Supplemental Information for Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects

Avidit Acharya,* Matthew Blackwell,† and Maya Sen‡

February 16, 2016

A Completely mediated effects and the CDE

VanderWeele (2011) stated but did not formally prove that if M and W completely mediate the effect of A on Y, then any controlled direct effect of A on Y with $M = m$, $ACDE(m)$, must be due to an indirect effect of A on Y through W. First, let us define what complete mediation entails.

Definition 1 (Complete mediation)**.** *A set of variables* $Z = \{Z_1, \ldots, Z_k\}$ completely mediates *the effect of* A *on* Y *if, for all values* $\{z_1, \ldots, z_k\} \in \mathcal{Z}$*,*

$$Y_i(a, z_1, \ldots, z_k) = Y_i(z_1, \ldots, z_k),$$

where $\mathcal{Z}$ *is the support of* Z*.*

Complete mediation is a common idea in the social sciences, where it is mostly seen in an instrumental variable design. There, the exclusion restriction assumes that the effect of

*Assistant Professor of Political Science, Stanford University. email: avidit@stanford.edu, web: http://www.stanford.edu/~avidit.

†Assistant Professor of Government, Harvard University. email: mblackwell@gov.harvard.edu, web: http://www.mattblackwell.org.

‡Assistant Professor of Public Policy, Harvard University. email: maya_sen@hks.harvard.edu, web: http://scholar.harvard.edu/msen.

the instrument on the outcome is completely mediated by the treatment. Philosophically, it is unclear if all effects are possibly completely mediated by some set of variables or whether there are effects for which no completely mediating set exists. We do not answer this question. We assume that such a set exists and that we can partition it into two subsets: M the potential mediator we wish to test and W, the set of all other mediators. The goal of this analysis is to show that there is some effect that is not through M—that is, that there is some indirect effect through W. Because W is possibly multivariate with unobserved components, it may not be possible to determine what part of W mediates the effect. Thus, this approach represents a falsification test of sorts.

Proposition 1. *If the effect of A on Y is completely mediated by $Z = M, W$ and the consistency assumption holds for all potential outcomes, then the average controlled direct effect with M fixed is also an indirect effect of W:*

$$E[Y_i(a, m) - Y_i(a', m)] = E[Y_i(a, m, W_i(a, m)) - Y_i(a, m, W_i(a', m))].$$

Proof.

$$\begin{aligned} E[Y_i(a, m) - Y_i(a', m)] &= E[Y_i(a, m, W_i(a, m)) - Y_i(a', m, W_i(a', m))] \\ &= E[Y_i(a, m, W_i(a, m)) - Y_i(a, m, W_i(a', m)) \\ &\quad + Y_i(a, m, W_i(a', m)) - Y_i(a', m, W_i(a', m))] \\ &= E[Y_i(a, m, W_i(a, m)) - Y_i(a, m, W_i(a', m))] \end{aligned}$$

The first equality follows from the consistency assumption and the last equality follows from the definition of complete mediation. □

Note that this is a type of natural indirect effect—natural because the other mediators, W, are allowed to take their natural value under a, a', and m. Of course this indirect effect fixes the value of M to m, so that it ignores any potential interaction between the indirect

effect size and the natural value of M. It is also for this reason that this result holds whether M affects W, W affects M, or they are independent.

B Consistent Variance estimation for the ACDE in linear blip-down estimation

Let W_i be the $1 \times k$ vector of variables in the first stage of the blip-down estimation. In the paper above, we took this to include A_i, M_i, X_i, and Z_i, but here we will be more general so as to allow interactions and possible non-nesting of the first and second stages. Let V_i be the $1 \times p$ vector of variables in the second stage, direct effect model. Obviously, this includes A_i, but might also include baseline covariates (and interactions with baseline covariates) as well. Let $M_i \subset W_i$ be the vector of mediators, functions of mediators, and interactions between the mediators and the treatment or baseline covariates. This vector will be the vector of variables defined by the blip-down function for m. Let M_i have dimension k_m. We gather each of these row vectors in matrices W, V, and M, so that W for instance is an $n \times k$ matrix.

Let α be the vector of regression coefficients for the first model, α_m be the subvector of coefficients for M_i, and β be the vector of coefficients for the direct effect model. Given linear models for each stage, we can write the regression errors $u_{i1}(\alpha) = Y_i - W_i\alpha$ and $u_{i2}(\beta, \alpha) = Y_i - M_i\alpha_m - V_i\beta$. Let $\widehat{\alpha}$ be the estimator for α based on the sample moment conditions $n^{-1}\sum_i W_i^T u_{i1}(\widehat{\alpha}) = 0$ and $\widehat{\beta} = \widehat{\beta}(\widehat{\alpha})$ be the estimator based on the sample moment condition $n^{-1}\sum_i V_i^T u_{i2}(\widehat{\beta}, \widehat{\alpha}) = 0$. These are simply the OLS estimates from the first and second stages and $u_{i1}(\widehat{\alpha})$ and $u_{i2}(\widehat{\beta}, \widehat{\alpha})$ are the residuals.

Under standard theory (Newey and McFadden, 1994), Assumptions 1 and 2, and the assumption of correct linear models, we can show that $\widehat{\beta}$ is asymptotically Normal with asymptotic variance:

$$\text{Var}\left[\widehat{\beta}\right] = \left(E[V_i^T V_i]\right)^{-1} E\left[g_i g_i^T\right] \left(E[V_i^T V_i]\right)^{-1}, \tag{1}$$

with

$$g_i = V_i^T u_{i2} - F\left(E[W_i^T W_i]\right)^{-1} W_i^T u_{i1} \tag{2}$$

Here, $F = E[-V_i^T \widetilde{W}_i]$, where $\widetilde{W}_i = [M_i \ \mathbf{0}]$ is the vector of W_i with all non-M_i entries set to 0. To prove this, one only need note that the population moment conditions here are $E[V_i^T u_{i2}] = 0$ and $E[W_i^T u_{i1}] = 0$. Using the above assumptions and these moment conditions into Theorem 6.1 of Newey and McFadden (1994, p. 2178) yields (1).

To derive a consistent estimator for the variance, we simply plug sample versions of the population expectations in (1). Under regularity conditions $V^T V/n \xrightarrow{p} E[V_i^T V_i]$ and $W^T W/n \xrightarrow{p} E[W_i^T W_i]$ and we can use

$$\widehat{F} = -\frac{1}{n}\sum_i V_i^T \widetilde{W}_i, \tag{3}$$

which is consistent for F. Finally, we plug in the residuals to form:

$$\widehat{g}_i = V_i^T u_{i2}(\widehat{\beta}, \widehat{\alpha}) - \widehat{F}\left(W^T W\right)^{-1} W_i^T u_{i1}(\widehat{\alpha}). \tag{4}$$

Finally, we can combine each of these to form consistent variance estimator:

$$\widehat{\text{Var}}\left[\widehat{\beta}\right] = (V^T V)^{-1}\left(n^{-1}\sum_i \widehat{g}_i \widehat{g}_i^T\right)(V^T V)^{-1} \tag{5}$$

Note that this variance estimator is "robust" in the sense that it is consistent even if there is heteroskedasticity in either model. Given the structure of g_i and F, the variance of $\widehat{\beta}$ with α estimated will always be higher than if we were to have knowledge about the true α.

C Bias formulas and sensitivity analysis details

Let $\Delta(V_i|W_i) \equiv V_i - \widehat{E}[V_i|W_i]$ be the residuals of a regression of V_i on W_i By the Frisch-Waugh Theorem, we can write the estimated coefficient of M_i on Y_i as the following:

$$\widehat{\alpha}_2 = \alpha_2 + \frac{\sum_{i=1}^n \varepsilon_{iy}\Delta(M_i|Z_i, A_i, X_i)}{\sum_{i=1}^n \Delta(M_i|Z_i, A_i, X_i)}$$

Note that $\frac{1}{n}\sum_i \Delta(M_i|Z_i, A_i, X_i)^2$ converges in probability to $\text{Var}[\varepsilon_{im}]$ and $\frac{1}{n}\sum_i \varepsilon_{iy}\Delta(M_i|Z_i, A_i, X_i)$ converges to $\text{Cov}[\varepsilon_{iy}, \varepsilon_{im}]$. Combining these two facts with Slutsky's theorem gives the following:

$$\text{plim}\ \widehat{\alpha}_2 = \alpha_2 + \frac{\text{Cov}[\varepsilon_{iy}, \varepsilon_{im}]}{\text{Var}[\varepsilon_{im}]} \tag{6}$$

$$= \alpha_2 + \frac{\rho\sigma_y\sigma_m}{\sigma_m^2} \tag{7}$$

$$= \alpha_2 + \frac{\rho\sigma_y}{\sigma_m} \tag{8}$$

Let the true blipped-down outcome be $\tilde{Y}_i = Y_i - \alpha_2 M_i$. We can write $\tilde{Y}_i = \beta_0 + \beta_1 A_i + X_i^T\beta_2 + \eta_i$, where β_1 is the ACDE. Let $\widehat{\beta}_1$ be the coefficient from the regression of $\widehat{\tilde{Y}}_i$ on A_i and X_i. By the Frisch-Waugh theorem, we have:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \widehat{\alpha}_2 M_i)\Delta(A_i|X_i)}{\sum_{i=1}^n \Delta(A_i|X_i)^2} \tag{9}$$

$$= \frac{\sum_{i=1}^n (\widetilde{Y}_i - \alpha_2 M_i + \widehat{\alpha}_2 M_i)\Delta(A_i|X_i)}{\sum_{i=1}^n \Delta(A_i|X_i)^2} \tag{10}$$

$$= \frac{\sum_{i=1}^n \widetilde{Y}_i\Delta(A_i|X_i)}{\sum_{i=1}^n \Delta(A_i|X_i)^2} - \frac{(\alpha_2 - \widehat{\alpha}_2)\sum_{i=1}^n M_i\Delta(A_i|X_i)}{\sum_{i=1}^n \Delta(A_i|X_i)^2} \tag{11}$$

Let $M_i = \widetilde{\delta}_0 + \widetilde{\delta}_1 A_i + X_i^T\widetilde{\delta}_2 + \widetilde{\varepsilon}_{im}$ be the regression of M_i on A_i and X_i. Basic regression results establish that $\sum_i \widetilde{Y}_i\Delta(A_i|X_i)/\sum_i \Delta(A_i|X_i)^2$ converges to β_1 and $\sum_i M_i\Delta(A_i|X_i)/\sum_i \Delta(A_i|X_i)^2$ converges to $\widetilde{\delta}_1$. And given our above results, we have that $\alpha - \widehat{\alpha}_2$ converges to $\rho\sigma_y/\sigma_m$. Again using repeated applications of Slutsky's theorem, we can derive the asymptotic bias:

$$\text{plim}\ \widehat{\beta}_1 = \beta_1 - \frac{\rho\sigma_y\widetilde{\delta}_1}{\sigma_m}$$

Of course, σ_y is not identified due to the confounding. We take a similar approach to Imai, Keele and Yamamoto (2010) and note the following relationships between the various parameters:

$$\text{Var}[\widetilde{\varepsilon}_{iy}] = \widetilde{\sigma}_y^2 = \alpha_2^2\sigma_m^2 + \sigma_y^2 + 2\rho\alpha_2\sigma_m\sigma_y \tag{12}$$

$$\text{Cov}[\widetilde{\varepsilon}_{iy}, \varepsilon_{im}] = \widetilde{\rho}\widetilde{\sigma}_y\sigma_m = \alpha_2\sigma_m^2 + \rho\sigma_m\sigma_y \tag{13}$$

Solving for σ_y, we find that $\sigma_y = \widetilde{\sigma}_y\sqrt{(1-\widetilde{\rho}^2)/(1-\rho^2)}$. Plugging this into above yields the asymptotic bias formula.

To complete the proof note that (12) implies that $\widetilde{\delta}_1$ is identified from a regression of M_i on A_i and X_i. Under the LSEM and (12), we can use the

$$\widehat{\widetilde{\sigma}}_y = \sqrt{\text{Var}[\widehat{\widetilde{\varepsilon}}_{iy}]} = \sqrt{\sum_i (Y_i - \widehat{\widetilde{\alpha}}_0 - \widehat{\widetilde{\alpha}}_1 A_i - X_i^T\widehat{\widetilde{\alpha}}_3 - Z_i^T\widehat{\widetilde{\alpha}}_4)^2},$$

which is consistent for $\widetilde{\sigma}_y$. Furthermore,

$$\widehat{\sigma}_m = \sqrt{\text{Var}[\widehat{\varepsilon}_{im}]} = \sqrt{\sum_i (M_i - \widehat{\delta}_0 - \widehat{\delta}_1 A_i - X_i^T\widehat{\delta}_2 - Z_i^T\widehat{\delta}_3)^2}$$

which is consistent for σ_m. Finally, we can estimate $\widetilde{\rho}$ with the correlation between the residuals $\widehat{\widetilde{\varepsilon}}_{iy}$ and $\widehat{\varepsilon}_{im}$. Thus, given ρ the asymptotic bias of $\widehat{\beta}_1$ is identified and we can use this to identify β_1.

To get standard errors and confidence intervals for the sensitivity analysis, it is easier to correct for the bias in $\widehat{\alpha}_2$ and pass this bias-corrected estimate to the second stage. Then, the variance estimator of B consistently estimates the variance of $\widehat{\beta}$ as if the first stage were correctly specified. That is, it is the correct variance under the assumption that we have correctly chosen ρ.

Finally, note that we can reparameterize this sensitivity analysis to be as a function of the residual variation explained by unmeasured confounding. To see this, we introduce an unmeasured confounder, U_i:

$$\varepsilon_{iy} = \alpha_u U_i + \varepsilon^*_{iy} \tag{14}$$

$$\varepsilon_{im} = \delta_u U_i + \varepsilon^*_{im} \tag{15}$$

With these in hand, we can define the partial R^2 for U_i in terms of the outcome and the mediator:

$$R^2_y = 1 - \frac{\text{Var}[\varepsilon^*_{iy}]}{\text{Var}[\varepsilon_{iy}]} \tag{16}$$

$$R^2_m = 1 - \frac{\text{Var}[\varepsilon^*_{im}]}{\text{Var}[\varepsilon_{im}]} \tag{17}$$

These values represent the share of the unexplained variance in Y_i and M_i, respectively, that U_i explains. As shown by Imai, Keele and Yamamoto (2010), we have the following relationship between ρ and these partial R^2 values: $\rho^2 = R_y^2 R_m^2$. Thus, we can vary these parameters, which imply differing values of ρ and consequently differing levels of bias. The advantage of this parameterization of the sensitivity analysis is that the partial R^2 may be more natural to interpret. For instance, we can compare them to the partial R^2 values of observed covariates in X_i and Z_i in order to gauge their relative magnitude (Imbens, 2004).

D Simulation setup

Here we present the simulation setup we discussed in Section and present results from entire set of draws as opposed to only a single draw. This simulation is not meant to prove any property of any estimator—these results are largely known and have been established analytically. Instead, we show these for illustrative purposes.

$$N = 500$$

$$A_i \sim N(50, 15^2)$$

$$Z_i \sim N(50, 15^2)$$

$$M_i \sim N(0.5A_i + 0.5Z_i, 5^2)$$

$$Y_i \sim N(75 - 0.5Z_i, 5^2)$$

We took 10,000 draws from this data generating process and ran three estimators on the samples. First, we ran a simple unconditional model of Y_i on A_i. Next, we ran a conditional model of Y_i on A_i and M_i. Finally, we applied the sequential g-estimation to estimate the direct effect of A_i, using Z_i as an intermediate confounder. We plot the results in Figure 1. Given the data generating process, it is unsurprising that both the unconditional and sequential g-estimation approach recover the truth on average, while conditioning on M_i induces very severe post-treatment bias. Note also that the sequential g-estimator has slightly

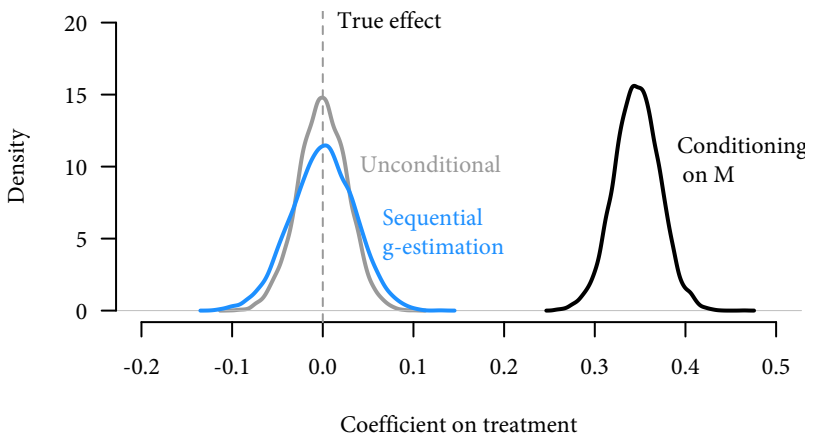

Figure 1: Simulated sampling distribution of the post-treatment bias estimator that conditions on M_i and the unbiased estimator that does not condition on M_i. Conditioning on a post-treatment covariate in this case produces serious bias.

higher variance than unconditional approach, which makes sense because the unconditional estimator is taking advantage of an addition restriction in this example: no effect of A_i on Z_i. If that were not true, then the unconditional estimator would also be biased.

Bibliography

Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1, February):51–71. **URL:** http://projecteuclid.org/euclid.ss/1280841733

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86(1, February):4–29.

Newey, Whitney K. and Daniel McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, ed. Robert F. Engle and Daniel L. McFadden. Elsevier pp. 2111–2245.

VanderWeele, Tyler J. 2011. "Controlled Direct and Mediated Effects: Definition, Identification and Bounds." *Scandinavian Journal of Statistics* 38(3, September):551–563.