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

August 23, 2018

Abstract

Political scientists are increasingly interested in causal mediation, and to this end, recent studies focus on estimating a quantity called the controlled direct effect (CDE). The CDE measures the strength of the causal relationship between a treatment and outcome when a mediator is fixed at a given value. To estimate the CDE, Vansteelandt (2009) and Joffe and Greene (2009) developed the method of sequential g-estimation, which was introduced to political science by Acharya et al. (2016). In this letter, we propose an alternative method called "regression-with-residuals" (RWR) for estimating the CDE. In special cases, we show that these two methods are algebraically equivalent. Yet, unlike sequential g-estimation, RWR can easily accommodate several types of effect moderation, including cases in which the effect of the mediator on the outcome is moderated by a post-treatment 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 estimating the CDE of negative media framing on public support for immigration, controlling for respondent anxiety.

*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 Matthew Blackwell, Felix Elwert and Gary King for helpful comments on previous versions of this work. Replication data are available in Zhou and Wodtke (2018).

1 Introduction

Over the past decade, interest in causal mediation has rapidly grown in 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. Although the study of causal mediation often rests on strong and untestable assumptions (VanderWeele and Vansteelandt 2009; Imai et al. 2011), these assumptions are relatively weak when we focus on a quantity called the controlled direct effect (CDE) (Pearl 2001; Robins 2003). The CDE measures the strength of the causal relationship between a treatment and outcome when a mediator is fixed at a given value for all units. A nonzero CDE implies that the causal effect of treatment on the outcome does not operate exclusively through the mediator of interest. The difference between the total effect and CDE can also be interpreted as the degree to which the mediator contributes to a causal mechanism that transmits the effect of treatment to the outcome (Acharya et al. 2016, 2018).

Identification of the CDE is not straightforward. Simply conditioning on the mediator (via stratification, matching, or regression adjustment) is 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 media framing on support for immigration at a given level of anxiety (the mediator), post-treatment variables, such as beliefs about the economic or cultural impact of immigration, may affect both anxiety and support for immigration (Imai and Yamamoto 2013). Following Acharya et al. (2016), we call these variables intermediate confounders. Intermediate confounders pose a dilemma for the identification and estimation of CDEs when they are affected by treatment. In this situation, omitting intermediate confounders would lead to bias in the estimated effects of the mediator on the outcome, and by extension, in estimates of the CDE. However, controlling for intermediate confounders using conventional regression or matching methods would also engender bias in estimates of the CDE because it would block causal pathways, and unblock noncausal pathways, from treatment to the outcome, which would also lead to bias in estimates of the CDE.

Fortunately, several approaches overcome this dilemma. First, we could estimate a model for the marginal mean of the potential outcomes under different levels of the treatment and mediator, known as a marginal structural model (MSM), using the method of inverse probability weighting (IPW) (Robins et al. 2000; VanderWeele 2009). This approach performs best when both the treatment and mediator are binary. When the treatment and/or mediator are continuous, it performs poorly because the weights involve conditional density estimates that are typically unreliable. Second, to overcome these limitations, we could instead estimate a structural nested mean model (SNMM) for the conditional mean of the potential outcomes given a set of both pretreatment and intermediate confounders using the method of sequential g-estimation (Vansteelandt 2009; Joffe and Greene 2009). This approach, however, is difficult 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 Acharya et al. (2016) 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 complications, 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 letter, we introduce an alternative method, termed "regression-with-residuals" (RWR), for estimating the CDE. Compared with sequential g-estimation, it is relatively easy to implement, even in the presence of intermediate interactions. In the absence of such interactions, we show that RWR is algebraically equivalent to sequential g-estimation. We illustrate RWR by reanalyzing data from a survey experiment conducted by Brader et al. (2008) to estimate the CDE of negative media framing on support for immigration while controlling for the level of anxiety triggered by negative media cues.

2 Notation, Assumptions, and Sequential G-estimation

We use A to denote 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. The CDE is defined as the *average* effect of changing treatment from a to a' while fixing the mediator at a given level m:

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

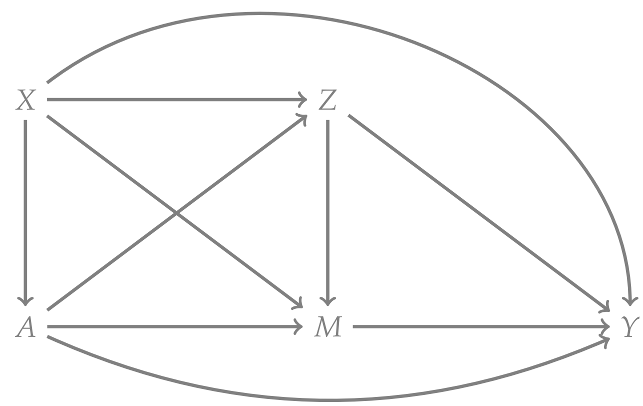

Figure 1: Causal Relationships under Sequential Ignorability Shown in Direct Acyclic Graph.
Note: A denotes the treatment, M denotes the mediator, Y denotes the outcome, X denotes pre-treatment confounders, Z denotes intermediate confounders.

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

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

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

Here, X denotes a vector of observed pretreatment confounders, while Z denotes a vector of observed intermediate confounders that affect both the mediator and 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. In this figure, Z is affected by A, and thus it is both an intermediate confounder and also a mediator. Because we focus on the CDE controlling for M only, we henceforth refer to Z exclusively as a confounder for clarity. Of course, it is possible to define a CDE controlling for M and Z jointly, which would illuminate the mediating role of both variables taken together. Estimands involving multiple mediators, however, are beyond the scope of this letter, although the methods we consider below can be generalized for more complex analyses of this type.

The CDE is distinct from several other estimands considered in analyses of causal mediation. For example, it is distinct from the average direct effect (ADE) considered in Imai et al. (2011), which is defined as

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

where $M(a)$ denotes the potential outcome for the mediator under treatment a. In contrast to the CDE, the ADE represents the average effect of changing treatment from a to a' while fixing the mediator for each unit at its value under treatment a. The ADE is equal to the difference between the total effect and the average causal mediation effect (ACME), which is defined as

$$\text{ACME}(a, a') = \mathbb{E}[Y(a', M(a)) - Y(a', M(a'))].$$

In general, the CDE differs from the ADE, and thus the difference between the total effect and CDE differs from the ACME, as long as the unit-level direct effect $Y(a, m) - Y(a', m)$ depends on *m for some units*. We focus on the CDE because it is identified under much weaker assumptions than the ADE and ACME. In particular, the CDE can still be identified in the presence of intermediate confounders affected by treatment, unlike the ADE and ACME (VanderWeele and Vansteelandt 2009).

Although the CDE is identified under sequential ignorability, additional modeling assumptions are needed to estimate the CDE in finite samples. Sequential g-estimation, for example, relies on a linear model for the conditional mean of the outcome given A, M, X, and Z. Moreover, because sequential g-estimation is difficult to implement in the presence of intermediate interactions, its application in practice also typically relies on an additional simplifying assumption that the effect of the mediator on the outcome is not moderated by intermediate confounders, which can be formally expressed as follows:

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

In words, this assumption states that among the units exposed to treatment a, the effect of the mediator on the outcome would not differ across the subgroups defined by the post-treatment confounders within levels of the pretreatment confounders.

Under this assumption, Acharya et al. (2016) illustrate sequential g-estimation of the CDE using the following model for the outcome:

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

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

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

This estimator is consistent under the assumptions of sequential ignorability and a correctly specified linear model for the outcome, which here requires that there must not be any effect moderation by the intermediate confounders. Standard errors can be obtained via the nonparametric bootstrap or a consistent variance estimator derived in Acharya et al. (2016).

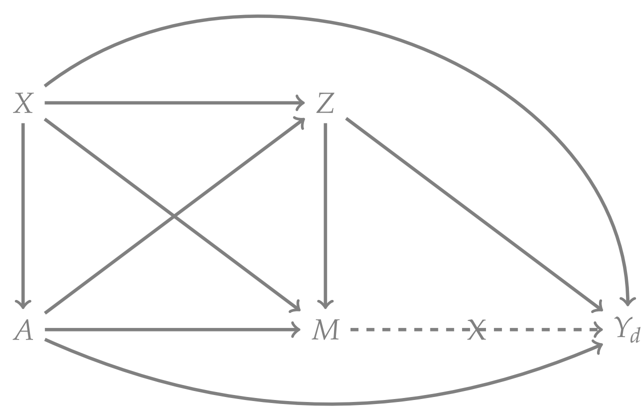

Figure 2: The Logic of Sequential G-estimation.
Note: A denotes the treatment, M denotes the mediator, Y denotes the outcome, X denotes pretreatment confounders, Z denotes intermediate confounders.

In Figure 2, we illustrate the logic of sequential g-estimation. First, under the identification and modeling assumptions outlined previously, 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 all other causal paths remain 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).

3 Regression-with-Residuals Estimation

RWR estimation was originally developed to assess how time-varying covariates moderate the effect of time-varying treatments (Almirall et al. 2010; Wodtke and Almirall 2017). In this section, we show how RWR can be adapted to estimate CDEs while properly adjusting for intermediate confounders. Specifically, RWR estimation of the CDE based on a model without intermediate interactions, such as equation (1), 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 shown in Supplementary Material A, RWR and sequential g-estimation are algebraically equivalent (i.e., $\hat{\kappa}_2 = \tilde{\beta}_2$; $\hat{\gamma}_2 = \tilde{\gamma}_2$) when there are no intermediate interactions. They rely on the same identification and modeling assumptions, and they share the same statistical properties.

In Figure 3, we illustrate the logic of RWR. First, residualizing the intermediate confounders in step 1 neutralizes the causal paths emanating from X and A to Z. Then, the residualized confounders can be included in an outcome regression to adjust for mediator-outcome confounding while avoiding the bias that normally results from conditioning on post-treatment variables. RWR estimation avoids post-treatment bias because $Z_{\perp}$ is no longer a consequence of A, and it avoids omitted variable bias because all confounders have been appropriately controlled in a model for the outcome.

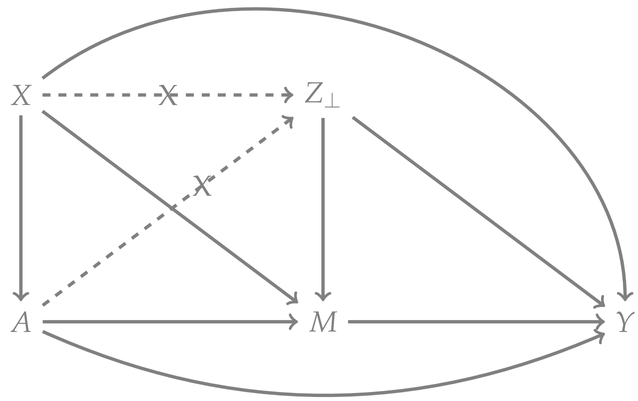

Figure 3: The Logic of Regression-with-residuals.
Note: A denotes the treatment, M denotes the mediator, Y denotes the outcome, X denotes pretreatment confounders, Z denotes intermediate confounders.

4 Intermediate Interactions

In the models considered previously, the effect of the mediator on the outcome is assumed to be invariant across all intermediate confounders. This is a strong and arguably implausible assumption in many social science applications. When it is not satisfied, estimates of the CDE may be biased and inconsistent. Thus, methods that accommodate, rather than naively assume away, intermediate interactions will make analyses of causal mediation more robust. The main advantage of RWR over sequential g-estimation is the ease with which RWR can accommodate intermediate interactions (Wodtke et al. 2018). Consider the following model, which extends equation (1) by including an interaction term between M and Z:

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

With this model, sequential g-estimation can still be used to estimate the CDE 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.

Although sequential g-estimation can still be used to estimate the CDE at $m = 0$ in the presence of intermediate interactions, we can no longer estimate the CDE for a general m using equation (2). This is because $\hat{\gamma}_2 m(a-a')$ is no longer a consistent estimate of the treatment-mediator interaction effect, as the inclusion of the term $\gamma_3^T z$ in equation (4) leads to post-treatment bias in $\gamma_2 a$. In other words, the de-mediation step only removes post-treatment bias in $\beta_2 a$ but not in $\gamma_2 a$.[1]

RWR estimation, by contrast, easily accommodates intermediate interactions, 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

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

where $\tilde{\gamma}_2$ remains a consistent estimate of the treatment-mediator interaction effect.[2]

As shown in Supplementary Material B, equation (5) is a consistent estimator of the CDE under the assumptions of sequential ignorability and no model misspecification. RWR estimation remains consistent even in the presence of intermediate interactions because, by appropriately residualizing the intermediate confounders, it removes any post-treatment bias from the main effect of treatment and from the treatment-mediator interaction effect. RWR can also accommodate "baseline interactions" between treatment A and the pretreatment confounders X. In this situation, we need only recenter the pretreatment confounders at their sample means and then include the appropriate interaction terms in the outcome regression. Standard errors can be computed using the nonparametric bootstrap.

[1] One way to circumvent this problem would be to recenter the mediator at different levels and then re-implement the sequential g-estimator. This approach, however, can be tedious when evaluating the CDE at a wide range of mediator values. Moreover, this approach does not allow us to directly estimate the treatment-mediator interaction effect, which is often of immediate scientific interest in analyses of causal mediation.

[2] In previous work (Almirall et al. 2010; Wodtke and Almirall 2017), where RWR has been used to estimate the moderated 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, 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.

5 The Effect of Media Framing on Support for Immigration

To illustrate RWR, we reanalyze data from Brader et al. (2008) to estimate the CDE of negative media framing on public support for immigration, controlling for respondent anxiety potentially triggered by negative media cues.[3] With a nationally representative sample of 354 white non-Hispanic adults, Brader et al. (2008) conducted a survey experiment in which respondents were asked to read a mock news report on immigration. In this report, both the ethnicity of the featured immigrant and the tone of the story were randomly manipulated using a 2×2 design. Specifically, respondents were presented with a story that featured either a white European immigrant or a Latino immigrant and that focused on either the benefits or the costs of immigration. After reading the story, respondents were asked to report their beliefs about the harms of immigration, their feelings about increased immigration, and their support for immigration. With these data, Brader et al. (2008) found that stories featuring both a Latino immigrant and a negative frame emphasizing the costs of immigration had a large negative effect on support for immigration. They also reported that a substantial proportion of this effect is mediated by respondents' anxiety about increased immigration and that beliefs about the harms of immigration, as opposed to negative emotions, do not play an important mediating role. However, Brader et al. (2008) assessed the mediating role of beliefs and emotions separately under the assumption that respondent anxiety is not affected by perceptions of the harms associated with immigration, which seems unlikely and appears to be inconsistent with their own data (Imai and Yamamoto 2013). Thus, we treat beliefs about the harm of immigration as an intermediate confounder and reassess the mediating role of respondent anxiety using RWR and, for comparative purposes, sequential g-estimation.

Specifically, we estimate the CDE of negative media framing on support for immigration, controlling for respondent anxiety, using several variants of the following model:

$$\mathbb{E}[Y|X = x, A = a, Z = z, M = m] = \beta_0 + \beta_1^T x + \beta_2 a + \beta_3 z + m(\gamma_0 + \gamma_1^T x + \gamma_2 a), \quad (6)$$

where the outcome, Y, is a measure of support for immigration on a five-point scale; the treatment, A, denotes receipt of a news story featuring both a Latino immigrant and a negative frame emphasizing the costs of immigration; the mediator, M, is the level of anxiety expressed by the respondent on a ten-point scale; the intermediate confounder, Z, is a measure of the perceived

[3] Replication data are available in Zhou and Wodtke (2018).

Table 1: Estimated CDE of Media Framing on Support for Immigration using Sequential G-estimation, RWR, and RWR with Intermediate Interactions

	Total Effect	Naive Regression	Sequential g-estimation (final step)	RWR	RWR with intermediate interactions
intercept	1.08	0.99	1.03	1.03	0.96
	(0.07)	(0.06)	(0.06)	(0.06)	(0.07)
negative Latino framing	-0.42	-0.21	-0.33	-0.33	-0.31
(i.e., $\widehat{\text{CDE}}(a, a+1, 0)$)	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)
level of anxiety		-0.19		-0.10	-0.12
		(0.02)		(0.03)	(0.03)
negative Latino framing * level of anxiety		0.06		0.06	0.07
		(0.04)		(0.04)	(0.04)
perceived harm				-0.20	-0.18
				(0.04)	(0.04)
perceived harm * level of anxiety					0.02
					(0.01)

Note: Numbers in parentheses are bootstrapped standard errors (500 replications). For ease of interpretation, all predictors except the treatment are centered at their means. Coefficients of pre-treatment covariates are omitted. Supplementary Material C presents the R code used to generate the results.

harm of immigration on a seven-point scale; and finally, the vector of pretreatment covariates, X, includes measures of gender, age, education, and income.[4] We control for a set of pretreatment covariates because, although treatment is randomly assigned, the mediator-outcome relationship may still be confounded by baseline factors in these data. To simplify interpretation, all variables except the treatment and outcome are centered at their sample means.

As a benchmark, the first two columns of Table 1 present an estimate of the total treatment effect from a regression of Y on X and A as well as a "naive" estimate of the CDE from a regression model similar to equation (6) but without adjustments for the intermediate confounder Z. Consistent with results reported by Brader et al. (2008), the estimated total effect indicates that negative media framing reduces support for immigration, and the naive estimate of the CDE suggests that about half of the total treatment effect is due to heightened anxiety.

The third and fourth columns of Table 1 present sequential g-estimates and RWR estimates, respectively, for the CDE based on model (6). As expected, the estimates given by these two

[4]Detailed definitions of these variables can be found in Brader et al. (2008).

methods are exactly the same (-0.33) because there are no intermediate interactions in this model. Contrary to the naive estimate of the CDE discussed previously, these results suggest that less than one-quarter of the total treatment effect may be due to heightened respondent anxiety. This finding is consistent with estimates of the ADE reported by Imai and Yamamoto (2013).

With sequential g-estimation, only the CDE at $m = 0$ (i.e., when the level of respondent anxiety is set at its sample mean) is reported in the final step. To construct the CDE at other levels of the mediator, the analyst must extract the coefficient on the treatment-mediator interaction from the regression in step 1 of the procedure. With RWR, by contrast, all the coefficients required for constructing the CDE at any level of the mediator are reported in the single regression for the outcome. This allows an analyst to construct any CDE of interest directly from the results in Table 1. For example, when respondent anxiety is one standard deviation (2.77) above the sample mean, the CDE is estimated to be $(-0.33 + 2.77 * 0.064) = -0.15$.

Thus far, the effect of anxiety on support for immigration has been assumed to be invariant across levels of the intermediate confounder, but if this effect is in fact moderated by beliefs about the harms of immigration, then estimates reported previously may be biased. We now relax this assumption by additionally including an interaction term between the level of anxiety and the perceived harm of immigration when implementing RWR. Results from this analysis are shown in the last column of Table 1. The estimated CDE from this model at $m = 0$ is -0.31, which is similar to that obtained from the model without intermediate interactions. In this example, it appears that our findings are fairly robust to the exclusion of intermediate interactions. Nevertheless, it is the flexibility of RWR that allows us to easily assess the sensitivity of results to different specifications.

6 Concluding Remarks

In this letter, we introduced RWR for estimating controlled direct effects. In the absence of intermediate interactions, RWR is algebraically equivalent to the sequential g-estimator. But unlike the sequential g-estimator, RWR can easily accommodate several different types of effect moderation, including intermediate interactions, which are likely common in the social sciences. In general, models with less stringent parametric constraints can be estimated more easily with RWR than with sequential g-estimation.

Nevertheless, RWR is still premised on a number of strong modeling assumptions. In particular, RWR requires a correctly specified linear model for the outcome. In applications with many confounders or complex patterns of effect heterogeneity, the modeling assumptions required of RWR may be difficult to satisfy, and when they are violated, RWR is biased. Moreover, in applications where a linear model may be inappropriate (e.g., in analyses with binary outcomes), RWR does not generalize in a straightforward manner for use with nonlinear models. Thus, semi-parametric methods, such as IPW estimation of MSMs or certain types of sensitivity analysis (e.g., Imai and Yamamoto 2013), may be preferable in applications with a large number of confounders, complex effect heterogeneity, or categorical outcomes.

These limitations notwithstanding, simulation studies indicate that g- and RWR estimation can still outperform IPW estimation even when the outcome model is misspecified, especially in applications with continuous treatments and/or mediators (Vansteelandt 2009; Wodtke 2018). RWR estimation can also be combined with a sensitivity analysis to assess the robustness of estimates to different violations of its motivating assumptions. Given its simplicity, flexibility, and relative efficiency, we expect that RWR will be widely used in causal mediation analyses.

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) in the main text and write the "naive" least squares regression of it as

$$Y = \hat{\beta}_0 + \hat{\beta}_1^T X + \hat{\beta}_2 A + \hat{\beta}_3^T Z + M(\hat{\gamma}_0 + \hat{\gamma}_1^T X + \hat{\gamma}_2 A) + Y_\perp, \tag{7}$$

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:

$$Z = \hat{\lambda}_0 + \hat{\Lambda}_1 X + \hat{\lambda}_2 A + Z_\perp, \tag{8}$$

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 (8) into equation (7), we have

$$Y = (\hat{\beta}_0 + \hat{\beta}_3^T \hat{\lambda}_0) + (\hat{\beta}_1^T + \hat{\beta}_3^T \hat{\Lambda}_1) X + (\hat{\beta}_2 + \hat{\beta}_3^T \hat{\lambda}_2) A + \hat{\beta}_3^T Z_\perp + M(\hat{\gamma}_0 + \hat{\gamma}_1^T X + \hat{\gamma}_2 A) + Y_\perp. \tag{9}$$

Since $Y_\perp$ is the least squares residual for regression (7), 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 (9) 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 (9), we also know that the de-mediated outcome can be written as

$$Y_d = (\hat{\beta}_0 + \hat{\beta}_3^T \hat{\lambda}_0) + (\hat{\beta}_1^T + \hat{\beta}_3^T \hat{\Lambda}_1) X + (\hat{\beta}_2 + \hat{\beta}_3^T \hat{\lambda}_2) A + \hat{\beta}_3^T Z_\perp + Y_\perp. \tag{10}$$

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 (10) 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.

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. Consider the following SNMM, which is analogous to the observed data regression in equation (7) except that it is a model for the potential outcomes:

$$\mathbb{E}[Y(a, m)|X, A = a, Z] = \beta_0 + \beta_1^T X + \beta_2 a + \beta_3^T Z + m(\gamma_0 + \gamma_1^T X + \gamma_2 a). \tag{11}$$

With the sequential g-estimator, the least squares regression in step 3 implies the linearity of $\mathbb{E}[Y(a, 0)|X, A = a]$ in X and a:

$$\mathbb{E}[Y(a, 0)|X, A = a] = \kappa_0 + \kappa_1^T X + \kappa_2 a. \tag{12}$$

Setting $m = 0$ in model (11), we have

$$\mathbb{E}[Y(a, 0)|X, A = a, Z] = \beta_0 + \beta_1^T X + \beta_2 a + \beta_3^T Z. \tag{13}$$

Taking the expectation of equation (13) over Z, conditional on X, yields

$$\mathbb{E}[Y(a, 0)|X, A = a] = \beta_0 + \beta_1^T X + \beta_2 a + \beta_3^T \mathbb{E}[Z|X, A = a]. \tag{14}$$

Comparing equations (12) and (14), we see that $\beta_3^T \mathbb{E}[Z|X, A = a]$ must be linear in X and a. Since β_3 represents model parameters that can vary freely in $\mathbb{R}^q$, the linearity of $\beta_3^T \mathbb{E}[Z|X, A = a]$ implies that each component of $\mathbb{E}[Z|X, A = a]$ must be linear in X and a. Conversely, when each component of $\mathbb{E}[Z|X, A = a]$ is linear in X and a, model 11 implies equation (12). Thus, the sequential g-estimator implicitly assumes each component of $\mathbb{E}[Z|X, A = a]$ is 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 $\mathbb{E}[Z|X, A = a]$,

$$\mathbb{E}[Z|X, A = a] = \lambda_0 + \Lambda_1 X + \lambda_2 a, \tag{15}$$

although this model can be specified more flexibly in practice by, for example, including higher-order or interaction terms involving X and a. Similar to equation (8) in Appendix A, λ_0 and λ_2 are both $q \times 1$ vectors and Λ_1 is a $q \times p$ matrix.

Next, to see the consistency of the RWR estimator in the presence of intermediate interactions, consider the following SNMM:

$$\mathbb{E}[Y(a, m)|X, A = a, Z] = \beta_0 + \beta_1^T X + \beta_2 a + \beta_3^T Z + m(\gamma_0 + \gamma_1^T X + \gamma_2 a + \gamma_3^T Z). \tag{16}$$

Given equation (15), the CDE can be expressed as

$$\begin{aligned}
\mathbb{E}[Y(a, m) - Y(a', m)] &= \mathbb{E}_X \mathbb{E}[Y(a, m)|X, A = a] - \mathbb{E}_X \mathbb{E}[Y(a', m)|X, A = a'] \quad (\text{because } Y(a, m) \perp\!\!\!\perp A|X, \forall a, m) \\
&= \mathbb{E}_X \mathbb{E}_{Z|X,A=a} \mathbb{E}[Y(a, m)|X, A = a, Z] - \mathbb{E}_X \mathbb{E}_{Z|X,A=a'} \mathbb{E}[Y(a', m)|X, A = a', Z] \\
&= \beta_2(a - a') + \gamma_2 m(a - a') + \beta_3^T \cdot \mathbb{E}_X[\mathbb{E}[Z|X, A = a] - \mathbb{E}[Z|X, A = a']] \\
&\quad + \gamma_3^T m \cdot \mathbb{E}_X[\mathbb{E}[Z|X, A = a] - \mathbb{E}[Z|X, A = a']] \\
&= \beta_2(a - a') + \gamma_2 m(a - a') + \beta_3^T \lambda_2 (a - a') + \gamma_3^T \lambda_2 m(a - a') \\
&= [(\beta_2 + \beta_3^T \lambda_2) + (\gamma_2 + \gamma_3^T \lambda_2)m](a - a')
\end{aligned}$$

It is easy to show that the RWR estimator based on model (4) is equal to

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

Thus, when linear models for both $\mathbb{E}[Y(a, m)|X, A = a, Z]$ and $\mathbb{E}[Z|X, A = a]$ are correctly specified and the potential outcomes are sequentially ignorable, all coefficient estimates are consistent. It follows that $\widehat{\text{CDE}}_{\text{RWR}}(a, a', m)$ is also consistent.

C: R Code for RWR

In this appendix, we illustrate the implementation of RWR in R for estimating the CDE of media framing on support for immigration. Replication data can be found at Teppei Yamamoto's Dataverse: https://hdl.handle.net/1902.1/19036

```
library(dplyr)
# load data
load("PA-ImaiYamamoto.RData")
# function for demeaning
demean <- function(x) x - mean(x, na.rm = TRUE)
# function for residualizing intermediate confounders
residualize <- function(formula, df) residuals(lm(formula, df))
# data preprocessing
Brader2 <- Brader %>%
  select(immigr, emo, p_harm, tone_eth, ppage, ppeducat, ppgender, ppincimp) %>% na.omit() %>%
  mutate(immigr = 4 - immigr,
         hs = (ppeducat == "high school"),
         sc = (ppeducat == "some college"),
         ba = (ppeducat == "bachelor's degree or higher"),
         female = (ppgender == "female")) %>%
  mutate_at(vars(emo, p_harm, ppage, female, hs, sc, ba, ppincimp), demean) %>%
  mutate(., p_harm_res = residualize(p_harm ~ ppage + female + hs + sc + ba + ppincimp + tone_eth, .))
# total effect model
total_mod <-  lm(immigr ~ ppage + female + hs + sc + ba + ppincimp + tone_eth,
                 data = Brader2)
# rwr without intermediate interactions
rwr1_mod <-  lm(immigr ~  ppage + female + hs + sc + ba + ppincimp + tone_eth + p_harm_res +
                emo * ( ppage + female + hs + sc + ba + ppincimp + tone_eth),
                data = Brader2)
# rwr with intermediate interactions
rwr2_mod <-  lm(immigr ~  ppage + female + hs + sc + ba + ppincimp + tone_eth + p_harm_res +
                emo * (ppage + female + hs + sc + ba + ppincimp + tone_eth + p_harm_res),
                data = Brader2)
# bootstrap
nboots <- 500
rwr1_hold <- matrix(NA, nrow = length(coef(rwr1_mod)), ncol = nboots)
rwr2_hold <- matrix(NA, nrow = length(coef(rwr2_mod)), ncol = nboots)
for (i in 1:nboots) {
```

```r
    star <- sample(1:nrow(Brader2), replace = TRUE)
    Brader2_star <- Brader2[star, ]
    Brader2_star <- Brader2_star %>% tbl_df() %>%
    mutate(., p_harm_res = residualize(p_harm ~ ppage + female + hs + sc + ba + ppincimp + tone_eth, .))
    rwr1_star <-  lm(immigr ~  ppage + female + hs + sc + ba + ppincimp + tone_eth + p_harm_res +
                  data = Brader2_star)
    rwr2_star <-  lm(immigr ~  ppage + female + hs + sc + ba + ppincimp + tone_eth + p_harm_res +
                  data = Brader2_star)
    rwr1_hold[, i] <- coef(rwr1_star)
    rwr2_hold[, i] <- coef(rwr2_star)
}
rownames(rwr1_hold) <- names(coef(rwr1_mod))
rownames(rwr2_hold) <- names(coef(rwr2_mod))

out_coefs <- c("(Intercept)", "tone_eth", "emo", "tone_eth:emo", "p_harm_res", "p_harm_res:emo")
rwr1_est <- coef(rwr1_mod)[out_coefs]
rwr2_est <- coef(rwr2_mod)[out_coefs]
rwr1_se <- apply(rwr1_hold, 1, sd)[out_coefs]
rwr2_se <- apply(rwr2_hold, 1, sd)[out_coefs]
```

References

Acharya, A., M. Blackwell, and M. Sen (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review 110*(3), 512–529.

Acharya, A., M. Blackwell, and M. Sen (2018). Analyzing causal mechanisms in survey experiments. *Political Analysis*.

Almirall, D., T. Ten Have, and S. A. Murphy (2010). Structural nested mean models for assessing time-varying effect moderation. *Biometrics 66*(1), 131–139.

Brader, T., N. A. Valentino, and E. Suhay (2008). What triggers public opposition to immigration? anxiety, group cues, and immigration threat. *American Journal of Political Science 52*(4), 959–978.

Imai, K., L. Keele, D. Tingley, and T. 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.

Imai, K. and T. Yamamoto (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis 21*(2), 141–171.

Joffe, M. M. and T. Greene (2009). Related causal frameworks for surrogate outcomes. *Biometrics 65*(2), 530–538.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann Publishers Inc.

Robins, J. (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, J. M. (1997). Causal inference from complex longitudinal data. *Latent Variable Modeling and Applications to Causality*, 69–117.

Robins, J. M. (2003). Semantics of causal dag models and the identification of direct and indirect effects. *Highly Structured Stochastic Systems*, 70–81.

Robins, J. M., M. A. Hernan, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*(5), 550–560.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology 20*(1), 18–26.

VanderWeele, T. J. and S. Vansteelandt (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface 2*(4), 457–468.

Vansteelandt, S. (2009). Estimating direct effects in cohort and case–control studies. *Epidemiology 20*(6), 851–860.

Wodtke, G. T. (2018). Regression-based adjustment for time-varying confounders. *Sociological Methods & Research*.

Wodtke, G. T., Z. Alaca, and X. Zhou (2018). Regression-with-residuals estimation of marginal effects: A method of adjusting for treatment-induced confounders that may also be moderators. URL: `https://arxiv.org/abs/1808.07795`.

Wodtke, G. T. and D. 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.

Zhou, X. and G. T. Wodtke (2018). Replication data for: A regression-with-residuals method for estimating controlled direct effects. https://doi.org/10.7910/DVN/BPU1XG, Harvard Dataverse, v1.