IsoLATEing:
Identifying Heterogeneous Effects of Multiple Treatments

Peter Hull[*]

December 2014

PRELIMINARY: Please do not cite or distribute without permission.

Please see www.mit.edu/~hull/research.html for the most current draft.

Abstract

This note explores identification of Local Average Treatment Effects (LATEs) by Instrumental Variables (IV) when treatment consists of multiple observable channels. I show that if there exists a stratification of the sample that induces variation in the composition of behavioral groups while maintaining independence of a binary instrument, an IV regression on multiple endogenous variables can identify channel-specific LATEs. In the simplest case of a binary stratification and two channels, mean-independence of one LATE with respect to the stratification is sufficient for point-identification of a weighted average of stratum-specific LATEs of the other causal channel. When this homogeneity condition fails, the IV estimands can yield intuitive bounds for this weighted average. I outline an application of these results for recovering treatment effects from Randomized Control Trials suffering from differential attrition.

[*] Department of Economics, Massachusetts Institute of Technology, email: hull@mit.edu. Joshua Angrist and Carl Jack Liebersohn provided valuable feedback.

1 Motivation

Applied empirical research is often concerned with quantifying the effect of a binary treatment D on an outcome Y for a set of individuals for which treatment is determined by randomized assignment, denoted by Z.[1] Write an individual's potential values of D under different realizations of Z by $\{D_1, D_0\}$ and of Y given realizations of Z and D by $\{Y_{11}, Y_{01}, Y_{10}, Y_{00}\}$.[2] As Imbens and Angrist (1994) show, under the following three assumptions,

Assumption IA1 (Independence): $\{Y_{11}, Y_{01}, Y_{10}, Y_{00}, D_1, D_0\} \perp\!\!\!\perp Z$,

Assumption IA2 (Exclusion): $P(Y_{0d} = Y_{1d}) = 1$, $d \in \{0, 1\}$,

Assumption IA3 (Monotonicity): $P(D_1 \geq D_0) = 1$,

the Instrumental Variables (IV) regression of Y on D instrumented by Z identifies:

$$\frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} = E[Y_1 - Y_0|D_1 > D_0] \tag{1}$$

where $Y_1 \equiv Y_{z1}$ and $Y_0 \equiv Y_{z0}$, well-defined by Assumption IA2. Imbens and Angrist (1994) refer to this IV estimand as the Local Average Treatment Effect (LATE), or the average effect of treatment D on the outcome for *compliers*, the subpopulation of individuals who only participate in treatment if assigned.

When treatment effects vary, it is of natural interest to characterize their heterogeneity. Consider a stratification of individuals with strata indexed by a covariate X. When Assumptions IA1-IA3 hold conditional on X (that is, within each stratum), stratum-specific LATEs are identified by conditional IV estimands. For example, in a Randomized Control Trial (RCT) in which offers for a particular program are randomly assigned to individuals, stratification on characteristics determined prior to randomization can reveal differential effects of the program on compliers with different baseline observables.

Often, however, the treatment effect heterogeneity of interest is unlikely to stem from stratification schemes satisfying Independence (Assumption IA1). For example, an individual's participation in an RCT may be measured as casual or intensive, and a researcher may wish to know just the effect among individuals that choose to participate intensely. Similarly, in an RCT with imperfect followup of a self-reported outcome, a researcher may be tempted to restrict the estimation of treatment effects to the sample where outcomes are observed. These "endogenous stratifications" on a

[1] Individual subscripts are omitted on all random variables in this note for notational clarity.

[2] That such potential values are well-defined random variables is a consequence of the Stable Treatment Unit Value Assumption (STUVA), discussed in detail in Angrist, Imbens, and Rubin (1996). I implicitly assume STUVA holds throughout this note.

covariate determined after assignment of Z (that is, an individual's participation level or decision to report Y) is likely to introduce selection bias in the estimation of LATEs (Angrist and Pischke, 2009). In general, point identification of average causal effects for individuals with particular post-assignment characteristics is notoriously difficult (Frangakis and Rubin, 2002), though certain dimensions of heterogeneity are identified under sufficiently strong assumptions.[3]

This note considers nonparametric identification of endogenous stratifications that can be modeled as distinguishing between average causal effects across multiple measurable treatment channels. That is, I allow a binary instrument Z to cause individuals to shift into two or more mutually-exclusive treatment states, and consider joint identification of state-specific Local Average Treatment Effects by potential treatment takeup among individuals who would comply to any treatment. Note that a multiple-treatment framework can also be applied to a setting with multiple "fallbacks;" in the differential attrition example above, those who are not treated can either be in an attrition state or can report their outcomes. The usual causal object of interest is then the LATE for compliers who would still produce an outcome if not treated.

The strategy I propose to "isoLATE" different endogenous causal channels is intuitive: if there exists an "exogenous" stratification across which the composition of individuals with different potential treatment takeup behavior varies but the average causal effects of interest do not, differences in strata-specific IV estimates can be attributed to differences in estimated complier shares in such a way that identifies each LATE. In particular, I show that an IV regression with multiple endogenous variables identified by interactions of the instrument with stratum indicators correctly solves for the causal parameters in this case. The requirement that average complier treatment effects are the same across strata may be too stringent, however. In general I show that IV can identify an interpretable weighted average of one treatment's LATE provided the other causal channel is mean-independent of the stratification, and that the bias due to failures of mean-independence can be bounded for bounded outcomes. I propose an application of these results to the prominent issue of differential attrition in an RCT, and show that the treatment effect of interest can be point-identified by IV when researchers follow conventional wisdom of randomly designating a subset of observed attriters for more intensive follow-up attempts.

The remainder of this note is organized as follows: in section 2 I extend the Imbens and Angrist (1994) framework to allow for multiple treatments, noting similarities and departures from existing models. Section 3 contains the main identification result and highlights special cases of interest. Section 4 concludes with a discussion of possible applications.

[3] See, for example, Heckman and Vytlacil (2005) and Jin and Rubin (2008).

2 Setting and Assumptions

We have a binary instrument Z, an outcome Y, a binary covariate X, and three possible treatment states $t \in \{0, a, b\}$.[4] Denote indicators for treatment a and b by A and B, respectively. Potential treatment indicators given Z are written A_z and B_z for $z \in \{0, 1\}$ and potential outcomes given Z and the treatment state are written Y_{zt}.

Consider the following four assumptions:

Assumption 1 (Independence): $\{Y_{1a}, Y_{0a}, Y_{1b}, Y_{0b}, Y_{10}, Y_{00}, A_1, A_0, B_1, B_0\} \perp\!\!\!\perp Z|X$,

Assumption 2 (Exclusion): $P(Y_{0t} = Y_{1t}) = 1$, $t \in \{0, a, b\}$,

Assumption 3 (Monotonicity): $P(A_1 \geq A_0) = P(B_1 \geq B_0) = 1$

Assumption 4 (No switching): $P(A_1 = B_0 = 1) = P(B_1 = A_0 = 1) = 0$.

Assumptions 1 and 2 extend the canonical Assumptions IA1 and IA2 to accomodate the expanded set of potential outcomes and treatment responses, with Assumption 1 also modified to require the instrument to satisfy independence within each stratum. Assumptions 3 and 4 restrict the behavioral response to Z: if the instrument affects an individual's treatment status, it is only allowed to *induce* treatment (Assumption 3), and may not induce switching *between* treatments (Assumption 4). These assumptions thus restrict the number of behavioral subpopulations to five:

1. Always-a-takers: $A_1 = A_0 = 1$,
2. Always-b-takers: $B_1 = B_0 = 1$,
3. Never-takers: $A_1 = A_0 = B_1 = B_0 = 0$,
4. a-compliers: $A_1 = 1$, $A_0 = 0$, $B_0 = 0$,
5. b-compliers: $B_1 = 1$, $B_0 = 0$, $A_0 = 0$

It is simple to verify that Assumptions 1-4 imply (conditional versions of) Assumptions IA1-IA3 on the combined treatment $D = A + B$, with the "no switching" condition given by Assumption 4 sufficient to ensure no violation of the exclusion restriction. When $P(Y_{za} \neq Y_{zb}) > 0$ (that is,

[4] It is straightforward to extend the assumptions and results of this note to allow for more than two treated states, provided the number is no more than the number of distinct elements in the support of X.

the treatments are actually distinct), Assumptions 1-4 are *equivalent* to the LATE assumptions for (Y, D, Z, X). In this case the strata-specific IV estimand is

$$E[Y_1 - Y_0|D_1 > D_0, X] = \frac{E[Y_a - Y_0|A_1 > A_0, X]P(A_1 > A_0|X) + E[Y_b - Y_0|B_1 > B_0, X]P(B_1 > B_0|X)}{P(A_1 > A_0|X) + P(B_1 > B_0|X)} \tag{2}$$

Here the goal is to recover the two causal channels from this weighted average.

Although Assumptions 1-4 ensure excludability of Z from Y given D, they do not rule out violations of exclusion when either A or B is considered in isolation. This is because an individual not observed in treatment state a may still be induced into treatment b by the instrument if she is a b-complier, and similarly for b. Thus the results of this note can also be thought of as addressing cases of "known" exclusion restriction violations, where the instrument affects untreated (from the perspective of a) individuals by shifting them to another observable state b.[5]

We can interpret Assumptions 1-4 as producing a slightly restricted version of the ordered treatment model first studied by Angrist and Imbens (1995). Namely, with $S \equiv A + 2B$ and potentials defined accordingly, the ordered treatment setting is equivalent to Assumptions 1-3 and Assumption 4 modified to only require $P(A_1 = B_0 = 1) = 0$. In this case Assumptions 3 and 4 are equivalent to $P(S_1 \geq S_0) = 1$ and the method outlined below concerns identification of, in the notation of Angrist and Imbens (1995):

$$E[Y_a - Y_0|A_1 > A_0] = E[Y_1 - Y_0|S_1 = 1, S_0 = 0] \tag{3}$$

$$E[Y_b - Y_0|B_1 > B_0] = E[Y_2 - Y_0|S_1 = 2, S_0 = 0], \tag{4}$$

which are average effects for the only two complier groups given the modified Assumption 4.

Finally, we can link Assumptions 1-4 to the two-treatment encouragement design considered by Behaghel, Crepon, and Gurgand (2013), who model three treatments as well as an instrument that takes on three values, $\tilde{Z} \in \{0, a, b\}$. With $Z \equiv \mathbf{1}\{\tilde{Z} \neq 0\}$ and the underlying value of $\tilde{Z}$ absorbed by the heterogeneity of individuals, Assumptions 1-4 are implied by their Assumptions 1-3. The identification results derived here can thus be thought of extensions to their approach when only one instrument for treatment is available.

[5] For another, more concrete example of "known" exclusion restriction violations, see Abdulkadiroğlu, Angrist, Hull, and Pathak (2014).

3 Main Result

Denote the covariate-specific LATE for each treatment as

$$\alpha(x) \equiv E[Y_a - Y_0 | A_1 > A_0, X = x] \tag{5}$$

$$\beta(x) \equiv E[Y_b - Y_0 | B_1 > B_0, X = x] \tag{6}$$

It will also be useful to define, for a given random variable V, the function

$$f_V(x) \equiv E[V|Z = 1, X = x] - E[V|Z = 0, X = x] \tag{7}$$

We then have the following result:

Proposition 1 *Consider the IV regression of Y on the pair (A, B) instrumented by the pair $(Z, Z \cdot X)$, controlling for X. Under Assumptions 1-4 the endogenous regressor coefficients identify*

$$\alpha = \omega\alpha(0) + (1 - \omega)\alpha(1) + \delta_a(\beta(0) - \beta(1)) \tag{8}$$

$$\beta = (1 - \omega)\beta(0) + \omega\beta(1) + \delta_b(\alpha(0) - \alpha(1)), \tag{9}$$

where

$$\omega = \left(1 - \frac{f_A(1)}{f_B(1)} / \frac{f_A(0)}{f_B(0)}\right)^{-1} \tag{10}$$

and where

$$\delta_a = \left(\frac{f_A(0)}{f_B(0)} - \frac{f_A(1)}{f_B(1)}\right)^{-1} \tag{11}$$

$$\delta_b = \left(\frac{f_B(0)}{f_A(0)} - \frac{f_B(1)}{f_A(1)}\right)^{-1} = -\frac{f_A(0)}{f_B(0)}\frac{f_A(1)}{f_B(1)}\delta_a. \tag{12}$$

Furthermore,

$$f_A(x) = P(A_1 > A_0 | X = x) \tag{13}$$

$$f_B(x) = P(B_1 > B_0 | X = x). \tag{14}$$

Proof: *See appendix.*

Proposition 1 states that the two-endogenous variable IV estimand is a function of covariate-specific Local Average Treatment Effects and first-stage coefficients for A and B, with the latter identifying the relative share of a-compliers and b-compliers, respectively, in each stratum.

We can unpack these functions further by considering special cases of Proposition 1. As a natural benchmark, suppose the stratification scheme is such that

$$E[Y_a - Y_0 | A_1 > A_0, X = 0] = E[Y_a - Y_0 | A_1 > A_0, X = 1] \tag{15}$$

$$E[Y_b - Y_0 | B_1 > B_0, X = 0] = E[Y_b - Y_0 | B_1 > B_0, X = 1]. \tag{16}$$

Equations 15 and 16 require a form of "LATE homogeneity:" that the average treatment effect for a-compliers and b-compliers are the same across strata. This homogeneity restriction is fairly mild – no other moments of the joint potential outcomes distribution are restricted – though clearly its plausibility is context-specific. In an application of Proposition 1 in Section 4 I consider a stratification that is independent of potential outcomes, in which case equations 15 and 16 hold by design. Trivially, under full LATE homogeneity we have $\alpha = \alpha(0) = \alpha(1)$ and $\beta = \beta(0) = \beta(1)$, so that the two-endogenous variable IV specification perfectly recovers both causal channels in this case.[6]

We can also consider the IV estimand when either equations 15 or 16 hold, but not both. When this is the case one of the two IV coefficients identifies a weighted average of strata-specific LATEs, with weights given by ω and $1 - \omega$. These are proper weights in the sense they sum to one, though they are also guaranteed to not be convex. This is because all of the first-stage coefficients identify positive quantities (the proportion of a- and b-compliers in each stratum) so that $\omega \in (-\infty, 0) \cup (1, \infty)$. By inspection, these weights become better-behaved (in the sense that $\omega \to 1$ or $\omega \to 0$) as the relative shares of compliers become more heterogeneous across strata.

Without LATE homogeneity, IV estimates this weighted average, plus a bias term that is proportional to the difference in LATEs across strata. The coefficients on this difference are again functions of complier shares, and are of opposite signs for the treatment effect of a and of b. When the outcome Y is guaranteed to lie in the interval $[m, M]$, moreover, we can use equations 8-9 and 11-12 to bound a weighted average of LATEs, as, for example,

$$\begin{aligned}
\bar{\alpha} &\equiv \omega\alpha(0) + (1 - \omega)\alpha(1) \\
&= \alpha - \delta_a(\beta(0) - \beta(1)) \\
&\in \left[\alpha - 2|\delta_a|(M - m), \alpha + 2|\delta_a|(M - m)\right],
\end{aligned} \tag{17}$$

[6] Note that one can extend Proposition 1 to consider multi-valued X and the overidentified IV regression instrumented by some combination of stratum interactions. When equations 15 and 16 hold across all values of X it is straightforward to show α and β are identified by any such regression, just as in the constant effects case. A test of overidentifying restrictions would thus be valid for jointly testing Assumptions 1-4 and equations 15 and 16. Furthermore one can verify that when LATE homogeneity fails, 2SLS continues to estimate interpretable functions of $\alpha(x)$ and $\beta(x)$. See the appendix for these formulas.

and δ_a may be consistently estimated by the estimated first-stage coefficients. Note that the width of the identified set is proportional to $|\delta_a|$, which is smaller when strata induce very different shares of a- and b-compliers. Thus, loosely speaking, the interpretability of α and β depends on the degree of heterogeneity in first stage coefficients across strata.

4 Discussion

Proposition 1 establishes a framework for using "exogenous" stratifications to isolate distinct "endogenous" causal channels when the effects of treatment are heterogeneous. Although the formula for the two-endogenous variable IV estimand holds in general, it is most useful when the stratification scheme satisfies LATE homogeneity for at least one of the two channels. When the homogeneity assumption fails, tighter bounds on average causal parameters are obtained by choosing a stratification with the most heterogeneity in first stages.

As an example of the idealized conditions for Proposition 1, consider again the setting in which a researcher conducts a Randomized Control Trial for a given program in which there imperfect followup of the outcome Y for untreated individuals. The researcher randomly assigns treatment by the binary instrument $\tilde{Z}$. Subjects that subsequently take the treatment and produce an outcome are classified by $D = 1$, while subjects that do not take the treatment but whose outcomes are still measured are classified by $D = 0$. All other subjects leave the sample without generating an outcome; denote this state by $D = a$ and set $Y = 0$ for all such individuals (the choice of this value is arbitrary but simplifies the exposition).

The issue of differential attrition offers a natural application of Proposition 1. Define $A = \mathbf{1}\{D = 0\}$, $B = \mathbf{1}\{D = a\}$, and $Z = 1 - \tilde{Z}$. Given any stratification by a pre-assignment variable X, Assumption 1 is satisfied by virtue of the randomization of Z. Exclusion of Z from Y given A and B is similarly defensible within the context of an RCT. The behavioral restrictions of Assumptions 3 and 4 imply that if any individual is affected by Z they are induced into treatment from either an attrition or a non-attrition state, but that Z does not cause any subject to stay in the sample but not get treated. This too seems plausible given the experimental design.

When there are b-compliers in the attrition example, the IV regression of Y on $\mathbf{1}\{D = 1\}$ instrumented by $\tilde{Z}$ identifies a LATE with a mixed counterfactual, as in equation 2. Here "isoLATEing" the average treatment effect just for a-compliers is of first-order importance, as the potential outcome Y_0 for b-compliers is set to zero (or some other value), rendering the combined LATE in equation 2 causally meaningless. In the light of Proposition 1, however, all that is needed is a covariate X satisfying LATE homogeneity to identify

$$E[Y_a - Y_0 | A_1 > A_0] = -E[Y_1 - Y_0 | D_1 = 1, D_0 = 0]$$
$$E[Y_b - Y_0 | B_1 > B_0] = -E[Y_1 | D_1 = 1, D_0 = a],$$

since we've flipped the multiple treatment setting to a multiple "fallback" setting, and since $Y = 0$ whenever $D = a$.

For any particular RCT, which stratification is most likely to maintain homogeneous LATEs while still producing variation in complier shares across strata depends on the context. One stratification that is guaranteed to satisfy LATE homogeneity and thus fully recover $E[Y_1 - Y_0 | D_1 = 1, D_0 = 0]$, however, exploits a practice that is often given as conventional wisdom in conducting an RCT with imperfect follow up (e.g. Duflo, Glennerster, and Kremer (2008), p. 3944). Suppose, upon measuring outcomes, a researcher selects among the attritors a random sample of individuals for more intensive follow-up attempts. Denote this set by $X = 1$ and let $X = 0$ for all other observed attritors. For the set of individuals for which outcomes are observed, let the binary X be randomly assigned with some probability. Since the designation of X across individuals is random, all latent variables across this stratification should have the same distribution. In particular, $E[Y_1 | D_1 = 1, D_0 = a, X]$ should not depend on X, nor should $E[Y_1 - Y_0 | D_1 = 1, D_0 = 0, X]$. Importantly, however, if the intensive follow-up is successful in attracting more outcomes from initial attritors, the subsample given by $X = 1$ should have a smaller proportion of b-compliers. One could thus point-identify the Local Average Treatment Effect of interest using Proposition 1, as well as the average treated outcome of potential attritors. This is one of perhaps many applications of this result for solving pervasive issues in program evaluation with multiple treatment states.

References

Abdulkadiroğlu, A., J. D. Angrist, P. D. Hull, and P. A. Pathak (2014): “Charters without Lotteries: Testing Takeovers in New Orleans and Boston,” NBER Working Paper 20792.

Angrist, J. D., and G. W. Imbens (1995): “Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity,” *Journal of the American Statistical Association*, 430(90), 431–442.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): “Identification of Causal Effects Using Instrumental Variables,” *Journal of the American Statistical Association*, 434(91), 444–455.

Angrist, J. D., and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.

Behaghel, L., B. Crepon, and M. Gurgand (2013): “Robustness of the Encouragement Design in a Two-Treatment Randomized Control Trial,” IZA Discussion Paper 7447.

Duflo, E., R. Glennerster, and M. Kremer (2008): “Using Randomization in Development Economics Research: A Toolkit,” Handbook of Development Economics 4, pp. 3895–3962. Elsevier.

Frangakis, C. E., and D. B. Rubin (2002): “Principal Stratification in Causal Inference,” *Biometrics*, 58, 21–29.

Heckman, J. J., and E. Vytlacil (2005): “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 73(3), 669–738.

Imbens, G. W., and J. D. Angrist (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62(2), 467–475.

Jin, H., and D. B. Rubin (2008): “Principal Stratification for Causal Inference with Extended Partial Compliance,” *Journal of the American Statistical Association*, 481(103), 101–111.

Appendix: Proof of Proposition 1

Consider the reduced-form regression of Y on Z, conditional on X. By the excludability of Z given treatment status (Assumption 2), we can write

$$Y = Y_0 + (Y_a - Y_0)A + (Y_b - Y_0)B,$$

so that

$$\begin{aligned}
E[Y|Z=1,X] - E[Y|Z=0,X] =& E[Y_0 + (Y_a - Y_0)A + (Y_b - Y_0)B|Z=1,X] \\
&- E[Y_0 + (Y_a - Y_0)A + (Y_b - Y_0)B|Z=0,X] \\
=& E[Y_0|Z=1,X] - E[Y_0|Z=0,X] \\
&+ E[(Y_a - Y_0)A_1|Z=1,X] - E[(Y_a - Y_0)A_0|Z=0,X] \\
&+ E[(Y_b - Y_0)B_1|Z=1,X] - E[(Y_b - Y_0)B_0|Z=0,X] \\
=& E[(Y_a - Y_0)(A_1 - A_0)|X] + E[(Y_b - Y_0)(B_1 - B_0)|X] \\
=& E[Y_a - Y_0|A_1 > A_0, X]P(A_1 > A_0|X) \\
&+ E[Y_b - Y_0|B_1 > B_0, X]P(B_1 > B_0|X)
\end{aligned}$$

where the third equality follows by independence of Z given X (Assumption 1) and the fourth by the behavioral restrictions of Assumptions 3 and 4. Furthermore, the first-stage regression for A is

$$\begin{aligned}
E[A|Z=1,X] - E[A|Z=0,X] &= E[A_1 - A_0|X] \\
&= P(A_1 > A_0|X)
\end{aligned}$$

and similarly for B. This again follows by the independence assumption and Assumptions 3-4. Thus we can write

$$f_Y(x) = \alpha(x) f_A(x) + \beta(x) f_B(x) \tag{18}$$

for each $x \in \{0,1\}$.

Now consider the multiple-endogenous variable IV regression of interest. With $\mathbf{Y}$ denoting the vector of observations of Y, $\mathbf{X}$ denoting the matrix of observations of A, B, and X, and $\mathbf{Z}$ denoting the matrix of observations of Z, $Z \cdot X$, and X we have

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = p\lim\left((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right) \tag{19}$$

where $\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{X}}$ are the residuals obtained by regressing $\mathbf{Z}$ and $\mathbf{X}$ on X and a constant. Furthermore, since the regression is just-identified,

$$\begin{aligned}\begin{bmatrix}\alpha \\ \beta\end{bmatrix} &= p\lim\left((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}})^{-1}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right) \\ &= p\lim\left(((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}})^{-1}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right).\end{aligned}$$

Note however that by construction,

$$\begin{aligned}p\lim\left((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}}\right) &= \begin{bmatrix}f_A(0) & f_B(0) \\ f_A(1) & f_B(1)\end{bmatrix} \\ p\lim\left((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right) &= \begin{bmatrix}f_Y(0) \\ f_Y(1)\end{bmatrix}.\end{aligned}$$

Thus, by the continuous mapping theorem, Slutsky's theorem, and equation 18:

$$\begin{aligned}\begin{bmatrix}\alpha \\ \beta\end{bmatrix} &= \begin{bmatrix}f_A(0) & f_B(0) \\ f_A(1) & f_B(1)\end{bmatrix}^{-1}\begin{bmatrix}f_Y(0) \\ f_Y(1)\end{bmatrix} \\ &= \begin{bmatrix}f_A(0) & f_B(0) \\ f_A(1) & f_B(1)\end{bmatrix}^{-1}\begin{bmatrix}\alpha(0)f_A(0) + \beta(0)f_B(0) \\ \alpha(1)f_A(1) + \beta(1)f_B(1)\end{bmatrix} \qquad (20)\end{aligned}$$

The proposition follows by simplifying this expression. □

Extension of Proposition 1 to the overidentified case

Let $\{X_j\}$ be the set of J mutually-exclusive and exhaustive indicators for the discrete covariate X taking on each value in its support, and consider the 2SLS regression of Y on the pair (A, B) instrumented by $\{Z \cdot X_j\}$ and controlling for $\{X_j\}$. Again with $\mathbf{Y}$ denoting the vector of observations of Y, $\mathbf{X}$ denoting the matrix of observations of A, B, and $\{X_j\}$, and $\mathbf{Z}$ denoting the matrix of observations of $\{Z \cdot X_j\}$, and $\{X_j\}$ we now have:

$$\begin{bmatrix}\alpha \\ \beta\end{bmatrix} = p\lim\Big((\widetilde{\mathbf{X}}'\widetilde{\mathbf{Z}}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{Z}}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\Big) \tag{21}$$

$$= p\lim\left(\left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{Z}}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\frac{\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}}{N}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{Z}}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\frac{\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}}{N}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right) \tag{22}$$

where, as before, $\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{X}}$ are the residuals obtained by regressing $\mathbf{Z}$ and $\mathbf{X}$ on $\{X_j\}$. Again:

$$p\lim\Big((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}}\Big) = \begin{bmatrix} f_A(1) & f_B(1) \\ \vdots & \vdots \\ f_A(J) & f_B(J)\end{bmatrix}$$

$$p\lim\Big((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\Big) = \begin{bmatrix} f_Y(1) \\ \vdots \\ f_Y(J)\end{bmatrix},$$

and furthermore

$$p\lim\left(\frac{\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}}{N}\right) = \begin{bmatrix} \pi(1)\sigma_Z^2(1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \pi(J)\sigma_Z^2(J)\end{bmatrix},$$

where we redefine $f_V(j) \equiv E[V|Z = 1, X_j = 1] - E[V|Z = 0, X_j = 1]$ and where $\sigma_Z^2(j) \equiv Var(Z|X_j = 1)$ and $\pi(j) \equiv P(X_j = 1)$. Under Assumptions 1-4 Equation 18 still holds with the definitions of $\alpha(x)$ and $\beta(x)$ similarly modified. Thus again by the continuous mapping theorem, Slutsky's theorem, and Equation 22:

$$\begin{bmatrix}\alpha \\ \beta\end{bmatrix} = \left(\begin{bmatrix} f_A(1) & \dots & f_A(J) \\ f_B(1) & \dots & f_B(J) \end{bmatrix} \begin{bmatrix} \pi(1)\sigma_Z^2(1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \pi(J)\sigma_Z^2(J) \end{bmatrix} \begin{bmatrix} f_A(1) & f_B(1) \\ \vdots & \vdots \\ f_A(J) & f_B(J) \end{bmatrix} \right)^{-1}$$
$$\begin{bmatrix} f_A(1) & \dots & f_A(J) \\ f_B(1) & \dots & f_B(J) \end{bmatrix} \begin{bmatrix} \pi(1)\sigma_Z^2(1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \pi(J)\sigma_Z^2(J) \end{bmatrix} \begin{bmatrix} \alpha(1)f_A(1) + \beta(1)f_B(1) \\ \vdots \\ \alpha(J)f_A(J) + \beta(J)f_B(J) \end{bmatrix} \tag{23}$$

Simplifying this expression for, say, α yields the desired extension of Proposition 1:

$$\alpha = \sum_j \frac{\omega_j^a \alpha(j)}{\sum_k \omega_k^a} + \sum_j \frac{\delta_j^a(\beta(j) - \beta(j^{++}))}{\sum_k \omega_k^a} \tag{24}$$

where $j^{++} \equiv j + 1$ for $j < J$ and 1 otherwise, and where

$$\omega_j^a = \pi(j)\sigma_Z^2(j) f_A(j) \left(\sum_k \pi(k)\sigma_Z^2(k) f_B(k) \left(f_A(j) f_B(k) - f_B(j) f_A(k) \right) \right) \tag{25}$$

$$\delta_j^a = \pi(j)\sigma_Z^2(j) f_B(j) \pi(j^{++}) \sigma_Z^2(j^{++}) f_B(j^{++}) \left(f_A(j) f_B(j^{++}) - f_B(j) f_A(j^{++}) \right). \tag{26}$$

Thus the overidentified 2SLS regression again estimates a weighted average of strata-specific LATEs plus a bias term which is a function of pairwise differences in average causal effects for compliers to the other causal channel. As the strata-specific variance of Z, histogram of X, and first stage coefficients can be consistently estimated, we can again compute bounds for the weighted averages of $\alpha(j)$ and $\beta(j)$ given an outcome with bounded support.