Das, Tirthatanmoy; Polachek, Solomon

Working Paper

A New Strategy to Identify Causal Relationships: Estimating a Binding Average Treatment Effect

IZA Discussion Papers, No. 12766

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Das, Tirthatanmoy; Polachek, Solomon (2019) : A New Strategy to Identify Causal Relationships: Estimating a Binding Average Treatment Effect, IZA Discussion Papers, No. 12766, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/215162

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12766

A New Strategy to Identify Causal Relationships: Estimating a Binding Average Treatment Effect

Tirthatanmoy Das
Solomon W. Polachek

NOVEMBER 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12766

A New Strategy to Identify Causal Relationships: Estimating a Binding Average Treatment Effect

Tirthatanmoy Das
Indian Institute of Management Bangalore and IZA

Solomon W. Polachek
State University of New York at Binghamton, Liverpool Hope University and IZA

NOVEMBER 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

A New Strategy to Identify Causal Relationships: Estimating a Binding Average Treatment Effect*

This paper proposes a new strategy to identify causal effects. Instead of finding a conventional instrumental variable correlated with the treatment but not with the confounding effects, we propose an approach which employs an instrument correlated with the confounders, but which itself is not causally related to the direct effect of the treatment. Utilizing such an instrument enables one to estimate the confounding endogeneity bias. This bias can then be utilized in subsequent regressions first to obtain a "binding" causal effect for observations unaffected by institutional barriers that eliminate a treatment's effectiveness, and second to obtain a population-wide treatment effect for all observations independent of institutional restrictions. Both are computed whether the treatment effects are homogeneous or heterogeneous. To illustrate the technique, we apply the approach to estimate sheepskin effects. We find the bias to be approximately equal to the OLS coefficient, meaning that the sheepskin effect is near zero. This result is consistent with Flores-Lagunes and Light (2010) and Clark and Martorell (2014). Our technique expands the econometrician's toolkit by introducing an alternative method that can be used to estimate causality. Further, one potentially can use both the conventional instrumental variable approach in tandem with our alternative approach to test the equality of the two estimators for a conventionally exactly identified causal model, should one claim to already have a valid conventional instrument.

JEL Classification: C18, C36, I26, J24, J33

Keywords: causality, OLS biases, sheepskin effects

Corresponding author:
Solomon W. Polachek
Department of Economics
State University of New York at Binghamton
Binghamton, NY 13902
USA

E-mail: polachek@binghamton.edu

* We thank Joshua Angrist, Robert Basmann, Alfonso Flores-Lagunes, Ivan Korolev, and David Slichter for extremely valuable comments.

Riddle: What do you call a student who graduates last in medical school? Answer: Doctor.

1. Introduction

Identification of causal effects (β_T) in the typical regression $y = \beta_0 + \beta_T x + \epsilon$ requires x to be exogenous, and thus uncorrelated with ϵ. But instead for many practical applications x is endogenous and thus correlated with ϵ because of simultaneity, omitted variables, and/or measurement errors in x. Rather than estimating the causal effect of x on y, the OLS estimator for the x coefficient measures a combined effect β consisting of both the true effect (β_T) as well as a bias θ so that $\beta = \beta_T + \theta$. As such, the OLS estimator in the presence of endogeneity is statistically inconsistent because it contains the confounding effect (θ) which prevents it from converging to the true population parameter representing the causal effect (β_T) when the sample size grows large.

The conventional approach to handle such an endogeneity problem is to find an exogenous instrumental variable correlated with x but uncorrelated with the confounders. Unfortunately finding valid instruments is difficult because many seemingly valid instruments violate this exclusion restriction, as they are often correlated with confounder variables. As an example, take the case of growing up close to a college, an often used instrument to estimate returns to education (Card, 1995). Carneiro and Heckman (2002) found that despite being correlated with college attendance, distance to a college is an invalid instrument because it is similarly correlated with a measure of ability, a confounding variable which also determines a respondent's earnings.

This paper proposes an alternative strategy which yields a new causal effects estimator. It relies on a unique identifying assumption where exogenous institutional factors make the treatment ineffective for a subset of the population. These observations receive zero treatment effect despite being treated. For these observations any observed correlation between the outcome and treatment must be due to confounders. This population subset enables one to estimate the confounding endogeneity bias, which then can be used to identify the causal effects.

This identifying assumption differentiates our approach from the existing literature in at least three ways. First, we identify two types of treatment effects. To our knowledge, one – the binding average treatment effect (BATE) – is new to the literature, as it explicitly differentiates between interpersonal heterogeneous treatment effects and treatment effects based on institutional considerations. Second, our identifying assumption enables us to use the treatment itself as an instrument whereas traditional approaches require finding exogenous variables or a valid instrument. Third, the approach enables us to test the validity of our identification strategy whereas traditional approaches only can do so with an exogenous variable or a potentially valid instrument.

The BATE estimator is particularly important when institutions have an impact. As a very simple example, take the case of the federal minimum wage. A number of empirical analyses (e.g., beginning with Moore, 1971) estimate the impact of federal minimum wage legislation on teenage unemployment. However, a federal minimum wage increase is not binding in states where the minimum wage is already higher than the mandated federal level. Incorporating such states in the analysis will bias estimates of the overall impact; but omitting these states as in many of the current

difference-in-differences (DID) studies could lead to selectivity biases. Further, estimates obtained either incorporating or omitting these states not bound by the federal minimum wage would lead to erroneous estimates should a new binding federal law exceed these formerly binding state levels. As such distinguishing between binding treatment effects and non-binding treatment effects can have consequences with regard to public policy.

We propose employing an instrument correlated with the confounders, but not causally related to the direct effect of the treatment. This contrasts with the conventional approach which instead finds a variable correlated with x but not with the confounding effects. As will be explained, our approach amounts to an instrument defining a situation where the treatment x has no direct causal effect on the outcome, but instead operates *only* through confounders. This entails finding an instrument that correlates with the outcome variable solely by way of the confounders, yet not because of the treatment per se. As such, our exclusion restriction is diametrically the opposite of a traditional instrument. Whereas the traditional IV requires a situation in which the instrument has a zero correlation with the unobservable confounders, our approach requires a situation in which the instrument is correlated with the confounders, but has no direct effect on the outcome. Similarly, our approach is methodologically distinct from traditional approaches such as DID, regression discontinuity, propensity score matching, and randomized control trials. Selection in such conventional methods can be viewed as randomly assigning observations into treatment and non-treatment groups to get a causal effect. In contrast, our approach models selection into a "zero-treatment effect" group to get at the OLS bias θ. Once the θ bias is obtained, it can then be used in another regression to identify the treatment effect either when the treatment effect is homogeneous or heterogeneous. As such, in our approach, identifying θ using our alternative instrument serves as a new strategy to obtain a consistent estimator of the true causal effects.[1]

Our approach is most related to the control function approach (Heckman and Robb 1985; Imbens and Woolridge 2007). Both approaches utilize existing sample data to compute an aspect of the confounding bias. The CFA approach *adjusts* for the endogeneity bias whereas our approach explicitly *estimates* the bias. Also, our approach is related to Angrist and Krueger (1995) and Flores and Flores-Lagunes (2009) because their approaches use a subsample of data to adjust for the bias, as does ours. However, what differentiates our approach from the CFA, the Angrist and Krueger, and the Flores and Flores-Lagunes applications, is our method identifies the causal effect without any need for exogenous or traditional instrumental variables. This is because our identifying assumption utilizes the endogenous variable itself as an instrument for confounders when an institutional constraint sets the treatment effect zero.

The main challenge to implementing the approach is to find an instrument for the confounding effects. This might seem difficult given confounders are typically unobservable. However, one way to do this is to use x itself as an instrument for the confounders if one can find a situation in which x does not influence y directly, but only does so through confounders. This identification strategy amounts to finding a circumstance where treatment x varies with outcome y, but by itself does not causally influence y. As it turns out, there are many such examples.

[1] To be more precise, our analysis yields two treatment effects. β_T represents the treatment when the treatment is effective, i.e. binding. We denote β_T to be the binding average treatment effect (BATE). We also define a population-wide average treatment effect β_T^p comparable to the commonly defined average treatment effect (ATE).

To illustrate, we apply the approach to estimate sheepskin effects, a topic for which there is an existing literature but with contradictory evidence. OLS estimates indicate employers pay large rewards to those with a high school or college diploma compared to those similarly schooled with 12 or 16 years of education. Many interpret this result to mean that a diploma per se is thus important for economic success. For obvious reasons, this is known as the sheepskin effect. However, these higher rewards observed in OLS regressions might come about *not* because having a degree is intrinsically valued by employers, but because those with a degree are motivated individuals with a high level of stick-to-itiveness, an unobserved characteristic affecting earnings. In this situation OLS results are biased and inconsistent because actually earning a degree could signify the effects of these unobserved individual characteristics rather than the intrinsic causal value of such a certification. Perhaps for this reason many employers require a diploma as a validation especially in occupations where they believe the skill of a worker must reach a certain threshold, determined by a baccalaureate. On the other hand, for seemingly comparable workers, a number of employers do not require a degree. Clearly those employers do not value a sheepskin, or otherwise they would have required the degree. Thus, in this case, diplomas cannot have an intrinsic value if degreed employees get higher wages in these jobs not requiring a diploma. Higher earnings for these workers must come about because of unobserved confounders rather than the innate value of a diploma. We use such a subsample to estimate the OLS confounding bias θ. As such, jobs not requiring a diploma motivate our instrument.

To implement the technique we compute θ based on a regression of incumbents in occupations not requiring a degree. We find θ to be approximately equal to the OLS coefficient for those in all occupations, meaning that the sheepskin effect is near zero. This result is consistent with Clark and Martorell (2014) and Flores-Lagunes and Light (2010) who find similar results respectively using a regression discontinuity strategy and an OLS with interaction terms between years of school and getting a degree.

Whereas we apply the technique to estimating sheepskin effects, there are numerous other applications for which our approach can be useful to estimate unbiased causal effects. These include estimating the causal impact of deploying additional police on marijuana use, the impact of parental inputs on child development, and the impact of school quality on earnings. So if, the answer to the above riddle is really true, and the grades one gets in medical school relative to one's classmates do not affect one's ability to practice medicine, then variations in economic success for medical doctors must be attributable to unobserved factors, in short the confounders. In this case, one can compare the earnings gradient of class rank based on grades for physicians to that of other comparable occupations to get at the true direct effect of class rank. These are but several of numerous possible applications for which our approach can be useful to estimate unbiased causal effects.

What is important is our approach identifies the *bias* contained in an OLS estimated causal relationship rather than attempting to isolate the effect via a potentially invalid instrument. Thus our technique expands the econometrician's toolkit by introducing an alternative method that can be used to estimate causality. Because in some circumstances it may be easier to identify the bias by devising an instrument correlated only with the confounding effects rather than only correlated with the direct effect of the treatment, our procedure should be valuable to assess causal relationships. Further, one potentially can use both the usual instrumental variable approach in tandem with our alternative approach to test the equality of the two estimators for a conventionally exactly identified causal model should one claim to

already have a valid traditional instrument. Obviously our approach is especially useful for policy related research since that literature requires unbiased causal effect estimates.

Before describing our method in detail, we go over the basic endogeneity problem in Section 2 and describe the current typically used IV strategy in Section 3. In Section 4 we describe our method, which denote to be a "bias excision" approach. Then in Section 5 we apply it to estimate potential sheepskin effects. We conclude in Section 6 citing several possible applications of the technique**.**

2. The Endogeneity Problem

Assume an economic relationship governed by the following equation

$$y = \beta_0 + \beta_T x + \epsilon \quad (1)$$

where $E[\epsilon] = 0$, but x is statistically dependent of ϵ so that $E[\epsilon|x] \neq 0$. The latter expectation implies (1) can be expressed as

$$E[y|x] = \beta_0 + \beta_T x + E[\epsilon|x] \quad (2)$$

The simple OLS estimate of β_T suffers from an omitted variable bias because it omits $E[\epsilon|x]$ which is a function of x. Running $y = \beta_0 + \beta_T x + \epsilon$ amounts to omitting the variable $E[\epsilon|x]$, implying that $E[\epsilon|x]$ goes into the error term ϵ. Because under general conditions $Cov[x, E[\epsilon|x]] \neq 0$, it can be shown that $Cov[x, \epsilon] \neq 0$. As a result OLS yields a biased and inconsistent estimator on β_T.

As such, ϵ can be represented by the following

$$\epsilon = a + \theta x + u \quad (3)$$

where $\theta \neq 0$, and u is an error vector such that $u \perp x$. See Appendix A for the derivation.

Substituting ϵ from (3) into (1) yields

$$y = (\beta_0 + a) + \beta_T x + \theta x + u \quad (4)$$

Rewriting (4) yields

$$y = \pi_0 + \beta x + u \quad (5)$$

where $\pi_0 = (\beta_0 + a)$ and $\beta = (\beta_T + \theta)$.

In this form the β coefficient of x is

$$\beta = \frac{Cov(y,x)}{Var(x)} = \beta_T + \theta \quad (6)$$

If $\theta \neq 0$, then $\beta \neq \beta_T$, indicating the OLS of y on x produces a biased and inconsistent estimator of the true effect.

3. **The Traditional Solution: An Instrumental Variable Method** (2SLS with IVs)

The traditional solution entails finding a single instrument or a vector of instrumental variables z such that each component of z is strongly correlated with x but uncorrelated with ϵ. Thus the following conditions have to be met:

$$Cov(x,z) \neq 0 \quad (7)$$

and

$$Cov(\epsilon,z) = 0 \quad (8)$$

Typically the estimation entails two stages. In stage 1, regress x on z such that

$$x = \delta_0 + \delta_1 z + \zeta \quad (9)$$

where the δs are the coefficients and ζ is random noise. This OLS yields estimates of $\hat{\delta}_0$ and $\hat{\delta}_1$ which are used to predict $\hat{x}$. In stage 2 one regresses y on $\hat{x}$ such that

$$y = \phi_0 + \phi_1 \hat{x} + \eta \quad (10)$$

where ϕ_0 and ϕ_1 are the coefficients and η is random noise. If z satisfies both the conditions (7) and (8), then the OLS estimator of $\hat{\phi}_1$ is consistent, so that $plim\ \hat{\phi}_1 = \beta_T$ (e.g., Basmann, 1957; Sargan, 1958; Theil, 1958; and Angrist and Imbens, 1995).

This method directly identifies the true causal effect. However, in practice it may be difficult to find appropriate z that satisfy the relevance and exclusion restriction conditions.

4. A Bias Excision Estimation Approach

Our approach utilizes a different type instrumental variable to identify and estimate the endogeneity bias often inherent in OLS estimation. We utilize the instrument in an initial first-stage regression to estimate the OLS bias and the resulting confounding effects. Then after netting out the confounding effects from the outcome data, we run two subsequent regressions to obtain two types of treatment effect parameters. One (β_T^P) is the population-wide average effect and the other (β_T) is the average treatment effect on the population for which the treatment is actually binding. We consider the bias and both parameters under homogenous and heterogeneous treatment regimes. When the treatment effects are homogeneous these parameters are constant across all observations. When the treatment effects are heterogeneous the parameters can vary across the observations.

We begin by defining the necessary criteria underlying the instrumental variable we propose. This instrumental variable can be a continuous or dichotomous variable. First, we present each of these for the homogeneous treatment case. Second, we present each for the heterogeneous case. Third, we assess potential selectivity biases. Finally, in the fourth subsection, we briefly relate our approach to aspects of the existing causality literature.

4.1 The Setup

Begin by rewriting (4) as

$$y = \beta_0 + \theta x + \epsilon' \qquad 11$$

where θ is the OLS bias coefficient and $\epsilon' = a + \beta_T x + u$ includes the direct treatment effect rather than the confounding effect in a typical regression, both previously defined in (3). Estimating (11) via an instrumental variable approach would require an instrument w (or a vector of variables W) such that w varies with the treatment x, i.e.,

$$Cov(x, w) \neq 0 \qquad 12$$

but is uncorrelated with the error, i.e.,

$$Cov(\epsilon', w) = 0. \qquad 13$$

Here w is an instrument for x that can be used to determine the bias θ. For (12) and (13) to hold, w cannot be correlated with the true effect ($\beta_T x$) which is in the error ϵ', but instead correlated with the confounding effect (θx).[2] Note, w differs from the traditional instrument. Indeed both are on opposite sides of the spectrum. The traditional IV requires a zero correlation between the instrument and the unobservables. Our bias excision approach IV requires a zero correlation between the instrument and

[2] One might argue such an instrument cannot exist since w is correlated with x, and x is a component of ϵ', but as we will show, under the right circumstances w can and does exist.

the direct effect of the endogenous variable. Clearly these exclusion restrictions differ from each other. Both are strong in their own ways, but surely there must be some circumstances in which our approach is more reasonable. This is the reason why in the conclusion we pitch our method as an alternative approach to identify causality without claiming superiority.[3]

Should one be able to find such an instrument, the same estimation procedure as in the regular IV case can be followed to obtain the estimator of θ, the bias component. Once the bias is identified, as will be shown, the bias component $\hat{\theta}x$ can be subtracted from (4) to obtain the two treatment effect estimates: β_T^P, the population-wide treatment effect, and β_T, the binding treatment effect, both mentioned above.

Just as with the typical IV estimation, there are various ways to actually implement the approach. One way is a two-step procedure. First, one finds an instrument (w) solely correlated with the biased effect of x on y. Here one employs w to estimate x so that

$$x = \gamma_0 + \gamma_1 w + \varsigma. \qquad 14$$

The second stage consists of using $\hat{x}$ in the main regression

$$y = \beta_0 + \psi\hat{x} + \epsilon'. \qquad 15$$

This yields $\hat{\psi}$ such that $plim\,\hat{\psi} = \theta$. It constitutes the confounding impact of x, which we denote to be the bias parameter. Finally, as alluded to above, and as will be explained shortly in more detail, one can use this estimate of θ to subtract the biased component $\hat{\psi}x$ from y, and then with the appropriate regressions get the two treatment effects $\hat{\beta}_T$ and $\hat{\beta}_T^P$.

The challenge with this implementation procedure is to find an instrument actually correlated with confounders, but which itself has no direct effect on the outcome. However, because confounders are unobservable in the data, it is potentially hard to find such an instrument. As such, one cannot rely on empirical strategies, but instead one must build upon prior institutional knowledge. [4]

Economic theory can help identify situations where such an instrument can be found. One such instance is when a particular institutional circumstance prohibits treatment x from having a direct causal effect on y, even though respondents are treated (have non-zero x). Here x does not affect y directly, but instead x affects y only through confounders.[5] For this reason one would think x should not enter the

[3] See Appendix B which more formally contrasts our IV approach and the traditional IV approach.

[4] This is consistent with Basmann's (2006:278) statement "Intuition of economic reality is central to conducting good econometrics."

[5] In a sense this is the flipside of "twins" studies. Analyses using twins data run a differenced regression of treatment outcome for monozygotic twins (Taubman, 1976; Ashenfelter and Rouse, 1998). In this case the causal effects are obtained because the treatment effect is measured holding constant confounding genetic characteristics. In our case, we estimate only the bias by holding the direct effect constant, which we then subsequently use to get the true effect. It is important to note

regression equation ($y = \beta_0 + \beta_T x + \epsilon$) because $\beta_T = 0$, but in reality x does enter since x is correlated with the confounders given that $\theta \neq 0$. As such, in this case, one can use x itself as an instrument for the confounders since x affects y only through the confounders and does not have a direct own effect.

In such a circumstance we identify the causal effect based on a design that assigns $\beta_T = 0$ to a representative subsample. Utilizing this subsample, we first identify the bias θ which later allows us to estimate the causal effect parameters β_T^P and β_T. In doing so two independence conditions are necessary, both of which are reasonable.

Condition 1: The treatment x is mean independent of u, implying $E[u|x] = E[u] = 0$.

Condition 1 simply indicates that the treatment assignment is mean independent of the unobserved u. Condition 1 is about independence and therefore different than the construction in (3) that $Cov[x, u] = 0$. Intuitively Condition 1 implies that once one nets out the OLS bias, the error structure adheres to the usual bias free OLS error structure.

Condition 2: The assignment of $\beta_T = 0$ is independent of the assignment of treatment implying $E[\beta_T|x] = E[\beta_T]$.

This condition means that those observations for which the treatment has no effect ($\beta_T = 0$) are not determined by the treatment status. This condition is reasonable given that an institutional circumstance dictates members of this population subgroup.[6] Nevertheless, it is permissible that assignment to a $\beta_T = 0$ group depends on other observed and unobserved characteristics. This condition is less stringent than the usually used complete random assignment employed when using a typical IV.

In short, one estimates the impact of x in two steps: One using only those respondents in the sample for which the treatment x is known to have no direct effect, and then second, using the entire sample. To implement this approach, we parcel the sample into three groups. We denote S to be the entire sample which contains the three subsamples: S_A, S_B, S_C. The first subsample $S_A \subset S$ is the set of individuals (observations) that do not receive any treatment. For these $x = 0$. The second subsample $S_B \subset S$ consists of individuals who receive treatment, but the sole effect of treatment is through confounders. Thus for individuals in S_B the treatment has no direct own effect. Finally the remaining subsample $S_C \subset S$ are those receiving treatment, but the effects arise partly from confounding forces, and partly from the treatment itself. Our first step entails a regression using the set $\{S_A, S_B\}$ and our second step utilizes the entire observation set $\{S_A, S_B, S_C\}$.[7]

---

that our identifying assumption that $\beta_T = 0$ arises because of institutional interventions, and not due to heterogeneous effects of x. As already indicated, later we distinguish between estimating the treatment effect for the whole population (β_T^P) which includes observations in which the treatment is and is not binding, and estimating the treatment effect only for the sample in which the treatment is binding (β_T).

[6] We emphasize that the presence of a subsample with $\beta_T = 0$ arises because of institutional considerations. However, β_T for these individuals would revert back to the binding β_T should these institutional considerations be eliminated.

[7] The closest approaches to ours we are aware of in the literature are the control function approach (Heckman and Robb1985, Imbens and Wooldridge, 2007 and Angrist and Krueger, 1995) and the mediation bias elimination approach of Flores and Flores-Lagunes (2009). Appendix C shows how our approach differs from the CFA approach. Section 4.3 of the Flores and Flores-Lagunes (2009) paper utilizes a population subsample that eliminates mediation effects caused by a treatment in order to obtain a net average treatment effect (NATE).

We distinguish two possibilities: (1) when the treatment is continuous, and (2) when the treatment is dichotomous. As mentioned, for each of these possibilities we identify the two treatment effect parameters.[8] One denoted β_T^P represents the average treatment effect for the whole population, which includes those population members for which the treatment is binding and those population members for which the treatment effect is institutionally zero and thus nonbinding. The second parameter denoted β_T is the average treatment effect for those population members for which the treatment is binding.[9]

4.1.A Continuous Treatment Variable

Identification of β_T^P: The Population Average Treatment Effect

In general (14) depicts the first-stage regression. However, when x affects y only through confounders, one can use x itself as an instrument, but for the subset of the data where x has no direct effect on y.[10] Because x is the instrument for itself, that is $\hat{x} = x$, this implies a second stage regression (16) rather than (15)

$$y = \pi_0 + \psi x + u \qquad 16$$

Estimated only for observations in S_A and S_B. This regression yields the estimate of the confounding effect $\hat{\theta}$ such that the $plim\ \hat{\psi} = \theta$ since $E[u|x] = 0$.[11]

Once $\hat{\theta}$ is obtained we subtract out the inherent bias in y for the entire sample $\tilde{y} = y - \hat{\theta}x$. We then run the following regression for the entire sample (i.e. $\{S_A, S_B, S_C\}$)

$$\tilde{y} = \pi_0 + \beta_T^P x + u. \qquad 17$$

Condition 1, $E[u|x] = 0$, then allows one to consistently estimate $\hat{\beta}_T^P$, that is the overall population treatment effect.[12] The coefficient $\hat{\beta}_T^P$ is a weighted average of β_T (for those with $\beta_T \neq 0$) and 0 (for those with $\beta_T = 0$).

Identification of β_T: The Average Treatment Effect for Those Whom the Treatment is Binding

[8] See Appendix B for details.

[9] As explained later, we call this the binding average treatment effect or BATE. Estimating this effect is new to the treatment effects literature because in that literature the treatment effect is binding for the whole population, albeit to different degrees in the heterogeneous treatment effect case.

[10] More formally, one would run a first-stage regression $x = \gamma_0 + \gamma_1 w + \varsigma$ where w is the instrument. But here $w = x$ for all observations in S_A, $S_B \subset S$. This implies $\gamma_0 = 0$ and $\varsigma = 0$ since x perfectly predicts itself. Thus $\hat{x} = x$.

[11] One should not confuse $\hat{\psi}$ with a placebo effect. A placebo effect results when a respondent seemingly exhibits an impact, but in reality receives *no* treatment. The $\hat{\psi}$ coefficient here measures the difference in outcomes between those treated who have a zero treatment effect and those untreated. As such, $\hat{\psi}$ measures the impact of the confounders. Thus it is an estimate of the bias.

[12] See Appendix D for details.

To identify β_T we construct $\tilde{y}$ following similar steps as before. However, we must up-weight our prior β_T^P estimate to measure the treatment solely for those when the treatment is binding, $\beta_T \neq 0$. Intuitively this means eliminating the zero impact of observations in the S_B group for which the treatment effect is institutionally zero. As explained in Appendix B, we do so by setting $x = 0$ for all S_B observations since the effect of the treatment $\beta_T x$ on $\tilde{y}$ is zero independent of *x*. We denote this new x as $\tilde{x}$, and then we run the following OLS regression for the entire sample (i.e. $\{S_{A,}\, S_B\, , S_C\}$) to obtain $\hat{\beta}_T$

$$\tilde{y} = \pi_0 + \beta_T \tilde{x} + u. \qquad 18$$

As before, *Condition 1*, $E[u|x] = 0$, allows one to consistently estimate $\hat{\beta}_T$, the true treatment effect parameter when the treatment effect is binding.

4.1.B Dichotomous Treatment Variable

Here, the basic methodological intuition remains the same. However, one needs to use a Wald (1940) type estimator to get the biased component θ because in this case the instrument (the dummy variable x in the subsample) is binary.

In a typical binary instrumental variable (say m) case, the IV Wald estimator is defined as

$$\theta = \frac{E[y|m=1] - E[y|m=0]}{E[d|m=1] - E[d|m=0]}. \qquad 19$$

As in the continuous case, one can employ an instrument m that is the treatment itself, which we now denote as d (i.e. $m = d$). In this first stage we examine the effect of d in the $\{S_{A,}\, S_B\}$ subsample where $\beta_T = 0$ or $d = 0$. Substituting $m = d$ in (19) yields

$$\theta = \frac{E[y|d=1] - E[y|d=0]}{E[d|d=1] - E[d|d=0]} = \frac{E[y|d=1] - E[y|d=0]}{1-0}$$
$$= E[y|d=1] - E[y|d=0]. \qquad 20$$

This is equivalent to the OLS coefficient of d (i.e. θ) in the regression

$$y = \beta_0 + \theta d_i + \epsilon' \;\; i \in \{S_{A,}\, S_B\} \;\; \#21$$

run only on the subsamples $S_{A,}$ and S_B where $\beta_T = 0$ or $d = 0$. As above, $\hat{\theta} x$ is subtracted from y to obtain $\tilde{y}$. Once $\tilde{y}$ is obtained one can estimate β_T^P and β_T from the following two OLS regressions with the entire sample (i.e. $\in \{S_{A,}\, S_B\, , S_C\}$).

$$\tilde{y} = \pi_0 + \beta_T^P d + u \quad (22)$$

$$\tilde{y} = \pi_0 + \beta_T \tilde{d} + u \quad (23)$$

where $\tilde{d}$ is a reconstructed version of d such that $\tilde{d} = 0$ for all observations where institutions dictate $\beta_T = 0$.

4.2 Heterogeneous Treatment Effects

The analysis so far assumed the bias and treatment parameters θ, β_T and β_T^P to be constant across all individuals. However, in reality these parameters could be heterogeneous, so that they vary across observations. Appendix B rigorously explains the details, but here we summarize. There, we interpret our estimate of the OLS bias parameter (θ) as well as our treatment effect parameters β_T and β_T^P in light of possible heterogeneity. We prove that our estimate of θ is the average OLS bias in the population as long as the assignment of observations into the $\beta_T = 0$ group does not depend on x. As such, $\hat{\theta}$ measures the average bias effect (ABE). Based on this we show β_T^P measures the average treatment effect (ATE) and $\hat{\beta}_T$ measures the average treatment effect when the treatment is binding. As such, we denote $\hat{\beta}_T$ as the binding average treatment effect (BATE). The β_T^P parameter is similar to the ATE defined in the existing treatment effects literature. However, to our knowledge, β_T is a parameter that is new to the causal effects literature as it measures the effect of the treatment only when the treatment is binding.

4.3 Selectivity

Interestingly, using two samples $\{S_A, S_B\}$ and $\{S_A, S_B, S_C\}$, one of which is a subset of the other, does not lead to a selectivity bias when estimating θ.[13] Unlike in the classic Heckman (1979) selectivity case, where the omitted variable defining selectivity is correlated with the treatment, our omitted variable defining selectivity is uncorrelated with the treatment x because institutions dictate $\beta_T = 0$ (i.e., our selection rule).[14] This institutional circumstance satisfies Condition 2. For this reason, we obtain an unbiased and consistent estimator of θ.[15] As a result, all potential biases due to sample selection goes into the intercept term, leaving the θ slope coefficient estimator unaltered. This result holds true for both the continuous and dichotomous cases. See Appendix E for details.

[13] However, there is a selectivity bias in estimating the constant β_0 in (16) or (21), but this selectivity bias is irrelevant in estimating β_T.

[14] As is shown in Appendix B, one can test whether assignment into the $\beta_T = 0$ group biases the estimator of θ. Obtaining a zero $E(\theta_i - \bar{\theta})$ coefficient when estimating $Eq\ (B.20)$ implies no selectivity bias when we estimate θ.

[15] As illustrated in Appendix E, when $\beta_T = 0, E(v|\beta_T = 0) = c,$ so that $E(Y|X) = (\beta_0 + c) + \theta x$ implying an unbiased and consistent estimate of θ.

4.4 Relating the Bias Excising Approach to the Current Causality Literature

We emphasize that our identification technique differs from the conventional IV approach in two major ways. First, standard IV requires finding a situation in which there are no confounders to bias the causal estimate (Koopman, 1953; Basmann, 1957, 1963; Imbens, 2014). Our approach is the opposite. We use an exclusion restriction that ensures that the instrument correlates with the confounders and thereby correlates with x, but because $\beta_T = 0$ the instrument is uncorrelated with the direct causal effect of x on y. This entails finding a situation in which the treatment has no direct causal effect, but instead only the confounders impact the outcome through the treatment variable. As such, our approach requires finding a circumstance in which the treatment has no causal effect, whereas the traditional approach requires finding a situation in which the treatment yields no confounding effect.

Second, the conventional methods (e.g. DiD, regression discontinuity, propensity score matching, randomized control trials, etc.) model selection into the treatment group to obtain a causal effect (e.g., Heckman 1979, Angrist and Krueger 1991, Card 1995). Similarly, the partial identification literature essentially does the same using data on always takers and never takers to bound the treatment effect (e.g., Manksi 1990; Manski and Pepper 2000; Blanco and Flores-Lagunes 2013 Chen et al., 2018). In contrast, our approach models selection based on a zero treatment effect. As a result of this different modelling strategy, our method is able to retrieve the bias component θ which then can be used to identify the true causal effect.

Third, our approach differs from the control function approach. Whereas both seek to remove a confounding bias, the CFA approach concentrates on a bias originating from an imperfect instrument. In our approach, the treatment itself is the instrument, thus there is no need for exogenous or traditional instrumental variables. This is because our identifying assumption utilizes the endogenous variable itself as an instrument for confounders when an institutional constraint sets the treatment effect zero.

5. An Application

To illustrate consider the literature on sheepskin effects. The question is whether or not actually getting a degree (namely a high school or college diploma) yields a monetary benefit over and above the 12 or 16 years of school required for each respective degree.

Early literature simply ran an OLS regression $y = \alpha S + \beta Degree + \epsilon$ where y is ln earnings, S is schooling, and *Degree* is a categorical dummy variable denoting whether one has a diploma (Jaeger and Page, 1996).[16] Typically β is positive, consistent with a sheepskin effect. We illustrate this result in Table 1 (columns 1 and 6) where the β coefficients are 0.46 and 0.08 respectively. However, *Degree* is not exogenous, as clearly intellectually able students with more stamina, an unobserved confounding factor, are likely to complete school and get a degree; whereas those less able students with less stamina

[16]Earlier literature such as Hungerford and Solon (1987) and Belman and Haywood (1991) do not actually use diploma per se, but instead a spline function with discontinuities denoting diploma levels at junior high, high school, and college. Other literature on sheepskin effects concentrate on particular countries. See Rodríguez and Muro (2015) and Omodunbi (2015) for a survey of such studies.

might not. Alfonso Flores-Lagunes and Audrey Light (2010) point this out by arguing the more able get their degree in fewer years. To illustrate, Flores-Lagunes and Light show the coefficient on schooling is *negative* for those holding a diploma. This negative coefficient means degree holders taking a longer number of years to get a degree are less able. However, the schooling coefficient is positive for those without a degree. This positive coefficient means an additional year of schooling is still valuable. Thus holding years of school constant, those getting a degree are more able. This finding is consistent with an endogeneity problem because ability is unobservable. To get around this hurdle, Damon Clark and Paco Martorell (2014) utilize a regression discontinuity approach to compare subsequent earnings of those students who just passed and just failed the high school exit exams required to get a diploma. Clark and Martorell find little evidence to support a sheepskin effect.

We use our approach to reexamine potential sheepskin effects. We do so first for college diplomas and second for high school diplomas.

We take all those respondents from the NLSY-79 data with 15 to 18 years of school when considering a college or university baccalaureate degree and all respondents with 11 and 12 years of schooling when considering a high school diploma. Some graduate with diplomas, whereas some others go out in the labor force but do not graduate. We run the regression $y_{it} = \beta_0 + \beta Diploma_{it} + \gamma M_{it} + \epsilon_{it}$ whereby *Diploma* denotes those individuals receiving a diploma and *M* denotes years of schooling and other characteristics (See Table 1, footnote #7), and where *i* depicts the individual and *t* the year. According to traditional models the β coefficient (on *x*) denotes a sheepskin effect. However, in reality the β coefficient could instead signify a combination of the sheepskin effect as well as the confounding effects. Typical 2SLS would require an instrument correlated with degree but not the error term. Devising such an instrument is difficult because most such instruments for a sheepskin are also usually correlated with confounders that affect earnings. For example, choosing location-specific unemployment rates as an instrument for receiving a diploma may seem reasonable since a low unemployment rate may cause one to quickly enter the labor force, thus foregoing actually finishing up and receiving a diploma. Yet a low local unemployment rate can put pressure on wages and thus violate the exclusion restriction.

Our new approach requires two steps. The first is to find an instrument correlated with the confounding effects of *x* on *y*, but uncorrelated with the direct causal effect of *x* on *y*. This enables us to estimate the confounding bias θ. We net out the biased effect ($\hat{\theta}x$) in a second set of regressions first to obtain estimates of the population-wide sheepskin effect and second to obtain the sheepskin effect for those whom the sheepskin effect is binding.

One such instrument is motivated by whether or not employers require a high school or college diploma for a particular occupation. If they do not value a college degree, then those employees in this occupation only get a diploma to somehow signal their stamina and hence their ability to stick to it, but the degree itself has no direct causal effect on earnings. As such, because the direct causal effect is zero, the treatment variable itself becomes an instrument for the confounders if used in a regression for individuals in occupations not requiring a degree.

To make use of this instrument we run a regression comprising only those individuals whose employer does not require a diploma.[17] This regression estimates the bias since it measures the impact of the

[17] Information on degree requirements were obtained from the Bureau of Labor Statistics Education and Training.

diploma in an occupation where the diploma is not needed to perform well. Further, there is no difference in learning since we control for schooling. That is, those with a diploma and those without a diploma have a similar amount of school. Thus the impact of a diploma can only result from employers viewing the degree as a reflection of potential employee stick to it type ability. Hence the coefficient is a measure of the bias in the sheepskin effect.

The second set of regressions comprise the same specification, but include the whole sample, with the confounding bias appropriately netted out. These regressions entail employees whose employers value a college degree as a requirement for the job as well as those employees who have a diploma even though it is not needed. For the latter, employers do not value the diploma as a job requirement, but they do value the diploma for the former, thus indicating a diploma has value at least for those employers. These latter regressions measure both the population-wide average treatment effect (ATE) and the effect for the affected population, the binding average treatment effect (BATE) .

Take college diplomas first. Table 1 contains both sets of regressions. But first, as noted earlier, column (1) contains the sheepskin effect based on the past traditional OLS estimation specification. The 0.46 coefficient is biased because it comprises both the true effect plus the endogeneity bias. The 0.42 coefficient in column (2), obtained from the first regression (eq. 11) using the sample of those not requiring a degree, is the biased effect of a diploma since it depicts the percent increase in real wages in occupations where employers do not value a sheepskin per se. As such, the 0.42 effect arises solely because incumbents themselves are more able. Column 3 gives the β_T^P coefficient for all incumbents whether or not they are in a job requiring a diploma. It is obtained from a second regression obtained by netting out the confounding effect $\hat{\theta}x$. The 0.04 population-wide average treatment effect (ATE) implies virtually no sheepskin effect for the whole population. After further netting out those individuals for whom the sheepskin effect is zero based on institutional reasons, we obtain (column 4) the sheepskin effect for those for whom the sheepskin effect can actually be binding. Though, for obvious reasons, the 0.16 β_T coefficient exceeds 0.04 it still remains statistically insignificant. Finally, the virtually zero (0.01) coefficient in column (6) indicates assignment into the zero treatment group does not result in a selectivity bias thereby corroborating the validity of our identification strategy.

Similarly for high school, the OLS coefficient is 0.08 (column 5). The average biased coefficient (ABE) θ is 0.06 (column 6). After appropriately netting out the bias, column 7 gives the 0.02 population-wide sheepskin effect β_T^P. Finally, column 8 contains the 0.03 β_T effect measured for those for whom the sheepskin effect is binding. Again, the sheepskin effect is statistically insignificant. Both the high school and college results are consistent with recent findings of a negligible, if any, sheepskin effect obtained by Flores-Lagunes and Light (2010) and Clark and Martorell (2014).

6. Conclusion

This paper proposes a way a potentially endogenous treatment variable can be used as an instrument for itself. The approach results in two estimators, one yielding a typical population-wide average treatment effect (ATE), and the other yielding a binding average treatment effect (BATE) which to our knowledge is new to the literature. This latter BATE estimator is particularly important when a treatment is ineffective in a subpopulation for institutional reasons. In this case the former conventional

ATE measure includes zero effects. This would lead to underestimates if law or institutional changes render effective the previously ineffective treatment for that segment of the population.

Rather than choosing an instrument correlated with the endogenous variable but not the error, as is conventionally done, we argue one is able to remove an endogeneity bias if one can find an institutional circumstance that for a subpopulation negates the effect of a treatment. For this subpopulation the treatment itself is correlated with the confounding but not the remaining effects. As such, it can be used as an instrument for itself. Indeed, finding such a situation may be even easier than finding a traditional instrument. But in any case, we hope our alternative method to get at endogeneity expands the profession's options for solving important often policy related problems.

We apply the technique to estimating the sheepskin effect, a topic on which there is already a small literature. An OLS regression yields a large effect, but our alternative approach yields no effect. In a sense Clark and Martorell's (2014) similar findings give credence to our results because the regression discontinuity approach they use is often thought to be a "gold standard" since those students that just pass and those students that just fail can be construed as random.

But the technique can potentially be used in many other applications. One such application is to estimate the causal impact of deploying additional police on marijuana use. The number of police deployed depends on multiple factors that are typically unobserved. If these unobserved factors are related to marijuana use, then an OLS regression would yield an inconsistent estimate of the causal effect. However, in some states access to marijuana is legal. In such environments, the number of police should have a zero causal effect on marijuana use since in these states police are irrelevant (in this regard). If in these states there is a correlation between access to marijuana and number of police, then this correlation must be due to the effect of confounders that also correlate with the number of police. In such a situation, OLS regression of marijuana usage on the number of police deployed yields the consistent estimator of θ. So, supposing rich areas hire more police, but at the same time rich areas use more marijuana, then regressing marijuana use on the number of police in areas where marijuana is legal will then yield a positive θ -- the bias. As was explained, this bias can then be used in subsequent regressions to obtain the causal effect population-wide effect (ATE) as well as the causal effect (BATE) in states where the police have an effect.

Another example is childhood development. Cunha and Heckman (2007) identify "critical periods" in which childhood interventions yield beneficial effects. This contrasts with interventions in other periods whereby there are no true effects. Typically parental child investments occur in both time periods, especially given that parents are either oblivious to such critical periods, or instead they want to hedge their bets. In this setting our method can potentially identify the causal effect of intervention by treating effects based on non-critical period interventions as arising from estimation biases. These are but a few examples. We leave to the reader to find others.

Table 1: Biased and unbiased sheepskin effects for college and high school diplomas (heterogeneous treatment effect model)

	College					High school				
	Incumbents in all occupations including those requiring degree	Incumbents in occupation not requiring a degree	Population average treatment effect (ATE)	Binding average treatment effect (BATE)	Average of the residual bias	Incumbents in all occupations including those requiring a diploma	Incumbents in occupation not requiring a diploma	Population average treatment effect (ATE)	Binding average treatment effect (BATE)	Average of the residual bias
	$\hat{\beta} = \hat{\beta}_T^P + \hat{\theta}$	$\hat{\theta}$	$Avg.\hat{\beta}_T^P$	$Avg.\hat{\beta}_T$	$Avg.(\theta_i - \bar{\theta})$	$\hat{\beta} = \hat{\beta}_T^P + \hat{\theta}$	$\hat{\theta}$	$Avg.\hat{\beta}_T^P$	$Avg.\hat{\beta}_T$	$Avg.(\theta_i - \bar{\theta})$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Coeff.	0.46	0.42	0.04	0.16	0.01	0.08	0.06	0.02	0.03	0.02
SE	(0.08)**	(0.08)**	(0.08)	(0.10)	(0.08)	(0.08)	(0.09)	(0.08)	(0.09)	(0.08)
R-sq.	0.15	0.13	0.09	0.09	0.09	0.15	0.18	0.15	0.15	0.15
N	3,299	2,403	3,299	3,299	3,299	4,960	3,191	4,960	4,960	4,960

Source: NLSY79 and BLS Education and Training Data occupational degree/diploma requirements.

Notes:

1. We only consider male workers for the regressions.

2. We run the regression specifications $y_{it} = \beta_0 + \beta \times Diploma_{it} + M_{it}\gamma + u_i t$ twice on the full sample, one for college degree and the other for high school diploma. Columns (1) and (6) represent the coefficients of the diploma variable. These include the biases. Here i represents individual and t represents year ($t = 2006, 2008, 2010, 2012, 2014, 2016$).

3. We also run the regressions specifications $y_{it} = \beta_0 + \theta \times Diploma_{it} + M_{it}\gamma + u_{it}$ twice on the sub samples sample where either occupations do not require diploma or workers do not have diploma. One regression is for college degree and the other for high school diploma. Column (2) and (7) represent the coefficients of the diploma variable. These represent estimates of the biases.

4. To obtain the ATE estimates we run the regression specifications $\tilde{y}_{it} = \beta_0 + [\beta_{Ti} + (\theta_i - \bar{\theta})] \times Diploma_{it} + M_{it}\gamma + u_{it}$ twice on the full sample, one for college degree and the other for high school diploma. Column (3) and (8) represent the coefficients of the diploma variable representing the ATEs.

5. To obtain the BATE estimates we run the regression specifications $\tilde{y}_{it} = \beta_0 + [\beta_{Ti} + (\theta_i - \bar{\theta})] \times \widetilde{Diploma}_{it} + (\theta_i - \bar{\theta}) \times \widetilde{Diploma}_{it} + M_{it}\gamma + u_{it}$ twice on the full sample, one for college degree and the other for high school diploma. Column (4) and (9) represent the coefficients of the diploma variables representing the BATEs.

6. Columns (5) and (10) present the average residual bias or $Avg.(\theta_i - \bar{\theta})$. Both of them are statistically insignificant at 5% level of significance.

7. The vector of control variables M includes occupational fixed effects, year fixed effects, years of schooling (15-18 years for college since some complete schooling and some late; 11-12 years of schooling for the high school graduate), years of work experience and its square, race and ability (AFQT).

References

Angrist, Joshua and Guido Imbens (1995) "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90(430):431-442.

Angrist, Joshua and Alan Krueger (1991) "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*, 106(4): 979-1014.

Angrist, Joshua and Alan Krueger (1994) "Why Do World War II Veterans Earn more than Nonveterans?" *Journal of Labor Economics*, 12(1): 74-97.

Ashenfelter, Orley and Cecilia Rouse (1998) "Income, Schooling, and Ability: Evidence From a New Sample of Identical Twins," *The Quarterly Journal of Economics*, 113(1): 253-284.

Basmann, Robert (1957) "A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation," *Econometrica* 25 (1), 77–83.

Basmann, Robert (1963) "The Causal Interpretation of Non-triangular Systems of Economic Relations," *Econometrica* 31: 439-448.

Basmann, Robert (2006) "Four Entrenched Notions Post Walrasians Should Avoid," in: Colander, D. (ed.), *Post Walrasian Macroeconomics: Beyond the Dynamic Stochastic General Equilibrium Model*. Cambridge University Press, pp. 277–286.

Belman, Dale and John S. Heywood (1991) "Sheepskin Effects in the Returns to Education: An Examination of Women and Minorities," *The Review of Economics and Statistics*, 73(4): 720-724.

Blanco, German, Xuan Chen, Carlos A. Flores, and Alfonso Flores-Lagunes (forthcoming) "Bounds on Average and Quantile Treatment Effects on Duration Outcomes under Censoring, Selection, and Noncompliance," *Journal of Business and Economic Statistics.*

Blanco, German, Carlos A. Flores, and Alfonso Flores-Lagunes (2013) "Bounds on Average and Quantile Treatment Effects of Job Corps Training on Wages," *Journal Human Resources* 48(3): 659-701.

Card, David (1995). "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," In: Christofides, Louis N., Grant E. Kenneth, Robert Swidinsky, eds. *Aspects of Labour Market Behaviour: Essays in Honor of John Vanderkamp*. University of Toronto Press; Toronto: pp. 201–222.

Carneiro, Pedro, and James Heckman (2002) "The Evidence on Credit Constraints in Post-Secondary Schooling," *Economic Journal* 112(482):705–734.

Chen, Xuan, Carlos A. Flores, and Alfonso Flores-Lagunes, (2018) "Going beyond LATE: Bounding Average Treatment Effects of Job Corps Training," *Journal of Human Resources*, 53(4), pages 1050-1099.

Clark, Damon and Paco Martorell (2014) "The Signaling Value of a High School Diploma," *Journal of Political Economy*, 122(2): 282-318.

Cunha, Flavio and James Heckman (2017) "The Technology of Skill Formation," *American Economic Review* 97(2): 31-47.

Flores-Lagunes, Alfonso and Audrey Light (2010) "Interpreting Degree Effects in the Returns to Education," *The Journal of Human Resources* 45(2) 439-467.

Flores, Carlos and Alfonso Flores-Lagunes (2009) "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness," IZA Discussion Paper No. 4237.

Heckman, James (1979) "Sample Selection Bias as a Specification Error," *Econometrica* 47(1): 153-161.

Heckman, James and Richard Robb (1985) "Alternative Methods for Evaluating the Impact of Interventions," in James Heckman and Burton Singer (eds.) *Longitudinal Analysis of Labor Market Data* (Cambridge: Cambridge University Press), 156-245.

Hungerford, Thomas and Gary Solon (1987) "Sheepskin Effects in the Returns to Education," *The Review of Economics and Statistics*, 69(1): 175-177.

Imbens, Guido (2014) "Instrumental Variables: An Econometrician's Perspective," *Statistical Science* 29(3): 323-358.

Imbens, Guido and Jeffrey M. Wooldridge (2007) "Control Functions and Related methods, Lecture 6 of the What's New in Econometrics NBER Summer Institute" http://www.nber.org/WNE/Slides7-31-07/Slides7-31-07.pdf.

Jaeger, David and Marianne E. Page (1996) "New Evidence on Sheepskin Effects in the Returns to Education," *The Review of Economics and Statistics*, 78(4): 733-740.

Koopmans T. (1953) "Identification Problems in Econometric Model Construction," In: Hood W., T. Koopmans, eds. *Studies in Econometric Method*. New York: Wiley; pp. 27–48.

Manski, Charles (1990) "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings* 80: 319-323.

Manski, Charles and John Pepper (2000) "Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(4): 997-1010.

Moore, Thomas G. (1971) "The Effect of Minimum Wages on Teenage Unemployment Rates," *Journal of Political Economy*, 79(4): 897-902.

Omodunbi, Oluwaropo (2015) "Investigations into Returns to Education in Sub Saharan Africa," Ph.D. Dissertation, State University of New York at Binghamton.

Rodríguez Jhon James Mora and Juan Muro (2015) "On the Size of Sheepskin Effects: A Meta-Analysis," *Meta-Analysis in Theory and Practice*, http://www.economics-ejournal.org/economics/discussionpapers/2015-31

Sargan, J. (1958) Estimation of Economic Relationships Using Instrumental Variables. *Econometrica* 67: 557–86.

Taubman, Paul (1976) "Earnings, Education, Genetics, and Environment," *The Journal of Human Resources*, 11(4): 447-461.

Theil, H. 1958. *Economic Forecasts and Policy,* Amsterdam: North-Holland.

Wald, A. (1940) "The Fitting of Straight Lines if Both Variables are Subject to Error." *The Annals of Mathematical Statistics* 11(3): 284-300.

Appendix A: Error Structure when the Treatment is Endogenous

Proposition A1: *If the* $Corr(x, \epsilon) \neq 0$, *then* $\exists$ *a unique* $(a,\ \theta) \in \mathbb{R}^2$ *and a noise vector* u *with mean 0 and standard deviation* σ_u *such that* $\epsilon = a + \theta x + u$.

Proof: Let the $Corr(x, \epsilon) = c$ such that $-1 \leq c \leq 1$. Then one can express the correlation as

$$c = \frac{Cov(x, \epsilon)}{\sigma_x \sigma_\epsilon} \qquad A.1$$

$$Cov(x, \epsilon) = c\sigma_x \sigma_\epsilon$$

Case 1: When the correlation between ϵ and x is perfect, that is $Corr(x, \epsilon) = c = 1$.

In this case the relationship between ϵ and x is linear which can be presented as follows

$$\epsilon = a + \theta x \qquad A.2$$

where a and θ are two scalars such that $\mu_\epsilon = a + \theta \mu_x$; and μ_ϵ and μ_x are the means of ϵ and x respectively. Thus, the covariance can be expressed as

$$Cov(x, \epsilon) = \sigma_x \sigma_\epsilon = Cov(a, x) + \theta Cov(x, x) = \theta \sigma_x^2 \qquad A.3$$

where $Cov(a, X) = 0$ since a is a constant. This yields the solution of $\theta = \frac{\sigma_\epsilon}{\sigma_x}$. Based on this, one can then compute $a = \mu_\epsilon - \frac{\sigma_\epsilon}{\sigma_x}\mu_x$ from $(A.2)$. Thus, there exists a unique $\theta = \frac{\sigma_\epsilon}{\sigma_x}$ such ϵ can be expressed as $\epsilon = a + \theta x$.

Case 2: When the correlation between ϵ and x is imperfect, that is $Corr(x, \epsilon) = c$ such that $|c| < 1$.

In this case let the following expression represent the relationship between ϵ and x

$$\epsilon = a + \theta x + u \qquad A.4$$

where u is a noise vector with 0 mean and variance σ_u^2. Additionally, $u \perp x$, that is $Corr(x, u) = Cov(x, u) = 0$. We need to show that there exists an unique $(a^*, \theta^*) \in \mathbb{R}^2$ such that $\epsilon = a + \theta x + u$

The covariance of $\epsilon\ and\ x$ can be written as

$$Cov(x, \epsilon) = c\sigma_x \sigma_\epsilon = Cov(a, x) + \theta^* Cov(x, x) + Cov(x, u) = \theta^* \sigma_x^2 \qquad A.5$$

Suppose these exists another distinct $(\tilde{a}, \tilde{\theta}) \in \mathbb{R}^2$such that the above equation is also satisfied. That is

$$Cov(x, \epsilon) = c\sigma_x\sigma_\epsilon = \tilde{\theta}\sigma_x^2$$

$$\tilde{\theta} = c\frac{\sigma_\epsilon}{\sigma_x} \qquad A.6$$

Because $\sigma_x, \sigma_\epsilon$ are unique constant scalars, $\theta^* = \tilde{\theta}$. Once θ is uniquely identified, one can show that a^* is also unique. Thus, there $\exists$ a unique $(a^*, \theta^*) \in \mathbb{R}^2$ such that ϵ can be expressed as

$\epsilon = a + \theta x + u$ …Q.E.D.

Appendix B: The IV Used to Identify the Bias Compared to the Traditional IV Used to Identify the Exogenous Variable

Conventional 2SLS-IV

Given the regression equation

$$y = \beta_0 + \beta_T x + \epsilon \qquad B.1$$

The conventional 2SLS-IV (instrument: z) requires the following two conditions be satisfied

$$Cov(x, z) \neq 0$$
$$Cov(\epsilon, z) = 0$$

Taking the covariance of (B.1) with respect to z yields

$$Cov(y, z) = \beta_T Cov(x, z) + Cov(\epsilon, z) \qquad B.2$$

Because the instrument is uncorrelated with ϵ, that is $Cov(\epsilon, z) = 0$, one can compute β_T as

$$\hat{\beta}_T = \frac{s_{yz}}{s_{xz}} \qquad B.3$$

where s_{yz} and s_{xz} represent sample covariances between (y, z) and (x, z). Given that the sample covariances are consistent estimators of the population covariances, one can express $plim\ \hat{\beta}_T$ as

$$plim\ \hat{\beta}_T = \frac{Cov(y, z)}{Cov(x, z)}.$$

Identifying the Bias

In our alternative approach we express the regression equation as

$$y = \beta_0 + \theta x + \epsilon' \qquad B.4$$

where $\epsilon' = a + \beta_T x + u$. Based on $(B.4)$, identification of θ with our alternative instrument (w) requires

$$Cov(x, w) \neq 0$$

and

$$Cov(\epsilon', w) = 0$$

Finding an instrument w that meets the latter covariance condition is difficult because ϵ' contains x which according to the first condition is correlated with w. However, both these covariance conditions can hold simultaneously if β_T or x equals zero and $Cov(u, w) = 0$. If one can find a subpopulation which for institutional reasons $\beta_T = 0$ or $x = 0$ and u and w are uncorrelated, then the condition $Cov(\epsilon', w) = 0$ is satisfied and w can be used as an instrument.

For this subpopulation defined in the text as $\{S_A, S_B\}$ one can obtain the estimator of θ by taking the covariance of with respect to w on the both sides

$$Cov(y, w) = \theta Cov(x, w) + Cov(\epsilon', w). \qquad B.5$$

If $Cov(\epsilon', w) = 0$, one can compute θ as

$$\hat{\theta} = \frac{s_{yw}}{s_{xw}} \qquad B.6$$

where s_{yw} and s_{xw} represent sample covariances between (y, w) and (x, w). If these sample covariances are consistent estimators of the population covariances, OLS yields a consistent estimator of θ, that is

$$plim\ \hat{\theta} = \frac{Cov(y, w)}{Cov(x, w)}. \qquad B.7$$

The variable x can be used as the instrument for x itself in the subsample $\{S_A, S_B\}$. In subsample $\{S_A, S_B\}$, $\beta_T = 0$ and x is uncorrelated with u by Proposition A.1. The instrument x in the $\{S_A, S_B\}$ subsample satisfies the above covariance conditions ($Cov(x, x) \neq 0$ and $Cov(x, \epsilon') = 0$). This allows one to estimate $\hat{\theta}$ by the following formula

$$\widehat{\theta} = \frac{s_{yx}}{s_{xx}}$$

Appendix E shows that estimating θ from $\{S_A, S_B\}$ rather than from $\{S_A, S_B, S_C\}$ does not result in a selectivity bias.

Identification of the treatment effects (β_T^P, β_T)

Treatment effects β_T^P and β_T are obtained by running a second regression on the whole sample $\{S_A, S_B, S_C\}$ in which $\hat{\theta}x$, the confounding effect, is removed from the outcome data for each observation in the entire sample $\{S_A, S_B, S_C\}$. Thus outcome y is transformed as follows

$$\tilde{y} = y - \hat{\theta}x \qquad B.8$$

B.1. The homogeneous treatment effect case

B.1.a. Continuous treatment variable

Identification of β_T^P

Given $\tilde{y}$, the population average treatment effect β_T^P is obtained from the following regression

$$\tilde{y} = \pi_0 + \beta_T^P x + u. \qquad B.9$$

β_T^P is the weighted average of the treatment effects in the subsample where $\beta_T \neq 0$ as well as the treatment effects in the subsample where $\beta_T = 0$. *Condition* 1 (i.e. $E[u|x] = E[u] = 0$) ensures that $Cov[x, u] = 0$ implying that the OLS estimator $\hat{\beta}_T^P$ is unbiased and consistent (See Appendix D for details).

Identification of β_T

As in Section 4 we partition the data into $S_A, S_B,$ and S_C. Subsample S_A contains observations in which either there is no treatment or in which the treatment effect is zero. Assume the latter has n_4 observations and the former n_2 observations. S_B contains observations in which the treatment effect is zero for institutional reasons. Assume there are n_3 of these observations. Finally S_C contains observations with a true and confounded treatment effect. We assume S_C contains n_1 observations. We represent the data structure as

$$\begin{aligned} n_1&: \tilde{y}_1 = \pi_0 + \beta_T \times x_1 + u_1 \\ n_2&: \tilde{y}_2 = \pi_0 + \beta_T \times 0 + u_2 \\ n_3&: \tilde{y}_3 = \pi_0 + 0 \times x_3 + u_3 \end{aligned} \qquad B.10$$

$$n_4: \tilde{y}_4 = \pi_0 + 0 \times 0 + u_4.$$

This is the original data structure but now rewritten to note explicitly the observations for which $\beta_T = 0$.

An algebraic manipulation of the B.10 yields

$$\begin{aligned} n_1&: \tilde{y}_1 = \pi_0 + \beta_T \times x_1 + 0 \times 0 + u_1 \\ n_2&: \tilde{y}_2 = \pi_0 + \beta_T \times 0 + 0 \times 0 + u_2 \\ n_3&: \tilde{y}_3 = \pi_0 + \beta_T \times 0 + 0 \times x_3 + u_3 \\ n_4&: \tilde{y}_4 = \pi_0 + \beta_T \times 0 + 0 \times 0 + u_4 \end{aligned}$$

or

$$\tilde{y} = \pi_0 + \beta_T \tilde{x} + 0 \times \breve{x} + u \qquad B.11$$

where

$$\tilde{x} = \begin{bmatrix} x_3 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } \breve{x} = \begin{bmatrix} 0 \\ 0 \\ x_3 \\ 0 \end{bmatrix}.$$

Because $0 \times \breve{x} = 0$ one can rewrite (B.11) as the estimable function that identifies β_T

$$\tilde{y} = \pi_0 + \beta_T \tilde{x} + u. \qquad B.12$$

B.12 transforms the *x* matrix so that each observation gets the same treatment effect.

B.1.b. Dichotomous Treatment Variable

Here, the methodological intuition remains the same. One simply substitutes the continuous variable x by the dichotomous treatment variable d. Under similar assumptions, similar subsamples, and the following estimation procedure we identify average of θ_i or ABE as

$$\theta = \frac{E[y|d_i = 1] - E[y_i|d_i = 0]}{E[d_i|d_i = 1] - E[d_i|d_i = 0]} = \frac{E[y_i|d_i = 1] - E[y_i|d_i) = 0]}{1 - 0}$$

Based on this $\hat{\theta}$, we construct $\tilde{y}_i = y_i - \hat{\theta} d_i$. Then utilizing $\tilde{y}$ and applying a parallel estimation methodology described in continuous treatment case, one obtains the estimators of average β_T^P (i.e. ATE) and average β_T (i.e. BATE).

B.2. The heterogeneous treatment effect case

Notation

We denote subscript i to depict the i-th observation in the data. Heterogeneous treatment effects imply β_T is subscripted yielding

$$y_i = \beta_0 + \beta_{Ti} x_i + \epsilon_i \qquad B.13$$

the counterpart to the homogeneous treatment effect case given in (1). As before we partition the data into S_A, S_B, and S_C. Recall observations in S_C receive treatment ($x \neq 0$) and the treatment is effective ($\beta_T \neq 0$). We assume there are n_1 observations inS_C. Subsample S_B consists of all treated observations for which, given institutional considerations, the treatment has a zero effect

($\beta_T = 0$ and $x \neq 0$). We assume n_3 observations in S_B. Finally, S_A consists of all observations for which there is no treatment $(x = 0)$. In theory these can be observations for which a treatment would or would not have an impact if actually treated. We assume this group has $n_2 + n_4$ observations, n_2 of which would have an effect and n_4 of which would not have a treatment effect, again for institutional reasons. Based on this notation we interpret the parameters for the continuous and discrete cases.

B.2.a. The Continuous Treatment Variable

Identification of Average Bias Effect

Based on this classification scheme, the subsample for which $\beta_{Ti} = 0$ is determined institutionally and can be represented under the heterogeneous paradigm as

$$n_3: y_{i_3} = \pi_0 + \theta_{i_3} x_{i_3} + u_{i_3}$$

$$n_4: y_{i_4} = \pi_0 + \theta_{i_4} x_{i_4} + u_{i_4}$$

alternatively expressed as

$$y_i = \pi_0 + \theta_i x_i + u_i \qquad B.14$$

where $i \in \{i_3, i_4\}$.

As already shown, *Condition* 1 and *Condition* 2 allows one to consistently estimate θ in the homogeneous treatment effect case. The mean independence in *Condition* 1 and $Cov[x, u] = 0$ ensures that x is an instrument for itself. Thus, the OLS regression with the third and fourth subsamples (with $[n_3 + n_4]$ observations) yields a consistent estimator of the average of θ ($\hat{\theta}$) for $i = i_3$ and i_4, that is, $plim\ \hat{\theta} = \bar{\theta}$. If *Condition* 2 holds, the OLS estimator $\hat{\theta}$ from subsample i_3 and i_4 is equal to the average of θ (i.e. $\bar{\theta}$) for the entire population i.e. for $i = i_1\ or\ i_2\ or\ i_3\ or\ i_4$. We call this the average bias effect or ABE.

Identification of the Population Average Treatment Effect (β_T^P)

To identify β_{Ti}^P we subtract $\hat{\theta} x_i$ from y_i to obtain $\tilde{y}_i$

$$\tilde{y}_i = y_i - \hat{\theta} x_i \qquad B.15$$

where $i \in \{i_1, i_2, i_3, i_4\}$.

Thus (B.15) can be expressed as

$$\tilde{y}_i = \pi_0 + \beta_{Ti} x_i + (\theta_i - \bar{\theta}) x_i + u_i \qquad B.16$$

Assuming that β_{Ti} and θ_i are independent of x_i, one can express the conditional expectations as

$$E[\tilde{y}_i | x_i] = \pi_0 + x_i E[\beta_{Ti} + (\theta_i - \bar{\theta})] + E[u_i | x_i]$$

Because *Condition 1* implies that $E[u_i | x_i] = 0$, the above expression can be rewritten as

$$E[\tilde{y}_i | x_i] = \pi_0 + E[\beta_{Ti} + (\theta_i - \bar{\theta})] x_i \ = \pi_0 + E[\beta_{Ti}] x_i + E\left[\theta_i - \hat{\theta}\right] x_i \qquad B.17$$

Given $plim\ \hat{\theta} = E[\theta_i] = \bar{\theta}$, so that $E\left[\theta_i - \hat{\theta}\right] = 0$, (B17) reduces to

$$E[\tilde{y}_i | x_i] = \pi_0 + x_i E[\beta_{Ti}]. \qquad B.18$$

The OLS regression of y on x utilizing the entire sample $n_{1,}, n_2, n_3,$ and n_4 yields an average of β_{Ti}, but this average comprises those with $\beta_{Ti} \neq 0$ and those with $\beta_{T_i} = 0$ for institutional reasons. Thus, $E[\beta_{Ti}]$ represents the ATE including observations for which β_{T_i} takes a zero value. We denote $E\left[\beta_{T_i}\right]\ as\ \beta_T^P$ since it is the population average treatment effect.

Identification of the Binding Average Treatment Effect (β_{T_i})

To get the treatment when the treatment is actually effective, that is binding, we rearrange the data matrix as follows

$$n_1 : \widetilde{y_{i_1}} = \pi_0 + [\beta_{Ti_1} + \left(\theta_{i_1} - \bar{\theta}\right)] \times x_{i_1} + u_{i_1}$$

$$n_2 : \widetilde{y_{i_2}} = \pi_0 + [\beta_{Ti_2} + \left(\theta_{i_2} - \bar{\theta}\right)] \times x_{i_2} + u_{i_2}$$

$$n_3 : \widetilde{y_{i_3}} = \pi_0 + [\beta_{Ti_3} + \left(\theta_{i_3} - \bar{\theta}\right)] \times x_{i_3} + u_{i_3}$$

$$n_4 : \widetilde{y_{i_4}} = \pi_0 + [\beta_{Ti_4} + \left(\theta_{i_4} - \bar{\theta}\right)] \times x_{i_4} + u_{i_4} \qquad \text{B.19}$$

Recalling that $x_{i_2} = x_{i_4} = 0$ and $\beta_{Ti_3} = \ \beta_{Ti_4} = 0$, yields

$$n_1 : \widetilde{y_{i_1}} = \pi_0 + [\beta_{Ti_1} + \left(\theta_{i_1} - \bar{\theta}\right)] \times x_{i_1} + u_{i_1}$$

$$n_2 : \widetilde{y_{i_2}} = \pi_0 + [\beta_{Ti_2} + \left(\theta_{i_2} - \bar{\theta}\right)] \times 0 + u_{i_2}$$

$$n_3: \widetilde{y_{i_3}} = \pi_0 + \quad 0 \quad + \left(\theta_{i_3} - \bar{\theta}\right) \times x_{i_3} + u_{i_3}$$

$$n_4: \widetilde{y_{i_4}} = \pi_0 + \quad 0 \quad + \left(\theta_{i_4} - \bar{\theta}\right) \times 0 + u_{i_4}$$

Rearranging yields

$$n_1: \widetilde{y_{i_1}} = \pi_0 + [\beta_{Ti_1} + \left(\theta_{i_1} - \bar{\theta}\right)] \times x_{i_1} + \left(\theta_{i_1} - \bar{\theta}\right) \times 0 + u_{i_1}$$

$$n_2: \widetilde{y_{i_2}} = \pi_0 + [\beta_{Ti_{1_2}} + \left(\theta_{i_2} - \bar{\theta}\right)] \times 0 + \left(\theta_{i_2} - \bar{\theta}\right) \times 0 + u_{i_2}$$

$$n_3: \widetilde{y_{i_3}} = \pi_0 + \left[\beta_{Ti_3} + \left(\theta_{i_3} - \bar{\theta}\right)\right] \times 0 + \left(\theta_{i_3} - \bar{\theta}\right) \times x_{i_3} + u_{i_3}$$

$$n_4: \widetilde{y_{i_4}} = \pi_0 + [\beta_{Ti_4} + \left(\theta_{i_4} - \bar{\theta}\right)] \times 0 + \left(\theta_{i_4} - \bar{\theta}\right)] \times 0 + u_{i_4}$$

which in condensed form can be written

$$\widetilde{y_i} = \pi_0 + [\beta_{Ti} + (\theta_i - \bar{\theta})]\tilde{x}_i + (\theta_i - \bar{\theta})\breve{x}_i + u_i \qquad B.20$$

where

$$\tilde{x} = \begin{bmatrix} x_{i_1} \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } \breve{x} = \begin{bmatrix} 0 \\ 0 \\ x_{i_3} \\ 0 \end{bmatrix}.$$

Because (B.20) is simply a rearrangement of (B.19), OLS estimation yields a consistent estimator of the average $[\beta_{Ti} + (\theta_i - \bar{\theta})]$. However, E$[\beta_{Ti} + (\theta_i - \bar{\theta})]$ = E$[\beta_{Ti}]$ because. E$(\theta_i - \bar{\theta}) = 0$. Thus, the OLS estimator of the $\tilde{x}$ coefficient in regression (B.19) represents the average of β_{T_i}, namely the average treatment effect when the treatment is binding, what we call the binding average treatment effect, BATE. In estimating B.20, the $(\theta_i - \bar{\theta})$ should be zero. A non-zero coefficient would imply a non-random selection of observations into the zero-treatment effect group, S_B.

B.2.b. Dichotomous Treatment Variable

The methodological intuition remains the same for a dichotomous treatment. One replaces the continuous variable x_i by the dichotomous treatment variable d_i. Under similar assumptions, similar subsamples, we identify the average of θ_i or ABE as

$$\theta = \frac{E[y_i|d_i = 1] - E[y_i|d_i = 0]}{E[d_i|d_i = 1] - E[d_i|d_i = 0]} = \frac{E[y_i|d_i = 1] - E[y_i|d_i) = 0]}{1 - 0}$$

Based on this $\hat{\theta}$, we construct $\tilde{y}_i = y_i - \hat{\theta} d_i$. Then utilizing $\tilde{y}$ and applying a parallel estimation methodology described for continuous treatment case, we obtain the estimators of the population

average treatment effect β_T^P (i.e. ATE) and average treatment effect when the treatment is binding β_T (i.e. BATE).

Appendix C: Relationship to the Control Function Approach

As stated in the text, our approach appears closest to the control function approach (CFA). However, there is a significant feature in our approach that differentiates it from the CFA approach. In a linear in parameter model, the control function approach requires exogenous variables or instrumental variables to construct the control function (Wooldridge 2007)[18]. The control function then can be included in the main regression as an explanatory variable to adjust for the endogeneity bias. In contrast, our approach does not require exogenous variables that are determined outside the model. Instead, our identifying assumption converts the endogenous variable into an exogenous variable within a representative subsample allowing for the estimation of the bias. To see this difference, consider the following regression equation.

$$y = \beta_0 + \beta_T x + \gamma w + \epsilon$$

The endogeneity problem arises because $Cov[x, \epsilon] \neq 0$. In this case the control function approach would require finding instruments or exogenous variables (z, w) such that $Cov[x, z] \neq 0$ and $Cov[z, \epsilon] = 0$, $Cov[w, \epsilon] = 0$. With these instruments or exogenous variables one can then construct the control function in the following manner

Regress

$$x = \alpha_0 + \alpha_z z + \alpha_w w + v$$

Given that z and w are uncorrelated with ϵ, any non-zero correlation between x and ϵ must be due to non-zero correlation between v and ϵ. Thus, the residual $\hat{v}$ from the above regression constitutes the control function. Inclusion of this in the main regression will now consistently estimate β_T, i.e.

$$y = \beta_0 + \beta_T x + \gamma w + \rho\, \hat{v} + u$$

where u is the new error term such that $Cov[x, u] = 0$. The key to this identification is the availability of an instruments or exogenous variables w. Without availability of instrumental variables or exogenous determinants of the outcome this method cannot identify the causal effects.

In contrast, our identifying condition identifies a representative subset of population where $\beta_T = 0$. This automatically makes x an exogenous variable for this subset of the population, so that the regression equation is

$$y = \beta_0 + \epsilon$$

Thus, if x is suspected to be endogenous, then it must be correlated with ϵ. Thus, a regression in a subsample with $\beta_T = 0$ yields the endogeneity bias, that is θ in the following regression

[18] See Imbens and Wooldridge (2007) at http://www.nber.org/WNE/Slides7-31-07/Slides7-31-07.pdf

$$y = \beta_0 + \theta x + u$$

Here x can be seen as an instrument for itself to identify θ. Unlike CFA, no additional instrument or exogenous variables are needed for the identification of θ. Moreover, since x serves as an instrument for itself, it identifies the population average bias θ and not the local average that a typical instrumental variable estimation would obtain.

Appendix D: Unbiasedness and Consistency of $\hat{\theta}$, $\hat{\beta}_T^P$ and $\hat{\beta}_T$

Proof:

Case 1: homogeneous treatment and bias effect

Unbiasedness and consistency of $\hat{\theta}$

The regression equation is

$$y = \beta_0 + \theta x + \epsilon'$$

where $\epsilon' = a + \beta_T x + u$. When the regression is run only for the subsamples where $\beta_T = 0$, $Cov[x, \epsilon'] = 0$. In the absence of the selectivity problem, this zero covariance condition ensures that $\hat{\theta}$ is an unbiased and consistent estimator of θ.

Unbiasedness of $\hat{\beta}_T^P$ and $\hat{\beta}_T$

Unbiasedness implies that

$E(\hat{\beta}_T^P) = \beta_T^P$ and $E(\hat{\beta}_T) = \beta_T$

As stated earlier, the estimator of $\hat{\beta}_T^P$ emerges from the following regression equation

$$\tilde{y} = \pi_0 + \beta_T^P x + u$$

$$\tilde{y} = M\gamma + u$$

where

$M = \begin{bmatrix} 1 & x_1 \\ 1 & 0 \\ 1 & x_3 \\ 1 & 0 \end{bmatrix}$ and $\gamma = \begin{bmatrix} \hat{\pi}_0 \\ \hat{\beta}_T^P \end{bmatrix}$

The OLS estimator $\hat{\gamma}$ is estimated as

$$\hat{\gamma} = (M'M)^{-1}M'\tilde{y} \qquad C.1$$

Substituting $y = M\gamma + u$ in (C.1) yields

$$\hat{\gamma} = (M'M)^{-1}M'M\gamma + (M'M)^{-1}M'u$$

Taking conditional expectations

$$E[\hat{\gamma}|M] = \gamma + (M'M)^{-1}M'E[u|M]$$

Because $E[u|M] = 0$ by *condition 1,* the conditional expectations above equals $E[\hat{\gamma}|M] = \gamma$.

Again, by the law of iterated expectations $E_M[E[\hat{\gamma}|M]] = E_M(\gamma)$. In other words $E[\hat{\gamma}] = \gamma$

That is $E\left[\hat{\beta}_T^P\right] = \beta_T^P$ (Q.E.D.)

Unbiasedness of $\hat{\beta}_T$

The regression equation in this case is

$$\tilde{y} = \pi_0 + \beta_T\, \tilde{x} + u$$

Following the same steps it can be shows that $E\left[\hat{\beta}_T\right] = \beta_T$ as long as $E[u|\tilde{x}] = 0$. As already stated above, $\tilde{x}$ is transformed based on the observations in which $\beta_T = 0$ for institutional reasons. Therefore, $\tilde{x}$ will also be mean independent of u since x is mean independent of u by *Condition 1*. Thus, $E\left[\hat{\beta}_T\right] = \beta_T$.

Consistency of of $\hat{\beta}_T^P$ *and* $\hat{\beta}_T$

Consistency requires that $plim\, \hat{\beta}_T^P = \beta_T^P$ and $plim\, \hat{\beta}_T = \beta_T$. The primary requirement for this condition is that the independent variables are uncorrelated to u. *Condition* 1 and *Condition* 2 jointly satisfy the mean independence conditions for both x and $\tilde{x}$. Thus both the treatments effects are consistent.

The same logic applies to the heterogeneous treatment effect cases. Thus, the estimators $\hat{\beta}_T^P$ and $\hat{\beta}_T$ are also consistent.

Simulation exercises (available upon request) show parameter estimates converge to the true values of θ, β_T^P and β_T as sample size increases. Also, as expected, the estimates are more precise as the sample size of the $\beta_T = 0$ bias identifying sample increases.

Appendix E: Sample Selection and the Unbiasedness and Consistency of θ

We examine two cases: (1) when x is continuous, and (2) when x is binary.

1. *Continuous x*

Proposition 1: When selection into $\beta_T = 0$ does not depend on x, omitting S_C from the sample to estimate θ causes a selection bias in the constant term, but does not bias the estimator of θ.

Proof: Ideally obtaining an unbiased and consistent estimator of θ entails estimating

$$y = (\beta_0 + a) + \theta x + u \qquad D.1$$

when $\beta_T = 0$, for all observations in$S_A, S_B,$ and S_C. However, $\beta_T \neq 0$ for observations in S_C. Thus we estimate D.1 using subsample $\{S_A, S_B\}$. Given that the selection rule $\beta_T = 0$ does not depend on x, one can express the regression equation as

$$E[y|\, \beta_T = 0, x] = (\beta_0 + a) + \theta x + E[u|\beta_T = 0, x]. \qquad D.2$$

Since u and x are statistically mean independent and x and β_T are independent, then $E[u|\beta_T = 0, x] = E\big[[u|\beta_T = 0]\big]$ yielding

$$E[y|\, \beta_T = 0, x] = (\beta_0 + a) + \theta x + E[u|\beta_T = 0]. \qquad D.3$$

Given that the selection rule $\beta_T = 0$ does not depend on x, the omitted term due to sample selection is $E[u|\beta_T = 0]$, a constant. Let $E[u|\beta_T = 0] = c$ to denote this constant term. Substituting $E[u|\beta_T = 0] = c$ into D.3 yields

$$E[y|x, \beta_T = 0] = (\beta_0 + a + c) + \theta x. \qquad D.4$$

Thus, the sample selection causes a bias in the intercept only when $c \neq 0$. It does not affect θ.

Proposition 2: When for some observations in $\{S_A, S_B\}$, selection into the $\beta_T = 0$ group occurs if and only if x=0, then omitting S_C from the sample to estimate θ yields a consistent and unbiased estimator if β_T and x are mean independent of u.

This selection rule gives rise to two distinct subsamples. In the first subsample selection into $\beta_T = 0$ does not depend on x. For these entities the regression equation is as before

$$E[y|\,\beta_T = 0, x] = (\beta_0 + a) + \theta x + E[u|\beta_T = 0]. \qquad D.3$$

However, for the second group where $\beta_T = 0$ only because $x = 0$, the regression equation is

$$E[y|\,\beta_T = 0, x = 0] = (\beta_0 + a) + \theta x + E[u|\beta_T = 0, x = 0] \qquad D.5$$

Let $E[u|\beta_T = 0, x = 0] = c_1$. The regression equation is

$$E[y|\,\beta_T = 0, x = 0] = (\beta_0 + a) + c_1 + \theta x$$

Bringing both subsamples together yields

For Sample 1: $E[y|\,\beta_T = 0, x] = (\beta_0 + a + c) + \theta x$

For sample 2: $E[y|\,\beta_T = 0, x = 0] = (\beta_0 + a + c_1) + \theta x$

However, in this case, identification of θ requires that u be independent of β_T and x, that is $E[u|\beta_T = 0, x = 0] = c_1 = 0$ and $E[u|\beta_T = 0] = c = 0$. This is a stronger assumption than selection independent of x. However, the selection rule can depend on the observables or unobservables that are independent of x and u.

Binary x: the treatment effect case

The proof of unbiasedness and consistency of θ is similar to the continuous variable case.