A Regression-with-Residuals Method for Estimating Controlled Direct Effects*

Xiang Zhou[1] and Geoffrey T. Wodtke[2]

[1]Harvard University
[2]University of Toronto

February 2, 2018

Abstract

In a recent contribution, Acharya, Blackwell and Sen (2016) described the method of sequential g-estimation for estimating the controlled direct effect (CDE). We propose an alternative method, which we call "regression-with-residuals" (RWR), for estimating the CDE. Compared with sequential g-estimation, the RWR method is easier to understand and to implement. More important, unlike sequential g-estimation, it can easily accommodate several different types of effect moderation, including cases in which the effect of the mediator on the outcome is moderated by a post-treatment, or intermediate, confounder. Although common in the social sciences, this type of effect moderation is typically assumed away in applications of sequential g-estimation, which may lead to bias if effect moderation is in fact present. We illustrate RWR by reanalyzing the effect of plough use on female political participation while allowing the effect of log GDP per capita (the mediator) to vary across levels of several intermediate confounders.

*Direct all correspondence to Xiang Zhou, Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA; email: xiang_zhou@fas.harvard.edu. The authors thank Felix Elwert and Gary King for helpful comments on previous versions of this work.

Introduction

Over the past decade, the use of causal mediation analysis has been rapidly growing in empirical political science. Many scholars are no longer satisfied with merely establishing the presence of a causal effect between one variable and another; rather, they now seek to additionally identify causal mechanisms that explain such effects (e.g., Abramson and Carter 2016; Hall 2017; Holbein 2017; Knutsen et al. 2017; Reese, Ruby and Pape 2017; Zhu 2017). The study of causal mediation, however, often rests on strong and untestable assumptions (VanderWeele and Vansteelandt 2009; Imai et al. 2011). In a recent contribution, Acharya, Blackwell and Sen (2016, henceforth ABS) show that these assumptions are relatively weak when we focus on a quantity called the controlled direct effect (CDE). The CDE measures the strength of a causal relationship between a treatment and outcome when a putative mediator is fixed at a given value for all units (Pearl 2001; Robins 2003). This estimand is useful because it helps to adjudicate between alternative causal explanations. A nonzero CDE, for example, would imply that the causal effect of treatment on the outcome does not operate exclusively through the mediator of interest. Moreover, as ABS suggest, the difference between the total effect and the CDE can be interpreted as the degree to which the mediator contributes to a causal mechanism that transmits the effect of treatment to the outcome.

Nevertheless, identification of the CDE is not straightforward. Simply conditioning on the mediator (via stratification, matching, or regression adjustment) is usually insufficient because the effect of the mediator on the outcome may be confounded, possibly by post-treatment variables. For example, when assessing the CDE of historical plough use on female political participation at a given level of log GDP per capita (the mediator), post-treatment variables, such as the industrial composition of the contemporary economy or level of democracy, may affect both log GDP per capita and female political participation. Following ABS, we call these variables intermediate confounders. Intermediate confounders pose a dilemma for the identification and estimation of CDEs. On the one hand, omitting intermediate confounders would bias estimates of the effect of the mediator on the outcome, and by extension, also estimates of the CDE. On the other hand, naively controlling for intermediate confounders may block causal pathways, and unblock noncausal pathways, from treatment to the outcome, which would also bias estimates of the CDE.

Fortunately, several different approaches have been developed to overcome this dilemma. First, we could estimate a model for the marginal expectation of the potential outcomes under different levels of the treatment and mediator, known as a marginal structural model (MSM), us-

ing the method of inverse probability weighting (IPW) (Robins, Hernan and Brumback 2000; VanderWeele 2009). This approach performs best when both the treatment and mediator are binary. When the treatment and/or mediator are many valued or continuous, it tends to perform poorly because the inverse probability weights involve conditional density estimates that are often unreliable. Second, to overcome these limitations, we could instead estimate a structural nested mean model (SNMM) for the conditional expectation of the potential outcomes given a set of both pretreatment and intermediate confounders using the method of sequential g-estimation as described in ABS (see also Vansteelandt 2009; Joffe and Greene 2009). Sequential g-estimation of an SNMM involves a two-stage regression-based procedure in which the variation in the outcome due to the causal effect of the mediator is removed, and then the "de-mediated" outcome is regressed on treatment and the pretreatment confounders.

Like IPW estimation, however, sequential g-estimation also suffers from several limitations. First, the underlying logic of g-estimation is not especially intuitive, which is perhaps why the application of this method remains infrequent (Vansteelandt and Joffe 2014). Second, sequential g-estimation is complicated to implement when there are "intermediate interactions," that is, when the effect of the mediator on the outcome is moderated by an intermediate confounder. As ABS note

> [I]f Assumption 2 [no intermediate interactions] is violated, it is still possible to estimate the ACDE in a second stage, but that requires (i) a model for the distribution of the intermediate covariates conditional on the treatment and (ii) the evaluation of the average of within-stratum ACDEs across the distribution of that model. The second part entails a high-dimensional integral that is computationally challenging, though Monte Carlo procedures have been developed (Robins 1986, 1997).

Because of these computational challenges, intermediate interactions are typically assumed away in applications of sequential g-estimation, but if this assumption is not met in practice, then estimates of the CDE may be biased.

In this paper, we introduce an alternative method, termed "regression-with-residuals" (RWR), for estimating the CDE that is both more intuitive and easier to implement than sequential g-estimation. In particular, RWR estimation is easily implemented even in the presence of intermediate interactions, while in the absence of such interactions, we show that this method is algebraically equivalent to sequential g-estimation. We illustrate the utility of RWR by reanalyzing

data considered in ABS to estimate the CDE of plough use on female political participation, now allowing the effect of log GDP per capita (the mediator) to vary across levels of the intermediate confounders, such as oil revenues per capita. We find evidence of a significant intermediate interaction that, when naively excluded, would appear to suppress estimates of the CDE.

Notation, Assumptions, and Sequential G-estimation

Following ABS, we use A to denote the treatment, M to denote the mediator, Y to denote the observed outcome, and $Y(a, m)$ to denote the potential outcome under treatment a and mediator m. With this notation, the CDE is formally defined as the *average* effect of changing treatment from a to a' while fixing the mediator at a given level m[1]:

$$\text{CDE}(a, a', m) = \mathbb{E}[Y(a, m) - Y(a', m)]$$

This quantity is nonparametrically identified under the assumption of sequential ignorability (Robins 1997; Vansteelandt 2009),[2] which can be formally expressed in two parts as follows:

1. $Y(a, m) \perp\!\!\!\perp A|X$ (i.e., no unmeasured treatment–outcome confounders)

2. $Y(a, m) \perp\!\!\!\perp M|X, A, Z$ (i.e., no unmeasured mediator-outcome confounders)

Here, X denotes a vector of observed pretreatment confounders that may affect both treatment and the outcome, while Z denotes a vector of observed post-treatment, or intermediate, confounders that may affect both the mediator and the outcome and that may be affected by treatment. The sequential ignorability assumption is satisfied in Figure 1, which contains a directed acyclic graph summarizing a set of hypothesized causal relationships between the variables outlined previously.

Although the sequential ignorability assumption is sufficient for nonparametric identification of the CDE, additional modeling assumptions are needed to estimate the CDE in finite samples. Sequential g-estimation, for example, typically relies on an unsaturated linear model for the conditional mean of the potential outcomes, $Y(a, m)$, given X and Z. Moreover, because sequential g-estimation is difficult to implement in the presence of intermediate interactions, its application

[1]The same quantity is defined as ACDE (i.e., the average CDE) in ABS, who use CDE_i to denote the individual-level controlled direct effect. We avoid this distinction for concision.

[2]The sequential ignorability assumption defined here is weaker than that stated in ABS (Assumption 1), as it does not require $M(a) \perp\!\!\!\perp A|X$ (i.e. no unmeasured treatment-mediator confounders). This condition, as discussed in VanderWeele and Vansteelandt (2009), is not required for identifying controlled direct effects.

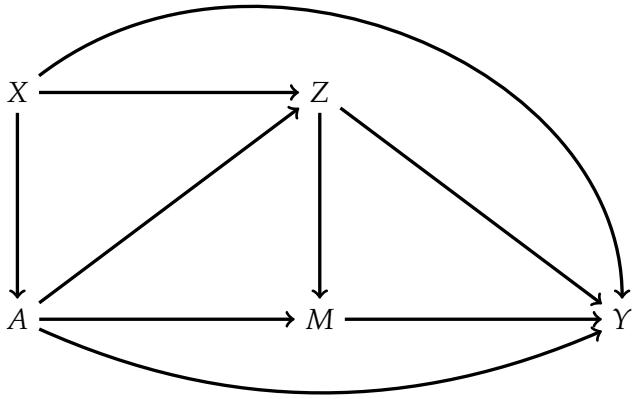

Figure 1: Causal Relationships under Sequential Ignorability Shown in Direct Acyclic Graph.

Note: A denotes the treatment, M denotes the mediator, Y denotes the outcome, X denotes pre-treatment confounders, Z denotes intermediate confounders.

in practice, as with ABS, relies on an additional simplifying assumption that the effect of the mediator on the outcome is not moderated by post-treatment confounders, which can be formally expressed as follows:

$$\mathbb{E}[Y(a,m) - Y(a,m')|X = x, Z = z] = \mathbb{E}[Y(a,m) - Y(a,m')|X = x] \quad \text{for any } a, m, m', x \text{ and } z$$

Under this assumption, ABS illustrate sequential g-estimation of the CDE using the following SNMM:

$$\mathbb{E}[Y(a,m)|X = x, Z = z] = \beta_0 + \beta_1^T x + \beta_2 a + \beta_3^T z + m(\gamma_0 + \gamma_1^T x + \gamma_2 a) \tag{1}$$

Specifically, with this model, sequential g-estimation proceeds in three steps:

1. Compute least squares estimates for equation (1) and save $\hat{\gamma}_2$
2. Construct a "de-mediated" outcome defined as $y_d = y - m(\hat{\gamma}_0 + \hat{\gamma}_1^T x + \hat{\gamma}_2 a)$
3. Compute least squares estimates for a linear regression of y_d on x and a, which can be expressed as $\hat{y}_d = \hat{\kappa}_0 + \hat{\kappa}_1^T x + \hat{\kappa}_2 a$

The sequential g-estimate of the CDE is then given by

$$\widehat{\text{CDE}}_{\text{SG}}(a, a', m) = (\hat{\kappa}_2 + \hat{\gamma}_2 m)(a - a') \tag{2}$$

Standard errors can be obtained via the nonparametric bootstrap or a consistent variance estimator derived in ABS.

In Figure 2, we attempt to illustrate the logic of sequential g-estimation. First, because the mediator-outcome relationship is unconfounded, the regression in step 1 identifies the causal effect of M on Y. Then, the "de-mediation" calculation in step 2 neutralizes the causal path from M to Y while keeping all other causal paths intact. Finally, the regression of the de-mediated outcome, Y_d, on X and A in step 3 identifies the controlled direct effect of A when $M = 0$, and because $\hat{\gamma}_2$ is a consistent estimate of the treatment-mediator interaction effect, the CDE when $M = m$ can be estimated with equation (2).

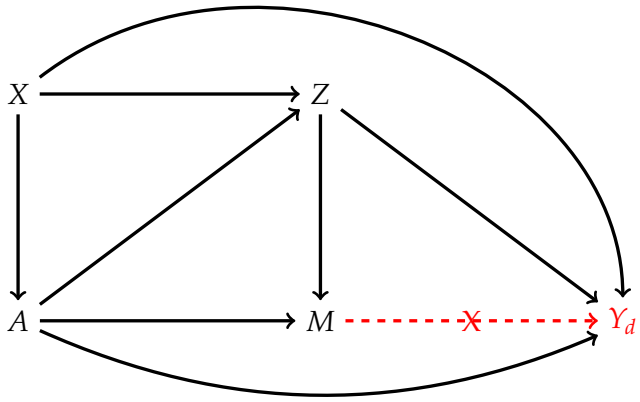

Figure 2: The Logic of Sequential G-estimation.

Regression-with-Residuals Estimation

Like g-estimation of SNMMs, RWR estimation was originally developed to assess how time-varying confounders moderate the effect of time-varying treatments (Almirall, Ten Have and Murphy 2010; Wodtke and Almirall 2017). In this section, we show how this method can be adapted to estimate CDEs while properly adjusting for intermediate confounders under the assumptions outlined previously. Specifically, RWR estimation of the CDE in a SNMM similar to that considered in ABS proceeds in two steps:

1. For each of the intermediate confounders, compute least squares estimates for a linear regression of z on x and a, and save the residuals, which we denote by $z_{\perp}$

2. Compute least squares estimates for a model similar to equation (1) but with z replaced by $z_\perp$, which can be expressed as $\hat{y} = \tilde{\beta}_0 + \tilde{\beta}_1^T x + \tilde{\beta}_2 a + \tilde{\beta}_3^T z_\perp + m(\tilde{\gamma}_0 + \tilde{\gamma}_1^T x + \tilde{\gamma}_2 a)$

The RWR estimate of the CDE is then given by

$$\widehat{\text{CDE}}_{\text{RWR}}(a, a', m) = (\tilde{\beta}_2 + \tilde{\gamma}_2 m)(a - a') \tag{3}$$

As we show in Appendix A, when there are no intermediate interactions, RWR and sequential g-estimation are algebraically equivalent (i.e., $\hat{\kappa}_2 = \tilde{\beta}_2$; $\hat{\gamma}_2 = \tilde{\gamma}_2$). They rely on the same identification and modeling assumptions, and they share the same statistical properties. But compared with sequential g-estimation, the logic of RWR estimation is somewhat more intuitive. As shown in Figure 3, residualizing the intermediate confounders in step 1 neutralizes the causal paths emanating from X and A to Z. The residualized confounders, which have been purged of their association with treatment, can then be included in the regression model for the outcome in order to adjust for mediator-outcome confounding while avoiding bias due to conditioning on post-treatment variables. In other words, RWR estimation avoids post-treatment bias because $Z_\perp$ is no longer a consequence of A, and it avoids omitted variable bias because all treatment-outcome and mediator-outcome confounders have been appropriately controlled in the model for the outcome.

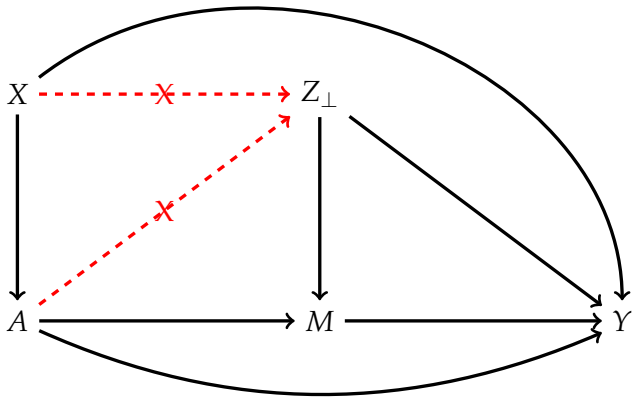

Figure 3: The Logic of Regression-with-residuals.

Intermediate Interactions

In the models considered previously, the effect of the mediator is assumed to be invariant across all intermediate confounders. This is a strong and arguably implausible assumption in many social science applications, and when it fails to hold, estimates of the CDE may be biased and inconsistent. Thus, accommodating, rather than naively assuming away, intermediate interactions will make analyses of causal mediation more robust. In the presence of intermediate interactions, the advantages of RWR over sequential g-estimation become more apparent. For example, consider the following model, which extends equation (1) by including an interaction term between M and Z:

$$\mathbb{E}[Y(a,m)|X=x,Z=z] = \beta_0 + \beta_1^T x + \beta_2 a + \beta_3^T z + m(\gamma_0 + \gamma_1^T x + \gamma_2 a + \gamma_3^T z) \tag{4}$$

With this model, sequential g-estimation can still be used to estimate the CDE, but only at $M = 0$. The only modification to the sequential g-estimator in this situation is that the de-mediated outcome, y_d, is obtained by subtracting $m(\hat{\gamma}_0 + \hat{\gamma}_1^T x + \hat{\gamma}_2 a + \hat{\gamma}_3^T z)$ instead of $m(\hat{\gamma}_0 + \hat{\gamma}_1^T x + \hat{\gamma}_2 a)$ from the observed outcome. Then, $\widehat{\text{CDE}}_{\text{SG}}(a, a', 0) = \hat{\kappa}_2(a - a')$, where $\hat{\kappa}_2$ is the coefficient on treatment from the regression of y_d on x and a. Unfortunately, however, we can no longer estimate the CDE in general for $M = m$ using $\hat{\gamma}_2 m(a - a')$ because this expression is no longer a consistent estimate of the treatment-mediator interaction effect. In equation (4), the inclusion of an intermediate interaction, $\gamma_3^T z$, leads to post-treatment bias in the treatment-mediator interaction, $\gamma_2 a$, just as the inclusion of main effects for the intermediate confounders, $\beta_3^T z$, leads to post-treatment bias in the main effect of treatment, $\beta_2 a$. With sequential g-estimation, the latter bias is removed by the de-mediation step, but the former is not.

By contrast, with RWR estimation, intermediate interactions can be easily accommodated, and its implementation in their presence remains almost exactly the same as before:

1. For each of the intermediate confounders, compute least squares estimates for a linear regression of z on x and a, and save the residuals, denoted by $z_\perp$

2. Compute least squares estimates for a model similar to equation (4) but with z replaced by $z_\perp$, which can be expressed as

$$\hat{y} = \tilde{\beta}_0 + \tilde{\beta}_1^T x + \tilde{\beta}_2 a + \tilde{\beta}_3^T z_\perp + m(\tilde{\gamma}_0 + \tilde{\gamma}_1^T x + \tilde{\gamma}_2 a + \tilde{\gamma}_3^T z_\perp)$$

The RWR estimate of the CDE is then given by[3]

$$\widehat{\text{CDE}}_{\text{RWR}}(a, a', m) = (\tilde{\beta}_2 + \tilde{\gamma}_2 m)(a - a') \tag{5}$$

As we show in Appendix B, equation (5) is a consistent estimator of the CDE under the identification assumptions outlined previously and provided that there is no model misspecification. In words, RWR estimation remains consistent even in the presence of intermediate interactions because, by appropriately residualizing the intermediate confounders, it removes any post-treatment bias for both the main effect of treatment and the treatment-mediator interaction effect. As before, standard errors can be computed using the nonparametric bootstrap.

The Effect of Plough Use on Female Political Participation

To illustrate the RWR method and demonstrate its utility, we reanalyze the CDE of historical plough use on female political participation. In their original study, Alesina, Giuliano and Nunn (2013) find that the total effect of plough use on female political participation is small and statistically insignificant, but they also find that the coefficient on plough use doubles and becomes statistically significant after controlling for log GDP per capita in the year 2000. Based on these results, Alesina, Giuliano and Nunn (2013) conclude that log GDP per capita is an important mediator and suggest that the small total effect of plough use on female political participation is due to the combination of a positive indirect effect (through log GDP per capita) and a negative direct effect. However, as highlighted by ABS, this conclusion may not be warranted because the analysis in Alesina, Giuliano and Nunn (2013) does not account for intermediate confounders that likely affect both log GDP per capita and women's participation in politics. To rectify this problem, ABS use sequential g-estimation to estimate the CDE of plough use on female political participation, controlling for log GDP per capita, and they find that this effect is even larger after appropriately adjusting for intermediate confounders.

To compare RWR with sequential g-estimation, we begin by considering the same model as in ABS:

$$\mathbb{E}[Y(a, m)|x, z] = \beta_0 + \beta_1^T x + \beta_2 a + \beta_3^T z + m(\gamma_0 + \gamma_1 a) + m^2(\gamma_2 + \gamma_3 a),$$

[3]In previous work (Almirall, Ten Have and Murphy 2010; Wodtke and Almirall 2017), where RWR has been used to estimate the moderated, or conditional, effects of time-varying treatments, the residualized confounders are only included as "main effects" and are not used in any cross-product terms. In our adaptation of RWR for estimating CDEs, however, the residualized confounders must be included both as "main effects" and in the relevant cross-product terms, which ensures that $\tilde{\beta}_2$ and $\tilde{\gamma}_2$ capture all the information needed to construct estimates of the CDE.

Table 1: Estimated CDE of Plough Use on Female Political Participation using Sequential G-estimation, RWR, and RWR with intermediate interactions

	Sequential g-estimation (final step)	RWR	RWR with intermediate interactions
intercept	8.53	8.53	7.6
	(5.42)	(5.42)	(6.33)
plough use (i.e., $\widehat{\text{CDE}}(a, a+1, 0)$)	-8.64**	-8.64**	-12.76***
	(3.14)	(3.14)	(3.83)
log GDP per capita		4.32**	5.76***
		(1.65)	(1.65)
log GDP per capita^2		0.72	0.46
		(0.71)	(0.64)
plough use * log GDP per capita		-3.5†	-2.69
		(1.83)	(1.86)
plough use * log GDP per capita^2		0.88	0.94
		(0.84)	(0.76)
years of civil conflict		0.13*	0.12*
		(0.06)	(0.06)
years of interstate conflict		-0.3*	-0.27†
		(0.14)	(0.14)
oil revenues per capita		-10.74	-72.39**
		(12.7)	(23.86)
proportion European descent		0.05	0.03
		(0.03)	(0.03)
former Communist rule		-0.16	0.53
		(2.33)	(2.22)
Polity score in 2000		-0.11	-0.23
		(0.17)	(0.17)
value added in Service as share of GDP in 2000		0.00	-0.07
		(0.08)	(0.09)
oil revenues per capita * log GDP per capita			26.56*
			(10.81)

Note: Numbers in parentheses are bootstrapped standard errors. †p<.1, *p<.05, **p<.01, ***p<.001 (two-tailed z tests). Coefficients of pretreatment confounders are omitted.

where the outcome, Y, is the share of political positions held by women in the year 2000; the treatment, a, is the relative proportion of ethnic groups in a country that traditionally used the plough; the mediator, m, is log GDP per capita in the year 2000, recentered at its mean; the vector of pre-treatment confounders, x, includes measures of agricultural suitability, tropical climate, presence of large animals, political hierarchy, economic complexity, and terrain ruggedness; and finally, the vector of intermediate confounders, z, includes measures of civil conflict, interstate conflict, oil revenues per capita, the proportion of population that is of European descent, former Communist rule, the policy score in 2000, and the value added of the service industry as share of GDP in the year 2000.[4] The first two columns of Table 1 present sequential g-estimates and RWR estimates, respectively, for the CDE of historical plough use based on this model.[5] As expected, the estimated CDE given by these two different methods is exactly the same. Note that, with sequential g-estimation, only the CDE at $m = 0$ is reported in the final step (in this case, the CDE when log GDP per capita is set at its mean). To construct the CDE at other levels of the mediator, the analyst must return to the regression in the first step of the sequential g-estimation procedure and extract the coefficients on the treatment-mediator interactions. With RWR, by contrast, all the necessary coefficients are reported in a single regression, which allows the analyst to quickly construct the CDE at any level of the mediator from the output in Table 1.

Thus far, the effect of log GDP per capita on female political participation has been assumed to be invariant across levels of the intermediate confounders, but if the effect of the mediator is in fact moderated by any of these post-treatment variables, then the estimates reported previously are likely biased and inconsistent. We now relax this assumption by additionally including an interaction term between log GDP and oil revenues per capita.[6] Results from this analysis are shown in the last column of Table 1, and Appendix C presents the R code used to generate them. As indicated by the additional interaction term, which is statistically significant at the 0.05 level, the effect of log GDP per capita is larger in countries with more oil revenue. When this intermediate interaction is appropriately modeled via RWR, the estimated CDE at $m = 0$ (i.e., when log GDP per capita is set at its mean) increases by almost 50%, from -8.64 to -12.76, while the treatment-mediator interaction between plough use and log GDP per capita becomes muted and no longer

[4] Detailed definitions of all these variables can be found in Alesina, Giuliano and Nunn (2013).

[5] Our results are slightly different from those reported in ABS because we handle missing values differently. Whereas ABS include countries with missing values on z in the third step of the sequential g-estimator, we use only complete observations throughout the analysis.

[6] We also estimated a model with all two-way interactions between log GDP per capita and the intermediate confounders, from which we obtained an RWR estimate of the CDE that is very similar to the one reported in Table 1.

statistically significant. Taken together, these results suggest that naively assuming away intermediate interactions when they do in fact exist could induce substantial bias in estimates of the CDE. Fortunately, this bias can be easily avoided with RWR.

Summary

In this paper, we introduced RWR for estimating controlled direct effects. In the absence of intermediate interactions, RWR is algebraically equivalent to the sequential g-estimator described in ABS. However, unlike the sequential g-estimator, RWR can easily accommodate several different types of effect moderation, including – as we have shown – intermediate interactions, which are likely ubiquitous in the social sciences. And in general, models with less stringent parametric constraints can be estimated more easily with RWR than with sequential g estimation. Given its simplicity, flexibility, and utility, we expect that RWR estimation will be used more widely in causal mediation analyses.

Appendix A: Equivalence between RWR and Sequential G-Estimation Under No Intermediate Interactions

To see the equivalence between RWR and sequential g-estimation, let us consider model (1) and write the "naive" least squares regression of it as

$$y = \hat{\beta}_0 + \hat{\beta}_1^T x + \hat{\beta}_2 a + \hat{\beta}_3^T z + m(\hat{\gamma}_0 + \hat{\gamma}_1^T x + \hat{\gamma}_2 a) + y_\perp, \tag{6}$$

where $y_\perp$ denotes the residual. Suppose x is a column vector of p pretreatment confounders and z is a column vector of q intermediate confounders. For each of the components in z, it has a least squares fit on x and a. These least squares fits can be combined in matrix form:

$$z = \hat{\lambda}_0 + \hat{\Lambda}_1 x + \hat{\lambda}_2 a + z_\perp, \tag{7}$$

where $\hat{\lambda}_0$ and $\hat{\lambda}_2$ are $q \times 1$ vectors, $\hat{\Lambda}_1$ is a $q \times p$ matrix, and $z_\perp$ is a $q \times 1$ vector of residuals. Substituting equation (7) into equation (6), we have

$$y = (\hat{\beta}_0 + \hat{\beta}_3^T \hat{\lambda}_0) + (\hat{\beta}_1^T + \hat{\beta}_3^T \hat{\Lambda}_1)x + (\hat{\beta}_2 + \hat{\beta}_3^T \hat{\lambda}_2)a + \hat{\beta}_3^T z_\perp + m(\hat{\gamma}_0 + \hat{\gamma}_1^T x + \hat{\gamma}_2 a) + y_\perp. \tag{8}$$

Since $y_\perp$ is the least squares residual for regression (6), it is orthogonal to the span of $\{1, x, a, z, m, mx, ma\}$. Because $z_\perp$ is a linear combination of x, a, and z, $\{1, x, a, z_\perp, m, mx, ma\}$ and $\{1, x, a, z, m, mx, ma\}$ span the same space. Thus equation (8) represents the least squares fit of y on $\{1, x, a, z_\perp, m, mx, ma\}$, meaning that the RWR estimator of the CDE is

$$\widehat{\text{CDE}}_{\text{RWR}}(a, a', m) = (\hat{\beta}_2 + \hat{\beta}_3^T \hat{\lambda}_2 + \hat{\gamma}_2 m)(a - a').$$

From equation (8), we also know that the de-mediated outcome can be written as

$$y_d = (\hat{\beta}_0 + \hat{\beta}_3^T \hat{\lambda}_0) + (\hat{\beta}_1^T + \hat{\beta}_3^T \hat{\Lambda}_1)x + (\hat{\beta}_2 + \hat{\beta}_3^T \hat{\lambda}_2)a + \hat{\beta}_3^T z_\perp + y_\perp. \tag{9}$$

Since $z_\perp$ and $y_\perp$ are both orthogonal to the span of $\{1, x, a\}$ (from the properties of least squares residuals), $\hat{\beta}_3^T z_\perp + y_\perp$ is also orthogonal to the span of $\{1, x, a\}$. Thus equation (9) represents the least squares fit of y_d on x and a, meaning that the sequential g-estimator of the CDE is

$$\widehat{\text{CDE}}_{\text{SG}}(a, a', m) = (\hat{\kappa}_2 + \hat{\gamma}_2 m)(a - a') = (\hat{\beta}_2 + \hat{\beta}_3^T \hat{\lambda}_2 + \hat{\gamma}_2 m)(a - a').$$

Obviously, the sequential g-estimator is the same as the RWR estimator.

Appendix B: Consistency of RWR in the Presence of Intermediate Interactions

First, we explain an implicit modeling assumption that underlies both the sequential g-estimator and the RWR estimator described in the main text. For the sequential g-estimator, the least squares regression in step 3 implies the linearity of $\mathbb{E}[Y(a, 0)|X = x]$ in x and a:

$$\mathbb{E}[Y(a, 0)|X = x] = \kappa_0 + \kappa_1^T x + \kappa_2 a. \tag{10}$$

Setting $m = 0$ in model (1) or (4), we have

$$\mathbb{E}[Y(a, 0)|X = x, Z = z] = \beta_0 + \beta_1^T x + \beta_2 a + \beta_3^T z. \tag{11}$$

Taking the expectation of equation (11) over z (given x and a) yields

$$\mathbb{E}[Y(a,0)|X=x] = \beta_0 + \beta_1^T x + \beta_2 a + \beta_3^T \mathbb{E}[z|x,a]. \tag{12}$$

Comparing equations (10) and (12), we can see that $\beta_3^T \mathbb{E}[z|x,a]$ must be linear in x and a. Since β_3 represents model parameters that can vary freely in $\mathbb{R}^q$, the linearity of $\beta_3^T \mathbb{E}[z|x,a]$ implies that each component of $\mathbb{E}[z|x,a]$ should be linear in x and a. Conversely, when each component of $\mathbb{E}[z|x,a]$ is linear in x and a, model (1) or (4) implies equation (10). Thus, the sequential g-estimator implicitly assumes each component of $\mathbb{E}[z|x,a]$ to be linear in x and a. This assumption is more explicit in the RWR estimator, which requires the user to fit a linear model for each of the intermediate confounders. Thus, both the sequential g-estimator and the RWR estimator are based on the linearity of $\mathbb{E}[z|x,a]$[7]

$$\mathbb{E}[z|x,a] = \lambda_0 + \Lambda_1 x + \lambda_2 a. \tag{13}$$

Paralleling equation (7) in Appendix A, λ_0 and λ_2 are both $q \times 1$ vectors and Λ_1 is a $q \times p$ matrix. To see the consistency of the RWR estimator in the presence of intermediate interactions, let us consider model (4). Given equation (13), the CDE can be written as

$$\begin{aligned}
\mathbb{E}[y(a,m) - y(a',m)] &= \mathbb{E}_x \mathbb{E}_{z|x,a} \mathbb{E}[y(a,m)|x,z] - \mathbb{E}_x \mathbb{E}_{z|x,a'} \mathbb{E}[y(a',m)|x,z] \\
&= \beta_2(a-a') + \gamma_2 m(a-a') + \beta_3^T \cdot \mathbb{E}_x[\mathbb{E}[z|x,a] - \mathbb{E}[z|x,a']] \\
&\quad + \gamma_3^T m \cdot \mathbb{E}_x[\mathbb{E}[z|x,a] - \mathbb{E}[z|x,a']] \\
&= \beta_2(a-a') + \gamma_2 m(a-a') + \beta_3^T \lambda_2 (a-a') + \gamma_3^T \lambda_2 m(a-a') \\
&= [(\beta_2 + \beta_3^T \lambda_2) + (\gamma_2 + \gamma_3^T \lambda_2) m](a-a')
\end{aligned}$$

It is easy to show that the RWR estimator for model (4) equals

$$\widehat{\text{CDE}}_{\text{RWR}}(a,a',m) = [(\hat{\beta}_2 + \hat{\beta}_3^T \hat{\lambda}_2) + (\hat{\gamma}_2 + \hat{\gamma}_3^T \hat{\lambda}_2) m](a-a').$$

Thus, when the models for $\mathbb{E}[Y(a,m)|x,z]$ and $\mathbb{E}[z|x,a]$ are both correctly specified, all coefficient estimates are consistent. It follows that $\widehat{\text{CDE}}_{\text{RWR}}(a,a',m)$ is also consistent.

[7] In practice, this model can be specified more flexibly, for example, by including higher-order or interaction terms of x and a.

Appendix C: R Code for RWR

In this appendix, we illustrate the implementation of RWR in R for estimating the CDE of plough use on female labor force participation. Replication data can be found at Matthew Blackwell's Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VNXEM6

```
library(foreign)
library(dplyr)
# load data
ploughs <- read.dta("crosscountry_dataset.dta")
# center log income at its mean, select variables, and keep complete cases
ploughs <- ploughs %>% tbl_df() %>%
  mutate(centered_ln_inc = ln_income - mean(ln_income, na.rm = TRUE),
         centered_ln_incsq = centered_ln_inc^2) %>%
  select(women_politics, plow, centered_ln_inc, centered_ln_incsq,
         agricultural_suitability, tropical_climate, large_animals,
         political_hierarchies, economic_complexity, rugged,
         years_civil_conflict, years_interstate_conflict, oil_pc,
         european_descent, communist_dummy, polity2_2000, serv_va_gdp2000) %>%
  na.omit()
# a function returning residualized intermediate confounders
residualize <- function(y){
        residuals(lm(y ~ plow + agricultural_suitability + tropical_climate + large_animals +
        political_hierarchies + economic_complexity + rugged, data = ploughs))
        }
# generate residualized intermediate confounders
ploughs <- ploughs %>%
        mutate_at(vars(years_civil_conflict:serv_va_gdp2000), funs(res = residualize))
# regression with residualized confounders
rwr_mod <- lm(women_politics ~ plow * centered_ln_inc + plow * centered_ln_incsq +
              agricultural_suitability +  tropical_climate +  large_animals +
              political_hierarchies + economic_complexity + rugged +
              years_civil_conflict_res +  years_interstate_conflict_res  + oil_pc_res +
              european_descent_res + communist_dummy_res + polity2_2000_res + serv_va_gdp2000_res +
              oil_pc_res * centered_ln_inc, data = ploughs)
```

References

Abramson, Scott F and David B Carter. 2016. "The Historical Origins of Territorial Disputes." *American Political Science Review* 110(4):675–698.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3):512–529.

Alesina, Alberto, Paola Giuliano and Nathan Nunn. 2013. "On the Origins of Gender Roles: Women and the Plough." *The Quarterly Journal of Economics* 128(2):469–530.

Almirall, Daniel, Thomas Ten Have and Susan A Murphy. 2010. "Structural Nested Mean Models for Assessing Time-Varying Effect Moderation." *Biometrics* 66(1):131–139.

Hall, Matthew EK. 2017. "Macro Implementation: Testing the Causal Paths from US Macro Policy to Federal Incarceration." *American Journal of Political Science* 61(2):438–455.

Holbein, John B. 2017. "Childhood Skill Development and Adult Political Participation." *American Political Science Review* 111(3):572–583.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4):765–789.

Joffe, Marshall M and Tom Greene. 2009. "Related Causal Frameworks for Surrogate Outcomes." *Biometrics* 65(2):530–538.

Knutsen, Carl Henrik, Andreas Kotsadam, Eivind Hammersmark Olsen and Tore Wig. 2017. "Mining and Local Corruption in Africa." *American Journal of Political Science* 61(2):320–334.

Pearl, Judea. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc. pp. 411–420.

Reese, Michael J, Keven G Ruby and Robert A Pape. 2017. "Days of Action or Restraint? How the Islamic Calendar Impacts Violence." *American Political Science Review* 111(3):439–459.

Robins, James. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period-Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7(9-12):1393–1512.

Robins, James M. 1997. "Causal Inference from Complex Longitudinal Data." *Latent Variable Modeling and Applications to Causality* pp. 69–117.

Robins, James M. 2003. "Semantics of Causal DAG models and the Identification of Direct and Indirect effects." *Highly Structured Stochastic Systems* pp. 70–81.

Robins, James M, Miguel Angel Hernan and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5):550–560.

VanderWeele, Tyler J. 2009. "Marginal Structural Models for the Estimation of Direct and Indirect effects." *Epidemiology* 20(1):18–26.

VanderWeele, Tyler J and Stijn Vansteelandt. 2009. "Conceptual Issues Concerning Mediation, Interventions and Composition." *Statistics and its Interface* 2(4):457–468.

Vansteelandt, Stijn. 2009. "Estimating Direct Effects in Cohort and Case–control Studies." *Epidemiology* 20(6):851–860.

Vansteelandt, Stijn and Marshall Joffe. 2014. "Structural Nested Models and g-estimation: The Partially Realized Promise." *Statistical Science* 29(4):707–731.

Wodtke, Geoffrey T and Daniel Almirall. 2017. "Estimating Moderated Causal Effects with Time-Varying Treatments and Time-Varying Moderators: Structural Nested Mean Models and Regression with Residuals." *Sociological Methodology* 47(1):212–245.

Zhu, Boliang. 2017. "MNCs, Rents, and Corruption: Evidence from China." *American Journal of Political Science* 61(1):84–99.