1)$ and $Y(0)$ among defiers and compliers can overlap. Moreover, it is weaker than Assumption 2.4 since under Assumption 6.2 there can exist points y_d^* in the support of $Y(d)$ where there are more defiers than compliers, provided there are also more defiers at all points (y_d^*, y_{1-d}) in the support of $y_d^* \times Y(1-d)$.
The following Theorem shows that it is possible to substitute Assumption 6.2 to Assumption 6.1 while maintaining Huber & Mellace (2012) identification results. Indeed, if Assumptions 2.1, 2.2 and 6.2 hold, it is always possible to exhibit a population T^*, larger than C^*, including both defiers and compliers, and such that the marginal distributions of each potential outcome within this population are identified from the data and have the same expression as in Huber & Mellace (2012).

Theorem 6.1 *Suppose Assumptions 2.1, 2.2 and 6.2 hold and* $P(D = 1|Z = 1) > P(D = 1|Z = 0)$. *Then, there exists a subpopulation of compliers and defiers* T^* *such that*

$$\begin{aligned} P(T^*) \;\; = \;\; & P(D = 1|Z = 1) - P(D = 1|Z = 0) \\ & +2 \int \max \left(f_{Y,D=1|Z=1}(y) - f_{Y,D=1|Z=0}(y), 0 \right) dy, \end{aligned}$$

and

$$\begin{aligned} f_{Y(d),T^*}(y) &= \max\left(f_{Y,D=d|Z=d}(y), f_{Y,D=d|Z=1-d}(y)\right) \\ &- \min\left(f_{Y,D=d|Z=d}(y), f_{Y,D=d|Z=1-d}(y)\right). \end{aligned}$$

Consequently, $E[Y(1)-Y(0)|T^*]$ *is identified from the data.*

Proof:

Intuitively, constructing T^* amounts to start from compliers and defiers, and to withdraw randomly a share $p(y_0,y_1)$ of compliers when $p(y_0,y_1)$ is lower than 1, and a share $\frac{1}{p(y_0,y_1)}$ of defiers when it is greater than 1.

Formally, let

$$p^c(y_0,y_1) = \min\left(\frac{f_{Y(0),Y(1),\overline{D}}(y_0,y_1)}{f_{Y(0),Y(1),C}(y_0,y_1)}, 1\right)$$

and

$$p^{\overline{d}}(y_0,y_1) = \min\left(\frac{f_{Y(0),Y(1),C}(y_0,y_1)}{f_{Y(0),Y(1),\overline{D}}(y_0,y_1)}, 1\right).$$

Let $(B^c(y_0,y_1))_{y_0,y_1}$ and $\left(B^{\overline{d}}(y_0,y_1)\right)_{y_0,y_1}$ denote two families of independent Bernoulli variables defined over $\mathcal{S}(Y(0)\times Y(1))$, independent of $(Y(0),Y(1))$ and $(D(0),D(1))$, and such that $P(B^c(y_0,y_1)=1)=p^c(y_0,y_1)$ and $P(B^{\overline{d}}(y_0,y_1)=1)=p^{\overline{d}}(y_0,y_1)$. Let C^* and $\overline{D}^*$ respectively denote the events $C\cap\{B^c(Y(0),Y(1))=0\}$ and $\overline{D}\cap\{B^{\overline{d}}(Y(0),Y(1))=0\}$. Finally, let $T^*=C^*\cup\overline{D}^*$. Note that in general, T^* is not uniquely defined, unless $p^c(y_0,y_1)$ and $p^{\overline{d}}(y_0,y_1)$ are both equal to 0 or 1 for every (y_0,y_1). We have

$$\begin{aligned}
f_{Y(0),Y(1),T^*}(y_0,y_1) &= f_{Y(0),Y(1),\left(C^*\cup\overline{D}^*\right)}(y_0,y_1) \\
&= f_{Y(0),Y(1),C^*}(y_0,y_1) + f_{Y(0),Y(1),\overline{D}^*}(y_0,y_1) \\
&= f_{Y(0),Y(1),C,B^c(Y(0),Y(1))=0}(y_0,y_1) + f_{Y(0),Y(1),\overline{D},B^{\overline{d}}(Y(0),Y(1))=0}(y_0,y_1) \\
&= f_{Y(0),Y(1),C,B^c(y_0,y_1)=0}(y_0,y_1) + f_{Y(0),Y(1),\overline{D},B^{\overline{d}}(y_0,y_1)=0}(y_0,y_1) \\
&= f_{Y(0),Y(1),C}(y_0,y_1)P(B^c(y_0,y_1)=0) \\
&+ f_{Y(0),Y(1),\overline{D}}(y_0,y_1)P(B^{\overline{d}}(y_0,y_1)=0) \\
&= f_{Y(0),Y(1),C}(y_0,y_1)\max\left(1-\frac{f_{Y(0),Y(1),\overline{D}}(y_0,y_1)}{f_{Y(0),Y(1),C}(y_0,y_1)}, 0\right) \\
&+ f_{Y(0),Y(1),\overline{D}}(y_0,y_1)\max\left(1-\frac{f_{Y(0),Y(1),C}(y_0,y_1)}{f_{Y(0),Y(1),\overline{D}}(y_0,y_1)}, 0\right) \\
&= \max\left(f_{Y(0),Y(1),C}(y_0,y_1), f_{Y(0),Y(1),\overline{D}}(y_0,y_1)\right) \\
&- \min\left(f_{Y(0),Y(1),C}(y_0,y_1), f_{Y(0),Y(1),\overline{D}}(y_0,y_1)\right). \qquad (6.1)
\end{aligned}$$

The first equality arises from the definition of T^*. The second one is due to the fact that the events C^* and $\overline{D}^*$ are disjoints since C and $\overline{D}$ are disjoints. The third arises from the definition of C^* and $\overline{D}^*$. The fourth comes from the fact that the event $Y(0) = y_0, Y(1) = y_1, B^t(Y(0), Y(1)) = 1$ is equal to $Y(0) = y_0, Y(1) = y_1, B^t(y_0, y_1) = 1$ for every t in $\{c, d\}$. The fifth follows from independence. Finally, the last is obtained considering separately the cases when $f_{Y(0),Y(1),\overline{D}}(y_0, y_1) \leq f_{Y(0),Y(1),C}(y_0, y_1)$ and when $f_{Y(0),Y(1),\overline{D}}(y_0, y_1) > f_{Y(0),Y(1),C}(y_0, y_1)$.

If $f_{Y(d),\overline{D}}(y_d) \leq f_{Y(d),C}(y_d)$, Assumption 6.2 implies that

$$f_{Y(0),Y(1),\overline{D}}(y_0, y_1) \leq f_{Y(0),Y(1),C}(y_0, y_1), \ \forall y_{1-d} \in \mathcal{S}(Y(1-d)).$$

Therefore,

$$\begin{aligned}
& \int_{\mathcal{S}(Y(1-d))} \max\left(f_{Y(0),Y(1),C}(y_0, y_1), f_{Y(0),Y(1),\overline{D}}(y_0, y_1)\right) dy_{1-d} \\
= & \int_{\mathcal{S}(Y(1-d))} f_{Y(0),Y(1),C}(y_0, y_1) dy_{1-d} \\
= & \ f_{Y(d),C}(y_d) \\
= & \max\left(f_{Y(d),C}(y_d), f_{Y(d),\overline{D}}(y_d)\right). \qquad (6.2)
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \int_{\mathcal{S}(Y(1-d))} \min\left(f_{Y(0),Y(1),C}(y_0, y_1), f_{Y(0),Y(1),\overline{D}}(y_0, y_1)\right) dy_{1-d} \\
= & \int_{\mathcal{S}(Y(1-d))} f_{Y(0),Y(1),\overline{D}}(y_0, y_1) dy_{1-d} \\
= & \ f_{Y(d),\overline{D}}(y_d) \\
= & \min\left(f_{Y(d),C}(y_d), f_{Y(d),\overline{D}}(y_d)\right). \qquad (6.3)
\end{aligned}$$

Combining Equations (6.2) and (6.3) shows that when $f_{Y(d),\overline{D}}(y_d) \leq f_{Y(d),C}(y_d)$, Assumption 6.2 implies that

$$\begin{aligned}
& \int_{\mathcal{S}(Y(1-d))} \max\left(f_{Y(0),Y(1),C}(y_0, y_1), f_{Y(0),Y(1),\overline{D}}(y_0, y_1)\right) dy_{1-d} \\
- & \int_{\mathcal{S}(Y(1-d))} \min\left(f_{Y(0),Y(1),C}(y_0, y_1), f_{Y(0),Y(1),\overline{D}}(y_0, y_1)\right) dy_{1-d} \\
= & \max\left(f_{Y(d),C}(y_d), f_{Y(d),\overline{D}}(y_d)\right) \\
- & \min\left(f_{Y(d),C}(y_d), f_{Y(d),\overline{D}}(y_d)\right). \qquad (6.4)
\end{aligned}$$

One can also show that when $f_{Y(d),\overline{D}}(y_d) \geq f_{Y(d),C}(y_d)$, Equation (6.4) also holds under Assumption 6.2, which implies that Equation (6.4) is always true.

Finally, note that Equation (3.1) implies that

$$f_{Y,D=d|Z=d}(y) \geq f_{Y,D=d|Z=1-d}(y) \Leftrightarrow f_{Y(d),C}(y) \geq f_{Y(d),\overline{D}}(y). \tag{6.5}$$

Therefore, integrating Equation (6.1) over $\mathcal{S}(Y(1-d)$, combining the result with (6.4), and then using (3.1) and (6.5) finally yields

$$\begin{aligned} f_{Y(d),T^*}(y) &= \max\left(f_{Y,D=d|Z=d}(y), f_{Y,D=d|Z=1-d}(y)\right) \\ &- \min\left(f_{Y,D=d|Z=d}(y), f_{Y,D=d|Z=1-d}(y)\right). \end{aligned}$$

This completes the proof, once noted that

$$\begin{aligned} & \max\left(f_{Y,D=1|Z=1}(y), f_{Y,D=1|Z=0}(y)\right) - \min\left(f_{Y,D=1|Z=1}(y), f_{Y,D=1|Z=0}(y)\right) \\ = \quad & f_{Y,D=1|Z=1}(y) - f_{Y,D=1|Z=0}(y) + 2\max\left(f_{Y,D=1|Z=0}(y) - f_{Y,D=1|Z=1}(y), 0\right). \end{aligned}$$

QED.

Note that when $f_{Y,D=d|Z=d}(y) \geq f_{Y,D=d|Z=1-d}(y)$, $T^* = C^*$. Moreover, since Assumption 6.2 is weaker than Assumption 2.4, Theorem 6.1 is a generalization of Theorem 4.1.

7 Concluding comments

The main result in this paper is that in the instrumental variable model with heterogeneous treatment effects of Angrist et al. (1996), the monotonicity condition can be replaced by a substantially weaker assumption which states that there must be less defiers than compliers conditional on every value of the potential outcomes. This condition is closely related to the testable implications of this model derived in Imbens & Rubin (1997) and Kitagawa (2008).

Using a similar type of analysis, one can show that this result can be extended to many treatment effects models which also rely on monotonicity. Important examples include fuzzy Regression Discontinuity Design and quantile IV. In the fuzzy Regression Discontinuity Design model in Hahn et al. (2001), let T be the forcing variable and t the cut-off value of this forcing variable. One can show that their Theorem 3 still holds if Assumption A.3, which is essentially a monotonicity assumption at the threshold, is replaced by a weaker LDTC type of condition which states at $T = t$, there are more compliers than defiers conditional on every value of the potential outcomes. Formally, their identification result still holds if almost surely,

$$P(\overline{D}|Y(0), Y(1), T = t) \leq P(C|Y(0), Y(1), T = t).$$

Another important example is the linear quantile IV regression model developed in Abadie et al. (2002). Estimation of this model relies on the "κ" identification results in Theorem 3.1 of Abadie (2003). One can show that this Theorem still holds if point iv) in Assumption 2.1 is replaced by a conditional LDTC assumption, i.e. by

$$P(\overline{D}|Y(0), Y(1), X) \leq P(C|Y(0), Y(1), X)$$

almost surely, where X is the set of covariates included in the regression.

References

Abadie, A. (2003), 'Semiparametric instrumental variable estimation of treatment response models', *Journal of Econometrics* **113**(2), 231 – 263.

Abadie, A., Angrist, J. & Imbens, G. (2002), 'Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings', *Econometrica* **70**(1), 91–117.

Angrist, J. D. & Evans, W. N. (1998), 'Children and their parents' labor supply: Evidence from exogenous variation in family size', *The American Economic Review* **88**(3), 450–477.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996), 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association* **91**(434), pp. 444–455.

Barua, R. & Lang, K. (2010), School entry, educational attainment and quarter of birth: A cautionary tale of late. Working Paper.

Chaisemartin, C. D. & D'Haultfoeuille, X. (2012), Late again with defiers, Pse working papers.

Hahn, J., Todd, P. & Van der Klaauw, W. (2001), 'Identification and estimation of treatment effects with a regression-discontinuity design', *Econometrica* **69**(1), 201–209.

Huber, M. & Mellace, G. (2012), Relaxing monotonicity in the identification of local average treatment effects. Working Paper.

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–75.

Imbens, G. W. & Rubin, D. B. (1997), 'Estimating outcome distributions for compliers in instrumental variables models', *Review of Economic Studies* **64**(4), 555–574.

Kitagawa, T. (2008), A bootstrap test for instrument validity in the heterogeneous treatment effect model. Working Paper.

Klein, T. J. (2010), 'Heterogeneous treatment effects: Instrumental variables without monotonicity?', *Journal of Econometrics* **155**(2).

Small & Tan (2007), A stochastic monotonicity assumption for the instrumental variables method. Working Paper.