All you need is LATE
Primary Job Market Paper *

Clément de Chaisemartin†

November 2012

Abstract

Instrumental variable (IV) is a standard tool to measure the effect of a treatment. However, it relies on a strong "no-defiers" condition, and it captures the effect of the treatment for compliers only. This paper shows that "no-defiers" can be replaced by a weaker condition, which requires only that conditional on their treatment effects, more subjects move in the compliant direction than otherwise. This weaker condition is sufficient to capture causal effects for a subgroup of same size as the standard population of compliers. It also shows that the effect of the treatment is the same for compliers as for a larger population G. As a result, IV also captures treatment effects for G. The size of G can be bounded. Those two results are used to reanalyze the internal and external validity of the estimates in Angrist & Evans (1998).

Keywords: instrumental variable, monotonicity, defiers, external validity.
JEL Codes: C21, C26

*This paper benefited from numerous discussions with Josh Angrist and Xavier D'Haultfoeuille. I am also very grateful to Alberto Abadie, Luc Behaghel, Stéphane Bonhomme, Laurent Davezies, Julien Grenet, Marc Gurgand, Martin Huber, Tobias Klein, Francis Kramarz, Kevin Lang, Eric Maurin, Giovanni Mellace and Edward Vytlacil for their helpful comments. I gratefully acknowledge financial support from région Ile de France.

†CREST and Paris School of Economics, clement.dechaisemartin@ensae.fr

1 Introduction

Instrumental variable (IV) is a standard tool to measure the effect of a treatment. However, IV relies on a strong "no-defiers" 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. Moreover, one usually wants to know the effect of the treatment within the entire population, but IV can identify only the effect for compliers (see Heckman & Urzúa, 2010 or Deaton, 2009).

This paper addresses those two limitations. First, it shows that the aforementioned monotonicity assumption can be replaced by a substantially weaker "more compliers than defiers" assumption, while leaving all the standard LATE theorems almost unchanged. Second, it shows that compliers have the same distribution of potential outcomes as a larger population G. As a result, IV estimates can be extrapolated to a population larger than compliers. The size of G can be bounded.

The first contribution of this paper is to weaken the "no-defiers" condition.

Imbens & Angrist (1994) and Angrist et al. (1996) show that instruments identify causal effects under three assumptions: random assignment, exclusion restriction, and monotonicity. Monotonicity means that the instrument will move all subjects in the sample in the same direction. Without monotonicity, IV estimates may be misleading; even in a world where all effects are positive, the IV estimand need not be positive.

Nevertheless, monotonicity 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. 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 girl, but might decide otherwise if their first two children are boys. Such parents would be defiers.

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, monotonicity 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 then 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". Month of birth has also an impact on grade repeating and grade skipping. Children born in December are the least mature of their cohort when they start school, and some of them repeat a grade. Conversely, children born in January are the most mature of their cohort when they start school, and some of them skip a grade.

This leaves us with three populations of defiers: children redshirted because they were born in December, children repeating a grade because they were born in December, and children skipping a grade because they were born in January. Those children would have taken national tests at a younger age had they been born in January. Likewise, children who are not redshirted are compliers, just as those who are always redshirted irrespective of when they are born. Compliers also include children who never skip or repeat a grade, those who always skip, and the ones who always repeat.

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. Monotonicity 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.

This paper shows that the Wald ratio still identifies a local average treatment effect (LATE) if monotonicity is replaced by a weaker condition. This condition states that more subjects are compliers than defiers in each stratum of the population with the same treatment effect $Y(1) - Y(0)$. In such instances, compliers can be partitioned into two subpopulations. The first subpopulation has the same size and the same average treatment effect as defiers, therefore I call them "comfiers" to emphasize the parallel. Treatment effects among comfiers and defiers cancel one another out in the Wald ratio, which finally identifies the average treatment effect among the remaining part of compliers. This second subpopulation consist in compliers who "out-survive" defiers, and accordingly I call them "comvivors". Comvivors have the same size as the population of compliers under monotonicity; substituting "more

compliers than defiers" to monotonicity does not lead to a loss in terms of external validity.

When there are defiers, Angrist et al. (1996) show that the Wald ratio still identifies a LATE if treatment effects are the same among defiers and compliers. Monotonicity and equal treatment effects among compliers and defiers are polar cases of this "more compliers than defiers" condition. Indeed, "more compliers than defiers" will hold if there are no defiers, or if the distribution of treatment effects is the same among defiers and compliers. In between those two polar cases, this condition will also hold in many intermediate cases, provided there are not too many defiers, or treatment effects are not too different among compliers and defiers.

This first "more compliers condition" ensures that the Wald ratio identifies a LATE. Notwithstanding, it is not sufficient to ensure that causal effects on the distribution of the outcome (quantile treatment effects: QTE) are identified for a subgroup, while QTE are identified for compliers under monotonicity. To that end, I introduce a stronger condition. It requires that there be more compliers than defiers in each stratum of the population with same potential outcomes $(Y(0), Y(1))$. I call this second condition the "strong more compliers than defiers condition" (SMC), while I call the first one the "weak more compliers than defiers condition" (WMC). As discussed below, SMC is a particular case of a condition studied by Small & Tan (2007).

The SMC condition has testable implications. The test derived in Kitagawa (2008) amounts to testing whether more subjects are compliers in each stratum with the same $Y(0)$, and in each stratum with the same $Y(1)$. This is close to testing SMC.

I shall now argue that the MC conditions are realistic in the first two examples outlined above. Whether they are also credible in the encouragement design example depends on the purpose of the flyer, as discussed subsequently.

In Angrist & Evans (1998), WMC seems plausible. In this paper, compliers are parents with a preference for diversity; defiers are parents who prefer boys or girls. Mothers participating in the labor market whether they have a third child or not are the two potential outcomes. Treatment effect only defines three strata: mothers such that $Y(1) - Y(0) = -1$, those such that $Y(1) - Y(0) = 0$, and those such that $Y(1) - Y(0) = 1$. WMC states that in each of these three strata, more parents have a preference for diversity than a preference for either sex. This seems realistic. When their first two children are of the same sex, more parents have a third child than when they are of a different sex. Accordingly, in the total population, more parents have a preference for diversity than a preference for either sex. Sex bias is probably correlated to some extent to labor market participation and treatment effect. For example, Asian parents might be more sex-biased and might also have different

labor market participation rates. However, for WMC not to hold, there needs to be one of the three $Y(1) - Y(0)$ strata with more sex-biased than diversity-loving parents. This seems unlikely.

In Bedard & Dhuey (2006), it is even clearer that SMC is likely to hold. Conditional on test scores around the median, few children are redshirted, and few children repeat or skip a grade. As a result, there are probably fewer defiers than compliers in this area of the distribution of test scores. Conditional on high test scores, there are probably fewer children skipping a grade because they were born in January than children skipping a grade irrespective of date of birth. Lastly, conditional on low test scores, there are probably fewer children redshirted or repeating a grade because they were born in December than children redshirted or repeating irrespective of when they were born. Thus, arguing that there are more compliers than defiers over the whole range of test scores seems credible.

In randomized experiments with an encouragement design, whether WMC is credible or not, depends on the purpose of the flyer. WMC is not credible if the objective of the flyer is to reveal to each subject the true benefit she can expect from treatment (provided this is possible). In such instances, there will probably be more defiers than compliers among subjects with low gains from treatment. Nevertheless, in most encouragement designs, the purpose of the flyer is to convey neutral information about the treatment, or even to advertise it in order to create the largest possible differential in take-up rates across groups. In such cases, WMC appears credible. In other words, the flyer is sufficiently well designed to ensure that in each $(Y(1) - Y(0))$ cell, at most, as many subjects have a worse than a better opinion of the treatment after reading it.

The structural interpretation of the Wald ratio remains the same under MC as under monotonicity. Vytlacil (2002) shows that monotonicity is 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 monotonicity is that the instrument affects the cost of treatment, but not the benefits subjects expect from it. In this framework, compliers get treated because the instrument reduces the cost of treatment. I show that MC is equivalent to a symmetric index model, in which the instrument can affect both the cost and the expected benefit of treatment. Within this structural framework, I can define benefit and cost compliers. Benefit compliers would comply even if the instrument had no impact on the cost of treatment. Cost compliers comply because the instrument lowers the cost of treatment. In this structural framework, comvivors, the subpopulation within which treatment effects are identified, correspond to cost compliers.

I extend this result to several other treatment effect models. For instance, Huber & Mellace (2012) identify treatment effects when the Kitagawa (2008) test rejects monotonicity and SMC in the data. I show that their approach is valid under a weaker condition than their local monotonicity assumption.

In a related paper, Small & Tan (2007) demonstrate that under SMC, the Wald ratio captures a weighted average treatment effect. Nevertheless, their weights can lie outside the unit interval, making their result hard to interpret. A related concern is that their weighting scheme fails to capture causal effects for a well-defined subpopulation of units. I show here that WMC is sufficient to capture average causal effects for a well-defined subpopulation 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 monotonicity, and shows that the bias of the Wald parameter can be well approximated if said violations are small. As mentioned above, Huber & Mellace (2012) identify average treatment effects on compliers and defiers under a local monotonicity 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 second contribution of this paper is to extrapolate the validity of IV estimates.

IV models identify the effect of the treatment among compliers only. Sometimes, the objective of a study is to assess whether a policy should be generalized, and the instrument used to answer this question corresponds to the policy change to be implemented if the policy were generalized. In such instances, external validity is not an issue: if the policy were to be generalized, only compliers would be affected. Therefore, knowing the effect of the policy on compliers is sufficient.

Nevertheless, the instrument is sometimes the only exogenous source of variation available to study a causal question relevant in the entire population. In such instances, findings derived from IV models lack external validity since they apply to compliers only, whereas one wants to know the effect of the treatment in the entire population.

A good example is Angrist & Evans (1998). In this paper, the percentage of women with a third child is only 6 percentage points higher when their first two children are of the same sex. Thus, compliers account for only 6% of the population. As a result, estimates from this paper apply solely to a very small fraction of the total population, whilst one would like to know the effect of childbearing on female labor supply in the entire population.

To tackle the question of external validity, I investigate whether compliers have the same distribution of potential outcomes as a larger population G. If the answer is positive, then treatment effects identified in IV models apply to a population whose size, $P(G)$, is larger than the size of compliers, $P(C)$.

First, I derive a simple formula for $P(G)$, the size of the largest population with the same distribution of potential outcomes as compliers. $P(G)$ is written as the ratio of the percentage of compliers in the total population, and the percentage of compliers in the $(Y(0), Y(1))$ stratum with the largest percentage of compliers. If potential outcomes are binary, they define four strata: subjects such that $(Y(0) = 0, Y(1) = 0)$, those such that $(Y(0) = 0, Y(1) = 1)$, those such that $(Y(0) = 1, Y(1) = 0)$, and those such that $(Y(0) = 1, Y(1) = 1)$. Assume for instance that the weight of compliers is the highest in the $(Y(0) = 0, Y(1) = 0)$ stratum. In this case, $P(G)$ is equal to $\frac{P(C)}{P(C|Y(0)=0,Y(1)=0)}$.

By consequence, the size of G is not point identified from the data. Computing $P(G)$ would require knowing the joint distribution of potential outcomes, while the data reveal the marginals only partially.

Notwithstanding, the formula still shows that $P(G) = 1$ if compliance status is independent of potential outcomes. It also shows that $P(G) = P(C)$ if one $(Y(0), Y(1))$ stratum bears compliers exclusively. In many applications, it seems unlikely that one $(Y(0), Y(1))$ stratum bears compliers exclusively. For instance, in Angrist & Evans (1998), compliers represent 6% of the total population. Potential outcomes are mothers participating in the labor market with or without a third child, which are binary variables. Accordingly, $(Y(0), Y(1))$ only define four strata. The scenario under which one of these four strata is made up of compliers only seems unlikely; this would require that one $(Y(0), Y(1))$ stratum bears 6% of the total population at most, and this is implausible.

Second, I derive sharp bounds for $P(G)$ under the standard IV assumptions. The sharp lower bound is trivial as it is equal to $P(C)$. This does not mean that $P(G) = P(C)$. It simply means that the IV assumptions are not strong enough to reject the claim according to which one $(Y(0), Y(1))$ stratum bears compliers only, even though in many applications it seems implausible that this is the case.

The sharp upper bound is not trivial as it can be lower than 1. Therefore, IV assumptions are sufficient to reject the claim that the Wald ratio is equal to the average treatment effect (ATE) in some applications.

Third, I introduce a supplementary assumption, under which I derive a non-trivial lower bound for $P(G)$. This assumption requires that there be covariates X such that $(Y(0), X)$

has more influence on compliance than $(Y(0), Y(1))$. I call this assumption a "strong instrument" assumption, since it requires that X be a strong determinant of compliance status.

Although it is not innocuous, this strong instrument condition is weaker than other assumptions used in the literature to extrapolate IV estimates. If, after conditioning for one potential outcome and for covariates, the effect of the treatment is independent of potential treatments, then my "strong instrument" assumption holds. Therefore, this condition is weaker than the "conditional effect ignorability" assumption used by Angrist & Fernandez-Val (2010) to extrapolate the LATE. Indeed, their assumption requires that the effect of the treatment be independent of potential treatments, after conditioning only for covariates. It is also weaker than the rank invariance assumption in Chernozhukov & Hansen (2005).

Lastly, I show how to estimate bounds of $P(G)$, and how to draw inference when the potential outcomes are discrete.

Many papers have already investigated whether IV estimates can be extrapolated to larger populations. To date, most have either relied on parametric restrictions (see e.g. Heckman et al., 2003), or on a non parametric identification at infinity condition, seldom met in practice (for a review, see Heckman, 2010). Few papers have succeeded in extrapolating IV estimates under credible non-parametric assumptions. As mentioned above, Angrist & Fernandez-Val (2010) use a covariate-based approach to retrieve the ATE from the LATE, but their approach is based on a more restrictive ignorability assumption. Two other related papers are Small & Tan (2007) and DiNardo & Lee (2011). They show that under SMC, the Wald ratio is rewritten as a weighted ATE. Nevertheless, this parameter could differ significantly from the unweighted ATE. If not properly understood, their result could even be misleading, since it could give the reader the impression that standard IV assumptions alone are sufficient to extrapolate the Wald ratio to larger populations. On the contrary, I show that IV assumptions alone are never sufficient to reject the claim that the Wald ratio applies to compliers only.

I use my results to reanalyze both the internal and the external validity of Angrist & Evans (1998). On internal validity, Kitagawa's test is not rejected in these data. Accordingly, there are more compliers than defiers in each $Y(0)$ stratum, and in each $Y(1)$ stratum. As a result, SMC likely holds; even though there might be defiers in the population, it does not seem to be a threat to the internal validity of their results.

On external validity, $P(G) = 1$ is not rejected in these data. I also estimate the non-trivial lower bound for $P(G)$ obtained under the "strong instrument" assumption mentioned above.

It lies between 16.3% and 23.2%, depending on the vector of covariates used. Under this supplementary assumption, results apply to at least 20% of the total population, instead of 6% only.

2 LATE with defiers

IV estimates might lack internal validity because they rely on a strong monotonicity assumption. This assumption is not realistic in all settings (see e.g. Freedman, 2006 and Barua & Lang, 2010). In this section, I show that monotonicity can be replaced by a substantially weaker condition, while keeping all the standard LATE results unchanged.

2.1 The "more compliers than defiers" assumptions

All the results which follow apply irrespective of whether potential outcomes are discrete or continuous, but results are simpler to understand with discrete potential outcomes than with continuous ones. Therefore, I will adopt the following conventions so as to write all the theorems "as if" the potential outcomes were discrete. For any continuous random variable or random vector X, $P(X = x)$ should be understood as $f_X(x)$ where f_X is the density of X. For any event A with positive probability, $P(X = x|A)$ means $f_{X|A}(x)$. Finally, $P(A|X = x)$ stands for $\frac{f_{X|A}(x)P(A)}{f_X(x)}$, and $P(A, X = x)$ stands for $f_{X|A}(x)P(A)$. Those conventions will ease the presentation of the results..

My framework is the same as in 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 the treatment and of the instrument. We only observe $D = D(Z)$ and $Y = Y(D, Z)$. 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) have $D(0) = 0$ and $D(1) = 1$, while defiers have $D(0) = 1$ and $D(1) = 0$. 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.

I assume that $P(D = 1|Z = 1) > P(D = 1|Z = 0)$. Under Assumption 2.1, this implies that more subjects are compliers than defiers: $P(C) > P(F)$. This is a mere normalization. If $P(D = 1|Z = 1) < P(D = 1|Z = 0)$, one can switch the words "defiers" and "compliers" in what follows.

Let me introduce the following notation. For any random variable or random vector T, let

$\mathcal{S}(T)$ denote the support of T. For instance, $\mathcal{S}(Y(0), Y(1))$ are all the values (y_0, y_1) such that $P(Y(0) = y_0, Y(1) = y_1) > 0$.

First, Angrist et al. (1996) assume that the instrument is independent of potential treatments and potential outcomes.

Assumption 2.1 *(Instrument independence)*

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

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

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

$$Y(d, 0) = Y(d, 1) = Y(d).$$

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

Assumption 2.3 *(Instrument monotonicity)*

$$D(1) \geq D(0).$$

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

Assumption 2.4 *(Weak more compliers than defiers assumption: WMC)*
For every δ in $\mathcal{S}(Y(1) - Y(0))$,

$$P(Y(1) - Y(0) = \delta, F) \leq P(Y(1) - Y(0) = \delta, C). \tag{2.1}$$

When monotonicity holds, the left hand side of Equation (2.1) is equal to 0, so Assumption 2.4 is verified. Therefore, WMC is weaker than monotonicity. Dividing both sides of Equation (2.1) by $P(Y(1) - Y(0) = \delta)$ shows that WMC is equivalent to

$$P(F|Y(1) - Y(0) = \delta) \leq P(C|Y(1) - Y(0) = \delta).$$

Angrist et al. (1996) show that the Wald ratio identifies a LATE if there are no defiers, or if treatment effects are the same among defiers and compliers. Multiplying each side of (2.1) by $\frac{P(C)}{P(F)}$ yields

$$\frac{P(Y(1) - Y(0) = \delta|F)}{P(Y(1) - Y(0) = \delta|C)} \leq \frac{P(C)}{P(F)}. \tag{2.2}$$

Therefore, the two assumptions Angrist et al. (1996) point out are polar cases of WMC. Remember that we have assumed, as a mere normalization, that $P(F)$ is smaller than $P(C)$. This implies that the right hand side of Equation (2.2) is greater than 1. $P(Y(1)-Y(0)=\delta|F)$ and $P(Y(1)-Y(0)=\delta|C)$ are the probability distributions of treatment effects among defiers and compliers. If those distributions are the same, meaning that treatment effects are the same in the two populations, WMC is verified since the left hand side of Equation (2.2) is equal to 1. On the contrary, when monotonicity is verified, $P(F)=0$. This implies that Equation (2.2) is also verified since its right hand side is equal to $+\infty$.

In-between those two polar cases, WMC holds in many intermediate cases. Equation (2.2) shows that the more defiers there are, the less compliers can differ from defiers for WMC to hold, and conversely. Indeed, when compliers and defiers differ a lot, i.e. when the highest value of

$$\frac{P(Y(1)-Y(0)=\delta|F)}{P(Y(1)-Y(0)=\delta|C)}$$

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 be "similar" for Equation (2.4) to hold. Therefore, WMC will hold if few subjects are defiers, or if defiers are reasonably similar to compliers. On the contrary, if many subjects are defiers and defiers differ a lot from compliers, it will not hold.

I also consider a stronger MC assumption, which is a special case of the stochastic monotonicity assumption considered in Small & Tan (2007). This stronger assumption is not necessary to show that the Wald ratio captures a LATE, but it is necessary to obtain results on QTE.

Assumption 2.5 *(Strong more compliers than defiers: SMC, Small & Tan (2007))*
For every (y_0, y_1) *in* $\mathcal{S}(Y(0), Y(1))$,

$$P(Y(0)=y_0, Y(1)=y_1, F) \leq P(Y(0)=y_0, Y(1)=y_1, C). \tag{2.3}$$

WMC is weaker than SMC. To see this, note that $P(Y(1)-Y(0)=\delta, F)$ and $P(Y(1)-Y(0)=\delta, C)$ are sums or integrals of $P(Y(0)=y_0, Y(1)=y_1, F)$ and $P(Y(0)=y_0, Y(1)=y_1, C)$ over all values of (y_0, y_1) such that $y_1-y_0=\delta$.

Multiplying each side of Equation (2.3) by $\frac{P(C)}{P(F)}$ yields:

$$\frac{P(Y(0)=y_0, Y(1)=y_1|F)}{P(Y(0)=y_0, Y(1)=y_1|C)} \leq \frac{P(C)}{P(F)}. \tag{2.4}$$

Equations (2.2) and (2.4) show that WMC requires that treatment effects be reasonably similar among compliers and defiers, while SMC require that potential outcomes be reasonably similar in those two populations. In many applications, treatment effects are less

heterogeneous across subpopulations than potential outcomes. Therefore, WMC is more likely to hold in practice than SMC.

SMC is implied by the stochastic monotonicity assumption studied in Small & Tan (2007). In their framework, SMC ensures that their equation (4) holds for $U = (Y(0), Y(1))$, and $U = (Y(0), Y(1))$ also ensures that their equations (2) and (3) hold. On the contrary, WMC is not implied by stochastic monotonicity.

2.2 Identification of local treatment effects under MC.

Let

$$P_0(y_0) = \frac{P(Y = y_0, D = 0|Z = 0) - P(Y = y_0, D = 0|Z = 1)}{P(D = 0|Z = 0) - P(D = 0|Z = 1)}$$
$$P_1(y_1) = \frac{P(Y = y_1, D = 1|Z = 1) - P(Y = y_1, D = 1|Z = 0)}{P(D = 1|Z = 1) - P(D = 1|Z = 0)}$$
$$W = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)}.$$

W is the Wald ratio; $P_0(.)$ $P_1(.)$ are functions appearing in Imbens & Rubin (1997). All those quantities only involve Y, D, and Z, and are thus identified from the data.

Angrist et al. (1996) and Imbens & Rubin (1997) show the following results.
LATE Theorems (Angrist et al. (1996) and Imbens & Rubin (1997))
Suppose Assumptions 2.1, 2.2, and 2.3 hold. Then,

$$P(C) = P(D = 1|Z = 1) - P(D = 1|Z = 0)$$
$$P(Y(0) = y_0|C) = P_0(y_0)$$
$$P(Y(1) = y_1|C) = P_1(y_1)$$
$$E\left[Y(1) - Y(0)|C\right] = W.$$

Under the standard LATE assumptions, the average treatment effect among compliers is identified from the data. The distribution of each potential outcome is also identified, which makes it possible to compute quantile treatment effects.

When WMC is substituted for monotonicity, the Wald ratio identifies the average effect of the treatment for a subgroup of compliers of size $P(D = 1|Z = 1) - P(D = 1|Z = 0)$ as shown below.

Theorem 2.1 *Suppose Assumptions 2.1, 2.2, and 2.4 hold. Then, one can partition compliers into two subpopulations C_F and C_V such that:*

1. *C_F satisfies*

$$P(C_F) = P(F)$$
$$E\left[Y(1) - Y(0)|C_F\right] = E\left[Y(1) - Y(0)|F\right].$$

2. *C_V satisfies*

$$P(C_V) = P(D = 1|Z = 1) - P(D = 1|Z = 0)$$
$$E\left[Y(1) - Y(0)|C_V\right] = W.$$

Proof of Theorem 2.1

I start proving the first point. Under Assumption 2.4, $P(C|Y(1) - Y(0) = \delta) = 0$ implies that $P(F|Y(1) - Y(0) = \delta) = 0$ as well. Therefore, when $P(C|Y(1) - Y(0) = \delta) = 0$,

$$p(\delta) = \frac{P(F|Y(1) - Y(0) = \delta)}{P(C|Y(1) - Y(0) = \delta)}$$

should be understood as 0. Assumption 2.4 also ensures that this ratio is smaller than 1.

For every δ in $\mathcal{S}((Y(1)-Y(0))$, let $(B(\delta))$ denote a collection of Bernoulli random variables independent of $(Y(1) - Y(0), D(0), D(1))$, and such that $P(B(\delta) = 1) = p(\delta)$. Let also $C_F = \{C, B(Y(1) - Y(0)) = 1\}$. For every δ in $\mathcal{S}(Y(1) - Y(0))$,

$$\begin{aligned}
& P(Y(1) - Y(0) = \delta, C_F) \\
= \; & P(Y(1) - Y(0) = \delta, C, B\left(Y(1) - Y(0)\right) = 1) \\
= \; & P(Y(1) - Y(0) = \delta, C, B\left(\delta\right) = 1) \\
= \; & P(Y(1) - Y(0) = \delta, C)P(B\left(\delta\right) = 1) \\
= \; & P(Y(1) - Y(0) = \delta, F).
\end{aligned} \tag{2.5}$$

Summing or integrating Equation (2.5) over $\mathcal{S}(Y(1)-Y(0))$ yields $P(C_F) = P(F)$. Dividing Equation (2.5) by $P(C_F) = P(F)$ yields $P(Y(1) - Y(0) = \delta|C_F) = P(Y(1) - Y(0) = \delta|F)$, which proves that

$$E\left[Y(1) - Y(0)|C_F\right] = E\left[Y(1) - Y(0)|F\right]. \tag{2.6}$$

I can now prove the second point. Let $C_V = C \setminus C_F = \{C, B(Y(0), Y(1)) = 0\}$.

$$P(C_V) = P(C) - P(C_F) = P(C) - P(F) = P(D = 1|Z = 1) - P(D = 1|Z = 0).$$

The last equality follows from the law of total probabilities and from Assumption 2.1.

Because C_V and C_F are included in C,

$$\begin{aligned} P(C_V|C) &= \frac{P(C_V)}{P(C)} = \frac{P(C)-P(F)}{P(C)} \\ P(C_F|C) &= \frac{P(C_F)}{P(C)} = \frac{P(F)}{P(C)}. \end{aligned} \tag{2.7}$$

Then, partitioning compliers into C_F and C_V yields

$$\begin{aligned} E\left[Y(1)-Y(0)|C\right] &= P(C_V|C)E\left[Y(1)-Y(0)|C_V\right] + P(C_F|C)E\left[Y(1)-Y(0)|C_F\right] \\ &= \frac{P(C)-P(F)}{P(C)}E\left[Y(1)-Y(0)|C_V\right] + \frac{P(F)}{P(C)}E\left[Y(1)-Y(0)|F\right]. \end{aligned} \tag{2.8}$$

The second equality is derived from Equations (2.6) and (2.7). When there are defiers, Angrist et al. (1996) show the following equality:

$$W = \frac{P(C)}{P(C)-P(F)}E\left[Y(1)-Y(0)|C\right] - \frac{P(F)}{P(C)-P(F)}E\left[Y(1)-Y(0)|F\right]. \tag{2.9}$$

Plugging Equation (2.8) into Equation (2.9) finally yields $W = E\left[Y(1)-Y(0)|C_V\right]$.
QED.

When there are defiers, Angrist et al. (1996) show that the Wald ratio is a weighted difference between the average treatment effect among compliers and defiers. But WMC ensures that a subgroup of compliers C_F has the same size and the same average treatment effect as defiers. Therefore, the average effect of the treatment for C_F and F cancel out, and only the average treatment effect among C_V remains. Hereafter, I call C_F the "comfiers", to emphasize the parallel between them and defiers. C_V are compliers who "out-survive" defiers. Henceforth, I call them the "comvivors".

To see how C_F is constructed, assume that $Y(0)$ and $Y(1)$ are binary. Then, $Y(1)-Y(0)$ defines three strata: subjects such that $Y(1)-Y(0)=-1$, those such that $Y(1)-Y(0)=0$, and those such that $Y(1)-Y(0)=1$. Assume also that $P(C)=0.75$ and $P(F)=0.25$. Finally, assume that the joint distribution of $(Y(1), Y(0), C, F)$ is as in Figure 1.

Y(1)-Y(0)	P(Y(1)-Y(0)=.,F)	P(Y(1)-Y(0)=,C)
-1	0.1	0.2
0	0.05	0.3
1	0.1	0.25

Figure 1: Joint distribution of $(Y(1), Y(0), C, F)$.

The WMC assumption is verified here since for every $\delta \in \{-1; 0; 1\}$,

$$P(Y(1)-Y(0)=\delta, F) \leq P(Y(1)-Y(0)=\delta, C).$$

To construct C_F, one needs to pick up

$$\frac{P(Y(1)-Y(0)=\delta, F)}{P(Y(1)-Y(0)=\delta, C)}\%$$

of compliers in each of the three $Y(1)-Y(0)=\delta$ strata. This is what the Bernoulli variables do in the proof. This amounts to picking up $\frac{0.1}{0.2}=\frac{1}{2}$ of compliers in the $Y(1)-Y(0)=-1$ stratum, $\frac{0.05}{0.30}=\frac{1}{6}$ of them in the $Y(1)-Y(0)=0$ stratum, and $\frac{0.10}{0.25}=\frac{2}{5}$ of them in the $Y(1)-Y(0)=1$ stratum. The joint distribution of $(Y(1), Y(0), C_V, C_F, F)$ resulting from this construction is displayed in Figure 2. One can check that C_F indeed satisfies all the requirements listed in point 1 of Theorem 2.1.

Y(1)-Y(0)	P(Y(1)-Y(0)=.,F)	P(Y(1)-Y(0)=.,CF)	P(Y(1)-Y(0)=.,CV)
-1	0.1	0.1	0.1
0	0.05	0.05	0.25
1	0.1	0.1	0.15

Figure 2: Constructing C_F and C_V.

But WMC is not sufficient to identify the effect of the treatment on the distribution of the potential outcomes for C_V, while SMC is as shown below.

Theorem 2.2 *Suppose Assumptions 2.1, 2.2, and 2.5 hold. Then, one can partition compliers into two subpopulations C_F and C_V*[1] *such that:*

1. *C_F satisfies*

$$\begin{aligned}
&P(C_F) = P(F)\\
&P(Y(0)=y_0|C_F) = P(Y(0)=y_0|F)\\
&P(Y(1)=y_1|C_F) = P(Y(1)=y_1|F).
\end{aligned}$$

2. *C_V satisfies*

$$\begin{aligned}
&P(C_V) = P(D=1|Z=1) - P(D=1|Z=0)\\
&P(Y(0)=y_0|C_V) = P_0(y_0)\\
&P(Y(1)=y_1|C_V) = P_1(y_1)\\
&E\left[Y(1)-Y(0)|C_V\right] = W.
\end{aligned}$$

[1]There is a slight abuse of notation here since those two populations are not necessarily the same as in Theorem 2.1, but I denote them C_F and C_V as well to keep the notation simple.

The proof of Theorem 2.2 is the same as the proof of Theorem 2.1, except that I use Bernoulli variables to pick up

$$\frac{P(Y(0) = y_0, Y(1) = y_1, F)}{P(Y(0) = y_0, Y(1) = y_1, C)}\%$$

of compliers in each $(Y(0) = y_0, Y(1) = y_1)$ stratum. This ensures that C_F and F have the same distribution of potential outcomes, and not only the same distribution of treatment effects, which proves point 1. Point 2 follows after some computations.

2.3 Who is in C_V? A structural interpretation.

The objective of this section is to extend the structural interpretation of the LATE developed in Vytlacil (2002) and Heckman & Vytlacil (2005). To that end, I start using a simple Roy selection model to convey the intuition, before turning to a more general model. I only consider the structural interpretation of the SMC condition, but similar results can be obtained on WMC.

To fix ideas, I discuss the example of a randomized experiment relying on an encouragement design protocol. In such protocols, the encouragement group typically receives a flyer describing the treatment (date, place, content...), and a financial incentive to get treated. To simplify, assume that $Y(1) - Y(0) > 0$.

Assume that selection into treatment is determined by the following Roy Model:

$$D = 1\{\alpha\,(Y(1) - Y(0)) + Z(\beta - \alpha)\,(Y(1) - Y(0)) \geq \varepsilon - \lambda Z\}. \tag{2.10}$$

α is a random coefficient equal to the ratio between the perceived and the true benefit of treatment before reading the flyer. β represents the same ratio after reading the flyer. λ is a positive constant, and represents the financial incentive given to subjects who undertake the treatment in the encouragement group. ε is a positive unobserved heterogeneity term, and represents the cost of treatment. $Z = 1$ for subjects in the encouragement group.

When $Z = 0$, an individual will undertake the treatment if the benefit she expects from treatment before reading the flyer is greater than the cost of getting treated. When $Z = 1$, she will get treated if the benefit she expects after reading the flyer is greater than the difference between the cost of treatment and the incentive.

Therefore, the instrument has a bi-dimensional effect on selection into treatment. It reduces the cost via the financial incentive; it also changes the benefit expected from treatment via the flyer. In particular, the flyer will reduce the left-hand-side of Equation (2.10) for

subjects whose expected benefit diminishes after reading the flyer, i.e. for subjects such that $\alpha > \beta$. This is the reason why defiers can exist in this model, even though Z reduces the right hand side for all subjects. If $P(\alpha = \beta) = 1$, meaning that the flyer has no impact on the benefits subjects expect from treatment, monotonicity holds.

Under (2.10), a few computations show that

$$P(F|Y(0) = y_0, Y(1) = y_1) \leq P\left(\alpha \geq \frac{\varepsilon}{y_1 - y_0} > \beta | Y(0) = y_0, Y(1) = y_1\right)$$

and

$$P(C|Y(0) = y_0, Y(1) = y_1) \geq P\left(\beta \geq \frac{\varepsilon}{y_1 - y_0} > \alpha | Y(0) = y_0, Y(1) = y_1\right).$$

Therefore, if

$$(\alpha, \beta, \varepsilon)|Y(0), Y(1) \sim (\beta, \alpha, \varepsilon)|Y(0), Y(1), \tag{2.11}$$

SMC holds. Indeed, if Equation (2.11) is verified, then

$$P\left(\alpha \geq \frac{\varepsilon}{y_1 - y_0} > \beta | Y(0) = y_0, Y(1) = y_1\right) = P\left(\beta \geq \frac{\varepsilon}{y_1 - y_0} > \alpha | Y(0) = y_0, Y(1) = y_1\right).$$

In mathematical terms, Equation (2.11) states that α and β are exchangeable conditional on $Y(0)$ and $Y(1)$. This means that for each value of $Y(0)$ and $Y(1)$, the number of people whose expected benefit from treatment increases after reading the flyer is at least as big as the number of people whose expected benefit decreases.

If the flyer reveals to each person the benefit she should expect from treatment (provided this is possible), SMC has little chances to hold: among people with low treatment effects, there will probably be more defiers than compliers. Still, in most encouragement designs, the flyer conveys neutral information about the treatment, or even advertises it to increase the difference in take-up rates across groups. In such instances, (2.11) is reasonably realistic. It means that the advertisement should be good enough to ensure that for each value of $Y(0)$ and $Y(1)$, there are at least as many people with a better opinion of the treatment after reading it as people with a worse opinion.

Let

$$BC = \left\{\beta \geq \frac{\varepsilon}{Y(1) - Y(0)}, \alpha < \frac{\varepsilon - \lambda}{Y(1) - Y(0)}\right\}.$$

BC are "benefit-compliers". They comply because the flyer increases the benefit they expect from treatment. Even if the encouragement group received no financial incentive ($\lambda = 0$), they would still comply. The remaining part of compliers is denoted CC, which stands for "cost-compliers".

One can show[2] that if Equation (2.11) holds, then

$$\begin{aligned}
&P(BC) = P(F) \\
&P(Y(0) = y_0|BC) = P(Y(0) = y_0|F) \\
&P(Y(1) = y_1|BC) = P(Y(1) = y_1|F).
\end{aligned}$$

This shows that the structural population BC corresponds to the reduced form population C_F defined in Theorem 2.2. Under SMC, the Wald ratio identifies the effect of the treatment among cost-compliers, by opposition to benefit-compliers.

I formalize this discussion in the next Theorem, which generalizes Vytlacil (2002). Let the DGP for potential treatments be as follows:

Assumption 2.6 *(Symmetric index model)*

$$D(z) = 1\{V_z \geq v_z\}$$

with

$$(V_0, V_1)|Y(0), Y(1) \sim (V_1, V_0)|Y(0), Y(1)$$

and

$$v_0 \geq v_1.$$

As Vytlacil's single index model, the symmetric index model (SIM) is a rational choice model. Indeed, $D(z) = 1$ if the benefit expected from treatment V_z is greater than the cost of treatment v_z. In the single index model, the instrument only affects the cost of treatment v_z; it does not affect the expected benefit: $V_0 = V_1 = V$. Here, the instrument might affect the expected benefit. In the Roy model, this corresponds to the informational effect of the flyer.

In the symmetric index model, V_0 and V_1 must be exchangeable conditional on $(Y(0), Y(1))$. For SMC to hold, this condition is sufficient, but it is not necessary. What makes this assumption appealing is rather the simple structural interpretation of the populations C_V and C_F which is obtained under it.

Let benefit-compliers (BC) verify

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

while cost-compliers (CC) verify

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

[2] I prove this statement in the proof of Theorem 2.3.

One can check that 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.

Theorem 2.3

1. *If Assumptions 2.1, 2.2, and 2.6 hold, then*

   (a) *Assumption 2.5 is verified.*

   (b) *BC satisfies*

$$\begin{aligned} &P(BC) = P(F) \\ &P(Y(0) = y_0|BC) = P(Y(0) = y_0|F) \\ &P(Y(1) = y_1|BC) = P(Y(1) = y_1|F). \end{aligned}$$

   (c) *CC satisfies*

$$\begin{aligned} &P(CC) = P(D = 1|Z = 1) - P(D = 1|Z = 0) \\ &P(Y(0) = y_0|CC) = P_0(y_0) \\ &P(Y(1) = y_1|CC) = P_1(y_1) \\ &E\left[Y(1) - Y(0)|CC\right] = W. \end{aligned}$$

2. *If Assumption 2.5 is verified, one can construct*

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

   which both rationalize the data and verify Assumption 2.6.

The first part of the theorem shows that if the DGP is a symmetric index model, then SMC is verified. Moreover, benefit-compliers and defiers have the same size and the same distribution of potential outcomes. In this structural framework, the Wald ratio identifies the effect of the treatment among cost-compliers only; the intuition of the simple Roy model extends to more general rational choice models.

In the single index model of Vytlacil (2002), compliers are necessarily cost-compliers since the instrument has no impact on the benefit expected of treatment. Therefore, C_V has the same structural interpretation under SMC than the standard population of compliers under monotonicity. To emphasize the parallel between the two populations, notice that Vytlacil's compliers satisfy $\{V \in [v_1, v_0)\}$, while cost-compliers satisfy $\{(V_0, V_1) \in [v_1, v_0)^2\}$.

The second part of the theorem shows that SMC is observationally equivalent to a SIM.[3] If the true DGP of potential treatments is a SIM, SMC holds. Conversely, if SMC holds, one can construct a SIM rationalizing the data, even though the true DGP might not be a SIM. This second statement is a mere generalization of the first point of Proposition 2.1 in Chaisemartin & D'Haultfoeuille (2012).

2.4 Testability.

Imbens & Rubin (1997) show that the standard LATE assumptions -random instrument, exclusion restriction, and monotonicity- have an implication which is testable from the data. Kitagawa (2008) develops the corresponding statistical test. The following lemma shows that the alternative LATE assumptions -random instrument, exclusion restriction, and SMC- have the same testable implication.

Lemma 2.1

1. *Assumptions 2.1, 2.2, and 2.3 imply that the following inequalities must hold in the data: for every* (y_0, y_1) *in* $\mathcal{S}(Y(0), Y(1))$,

$$\begin{aligned} P(Y = y_0, D = 0|Z = 1) &\leq P(Y = y_0, D = 0|Z = 0) \\ P(Y = y_1, D = 1|Z = 0) &\leq P(Y = y_1, D = 1|Z = 1). \end{aligned} \tag{2.12}$$

2. *Assumptions 2.1, 2.2, and 2.5 also imply that Equation* (2.12) *must hold in the data.*

Proof of Lemma 2.1:
The first point of the lemma is shown in Imbens & Rubin (1997).

To prove the second point, notice that under Assumptions 2.1, 2.2, and 2.5, Theorem 2.2 states that

$$\begin{aligned} P(Y(0) = y_0|C_V) &= P_0(y_0) \\ P(Y(1) = y_1|C_V) &= P_1(y_1). \end{aligned}$$

This implies that we must have

$$\begin{aligned} 0 &\leq P_0(y_0) \\ 0 &\leq P_1(y_1). \end{aligned}$$

[3] One can also show that WMC is observationally equivalent to a SIM in which V_0 and V_1 are exchangeable conditional on $Y(1) - Y(0)$.

This yields the result.
QED.

This means that both the standard and the alternative LATE assumptions are partly testable. Still, it remains unclear whether Kitagawa's test is likely to detect violations of those assumptions or not. Indeed, those two sets of assumptions imply that Equation (2.12) must hold in the data, but the converse is false; Equation (2.12) might hold, while those assumptions are violated, leading to type II error. This is not a statistical power issue, but rather a logical power issue.

The following Lemma is a first attempt to clarify the logical power of Kitagawa's test. It shows that this test has more logical power to test SMC than to test monotonicity, as the implication tested is closer to SMC than to monotonicity.

Lemma 2.2 *Suppose Assumptions 2.1 and 2.2 hold. Then Equation* (2.12) *is verified if and only if for every* (y_0, y_1) *in* $\mathcal{S}(Y(0), Y(1))$

$$\begin{aligned} P(Y(0) = y_0, F) &\leq P(Y(0) = y_0, C) \\ P(Y(1) = y_1, F) &\leq P(Y(1) = y_1, C). \end{aligned} \tag{2.13}$$

Proof of Lemma 2.2:
Notice that

$$\begin{aligned} & P(Y = y_0, D = 0|Z = 0) - P(Y = y_0, D = 0|Z = 1) \\ = \; & P(Y(0) = y_0, D(0) = 0) - P(Y(0) = y_0, D(1) = 0) \\ = \; & P(Y(0) = y_0, D(0) = 0, D(1) = 1) - P(Y(0) = y_0, D(0) = 1, D(1) = 0) \\ = \; & P(Y(0) = y_0, C) - P(Y(0) = y_0, F). \end{aligned}$$

The first equality follows from Assumptions 2.1 and 2.2. The second one follows from the law of total probabilities.

Therefore,

$$P(Y = y_0, D = 0|Z = 1) \leq P(Y = y_0, D = 0|Z = 0)$$

if and only if

$$P(Y(0) = y_0, F) \leq P(Y(0) = y_0, C).$$

This proves the first equivalence. The second one is obtained through similar arguments.
QED.

Assume one is ready to take instrument independence and exclusion restriction for granted. Then, this lemma shows that Kitagawa's test has more logical power to test SMC than to

test monotonicity, because the testable implication is closer to SMC than to monotonicity. Assume that $Y(0)$ and $Y(1)$ are binary. Then, each $Y(0)$ and $Y(1)$ stratum is the aggregation of two $(Y(0), Y(1))$ strata. Kitagawa's test amounts to testing whether more subjects are compliers than defiers in each $Y(0)$ and $Y(1)$ stratum. If this test tells us that there are more compliers than defiers in, say, the $Y(0) = 0$ stratum, we can be reasonably confident that there are also more compliers in the $(Y(0) = 0, Y(1) = 0)$ and $(Y(0) = 0, Y(1) = 1)$ strata. On the contrary, the test does not tell us much on whether $P(F) = 0$ or not.

Finally, as Equation (2.13) is close to SMC, one might wonder whether Equation (2.13) is sufficient for Theorem 2.2 to hold. This would be great news, as this would mean that one can substitute a fully testable assumption to monotonicity.

The answer is no. Figure 3 presents a counterexample. Equation (2.13) is verified. Indeed, the sum of each line and each column is smaller for defiers than compliers. Still, it is impossible to construct a subpopulation C_F of compliers with the same marginal distributions of potential outcomes as defiers.

Indeed, $P(Y(0) = 1, F) = P(Y(1) = 1, F) = 0.2$. Therefore, to ensure that the marginal distributions are the same in C_F and F, one must have

$$P(Y(0) = 1, Y(1) = 1, C_F) + P(Y(0) = 1, Y(1) = 0, C_F) = 0.2 \tag{2.14}$$

and

$$P(Y(0) = 1, Y(1) = 1, C_F) + P(Y(0) = 0, Y(1) = 1, C_F) = 0.2. \tag{2.15}$$

Since $P(Y(0) = 1, Y(1) = 1, C) = 0.05$, $P(Y(0) = 1, Y(1) = 1, C_F)$ must be smaller than 0.05. As a result, one must set

$$P(Y(0) = 1, Y(1) = 0, C_F) = P(Y(0) = 0, Y(1) = 1, C_F) = 0.15.$$

for Equations (2.14) and (2.15) to hold.

Then,

$$P(Y(0) = 0, F) = P(Y(1) = 0, F) = 0.05.$$

Given the values of $P(Y(0) = 1, Y(1) = 0, C_F)$ and $P(Y(0) = 0, Y(1) = 1, C_F)$, having

$$P(Y(0) = 0, C_F) = P(Y(1) = 0, C_F) = 0.05$$

as well would require to set

$$P(Y(0) = 0, Y(1) = 0, C_F) = -0.10,$$

which is impossible.

Y(0) \ Y(1)	0	1
0	P(Y(0)=0,Y(1)=0,C)=0.40 P(Y(0)=0,Y(1)=0,F)=0.00 P(Y(0)=0,Y(1)=0,CF)= -0.10	P(Y(0)=0,Y(1)=1,C)=0.15 P(Y(0)=0,Y(1)=1,F)=0.05 P(Y(0)=0,Y(1)=1,CF)=0.15
1	P(Y(0)=1,Y(1)=0,C)=0.15 P(Y(0)=1,Y(1)=0,F)=0.05 P(Y(0)=1,Y(1)=0,CF)=0.15	P(Y(0)=1,Y(1)=1,C)=0.05 P(Y(0)=1,Y(1)=1,F)=0.15 P(Y(0)=1,Y(1)=1,CF)=0.05

Figure 3: Equation (2.13) is not sufficient.

Equation (2.13) ensures that there exists two subpopulations of compliers C_V^0 and C_V^1 such that the distribution of $Y(0)$ is the same in C_V^0 and F, the distribution of $Y(1)$ is the same in C_V^1 and F, and both C_V^0 and C_V^1 have the same size as defiers. But the preceding counterexample shows that it does not ensure that there exists a unique population C_V with the same size and the same marginal distributions of $Y(0)$ and $Y(1)$ as F.

2.5 Extensions

2.5.1 Characterizing C_V.

LATE are by definition local effects. Therefore, it is important to describe the population on which those effects are measured, so as to assess whether the findings of the analysis generalize 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, SMC does not ensure that the distribution of covariates in C_V is identified. Still, a weak strengthening of SMC ensures that the distribution of X and the effect of the treatment are identified within a subpopulation a compliers. Let X denote a vector of covariates.

Assumption 2.7 *(SMC-X)*

For every (y_0, y_1, x) in $\mathcal{S}((Y(0), Y(1), X))$,

$$P(Y(0) = y_0, Y(1) = y_1, X = x, F) \leq P(Y(0) = y_0, Y(1) = y_1, X = x, C). \tag{2.16}$$

It suffices to sum or to integrate (2.16) over $\mathcal{S}(X)$ to see that Assumption 2.7 is stronger than SMC. But it also has more testable implications. Indeed, under instrument independence and exclusion restriction, it implies that the following equations must hold in the

data: for every (y_0, y_1, x) in $\mathcal{S}(Y(0), Y(1), X)$,

$$\begin{aligned} P(Y = y_0, D = 0, X = x|Z = 1) &\leq P(Y = y_0, D = 0, X = x|Z = 0) \\ P(Y = y_1, D = 1, X = x|Z = 0) &\leq P(Y = y_1, D = 1, X = x|Z = 1). \end{aligned}$$

SMC-X ensures that there is a subpopulation of compliers for which treatment effects and the distribution of covariates are identified as shown below. Let

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

Theorem 2.4 *Suppose Assumptions 2.1, 2.2, and 2.7 hold. Then, one can partition compliers into two subpopulations C_F and C_V*[4] *such that:*

1. *C_F satisfies*

$$\begin{aligned} &P(C_F) = P(F) \\ &P(Y(0) = y_0, Y(1) = y_1, X = x|C_F) = P(Y(0) = y_0, Y(1) = y_1, X = x|F). \end{aligned}$$

2. *C_V satisfies*

$$\begin{aligned} &P(C_V) = P(D = 1|Z = 1) - P(D = 1|Z = 0) \\ &P(Y(0) = y_0|C_V) = P_0(y_0) \\ &P(Y(1) = y_1|C_V) = P_1(y_1) \\ &E\left[Y(1) - Y(0)|C_V\right] = W \\ &P(X = x|C_V) = P_X(x). \end{aligned}$$

2.5.2 Identification of a LATE under Local Stochastic Monotonicity.

When Kitagawa's test is rejected in the data, implying that monotonicity and SMC are violated, Huber & Mellace (2012) replace those assumptions by a local monotonicity condition. Under this condition, treatment effects are identified among compliers and defiers.

Assumption 2.8 *(Local monotonicity)*
For every (y_0, y_1) in $\mathcal{S}((Y(0), Y(1)))$, either

$$P(D(1) \geq D(0)|Y(0) = y_0, Y(1) = y_1) = 1,$$

or

$$P(D(0) \geq D(1)|Y(0) = y_0, Y(1) = y_1) = 1.$$

[4] Again, there is a slight abuse of notation here.

This condition is weaker than monotonicity; under local monotonicity there can be defiers in the population. But among subjects such that $Y(0) = y_0$ and $Y(1) = y_1$, there cannot be both defiers and compliers. For local monotonicity to hold, the distribution of potential outcomes among compliers and defiers must differ a lot, as their supports cannot overlap.

Assumption 2.9 *(Local Stochastic Monotonicity)*
For every $d \in \{0; 1\}$, for every $y_d \in \mathcal{S}(Y(d))$,

$$P(Y(d) = y_d, F) \leq P(Y(d) = y_d, C) \quad \Rightarrow \quad P(Y(0) = y_0, Y(1) = y_1, F) \leq P(Y(0) = y_0, Y(1) = y_1, C)$$
$$P(Y(d) = y_d, F) \geq P(Y(d) = y_d, C) \quad \Rightarrow \quad P(Y(0) = y_0, Y(1) = y_1, F) \geq P(Y(0) = y_0, Y(1) = y_1, C).$$

This condition is weaker than local monotonicity; the support of $(Y(0), Y(1))$ among compliers and defiers can overlap. It is also weaker than SMC; under Assumption 2.9, there can be $Y(0) = y_0^*$ strata with more defiers than compliers, provided there are also more defiers in all the corresponding $(Y(0), Y(1)) = (y_0^*, y_1)$ strata.

One can substitute Assumption 2.9 to Assumption 2.8, while maintaining Huber & Mellace (2012) identification results as shown below.

Theorem 2.5 *Suppose Assumptions 2.1, 2.2, and 2.9 hold. Then, there is a population HM which includes both compliers and defiers, and which is such that*

$$\begin{aligned}
P(Y(0) = y_0, HM) &= \max\left(P(Y = y_0, D = 0|Z = 0), P(Y = y_0, D = 0|Z = 1)\right) \\
&- \min\left(P(Y = y_0, D = 0|Z = 0), P(Y = y_0, D = 0|Z = 1)\right) \\
P(Y(1) = y_1, HM) &= \max\left(P(Y = y_1, D = 1|Z = 1), P(Y = y_1, D = 1|Z = 0)\right) \\
&- \min\left(P(Y = y_1, D = 1|Z = 1), P(Y = y_1, D = 1|Z = 0)\right).
\end{aligned}$$

Summing or integrating the first equality over $\mathcal{S}(Y(0))$ yields a formula for $P(HM)$. Dividing those two equalities by $P(HM)$ yields formulas for $P(Y(0) = y_0|HM)$ and $P(Y(1) = y_1|HM)$, from which one can finally derive $E(Y(1) - Y(0)|HM)$.

2.5.3 Other treatment effects models

Those results extend to many other treatment effects models relying on monotonicity. Important examples include fuzzy Regression Discontinuity (RD) and quantile IV. In the fuzzy (RD) model in Hahn et al. (2001), let T be the forcing variable, and let t be the cut-off value of this forcing variable. Their Theorem 3 still holds if their Assumption A.3, a "monotonicity at the threshold" assumption, is replaced by a weaker "SMC at the threshold" condition: for every $(y_0, y_1) \in \mathcal{S}(Y(0), Y(1))$,

$$P(Y(0) = y_0, Y(1) = y_1, T = t, F) \leq P(Y(0) = y_0, Y(1) = y_1, T = t, C).$$

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). This Theorem still holds if in Assumption 2.1, point iv) is replaced by an SMC-X type of assumption writing as follows. For every $(y_0, y_1, x) \in \mathcal{S}(Y(0), Y(1), X)$,

$$P(Y(0) = y_0, Y(1) = y_1, X = x, F) \leq P(Y(0) = y_0, Y(1) = y_1, X = x, C),$$

where X is the set of covariates included in the regression.

3 A more global LATE

IV identifies the effect of the treatment for compliers only, while one might be interested in knowing the effect of the treatment for the entire population (see Deaton, 2009 and Heckman & Urzúa, 2010). In this section, I address this concern.

3.1 Three definitions of external validity

The following results hold under "more compliers than defiers", but to simplify the presentation I assume monotonicity. In a world without defiers, IV identify treatment effects for compliers, and IV estimates apply to a population of size

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

Therefore, the external validity of IV is at least equal to $P(C)$, but might be greater if a population larger than compliers has the same treatment effects as compliers. In what follows, I define the external validity of IV estimates as the size of the largest population with same treatment effects as compliers.

Let G^J be the largest population with the same joint distribution of $(Y(0), Y(1))$ as compliers. G^J is the largest population satisfying

$$P(Y(0) = y_0, Y(1) = y_1|G^J) = P(Y(0) = y_0, Y(1) = y_1|C). \tag{3.1}$$

Similarly, let G^M be the largest population with the same marginal distributions of $Y(0)$ and $Y(1)$ as compliers. G^M is the largest population satisfying

$$\begin{aligned} P(Y(0) = y_0|G^J) &= P(Y(0) = y_0|C) \\ P(Y(1) = y_1|G^J) &= P(Y(1) = y_1|C). \end{aligned} \tag{3.2}$$

Finally, let G^{Δ} be the largest population with the same average treatment effect as compliers. G^{Δ} is the largest population satisfying

$$E(Y(1) - Y(0)|G^{\Delta}) = E(Y(1) - Y(0)|C). \tag{3.3}$$

$P(G^J)$, $P(G^M)$, and $P(G^{\Delta})$ satisfy

$$P(G^J) \leq P(G^M) \leq P(G^{\Delta}).$$

$P(G^{\Delta})$ has two strong drawbacks with respect to $P(G^J)$ and $P(G^M)$. First, $P(G^{\Delta})$ is the external validity of the Wald ratio alone, while $P(G^J)$ and $P(G^M)$ are the external validity of both the Wald ratio and the QTE estimated in IV studies. Second, $P(G^{\Delta})$ is not robust to a monotonous transform of the outcome, while $P(G^J)$ and $P(G^M)$ are. For instance,

$$E(Y(1) - Y(0)|G^{\Delta}) = E(Y(1) - Y(0)|C)$$

does not imply

$$E(\ln(Y(1)) - \ln(Y(0))|G^{\Delta}) = E(\ln(Y(1)) - \ln(Y(0))|C).$$

On the contrary

$$P(Y(1) = y_1, Y(0) = y_0|G^J) = P(Y(1) = y_1, Y(0) = y_0|C)$$

implies

$$E(\ln(Y(1)) - \ln(Y(0))|G^J) = E(\ln(Y(1)) - \ln(Y(0))|C).$$

This leads me to focus on $P(G^J)$ and $P(G^M)$ in what follows.

3.2 A simple formula for external validity

This section derives a simple formula for $P(G^J)$.[5] For any random vector or random variable R with support $\mathcal{S}(R)$, let

$$P^+(C|R) = \sup_{r \in \mathcal{S}(R)} \{P(C|R = r)\}.$$

$P^+(C|Y(0), Y(1))$ is the share of compliers in the $(Y(0), Y(1))$ stratum in which this share is the highest. If

$$P(C|Y(0) = 0, Y(1) = 0) = 0.2$$
$$P(C|Y(0) = 0, Y(1) = 1) = 0.1$$
$$P(C|Y(0) = 1, Y(1) = 0) = 0.3$$
$$P(C|Y(0) = 1, Y(1) = 1) = 0.7,$$

[5] In the appendix, I also derive an explicit formula for $P(G^M)$.

then $P^+(C|Y(0), Y(1)) = 0.7$. One can show that $P^+(C|R)$ is included between $P(C)$ and 1. Let 1_C denote the compliance dummy, equal to 1 for compliers only.

Theorem 3.1

1. $P(G^J) = \frac{P(C)}{P^+(C|Y(0),Y(1))}$.
2. $P(G^J) = 1$ *if and only if* $(Y(0), Y(1)) \perp\!\!\!\perp 1_C$.
3. $P(G^J) = P(C)$ *if and only if* $P^+(C|Y(0), Y(1)) = 1$.

Proof of Theorem 3.1

Let

$$p(y_0, y_1) = \frac{P(C|Y(0) = y_0, Y(1) = y_1)}{P^+(C|Y(0), Y(1))}.$$

It follows from the definition of $P^+(C|Y(0), Y(1))$ that this ratio is between 0 and 1. For every (y_0, y_1) in $\mathcal{S}((Y(0), Y(1))$, let $(B(y_0, y_1))$ denote a collection of Bernoulli variables independent of $(Y(0), Y(1))$, and such that $P(B(y_0, y_1) = 1) = p(y_0, y_1)$. Let also $G^J = \{B(Y(0), Y(1)) = 1\}$. For every (y_0, y_1) in $\mathcal{S}((Y(0), Y(1))$,

$$\begin{aligned}
& P(Y(0) = y_0, Y(1) = y_1, G^J) \\
= \; & P(Y(0) = y_0, Y(1) = y_1, B\left(Y(0), Y(1)\right) = 1) \\
= \; & P(Y(0) = y_0, Y(1) = y_1, B\left(y_0, y_1\right) = 1) \\
= \; & P(Y(0) = y_0, Y(1) = y_1) P(B\left(y_0, y_1\right) = 1) \\
= \; & \frac{P(Y(0) = y_0, Y(1) = y_1, C)}{P^+(C|Y(0), Y(1))}.
\end{aligned} \tag{3.4}$$

Summing or integrating Equation (3.4) over $\mathcal{S}(Y(0), Y(1))$ yields

$$P(G^J) = \frac{P(C)}{P^+(C|Y(0), Y(1))}.$$

Dividing Equation (3.4) by $P(G^J) = \frac{P(C)}{P^+(C|Y(0),Y(1))}$ yields

$$P(Y(0) = y_0, Y(1) = y_1|G^J) = P(Y(0) = y_0, Y(1) = y_1|C).$$

This proves that the largest population with the same distribution of $(Y(0), Y(1))$ as compliers is at least of size $\frac{P(C)}{P^+(C|Y(0),Y(1))}$.

To complete the proof, I show that there cannot exist a population G' larger than G^J and such that

$$P(Y(0) = y_0, Y(1) = y_1|G') = P(Y(0) = y_0, Y(1) = y_1|C). \tag{3.5}$$

Assume such a population exists. Since G' is larger than G^J, we must have

$$P(G') > (1+\varepsilon)\frac{P(C)}{P^+(C|Y(0),Y(1))} \tag{3.6}$$

for some $\varepsilon > 0$. Combining Equations (3.5) and (3.6) yields (after a few computations)

$$P(G'|Y(0)=y_0,Y(1)=y_1) > P(C|Y(0)=y_0,Y(1)=y_1)\frac{1+\varepsilon}{P^+(C|Y(0),Y(1))}. \tag{3.7}$$

Then, it follows from the definition of $P^+(C|Y(0),Y(1))$ that for some (y_0^*, y_1^*) in $\mathcal{S}(Y(0),Y(1))$,

$$P(C|Y(0)=y_0^*,Y(1)=y_1^*) > \frac{P^+(C|Y(0),Y(1))}{1+\varepsilon}.$$

Plugging this into Equation (3.7) finally implies

$$P(G'|Y(0)=y_0^*,Y(1)=y_1^*) > 1,$$

a contradiction. This completes the proof of the first point. The proofs of the second and third points are straightforward.

QED.

Let me illustrate through an example how G^J is constructed. Assume that $Y(0)$ and $Y(1)$ are binary, and assume that the joint distribution of $(Y(1),Y(0),C)$ is as presented in Figure 4. $\frac{2}{3}$ of the $(Y(0)=0,Y(1)=0)$ stratum are compliers. In the three remaining strata, compliers represent a lower fraction of the population. As a result, $P^+(C|Y(0),Y(1)) = P(C|Y(0)=0,Y(1)=0) = \frac{2}{3}$.

Y(0) \ Y(1)	0	1
0	P(Y(0)=0,Y(1)=0)=0.30 P(Y(0)=0,Y(1)=0,C)=0.20	P(Y(0)=0,Y(1)=1)=0.30 P(Y(0)=0,Y(1)=1,C)=0.10
1	P(Y(0)=1,Y(1)=0)=0.20 P(Y(0)=1,Y(1)=0,C)=0.10	P(Y(0)=1,Y(1)=1)=0.20 P(Y(0)=1,Y(1)=1,C)=0.10

Figure 4: Joint distribution of $(Y(1),Y(0),C)$.

To construct G^J, one needs to pick up

$$\frac{P(C|Y(0)=y_0,Y(1)=y_1)}{P(C|Y(0)=0,Y(1)=0)}\%$$

of units in each of the four $(Y(0)=y_0,Y(1)=y_1)$ strata. This amounts to picking up 100% of units in the $(Y(0)=0,Y(1)=0)$ stratum, 50% of them in the $(Y(0)=0,Y(1)=1)$

stratum, and 75% of them in the two remaining strata. This is what the Bernoullis do in the proof. This yields a population G^J such that $P(G^J) = \frac{3}{2}P(C)$. The relative sizes of the $(Y(0) = y_0, Y(1) = y_1)$ strata are the same for G^J as for compliers, which ensures that

$$P(Y(0) = y_0, Y(1) = y_1|G^J) = P(Y(0) = y_0, Y(1) = y_1|C).$$

The distribution of $(Y(1), Y(0), C, G^J)$ resulting from that construction is presented in Figure 5.

Y(0) \ Y(1)	0	1
0	P(Y(0)=0,Y(1)=0)=0.30 P(Y(0)=0,Y(1)=0,C)=0.20 P(Y(0)=0,Y(1)=0,G)=0.30	P(Y(0)=0,Y(1)=1)=0.30 P(Y(0)=0,Y(1)=1,C)=0.10 P(Y(0)=0,Y(1)=1,G)=0.15
1	P(Y(0)=1,Y(1)=0)=0.20 P(Y(0)=1,Y(1)=0,C)=0.10 P(Y(0)=1,Y(1)=0,G)=0.15	P(Y(0)=1,Y(1)=1)=0.20 P(Y(0)=1,Y(1)=1,C)=0.10 P(Y(0)=1,Y(1)=1,G)=0.15

Figure 5: Constructing G^J.

There cannot be a population G' larger than G^J, and such that the joint distribution of potential outcomes is the same as for compliers. Constructing such a population would require picking up more than 100% of units in the $(Y(0) = 0, Y(1) = 0)$ stratum, which is not possible.

$P(G^J) = 1$ if and only if $P^+(C|Y(0), Y(1)) = P(C)$, i.e. if and only if compliance is independent of potential outcomes. This is not a surprising result; if compliers are representative of the entire population, IV results apply to the entire population.

From point 1 of Theorem 3.1, it is also appears that $P(G^J) = P(C)$ if and only if one $(Y(0), Y(1))$ stratum only bears compliers. This result is more interesting; in many applications, it seems unlikely that one $(Y(0), Y(1))$ stratum only bears compliers. In Angrist & Evans (1998), compliers only represent 6% of the total population. Potential outcomes are females participation to the labor market with and without a third child, which are binary variables. As a result, $(Y(0), Y(1))$ only define four strata. It seems unlikely that one of these four strata is only made up of compliers; this would require that one $(Y(0), Y(1))$ stratum bear at most 6% of the total population, which is implausible. This shows that in many applications, instrumental variables estimates apply to a population larger than compliers.

But one cannot use Theorem 3.1 to compute $P(G^J)$ from the data, as $P^+(C|Y(0), Y(1))$ is not point identified under the standard IV assumptions. Computing $P^+(C|Y(0), Y(1))$ would require knowing the distribution of $(Y(0), Y(1), 1_C)$, while under those assumptions the data only (partially) reveals the distributions of $(Y(0), 1_C)$ and $(Y(1), 1_C)$. But $P(G^J)$ and $P(G^M)$ can still be bounded, as shown in the two following subsections.

3.3 Identification of external validity under IV assumptions

Before stating the theorem, I must introduce quite a bit of new notations. Under the standard IV assumptions,

$$\begin{aligned} P(C|Y(1) = y_1) &= \frac{P(Y(1) = y_1, C)}{P(Y(1) = y_1)} \\ &= \frac{P(Y(1) = y_1, D(1) = 1) - P(Y(1) = y_1, D(0) = 1)}{P(Y(1) = y_1, D(1) = 1) + P(Y(1) = y_1, D(1) = 0)} \\ &= \frac{P(Y = y_1, D = 1|Z = 1) - P(Y = y_1, D = 1|Z = 0)}{P(Y = y_1, D = 1|Z = 1) + P(Y(1) = y_1, D(1) = 0)}. \end{aligned}$$

This shows that $P(C|Y(1) = y_1)$ is not point identified, because the data does not reveal $P(Y(1) = y_1, D(1) = 0)$, i.e. the distribution of $Y(1)$ for never takers. But $P(C|Y(1) = y_1)$ can be bounded, as $P(Y(1) = y_1, D(1) = 0)$ is included between 0 and $P(D = 0|Z = 1)$, the size of the population of never takers. Similarly, $P(C|Y(0) = y_0)$ is not point identified, because the data does not reveal the distribution of $Y(0)$ for always takers, but this quantity can also be bounded.

For every $(y_0, y_1) \in \mathcal{S}((Y(0), Y(1)))$, let

$$\begin{aligned} \overline{P}(C|Y(0) = y_0) &= \frac{P(Y = y_0, D = 0|Z = 0) - P(Y = y_0, D = 0|Z = 1)}{P(Y = y_0, D = 0|Z = 0)} \\ \overline{P}(C|Y(1) = y_1) &= \frac{P(Y = y_1, D = 1|Z = 1) - P(Y = y_1, D = 1|Z = 0)}{P(Y = y_1, D = 1|Z = 1)} \end{aligned}$$

be the upper bounds of $P(C|Y(0) = y_0)$ and $P(C|Y(1) = y_1)$ obtained by setting

$$\begin{aligned} P(Y(0) = y_0, D(0) = 1) &= 0 \\ P(Y(1) = y_1, D(1) = 0) &= 0. \end{aligned}$$

For every subset of $\mathcal{S}(Y(0))$ and $\mathcal{S}(Y(1))$, denoted E_0 and E_1, let[6]

$$\begin{aligned}
\underline{P}(C|Y(0) \in E_0) &= \frac{P(Y \in E_0, D = 0|Z = 0) - P(Y \in E_0, D = 0|Z = 1)}{P(Y \in E_0, D = 0|Z = 0) + P(D = 1|Z = 0)} \\
\underline{P}(C|Y(1) \in E_1) &= \frac{P(Y \in E_1, D = 1|Z = 1) - P(Y \in E_1, D = 1|Z = 0)}{P(Y \in E_1, D = 1|Z = 1) + P(D = 0|Z = 1)}
\end{aligned}$$

be the lower bounds of $P(C|Y(0) \in E_0)$ and $P(C|Y(1) \in E_1)$ obtained by setting

$$\begin{aligned}
P(Y(0) \in E_0, D(0) = 1) &= P(D = 1|Z = 0) \\
P(Y(1) \in E_1, D(1) = 0) &= P(D = 0|Z = 1).
\end{aligned}$$

For every d in $\{0; 1\}$, let $E_d^0 = \mathcal{S}(Y(d))$. Then, define by iteration

$$E_d^{m+1} = \{y_d \in \mathcal{S}(Y(d)) : \overline{P}(C|Y(d) = y_d) \geq \underline{P}(C|Y(d) \in E_d^m)\}. \tag{3.8}$$

We have

$$\begin{aligned}
\underline{P}(C|Y(0) \in E_0^0) &= \underline{P}(C|Y(0) \in \mathcal{S}(Y(0))) \\
&= \frac{P(Y \in \mathcal{S}(Y(0)), D = 0|Z = 0) - P(Y \in \mathcal{S}(Y(0)), D = 0|Z = 1)}{P(Y \in \mathcal{S}(Y(0)), D = 0|Z = 0) + P(D = 1|Z = 0)} \\
&= \frac{P(D = 0|Z = 0) - P(D = 0|Z = 1)}{P(D = 0|Z = 0) + P(D = 1|Z = 0)} \\
&= P(C).
\end{aligned}$$

Therefore, Equation (3.8) says that E_0^1 is the set of all the points in $\mathcal{S}(Y(0))$ such that $\overline{P}(C|Y(0) = y_0) \geq P(C)$. Determining E_0^1 requires comparing $\overline{P}(C|Y(0) = y_0)$ and $P(C)$ for every point in $\mathcal{S}(Y(0))$. If $E_0^1 = \mathcal{S}(Y(0))$, the iteration stops. Otherwise, one must compute $\underline{P}(C|Y(0) \in E_0^1)$ in order to determine E_0^2, and so on and so forth.

Finally, for every $d \in \{0; 1\}$, let $E_d^* = \cap_{m \in \mathbb{N}} E_d^m$ be the intersection of all events E_d^m.

Theorem 3.2 *Suppose Assumptions 2.1, 2.2, and 2.3 hold, and only the triple* (Y, D, Z) *is observed. Then,*

$$B_{IV}^- = P(C) \leq P(G^J) \leq P(G^M) \leq B_{IV}^+ = \frac{P(C)}{\max\left(\underline{P}(C|Y(0) \in E_0^*), \underline{P}(C|Y(1) \in E_1^*)\right)}.$$

The bounds are sharp.

[6]The following statement is slightly abusive: when $\mathcal{S}(Y(0))$ and $\mathcal{S}(Y(1))$ are continuous, some subsets of $\mathcal{S}(Y(0))$ and $\mathcal{S}(Y(1))$ might not be events, i.e. elements of the sigma-algebra associated to $\mathcal{S}(Y(0))$ and $\mathcal{S}(Y(1))$.

Under the IV assumptions, the sharp lower bound for $P(G^J)$ and $P(G^M)$ is equal to $P(C)$; $P(G^J)$ and $P(G^M)$ might be greater than $P(C)$, but IV assumptions are not strong enough to reject with certainty that $P(C) = P(G^M) = P(G^J)$.

To understand this point, consider the example of Angrist & Evans (1998). In this paper, the outcome Y is a dummy variable for females participation to the labor market, the treatment D is a dummy for having three children or more, and the instrument Z is a dummy equal to one when the first two children in a couple have the same sex. The joint distributions of Y and D conditional on each value of Z are displayed in Table 1. The first figure in the table should be read as $\widehat{P}(Y = 0, D = 0|Z = 0) = 24.3\%$; among women whose first two children do not have the same sex ($Z = 0$), 24.3% have two children ($D = 0$) and do not work ($Y = 0$).

Table 1: Distribution of Y and D conditional on Z.

$Z = 0$	$Y = 0$	$Y = 1$	$Z = 1$	$Y = 0$	$Y = 1$
$D = 0$	24.3%	38.5%	$D = 0$	21.9%	34.9%
$D = 1$	18.8%	18.4%	$D = 1$	21.9%	21.2%

Under the IV assumptions, the data only reveal the following information on the distribution of $(Y(0), Y(1), 1_C)$:

$$\begin{aligned}
&P(Y(0) = 0, C) = P(Y = 0, D = 0|Z = 0) - P(Y = 0, D = 0|Z = 1) = 24.3\% - 21.9\% = 2.4\% \\
&P(Y(0) = 1, C) = P(Y = 1, D = 0|Z = 0) - P(Y = 1, D = 0|Z = 1) = 38.5\% - 34.9\% = 3.6\% \\
&P(Y(1) = 0, C) = P(Y = 0, D = 1|Z = 1) - P(Y = 0, D = 1|Z = 0) = 21.9\% - 18.8\% = 3.1\% \\
&P(Y(1) = 1, C) = P(Y = 1, D = 1|Z = 1) - P(Y = 1, D = 1|Z = 0) = 21.2\% - 18.4\% = 2.9\% \\
&P(Y(0) = 0) \geq P(Y(0) = 0, D(0) = 0) = P(Y = 0, D = 0|Z = 0) = 24.3\% \\
&P(Y(0) = 1) \geq P(Y(0) = 1, D(0) = 0) = P(Y = 1, D = 0|Z = 0) = 38.5\% \\
&P(Y(1) = 0) \geq P(Y(1) = 0, D(1) = 1) = P(Y = 0, D = 1|Z = 1) = 21.9\% \\
&P(Y(1) = 1) \geq P(Y(1) = 1, D(1) = 1) = P(Y = 1, D = 1|Z = 1) = 21.2\%. \qquad (3.9)
\end{aligned}$$

The IV assumptions only tell that the distribution of $(Y(0), Y(1), 1_C)$ must satisfy Equation (3.9). For instance, we must have

$$P(Y(0) = 0, Y(1) = 0) + P(Y(0) = 0, Y(1) = 1) \geq 24.3\%.$$

As a result, the IV assumptions are not strong enough to reject with certainty that $P(G^J) = P(C)$. The distribution P^* of $(Y(0), Y(1), 1_C)$ presented in Figure 6 satisfies Equation

(3.9). This distribution is implausible; it seems unlikely that the $(Y(0) = 0, Y(1) = 0)$ stratum only represents 2.4% of the total population. But as it satisfies Equation (3.9), the IV assumptions are not sufficient to assert that this cannot be the true distribution of $(Y(0), Y(1), 1_C)$. Under this distribution P^*, it follows from the third point of Theorem 3.1 that IV results only apply to compliers, as the $(Y(0) = 0, Y(1) = 0)$ stratum only bears compliers.

	P*(Y(1)=0)=42.4% P*(Y(1)=0,C)=3.1%	P*(Y(1)=1)=57.6% P*(Y(1)=1,C)=2.9%
P*(Y(0)=0)=32.4% P*(Y(0)=0,C)=2.4%	P*(Y(0)=0,Y(1)=0)=2.4% P*(Y(0)=0,Y(1)=0,C)=2.4%	P*(Y(0)=0,Y(1)=1)=30% P*(Y(0)=0,Y(1)=1,C)=0%
P*(Y(0)=1)=67.6% P*(Y(0)=1,C)=3.6%	P*(Y(0)=1,Y(1)=0)=40% P*(Y(0)=1,Y(1)=0,C)=0.7%	P*(Y(0)=1,Y(1)=1)=27.6% P*(Y(0)=1,Y(1)=1,C)=2.9%

Figure 6: P^* is consistent with the IV assumptions, and $P^*(G^J) = P^*(C)$.

The theorem does not say that $P(G^J) = P(C)$, it only says that the IV assumptions are not strong enough to reject with certainty that $P(G^J) = P(C)$. IV assumptions alone are never sufficient to reject that IV estimates only apply to compliers. Increasing the external validity of IV estimates requires introducing supplementary assumptions; there is no such thing as a free lunch. But under a reasonably credible assumption, it is possible to discard implausible distributions of the data such as P^*, so as to increase the external validity of IV estimates, as shown in the next subsection.

Contrary to B^-_{IV}, B^+_{IV} is not trivial as it can be lower than 1. When $B^+_{IV} < 1$, IV estimates do not apply to the entire population; IV assumptions are sufficient to reject that the Wald ratio is equal to the ATE in some applications.

Testing whether the Wald ratio is equal to the ATE relies on an intuitive condition. $B^+_{IV} = 1$ is equivalent to

$$\begin{aligned} \underline{P}(C|Y(0) = y_0) \leq P(C) \leq \overline{P}(C|Y(0) = y_0) \\ \underline{P}(C|Y(1) = y_1) \leq P(C) \leq \overline{P}(C|Y(1) = y_1). \end{aligned} \tag{3.10}$$

By construction, we also have

$$\begin{aligned} \underline{P}(C|Y(0) = y_0) \leq P(C|Y(0) = y_0) \leq \overline{P}(C|Y(0) = y_0) \\ \underline{P}(C|Y(1) = y_1) \leq P(C|Y(1) = y_1) \leq \overline{P}(C|Y(1) = y_1). \end{aligned}$$

Therefore, when equation (3.10) is satisfied, we cannot reject $P(C|Y(0) = y_0) = P(C)$ and $P(C|Y(1) = y_1) = P(C)$, which implies that we cannot reject $1_C \perp\!\!\!\perp (Y(0), Y(1))$. On the contrary, when Equation (3.10) is not satisfied, we can reject $1_C \perp\!\!\!\perp (Y(0), Y(1))$.

B_{IV}^{+} is obtained in two steps. First, reasoning as in the proof of Theorem 3.1, one can show that $\frac{P(C)}{P^{+}(C|Y(0))}$ is the size of the largest population G^0 such that

$$P(Y(0) = y_0|G^0) = P(Y(0) = y_0|C). \tag{3.11}$$

Similarly, $\frac{P(C)}{P^{+}(C|Y(1))}$ is the size of the largest population G^1 such that

$$P(Y(1) = y_1|G^1) = P(Y(1) = y_1|C). \tag{3.12}$$

Since G^M must verify both Equation (3.11) and Equation (3.12), this implies that G^M must be smaller than G^0 and G^1. As a result, we must have

$$P(G^M) \leq \frac{P(C)}{\max(P^{+}(C|Y(0)), P^{+}(C|Y(1)))}.$$

Second, I show that $\underline{P}(C|Y(0) \in E_0^*)$ and $\underline{P}(C|Y(1) \in E_1^*)$ are sharp lower bounds of $P^{+}(C|Y(0))$ and $P^{+}(C|Y(1))$, which yields the result.

When $\mathcal{S}(Y(0))$ and $\mathcal{S}(Y(1))$ are finite, computing B_{IV}^{+} requires performing at most $k(k+1)$ pairwise comparisons, where k is the size of $\mathcal{S}(Y(0))$ and $\mathcal{S}(Y(1))$. Indeed, to compute $\underline{P}(C|Y(0) \in E_0^*)$, the first step is to determine E_0^1. To that end, one must compare $\overline{P}(C|Y(0) = y_0)$ to $\underline{P}(C|Y(0) \in \mathcal{S}(Y(0)))$ for every y_0 in $\mathcal{S}(Y(0))$. This amounts to performing k pairwise comparisons. If $E_0^1 = E_0^0 = \mathcal{S}(Y(0))$, the algorithm stops and

$$\underline{P}(C|Y(0) \in E_0^*) = \underline{P}(C|Y(0) \in \mathcal{S}(Y(0))) = P(C).$$

Otherwise, one must compare $\overline{P}(C|Y(0) = y_0)$ to $\underline{P}(C|Y(0) \in E_0^1)$ for every y_0 in E_0^1 so as to determine E_0^2. This requires performing at most $k-1$ pairwise comparisons. As a result, the algorithm bears at most k steps, and requires performing at most $\frac{k(k+1)}{2}$ pairwise comparisons.

When $\mathcal{S}(Y(0))$ is infinite, computing B_{IV}^{+} is more complicated. But I show that $\underline{P}(C|Y(0) \in E_0^m)$ and $\underline{P}(C|Y(0) \in E_1^m)$ are increasing sequences, converging respectively towards $\underline{P}(C|Y(0) \in E_0^*)$ and $\underline{P}(C|Y(1) \in E_1^*)$. Therefore, for m large enough, one can use

$$\frac{P(C)}{\max\left(\underline{P}(C|Y(0) \in E_0^m), \underline{P}(C|Y(1) \in E_1^m)\right)}$$

as an approximation of B_{IV}^{+}.

Overall, Theorem 3.2 shows that standard IV assumptions are not sufficient to rule out that IV estimates only apply to a population of size $P(C)$. On the contrary, they are sufficient to reject that IV estimates apply to the entire population.

3.4 Identification of external validity under a strong instrument assumption

I study now an assumption under which it is possible to extend the external validity of IV estimates, and which is weaker than other conditions previously used in the literature. This assumption requires that there be covariates X such that $(Y(d), X)$ is a stronger determinant of compliance than $(Y(0), Y(1))$.

Assumption 3.1 *(Strong "instrument")*
For every $d \in \{0; 1\}$,

$$P^+(C|Y(0), Y(1)) \leq P^+(C|Y(d), X). \tag{3.13}$$

$P^+(C|Y(0), Y(1))$ is the largest value of $P(C|Y(0) = y_0, Y(1) = y_1)$, while $P^+(C|Y(d), X)$ is the largest value of $P(C|Y(d) = y_d, X = x)$. Therefore, Assumption 3.1 requires that X induce more variation in the share of compliers than $Y(1-d)$. X should be a stronger "determinant" of compliance than $Y(1-d)$.

This strong instrument condition is somewhat unusual in the literature. To ease the comparison of Assumption 3.1 with standard assumptions, I consider a condition which is sufficient for Assumption 3.1 to hold, and which is easier to relate to conditions used previously in the literature.

Assumption 3.2 *(Conditional independence)*
For every $d \in \{0; 1\}$,

$$Y(1-d) \perp\!\!\!\perp 1_C | Y(d), X. \tag{3.14}$$

Lemma 3.1 *Assumption 3.2 is stronger than Assumption 3.1.*

Proof of Lemma 3.1:
Assume that Assumption 3.2 holds. $P(C|Y(0) = y_0, Y(1) = y_1)$ is equal to the expectation of $P(C|Y(0) = y_0, Y(1) = y_1, X = x)$ over the distribution of $X|Y(0) = y_0, Y(1) = y_1$. As a result, the largest value of $P(C|Y(0) = y_0, Y(1) = y_1)$ must necessarily be smaller than the largest value of $P(C|Y(0) = y_0, Y(1) = y_1, X = x)$. This implies that

$$P^+(C|Y(0), Y(1)) \leq P^+(C|Y(0), Y(1), X). \tag{3.15}$$

Then, Assumption 3.2 implies that

$$P(C|Y(0) = y_0, Y(1) = y_1, X = x) = P(C|Y(d) = y_d, X = x),$$

which in turn implies that

$$P^+(C|Y(0), Y(1), X) = P^+(C|Y(d), X). \tag{3.16}$$

Combining Equations (3.15) and (3.16) proves that Assumption 3.1 is satisfied.
QED.

Assumption 3.2 states that compliers and non compliers with the same covariates and the same $Y(0)$ must also have the same distribution of $Y(1)$. Putting it in other words, compliers and non compliers with the same covariates and the same $Y(0)$ must also have the same distribution of treatment effects.

This conditional independence condition is not innocuous, but it is weaker than several assumptions which have been used in the literature. It is weaker than the "conditional effect ignorability" assumption used by Angrist & Fernandez-Val (2010) to extrapolate the LATE. Their condition states that compliers and non compliers with the same covariates must have the same treatment effects. It is also weaker than the rank invariance assumption in Chernozhukov & Hansen (2005). They require that $Y(d) = h_d(U, X)$, with h_d increasing in U. This implies that

$$U = h_0^{-1}(Y(0); X) = h_1^{-1}(Y(1); X),$$

which in turn implies that

$$\begin{aligned} Y(0) &= h_0(h_1^{-1}(Y(1); X), X) \\ Y(1) &= h_1(h_0^{-1}(Y(0); X), X). \end{aligned}$$

Conditional on $Y(0)$ and X, $Y(1)$ is a constant, independent of the compliance dummy. Conversely, conditional on $Y(1)$ and X, $Y(0)$ is also a constant, independent of the compliance dummy. As a result, Assumption 3.2 holds.

Assumption 3.1 requires that X be a strong determinant of potential treatments. Equation (3.14) and the comparison with the rank invariance assumption of Chernozhukov & Hansen (2005) also show that Assumption 3.1 is more likely to be satisfied if X is a strong determinant of potential outcomes; the less variation there is in $Y(d)$ after conditioning for X, the more likely it is that Equation (3.14) holds. As a result, when looking for a strong instrument, one should look for a vector of covariates X strongly correlated to both the outcome and the treatment.

Before stating the theorem, I must introduce new notations. $P(C|Y(0) = y_0, X = x)$ is not identified, because the data does not reveal the distribution of $(Y(0), X)$ for always takers. Similarly, $P(C|Y(1) = y_1, X = x)$ is not identified, because the data does not reveal the distribution of $(Y(1), X)$ for never takers. As a result, $P^+(C|Y(0), X)$ and $P^+(C|Y(1), X)$ are not identified. However, those quantities can be bounded, following the same steps as in the previous subsection.

Let

$$\begin{aligned}
\overline{P}(C|Y(0)=y_0, X=x) &= \frac{P(Y=y_0, X=x, D=0|Z=0) - P(Y=y_0, X=x, D=0|Z=1)}{P(Y=y_0, X=x, D=0|Z=0)} \\
\overline{P}(C|Y(1)=y_1, X=x) &= \frac{P(Y=y_1, X=x, D=1|Z=1) - P(Y=y_1, X=x, D=1|Z=0)}{P(Y=y_1, X=x, D=1|Z=1)}
\end{aligned}$$

be the upper bounds of $P(C|Y(0)=y_0, X=x)$ and $P(C|Y(1)=y_1, X=x)$ which are obtained setting

$$\begin{aligned}
&P(Y(0)=y_0, X=x, D(0)=1)=0 \\
&P(Y(1)=y_1, X=x, D(1)=0)=0.
\end{aligned}$$

Let

$$\begin{aligned}
\overline{P}^{+}(C|Y(0), X) &= \sup_{(y_0,x)\in\mathcal{S}((Y(0),X))} \{\overline{P}(C|Y(0)=y_0, X=x)\} \\
\overline{P}^{+}(C|Y(1), X) &= \sup_{(y_1,x)\in\mathcal{S}((Y(1),X))} \{\overline{P}(C|Y(1)=y_1, X=x)\}
\end{aligned}$$

be the corresponding upper bounds for $P^{+}(C|Y(0), X)$ and $P^{+}(C|Y(1), X)$.

Theorem 3.3 *Suppose Assumptions 2.1, 2.2, 2.3, and 3.2 hold. Then,*

$$B_{CI}^{-} = \frac{P(C)}{\min\left(\overline{P}^{+}(C|Y(0), X), \overline{P}^{+}(C|Y(1), X)\right)} \leq P(G^J) \leq P(G^M).$$

Proof of Theorem 3.3

Combining the first point of Theorem 3.1 with Assumption 3.1 yields

$$\frac{P(C)}{P^{+}(C|Y(d), X)} \leq P(G^J)$$

for every $d \in \{0; 1\}$. Then it suffices to notice that

$$P^{+}(C|Y(d), X) \leq \overline{P}^{+}(C|Y(d), X)$$

to obtain the result.

QED.

Under Assumption 3.1 IV estimates apply to a population which represents at least

$$\frac{P(C)}{\min\left(\overline{P}^{+}(C|Y(0), X), \overline{P}^{+}(C|Y(1), X)\right)}\%$$

of the total population.

3.5 Inference

I draw inference on $P^+(C|Y(0))$ and $P^+(C|Y(1))$ when $\mathcal{S}(Y(0))$ and $\mathcal{S}(Y(1))$ are finite.[7] Inference in the continuous case is left for future work.

Even with $\mathcal{S}(Y(d))$ finite, inference is not straightforward because the limit distributions of

$$\sqrt{n}\left(\underline{P}^+\widehat{(C|Y}(d)) - \underline{P}^+(C|Y(d))\right)$$
$$\sqrt{n}\left(\overline{P}^+\widehat{(C|Y}(d)) - \overline{P}^+(C|Y(d))\right)$$

are discontinuous functions of the data distribution, which raises difficulties (see Hirano & Porter, 2009). But $P^+(C|Y(d))$ satisfies a moment inequality model as in Andrews & Soares (2010); it is possible to draw inference on $P^+(C|Y(d))$ using their method.

Theorem 3.4 *Suppose that for every $d \in \{0;1\}$, $P(Z = d) \geq \varepsilon$ for some $\varepsilon > 0$. Then, $P^+(C|Y(d))$ defines a moment inequality model verifying all the technical conditions defined in Andrews & Soares (2010). As a result, one can one use their results to draw inference on $P^+(C|Y(d))$.*

The proof of this theorem is straightforward, even though it requires introducing cumbersome notations. It amounts to verifying that all the conditions defined in Andrews & Soares (2010) apply. Then, confidence intervals are obtained inverting a test statistic defined by those authors. A formula for this test statistic is presented in the proofs.

4 Application

I use the results obtained above to re-analyze the internal and external validity of the findings in Angrist & Evans (1998). Their main outcome Y is a dummy variable for females participation to the labor market. The 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 couple have the same sex. I use the 1980 PUMS sample, which bears 394 840 observations.

[7] Inference on $P^+(C|Y(0), X)$ and $P^+(C|Y(1), X)$ can be conducted similarly when $\mathcal{S}(X)$ is finite as well.

4.1 Internal validity

As mentioned in the introduction, there might be defiers in this application. When their first two children are girls, the share of females having a third child is 1.5 percentage points higher than when their first two children are boys, and the difference is significant (P-value < 0.001). 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 girl, but might decide otherwise if their first two children are boys. Such parents would be defiers.

I use the results of Section 2 to assess whether this is a serious threat to the internal validity of the results. To that end, the joint distributions of Y and D conditional on each value of Z are displayed in Table 2. The first figure in the table should be read as $\widehat{P}(Y = 0, D = 0|Z = 0) = 24.3\%$; among women whose first two children do not have the same sex ($Z = 0$), 24.3% have two children ($D = 0$) and do not work ($Y = 0$).

Table 2: Distribution of Y and D conditional on Z.

$Z = 0$	$Y = 0$	$Y = 1$	$Z = 1$	$Y = 0$	$Y = 1$
$D = 0$	24.3%	38.5%	$D = 0$	21.9%	34.9%
$D = 1$	18.8%	18.4%	$D = 1$	21.9%	21.2%

For every (y_0, y_1) in $\{0; 1\}^2$, we have

$$\begin{aligned} \widehat{P}(Y = y_0, D = 0|Z = 1) &\leq \widehat{P}(Y = y_0, D = 0|Z = 0) \\ \widehat{P}(Y = y_1, D = 1|Z = 0) &\leq \widehat{P}(Y = y_1, D = 1|Z = 1), \end{aligned} \tag{4.1}$$

i.e. the green figures are always greater than their red counterparts. I test the four opposite inequalities. The lowest F-statistic is 305.88 and the four Holm-Bonferronni adjusted P-values are lower than 0.0001.

As Equation (4.1) is not rejected, there are more compliers than defiers in each $Y(0)$ and $Y(1)$ stratum. Indeed, as shown in Lemma 2.2, Equation (4.1) is equivalent to

$$\begin{aligned} \widehat{P}(Y(0) = y_0, F) &\leq \widehat{P}(Y(0) = y_0, C) \\ \widehat{P}(Y(1) = y_1, F) &\leq \widehat{P}(Y(1) = y_1, C). \end{aligned}$$

Therefore, we can be reasonably confident that there are also more compliers than defiers in each $(Y(0), Y(1))$ stratum. Even if there may be defiers in the sample, it seems that this is not a threat to the internal validity of the results.

4.2 External validity

Estimates in this paper have low external validity. The fraction of women having three children or more is 6.0 percentage points higher among women whose first two children have the same sex, than among women whose first two children are of a different sex. This implies that compliers only represent 6% of the total population. I use now the results of Section 3 to assess whether the external validity of those estimates can be extended or not.

First, the upper and lower bounds of $P(C|Y(d) = y)$ are displayed in Table 3. For every d and y,

$$\widehat{P}(C) \in \left(\underline{\widehat{P}}(C|Y(d) = y), \overline{\widehat{P}}(C|Y(d) = y)\right).$$

One cannot reject this equation for the four possible values of d and y; the lowest F-statistic is equal to 58.10, and the four Holm-Bonferronni adjusted P-values are lower than 0.0001.

Table 3: Bounds for $P(C|Y(d) = y)$.

$P(C\|Y(0) = 0) \in [3.8\%, 9.7\%]$	$P(C\|Y(1) = 0) \in [4.0\%, 14.3\%]$
$P(C\|Y(0) = 1) \in [4.8\%, 9.4\%]$	$P(C\|Y(1) = 1) \in [3.6\%, 13.3\%]$

As a result, $\widehat{B}^{+}_{IV} = 1$; we cannot reject that estimates in this paper apply to the entire population.

Second, I estimate B^{-}_{CI}. To that end, I must choose a vector of covariates X strongly correlated to the decision of having a third kid, and to mothers participation to the labor market. On top of data on fertility, the data set includes information on parents education, labor market participation, income, and ethnicity. It appears from simple OLS regressions displayed in Table 4 that among those variables, mothers years of education and ethnicity are very strong predictors of their propensity to have a third child. Fathers income, mothers education, and mothers ethnicity are very strong predictors of mothers participation to the labor market. T-stats are displayed between parenthesis.

Table 4: Determinants of D(z) and Y(d).

	Mother has a third child	Mother is employed
Logarithm of father's income	0.001	−0.009
	(8.21)	(−44.20)
Mother's years of education	−0.029	0.020
	(−92.16)	(60.23)
Mother white	−0.096	−0.069
	(−45.72)	(−32.47)
Mother has a third child		−0.113
		(−70.13)
First two children have same sex	0.059	
	(38.73)	
R^2	0.03	0.03
N	394 840	394 840

X should also be discrete, and there should be a sufficient number of observations in each (Y, D, X) stratum to ensure that $\widehat{\overline{P}}(C|Y = y, D = d, X = x)$ is well estimated in each stratum.

I consider a first set of covariates X_1. X_1 includes 10 dummies for father income (one for each decile), and 3 dummies for mother education (less than high school, high school graduate, more than high school) which partition the population into three thirds. This yields 30 different strata, each of them bearing at least 791 women.

X_1 induces substantial variation in $P(C|X_1 = x_1)$. The lowest value of $\widehat{P}(C|X_1 = x_1)$ in those 30 strata is 3.3%, while its highest value is 8.7%. $\widehat{P}^+(C|X_1) = 8.7\%$ is almost equal to $\widehat{\overline{P}}^+(C|Y(0)) = 9.7\%$. Because there are many always takers ($P(D = 1|Z = 0) = 33.2\%$), $\widehat{\overline{P}}^+(C|Y(0))$ is a fairly conservative upper bound of $P^+(C|Y(0))$. Overall, this suggests that X_1 is at least as strong a determinant of compliance status as $Y(0)$.

$\widehat{B}^-_{CI^1}$ is larger than $\widehat{P}(C)$. The 120 (Y, D, X_1) strata all bear at least 145 females. Across those 120 strata, the maximum value of $\widehat{\overline{P}}(C|Y = y, D = d, X_1 = x_1)$ is 25.7%, which yields $\widehat{B}^-_{CI^1} = \frac{6.0}{25.7} = 23.2\%$. This quantity is well estimated, as the stratum in which $\widehat{\overline{P}}(C|Y = y, D = d, X_1 = x_1)$ reaches its maximum value bears 1124 females. The 95% confidence interval for $B^-_{CI^1}$ is $[15.1\%; 31.2\%]$. This confidence interval is obtained using Andrews & Soares (2010)'s method described in the preceding section.

As only 18% of mothers are not white in the sample, one gets some very small strata when combining this variable with father income and mother education. I consider an

alternative set of covariates, in which I introduce the ethnicity dummy while reducing the number of dummies for father income. X_2 includes 5 dummies for father income (one for each quintile), 3 dummies for mother education, and the ethnicity dummy. The 30 X_2 strata all bear at least 771 females. The lowest value of $\widehat{P}(C|X_1 = x_1)$ is -0.004%, while its highest value is 10.1%; X_2 induces even more variation in the proportion of compliers than X_1.

$\widehat{B}^-_{CI^2}$ is also larger than $\widehat{P}(C)$. The 120 (Y, D, X_2) strata all bear at least 150 females. Across those strata, the maximum value of $\widehat{\overline{P}}(C|Y = y, D = d, X_2 = x_2)$ is 36.4%, which yields $\widehat{B}^-_{CI^2} = \frac{6.0}{36.4} = 16.4\%$. The stratum in which $\widehat{\overline{P}}(C|Y = y, D = d, X_1 = x_1)$ reaches its maximum value only bears 165 females, which implies that there is much more sampling variance in $\widehat{B}^-_{CI^2}$ than in $\widehat{B}^-_{CI^1}$. The 95% confidence interval for $B^-_{CI^1}$ is $[7.0\%; 25.7\%]$.

Results are summarized in Table 5.

Table 5: Bounds for $P(G^J)$ and $P(G^M)$.

$\widehat{B}^+_{IV}$	100%
$\widehat{B}^-_{CI^1}$	23.15%
$\widehat{B}^-_{CI^2}$	16.36%

5 Concluding comments

In the instrumental variable model with heterogeneous treatment effects, the monotonicity condition can be replaced by a substantially weaker assumption. This assumption requires that there be more compliers than defiers conditional on treatment effects. Under this condition, treatment effects are identified among a sub-population of compliers C_V, of same size as the population of compliers under monotonicity.

IV estimates also apply to a larger population than compliers. The size of this larger population G is not identified. I give a first condition under which $P(G) = 1$, and a second one under which $P(G) = P(C)$. Both conditions seem restrictive; in most applications, $P(G)$ should lie somewhere in-between $P(C)$ and 1. I derive sharp bounds for $P(G)$ under the standard IV assumptions. The lower bound is trivial since it is equal to $P(C)$; the upper bound is not trivial. Under a supplementary assumption, I derive a non trivial lower bound for $P(G)$.

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.

Andrews, D. W. K. & Soares, G. (2010), 'Inference for parameters defined by moment inequalities using generalized moment selection', *Econometrica* **78**(1), 119–157.

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

Angrist, J. & Fernandez-Val, I. (2010), Extrapolate-ing: External validity and overidentification in the late framework, Working Paper 16566, National Bureau of Economic Research.

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.

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

Chernozhukov, V. & Hansen, C. (2005), 'An iv model of quantile treatment effects', *Econometrica* **73**(1), 245–261.

Deaton, A. S. (2009), Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development, NBER Working Papers 14690, National Bureau of Economic Research, Inc.

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

Freedman, D. (2006), 'Statistical models for causation: what inferential leverage do they provide?', *Evaluation review* **30**(6), 691.

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.

Heckman, J. J. (2010), 'Building bridges between structural and program evaluation approaches to evaluating policy', *Journal of Economic Literature* **48**(2), 356–98.

Heckman, J. J. & Urzúa, S. (2010), 'Comparing iv with structural models: What simple iv can and cannot identify', *Journal of Econometrics* **156**(1), 27 – 37.

Heckman, J. J. & Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation', *Econometrica* **73**(3), 669–738.

Heckman, J., Tobias, J. L. & Vytlacil, E. (2003), 'Simple estimators for treatment parameters in a latent-variable framework', *The Review of Economics and Statistics* **85**(3), 748–755.

Hirano, K. & Porter, J. (2009), Impossibility results for nondifferentiable functionals, Mpra paper, University Library of Munich, Germany.

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

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, 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.

Appendix 1: Inference on $\underline{P}(G)$

For every $i \in \{0;1\}$, let $(Y = y_*, D = d_*, X_i = x_{i*})$ denote the cell in which the maximum value of

$$\widehat{\overline{P}}(C|Y = y, D = d, X_i = x_i)$$

is attained. For every other possible values of y, d, x_i, we have

$$\widehat{\overline{P}}(C|Y = y, D = d, X_i = x_i) + \sqrt{\frac{2\ln(\ln(n))}{n}} < \widehat{\overline{P}}(C|Y = y_*, D = d_*, X_i = x_{i*}),$$

where n denotes the number of observations in the sample. Following Andrews & Soares (2010), this means that one can draw inference on $B^-_{CI^i}$ "as if" we had

$$B^-_{CI^i} = \frac{P(C)}{\overline{P}(C|Y = y*, D = d*, X_i = x*_i)}.$$

Then, it is straightforward to establish the asymptotic normality of

$$\sqrt{n}\left(\widehat{B}^-_{CI^i} - B^-_{CI^i}\right)$$

through simple delta methods.

Appendix 2: Proofs

Proof of Theorem 2.2:

I start proving the first point. Under Assumption 2.5, $P(Y(0) = y_0, Y(1) = y_1, C) = 0$ implies that $P(Y(0) = y_0, Y(1) = y_1, F) = 0$ as well. Therefore, when $P(Y(0) = y_0, Y(1) = y_1, C) = 0$,

$$p(y_0, y_1) = \frac{P(Y(0) = y_0, Y(1) = y_1, F)}{P(Y(0) = y_0, Y(1) = y_1, C)}$$

should be understood as 0. Assumption 2.5 also ensures that this ratio is smaller than 1.

For every (y_0, y_1) in $\mathcal{S}((Y(0), Y(1)))$, let $(B(y_0, y_1))$ denote a collection of Bernoulli variables independent of $(Y(0), Y(1), D(0), D(1))$, and such that $P(B(y_0, y_1) = 1) = p(y_0, y_1)$. Let also $C_F = \{C, B(Y(0), Y(1)) = 1\}$. Using the same steps as in the proof of Theorem 2.1, one can show that

$$P(Y(0) = y_0, Y(1) = y_1, C_F) = P(Y(0) = y_0, Y(1) = y_1, F). \tag{5.1}$$

Summing or integrating Equation (5.1) over $\mathcal{S}((Y(0), Y(1)))$ yields $P(C_F) = P(F)$. Dividing Equation (2.5) by $P(C_F) = P(F)$ yields

$$P(Y(0) = y_0, Y(1) = y_1 | C_F) = P(Y(0) = y_0, Y(1) = y_1 | F).$$

Integrating or summing this last equation over $\mathcal{S}(Y(1))$ (resp. $\mathcal{S}(Y(0))$) proves the second (resp. the third) statement on C_F. This completes the proof of the first point.

I prove now the second point. Let

$$C_V = C \setminus C_F = \{C, B(Y(0), Y(1)) = 0\}.$$

As in Theorem 2.1, one can show that $P(C_V) = P(D = 1|Z = 1) - P(D = 1|Z = 0)$.

Integrating or summing Equation (5.1) over $\mathcal{S}(Y(1))$ yields

$$P(Y(0) = y_0, C_F) = P(Y(0) = y_0, F).$$

From the definition of C_V and this last equation, it follows that

$$\begin{aligned} P(Y(0) = y_0, C_V) &= P(Y(0) = y_0, C) - P(Y(0) = y_0, C_F) \\ &= P(Y(0) = y_0, C) - P(Y(0) = y_0, F) \\ &= P(Y(0) = y_0, D(0) = 0) - P(Y(0) = y_0, D(1) = 0) \\ &= P(Y = y_0, D = 0|Z = 0) - P(Y = y_0, D = 0|Z = 1). \end{aligned} \tag{5.2}$$

The third equality follows from the law of total probability. The fourth is obtained under Assumptions 2.1 and 2.2. Notice that $P(D = 1|Z = 1) - P(D = 1|Z = 0)$ is equal to $P(D = 0|Z = 0) - P(D = 0|Z = 1)$. Therefore, dividing each side of (5.2) by

$$P(C_V) = P(D = 0|Z = 0) - P(D = 0|Z = 1)$$

yields the second equation of point 2. One can obtain the third equation similarly.

Finally, to obtain the result on the average treatment effect, one can merely follow the same steps as for the proof of Theorem 2.1.

QED.

Proof of Theorem 2.3

I start proving the first point.

$$\begin{aligned} P(F|Y(0) = y_0, Y(1) = y_1) &= P(V_0 \geq v_0, V_1 < v_1|Y(0) = y_0, Y(1) = y_1) \\ &= P(V_1 \geq v_0, V_0 < v_1|Y(0) = y_0, Y(1) = y_1) \\ &\leq P(V_1 \geq v_1, V_0 < v_0|Y(0) = y_0, Y(1) = y_1) \\ &= P(C|Y(0) = y_0, Y(1) = y_1). \end{aligned}$$

The second equality holds because of the exchangeability assumption. $v_0 \geq v_1$ implies that $\{V_1 \geq v_0, V_0 < v_1\} \subseteq \{V_1 \geq v_1, V_0 < v_0\}$, which proves the inequality. This proves a).

Then,

$$\begin{aligned} P(BC|Y(0) = y_0, Y(1) = y_1) &= P(V_0 < v_1, V_1 \geq v_0|Y(0) = y_0, Y(1) = y_1) \\ &= P(V_1 < v_1, V_0 \geq v_0|Y(0) = y_0, Y(1) = y_1) \\ &= P(F|Y(0) = y_0, Y(1) = y_1). \end{aligned}$$

The second equality holds because of the exchangeability assumption. Multiplying both sides of this equality by $P(Y(0) = y_0, Y(1) = y_1)$ implies

$$P(Y(0) = y_0, Y(1) = y_1, BC) = P(Y(0) = y_0, Y(1) = y_1, F). \tag{5.3}$$

Integrating or summing Equation (5.3) over $\mathcal{S}((Y(0), Y(1)))$ implies that $P(BC) = P(F)$ which proves the first equation of point b).

Then, dividing Equation (5.3) by $P(BC) = P(F)$ yields

$$P(Y(0) = y_0, Y(1) = y_1|BC) = P(Y(0) = y_0, Y(1) = y_1|F).$$

Integrating or summing this last equation over $\mathcal{S}(Y(1))$ (resp. $\mathcal{S}(Y(0))$) yields the second (resp. the third) equation of point b). This completes the proof of point b).

This also proves c): BC verifies all the points listed in the first point of Theorem 2.2. As a result, one can show that the remaining part of compliers verifies all the equations listed in the second point of this theorem, which correspond to those listed in point c).

Now, I prove the second point. The proof is a generalization of the proof of the first point of Proposition 2.1 in Chaisemartin & D'Haultfoeuille (2012). Assume that Assumption 2.5 holds. Let $v_z = P(D(z) = 0)$. Let

$$\begin{aligned}
t_0 &= 0 \\
t_1 &= P(D(0) = 0, D(1) = 0) \\
t_2 &= P(D(1) = 0) \\
t_3 &= P(D(0) = 0) \\
t_4 &= 1 - P(D(0) = 1, D(1) = 1) \\
t_5 &= 1
\end{aligned}$$

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 $(Y(0), Y(1), D(0), D(1))$ and of $(B(y_0, y_1))$, the family of Bernoulli random variables defined in the proof of Theorem 2.2. Take W_i uniform on $[t_i, t_{i+1}]$. Finally, let

$$\begin{aligned}
V_0 &= (1 - D(0))(1 - D(1))W_0 + (1 - D(0))D(1)(B(Y(0), Y(1))W_1 + (1 - B(Y(0), Y(1)))W_2) \\
&+ D(0)(1 - D(1))W_3 + D(0)D(1)W_4 \\
V_1 &= (1 - D(0))(1 - D(1))W_0 + (1 - D(0))D(1)(B(Y(0), Y(1))W_3 + (1 - B(Y(0), Y(1)))W_2) \\
&+ D(0)(1 - D(1))W_1 + D(0)D(1)W_4.
\end{aligned}$$

By construction, $D(z) = 1\{V_z \leq v_z\}$ almost surely, so it suffices to check that (V_0, V_1) are exchangeable conditional on $(Y(0), Y(1))$. This amounts to checking that for every $v \leq v'$,

$$P(V_0 = v, V_1 = v' | Y(0) = y_0, Y(1) = y_1) = P(V_0 = v, V_1 = v' | Y(0) = y_0, Y(1) = y_1).$$

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

$$P(V_0 = v, V_1 = v' | Y(0) = y_0, Y(1) = y_1) = P(V_0 = v, V_1 = v' | Y(0) = y_0, Y(1) = y_1) = 0,$$

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

$$\begin{aligned}
& P(V_0 = v, V_1 = v' | Y(0) = y_0, Y(1) = y_1) \\
= \; & P(C | Y(0) = y_0, Y(1) = y_1) P(B(y_0, y_1) = 1) P(W_1 = v) P(W_3 = v') \\
= \; & P(F | Y(0) = y_0, Y(1) = y_1) P(W_3 = v') P(W_1 = v) \\
= \; & P(V_0 = v', V_1 = v | Y(0) = y_0, Y(1) = y_1).
\end{aligned}$$

The first equality follows from the definition of V_0 and V_1. 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 $B(y_0, y_1) = 1$. 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.

Proof of Theorem 2.4:
Assumption 2.7 ensures that

$$\frac{P(Y(0) = y_0, Y(1) = y_1, X = x, F)}{P(Y(0) = y_0, Y(1) = y_1, X = x, C)}$$

is included between 0 and 1. For every (y_0, y_1, x) in $\mathcal{S}((Y(0), Y(1), X)$, let $(B(y_0, y_1, x))$ denote a family of Bernoulli variables independent of $(Y(0), Y(1), D(0), D(1), X)$ and such that

$$P(B(y_0, y_1, x) = 1) = \frac{P(Y(0) = y_0, Y(1) = y_1, X = x, F)}{P(Y(0) = y_0, Y(1) = y_1, X = x, C)}.$$

Then, let $C_F = \{C, B(Y(0), Y(1), X) = 1\}$. Using the same steps as in the proof of Theorem 2.2, one can show that for every (y_0, y_1, x) in $\mathcal{S}((Y(0), Y(1), X)$,

$$P(Y(0) = y_0, Y(1) = y_1, X = x, C_F) = P(Y(0) = y_0, Y(1) = y_1, X = x, F), \tag{5.4}$$

which in turn implies that

$$\begin{aligned} &P(C_F) = P(F) \\ &P(Y(0) = y_0, Y(1) = y_1, X = x | C_F) = P(Y(0) = y_0, Y(1) = y_1, X = x | F). \end{aligned} \tag{5.5}$$

This proves the first point.

Now, I prove the second point. Let

$$C_V = C \setminus C_F = \{C, B(Y(0), Y(1), X) = 0\}.$$

Integrating or summing the second equality of Equation (5.5) over $\mathcal{S}((Y(1), X))$ and over $\mathcal{S}((Y(0), X))$ proves that C_F verifies all the equations listed in point 1 of Theorem 2.2. This implies that C_V verifies the four equations listed in point 2 of this theorem. As a result, the only equation left to be proven is

$$P(X = x | C_V) = P_X(x).$$

Integrating or summing Equation (5.4) over $\mathcal{S}((Y(0), Y(1)))$ yields

$$P(X = x, C_F) = P(X = x, F).$$

Then, it follows from the definition of C_V and from this last equation that

$$P(X = x, C_V) = P(X = x, C) - P(X = x, C_F) = P(X = x, C) - P(X = x, F).$$

Under Assumptions 2.1 and 2.2, we have

$$\begin{aligned} & P(X = x, D = 1|Z = 1) - P(X = x, D = 1|Z = 0) \\ = \; & P(X = x, D(1) = 1) - P(X = x, D(0) = 1) \\ = \; & P(X = x, C) - P(X = x, F). \end{aligned}$$

Combining those last two equations yields

$$P(X = x, C_V) = P(X = x, D = 1|Z = 1) - P(X = x, D = 1|Z = 0).$$

Finally, dividing each side by

$$P(C_V) = P(D = 1|Z = 1) - P(D = 1|Z = 0).$$

yields the result.
QED.

Proof of Theorem 2.5
Let $(B^c(y_0, y_1))$ and $\left(B^f(y_0, y_1)\right)$ denote two families of Bernoulli variables independent of $(Y(0), Y(1), D(0), D(1))$ and such that

$$\begin{aligned} P(B^c(y_0, y_1) = 1) &= \min\left(\frac{P(Y(0) = y_0, Y(1) = y_1, F)}{P(Y(0) = y_0, Y(1) = y_1, C)}, 1\right) \\ P(B^f(y_0, y_1) = 1) &= \min\left(\frac{P(Y(0) = y_0, Y(1) = y_1, C)}{P(Y(0) = y_0, Y(1) = y_1, F)}, 1\right). \end{aligned}$$

Let $C_V = \{C, B^c(Y(0), Y(1)) = 0\}$ and $F_V = \{F, B^f(Y(0), Y(1)) = 0\}$. Finally, let $HM = C_V \cup F_V$.

After a few computations, one can show that

$$\begin{aligned} & P(Y(0) = y_0, Y(1) = y_1, C_V) \\ = \; & \max\left(P(Y(0) = y_0, Y(1) = y_1, C) - P(Y(0) = y_0, Y(1) = y_1, F), 0\right) \end{aligned} \tag{5.6}$$

and

$$\begin{aligned} & P(Y(0) = y_0, Y(1) = y_1, F_V) \\ = \; & \max\left(P(Y(0) = y_0, Y(1) = y_1, F) - P(Y(0) = y_0, Y(1) = y_1, C), 0\right). \end{aligned} \tag{5.7}$$

As a result,

$$\begin{aligned}
P(Y(0)=y_0, Y(1)=y_1, HM) &= P(Y(0)=y_0, Y(1)=y_1, C_V) + P(Y(0)=y_0, Y(1)=y_1, F_V) \\
&= \max\left(P(Y(0)=y_0, Y(1)=y_1, C) - P(Y(0)=y_0, Y(1)=y_1, F), 0\right) \\
&+ \max\left(P(Y(0)=y_0, Y(1)=y_1, F) - P(Y(0)=y_0, Y(1)=y_1, C), 0\right) \\
&= \max\left(P(Y(0)=y_0, Y(1)=y_1, C), P(Y(0)=y_0, Y(1)=y_1, F)\right) \\
&- \min\left(P(Y(0)=y_0, Y(1)=y_1, C), P(Y(0)=y_0, Y(1)=y_1, F)\right). \quad (5.8)
\end{aligned}$$

The first equality arises from the definition of HM and from the fact that the events C_V and F_V are disjoints since C and F are disjoints. The second arises from Equations (5.6) and (5.7). One can check that the last one is true by considering separately the only two possible cases:

$$P(Y(0)=y_0, Y(1)=y_1, F) \leq P(Y(0)=y_0, Y(1)=y_1, C)$$

and

$$P(Y(0)=y_0, Y(1)=y_1, C) \leq P(Y(0)=y_0, Y(1)=y_1, F).$$

If

$$P(Y(0)=y_0, F) \leq P(Y(0)=y_0, C),$$

Assumption 2.9 ensures that for every y_1 in $\mathcal{S}(Y(1))$,

$$P(Y(0)=y_0, Y(1)=y_1, F) \leq P(Y(0)=y_0, Y(1)=y_1, C).$$

In this case, it follows from Equation (5.8) that

$$P(Y(0)=y_0, Y(1)=y_1, HM) = P(Y(0)=y_0, Y(1)=y_1, C) - P(Y(0)=y_0, Y(1)=y_1, F).$$

Integrating or summing this last equation over $\mathcal{S}(Y(1))$ yields

$$P(Y(0)=y_0, HM) = P(Y(0)=y_0, C) - P(Y(0)=y_0, F)$$

which is also equal to

$$\max\left(P(Y(0)=y_0, C), P(Y(0)=y_0, F)\right) - \min\left(P(Y(0)=y_0, C), P(Y(0)=y_0, F)\right),$$

since by assumption

$$P(Y(0)=y_0, F) \leq P(Y(0)=y_0, C).$$

If

$$P(Y(0)=y_0, C) \leq P(Y(0)=y_0, F),$$

one can use a similar reasoning to show that in this case we also have

$$\begin{aligned} P(Y(0)=y_0, HM) &= P(Y(0)=y_0, F) - P(Y(0)=y_0, C) \\ &= \max\left(P(Y(0)=y_0, C), P(Y(0)=y_0, F)\right) \\ &- \min\left(P(Y(0)=y_0, C), P(Y(0)=y_0, F)\right). \end{aligned}$$

This proves that we always have

$$\begin{aligned} P(Y(0)=y_0, HM) &= \max\left(P(Y(0)=y_0, C), P(Y(0)=y_0, F)\right) \\ &- \min\left(P(Y(0)=y_0, C), P(Y(0)=y_0, F)\right). \end{aligned} \tag{5.9}$$

Finally, notice that under Assumptions 2.1 and 2.2, we have

$$\begin{aligned} & P(Y=y_0, D=0|Z=0) - P(Y=y_0, D=0|Z=1) \\ = \; & P(Y(0)=y_0, D(0)=0) - P(Y(0)=y_0, D(1)=0) \\ = \; & P(Y(0)=y_0, C) - P(Y(0)=y_0, F). \end{aligned} \tag{5.10}$$

Equation (5.10) implies that

$$P(Y(0)=y_0, F) \leq P(Y(0)=y_0, C)$$

if and only if

$$P(Y=y_0, D=0|Z=1) \leq P(Y=y_0, D=0|Z=0),$$

which in turn implies that

$$\begin{aligned} & \max\left(P(Y(0)=y_0, C), P(Y(0)=y_0, F)\right) \\ - \; & \min\left(P(Y(0)=y_0, C), P(Y(0)=y_0, F)\right) \\ = \; & \max\left(P(Y=y_0, D=0|Z=0), P(Y=y_0, D=0|Z=1)\right) \\ - \; & \min\left(P(Y=y_0, D=0|Z=0), P(Y=y_0, D=0|Z=1)\right). \end{aligned} \tag{5.11}$$

Combining Equations (5.10) and (5.11) proves the result for $Y(0)$. The result for $Y(1)$ can be obtained through a similar reasoning.
QED.

Proof of Theorem 3.2
This proof is long and somewhat complicated. It is organized in two steps. First, I prove the result for the upper bound, then, I prove the result for the lower bound. To that end, I must introduce measure theoretic concepts. Let $\mathcal{E}_0$ and $\mathcal{E}_1$ denote sigma-algebras

associated to $Y(0)$ and $Y(1)$. Let $\mathcal{E}$ denote the sigma-algebra associated to $(Y(0), Y(1))$ generated by the products of all event E_0 and E_1 in $\mathcal{E}_0$ and $\mathcal{E}_1$.

Step 1: validity and sharpness of B_{IV}^{+}

This step itself is organized in two steps; I first prove the validity of B_{IV}^{+}, and then I show that this bound is sharp.

Step 1.a): validity of B_{IV}^{+}

Let G^0 denote the largest population with the same distribution of $Y(0)$ as compliers. Let G^1 denote the largest population with the same distribution of $Y(1)$ as compliers. Reasoning as in the proof of the first point of Theorem 3.1, one can show that

$$\begin{aligned} P(G^0) &= \frac{P(C)}{P^{+}(C|Y(0))} \\ P(G^1) &= \frac{P(C)}{P^{+}(C|Y(1))}. \end{aligned}$$

Since G^J and G^M verify stronger requirements than G^0 and G^1, they must both be smaller than G^0 and G^1. This yields

$$P(G^J) \leq P(G^M) \leq \frac{P(C)}{\max\left(P^{+}(C|Y(0)), P^{+}(C|Y(1))\right)}. \tag{5.12}$$

Then, notice that for every $d \in \{0; 1\}$,[8]

$$P^{+}(C|Y(d)) = \sup_{E_d \in \mathcal{E}_d} \{P(C|Y(d) \in E_d)\} \tag{5.13}$$

$P(C|Y(d) \in E_d)$ is the expectation of $P(C|Y(d) = y_d)$ over all the values of y_d in E_d; consequently, the largest value of $P(C|Y(d) \in E_d)$ must be smaller than the largest value of $P(C|Y(d) = y_d)$. Then, $\{y_d\}$ belongs to $\mathcal{E}_d$; consequently the largest value of $P(C|Y(d) \in E_d)$ must be larger than the largest value of $P(C|Y(d) = y_d)$. This proves the equality.

Then, for every $d \in \{0; 1\}$, let

$$\underline{P}^{+}(C|Y(d)) = \sup_{E_d \in \mathcal{E}_d} \left\{ \frac{P(Y \in E_d, D = d|Z = d) - P(Y \in E_d, D = d|Z = 1 - d)}{P(Y \in E_d, D = d|Z = d) + P(D = 1 - d|Z = d)} \right\}.$$

For every $E_d \in \mathcal{E}_d$,

$$\begin{aligned} P(C|Y(d) \in E_d) &= \frac{P(Y(d) \in E_d, C)}{P(Y(d) \in E_d)} \\ &= \frac{P(Y \in E_d, D = d|Z = d) - P(Y \in E_d, D = d|Z = 1 - d)}{P(Y(d) \in E_d, D(d) = d) + P(Y(d) \in E_d, D(d) = 1 - d)} \\ &\geq \frac{P(Y \in E_d, D = d|Z = d) - P(Y \in E_d, D = d|Z = 1 - d)}{P(Y \in E_d, D = d|Z = d) + P(D = 1 - d|Z = d)}. \end{aligned}$$

[8]The convention adopted throughout the paper to define conditional probabilities with respect to 0 probability events ensures that the following quantity is well-defined even when E_d reduces to a singleton.

For the numerator, the first equality follows from Theorem 2.2. For the denominator, it merely follows from the law of total probabilities. Therefore,

$$\underline{P}^{+}(C|Y(d)) \leq P^{+}(C|Y(d)). \tag{5.14}$$

Combining Equations (5.12), (5.13), (5.14) and the second point of the next lemma establishes the validity of the upper bound.

Lemma 5.1 *Suppose Assumptions 2.1, 2.2 and 2.3 hold. Then,*

1. *The sequence $\underline{P}(C|Y(d) \in E_d^n)$ is increasing and bounded. As a result, its limit exists. Moreover, the sequence of events E_d^n is decreasing.*
2. $\underline{P}^{+}(C|Y(d)) = \underline{P}(C|Y(d) \in E_d^*) = \lim_{n \to +\infty} \underline{P}(C|Y(d) \in E_d^n)$.

Proof of Lemma 5.1

The proof of this lemma relies on the two following lemmas. For every x included between 0 and 1, let

$$E_x = \{y_d \in \mathcal{S}(Y(d)) : \overline{P}(C|Y(d) = y_d) \geq x\}.$$

Lemma 5.2

1. *For every $E_d \in \mathcal{E}_d$ such that $E_d \subseteq E_x$,*

$$\overline{P}(C|Y(d) \in E) \geq x.$$

2. *For every $E_d \in \mathcal{E}_d$ such that $E \subseteq E_x^c$,*

$$\overline{P}(C|Y(d) \in E) < x.$$

Proof of Lemma 5.2

I only prove the first statement. For every $E_d \subseteq E_x$,

$$\begin{aligned}
& P(Y \in E_d, D = d|Z = d) - P(Y \in E_d, D = d|Z = 1-d) \\
= & \int_{E_d} P(Y = y_d, D = d|Z = d) - P(Y = y_d, D = d|Z = 1-d) d\lambda(y_d) \\
\geq & \; x \int_{E_d} P(Y = y_d, D = d|Z = d) d\lambda(y_d) \\
= & \; xP(Y \in E_d, D = d|Z = d).
\end{aligned}$$

The inequality follows from the definition of E_x and from the fact that $E_d \subseteq E_x$. This proves the result.

QED.

Lemma 5.3 *Let* $(a, c, e, b, d, f) \in \mathbb{R}_+^3 \times (\mathbb{R}_+ \setminus \{0\})^3$ *be such that* $\frac{a}{b} = \frac{c+e}{d+f}$.

1. $\frac{a}{b} \leq \frac{c}{d} \Leftrightarrow \frac{a}{b} \geq \frac{e}{f}$.
2. $\frac{a}{b} < \frac{c}{d} \Leftrightarrow \frac{a}{b} > \frac{e}{f}$.

The proof of this last Lemma is straightforward.

I can now prove Lemma 5.1. In the proof, I drop the subscript d in E_d^n and E_d^* to alleviate the notational burden.

Proof of 1

I start proving that $\underline{P}(C|Y(d) \in E^n)$ is increasing. Assume first that $P(Y \in E^n \cap (E^{n+1})^c, D = d|Z = d) > 0$. It follows from the definition of the sequence of events $(E^n)_{n\in\mathbb{N}}$ that for every $n \in \mathbb{N}$ we either have $E^{n+1} \subseteq E^n$ or $E^n \subset E^{n+1}$.

Assume that $E^n \subset E^{n+1}$. It follows from the fact that

$$E^n = \left(E^n \cap E^{n+1}\right) \cup \left(E^n \cap \left(E^{n+1}\right)^c\right)$$

that

$$\begin{aligned}
& \underline{P}(C|Y(d) \in E^n) \\
= \; & \frac{P(Y \in E^n, D = d|Z = d) - P(Y \in E^n, D = d|Z = 1-d)}{P(Y \in E^n, D = d|Z = d) + P(D = 1-d|Z = d)} \\
= \; & \frac{P(Y \in E^n \cap E^{n+1}, D = d|Z = d) - P(Y \in E^n \cap E^{n+1}, D = d|Z = 1-d)}{P(Y \in E^n \cap E^{n+1}, D = d|Z = d) + P(D = 1-d|Z = d) + P(Y \in E^n \cap (E^{n+1})^c, D = d|Z = d)} \\
+ \; & \frac{P(Y \in E^n \cap \left(E^{n+1}\right)^c, D = d|Z = d) - P(Y \in E^n \cap \left(E^{n+1}\right)^c, D = d|Z = 1-d)}{P(Y \in E^n \cap E^{n+1}, D = d|Z = d) + P(D = 1-d|Z = d) + P(Y \in E^n \cap (E^{n+1})^c, D = d|Z = d)}.
\end{aligned} \tag{5.15}$$

Since $E^n \subset E^{n+1}$ by assumption, Equation (5.15) is equivalent to $\frac{a}{b} = \frac{a+e}{b+f}$, with

$$\begin{aligned}
\frac{a}{b} &= \underline{P}(C|Y(d) \in E^n) \\
\frac{e}{f} &= \overline{P}(C|Y(d) \in E^n \cap \left(E^{n+1}\right)^c).
\end{aligned}$$

Since $E^n \cap (E^{n+1})^c \subseteq (E^{n+1})^c$, and since $P(Y \in E^n \cap (E^{n+1})^c, D = d|Z = d) > 0$ by assumption, it follows from the second point of Lemma 5.2 that

$$\overline{P}(C|Y(d) \in E^n \cap \left(E^{n+1}\right)^c) < \underline{P}(C|Y(d) \in E^n).$$

This implies that $\frac{a}{b} = \frac{a+e}{b+f}$ and $\frac{e}{f} < \frac{a}{b}$, a contradiction as per the second point of Lemma 5.3. This proves that $E^{n+1} \subseteq E^n$.

As a result, Equation (5.15) is actually equivalent to $\frac{a}{b} = \frac{c+e}{d+f}$, with

$$\underline{P}(C|Y(d) \in E^{n+1}) = \frac{c}{d}.$$

$\frac{e}{f} < \frac{a}{b}$ as shown above using the second point of Lemma 5.2. Therefore, it follows from the second point of Lemma 5.3 that

$$\underline{P}(C|Y(d) \in E^n) < \underline{P}(C|Y(d) \in E^{n+1}).$$

When $P(Y \in E^n \cap (E^{n+1})^c, D = d|Z = d) = 0$, it is easy to check from Equation (5.15) that under Assumption 2.3, we have

$$\underline{P}(C|Y(d) \in E^n) = \underline{P}(C|Y(d) \in E^{n+1}).$$

This proves that we always have

$$\underline{P}(C|Y(d) \in E^n) \leq \underline{P}(C|Y(d) \in E^{n+1}),$$

which implies that the sequence $(\underline{P}(C|Y(d) \in E^n))_{n\in\mathbb{N}}$ is increasing. Since it is bounded by 1, its limit exists. This also implies that the sequence of events $(E^n)_{n\in\mathbb{N}}$ is decreasing.

Proof of 2

One can use the fact that

$$E^n = (E^n \cap E^*) \cup (E^n \cap (E^*)^c)$$

to derive an $\frac{a}{b} = \frac{c+e}{d+f}$ type of formula as in Equation (5.15). Since E^n is a decreasing sequence of events, it is easy to see from that decomposition that

$$\underline{P}\left(C|Y(d) \in E^*\right) = \lim_{n\to+\infty} \underline{P}(C|Y(d) \in E^n).$$

It follows from Equation (5.13) that

$$\underline{P}\left(C|Y(d) \in E^*\right) \leq \underline{P}^+(C|Y(d)).$$

Assume that

$$\underline{P}\left(C|Y(d) \in E^*\right) < \underline{P}^+(C|Y(d)).$$

Equation (5.13) implies that there must exist an event A such that

$$\underline{P}\left(C|Y(d) \in E^*\right) < \underline{P}\left(C|Y(d) \in A\right).$$

Then, we can use the fact that

$$A = (E^* \cap A) \cup ((E^*)^c \cap A)$$

to derive an $\frac{a}{b} = \frac{c+e}{d+f}$ type of formula, with

$$\begin{aligned}
\frac{a}{b} &= \underline{P}(C|Y(d) \in A) \\
\frac{c}{d} &= \underline{P}\,(C|Y(d) \in E^* \cap A) \\
\frac{e}{f} &= \overline{P}\,(C|Y(d) \in (E^*)^c \cap A)\,.
\end{aligned}$$

Since $A \cap (E^*)^c \subseteq (E^*)^c$, it follows from the second point of Lemma 5.2 that

$$\overline{P}\,(C|Y(d) \in (E^*)^c \cap A) < \underline{P}\,(C|Y(d) \in E^*)\,.$$

Combined with the fact that $\underline{P}\,(C|Y(d) \in E^*) < \underline{P}(C|Y(d) \in A)$ by assumption, and with the second point of Lemma 5.3, this yields

$$\underline{P}(C|Y(d) \in E^*) < \underline{P}\,(C|Y(d) \in E^* \cap A)\,. \tag{5.16}$$

Finally, we can use the fact that

$$E^* = (E^* \cap A) \cup (E^* \cap (A)^c)$$

to derive another $\frac{a}{b} = \frac{c+e}{d+f}$ type of formula, with

$$\begin{aligned}
\frac{a}{b} &= \underline{P}\,(C|Y(d) \in E^*) \\
\frac{c}{d} &= \underline{P}\,(C|Y(d) \in E^* \cap A) \\
\frac{e}{f} &= \overline{P}\,(C|Y(d) \in E^* \cap (A)^c)\,.
\end{aligned}$$

Since $E^* \cap (A)^c \subseteq E^*$, it follows from the first point of Lemma 5.2 that

$$\overline{P}\,(C|Y(d) \in E^* \cap (A)^c) \geq \underline{P}\,(C|Y(d) \in E^*)\,.$$

Combined with the first point of Lemma 5.3, this yields

$$\underline{P}\,(C|Y(d) \in E^* \cap A) \leq \underline{P}(C|Y(d) \in E^*),$$

hence a contradiction with Equation (5.16).
This proves that

$$\underline{P}\,(C|Y(d) \in E^*) = \underline{P}^+(C|Y(d)).$$

QED.

Step 1.b): sharpness of B^+_{IV}

To prove that B_{IV}^{+} is a sharp upper bound of $P(G^J)$ and $P(G^M)$, I only need to prove that B_{IV}^{+} is a sharp upper bound of $P(G^J)$. Indeed, since for every DGP, $P(G^J) \leq P(G^M) \leq B_{IV}^{+}$, if there exists a DGP consistent with the data such that $\tilde{P}(G^J) = B_{IV}^{+}$, the previous inequality implies that we necessarily have $\tilde{P}(G^M) = B_{IV}^{+}$ as well.

The proof of this step divided into two parts. I first show that $\max(P^{+}(C|Y(0)), P^{+}(C|Y(1)))$ is a sharp lower bound of $P^{+}(C|Y(0), Y(1))$ **conditional on the marginal distributions of** $Y(0)$ **and** $Y(1)$. This means that even if those marginal distributions were identified, $\max(P^{+}(C|Y(0)), P^{+}(C|Y(1)))$ would still be a sharp lower bound. Then, I show that $\underline{P}^{+}(C|Y(d))$ is a sharp lower bound of $P^{+}(C|Y(d))$, conditional on what is actually identified from the data.

Part 1.

I show that if $P^{+}(C|Y(0)) \neq P^{+}(C|Y(1))$,[9] then $\max(P^{+}(C|Y(0)), P^{+}(C|Y(1)))$ is a sharp lower bound of $P^{+}(C|Y(0), Y(1))$. On that purpose, I construct one sequence of joint distributions of $(Y(0), Y(1))$ denoted $\tilde{P}^n$, and one sequence of joint distributions of $(Y(0), Y(1), C)$, denoted $\tilde{P}_C^n$, potentially different from the true ones, consistent with the true marginal probability measures, and such that $\tilde{P}_n^{+}(C|Y(0), Y(1)) \to \max(P^{+}(C|Y(0)), P^{+}(C|Y(1)))$.

The definition of $P^{+}(C|Y(0))$ in Equation (5.13) implies that there is a sequence of events $E_0^n \in \mathcal{E}_0$ such that $P(C|Y(0) \in E_0^n) \to P^{+}(C|Y(0))$. Since $P^{+}(C|Y(0)) > 0$, there exists n_0 such that $P(C|Y(0) \in E_0^n) > 0$ for every $n \geq n_0$. Without loss of generality, assume that $P^{+}(C|Y(1)) < P^{+}(C|Y(0))$. This implies that there exists n_1 such that for every $n \geq n_1$ and for every $E_1 \in \mathcal{E}_1$,

$$P(C|Y(0) \in E_0^n) \geq P(C|Y(1) \in E_1).$$

Let $n_2 = \max(n_0, n_1)$.

For every $n \geq n_2$, let

$$\tilde{P}_C^n((Y(0), Y(1)) \in E, C) = P((Y(0), Y(1)) \in E, C)$$

for every $E \in \mathcal{E}$, All the assumptions on $\tilde{P}_C^n$ are verified since it is equal to the true probability measure.

[9] Assuming that $P^{+}(C|Y(0)) \neq P^{+}(C|Y(1))$ is without loss of generality: $P^{+}(C|Y(d))$ are not point identified and their identification regions never reduce to a point so that one can never reject $P^{+}(C|Y(0)) \neq P^{+}(C|Y(1))$.

For every $(E_0, E_1) \in \mathcal{E}_0 \times \mathcal{E}_1$, let

$$\begin{aligned}
& \tilde{P}^n((Y(0), Y(1)) \in E_0 \times E_1) \\
= \; & P(Y(0) \in E_0 | Y(0) \in E_0^n) \frac{P((Y(0), Y(1)) \in E_0^n \times E_1, C)}{P(C | Y(0) \in E_0^n)} \\
+ \; & P(Y(0) \in E_0 | Y(0) \in (E_0^n)^c) \left(P(Y(1) \in E_1) - \frac{P((Y(0), Y(1)) \in E_0^n \times E_1, C)}{P(C | Y(0) \in E_0^n)} \right).
\end{aligned}$$

I show that $\tilde{P}^n$ defines a sequence of probability measures on $\mathcal{E}_0 \times \mathcal{E}_1$. To show that $\tilde{P}^n \geq 0$, it is sufficient to check that for every $E_1 \in \mathcal{E}_1$,

$$0 \leq \frac{P((Y(0), Y(1)) \in E_0^n \times E_1, C)}{P(C | Y(0) \in E_0^n)} \leq P(Y(1) \in E_1).$$

It is straightforward that the left-hand side inequality holds. Consider now the right-hand side inequality. If $P(Y(1) \in E_1) = 0$, this inequality is trivially verified. When $P(Y(1) \in E_1) > 0$,

$$\begin{aligned}
\frac{P((Y(0), Y(1)) \in E_0^n \times E_1, C)}{P(C | Y(0) \in E_0^n)} \; & \leq \; \frac{P(Y(1) \in E_1, C)}{P(C | Y(0) \in E_0^n)} \\
& = \; \frac{P(C | Y(1) \in E_1)}{P(C | Y(0) \in E_0^n)} P(Y(1) \in E_1) \\
& \leq \; P(Y(1) \in E_1).
\end{aligned}$$

The last inequality holds because $n \geq n_2$ and $P(Y(1) \in E_1) > 0$.

Then, we also have

$$\begin{aligned}
& \tilde{P}^n((Y(0), Y(1)) \in \mathcal{S}(Y(0)) \times \mathcal{S}(Y(1))) = 1 \\
& \tilde{P}^n((Y(0), Y(1)) \in \emptyset) = 0.
\end{aligned}$$

Then, one can check that $(E_0 \times E_1) \cap (E_0' \times E_1') = \emptyset$ and $(E_0 \times E_1) \cup (E_0' \times E_1') \in \mathcal{E}_0 \times \mathcal{E}_1$ if and only if $E_0 = E_0'$ and $E_1 \cap E_1' = \emptyset$, or $E_1 = E_1'$ and $E_0 \cap E_0' = \emptyset$. Therefore, to show that $\tilde{P}^n$ is σ-additive on $\mathcal{E}_0 \times \mathcal{E}_1$, it is sufficient to show it for $(E_0 \times E_1) \cap (E_0' \times E_1')$ with $E_0 = E_0'$ and $E_1 \cap E_1' = \emptyset$, or $E_1 = E_1'$ and $E_0 \cap E_0' = \emptyset$. Assume for instance that $E_0 = E_0'$ and $E_1 \cap E_1' = \emptyset$.

$$\begin{aligned}
& \tilde{P}^n((Y(0), Y(1)) \in (E_0 \times E_1) \cup (E_0' \times E_1')) \\
= \; & \tilde{P}^n((Y(0), Y(1)) \in (E_0 \times (E_1 \cup E_1')) \\
= \; & \tilde{P}^n((Y(0), Y(1)) \in (E_0 \times E_1)) + \tilde{P}^n((Y(0), Y(1)) \in (E_0 \times E_1')) \\
= \; & \tilde{P}^n((Y(0), Y(1)) \in (E_0 \times E_1)) + \tilde{P}^n((Y(0), Y(1)) \in (E_0' \times E_1')).
\end{aligned}$$

The second equality follows from the definition of $\tilde{P}^n$ and from the fact that the true measures $P(Y(1) \in E_1)$ and $P((Y(0), Y(1)) \in E_0 \times E_1, C)$ are additive.

Finally, σ-additivity on $\mathcal{E}_0 \times \mathcal{E}_1$ combined with the fact that $\tilde{P}^n((Y(0), Y(1)) \in \mathcal{S}(Y(0)) \times \mathcal{S}(Y(1))) = 1$ implies that for every $(E_0, E_1) \in \mathcal{E}_0 \times \mathcal{E}_1$,

$$\begin{aligned}\tilde{P}^n((Y(0), Y(1)) \in E_0 \times E_1) &= 1 - \tilde{P}^n((Y(0), Y(1)) \in (E_0)^c \times \mathcal{S}(Y(1))) \\ &- \tilde{P}^n((Y(0), Y(1)) \in E_0 \times (E_1)^c),\end{aligned}$$

which is smaller than 1. This proves that $\tilde{P}^n$ defines a sequence of probability measures on $\mathcal{E}_0 \times \mathcal{E}_1$.

Now, notice that $\mathcal{E}_0 \times \mathcal{E}_1$ is a semi-ring. Therefore, by Carathéodory extension Theorem, $\tilde{P}^n$ can be extended to $\mathcal{E}$.

$\tilde{P}^n$ is compatible with the true marginal probabilities. Indeed, for every $E_0 \in \mathcal{E}_0$,

$$\begin{aligned}
& \tilde{P}^n((Y(0)) \in E_0) \\
= \; & \tilde{P}^n((Y(0), Y(1)) \in E_0 \times \mathcal{S}(Y(1))) \\
= \; & P(Y(0) \in E_0 | Y(0) \in E_0^n) \frac{P((Y(0), Y(1)) \in E_0^n \times \mathcal{S}(Y(1)), C)}{P(C | Y(0) \in E_0^n)} \\
+ \; & P(Y(0) \in E_0 | Y(0) \in (E_0^n)^c) \left(P(Y(1) \in \mathcal{S}(Y(1))) - \frac{P((Y(0), Y(1)) \in E_0^n \times \mathcal{S}(Y(1)), C)}{P(C | Y(0) \in E_0^n)} \right) \\
= \; & P(Y(0) \in E_0 | Y(0) \in E_0^n) P(Y(0) \in E_0^n) \\
+ \; & P(Y(0) \in E_0 | Y(0) \in (E_0^n)^c) P(Y(0) \in (E_0^n)^c) \\
= \; & P(Y(0) \in E_0).
\end{aligned}$$

Moreover, for every $E_1 \in \mathcal{E}_1$,

$$\begin{aligned}
& \tilde{P}^n((Y(1)) \in E_1) \\
= \; & \tilde{P}^n((Y(0), Y(1)) \in \mathcal{S}(Y(0)) \times E_1) \\
= \; & P(Y(0) \in \mathcal{S}(Y(0)) | Y(0) \in E_0^n) \frac{P((Y(0), Y(1)) \in E_0^n \times E_1, C)}{P(C | Y(0) \in E_0^n)} \\
+ \; & P(Y(0) \in \mathcal{S}(Y(0)) | Y(0) \in (E_0^n)^c) \left(P(Y(1) \in E_1) - \frac{P((Y(0), Y(1)) \in E_0^n \times E_1, C)}{P(C | Y(0) \in E_0^n)} \right) \\
= \; & P(Y(1) \in E_1).
\end{aligned}$$

Finally, I show that

$$\tilde{P}_n^+(C | Y(0), Y(1)) = P(C | Y(0) \in E_0^n).$$

First, notice that

$$\frac{\tilde{P}^n((Y(0), Y(1)) \in E_0^n \times \mathcal{S}(Y(1)), C)}{\tilde{P}^n((Y(0), Y(1)) \in E_0^n \times \mathcal{S}(Y(1)))} = P(C | Y(0) \in E_0^n),$$

which proves that

$$\tilde{P}_n^+(C | Y(0), Y(1)) \geq P(C | Y(0) \in E_0^n).$$

Then, notice that for every $(E_0, E_1) \in \mathcal{E}_0 \times \mathcal{E}_1$,

$$\begin{aligned}\tilde{P}^n((Y(0), Y(1)) \in E_0 \times E_1) &= \tilde{P}^n((Y(0), Y(1)) \in (E_0 \cap E_0^n) \times E_1) \\ &+ \tilde{P}^n((Y(0), Y(1)) \in (E_0 \cap (E_0^n)^c) \times E_1).\end{aligned}$$

Therefore, it is sufficient to prove that

$$\begin{aligned}\tilde{P}^n((Y(0), Y(1)) \in (E_0 \cap E_0^n) \times E_1) &\geq \frac{\tilde{P}^n((Y(0), Y(1)) \in (E_0 \cap E_0^n) \times E_1, C)}{P(C|Y(0) \in E_0^n)} \\ \tilde{P}^n((Y(0), Y(1)) \in (E_0 \cap (E_0^n)^c) \times E_1) &\geq \frac{\tilde{P}^n((Y(0), Y(1)) \in (E_0 \cap (E_0^n)^c) \times E_1, C)}{P(C|Y(0) \in E_0^n)}\end{aligned}$$

to obtain that for every $E_0 \times E_1 \in \mathcal{E}_0 \times \mathcal{E}_1$,

$$\tilde{P}^n((Y(0), Y(1)) \in E_0 \times E_1) \geq \frac{\tilde{P}^n((Y(0), Y(1)) \in E_0 \times E_1, C)}{P(C|Y(0) \in E_0^n)}. \tag{5.17}$$

Consider the first inequality:

$$\begin{aligned}\tilde{P}^n((Y(0), Y(1)) \in (E_0 \cap E_0^n) \times E_1) &= \frac{P((Y(0), Y(1)) \in E_0^n \times E_1, C)}{P(C|Y(0) \in E_0^n)} \\ &\geq \frac{P((Y(0), Y(1)) \in (E_0 \cap E_0^n) \times E_1, C)}{P(C|Y(0) \in E_0^n)} \\ &= \frac{\tilde{P}^n((Y(0), Y(1)) \in (E_0 \cap E_0^n) \times E_1, C)}{P(C|Y(0) \in E_0^n)}.\end{aligned}$$

This proves the first inequality. Consider now the second inequality.

$$\begin{aligned}& \tilde{P}^n((Y(0), Y(1)) \in (E_0 \cap (E_0^n)^c) \times E_1) \\ - & \frac{\tilde{P}^n((Y(0), Y(1)) \in (E_0 \cap (E_0^n)^c) \times E_1, C)}{P(C|Y(0) \in E_0^n)} \\ = & P(Y(1) \in E_1) - \frac{P((Y(0), Y(1)) \in E_0^n \times E_1, C)}{P(C|Y(0) \in E_0^n)} \\ - & \frac{P((Y(0), Y(1)) \in (E_0 \cap (E_0^n)^c) \times E_1, C)}{P(C|Y(0) \in E_0^n)} \\ \geq & P(Y(1) \in E_1) - \frac{P((Y(0), Y(1)) \in E_0^n \times E_1, C)}{P(C|Y(0) \in E_0^n)} \\ - & \frac{P((Y(0), Y(1)) \in (E_0^n)^c \times E_1, C)}{P(C|Y(0) \in E_0^n)} \\ \geq & P(Y(1) \in E_1) - \frac{P(Y(1) \in E_1, C)}{P(C|Y(0) \in E_0^n)} \\ \geq & 0.\end{aligned}$$

The last inequality is trivially true if $P(Y(1) \in E_1) = 0$. It is also verified if $P(Y(1) \in E_1) > 0$ since $n \geq n_2$. This proves Equation (5.17). Using the definition of Carathéodory's

extension of $\tilde{P}^n$ to $\mathcal{E}$, one can show that Equation (5.17) extends to $\mathcal{E}$. Moreover, Equation (5.17) implies that

$$P(C|Y(0) \in E_0^n) \geq \tilde{P}^n(C|(Y(0), Y(1)) \in E).$$

This proves that

$$\tilde{P}_n^+(C|Y(0), Y(1)) \leq P(C|Y(0) \in E_0^n).$$

This completes the proof.

Part 2.

To prove that the bound is sharp, I construct a probability measure $\tilde{P}$, potentially different from the true one but consistent with the data, and such that

$$\tilde{P}^+(C|Y(d)) = \underline{P}^+(C|Y(d)).$$

As shown in Lemma 5.1,

$$\underline{P}^+(C|Y(d)) = \underline{P}(C|Y(d) \in E_d^*).$$

As a result, $\tilde{P}$ must be such that

$$\tilde{P}^+(C|Y(d)) = \underline{P}(C|Y(d) \in E_d^*).$$

For every $E_d \in \mathcal{E}_d$, let

$$\begin{aligned} & \tilde{P}(Y(d) \in E_d, D(d) = 1-d) \\ = \; & \frac{P(Y \in E_d \cap E_d^*, D = d|Z = d) - P(Y \in E_d \cap E_d^*, D = d|Z = 1-d)}{P(Y \in E_d^*, D = d|Z = d) - P(Y \in E_d^*, D = d|Z = 1-d)} (P(Y \in E_d^*, D = d|Z = d) + P(D = 1-d|Z = d)) \\ - \; & P(Y \in E_d \cap E_d^*, D = d|Z = d). \end{aligned}$$

It is easy to check that this defines a σ-additive measure.

I prove now that

$$0 \leq \tilde{P}(Y(d) \in E_d, D(d) = 1-d) \leq P(D = 1-d|Z = d).$$

Consider first $E_d \subseteq E_d^*$. $\tilde{P}(Y(d) \in E_d, D(d) = 1-d)$ is greater than 0 if and only if

$$\overline{P}^+(C|Y(d) \in E_d) \geq \underline{P}^+(C|Y(d) \in E_d^*),$$

which is true by the definition of E_d^* and by the first point of Lemma 5.2. Notice that $\tilde{P}(Y(d) \in E_d^*, D(d) = 1-d) = P(D = 1-d|Z = d)$. This combined with σ-additivity implies that we also have $\tilde{P}(Y(d) \in E_d, D(d) = 1-d) \leq P(D = 1-d|Z = d)$. Consider now $E_d \subseteq (E_d^*)^c$. It is easy to see $\tilde{P}(Y(d) \in E_d, D(d) = 1-d) = 0$. Combining all of this

with σ-additivity and with the fact that for every $E_d \in \mathcal{E}_d$, $E_d = (E_d \cap E_d^*) \cup (E_d \cap (E_d^*)^c)$ finally yields

$$0 \leq \tilde{P}(Y(d) \in E_d, D(d) = 1-d) \leq P(D = 1-d|Z = d).$$

Then, let

$$\begin{aligned}
\tilde{P}(Y(d) \in E_d) &= P(Y \in E_d, D = d|Z = d) + \tilde{P}(Y(d) \in E_d, D(d) = 1-d) \\
\tilde{P}(Y(d) \in E_d, C) &= P(Y \in E_d, D = d|Z = d) - P(Y \in E_d, D = d|Z = 1-d).
\end{aligned}$$

It follows from what precedes that this defines a probability measure consistent with the data and with Assumptions 2.1, 2.2 and 2.3.

Finally, I show that

$$\tilde{P}^+(C|Y(d)) = \underline{P}(C|Y(d) \in E_d^*).$$

First, notice that for every $E_d \subseteq E_d^*$,

$$\tilde{P}(C|Y(d) \in E_d) = \underline{P}(C|Y(d) \in E_d^*). \tag{5.18}$$

Then, for every $E_d \supseteq E_d^*$,

$$\tilde{P}(C|Y(d) \in E_d) = \underline{P}(C|Y(d) \in E_d) \leq \underline{P}(C|Y(d) \in E_d^*), \tag{5.19}$$

where the inequality follows the definition of E_d^*. Finally, using the fact that for every E_d,

$$E_d = (E_d \cap E_d^*) \cup (E_d \cap (E_d^*)^c),$$

one can show that

$$\begin{aligned}
\tilde{P}(C|Y(d) \in E_d) &= \frac{\tilde{P}(Y(d) \in E_d, C)}{\tilde{P}(Y(d) \in E_d)} \\
&= \frac{\tilde{P}(Y(d) \in E_d \cap E_d^*, C) + \tilde{P}(Y(d) \in E_d \cap (E_d^*)^c, C)}{\tilde{P}(Y(d) \in E_d \cap E_d^*) + \tilde{P}(Y(d) \in E_d \cap (E_d^*)^c)}.
\end{aligned}$$

This is what I hereafter to as an $\frac{a}{b} = \frac{c+e}{d+f}$ type of formula, with

$$\begin{aligned}
\frac{a}{b} &= \tilde{P}(C|Y(d) \in E_d) \\
\frac{c}{d} &= \tilde{P}(C|Y(d) \in E_d \cap E_d^*) \\
\frac{e}{f} &= \tilde{P}(C|Y(d) \in E_d \cap (E_d^*)^c).
\end{aligned}$$

Combining the second point of Lemma 5.3 with Equations (5.18) and (5.19) proves that

$$\tilde{P}(C|Y(d) \in E_d) \leq \underline{P}(C|Y(d) \in E_d^*).$$

This completes the proof.

Step 2: validity and sharpness of B_{IV}^{-}

The validity of the lower bound is straightforward. To prove that B_{IV}^{-} is a sharp lower bound of $P(G^J)$ and $P(G^M)$, I only need to prove that B_{IV}^{-} is a sharp lower bound of $P(G^M)$. Indeed, since for every DGP, $B_{IV}^{+} \leq P(G^J) \leq P(G^M)$, if there exists a DGP consistent with the data such that $\tilde{P}(G^M) = B_{IV}^{-}$, the previous inequality implies that we necessarily have $\tilde{P}(G^J) = B_{IV}^{-}$ as well.

To prove that B_{IV}^{-} is a sharp lower bound of $P(G^M)$, I rely on a preliminary lemma, in which I derive an explicit formula for $P(G^M)$. The true distribution of $(Y(0), Y(1), 1_C)$ is denoted P. Let $\mathcal{P}$ be the set of all the probability distributions of $(Y(0), Y(1))$ for compliers compatible with the data and compatible with the true distribution of $(Y(0), Y(1))$ in the entire population. $\mathcal{P}$ is the set of all the probability distributions P' verifying the three following requirements:

1. $P'((Y(0) = y_0, Y(1) = y_1, C) \leq P(Y(0) = y_0, Y(1) = y_1)$
2. $P'(Y(0) = y_0, C) = PY(0) = y_0, C)$
3. $P'(Y(1) = y_1, C) = P(Y(1) = y_1, C)$.

Requirement 1 ensures that in each $(Y(0), Y(1))$ stratum, P' does not requires that there be more compliers than the total number of units in that stratum. This ensures that P' is compatible with the joint distribution of $(Y(0), Y(1))$ in the total population. Requirement 2 ensures that P' is compatible with the true distribution of $(Y(0), C)$, as this true distribution is identified from the data. Requirement 3 ensures that P' is compatible with the true distribution of $(Y(1), C)$, as this true distribution is also identified from the data.

Then, let

$$P'(C|Y(0) = y_0, Y(1) = y_1) = \frac{P'(Y(0) = y_0, Y(1) = y_1, C)}{P(Y(0) = y_0, Y(1) = y_1)}$$

denote the shares of compliers in each $(Y(0), Y(1))$ stratum, as per P'. If P' was the true distribution of $(Y(0), Y(1))$ for compliers, then the share of compliers in each $(Y(0), Y(1))$ stratum would be equal to $P'(C|Y(0) = y_0, Y(1) = y_1)$. Let also

$$P'^{+}(C|(Y(0), Y(1))) = \sup_{(y_0,y_1)\in\mathcal{S}((Y(0),Y(1)))} \{P'(C|Y(0) = y_0, Y(1) = y_1)\}$$

be the largest share of compliers in a $(Y(0), Y(1))$ stratum, as per P'. If P' was the true distribution of $(Y(0), Y(1))$ for compliers, then the largest share of compliers in a $(Y(0), Y(1))$ stratum would be equal to $P'^{+}(C|(Y(0), Y(1)))$. Finally, let

$$P^{+}(C|Y(0); Y(1)) \quad = \quad \inf_{P'\in\mathcal{P}} \left\{P'^{+}(C|(Y(0), Y(1)))\right\}.$$

$P^+(C|Y(0);Y(1))$ is the lowest value of $P'^+(C|(Y(0),Y(1))$, for all P' belonging to $\mathcal{P}$.

Lemma 5.4

$$P(G^M) = \frac{P(C)}{P^+(C|Y(0);Y(1))}.$$

Proof of Lemma 5.4

For every P' in $\mathcal{P}$, one can create a population of units C' such that

$$P(Y(0) = y_0, Y(1) = y_1, C') = P'(Y(0) = y_0, Y(1) = y_1, C). \tag{5.20}$$

This is achieved picking up

$$\frac{P'(Y(0) = y_0, Y(1) = y_1, C)}{P(Y(0) = y_0, Y(1) = y_1)}\%$$

of units in each $(Y(0) = y_0, Y(1) = y_1)$ stratum. It follows from the definition of P' that the marginal distributions of $Y(0)$ and $Y(1)$ are the same in C and C', even though the joint distributions of $(Y(0), Y(1))$ can be different in the two populations.

Then, using the same reasoning as for the proof of Theorem 3.1, one can show that it is possible to construct a population $G^M_{P'}$ of size

$$\frac{P(C)}{P^+(C'|(Y(0),Y(1))},$$

with the same distribution of $(Y(0), Y(1))$ as C'. Equation (5.20) implies that

$$P(G^M_{P'}) = \frac{P(C)}{P'^+(C|(Y(0),Y(1))}.$$

Moreover, as C and C' have the same marginal distributions of $Y(0)$ and $Y(1)$, and as C' and $G^M_{P'}$ have the same joint distributions of $(Y(0), Y(1))$, it follows that C and $G^M_{P'}$ have the same marginal distributions of $Y(0)$ and $Y(1)$.

This shows that G^M, the largest population with same marginal distributions of $Y(0)$ and $Y(1)$ as compliers is at least of size

$$\frac{P(C)}{P'^+(C|(Y(0),Y(1))}.$$

Since this is true for every P' in $\mathcal{P}$, we finally have

$$P(G^M) \geq \frac{P(C)}{P^+(C|Y(0);Y(1))}.$$

Conversely, assume that there exists a population G' strictly larger than all the populations $G^M_{P'}$, and with the same marginal distributions of $Y(0)$ and $Y(1)$ as compliers. As a result there must be a strictly positive number ε such that

$$P(G') > (1+\varepsilon)\frac{P(C)}{P^+(C|Y(0);Y(1))}. \tag{5.21}$$

Then, let

$$P'_{G'}(Y(0)=y_0, Y(1)=y_1, C) = P(Y(0)=y_0, Y(1)=y_1, G')\frac{P(C)}{P(G')}. \tag{5.22}$$

Using the fact that G' has the same marginal distributions of $Y(0)$ and $Y(1)$ as compliers, one can show that $P'_{G'}$ belongs to $\mathcal{P}$. Combining (5.21) and (5.22), we obtain

$$P(G'|Y(0)=y_0, Y(1)=y_1) > \frac{1+\varepsilon}{P^+(C|Y(0);Y(1))}P'_{G'}(C|Y(0)=y_0, Y(1)=y_1). \tag{5.23}$$

Now, let (y_0^*, y_1^*) be such that

$$P'_{G'}(C|Y(0)=y_0^*, Y(1)=y_1^*) > \frac{P'^+_{G'}(C|Y(0),Y(1))}{1+\varepsilon}.$$

This also implies that

$$P'_{G'}(C|Y(0)=y_0^*, Y(1)=y_1^*) > \frac{P^+(C|Y(0);Y(1))}{1+\varepsilon}.$$

Combining this last equation with (5.23) yields

$$P(G'|Y(0)=y_0, Y(1)=y_1^*) > 1,$$

a contradiction. This proves the result.

QED.

It follows from Lemma 5.4 that proving that B^-_{IV} is a sharp lower bound $P(G^M)$ amounts to proving that 1 is a sharp upper bound of $P^+(C|Y(0);Y(1))$. On that purpose, I am going to construct a probability distribution $\tilde{P}$, potentially different from the true one but consistent with the data, and such that $\tilde{P}^+(C|Y(0);Y(1)) = 1$.

Let $\underline{E}_0$ and $\underline{E}_1$ be two elements of $\mathcal{E}_0$ and $\mathcal{E}_1$ verifying

$$P(Y(0)\in(\underline{E}_0)^c, C) \leq P(Y(1)\in\underline{E}_1, C) \leq P(Y(0)\in\underline{E}_0, C). \tag{5.24}$$

There always exists two such elements: since

$$P(Y(0)\in\underline{E}_0, C) + P(Y(0)\in(\underline{E}_0)^c, C) = P(Y(1)\in\underline{E}_1, C) + P(Y(1)\in(\underline{E}_1)^c, C),$$

either $P(Y(1) \in \underline{E}_1, C)$ or $P(Y(1) \in (\underline{E}_1)^c, C)$ is included between the min and the max of $P(Y(0) \in \underline{E}_0, C)$ and $P(Y(0) \in (\underline{E}_0)^c, C)$. Therefore, this is without loss of generality to assume that $P(Y(0) \in \underline{E}_0, C)$ is greater than $P(Y(0) \in (\underline{E}_0)^c, C)$, and that $P(Y(1) \in \underline{E}_1, C)$ is included between those two quantities. If it is not the case, one can merely consider $(\underline{E}_0)^c$ instead of $\underline{E}_0$, and / or $(\underline{E}_1)^c$ instead of $\underline{E}_1$.

Assume also that

$$P(Y(1) \in \underline{E}_1) - P(Y(1) \in \underline{E}_1, C) \geq P(Y(0) \in \underline{E}_0) - P(Y(0) \in \underline{E}_0, C). \tag{5.25}$$

This is also without loss of generality, given what is identified from the data. Indeed, under Assumptions 2.1, 2.2 and 2.3, for every $d \in \{0; 1\}$,

$$\begin{aligned} & P(Y(d) \in \underline{E}_d) - P(Y(d) \in \underline{E}_d, C) \\ = \; & P(Y(d) \in \underline{E}_d, D(d) = 1 - d) + P(Y \in \underline{E}_d, D = d | Z = 1 - d). \end{aligned}$$

The first term of this last inequality is not identified. Therefore, one can set for instance $P(Y(1) \in \underline{E}_1, D(1) = 0) = P(D = 0|Z = 1)$ and $P(Y(0) \in \underline{E}_0, D(0) = 1) = 0$ to have that Equation (5.25) holds.

Let

$$\begin{aligned}
&\tilde{P}((Y(0), Y(1)) \in \underline{E}_0 \times \underline{E}_1) = P(Y(1) \in \underline{E}_1, C) + P(Y(0) \in \underline{E}_0) - P(Y(0) \in \underline{E}_0, C) \\
&\tilde{P}((Y(0), Y(1)) \in \underline{E}_0 \times (\underline{E}_1)^c) = P(Y(0) \in \underline{E}_0, C) - P(Y(1) \in \underline{E}_1, C) \\
&\tilde{P}((Y(0), Y(1)) \in (\underline{E}_0)^c \times \underline{E}_1) = P(Y(1) \in \underline{E}_1) - P(Y(1) \in \underline{E}_1, C) - (P(Y(0) \in \underline{E}_0) - P(Y(0) \in \underline{E}_0, C)) \\
&\tilde{P}((Y(0), Y(1)) \in (\underline{E}_0)^c \times (\underline{E}_1)^c) = P(Y(1) \in (\underline{E}_1)^c) + P(Y(1) \in \underline{E}_1, C) - P(Y(0) \in \underline{E}_0, C).
\end{aligned}$$

Then, I extend $\tilde{P}$ to $\mathcal{E}_0 \times \mathcal{E}_1$ by setting

$$\begin{aligned}
& \tilde{P}((Y(0), Y(1)) \in E_0 \times E_1) \\
= \; & P(Y(0) \in E_0 | Y(0) \in \underline{E}_0) \times P(Y(1) \in E_1 | Y(1) \in \underline{E}_1) \times \tilde{P}((Y(0), Y(1)) \in \underline{E}_0 \times \underline{E}_1) \\
+ \; & P(Y(0) \in E_0 | Y(0) \in \underline{E}_0) \times P(Y(1) \in E_1 | Y(1) \in (\underline{E}_1)^c) \times \tilde{P}((Y(0), Y(1)) \in \underline{E}_0 \times (\underline{E}_1)^c) \\
+ \; & P(Y(0) \in E_0 | Y(0) \in (\underline{E}_0)^c) \times P(Y(1) \in E_1 | Y(1) \in \underline{E}_1) \times \tilde{P}((Y(0), Y(1)) \in (\underline{E}_0)^c \times \underline{E}_1) \\
+ \; & P(Y(0) \in E_0 | Y(0) \in (\underline{E}_0)^c) \times P(Y(1) \in E_1 | Y(1) \in (\underline{E}_1)^c) \times \tilde{P}((Y(0), Y(1)) \in (\underline{E}_0)^c \times (\underline{E}_1)^c).
\end{aligned}$$

One can check that under Equations (5.24) and (5.25), $\tilde{P}((Y(0), Y(1)) \in E_0 \times E_1)$ defines a probability measure on $\mathcal{E}_0 \times \mathcal{E}_1$.[10] Since $\mathcal{E}_0 \times \mathcal{E}_1$ is a semi-ring, $\tilde{P}((Y(0), Y(1)) \in E_0 \times E_1)$ can be extended to $\mathcal{E}$ using Carathéodory's extension Theorem. One can also check that $\tilde{P}$ is consistent with the true marginal distributions of $Y(0)$ and $Y(1)$.

[10] To prove σ-additivity, one can use similar arguments than those used in the proof of the sharpness of B^+_{IV}.

Let P' be a measure belonging to the set of measures $\mathcal{P}(\tilde{P})$. $\mathcal{P}(\tilde{P})$ denotes the set of measures satisfying requirements 1, 2 and 3 listed above, except that $\tilde{P}$ is substituted to P in the first requirement. $\mathcal{P}(\tilde{P})$ is the set of all the probability distributions of $(Y(0), Y(1))$ for compliers, which are compatible with the data, and which would be compatible with the true distribution of $(Y(0), Y(1))$ in the entire population, if this true distribution was $\tilde{P}$.

To satisfy requirement 1, P' must verify (among other things):

$$P'((Y(0), Y(1)) \in \underline{E}_0 \times (\underline{E}_1)^c, C) \leq P(Y(0) \in \underline{E}_0, C) - P(Y(1) \in \underline{E}_1, C).$$

Assume that

$$P'((Y(0), Y(1)) \in (\underline{E}_0)^c \times \underline{E}_1, C) > 0.$$

For requirement 3 to hold, this implies that we must have

$$P'((Y(0), Y(1)) \in \underline{E}_0 \times \underline{E}_1, C) < P(Y(1) \in \underline{E}_1, C),$$

which in turn implies that we must have

$$P'((Y(0), Y(1)) \in \underline{E}_0 \times (\underline{E}_1)^c, C) > P(Y(0) \in \underline{E}_0, C) - P(Y(1) \in \underline{E}_1, C),$$

for requirement 2 to hold. But this last equation would violate requirement 1. This implies that for requirements 1, 2 and 3 to hold, we must have

$$\begin{aligned}
&P'((Y(0), Y(1)) \in \underline{E}_0 \times \underline{E}_1, C) = P(Y(1) \in \underline{E}_1, C) \\
&P'((Y(0), Y(1)) \in \underline{E}_0 \times (\underline{E}_1)^c, C) = P(Y(0) \in \underline{E}_0, C) - P(Y(1) \in \underline{E}_1, C) \\
&P'((Y(0), Y(1)) \in (\underline{E}_0)^c \times \underline{E}_1, C) = 0 \\
&P'((Y(0), Y(1)) \in (\underline{E}_0)^c \times (\underline{E}_1)^c, C) = P(Y(0) \in (\underline{E}_0)^c, C).
\end{aligned} \tag{5.26}$$

Then, one can extend P', first to $\mathcal{E}_0 \times \mathcal{E}_1$, and then to $\mathcal{E}$, following the same steps as for $\tilde{P}$. This proves that $\mathcal{P}(\tilde{P})$ is not empty, and that all its elements must verify equation (5.26).

We either have $P(Y(0) \in \underline{E}_0, C) > P(Y(1) \in \underline{E}_1, C)$ or $P(Y(1) \in \underline{E}_1, C) > 0$. Indeed, $P(Y(0) \in \underline{E}_0, C) = P(Y(1) \in \underline{E}_1, C) = 0$ combined with equation (5.24) would imply $P(C) = 0$. If $P(Y(0) \in \underline{E}_0, C) > P(Y(1) \in \underline{E}_1, C)$,

$$P'(C|(Y(0), Y(1)) \in \underline{E}_0 \times (\underline{E}_1)^c) = 1.$$

If $P(Y(0) \in \underline{E}_0, C) = P(Y(1) \in \underline{E}_1, C)$ and $P(Y(1) \in \underline{E}_1, C) > 0$,

$$P'(C|(Y(0), Y(1)) \in \underline{E}_0 \times \underline{E}_1) = 1.$$

This proves that for every $P' \in \mathcal{P}(\tilde{P})$, we always have $P'^{+}(C|(Y(0), Y(1)) \in E) = 1$. This implies that $\tilde{P}^{+}(C|Y(0); Y(1)) = 1$. This proves the result.

Proof of Theorem 3.4

Let k be the cardinal of $\mathcal{S}(Y(d))$. For every y_d in $\mathcal{S}(Y(d))$ and every strictly positive real number p, let

$$\begin{aligned}
\underline{m}(p, y_d) &= \left(\frac{1\{Y = y_d, D = d, Z = d\}}{P(Z = d)} - \frac{1\{Y = y_d, D = d, Z = 1-d\}}{P(Z = 1-d)}\right) \\
&- \frac{1\{Y = y_d, D = d, Z = d\}}{P(Z = d)} p \\
\overline{m}(p, E_d) &= \frac{1\{Y = y_d, D = d, Z = d\} + 1\{D = 1-d, Z = d\}}{P(Z = d)} p \\
&- \left(\frac{1\{Y = y_d, D = d, Z = d\}}{P(Z = d)} - \frac{1\{Y = y_d, D = d, Z = 1-d\}}{P(Z = 1-d)}\right).
\end{aligned}$$

$P^+(C|Y(d))$ satisfies the $2k$ following moment inequalities: for every y_d in $\mathcal{S}(Y(d))$,

$$\begin{aligned}
E\left(\underline{m}(P^+(C|Y(d)), E_d)\right) &\geq 0 \\
E\left(\overline{m}(P^+(C|Y(d)), E_d)\right) &\geq 0.
\end{aligned}$$

Let

$$\begin{aligned}
\underline{\sigma}^2_{y_d}(p) &= Var(\underline{m}(p, y_d)) \\
\overline{\sigma}^2_{y_d}(p) &= Var(\overline{m}(p, y_d)).
\end{aligned}$$

Let also

$$T_n(p) = \sum_{y_d \in \mathcal{S}(Y(d))} \left[\sqrt{n}\frac{\widehat{E}\left(\underline{m}(p), y_d)\right)}{\widehat{\underline{\sigma}}_{y_d}(p)}\right]^2_- + \sum_{y_d \in \mathcal{S}(Y(d))} \left[\sqrt{n}\frac{\widehat{E}\left(\overline{m}(p, y_d)\right)}{\widehat{\overline{\sigma}}_{y_d}(p)}\right]^2_-,$$

where $[x]_- = x$ if $x < 0$ and 0 otherwise. Let $\underline{K}(p, y_d)$ be an indicator variable equal to 1 when $\sqrt{n}\frac{\widehat{E}(\underline{m}(p,y_d))}{\widehat{\underline{\sigma}}_{y_d}(P^+(C|Y(d)))} \leq \sqrt{2\ln(\ln(n))}$. Conversely, let $\overline{K}(p, y_d)$ be an indicator variable equal to 1 when $\sqrt{n}\frac{\widehat{E}(\overline{m}(p,y_d))}{\widehat{\overline{\sigma}}_{y_d}(p)} \leq \sqrt{2\ln(\ln(n))}$. Let CS_n be the confidence interval obtained inverting $T_n(p)$, using as critical values $c_{1-\alpha}(p)$ the $(1-\alpha)^{th}$ quantile of the distribution of

$$\sum_{y_d \in \mathcal{S}(Y(d)), \underline{K}(p,y_d)=1} [\underline{N}_{y_d}]^2_- + \sum_{y_d \in \mathcal{S}(Y(d)), \overline{K}(p,y_d)} [\overline{N}_{y_d}]^2_-,$$

where $\left(\underline{N}_{y_d}, \overline{N}_{y_d}\right)_{y_d \in \mathcal{S}(Y(d))}$ is a vector of $\mathcal{N}(0,1)$ random variables with a variance Ω equal to the asymptotic variance of

$$\left(\sqrt{n}\frac{\widehat{E}\left(\underline{m}(p, y_d)\right) - E\left(\underline{m}(p, y_d)\right)}{\widehat{\underline{\sigma}}_{y_d}(p)}, \sqrt{n}\frac{\widehat{E}\left(\overline{m}(p, y_d)\right) - E\left(\overline{m}(p, y_d)\right)}{\widehat{\overline{\sigma}}_{y_d}(p)}\right)_{y_d \in \mathcal{S}(Y(d))}.$$

Formally,

$$CS_n = \{p \,/\, T_n(p) \leq c_{1-\alpha}(p)\}.$$

Since for every $d \in \{0;1\}$, $P(Z = d) \geq \varepsilon$, all the technical conditions in Andrews & Soares (2010) apply here since $\underline{m}(p, y_d)$ and $\overline{m}(p, y_d)$ are bounded. Therefore, I invoke their Theorem 1 to assert that CS_n is a uniformly valid and not conservative confidence interval for $P^+(C|Y(d))$ with coverage rate $1 - \alpha$.

QED.