NBER WORKING PAPER SERIES

ON HECKITS, LATE, AND NUMERICAL EQUIVALENCE

Patrick M. Kline
Christopher R. Walters

Working Paper 24477
http://www.nber.org/papers/w24477

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2018

We thank Josh Angrist, David Card, James Heckman, Magne Mogstad, Parag Pathak, Demian Pouzo, Raffaele Saggio, and Andres Santos for helpful discussions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

On Heckits, LATE, and Numerical Equivalence
Patrick M. Kline and Christopher R. Walters
NBER Working Paper No. 24477
April 2018
JEL No. C01,C26,C31,C52

ABSTRACT

Structural econometric methods are often criticized for being sensitive to functional form assumptions. We study parametric estimators of the local average treatment effect (LATE) derived from a widely used class of latent threshold crossing models and show they yield LATE estimates algebraically equivalent to the instrumental variables (IV) estimator. Our leading example is Heckman's (1979) two-step ("Heckit") control function estimator which, with two-sided non-compliance, can be used to compute estimates of a variety of causal parameters. Equivalence with IV is established for a semi-parametric family of control function estimators and shown to hold at interior solutions for a class of maximum likelihood estimators. Our results suggest differences between structural and IV estimates often stem from disagreements about the target parameter rather than from functional form assumptions per se. In cases where equivalence fails, reporting structural estimates of LATE alongside IV provides a simple means of assessing the credibility of structural extrapolation exercises.

Patrick M. Kline
Department of Economics
University of California, Berkeley
530 Evans Hall #3880
Berkeley, CA 94720
and NBER
pkline@econ.berkeley.edu

Christopher R. Walters
Department of Economics
University of California, Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
crwalters@econ.berkeley.edu

1 Introduction

In a seminal paper, Imbens and Angrist (1994) studied conditions under which the instrumental variables (IV) estimator is consistent for the Local Average Treatment Effect (LATE) – an average effect for a subpopulation of "compliers" compelled to change treatment status by an external instrument. The plausibility and transparency of these conditions is often cited as an argument for preferring IV to nonlinear estimators based on parametric models (Angrist and Pischke, 2009, 2010). On the other hand, LATE itself has been criticized as difficult to interpret, lacking in policy relevance, and problematic for generalization (Heckman, 1997; Deaton, 2009; Heckman and Urzua, 2010). Adherents of this view favor estimators motivated by joint models of treatment choice and outcomes with structural parameters defined independently of the instrument at hand.

This note develops some connections between IV and structural estimators intended to clarify how the choice of estimator affects the conclusions researchers obtain in practice. Our primary result is that, in the familiar binary instrument/binary treatment setting with imperfect compliance, a wide array of structural "control function" estimators derived from parametric threshold-crossing models yield LATE estimates numerically identical to IV. Notably, this equivalence applies to appropriately parameterized variants of Heckman's (1976; 1979) classic two-step ("Heckit") estimator that are nominally predicated on bivariate normality. Differences between structural and IV estimates therefore stem in canonical cases entirely from disagreements about the target parameter rather than from functional form assumptions.

After considering how our results extend to settings with instruments taking multiple values, we briefly probe the limits of our findings by examining some estimation strategies where equivalence fails. First, we revisit a control function estimator considered by LaLonde (1986) and show that it produces results identical to IV only under a symmetry condition on the estimated probability of treatment. Next, we study an estimator motivated by a selection model that violates the monotonicity condition of Imbens and Angrist (1994) and establish that it yields a LATE estimate different from IV, despite fitting the same sample moments. Standard methods of introducing observed covariates also break the equivalence of control function and IV estimators, but we discuss a reweighting approach that ensures equivalence is restored. We then consider full information maximum likelihood (FIML) estimation of some generalizations of the textbook bivariate probit model and show that this yields LATE estimates that coincide with IV at interior solutions. However, FIML diverges from IV when the likelihood is maximized on the boundary of the structural parameter space, which serves as the basis of recent proposals for testing instrument validity in just-identified settings (Huber and Mellace, 2015; Kitagawa, 2015). Finally, we discuss why estimation of over-identified models generally yields LATE estimates different from IV.

The equivalence results developed here provide a natural benchmark for assessing the credibility of structural estimators, which typically employ a number of over-identifying restrictions in practice. As Angrist and Pischke (2010) note: "A good structural model might tell us something about economic mechanisms as well as causal effects. But if the information about mechanisms is to be worth anything, the structural estimates should line up with those derived under weaker assumptions." Comparing the model-based LATEs implied by structural

estimators with unrestricted IV estimates provides a transparent assessment of how conclusions regarding a common set of behavioral parameters are influenced by the choice of estimator. A parsimonious structural estimator that rationalizes a variety of IV estimates may reasonably be deemed to have survived a "trial by fire," lending some credibility to its predictions.

2 Two views of LATE

We begin with a review of the LATE concept and its link to IV estimation. Let Y_i represent an outcome of interest for individual i, with potential values Y_{i1} and Y_{i0} indexed against a binary treatment D_i. Similarly, let D_{i1} and D_{i0} denote potential values of the treatment indexed against a binary instrument Z_i. Realized treatments and outcomes are linked to their potential values by the relations $D_i = Z_i D_{i1} + (1 - Z_i) D_{i0}$ and $Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0}$. Imbens and Angrist (1994) consider instrumental variables estimation under the following assumptions:

IA.1 Independence/Exclusion: $(Y_{i1}, Y_{i0}, D_{i1}, D_{i0}) \perp\!\!\!\perp Z_i$.

IA.2 First Stage: $Pr[D_i = 1|Z_i = 1] > Pr[D_i = 1|Z_i = 0]$.

IA.3 Monotonicity: $Pr[D_{i1} \geq D_{i0}] = 1$.

Assumption IA.1 requires the instrument to be as good as randomly assigned and to influence outcomes only through its effect on D_i. Assumption IA.2 requires the instrument to increase the probability of treatment, and assumption IA.3 requires the instrument to weakly increase treatment for all individuals.

Imbens and Angrist (1994) define LATE as the average treatment effect for "compliers" induced into treatment by the instrument (for whom $D_{i1} > D_{i0}$). Assumptions IA.1-IA.3 imply that the population Wald (1940) ratio identifies LATE:

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[Y_{i1} - Y_{i0}|D_{i1} > D_{i0}] \equiv LATE.$$

Suppose we have access to an iid vector of sample realizations $\{Y_i, D_i, Z_i\}_{i=1}^n$ obeying the following assumption:

Assumption 1. $\frac{1}{\sum_i Z_i} \sum_i Z_i D_i \neq \frac{1}{\sum_i (1-Z_i)} \sum_i (1 - Z_i) D_i$.

When IA.2 is satisfied the probability of Assumption 1 being violated approaches zero at an exponential rate in n. The analogy principle suggests estimating LATE with:

$$\widehat{LATE}^{IV} = \frac{\frac{1}{\sum_i Z_i} \sum_i Z_i Y_i - \frac{1}{\sum_i (1-Z_i)} \sum_i (1 - Z_i) Y_i}{\frac{1}{\sum_i Z_i} \sum_i Z_i D_i - \frac{1}{\sum_i (1-Z_i)} \sum_i (1 - Z_i) D_i}.$$

This IV estimator is well-defined under Assumption 1, and is consistent for $LATE$ under assumptions IA.1-IA.3 and standard regularity conditions.

Threshold-crossing representation

Vytlacil (2002) showed that the LATE model can be written as a joint model of potential outcomes and self-selection in which treatment is determined by a latent index crossing a threshold. Suppose treatment status is generated by the equation

$$D_i = 1\{\psi(Z_i) \geq V_i\},$$

where the latent variable V_i is independently and identically distributed according to some continuous distribution with cumulative distribution function $F_V(.) : \mathbb{R} \to [0,1]$, and $\psi(.) : \{0,1\} \to \mathbb{R}$ defines instrument-dependent thresholds below which treatment ensues. Typically $F_V(.)$ is treated as a structural primitive describing a stable distribution of latent costs and benefits influencing program participation that exists independently of a particular instrument, as in the classic selection models of Roy (1951) and Heckman (1974). We follow Heckman and Vytlacil (2005) and work with the equivalent transformed model

$$D_i = 1\{P(Z_i) \geq U_i\}, \tag{1}$$

where $U_i \equiv F_V(V_i)$ follows a uniform distribution and $P(Z_i) \equiv F_V(\psi(Z_i))$ is the propensity score. The instrument Z_i is presumed to increase the likelihood of treatment ($P(1) > P(0)$), and to be independent of U_i and potential outcomes:

$$(Y_{i1}, Y_{i0}, U_i) \perp\!\!\!\perp Z_i. \tag{2}$$

The selection model defined by (1) and (2) is equivalent to the treatment effects model described by assumptions IA.1-IA.3. Equation (1) merely translates the behavioral responses that are permitted in the LATE model into a partition of the unit interval. In the terminology of Angrist et al. (1996), assumption IA.3 implies that the population consists of compliers with $D_{i1} > D_{i0}$, "always takers" with $D_{i1} = D_{i0} = 1$, and "never takers" with $D_{i1} = D_{i0} = 0$. The latent variable U_i is defined such that always takers have $U_i \in [0, P(0)]$, compliers have $U_i \in (P(0), P(1)]$, and never takers have $U_i \in (P(1), 1]$. Condition (2) implies that potential outcomes and treatment choices are independent of the instrument and imposes no further restrictions on the joint distribution of these quantities. It follows that we can equivalently define $LATE = E[Y_{i1} - Y_{i0} | P(0) < U_i \leq P(1)]$.

Though Vytlacil's (2002) results establish equivalence between a non-parametric latent index model and the LATE model, the fully non-parametric model is typically not used for estimation. Rather, to motivate alternatives to IV estimation, it is conventional to make additional assumptions regarding the joint distribution of the latent cost U_i and the potential outcomes (Y_{i1}, Y_{i0}). The goal of this note is to investigate the consequences of such assumptions for empirical work.

3 Control function estimation

We begin by considering estimators predicated on the existence of a parametric "control function" capturing the endogeneity in the relationship between outcomes and treatment (Heckman and Robb, 1985; Blundell and

Matzkin, 2014; Wooldridge, 2015). The workhorse models in this literature obey the following semi-parametric restriction:

$$E\left[Y_{id}|U_i\right] = \alpha_d + \gamma_d \times \left(J(U_i) - \mu_J\right),\ d \in \{0,1\}, \tag{3}$$

where $J(\cdot) : [0,1] \to \mathbb{R}$ is a strictly increasing continuous function and $\mu_J \equiv E\left[J(U_i)\right]$. Lee (1982) studied this dependence structure in the context of classic "one-sided" selection problems where outcomes are only observed when $D_i = 1$. Setting $J(\cdot)$ equal to the inverse normal CDF yields the canonical Heckman (1976; 1979) sample selection ("Heckit") model, while choosing $J(u) = u$ yields the linear selection model studied by Olsen (1980), and choosing the inverse logistic CDF for $J\left(.\right)$ yields the logit selection model considered by Mroz (1987).

Subsequent work applies versions of (3) to policy evaluation by modeling program participation as a "two-sided" sample selection problem with coefficients indexed by the treatment state d. For example, Bjorklund and Moffitt (1987) build on the Heckit framework by assuming $J(\cdot)$ is the inverse normal CDF and allowing $\alpha_1 \neq \alpha_0$, $\gamma_1 \neq \gamma_0$. Likewise, the estimator of Brinch et al. (2017) is a two-sided variant of Olsen's (1980) approach that imposes an identity $J(\cdot)$ function with coefficients indexed by d. Interestingly, Dubin and McFadden's (1984) classic multinomial selection model collapses in the binary treatment effects case to a two-sided version of Mroz's (1987) logit model.

Assumption (3) nullifies Vytlacil's (2002) equivalence result by imposing restrictions on the relationships between mean potential outcomes of subgroups that respond differently to the instrument Z_i. Let μ_{dg} denote the mean of Y_{id} for group $g \in \{at, nt, c\}$, representing always takers, never takers and compliers. For any strictly increasing $J(\cdot)$, equation (3) implies $sgn(\mu_{dat} - \mu_{dc}) = sgn(\mu_{dc} - \mu_{dnt})$ for $d \in \{0,1\}$. In contrast, the nonparametric model defined by assumptions IA.1-IA.3 is compatible with any arrangement of differences in mean potential outcomes for the three subgroups. We next consider whether these additional restrictions are consequential for estimation of LATE.

LATE

Equation (3) implies that mean outcomes conditional on treatment status are

$$E\left[Y_i|Z_i, D_i = d\right] = \alpha_d + \gamma_d \lambda_d\left(P(Z_i)\right),$$

where $\lambda_1(p)$ and $\lambda_0(p)$ are control functions giving the means of $\left(J\left(U_i\right) - \mu_J\right)$ when U_i is truncated from above and below at p:

$$\lambda_1(p) = E\left[J(U_i) - \mu_J | U_i \leq p\right],\ \lambda_0(p) = E\left[J(U_i) - \mu_J | U_i > p\right].$$

While attention in parametric selection models often focuses on the population average treatment effect $\alpha_1 - \alpha_0$ (Garen, 1984; Heckman, 1990; Wooldridge, 2015), equation (3) can also be used to compute treatment effects for other subgroups. The average effect on compliers can be written

$$LATE = \alpha_1 - \alpha_0 + \left(\gamma_1 - \gamma_0\right)\Gamma\left(P(0), P(1)\right), \tag{4}$$

where $\Gamma(p, p')$ gives the mean of $J(U_i) - \mu_J$ truncated between p and p':

$$\Gamma(p, p') = E\left[J(U_i) - \mu_J | p < U_i \leq p'\right] = \frac{p'\lambda_1(p') - p\lambda_1(p)}{p' - p}.$$

The last term in (4) adjusts the average treatment effect to account for non-random selection into compliance with the instrument; this quantity is well-defined when IA.2 holds and non-compliance is "two-sided" so that $0 < P(0) < P(1) < 1$.

Estimation

To motivate control function estimation, suppose that the sample exhibits two-sided non-compliance as follows:

Assumption 2. $0 < \sum_i 1\{D_i = d\}Z_i < \sum_i 1\{D_i = d\}$ *for* $d \in \{0, 1\}$.

This condition requires at least one observation with every combination of Z_i and D_i. Assumption 2 is satisfied with probability approaching one at an exponential rate in n whenever $0 < Pr[Z_i = 1] < 1$ and $0 < P(z) < 1$ for $z \in \{0, 1\}$.

Control function estimation typically proceeds in two steps, both for computational reasons and because of the conceptual clarity of plug-in estimation strategies (Heckman, 1979; Smith and Blundell, 1986). Deferring a discussion of one-step estimation approaches to later sections, we define the control function estimator as a procedure which first fits the choice model in equation (1) by maximum likelihood, then builds estimates of $\lambda_1(\cdot)$ and $\lambda_0(\cdot)$ to include in second-step ordinary least squares (OLS) regressions for each treatment category. The first step estimates can be written

$$\left(\hat{P}(0), \hat{P}(1)\right) = \arg \max_{P(0),P(1)} \sum_i D_i \log P(Z_i) + \sum_i (1 - D_i) \log\left(1 - P(Z_i)\right). \tag{5}$$

The second step OLS estimates are

$$(\hat{\alpha}_d, \hat{\gamma}_d) = \arg \min_{\alpha_d, \gamma_d} \sum_i 1\{D_i = d\} \left[Y_i - \alpha_d - \gamma_d \lambda_d(\hat{P}(Z_i))\right]^2, \; d \in \{0, 1\}. \tag{6}$$

Under Assumption 2, $\hat{\alpha}_d$ and $\hat{\gamma}_d$ can be computed for each value of d. The analogy principle then suggests the following plug-in estimator of LATE:

$$\widehat{LATE}^{CF} = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\gamma}_1 - \hat{\gamma}_0)\, \Gamma(\hat{P}(0), \hat{P}(1)).$$

4 Equivalence results

Compared to $\widehat{LATE}^{IV}$, $\widehat{LATE}^{CF}$ would seem to be highly dependent upon the functional form assumed for $J(\cdot)$ and the linearity of equation (3). Our first result shows that this is not the case.

Theorem 1. *If Assumptions 1 and 2 hold then* $\widehat{LATE}^{CF} = \widehat{LATE}^{IV}$.

Proof: The maximum likelihood procedure in (5) yields the empirical treatment rates $\hat{P}(z) = \frac{\sum_i 1\{Z_i=z\}D_i}{\sum_i 1\{Z_i=z\}}$ for $z \in \{0,1\}$. The second-step OLS regressions can be rewritten

$$(\hat{\alpha}_d, \hat{\gamma}_d) = \arg\min_{\alpha_d,\gamma_d} \textstyle\sum_i 1\{D_i = d\} \left(Y_i - \left[\alpha_d + \gamma_d \lambda_d(\hat{P}(0))\right] - \gamma_d \left[\lambda_d(\hat{P}(1)) - \lambda_d(\hat{P}(0))\right] Z_i\right)^2.$$

This is a least squares fit of Y_i on an intercept and the indicator Z_i in the subsample with $D_i = d$. Such regressions can be estimated as long as there is two-sided non-compliance with the instrument Z_i, which follows from Assumption 2. Defining $\bar{Y}_d^z \equiv \frac{\sum_i 1\{D_i=d\}1\{Z_i=z\}Y_i}{\sum_i 1\{D_i=d\}1\{Z_i=z\}}$, we have

$$\bar{Y}_d^0 = \hat{\alpha}_d + \hat{\gamma}_d \lambda_d(\hat{P}(0)),\ \bar{Y}_d^1 - \bar{Y}_d^0 = \hat{\gamma}_d \left[\lambda_d(\hat{P}(1)) - \lambda_d(\hat{P}(0))\right].$$

Under Assumption 1, we have $\lambda_d(\hat{P}(1)) \neq \lambda_d(\hat{P}(0))$, and this pair of equations can be solved for $\hat{\gamma}_d$ and $\hat{\alpha}_d$ as

$$\hat{\gamma}_d = \tfrac{\bar{Y}_d^1 - \bar{Y}_d^0}{\lambda_d(\hat{P}(1)) - \lambda_d(\hat{P}(0))},\ \hat{\alpha}_d = \tfrac{\lambda_d(\hat{P}(1))\bar{Y}_d^0 - \lambda_d(\hat{P}(0))\bar{Y}_d^1}{\lambda_d(\hat{P}(1)) - \lambda_d(\hat{P}(0))}.$$

We can therefore rewrite the control function estimate of LATE as

$$\widehat{LATE}^{CF} = \left(\left[\tfrac{\lambda_1(\hat{P}(1))\bar{Y}_1^0 - \lambda_1(\hat{P}(0))\bar{Y}_1^1}{\lambda_1(\hat{P}(1)) - \lambda_1(\hat{P}(0))}\right] - \left[\tfrac{\lambda_0(\hat{P}(1))\bar{Y}_0^0 - \lambda_0(\hat{P}(0))\bar{Y}_0^1}{\lambda_0(\hat{P}(1)) - \lambda_0(\hat{P}(0))}\right]\right)$$
$$+ \left(\left[\tfrac{\bar{Y}_1^1 - \bar{Y}_1^0}{\lambda_1(\hat{P}(1)) - \lambda_1(\hat{P}(0))}\right] - \left[\tfrac{\bar{Y}_0^1 - \bar{Y}_0^0}{\lambda_0(\hat{P}(1)) - \lambda_0(\hat{P}(0))}\right]\right) \times \left(\tfrac{\hat{P}(1)\lambda_1(\hat{P}(1)) - \hat{P}(0)\lambda_1(\hat{P}(0))}{\hat{P}(1) - \hat{P}(0)}\right).$$

Using the fact that $\lambda_0(p) = -\lambda_1(p)p/(1-p)$, this simplifies to

$$\widehat{LATE}^{CF} = \frac{\left[\hat{P}(1)\bar{Y}_1^1 + (1 - \hat{P}(1))\bar{Y}_0^1\right] - \left[\hat{P}(0)\bar{Y}_1^0 + (1 - \hat{P}(0))\bar{Y}_0^0\right]}{\hat{P}(1) - \hat{P}(0)},$$

which is $\widehat{LATE}^{IV}$. ■

Remark 1. An immediate consequence of Theorem 1 is that $\widehat{LATE}^{CF}$ is also equivalent to the coefficient on D_i associated with a least squares fit of Y_i to D_i and a first stage residual $D_i - \hat{P}(Z_i)$. Blundell and Matzkin (2014) attribute the first proof of the equivalence between this estimator and IV to Telser (1964).

Remark 2. Theorem 1 extends the analysis of Brinch et al. (2017) who observe that an attractive feature of their linear control function estimator is that it produces a LATE estimate numerically equivalent to IV. In fact, a wide class of non-linear control function estimators yield LATE estimates numerically equivalent to IV. With a binary treatment and instrument, an instrumental variables estimate can always be viewed as the numerical output of a variety of parametric control function estimators.

Potential outcome means

Corresponding equivalence results hold for estimators of other parameters identified in the LATE framework. Imbens and Rubin (1997) and Abadie (2002) discuss identification and estimation of the treated outcome distribution for always takers, the untreated distribution for never takers, and both marginal distributions for compliers. Nonparametric estimators of the four identified marginal mean potential outcomes are given by

$$\hat{\mu}_{1at}^{IV} = \bar{Y}_1^0,\ \hat{\mu}_{0nt}^{IV} = \bar{Y}_0^1,$$

$$\hat{\mu}_{1c}^{IV} = \tfrac{\hat{P}(1)\bar{Y}_1^1 - \hat{P}(0)\bar{Y}_1^0}{\hat{P}(1)-\hat{P}(0)},\ \hat{\mu}_{0c}^{IV} = \tfrac{(1-\hat{P}(0))\bar{Y}_0^0 - (1-\hat{P}(1))\bar{Y}_0^1}{\hat{P}(1)-\hat{P}(0)}.$$

The corresponding control function estimators are:

$$\hat{\mu}_{1at}^{CF} = \hat{\alpha}_1 + \hat{\gamma}_1\lambda_1(\hat{P}(0)),\ \hat{\mu}_{0nt}^{CF} = \hat{\alpha}_0 + \hat{\gamma}_0\lambda_0(\hat{P}(1)),$$

$$\hat{\mu}_{dc}^{CF} = \hat{\alpha}_d + \hat{\gamma}_d\Gamma(\hat{P}(0), \hat{P}(1)),\ d \in \{0,1\}.$$

The following proposition shows that these two estimation strategies produce algebraically identical results.

Proposition 1. *If Assumptions 1 and 2 hold then*

$$\hat{\mu}_{dg}^{CF} = \hat{\mu}_{dg}^{IV} \text{ for } (d,g) \in \{(1,at),(0,nt),(1,c),(0,c)\}.$$

Proof: Using the formulas from the proof of Theorem 1, the control function estimate of μ_{1at} is

$$\hat{\mu}_{1at}^{CF} = \left(\tfrac{\lambda_1(\hat{P}(1))\bar{Y}_1^0 - \lambda_1(\hat{P}(0))\bar{Y}_1^1}{\lambda_1(\hat{P}(1))-\lambda_1(\hat{P}(0))}\right) + \left(\tfrac{\bar{Y}_1^1-\bar{Y}_1^0}{\lambda_1(\hat{P}(1))-\lambda_1(\hat{P}(0))}\right)\lambda_1(\hat{P}(0)) = \bar{Y}_1^0,$$

which is $\hat{\mu}_{1at}^{IV}$. Likewise,

$$\hat{\mu}_{0nt}^{CF} = \left(\tfrac{\lambda_0(\hat{P}(1))\bar{Y}_0^0 - \lambda_0(\hat{P}(0))\bar{Y}_1^1}{\lambda_0(\hat{P}(1))-\lambda_0(\hat{P}(0))}\right) + \left(\tfrac{\bar{Y}_0^1-\bar{Y}_0^0}{\lambda_0(\hat{P}(1))-\lambda_0(\hat{P}(0))}\right)\lambda_0(\hat{P}(1)) = \bar{Y}_0^1\ ,$$

which is $\hat{\mu}_{0nt}^{IV}$. The treated complier mean estimate is

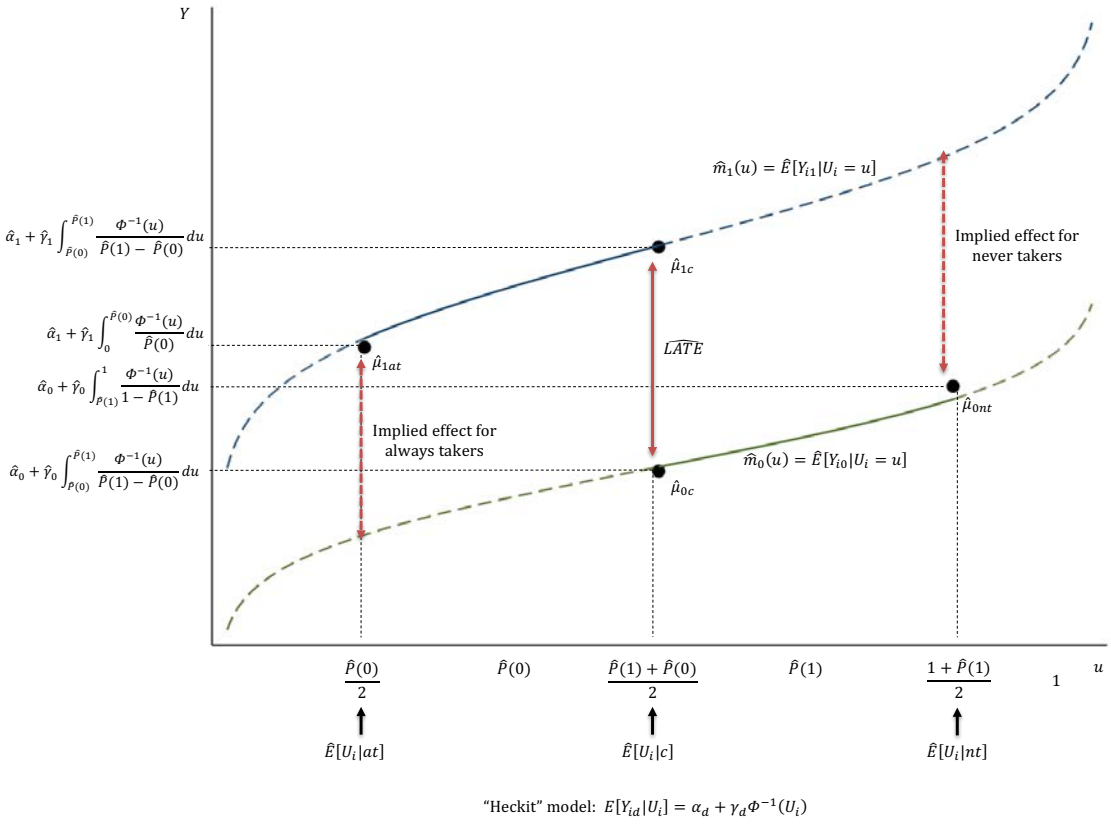
$$\begin{aligned}\hat{\mu}_{1c}^{CF} &= \left(\tfrac{\lambda_1(\hat{P}(1))\bar{Y}_1^0 - \lambda_1(\hat{P}(0))\bar{Y}_1^1}{\lambda_1(\hat{P}(1))-\lambda_1(\hat{P}(0))}\right) + \left(\tfrac{\bar{Y}_1^1-\bar{Y}_1^0}{\lambda_1(\hat{P}(1))-\lambda_1(\hat{P}(0))}\right) \times \left(\tfrac{\hat{P}(1)\lambda_1(\hat{P}(1)) - \hat{P}(0)\lambda_1(\hat{P}(0))}{\hat{P}(1)-\hat{P}(0)}\right) \\ &= \tfrac{\left(\lambda_1(\hat{P}(1))-\lambda_1(\hat{P}(0))\right)\hat{P}(1)\bar{Y}_1^1 - \left(\lambda_1(\hat{P}(1))-\lambda_1(\hat{P}(0))\right)\hat{P}(0)\bar{Y}_1^0}{\left(\lambda_1(\hat{P}(1))-\lambda_1(\hat{P}(0))\right)\left(\hat{P}(1)-\hat{P}(0)\right)} = \tfrac{\hat{P}(1)\bar{Y}_1^1 - \hat{P}(0)\bar{Y}_1^0}{\hat{P}(1)-\hat{P}(0)},\end{aligned}$$

which is $\hat{\mu}_{1c}^{IV}$. Noting that $\widehat{LATE}^{IV} = \hat{\mu}_{1c}^{IV} - \hat{\mu}_{0c}^{IV}$ and $\widehat{LATE}^{CF} = \hat{\mu}_{1c}^{CF} - \hat{\mu}_{0c}^{CF}$, it then follows by Theorem 1 that $\hat{\mu}_{0c}^{CF} = \hat{\mu}_{0c}^{IV}$. ■

5 Equivalence and extrapolation

Proposition 1 establishes that all control function estimators based on equation (3) produce identical estimates of the potential outcome means that are nonparametrically identified in the LATE framework. Different functional form assumptions generate different estimates of quantities that are under-identified, however. For example, the choice of $J(\cdot)$ in equation (3) determines the shapes of the curves that the model uses to extrapolate from estimates of the four identified average potential outcome means $(\mu_{1at}, \mu_{0nt}, \mu_{1c}, \mu_{0c})$ to the two under-identified potential outcome means (μ_{0at}, μ_{1nt}).

Figure 1: "Heckit" extrapolation



Figures 1 and 2 illustrate this extrapolation in a hypothetical example. The horizontal axis plots values u of the unobserved treatment cost U_i, while the vertical axis plots mean potential outcomes $m_d(u) = E[Y_{id}|U_i = u]$ as functions of this cost. Estimates of these functions are denoted $\hat{m}_d(u) = \hat{\alpha}_d + \hat{\gamma}_d \times (J(u) - \mu_J)$ and their difference $\hat{m}_1(u) - \hat{m}_0(u)$ provides an estimate of the marginal treatment effect (Bjorklund and Moffitt, 1987; Heckman and Vytlacil, 2005; Heckman et al., 2006) for an individual with latent cost u.

Assumptions IA.1-IA3 ensure two averages of $m_d(U_i)$ are identified for each potential outcome: the treated means for always takers and compliers, and the untreated means for never takers and compliers. The control function estimator chooses $\hat{\alpha}_d$ and $\hat{\gamma}_d$ so that averages of $\hat{m}_d(U_i)$ over the relevant ranges match the corresponding nonparametric estimates for each compliance group. The coefficient $\hat{\gamma}_1$ parameterizes the difference in mean treated outcomes between compliers and always takers, while $\hat{\gamma}_0$ measures the difference in mean untreated outcomes between compliers and never takers. Several tests of endogeneous treatment assignment (see, e.g., Angrist, 2004; Battistin and Rettore, 2008; Bertanha and Imbens, 2014; and Kowalski, 2016) amount to testing whether $(\hat{\gamma}_0, \hat{\gamma}_1)$ are significantly different from zero.

Figure 2: Linear extrapolation

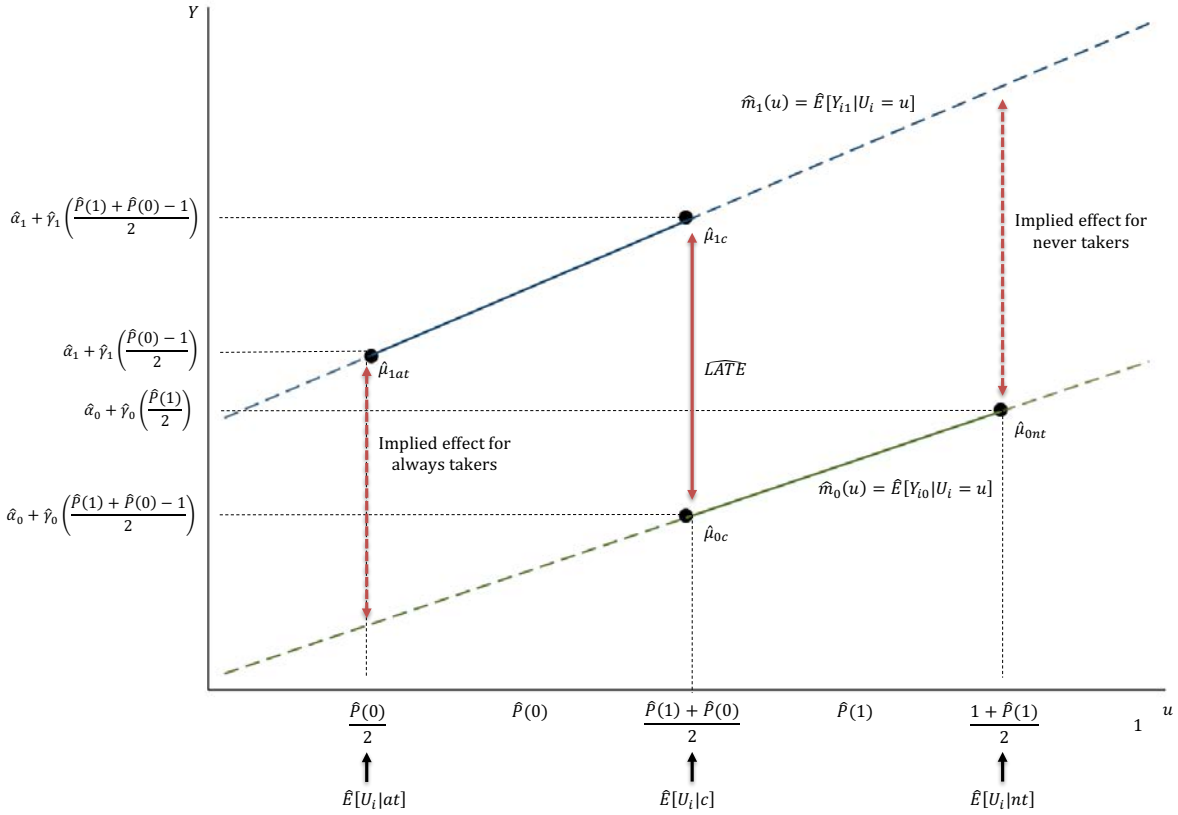


Figure 1 depicts the results of parametric extrapolation based on the Heckit model, while Figure 2 shows results for the linear control function model discussed by Brinch et al. (2017). Both models match the same four estimated mean potential outcomes, thereby generating identical estimates of LATE. Note that by Jensen's inequality, the nonlinear $\hat{m}_d(u)$ curves in Figure 1 do not pass directly through the group mean potential outcomes. The two models yield different imputations for the missing potential outcomes of always takers and never takers, and therefore also different estimates of the ATE, which averages over all three subpopulations. This sensitivity to functional form is intuitive: treatment effects for always and never takers are fundamentally under-identified, an insight that has led to consideration of worst case bounds on these quantities (Manski, 1990; Balke and Pearl, 1997; Mogstad et al., 2016).

6 Multi-valued instruments

Consider an instrument Z_i taking values in $\{0, 1, .., K\}$, and suppose that $0 < \hat{P}(z-1) < \hat{P}(z) < 1$ for $z \in \{1, 2, ..., K\}$. Let D_{iz} denote i's treatment choice when $Z_i = z$. If assumptions IA.1-IA.3 hold for every pair of instrument values, Wald ratios of the form $\frac{E[Y_i|Z_i=z]-E[Y_i|Z_i=z-1]}{E[D_i|Z_i=z]-E[D_i|Z_i=z-1]}$ identify the average treatment effect among

compliers indexed by a unit increment in the instrument, which we denote $LATE_z \equiv E\left[Y_{i1} - Y_{i0} | D_{iz} > D_{iz-1}\right]$. Analog estimators of $LATE_z$ are given by the following pairwise IV estimator:

$$\widehat{LATE}_z^{IV} = \frac{\frac{1}{\sum_i 1\{Z_i=z\}} \sum_i 1\{Z_i = z\} Y_i - \frac{1}{\sum_i 1\{Z_i=z-1\}} \sum_i 1\{Z_i = z-1\} Y_i}{\frac{1}{\sum_i 1\{Z_i=z\}} \sum_i 1\{Z_i = z\} D_i - \frac{1}{\sum_i 1\{Z_i=z-1\}} \sum_i 1\{Z_i = z-1\} D_i}.$$

From Theorem 1, $\widehat{LATE}_z^{IV}$ is numerically equivalent to the corresponding pairwise control function estimator of $LATE_z$. However, to improve precision, it is common to impose additional restrictions on the $LATE_z$.

Consider the following restriction on potential outcomes:

$$E\left[Y_{id} | U_i\right] = \alpha_d + \sum_{\ell=1}^{L} \gamma_{d\ell} \times \left(J(U_i) - \mu_J\right)^\ell, \; d \in \{0, 1\}. \tag{7}$$

Polynomial models of this sort were considered by Brinch et al. (2017) with $J(.)$ restricted to be the identity function. Letting $\lambda_{1\ell}(p) = E\left[(J(U_i) - \mu_J)^\ell | U_i \leq p\right]$ and $\lambda_{0\ell}(p) = E\left[(J(U_i) - \mu_J)^\ell | U_i > p\right]$, a two-step control function estimator of the parameters of equation (7) is

$$(\hat{\alpha}_d, \hat{\gamma}_{d1}, ..., \hat{\gamma}_{dL}) = \arg \min_{\alpha_d, \gamma_{d1}, ..., \gamma_{dL}} \sum_i 1\left\{D_i = d\right\} \left[Y_i - \alpha_d - \textstyle\sum_{\ell=1}^{L} \gamma_{d\ell} \lambda_{d\ell}(\hat{P}(Z_i))\right]^2.$$

The resulting control function estimator of $LATE_z$ is then

$$\widehat{LATE}_z^{CF} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \sum_{\ell=1}^{L} (\hat{\gamma}_{1\ell} - \hat{\gamma}_{0\ell}) \Gamma_\ell(\hat{P}(z-1), \hat{P}(z)), \tag{8}$$

where $\Gamma_\ell(p, p') = [p' \lambda_{1\ell}(p') - p \lambda_{1\ell}(p)]/[p' - p]$. The following proposition establishes that this estimator is identical to $\widehat{LATE}_z^{IV}$ when $L = K$.

Proposition 2. *If Assumptions 1 and 2 hold for every pair of instrument values and the polynomial order L equals K then $\widehat{LATE}_z^{CF} = \widehat{LATE}_z^{IV} \; \forall z \in \{1, 2, ..., K\}$.*

Proof: See the Appendix. ■

Remark 3. Instrumenting D_i with a scalar function $g(Z_i)$ generates an IV estimate equal to a convex weighted average of the $\widehat{LATE}_z^{IV}$ (Imbens and Angrist, 1994). From Proposition 2, applying these weights to the $\widehat{LATE}_z^{CF}$ when $L = K$ will yield an identical result. By contrast, the set of $\widehat{LATE}_z^{CF}$ that result from imposing $L < K$ need not correspond to weighted averages of the $\widehat{LATE}_z^{IV}$, but are likely to exhibit reduced sampling variability.

Remark 4. When $L < K - 1$, the restriction in (7) can be used to motivate estimators of particular LATEs that are convex combinations of IV estimators. In the case where $K = 3$ and $L = 1$, one can show that:

$$LATE_2 = \left(\tfrac{\Gamma(P(1),P(2)) - \Gamma(P(0),P(1))}{\Gamma(P(2),P(3)) - \Gamma(P(0),P(1))}\right) LATE_3 + \left(\tfrac{\Gamma(P(2),P(3)) - \Gamma(P(1),P(2))}{\Gamma(P(2),P(3)) - \Gamma(P(0),P(1))}\right) LATE_1.$$

This representation suggests combination estimators of the form

$$\widehat{LATE}_2^{\xi} = \xi \widehat{LATE}_2^{IV} + (1-\xi)\left[\left(\frac{\Gamma(\hat{P}(1),\hat{P}(2))-\Gamma(\hat{P}(0),\hat{P}(1))}{\Gamma(\hat{P}(2),\hat{P}(3))-\Gamma(\hat{P}(0),\hat{P}(1))}\right)\widehat{LATE}_3^{IV} + \left(\frac{\Gamma(\hat{P}(2),\hat{P}(3))-\Gamma(\hat{P}(1),\hat{P}(2))}{\Gamma(\hat{P}(2),\hat{P}(3))-\Gamma(\hat{P}(0),\hat{P}(1))}\right)\widehat{LATE}_1^{IV}\right],$$

for $\xi \in (0,1)$. To maximize precision, one can set $\xi = [\hat{v}_2 - \hat{v}_{12}]/[\hat{v}_1 + \hat{v}_2 - 2\hat{v}_{12}]$, where $\hat{v}_1$ and $\hat{v}_2$ are estimated variances of $\widehat{LATE}_2^{IV}$ and the term in brackets, respectively, and $\hat{v}_{12}$ is their covariance. By construction, $\widehat{LATE}_2^{\xi}$ provides an estimate of $LATE_2$ more precise than $\widehat{LATE}_2^{IV}$. Though $\widehat{LATE}_2^{\xi}$ will tend to be less precise than $\widehat{LATE}_2^{CF}$ when restriction (7) is true, the probability limit of $\widehat{LATE}_2^{\xi}$ retains an interpretation as a weighted average of causal effects for complier subpopulations when (7) is violated, a robustness property emphasized elsewhere by Angrist and Pischke (2009).

7 Equivalence failures

Though Theorem 1 establishes equivalence between IV and a wide class of control function estimates of LATE, other control function estimators fail to match IV even with a single binary instrument. LaLonde (1986) considered OLS estimation of the following model:

$$Y_i = \alpha + \beta D_i + \gamma\left[D_i \times \left(-\frac{\phi(\Phi^{-1}(\hat{P}(Z_i)))}{\hat{P}(Z_i)}\right) + (1-D_i)\times\left(\frac{\phi(\Phi^{-1}(\hat{P}(Z_i)))}{1-\hat{P}(Z_i)}\right)\right] + \epsilon_i. \tag{9}$$

By imposing a common coefficient γ on the Mills ratio terms for the treatment and control groups, this specification allows for selection on levels but rules out selection on treatment effects.

The term in brackets in equation (9) simplifies to $(D_i - \hat{P}(Z_i)) \times \{-\phi(\Phi^{-1}(\hat{P}(Z_i)))/[\hat{P}(Z_i)(1-\hat{P}(Z_i))]\}$. When $\hat{P}(1) = 1 - \hat{P}(0)$ this term is proportional to the first stage residual and least squares estimation of (9) yields an estimate of β numerically identical to IV. This is a finite sample analogue of Heckman and Vytlacil's (2000) observation (elaborated upon in Angrist, 2004) that LATE equals ATE when both the first stage and the error distribution are symmetric. When $\hat{P}(1) \neq 1 - \hat{P}(0)$, however, the control function in equation (9) differs from the first stage residual and the estimate of β will not match IV.

Remark 5. When $\hat{P}(1) = 1 - \hat{P}(0)$, the ATE estimate $\hat{\alpha}_1 - \hat{\alpha}_0$ from a control function estimator of the form given in (6) coincides with IV whenever $J(U_i)$ is presumed to follow a symmetric distribution.

Moments and monotonicity

Theorem 1 relied upon the fact that equation (3) includes enough free parameters to allow the control function estimator to match the sample mean of Y_i for every combination of D_i and Z_i. One might be tempted to conclude that any structural estimator that fits these moments will produce a corresponding LATE estimate equal to IV. We now show that this is not the case.

Suppose that treatment status is generated by a heterogeneous threshold crossing model:

$$D_i = 1\{\kappa + \delta_i Z_i \geq U_i\}, \tag{10}$$

where U_i is uniformly distributed and the random coefficient δ_i is a mixture taking values in $\{-\eta, \eta\}$ for some known positive constant η. Define $\upsilon \equiv Pr[\delta_i = \eta]$, and suppose that δ_i is independent of $(Y_{i1}, Y_{i0}, U_i, Z_i)$. Note that this model does not admit a representation of the form of equation (1) as it allows $D_{i1} < D_{i0}$.

Model (10) has two unknown parameters, κ and υ, and can therefore rationalize the two observed choice probabilities by choosing $\hat{\kappa} = \hat{P}(0)$ and $\hat{\upsilon} = (\eta + \hat{P}(1) - \hat{P}(0))/2\eta$. Equations (3) and (10) imply

$$E[Y_i|D_i = d, Z_i] = \alpha_d + \gamma_d \times [\upsilon\lambda_d(\kappa + \eta Z_i) + (1 - \upsilon)\lambda_d(\kappa - \eta Z_i)].$$

As before, we can use $\hat{\kappa}$ and $\hat{\upsilon}$ to construct control functions to include in a second-step regression, producing estimates $\hat{\alpha}_d$ and $\hat{\gamma}_d$ that exactly fit $\bar{Y}_d^1$ and $\bar{Y}_d^0$.

Though this estimator matches all choice probabilities and conditional mean outcomes, it produces an estimate of LATE different from IV. The model's implied LATE is

$$E[Y_{i1} - Y_{i0}|D_{i1} > D_{i0}] = (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \times E[J(U_i) - \mu_J|\delta_i = \eta, \kappa < U_i \leq \kappa + \eta].$$

The corresponding control function estimator of this quantity is

$$\widehat{LATE}^* = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\gamma}_1 - \hat{\gamma}_0) \times \left(\frac{(\hat{\kappa}+\eta)\lambda_1(\hat{\kappa}+\eta) - \hat{\kappa}\lambda_1(\hat{\kappa})}{\eta}\right). \tag{11}$$

It is straightforward to verify that $\widehat{LATE}^*$ is not equal to $\widehat{LATE}^{IV}$. Equivalence fails here because the selection model implies the presence of "defiers" with $D_{i1} < D_{i0}$. IV does not identify LATE when there are defiers; hence, the model suggests using a different function of the data to estimate the LATE.

Covariates

It is common to condition on a vector of covariates X_i either to account for possible violations of the exclusion restriction or to increase precision. Theorem 1 implies that IV and control function estimates of LATE coincide if computed separately for each value of the covariates, but this may be impractical or impossible when X_i can take on many values.

A standard approach to introducing covariates is to enter them additively into the potential outcomes model (see, e.g., Cornelissen et al., 2016; Kline and Walters, 2016; and Brinch et al., 2017). Suppose treatment choice is given by $D_i = 1\{P(X_i, Z_i) \geq U_i\}$ with U_i independent of (X_i, Z_i), and assume

$$E[Y_{id}|U_i, X_i] = \alpha_d + \gamma_d \times (J(U_i) - \mu_J) + X_i'\tau, \ d \in \{0, 1\}. \tag{12}$$

Letting $\hat{P}(X_i, Z_i)$ denote an estimate of $Pr[D_i = 1|X_i, Z_i]$, the control function estimates for this model are

$$(\hat{\alpha}_1, \hat{\gamma}_1, \hat{\alpha}_0, \hat{\gamma}_0, \hat{\tau}) = \arg\min_{\alpha_1, \gamma_1, \alpha_0, \gamma_0, \tau} \sum_i \sum_{d \in \{0,1\}} 1\{D_i = d\} \left[Y_i - \alpha_d - \gamma_d\lambda_d(\hat{P}(X_i, Z_i)) - X_i'\tau\right]^2. \tag{13}$$

To ease exposition, we will study the special case of a single binary covariate $X_i \in \{0, 1\}$. Define $LATE(x) \equiv E[Y_{i1} - Y_{i0}|P(x, 0) < U_i \leq P(x, 1), X_i = x]$ as the average treatment effect for compliers with $X_i = x$, and let

$\hat{\alpha}_d(x)$ and $\hat{\gamma}_d(x)$ denote estimates from unrestricted control function estimation among the observations with $X_i = x$. The additive separability restriction in (12) suggests the following two estimators of $LATE(1)$:

$$\widehat{LATE}_x^{CF}(1) = (\hat{\alpha}_1(x) - \hat{\alpha}_0(x)) + (\hat{\gamma}_1(x) - \hat{\gamma}_0(x))\,\Gamma(\hat{P}(1,0), \hat{P}(1,1)),\ x \in \{0,1\}.$$

By Theorem 1 $\widehat{LATE}_1^{CF}(1)$ is a Wald estimate for the $X_i = 1$ sample. $\widehat{LATE}_0^{CF}(1)$ gives an estimated effect for compliers with $X_i = 1$ based upon control function estimates for observations with $X_i = 0$. The following proposition describes the relationship between these two estimators and the restricted estimator of $LATE\,(1)$ based upon (13).

Proposition 3. *Suppose Assumptions 1 and 2 hold for each value of* $X_i \in \{0,1\}$ *and let* $\widehat{LATE}_r^{CF}(1) = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\gamma}_1 - \hat{\gamma}_0)\Gamma(\hat{P}(1,0), \hat{P}(1,1))$ *denote an estimate of* $LATE(1)$ *based on (13). Then*

$$\widehat{LATE}_r^{CF}(1) = w\widehat{LATE}_1^{CF}(1) + (1-w)\widehat{LATE}_0^{CF}(1) + b_1\,(\hat{\gamma}_1(1) - \hat{\gamma}_1(0)) + b_0\,(\hat{\gamma}_0(1) - \hat{\gamma}_0(0)).$$

The coefficients w, b_1, *and* b_0 *depend only on the joint empirical distribution of* D_i, X_i, *and* $\hat{P}(X_i, Z_i)$.

Proof: See the Appendix. ■

Remark 6. Proposition 3 demonstrates that control function estimation under additive separability gives a linear combination of covariate-specific estimates plus terms that equal zero when the separability restrictions hold exactly in the sample. One can show that the coefficient w need not lie between 0 and 1. By contrast, two-stage least squares estimation of a linear model with an additive binary covariate using all interactions of X_i and Z_i as instruments generates a weighted average of covariate-specific IV estimates (Angrist and Pischke, 2009).

Remark 7. Consider the following extension of equation (12):

$$E\,[Y_{id}|U_i, X_i] = \alpha_d + \gamma_d \times (J(U_i) - \mu_J) + X_i'\tau_{dc} + 1\{U_i \leq P(X_i, 0)\}X_i'\tau_{at} + 1\{U_i > P(X_i, 1)\}X_i'\tau_{nt},\ d \in \{0,1\}.$$

This equation allows different coefficients on X_i for always takers, never takers, and compliers by interacting X_i with indicators for thresholds of U_i, and also allows the complier coefficients to differ for treated and untreated outcomes. When X_i includes a mutually exclusive and exhaustive set of indicator variables and $\hat{P}(X_i, Z_i)$ equals the sample mean of D_i for each (X_i, Z_i), control function estimation of this model produces the same estimate of $E[Y_i|X_i, D_i, D_{i1} > D_{i0}]$ as the semi-parametric procedure of Abadie (2003). Otherwise the estimates may differ even asymptotically as the control function estimator employs a different set of approximation weights when the model is misspecified.

Remark 8. A convenient means of adjusting for covariates that maintains the numerical equivalence of IV and control function estimates is to weight each observation by $\omega_i = Z_i/\hat{e}(X_i) + (1 - Z_i)/(1 - \hat{e}(X_i))$ where $\hat{e}(x) \in (0,1)$ is a first step estimate of $Pr\,[Z_i = 1|X_i = x]$. It is straightforward to show that the ω_i-weighted IV and control function estimates of the unconditional LATE will be identical, regardless of the propensity score estimator $\hat{e}(X_i)$ employed. See Hull (2016) for a recent application of this approach to covariate adjustment of a selection model.

8 Maximum likelihood

A fully parametric alternative to two-step control function estimation is to specify a joint distribution for the model's unobservables and estimate the parameters in one step via full information maximum likelihood (FIML). Consider a model that combines (1) and (2) with the distributional assumption

$$Y_{id}|U_i \sim F_{Y|U}\left(y|U_i;\theta_d\right), \tag{14}$$

where $F_{Y|U}(y|u;\theta)$ is a conditional CDF indexed by a finite dimensional parameter vector θ. For example, a fully parametric version of the Heckit model is $Y_{id}|U_i \sim N\left(\alpha_d + \gamma_d \Phi^{-1}(U_i), \sigma_d^2\right)$. Since the marginal distribution of U_i is also known, this model provides a complete description of the joint distribution of (Y_{id}, U_i). FIML exploits this distributional knowledge, estimating the model's parameters as

$$\begin{aligned}\left(\hat{P}(0)^{ML}, \hat{P}(1)^{ML}, \hat{\theta}_0^{ML}, \hat{\theta}_1^{ML}\right) = \arg \max_{(P(0),P(1),\theta_0,\theta_1)} & \sum_i D_i \log\left(\int_0^{P(Z_i)} f_{Y|U}\left(Y_i|u;\theta_1\right) du\right) \\ & + \sum_i (1-D_i)\log\left(\int_{P(Z_i)}^1 f_{Y|U}(Y_i|u;\theta_0) du\right),\end{aligned} \tag{15}$$

where $f_{Y|U}(\cdot|u;\theta_d) \equiv dF_{Y|U}\left(.|u;\theta_d\right)$ denotes the density (or probability mass function) of Y_{id} given $U_i = u$. The corresponding FIML estimates of treated and untreated complier means are

$$\hat{\mu}_{dc}^{ML} = \frac{\int_{\hat{P}(0)^{ML}}^{\hat{P}(1)^{ML}} \int_{-\infty}^{\infty} y f_{Y|U}(y|u;\theta_d) dy du}{\hat{P}(1)^{ML} - \hat{P}(0)^{ML}},$$

and the FIML estimate of LATE is $\widehat{LATE}^{ML} = \hat{\mu}_{1c}^{ML} - \hat{\mu}_{0c}^{ML}$.

Binary outcomes

We illustrate the relationship between FIML and IV estimates of LATE with the special case of a binary Y_i. A parametric model for this setting is given by

$$\begin{aligned} Y_{id} =& 1\left\{\alpha_d \geq \epsilon_{id}\right\}, \\ \epsilon_{id}|U_i \sim& F_{\epsilon|U}\left(\epsilon|U_i;\rho_d\right), \end{aligned} \tag{16}$$

where $F_{\epsilon|U}(\epsilon|u;\rho)$ is a conditional CDF characterized by the single parameter ρ. Equations (1) and (16) include six parameters, which matches the number of observed linearly independent probabilities (two values of $Pr\left[D_i = 1|Z_i\right]$, and four values of $Pr\left[Y_i = 1|D_i, Z_i\right]$). The model is therefore "saturated" in the sense that a model with more parameters would be under-identified.

The following result establishes the conditions under which maximum likelihood estimates of complier means (and therefore LATE) coincide with IV.

Proposition 4. *Consider the model defined by (1), (2) and (16). Suppose that Assumptions 1 and 2 hold, and that the maximum likelihood problem (15) has a unique solution. Then* $\hat{\mu}_{dc}^{ML} = \hat{\mu}_{dc}^{IV}$ *for* $d \in \{0,1\}$ *if and only if* $\hat{\mu}_{dc}^{IV} \in [0,1]$ *for* $d \in \{0,1\}$.

Proof: See the Appendix. ■

Remark 9. The intuition for Proposition 4 is that the maximum likelihood estimation problem can be rewritten in terms of the six identified parameters of the LATE model: $(\mu_{1at}, \mu_{0nt}, \mu_{1c}, \mu_{0c}, \pi_{at}, \pi_c)$, where π_g is the population share of group g. Unlike the IV and control function estimators, the FIML estimator accounts for the binary nature of Y_{id} by constraining all probabilities to lie in the unit interval. When these constraints do not bind the FIML estimates coincide with nonparametric IV estimates, but the estimates differ when the nonparametric approach produces complier mean potential outcomes outside the logically possible bounds. Logical violations of this sort have been proposed elsewhere as a sign of failure of instrument validity (Balke and Pearl, 1997; Imbens and Rubin, 1997; Huber and Mellace, 2015; Kitagawa, 2015).

Remark 10. A simple "limited information" approach to maximum likelihood estimation is to estimate $P(0)$ and $P(1)$ in a first step and then maximize the plug-in conditional log-likelihood function

$$\sum_i D_i \log\left(\int_0^{\hat{P}(Z_i)} f_{Y|U}\left(Y_i|u;\theta_1\right) du\right) + \sum_i (1-D_i) \log\left(\int_{\hat{P}(Z_i)}^1 f_{Y|U}(Y_i|u;\theta_0) du\right)$$

with respect to (θ_0, θ_1) in a second stage. One can show that applying this less efficient estimator to a saturated model will produce an estimate of LATE equivalent to IV under Assumptions 1 and 2. This broader domain of equivalence results from some cross-equation parameter restrictions being ignored by the two-step procedure. For example, the FIML estimator may choose an estimate of π_c other than $\hat{P}(1) - \hat{P}(0)$ in order to enforce the constraint that $(\mu_{1c}, \mu_{0c}) \in [0,1]^2$.

Overidentified models

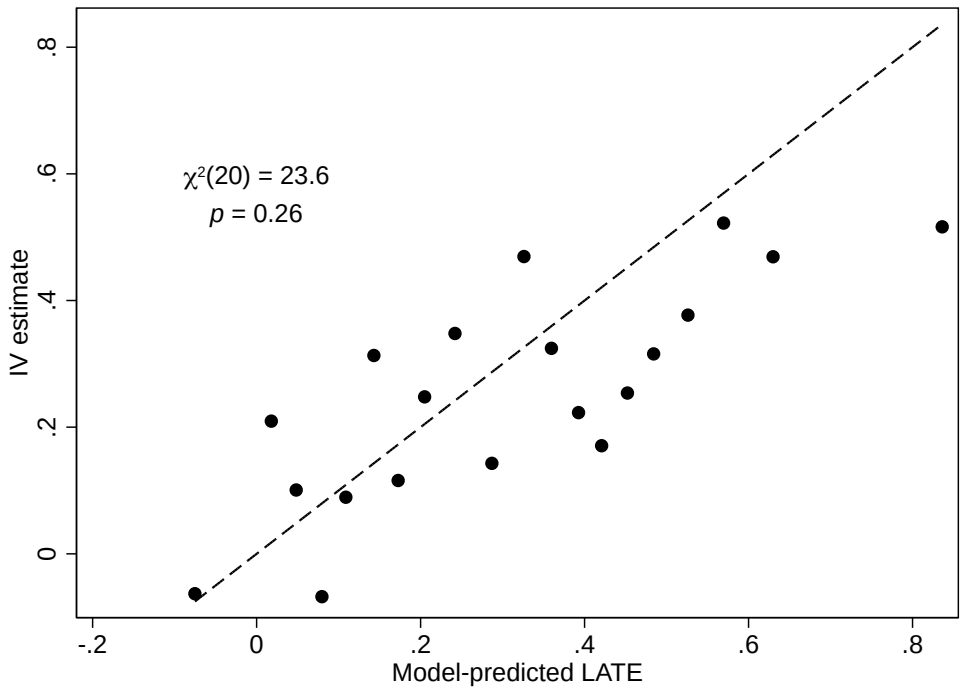
Equivalence of FIML and IV estimates at interior solutions in our binary example follows from the fact that the model satisfies monotonicity and includes enough parameters to match all observed choice probabilities. Similar arguments apply to FIML estimators of sufficiently flexible models for multi-valued outcomes. When the model includes fewer parameters than observed choice probabilities, overidentification ensues. For example, the standard bivariate probit model is a special case of (16) that uses a normal distribution for $F_{\epsilon|U}(\cdot)$ and imposes $\epsilon_{i1} = \epsilon_{i0}$ and therefore $\rho_1 = \rho_0$ (see Greene, 2007). Hence, only five parameters are available to rationalize six linearly independent probabilities.

Maximum likelihood estimation of this more parsimonious model may yield an estimate of LATE that differs from IV even at interior solutions. This divergence stems from the model's overidentifying restrictions which, if correct, may yield efficiency gains but if wrong can compromise consistency. Though maximum likelihood estimation of misspecified models yields a global best approximation to the choice probabilities (White, 1982), there is no guarantee that it will deliver a particularly good approximation to the LATE.

9 Model evaluation

In practice researchers often estimate selection models that impose additive separability assumptions on exogenous covariates, combine multiple instruments, and employ additional smoothness restrictions that break the algebraic equivalence of structural LATE estimates with IV. The equivalence results developed above provide a useful conceptual benchmark for assessing the performance of structural models in such applications. An estimator derived from a properly specified model of treatment assignment and potential outcomes should come close to matching a nonparametric IV estimate of the same parameter. Significant divergence between these estimates would signal that the restrictions imposed by the structural model are violated.

Figure 3: Model-based and IV estimates of LATE



Notes: This figure reproduces Figure A.III from Kline and Walters (2016). The figure is constructed by splitting the Head Start Impact Study sample into vingtiles of the predicted LATE based on the control function estimates reported in Section VIII of the paper. The horizontal axis displays the average predicted LATE in each group, and the vertical axis shows corresponding IV estimates. The dashed line is the 45-degree line. The chi-squared statistic and p-value come from a bootstrap Wald test of the hypothesis that the 45 degree line fits all points up to sampling error. See Appendix F of Kline and Walters (2016) for more details.

Figure 3 shows an example of this approach to model assessment from Kline and Walters' (2016) reanalysis of the Head Start Impact Study (HSIS) – a randomized experiment with two-sided non-compliance (Puma et al., 2012). On the vertical axis are non-parametric IV estimates of the LATE associated with participating in the Head Start program relative to a next best alternative for various subgroups in the HSIS defined by experimental sites and baseline child and parent characteristics. On the horizontal axis are two-step control function estimates of the same parameters derived from a heavily over-identified selection model involving

multiple endogenous variables, baseline covariates, and excluded instruments. Had this model been saturated, all of the points would lie on the 45 degree line. In fact, a Wald test indicates these deviations from the 45 degree line cannot be distinguished from noise at conventional significance levels, suggesting that the approximating model is not too far from the truth.

Passing a specification test does not obviate the fundamental identification issues inherent in interpolation and extrapolation exercises. As philosophers of science have long argued, however, models that survive empirical scrutiny deserve greater consideration then those that do not (Popper, 1959; Lakatos, 1976). Demonstrating that a tightly restricted model yields a good fit to IV estimates not only bolsters the credibility of the model's counterfactual predictions, but serves to clarify what the estimated structural parameters have to say about the effects of a research design as implemented. Here the control function estimates reveal that Head Start had very different effects on different sorts of complying households, a finding rationalized by estimated heterogeneity in both patterns of selection into treatment and potential outcome distributions.

10 Conclusion

This paper shows that two-step control function estimators of LATE derived from a wide class of parametric selection models coincide with the instrumental variables estimator. Control function and IV estimates of mean potential outcomes for compliers, always takers, and never takers are also equivalent. While many parametric estimators produce the same estimate of LATE, different parameterizations can produce dramatically different estimates of population average treatment effects and other under-identified quantities. The sensitivity of average treatment effect estimates to the choice of functional form may be the source of the folk wisdom that structural estimators are less robust than instrumental variables estimators. Our results show that this view confuses robustness for a given target parameter with the choice of target parameter.

Structural estimators that impose overidentifying restrictions may generate LATE estimates different from IV. Reporting the LATEs implied by such estimators facilitates comparisons with unrestricted IV estimates and is analogous to the standard practice of reporting average marginal effects in binary choice models (Wooldridge, 2001). Such comparisons provide a convenient tool for assessing the behavioral restrictions imposed by structural models. Model-based estimators that cannot rationalize unrestricted IV estimates of LATE are unlikely to fare much better at extrapolating to fundamentally under-identified quantities. On the other hand, a tightly constrained structural estimator that fits a collection of disparate IV estimates enjoys some degree of validation that bolsters the credibility of its counterfactual predictions.

References

Abadie, A. (2002): "Bootstrap tests for distributional treatment effects in instrumental variable models," *Journal of the American Statistical Association*, 97, 284–292.

——— (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 113, 231–263.

Angrist, J. D. (2004): "Treatment effect heterogeneity in theory and practice," *The Economic Journal*, 114, C52–C83.

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

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

——— (2010): "The credibility revolution in empirical economics: how better research design is taking the con out of econometrics," *Journal of Economic Perspectives*, 24, 3–30.

Balke, A. and J. Pearl (1997): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Battistin, E. and E. Rettore (2008): "Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs," *Journal of Econometrics*, 142, 715–730.

Bertanha, M. and G. W. Imbens (2014): "External validity in fuzzy regression discontinuity designs," NBER working paper no. 20773.

Bjorklund, A. and R. Moffitt (1987): "The estimation of wage gains and welfare gains in self-selection models," *Review of Economics and Statistics*, 69, 42–49.

Blundell, R. and R. L. Matzkin (2014): "Control functions in nonseparable simultaneous equations models," *Quantitative Economics*, 5, 271–295.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a discrete instrument," *Journal of Political Economy*, 125, 985–1039.

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2016): "From LATE to MTE: alternative methods for the evaluation of policy interventions," *Labour Economics*, 41, 47–60.

Deaton, A. S. (2009): "Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development," NBER working paper no. 14690.

Dubin, J. A. and D. L. McFadden (1984): ""An econometric analysis of residential electric appliance holdings and consumption," *Econometrica*, 52, 345–362.

Garen, J. (1984): "The returns to schooling: a selectivity bias approach with a continuous choice variable," *Econometrica*, 52, 1199–1218.

Greene, W. H. (2007): *Econometric Analysis*, Upper Saddle River, New Jersey: Prentice Hall, 7th ed.

Heckman, J. J. (1974): "Shadow prices, market wages, and labor supply," *Econometrica*, 42, 679–694.

——— (1976): "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement*, 5, 475–492.

——— (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153–161.

——— (1990): "Varieties of selection bias," *American Economic Review: Papers & Proceedings*, 80, 313–318.

——— (1997): "Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations," *Journal of Human Resources*, 32, 441–462.

Heckman, J. J. and R. Robb (1985): "Alternative methods for evaluating the impact of interventions: an overview," *Journal of Applied Econometrics*, 30, 239–267.

Heckman, J. J. and S. Urzua (2010): "Comparing IV with structural models: what simple IV can and cannot identify," *Journal of Econometrics*, 156, 27–37.

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding instrumental variables estimates in models with essential heterogeneity," *Review of Economics and Statistics*, 88, 389–432.

Heckman, J. J. and E. Vytlacil (2000): "Local instrumental variables," NBER technical working paper no. 252.

——— (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73, 669–738.

Huber, M. and G. Mellace (2015): "Testing instrument validity for LATE identification based on inequality moment constraints," *Review of Economics and Statistics*, 97, 398–411.

Hull, P. D. (2016): "Estimating hospital quality with quasi-experimental data," Working paper.

Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

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

Kitagawa, T. (2015): "A test for instrument validity," *Econometrica*, 83, 2043–2063.

Kline, P. and C. R. Walters (2016): "Evaluating public programs with close substitutes: the case of Head Start," *Quarterly Journal of Economics*, 131, 1795–1848.

Kowalski, A. (2016): "Doing more when you're running LATE: applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments," NBER working paper no. 22363.

Lakatos, I. (1976): "Falsification and the methodology of scientific research programs," in *Can Theories Be Refuted?*, ed. by S. G. Harding, Dordecht, Holland: D. Reidel Publishing Company.

LaLonde, R. J. (1986): "Evaluating the econometric evaluations of training programs with experimental data," *American Economic Review*, 76, 604–620.

Lee, L.-F. (1982): "Some approaches to the correction of selectivity bias," *Review of Economic Studies*, 49, 355–372.

Manski, C. F. (1990): "Nonparametric bounds on treatment effects," *American Economic Review: Papers & Proceedings*, 80, 319–323.

Mogstad, M., A. Santos, and A. Torgovitsky (2016): "Using instrumental variables for inference about policy relevant treatment parameters," Working paper.

Mroz, T. A. (1987): "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions," *Econometrica*, 55, 765–799.

Olsen, R. J. (1980): "A least squares correction for selectivity bias," *Econometrica*, 48, 1815–1820.

Popper, K. (1959): *The Logic of Scientific Discovery*, London, UK: Hutchinson and Co.

Puma, M., S. Bell, R. Cook, and C. Heid (2012): "Head Start Impact Study: Final Report," US Department of Health and Human Services, Administration for Children and Families.

Roy, A. (1951): "Some thoughts on the distribution of earnings," *Oxford Economic Papers*, 3, 135–146.

Smith, R. J. and R. W. Blundell (1986): "An exogeneity test for a simultaneous equations Tobit model with an application to labor supply," *Econometrica*, 54, 679–685.

Telser, L. G. (1964): "Advertising and competition," *Journal of Political Economy*, 72, 537–562.

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

Wald, A. (1940): "The fitting of straight lines if both variables are subject to error," *The Annals of Mathematical Statistics*, 11, 284–300.

White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–25.

Wooldridge, J. M. (2001): *Econometric Analysis of Cross-Section and Panel Data*, MIT Press.

——— (2015): "Control function methods in applied econometrics," *Journal of Human Resources*, 50, 420–445.

Appendix

Proof of Proposition 2

We begin by rewriting the IV and control function estimates of $LATE_z$ in matrix form. The IV estimator is given by

$$\widehat{LATE}_z^{IV} = \frac{\left[\hat{P}(z)\bar{Y}_1^z + (1-\hat{P}(z))\bar{Y}_0^z\right] - \left[\hat{P}(z-1)\bar{Y}_1^{z-1} + (1-\hat{P}(z-1))\bar{Y}_0^{z-1}\right]}{\hat{P}(z) - \hat{P}(z-1)}$$

$$= \Psi_1^z(\hat{P})'\bar{\mathbf{Y}}_1 - \Psi_0^z(\hat{P})'\bar{\mathbf{Y}}_0,$$

where $\bar{\mathbf{Y}}_d \equiv (\bar{Y}_d^0, \bar{Y}_d^1, ..., \bar{Y}_d^K)'$ is the $(K+1)\times 1$ vector of sample average outcomes for each value of z conditional on $D_i = d$ and $\hat{P}$ is the vector of propensity score estimates. The $(K+1)\times 1$ vector $\Psi_1^z(\hat{P})$ has $-\hat{P}(z-1)/[\hat{P}(z) - \hat{P}(z-1)]$ at entry $z-1$, $\hat{P}(z)/[\hat{P}(z) - \hat{P}(z-1)]$ at entry z, and zeros elsewhere, and the $(K+1)\times 1$ vector $\Psi_0^z(\hat{P})$ has $(1-\hat{P}(z-1))/[\hat{P}(z) - \hat{P}(z-1)]$ at entry $z-1$, $-(1-\hat{P}(z))/[\hat{P}(z) - \hat{P}(z-1)]$ at entry z, and zeros elsewhere:

$$\Psi_1^z(\hat{P}) = \left(0, ..0, \tfrac{-\hat{P}(z-1)}{\hat{P}(z)-\hat{P}(z-1)}, \tfrac{\hat{P}(z)}{\hat{P}(z)-\hat{P}(z-1)}, 0, ..., 0\right)',$$

$$\Psi_0^z(\hat{P}) = \left(0, ..0, \tfrac{(1-\hat{P}(z-1))}{\hat{P}(z)-\hat{P}(z-1)}, \tfrac{-(1-\hat{P}(z))}{\hat{P}(z)-\hat{P}(z-1)}, 0, ..., 0\right)'.$$

The second-step control function estimates with $L = K$ can be rewritten

$$(\hat{\alpha}_d, \hat{\gamma}_{d1}, ..., \hat{\gamma}_{dK}) = \arg \min_{\alpha_d, \gamma_{d1}, ..., \gamma_{dK}} \textstyle\sum_i 1\{D_i = d\} \left[Y_i - \alpha_d - \sum_{\ell=1}^K 1\{Z_i = z\}\gamma_{d\ell}\lambda_{d\ell}(\hat{P}(z))\right]^2.$$

This is a saturated OLS regression of Y_i on Z_i for each treatment category. The coefficient estimates satisfy

$$\hat{\alpha}_d + \sum_{\ell=1}^K \hat{\gamma}_{d\ell}\lambda_{d\ell}(\hat{P}(z)) = \bar{Y}_d^z,\ d \in \{0,1\},\ z \in \{0,1,...K\}.$$

Letting $\hat{\Delta}_d = (\hat{\alpha}_d, \hat{\gamma}_{d1}, ..., \hat{\gamma}_{dK})'$ denote the control function estimates for treatment value d, we can write this system in matrix form as

$$\Lambda_d(\hat{P})\hat{\Delta}_d = \bar{\mathbf{Y}}_d,$$

where the matrix $\Lambda_d(\hat{P})$ has ones in its first column and $\lambda_{dj-1}(\hat{P}(k))$ in row k and column $j > 1$:

$$\Lambda_d(\hat{P}) = \begin{bmatrix} 1 & \lambda_{d1}(\hat{P}(1)) & \cdots & \lambda_{dK}(\hat{P}(1)) \\ 1 & \lambda_{d1}(\hat{P}(2)) & \cdots & \lambda_{dK}(\hat{P}(2)) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{d1}(\hat{P}(K)) & \cdots & \lambda_{dK}(\hat{P}(K)) \end{bmatrix}.$$

The control function estimates are therefore given by

$$\hat{\Delta}_d = \Lambda_d(\hat{P})^{-1}\bar{\mathbf{Y}}_d.$$

The values of $\lambda_{dk}(\hat{P}(z))$ are well-defined for all (k,z) whenever $0 < \hat{P}(z) < 1\ \forall z$, and $\Lambda_d(\hat{P})$ is full rank if $\hat{P}(z) \neq \hat{P}(z')$ whenever $z \neq z'$. These conditions hold if Assumptions 1 and 2 are true for every pair of instrument values, so the matrix $\Lambda_d(\hat{P})$ is invertible under the conditions of Proposition 2 and the control function estimate $\hat{\Delta}_d$ exists.

In matrix form, the control function estimate of $LATE_z$ is given by

$$\widehat{LATE}_z^{CF} = \Upsilon^z(\hat{P})' \left(\hat{\Delta}_1 - \hat{\Delta}_0\right),$$

where the $(K+1)\times 1$ vector $\Upsilon^z(\hat{P})$ has first entry equal to unity and kth entry $\Gamma_{k-1}(\hat{P}(z-1), \hat{P}(z))$ for $k > 1$:

$$\Upsilon^z(\hat{P}) = \left(1, \tfrac{\hat{P}(z)\lambda_{11}(\hat{P}(z)) - \hat{P}(z-1)\lambda_{11}(\hat{P}(z-1))}{\hat{P}(z)-\hat{P}(z-1)}, ..., \tfrac{\hat{P}(z)\lambda_{1K}(\hat{P}(z)) - \hat{P}(z-1)\lambda_{1K}(\hat{P}(z-1))}{\hat{P}(z)-\hat{P}(z-1)}\right)'.$$

Plugging in the formulas for $\hat{\Delta}_1$ and $\hat{\Delta}_0$ yields

$$\widehat{LATE}_z^{CF} = \Upsilon^z(\hat{P})'\Lambda_1(\hat{P})^{-1}\bar{\mathbf{Y}}_1 - \Upsilon^z(\hat{P})'\Lambda_0(\hat{P})^{-1}\bar{\mathbf{Y}}_0.$$

The IV and control functions are therefore identical if $\Psi_d^z(\hat{P})' = \Upsilon^z(\hat{P})'\Lambda_d(\hat{P})^{-1}$ for $d \in \{0,1\}$, or equivalently, if $\Lambda_d(\hat{P})'\Psi_d^z(\hat{P}) = \Upsilon^z(\hat{P})$ for $d \in \{0,1\}$.

For $d = 1$, we have

$$\begin{aligned}\Lambda_1(\hat{P})'\Psi_1^z(\hat{P}) &= \left(1, \tfrac{\hat{P}(z)\lambda_{11}(\hat{P}(z)) - \hat{P}(z-1)\lambda_{11}(\hat{P}(z-1))}{\hat{P}(z)-\hat{P}(z-1)}, ..., \tfrac{\hat{P}(z)\lambda_{1K}(\hat{P}(z)) - \hat{P}(z-1)\lambda_{1K}(\hat{P}(z-1))}{\hat{P}(z)-\hat{P}(z-1)}\right)' \\ &= \Upsilon^z(\hat{P}).\end{aligned}$$

For $d = 0$, we have

$$\begin{aligned}\Lambda_0(\hat{P})'\Psi_0^z(\hat{P}) &= \left(1, \tfrac{\lambda_{01}(\hat{P}(z-1))(1-\hat{P}(z-1)) - \lambda_{01}(\hat{P}(z))(1-\hat{P}(z))}{\hat{P}(z)-\hat{P}(z-1)}, ..., \tfrac{\lambda_{01}(\hat{P}(z-1))(1-\hat{P}(z-1)) - \lambda_{01}(\hat{P}(z))(1-\hat{P}(z))}{\hat{P}(z)-\hat{P}(z-1)}\right)' \\ &= \left(1, \tfrac{\lambda_{11}(\hat{P}(z))\hat{P}(z) - \lambda_{11}(\hat{P}(z-1))\hat{P}(z-1)}{\hat{P}(z)-\hat{P}(z-1)}, ..., \tfrac{\lambda_{1K}(\hat{P}(z))\hat{P}(z) - \lambda_{1K}(\hat{P}(z-1))\hat{P}(z-1)}{\hat{P}(z)-\hat{P}(z-1)}\right)' \\ &= \Upsilon^z(\hat{P}),\end{aligned}$$

where the second equality follows from the fact that $p'\lambda_{1\ell}(p') - p\lambda_{1\ell}(p) = (1-p)\lambda_{0\ell}(p) - (1-p')\lambda_{0\ell}(p')$ for any p, p' and ℓ. This implies that $\widehat{LATE}_z^{IV}$ and $\widehat{LATE}_z^{CF}$ are equal to the same linear combination of $\bar{\mathbf{Y}}_1$ and $\bar{\mathbf{Y}}_0$, so these estimates are identical for any z.

Proof of Proposition 3

The unrestricted control function estimates come from the regression

$$\begin{aligned}Y_i = &\alpha_0(0)(1-D_i)(1-X_i) + \gamma_0(0)(1-D_i)(1-X_i)\lambda_0(\hat{P}(0,Z_i)) \\ &+\alpha_0(1)(1-D_i)X_i + \gamma_0(1)(1-D_i)X_i\lambda_0(\hat{P}(1,Z_i)) \\ &+\alpha_1(0)D_i(1-X_i) + \gamma_1(0)D_i(1-X_i)\lambda_1(\hat{P}(0,Z_i))\end{aligned}$$

$$+\alpha_1(1)D_iX_i + \gamma_1(1)D_iX_i\lambda_1(\hat{P}(1,Z_i)) + \epsilon_i.$$

We can write this equation in matrix form as

$$Y = W\Delta + \epsilon,$$

where W is the matrix of regressors and $\Delta = (\alpha_0(0), \gamma_0(0), \alpha_0(1), \gamma_0(1), \alpha_1(0), \gamma_1(0), \alpha_1(1), \gamma_1(1))'$ collects the control function coefficients. Under the conditions of Proposition 3, $W'W$ has full rank and the unrestricted control function estimates are

$$\hat{\Delta}_u = (W'W)^{-1}W'Y.$$

The estimator in equation (13) imposes three restrictions: $\alpha_1(1) - \alpha_1(0) = \alpha_0(1) - \alpha_0(0)$, $\gamma_1(1) = \gamma_1(0)$, and $\gamma_0(1) = \gamma_0(0)$. The resulting estimates can be written

$$(\hat{\Delta}_r, \hat{\varrho}) = \arg\min_{\Delta,\varrho} (Y - W\Delta)'(Y - W\Delta) - \varrho C\Delta,$$

where

$$C = \begin{bmatrix} -1 & 0 & 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and ϱ is a Lagrange multiplier. Then

$$\hat{\Delta}_r = \hat{\Delta}_u + (W'W)^{-1}C'\hat{\varrho},$$

$$\hat{\varrho} = -(C(W'W)^{-1}C')^{-1}C\hat{\Delta}_u.$$

For any estimate $\hat{\Delta}$, the corresponding estimate of LATE for compliers with $X_i = 1$ is $\Upsilon(\hat{P})'\hat{\Delta}$, with

$$\Upsilon(\hat{P}) = (0, 0, -1, -\Gamma(\hat{P}(1,0), \hat{P}(1,1)), 0, 0, 1, \Gamma(\hat{P}(1,0), \hat{P}(1,1)))'.$$

The restricted estimate of $LATE(1)$ is therefore

$$\widehat{LATE}_r^{CF}(1) = \Upsilon(\hat{P})' \left(\hat{\Delta}_u + (W'W)^{-1}C'\hat{\varrho}\right)$$

$$= \widehat{LATE}_1^{CF}(1) + \Upsilon(\hat{P})'(W'W)^{-1}C' \left(C(W'W)^{-1}C'\right)^{-1} \zeta,$$

where $\zeta = -C\hat{\Delta}_u$ is the constraint vector evaluated at the unrestricted estimates:

$$\zeta = ([\hat{\alpha}_0(0) - \hat{\alpha}_0(1)] - [\hat{\alpha}_1(0) - \hat{\alpha}_1(1)], \hat{\gamma}_1(0) - \hat{\gamma}_1(1), \hat{\gamma}_0(0) - \hat{\gamma}_0(1))'.$$

Write $\Omega = (W'W)^{-1}C' \left(C(W'W)^{-1}C'\right)^{-1}$, and let ν_k denote the 3×1 vector equal to the transpose of the kth row of Ω. Using the fact that a scalar is equal to its trace, we can then write the difference in restricted and unrestricted LATE estimates as

$$\widehat{LATE}_r^{CF}(1) - \widehat{LATE}_1^{CF}(1) = tr\left(\Upsilon(\hat{P})'\Omega\zeta\right)$$

$$= tr\left(\Omega\zeta\Upsilon(\hat{P})'\right)$$

$$= \varphi'\zeta,$$

where $\varphi = \nu_7 - \nu_3 + \Gamma(\hat{P}(1,0), \hat{P}(1,1))(\nu_8 - \nu_4) \equiv (\varphi_1, \varphi_2, \varphi_3)'$. Then

$$\widehat{LATE}_r^{CF}(1) - \widehat{LATE}_1^{CF}(1) = \varphi_1\left([\hat{\alpha}_0(0) - \hat{\alpha}_0(1)] - [\hat{\alpha}_1(0) - \hat{\alpha}_1(1)]\right) + \varphi_2\left(\hat{\gamma}_1(0) - \hat{\gamma}_1(1)\right) + \varphi_3\left(\hat{\gamma}_0(0) - \hat{\gamma}_0(1)\right)$$

$$= \varphi_1\left(\widehat{LATE}_1^{CF}(1) - \widehat{LATE}_0^{CF}(1)\right) + \left(\hat{\gamma}_1(0) - \hat{\gamma}_1(1)\right)\left(\varphi_2 + \varphi_1\Gamma(\hat{P}(1,0), \hat{P}(1,1))\right)$$

$$+\left(\hat{\gamma}_0(0) - \hat{\gamma}_0(1)\right)\left(\varphi_3 - \varphi_1\Gamma(\hat{P}(1,0), \hat{P}(1,1))\right).$$

This implies

$$\widehat{LATE}_r^{CF}(1) = w\widehat{LATE}_1^{CF}(1) + (1-w)\widehat{LATE}_0^{CF}(1) + b_1\left(\hat{\gamma}_1(1) - \hat{\gamma}_1(0)\right) + b_0\left(\hat{\gamma}_0(1) - \hat{\gamma}_0(0)\right),$$

where $w = 1 + \varphi_1$, $b_1 = -(\varphi_2 + \varphi_1\Gamma(\hat{P}(1,0), \hat{P}(1,1)))$, and $b_0 = \varphi_1\Gamma(\hat{P}(1,0), \hat{P}(1,1)) - \varphi_3$. Furthermore, note that the elements of φ only depend on sample moments of D_i, X_i, and $\hat{P}(X_i, Z_i)$, so the proposition follows.

Proof of Proposition 4

The log likelihood function for model (16) is

$$\log\mathcal{L}\left(P(0), P(1), \alpha_0, \alpha_1, \rho_0, \rho_1\right) = \sum_i D_i \log\left(\int_0^{P(Z_i)} \left[Y_i F_{\epsilon|U}(\alpha_1|u;\rho_1) + (1-Y_i)(1 - F_{\epsilon|U}(\alpha_1|u;\rho_1))\right] du\right)$$

$$+\sum_i (1-D_i)\log\left(\int_{P(Z_i)}^1 \left[Y_i F_{\epsilon|U}(\alpha_0|u;\rho_0) + (1-Y_i)(1 - F_{\epsilon|U}(\alpha_0|u;\rho_0))\right] du\right).$$

We first rewrite this likelihood in terms of the six identified parameters of the LATE model, which are given by

$$\pi_{at} = P(0),$$

$$\pi_c = P(1) - P(0),$$

$$\mu_{1at} = \frac{\int_0^{P(0)} F_{\epsilon|U}(\alpha_1|u;\rho_1)du}{P(0)},$$

$$\mu_{0nt} = \frac{\int_{P(1)}^1 F_{\epsilon|U}\left(\alpha_0|u;\rho_0\right)du}{1 - P(1)},$$

$$\mu_{dc} = \frac{\int_{P(0)}^{P(1)} F_{\epsilon|U}(\alpha_d|u;\rho_d)du}{P(1) - P(0)},\ d \in \{0,1\}.$$

Note that since $F_{\epsilon|U}(\cdot|u;\rho)$ is a CDF, we have $\mu_{dg} \in [0,1]\ \forall(d,g)$. Substituting these parameters into the likelihood function yields

$$\begin{aligned}\log \mathcal{L}(\pi_{at},\pi_c,\mu_{1at},\mu_{0nt},\mu_{1c},\mu_{0c}) =& \sum_i D_i Z_i \log\left(\pi_{at}\left[Y_i\mu_{1at}+(1-Y_i)(1-\mu_{1at})\right]+\pi_c\left[Y_i\mu_{1c}+(1-Y_i)(1-\mu_{1c})\right]\right)\\ &+\sum_i D_i(1-Z_i)\log\left(\pi_{at}\left[Y_i\mu_{1at}+(1-Y_i)(1-\mu_{1at})\right]\right)\\ &+\sum_i (1-D_i)Z_i\log\left((1-\pi_{at}-\pi_c)\left[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})\right]\right)\\ &+\sum_i (1-D_i)(1-Z_i)\log\left((1-\pi_{at}-\pi_c)\left[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})\right]+\pi_c\left[Y_i\mu_{0c}+(1-Y_i)(1-\mu_{0c})\right]\right).\end{aligned}$$

We first consider interior solutions. The first-order conditions are

$$[\mu_{1at}]:\ \sum_i\left(\frac{D_i(2Y_i-1)Z_i\pi_{at}}{\pi_{at}[Y_i\mu_{1at}+(1-Y_i)(1-\mu_{1at})]+\pi_c[Y_i\mu_{1c}+(1-Y_i)(1-\mu_{1c})]}+\frac{D_i(2Y_i-1)(1-Z_i)\pi_{at}}{\pi_{at}[Y_i\mu_{1at}+(1-Y_i)(1-\mu_{1at})]}\right)=0,$$

$$[\mu_{0nt}]:\ \sum_i\left(\frac{(1-D_i)(2Y_i-1)Z_i(1-\pi_{at}-\pi_c)}{(1-\pi_{at}-\pi_c)[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})]}+\frac{(1-D_i)(2Y_i-1)(1-Z_i)(1-\pi_{at}-\pi_c)}{(1-\pi_{at}-\pi_c)[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})]+\pi_c[Y_i\mu_{0c}+(1-Y_i)(1-\mu_{0c})]}\right)=0,$$

$$[\mu_{1c}]:\ \sum_i\frac{D_iZ_i(2Y_i-1)\pi_c}{\pi_{at}[Y_i\mu_{1at}+(1-Y_i)(1-\mu_{1at})]+\pi_c[Y_i\mu_{1c}+(1-Y_i)(1-\mu_{1c})]}=0,$$

$$[\mu_{0c}]:\ \sum_i\frac{(1-D_i)(1-Z_i)(2Y_i-1)\pi_c}{(1-\pi_{at}-\pi_c)[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})]+\pi_c[Y_i\mu_{0c}+(1-Y_i)(1-\mu_{0c})]}=0,$$

$$\begin{aligned}[\pi_{at}]:\ &\sum_i\frac{D_iZ_i[Y_i\mu_{1at}+(1-Y_i)(1-\mu_{1at})]}{\pi_{at}[Y_i\mu_{1at}+(1-Y_i)(1-\mu_{1at})]+\pi_c[Y_i\mu_{1c}+(1-Y_i)(1-\mu_{1c})]}+\sum_i\frac{D_i(1-Z_i)}{\pi_{at}}\\ &-\sum_i\frac{(1-D_i)Z_i[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})]}{(1-\pi_{at}-\pi_c)[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})]}-\sum_i\frac{(1-D_i)(1-Z_i)[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})]}{(1-\pi_{at}-\pi_c)[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})]+\pi_c[Y_i\mu_{0c}+(1-Y_i)(1-\mu_{0c})]}=0,\end{aligned}$$

$$\begin{aligned}[\pi_c]:\ &\sum_i\frac{D_iZ_i[Y_i\mu_{1c}+(1-Y_i)(1-\mu_{1c})]}{[\pi_{at}[Y_i\mu_{1at}+(1-Y_i)(1-\mu_{1at})]+\pi_c[Y_i\mu_{1c}+(1-Y_i)(1-\mu_{1c})]]}-\sum_i\frac{(1-D_i)Z_i[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})]}{(1-\pi_{at}-\pi_c)[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})]}\\ &-\sum_i\frac{(1-D_i)(1-Z_i)(2Y_i-1)(\mu_{0nt}-\mu_{0c})}{(1-\pi_{at}-\pi_c)[Y_i\mu_{0nt}+(1-Y_i)(1-\mu_{0nt})]+\pi_c[Y_i\mu_{0c}+(1-Y_i)(1-\mu_{0c})]}=0.\end{aligned}$$

Under Assumptions 1 and 2 we can compute $\hat{\mu}_{1at}^{IV}$, $\hat{\mu}_{0nt}^{IV}$, $\hat{\mu}_{1c}^{IV}$, and $\hat{\mu}_{0c}^{IV}$. Setting $\hat{\pi}_c^{IV}=\hat{P}(1)-\hat{P}(0)$ and $\hat{\pi}_{at}^{IV}=\hat{P}(0)$ and plugging the IV parameter estimates into the FIML first order conditions, we see that these conditions are satisfied. Thus at interior solutions maximum likelihood and IV estimators of all parameters are equal, and it follows that $\hat{\mu}_{dc}^{ML}=\hat{\mu}_{dc}^{IV}$ for $d\in\{0,1\}$.

Next, we consider corner solutions, which occur when at least one parameter lies outside [0,1] at the unconstrained solution to the first order conditions. Note that $\widehat{\mu}_{1at}^{IV}$, $\widehat{\mu}_{0nt}^{IV}$, and $\hat{\pi}_{at}^{IV}$ are sample means of binary variables, so these estimates are always in the unit interval. $\hat{\pi}_c^{IV}$ is the difference in empirical treatment rates between the two values of Z_i; without loss of generality we assume that $Z_i=1$ refers to the group with the higher treatment rate, so $\hat{\pi}_c^{IV}\in(0,1)$. Thus a constraint binds if and only if $\hat{\mu}_{dc}^{IV}$ is outside [0,1] for $d=0$, $d=1$, or both. In these cases at least one of the maximum likelihood complier means fails to match the corresponding IV estimate because the IV estimate is outside the FIML parameter space. This establishes that the FIML and IV estimates match if and only if both $\hat{\mu}_{1c}^{IV}$ and $\hat{\mu}_{0c}^{IV}$ are in [0,1], which completes the proof.