Defying the LATE? Identification of local treatment effects when the instrument violates monotonicity

Clement de Chaisemartin

No 1020

WARWICK ECONOMIC RESEARCH PAPERS

DEPARTMENT OF ECONOMICS

THE UNIVERSITY OF WARWICK

Defying the LATE? Identification of local treatment effects when the instrument violates monotonicity.*

Clément de Chaisemartin†

August 2013

Abstract

The instrumental variable method relies on a strong "no-defiers" condition, which requires that the instrument affect every subject's treatment decision in the same direction. This paper shows that "no-defiers" can be replaced by a weaker "compliers-defiers" condition, which requires that a subgroup of compliers have the same size and the same distribution of potential outcomes as defiers. This condition is necessary and sufficient for IV to capture causal effects for the remaining part of compliers. In many applications, "compliers-defiers" is a very weak condition. For instance, in Angrist & Evans (1998), 94% of DGPs compatible with the data satisfy "compliers-defiers", while 0% satisfy "no-defiers".

Keywords: instrumental variable, heterogeneous effects, defiers, single index model.
JEL Codes: C21, C26

*I benefited from numerous discussions with Josh Angrist and Xavier D'Haultfœuille. I am also very grateful to Alberto Abadie, Luc Behaghel, Stéphane Bonhomme, Federico Bugni, Laurent Davezies, Julien Grenet, Marc Gurgand, Martin Huber, Toru Kitagawa, Tobias Klein, Kevin Lang, Giovanni Mellace, Dennis Novy, Petra Persson, Roland Rathelot, and Edward Vytlacil for their helpful comments. I am also very grateful to seminar participants at Crest, PSE, CORE, Brown, Harvard, MIT, Boston College, Columbia, Warwick University, Bocconi University, Université de Montréal, HEC Montréal, Duke University, Chicago Booth School of Business, Cambridge University, Bristol University, Copenhagen Business School, Pompeu Fabra University, University of St-Gallen, and IFS. I gratefully acknowledge financial support from région Île de France.

†Department of Economics, University of Warwick, clementdechaisemartin@gmail.com

1 Introduction

Instrumental variable (IV) is a standard tool to measure the effect of a treatment. However, IV relies on a strong "no-defiers" (ND) condition which might not always be credible. Barua & Lang (2010) highlight several empirical papers in which it is fairly clear that some subjects are defiers. This paper addresses this limitation. It shows that the aforementioned ND assumption can be replaced by a substantially weaker condition, while leaving all the standard theorems almost unchanged.

Imbens & Angrist (1994) and Angrist et al. (1996) show that IV captures a causal effect under three assumptions: random assignment, exclusion restriction, and no-defiers. The parameter it identifies is the effect of the treatment among subjects who are induced to get treated because of the instrument, whom they call "compliers". This is the so-called local average treatment effect (LATE). ND means that the instrument will never induce a subject not to get treated. If it is violated, the IV estimate may be misleading, as it is equal to the average effect of the treatment for compliers minus the average effect for defiers; even in a world where all effects are positive, the IV estimand need not be positive.

Nevertheless, ND may be problematic in some applications. A first example consists in papers using sibling-sex composition as an IV to study the effect of childbearing on labor supply, as in Angrist & Evans (1998). When their first two children are of the same sex, the amount of parents having a third child is 6 percentage points higher than when their first two children are of a different sex. This implies that some parents have a preference for diversity. Those parents are compliers: the instrument induces them to have a third child. But when their first two children are girls, the amount of parents having a third child is 1.5 percentage points higher than when their first two children are boys. This implies that some parents are sex biased, either because they have a preference for boys, or because they find it more tiring to raise boys than girls. Among sex-biased parents, some might decide to have a third child if their first two children are a boy and a girl, while they would have decided otherwise if their first two children had been boys. Such parents would be defiers, as the instrument would induce them not to have a third child.

A second example consists in papers using month of birth as an instrument for school entry age, or for age when taking a test. As Barua & Lang (2010) argue, ND is problematic in

those papers. A well-known example is Bedard & Dhuey (2006). This paper compares test scores of children born before and after the cutoff date for school eligibility. In most countries, this cut-off date is January, 1. Children born in December should thus enter school at a younger age than children born in January, and should subsequently take national tests at a younger age as well. Nonetheless, as is well-known, parents are more prone to delaying school entry of children born late in the year, so-called "redshirting". Children redshirted because they were born in December are defiers, as they would have taken national tests at a younger age had they been born in January.

A third example consists in randomized experiments relying on an encouragement design. In such designs, the encouragement group typically receives a financial incentive to get treated, or a flyer describing the treatment, or both. ND means that the incentive and the flyer will positively impact every subject's treatment decision. In practice, those incentives might have the expected positive effect on a majority of subjects, but might still discourage some of them from getting treated. For instance, the flyer might lead some subjects to think that the benefit they can expect from treatment is lower than what they initially thought.

A fourth example consists in papers using randomly assigned judges with different sentencing propensities as an instrument for incarceration length, as in Kling (2006). In those papers, ND requires that, for every defendant, a judge with a high average of strictness will hand down a sentence that is more severe than that of a judge who is on average more lenient. This might be violated. Assume judge A does not take into account the defendant's socio-economic background, while judge B tends to be more lenient towards defendants from disadvantaged environments, and more severe with defendants from well-off backgrounds. If the pool of defendants bears more poor than rich individuals, judge B will be on average more lenient than judge A, but he will be more severe with socio-economically privileged defendants.

This paper shows that instruments still identify causal effects if ND is replaced by a weaker "compliers-defiers" (CD) condition. Let Y_0 and Y_1 respectively denote a subject's potential outcome without and with the treatment. The CD assumption requires that a subgroup of compliers, which I call comfiers, have the same size and the same probability distribution of Y_0 and Y_1 as defiers. Then, treatment effects among comfiers and defiers cancel one another

out, and the IV estimate finally identifies the average treatment effect among the remaining part of compliers. This second subgroup consists in compliers who "out-survive" defiers, and accordingly I call them comvivors. Comvivors have the same size as the population of compliers under ND; substituting CD to ND does not lead to a loss in terms of the external validity of the identified parameter. Moreover, quantile treatment effects among comvivors are also identified under that condition. Finally, the distribution of any vector of covariates among comvivors is identified as well, under a mild strengthening of the CD condition.

CD is therefore sufficient for IV to capture causal effect among a subpopulation of compliers. It is also necessary: if IV captures treatment effects for a subgroup of compliers, then CD must hold.

Albeit somewhat abstract, CD stands in between two easily interpretable conditions. For CD to be true, it suffices that there be more compliers than defiers in each stratum of the population with the same value of (Y_0, Y_1), and it is necessary that there be more compliers than defiers in each stratum with the same value of Y_0 and in each stratum with the same value of Y_1. I call the former condition the "more compliers than defiers" (MC) condition, and I call the latter the "weak more compliers than defiers" (WMC) condition.

The CD condition is close to being fully testable, at least when potential outcomes are binary. WMC is testable from the data, for instance using the test developed in Kitagawa (2008). CD implies WMC, but the converse is not true, as some DGPs satisfy WMC but violate CD. Therefore, CD cannot be consistently tested. Notwithstanding, I conduct a simulation exercise to show that when potential outcomes are binary, a very low fraction of DGPs satisfying WMC violate CD. In Angrist & Evans (1998), Kitagawa's test is not rejected. I put a uniform measure on the set of DGPs compatible with the data, and find that 94% of them satisfy CD, while 32% satisfy MC, and 0% satisfy ND. As a result, CD is very likely to hold in this application; even an "agnostic" decision maker with this uninformative uniform prior over the set of DGPs would choose to believe in it.

The CD condition does not arise from a structural model in which subjects make treatment decisions after solving a utility maximization problem. Embedding this condition in such a model would be useful, to assess whether it is significantly weaker than ND not

only from the perspective of an agnostic decision maker, but also from the perspective of economic theory. Nonetheless, giving a structural interpretation to the CD condition is not easy. Instead, I derive a structural interpretation of the MC condition, and discuss how it compares with the structural interpretation of ND derived in Vytlacil (2002).

Vytlacil (2002) shows that ND is observationally equivalent to a single index model for potential treatments. In this model, the instrument affects the threshold above which an observation gets treated, but it does not affect the unobserved heterogeneity index. Therefore, one interpretation of ND is that the instrument affects the cost of treatment, but not the benefits subjects expect from it. This might be violated in the encouragement design example, as subjects' expectations of treatment benefits may be altered after reading the flyer.

MC is observationally equivalent to an asymmetric index model. In this model, the instrument can affect both the cost and the benefit expected from treatment, provided that in each (Y_0, Y_1) stratum there be more subjects whose expected benefit is greater with than without the instrument. In the encouragement design example, MC requires that the flyer be sufficiently appealing to ensure that in each (Y_0, Y_1) cell, the number of subjects having a worse opinion of the treatment is equal to or less than the number of subjects having a better opinion of it after reading the flyer.

In Vytlacil's single index model, compliers are "marginal" subjects; their expected benefit from treatment is too low for them to get treated if they do not receive any incentive, but high enough for them to do so if they do receive one. Comvivors can receive a similar interpretation, under a symmetric index model which is more restrictive than the asymmetric one.

In a related paper, Small & Tan (2007) demonstrate that under MC, the Wald ratio captures a weighted average treatment effect. Nevertheless, their weights can lie outside the unit interval. As a result, their weighting scheme does not capture causal effects for a well-defined subpopulation of units. I show here that CD is necessary and sufficient to capture causal effects for a subgroup of compliers of known size, that I also describe in the context of specific assignment models.

Other related papers include DiNardo & Lee (2011), who derive a result very similar to

Small & Tan (2007). Klein (2010) introduces "local" violations of ND, and shows that the bias of the Wald parameter can be well approximated if said violations are small. When Kitagawa's test is rejected, meaning that both ND and CD are violated, Huber & Mellace (2012) identify average treatment effects on compliers and defiers under a local ND condition. Finally, Hoderlein & Gautier (2012) introduce a selection model with random coefficients where there can be both defiers and compliers. But their instrument is continuous, while mine is binary.

The remainder of the paper is organized as follows. Section 2 presents the model and introduces the MC, CD, and WMC conditions. Section 3 presents the main identification result of the paper. Section 4 shows that the distribution of any vector of covariates among comvivors is identified under a weak strengthening of CD. Section 5 discusses the testability of CD. Section 6 shows that even a bayesian decision maker with an uninformative prior over the set of DGPs would choose to believe in CD in most applications in which it is not rejected. Section 7 gives a structural interpretation to the MC condition and to the comvivors population. Section 8 shows how the ideas presented in this paper extend to other microeconometrics models relying on ND type of conditions. Section 9 concludes.

2 Framework and assumptions

For any random variable X, let $\mathcal{S}(X)$ denote the support of X. Let Z be a binary instrument. Let $D_z \in \{0;1\}$ denote a subject's potential treatment when $Z = z$.[1] Let Y_{dz} denote her potential outcomes as functions of the treatment and of the instrument. Only Z, $D = D_Z$ and $Y = Y_{DZ}$ are observed. Following Imbens & Angrist (1994), never takers (NT) are subjects such that $D_0 = 0$ and $D_1 = 0$. Always takers (AT) are such that $D_0 = 1$ and $D_1 = 1$. Compliers (C) satisfy $D_0 = 0$ and $D_1 = 1$, while defiers (F)[2] satisfy $D_0 = 1$ and $D_1 = 0$. Let $T \in \{NT; AT; C; F\}$ be a random variable describing the type of a subject.

[1] For now, I focus on binary instrument and treatment, but Section 8 extends some results of the paper to multivariate instrument and treatment.

[2] In most of the treatment effect literature, treatment is denoted by D. To avoid confusion, defiers are denoted by the letter F throughout the paper.

Angrist et al. (1996) assume first that $P(D = 1|Z = 1) > P(D = 1|Z = 0)$, which means that the instrument has an effect on the treatment rate. This is equivalent to the rank condition in standard IV models. Under Assumption 2.1 (see below), this condition implies that more subjects are compliers than defiers: $P(C) > P(F)$.[3] Then, they make three assumptions. First, they assume that the instrument is independent of potential treatments and outcomes.

Assumption 2.1 *(Instrument independence)*

$$(Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1) \perp\!\!\!\perp Z.$$

Second, they assume that the instrument has an impact on the outcome only through its impact on treatment.

Assumption 2.2 *(Exclusion restriction)*
$\forall d \in \{0, 1\}$,

$$Y_{d0} = Y_{d1} = Y_d.$$

Third, they assume that the instrument moves all subjects into the same direction.

Assumption 2.3 *(No-defiers: ND)*

$$D_1 \geq D_0.$$

This paper shows that Assumption 2.3 can be substantially weakened while keeping results in Angrist et al. (1996) and Imbens & Rubin (1997) essentially unchanged. I consider a first weakening of ND, which is a special case of the stochastic monotonicity assumption in Small & Tan (2007).

Assumption 2.4 *(More compliers than defiers: MC, Small & Tan (2007))*
Almost surely,

$$P(F|Y_0, Y_1) \leq P(C|Y_0, Y_1). \tag{2.1}$$

[3] If $P(D = 1|Z = 1) < P(D = 1|Z = 0)$, one can switch the words "defiers" and "compliers" in what follows.

MC requires that each subgroup of the population with the same value of (Y_0, Y_1) comprise more compliers than defiers. It is automatically satisfied when $P(F) = 0$.

Angrist et al. (1996) show that the Wald ratio identifies a LATE if there are no defiers, or if defiers and compliers have the same distribution of (Y_0, Y_1). Those two assumptions are "polar cases" of MC. Assume to simplify that Y_0 and Y_1 are discrete. Then, MC is equivalent to

$$\frac{P(Y_0 = y_0, Y_1 = y_1|F)}{P(Y_0 = y_0, Y_1 = y_1|C)} \leq \frac{P(C)}{P(F)}. \tag{2.2}$$

Therefore, MC holds when defiers and compliers have the same distribution of (Y_0, Y_1), as the left hand side of Equation (2.2) is then equal to 1, while its right hand side is greater than 1.[4] And MC also holds when there are no defiers, as $P(F) = 0$ implies that the right hand side of Equation (2.2) is equal to $+\infty$.

In between those two polar cases, MC holds in many intermediate cases. Equation (2.2) shows that under MC, the more defiers there are, the less the distribution of (Y_0, Y_1) among compliers and defiers can differ, and conversely. Indeed, when compliers and defiers have very different distributions of (Y_0, Y_1), i.e. when the highest value of

$$\frac{P(Y_0 = y_0, Y_1 = y_1|F)}{P(Y_0 = y_0, Y_1 = y_1|C)}$$

is large, $P(F)$ should be close to 0 for Equation (2.2) to hold. On the contrary, when $P(F)$ is close to $P(C)$, compliers and defiers should have very similar distributions of potential outcomes for Equation (2.2) to hold. Therefore, MC will hold if few subjects are defiers, or if defiers and compliers have reasonably similar distributions of potential outcomes. On the contrary, if many subjects are defiers and defiers and compliers have very different distributions of potential outcomes, then MC will be violated.

MC is weaker than ND, but identification results derived in this paper are obtained under an even weaker condition. Before stating it, I introduce a very mild technical condition which will be maintained throughout the paper.

Assumption 2.5 *(Absolute continuity)*

The vector (Y_0, Y_1, T) has a density with respect to some σ-finite measure λ.

[4] I have assumed, as a mere normalization, that $P(F)$ is smaller than $P(C)$.

For any vector of random variables X absolutely continuous with respect to λ, let f_X denote the density of X. For any event A, let $f_{X|A}$ denote the density of X conditional on A.

Assumption 2.6 *(Compliers-defiers: CD)*[5]
There is a real-valued function g defined on $\mathcal{S}(Y_0) \times \mathcal{S}(Y_1)$ such that almost everywhere:

$$0 \leq \quad g(y_0, y_1) \quad \leq f_{Y_0,Y_1,T}(y_0, y_1, C) \tag{2.3}$$

$$\int_{\mathcal{S}(Y_1)} \quad g(y_0, y_1) \quad d\lambda(y_1) = f_{Y_0,T}(y_0, F) \tag{2.4}$$

$$\int_{\mathcal{S}(Y_0)} \quad g(y_0, y_1) \quad d\lambda(y_0) = f_{Y_1,T}(y_1, F). \tag{2.5}$$

As I show in the proofs, CD ensures that a subgroup of compliers, hereafter referred to as "compliers-defiers", or "comfiers", has the same size and the same marginal distributions of Y_0 and Y_1 as defiers. Conversely, if a subpopulation of compliers has the same size and the same marginal distributions of Y_0 and Y_1 as defiers, then CD must hold.

CD is weaker than MC. Indeed, MC implies

$$0 \leq f_{Y_0,Y_1,T}(y_0, y_1, F) \leq f_{Y_0,Y_1,T}(y_0, y_1, C)$$

almost everywhere.[6] Then, $f_{Y_0,Y_1,T}(y_0, y_1, F)$ satisfies Equations (2.3), (2.4), and (2.5).

Albeit weaker than MC, CD is more difficult to interpret. To ease its interpretation, I consider a last assumption.

Assumption 2.7 *(Weak more compliers than defiers: WMC)*
Almost surely,

$$P(F|Y_0) \leq P(C|Y_0)$$

$$P(F|Y_1) \leq P(C|Y_1). \tag{2.6}$$

[5]With a slight abuse of notation, λ should be understood in Equation (2.4) as the "marginal" measure defined on the σ-algebra associated to $\mathcal{S}(Y_1)$. Similar abuses of notations appear at a few other places in the paper, for instance in Equation (2.5).

[6]To see it, assume that $f_{Y_0,Y_1,T}(y_0, y_1, F) > f_{Y_0,Y_1,T}(y_0, y_1, C)$ over a set A with strictly positive probability. This implies $P(F|(Y_0, Y_1) \in A) > P(C|(Y_0, Y_1) \in A)$ which violates MC.

For WMC to hold, each subgroup of the population with the same value of Y_0 should comprise more compliers than defiers, and each subgroup with the same value of Y_1 should also comprise more compliers than defiers.

CD is stronger than WMC. Indeed, integrating

$$g(y_0, y_1) \leq f_{Y_0,Y_1,T}(y_0, y_1, C)$$

over $\mathcal{S}(Y_1)$ and dividing each side of the resulting inequality by $f_{Y_0}(y_0)$ yields $P(F|Y_0 = y_0) \leq P(C|Y_0 = y_0)$. Integrating the same inequality over $\mathcal{S}(Y_0)$ and dividing each side by $f_{Y_1}(y_1)$ yields $P(F|Y_1 = y_1) \leq P(C|Y_1 = y_1)$.

As MC $\Rightarrow$ CD $\Rightarrow$ WMC, CD is stronger than having more compliers than defiers in each Y_0 and each Y_1 subgroup, but it is weaker than having more compliers than defiers in each (Y_0, Y_1) subgroup.

3 Identification of local treatment effects under CD

Let

$$\begin{aligned}
f_0(y_0) &= \frac{f_{Y,D|Z=0}(y_0, 0) - f_{Y,D|Z=1}(y_0, 0)}{P(D = 0|Z = 0) - P(D = 0|Z = 1)} \\
f_1(y_1) &= \frac{f_{Y,D|Z=1}(y_1, 1) - f_{Y,D|Z=0}(y_1, 1)}{P(D = 1|Z = 1) - P(D = 1|Z = 0)} \\
W &= \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)}.
\end{aligned}$$

W is the Wald ratio; f_0 f_1 are functions appearing in Imbens & Rubin (1997). All those quantities are identified from the data, as they only involve Y, D, and Z.

Imbens & Angrist (1994), Angrist et al. (1996), and Imbens & Rubin (1997) show the following results.

LATE Theorems

Suppose Assumptions 2.1, 2.2, and 2.3 hold. Then,

$$P(C) = P(D = 1|Z = 1) - P(D = 1|Z = 0)$$
$$f_{Y_0|C}(y_0) = f_0(y_0)$$
$$f_{Y_1|C}(y_1) = f_1(y_1)$$
$$E[Y_1 - Y_0|C] = W.$$

Under the standard LATE assumptions, the size of compliers and their marginal distributions of Y_0 and Y_1 are identified from the data. As a result, the average effect of the treatment as well as quantile treatment effects are identified within that subpopulation.

Theorem 3.1 generalizes this result, by showing that when CD is substituted to ND, the Wald ratio identifies the average effect of the treatment for a subgroup of compliers of size $P(D = 1|Z = 1) - P(D = 1|Z = 0)$, while f_0 and f_1 identify the marginal distributions of Y_0 and Y_1 for the same subpopulation.

Theorem 3.1 *Suppose Assumptions 2.1 and 2.2 hold.*
If Assumption 2.6 holds, then there is a subpopulation of compliers C_V such that:

$$P(C_V) = P(D = 1|Z = 1) - P(D = 1|Z = 0) \tag{3.1}$$
$$f_{Y_0|C_V}(y_0) = f_0(y_0) \tag{3.2}$$
$$f_{Y_1|C_V}(y_1) = f_1(y_1) \tag{3.3}$$
$$E[Y_1 - Y_0|C_V] = W. \tag{3.4}$$

Conversely, if there is a subpopulation of compliers C_V satisfying Equations (3.1), (3.2), (3.3), *and* (3.4), *then Assumption 2.6 must hold.*

When there are defiers, Angrist et al. (1996) show that the Wald ratio is equal to a weighted difference between the average effect of the treatment among compliers and defiers:

$$W = \frac{P(C)E[Y_1 - Y_0|C] - P(F)E[Y_1 - Y_0|F]}{P(C) - P(F)}. \tag{3.5}$$

The numerator of the Wald ratio is the difference between $E(Y|Z = 1)$ and $E(Y|Z = 0)$. Always takers and never takers vanish away in this numerator, as the same potential

outcome is observed for them irrespective of whether $Z = 1$ or $Z = 0$. Only compliers and defiers remain. Compliers go from treatment to non-treatment when the instrument moves from 1 to 0, which is the reason why their average treatment effect appears. Defiers go from non-treatment to treatment, so that the opposite of their treatment effect appears. Therefore, Angrist et al. (1996) argue that the Wald ratio has no causal interpretation when there are defiers; it could for instance be positive while the average treatment effect is negative both among compliers and defiers.

But if a subpopulation of comfiers C_F has the same size and the same marginal distributions of potential outcomes as defiers, then the Wald ratio captures the average effect of the treatment among the remaining part of compliers, C_V. In such instances, the average effect of the treatment among compliers can be decomposed into

$$\begin{aligned} & E\left[Y_1 - Y_0|C\right] \\ = \; & P(C_V|C)E\left[Y_1 - Y_0|C_V\right] + P(C_F|C)E\left[Y_1 - Y_0|C_F\right] \\ = \; & \frac{P(C) - P(F)}{P(C)}E\left[Y_1 - Y_0|C_V\right] + \frac{P(F)}{P(C)}E\left[Y_1 - Y_0|F\right]. \end{aligned} \tag{3.6}$$

Plugging Equation (3.6) into (3.5), yields

$$W = E\left[Y_1 - Y_0|C_V\right].$$

Intuitively, this is because the average treatment effect among comfiers and defiers cancel one another out in Equation (3.5). Therefore, C_V are hereafter referred to as "compliers-survivors", or "comvivors", as they are compliers who "out-survive" defiers.

As a result, proving the first statement of Theorem 3.1 amounts to proving that if CD holds, then a subpopulation of compliers C_F has the same size and the same marginal distributions of potential outcomes as defiers. I use a constructive argument to prove this, by showing that one can always construct such a subgroup C_F when CD is satisfied. To convey the intuition, I consider first an example in which MC holds, as is the case in Figure 1. Y_0 and Y_1 are binary. The population bears 20 subjects. 13 of them are compliers, while 7 are defiers. Those 20 subjects are scattered over the 4 (Y_0, Y_1) cells as shown in Figure 1. MC holds as there are more compliers than defiers in each cell.

Y(0) \ Y(1)	0	1
0	C={c1 c2} F={f1}	C={c3 c4 c5} F={f2 f3}
1	C={c6 c7 c8 c9} F={f4 f5}	C={c10 c11 c12 c13} F={f6 f7}

Figure 1: An example in which MC is satisfied.

To construct C_F, it suffices to choose at random

$$\frac{f_{Y_0,Y_1,T}(y_0,y_1,F)}{f_{Y_0,Y_1,T}(y_0,y_1,C)}\%$$

of compliers in each of the four (Y_0,Y_1) strata. This amounts to picking up $\frac{0.05}{0.1}=\frac{1}{2}$ of compliers in the $(Y_0,Y_1)=(0,0)$ stratum, $\frac{0.1}{0.15}=\frac{2}{3}$ of them in the $(Y_0,Y_1)=(0,1)$ stratum, and $\frac{0.1}{0.2}=\frac{1}{2}$ of them in the $(Y_0,Y_1)=(1,0)$ and $(Y_0,Y_1)=(1,1)$ strata. The populations C_F and C_V resulting from this construction are displayed in Figure 2. C_F has both the same size and the same joint distribution of (Y_0,Y_1) as defiers, which ensures that it has the same marginal distributions of Y_0 and Y_1.

Y(0) \ Y(1)	0	1
0	CV={c2} CF={c1} F={f1}	CV={c4} CF={c3 c5} F={f2 f3}
1	CV={c6 c7} CF={c8 c9} F={f4 f5}	CV={c11 c12 } CF={c10 c13} F={f6 f7}

Figure 2: Constructing C_F and C_V in this first example.

But MC is not necessary for Theorem 3.1 to hold. To see it, consider Figure 3. In this second example, MC is not satisfied as there are more defiers than compliers in the $(Y_0,Y_1)=(0,1)$ cell. On the contrary, CD holds. One can for instance set $g(0,0)=g(0,1)=g(1,1)=\frac{3}{25}$ and $g(1,0)=\frac{1}{25}$.

Y(0) \ Y(1)	0	1
0	C={c1 c2 c3 c4 c5} F={f1 f2}	C={c6 c7 c8} F={f3 f4 f5 f6}
1	C={c9 c10 c11} F={f7 f8}	C={c12 c13 c14 c15} F={f9 f10}

Figure 3: An example in which MC is not satisfied.

Then, to construct C_F, it suffices to choose at random

$$\frac{g(y_0, y_1)}{f_{Y_0,Y_1,T}(y_0, y_1, C)}\%$$

of compliers in each of the four (Y_0, Y_1) strata. This amounts to picking up $\frac{3}{5}$ of compliers in the $(Y_0, Y_1) = (0, 0)$ stratum, all of them in the $(Y_0, Y_1) = (0, 1)$ stratum, $\frac{1}{3}$ of them in the $(Y_0, Y_1) = (1, 0)$ stratum, and $\frac{3}{4}$ of them in the $(Y_0, Y_1) = (1, 1)$ stratum. One can check from Figure 4 that C_F has the same size and the same marginal distributions of Y_0 and Y_1 as defiers.

Y(0) \ Y(1)	0	1
0	CV={c2 c4} CF={c1 c3 c5} F={f1 f2}	CV={ø} CF={c6 c7 c8} F={f3 f4 f5 f6}
1	CV={c10 c11} CF={c9} F={f7 f8}	CV={c13} CF={c12 c14 c15} F={f9 f10}

Figure 4: Constructing C_F and C_V in this second example.

4 Who belongs to C_V?

The LATE I exhibit applies to a different population than the one identified in Imbens & Angrist (1994). This C_V population is more abstract, and could potentially be even more specific than the standard population of compliers. Hence the need to be able to describe it, so as to assess whether this LATE is likely to apply to other populations or not. Under

ND, the distribution of any vector of covariates X among compliers is identified. This is a consequence of the κ result in Abadie (2003). A weak strengthening of CD is sufficient for the distribution of X and the effect of the treatment to be identified within a subpopulation a compliers.

Assumption 4.1 *(CD-X)*

There is a real-valued function g defined on $\mathcal{S}(Y_0) \times \mathcal{S}(Y_1) \times \mathcal{S}(X)$ such that almost everywhere:

$$0 \leq \quad g(y_0, y_1, x) \quad \leq f_{Y_0,Y_1,X,T}(y_0, y_1, x, C) \tag{4.1}$$

$$\int_{\mathcal{S}(Y_1)\times\mathcal{S}(X)} g(y_0, y_1, x) \quad d\lambda(y_1, x) = f_{Y_0,T}(y_0, F) \tag{4.2}$$

$$\int_{\mathcal{S}(Y_0)\times\mathcal{S}(X)} g(y_0, y_1, x) \quad d\lambda(y_0, x) = f_{Y_1,T}(y_1, F) \tag{4.3}$$

$$\int_{\mathcal{S}(Y_0)\times\mathcal{S}(Y_1)} g(y_0, y_1, x) \quad d\lambda(y_0, y_1) = f_{X,T}(x, F). \tag{4.4}$$

A sufficient condition for CD-X to hold is that there be more compliers than defiers in each subgroup of the population with the same value of (Y_0, Y_1, X).

CD-X is stronger than CD. If $g(y_0, y_1, x)$ satisfies Equations (4.1), (4.2), and (4.3), then

$$h(y_0, y_1) = \int_{\mathcal{S}(X)} g(y_0, y_1, x) d\lambda(x)$$

satisfies (2.3), (2.4), and (2.5).[7]

CD-X is sufficient to ensure that the distribution of X and the effect of the treatment are identified within a subpopulation a compliers. Let

$$l(x) = \frac{f_{X,D|Z=0}(x, 0) - f_{X,D|Z=1}(x, 0)}{P(D = 0|Z = 0) - P(D = 0|Z = 1)}.$$

Theorem 4.1 *Suppose Assumptions 2.1, 2.2, and 4.1 hold. Then there is a subpopulation*

[7] It is straightforward that h satisfies (2.3). To see that it satisfies for instance (2.4), it suffices to integrate h over $\mathcal{S}(Y_1)$, use the Fubini-Tonelli theorem, and then use Equation (4.2).

of compliers C_V such that:

$$P(C_V) = P(D = 1|Z = 1) - P(D = 1|Z = 0) \tag{4.5}$$

$$f_{Y_0|C_V}(y_0) = f_0(y_0) \tag{4.6}$$

$$f_{Y_1|C_V}(y_1) = f_1(y_1) \tag{4.7}$$

$$E\left[Y_1 - Y_0|C_V\right] = W \tag{4.8}$$

$$f_{X|C_V}(x) = l(x). \tag{4.9}$$

This theorem shows that under CD-X, there is a population C_V for which both treatment effects and the distribution of X are identified. Therefore, one can describe C_V, to assess whether it differs from the remaining part of the population. Under ND, $l(.)$ captures the distribution of X among the entire population of compliers. As a result, the distribution of X for C_V under the CD-X condition is the same as the distribution of X for compliers under the ND condition. Here again, substituting CD-X to ND does not lead to a change in the quantities to be estimated, but it changes their interpretation.

5 Testability

As pointed out in Imbens & Rubin (1997), the standard LATE assumptions - random instrument, exclusion restriction, and ND - have testable implications. Kitagawa (2008) develops the corresponding statistical test. As shown in the next lemma, the alternative LATE assumptions - random instrument, exclusion restriction, and CD - have the same testable implications. Moreover, if Assumptions 2.1 and 2.2 are maintained hypothesis, performing Kitagawa's test amounts to testing WMC.

Lemma 5.1

1. *Assumptions 2.1, 2.2, and 2.6 imply that the following inequalities must hold almost everywhere:*

$$\begin{aligned} f_{Y,D|Z=1}(y_0, 0) &\leq f_{Y,D|Z=0}(y_0, 0) \\ f_{Y,D|Z=0}(y_1, 1) &\leq f_{Y,D|Z=1}(y_1, 1). \end{aligned} \tag{5.1}$$

2. Under Assumptions 2.1 and 2.2, Equation (5.1) *is equivalent to WMC.*

Kitagawa's test is not a consistent test of CD, as some DGPs satisfy WMC but violate CD. To see this, consider the joint distribution of (Y_0, Y_1, D_0, D_1) presented in Figure 5. This joint distribution satisfies WMC, as the sum of each line and each column is smaller for defiers than compliers. Still, it is impossible to construct a function g satisfying Equations (2.3), (2.4), and (2.5). $P(Y_0 = 1, F) = P(Y_1 = 1, F) = 0.2$. To ensure that g satisfies Equations (2.4) and (2.5), we must have

$$g(1,1) + g(1,0) = 0.2 \tag{5.2}$$

and

$$g(1,1) + g(0,1) = 0.2. \tag{5.3}$$

Since $P(Y_0 = 1, Y_1 = 1, C) = 0.05$, $g(1,1)$ must be smaller than 0.05 for Equation (2.3) to hold. As a result, one must set

$$g(0,1) = g(1,0) \geq 0.15.$$

Then,

$$P(Y_0 = 0, F) = P(Y_1 = 0, F) = 0.05.$$

Given $g(0,1)$ and $g(1,0)$, having

$$g(0,0) + g(0,1) = g(0,0) + g(1,0) = 0.05$$

would require setting

$$g(0,0) \leq -0.10,$$

which would violate Equation (2.3).

Y(0) \ Y(1)	0	1
0	P(Y(0)=0,Y(1)=0,C)=0.30 P(Y(0)=0,Y(1)=0,F)=0.00	P(Y(0)=0,Y(1)=1,C)=0.20 P(Y(0)=0,Y(1)=1,F)=0.05
1	P(Y(0)=1,Y(1)=0,C)=0.20 P(Y(0)=1,Y(1)=0,F)=0.05	P(Y(0)=1,Y(1)=1,C)=0.05 P(Y(0)=1,Y(1)=1,F)=0.15

Figure 5: WMC does not imply CD.

As shown in Kitagawa (2008), Equation (5.1) is the strongest testable implication of ND: if it is satisfied in the data, it is possible to rationalize ND. As ND is stronger than MC or CD, Equation (5.1) is also the strongest testable implication of those two assumptions. Therefore, there is no consistent test of ND, MC, or CD.

6 Should we believe in CD? The perspective of an agnostic decision maker

Consider a bayesian decision maker who has to decide whether to believe or not in CD, MC, and ND, in an application where potential outcomes are binary. Let J denote the joint distribution of (Y_0, Y_1, D_0, D_1). As potential outcomes are binary,

$$J = (P(Y_0 = y_0, Y_1 = y_1, D_0 = d_0, D_1 = d_1))_{(y_0,y_1,d_0,d_1)\in\{0;1\}^4}$$

is merely a vector in $[0,1]^{16}$ whose coordinates sum up to 1. Let $\mathcal{J}_{pr}$ denote the set of all such vectors. Before the data is revealed, J can be any element of that set.

The prior distribution of that decision maker for J is the same as the distribution of $U|U \in \mathcal{J}_{pr}$, where U denotes a vector of 16 independent random variables uniformly distributed on $[0,1]$. As it is symmetric, this prior is invariant to relabeling the parameters of the DGP. It is therefore uninformative, both on the size of each type (always takers, never takers, compliers, and defiers), and on their distribution of potential outcomes. This implies, for instance, that before the data reveals the distribution of (Y, D, Z), the decision maker assigns the same weight to the two following statements: "the population comprises 30% of compliers, 20% of defiers, 40% of never takers, and 10% of always takers", and "the population comprises 10% of compliers, 30% of defiers, 20% of never takers, and 40% of always takers." *À priori*, she does not believe that some types are likely to outnumber the others in the population.

Then, the data (asymptotically) reveals the distribution of (Y, D, Z). This imposes some

equality and inequality constraints on the 16 coordinates of J. For instance, we must have

$$\begin{aligned} & P(Y=0,D=0|Z=0) - P(Y=0,D=0|Z=1) \\ = \; & P(Y_0=0,Y_1=0,D_0=0,D_1=1) + P(Y_0=0,Y_1=1,D_0=0,D_1=1) \\ - \; & (P(Y_0=0,Y_1=0,D_0=1,D_1=0) + P(Y_0=0,Y_1=1,D_0=1,D_1=0)). \end{aligned}$$

This reduces the set of potential DGPs to $\mathcal{J}_{po}$, the set of elements in $[0,1]^{16}$ compatible with all those constraints. The posterior distribution of the decision maker for J is the distribution of $U|U \in \mathcal{J}_{po}$.

Let $\mathcal{ND}$, $\mathcal{MC}$, and $\mathcal{CD}$ be the set of joint distributions of (Y_0, Y_1, D_0, D_1) satisfying respectively ND, MC, and CD. If the decision maker's loss function is symmetric in Type 1 and Type 2 error, she will choose to believe, for instance, in CD if and only if $P_{po}(J \in \mathcal{CD}) \geq \frac{1}{2}$. $P_{po}(J \in \mathcal{CD})$, $P_{po}(J \in \mathcal{MC})$, and $P_{po}(J \in \mathcal{ND})$ can be approximated using a Monte-Carlo method, by drawing realizations of vectors U belonging to $\mathcal{J}_{po}$ and by computing the share of them which respectively belong to $\mathcal{CD}$, $\mathcal{MC}$, and $\mathcal{ND}$.[8]

With a finite sample, the data does not reveal the joint distribution of (Y, D, Z). Therefore, we can only estimate $P_{po}(J \in \mathcal{CD})$, $P_{po}(J \in \mathcal{MC})$, and $P_{po}(J \in \mathcal{ND})$, by substituting $(\widehat{P}(Y=y,D=d|Z=z))_{(y,d,z)\in\{0;1\}^3}$ with $(P(Y=y,D=d|Z=z))_{(y,d,z)\in\{0;1\}^3}$ when conducting the Monte-Carlo simulations described above. The resulting estimator is consistent.[9]

In Angrist & Evans (1998), this agnostic decision maker with a symmetric loss function will choose to believe in CD but not in MC or ND. Their main outcome Y is a binary variable for female participation to the labor market. Treatment D is a dummy for having three children or more. The instrument Z is a dummy equal to one when the first two children in a family are of the same sex. Computations are based on the 1980 PUMS sample, which bears 394 840 observations. As shown in Huber (2012), Kitagawa's test is not rejected in this data. Monte-Carlo simulations yield $\widehat{P}_{po}(J \in \mathcal{CD}) \approx 94\%$, $\widehat{P}_{po}(J \in \mathcal{MC}) \approx 32\%$, and

[8] When potential outcomes are binary, it is easy to determine whether a joint distribution of (Y_0, Y_1, D_0, D_1) satisfies CD or not.

[9] This follows from the continuous mapping theorem, once noted that $\mathcal{J}_{po}$ is a continuous function of $(P(Y=y,D=d|Z=z))_{(y,d,z)\in\{0;1\}^3}$, and that $P(U \in A|U \in B)$ is continuous in B by continuity of the cumulative distribution function of U.

$\widehat{P}_{po}(J \in \mathcal{ND}) = 0\%$.[10] As the sample size is large, those estimates can be regarded as fairly precise approximations of the true underlying parameters.

Angrist & Evans (1998) is not an exception: for most distributions of observables

$$(P(Y = y, D = d|Z = z))_{(y,d,z)\in\{0;1\}^3}$$

satisfying WMC, a very large share of the corresponding set of admissible DGPs also satisfies CD. To show this, I draw 200 distributions of observables from uniform distributions, imposing that

$$P(Y = y, D = d|Z = d) \geq P(Y = y, D = d|Z = 1 - d),$$

for every $(y, d) \in \{0; 1\} \times \{0; 1\}$,[11] and

$$\begin{aligned} & P(Y = 0, D = 0|Z = z) + P(Y = 1, D = 0|Z = z) \\ + \quad & P(Y = 0, D = 1|Z = z) + P(Y = 1, D = 1|Z = z) \end{aligned}$$

be equal to 1 for every $z \in \{0; 1\}$. For each of those distributions of observables, the corresponding value of $P_{po}(J \in \mathcal{CD})$ is approximated using a Monte-Carlo simulation. The average value of $P_{po}(J \in \mathcal{CD})$ is 81%, while its median is 88%. $P_{po}(J \in \mathcal{CD})$ is greater than 50% in 90% of cases. It sometimes takes very low values (the lowest value of $P_{po}(J \in \mathcal{CD})$ is 0%), but only when

$$P(Y = y, D = d|Z = d) - P(Y = y, D = d|Z = 1 - d)$$

is very low for some value of (y, d), meaning that the WMC assumption is close to being violated. Among vectors such that for every (y, d)

$$P(Y = y, D = d|Z = d) - P(Y = y, D = d|Z = 1 - d) \geq 0.02,$$

the lowest value of $P_{po}(J \in \mathcal{CD})$ is 41%.

When interested in a binary outcome, applied researchers can run simulations similar to those I conducted to assess how credible CD is in the application they consider.[12] This

[10] This is merely because ND constrains 4 coordinates of J to be equal to 0. As a result, $\mathcal{ND}$ is of a strictly lower dimension than $\mathcal{J}_{po}$, which implies that the Lebesgue measure of $\mathcal{ND}$ relative to $\mathcal{J}_{po}$ is 0.

[11] This ensures that WMC is satisfied.

[12] The Stata code is available upon request.

bayesian analysis could be extended to multivariate outcomes with finite support, even though the simulations will become more cumbersome as the support of the outcome grows. It would be more challenging to extend this type of analysis to outcomes with infinite support.

As per this uniform prior, DGPs satisfying ND will always account for 0% of DGP compatible with the data. This does not mean ND is always an incredible assumption. In some applications, economic theory gives strong arguments supporting ND. In such cases, it might not be insightful to adopt the perspective of an agnostic decision maker with an uninformative prior, as we have good reasons not to be agnostic. The following section reviews the standard structural interpretation of ND, and derives structural interpretations of the MC and WMC conditions, to discuss how those assumptions compare from the perspective of economic theory.

7 A structural interpretation of the MC and WMC conditions

The CD condition does not arise from a structural model in which subjects make treatment decisions after solving a utility maximization problem. Embedding this condition in such a model would be useful, to assess whether it is significantly weaker than ND not only from the perspective of an agnostic decision maker, but also from the perspective of economic theory. Nonetheless, giving a structural interpretation to the CD condition is not easy. Instead, I give structural interpretations of both the MC and WMC conditions. As CD stands in between the two, its structural interpretation will also stand in between the interpretation of MC and WMC.

Without any loss of generality, one can always write

$$D_z = 1\{V_z \geq v_z\}, \tag{7.1}$$

with $v_0 \geq v_1$. This can be interpreted as a rational choice model: when $Z = z$, a subject gets treated if and only if the benefit she expects from treatment (V_z) is greater than its cost (v_z).

Then, let benefit-compliers (BC) satisfy

$$\{V_0 < v_1, V_1 \geq v_0\},$$

while cost-compliers (CC) satisfy

$$\{(V_0, V_1) \in [v_1, v_0)^2\}.$$

BC and CC define a partition of compliers. Benefit compliers would comply even if the instrument had no impact on the cost of treatment, i.e. if $v_1 = v_0$. Cost compliers are the remaining part of compliers.

As shown in Vytlacil (2002), the restriction ND imposes on the fully general model presented in Equation (7.1) is

Assumption 7.1 *(Single Index)*

$$V_0 = V_1 = V. \tag{7.2}$$

Under this restriction, the fully general model becomes a single index model. Compliers can only be cost compliers, and they satisfy $V \in [v_1, v_0)$. They are "marginal" subjects: their expected benefit from treatment is too low for them to get treated if they do not receive an incentive, but high enough for them to get treated if they do receive one.

Following the rational choice interpretation outlined above, single index amounts to assuming that the instrument has no effect on the expected benefit from treatment. This might not always be credible. For instance, in randomized controlled trials relying on an encouragement design, the encouragement group typically receives a flyer describing the treatment, and a financial incentive to get treated. As the flyer conveys information on the treatment, it might change the benefits subjects expect from it.

MC imposes a weaker restriction than ND on the fully general model. Consider the following condition:

Assumption 7.2 *(Asymmetric indexes: AI)*
For every $h \geq l$,

$$P(V_0 < l, V_1 \geq h | Y_0, Y_1) \geq P(V_0 \geq h, V_1 < l | Y_0, Y_1) \tag{7.3}$$

almost surely.

In the encouragement design example, the AI condition means that even though the flyer might decrease the benefits that some subjects expect from treatment, it is still appealing enough to ensure that in each (Y_0, Y_1) subgroup there are at least as many subjects whose expected benefit from treatment increases from l to h after reading it, as there are subjects whose expected benefit decreases from h to l.

As shown in the next theorem, the AI condition is sufficient for MC to hold, but it does not ensure that the C_V population can receive a simple structural interpretation. This is achieved under the following more restrictive condition:

Assumption 7.3 *(Symmetric indexes: SI)*
For every $h \geq l$,

$$P(V_0 < l, V_1 \geq h | Y_0, Y_1) = P(V_0 \geq h, V_1 < l | Y_0, Y_1) \tag{7.4}$$

almost surely.

In the encouragement design example, the SI condition requires that there be exactly as many subjects whose expected benefit from treatment increases from l to h after reading the flyer, as there as subjects for whom it decreases from h to l.

Theorem 7.1 *Suppose potential treatments are generated following Equation* (7.1).

1. *If Assumption 7.2 holds, then Assumption 2.4 is satisfied.*
2. *If Assumption 7.3 holds, then CC satisfies Equations* (3.1), (3.2), (3.3), *and* (3.4).
3. *If Assumption 2.4 holds, one can construct*

$$(V_0^*, V_1^*, v_0^*, v_1^*)$$

such that Equation (7.1) *holds and Assumption 7.2 is satisfied.*

The first point of the theorem shows that if the true DGP of potential treatments satisfies the AI condition, then MC holds. The second part shows that if this DGP satisfies the more restrictive SI condition, then the C_V population for which treatment effects are identified corresponds to cost-compliers. Finally, the third part shows that if MC holds, one can

construct a DGP satisfying the AI condition. Therefore, the AI, SI, and MC conditions are observationally equivalent. This generalizes Vytlacil (2002).

Finally, I sketch a structural interpretation of the CD condition. As discussed above, CD stands in between MC and WMC. One can show that WMC is observationally equivalent to a weaker AI condition: for every $h \geq l$,

$$\begin{aligned} P(V_0 < l, V_1 \geq h|Y_0) \geq P(V_0 \geq h, V_1 < l|Y_0) \\ P(V_0 < l, V_1 \geq h|Y_1) \geq P(V_0 \geq h, V_1 < l|Y_1), \end{aligned} \tag{7.5}$$

almost surely. As a result, CD is weaker than imposing Equation (7.3), but it is stronger than imposing only Equation (7.5).

8 Extensions

8.1 Extension to IV with multivariate treatment and instrument

Results on average effects presented in this paper extend to applications where treatment is multivariate. Assume that for every $z \in \{0; 1\}$, $D_z \in \{0, 1, 2, ..., J\}$ for some integer J. Let $(Y_j)_{j\in\{0,1,2,...,J\}}$ denote the corresponding potential outcomes. Assume the following MC type of condition is verified: for every $j \in \{1, 2, ..., J\}$,

$$P(D_0 \geq j > D_1|Y_j, Y_{j-1}) \leq P(D_1 \geq j > D_0|Y_j, Y_{j-1}).$$

This condition means that in each (Y_j, Y_{j-1}) subgroup, there are less subjects who are induced to go from a treatment greater than j to a treatment strictly lower than j because of the instrument, than subjects who are induced to go from strictly below to above j. Then, the Wald ratio captures the following average causal response:

$$\sum_{j=1}^{J} w_j E(Y_j - Y_{j-1}|C_V^j).$$

C_V^j is a subset of the population satisfying $\{D_1 \geq j > D_0\}$ of size $P(D_1 \geq j > D_0) - P(D_0 \geq j > D_1)$, and

$$w_j = \frac{P(D_1 \geq j > D_0) - P(D_0 \geq j > D_1)}{\sum_{j=1}^{J} P(D_1 \geq j > D_0) - P(D_0 \geq j > D_1)}$$

are positive weights summing up to 1. This generalizes Theorem 1 in Angrist & Imbens (1995).

Their Theorem 2 states that with a multivariate instrument $Z \in \{0, 1, 2, ..., K\}$, if $D_{z'} \geq D_z$ for every $z' \geq z$, then the Wald ratio captures a weighted average of average causal responses. This result can also be generalized under the following MC condition:[13] for every $0 \leq z < z' \leq K$ and for every $j \in \{0, 1, 2, ..., J\}$,

$$P(D_z \geq j > D_{z'}|Y_j, Y_{j-1}) \leq P(D_{z'} \geq j > D_z|Y_j, Y_{j-1}).$$

8.2 Extension to other microeconometrics models

My results also extend to many treatment effect models relying on ND type of conditions. An important example is the fuzzy regression discontinuity (RD) model studied in Hahn et al. (2001). Their Theorem 3 still holds if their Assumption A.3, a "ND at the threshold" assumption, is replaced for instance by a weaker "MC at the threshold" condition: almost surely

$$P(F|Y_0, Y_1, S = s) \leq P(C|Y_0, Y_1, S = s),$$

where S denotes the forcing variable, and s is the cut-off.

Finally, ND type of conditions have also been used in the sample selection and survey non-response literatures, for instance in Lee (2009) or Behaghel et al. (2012). Let R_0 and R_1 denote a subject's potential responses to a survey without and with the treatment. Under the assumption that there are no "response-defiers" ($R_1 \geq R_0$), Lee (2009) derives sharp bounds for $E(Y_1 - Y_0|R_0 = 1, R_1 = 1)$, the average effect of the treatment among always-respondents (AR).

Let RC and RF respectively denote "reponse-compliers" (subjects who satisfy $R_0 = 0$ and $R_1 = 1$) and "response-defiers" (subjects who satisfy $R_0 = 1$ and $R_1 = 0$). Under the assumption that there are more response-compliers than response-defiers conditional on (Y_0, Y_1), or under a weaker CD type of assumption, one can show that the bounds derived in Lee (2009) are valid bounds for $E(Y_1 - Y_0|AR \cup RC_F)$, where RC_F denotes a

[13] Theorem 1 and 2 in Angrist & Imbens (1995) could actually be generalized under even weaker CD type of conditions.

subpopulation of response compliers with the same marginal distributions of Y_0 and Y_1 as response-defiers.

9 Concluding comments

In the IV model with heterogeneous effects, the "no-defiers" condition can be replaced by a weaker "compliers-defiers" condition, while maintaining the interpretation of IV estimates as local effects. Like "no-defiers", "compliers-defiers" can be partly tested, through a procedure developed in Kitagawa (2008). Few DGPs violate "compliers-defiers" and satisfy the null hypothesis in Kitagawa's test. As a result, applied researchers should conduct this test whenever they suspect there could be defiers in their application. If this test is not rejected and the outcome they consider is binary, they can also conduct Monte-Carlo simulations to assess which share of DGPs compatible with their data satisfy the CD assumption. This will enable them to assess whether "compliers-defiers" is a strong or a weak assumption in their application, and whether they can be highly confident or not in interpreting their results as local effects.

References

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

Angrist, J. D. & Evans, W. N. (1998), 'Children and their parents' labor supply: Evidence from exogenous variation in family size', *American Economic Review* **88**(3), 450–77.

Angrist, J. D. & Imbens, G. W. (1995), 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American Statistical Association* **90**(430), pp. 431–442.

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.

Bedard, K. & Dhuey, E. (2006), 'The persistence of early childhood maturity: International evidence of long-run age effects', *The Quarterly Journal of Economics* **121**(4), 1437–1472.

Behaghel, L., Crépon, B., Gurgand, M. & Le Barbanchon, T. (2012), Please call again: Correcting non-response bias in treatment effect models, IZA Discussion Papers 6751, Institute for the Study of Labor (IZA).

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

DiNardo, J. & Lee, D. S. (2011), *Program Evaluation and Research Designs*, Vol. 4 of *Handbook of Labor Economics*, Elsevier, chapter 5, pp. 463–536.

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.

Hoderlein, S. & Gautier, E. (2012), Estimating treatment effects with random coefficients in the selection equation, Technical report.

Huber, M. (2012), Statistical verification of a natural "natural experiment": Tests and sensitivity checks for the sibling sex ratio instrument, Economics Working Paper Series 1219, University of St. Gallen, School of Economics and Political Science.

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).

Kling, J. R. (2006), 'Incarceration length, employment, and earnings', *American Economic Review* **96**(3), 863–876.

Lee, D. S. (2009), 'Training, wages, and sample selection: Estimating sharp bounds on treatment effects', *Review of Economic Studies* **76**(3), 1071–1102.

Small, D. & Tan, Z. (2007), A stochastic monotonicity assumption for the instrumental variables method, Working paper, department of statistics, wharton school, university of pennsylvania.

Vytlacil, E. (2002), 'Independence, monotonicity, and latent index models: An equivalence result', *Econometrica* **70**(1), 331–341.

A Proofs

Throughout the proofs, it is assumed that (Y_0, Y_1) takes values in an arbitrary measurable space $\mathcal{S}(Y_0) \times \mathcal{S}(Y_1)$.

The proof of Theorem 3.1 relies on the following lemma:

Lemma A.1 *There is a subpopulation of compliers C_F such that:*

$$P(C_F) = P(F) \tag{A.1}$$

$$f_{Y_0|C_F}(y_0) = f_{Y_0|F}(y_0) \tag{A.2}$$

$$f_{Y_1|C_F}(y_1) = f_{Y_1|F}(y_1) \tag{A.3}$$

if and only if Assumption 2.6 holds.

Proof of Lemma A.1:

Assume that Assumption 2.6 holds. Consider a function g which satisfies Equations (2.3), (2.4), and (2.5). Densities being uniquely defined up to 0 probability sets, I can assume without loss of generality that those three equations hold everywhere. Then, let

$$p(y_0, y_1) = \frac{g(y_0, y_1)}{f_{Y_0,Y_1,T}(y_0, y_1, C)} 1\{f_{Y_0,Y_1,T}(y_0, y_1, C) > 0\}.$$

Equation (2.3) ensures that this ratio is always included between 0 and 1. Then, let B be a Bernoulli random variable such that $P(B = 1|Y_0 = y_0, Y_1 = y_1, C) = p(y_0, y_1)$. Finally, let $C_F = \{C, B = 1\}$.

For every $(y_0, y_1) \in \mathcal{S}(Y_0) \times \mathcal{S}(Y_1)$,

$$\begin{aligned}
& f_{Y_0,Y_1,1_{C_F}}(y_0, y_1, 1) \\
= \; & f_{Y_0,Y_1,T,B}(y_0, y_1, C, 1) \\
= \; & f_{Y_0,Y_1,T}(y_0, y_1, C) P(B = 1|Y_0 = y_0, Y_1 = y_1, C) \\
= \; & f_{Y_0,Y_1,T}(y_0, y_1, C) \frac{g(y_0, y_1)}{f_{Y_0,Y_1,T}(y_0, y_1, C)} 1\{f_{Y_0,Y_1,T}(y_0, y_1, C) > 0\} \\
= \; & g(y_0, y_1).
\end{aligned} \tag{A.4}$$

The first equality arises from the definition of C_F. The second follows from Bayes rule. The third follows from the definition of B. The fourth holds because under Equation (2.3),

$$f_{Y_0,Y_1,T}(y_0, y_1, C) = 0 \Rightarrow g(y_0, y_1) = 0.$$

Therefore,

$$\begin{aligned} f_{Y_0,1_{C_F}}(y_0,1) &= \int_{\mathcal{S}(Y_1)} f_{Y_0,Y_1,1_{C_F}}(y_0,y_1,1)d\lambda(y_1) \\ &= \int_{\mathcal{S}(Y_1)} g(y_0,y_1)d\lambda(y_1) \\ &= f_{Y_0,T}(y_0,F). \end{aligned} \tag{A.5}$$

One can similarly show that

$$f_{Y_1,1_{C_F}}(y_1,1) = f_{Y_1,T}(y_1,F). \tag{A.6}$$

Integrating (A.5) over $\mathcal{S}(Y_0)$ yields (A.1). Dividing (A.5) and (A.6) by $P(C_F) = P(F)$ yields (A.2) and (A.3).

To prove the converse statement, it suffices to notice that if a subpopulation of compliers C_F satisfies Equations (A.1), (A.2), and (A.3), then $f_{Y_0,Y_1,1_{C_F}}$ must satisfy Equations (2.3), (2.4), and (2.5).

QED.

Proof of Theorem 3.1:

First, notice that a subgroup of compliers C_V satisfies Equations (3.1), (3.2) , (3.3), and (3.4) if and only if it satisfies

$$f_{Y_0,1_{C_V}}(y_0,1) = f_{Y,D|Z=0}(y_0,0) - f_{Y,D|Z=1}(y_0,0) \tag{A.7}$$

$$f_{Y_1,1_{C_V}}(y_1,1) = f_{Y,D|Z=1}(y_1,1) - f_{Y,D|Z=0}(y_1,1). \tag{A.8}$$

Similarly, as shown in the proof of Lemma A.1, a subgroup of compliers C_F satisfies Equations (A.1), (A.2) , and (A.3) if and only if it satisfies (A.5) and (A.6).

Combining those two remarks with Lemma A.1 shows that to prove Theorem 3.1, it suffices to prove that there is a subgroup of compliers C_V satisfying Equations (A.7) and (A.8) if and only if there is a subgroup of compliers C_F satisfying Equations (A.5) and (A.6).

Then, notice also that

$$\begin{aligned} & f_{Y,D|Z=0}(y_0,0) - f_{Y,D|Z=1}(y_0,0) \\ = \; & f_{Y_0,D_0}(y_0,0) - f_{Y_0,D_1}(y_0,0) \\ = \; & f_{Y_0,D_0,D_1}(y_0,0,0) + f_{Y_0,D_0,D_1}(y_0,0,1) \\ - \; & f_{Y_0,D_0,D_1}(y_0,0,0) - f_{Y_0,D_0,D_1}(y_0,1,0) \\ = \; & f_{Y_0,T}(y_0,C) - f_{Y_0,T}(y_0,F). \end{aligned} \tag{A.9}$$

The first equality follows from Assumptions 2.1 and 2.2, the second one follows from the law of total probabilities. Similarly one can show that

$$f_{Y,D|Z=1}(y_1,1) - f_{Y,D|Z=0}(y_1,1) = f_{Y_1,T}(y_1,C) - f_{Y_1,T}(y_1,F). \tag{A.10}$$

Now, assume that a subgroup of compliers C_F satisfies Equations (A.5) and (A.6). Let $C_V = C \setminus C_F$.

$$\begin{aligned} f_{Y_0,1_{C_V}}(y_0,1) &= f_{Y_0,T}(y_0,C) - f_{Y_0,1_{C_F}}(y_0,1) \\ &= f_{Y_0,T}(y_0,C) - f_{Y_0,T}(y_0,F) \\ &= f_{Y,D|Z=0}(y_0,0) - f_{Y,D|Z=1}(y_0,0). \end{aligned}$$

The first equality follows from the law of total probabilities, the second one follows from C_F satisfying Equation (A.5), the last one follows from Equation (A.9). Similarly, one can show that

$$f_{Y_1,1_{C_V}}(y_1,1) = f_{Y,D|Z=1}(y_1,1) - f_{Y,D|Z=0}(y_1,1).$$

This proves that C_V satisfies Equations (A.7) and (A.8).

Conversely, assume that a subgroup of compliers C_V satisfies Equations (A.7) and (A.8). Let $C_F = C \setminus C_V$.

$$\begin{aligned} f_{Y_0,1_{C_F}}(y_0,1) &= f_{Y_0,T}(y_0,C) - f_{Y_0,1_{C_V}}(y_0,1) \\ &= f_{Y_0,T}(y_0,C) - \big(f_{Y,D|Z=0}(y_0,0) - f_{Y,D|Z=1}(y_0,0)\big) \\ &= f_{Y_0,T}(y_0,F). \end{aligned}$$

The first equality follows from the law of total probabilities, the second one follows from C_V satisfying Equation (A.7), the last one follows from Equation (A.9). Similarly, one can

show that

$$f_{Y_1,1_{C_F}}(y_1,1) = f_{Y_1,T}(y_1,F).$$

This proves that C_F satisfies Equations (A.5) and (A.6).

QED.

Proof of Theorem 4.1:

I sktech this proof only, as it is very similar to the proof of Theorem 3.1. Consider a function g satisying Equations (4.1), (4.2), (4.3), and (4.4). Let

$$q(y_0,y_1,x) = \frac{g(y_0,y_1,x)}{f_{Y_0,Y_1,X,T}(y_0,y_1,x,C)} 1\{f_{Y_0,Y_1,X,T}(y_0,y_1,x,C) > 0\}.$$

Equation (4.1) ensures that this ratio is always included between 0 and 1. Then, let B denote a Bernoulli variable such that $P(B=1|Y_0=y_0,Y_1=y_1,X=x,C) = q(y_0,y_1,x)$. Finally, let $C_F = \{C, B=1\}$.

Following the same steps as for the proof of Lemma A.1, one can show that

$$\begin{aligned}
P(C_F) &= P(F) \\
f_{Y_0|C_F}(y_0) &= f_{Y_0|F}(y_0) \\
f_{Y_1|C_F}(y_1) &= f_{Y_1|F}(y_1) \\
f_{X|C_F}(x) &= f_{X|F}(x).
\end{aligned}$$

Following the same steps as for the proof of Theorem 3.1, one can then show that $C_V = C \setminus C_F$ satisfies Equations (4.5), (4.6), (4.7), (4.8), and (4.9).

QED.

Proof of Lemma 5.1

The first point of the Lemma has already been proven elsewhere (see for instance Proposition 1 in Kitagawa (2008)). I prove only the second point. Under Assumptions 2.1 and 2.2,

$$\begin{aligned}
f_{Y,D|Z=0}(y_0,0) - f_{Y,D|Z=1}(y_0,0) &= f_{Y_0,T}(y_0,C) - f_{Y_0,T}(y_0,F) \\
f_{Y,D|Z=1}(y_1,1) - f_{Y,D|Z=0}(y_1,1) &= f_{Y_1,T}(y_1,C) - f_{Y_1,T}(y_1,F)
\end{aligned}$$

as shown above. Therefore, under those two assumptions (5.1) is equivalent to

$$\begin{aligned}
f_{Y_0,T}(y_0,F) \leq f_{Y_0,T}(y_0,C) \\
f_{Y_1,T}(y_1,F) \leq f_{Y_1,T}(y_1,C).
\end{aligned}$$

It suffices to divide the first inequality by $f_{Y_0}(y_0)$ and the second one by $f_{Y_1}(y_1)$ to see that they are equivalent to

$$P(F|Y_0 = y_0) \leq P(C|Y_0 = y_0)$$
$$P(F|Y_1 = y_1) \leq P(C|Y_1 = y_1).$$

QED.

Proof of Theorem 7.1

Proof of 1

Suppose Assumption 7.2 holds. For every $(y_0, y_1) \in \mathcal{S}(Y_0) \times \mathcal{S}(Y_1)$,

$$\begin{aligned} P(F|Y_0 = y_0, Y_1 = y_1) &= P(V_0 \geq v_0, V_1 < v_1|Y_0 = y_0, Y_1 = y_1) \\ &\leq P(V_0 < v_1, V_1 \geq v_0|Y_0 = y_0, Y_1 = y_1) \\ &\leq P(V_0 < v_0, V_1 \geq v_1|Y_0 = y_0, Y_1 = y_1) \\ &= P(C|Y_0 = y_0, Y_1 = y_1). \end{aligned}$$

The first and last equality follow from Equation (7.1). The first inequality follows from the AI condition, as $v_0 \geq v_1$. The second one follows from the fact that $\{V_0 < v_1, V_1 \geq v_0\} \subseteq \{V_0 < v_0, V_1 \geq v_1\}$, once again because $v_0 \geq v_1$.

Proof of 2

Suppose Assumption 7.3 holds. For every $(y_0, y_1) \in \mathcal{S}(Y_0) \times \mathcal{S}(Y_1)$,

$$\begin{aligned} P(BC|Y_0 = y_0, Y_1 = y_1) &= P(V_0 < v_1, V_1 \geq v_0|Y_0 = y_0, Y_1 = y_1) \\ &= P(V_0 \geq v_0, V_1 < v_0|Y_0 = y_0, Y_1 = y_1) \\ &= P(F|Y_0 = y_0, Y_1 = y_1). \end{aligned}$$

The first equality follows from the definition of benefit-compliers. The second one follows from the SI condition. The last one follows from Equation (7.1).

Multiplying both sides of the last equality by $f_{Y_0,Y_1}(y_0, y_1)$ and integrating it over $\mathcal{S}(Y_1)$ implies that BC satisfies Equation (A.5). One can similarly show that BC satisfies Equation (A.6). Therefore, BC satisfies Equations (A.1), (A.2), and (A.3). Using the same steps as for the proof of Theorem 3.1, one can then show that CC satisfies Equations (3.1), (3.2), (3.3), and (3.4).

Proof of 3

The proof is a generalization of the proof of the first point of Proposition 2.1 in Chaisemartin & D'Haultfœuille (2012). Assume that Assumption 2.4 holds. Let

$$r(y_0, y_1) = \frac{P(F|Y_0 = y_0, Y_1 = y_1)}{P(C|Y_0 = y_0, Y_1 = y_1)} 1\{P(C|Y_0 = y_0, Y_1 = y_1) > 0\}.$$

$r(y_0, y_1)$ is smaller than 1 under MC. Let B denote a Bernoulli random variable such that $P(B = 1|Y_0 = y_0, Y_1 = y_1, C) = r(y_0, y_1)$.

Let $v_z = P(D_z = 0)$. Let $t_0 = 0, t_1 = P(D_0 = 0, D_1 = 0), t_2 = P(D_1 = 0), t_3 = P(D_0 = 0)$, $t_4 = 1 - P(D_0 = 1, D_1 = 1)$, and $t_5 = 1$. One can check that $t_0 \leq t_1 \leq t_2 \leq t_3 \leq t_4 \leq t_5$. Then, consider $(W_0, ..., W_4)$, mutually independent random variables, also independent of (Y_0, Y_1, D_0, D_1, B). Take W_i uniform on $[t_i, t_{i+1}]$. Finally, let

$$\begin{aligned} V_0 &= (1 - D_0)(1 - D_1)W_0 + (1 - D_0)D_1(BW_1 + (1 - B)W_2) \\ &+ D_0(1 - D_1)W_3 + D_0D_1W_4 \\ V_1 &= (1 - D_0)(1 - D_1)W_0 + (1 - D_0)D_1(BW_3 + (1 - B)W_2) \\ &+ D_0(1 - D_1)W_1 + D_0D_1W_4. \end{aligned}$$

By construction, $D_z = 1\{V_z \leq v_z\}$ almost surely. Now, I am going to prove that (V_0, V_1) are exchangeable conditional on (Y_0, Y_1), which will ensure that both SI and AI are verified. This amounts to checking that for every $v \leq v'$,

$$f_{V_0,V_1|Y_0,Y_1}(v, v'|y_0, y_1) = f_{V_0,V_1|Y_0,Y_1}(v', v|y_0, y_1).$$

This equality is obviously true when $v = v'$. When $v < v'$,

$$f_{V_0,V_1|Y_0,Y_1}(v, v'|y_0, y_1) = f_{V_0,V_1|Y_0,Y_1}(v', v|y_0, y_1) = 0,$$

except when $(v, v') \in [t_1, t_2] \times [t_3, t_4]$. Let $(v, v') \in [t_1, t_2] \times [t_3, t_4]$.

$$\begin{aligned} & f_{V_0,V_1|Y_0,Y_1}(v, v'|y_0, y_1) \\ = \; & P(C|Y_0 = y_0, Y_1 = y_1)P(B = 1|Y_0 = y_0, Y_1 = y_1, C)f_{W_1}(v)f_{W_3}(v') \\ = \; & P(F|Y_0 = y_0, Y_1 = y_1)f_{W_1}(v)f_{W_3}(v') \\ = \; & f_{V_0,V_1|Y_0,Y_1}(v', v|y_0, y_1). \end{aligned}$$

The first equality follows from the definition of V_0 and V_1 and from Bayes rule. Indeed, one can check that observations verifying $(V_0, V_1) \in [t_1, t_2] \times [t_3, t_4]$ must be compliers such that $B = 1$. The last equality also follows from the definition of V_0 and V_1: observations verifying $(V_1, V_0) \in [t_1, t_2] \times [t_3, t_4]$ must be defiers. This proves the result.

QED.