Running Head: Regression-with-residuals for Marginal Effects

REGRESSION-WITH-RESIDUALS ESTIMATION OF MARGINAL EFFECTS: A METHOD OF ADJUSTING FOR TREATMENT-INDUCED CONFOUNDERS THAT MAY ALSO BE EFFECT MODIFIERS

Geoffrey T. Wodtke

University of Toronto

Zahide Alaca

University of Toronto

Xiang Zhou

Harvard University

Corresponding Author: Geoffrey T. Wodtke, Department of Sociology, University of Toronto, 725 Spadina Avenue, Toronto, ON M5S 2J4, Canada. Email: geoffrey.wodtke@utoronto.ca

Acknowledgements: This research was supported by an Insight Grant from the Social Sciences and Humanities Research Council of Canada (Grant No. 435-2018-0736).

Abstract

When making causal inferences, treatment-induced confounders complicate analyses of time-varying treatment effects. Conditioning on these variables naively to estimate marginal effects may inappropriately block causal pathways and may induce spurious associations between treatment and the outcome, leading to bias. Although several methods for estimating marginal effects avoid these complications, including inverse-probability-of-treatment-weighted estimation of marginal structural models as well as g- and regression-with-residuals estimation of highly constrained structural nested mean models, each suffers from a set of nontrivial limitations, among them an inability to accommodate effect modification. In this study, we adapt the method of regression-with-residuals to estimate marginal effects with a set of moderately constrained structural nested mean models that easily accommodate several types of treatment-by-confounder interaction. With this approach, the confounders at each time point are first residualized with respect to the observed past and then the outcome is regressed on all prior variables, including a set of treatment-by-confounder interaction terms, with the residual terms substituted for the untransformed confounders both as “main effects” and as part of any interaction terms. Through a series of simulation experiments and empirical examples, we show that this approach outperforms other methods for estimating the marginal effects of time-varying treatments.

Keywords: treatment-induced confounding, structural nested mean models, regression-with-residuals, g-estimation, marginal structural models, inverse probability of treatment weighting

1. Introduction

In analyses of time-varying treatment effects, social scientists must often contend with the complications posed by treatment-induced confounders (e.g., Acharya et al. 2016; Elwert and Winship 2014; Wodtke et al. 2011). A treatment-induced confounder is a variable that is affected by a prior treatment and affects both selection into future treatment and the outcome. For example, we consider whether living in a disadvantaged neighborhood throughout childhood and adolescence affects academic achievement (e.g., Sampson et al. 2008; Wodtke et al. 2011, 2016). In studies of neighborhood effects, parental income is likely affected by prior neighborhood conditions and also likely affects both future residential choices and child educational outcomes.

If left uncontrolled, treatment-induced confounders lead to bias in estimates of marginal effects, such as the cumulative treatment effect (CTE) in analyses of time-varying treatments. At the same time, adjusting naively for treatment-induced confounders by, for example, including them as predictors in a conventional regression model or matching on them via the propensity score also leads to bias. Specifically, conditioning on a treatment-induced confounder with conventional regression or matching methods leads to bias from over-control of intermediate pathways and endogenous selection (Elwert and Winship 2014; Robins et al. 2000; VanderWeele 2015). Thus, even if all relevant confounders are observed, which is a necessary condition for drawing causal inferences from any non-experimental study, treatment-induced confounders pose additional challenges for the most common approaches to covariate adjustment. Alternative methods are therefore required when estimating marginal effects in the presence of these variables.

Fortunately, there are several methods that avoid the complications outlined previously and that are capable of consistently estimating marginal effects, even when adjustment is

required for treatment-induced confounders. These include inverse-probability-of-treatment-weighted (IPTW) estimation of marginal structural models (MSMs; Robins et al. 1994, 2000), g-estimation of highly constrained structural nested mean models (SNMMs; Naimi et al. 2017; Vansteelandt 2009; Vansteelandt and Sjolander 2016), and regression-with-residuals (RWR) estimation of highly constrained SNMMs (Wodtke 2018).

Each of these methods, however, suffers from a set of nontrivial limitations. IPTW estimation is relatively inefficient, is difficult to use with continuous treatments, and may suffer from finite-sample bias when confounders strongly predict treatment (Lunceford and Davidian 2004; Naimi et al. 2014; Robins et al. 1994). G- and RWR estimation of highly constrained SNMMs for marginal effects avoid the limitations of IPTW, but they are premised on the strong assumption of no effect modification (e.g., Vansteelandt 2009; Wodtke 2018), which is unrealistic in most social science applications.[1] If, for example, a treatment-induced confounder also modifies the effect of a future treatment on the outcome, then these methods suffer from model misspecification bias. Because effect modification is ubiquitous in the social sciences (Morgan and Winship 2015; Xie 2007), this assumption may limit the utility of these methods in practice.

In this study, we adapt the method of RWR to estimate a set of moderately constrained SNMMs for marginal effects that accommodate several types of treatment-by-confounder interaction. Briefly, RWR estimation of marginal effects in a moderately constrained SNMM proceeds in two stages. First, the confounders at each time point are regressed on all prior variables and then residualized. Second, the outcome is regressed on all prior variables, including

[1] Effect modification is also sometimes referred to as *effect moderation* (e.g., Almirall et al. 2010; Wodtke and Almirall 2017).

a set of treatment-by-confounder interaction terms, with the residuals from the first stage substituted for the untransformed confounders both as "main effects" and as part of the interaction terms. Our adaptation differs from previous implementations of RWR (e.g., Almirall et al. 2010; Wodtke and Almirall 2017; Wodtke 2018) by additionally including the residualized confounders in interaction terms with treatment, which accommodates several types of effect modification while neatly isolating the marginal effects of interest in a single, possibly vector-valued, parameter.

Under the assumptions of sequential ignorability and no model misspecification, the proposed method is consistent for marginal effects, like the CTE, even in the presence of treatment-induced confounders. It avoids the biases that arise with naive adjustments for treatment-induced confounders because the residualized confounders are purged of their association with prior treatment and thus including them in a regression model for the outcome is unproblematic. In addition, because it does not involve weighting by a function of the conditional probability of treatment, the proposed method avoids the limitations associated with IPTW estimation. Finally, because it accommodates several types of treatment-by-confounder interaction, the proposed method also mitigates the limitations associated with both g- and RWR estimation of marginal effects using a highly constrained SNMM in which effect modification is assumed away entirely.

In the sections that follow, we begin by considering the problem of estimating marginal effects for a time-varying treatment. First, we formally define the effects of interest, explain when they are identified from observed data, and illustrate the problems that afflict conventional estimation methods in the presence of treatment-induced confounding. Second, we present an SNMM for the conditional, rather than marginal, effects of treatment, but we then show how

these conditional effects can be additively decomposed into a set of functions that capture the marginal effects of interest and another set of functions that capture effect modification. Third, we show how to appropriately parameterize these functions and adapt the method of RWR to estimate marginal effects with an SNMM under this alternative parameterization. Next, we briefly consider analyses of causal mediation and the problem of estimating controlled direct effects in the presence of mediator-outcome confounders that are affected by treatment, which we show can be accomplished with the same methods used for estimating marginal effects in the time-varying setting. Finally, with a series of simulation experiments and empirical examples, we illustrate several applications of our proposed method and show that it outperforms other common approaches.

2. Background

2.1. Notation, Estimands, and Identification

In this section, we formally define the marginal, or population average, effects of interest and explain when they can be identified from observed data, drawing on the potential outcomes framework (Holland 1986; Rubin 1974) and directed acyclic graphs (DAGs; Pearl 2009) throughout. For expositional clarity, we focus on a simplified example with a binary treatment measured at two time points, a binary confounder measured at two time points, and a continuous outcome measured at the end of follow-up, although these methods can be easily adapted for more complex analyses.

First, let $a_t = 1$ denote exposure to treatment, and $a_t = 0$ denote the absence of treatment, at time $t \in \{1,2\}$. Second, let $Y_i(a_1, a_2)$ denote the potential outcome for subject i had she previously been exposed to the treatment sequence $\{a_1, a_2\}$. For example, $Y_i(0,0)$ is the

potential outcome for subject i had she never received treatment, $Y_i(1,0)$ is her potential outcome had she received treatment only at time $t = 1$, and so on. In this framework, each subject is conceived to have a potential outcome corresponding to each of the four possible treatment sequences, but only the single potential outcome corresponding to the treatment sequence actually received is ever observed in reality, and the others are so-called "counterfactuals." Third, let C_{i1} denote the confounder for subject i measured just prior to treatment at time $t = 1$, and let $C_{i2}(a_1)$ denote the confounder for subject i measured just before treatment at time $t = 2$, which is indexed by a_1 as a potential outcome to reflect that it is affected by prior treatment. In other words, $C_{i2}(a_1)$ is a treatment-induced confounder. Finally, let the set $\{C_{i1}, A_{i1}, C_{i2}, A_{i2}, Y_i\}$ denote the observed data in temporal order.

In general, marginal effects are contrasts between different potential outcomes averaged over a population of individuals. Specifically, they give the average difference in the end-of-study outcome had everyone in the target population received one rather than another treatment sequence. With two time points, several different marginal effects may be of interest. The first is the distal treatment effect, or DTE, which can be formally defined as

$$DTE(a_1) = E\big(Y_i(a_1, 0) - Y_i(0,0)\big). \quad (1)$$

It gives the average effect of receiving treatment only at time $t = 1$ rather than never receiving treatment. The second is the proximal treatment effect, or PTE, which can be formally defined as

$$PTE(a_1, a_2) = E\big(Y_i(a_1, a_2) - Y_i(a_1, 0)\big). \quad (2)$$

When $a_1 = 0$, it gives the average effect of receiving treatment only at time $t = 2$ rather than never receiving treatment, and when $a_1 = 1$, it gives the average effect of always receiving treatment rather than receiving treatment only at time $t = 1$. The third is the cumulative treatment effect, or CTE. This effect is equal to the sum of $DTE(1)$ and $PTE(1,1)$,

$$CTE = DTE(1) + PTE(1,1) =$$

$$E\big(Y_i(1,0) - Y_i(0,0)\big) + E\big(Y_i(1,1) - Y_i(1,0)\big) = E\big(Y_i(1,1) - Y_i(0,0)\big), \quad (3)$$

which gives the average effect of being always versus never treated. Finally, the last is the interaction effect, or INE. This effect can be formally defined as

$$INE = PTE(1,1) - PTE(0,1) = E\big(Y_i(1,1) - Y_i(1,0)\big) - E\big(Y_i(0,1) - Y_i(0,0)\big), \quad (4)$$

which describes how the effect of receiving treatment at time $t = 2$ differs depending on whether an individual had previously received treatment at time $t = 1$.

All of these effects can be non-parametrically identified from the observed data under the assumptions of stable unit treatment values, consistency, positivity, and sequential ignorability (Robins et al. 1994, 2000; Rubin 1986). The stable unit treatment value assumption (SUTVA) requires that the potential outcomes for any given individual be unaffected by the mechanisms used to assign treatment status or by the treatments to which other individuals are exposed. The consistency assumption requires that the observed outcome Y_i be equal to $Y_i(a_1, a_2)$ whenever $A_{i1} = a_1$ and $A_{i2} = a_2$. The positivity assumption requires that there not be any subgroups within the target population that are treated or untreated with certainty at any time point. The sequential ignorability assumption requires that the potential outcomes are independent of treatment at each time point conditional on the observed past. Formally, this assumption can be expressed as

$$Y_i(a_1, a_2) \perp A_{i1} | C_{i1} \; \forall \; (a_1, a_2) \text{ and } Y_i(a_1, a_2) \perp A_{i2} | C_{i1}, A_{i1}, C_{i2} \; \forall \; (a_1, a_2), \quad (5)$$

where $\perp$ denotes statistical independence. It is satisfied when there are not any unobserved variables that directly affect both selection into treatment at each time point and the outcome.

Panel A of Figure 1 presents a DAG illustrating a set of causal relationships between the variables outlined previously in which the sequential ignorability assumption is satisfied. It

shows that both treatments, A_{i1} and A_{i2}, directly affect the outcome, Y_i, and that A_{i1} also indirectly affects the outcome through C_{i2}. In addition, it shows that C_{i1} confounds the effect of A_{i1} on Y_i and that C_{i2} confounds the effect of A_{i2} on Y_i. Treatment assignment is sequentially ignorable in this figure because treatment at each time point is not directly affected by any unobserved variables; rather, the only unobserved variables, denoted by U_{i1} and U_{i2}, directly affect the observed confounders and the outcome but not either treatment. The marginal effects outlined previously can be consistently estimated from the observed data by appropriately adjusting for all variables that directly affect both treatment and the outcome—in this case, C_{i1} and C_{i2}.

2.2. The Problem of Treatment-induced Confounding

Because C_{i2} is affected by A_{i1} and confounds the effect of A_{i2} on Y_i, it is a treatment-induced confounder. Treatment-induced confounders pose several challenges for estimating marginal effects of a time-varying treatment (Daniel et al. 2013). In particular, conventional methods of covariate adjustment, including conditioning, stratifying, or matching directly on a treatment-induced confounder, lead to several types of bias, even when the effects of interest are identified under sequential ignorability. At the same time, failing to appropriately adjust for a treatment-induced confounder also leads to bias. Thus, treatment-induced confounders seemingly present a "damned if you do and damned if you don't" dilemma with regard to covariate adjustment.

To appreciate this, first consider the causal graph in Panel B of Figure 1, and recall that a path in a DAG is "blocked" when it contains (a) an outcome of two or more variables, known as a collider, that has not been conditioned upon or (b) a non-collider that has been conditioned upon; otherwise, it is "unblocked" (Pearl 2009). Panel B of Figure 1 shows that conditioning

naively on the treatment-induced confounder, C_{i2}, blocks the causal pathway $A_{i1} \rightarrow C_{i2} \rightarrow Y_i$ emanating from treatment at time $t = 1$ to the outcome, which leads to bias from over-control of intermediate pathways. Next consider the causal graph in Panel C of Figure 1. This graph shows that conditioning naively on C_{i2} also unblocks the non-causal pathway $A_{i1} \rightarrow C_{i2} \leftarrow U_{i2} \rightarrow Y_i$ emanating from treatment at time $t = 1$ to the outcome, which leads to bias from so-called "endogenous selection" or "collider stratification" (Elwert and Winship 2014). Specifically, it shows that C_{i2} is a collider of A_{i1} and U_{i2}, and because U_{i2} affects Y_i, conditioning on C_{i2} induces a spurious association between treatment at time $t = 1$ and the outcome. Finally, consider the graph in Panel D of Figure 1. This graph shows that when C_{i2} has not been conditioned upon, the non-causal pathways emanating from treatment at time $t = 2$ to the outcome, $A_{i2} \leftarrow C_{i2} \rightarrow Y_i$ and $A_{i2} \leftarrow C_{i2} \leftarrow U_{i2} \rightarrow Y_i$, remain unblocked, which leads to bias from uncontrolled confounding. Thus, conventional methods of covariate adjustment inevitably lead to bias in estimates of marginal effects when there is treatment-induced confounding, and alternative methods are required.

3. Regression-with-residuals for Marginal Effects

3.1. Estimating the Marginal Effects of a Time-varying Treatment

A structural nested mean model (SNMM) is a model for the conditional, or subpopulation average, effects of a time-varying treatment given past confounder and treatment history (Almirall et al. 2010; Robins 1994; Robins et al. 2007; Wodtke and Almirall 2017). In this section, we show that conditional effects modeled with an SNMM can be additively decomposed into a set of functions that capture the marginal, or population average, effects of interest and another set of functions that capture effect modification, that is, whether and how the conditional

effects vary around the marginal effects. We then show how to appropriately parameterize these functions and adapt the method of regression-with-residuals (RWR) to consistently estimate them.

An SNMM is based on the following decomposition of the conditional mean of the potential outcomes given the confounders into a set of conditional treatment effects and a set of so-called "nuisance" associations:

$$E\big(Y_i(a_1,a_2)\big|C_{i1},C_{i2}(a_1)\big)=\beta_{00}+\varepsilon_1(C_{i1})+\mu_1(C_{i1},a_1)+\varepsilon_2\big(C_{i1},a_1,C_{i2}(a_1)\big)+$$
$$\mu_2(C_{i1},a_1,C_{i2}(a_1),a_2), \quad (6)$$

where $\beta_{00}=E\big(Y_i(0,0)\big)$ is the marginal mean of the potential outcomes under no treatment; $\varepsilon_1(C_{i1})=\big[E(Y_i(0,0)|C_{i1})-E\big(Y_i(0,0)\big)\big]$ is a nuisance association that captures the relationship between the confounder at time $t=1$ and the outcome under no treatment; $\mu_1(C_{i1},a_1)=E(Y_i(a_1,0)-Y_i(0,0)|C_{i1})$ is a causal function that captures the conditional effects of treatment at time $t=1$ given C_{i1}; $\varepsilon_2\big(C_{i1},a_1,C_{i2}(a_1)\big)=\big[E\big(Y_i(a_1,0)\big|C_{i1},C_{i2}(a_1)\big)-E(Y_i(a_1,0)|C_{i1})\big]$ is another nuisance association that captures the relationship between the confounder at time $t=2$ and the outcome under treatment sequence $\{a_1,0\}$; and $\mu_2(C_{i1},a_1,C_{i2}(a_1),a_2)=E\big(Y_i(a_1,a_2)-Y_i(a_1,0)\big|C_{i1},C_{i2}(a_1)\big)$ is another causal function that captures the conditional effects of treatment at time $t=2$ given both prior confounders. The functions $\varepsilon_1(C_{i1})$ and $\varepsilon_2\big(C_{i1},a_1,C_{i2}(a_1)\big)$ are called "nuisance" associations because they do not contain any information about the causal effects of treatment (Wodtke and Almirall 2017).

The first causal function, $\mu_1(C_{i1},a_1)$, can be further decomposed into a marginal effect of interest and a term that captures effect modification as follows:

$$\mu_1(C_{i1},a_1)=\mu_{11}(a_1)+\mu_{12}(C_{i1},a_1), \quad (7)$$

where $\mu_{11}(a_1) = E\big(Y_i(a_1, 0) - Y_i(0,0)\big)$ is equal to the $DTE(a_1)$ and $\mu_{12}(C_{i1}, a_1) = \big[E(Y_i(a_1, 0) - Y_i(0,0)|C_{i1}) - E\big(Y_i(a_1, 0) - Y_i(0,0)\big)\big]$ captures how the effect of treatment at time $t = 1$ differs across levels of C_{i1}.

Similarly, the second causal function, $\mu_2(C_{i1}, a_1, C_{i2}(a_1), a_2)$, can also be further decomposed as follows:

$$\mu_2(C_{i1}, a_1, C_{i2}(a_1), a_2) = \mu_{21}(a_1, a_2) + \mu_{22}(C_{i1}, a_1, a_2) + \mu_{23}(C_{i1}, a_1, C_{i2}(a_1), a_2), \quad (8)$$

where $\mu_{21}(a_1, a_2) = E\big(Y_i(a_1, a_2) - Y_i(a_1, 0)\big)$ is equal to another marginal effect of interest, the $PTE(a_1, a_2)$; $\mu_{22}(C_{i1}, a_1, a_2) = \big[E(Y_i(a_1, a_2) - Y_i(a_1, 0)|C_{i1}) - E\big(Y_i(a_1, a_2) - Y_i(a_1, 0)\big)\big]$ captures how the effect of treatment at $t = 2$ differs across levels of C_{i1}; and $\mu_{23}(C_{i1}, a_1, C_{i2}(a_1), a_2) = \big[E\big(Y_i(a_1, a_2) - Y_i(a_1, 0)|C_{i1}, C_{i2}(a_1)\big) - E(Y_i(a_1, a_2) - Y_i(a_1, 0)|C_{i1})\big]$ captures how the effect of treatment at $t = 2$ differs across levels of $C_{i2}(a_1)$ within levels of C_{i1}.

Any parameterization of the marginal effects, $\mu_{11}(a_1)$ and $\mu_{21}(a_1, a_2)$, must satisfy the constraint that they are equal to zero when contemporaneous treatment is equal to zero. With a binary treatment, a saturated parameterization for $\mu_{11}(a_1)$ is

$$\mu_{11}(a_1) = \beta_{10} a_1, \quad (9)$$

and a saturated parameterization for $\mu_{21}(a_1, a_2)$ is

$$\mu_{21}(a_1, a_2) = (\beta_{20} + \beta_{21} a_1) a_2, \quad (10)$$

where $\beta_{10} = DTE(1)$, $\beta_{20} = PTE(0,1)$, and $\beta_{20} + \beta_{21} = PTE(1,1)$. In addition, note that $\beta_{10} + \beta_{20} + \beta_{21} = CTE$ and $\beta_{21} = INE$. Thus, all of the marginal effects defined previously are given by the parameter vector $\{\beta_{10}, \beta_{20}, \beta_{21}\}$.

Any parameterization of $\mu_{12}(C_{i1}, a_1)$ must satisfy the constraints that it is equal to zero when $a_1 = 0$ and that it has mean zero. With a treatment and confounder that are both binary, a saturated parameterization for this function is

$$\mu_{12}(C_{i1}, a_1) = \theta_{10} a_1 C_{i1}^{\perp}, \quad (11)$$

where $C_{i1}^{\perp} = \left(C_{i1} - E(C_{i1})\right)$ is a residual transformation of C_{i1} with respect to its marginal mean. This parameterization satisfies the zero mean constraint because $E(\theta_{10} a_1 C_{i1}^{\perp}) = \theta_{10} a_1 E(C_{i1}^{\perp}) = \theta_{10} a_1 E\left(\left(C_{i1} - E(C_{i1})\right)\right) = 0$.

Similarly, any parameterization of $\mu_{22}(C_{i1}, a_1, a_2)$ must satisfy the constraints that it is equal to zero when $a_2 = 0$ and that it has mean zero. A saturated parameterization for this function is

$$\mu_{22}(C_{i1}, a_1, a_2) = (\theta_{20} + \theta_{21} a_1) a_2 C_{i1}^{\perp}, \quad (12)$$

which has mean zero because the expectation function is a linear operator and because $E(C_{i1}^{\perp}) = 0$, as above.

Finally, any parameterization of $\mu_{23}(C_{i1}, a_1, C_{i2}(a_1), a_2)$ must satisfy the constraints that it is equal to zero when $a_2 = 0$ and that it has mean zero conditional on C_{i1}. A saturated parameterization for this function is

$$\mu_{23}(C_{i1}, a_1, C_{i2}(a_1), a_2) = (\theta_{22} + \theta_{23} a_1 + (\theta_{24} + \theta_{25} a_1) C_{i1}^{\perp}) a_2 C_{i2}^{\perp}(a_1), \quad (13)$$

where $C_{i2}^{\perp}(a_1) = \left(C_{i2}(a_1) - E(C_{i2}(a_1)|C_{i1})\right)$ is a residual transformation of $C_{i2}(a_1)$ with respect to its conditional mean given C_{i1}. This parameterization satisfies the zero mean constraint because the expectation function is a linear operator and because $E(C_{i2}^{\perp}(a_1)|C_{i1}) = E\left(\left(C_{i2}(a_1) - E(C_{i2}(a_1)|C_{i1})\right)|C_{i1}\right) = E(C_{i2}(a_1)|C_{i1}) - E(C_{i2}(a_1)|C_{i1}) = 0$. The parameter

vector $\{\theta_{10}, \theta_{20}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}, \theta_{25}\}$ captures how the confounders modify the effect of treatment at each time point.

The nuisance associations, $\varepsilon_1(C_{i1})$ and $\varepsilon_2\big(C_{i1}, a_1, C_{i2}(a_1)\big)$, must also be parameterized under the constraint that they have mean zero given the past, which can be accomplished using the same residualized confounders as defined previously. Specifically, a saturated parameterization for the first nuisance association is

$$\varepsilon_1(C_{i1}) = \gamma_{10} C_{i1}^{\perp}, \quad (14)$$

and a saturated parameterization for the second nuisance association is

$$\varepsilon_2\big(C_{i1}, a_1, C_{i2}(a_1)\big) = (\gamma_{20} + \gamma_{21} a_1 + (\gamma_{22} + \gamma_{23} a_1) C_{i1}^{\perp}) C_{i2}^{\perp}(a_1), \quad (15)$$

where the parameter vector $\{\gamma_{10}, \gamma_{20}, \gamma_{21}, \gamma_{22}, \gamma_{23}\}$ captures the associational (i.e., causal and possibly non-causal) effects of the confounders on the outcome.

Combining parametric expressions for the causal functions and nuisance associations yields the following saturated SNMM:

$$E\big(Y_i(a_1, a_2) | C_{i1}, C_{i2}(a_1)\big) = \beta_{00} + \gamma_{10} C_{i1}^{\perp} + \beta_{10} a_1 + \theta_{10} a_1 C_{i1}^{\perp} + (\gamma_{20} + \gamma_{21} a_1 + (\gamma_{22} + \gamma_{23} a_1) C_{i1}^{\perp}) C_{i2}^{\perp}(a_1) + (\beta_{20} + \beta_{21} a_1) a_2 + (\theta_{20} + \theta_{21} a_1) a_2 C_{i1}^{\perp} + (\theta_{22} + \theta_{23} a_1 + (\theta_{24} + \theta_{25} a_1) C_{i1}^{\perp}) a_2 C_{i2}^{\perp}(a_1). \quad (16)$$

This model differs from those outlined in Almirall et al. (2010), Wodtke and Almirall (2017), and Vansteelandt and Sjolander (2016; e.g., their Equation 6) in that the residualized confounders, rather than the untransformed values of these variables, are included as part of interaction terms in the causal functions. It also differs from the highly constrained SNMMs outlined in Vansteelandt and Sjolander (2016; e.g., their Equation 3) and Wodtke (2018) in that effect modification is not assumed to be absent but rather is explicitly modeled, or in other words, $\{\theta_{10}, \theta_{20}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}, \theta_{25}\}$ are free parameters that are not assumed to be zero.

An SNMM parameterized as above can be estimated using RWR, which proceeds in two stages. In the first stage, residual terms are estimated by centering C_{i1} around its sample mean and by centering C_{i2} around its estimated conditional mean given C_{i1} and A_{i1}. Specifically, $\hat{C}_{i1}^{\perp} = C_{i1} - \hat{E}(C_{i1})$ and $\hat{C}_{i2}^{\perp} = C_{i2} - \hat{E}(C_{i2}|C_{i1}, A_{i1})$, where $\hat{E}(C_{i1}) = \frac{1}{n}\sum_i C_{i1}$ and $\hat{E}(C_{i2}|C_{i1}, A_{i1})$ is estimated from a generalized linear model with, for example, the logit or probit link when C_{i2} is binary. Then, in the second stage, least squares estimates are computed for a linear regression of the outcome on prior treatments, the residualized confounders, and interactions involving the prior treatments and residualized confounders. This regression can be expressed as follows:

$$\hat{E}(Y_i|C_{i1}, A_{i1}, C_{i2}, A_{i2}) = \hat{\beta}_{00} + \hat{\gamma}_{10}\hat{C}_{i1}^{\perp} + \hat{\beta}_{10}A_{i1} + \hat{\theta}_{10}A_{i1}\hat{C}_{i1}^{\perp} + \big(\hat{\gamma}_{20} + \hat{\gamma}_{21}A_{i1} + (\hat{\gamma}_{22} + \hat{\gamma}_{23}A_{i1})\hat{C}_{i1}^{\perp}\big)\hat{C}_{i2}^{\perp} + \big(\hat{\beta}_{20} + \hat{\beta}_{21}A_{i1}\big)A_{i2} + \big(\hat{\theta}_{20} + \hat{\theta}_{21}A_{i1}\big)A_{i2}\hat{C}_{i1}^{\perp} + \big(\hat{\theta}_{22} + \hat{\theta}_{23}A_{i1} + \big(\hat{\theta}_{24} + \hat{\theta}_{25}A_{i1}\big)\hat{C}_{i1}^{\perp}\big)A_{i2}\hat{C}_{i2}^{\perp}. \quad (17)$$

where different combinations of the estimated beta coefficients, $\{\hat{\beta}_{00}, \hat{\beta}_{10}, \hat{\beta}_{20}, \hat{\beta}_{21}\}$, are consistent for the marginal effects of interest under the identification assumptions outlined previously and under the assumption that the model is correctly specified, which is here assured by saturating it. This approach is nearly identical to conventional least squares regression except that the confounders at each time point are first residualized with respect to the observed past.

Figure 2 displays a stylized graph that illustrates the logic of RWR estimation. It shows that residualizing the confounders at each time point with respect to the observed past purges the treatment-induced confounder, C_{i2}, of its association with prior treatment, A_{i1}. As a result, including the residual transformation of C_{i2} in a model for the outcome avoids bias due to over-control and endogenous selection. In addition, because RWR adjusts for observed confounding by conditioning on residual transformations of the confounders in an outcome regression rather than by re-weighting the data to appropriately balance the confounders across future treatments,

it also avoids the limitations associated with inverse-probability-of-treatment-weighted estimation, such as the difficulty associated with constructing well-behaved weights for continuous treatments. Finally, by including the residualized confounders as part of interaction terms with treatment, RWR can accommodate effect modification while neatly isolating the marginal effects of interest in a single parameter vector.

3.2. Model Specification and Other Considerations

In practice, estimating a saturated SNMM is often impractical, or even impossible, either because the confounders or treatments are continuous, because there are a large number of time periods, or because there are many confounders. In this situation, a set of parametric constraints must be imposed on the SNMM to facilitate estimation. For example, an analyst might consider excluding all higher-order interactions involving confounders at multiple time periods, in which case regression-with-residual (RWR) estimation would proceed exactly as outlined previously except with the outcome regression in the second stage simplified as follows:

$$\hat{E}(Y_i|C_{i1}, A_{i1}, C_{i2}, A_{i2}) = \hat{\beta}_{00} + \hat{\gamma}_{10}\hat{C}_{i1}^{\perp} + \hat{\beta}_{10}A_{i1} + \hat{\theta}_{10}A_{i1}\hat{C}_{i1}^{\perp} + (\hat{\gamma}_{20} + \hat{\gamma}_{21}A_{i1})\hat{C}_{i2}^{\perp} + \left(\hat{\beta}_{20} + \hat{\beta}_{21}A_{i1}\right)A_{i2} + \left(\hat{\theta}_{20} + \hat{\theta}_{21}A_{i1}\right)A_{i2}\hat{C}_{i1}^{\perp} + \left(\hat{\theta}_{22} + \hat{\theta}_{23}A_{i1}\right)A_{i2}\hat{C}_{i2}^{\perp}. \quad (18)$$

Of course, many other constraints are possible, but recall that RWR requires a correctly specified model for the outcome. Thus, if any of these modeling constraints are inappropriate, then RWR is biased, even when the effects of interest are identified under sequential ignorability.

Additional modeling considerations may be required of RWR when there are many time periods. Specifically, it may be necessary to impose rather stringent constraints on the SNMM in order to reduce its complexity. For example, an analyst might consider not merely excluding all

higher-order interactions but also constraining parameters to be invariant over time, in which case the outcome regression would be simplified as follows:

$$\hat{E}\left(Y_i|\overline{C}_i, \overline{A}_i, \right) = \hat{\beta}_{00} + \hat{\gamma}\sum_t \hat{C}_{it}^{\perp} + \hat{\beta}\sum_t A_{it} + \hat{\theta}\sum_t A_{it}\hat{C}_{it}^{\perp}, \quad (19)$$

where overbars denote variable "histories," that is, $\overline{C}_i = \{C_{i1}, \ldots, C_{iT}\}$ and $\overline{A}_i = \{A_{i1}, \ldots, A_{iT}\}$. Relatedly, with many time periods, it may also be necessary to constrain the models for the confounders in the first stage, and this might be accomplished by relying on some of the same simplifying constraints considered here (e.g., no higher-order interactions, time-invariant coefficients, and so forth). Nevertheless, RWR requires correctly specified models, which may be challenging to achieve in applications with a large number of time periods, as the potential complexity of these models increases with the dimension of the data.

Additional modeling considerations are also required with RWR when there are multiple different confounders for which adjustment is necessary. First, all of the different confounders must be appropriately residualized in the first stage. This is accomplished by fitting a model for each confounder at each time point using only prior variables as predictors and then extracting their residuals. Thus, these variables are residualized exactly as outlined previously without needing to specify the causal ordering of different confounders measured concurrently in time. Second, all of the residualized confounders must be included in the second-stage regression for the outcome, which may now involve additional interaction terms between treatment and the residualized confounders.

When estimating marginal effects with RWR and multiple different confounders, the method can accommodate all types of treatment-by-confounder interaction except for higher-order (i.e., three-way and above) interactions involving treatment and two or more different confounders measured contemporaneously. In the presence of such higher-order interactions, the

conditional effects of treatment cannot be conveniently decomposed and parameterized with residual terms. As an approximation, RWR could still be implemented in this situation by residualizing the different confounders and their cross-products using separate models for each term, but with this approach, it is possible that the models could be incompatible. Thus, with multiple different confounders, RWR estimation of marginal effects is most appropriate for a moderately constrained SNMM in which some especially complex forms of effect modification are assumed to be absent. Although somewhat limiting, this modeling constraint is still considerably weaker than that required of other methods for estimating marginal effects with an SNMM (e.g., Wodtke 2018).

RWR estimation is also only appropriate for a *linear* SNMM. It is therefore best suited for use with continuous outcomes. In certain situations, it may also be suitable for use with other types of outcomes for which a linear model is sometimes a reasonable approximation, such as linear probability models for binary outcomes. Relatedly, with RWR, the outcome model is typically fit via ordinary least squares. If the data come from a complex sampling design, RWR may also be implemented using weighted or generalized least squares. The method therefore inherits the strengths and also some of the limitations of least squares estimators, such as sensitivity to outliers, and its performance should be assessed with regression diagnostics.

In sum, RWR estimation of a moderately constrained SNMM for marginal effects is a relatively simple adaptation of conventional least squares regression. It proceeds as follows: first, the confounders at each time point are regressed on all prior variables and then residualized, and second, the outcome is regressed on prior treatments, the residualized confounders, and to accommodate effect modification, an admissible set of interaction terms involving prior treatments and the residualized confounders. The proposed method can accommodate all types of

effect modification except for those involving higher-order interactions between treatment and two or more different confounders measured at the same point in time. RWR is consistent under the identification assumptions outlined previously along with the assumption of correct model specification. Because identifying correct models is challenging with many time periods and/or confounders, the method may be best suited for applications that have fewer time periods and that require adjustment for a smaller number of covariates. At the same time, however, the complexity of the modeling required with other methods, such as IPTW and g-estimation, also increases with the dimension of the data, as a large number of time periods and/or confounders present a set of generic challenges for causal inference in the time-varying setting.

Valid standard errors for RWR estimates can be obtained using the nonparametric bootstrap (Almirall et al. 2014; Efron and Tibshirani 1994). This involves repeatedly sampling n observations from the observed data with replacement in order to construct b replicate samples. RWR estimates are then computed from each of the b samples, and their standard deviation across them gives the bootstrap standard error. Efron and Tibshirani (1994) suggest using $b \geq 200$ replications when estimating standard errors.

3.3. Estimating Controlled Direct Effects in Analyses of Causal Mediation

In this section, we briefly demonstrate that the methods outlined previously can also be used to estimate controlled direct effects in analyses of causal mediation when there are treatment-induced mediator-outcome confounders. To appreciate this, first let d denote exposure to a binary treatment, and let m denote a putative mediator that is also binary. In addition, let $Y_i(d, m)$ denote the potential outcome for subject i had she previously been exposed to treatment d and the mediator m. Finally, let X_i denote a treatment-outcome confounder for subject i

measured at baseline, and let $Z_i(d)$ denote a post-treatment confounder of the mediator-outcome relationship, which is indexed as a potential outcome by d to reflect that it is a treatment-induced confounder.

The controlled direct effect (CDE) measures the causal relationship between treatment and the outcome when the putative mediator is fixed at the same value for all individuals. It can be formally defined as

$$CDE(d,m) = E\big(Y_i(d,m) - Y_i(0,m)\big). \quad (20)$$

This quantity can be non-parametrically identified from the observed data under the assumptions of stable unit treatment values, consistency, positivity, and sequential ignorability (VanderWeele 2009, 2015). In this context, sequential ignorability is satisfied when there are no unobserved treatment-outcome or mediator-outcome confounders.

Although conventional methods are biased in the presence of treatment-induced confounders, the CDE can still be consistently estimated using a structural nested mean model (SNMM) and regression-with-residuals (RWR; Zhou and Wodtke 2019). For example, consider the following moderately constrained SNMM for the joint effects of treatment and the mediator on the outcome,

$$E\big(Y_i(d,m)|X_i, Z_i(d)\big) = \beta_{00} + \gamma_{10}X_i^{\perp} + \beta_{10}d + \theta_{10}dX_i^{\perp} + (\gamma_{20} + \gamma_{21}d + \gamma_{22}X_i^{\perp})Z_i^{\perp}(d) +$$

$$(\beta_{20} + \beta_{21}d)m + (\theta_{20} + \theta_{21}d)mX_i^{\perp} + (\theta_{22} + \theta_{23}d)mZ_i^{\perp}(d), \quad (21)$$

where $X_i^{\perp} = X_i - E(X_i)$, $Z_i^{\perp}(d) = Z_i(d) - E(Z_i(d)|X_i)$, and, for simplicity, higher-order interactions involving both confounders are excluded. With this model, the CDE is given by $CDE(d,m) = (\beta_{10} + \beta_{21}m)d$, any potential modification of the treatment effect by the baseline confounder, X_i, is captured by θ_{10}, and any potential modification of the mediator effect by X_i or the post-treatment confounder, $Z_i(d)$, is captured by $\{\theta_{20}, \theta_{21}, \theta_{22}, \theta_{23}\}$. This model can be

estimated via RWR by, first, centering X_i around its sample mean and centering Z_i around its estimated conditional mean given X_i and D_i, and then second, fitting a regression of the outcome on treatment, the mediator, the residualized confounders, and a set of interaction terms between treatment, the mediator, and the residualized confounders. As in the time-varying setting, valid standard errors can be obtained from the nonparametric bootstrap.

4. Simulation Experiments

We use a series of simulation experiments to evaluate the performance of regression-with-residuals (RWR) estimation for marginal effects relative to other methods. Specifically, we use 10,000 simulations of $\mathrm{n} = 500$ to estimate the cumulative treatment effect (CTE) of a time-varying exposure measured at two time points. In each simulation, we generate an "unobserved" continuous variable U_i, an observed continuous time-varying confounder $\{C_{i1}, C_{i2}\}$, a binary time-varying treatment $\{A_{i1}, A_{i2}\}$, and a continuous end-of-study outcome, Y_i. In these simulations, $[U_i]\sim N\left(\mu_{U_i} = 0, \sigma^2_{U_i} = 1\right)$; $[C_{i1}]\sim N\left(\mu_{C_{i1}} = 0, \sigma^2_{C_{i1}} = 1\right)$; $[C_{i2}|U_i, C_{i1}, A_{i1}]\sim N\left(\mu_{C_{i2}|U_i,C_{i1},A_{i1}} = 0.5U_i + 0.5C_{i1} + 0.5A_{i1}, \sigma^2_{C_{i2}|U_i,C_{i1},A_{i1}} = 1\right)$; $[A_{i1}|C_{i1}]\sim \text{Bernoulli}\left(p_{A_{i1}} = \Phi(\gamma C_{i1})\right)$; $[A_{i2}|C_{i1}, A_{i1}, C_{i2}]\sim \text{Bernoulli}\left(p_{A_{i2}|C_{i1},A_{i1},C_{i2}} = \Phi(\gamma C_{i1} + 0.5A_{i1} + \gamma C_{i2})\right)$; $[Y_i|U_i, C_{i1}, A_{i1}, C_{i2}, A_{i2}]\sim N\left(\mu_{Y_i|U_i,C_{i1},A_{i1},C_{i2},A_{i2}} = 0.5U_i + \gamma\left(C_{i1} - \mu_{C_{i1}}\right) + A_{i1}\left(0.2 + \theta\left(C_{i1} - \mu_{C_{i1}}\right)\right) + \left(C_{i2} - \mu_{C_{i2}}\right)\left(\gamma + \left(C_{i1} - \mu_{C_{i1}}\right)\eta\right) + A_{i2}\left(0.2 + 0.1A_{i1} + \theta\left(\left(C_{i1} - \mu_{C_{i1}}\right) + \left(C_{i2} - \mu_{C_{i2}}\right)\right)\right), \sigma^2_{Y_i|U_i,C_{i1},A_{i1},C_{i2},A_{i2}} = 1\right)$, where Φ is the standard normal cumulative distribution function, γ is a parameter used to control the level of treatment-outcome confounding, θ is a parameter used to control the level of treatment effect

modification, and η is a parameter used to control the degree to which the associational effect of C_{i2} on Y_i is modified by C_{i1}. In all simulations, the CTE is identified and its true value is 0.5.

We compare the performance of RWR estimation of a moderately constrained SNMM for marginal effects (henceforth "RWR with interactions") to the performance of conventional least squares regression, IPTW estimation of a marginal structural model, g-estimation of a highly constrained SNMM in which effect modification is assumed to be absent, and RWR estimation of the same highly constrained SNMM (henceforth "RWR without interactions"). To compute conventional regression estimates, we fit by least squares a linear regression of the outcome on prior treatments, the observed confounders, and a treatment-by-treatment interaction. The estimated CTE is then given by the sum of the coefficients on prior treatments and the interaction term.

To compute IPTW estimates (Robins et al. 1994, 2000), we fit a linear regression of the outcome on prior treatments and their interaction using weighted least squares, with weights equal to

$$w_i = \frac{P(A_{i1}=a_{i1})}{P(A_{i1}=a_{i1}|C_{i1})} \times \frac{P(A_{i2}=a_{i2}|A_{i1}=a_{i1})}{P(A_{i2}=a_{i2}|C_{i1},A_{i1}=a_{i1},C_{i2})}, \quad (22)$$

where w_i is estimated from a series of probit models for the conditional probabilities in the numerator and denominator of the weight. At each time point, weighting by w_i balances (in expectation) prior confounders across future treatments by giving more weight to subjects with confounder histories that are underrepresented in a treatment group and less weight to subjects with confounder histories that are overrepresented in a treatment group. The estimated CTE is the sum of the coefficients on prior treatments and their interaction.

To compute g-estimates of marginal effects using a highly constrained SNMM without any effect modification, we use the g-estimator proposed by Vansteelandt and Sjolander (2016).

Specifically, we first fit a linear regression of the outcome on prior treatments and their interaction, estimated propensity scores for treatment at each time point, an interaction between treatment at time $t = 1$ and the estimated propensity score for treatment at time $t = 2$, and the observed confounders at both time points. The coefficients on treatment at time $t = 2$ and its interaction with treatment at time $t = 1$ from this model provide estimates of the proximal treatment effect. Then, we subtract the estimated proximal treatment effect from the outcome for each respondent and regress this transformed outcome on the treatment, propensity score, and the observed confounder at time $t = 1$. The coefficient on treatment from this model provides an estimate of the distal treatment effect, and then the sum of the distal and proximal treatment effects computed as above give the estimated CTE. Vansteelandt and Sjolander (2016) show that this estimator is asymptotically equivalent to the doubly robust g-estimator considered in Robins et al. (1992).

To compute estimates based on RWR without interactions, we first residualize the observed confounders at each time point by regressing them on all prior variables and then centering them around their estimated conditional means. Second, we regress the outcome on prior treatments and their interaction as well as the residualized confounders. The estimated CTE is the sum of the coefficients on prior treatments and their interaction. Computing estimates based on RWR with interactions proceeds in almost exactly the same manner except that all two-way interactions between the treatments and residualized confounders are additionally included in the outcome regression. Part A of the Online Supplement presents the R code used to execute all of the simulations outlined previously.

We compare the performance of these methods in terms of their bias, standard deviation, and root mean squared error (RMSE) under different levels of treatment-outcome confounding

and under different levels of effect modification. Because treatment-induced confounding is present in all simulations, we expect conventional regression to perform poorly across all scenarios. Because IPTW estimation is relatively inefficient and susceptible to finite-sample bias when confounders strongly predict treatment, we expect its performance to suffer in simulations with high levels of treatment-outcome confounding. Because g- and RWR estimation of marginal effects using a highly constrained SNMM require that the confounders must not modify the effects of treatment, we expect their performance to deteriorate in simulations with high levels of treatment effect modification. Finally, because RWR with interactions accommodates this type of effect modification, we expect it to perform well across all simulations.

Table 1 presents results from a first set of simulation experiments, wherein we varied the level of treatment-outcome confounding in the absence of effect modification. Conventional regression is badly biased at all levels of confounding, as expected. IPTW estimation performs well at lower levels of confounding but suffers from finite-sample bias at higher levels and is relatively inefficient, also as expected. G- and both types of RWR estimation perform similarly in these simulations: they are all unbiased and achieve comparable efficiency gains relative to IPTW.

Table 2 presents results from a second set of simulation experiments, wherein we varied the level of treatment effect modification after setting the level of treatment-outcome confounding at a moderate-to-high level. As expected, both conventional regression and IPTW estimation perform poorly. Although IPTW estimation accommodates effect modification, it still suffers from finite-sample bias due to the high level of confounding and is relatively inefficient. Also as expected, G-estimation and RWR estimation without interactions are increasingly biased

as the magnitude of treatment effect modification rises, whereas RWR with interactions is unbiased and achieves the lowest RMSE across all scenarios.

Finally, because RWR requires strong modeling assumptions that may be difficult to satisfy in some applications, we evaluate the method's performance with an SNMM that has an incorrectly specified nuisance association. Specifically, Table 3 presents results from a third set of simulation experiments, wherein we varied the degree to which the associational effect of C_{i2} on Y_i is modified by C_{i1} after setting the level of both treatment-outcome confounding and treatment effect modification at moderate-to-high levels. Because all the SNMMs considered in these simulations constrain the associational effects of the confounders to be invariant, they are all incorrectly specified when this type of effect modification is present.

In Table 3, both conventional regression and IPTW estimation perform poorly because, as before, these simulations involve a moderate-to-high level of treatment-induced confounding. G-estimation and RWR without interactions also do not perform very well because these simulations involve a nontrivial level of effect modification. Note, however, that the performance of g-estimation is similar regardless of the degree to which the nuisance associations are incorrectly specified, which reflects its doubly robust property. The performance of RWR without interactions, by contrast, further deteriorates with greater misspecification of the nuisance associations. Similarly, the bias and RMSE of RWR with interactions also steadily increases with the degree to which the nuisance associations are incorrectly specified. Nevertheless, RWR with interactions still appears to outperform other methods, even when it is based on an SNMM with misspecified nuisance associations.

5. Empirical Examples

5.1. The CTE of Neighborhood Poverty on Academic Achievement

Does growing up in a disadvantaged neighborhood inhibit academic achievement? The effects of neighborhood composition on child development have long concerned social scientists across several different disciplines (e.g., Chetty et al. 2016; Leventhal and Brooks-Gunn 2000; Sampson et al. 2008; Wodtke et al. 2011). To illustrate how the proposed method can be used with time-varying treatments, we estimate the CTE of residence in a disadvantaged neighborhood throughout the early life course on adolescent math achievement using data from $n = 1{,}135$ individuals in the Panel Study of Income Dynamics – Child Development Supplement (PSID-CDS; Michigan Survey Research Center 2014).[2]

In these data, the outcome, Y_i, represents standardized scores on the Woodcock-Johnson applied problems test measured at the end of follow-up when individuals were age 13 to 17 (Woodcock and Johnson 1989). The time-varying treatment, a_t, is a standardized index of neighborhood disadvantage generated via a principal component analysis of multiple census tract characteristics, such as the poverty rate, unemployment rate, and median household income. Treatment is first measured during childhood when individuals were age 5 to 9 and then again during adolescence when they were age 11 to 15. The vector of baseline confounders, C_{i1}, contains a set of time-invariant factors, such as race, gender, and birth cohort, as well as a set of time-varying characteristics, including equivalized family income, parental marital status, and

[2] Some of the data used in this analysis are based on "sensitive data files" from the PSID-CDS, which were obtained under special contractual arrangements designed to protect the anonymity of respondents. These data are not available from the authors. Persons interested in obtaining sensitive data files from the PSID-CDS should contact psidhelp@isr.umich.edu. A set of replication files for this analysis, without any sensitive data, is provided as part of the online supplementary material.

lagged achievement test scores, all measured during early childhood. Another vector of confounders, C_{i2}, contains the same set of time-varying characteristics only now measured just before the onset of adolescence.

Previously, Wodtke (2018) estimated the CTE of residence in a disadvantaged neighborhood with data from the PSID-CDS by fitting a conventional regression model using least squares, a marginal structural model using inverse probability of treatment weighting (IPTW), and a highly constrained structural nested mean model (SNMM) without any effect modification using regression-with-residuals (RWR). In that analysis, RWR estimates indicated that long-term residence in a disadvantaged neighborhood has a severe negative effect on math achievement—an effect that is obscured by bias in conventional regression models and imprecisely captured by IPTW. These estimates, however, are premised on the strong and arguably unrealistic assumption of no effect modification.

We overcome this limitation by estimating the CTE using RWR and an SNMM that includes all two-way treatment-by-confounder interactions. Specifically, we model the distal, proximal, and cumulative marginal effects of exposure to a disadvantaged neighborhood on adolescent math achievement using the following SNMM:

$$E\big(Y_i(a_1,a_2)|C_{i1},C_{i2}(a_1)\big) = \beta_{00} + \gamma_{10}^T C_{i1}^{\perp} + \beta_{10}a_1 + \theta_{10}^T a_1 C_{i1}^{\perp} + \gamma_{20}^T C_{i2}^{\perp}(a_1) + \beta_{20}a_2 + \theta_{20}^T a_2 C_{i1}^{\perp} + \theta_{21}^T a_2 C_{i2}^{\perp}(a_1), \quad (23)$$

where $\beta_{10} = DTE(1)$, $\beta_{20} = PTE(a_1, 1)$, and $\beta_{10} + \beta_{20} = CTE$. Note that this model is just a moderately constrained version of Equation (16).

The first row of Table 4 reports RWR estimates for the distal, proximal, and cumulative effects of living in a disadvantaged neighborhood. We compute these estimates by, first, centering the elements of C_{i1} around their sample means and centering the elements of C_{i2}

around their estimated conditional means, which come from least squares regressions of C_{i2} on the treatment and confounders measured earlier during childhood. Second, we compute the marginal effects of interest by regressing the outcome on both treatments, the residualized confounders, and all two-way interactions between the treatments and residualized confounders. For comparative purposes, the second and third rows of Table 4 report RWR and g-estimates of marginal effects from a highly constrained SNMM in which all treatment-by-confounder interactions are excluded. Part B of the Online Supplement presents the R code used to generate the results in this table.

All of the estimates in Table 4 indicate that the distal effect of childhood exposure to a disadvantaged neighborhood on adolescent math achievement is substantively small and fails to reach conventional significance thresholds, that the proximal effect of adolescent exposure is larger and statistically significant, and that the cumulative effect of sustained exposure is substantively large and highly significant. For example, according to these estimates, sustained exposure to a poor neighborhood one standard deviation above the national mean of the disadvantage index, rather than sustained exposure to a wealthy neighborhood one standard deviation below the national mean, is estimated to reduce adolescent math achievement by about $0.127 \times 2 = 0.254$ standard deviations.

The results in Table 4 are similar across the different methods employed. Consistent with this finding, a Wald test of the null hypothesis that $\theta_{10} = \theta_{20} = \theta_{21} = 0$ provides little evidence that the effects of living in a disadvantaged neighborhood are modified by any of the confounders ($\chi^2 = 10.7, df = 17, p = 0.873$). In this application, it therefore appears that our findings are not particularly sensitive to the presence or absence of effect modification.

Nevertheless, it is the flexibility of RWR that allows for an easy assessment of marginal effects under different specifications.

5.2. The CDE of Education on Mental Health

Does income explain the effect of education on depression? A number of prior studies have investigated the causal relationship between education and mental health (e.g., Cutler and Lleras-Muney 2006; Heckman et al. 2018; Lee 2011), but the mechanisms underlying this causal link remain unclear. Education may improve mental health by providing access to higher economic status and greater resources, or it may affect mental health through other channels—for example, by providing greater access to health information and improving health behaviors. To illustrate the utility of regression-with-residuals (RWR) for analyses of causal mediation, we estimate the controlled direct effect (CDE) of college completion on mental health controlling for family income as a putative mediator. In this example, a comparison of the total effect with the CDE helps to adjudicate whether family economic status explains the mental health benefits of college completion.

We use data from $n = 2{,}719$ individuals in the National Longitudinal Survey of Youth 1979 (NLSY79) who were age 14-17 when they were first interviewed in 1979. In these data, the outcome, Y_i, represents scores on the Center for Epidemiologic Studies Depression Scale (CES-D) when respondents were age 40. We standardize CES-D scores to have mean zero and unit variance, where a higher score implies greater depression. The treatment, d, is defined as completion of a four-year college degree by age 25, while the mediator of interest, m, is the percentile rank of equivalized family income averaged over ages 36-40. The vector of baseline confounders, X_i, include gender, race, Hispanic ethnicity, mother's years of schooling, father's

presence in the home, number of siblings, urban residence, educational expectations, and percentile scores on the Armed Forces Qualification Test (AFQT). Finally, the vector of post-treatment confounders, Z_i, include CES-D scores measured when respondents were age 27-30, the proportion of time married between 1990 and 1998, and the number of relationship transitions between 1990 and 1998. These variables capture mental health and family stability during young adulthood, which may be affected by treatment (college completion by age 25) and also affect both the mediator (family income between age 36 and 40) and the outcome (depression at age 40).

With these data, we first estimate the total effect of college completion using the following structural nested mean model (SNMM):

$$E(Y_i(d)|X_i) = \beta_{00} + \gamma_{10}^T X_i^\perp + \beta_{10} d + \theta_{10}^T d X_i^\perp. \quad (24)$$

Under this specification, β_{10} captures the total effect of college completion on depression at age 40. RWR estimates of this model yield a sizable and statistically significant total effect of education on mental health, where completing college is estimated to lower depression scores by 0.165 standard deviations ($SE = 0.066, p < 0.05$) on average.

We then model the joint effects of college completion and family income on depression using the following SNMM:

$$E\big(Y_i(d,m)|X_i, Z_i(d)\big) = \beta_{00} + \gamma_{10}^T X_i^\perp + \beta_{10} d + \theta_{10}^T d X_i^\perp + \gamma_{20}^T Z_i^\perp(d) + (\beta_{20} + \beta_{21} d)m + \theta_{20}^T m X_i^\perp + \theta_{21}^T m Z_i^\perp(d), \quad (25)$$

where the controlled direct effect is given by $CDE(m) = \beta_{10} + \beta_{21} m$. Note that this model is just a moderately constrained version of Equation (21).

The first panel of Figure 3 reports estimates for the controlled direct effects of college completion on depression computed using RWR with interactions. These estimates are obtained

by, first, computing residuals for each of the baseline confounders X_i and post-treatment confounders Z_i, which involves centering the elements of X_i around their sample means and centering the elements of Z_i around their estimated conditional means given the past. Second, the controlled direct effects are then estimated by fitting the model described previously using these residual terms. Because the $CDE(m)$ may vary with m, we estimate and plot controlled direct effects across the support of the mediator. For comparative purposes, the second and third panels of Figure 3 report RWR and g-estimates based on a highly constrained SNMM in which all treatment-by-confounder and mediator-by-confounder interactions are excluded. Part C of the Online Supplement presents the R code used to generate these results.

These estimates provide some evidence that the effect of education on depression is mediated by family income. For example, point estimates of the CDE when family income is fixed at or above its sample median are substantially smaller in magnitude than the estimated total effect, and the 95% confidence intervals contain 0 at every value of family income. Thus, these results suggest that at least some portion of the total effect operates through pathways involving family economic resources.

The estimates in Figure 3 are fairly consistent across the different methods employed. Moreover, a Wald test of the null hypothesis that $\theta_{10} = \theta_{20} = \theta_{21} = 0$ does not provide much evidence of effect modification ($\chi^2 = 26.9, df = 21, p = 0.175$). Thus, it appears that our findings are insensitive to the inclusion of treatment-by-confounder interactions in this application as well, but recall that the flexibility of RWR is what enables us to easily assess whether estimated marginal effects are robust to different specifications.

6. Discussion and Conclusions

In analyses of causal mediation and time-varying treatment effects, treatment-induced confounders often complicate efforts to estimate marginal effects. Several available methods avoid these complications, including marginal structural models and inverse probability of treatment weighting (IPTW) as well as g- and regression-with-residuals (RWR) estimation of highly constrained structural nested mean models (SNMMs), but they are not without limitations. Specifically, the performance of IPTW is poor with continuous treatments and/or mediators, a high degree of confounding, and small samples, while both g- and RWR estimation of highly constrained SNMMs are biased for the marginal effects of interest when effect modification is present. To overcome these limitations, we adapt the method of RWR to estimate marginal effects with a set of moderately constrained SNMMs that easily accommodate several types of effect modification as well as continuous treatments and/or mediators. A series of simulation experiments indicate that the proposed method outperforms IPTW estimation of MSMs in general and that it outperforms both g- and RWR estimation of highly constrained SNMMs in the presence of effect modification. Because the proposed method involves only simple and familiar computations, it is easily implemented with standard software, as we demonstrate across two empirical illustrations.

Nevertheless, despite its many advantages, RWR estimation of marginal effects is premised on a number of strong modeling assumptions. Specifically, it requires a correctly specified SNMM, which in turn requires that all of the causal functions and nuisance associations that compose this model are correctly specified. It also requires the absence of more complex forms of effect modification involving two or more confounders measured contemporaneously, which complicates the decomposition and parameterization of the SNMM

causal functions using residual terms. The assumption of a correctly specified SNMM may be reasonable with a relatively small number of confounders and time periods, but identifying a correct model may be challenging with high dimensional data.

In this situation, researchers might consider combining the methods proposed in this study with either IPTW or g-estimation to leverage their strengths while mitigating their weaknesses. For example, RWR could be used to adjust for a subset of the time-varying confounders that prove difficult to appropriately balance using IPTW. Then, a simplified SNMM involving only this subset of confounders and a more limited set of interaction terms could be fit by RWR to an appropriately weighted sample in which the remaining confounders have all been balanced. Alternatively, the confounders could first be residualized with respect to the observed past and then included in interaction terms with treatment and/or a mediator at each stage of the g-estimation procedure outlined by Vansteelandt and Sjolander (2016). This may provide some protection against bias due to misspecification of the nuisance associations in an SNMM, as g-estimation is doubly robust, while simultaneously accommodating several types of effect modification in analyses of marginal effects. RWR might also be combined with variable selection and regularization techniques, such as the LASSO, in an effort to identify sufficiently accurate yet parsimonious models in applications with many time periods or confounders.

In sum, RWR estimation of a moderately constrained SNMM for marginal effects provides an appealing alternative to IPTW estimation of MSMs and to both g- and RWR estimation of highly constrained SNMMs in which effect modification is assumed away. The proposed method improves upon IPTW estimation in that it is more efficient, easy to use with continuous treatments and/or mediators, and avoids finite-sample bias when the magnitude of observed confounding is strong. It improves upon g- and RWR estimation of highly constrained

SNMMs in that it can accommodate all but highly complex forms of effect modification while still neatly isolating the marginal effects of interest in a single set of parameters. Although the proposed method is premised on a number of strong modeling assumptions, it can be integrated with IPTW or g-estimation in situations where these assumptions are questionable to enhance its robustness. Given their flexibility, efficiency, and ease of use, we expect moderately constrained SNMMs along with the associated method of RWR to be frequently used in future studies of causal mediation and time-varying treatment effects.

References

Acharya, A., Blackwell, M., and Sen, M. (2016) Explaining causal findings without bias: detecting and assessing direct effects. *Am. Polit. Sci. Rev.*, 110, 512-529.

Almirall, D., Griffin, B. A., McCaffrey, D. F., Ramchand, R., Yuen, R. A., and Murphy, S. A. (2014) Time-varying effect moderation using the structural nested mean model: estimation using inverse-weighted regression-with-residuals. *Stat. Med.*, 33, 3466-3487.

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

Chetty, R., Hendren, N., and Katz, L. (2016) The effects of exposure to better neighborhoods on children: new evidence from the Moving to Opportunity project. *Am. Econ. Rev.*, 106, 855-902.

Cutler, D. M., and Lleras-Muney, A. (2006) Education and health: evaluating theories and evidence. Working Paper 12352, National Bureau of Economic Research.

Daniel, R. M., S. N. Cousens, B. L. De Stavola, M. G. Kenward, J. A. C. Sterne. (2013) Methods for dealing with time-dependent confounding. *Stat. Med.*, 32, 1584-1618.

Efron, B. and R. J. Tibshirani. 1994. *An introduction to the bootstrap.* New York: Chapman and Hall/CRC.

Elwert, F. and Winship, C. (2014) Endogenous selection bias: the problem of conditioning on a collider variable. *Annu. Rev. Soc.*, 40, 31-53.

Heckman, J. J., Humphries, J. E., and Veramendi, G. (2018) The nonmarket benefits of education and ability. *J. Hum. Cap.*, 12, 282-304.

Holland, P. W. (1986) Statistics and causal inference. *J. Am. Stat. Assoc.*, 81, 945-960.

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

Lee, J. (2011) Pathways from education to depression. *J. Cross Cult. Geron.*, 26, 121-135.

Leventhal, T., and Brooks-Gunn, J. (2000) The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes. *Psychol. Bull.*, 126, 309-337.

Lunceford, J. K., and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study." *Stat. Med.*, 23, 2937-2960.

Michigan Survey Research Center. (2014) Panel Study of Income Dynamics Public- and Restricted-Use Datasets. Ann Arbor, MI: Institute for Social Research.

Morgan, S. L., and Winship, C. (2015) *Counterfactuals and Causal Inference: Methods and Principles for Social Research (2nd Ed.)*. New York: Cambridge University Press.

Naimi, A. I., Moodie, E. E. M., Auger, N., and Kaufman, J. S. (2014) Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology*, 25, 292-299.

Naimi, A. I., Cole, S. R., and Kennedy, E. H. (2017) An introduction to G methods. *Int. J. Epidemiol.*, 46, 756-762.

Pearl, J. (2009) *Causality (2nd Edition).* New York: Cambridge University Press.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992) Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48, 479-495.

Robins, J. M. (1994) Correcting for noncompliance in randomized trials using structural nested mean models. *Commun. Stat. A-Theor.*, 23, 2379-2412.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.*, 89, 846-866.

Robins, J. M., Hernán, M. A., and Rotnitzky, A. (2007) Effect modification by time-varying covariates. *Am. J. Epidemiol.*, 166, 994-1002.

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

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66, 688-701.

Rubin, D. B. (1986) Which ifs have causal answers. *J. Am. Stat. Assoc.*, 81, 961-962.

Sampson, R. J., Sharkey, P., and Raudenbush, S. W. (2008) Durable effects of concentrated disadvantage on verbal ability among African-American children. *P. Natl. A. Sci.*, 105, 845-852.

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

VanderWeele, T. (2015) *Explanation in Causal Inference: Methods for Mediation and Interaction.* New York: Oxford University Press.

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

Vansteelandt, S., and Sjolander, A. (2016) Revisiting G-estimation of the effect of a time-varying exposure subject to time-varying confounding. *Epidemiol. Meth.*, 5, 37-46.

Wodtke, G. T. (2018) Regression-based adjustment for time-varying confounders. *Sociol. Method. Res*. E-pub ahead of print.

Wodtke, G. T., and Almirall, D. (2017) Estimating moderated causal effects with time-varying treatments and time-varying moderators: structural nested mean models and regression-with-residuals. *Sociol. Methodol.*, 47, 212-245.

Wodtke, G. T., Harding, D. J., and Elwert, F. (2011) Neighborhood effects in temporal perspective: the impact of long-term exposure to concentrated disadvantage on high school graduation. *Am. Sociol. Rev.*, 76, 713-736.

Wodtke, G. T., Elwert, F., and Harding, D. J. (2016) Neighborhood effect heterogeneity by family income and developmental period. *Am. J. Sociol.*, 121, 1168-1222.

Woodcock, R. W., and Bonner Johnson, M. (1989) *Tests of Achievement, Standard Battery (Form B)*. Chicago, IL: Riverside.

Xie, Y. (2007) Otis Dudley Duncan's legacy: the demographic approach to quantitative reasoning in social science. *Res. Soc. Strat. Mobil.*, 25, 141-156.

Zhou, X., and Wodtke, G. T. (2019) A regression-with-residuals method for estimating controlled direct effects. *Polit. Anal*.

Figures

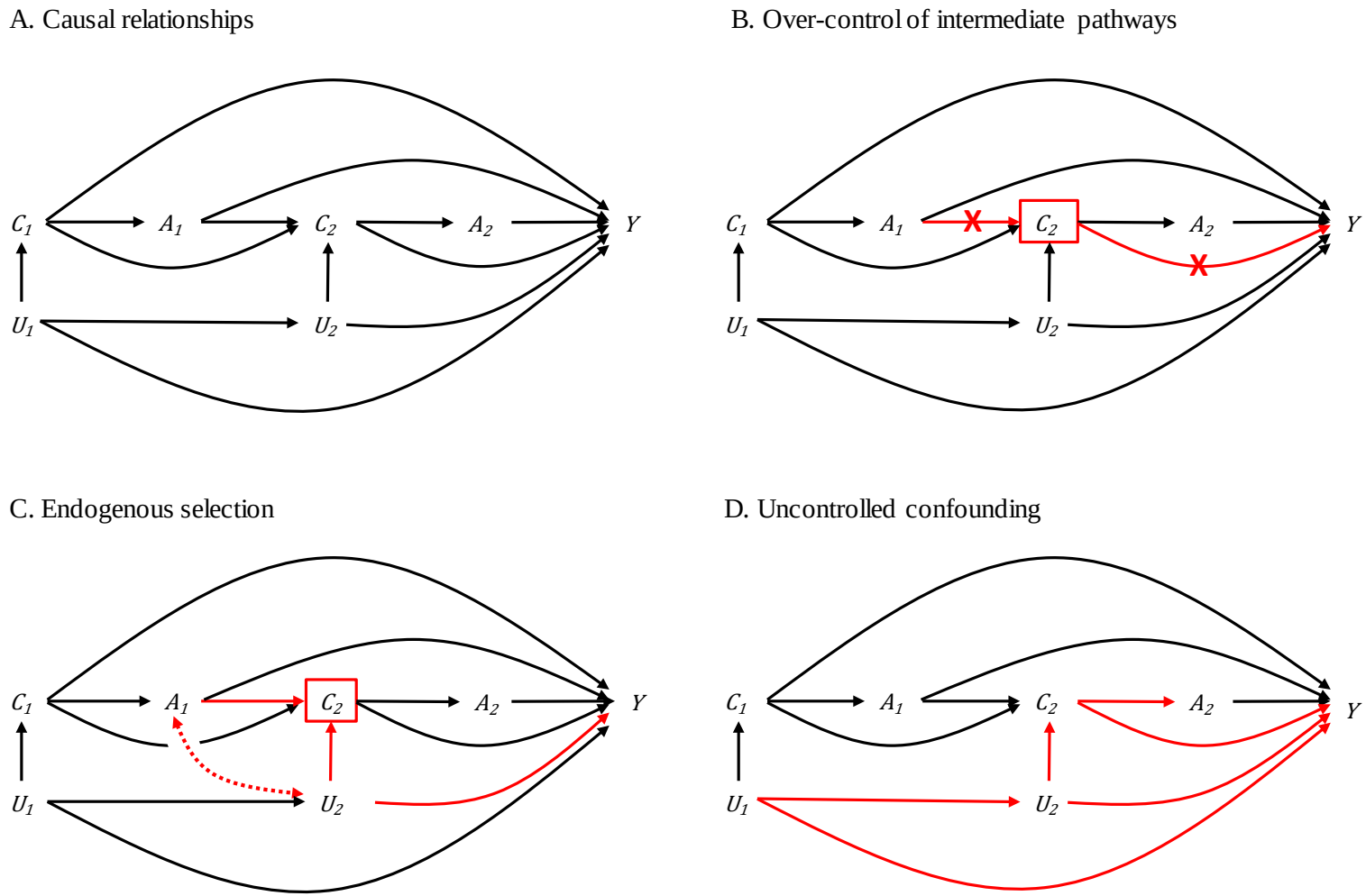

Figure 1. Directed acyclic graphs illustrating a set of causal relationships between a time-varying treatment (A_t), an observed time-varying confounder (C_t), an unobserved time-varying covariate (U_t), and an outcome (Y)

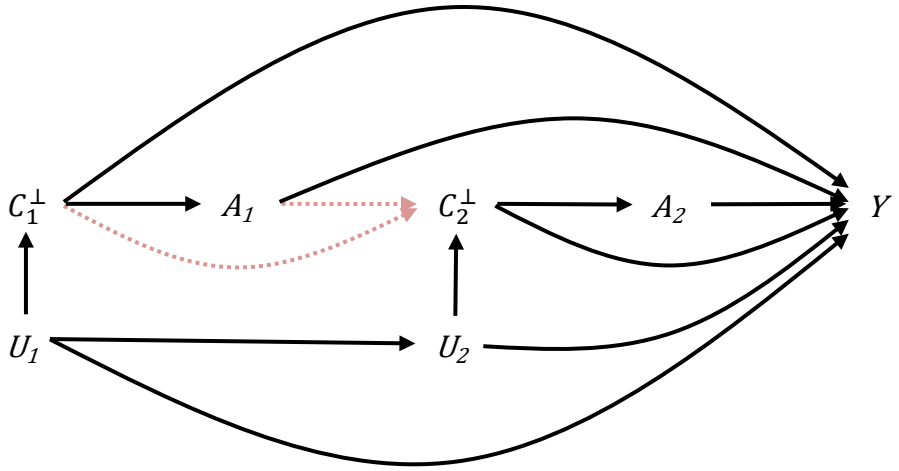

Figure 2. A stylized graph illustrating the logic of regression-with-residuals

Notes: A_t denotes a time-varying treatment, $C_1^{\perp} = C_1 - E(C_1)$ and $C_2^{\perp} = C_2 - E(C_2|C_1, A_1)$ denote residualized time-varying confounders, U_t denotes an unobserved time-varying covariate, and Y denotes an end-of-study outcome.

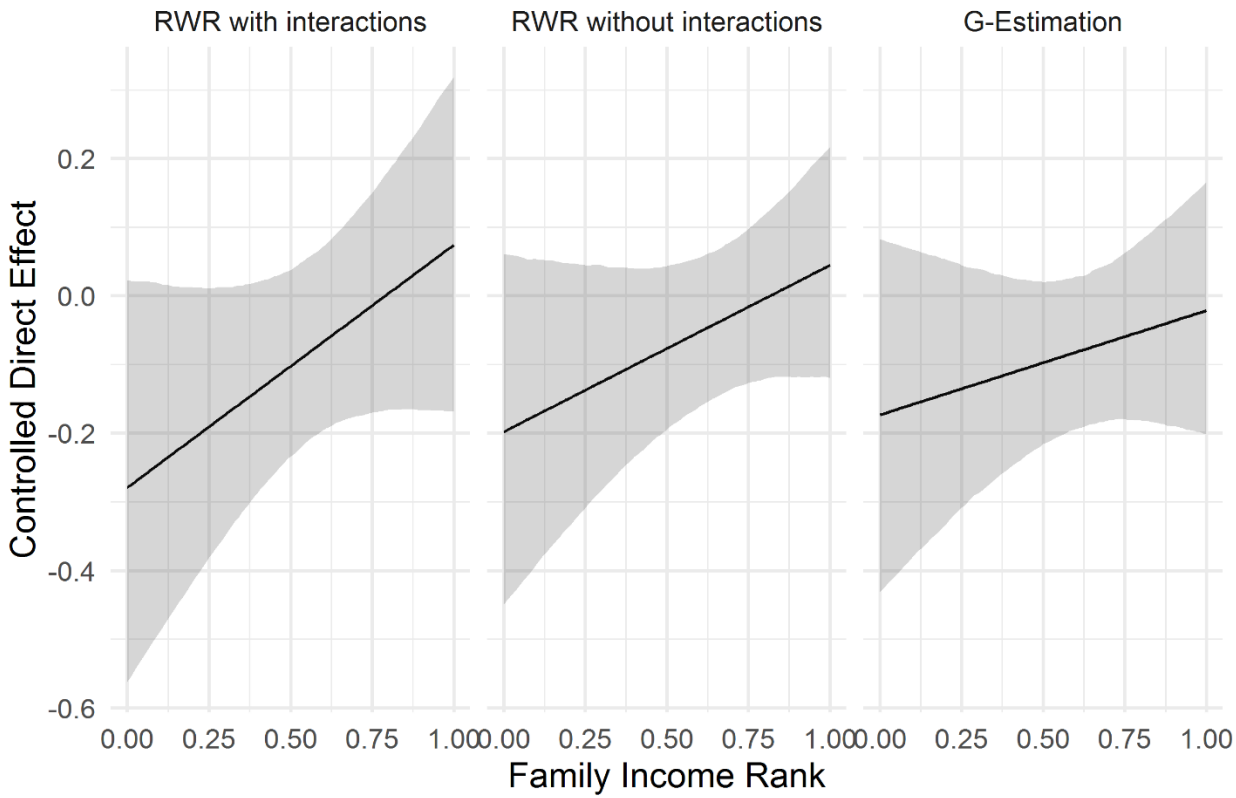

Figure 3. Estimated controlled direct effects of college completion on depression (controlling for family income)

Notes: Sample includes respondents to the 1979 National Longitudinal Survey of Youth who were age 13-17 when first interviewed. Confidence intervals are based on the nonparametric bootstrap with 1,000 replications. RWR = regression-with-residuals.

Tables

Table 1. Results from simulation experiments evaluating the performance of RWR relative to other estimators under different levels of treatment-outcome confounding

Estimator/statistic	Magnitude of treatment-outcome confounding				
	γ=0.1	γ=0.2	γ=0.3	γ=0.4	γ=0.5
Conventional regression					
Bias	-0.150	-0.200	-0.252	-0.299	-0.351
SD	0.134	0.137	0.141	0.145	0.151
RMSE	0.201	0.242	0.288	0.332	0.382
IPTW estimation					
Bias	0.000	0.002	0.005	0.020	0.057
SD	0.135	0.147	0.176	0.228	0.296
RMSE	0.135	0.147	0.176	0.229	0.302
G-estimation					
Bias	0.000	0.000	-0.001	0.002	0.000
SD	0.134	0.139	0.145	0.152	0.163
RMSE	0.134	0.139	0.145	0.152	0.163
RWR w/o interactions					
Bias	0.000	0.000	-0.001	0.002	0.000
SD	0.134	0.139	0.145	0.151	0.161
RMSE	0.134	0.139	0.145	0.151	0.161
RWR w/ interactions					
Bias	0.000	0.000	-0.001	0.002	-0.001
SD	0.134	0.140	0.146	0.154	0.164
RMSE	0.134	0.140	0.146	0.154	0.164

Notes: SD = standard deviation; RMSE = root mean squared error; IPTW = inverse-probability-of-treatment-weighted; RWR = regression-with-residuals. Across all simulations, Θ=0 and η=0. Results are based on 10,000 simulations. See the Online Supplement for details.

Table 2. Results from simulation experiments evaluating the performance of RWR relative to other estimators under different levels of treatment effect modification and a moderate-to-high level of treatment-outcome confounding

Estimator/statistic	Magnitude of treatment effect modification				
	θ=0.1	θ=0.2	θ=0.3	θ=0.4	θ=0.5
Conventional regression					
Bias	-0.369	-0.439	-0.508	-0.575	-0.645
SD	0.145	0.149	0.151	0.155	0.163
RMSE	0.396	0.463	0.530	0.595	0.665
IPTW estimation					
Bias	0.022	0.024	0.023	0.028	0.021
SD	0.235	0.246	0.261	0.274	0.299
RMSE	0.236	0.247	0.262	0.275	0.300
G-estimation					
Bias	-0.023	-0.047	-0.071	-0.094	-0.119
SD	0.155	0.161	0.164	0.168	0.177
RMSE	0.157	0.168	0.179	0.193	0.213
RWR w/o interactions					
Bias	-0.037	-0.076	-0.115	-0.151	-0.192
SD	0.154	0.161	0.166	0.171	0.182
RMSE	0.159	0.178	0.202	0.228	0.264
RWR w/ interactions					
Bias	0.001	0.001	0.001	0.000	-0.001
SD	0.156	0.161	0.164	0.167	0.175
RMSE	0.156	0.161	0.164	0.167	0.175

Notes: SD = standard deviation; RMSE = root mean squared error; IPTW = inverse-probability-of-treatment-weighted; RWR = regression-with-residuals. Results are based on 10,000 simulations. Across all simulations, γ=0.4 and η=0. See the Online Supplement for details.

Table 3. Results from simulation experiments evaluating the performance of RWR relative to other estimators under different levels of nuisance model misspecification and moderate-to-high levels of both treatment-outcome confounding and treatment effect modification

Estimator/statistic	Magnitude of associational effect modification				
	η=0.1	η=0.2	η=0.3	η=0.4	η=0.5
Conventional regression					
Bias	-0.586	-0.597	-0.608	-0.616	-0.627
SD	0.158	0.162	0.164	0.166	0.174
RMSE	0.606	0.618	0.630	0.638	0.651
IPTW estimation					
Bias	0.030	0.031	0.029	0.033	0.027
SD	0.276	0.280	0.292	0.296	0.318
RMSE	0.277	0.282	0.294	0.297	0.319
G-estimation					
Bias	-0.093	-0.094	-0.095	-0.094	-0.095
SD	0.169	0.173	0.175	0.176	0.183
RMSE	0.193	0.197	0.199	0.200	0.206
RWR w/o interactions					
Bias	-0.160	-0.171	-0.182	-0.189	-0.201
SD	0.173	0.176	0.179	0.179	0.188
RMSE	0.236	0.246	0.255	0.261	0.275
RWR w/ interactions					
Bias	0.015	0.026	0.039	0.050	0.063
SD	0.169	0.172	0.174	0.176	0.185
RMSE	0.169	0.174	0.179	0.183	0.195

Notes: SD = standard deviation; RMSE = root mean squared error; IPTW = inverse-probability-of-treatment-weighted; RWR = regression-with-residuals. Results are based on 10,000 simulations. Across all simulations, γ=0.4 and θ=0.4. See the Online Supplement for details.

Table 4. Estimated marginal effects of exposure to disadvantaged neighborhoods on end-of-study math achievement

Estimator/statistic	$DTE\,(1,0)$		$PTE\,(a_1,1)$			CTE		
	Est	SE	Est	SE		Est	SE	
RWR with interactions	-0.034	(0.049)	-0.094	(0.046)	*	-0.127	(0.038)	***
RWR without interactions	-0.030	(0.044)	-0.097	(0.040)	*	-0.127	(0.038)	***
G-estimation	-0.032	(0.040)	-0.096	(0.041)	*	-0.127	(0.047)	**

Notes: Sample includes respondents who were interviewed at the 1997 wave of the Child Development Supplement between age 3 and 7. Results are combined estimates from 100 imputations. The outcome is standardized to have zero mean and unit variance. Standard errors are based on the block boostrap with 1,000 replications. Est = point estimate; SE = standard error; RWR = regression-with-residuals; DTE = distal treatment effect; PTE = proximal treatment effect; CTE = cumulative treatment effect.

†$p < 0.10$, *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$ for two-sided tests of no effect.