A regression-with-residuals method for analyzing causal mediation: The rwrmed package

Ariel Linden
Linden Consulting Group, LLC
San Francisco, CA, USA
alinden@lindenconsulting.org

Chuck Huber
StataCorp, LLC
chuber@stata.com

Geoffrey T. Wodtke
Department of Sociology
University of Chicago
wodtke@uchicago.edu

Abstract. This article introduces the `rwrmed` package, which performs mediation analysis using the methods proposed by Wodtke and Zhou (2020, *Epidemiology*, 31: 369-375). Specifically, `rwrmed` estimates interventional direct and indirect effects in the presence of treatment-induced confounding by fitting models for (1) the conditional mean of the mediator given the treatment and a set of baseline confounders and (2) the conditional mean of the outcome given the treatment, mediator, baseline confounders, and a set of treatment-induced confounders that have been residualized with respect to the observed past. Interventional direct and indirect effects are simple functions of the parameters in these models when they are correctly specified and when there are no unobserved variables that confound the treatment–outcome, treatment–mediator, or mediator–outcome relationships. When no treatment-induced confounders are specified, `rwrmed` produces natural direct and indirect effect estimates.

Keywords: mediation, effect decomposition, causal inference, confounding

1 Introduction

When evaluating the effectiveness of an intervention, many analysts focus exclusively on estimating the overall association between a treatment and outcome. A common criticism of this narrow focus is that it precludes the discovery of causal mechanisms by which the intervention is hypothesized to affect the outcome. Mediation analysis, by contrast, aims to identify intermediate variables that transmit the effect of treatment on the outcome (Linden and Karlson 2013;

VanderWeele 2015). To this end, analyses of causal mediation typically focus on decomposing an overall effect of treatment into an indirect component operating through a mediator of interest and a direct component operating through alternative pathways.

In Stata, users can perform mediation analysis using `sem` followed by `estat teffects` to compute direct, indirect, and total effects (see [SEM] estat teffects). Additionally, UCLA's Statistical Consulting Group has written a series of FAQs extending basic mediation analysis to more complicated cases using official Stata commands (UCLA Statistical Consulting Group 2020). Several community-contributed packages are also currently available for analyzing causal mediation, including `medeff` (Hicks and Tingley 2011), `khb` (Kohler, Karlson, and Holm 2011), `paramed` (Emsley and Liu 2013), `ldecomp` (Buis 2010) `gformula` (Daniel, De Stavola, and Cousens 2011), `cta` (Linden 2020) which implements the approach described by Linden and Yarnold (2018) and `ivmediate` (Dippel et al. 2020). Additionally, `riskplot` (Falcaro and Pickles 2010) may be considered as a graphical aid when examining possible mediation of effects.

In this paper, we introduce the `rwrmed` package, which performs mediation analysis using the methods proposed by Wodtke and Zhou (2020) for decomposing treatment effects in the presence of treatment-induced confounding. Specifically, `rwrmed` decomposes an overall effect into *interventional* direct and indirect effects (defined in Section 2) using a "regression-with-residuals" approach. With this approach, interventional effects are estimated by fitting models for (1) the conditional mean of the mediator given the treatment and a set of baseline confounders and (2) the conditional mean of the outcome given the treatment, mediator, baseline confounders, and a set of treatment-induced confounders that have been residualized with respect to the observed past. Regression-with-residuals estimates are consistent and asymptotically normal when these models are correctly specified and there are no unobserved variables that confound the treatment–outcome, treatment–mediator, or mediator–outcome relationships (Wodtke and Zhou 2020; Zhou and Wodtke 2019).

2 Method and Formulas

Evaluating causal mediation is typically accomplished by decomposing the average total effect of a treatment on an outcome into the sum of so-called "natural" direct and indirect effects (VanderWeele 2015). The natural direct effect is the expected difference in the outcome if each unit in the target population were exposed, rather than unexposed, to treatment and then were exposed to the level of the mediator they would have naturally experienced had they not received treatment. The natural indirect effect is the expected difference in the outcome if each unit were exposed to treatment and then were exposed to the level of the mediator they experience under this treatment rather than the level of the mediator they would have experienced had they not received treatment. Taken together, natural direct and indirect effects neatly separate the total effect into components operating through the mediator of interest versus alternative pathways. They can be estimated in Stata using `paramed` (Emsley and Liu 2013) or `gsem` followed by `nlcom`, among several other options.

A key limitation of mediation analyses focused on natural direct and indirect effects, however, is that these estimands are not identified in the presence of treatment-induced confounders – at least not without invoking strong parametric assumptions (VanderWeele 2015; VanderWeele, Vansteelandt, and Robins 2014). Treatment-induced confounders are posttreatment variables that affect both the mediator and outcome and that are also affected by treatment. They are empirically common in analyses of causal mediation, wherever the effects of treatment operate through multiple intermediate variables that may influence one another. Thus, in many applications, a focus on natural direct and indirect effects is not warranted, and alternative estimands that must be learned from the data using alternative methods are needed.

In this section, we introduce a set of interventional direct and indirect effects that can be identified in the presence of treatment-induced confounders. We then explain how they can be estimated using the method of regression-with-residuals. We conclude by discussing several situations where natural and interventional effects are equivalent.

2.1 Interventional Direct and Indirect Effects

Let Y denote an outcome variable, A the treatment, M a mediator of interest, C a set of pretreatment covariates, and L a set of posttreatment covariates.

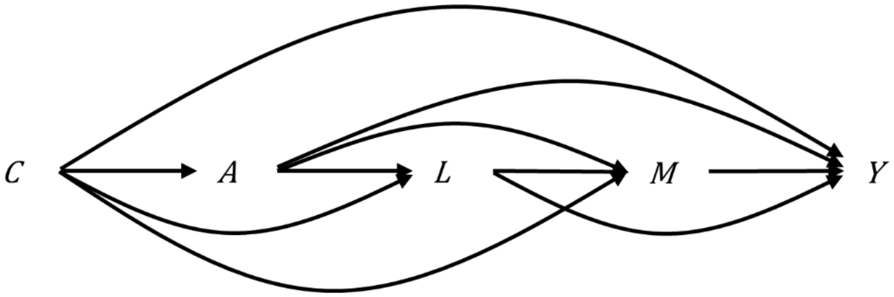


Figure 1. A directed acyclic graph summarizing causal relationships between the baseline covariates (C), treatment (A), posttreatment covariates (L), mediator (M), and outcome (Y).

In Figure 1, we summarize the causal relationships between these variables using a directed acyclic graph (Pearl 2009), where arrows between nodes represent direct causal effects of arbitrary functional form. This graph shows that M mediates the effect of A on Y, as indicated by the $A \to M \to Y$ and $A \to L \to M \to Y$ paths. It also shows that A affects Y through mechanisms that do not involve the mediator of interest, M, as indicated by the $A \rightarrow Y$ and $A \rightarrow L \rightarrow Y$ paths. Finally, it shows, via the $Y \leftarrow L \rightarrow M$ and $A \rightarrow L$ paths, that L confounds the effect of M on Y and is also affected by A. The covariates in L are therefore treatment-induced confounders of the mediator-outcome relationship. The `rwrmed` package is designed for analyzing whether and to what extent an overall effect of a treatment, A, on an outcome, Y, is mediated via a focal intermediate variable, M, when the data are generated by a process resembling this graph, where there are treatment-induced confounders, such as L. To this end, it focuses on interventional direct and indirect effects.

Interventional direct and indirect effects are defined using potential outcomes notation (VanderWeele 2015; VanderWeele et al. 2014). Specifically, let Y_a and M_a denote the potential values of the outcome and mediator, respectively, that would have been observed under exposure to treatment a. Similarly, let Y_{am} denote the potential value of the outcome under exposure to treatment a and the level of the mediator given by m. Finally, let $\mathcal{M}_{a|c}$ denote a value of the mediator randomly selected from its population distribution under exposure to treatment a conditional on the pretreatment covariates $C = c$.

With this notation, a randomized intervention analogue to the average total effect of treatment on the outcome can be defined as follows:

$$\mathrm{r}ATE(c) = E\left(Y_{a^*\mathcal{M}_{a^*|c}} - Y_{a\mathcal{M}_{a|c}} | C = c\right),$$

This effect represents the expected difference in the outcome if units were exposed to treatment a^* rather than a and to a value of the mediator randomly selected from its distribution under each of these alternative treatments among the subpopulation defined by $C = c$. It is similar to a conventional total effect except that it is defined in terms of both a contrast between different treatments and a randomized intervention on the population distribution of the mediator.

The $\mathrm{r}ATE(c)$ can be additively decomposed into interventional direct and indirect effects as follows:

$$\begin{aligned}\mathrm{r}ATE(c) &= \mathrm{r}NIE(c) + \mathrm{r}NDE(c) \\ &= E\left(Y_{a^*\mathcal{M}_{a^*|c}} - Y_{a^*\mathcal{M}_{a|c}} | C = c\right) + E\left(Y_{a^*\mathcal{M}_{a|c}} - Y_{a\mathcal{M}_{a|c}} | C = c\right).\end{aligned}$$

The first expression in this decomposition, $\mathrm{r}NIE(c) = E\left(Y_{a^*\mathcal{M}_{a^*|c}} - Y_{a^*\mathcal{M}_{a|c}} | C = c\right)$, is a randomized intervention analogue of the natural indirect effect. This effect represents the expected difference in the outcome if units with covariates $C = c$ were exposed to treatment a^* and then were exposed to a level of the mediator randomly selected from its distribution under treatment a^* rather than a. It captures an effect of A on Y that is mediated through M by fixing treatment at a^* and then comparing outcomes with the mediator randomly selected from its distribution under different levels of treatment.

The second expression in this decomposition, r$NDE(c) = E\left(Y_{a^*\mathcal{M}_{a|c}} - Y_{a\mathcal{M}_{a|c}}|C = c\right)$, is a randomized intervention analogue of the natural direct effect. This effect represents the expected difference in the outcome if units with covariates $C = c$ were exposed to treatment a^* rather than a and then were exposed to a level of the mediator randomly selected from its distribution under treatment a. It captures an effect of A on Y that is not mediated through M by fixing the distribution from which the mediator is assigned and then comparing outcomes under different treatments.

The r$NDE(c)$ can also be expressed as a function of the controlled direct effect, which is another interventional estimand that is frequently targeted in analyses of causal mediation (VanderWeele 2015). Specifically, the r$NDE(c)$ can be further decomposed as follows:

$$\begin{aligned}\text{r}NDE(c) &= CDE(c,m) + \text{rINT}_{\text{ref}}(c,m)\\ &= E(Y_{a^*m} - Y_{am}|C = c)\\ &\quad + \left\{E\left(Y_{a^*\mathcal{M}_{a|c}} - Y_{a\mathcal{M}_{a|c}}|C = c\right) - E(Y_{a^*m} - Y_{am}|C = c)\right\},\end{aligned}$$

where the controlled direct effect, $CDE(c,m) = E(Y_{a^*m} - Y_{am}|C = c)$, captures the expected difference in the outcome if units with covariates $C = c$ were exposed to treatment a^* rather than a after fixing the level of mediator at the same value m for all. The r$NDE(c)$ differs from the $CDE(c,m)$ conceptually in that the former involves an intervention to fix the population distribution of the mediator whereas the later involves an intervention to fix the mediator at the same specific value for each unit. The r$NDE(c)$ differs from the $CDE(c,m)$ computationally depending on the magnitude of a treatment-mediator interaction effect, $\text{rINT}_{\text{ref}}(c,m) = \left\{E\left(Y_{a^*\mathcal{M}_{a|c}} - Y_{a\mathcal{M}_{a|c}}|C = c\right) - E(Y_{a^*m} - Y_{am}|C = c)\right\}$, occurring in the absence of mediation. If there is no treatment-mediator interaction, the r$NDE(c)$ is equal to the $CDE(c,m)$.

All of the interventional effects defined previously can be identified from the observed data under the following conditional independence assumptions, which involve restrictions on the joint distribution of the potential outcomes and then the observed treatment and mediator:

i. $Y_{am} \perp A|C$
ii. $M_a \perp A|C$
iii. $Y_{am} \perp M|C, A, L,$

where $\perp$ denotes statistical independence (VanderWeele et al. 2014). Informally, assumption (i) requires that there must not be any unobserved confounding of the treatment-outcome relationship, conditional on the pretreatment covariates. Assumption (ii) requires that there must not be any unobserved confounding of the treatment-mediator relationship, conditional on the pretreatment covariates. And assumption (iii) requires that there must not be any unobserved confounding of the mediator-outcome relationship, conditional on treatment and then both the pretreatment and posttreatment covariates. These assumptions are satisfied in Figure 1, for example, because there are no unobserved variables that jointly affect A and Y, M and Y, or A and M.

2.2 Regression-with-residuals

Although interventional direct and indirect effects can be identified when there is no unobserved confounding of the treatment-outcome, treatment-mediator, and mediator-outcome relationships, estimating them from the observed data poses special challenges even when these conditions are met. The central challenge arises from the need to adjust for posttreatment covariates L. Failing to adjust for these covariates may lead to bias from uncontrolled confounding of the mediator-outcome relationship. At the same time, adjusting for them naively using conventional methods of covariate control may also lead to bias. In particular, naively adjusting for L may lead to bias from over-control of mediating pathways, as any posttreatment covariate may also transmit the effects of A on Y. It may also lead to bias from endogenous selection, as naively adjusting for posttreatment covariates can induce a spurious association between treatment and the outcome (Elwert and Winship 2014).

Regression-with-residuals avoids these biases by adjusting for posttreatment covariates only after they have been residualized with respect to the observed past. Adjusting for these residualized covariates appropriately controls for mediator-outcome confounding while circumventing the problems of over-control and endogenous selection, as the residuals are purged of their association with treatment by design. Specifically, regression-with-residuals estimates of interventional direct and indirect effects come from the following set of models for L, M, and Y:

$$g\big(E(L|C,A)\big) = \tau_0 + \tau_1^T C^\perp + \tau_2 A$$

$$h\big(E(M|C,A)\big) = \theta_0 + \theta_1^T C^\perp + \theta_2 A$$

$$E(Y|C,A,L,M) = \beta_0 + \beta_1^T C^\perp + \beta_2 A + \beta_3^T L^\perp + \beta_4 M + \beta_5 AM,$$

where $g(\cdot)$ and $h(\cdot)$ are smooth and invertible link functions, $C^\perp = C - E(C)$ is a vector of mean-centered pretreatment covariates, and $L^\perp = L - E(L|C,A) = L - g^{-1}(\tau_0 + \tau_1^T C^\perp + \tau_2 A)$ is a vector of residual terms for the posttreatment covariates. The models for L and M are nearly identical to conventional generalized linear models except that the pretreatment covariates C have been centered around their marginal means. Similarly, the model for Y is nearly identical to a conventional linear regression except that the posttreatment covariates L have additionally been centered around their conditional means given C and A.

Under assumptions (i), (ii), and (iii) and provided that all of the models outlined previously are correctly specified, the randomized intervention analogues of natural direct and indirect effects are equal to the following parametric expressions:

$$\mathrm{r}NIE(c) = (\beta_4 + \beta_5 a^*)\ [h^{-1}(\theta_0 + \theta_1^T c^\perp + \theta_2 a^*) - h^{-1}(\theta_0 + \theta_1^T c^\perp + \theta_2 a)]$$

$$\text{r}NDE(c) = [\beta_2 + \beta_5 h^{-1}(\theta_0 + \theta_1^T c^{\perp} + \theta_2 a)](a^* - a),$$

and the rATE(c) is equal to their sum. Under assumptions (i) and (iii) and provided that the models for L and Y are correctly specified, the controlled direct effect is equal to:

$$CDE(c, m) = [\beta_2 + \beta_5 m](a^* - a).$$

Although the models outlined previously omit treatment-covariate and mediator-covariate interactions, such that the $CDE(c, m)$ is not in fact a function of c, these terms can be easily incorporated to allow for different patterns of effect moderation. By default, `rwrmed` evaluates all interventional effects at the sample means of the pretreatment covariates.

`rwrmed` computes regression-with-residuals estimates of these quantities as follows. First, the models for L are estimated using `regress` for continuous variables and `logit` for binary variables (for multi-categorical variables, indicator [dummy] variables are created and then passed on to `logit`) followed by the generation of residual terms. Second, the models for M and Y are estimated simultaneously with `gsem` using the residual terms from the first step where appropriate. Third, the coefficients from these models are used to construct estimates for the interventional effects of interest with the expressions outlined previously. Linear, logistic, or Poisson models with their canonical link functions may be used for the mediator, as appropriate depending on its level of measurement. A linear model is required for Y, and thus regression-with-residuals is best suited for applications with metric outcomes. Nevertheless, it may still be used with binary or ordinal outcomes if a linear model provides a defensible approximation to the true but unknown conditional expectation function (Kohler et al. 2011; Wodtke and Almirall 2017). In this situation, effect estimates would be based on a linear probability or ordinal mean model for Y, with all their attendant limitations (Agresti 2002).

The outcome model described above includes a treatment-mediator interaction. This interaction may be constrained to equal zero, in which case the r$NDE(c)$ is equal to the $CDE(c)$. Moreover, when there is no treatment-mediator interaction, natural direct and indirect effects are equal to their interventional analogues. Natural direct and indirect effects are also equal to their interventional analogues when there are not any treatment-induced confounders. Thus, `rwrmed` computes and reports natural direct and indirect effects if the treatment-mediator interaction is suppressed or no posttreatment covariates are supplied. Otherwise, it provides their interventional analogues.

Valid standard errors for regression-with-residuals estimates can be computed using the nonparametric bootstrap (Almirall et al. 2010; Wodtke and Almirall 2017). Analytic standard errors can also be approximated via the delta method using the combined variance-covariance matrix of the parameter estimates from models for M and Y. These standard errors are approximations because they do not account for the fact the residual terms used to fit the outcome model are themselves estimated and thus subject to their own sampling variability. Nevertheless, rwrmed provides them because they are computationally expedient, and in simulation studies, they appear to provide a reasonable approximation to the true standard deviations of the effect estimates under repeated sampling across several different scenarios. Strictly valid inferences, however, can only be obtained via bootstrapping at present.

3 The rwrmed package

This section describes the syntax of the rwrmed package and various options available.

3.1 Syntax

```
rwrmed depvar [lvars] [if] [in] [pweight] , avar(varname)
 mvar(varname) a(#) astar(#) m(#) [ mreg(string) cvar(varlist)
 cat(varlist) nointeraction cxa cxm lxm noisily
 bootstrap[(bootstrap_options)] model_options ]
```

lvars are post-treatment covariates

vce(robust) and vce(cluster clustvar) are allowed

3.2 Options

avar(varname) specifies the treatment (exposure) variable; **avar()** is required.

mvar(varname) specifies the mediator variable; **mvar()** is required.

a(#) sets the reference level of treatment; **a()** is required.

astar(#) sets the alternative level of treatment. Together, (astar - a) defines the treatment contrast of interest; **astar()** is required.

m(#) sets the level of the mediator at which the controlled direct effect is evaluated. If there is no treatment-mediator interaction, then the controlled direct effect is the same at all levels of the mediator and thus an arbitrary value can be chosen; **m()** is required.

mreg(string) specifies the form of regression model to be estimated for the mediator variable. Options include regress, logit, and poisson; default is regress.

cvars(varlist) specifies the pre-treatment covariates to be included in the analysis.

cat(varlist) specifies which of the *cvars* and *lvars* should be handled as categorical variables. For categorical variables with more than two levels, rwrmed generates dummy variables for each level and then residualizes them individually. A warning message will be issued if the logit model produces perfect predictions, resulting in dropped observations. The program will terminate if the logit model cannot converge. In both of these cases (dropped observations or model non-convergence), the user should consider either collapsing the multi-categorical variable into fewer categories or specifying it as a continuous variable by not adding it to **cat()** if appropriate.

`nointeraction` specifies that a treatment-mediator interaction should not be included in the outcome model. When not specified, `rwrmed` will generate a treatment-mediator interaction term.

`cxa` specifies that all two-way interactions between treatment and the baseline covariates be included in the mediator and outcome models.

`cxm` specifies that all two-way interactions between the mediator and the baseline covariates be included in the outcome model.

`lxm` specifies that all two-way interactions between the mediator and the posttreatment covariates be included in the outcome model.

`noisily` displays the GSEM output tables; this option is not available with bootstrap.

`bootstrap[(`*`bootstrap_options`*`)]` specifies that bootstrap replications are to be used to estimate the variance-covariance matrix. All bootstrap options are available. Specifying bootstrap without options uses the default bootstrap settings.

`model_options` allows the user to specify any option available for gsem

4 Examples

In this section, we use `rwrmed` to decompose the effect of negative media framing on support for immigration into direct and indirect components using data from Brader, Valentino, and Suhay (2008). These researchers conducted a survey experiment on a nationally representative sample of white non-Hispanic adults. For this experiment, respondents were asked to read a mock news report on immigration where both the ethnicity of the featured immigrant and the tone of the story were randomly manipulated. 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 emotional reactions toward the prospect of increased immigration, and their support for further immigration to the U.S. With these data, Brader et al.

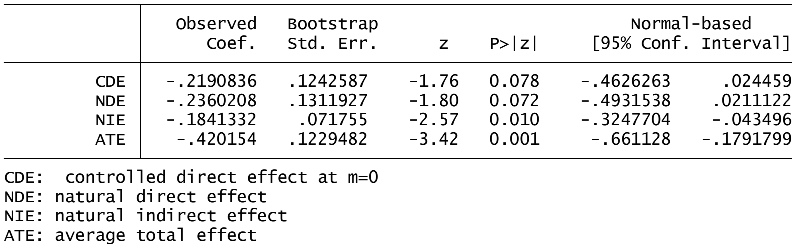
(2008) found that stories featuring a Latino immigrant and a frame emphasizing the costs of immigration had a large negative effect on support for immigration. We extend this analysis by using RWR to examine whether this effect may be mediated by respondents' emotional reactions to the news story, adjusting for their beliefs about the harms of immigration.

The outcome, *depvar*, is a measure of support for immigration on a five-point scale, with response categories ranging from the statement that immigration should be "decreased a lot" to the statement that it should be "increased a lot." The treatment, *avar*, denotes receipt of a news story featuring both a Latino immigrant and a negative frame emphasizing the costs of immigration. The mediator, *mvar*, is the level of anxiety expressed by the respondent about the prospect of increased immigration on a ten-point scale. The vector of baseline confounders, *cvars*, includes measures of gender, age, education, and income. Finally, a potentially important posttreatment confounder, included as an *lvar*, is respondents' beliefs about the harms of immigration on a seven-point scale, which was constructed from questions asking respondents about the likelihood that immigration will negatively impact the "finances" and "way of life" of American communities.

4.1 A Conventional Analysis of Mediation

To begin, we implement `rwrmed` without including any posttreatment covariates, in which case the command will produce estimates of the natural direct effect (NDE), natural indirect effect (NIE), and average total effect (ATE). In the following syntax, we specify the outcome (immigr), the mediator (emo), the binary treatment (tone_eth), four pre-treatment covariates (of which two are categorical), an interaction between treatment and mediator (indicated by not specifying the *nointer* option), and the bootstrap method for variance estimation with 10,000 repetitions and a seed (1234) to allow for replication of results:

```
. use immigration.dta

. rwrmed immigr, avar(tone_eth) mvar(emo) mreg(reg) ///

       a(0) astar(1) m(0) cvar(ppage female ppeducat ppincimp) ///

       cat(female ppeducat) boot(reps(10000) seed(1234))

(output omitted)
```



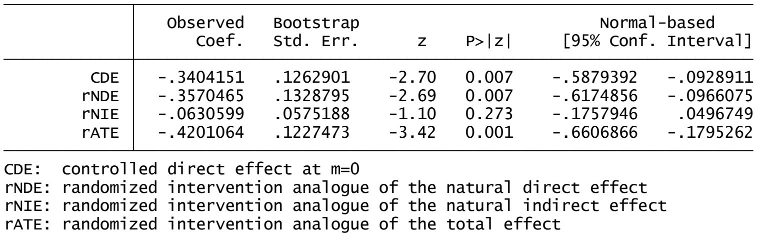
```
             Observed   Bootstrap                         Normal-based
                Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

        CDE  -.2190836   .1242587   -1.76   0.078    -.4626263     .024459
        NDE  -.2360208   .1311927   -1.80   0.072    -.4931538    .0211122
        NIE  -.1841332    .071755   -2.57   0.010    -.3247704    -.043496
        ATE   -.420154   .1229482   -3.42   0.001     -.661128   -.1791799

CDE:  controlled direct effect at m=0
NDE: natural direct effect
NIE: natural indirect effect
ATE: average total effect
```

As shown in the output, we estimate that negative media framing has a sizable total effect on support for immigration. Specifically, receipt of a news story featuring a Latino immigrant and emphasizing the costs of immigration is estimated to lower support for immigration by 0.42 points on average (95% CI: [–0.66, –0.18]). The NDE and NIE are estimated to be –0.24 (95% CI: [–0.49, 0.02]) and –0.18 (95% CI: [–0.32, -0.04]), respectively. This suggests that about 44% (NIE / ATE ≈ 0.44) of the overall effect is mediated by a negative emotional reaction to immigration All of the estimates reported here, however, are based on the assumption of no treatment-induced confounding, which may not be appropriate in this analysis. This is because beliefs about the harms of immigration likely affect both emotional reactions and levels of support for immigration, and they may also be affected by treatment.

4.2 Adjusting for Posttreatment Confounding

In the following syntax, we specify the outcome (immigr), the mediator (emo), the binary treatment (tone_eth), four pre-treatment covariates (of which two are categorical), one post-treatment covariate, an interaction between treatment and mediator (indicated by not specifying the *nointer* option), and the bootstrap method for estimation with 10,000 repetitions and a seed (1234) to allow for replication of results:

```
. rwrmed immigr p_harm, avar(tone_eth) mvar(emo) mreg(reg) ///

       a(0) astar(1) m(0) cvar(ppage female ppeducat ppincimp) ///

       cat(female ppeducat) boot(reps(10000) seed(1234))

(output omitted)
```



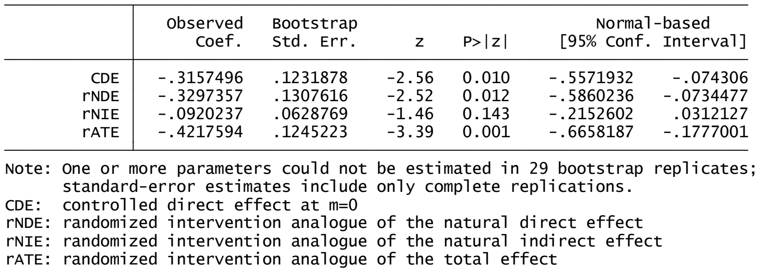
```
-------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
         CDE |  -.3404151   .1262901    -2.70   0.007    -.5879392   -.0928911
        rNDE |  -.3570465   .1328795    -2.69   0.007    -.6174856   -.0966075
        rNIE |  -.0630599   .0575188    -1.10   0.273    -.1757946    .0496749
        rATE |  -.4201064   .1227473    -3.42   0.001    -.6606866   -.1795262
-------------------------------------------------------------------------------
CDE:  controlled direct effect at m=0
rNDE: randomized intervention analogue of the natural direct effect
rNIE: randomized intervention analogue of the natural indirect effect
rATE: randomized intervention analogue of the total effect
```

As shown in the output, the regression-with-residuals estimates for the overall effect of exposure to a news story featuring a Latino immigrant and emphasizing the costs of immigration is similar to the total effect estimate from Section 4.1, which was based on conventional models that did not adjust for posttreatment confounding. Specifically, the estimate of the rATE indicates that the overall effect of treatment is to reduce support for immigration by about 0.42 points, on average (95% CI: [–0.66, –0.18]). The rNDE and rNIE estimates are very different, however. The rNDE estimate is –0.36 (95% CI: [–0.62, -0.10]), and the rNIE estimate is –0.06 (95% CI: [–0.18, 0.05]). This suggests that, after properly adjusting for posttreatment confounding by beliefs about the costs of immigration, only about 15% (rNIE / rATE ≈ 0.15) of the overall effect is mediated by a negative emotional reaction to immigration.

4.3 Adding Covariate Interactions

In the following syntax, we extend the model from Section 4.2 by adding two-way interactions between the treatment indicator and all pretreatment covariates, the mediator and all pretreatment covariates, and the mediator and all posttreatment covariates using the `cxa`, `cxm`, and `lxm` options, respectively.

```
. rwrmed immigr p_harm, avar(tone_eth) mvar(emo) mreg(reg) ///

       a(0) astar(1) m(0) cvar(ppage female ppeducat ppincimp) ///

       cat(female ppeducat) boot(reps(10000) seed(1234)) cxa cxm lxm

(output omitted)
```



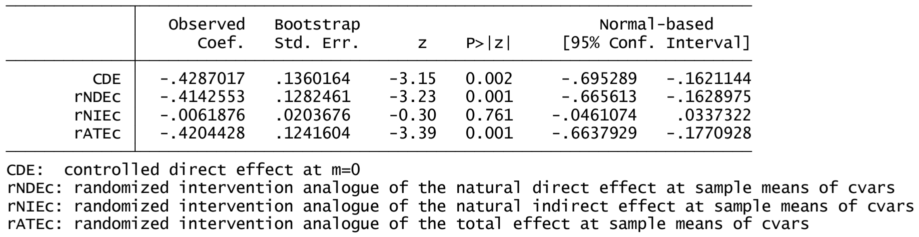
```
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         CDE |  -.3157496    .1231878    -2.56   0.010    -.5571932    -.074306
        rNDE |  -.3297357    .1307616    -2.52   0.012    -.5860236   -.0734477
        rNIE |  -.0920237    .0628769    -1.46   0.143    -.2152602    .0312127
        rATE |  -.4217594    .1245223    -3.39   0.001    -.6658187   -.1777001
------------------------------------------------------------------------------
Note: One or more parameters could not be estimated in 29 bootstrap replicates;
      standard-error estimates include only complete replications.
CDE:  controlled direct effect at m=0
rNDE: randomized intervention analogue of the natural direct effect
rNIE: randomized intervention analogue of the natural indirect effect
rATE: randomized intervention analogue of the total effect
```

As shown in the output, the interventional effect estimates are similar to those from the model in Section 4.2, with the indirect effect now accounting for about 22% of the overall effect (rNIE / rATE ≈ 0.22). That the estimates are stable when comparing a parsimonious model to one that accommodates several different types of effect moderation gives us more confidence that the results are not distorted by model misspecification.

4.4 Analyzing a Dichotomized Mediator

In this example, we present a parallel analysis of the immigration data in which the *mvar* is coded as a binary variable for illustrative purposes. Specifically, *mvar* is coded 1 if a respondent's anxiety about increased immigration is in the top quintile of the sample distribution, indicating he or she is among the 20% most anxious, and 0 otherwise. All other variables are defined as in the prior examples.

```
. egen p80_emo = pctile( emo ), p(80)

. gen emo_bin = cond(emo>=p80_emo,1,0)

. rwrmed immigr p_harm, avar(tone_eth) mvar(emo_bin) mreg(logit) ///

       a(0) astar(1) m(0) cvar(ppage female ppeducat ppincimp) ///

       cat(female ppeducat) boot(reps(10000) seed(1234))

(output omitted)
```



```
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         CDE |  -.4287017    .1360164    -3.15   0.002     -.695289   -.1621144
       rNDEc |  -.4142553    .1282461    -3.23   0.001     -.665613   -.1628975
       rNIEc |  -.0061876    .0203676    -0.30   0.761    -.0461074    .0337322
       rATEc |  -.4204428    .1241604    -3.39   0.001    -.6637929   -.1770928
------------------------------------------------------------------------------
CDE:  controlled direct effect at m=0
rNDEc: randomized intervention analogue of the natural direct effect at sample means of cvars
rNIEc: randomized intervention analogue of the natural indirect effect at sample means of cvars
rATEc: randomized intervention analogue of the total effect at sample means of cvars
```

The output indicates that the rATE estimate is almost identical to those reported previously. Nevertheless, it also shows that, with a dichotomized mediator, nearly the entire effect appears to operate through a direct pathway, while the indirect effect operating through respondent anxiety is negligible. The difference between the results in this example and those reported in Sections 4.2 and 4.3 may reflect the problems that arise when dichotomizing a continuous variable (Royston, Altman, and Sauerbrei 2006). Nevertheless, they are generally consistent with those based on the mediator when measured in its original metric.

5 Discussion

In this paper, we introduced the `rwrmed` package to perform mediation analysis using regression-with-residuals, a method for decomposing an overall effect of treatment into direct and indirect components when treatment-induced confounding is present. Because regression-with-residuals involves only a minor adaption of conventional methods, its computations should be familiar to most applied researchers. We therefore expect that it will find wide application in analyses of causal mediation.

The method's simplicity, however, is premised on a set of strong modeling assumptions. In particular, regression-with-residuals requires a correct linear model for the outcome and then correct generalized linear models for the mediator and each of the posttreatment confounders. If any of these models are incorrectly specified, then regression-with-residuals estimates of direct and indirect effects may be biased. Regression-with-residuals is also premised on a set of strong identification assumptions. These assumptions require that all relevant confounders of the treatment–outcome, treatment–mediator, and mediator–outcome relationships have been observed and appropriately controlled. In observational studies where treatment has not been randomly assigned, all of these assumptions must be carefully scrutinized. If any are violated, then regression-with-residuals estimates of direct and indirect effects will be biased. Moreover, even in experimental studies where treatment has been randomly assigned -- like that considered in our empirical illustrations -- it is still important to consider the possibility that the assumption of no mediator-outcome confounding may be violated. Thus, analysts should always consider implementing a formal sensitivity analysis, as outlined in Wodtke and Zhou (2020), to assess the degree to which their causal inferences are sensitive to different patterns of bias due to unobserved confounding.

6 References

Agresti, A. 2002. *Categorical Data Analysis.* Hoboken, NJ: John Wiley & Sons.

Almirall, D., T. Ten Have, and S. A. Murphy. 2010. "Structural Nested Mean Models for Assessing Time-Varying Effect Moderation." *Biometrics* 66:131-139.

Brader, T., Valentino, N. A., and E. Suhay. 2008. "What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration." *American Journal of Political Science* 52: 959–78.

Buis, M. L. 2010. Direct and indirect effects in a logit model. *Stata Journal* 10: 11-29.

Daniel, R. M., De Stavola, B. L., and S. N. Cousens. 2011. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *Stata Journal* 11: 479-517.

Dippel, C., Ferrara, A., and S. Heblich. 2020. Causal mediation analysis in instrumental-variables regressions. *Stata Journal* 20: 613-626.

Elwert, F. and C. Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40:31-53.

Emsley, R., and H. Liu. 2013. paramed: Stata module to perform causal mediation analysis using parametric regression models. Statistical Software Components S457581, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s457581.html.

Falcaro, M. and A. Pickles. 2010. riskplot: A graphical aid to investigate the effect of multiple categorical risk factors. *Stata Journal* 10: 61-68.

Hicks, R., and D. Tingley. 2011. Casual mediation analysis. *Stata Journal* 11: 605–619.

Kohler, U., Karlson, K. B., and A. Holm. 2011. Comparing coefficients of nested nonlinear probability models. *Stata Journal* 11: 420–438.

Lee, J. 2011. Pathways from education to depression. *Journal of Cross Cultural Gerontology* 26: 121–135.

Linden A. 2020. cta: Stata module for conducting Classification Tree Analysis. Statistical Software Components S458729, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458729.html

Linden, A., and K. B. Karlson. 2013. Using mediation analysis to identify causal mechanisms in disease management interventions. *Health Services and Outcomes Research Methodology* 13: 86-108.

Linden, A., and P. R. Yarnold. 2018. Identifying causal mechanisms in health care interventions using classification tree analysis. *Journal of Evaluation in Clinical Practice* 24: 353-361.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Royston, P., Altman, D. G., and W. Sauerbrei. 2006. Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine* 25:127-41.

UCLA: Statistical Consulting Group. How can I do mediation analysis with the sem command? from https://stats.idre.ucla.edu/stata/faq/how-can-i-do-mediation-analysis-with-the-sem-command/ (accessed June 24, 2020)

UCLA: Statistical Consulting Group. How can I get Monte Carlo standard errors for indirect effects? from https://stats.idre.ucla.edu/stata/faq/how-can-i-get-monte-carlo-standard-errors-for-indirect-effects/ (accessed June 24, 2020)

UCLA: Statistical Consulting Group. How can I do mediation analysis with a categorical IV in Stata? from https://stats.idre.ucla.edu/stata/faq/how-can-i-do-mediation-analysis-with-a-categorical-iv-in-stata/ (accessed June 24, 2020)

UCLA: Statistical Consulting Group. How can I analyze multiple mediators in Stata? from https://stats.idre.ucla.edu/stata/faq/how-can-i-analyze-multiple-mediators-in-stata/ (accessed June 24, 2020)

UCLA: Statistical Consulting Group. How can I perform mediation with multilevel data? (Method 1) from https://stats.idre.ucla.edu/stata/faq/how-can-i-perform-mediation-with-multilevel-data-method-1/ (accessed June 24, 2020)

UCLA: Statistical Consulting Group. How can I perform mediation with multilevel data? (Method 2) from https://stats.idre.ucla.edu/stata/faq/how-can-i-perform-mediation-with-multilevel-data-method-2/ (accessed June 24, 2020)

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

VanderWeele, T. J., S. Vansteelandt, and J. M. Robins. 2014. "Effect Decomposition in the Presence of an Exposure-induced Mediator-outcome Confounder." *Epidemiology* 25: 300-306.

Wodtke, G.T., 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* 47:212-245.

Wodtke, G. T., and X. Zhou. 2020. Effect Decomposition in the Presence of Treatment-induced Confounding: A Regression-with-Residuals Approach. *Epidemiology* 31: 369-375.

Zhou, X., and G. T. Wodtke. 2019. A Regression-with-residuals Method for Estimating Controlled Direct Effects. *Political Analysis* 27: 360-369.

7 Acknowledgement

We thank Enrique Pinzon of StataCorp for his helpful suggestion to use margins with the vce(unconditional) option for estimating analytic standard errors. We also thank Xiang Zhou for reviewing a draft of the manuscript. Finally, we thank the anonymous reviewer and chief editor for their thoughtful reviews and recommendations for improving the article.

About the authors

Ariel Linden is a health services researcher specializing in the evaluation of health care interventions and policy changes. He is both an independent consultant and a research scientist in the Department of Medicine, at the University of California, San Francisco. Thus far he has written 45 community-contributed packages for Stata.

Chuck Huber is the Associate Director of Statistical Outreach, at StataCorp, LLC

Geoffrey T. Wodtke is an associate professor in the Department of Sociology at the University of Chicago. His research is in the areas of neighborhood effects and urban poverty, group conflict and racial attitudes, class structure and income inequality, and methods of causal inference in observational research. He is currently working on several projects dealing with the impact of neighborhood poverty on child development, the link between privately held business assets and growing income inequality, and new methods for estimating causal effects in longitudinal studies.