On Estimating Causal Mediation Effects from a Single Regression Equation

By

Christina Tripp Saunders

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

May 11, 2018

Nashville, Tennessee

Approved:

Jonathan S. Schildcrout, Ph.D.

Jeffrey D. Blume, Ph.D.

Robert E. Johnson, Ph.D.

Melinda C. Aldrich, Ph.D.

To my inspirational parents, Maria and Robert
and to my loving husband, Scott

ACKNOWLEDGEMENTS

I am humbled by and grateful for the outpouring of support I received from so many people. Without you, this dissertation would exist with probability zero.

To my advisor, Dr. Jeffrey Blume, for being a mentor during my time as a graduate student. You fostered my professional growth by encouraging me to think critically, to question the status quo, and to believe in the value of my own ideas. Thank you for providing timely pep talks and for sharing your wisdom with enthusiasm.

To my dissertation committee, Dr. Jonathan Schildcrout, Dr. Robert Johnson, and Dr. Melinda Aldrich, for your constructive feedback throughout the writing of my dissertation. Thank you for giving me your valuable time and thoughtful insights.

To Dr. Edward Siew, for your financial support of my research assistantship in the Vanderbilt Center for Kidney Disease. Our collaboration was an enriching experience that sharpened my skills as an applied biostatistician. Thank you for your kindness.

To my brilliant cohort, David, Mark, Derek, and Minchun, for all the times you helped me understand a difficult concept, homework problem, programming question, and for lifting me up when I felt discouraged. I am proud to be among your company.

To my undergraduate mathematics and statistics professors, Dr. Johnny Henderson, Dr. Matthew Beauregard, Dr. Jeanne Hill, and Dr. Edward Burger, for your enlightening teaching and for encouraging me to apply to a doctoral program in biostatistics.

To mom and dad, the catalyst of my high hopes. Thank you for teaching me to live with integrity, nurturing my love of learning, and emboldening me to face challenges and think independently. I am forever grateful for your guidance and unconditional love.

To my husband, Scott, for your abundant patience, your endless supply of hugs, and your steadfast confidence in me. You remind me that perseverance means putting one foot in front of the other even if I can't see my destination. Thank you for walking this road with me and reassuring me that there is always sunshine above the clouds.

To my sister, Alex, for empathizing with me over school-related stresses and understanding the comfort a cup of coffee, a square of dark chocolate, and a snuggle from Lily can bring.

To my biostatistics big sister, Sarah, for helping me debug code and making me feel normal when I felt like a fish out of water.

To God, my refuge and my strength, for giving me the ability to complete this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Mediation analysis estimates how much of an exposure's effect on an outcome is transmitted through variables along the causal pathway, which can be biological, psychological, behavioral, or social constructs. Psychologists and social scientists concerned with dynamic relations and causal mechanisms have been studying mediation processes for decades (Woodworth 1928; Alwin and Hauser 1975; Baron and Kenny 1986). Modern scientific investigations, such as genetic pathway analysis and disease prevention research, have led to widespread use of mediation analysis and the need for methodological advancement. A variety of methods exist for conducting mediation analyses; the Baron-Kenny causal steps approach (Baron and Kenny 1986; Zhao et al. 2010), the structural equation modeling approach (Gunzler et al. 2013), and the potential outcomes approach (Robins and Greenland 1992; Pearl 2001; Imai, Keele and Tingley 2010; VanderWeele 2015) are well-known frameworks.

In Chapter 2, we begin by proposing a classical regression framework for mediation analysis with linear models that allows us to estimate the portion eliminated from the fit of a single well-specified regression equation, rather than from the fit of several equations. We introduce the essential mediation components (EMCs), a general form for the difference in the exposure pathway coefficients between the marginal and full outcome models. Using multivariate normal theory, we derive a single-model formula for the EMCs. The portion eliminated, which is the difference between the total effect and the controlled direct effect, is a function of the EMCs and can be obtained without fitting any additional models. A closed-form expression for the model-based variance of the portion eliminated follows directly.

The portion eliminated can be used to evaluate the reduction in the total effect when indirect paths are blocked and is important for health policy research (Pearl 2012*a*; VanderWeele 2013; Naimi et al. 2014; VanderWeele 2015). For example, when studying how an intervention can prevent adverse health outcomes, the portion eliminated measures the maximum preventive effect of said therapy on the mediating pathways. If there is evidence of a strong portion eliminated, one could adapt policies to intervene on the mediator to limit harmful exposure effects. One can use our simple formula to estimate mediation effects from the fit of only the full outcome model, rather than having to fit a system of equations and aggregate coefficient estimates. Our formula for the portion eliminated incorporates an exposure-mediator

relationship that is as flexible as the full model exposure effect (without having to fit a separate mediator model). Furthermore, this approach extends to settings with multiple mediators, interactions, and nonlinearities, and advanced regression tools are more easily applied to a single model than to a system of equations. In a series of examples using data from the BRAIN-ICU study (Pandharipande et al. 2013), we illustrate our method and compare it to existing regression-based approaches.

Chapter 3 extends our single-model approach to non-nested mediation systems. We present a way to visualize mediation effects and propose a measure of joint mediation. We highlight situations in which using the difference and product of coefficients approaches do not yield the same estimate of the total exposure effect, which is surprising in the linear model context. This finding suggests that discrepancies between these two approaches' estimates of mediation effects depends on the specification of the marginal model and the estimation of the total effect. We conclude this chapter with extensive examples to illustrate how the proposed approach can be used to address complex behavioral research questions.

Chapter 4 considers the new framework's extension to generalized linear models and discusses its implications in the context of existing methods. We show how mediation effects can be defined on the outcome variable scale (in terms of changes in expected values) or on the link function scale (in terms of changes in a transformed space). Using a large-scale example from genetic epidemiology, we investigate whether smoking mediates the effects of genetic variants on risk of lung cancer, and we compare the results obtained using our formula to existing methods. The single-model framework imparts substantial gains in computational efficiency and meaningful insight into the formation and evaluation of complex mediation hypotheses. This chapter concludes with our ideas for future research.

CHAPTER 2

A CLASSICAL REGRESSION FRAMEWORK FOR CONDUCTING MEDIATION ANALYSES: ESTIMATING THE ESSENTIAL MEDIATION COMPONENTS FROM A SINGLE REGRESSION MODEL

2.1 Introduction

Mediators are behavioral, biological, psychological, or social constructs that transmit the effect of one variable to another. Mediation analysis seeks to understand how much of an exposure's effect on an outcome is diverted through a mediating variable (Woodworth 1928; Alwin and Hauser 1975; Baron and Kenny 1986). Background information on mediation analysis can be found in Baron and Kenny 1986; MacKinnon 2008; Hayes 2013; Preacher 2015; VanderWeele 2015. Modern scientific investigations, such as genetic pathway analysis and disease prevention research, require a sophisticated framework for conducting mediation analysis.

The literature on mediation analysis is largely comprised of the approach popularized by Baron and Kenny (1986), the causal inference framework (Robins and Greenland 1992; Pearl 2001; Imai, Keele and Tingley 2010; VanderWeele 2015), and the structural equation modeling approach (Gunzler et al. 2013). Having been cited over 70,000 times (Google Scholar), the Baron-Kenny causal steps approach is ubiquitous in the social sciences and considered to be a cornerstone of mediation analysis. However, a growing technical literature has pointed out its inability to handle complex mediation hypotheses (MacKinnon et al. 2002; Fritz and MacKinnon 2007; Preacher and Hayes 2008*b*; Zhao et al. 2010; Hayes 2013).

Here we propose a classical regression framework for conducting mediation analysis with linear models. We introduce the essential mediation components (EMCs), a general form for the difference in the exposure pathway coefficients. A formula for the EMCs and their model-based variance are derived from the fit of a single well-specified regression model. For the simple mediation model, the indirect effect for a unit change in the exposure is mathematically equivalent to the EMC; in general, however, causal mediation estimands (e.g., portion eliminated, natural indirect effect) and their variance are *functions* of the EMCs, a critical distinction. A closed-form expression for the variance is a welcome advance of the framework, eliminating the need for delta method or resampling approximations.

This approach extends to settings with multiple mediators, interactions, and nonlinearities. Fitting a single model allows for a clean application of regression tools

(e.g., imputation of missing data, cross-validation, and penalized likelihood methods) that are not easily implemented in a system of three equations. In a series of examples using data from the BRAIN-ICU study (Pandharipande et al. 2013), we illustrate our method and compare it to existing regression-based approaches. This chapter focuses on the setting of continuous outcomes (i.e. linear models).

2.2 Background and notation

2.2.1 The simple mediation model

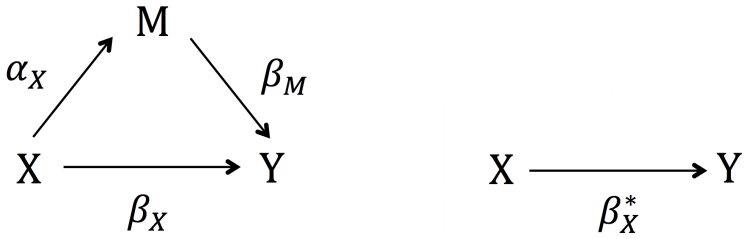


Figure 2.1: Simple mediation model for exposure X, continuous mediator M, and continuous outcome Y. The coefficients $\alpha_X, \beta_X, \beta_M$, and β_X^* are estimated from the system of three regression equations.

Mediation analyses generally seek to partition the *total effect* of an exposure into its *direct* and *indirect* components. For exposure X, continuous mediator M, and continuous outcome Y, the classic Baron-Kenny *simple mediation model* is illustrated in Figure 2.1 and represented by the following three regression equations. Errors are assumed to be normally distributed.

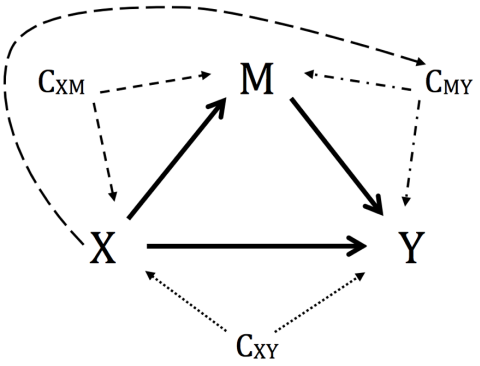
$$\mathrm{E}[Y|X, M] = \beta_0 + \beta_X X + \beta_M M \tag{2.1}$$

$$\mathrm{E}[M|X] = \alpha_0 + \alpha_X X \tag{2.2}$$

$$\mathrm{E}[Y|X] = \beta_0^* + \beta_X^* X \tag{2.3}$$

The estimated total and direct effects for a unit change in X are $\hat{\beta}_X^*$ and $\hat{\beta}_X$, respectively. The indirect effect of X is commonly estimated using the difference of coefficients, $\hat{\beta}_X^* - \hat{\beta}_X$, or the product of coefficients, $\hat{\alpha}_X \hat{\beta}_M$. For the simple mediation model, the two approaches agree and the total effect of X on Y is the sum of the direct and indirect effects: $\hat{\beta}_X^* = \hat{\beta}_X + \hat{\alpha}_X \hat{\beta}_M$. To infer causality, one must assume the relevant confounders (enumerated in Figure 2.2) have been accounted for (VanderWeele 2015). Although the original Baron-Kenny model did not include confounders, we emphasize the importance of adjusting for confounders of the type listed in Figure

2.2 so that mediation effects are identifiable. One can simply add the set of relevant confounders to each model. Furthermore, the simple mediation model above assumes linear relationships and no interaction among the variables, an assumption that can be relaxed. As in any scientific investigation, the assumed causal relationships in a mediation model rely on theory and empirical evidence.



No unmeasured confounding assumptions:

i) No unmeasured X→Y confounding (i.e., adjusted for C_{XY})

ii) No unmeasured M→Y confounding (i.e., adjusted for C_{MY})

iii) No unmeasured X→M confounding (i.e., adjusted for C_{XM})

iv) No M→Y confounders caused by X

Figure 2.2: Assumptions required for estimating causal mediation effects from a model with exposure X, mediator M, and outcome Y. C_{XY} represents confounding variables of the $X \to Y$ relationship, C_{MY} represents confounders of the $M \to Y$ relationship, and C_{XM} represents confounders of the $X \to M$ relationship. The researcher assumes that he has controlled for C_{XY}, C_{MY}, C_{XM} and that there are no $M \to Y$ confounders caused by X. Identifying controlled direct effects requires that assumptions i) and ii) be met. Identifying natural direct and indirect effects requires that all four assumptions be met. Figure adapted from VanderWeele (2015).

2.2.2 The difference of coefficients approach

There is disagreement as to whether the difference or product of coefficients approach is preferable (Alwin and Hauser 1975; Preacher and Hayes 2008*b*; Imai, Keele and Tingley 2010). Although the two approaches agree for linear models, in general they "represent legitimate intuitions in pursuit of two distinct causal quantities" and are not equivalent (Pearl 2012*b*). In Section 2.3.1, we provide a general formula for the difference of coefficients approach (i.e., the portion eliminated), which seeks to evaluate the reduction in the total effect if indirect paths were blocked. This approach is recognized as being of great importance to public health policy research (Pearl 2012*a*; VanderWeele 2013; Naimi et al. 2014; VanderWeele 2015). For example, when studying how an intervention can prevent adverse health outcomes, the difference of coefficients measures the maximum preventive effect of any such intervention on the mediating pathways (Pearl 2012*a*; VanderWeele 2015).

2.2.3 The causal inference framework for mediation analysis

The causal inference framework for mediation analysis defines mediation effects as contrasts in average potential outcomes (Holland 1986; Robins and Greenland 1992; Pearl 2001). Let $Y(x, m)$ be the potential outcome that would be observed if the exposure X were equal to x and the mediator M were equal to m. Let $Y(x, M(x_\circ))$ be the potential outcome that would be observed if the exposure were equal to x but the mediator M were equal to the value it *would have been if the exposure were equal to* $x_\circ$. Note that the counterfactuals $Y(x, M(x_\circ))$ and $Y(x_\circ, M(x))$ can never be observed. Table 2.1 provides the counterfactual definitions of causal mediation effects. We use the $(x, x_\circ)$ notation to make explicit that causal mediation effects are defined for any two levels of the exposure. When X is a binary variable, the only possible pair of values is $(0, 1)$.

The causal mediation literature distinguishes between *controlled* and *natural* effects. The controlled direct effect (CDE) measures the effect of X on Y while holding the mediator fixed at level m for everyone in the population. The natural direct effect (NDE) measures the effect of the exposure on the outcome when each individual's mediator is fixed to $M(x_\circ)$, what it would have been "naturally" had the exposure been absent (or equal to some referent value). The natural indirect effect (NIE) represents the difference in the outcome if one holds the exposure at level x and changes the mediator from the value that would have been observed under the referent exposure, $M(x_\circ)$, to the value that would have been observed under treatment, $M(x)$. The natural indirect effect is the difference between the total and natural direct effects: NIE = TE − NDE. Another important quantity is the portion eliminated (PE), which is the difference between the total and controlled direct effects: PE = TE − CDE (VanderWeele 2015).

Pearl's *mediation formula* (see Appendix 2.9) is a generalization of the product of coefficients approach and can be used to estimate causal mediation effects from any type of model (Pearl 2001; Imai, Keele and Tingley 2010; Pearl 2012*a*). For illustrative purposes, Table 2.1 displays the regression estimand obtained from applying the mediation formula to the simple mediation model: TE $= \beta^*_X(x - x_\circ)$, NDE $= \beta_X(x - x_\circ) =$ CDE, and NIE $= \alpha_X\beta_M(x - x_\circ) =$ PE. In this simple case, the estimated effects are identical to those obtained in the Baron-Kenny approach for a unit change in the exposure.

Causal Mediation Effects		Simple Mediation Model
Causal Effect	Potential Outcome	Regression Estimand
$\text{CDE}(x, x_\circ, m)$	$\text{E}[Y(x, m) - Y(x_\circ, m)]$	$\beta_X(x - x_\circ)$
$\text{NDE}(x, x_\circ)$	$\text{E}[Y(x, M(x_\circ)) - Y(x_\circ, M(x_\circ))]$	$\beta_X(x - x_\circ)$
$\text{TE}(x, x_\circ)$	$\text{E}[Y(x) - Y(x_\circ)]$	$\beta^*_X(x - x_\circ)$
$\text{NIE}(x, x_\circ)$	$\text{E}[Y(x, M(x)) - Y(x, M(x_\circ))]$	$(\beta^*_X - \beta_X)(x - x_\circ)$
$\text{PE}(x, x_\circ)$	$\text{E}[Y(x) - Y(x_\circ) - (Y(x, m) - Y(x_\circ, m))]$	$(\beta^*_X - \beta_X)(x - x_\circ)$

Table 2.1: Counterfactual definitions of causal mediation effects and the corresponding regression-based estimands obtained from the simple mediation model.

2.2.4 Estimating the variance of mediation effects

Currently, estimating the variance of the indirect effect relies on approximations; a closed-form solution has not been discovered until now. Sobel's (1982) delta method approximation for the variance of the product of coefficients is $\widehat{\text{Var}}(\hat{\alpha}_X\hat{\beta}_M) = \hat{\alpha}^2_X s^2_{\beta_M} + \hat{\beta}^2_M s^2_{\alpha_X}$. Even though the sampling distribution of $\hat{\alpha}_X\hat{\beta}_M$ tends to be skewed and highly leptokurtic, inference procedures rely on a large sample normal approximation; as a result, Sobel confidence intervals tend to lie to the left of the true value for positive indirect effects and to the right for negative indirect effects (Stone and Sobel 1990; MacKinnon et al. 1995; MacKinnon et al. 2004). VanderWeele has derived delta method variance approximations for more complex mediation models (2015). Bootstrapping handles asymmetric sampling distributions better than the delta method and thus improves the accuracy of confidence limits (Preacher and Hayes 2008*b*). Monte Carlo methods estimate the variance by simulating the sampling distribution of mediation effects (MacKinnon et al. 2004) and are implemented in the software by Imai, Keele and Tingley (2010). Now that an analytical solution for the variance exists, it is of interest to re-examine the behavior of these approximations. Simulations in Section 2.5.2 shed light on these considerations and the efficiency gains inherent in avoiding conservative approximations.

2.3 A classical regression framework

We define an intermediate inferential target called the essential mediation components (EMCs), which is the vector of changes in the exposure coefficients. Analytical estimates of the EMCs and their model-based variance are derived from the fit of a single regression model. Inference for causal mediation effects, which are functions of the EMCs, follows naturally. Furthermore, because the fit of only one model is required, it is straightforward to incorporate multiple mediators, exposure-exposure

interactions, and mediator-mediator interactions.

2.3.1 The essential mediation components

Recall that the simple mediation model (2.1-2.3) assumes X is linearly related to Y. A more general formulation allows the effect of X to be nonlinear: $\mathrm{E}[Y|X,M] = \beta_0 + \beta_X h(X) + \beta_M M$, where $h(X)$ is a flexible function of X (e.g., $\log(X)$). For p exposures $\boldsymbol{X}$ and j mediators $\boldsymbol{M}$, the full model and its implied reduced model are

$$\mathrm{E}[Y|\boldsymbol{X},\boldsymbol{M}] = \beta_0 + \boldsymbol{h(X)\beta_X} + \boldsymbol{M\beta_M} \tag{2.4}$$

$$\mathrm{E}[Y|\boldsymbol{X}] = \beta_0^* + \boldsymbol{h(X)\beta_X^*} \tag{2.5}$$

where $\boldsymbol{h}(\boldsymbol{X})$ is a vector that captures the non-linear trends in $\boldsymbol{X}$, such as a spline basis. We call the vector of differences $\Delta = \boldsymbol{\beta_X^* - \beta_X}$ the essential mediation components.

Using properties of the multivariate Gaussian distribution, we obtain estimates of the EMCs and their variance using functionals from the fitted full model (2.4). Under well-known conditions on the linear model, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathrm{MVN}_k(0, \boldsymbol{\Sigma})$, where $k = p + j$ is the number of parameters in the full model. Without loss of generality, we consider a model with no intercept. Partition $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{\boldsymbol{X}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{M}})^{\mathrm{T}}$, where $\hat{\boldsymbol{\beta}}_{\boldsymbol{X}}$ is the p-vector of exposure coefficients and $\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}$ is the j-vector of mediator coefficients such that

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{X}} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{M}} \end{bmatrix} \sim \mathrm{MVN}_k \left(\begin{bmatrix} \boldsymbol{\beta}_{\boldsymbol{X}} \\ \boldsymbol{\beta}_{\boldsymbol{M}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{V_X} & \boldsymbol{V_{XM}} \\ \boldsymbol{V_{MX}} & \boldsymbol{V_M} \end{bmatrix} \right)$$

The conditional distribution of $\hat{\boldsymbol{\beta}}_{\boldsymbol{X}}$ given $\hat{\boldsymbol{\beta}}_{\boldsymbol{M}} = \boldsymbol{b_M}$ is $(\hat{\boldsymbol{\beta}}_{\boldsymbol{X}} | \hat{\boldsymbol{\beta}}_{\boldsymbol{M}} = \boldsymbol{b_M}) \sim \mathrm{MVN}_p(\boldsymbol{\beta_{X|M}}, \boldsymbol{V_{X|M}})$, where $\boldsymbol{\beta_{X|M}} = \boldsymbol{\beta_X} + \boldsymbol{V_{XM}V_M^{-1}}(\boldsymbol{b_M} - \boldsymbol{\beta_M})$ and $\boldsymbol{V_{X|M}} = \boldsymbol{V_X} - \boldsymbol{V_{XM}V_M^{-1}V_{MX}}$. If $\boldsymbol{b_M} = 0$, we obtain $(\hat{\boldsymbol{\beta}}_{\boldsymbol{X}} | \hat{\boldsymbol{\beta}}_{\boldsymbol{M}} = 0) \sim \mathrm{MVN}_p(\boldsymbol{\beta_X^*}, \boldsymbol{V_X^*})$, where $\boldsymbol{\beta_X^*} = \boldsymbol{\beta_X} - \boldsymbol{V_{XM}V_M^{-1}\beta_M}$. Thus, a general formula for the difference in exposure pathway coefficients $\boldsymbol{\beta_X^* - \beta_X}$, which we call the essential mediation components, is

$$\Delta \equiv \boldsymbol{\beta_X^* - \beta_X} = -\boldsymbol{V_{XM}V_M^{-1}\beta_M} \tag{2.6}$$

This formula allows us to estimate multidimensional mediation effects ($\boldsymbol{X}$ and $\boldsymbol{M}$ can be multivariate) from a single regression model (2.4), rather than fitting separate models and aggregating effect estimates. Notice that for the simple mediation model (2.1-2.3), $\Delta = \beta_X^* - \beta_X$ is exactly equal to the portion eliminated for a unit change in X (which equals the causal natural indirect effect and the Baron-Kenny product of coefficients estimand). In general, when the exposure or mediator effects are non-

scalar, the portion eliminated is a function of Δ:

$$\text{PE}(x, x_\circ) = [\boldsymbol{h}(x) - \boldsymbol{h}(x_\circ)]\Delta \tag{2.7}$$

Table 2.2 provides a list of commonly encountered mediation models for which the controlled and natural direct effects are equivalent, and as a result the portion eliminated and the natural indirect effect are the same. For these models, the natural indirect effect can be estimated using our formula: $\text{NIE}(x, x_\circ) = [\boldsymbol{h}(x) - \boldsymbol{h}(x_\circ)]\Delta$. We will consider the case of interactions in Section 2.3.5.

2.3.2 Illustration

To illustrate the relationship between the EMCs and causal mediation effects, consider the full model with a quadratic effect of the exposure, $\boldsymbol{h}(\boldsymbol{X}) = [X, X^2]$, so that the full model is given by $\text{E}[Y|X, M] = \beta_0 + \beta_X X + \beta_{X^2} X^2 + \beta_M M$, the model for M is $\text{E}[M|X] = \alpha_0 + \alpha_X X$, and the reduced model for the total effect of X is $\text{E}[Y|X] = \beta^*_0 + \beta^*_X X + \beta^*_{X^2} X^2$. The controlled and natural direct effects both equal $\beta_X(x - x_\circ) + \beta_{X^2}(x - x_\circ)^2$, which are estimated from the full model. The EMCs $\Delta = \begin{bmatrix} \beta^*_X - \beta_X \\ \beta^*_{X^2} - \beta_{X^2} \end{bmatrix} = -\boldsymbol{V_{XM}}\boldsymbol{V_M^{-1}}\boldsymbol{\beta_M}$ and the natural indirect effect $[\boldsymbol{h}(x) - \boldsymbol{h}(x_\circ)]\Delta = [x - x_\circ, x^2 - x^2_\circ] \begin{bmatrix} \beta^*_X - \beta_X \\ \beta^*_{X^2} - \beta_{X^2} \end{bmatrix} = (\beta^*_X - \beta_X)(x - x_\circ) + (\beta^*_{X^2} - \beta_{X^2})(x^2 - x^2_\circ)$ are functionals that can be estimated from components of the full model. Notice that the causal mediation effects depend on the choice of $(x, x_\circ)$, while if X is binary this reduces to $(\beta^*_X + \beta^*_{X^2}) - (\beta_X + \beta_{X^2})$.

Now suppose the full model includes an exposure-mediator interaction so that the full model is $\text{E}[Y|X, M] = \beta_0 + \beta_X X + \beta_{X^2} X^2 + \beta_M M + \beta_{XM} XM$. The implied reduced model is $\text{E}[Y|X] = \gamma_0 + \gamma_X X + \gamma_{X^2} X^2$. The EMCs $\Delta = \begin{bmatrix} \gamma_X - \beta_X \\ \gamma_{X^2} - \beta_{X^2} \end{bmatrix}$ and the NIE is $[\boldsymbol{h}(x) - \boldsymbol{h}(x_\circ)]\Delta = (\gamma_X - \beta_X)(x - x_\circ) + (\gamma_{X^2} - \beta_{X^2})(x^2 - x^2_\circ)$. For a unit change in X, this reduces to $(\gamma_X + \gamma_{X^2}) - (\beta_X + \beta_{X^2})$. Notice that in both examples, the implied reduced model $\text{E}[Y|X]$ has the same form. As a result, the total effects $\text{TE}_1 = \beta^*_X(x - x_\circ) + \beta^*_{X^2}(x^2 - x^2_\circ)$ and $\text{TE}_2 = \gamma_X(x - x_\circ) + \gamma_{X^2}(x^2 - x^2_\circ)$ would have the same empirical estimate even though the system of equations is different.

Scenarios where PE = NIE	Causal Mediation Models*
Simple mediation model	$\mathbf{E[Y\|X,M] = \beta_0 + \beta_X X + \beta_M M}$ $\mathrm{E}[M\|X] = \alpha_0 + \alpha_X X$
Confounders	$\mathbf{E[Y\|X,M,C] = \beta_0 + \beta_X X + \beta_M M + \beta_C C}$ $\mathrm{E}[M\|X,C] = \alpha_0 + \alpha_X X + \alpha_C C$
Exposure-confounder interaction	$\mathbf{E[Y\|X,M,C] = \beta_0 + \beta_X X + \beta_M M + \beta_C C + \beta_{XC} XC}$ $\mathrm{E}[M\|X,C] = \alpha_0 + \alpha_X X + \alpha_C C + \alpha_{XC} XC$
Multiple mediators	$\mathbf{E[Y\|X,M_1,M_2] = \beta_0 + \beta_X X + \beta_{M_1} M_1 + \beta_{M_2} M_2}$ $\mathrm{E}[M_1\|X] = \alpha_{01} + \alpha_1 X$ $\mathrm{E}[M_2\|X] = \alpha_{02} + \alpha_2 X$
Multiple mediators with confounders	$\mathbf{E[Y\|X,M_1,M_2,C] = \beta_0 + \beta_X X + \beta_{M_1} M_1 + \beta_{M_2} M_2 + \beta_C C}$ $\mathrm{E}[M_1\|X,C] = \alpha_{01} + \alpha_1 X + \alpha_{C1} C$ $\mathrm{E}[M_2\|X,C] = \alpha_{02} + \alpha_2 X + \alpha_{C2} C$
Mediator-mediator interactions	$\mathbf{E[Y\|X,M_1,M_2] = \beta_0 + \beta_X X + \beta_{M_1} M_1 + \beta_{M_2} M_2 + \beta_{M_1 M_2} M_1 M_2}$ $\mathrm{E}[M_1\|X] = \alpha_{01} + \alpha_1 X$ $\mathrm{E}[M_2\|X] = \alpha_{02} + \alpha_2 X$ $\mathrm{E}[M_1 M_2\|X] = \alpha_{03} + \alpha_3 X$

* Bolded equation represents the fitted model used in the proposed framework

Table 2.2: Commonly encountered mediation models for exposure X, mediator M, outcome Y, confounders C where the portion eliminated (PE) and the natural indirect effect (NIE) are equal. The models used to estimate mediation effects in the traditional causal framework are shown in the second column. Using the proposed single model framework requires fitting only the first model listed under for each scenario, shown in bold. Unless otherwise specified, the exposure can be any type of variable and the mediator and outcome are continuous variables.

2.3.3 The conditional and the unconditional variance of the indirect effect

A closed form expression for the fully conditional variance of the EMCs follows directly as $\text{Var}(\hat{\Delta}|\boldsymbol{X}, \boldsymbol{M}) = \boldsymbol{V_{XM}V_M^{-1}V_{MX}}$. The variance of the natural indirect effect (and more generally, the portion eliminated) is trivial to obtain using $[\boldsymbol{h}(x) - \boldsymbol{h}(x_\circ)]\text{Var}(\hat{\Delta}|\boldsymbol{X}, \boldsymbol{M})[\boldsymbol{h}(x)-\boldsymbol{h}(x_\circ)]^{\text{T}}$, which requires fitting only one model (2.4). From properties of a regression model, for a scalar portion eliminated with a unit change in the exposure we have $\frac{\widehat{\text{PE}}-\text{PE}}{\sqrt{\widehat{\text{Var}(\text{PE})}}} \sim t(\text{df}{=}n-k, \text{scale} = -1)$ and the 95% confidence interval is $\widehat{\text{PE}} \pm t_{.975,n-k} \times \hat{V}_{XM}\hat{V}_M^{-1}\sqrt{\widehat{\text{Var}}(\hat{\beta}_M)}$.

In equations (2.4) and (2.5), Y is a random variable and X and M are fixed covariates. One may wish to treat the mediator as a random variable and marginalize over M. The causal inference framework uses the marginal variance for inference (VanderWeele 2015). Using the law of total probability,

$$\begin{aligned}\text{Var}(\hat{\Delta}|X) &= \text{E}_{M|X}[\text{Var}(\hat{\Delta}|X, M)] + \text{Var}_{M|X}[\text{E}(\hat{\Delta}|X, M)] \\ &= \text{E}_{M|X}\left[\frac{n^2 r_{XM}^2 \hat{\sigma}_M^2 \hat{\sigma}_{Y|X,M}^2}{|\boldsymbol{D}^{\text{T}}\boldsymbol{D}|}\right] + \beta_M^2\left[\frac{\sigma_{M|X}^2}{n\hat{\sigma}_X^2}\right] \qquad (2.8)\end{aligned}$$

where $\boldsymbol{D} = (1, X, M)$ is the $n \times 3$ design matrix. This quantity can be estimated by plugging in the sample correlation r, the maximum likelihood estimates of the variances of X and M, estimates of the mean square error of Y from (2.1) and of M from (2.2), and $\hat{\beta}_M$. Notice that the second term in (2.8) is an increasing function of β_M and a decreasing function of the sample size n. We used simulations to empirically verify (2.8) under various sample sizes ($n = 100, 200, 400, 1000$) and magnitudes of $\beta_M = 2, 4$. Although the second term in $\text{Var}(\hat{\Delta}|X)$ requires estimating the variance of the residuals from the regression of M on X, the contribution is of order $1/n$ and becomes negligible in moderate sample sizes. The marginal variance of mediation effects follows.

Because the mediator is (in theory) a consequent of the exposure, M cannot be randomized and one could argue in favor of treating both X and M as fixed (Pearl 2012*a*). In classical regression settings, the conditional variance is frequently used for inference even when the covariate changes in a population. As we note above, the distinction between the two variances becomes semantic in large samples. Which variance is to be preferred deserves consideration, but further discussion is beyond the scope of this paper. Note that the nonparametric bootstrap, which samples with replacement from pairs of X and M, approximates the fully unconditional variance (marginalized over both predictor and mediator).

2.3.4 Multiple mediators

Suppose the exposure's effect on the outcome is transmitted through several mediators. Estimating the *total indirect effect* in a multiple mediator model aims to determine if the *set* of j mediators $\boldsymbol{M}$ transmits the effect of X to Y. To identify natural direct and indirect effects from multiple mediator models, all four no unmeasured confounding assumptions outlined in Figure 2.2 must hold with respect to $\boldsymbol{M}$. Existing approaches in the context of multiple mediators include the *single-step multiple mediator model* (MacKinnon 2008), also termed the *parallel multiple mediator model* (Hayes 2013), and the *serial multiple mediator model* (Hayes 2013). The single-step approach specifies a separate outcome model for each mediator in which they independently affect the outcome (see panel C of Figure 2.3). The serial model relies on assumptions about the directionality of the mediators, which can be unverifiable with cross-sectional data (see panel B of Figure 2.3). VanderWeele and Vansteelandt (2013) provide both regression-based and weighting approaches that allow mediators to be interdependent (see panel A of Figure 2.3). The simulation-based approach by Imai, Keele and Tingley (2010) handles multiple mediator models of all types, but the software currently accommodates only two mediators and the user must specify one mediator as "main" and the other as "alternative."

Within our framework, incorporating multiple mediators is simple and efficient. Our formulation allows the mediators to covary, a more realistic assumption than assuming the mediators do not affect each other (as is required for the single-step models), or that we know the order in which they affect each other (as is required for serial models). The advantage of using our approach is that it requires fitting only one model to obtain causal mediation estimands (compared to three or more models required by existing approaches), and it yields model-based variance estimates that do not require the computation time of resampling methods.

If we posit j mediators such that the full mediation model is $\mathrm{E}[Y|X,M] = \beta_0 + \beta_X X + \Sigma_{i=1}^{j}\beta_{M_i}M_i$ and the corresponding reduced model for the total effect of X is $\mathrm{E}[Y|X] = \beta_0^* + \beta_X^* X$, then the total indirect effect through $\boldsymbol{M}$ is estimated by $[\boldsymbol{h}(x)-\boldsymbol{h}(x_\circ)]\hat{\Delta} = -\hat{\boldsymbol{V}}_{X\boldsymbol{M}}\hat{\boldsymbol{V}}_{\boldsymbol{M}}^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}(x-x_\circ)$, and its variance by $\widehat{\mathrm{Var}}([\boldsymbol{h}(x)-\boldsymbol{h}(x_\circ)]\hat{\Delta}) = (x-x_\circ)^2\hat{\boldsymbol{V}}_{X\boldsymbol{M}}\hat{\boldsymbol{V}}_{\boldsymbol{M}}^{-1}\hat{\boldsymbol{V}}_{\boldsymbol{M}X}$. The *mediator-specific indirect effect* represents the ability of M_i to mediate the effect of X on Y above and beyond the other $j-1$ mediators. The specific indirect effect through $M_{i'}$ is estimated using $-\hat{V}_{XM_{i'}}\hat{V}_{M_{i'}}^{-1}\hat{\beta}_{M_{i'}}(x-x_\circ)$. The variance is estimated with $\hat{V}_{XM_{i'}}\hat{V}_{M_{i'}}^{-1}\hat{V}_{M_{i'}X}(x-x_\circ)^2$. Importantly, formula (2.6) gives us these effects and their variances without having to fit any reduced models.

If two or more mediators share a role in transmitting the effect of X to Y, then

the effect attributed to a specific mediator M_i may exclude this overlapping effect. Additionally, specific indirect effects might have different signs, leading to inconsistent mediation. As a result, the specific indirect effects attributed to each mediator do not necessarily sum to the total indirect effect mediated by the set of mediators. We emphasize that the estimated total indirect effect through $\boldsymbol{M}$ comes from the full model containing all of the mediators and does not suffer bias from the misspecification of inter-mediator relationships. Thus, we recommend the researcher's primary interest lie in the total indirect effect, rather than the amount mediated by a specific mediator. We present relevant examples in Section 2.6.

The regression-based approach by VanderWeele and Vansteelandt (2013) uses one outcome model for all of the mediators but also requires a separate model for each mediator and each mediator-mediator interaction. Including covariates C can lead to compatibility issues between the models for M_i, M_k, and their product M_iM_k. Their alternative inverse probability weighting approach circumvents this issue in settings with mediator-mediator interactions. The weighting approach allows the mediators to affect each other and does not require modeling the mediators, but it does require fitting several logistic regression models to estimate $\mathrm{P}[X = x]$, $\mathrm{P}[X = x_\circ]$, $\mathrm{P}[X = x|C = c]$, $\mathrm{P}[X = x_\circ|C = c]$ for the weights. It should be noted that the weighting method performs best when the exposure has only a few levels (e.g., binary or discrete) (VanderWeele and Vansteelandt 2013).

2.3.5 Interactions and moderated mediation

Our framework accommodates exposure-exposure and mediator-mediator interactions, as well as interactions with confounders. Simply include the interaction terms of interest in the full model and use formulas (2.6) and (2.7) to estimate the EMCs and causal mediation effects. We now consider the more complex setting of exposure-mediator interactions (so-called *moderated mediation*).

The causal mediation literature often considers exposure-mediator interactions with binary X, such that the full model is $\mathrm{E}[Y|X, M] = \beta_0+\beta_X X+\beta_M M+\beta_{XM}XM$ and the reduced model is $\mathrm{E}[Y|X] = \beta_0^* + \beta_X^* X$. The portion eliminated $\mathrm{PE} = \mathrm{TE}-\mathrm{CDE}(m) = [\beta_X^*-(\beta_X+\beta_{XM}m)](x-x_\circ)$ is estimated from the fit of the full model using $[\Delta - \beta_{XM}m](x - x_\circ)$. One could plot the PE as a function of M or report the PE for a point of interest m, such as the sample mean. Its variance follows by direct calculation: $(x - x_\circ)^2 \left[V_{XM}V_M^{-1}V_{MX} + m^2\mathrm{Var}(\hat{\beta}_{XM}) + 2mV_{XM}V_M^{-1}\mathrm{Cov}(\hat{\beta}_M, \hat{\beta}_{XM})\right]$.

If the exposure is continuous, then the exposure-mediator interaction model above

implies a marginal model that includes an X^2 term. In this case, the marginal model exposure coefficient cannot be estimated using formula (2.6) because it is not nested within the full model. This formula requires that the full model also include an X^2 term (see Section 2.3.2 for an example), which can be viewed as relaxing the assumption that all nonlinear effects of the exposure act through the mediator. In this sense, a broader full model is desirable. Despite the non-nested reduced model, it is possible to use the mediation formula to proceed with estimation in this setting. Further thoughts are included in Remark D.

Notice that exposure-mediator interactions lead to mediation effects that are less clearly defined. Because M acts simultaneously as a moderator and a mediator, both the direct and indirect effects are affected by β_{XM}. As a result, there is more than one way to decompose the total effect depending on how the interaction effect is accounted for (Robins and Greenland 1992; Pearl 2001). If one attributes β_{XM} to the indirect effect, then the total effect decomposes into the natural direct and total indirect effects. Conversely, if one attributes β_{XM} to the direct effect, then the total effect decomposes into the total direct and pure indirect effects. Thus, mediation and moderation are "inextricably intertwined and cannot be assessed separately" (Pearl 2012*b*).

With exposure-mediator interactions, the controlled and natural direct effects diverge: $\text{CDE}(m) = (\beta_X + \beta_{XM}m)(x - x_\circ)$ and $\text{NDE} = (\beta_X + \beta_{XM}\text{E}[M|x_\circ])(x - x_\circ)$. Because the controlled direct effect is a function of m, the portion eliminated and its variance are functions of the mediator as well. The natural direct effect marginalizes over $\text{E}[M|x_\circ]$ and is a function of the exposure that can be defined for any levels of $(x, x_\circ)$ (Naimi et al. 2014). The portion eliminated does not depend on the choice of decomposition because it is the portion of the total effect attributed to *both* interaction and mediation. As such, the PE is a comprehensive estimate of moderated mediation in these complex settings.

2.4 Impact of omitted covariates on estimating the indirect effect

What happens to $\hat{\Delta}$ when we omit an important covariate in the specification of the full model? If the omitted covariate is orthogonal to X or to M, then $\hat{\Delta}$ does not incur additional bias. To fix ideas, consider the simple setting in which we have one exposure X, one mediator M, and a third omitted covariate W. Suppose the true data generating mechanism is $Y = \gamma_0 + \gamma_X X + \gamma_M M + \gamma_W W + \varepsilon$, but we don't know W so we *incorrectly* specify the full model as $Y = \beta_0 + \beta_X X + \beta_M M + \varepsilon$ and the reduced

model as $Y = \beta_0^* + \beta_X^* X + \varepsilon$. The estimates $\hat{\beta}_X$ and $\hat{\beta}_X^*$ will be biased estimates of γ_X, the "true" effect of X. The expected values of the parameters from the full model and reduced model are $\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_X \\ \hat{\beta}_M \end{bmatrix} \rightarrow E_{true} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_X \\ \hat{\beta}_M \end{bmatrix} = \begin{bmatrix} \gamma_0 \\ \gamma_X \\ \gamma_M \end{bmatrix} + \begin{bmatrix} \delta_0 \\ r_{WX.M}\sigma_W/\sigma_X \\ r_{WM.X}\sigma_W/\sigma_M \end{bmatrix} \gamma_W$ and $\begin{bmatrix} \hat{\beta}_0^* \\ \hat{\beta}_X^* \end{bmatrix} \rightarrow E_{true} \begin{bmatrix} \hat{\beta}_0^* \\ \hat{\beta}_X^* \end{bmatrix} = \begin{bmatrix} \gamma_0 \\ \gamma_X \end{bmatrix} + \begin{bmatrix} \alpha_0 \\ r_{XM}\sigma_M/\sigma_X \end{bmatrix} \gamma_M + \begin{bmatrix} \kappa_0 \\ r_{XW}\sigma_W/\sigma_X \end{bmatrix} \gamma_W$, respectively. Thus, when we omit W, the expected difference in the estimated total and direct effects for a unit change in X is $\mathrm{E}[\hat{\Delta}_1] = E_{\mathrm{true}}[\hat{\beta}_X^* - \hat{\beta}_X] = r_{XM}\frac{\sigma_M}{\sigma_X}\gamma_M + (r_{XW} - r_{XW.M})\frac{\sigma_W}{\sigma_X}\gamma_W$.

Next, suppose that Y does not depend on W and the true data generating mechanism is $Y = \gamma_0 + \gamma_X X + \gamma_M M + \varepsilon$. If we *correctly* specify the full model as $Y = \beta_0 + \beta_X X + \beta_M M + \varepsilon$ and the reduced model as $Y = \beta_0^* + \beta_X^* X + \varepsilon$, then $\mathrm{E}[\hat{\Delta}_2] = \mathrm{E}[\hat{\beta}_X^* - \hat{\beta}_X] = \gamma_X + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}M\gamma_M - \gamma_X = r_{XM}\frac{\sigma_M}{\sigma_X}\gamma_M$. The bias in the estimated indirect effect when the full model omits W is given by $\mathrm{E}[\hat{\Delta}_1 - \hat{\Delta}_2] = (r_{XW} - r_{XW.M})\frac{\sigma_W}{\sigma_X}\gamma_W$. As a result, if W is orthogonal to X or M ($r_{XW} = r_{XW.M}$) or $\gamma_W = 0$ (a trivial case), then the estimated indirect effect under the incorrectly specified full model is robust to misspecification. That is, omitting W will not change the estimate of the indirect effect.

Note that if W is *not* orthogonal to either X or M such that W is a confounder of the exposure-mediator relationship, then assumption (iii) of the no-unmeasured confounding assumptions is violated and natural direct and indirect effects are not identifiable. Controlled direct effects are still identifiable in this setting, provided there are no unmeasured confounders of the exposure-outcome and mediator-outcome relationship (see Figure 2.2).

2.5 Simulations

2.5.1 Setup

We use simulations to provide empirical support for the proposed approach to mediation analysis with a simple mediation model. We simulated 5,000 datasets of sample size $n \in \{50, 100, 200\}$ with a "true" indirect effect of 1.5. The "true" full model was $Y = \beta_0 + \beta_X X + \beta_M M + \varepsilon_Y$ where $\varepsilon \sim N(0, \sigma_Y^2)$. The exposure $X \sim N(0, \sigma_X^2)$ and mediator $M = \alpha_0 + \alpha_X X + \varepsilon_M$, where $\varepsilon_M \sim N(0, \sigma_M^2)$. When comparing methods, 10,000 bootstrap replications and 10,000 Monte Carlo draws were used. For each replication, we computed each method's estimated indirect effect and estimated variance and compared these to the true effect and the empirical ("true") variance. The bias of the estimated indirect effect was captured when X and M were

both fixed and when M was random. These simulations demonstrate the performance of our formulas and are not intended to be exhaustive. Varying parameter values impacted the magnitude of the results but not the general patterns.

2.5.2 How our variance measure compares to existing measures

The results of estimating the variance of the indirect effect using the analytical regression-based formula, bootstrapping, Sobel's formula, and Monte Carlo methods are shown in Figure 2.4 and Table 2.3 in Appendix 2.9. The analytical variance formula appears unbiased for the true variance. Sobel's variance performs similarly to the case bootstrap. As expected, the estimated variance from bootstrapping cases is greater than that from the residual bootstrap (see Section 2.3.3). As the sample size increases, the variances of the estimates of $\widehat{\text{Var}}([h(x) - h(x_\circ)]\hat{\Delta})$ from the cases bootstrap, Sobel's formula, and the MC methods decrease but remain biased.

2.5.3 The bias of indirect effect estimates depends on the conditioning set

Under the full model, the expectation of $\hat{\Delta}$ is $\text{E}[\hat{\beta}^*_X - \hat{\beta}_X] = (\beta_X + P_{X.M}\beta_M) - \beta_X = P_{X.M}\beta_M$, where $P_{X.M}$ is the projection of M onto X. For the simple mediation model, $P_{X.M}\beta_M = \rho_{XM}\left(\frac{\rho_{MY} - \rho_{YX}\rho_{XM}}{1-\rho^2_{XM}}\frac{\sigma_Y}{\sigma_X}\right)$. To estimate $\hat{\Delta}$, we replace ρ and σ^2 with their sample estimates r and s^2 to obtain $r_{XM}\left(\frac{r_{MY} - r_{YX}r_{XM}}{1-r^2_{XM}}\frac{s_Y}{s_X}\right)$, which is biased per Jensen's inequality. This is not surprising because the sample correlation r is a biased estimate of ρ, a result given by Fisher (1915): $\text{E}[r] = \rho - \rho(1-\rho^2)/2N$. Since $r \to \rho$ and $s^2 \to \sigma^2$ as $N \to \infty$, $\hat{\Delta}$ is biased but consistent for the true Δ by the Law of Large Numbers and the Continuous Mapping Theorem.

The distributions of $\hat{\Delta}$ when X and M are both fixed and when M varies are shown in Figure 2.5 in Appendix 2.9 (note that $\hat{\Delta}$ equals the indirect effect for a unit change in X from the simple mediation model). When X and M are both fixed, $\hat{\Delta}$ is biased as a function of the bias of r_{XM}. If we allow M to vary, the bias is reduced because r_{XM} is no longer fixed and it tends to approximate ρ_{XM} better on average. Therefore, because of the bias-variance tradeoff, coverage probability alone is not the proper performance measure when r_{XM} poorly approximates ρ_{XM}.

2.6 Examples with data from Vanderbilt ICU patients

We illustrate our method and existing approaches using data from a prospective cohort of 217 ICU patients at Vanderbilt University Medical Center with acute

respiratory failure and/or cardiogenic or septic shock (Pandharipande et al. 2013). The goal is to examine the cognitive effects of critical illness. We use measurements of creatinine (mg/dL) and estimated glomerular filtration rate (eGFR) measured at baseline, benzodiazepine dose (mg), Sequential Organ Failure Assessment (SOFA) score, mental status (delirious or normal) assessed with the Confusion Assessment Method for the ICU and Richmond Agitation-Sedation Scale, and Repeatable Battery for the Assessment of Neuropsychological Status (RBANS), a global cognitive score measured three months post discharge. Biomarker S100B levels were measured for 121 of these patients.

We present several simple examples of mediation models to illustrate the efficiency and coherence of our proposed framework. We compare variance estimates obtained from the model-based formula, Sobel's formula, and percentiles of 10,000 bootstrap replications. The first two examples assume there are no unmeasured confounders, and the third assumes the covariate C sufficiently controls for confounding. We do not intend for the examples and their results to be interpreted scientifically; rather, they are meant to illustrate the methods discussed throughout the paper. All models assume errors $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. Unless otherwise specified, we compare unit changes in the exposure so that $(x - x_\circ) = 1$.

Example 1 (Simple mediation model): Does severity of illness (SOFA) mediate the effect of creatinine on S100B levels? We specify the full model as $\texttt{S100B} = \beta_0 + \beta_X \texttt{Cr} + \beta_M \texttt{SOFA} + \varepsilon$ to estimate the EMC $\hat{\Delta} = -\hat{V}_{XM}\hat{V}_M^{-1}\hat{\beta}_M$, where $X = \texttt{Cr}$ and $M = \texttt{SOFA}$. The mediated effect of creatinine on S100B is $\hat{\Delta}(x - x_\circ)$=28.64 (SE=7.12) with 95% CI 14.54 to 42.74. The residual bootstrap SE=7.06, Sobel's SE=17.25, and the case bootstrap SE=18.74. Importantly, the model-based variance is five times smaller than Sobel's and the case bootstrap, which yield 95% CIs that include zero. Although the residual bootstrap variance gives essentially the same answer as the model-based formula, the formula avoids the computation time and effort.

To allow for a *quadratic* relationship between creatinine and S100B, simply specify the full model as $\texttt{S100B} = \beta_0 + \beta_{X_1}\texttt{Cr} + \beta_{X_2}\texttt{Cr}^2 + \beta_M \texttt{SOFA} + \varepsilon$. The EMCs $\hat{\Delta}$= $-\hat{V}_{\boldsymbol{X}M}\hat{V}_M^{-1}\hat{\beta}_M = [28.33, -11.08]^{\mathrm{T}}$ is now a vector of the linear and quadratic effects of creatinine. The portion eliminated (which equals the NIE) is $28.33(x - x_\circ) - 11.08(x^2 - x_\circ^2) = 17.25$. Using the mediation package gives an estimated NIE of 17.25 (exactly equal to our estimate, as expected) and requires 111.94 seconds of computation time compared to 0.01 seconds using our approach. In the remaining examples we consider only linear effects, but allowing for nonlinear relationships in practice is strongly advised and easily implemented within the proposed framework.

Example 2 (Simple mediation model where Sobel's approximation holds): It is not always true that we see such large efficiency gains. For instance, our method yields similar results to standard approaches when we investigate whether the relationship between creatinine and overall cognitive function (RBANS) is mediated by severity of illness. The estimated indirect effect is 0.01 (SE=0.27), the residual bootstrap SE=0.27, Sobel's SE=0.27, and the case bootstrap SE=0.29.

Example 3 (Exposure-confounder interactions): Recall that identifying mediation effects relies on a strict set of no unmeasured confounding assumptions (outlined in Figure 2.2). To keep this example simple, we assume that adjusting for C = Charlson score is sufficient to satisfy these assumptions. We also include an exposure-confounder interaction so that the full model is $\texttt{RBANS} = \beta_0 + \beta_X \texttt{Cr} + \beta_M \texttt{SOFA} + \beta_C \texttt{Charlson} + \beta_{XC} \texttt{Cr:Charlson} + \varepsilon$. The indirect effect marginalized over the confounder is $\mathrm{E}[h(x) - h(x_\circ)]\Delta|C] = (\beta_X^* - \beta_X)(x - x_\circ) + (\beta_{XC}^* - \beta_{XC})(x - x_\circ)\mathrm{E}[C]$. The variance is estimated using $(x - x_\circ)^2\mathrm{Var}(\Delta_1) + (x - x_\circ)^2\mathrm{E}[C]^2\mathrm{Var}(\Delta_2) + 2(x - x_\circ)^2\mathrm{E}[C]\mathrm{Cov}(\Delta_1, \Delta_2)$.

For a unit change in creatinine, the estimated indirect effect is 0.0028 (SE=0.22). The regression-based approach by VanderWeele requires fitting the mediator model $\texttt{SOFA} = \alpha_0 + \alpha_X \texttt{Cr} + \alpha_C \texttt{Charlson} + \beta_{XC} \texttt{Cr:Charlson} + \varepsilon$ in addition to the full model. The mediation formula estimates the indirect effect using $\mathrm{E}[\beta_M(\mathrm{E}[M|x] - \mathrm{E}[M|x_\circ])|C] = \beta_M(\alpha_X + \alpha_{XC}\mathrm{E}[C])(x - x_\circ) = 0.0028$ (SE=0.24).

To examine the indirect effect comparing the 75th percentile to the median value of Cr, one simply plugs in these values for x and $x_\circ$. Using the single-model approach took 0.02 seconds to estimate the total, direct, and indirect effects, and an additional 0.003 seconds to recalculate the indirect effect for the new pair of exposure values. Using the simulation-based mediation package required 24.89 seconds, and a new simulation must be run for each additional pair of exposure values. Although this difference may seem inconsequential for this simple example, using (2.6) and (2.7) reduces the computation time by several orders of magnitude when applied to big data. For example, with the current sample size (N=217), if one were to study mediation across 10,000 SNPs and three different pairs of the exposure were of interest, the simulation-based approach would require around $10,000 \times 24.89 \times 3$ seconds (over 8 days) to run. The proposed approach would take $(10,000 \times .02) + (0.003 \times 3)$ seconds (under 4 minutes).

Example 4 (Multiple mediator model): Is the effect of creatinine on cognitive function mediated by severity of illness and benzodiazepine dose? The conceptual diagrams in Figure 2.3 depict the single-model approach for multiple mediators, the

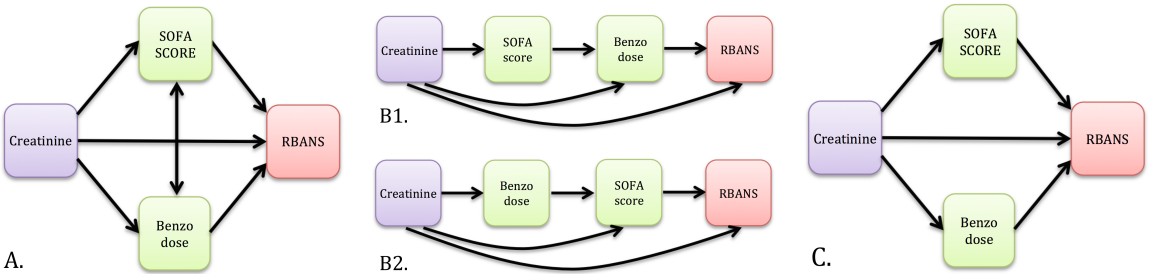


Figure 2.3: Comparison of multiple mediator models. Model A depicts the proposed single-model framework for assessing mediation with multiple mediators. Model B depicts the serial multiple mediator model. Model C depicts the single-step or parallel multiple mediator model. The directions of arrows indicate the assumed causal pathways.

serial multiple mediator model, and the parallel multiple mediator model. For all three methods, the full model is $\texttt{RBANS} = \beta_0 + \beta_X \texttt{Cr} + \beta_{M_1} \texttt{SOFA} + \beta_{M_2} \texttt{Benz} + \varepsilon$ and the reduced model for the total effect of creatinine on cognitive function is $\texttt{RBANS} = \beta_0^* + \beta_X^* \texttt{Cr} + \varepsilon$. Thus, the direct effect of creatinine is $\beta_X(x - x_\circ)$, the total effect of creatinine is $\beta_X^*(x - x_\circ)$, and the total indirect effect of creatinine through SOFA and benzodiazepine is $(\beta_X^* - \beta_X)(x - x_\circ)$, regardless of which method you use. It is important to recognize that the total indirect effect does not depend on the order or directionality of the mediators, whereas the amount of mediation attributed *specifically* to SOFA or benzodiazepine will differ across methods due to their varying assumptions about inter-mediator relationships.

We estimate how much severity of illness and benzodiazepine dose mediate the relationship between creatinine and cognitive function using only the full model and formula (2.7). For $X = \texttt{Cr}$ and $\boldsymbol{M} = \{\texttt{SOFA}, \texttt{Benz}\}$, the total indirect effect through $\boldsymbol{M}$ is estimated using $[h(x) - h(x_\circ)]\hat{\Delta} = -\hat{V}_{X\boldsymbol{M}}\hat{V}_{\boldsymbol{M}}^{-1}\hat{\beta}_{\boldsymbol{M}}(x - x_\circ) =$ -0.34 and its empirical variance $\hat{V}_{X\boldsymbol{M}}\hat{V}_{\boldsymbol{M}}^{-1}\hat{V}_{\boldsymbol{M}X}(x - x_\circ)^2 = 0.139$ (SE = 0.37).

Now suppose we are interested in mediator-specific effects. Since we have already fit the full model, to estimate how much is mediated specifically through $M_1 = \texttt{SOFA}$ we simply apply (2.6): $-\hat{V}_{XM_1}\hat{V}_{M_1}^{-1}\hat{\beta}_{M_1}(x - x_\circ) =$ -0.046. The variance follows directly: $\hat{V}_{XM_1}\hat{V}_{M_1}^{-1}\hat{V}_{M_1X}(x - x_\circ)^2 = 0.092$ (SE=0.30). Similarly, to estimate how much the effect of creatinine is mediated through $M_2 = \texttt{Benz}$, use $-\hat{V}_{XM_2}\hat{V}_{M_2}^{-1}\hat{\beta}_{M_2}(x - x_\circ) =$ -0.352 which has an estimated variance of $\hat{V}_{XM_2}\hat{V}_{M_2}^{-1}\hat{V}_{M_2X}(x - x_\circ)^2 = 0.066$ (SE = 0.26). Keeping in mind SOFA is correlated with benzodiazepine dose, notice that the mediator-specific indirect effects sum to -0.398, which does not equal the total indirect effect of -0.34.

The parallel approach (MacKinnon 2008; Hayes 2013) and the causal regression-

based approach (VanderWeele and Vansteelandt 2013) specify the same full model as above $\texttt{RBANS} = \beta_0 + \beta_X\texttt{Cr} + \beta_{M_1}\texttt{SOFA} + \beta_{M_2}\texttt{Benz} + \varepsilon$ and an additional model for each mediator: $\texttt{SOFA} = \alpha_{01} + \alpha_{X1}\texttt{Cr} + \varepsilon$ and $\texttt{Benz} = \alpha_{02} + \alpha_{X2}\texttt{Cr} + \varepsilon$. This specification assumes the mediators "act in parallel" (see Figure 2.3 panel C). The estimated indirect effect through SOFA is $\hat{\alpha}_{X1}\hat{\beta}_{M_1}$ = -0.042 (SE=0.27) and the estimated indirect effect through benzodiazepine is $\hat{\alpha}_{X2}\hat{\beta}_{M_2}$ = -0.299 (SE=0.25), which sum to the total indirect effect. Delta method approximations are used to estimate standard errors. When the exposure is continuous and there are no mediator-mediator interactions, the weighting approach is not recommended (VanderWeele and Vansteelandt 2013).

The serial model given by Hayes (2013) requires specifying the order in which the mediators affect each other. Suppose we assume $\texttt{Cr} \rightarrow \texttt{SOFA} \rightarrow \texttt{Benz} \rightarrow \texttt{RBANS}$ (see Figure 2.3 panel B1). The full model is specified as $\texttt{RBANS} = \beta_0 + \beta_X\texttt{Cr} + \beta_1\texttt{SOFA} + \beta_2\texttt{Benz} + \varepsilon$ (the same as above), the first reduced model is $\texttt{Benz} = \alpha_{02} + \alpha_2\texttt{Cr} + \delta_{21}\texttt{SOFA} + \varepsilon$, and the second reduced model is $\texttt{SOFA} = \alpha_{01} + \alpha_1\texttt{Cr} + \varepsilon$. There are three estimated indirect effects: $\hat{\alpha}_1\hat{\beta}_1$ = -0.04 (SE=0.27) is the indirect effect of creatinine through SOFA to RBANS, $\hat{\alpha}_2\hat{\beta}_2$ = -0.35 (SE=0.29) is the indirect effect of creatinine through benzodiazepine to RBANS, and $\hat{\alpha}_1\hat{\delta}_{21}\hat{\beta}_2$ = 0.05 (SE=0.36) is the indirect effect of creatinine through SOFA to benzodiazepine to RBANS. The variances of $\hat{\alpha}_1\hat{\beta}_1$ and $\hat{\alpha}_2\hat{\beta}_2$ are estimated using Sobel's formula and $\widehat{\text{Var}}(\hat{\alpha}_1\hat{\delta}_{21}\hat{\beta}_2) = \hat{\alpha}_1^2\hat{\delta}_{21}^2 s^2_{\beta_2} + \hat{\alpha}_1^2\hat{\beta}_2^2 s^2_{\delta_{21}} + \hat{\delta}_{21}^2\hat{\beta}_2^2 s^2_{\alpha_1}$ (Hayes 2013).

To demonstrate how mediator-specific indirect effects depend on the specified order in a serial model, suppose we change the order of mediation to $\texttt{Cr} \rightarrow \texttt{Benz} \rightarrow \texttt{SOFA} \rightarrow \texttt{RBANS}$ (see Figure 2.3 panel B2). The total indirect effect remains unchanged, but now the indirect effect of creatinine through benzodiazepine to RBANS is -0.299, the indirect effect of creatinine through SOFA to RBANS is -0.046, and the indirect effect of creatinine through benzodiazepine to SOFA to RBANS is 0.0046. Notice that in either case, the serially mediated indirect effects sum to the total indirect effect. Estimating indirect effects from the serial model is analogous to examining sequential sums of squares, whereas estimating effects from the proposed framework is analogous to examining partial sums of squares. Just as partial sums of squares do not necessarily sum to the total, mediator-specific indirect effects do not necessarily sum to the total indirect effect. In contrast, the serial indirect effects do sum to the total indirect effect, but their estimation depends heavily on the assumed order of the mediators.

2.7 Remarks

The statistical literature abounds with methods for estimating the indirect effect and its variance from the simple mediation model. For sophisticated mediation analyses involving interactions, splines, and any combination of continuous, binary, and categorical mediators, the proposed single-model approach is straightforward to implement.

Remark A: Straightforward application of modeling tools

The proposed framework can be viewed as having two key steps: first, estimation of a single fully-conditional model for the outcome and second, estimation of mediation functionals from that model. As a result, this approach allows for mediation analysis with a straightforward application of regression modeling tools - e.g., penalization procedures such as the elastic net or lasso, multiple imputation, and cross-validation. One simply applies these techniques to the single well-specified full model and their impact is automatically incorporated in the mediation functionals.

Remark B: Advantage of using one outcome model in multiple mediator settings

As pointed out by VanderWeele and Vansteelandt (2013), the approach of using one outcome model for all of the mediators is "robust to unmeasured common causes $[C]$ of two or more mediators," whereas having separate outcome models for each mediator is not. When the outcome model contains all the mediators, C only affects the outcome through the set of mediators, so C does not confound the joint effect of $\boldsymbol{M}$ on Y. If, instead, one specifies a separate outcome model for each mediator, C affects M_i and it affects Y through $\boldsymbol{M}_{i' \neq i}$, which leads to biased estimates of the the effect M_i on Y. Thus, it is recommended to specify one full outcome model that contains all of the mediators.

Remark C: Controlled indirect effect

"Controlled indirect effects are notably difficult to conceptualize, and instead are defined as some contrast between the total and controlled direct effects in the absence of exposure-mediator interactions" (Naimi et al. 2014). Our approach provides a general formula for estimating the difference between the total and controlled direct effect, i.e., the so-called controlled indirect effect. By contrast, the mediation formula provides a general formula for estimating the natural indirect effect, the difference between the total and natural direct effect (Pearl 2001).

Remark D: Non-nested reduced models

In order to use the formula for the EMCs, the implied reduced model must be nested within the full model. The simple exposure-mediator interaction model is a commonly encountered example of a marginal model that is not nested. The full

model $\mathrm{E}[Y|X,M] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM$ has the implied reduced model $\mathrm{E}[Y|X] = \beta_0^* + \beta_X^* X + \beta_{X^2}^* X^2$. This full model contains a linear term for X, which implies the entire non-linear effect of the exposure is captured by the mediator (via the interaction). This is an impactful assumption that we would prefer to relax by including the nonlinear exposure effects $h(X)$ in the full model.

Remark E: Fitted versus implied total effect

The standard approach in the causal inference literature is to use the implied total effect that results from fitting the full outcome model and the model for the mediator. We say "implied" here because the marginal model $\mathrm{E}[Y|X]$ is never actually fit to the data. Instead, the sum of the estimated natural direct and indirect effects is used as the total effect estimate (for instance, this is how the mediation package in R estimates the total effect). In contrast, our approach estimates the total effect directly from the fitted marginal model.

Importantly, the estimated total effect obtained from *fitting* the marginal model $\mathrm{E}[Y|X]$ does not necessarily equal the sum of the estimated natural direct and indirect effects, an unexpected finding. We found this to be the case when fitting the full model $\mathrm{E}[Y|X,M] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM$, the mediator model $\mathrm{E}[M|X] = \alpha_0 + \alpha_X X$, and the implied marginal model $\mathrm{E}[Y|X] = \gamma_0 + \gamma_X X + \gamma_X^2 X^2$. To be clear, our empirical estimate of the total effect $\gamma_X(x - x_\circ) + \gamma_X^2(x^2 - x_\circ^2)$ did not equal the sum of the natural direct and indirect effects. We can only speculate that the maximum likelihood fit of the reduced model is not equivalent to the implied reduced model derived from the maximum likelihood fits of the first two models. One explanation is that several different systems of equations will yield the same reduced model, but only one reduced model is implied once the outcome model and mediator model are fit. This is an interesting finding that merits further study.

2.8 Software

While the BRAIN-ICU data used for the examples is not publically available, software in the form of R code and documentation is available at https://github.com/trippcm/Biostatistics-Mediation-R-Code.

Acknowledgments

Example data was derived from the "Bringing to Light the Risk Factors and Incidence of Neuropsychological Dysfunction in ICU Survivors (BRAIN-ICU) Study" funded by the NIH (AG027472) and published in NEJM (Pandharipande et al. 2013).

We appreciate the contributions and comments of BRAIN-ICU investigators Timothy Gerard, Wes Ely, and Pratik Pandharipande. The authors thank Professors Kristopher Preacher, Jonathan Schildcrout, Robert Johnson, Melinda Aldrich, Bryan Shepherd, and Robert Greevy for critical reading, helpful suggestions, and valuable feedback on the original version of the paper. We are also grateful for the constructive comments from the associate editor and two anonymous referees which ultimately improved the presentation of our ideas.

2.9 Appendix

Pearl's *mediation formula* is a generalization of the product of coefficients approach and can be used to estimate causal mediation effects from any type of model Pearl (2001); Imai, Keele and Tingley (2010); Pearl (2012*a*).

$$\begin{aligned}
\text{NDE}(x, x_\circ) &= \text{E}[Y(x, M(x_\circ)) - Y(x_\circ, M(x_\circ))] \\
&= \Sigma_{c,m} \left(\text{E}[Y|x, m, c] - \text{E}[Y|x_\circ, m, c]\right) \text{P}[m|x_\circ, c]\text{P}[c] \\
\text{NIE}(x, x_\circ) &= \text{E}[Y(x, M(x)) - Y(x, M(x_\circ)] \\
&= \Sigma_{c,m}\text{E}[Y|x, m, c] \left(\text{P}[m|x, c] - \text{P}[m|x_\circ, c]\right) \text{P}[c] \\
\text{TE}(x, x_\circ) &= \text{E}[Y(x) - Y(x_\circ)] \\
&= \Sigma_c \left(\text{E}[Y|x, c] - \text{E}[Y|x_\circ, c]\right) \text{P}[c]
\end{aligned}$$

N=50	Avg $\widehat{\text{IDE}}$	Bias $\widehat{\text{IDE}}$	Var $\widehat{\text{IDE}}$	Avg $\widehat{\text{Var}}(\widehat{\text{IDE}})$	Bias $\widehat{\text{Var}}(\widehat{\text{IDE}})$
Regression formula	1.5541	0.0541	0.0217	0.0217	-0.0001
Sobel's formula	1.5541	0.0541	0.0217	0.0975	0.0758
Bootstrap cases	1.5375	0.0375	0.0213	0.0931	0.0714
Bootstrap residuals	1.5541	0.0541	0.0217	0.0203	-0.0014
MC difference of coef	1.5540	0.0540	0.0217	0.1616	0.1398
MC product of coef	1.5540	0.0540	0.0217	0.0982	0.0765

N=100	Avg $\widehat{\text{IDE}}$	Bias $\widehat{\text{IDE}}$	Var $\widehat{\text{IDE}}$	Avg $\widehat{\text{Var}}(\widehat{\text{IDE}})$	Bias $\widehat{\text{Var}}(\widehat{\text{IDE}})$
Regression formula	1.8430	0.3430	0.0185	0.0171	-0.0015
Sobel's formula	1.8430	0.3430	0.0185	0.0417	0.0231
Bootstrap cases	1.8462	0.3462	0.0188	0.0533	0.0347
Bootstrap residuals	1.8430	0.3430	0.0185	0.0165	-0.0020
MC difference of coef	1.8430	0.3430	0.0185	0.0656	0.0471
MC product of coef	1.8430	0.3430	0.0185	0.0418	0.0232

N=200	Avg $\widehat{\text{IDE}}$	Bias $\widehat{\text{IDE}}$	Var $\widehat{\text{IDE}}$	Avg $\widehat{\text{Var}}(\widehat{\text{IDE}})$	Bias $\widehat{\text{Var}}(\widehat{\text{IDE}})$
Regression formula	1.3788	-0.1212	0.0043	0.0043	-0.0001
Sobel's formula	1.3788	-0.1212	0.0043	0.0171	0.0128
Bootstrap cases	1.3773	-0.1227	0.0043	0.0175	0.0131
Bootstrap residuals	1.3787	-0.1213	0.0043	0.0042	-0.0001
MC difference of coef	1.3787	-0.1213	0.0043	0.0286	0.0242
MC product of coef	1.3787	-0.1213	0.0043	0.0172	0.0128

Table 2.3: Results of 5000 simulations (10,000 bootstrap and 10,000 Monte Carlo replications per simulation) of the estimated indirect effect and its estimated variance under the simple mediation model with a true indirect effect of 1.5 for sample sizes N = 50, 100, 200. Note that for the simple mediation model, the (natural) indirect effect for a unit change in the exposure equals the essential mediation component Δ. We use $\widehat{\text{IDE}}$ as an acronym for the indirect effect.

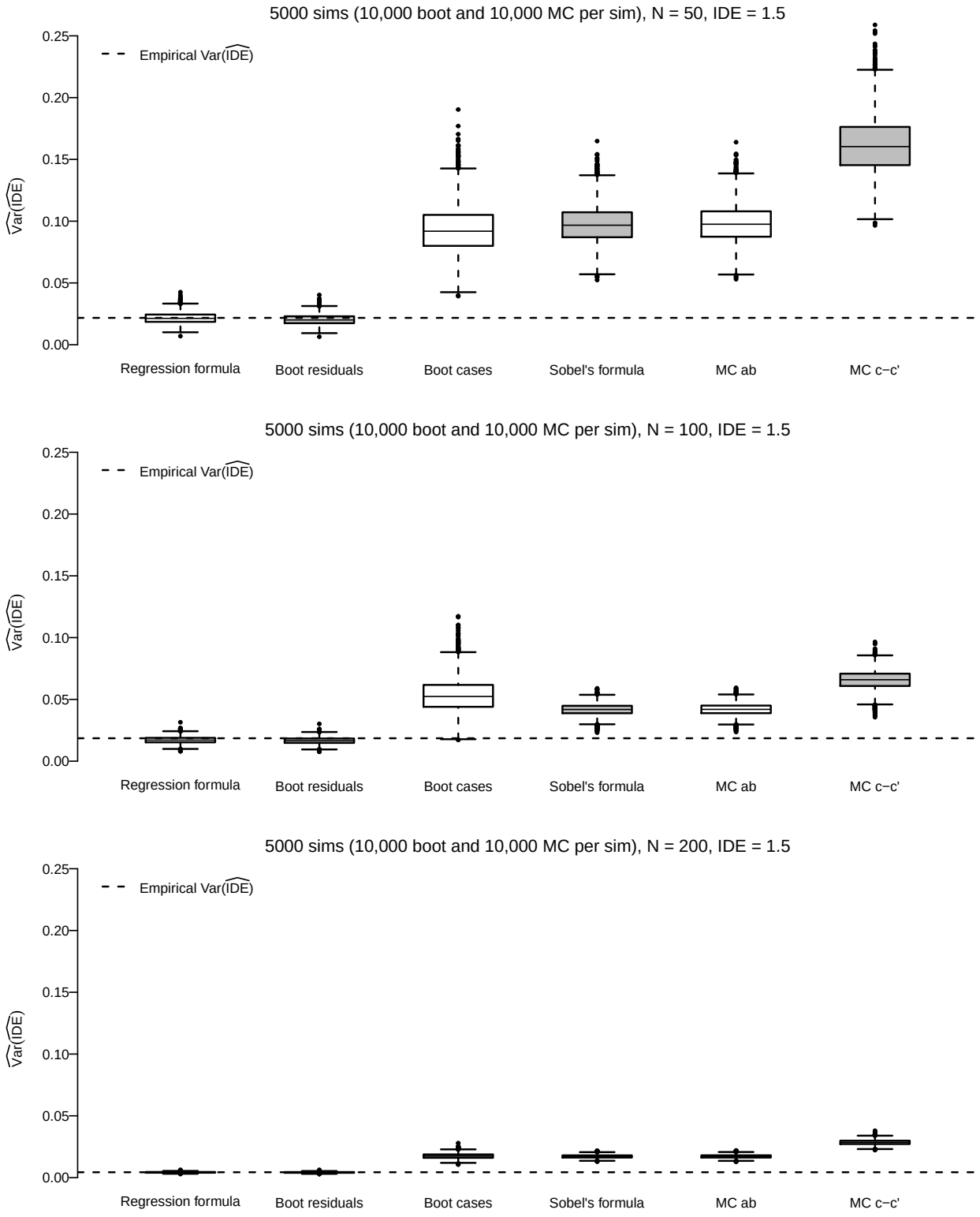


Figure 2.4: Results of simulating the estimated variance of indirect effect estimates from the simple mediation model for sample sizes $N = 50, 100, 200$. Note that for the simple mediation model, the (natural) indirect effect for a unit change in the exposure equals the essential mediation component Δ. We use $\widehat{\text{IDE}}$ as an acronym for the indirect effect. We compare $\widehat{\text{Var}}(\widehat{\text{IDE}})$ using the analytical regression formula, bootstrapping residuals and cases, Sobel's formula, and Monte Carlo methods for the product and difference of coefficients.

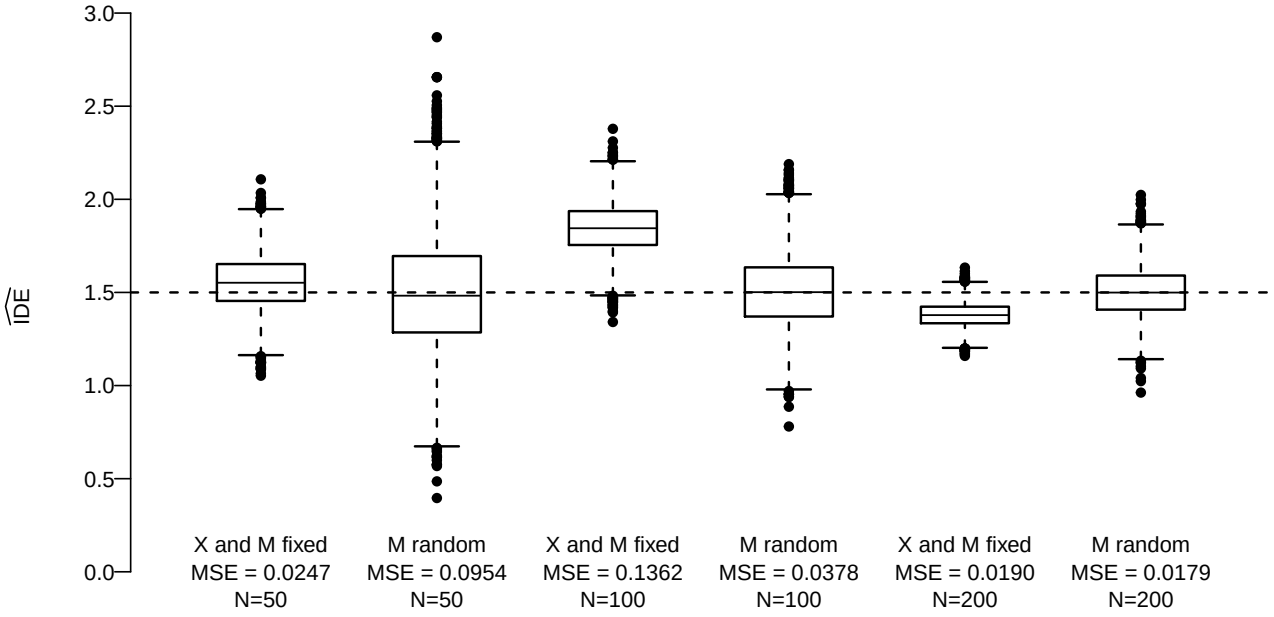


Figure 2.5: We use $\widehat{\text{IDE}}$ as an acronym for the estimated (natural) indirect effect. Results of 5000 simulations of $\widehat{\text{IDE}}$ from the simple mediation model for sample sizes $N = 50, 100, 200$. Note that using the product of coefficients or difference of coefficients yields equivalent estimates of $\widehat{\text{IDE}}$ for this simple model. We compare the effect of treating both X and M as fixed covariates (i.e. at every simulation a new Y is generated conditional on the same X and M) and treating X as fixed and M as a random variable (i.e. at every simulation a new M and a new Y are generated). Notice that when M is treated as a random variable, the bias of $\widehat{\text{IDE}}$ is reduced. Intuitively, when we let M vary, r_{XM} varies and better approximates ρ_{XM} than the sample correlation from a single draw from the distribution of M. We notice the bias-variance tradeoff since when M is random we also see an increased variance in the distribution of $\widehat{\text{IDE}}$. For N = 50, $\text{cov}(x, m) = 2.606$, $\text{var}(x) = 2.514$, $\text{var}(m) = 6.740$. For N = 100, $\text{cov}(x, m) = 4.785$, $\text{var}(x) = 4.156$, $\text{var}(m) = 9.316$. For N = 200, $\text{cov}(x, m) = 3.227$, $\text{var}(x) = 3.510$, $\text{var}(m) = 6.933$.

CHAPTER 3

EXTENSIONS, VISUALIZATIONS, AND APPLICATIONS TO BEHAVIORAL SCIENCE

3.1 Introduction

Psychologists and social scientists concerned with dynamic relations and the mechanisms by which an exposure affects an outcome have been studying mediation processes for decades (Woodworth 1928; Alwin and Hauser 1975). A large body of literature and a variety of methods exist for conducting mediation analyses; the Baron-Kenny causal steps approach (Baron and Kenny 1986; Zhao et al. 2010), the structural equation modeling approach (Gunzler et al. 2013), and the potential outcomes approach (Robins and Greenland 1992; Pearl 2001; Imai, Keele and Tingley 2010; VanderWeele 2015) are well-known frameworks. Additional background information can be found in MacKinnon 2008; Gelfand et al. 2009; Preacher 2015 and in Appendix 3.9.1.

This paper further develops our recently-proposed classical regression approach to mediation analysis (Saunders and Blume 2017) and discusses its implications in the context of existing methods. Instead of fitting a system of equations to estimate the total, direct, and indirect effects, the classical regression approach uses a simple formula to estimate mediation effects from the fit of only one model. It is essentially a generalization of the difference of coefficients approach, which seeks to evaluate the reduction in the total effect when indirect paths are blocked. The difference approach is recognized as being of great importance in health policy research (Pearl 2012*a*; VanderWeele 2013; Naimi et al. 2014; VanderWeele 2015). For example, when studying how an intervention (such as seeing a psychologist for help with depression) can prevent adverse mental health outcomes (such as suicide), the difference of coefficients measures the maximum preventive effect of said therapy on the mediating pathways. Furthermore, this new approach yields a closed-form expression for the model-based variance, an improvement over widely-used approximations (e.g., delta method, bootstrap, Monte Carlo).

Our approach extends to settings with multiple mediators, interactions, and nonlinearities. Advanced regression tools are then easily applied to a single model rather than to the system of equations. We illustrate the new approach and compare it to existing methods in a series of detailed, reproducible examples.

3.2 Background and notation

3.2.1 What is a mediator?

Several types of variables may be present when analyzing the relationship between an exposure and some outcome of interest. A confounder is a variable related to two factors of interest that falsely obscures or accentuates the relationship between them (Meinert 1986). We want to adjust for appropriate confounders to obtain an unbiased estimate of the relationship between the exposure and the outcome. By contrast, a moderator (also known as "effect-modifier" in the epidemiologic literature) is a variable that affects the direction or strength of the relationship between the exposure and outcome (Baron and Kenny 1986). In regression analysis, we typically account for moderators by including interaction terms in the model. Lastly, a mediator represents "the generative mechanism through which the focal independent variable is able to influence the dependent variable" (Baron and Kenny 1986) or "a variable that occurs in a causal pathway from an independent variable to a dependent variable" (Last 1988). We will return to the importance of distinguishing between these types of variables when we discuss the assumptions of mediation models. Next, we introduce the simple mediation model.

3.2.2 The simple mediation model

By partitioning the *total effect* of an exposure into its *direct* and *indirect* components, mediation analysis seeks to understand how much of an exposure's effect on an outcome is transmitted through intermediate pathways. The total effect (TE) of the exposure variable X on the outcome Y represents how much a change in X results in a change Y, irrespective of the mechanisms by which the change occurs; the part of the total effect that is not transmitted through intervening variables is called the direct effect (Alwin and Hauser 1975); the indirect effect is the part of a variable's total effect that is transmitted to the outcome via a mediating variable(s) M. Suppose we have one exposure, one continuous mediator, and one continuous outcome. The Baron-Kenny *simple mediation model* is illustrated in Figure 2.1 and represented by the following three regression equations.

$$\mathrm{E}[Y|X, M] = \beta_0 + \beta_X X + \beta_M M \tag{3.1}$$

$$\mathrm{E}[M|X] = \alpha_0 + \alpha_X X \tag{3.2}$$

$$\mathrm{E}[Y|X] = \beta_0^* + \beta_X^* X \tag{3.3}$$

This model assumes linear relationships, no interaction among the variables, and normally distributed errors. Although the original Baron-Kenny model did not include confounders, it is important to adjust for confounders of the type listed in Figure 2.2 so that mediation effects are identifiable (simply add the set of relevant confounders to each model).

3.2.3 Assumptions for causal inference

Critical assumptions concerning the relationships in a proposed mediation model rely on theory and empirical support. To infer causality from a mediation analysis, one must assume the confounders enumerated in Figure 2.2 have been accounted for (VanderWeele 2015). In addition, mediation analysis assumes the temporal order of the variables was correctly specified (Judd and Kenny 1981; Stone and Sobel 1990). A mediator must truly be a dependent variable relative to the exposure and an independent variable relative to the outcome.

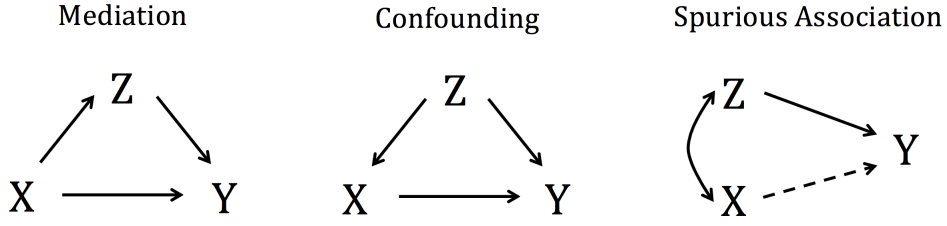


Figure 3.1: Statistically indistinguishable three-variable systems: The arrows represent the causal direction of the effects between variables and dashed lines represent a spurious effect between X and Y. The left panel represents the simple mediation model where Z mediates the effect of X on Y. The middle panel shows Z confounding the relationship between X and Y. The right panel shows X and Z as two covariates having a reciprocal relationship.

Mediation analysis also relies on correctly specified causal directions (McDonald 1997). Consider the diagrams in Figure 3.1 where X is the exposure, Y is the outcome, and Z is a third variable influencing the effect of X on Y. The dashed lines represent a spurious $X \rightarrow Y$ effect. The first panel displays Z acting as a mediator; the second panel shows Z influencing both X and Y, leading to a spurious $X \rightarrow Y$ effect; the third panel shows a reciprocal relationship between X and Z, with Z affecting Y without being causally intermediate. Although these models are conceptually distinct, they are mathematically equivalent and cannot be empirically distinguished from one another with cross-sectional data (Cole and Maxwell 2003).

3.2.4 Intersection of existing frameworks

Many articles on mediation analysis tout the benefits of one framework over another for making causal inferences. Social and behavioral scientists tend to adopt the structural equation modeling (SEM) language, while statisticians more often use the potential outcomes (PO) approach (and may even be unfamiliar with the basic tenets of SEM). Although one framework may lend itself to a specific research question, the SEM and PO frameworks are "logically equivalent," a result proven formally by (Galles and Pearl 1998).

> "The essential difference between the SEM and PO frameworks is that the former encodes causal knowledge in the form of functional relationships among ordinary variables, observable as well as latent, while the latter encodes such knowledge in the form of statistical relationships among hypothetical (or counterfactual) variables, whose value is determined only after a treatment is enacted... A systematic analysis of the syntax and semantics of the two notational systems reveals that they are logically equivalent; a theorem in one is a theorem in the other, and an assumption in one has a parallel interpretation in the other (Bollen and Pearl 2013)."

Providing definitions of causal mediation effects and their regression-based estimands, Table 3.1 is a crosswalk between the SEM and PO nomenclature.

$E[Y|X,M,C] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM + \beta_C C$
$E[M|X,C] = \alpha_0 + \alpha_X X + \alpha_C C$
$E[Y|X,C]^1 = \beta^*_0 + \beta^*_X X + \beta^*_{X^2} X^2 + \beta^*_C C$

SEM Name	PO Name(s)	PO Definition	Regression-based estimand for continuous X	for (0,1) X
Total Effect[2] of X		$Y(x) - Y(x_\circ)$	$(\beta^*_X + \beta^*_{X^2})(x - x_\circ)$	β^*_X
Direct Effect of X	$(\beta_X + \beta_{XM}M)(x - x_\circ)$			
a) set $M = m$	Controlled direct effect	$Y(x,m) - Y(x_\circ,m)$	$(\beta_X + \beta_{XM}m)(x - x_\circ)$	$(\beta_X + \beta_{XM}m)$
b) set $M = E[M\|x_\circ,c]$	Natural direct effect †			
	Pure direct effect §	$Y(x,M(x_\circ)) - Y(x_\circ,M(x_\circ))$	$(\beta_X + \beta_{XM}E[M\|x_\circ,c])(x - x_\circ) = (\beta_X + \beta_{XM}(\alpha_0 + \alpha_X x_\circ + \alpha_C c))(x - x_\circ)$	$\beta_X + \beta_{XM}(\alpha_0 + \alpha_C c)$
	Average direct effect (control) ⋆			
c) set $M = E[M\|x,c]$	Total direct effect §	$Y(x,M(x)) - Y(x_\circ,M(x))$	$(\beta_X + \beta_{XM}E[M\|x,c])(x - x_\circ) = (\beta_X + \beta_{XM}(\alpha_0 + \alpha_X x + \alpha_C c))(x - x_\circ)$	$\beta_X + \beta_{XM}(\alpha_0 + \alpha_X + \alpha_C c)$
	Average direct effect (treatment) ⋆			
Indirect Effect of X				
a) set $X = x_\circ$	Pure indirect effect §	$Y(x_\circ,M(x)) - Y(x_\circ,M(x_\circ))$	$(\beta_M + \beta_{XM}x_\circ)(E[M\|x,c] - E[M\|x_\circ,c]) = (\beta_M + \beta_{XM}x_\circ)\alpha_X(x - x_\circ)$	$\beta_M\alpha_X$
	Average causal mediation effect (control) ⋆			
b) set $X = x$	Natural indirect effect †			
	Total indirect effect §	$Y(x,M(x)) - Y(x,M(x_\circ))$	$(\beta_M + \beta_{XM}x)(E[M\|x,c] - E[M\|x_\circ,c]) = (\beta_M + \beta_{XM}x)\alpha_X(x - x_\circ)$	$(\beta_M + \beta_{XM})\alpha_X$
	Average causal mediation effect (treat) ⋆			
Portion Eliminated	PE = Total effect - controlled direct effect	$Y(x) - Y(x_\circ) - (Y(x,m) - Y(x_\circ,m))$	$(\beta^*_X + \beta^*_{X^2} - (\beta_X + \beta_{XM}m))(x - x_\circ)$	$\beta^*_X - (\beta_X + \beta_{XM}m)$

[1]The marginal model E$[Y|X,C]$ is obtained by $\mathrm{E}_{M|X,C}\mathrm{E}[Y|X,M,C]$. Notice that the marginal model has an X^2 term and the full model does not.

[2]The maximum likelihood fit of the total effect $(\hat{\beta}^*_X + \hat{\beta}^*_{X^2})(x - x_\circ)$ does not always equal the sum of the estimated natural direct and indirect effects.

If X is a binary variable, then fitting the model E$[Y|X,C]$ will drop the X^2 term and the total effect estimate $\hat{\beta}^*_X$ will equal the sum of the natural direct and indirect effects.

†VanderWeele

§Robins and Greenland

⋆Imai et al.

Table 3.1: Nomenclature and definitions of causal mediation effects for exposure X, mediator M, confounders C, and outcome Y. Effects compare the value $X = x$ to $X = x_\circ$ (the referent exposure).

3.3 The recently proposed classical regression framework

Recall that the simple mediation model (3.1-3.3) assumes X is linearly related to Y. A more general formulation allows the effect of the exposure to be nonlinear and includes additional confounders C: $\mathrm{E}[Y|X, M, C] = \beta_0 + \beta_X h(X) + \beta_M M + \beta_C C$, where $h(X)$ is a flexible function of X (e.g., $\log(X)$). Consider p exposures $\boldsymbol{X}$, j mediators $\boldsymbol{M}$, and l confounders $\boldsymbol{C}$ such that the full model and its implied reduced model are

$$\mathrm{E}[Y|\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{C}] = \beta_0 + \boldsymbol{\beta_X h(X)} + \boldsymbol{\beta_M M} + \boldsymbol{\beta_C C} \tag{3.4}$$

$$\mathrm{E}[Y|\boldsymbol{X}, \boldsymbol{C}] = \beta_0^* + \boldsymbol{\beta_X^* h(X)} + \boldsymbol{\beta_C^* C} \tag{3.5}$$

The vector $\boldsymbol{h(X)}$, such as $[X,X^2]$ or $[X,XC]$, captures the non-linear trends in $\boldsymbol{X}$. In a recent paper (Saunders and Blume 2017), we named the difference in exposure pathway coefficients $\Delta = \boldsymbol{\beta_X^* - \beta_X}$ the essential mediation components (EMCs) of $\boldsymbol{X}$. We derived analytical estimates of the EMCs and their model-based variance from the fit of a single regression model. Because the fit of only one model is required, inference for causal mediation effects (which are functions of the EMCs) follows naturally.

Our method uses the "full" outcome model (3.4) and the model for the total effect of the exposure (3.5), in which the effect of the mediator is blocked. The general idea of our approach is to use the sweep operator on the full model to obtain coefficients from any nested reduced model, without having to actually fit said reduced model (Goodnight 1979). This allows us to obtain the EMCs from the fit of the full model alone. Saunders and Blume discuss the advantages that result from having to fit only one model (e.g., simplified application of regression tools and reduced computation time).

3.3.1 The essential mediation components

A general formula for estimating the EMCs from the fit of the full regression model (3.4) is

$$\begin{aligned} \hat{\Delta} \equiv \hat{\boldsymbol{\beta}}_{\boldsymbol{X}}^* - \hat{\boldsymbol{\beta}}_{\boldsymbol{X}} &= -\hat{\boldsymbol{V}}_{\boldsymbol{XM}} \hat{\boldsymbol{V}}_{\boldsymbol{M}}^{-\mathbf{1}} \hat{\boldsymbol{\beta}}_{\boldsymbol{M}} \\ &= -\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{X}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{M}}) \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{M}})^{-1} \hat{\boldsymbol{\beta}}_{\boldsymbol{M}}, \end{aligned} \tag{3.6}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{X}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}$ are the vectors of estimated exposure and mediator coefficients from the full model, $\hat{\boldsymbol{V}}_{\boldsymbol{XM}}$ is the covariance between $\hat{\boldsymbol{\beta}}_{\boldsymbol{X}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}$, and $\hat{\boldsymbol{V}}_{\boldsymbol{M}}^{-\mathbf{1}}$ is the inverse variance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}$. For the simple mediation model, $-\widehat{\mathrm{Cov}}(\hat{\beta}_X, \hat{\beta}_M)\widehat{\mathrm{Var}}(\hat{\beta}_M)^{-1}\hat{\beta}_M$

equals the Baron-Kenny product of coefficients ab, where $-\widehat{\text{Cov}}(\hat{\beta}_X, \hat{\beta}_M)\widehat{\text{Var}}(\hat{\beta}_M)^{-1}$ equals a and $\hat{\beta}_M$ equals b.

The distinction between changes in the exposure pathway coefficients (the EMCs) and causal mediation estimands (e.g., portion eliminated, natural indirect effect) is critical. For the simple mediation model (3.1-3.3), the EMC $\hat{\Delta} = \hat{\beta}_X^* - \hat{\beta}_X$ is mathematically equivalent to the indirect effect *for a unit change in the exposure*; more generally, causal mediation estimands are *functions* of the EMCs. To estimate the portion eliminated comparing some referent level of the exposure $x_\circ$ to x, use

$$\text{PE}(x, x_\circ) = [\boldsymbol{h}(x) - \boldsymbol{h}(x_\circ)]\Delta \tag{3.7}$$

When the exposure or mediator effects are non-scalar, formulas (3.6) and (3.7) allow for estimation of multidimensional mediation effects from the fit of a single fitted regression model (3.4), rather than fitting separate models and aggregating effect estimates. A list of commonly encountered mediation models for which the controlled and natural direct effects are equivalent (and as a result the portion eliminated and the natural indirect effect are the same) can be found in Table 2.2. For these models, the natural indirect effect can be estimated using the formula: $\text{NIE}(x, x_\circ) = [\boldsymbol{h}(x) - \boldsymbol{h}(x_\circ)]\Delta$.

3.3.2 The model-based variance

The closed-form expression for the fully conditional variance of the EMCs is given by $\text{Var}(\hat{\Delta}|\boldsymbol{X}, \boldsymbol{M}) = \boldsymbol{V_{XM}V_M^{-1}V_{MX}}$. The model-based variance of the natural indirect effect (and more generally, the portion eliminated) is simply $[\boldsymbol{h}(x) - \boldsymbol{h}(x_\circ)]\text{Var}(\hat{\Delta}|\boldsymbol{X}, \boldsymbol{M})[\boldsymbol{h}(x) - \boldsymbol{h}(x_\circ)]^{\text{T}}$, which requires fitting only model (3.4). In classical regression settings, the standard conditional variance is used for inference and one could argue in favor of treating both the exposure and the mediator as fixed since the mediator is a theoretic consequent of the exposure (Pearl 2012*a*). Alternately, one can marginalize over M as discussed in Section 2.3.3. A parametric (residual-based) bootstrap approximates the conditional model-based variance, while the nonparametric (case-based) bootstrap approximates the fully unconditional variance (marginalized over both exposure and mediator). Information on commonly used approximations to the variance is provided in Appendix 3.9.2.

3.3.3 Extensions to non-nested mediation systems

A direct application of the EMC approach is not applicable to mediation models in which the marginal model is not nested within the full model. Here we show how a recursive sweep algorithm solves this problem. Simply specify a global model under which (3.4) and (3.5) are nested, and use formula (3.6) to "sweep" from the global model to the full model and also from the global model to the marginal model. After this "double sweep", subtract the corresponding Δs to get the proper EMCs. This expands the applicability of the classical regression framework to all possible mediation systems.

Suppose the outcome model, mediator model, and the implied marginal model are $\mathrm{E}[Y|X,M] = \delta_0 + \delta_X X + \delta_M M$, $\mathrm{E}[M|X,C] = \alpha_0 + \alpha_X X + \alpha_C C$, and $\mathrm{E}[Y|X,C] = \kappa_0 + \kappa_X X + \kappa_C C$, respectively. Estimating the EMCs $\kappa_X - \delta_X$ using formula (3.6) requires the marginal model to be a defacto submodel of the full model. However, by fitting a global model $\mathrm{E}[Y|X,M,C] = \lambda_0 + \lambda_X X + \lambda_M M + \lambda_C C$ and using (3.6) to estimate $\kappa_X - \lambda_X$ and $\delta_X - \lambda_X$, we can estimate the EMCs using functionals of only the global model: $\kappa_X - \delta_X = (\kappa_X - \lambda_X) - (\delta_X - \lambda_X)$.

Now consider the model $\mathrm{E}[Y|X,M] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM$. If $\mathrm{E}[M|X] = \alpha_0 + \alpha_X X$, the marginal model for the total effect is $\mathrm{E}[Y|X] = \beta_0^* + \beta_X^* X + \beta_{X^2}^* X^2$. To estimate the EMCs $\Delta^{\mathrm{T}} = [\beta_X^* - \beta_X, \beta_{X^2}^* - 0]$, fit a global model $\mathrm{E}[Y|X,M] = \gamma_0 + \gamma_X X + \gamma_M M + \gamma_{XM} XM + \gamma_{X^2} X^2$ and use (3.6) to estimate $[\beta_X^* - \gamma_X, \beta_{X^2}^* - \gamma_{X^2}] - [\beta_X - \gamma_X, 0 - \gamma_{X^2}] = [\beta_X^* - \beta_X, \beta_{X^2}^* - 0]$. Thus, one can use the difference of coefficients approach and fit only one model to estimate mediation effects from systems where the marginal model is not a submodel of the full model.

It is simple to obtain the variance of the double sweep estimator. If Δ_1 is the change in exposure coefficients between the marginal model and the global model and Δ_2 is the change in exposure coefficients between the full model and the global model, then $\mathrm{Var}(\Delta_1 - \Delta_2) = \mathrm{Var}(\Delta_1) + \mathrm{Var}(\Delta_2) - 2\mathrm{Cov}(\Delta_1, \Delta_2)$. In the first example above, $\Delta_1 = -V_{XM} V_M^{-1} \lambda_M$ to estimate $\kappa_X - \lambda_X$, the change in exposure coefficients between the reduced model (excluding M) and the global model. Then, we use $\Delta_2 = -V_{XC} V_C^{-1} \lambda_C$ to estimate $\delta_X - \lambda_X$, the change in exposure coefficients between the full model (excluding C) and the global model. The variance is thus $(V_{XM} V_M^{-1} V_{MX}) + (V_{XC} V_C^{-1} V_{CX}) - V_{XM} V_M^{-1} V_{XC} V_C^{-1} \mathrm{Cov}(\lambda_M, \lambda_C)$, which can be estimated using functionals of the fitted global model.

3.3.4 Visualizing mediation effects

Mediation effects can be visualized as functions of the exposure and mediator values. Recall that for the simple mediation model specified in (3.1-3.3), the EMC $\Delta = \beta_X^* - \beta_X$ and the indirect effect $[h(x) - h(x_\circ)]\Delta = (\beta_X^* - \beta_X)(x - x_\circ)$. Figure 3.2 panel A shows the indirect effect as the *distance* between the line $h(X)\Delta$ evaluated at x and $x_\circ$.

Now consider the full model with a quadratic exposure effect so that $\boldsymbol{h(X)} = [X, X^2]$, $\mathrm{E}[Y|X, M] = \lambda_0 + \lambda_X X + \lambda_{X^2} X^2 + \lambda_M M$, and the reduced model $\mathrm{E}[Y|X] = \lambda_0^* + \lambda_X^* X + \lambda_{X^2}^* X^2$. The EMCs $[\Delta_1, \Delta_2]^{\mathrm{T}} = [\lambda_X^* - \lambda_X, \lambda_{X^2}^* - \lambda_{X^2}]$ and the indirect effect is $[\boldsymbol{h}(x) - \boldsymbol{h}(x_\circ)]\boldsymbol{\Delta} = [x - x_\circ x^2 - x_\circ^2]\Delta = (\lambda_X^* - \lambda_X)(x - x_\circ) + (\lambda_{X^2}^* - \lambda_{X^2})(x^2 - x_\circ^2)$. The indirect effect is simply the distance between the parabola $\boldsymbol{h(X)\Delta} = \Delta_1 X + \Delta_2 X^2$ evaluated at x and $x_\circ$, as shown in Figure 3.2 panel B. This is a helpful way to illustrate complex indirect effect behavior.

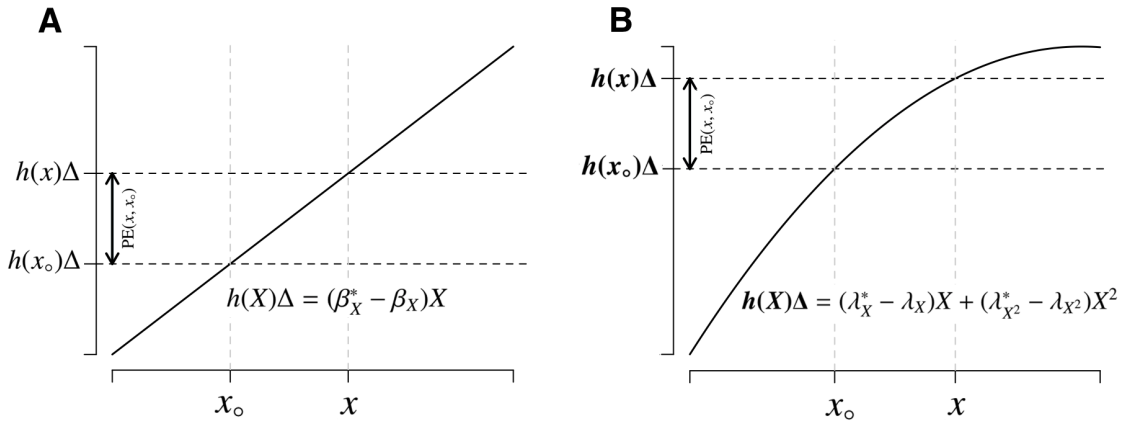


Figure 3.2: Visualizing the indirect effect as a function of the exposure values $(x, x_\circ)$. Panel A shows the indirect effect as the distance (arrow $\leftrightarrow$) between the line $h(X)\Delta$ evaluated at x and $x_\circ$. Panel B shows the indirect effect as the distance between the parabola $\boldsymbol{h(X)\Delta} = \Delta_1 X + \Delta_2 X^2$ evaluated at x and $x_\circ$.

3.3.5 Proposed measure of joint mediation

Until now, we have considered mediation of a single exposure through one mediator. Suppose the mediation model is more complex, such that we are interested in the mediation of a *set* of p exposures $\boldsymbol{X} = (X_1, \ldots, X_p)$ by a set of j mediators $\boldsymbol{M} = (M_1, \ldots, M_j)$. One can fit the full model and obtain separate estimates of

each exposure's indirect effect using formula (3.6). However, summing separate indirect effect estimates to measure total mediation may fail to account for overlapping mediation effects when exposures are mediated jointly.

To address this problem, we propose a general measure of the *joint mediation effect* (JME). This measure aims to capture the amount of mediation that a group of exposure variables is responsible for as a whole, which is not necessarily the sum of the indirect effects. Let $R^2_{Y.\boldsymbol{XMC}}$ and $R^2_{Y.\boldsymbol{XC}}$ be the coefficients of determination from the full model $\mathrm{E}[Y|\boldsymbol{X},\boldsymbol{M},\boldsymbol{C}] = \beta_0 + \boldsymbol{\beta_X X} + \boldsymbol{\beta_M M} + \boldsymbol{\beta_C C}$ and the reduced model $\mathrm{E}[Y|\boldsymbol{X},\boldsymbol{C}] = \beta^*_0 + \boldsymbol{\beta^*_X X} + \boldsymbol{\beta^*_C C}$, respectively. To measure the joint mediation effect of multiple exposures through multiple mediators, it is helpful to scale the variables to have unit variance. We define the JME as a linear combination of p individual EMCs (which, when standardized, are on the same scale), where the EMC $\hat{\Delta}_i$ of exposure X_i is multiplied by X_i's correlation with the outcome Y:

$$r^{\mathrm{T}}_{Y\boldsymbol{X}}\hat{\boldsymbol{\Delta}} = \sum_{i=1}^{p} r_{YX_i}(\hat{\beta}^*_{X_i} - \hat{\beta}_{X_i}) = \sum_{k=1}^{j} r_{YM_k}\hat{\beta}_{M_k} - (R^2_{Y.\boldsymbol{XMC}} - R^2_{Y.\boldsymbol{XC}}) - \sum_{h=1}^{l} r_{YC_h}(\hat{\beta}^*_{C_h} - \hat{\beta}_{C_h})$$

The JME is unitless and will give the same numerical value whether the data are standardized or unstandardized. Note that when the data are unscaled, the JME is $\sum_{i=1}^{p} \frac{\mathrm{Cov}(Y,X_i)}{\mathrm{Var}(Y)}\hat{\Delta}_i = \sum_{k=1}^{j} \frac{\mathrm{Cov}(Y,M_k)}{\mathrm{Var}(Y)}\hat{\beta}_{M_k} - (R^2_{Y.\boldsymbol{XMC}} - R^2_{Y.\boldsymbol{XC}}) - \sum_{h=1}^{l} \frac{\mathrm{Cov}(Y,C_h)}{\mathrm{Var}(Y)}(\hat{\beta}^*_{C_h} - \hat{\beta}_{C_h})$, which may appear incongruous with the unscaled $\hat{\Delta}_i$s because of the difference in units among exposures. We can show this measure is bounded between (-2,2), although there may be tighter achievable bounds.

MacKinnon provides an R^2 measure "designed to localize the amount of variance in Y that is explained by M specific to the mediated effect... by identifying the variance in Y explained by both M and X but not by X alone or M alone:" $R^2_{\text{y.med}} = r^2_{YM} - (R^2_{Y.XM} - R^2_{Y.X})$ (MacKinnon 2008; Fairchild et al. 2009). de Heus (2012) argues that $R^2_{\text{y.med}}$ assigns all overlap between the direct and indirect effects to the indirect effect, which is problematic because they are "heavily interdependent." For the simple mediation model, the JME replaces the correlation r_{YM} with the semi-partial correlation $\hat{\beta}_M = r_{YM.X}$, giving $r_{YX}\hat{\Delta} = r_{YM}\hat{\beta}_M - (R^2_{Y.XM} - R^2_{Y.X})$. The first term $r_{YM}\hat{\beta}_M$ will be less than r^2_{YM} if $r_{YM.X} < r_{YM}$. Our R^2 measure matches MacKinnon's in two rather extreme cases: $\hat{\beta}_M = r_{YM}$ if either X and M are uncorrelated or the effect of X on Y adjusted for M in the full model is zero (in the Baron and Kenny framework, this is the definition of "complete" mediation). The Venn diagrams in Figure 3.3 help intuit the similarity between our measure of the joint mediation effect and MacKinnon's measure for the simple mediation model.

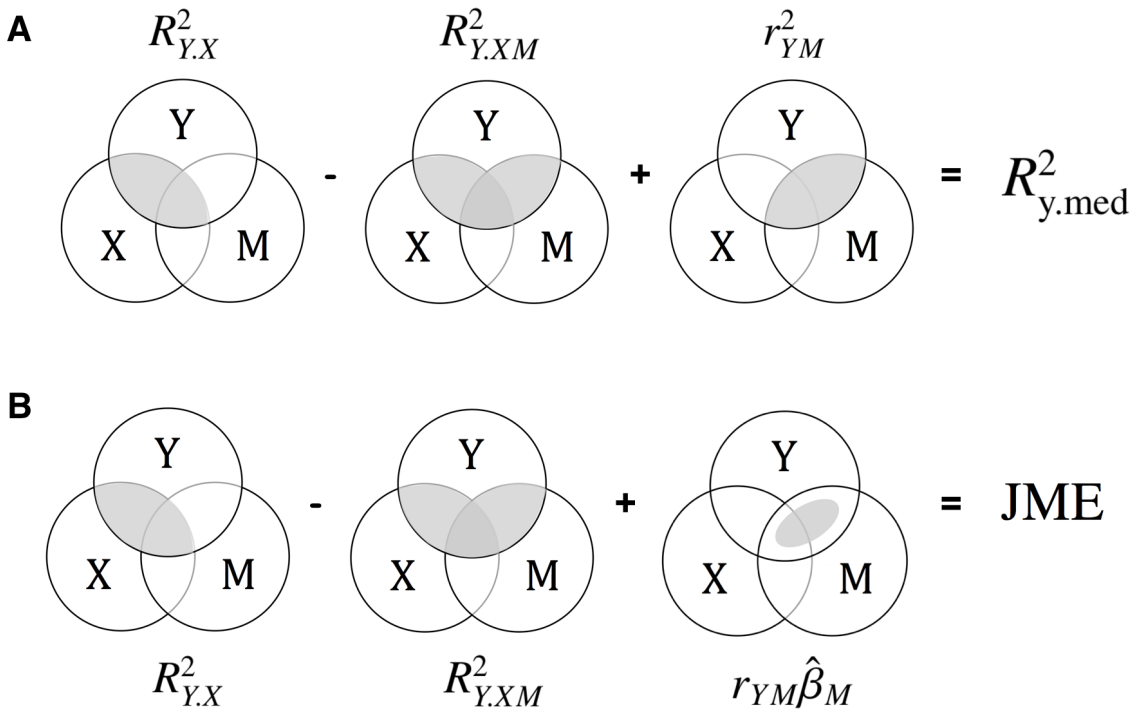


Figure 3.3: A comparison of MacKinnon's $R^2_{\text{y.med}}$ measure of mediation (panel A) to the joint mediation effect (panel B) for the simple mediation model with exposure X, mediator M, and outcome Y.

3.4 Estimating the total effect

The equations in a mediation model form an interdependent system and need to be specified so that they remain congruous. For example, suppose one believes M depends on X quadratically and writes the following model:

$$\begin{aligned} \mathrm{E}[Y|X,M] &= \beta_0 + \beta_X X + \beta_M M \\ \mathrm{E}[M|X] &= \alpha_0 + \alpha_X X + \alpha_{X^2} X^2 \\ \mathrm{E}[Y|X] &= \beta_0^* + \beta_X^* X \end{aligned}$$

This system is not congruous because marginalizing over the full outcome model $\mathrm{E}_{M|X}[Y|X,M]$ does not yield the specified marginal model $\mathrm{E}[Y|X]$. A properly specified system of equations would include an X^2 term in the third equation. System congruency is important because the difference of coefficients approach relies on the full model and the marginal model, while the product of coefficients approach relies on the full model and the mediator model. When the specified system is not congruent, the difference and product approaches are no longer comparable beccause they

are actually using *different systems* to estimate mediation effects. Thus, discrepancies between the two approaches may be due to differences in their underlying mediation systems.

At its core, mediation analysis aims to decompose the total effect into direct and indirect components. Our approach uses the maximum likelihood estimate of the exposure's total effect on the outcome from the marginal model. In contrast, the product of coefficients approach and the potential outcomes approach estimate the direct and indirect effects from fitting the full outcome model and the mediator model, and then sum these to estimate the implied total effect. This estimate of the total effect is "implied" because the marginal model $\mathrm{E}[Y|X]$ is never actually fit to the data. The sum of the estimated natural direct and indirect effects is routinely used as the total effect estimate in the potential outcomes approach (for instance, this is how the R mediation package by Imai, Keele and Tingley (2010) estimates the total effect). Identifying the portion eliminated as the difference between the total and controlled direct effects relies on the first two assumptions in Figure 2.2, whereas estimating the natural direct and indirect effects (and subsequently summing them to estimate the total effect) requires all four assumptions be met (Robins and Greenland 1992). As a result, assumptions about potential mediation pathways play an outsized role in determining the implied total effect.

The difference and product approaches often yield the same conclusion. For the simple mediation model (3.1-3.3), it is well-known that $\mathrm{E}[Y|X] = \mathrm{E}_{M|X}\mathrm{E}[Y|X,M] = \beta_0 + \beta_X X + \beta_M(\alpha_0 + \alpha_X X) = \beta_0^* + \beta_X^* X$, which proves $\beta_X^* = \beta_X + \alpha_X\beta_M$. That is, the estimated total effect $\hat{\beta}_X^*$ obtained from *fitting* the marginal model $\mathrm{E}[Y|X]$ equals the sum of the estimated natural direct and indirect effects. However, this is not always the case. In our experience, the discrepancies present themselves in settings with exposure-mediator interactions (exposure-covariate interactions do not present this problem). When fitting the full model $\mathrm{E}[Y|X,M] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM$, the mediator model $\mathrm{E}[M|X] = \alpha_0 + \alpha_X X$, and the implied marginal model $\mathrm{E}[Y|X] = \gamma_0 + \gamma_X X + \gamma_X^2 X^2$, the empiric estimate of the total effect $\hat{\gamma}_X(x - x_\circ) + \hat{\gamma}_X^2(x^2 - x_\circ^2)$ does not equal the sum of the estimated natural direct and indirect effects. Thus, even when the marginal models are theoretically equivalent, the estimates of the total effect can differ. This can lead to conflicting results since the total effect is used to gauge the overall decomposition of the exposure effect.

Why does this happen? Once the full outcome model and the mediator model are specified, there is one theoretical marginal model. However, there can be several mediator models that imply the same form for the marginal model. For example,

consider the full model $\mathrm{E}[Y|X,M,W] = \beta_0 + \beta_X X + \beta_M M + \beta_W W + \beta_{XW} XW$ and two mediator models: (a) $\mathrm{E}[M|X,W] = \alpha_0 + \alpha_X X + \alpha_W W$ and (b) $\mathrm{E}[M|X,W] = \delta_0 + \delta_X X + \delta_W W + \delta_{XW} XW$. Notice that model (b) includes an XW interaction and model (a) does not. Marginalizing the full model over $M|X,W$ using (a) gives $\mathrm{E}[Y|X,W] = \mathrm{E}_{M|X,W}[Y|X,M,W] = (\beta_0+\alpha_0\beta_M)+(\beta_X+\alpha_X\beta_M)X+(\beta_W+\alpha_W\beta_M)W+\beta_{XW}XW$. Marginalizing over M using (b) gives $\mathrm{E}[Y|X,W] = (\beta_0 + \delta_0\beta_M) + (\beta_X + \delta_X\beta_M)X + (\beta_W + \delta_W\beta_M)W + (\beta_{XW} + \delta_{XW}\beta_M)XW$. *Both marginal models have the form* $\mathrm{E}[Y|X,W] = \gamma_0 + \gamma_X X + \gamma_W W + \gamma_{XW} XW$. As a result, and importantly, the fitted estimates of the marginal model effects $\hat{\boldsymbol{\gamma}}$ will be the same (even though the implied coefficients from the two marginal models are different).

In a conditional process model meant to represent "moderation of only the direct effect," Hayes (2013) (p 335) specifies the same full model as above and a mediator model that omits W and XW: $E[M|X] = \alpha_0 + \alpha_X X$. If $\mathrm{E}[M|X,W] = \delta_0 + \delta_X X + \delta_W W + \delta_{XW} XW$, then $\mathrm{E}[M|X] = E_{W|X}[E[M|X,W]] = E_{W|X}[\delta_0 + \delta_X X + \delta_W W + \delta_{XW} XW] = \delta_0 + \delta_X X + (\delta_W + \delta_{XW} X)E[W|X]$. So writing $E[M|X] = \alpha_0 + \alpha_X X$ assumes $\delta_W = \delta_{XW} = 0$. From this example, we see that the "implied" total effect (obtained from summing the estimated direct and indirect effects) may rely on hidden assumptions about the mediation mechanism. Estimating the marginal model directly can be used to assess the degree to which these assumptions are supported by the data.

3.5 Multiple mediators

Multiple mediator models are useful when researchers hypothesize that the exposure affects the outcome through several intermediate pathways. Consider the full model that contains j mediators $\mathrm{E}[Y|X,M,C] = \beta_0 + \beta_X X + \Sigma_{i=1}^{j}\beta_i M_i + \beta_C C$ and the corresponding reduced model $\mathrm{E}[Y|X,C] = \beta_0^* + \beta_X^* X + \beta_C^* C$. The total and direct effects of X on Y are given by $\beta_X^*(x - x_\circ)$ and $\beta_X(x - x_\circ)$, respectively. To identify mediation effects from multiple mediator models, all four no unmeasured confounding assumptions outlined in Figure 2.2 must hold with respect to the set of mediators $\boldsymbol{M}$.

Estimating the *total indirect effect* aims to determine if the *set* of j mediators transmits the effect of X to Y. This is analogous to conducting a regression analysis with several exposures, with the aim of determining if an overall effect exists. The *mediator-specific indirect effect* represents the amount of the exposure's effect on the outcome that is mediated by M_i above and beyond the other $j-1$ mediators and adjusted for the confounders in the model. The mediator-specific effects are often

attenuated due to collinearity among the mediators (Preacher and Hayes 2008*a*); that is, if two or more mediators share a role in transmitting the effect of X to Y, then the effect attributed uniquely to mediator M_i may exclude this overlapping effect. Additionally, specific indirect effects might have different signs, leading to inconsistent mediation. As a result, mediator-specific indirect effects do not necessarily sum to the total indirect effect.

A clear advantage of our approach in the presence of multiple mediators is that it requires fitting only one model to obtain an estimate of the total indirect effect through M and estimates of mediator-specific indirect effects. It also yields model-based variance estimates that do not require the computation time of resampling methods. Using formula (3.7), the total indirect effect through $\boldsymbol{M}$ and the mediator-specific indirect effect of X through M_i are easily estimated using $-\hat{\boldsymbol{V}}_{X\boldsymbol{M}}\hat{\boldsymbol{V}}_{\boldsymbol{M}}^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}(x - x_\circ)$ and $-\hat{V}_{XM_i}\hat{V}_{M_iM_i}^{-1}\hat{\beta}_{M_i}(x - x_\circ)$, respectively. The corresponding variances are estimated by $\widehat{\text{Var}}([h(x) - h(x_\circ)]\hat{\Delta}) = (x - x_\circ)^2\hat{\boldsymbol{V}}_{X\boldsymbol{M}}\hat{\boldsymbol{V}}_{\boldsymbol{M}}^{-1}\hat{\boldsymbol{V}}_{\boldsymbol{M}X}$ and $(x - x_\circ)^2\hat{V}_{XM_i}\hat{V}_{M_i}^{-1}\hat{V}_{M_iX}$. This approach does not assume the mediators act independently nor does it assume a particular order of effects.

3.5.1 Comparison of multiple mediator models

Existing approaches in the context of multiple mediators include the *single-step multiple mediator model* (MacKinnon 2008), also termed the *parallel multiple mediator model* (Hayes 2013), and the *serial multiple mediator model* (Hayes 2013). The single-step approach specifies a separate model for each mediator in which they independently affect the outcome (see panel A of Figure 3.4). The serial model relies on assumptions about the directionality of the mediators (see panel B of Figure 3.4). Our approach as well as regression-based and weighting approaches by 2013 allow mediators to be interdependent (see panel C of Figure 3.4). Appendix 3.9.3 describes these approaches to multiple mediators in more detail.

3.5.2 Advantages of estimating the total indirect effect through the set of mediators

We recommend the researcher's primary interest lie in the total indirect effect rather than the amount mediated by a specific mediator. Importantly, both our framework and existing approaches to mediation with multiple mediators specify the *same full model* for the outcome Y, and as a result yield the same estimate of the direct effect of X and of the total indirect effect of X through $\boldsymbol{M}$. The estimated total indirect effect through $\boldsymbol{M}$ comes from the full model containing all of the mediators

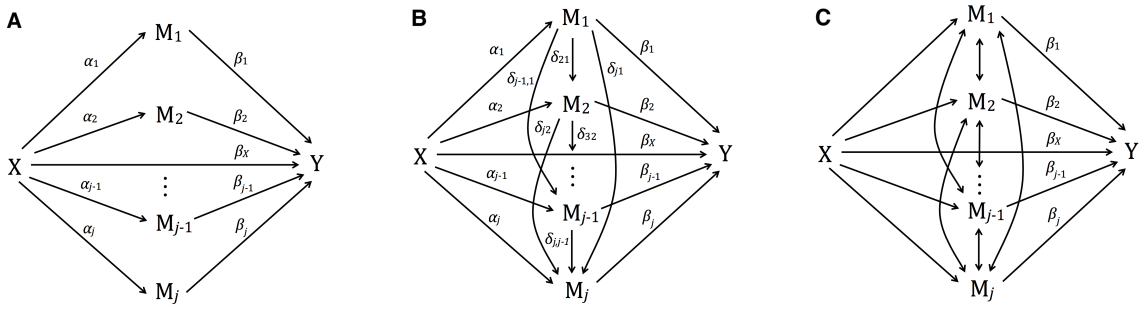


Figure 3.4: Comparison of multiple mediator models. Panel A depicts the single-step or parallel multiple mediator model. Panel B depicts the serial multiple mediator model. Panel C depicts our framework for assessing mediation with multiple mediators. The directions of arrows indicate the assumed causal pathways.

and does not suffer bias from the misspecification of inter-mediator relationships, whereas estimates of mediator-specific indirect effects rely heavily on assumed inter-mediator relationships (such as the order of causal effects in serial multiple mediator models, which can be unverifiable with cross-sectional data). We present relevant examples of this in Section 3.7.3.

3.6 Interactions and moderated mediation

To use the proposed framework with exposure-exposure, mediator-mediator interactions, and exposure-confounder interactions, simply include the interaction terms in the full model and use formulas (3.6) and (3.7) to estimate the EMCs and causal mediation effects. We provide examples in Section 3.7.5.

Exposure-mediator interactions, on the other hand, lead to mediation effects that are less clearly defined. Judd and Kenny (1981) used the term *moderated mediation* to describe when X moderates its own indirect effect on Y through M by moderating the effect of M on Y. Although the causal-steps approach does not accommodate interactions, Baron and Kenny suggested the indirect effect could be conditional on a moderator in their 1986 paper. Preacher et al. (2007) addressed moderated mediation by considering *conditional indirect effects*; Hayes (2013) calls models in which the mediated effects are conditional on moderator variable(s) *conditional process models*. The potential outcomes approach provides decompositions of the total effect into mediated and moderated components. Figure 3.5 depicts the various two, three, and four-way decompositions of the total effect (VanderWeele 2015).

Consider the full model $\mathrm{E}[Y|X,M] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM$. With binary X, the reduced model is $\mathrm{E}[Y|X] = \beta_0^* + \beta_X^* X$. The controlled direct effect ($\beta_X +$

$\beta_{XM}m$) and its variance $\text{Var}(\beta_X) + m^2\text{Var}(\beta_{XM}) + 2m\text{Cov}(\beta_X, \beta_{XM})$ are functions of m. The portion eliminated and its variance are functions of the mediator as well. The portion eliminated $\text{PE} = \text{TE}-\text{CDE}(m) = [\beta_X^* - (\beta_X + \beta_{XM}m)](x - x_\circ)$ can be estimated using $[\hat{\Delta} - \hat{\beta}_{XM}m](x - x_\circ)$. The variance follows by direct calculation: $(x - x_\circ)^2 \left[V_{XM}V_M^{-1}V_{MX} + m^2\text{Var}(\hat{\beta}_{XM}) + 2mV_{XM}V_M^{-1}\text{Cov}(\hat{\beta}_M, \hat{\beta}_{XM})\right]$. We suggest reporting mediation effects for meaningful values of the moderator, such as the sample mean or quartiles.

<table>
<tr><td colspan="4">TE</td></tr>
<tr><td>CDE</td><td>INT_{ref}</td><td>INT_{med}</td><td>PIE</td></tr>
<tr><td>CDE</td><td colspan="2">PAI</td><td>PIE</td></tr>
<tr><td colspan="2">PDE</td><td>INT_{med}</td><td>PIE</td></tr>
<tr><td colspan="3">TDE</td><td>PIE</td></tr>
<tr><td colspan="2">PDE</td><td colspan="2">TIE</td></tr>
<tr><td>CDE</td><td>INT_{ref}</td><td colspan="2">TIE</td></tr>
<tr><td>CDE</td><td colspan="3">PE</td></tr>
</table>

Figure 3.5: The two, three, and four-way decompositions of the total effect (TE) into the controlled direct effect (CDE), reference interaction (INT_{ref}), mediated interaction (INT_{med}), portion attributable to interaction (PAI), pure indirect effect (PIE), pure direct effect (PDE), total direct effect (TDE), total indirect effect (TIE), and the portion eliminated (PE). Formulas for these decompositions are provided and proven formally in VanderWeele (2015).

If the exposure is continuous, then the exposure-mediator interaction model above implies a marginal model that includes an X^2 term: $\text{E}[Y|X] = \lambda_0 + \lambda_X X + \lambda_{X^2}X^2$. To estimate the EMCs $\boldsymbol{\Delta}^{\text{T}} = [\lambda_X - \beta_X, \lambda_{X^2} - 0]$, use the "double sweep" approach described in Section 3.3.3. Alternatively, one can use the mediation formula (provided in Appendix 2.9) to proceed with estimation in this setting. Recall that the estimate of the total effect from fitting the marginal model may not equal the implied total effect from summing the natural direct and indirect effects.

Notice that because M acts simultaneously as a moderator and a mediator, both the direct and indirect effects are affected by the interaction term β_{XM}. As a result, there is more than one way to decompose the total effect. If one attributes β_{XM} to the indirect effect, then the total effect decomposes into the natural (or pure) direct and total indirect effects; if one attributes β_{XM} to the direct effect, then the total effect decomposes into the total direct and pure indirect effects (Robins and Greenland 1992; Pearl 2001). The portion eliminated does not depend on the choice of

decomposition because it is the portion of the total effect attributed to *both* interaction and mediation. Figure 3.6 shows how the pure indirect effect and the portion eliminated measures account for the interaction effect β_{XM} in the presence of an exposure-mediator interaction.

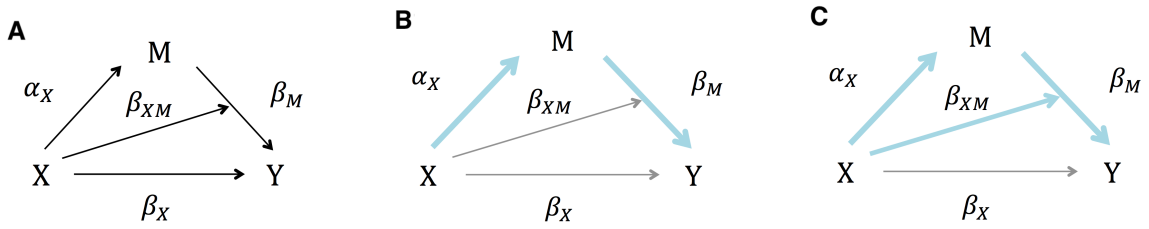


Figure 3.6: Panel A provides a conceptual diagram of the mediation model with an exposure-mediator interaction: $\mathrm{E}[Y|X, M] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM$. The interaction term β_{XM} can be attributed to either the *moderated effect* or the *mediated effect*. In order to obtain the so-called *pure indirect effect*, one attributes the interaction term to the direct effect of the exposure X. The pure indirect effect is depicted by the bolded lines in panel B. The *portion eliminated*, which is the portion of the total effect attributed to both interaction and mediation, is depicted by the bolded lines in panel C.

3.7 Examples using data from social science research

We provide several examples of mediation models to illustrate the efficiency and coherence of the proposed framework. We compare variance estimates obtained from the model-based formula and percentiles of 5000 bootstrap replications. We also include results from the mediation software by 2014; in models with interactions, we found that 5000 simulations were required to yield results matching the analytical solution (the default is 1000). These examples do not cover all aspects of the analysis process and the results are not meant to be interpreted scientifically. Rather, they are intended to demonstrate how to use the methods discussed throughout the paper and to aid researchers who want to implement the newly proposed approach to mediation analysis. All models assume errors $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. Unless otherwise specified, we consider unit changes in continuous exposures so that $(x - x_\circ) = 1$.

3.7.1 Data accessibility

Section 3.7.2 uses a subset of data from the Jobs Search Intervention Study (JOBS II) (Vinokur and Schul 1997) that can be downloaded using the R package mediation http://www.jstatsoft.org/v59/i05/ (Tingley et al. 2014). The remaining examples use data from several studies described in 2017. The data is available for download at http://afhayes.com/introduction-to-mediation-moderation-and-

conditional-process-analysis.html. My R code and documentation is available at https://github.com/trippcm/Mediation-SocialScienceExamples-RCode.

3.7.2 Example: the simple mediation model

Consider N=899 subjects from the JOBS II study, an experiment that randomly assigned unemployed workers to either treatment (job skills workshops) or control (a booklet with job-search tips). Researchers hypothesized that the workshops would lead to reduced depression score by enhancing unemployed workers' confidence in their ability to find a job. Let $X = \texttt{treat}$ be an indicator of whether the patient was randomized to receive treatment or control, $M = \texttt{job-seek}$ be a continuous measure of job search self-efficacy, and $Y = \texttt{depress2}$ be a continuous measure of depressive symptoms. Baseline confounders include pre-treatment depression, age, gender, race, education level, and level of economic hardship.

Does confidence in job-finding (`job-seek`) mediate the effect of job-skills workshops (`treat`) on depression levels (`depress2`)? Consider the simple mediation model that includes baseline depression (`depress1`). The full model is $\texttt{depress2} = \beta_0 + \beta_X\texttt{treat} + \beta_M\texttt{job-seek} + \beta_C\texttt{depress1} + \varepsilon$ and the estimated EMC $\hat{\Delta} = -\hat{V}_{XM}\hat{V}_M^{-1}\hat{\beta}_M = -\widehat{\text{Cov}}(\hat{\beta}_X, \hat{\beta}_M)\widehat{\text{Var}}(\hat{\beta}_M)^{-1}\hat{\beta}_M$. The mediated effect of treatment on depression is $[h(x) - h(x_\circ)]\hat{\Delta} = (x - x_\circ)\hat{\Delta} = -0.0103$ (SE=0.0016) with 95% CI -0.0135 to -0.0071. The residual bootstrap SE=0.0017, Sobel's SE=0.0087, and the case bootstrap SE=0.0088. Notice that the fully conditional standard error estimate aligns with the residual bootstrap, whereas Sobel's estimate aligns with the case-based bootstrap.

To strengthen the validity of the causal mediation analysis assumptions (outlined in Figure 2.2), we include additional pre-treatment confounders: age, gender, race, education level, and economic hardship. The full model is now specified as $\texttt{depress2} = \gamma_0 + \gamma_X\texttt{treat} + \gamma_M\texttt{job-seek} + \gamma_C\texttt{depress1} + \gamma_{C2}\texttt{age} + \gamma_{C3}\texttt{gender} + \gamma_{C4}\texttt{race} + \gamma_{C5}\texttt{educ} + \gamma_{C6}\texttt{econ hard} + \varepsilon$. The new estimated indirect effect is $(x - x_\circ)\hat{\Delta} = -0.0107$ (SE = 0.0016). The 95% CI is (-0.0138, -0.0075), compared to the residual bootstrap (-0.0138, -0.0075), the case-based bootstrap (-0.0304, 0.0067), and the mediation function (-0.0297, 0.0074).

3.7.3 Example: multiple mediators

We now consider data from the Media Influence Study, which analyzed subjects' reactions to a newspaper article about a likely sugar shortage (Tal-Or et al. 2010). Half of the subjects were told the article would be published on the front page and the

other half were told it would be published in an internal supplement. After reading the article, researchers measured the subjects' beliefs about the article's influence and importance. The presumed media influence (`PMI`) and the perceived issue importance (`import`) were two beliefs hypothesized to mediate the relationship between the article's (`location`) and intentions to buy sugar (`reaction`). We fit a multiple mediator model that adjusts for confounders $\boldsymbol{C} = \{\texttt{gender, age}\}$.

The conceptual diagrams in Figure 3.7 depict the single-model approach for multiple mediators, the serial multiple mediator model, and the parallel multiple mediator model. For all three methods, the full model is $\texttt{reaction} = \beta_0 + \beta_X\texttt{location} + \beta_{M_1}\texttt{import} + \beta_{M_2}\texttt{PMI} + \beta_C\texttt{gender} + \beta_{C2}\texttt{age} + \varepsilon$ and the reduced model for the total effect of article location is $\texttt{reaction} = \beta_0^* + \beta_X^*\texttt{location} + \beta_C^*\texttt{gender} + \beta_{C2}^*\texttt{age} + \varepsilon$. Thus, the direct effect of location is $\beta_X(x - x_\circ)$, the total effect is $\beta_X^*(x - x_\circ)$, and the total indirect effect of location mediated through perceived importance and presumed media influence is $(\beta_X^* - \beta_X)(x - x_\circ)$, regardless of whether the analyst uses the single-model, parallel, or serial approach. Importantly, the total indirect effect does not depend on the order or directionality of the mediators, whereas the amount of mediation attributed *specifically* to perceived importance or PMI will differ across methods due to their varying assumptions about inter-mediator relationships.

The single-model approach allows us to estimate how much perceived importance and presumed media influence mediate the relationship between location and reaction using only the full model and formulas (3.6) and (3.7). For $X = \texttt{location}$ and $\boldsymbol{M} = \{\texttt{import}, \texttt{PMI}\}$, the EMCs are

$$\begin{aligned}\Delta &= -\boldsymbol{V_{XM}}\boldsymbol{V_M^{-1}}\boldsymbol{\beta_M} \\ &= -\begin{bmatrix}\text{Cov}(\beta_X, \beta_{M_1}) & \text{Cov}(\beta_X, \beta_{M_2})\end{bmatrix}\begin{bmatrix}\text{Var}(\beta_{M_1}) & \text{Cov}(\beta_{M_1}, \beta_{M_2}) \\ \text{Cov}(\beta_{M_2}, \beta_{M_1}) & \text{Var}(\beta_{M_2})\end{bmatrix}^{-1}\begin{bmatrix}\beta_{M_1} \\ \beta_{M_2}\end{bmatrix}\end{aligned}$$

The total indirect effect through $\boldsymbol{M}$ is estimated by $[h(x) - h(x_\circ)]\hat{\Delta} = 0.4053$ and its empirical variance $\hat{V}_{X\boldsymbol{M}}\hat{V}_{\boldsymbol{M}}^{-1}\hat{V}_{\boldsymbol{M}X}(x - x_\circ)^2 = 0.0031$ (SE = 0.0553).

To estimate how much is mediated specifically through $M_1 = \texttt{import}$, we again apply formulas (3.6) and (3.7): $-\hat{V}_{XM_1}\hat{V}_{M_1}^{-1}\hat{\beta}_{M_1}(x - x_\circ) = 0.1643$. The variance follows directly: $\hat{V}_{XM_1}\hat{V}_{M_1}^{-1}\hat{V}_{M_1X}(x - x_\circ)^2 = 0.0012$ (SE=0.0350). Similarly, to estimate how much the effect of location is mediated through $M_2 = \texttt{PMI}$, use $-\hat{V}_{XM_2}\hat{V}_{M_2}^{-1}\hat{\beta}_{M_2}(x - x_\circ)$ = 0.1359, which has an estimated variance of $\hat{V}_{XM_2}\hat{V}_{M_2}^{-1}\hat{V}_{M_2X}(x - x_\circ)^2 = 0.0010$ (SE = 0.0322). Notice that the mediator-specific indirect effects sum to 0.3002, which is less than the total indirect effect of 0.4053 (perceived importance is correlated with

presumed media influence, $r = 0.28$).

The parallel approach (MacKinnon 2008; Hayes 2013) and the causal regression-based approach (VanderWeele and Vansteelandt 2013) specify the same full model as above, $\texttt{reaction} = \beta_0 + \beta_X \texttt{location} + \beta_{M_1} \texttt{import} + \beta_{M_2} \texttt{PMI} + \beta_C \texttt{gender} + \beta_{C2} \texttt{age} + \varepsilon$, and an additional model for each mediator: $\texttt{import} = \alpha_{01} + \alpha_1 \texttt{location} + \alpha_G \texttt{gender} + \alpha_A \texttt{age} + \varepsilon$ and $\texttt{PMI} = \alpha_{02} + \alpha_2 \texttt{location} + \alpha_{G2} \texttt{gender} + \alpha_{A2} \texttt{age} + \varepsilon$. This specification assumes the mediators "act in parallel" (see Figure 3.7 panel A). The delta method is commonly used to estimate standard errors under this approach. The estimated indirect effect through perceived importance is $\hat{\alpha}_1 \hat{\beta}_{M_1} = 0.2184$ (SE=0.1144) and the estimated indirect effect through PMI is $\hat{\alpha}_2 \hat{\beta}_{M_2} = 0.1869$ (SE=0.1039), which sum to the total indirect effect of 0.4053 (SE = 0.1510).

The serial model given by 2013 requires specifying the order in which the mediators affect each other. Suppose we assume $\texttt{location} \rightarrow \texttt{import} \rightarrow \texttt{PMI} \rightarrow \texttt{reaction}$ (see Figure 3.7 panel B1). The full model is specified as $\texttt{reaction} = \beta_0 + \beta_X \texttt{location} + \beta_1 \texttt{import} + \beta_2 \texttt{PMI} + \beta_C \texttt{gender} + \beta_{C2} \texttt{age} + \varepsilon$ (the same as above), the first reduced model is $\texttt{PMI} = \alpha_{02} + \alpha_2 \texttt{location} + \delta_{21} \texttt{import} + \delta_C \texttt{gender} + \delta_{C2} \texttt{age} + \varepsilon$, and the second reduced model is $\texttt{import} = \alpha_{01} + \alpha_1 \texttt{location} + \alpha_C \texttt{gender} + \alpha_{C2} \texttt{age} + \varepsilon$. There are three estimated indirect effects: $\hat{\alpha}_1 \hat{\beta}_1 = 0.2184$ (SE=0.1144) is the indirect effect of location through import to reaction, $\hat{\alpha}_2 \hat{\beta}_2 = 0.1359$ (SE=0.0982) is the indirect effect of location through PMI to reaction, and $\hat{\alpha}_1 \hat{\delta}_{21} \hat{\beta}_2 = 0.0510$ (SE=0.3235) is the indirect effect of location through importance to PMI to reaction.

To demonstrate how mediator-specific indirect effects depend on the specified order in a serial model, suppose we change the order of mediation to $\texttt{location} \rightarrow \texttt{PMI} \rightarrow \texttt{import} \rightarrow \texttt{reaction}$ (see Figure 3.7 panel B2). The total indirect effect remains unchanged, but now the indirect effect of location through PMI to reaction is 0.1869, the indirect effect of location through importance to reaction is 0.1643, and the indirect effect of location through PMI to importance to reaction is 0.0541. Notice that in either case, the serially mediated indirect effects sum to the total indirect effect of 0.4053.

Estimating mediator-specific indirect effects from the serial model is analogous to examining sequential sums of squares. Although the amount of mediation attributed to specific mediators depends heavily on their assumed order, the serially mediated indirect effects always sum to the total indirect effect (as shown in the example above). In contrast, estimating effects from our proposed framework is analogous to examining partial sums of squares. Just as partial sums of squares do not necessarily sum to the total, the mediator-specific indirect effects from our framework do not

necessarily sum to the total indirect effect. If the mediators are in fact independent, the mediator-specific indirect effects will sum to the total indirect effect.

3.7.4 Example: the joint mediation effect

In order to quantify how article location is jointly mediated through perceived importance and presumed media influence, we look at the joint mediation effect (JME). The standardized total indirect effect of article location through $\boldsymbol{M}$ is 0.131. The joint mediation effect: $r_{YX}\hat{\Delta} = r_{Y\boldsymbol{M}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\boldsymbol{M}} - (R^2_{Y.X\boldsymbol{MZ}} - R^2_{Y.X\boldsymbol{Z}}) - r_{Y\boldsymbol{Z}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\boldsymbol{Z}} = 0.0210$. The fraction of the coefficient of determination from the full model ($R^2_{Y.X\boldsymbol{MZ}}$=0.3377) accounted for by the JME is 0.0622. That is, the JME accounts for about 6% of the total variation in Y explained by the full model. Notice that $R^2_{Y.X\boldsymbol{MZ}}$ (the proportion of variance in the outcome explained by the exposure, mediator, and confounders) is small to begin with, as is the joint mediation effect.

3.7.5 Example: interactions

In this section we consider a study looking at how beliefs about sexism impact women's reactions to discriminatory treatment (Garcia et al. 2010). Female study participants (N = 129) were told that a female attorney lost a promotion to a male candidate who was less qualified due to discriminatory practices of the senior partners. The participants were told either that the attorney confronted the partners or that she did not take action. Researchers then measured how much participants "liked" the attorney, their "perceived appropriateness of the response," and their belief about how widespread sex discrimination is. The hypothesis is that whether or not the attorney protested (X = `protest`) affected participants' perceptions of her (Y = `liking`), and that this association could be mediated by perceived appropriateness of the response (M = `appropriate`). Furthermore, we'll look at whether the mediated effect is moderated by beliefs about the pervasiveness of sex discrimination (W = `sexism`).

3.7.5.1 Exposure-moderator interaction

We include an exposure-moderator interaction so that the full model is $\texttt{liking} = \beta_0 + \beta_X\texttt{protest} + \beta_M\texttt{appropriate} + \beta_W\texttt{sexism} + \beta_{XW}\texttt{protest:sexism} + \varepsilon$. With $h(X) = [X, XW]$, the controlled direct effect is conditional on `sexism`: $(\beta_X + \beta_{XW}w)(x - x_\circ) = (-2.8075 + 0.5426w)(x - x_\circ)$. The EMCs $\Delta^{\mathrm{T}} = [\beta^*_X - \beta_X, \beta^*_{XW} - \beta_{XW}] = -[\mathrm{Cov}(\beta_X, \beta_M), \mathrm{Cov}(\beta_{XW}, \beta_M)]^{\mathrm{T}}\mathrm{Var}(\beta_M)^{-1}\beta_M$ are estimated to be $\hat{\boldsymbol{\Delta}}^{\mathrm{T}} =$

$[-0.9652, 0.2910]$ and the conditional indirect effect $(x - x_\circ, xw - x_\circ w)\Delta = (\beta^*_X - \beta_X)(x - x_\circ) + (\beta^*_{XW} - \beta_{XW})(xw - x_\circ w)$ is estimated to be $-0.9652 + 0.2910w$. Figure 3.8 shows the conditional mediation effects as a function of the moderator $W =$ `sexism`.

The indirect effect marginalized over sexism is $\mathrm{E}[(x - x_\circ, xw - x_\circ w)\Delta|W] = (\beta^*_X - \beta_X)(x - x_\circ) + (\beta^*_{XW} - \beta_{XW})(x - x_\circ)\mathrm{E}[W]$. The variance is estimated using $(x - x_\circ)^2\mathrm{Var}(\Delta_1) + (x - x_\circ)^2\mathrm{E}[W]^2\mathrm{Var}(\Delta_2) + 2(x - x_\circ)^2\mathrm{E}[W]\mathrm{Cov}(\Delta_1, \Delta_2)$. The estimated indirect effect for $W = \bar{w}$ is -0.9652 + 0.2910 * 5.1170 = 0.5238 (SE=0.1029). The regression-based approach by VanderWeele requires fitting the mediator model $\texttt{appropriate} = \alpha_0 + \alpha_X\texttt{protest} + \alpha_W\texttt{sexism} + \beta_{XW}\texttt{protest:sexism} + \varepsilon$ in addition to the full model. The mediation formula estimates the indirect effect using $\mathrm{E}[\hat{\beta}_M(\mathrm{E}[M|x] - \mathrm{E}[M|x_\circ])|W] = \hat{\beta}_M(\hat{\alpha}_X + \hat{\alpha}_{XW}\mathrm{E}[W])(x - x_\circ) = 0.5238$ (SE=0.1295).

3.7.5.2 Exposure-mediator interaction

Now consider a model with an exposure-mediator interaction: $\texttt{liking} = \kappa_0 + \kappa_X\texttt{protest} + \kappa_M\texttt{appropriate} + \kappa_{XM}\texttt{protest:appropriate} + \varepsilon$. We estimate the portion eliminated (the difference between the total and controlled direct effects) using $(\hat{\Delta} - \hat{\kappa}_{XM}m)(x - x_\circ)$. Figure 3.9 shows a plot of $\mathrm{CDE}(m) = \kappa_X + \kappa_{XM}m$ and $\mathrm{PE}(m)$ for a unit change in X. One could also report the mediation effects for specific values of m, such as the sample mean or quartiles. The estimated controlled direct effect for the 25th percentile ($m = 4$) is -0.1616 (SE = 0.2167), and the estimated portion eliminated is 0.6402 (SE = 0.6469).

3.7.6 Example: nonlinear exposure effects

To demonstrate how to include nonlinear exposure effects, we use data from a study on economic stress among N=262 entrepreneurs (Pollack et al. 2012). The hypothesis is that economic stress `stress` leads to a depressed `affect`, which can in turn lead business-persons to `withdraw` from "entrepreneurial activities." We adjust for subjects' `age`, `gender`, business `tenure`, and a self-confidence measure called entrepreneurial self-efficacy (`ESE`).

3.7.6.1 Quadratic exposure effect

To allow for a *quadratic* relationship between stress and withdrawal symptoms, specify the full model as $\texttt{withdraw} = \beta_0 + \beta_{X_1}\texttt{stress} + \beta_{X_2}\texttt{stress}^2 + \beta_M\texttt{affect} + \beta_Z\texttt{gender} + \beta_{Z2}\texttt{age} + \beta_{Z3}\texttt{ESE} + \beta_{Z4}\texttt{tenure} + \varepsilon$. With $\boldsymbol{h}(X) = [X, X^2]$, the EMCs $\boldsymbol{\Delta} = -\boldsymbol{V_{XM}}\boldsymbol{V_M^{-1}}\boldsymbol{\beta_M} = -[\mathrm{Cov}(\beta_X, \beta_M), \mathrm{Cov}(\beta^2_X, \beta_M)]^{\mathrm{T}}\mathrm{Var}(\beta_M)^{-1}\beta_M$. The EMCs are estimated to be $[-0.4620, 0.0651]^{\mathrm{T}}$ and the estimated portion eliminated (which equals

the NIE) is $[(x - x_\circ, x^2 - x_\circ^2)]\hat{\Delta} = -0.4620(x - x_\circ) - 0.0651(x^2 - x_\circ^2) = -0.3970$ (SE = 0.0616). Using the mediation package gives an estimated NIE of -0.3970 (exactly equal to our estimate, as expected) with SE = 0.2023.

3.7.6.2 Splined exposure effect

Suppose instead we wish to model the effect of stress using restricted cubic splines with 4 knots $k_1, k_2, k_3, k_4 = (2.5, 4, 5.5, 7)$ at the 5th, 35th, 65th, and 95th percentiles of `stress`. Consider the full model $\texttt{withdraw} = \beta_0 + \boldsymbol{h}(\texttt{stress})\boldsymbol{\beta_X} + \beta_M\texttt{affect} + \boldsymbol{\beta_Z Z} + \varepsilon$ and the marginal model $\texttt{withdraw} = \beta_0^* + \boldsymbol{h}(\texttt{stress})\boldsymbol{\beta_X^*} + \boldsymbol{\beta_Z^* Z} + \varepsilon$, where $\boldsymbol{Z}$ is the vector of confounders specified in the previous example and $\boldsymbol{h(X)} = [S_1(X), S_2(X), S_3(X)]$ are the splined components of X given by (Harrell 2015):

$$
\begin{aligned}
S_1(X) &= X \\
S_2(X) &= (X - k_1)_+^3 - \frac{(X - k_3)_+^3(k_4 - k_1)}{k_4 - k_3} + \frac{(X - k_4)_+^3(k_3 - k_1)}{(k_4 - k_3)} \\
S_3(X) &= (X - k_2)_+^3 - \frac{(X - k_3)_+^3(k_4 - k_2)}{k_4 - k_3} + \frac{(X - k_4)_+^3(k_3 - k_2)}{(k_4 - k_3)}
\end{aligned}
$$

To put all basis functions for X on the same scale, by default the R function divides the terms $S_j(X), (j > 1)$ by $\tau = (k_4 - k_1)^{2/3}$. The EMCs $\boldsymbol{\Delta} = \boldsymbol{\beta_X^*} - \boldsymbol{\beta_X} = [\beta_1^* - \beta_1, \beta_2^* - \beta_2, \beta_3^* - \beta_3]^{\mathrm{T}}$ can be estimated from the fit of only the full model using formula (3.6). Thus, the indirect effect is given by

$$
[\boldsymbol{h(x)} - \boldsymbol{h(x_\circ)}]\boldsymbol{\Delta} = [S_1(x) - S_1(x_\circ), S_2(x) - S_2(x_\circ), S_3(x) - S_3(x_\circ)] \begin{bmatrix} \beta_1^* - \beta_1 \\ \beta_2^* - \beta_2 \\ \beta_3^* - \beta_3 \end{bmatrix}.
$$

To estimate the indirect effect comparing $x =$ the 75th quantile to $x_\circ =$ the median of economic stress, we have $[h(x) - h(x_\circ)]\hat{\boldsymbol{\Delta}}^{\mathrm{T}} = [5.5 - 4.5, 1.3333 - 0.3951, 0.1667 - 0.0062][-0.0609, 0.3011, -0.1526]^{\mathrm{T}} = 0.1971$. The standard error = 0.0266.

3.8 Summary

In this paper, we defined the essential mediation components, provided formulas for estimating mediation effects and their variance from the fit of a single regression model, showed how to visualize mediation effects, and presented a measure of joint mediation. We highlighted situations in which using the difference and product of coefficients approaches do not yield the same estimate of the total effect of the exposure.

This suggests that discrepancies between these two approaches' estimates of mediation effects depends on the specification of the marginal model and the estimation of the total effect. Last, we provided extensive examples to illustrate our approach and how it can be applied to complex mediation hypotheses, including models with multiple mediators, interactions, and nonlinearities.

The statistical literature abounds with methods for measuring mediation in the simple setting of one exposure, one mediator, and one outcome. However, scientific mediation hypotheses typically involve a more complicated interplay between several variables. Rather than estimating the total, direct, and indirect effects from separate regression equations, one can use our simple formulas to estimate mediation effects and their variance. We recommend using our formula to obtain estimates of the portion eliminated (and in several settings, the natural indirect effect). This approach provides an analytical variance and reduces computation time. For estimating the pathway decompositions displayed in Figure 3.5, we recommend using the formulas given by 2015. When estimating mediation effects, one should thoughtfully consider the plausibility of the assumptions required for causal inference.

3.9 Appendix

3.9.1 Approaches to estimating the indirect effect

3.9.1.1 Baron and Kenny's causal steps

Baron and Kenny published their landmark paper on assessing mediation from the simple mediation model using the *causal steps approach* in 1986. Their approach says that before estimating the indirect effect and its variance, the variables must be significant in a series of hypothesis tests: X must affect M in equation (3.2), X must affect Y in equation (3.3), and M must affect Y in equation (3.1). If the causal steps are established, one estimates the indirect effect and tests for its significance using the Sobel test (Sobel 1982). If the exposure has no effect when the mediator is controlled (if $\beta_X = 0$), then there is "strong evidence for a single, dominant mediator," or so-called *perfect mediation*. If $\beta_X \neq 0$, this is termed *partial mediation* and indicates "the operation of multiple mediating factors" (Baron and Kenny 1986).

Although Baron and Kenny's 1986 paper is considered a cornerstone of mediation analysis (three decades later, researchers have cited their method over 70,000 times), flaws in the causal steps approach have been presented (MacKinnon et al. 2002; Fritz and MacKinnon 2007; Preacher and Hayes 2008*b*; Gelfand et al. 2009; Hayes 2009; Zhao et al. 2010; Hayes 2013; Vansteelandt et al. 2012). If an indirect effect exists

and there is *inconsistent mediation* (the direct effect and indirect effects of X on Y have similar magnitudes and opposite signs, leading to a total effect near zero), using the causal steps approach would stop the analysis at the second step. Plus, a nonzero association between X and Y reducing to zero when a third variable Z is covaried does not necessarily mean that Z mediates the effect of X on Y. Furthermore, the terms "partial" and "complete" mediation are defined in terms of statistical significance (Preacher and Hayes 2008*b*) and concluding "complete mediation" may inhibit future research into other possible mediators. We recommend disregarding the causal steps approach (because it places too much emphasis on statistical significance); instead, one should carefully construct a mediation hypothesis, consider the assumptions required for causal inference, and report mediation effects and their confidence intervals.

3.9.1.2 The product and difference of coefficients approaches

For a unit change in the exposure, the estimated total effect of X is $\hat{\beta}_X^*$, the coefficient for X in equation (3.3), and the estimated direct effect of X is $\hat{\beta}_X$, the coefficient for X in equation (3.1). The difference of coefficients approach estimates the indirect effect for a unit change in X by subtracting the direct effect from the total effect: $\hat{\beta}_X^* - \hat{\beta}_X$. The product of coefficients approach estimates the indirect effect by multiplying the coefficient for X in equation (3.2) by the coefficient for M in equation (3.1): $\hat{\alpha}_X \hat{\beta}_M$. For the simple mediation model with continuous M and Y, the product and difference of coefficients approaches agree and $\hat{\beta}_X^* - \hat{\beta}_X = \hat{\alpha}_X \hat{\beta}_M$. This leads to a nice interpretation of $\beta_X^* = \beta_X + \alpha_X \beta_M$: the total effect of X on Y equals the sum of the direct and indirect effects. Although the product and difference of coefficients approaches agree for linear models, in general the two approaches and their interpretation may differ (Pearl 2012*b*) and there is disagreement as to which approach is preferable (Alwin and Hauser 1975; Preacher and Hayes 2008*b*; Imai, Keele and Tingley 2010).

3.9.1.3 The potential outcomes framework

A formal approach to mediation analysis based on the potential outcomes framework has been developed (Holland 1986; Robins and Greenland 1992; Pearl 2001). Causal mediation effects are defined as contrasts in average potential outcomes that depend on both the exposure and mediator variables. Let $Y(x, m)$ be the potential outcome that would be observed if the exposure X were equal to x and the mediator M were equal to m. Let $Y(x, M(x_\circ))$ be the potential outcome that would be ob-

served if the exposure were equal to x but the mediator M were equal to the value it *would have been if the exposure were equal to* $x_\circ$. Note that the potential outcomes $Y(x, M(x))$ and $Y(x_\circ, M(x_\circ))$ are observable, but $Y(x, M(x_\circ))$ and $Y(x_\circ, M(x))$ can never be observed, and thus are always counterfactual. The counterfactual definitions of causal mediation effects are listed in Table 1. By using the $(x, x_\circ)$ notation, we make explicit that causal mediation effects are generally defined for any two levels of the exposure. When X is a binary exposure, the only possible pair of values is $(0, 1)$.

Until now, we have discussed total, direct, and indirect effects. However, the causal mediation literature distinguishes between *controlled* and *natural* effects. The controlled direct effect measures the effect of X on Y while holding the mediator fixed at level m for everyone in the population: $\text{CDE}(x, x_\circ, m) \equiv Y(x, m) - Y(x_\circ, m)$. The natural direct effect measures the effect of the exposure on the outcome when each individual's mediator is fixed to $M(x_\circ)$, what it would have been "naturally" had the exposure been absent (or equal to some referent value): $\text{NDE}(x, x_\circ) \equiv Y(x, M(x_\circ)) - Y(x_\circ, M(x_\circ))$. The natural indirect effect represents the difference in the outcome if one holds the exposure at level x and changes the mediator from the value that would have been observed under the referent exposure, $M(x_\circ)$, to the value that would have been observed under treatment, $M(x)$: $\text{NIE}(x, x_\circ) \equiv Y(x, M(x)) - Y(x, M(x_\circ))$. Regardless of how the direct and indirect effects are defined, the total effect of X on Y is $\text{TE}(x, x_\circ) \equiv Y(x) - Y(x_\circ)$. Causal effects cannot be estimated at the individual level, but one may estimate average causal effects by taking the expectation of the causal contrasts.

The controlled and natural direct effects diverge in the presence of exposure-mediator interactions. From $\text{E}[Y|X, M, C] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM + \beta_C C$, the controlled direct effect of X is estimated by $E[Y(x, m) - Y(x_\circ, m)|C] = (\beta_X + \beta_{XM} m)(x - x_\circ)$. The natural direct effect is estimated by $E[Y(x, M(x_\circ)) - Y(x_\circ, M(x_\circ))|C] = (\beta_X + \beta_{XM}\text{E}[M|x_\circ])(x - x_\circ) = (\beta_X + \beta_{XM}(\alpha_0 + \alpha_X x_\circ + \alpha_C C))(x - x_\circ)$. To estimate the natural indirect effect, one uses $E[Y(x, M(x)) - Y(x, M(x_\circ))|C] = \alpha_X(\beta_M + \beta_{XM} x)(x - x_\circ)$.

Because the total effect can always be broken down into the natural direct and indirect effects, the natural indirect effect can be written as the difference between the total and natural direct effects: $\text{NIE} = \text{TE} - \text{NDE}$. Another important quantity is the portion eliminated (PE), which is the difference between the total and controlled direct effects: $\text{PE} = \text{TE} - \text{CDE}$ (VanderWeele 2015). For the simple mediation model, the portion eliminated can be estimated using $\text{E}[Y|x] - \text{E}[Y|x_\circ] - (\text{E}[Y|x, m] - \text{E}[Y|x_\circ, m]) = (\beta_X^* - \beta_X)(x - x_\circ)$. In this case, $\widehat{\text{CDE}} = \widehat{\text{NDE}}$, so $\widehat{\text{PE}} = \widehat{\text{NIE}}$. In general,

the difference between the total effect and the controlled direct effect is not equal to the indirect effect. If the full model includes an exposure-mediator interaction, the difference may be nonzero due to an interaction effect, not due to mediation. As shown in Figure 3.5, the portion eliminated equals the PIE (pure indirect effect) + PAI (portion attributable to interaction), or alternatively, TIE (total indirect effect) + INT_{ref} (reference interaction). Even if the exposure does not affect the mediator (such that there is no mediation of X by M), there could be interaction effects that the portion eliminated will capture.

3.9.1.4 The structural equation modeling framework

The language of structural equation modeling is often used by social scientists for conducting mediation analysis. We briefly mention a few basic characteristics of SEM; a thorough and technical treatment of using SEM for mediation analysis is given in Bollen (1987).

SEMs distinguish between observed and latent (unobserved) variables, as well as endogenous and exogenous variables. Endogenous variables are affected by other variables, whereas exogenous variables only affect other variables, without being affected themselves. Furthermore, structural equation modeling makes use of a measurement model and a structural model, from which effects are estimated simultaneously. The measurement model specifies the relationship between latent variables and measured indicator variables, and the structural model specifies the causal relationships among the variables and their covariance structure.

SEM uses path diagrams to graphically display the theoretical causal relationships: rectangles represent observed variables, ovals represent latent variables, straight unidirectional arrows show causal effects between variables, and curved bidirectional arrows represent covariance between two variables. The absence of a link between two variables is important- it represents an assumed *lack* of a causal relationship. The simple mediation model is an example of a structural equation model with observed exposure, mediator, and outcome variables and uncorrelated errors. The exposure is exogenous, the mediator is endogenous with respect to the exposure and exogenous with respect to the outcome, and the outcome variable is endogenous.

3.9.2 Existing approaches to estimating the variance of mediation effects

The methods commonly used for approximating the variance of the estimated indirect effect are based on the multivariate delta method, bootstrapping, and Monte

Carlo simulation. Now that an analytical solution for the variance exists, it is of interest to re-examine the behavior of these approximations. Simulations in Saunders and Blume 2017 shed light on the efficiency gains inherent in avoiding conservative approximations.

3.9.2.1 Delta method approximations

Sobel (1982) proposed the multivariate delta method (or first order Taylor series) approximation to the variance of the indirect effect for the simple mediation model: $\widehat{\text{Var}}(\hat{\alpha}_X\hat{\beta}_M)_{\text{Sobel}} = \hat{\alpha}_X^2 s_{\beta_M}^2 + \hat{\beta}_M^2 s_{\alpha_X}^2$. The second order Taylor series approximation is $\widehat{\text{Var}}(\hat{\alpha}_X\hat{\beta}_M)_{\text{Exact}} = \hat{\alpha}_X^2 s_{\beta_M}^2 + \hat{\beta}_M^2 s_{\alpha_X}^2 + s_{\alpha_X}^2 s_{\beta_M}^2$, but the $s_{\alpha_X}^2 s_{\beta_M}^2$ term tends to be trivially small in practice and is often omitted from the standard error calculation (MacKinnon et al. 1995). An unbiased variance estimator subtracts rather than adds $s_{\alpha_X}^2 s_{\beta_M}^2$ in the equation above (Goodman 1960) but can result in a negative value for the standard error (so it is not recommended) (MacKinnon et al. 2002). The derivation of these three methods assumes independence of α_X and β_M. VanderWeele (2015) has derived delta method variance approximations for several more complex mediation models.

A disadvantage of the above delta method approximations is their reliance on the sampling distribution of $\hat{\alpha}_X\hat{\beta}_M$ being normal. In practice $\hat{\alpha}_X\hat{\beta}_M$ tends to be skewed and highly leptokurtic (MacKinnon et al. 2002; MacKinnon et al. 2004). The *Sobel test* of the indirect effect compares the statistic $(\hat{\alpha}_X\hat{\beta}_M)/\text{SE}(\hat{\alpha}_X\hat{\beta}_M)$ to a standard normal distribution. The Sobel confidence intervals tend to lie to the left of the true value for positive indirect effects, and to the right for negative indirect effects (MacKinnon et al. 1995; Stone and Sobel 1990). As a result, the Sobel test has less power than expected to detect a true indirect effect because the 95% CI will often improperly include zero (MacKinnon et al. 2004).

3.9.2.2 Distribution of the product method

Rather than assuming the product $\hat{\alpha}_X\hat{\beta}_M$ is normally distributed, the *distribution of the product* method assumes $\hat{\alpha}_X \sim \text{N}(\alpha_X, \sigma_{\alpha_X}^2)$ and $\hat{\beta}_M \sim \text{N}(\beta_M, \sigma_{\beta_M}^2)$, and uses the analytical distribution of the product of two normal random variables (MacKinnon et al. 2004). The assumption of normality of the sampling distributions of $\hat{\alpha}_X$ and $\hat{\beta}_M$ is arguably more realistic than the assumption of normality of the distribution of their product. After all, the coefficient estimates properly scaled have a t-distribution. The form of the distribution of the product is highly complex, but values of the function under the null condition that $\hat{\alpha}_X = \hat{\beta}_M = 0$ are tabulated in (Springer and

Thompson 1966). Although there are tables that do not assume both $\hat{\alpha}_X$ and $\hat{\beta}_M$ are zero, for hypothesis testing their use still requires assumptions about their true value, information that is not usually available. The distribution of the product approach relies on large samples for accurate approximation.

3.9.2.3 Bootstrapping

The bootstrap method estimates the variance of mediation effects from the empirical sampling distribution of the estimates. Bootstrapping handles asymmetric sampling distributions better than the delta method and thus improves the accuracy of confidence limits (Preacher and Hayes 2008*b*). There are two approaches to bootstrap resampling in regression, observation resampling and residual resampling. Observation resampling is not model dependent and treats the design matrix as random by resampling cases (i.e. the rows in a design matrix). By contrast, residual resampling treats the design matrix as fixed, is model dependent, and does not maintain the (X,M,Y) association. Bootstrapping cases usually gives a larger estimate of the variance since it allows for more sources of variation from the randomness in the design matrix. As the sample size grows, both methods become similar, assuming the model is correctly specified.

The case-based bootstrap approach to estimating the variance of the indirect effect proceeds as follows. From the data (X, M, Y) of sample size N, draw with replacement N observations to create a bootstrap sample $B^* = (X^*, M^*, Y^*)$. From B^*, estimate the indirect effect using either the product or difference of coefficients approach. Repeat this process $M > 5000$ times. The distribution of the M bootstrap estimates of the indirect effect provides an empirical, nonparametric approximation to the sampling distribution of the indirect effect. Obtain the point estimate and the standard deviation of the indirect effect from the mean and standard error of the M mediation effect estimates, respectively. A 95% percentile confidence interval is constructed from the 2.5^{th} and 97.5^{th} percentiles of the empirical distribution.

Under the three-equation system of the simple mediation model, bootstrapping residuals is complicated. Since there are three equations, one might think to bootstrap the residuals from each model separately. This, however, leads to inconsistent results. To bootstrap residuals, fit the full model and save the fitted values $\hat{Y}$ and residuals e. Sample with replacement from the residuals e to get e^* and a new outcome variable $Y^* = \hat{Y} + e^*$. To estimate the indirect effect using the difference of coefficients approach, re-fit the full and reduced models as follows: $Y^* = \beta_0 + \beta_X X + \beta_M M$ and $Y^* = \gamma_0 + \gamma_X X$ and store $\gamma_x - \beta_x$. Use the distribution of $\gamma_x - \beta_x$ for inference.

To estimate the indirect effect using the product of coefficients approach, fit $M = \alpha_0 + \alpha_X X$ and multiply α_X by β_M from the bootstrapped full model. For the simple mediation model, the residual bootstrap distributions of the estimated indirect effect will be identical under both approaches.

3.9.2.4 The Monte Carlo method

Monte Carlo methods estimate the variance by simulating the sampling distribution of mediation effects (MacKinnon et al. 2004). First, estimate the coefficients used in calculating the indirect effect and their standard errors. For example, if using the product of coefficients approach to estimate the indirect effect, obtain the estimates $\hat{\alpha}_X, \hat{\beta}_M, s^2_{\alpha_X}$ and $s^2_{\beta_M}$. Next, generate $S > 5000$ random samples of the product $\alpha^*_X \beta^*_M$ based on $\alpha^*_X \sim \mathrm{N}(\hat{\alpha}_X, s^2_{\alpha_X})$ and $\beta^*_M \sim \mathrm{N}(\hat{\beta}_M, s^2_{\beta_M})$. To allow $\hat{\alpha}_X$ and $\hat{\beta}_M$ to covary, specify a bivariate normal distribution with some covariance. Obtain the lower and upper confidence limits for the indirect effect from the percentiles of the simulated sampling distribution of the indirect effect. The same general procedure holds for the difference of coefficients approach. We do not recommend using the Monte Carlo approach to estimate the variance unless one has the coefficient and standard error estimates but the raw data are unavailable.

3.9.3 Existing approaches to multiple mediator models

3.9.3.1 Parallel (or single-step) models

The parallel multiple mediator model specifies a separate model for each mediator in which they independently affect the outcome (see panel A of Figure 3.4):

$$\begin{aligned} \mathrm{E}[Y|X, \boldsymbol{M}] &= \beta_0 + \beta_X X + \beta_1 M_1 + \beta_2 M_2 + \cdots + \beta_j M_j \\ \mathrm{E}[M_i|X] &= \alpha_{0i} + \alpha_i X, i = 1, \ldots, j \\ \mathrm{E}[Y|X] &= \beta^*_0 + \beta^*_X X \end{aligned}$$

Analagous to the simple mediation model, the total and direct effects for a unit change in X are given by the coefficients β^*_X and β_X, respectively. This approach assumes no mediators affect each other. The specific indirect effect of X on Y through M_i is quantified as $\alpha_i \beta_i$ (MacKinnon 2008). If the independence of mediators assumption holds, the total indirect effect of X on Y is the sum of the specific indirect effects, $\Sigma_i(\alpha_i \beta_i)$, $i = 1 \ldots j$, which equals $\beta^*_X - \beta_X$ for ordinary least squares regression with continuous M and Y. In this case, the total effect of X on Y can be written as the sum

of the direct effect and all j mediator-specific indirect effects: $\beta_X^* = \beta_X + \Sigma_i(\alpha_i\beta_i), i = 1 \ldots j$.

This approach traditionally uses delta method approximations to estimate the variance of mediator-specific indirect effects and the total indirect effect. The formulas for a two-mediator model are $\hat{\alpha}_1^2 s_{\beta_{M_1}}^2 + \hat{\beta}_{M_1}^2 s_{\alpha_1}^2$, $\hat{\alpha}_2^2 s_{\beta_{M_2}}^2 + \hat{\beta}_{M_2}^2 s_{\alpha_2}^2$, and $s_{\alpha_1}^2 \hat{\beta}_{M_1}^2 + s_{\beta_{M_1}}^2 \hat{\alpha}_1^2 + s_{\alpha_2}^2 \hat{\beta}_{M_2}^2 + s_{\beta_{M_2}}^2 \hat{\alpha}_2^2 + 2\hat{\alpha}_1\hat{\alpha}_2 s_{\beta_{M_1}\beta_{M_2}}$, respectively.

2013 provide a regession-based approach for multiple mediators that is similar in spirit to the single-step multiple mediator model. This approach specifies one regression for the outcome Y (regress Y on X, $\boldsymbol{M}$, and Z), and a separate model for each mediator and each mediator-mediator interaction. Both the natural and controlled direct effects are given by the coefficient for the X in the full model. The natural indirect effect is equal to the sum over the j mediators of the product of the coefficient for the exposure in the model for the ith mediator and the coefficient for the ith mediator in the full model. Including confounders C can lead to compatibility issues between the models for M_i, M_k, and their product M_iM_k. Their alternative inverse probability weighting approach circumvents this issue in settings with mediator-mediator interactions.

3.9.3.2 Weighting approach

The weighting approach does not require modeling the mediators, allows the mediators to affect each other, and can be used for essentially any type of outcome and mediators, although it performs best when the exposure has only a few levels (e.g., binary or discrete) (VanderWeele and Vansteelandt 2013). Obtaining the weights requires fitting several logistic regression models to estimate $\mathrm{P}[X = x]$, $\mathrm{P}[X = x_\circ]$, $\mathrm{P}[X = x|C = c]$, $\mathrm{P}[X = x_\circ|C = c]$. They recommend bootstrapping the variance for both the regression-based and weight-based approaches.

3.9.3.3 Serial models

The *serial multiple mediator model* requires the researcher to specify the order in which the mediators affect each other. Like the single-step multiple mediator model, this approach specifies a separate model for each mediator, although now each mediator depends on those that precede them temporally in the causal chain.

The model is specified as

$$\begin{aligned}
\mathrm{E}[Y|X, \boldsymbol{M}] &= \beta_0 + \beta_X X + \beta_1 M_1 + \beta_2 M_2 + \cdots + \beta_j M_j \\
\mathrm{E}[M_i|X] &= \alpha_{0i} + \alpha_i X + \Sigma_{k=1}^{i-1} \delta_{ik} M_k, i = 2, \ldots, j \\
\mathrm{E}[M_1|X] &= \alpha_{01} + \alpha_1 X \\
\mathrm{E}[Y|X] &= \beta_0^* + \beta_X^* X
\end{aligned}$$

As before, the total indirect effect for a unit change in X is given by $\beta_X^* - \beta_X$. The indirect effect through M_1 is given by $\alpha_1\beta_1$. For $i = 2, \ldots, j$, the indirect effect through M_i only is $\alpha_i\beta_i$, and the indirect effect through $M_1 \to \cdots \to M_{i-1}$ in serial is $\alpha_1 \times \delta_{21} \ldots \delta_{i-1,i-2}\delta_{i,i-1} \times \beta_i$. If these relationships are correctly specified, then the total indirect effect can be written as the sum of the serially mediated indirect effects. For the two-mediator example, we have $\beta_X^* - \beta_X = \alpha_1\beta_1 + \alpha_2\beta_2 + \alpha_1\delta_{21}\beta_2$. The variances of $\hat{\alpha}_1\hat{\beta}_1$ and $\hat{\alpha}_2\hat{\beta}_2$ are estimated using Sobel's formula and $\widehat{\mathrm{Var}}(\hat{\alpha}_1\hat{\delta}_{21}\hat{\beta}_2) = \hat{\alpha}_1^2\hat{\delta}_{21}^2 s_{\beta_2}^2 + \hat{\alpha}_1^2\hat{\beta}_2^2 s_{\delta_{21}}^2 + \hat{\delta}_{21}^2\hat{\beta}_2^2 s_{\alpha_1}^2$ (Hayes 2013). When the ordering of the mediators is known, 2013 provide a potential outcomes approach in which indirect effects are estimated sequentially, similar to the serial multiple mediator model.

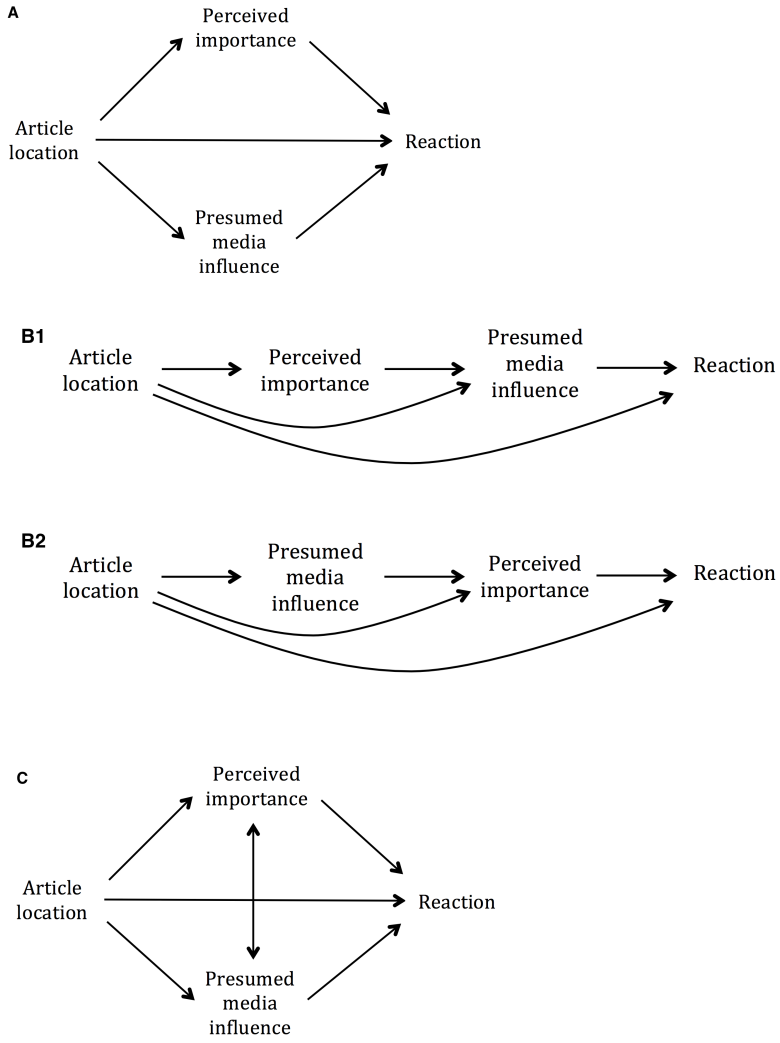


Figure 3.7: Comparison of multiple mediator models. Model A depicts the single-step or parallel multiple mediator model. Models B1 and B2 depict serial multiple mediator models. Model C depicts the proposed framework for assessing mediation with multiple mediators from the fit of a single model. The directions of arrows indicate the assumed causal pathways.

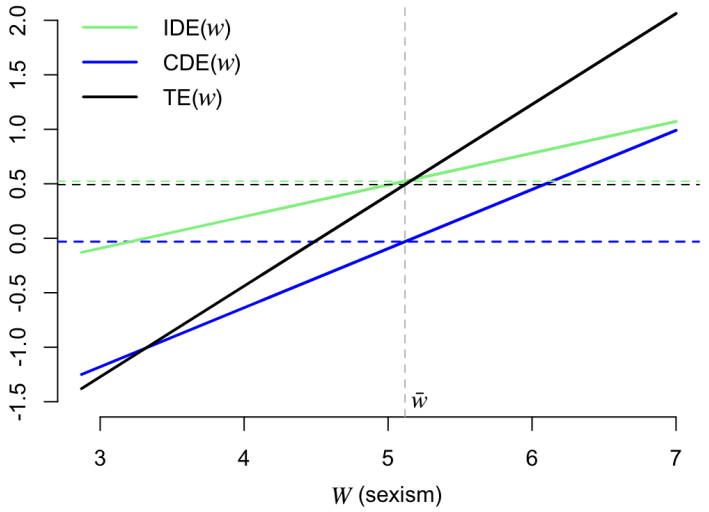


Figure 3.8: Plot of the controlled direct effect (blue line), the indirect effect (green line), and the total effect (black line) for a unit change in the exposure ($x - x_\circ = 1$ since the exposure is binary) from a mediation model with an exposure-moderator interaction. The solid lines show the mediation effects conditional on the moderator W and the dashed lines show the effects given the average value of $W = \bar{w}$. Notice that for any value of W, the conditional direct and indirect effects sum to the total effect.

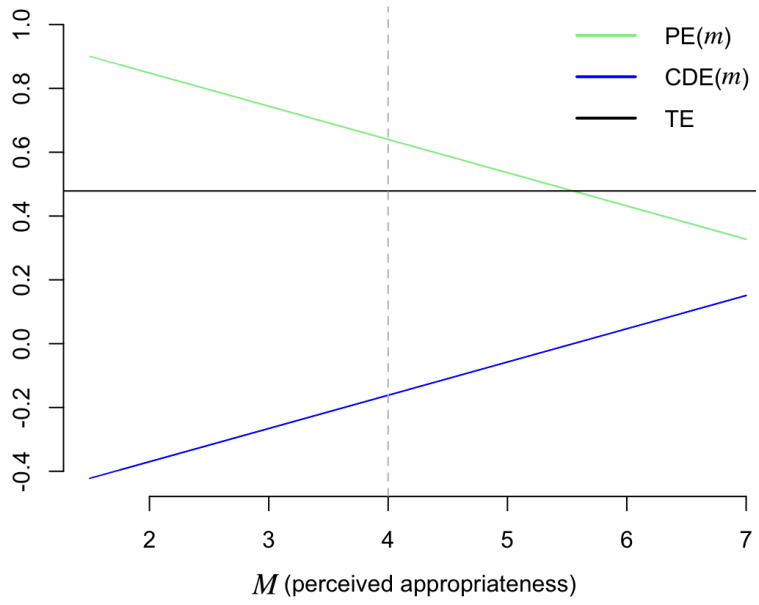


Figure 3.9: Plot of the controlled direct effect (blue line), the portion eliminated (green line), and the total effect (black line) for a unit change in the exposure ($x - x_\circ = 1$ since the exposure is binary) from a mediation model with an exposure-mediator interaction. Notice that the portion eliminated and controlled direct effects sum to the total effect even as they vary as functions of M.

CHAPTER 4

EXTENSIONS TO GENERALIZED LINEAR MODELS

4.1 Introduction

Mediation hypotheses in clinical research, epidemiological studies, and social psychology often involve mediator and outcome variables that are discrete, in which case linear regression models may no longer be appropriate. This chapter extends our single-model framework to generalized linear models (GLMs). Since estimated regression coefficients are maximum likelihood estimates with an approximate normal distribution, our single-model estimator readily extends to GLMs on the link scale. However, extensions to logistic regression require careful attention to issues of collapsibility (Greenland et al. 1999). This chapter focuses on GLMs without exposure-mediator interactions. We use an example from genetic epidemiology to apply our single-model formula, and we compare our results to those obtained from existing approaches.

4.2 Estimating causal mediation effects from GLMs

In this section, we introduce notation, define causal mediation effects for generalized linear models, and note the assumptions needed to make causal inferences. We provide an approximation to and an approach to visualizing the portion eliminated from generalized linear models. We then discuss the distinction between the reduced outcome model and the true marginal model.

4.2.1 The generalized linear model set-up

Let X denote the exposure, M the mediator, C the confounders, and Y the outcome variables of interest. Let $g(\cdot)$ be the link function (e.g., log, logit, probit) for the "full" outcome model (4.1) and the "reduced" outcome model (4.2) which excludes the mediator. Consider the following system:

$$g(\mathrm{E}[Y|X,M,C]) = \beta_0 + h(X)\beta_X + \beta_M M + \beta_C C \tag{4.1}$$

$$g(\mathrm{E}[Y|X,C]) = \beta_0^* + h(X)\beta_X^* + \beta_C^* C \tag{4.2}$$

Let $\eta_F(X,M,C) = \beta_0 + h(X)\beta_X + \beta_M M + \beta_C C$ and $\eta_R(X,C) = \beta_0^* + h(X)\beta_X^* + \beta_C^* C$ be the corresponding linear predictors. Then $\mathrm{E}[Y|x,m,c] = g^{-1}(\eta_F(x,m,c))$ and

$\mathrm{E}[Y|x,c] = g^{-1}(\eta_R(x,c))$. Note that $h(X)$ can be a vector of exposure terms (e.g., $[X, X^2]$ or $[X$,$XC]$).

Causal mediation effects are defined on the link scale (Huang et al. 2004; Kohler et al. 2011; Tchetgen Tchetgen 2014; VanderWeele 2015) and on the expected value scale (Imai, Keele and Yamamoto 2010). Assuming equations (4.1) and (4.2) are correctly specified, the total effect (TE) of X on the link scale of Y is $g(\mathrm{E}[Y|x,c]) - g(\mathrm{E}[Y|x_\circ,c]) = [h(x) - h(x_\circ)]\beta_X^*$ and the controlled direct effect (CDE) of X on the link scale of Y is $g(\mathrm{E}[Y|x,m,c]) - g(\mathrm{E}[Y|x_\circ,m,c]) = [h(x) - h(x_\circ)]\beta_X$. The portion eliminated (PE) is the difference between the total and the controlled direct effects of X, and measures the reduction in the total effect if indirect paths were blocked. The PE on the link scale is $g(\mathrm{E}[Y|x,c]) - g(\mathrm{E}[Y|x_\circ,c]) - (g(\mathrm{E}[Y|x,m,c]) - g(\mathrm{E}[Y|x_\circ,m,c]))$ $= [h(x) - h(x_\circ)](\beta_X^* - \beta_X)$. Without an exposure-mediator interaction, the controlled direct effect is constant for all values of M and thus equals the natural direct effect (NDE). In this setting, subtracting the controlled direct effect from the total effect yields the natural indirect effect (NIE).

The average causal mediation effects from equations (4.1) and (4.2) with $h(X) = X$ are presented on both the link and expected value scales in Table 4.1 and Table 4.2, respectively. Often researchers are interested in reporting mediation effects in terms of odds ratios or risk ratios, as presented in Table 4.1. To obtain the values in Table 4.1, fit the corresponding regression models, take differences on the link scale, then transform to get the association measure (e.g., odds ratio, risk ratio) of interest. Notice that these mediation effects are functions of β_X^* and β_X. As a result, the portion eliminated on the link scale can be estimated using a function of the essential mediation components (EMCs). By contrast, mediation effects on the probability scale are functions of the linear predictors η_F and η_R (and can't be written strictly in terms of the EMCs). To obtain the values in Table 4.2, fit the corresponding regression models, use the transformations $g^{-1}(\eta_F)$ and $g^{-1}(\eta_R)$, and calculate differences in predicted values.

4.2.2 Assumptions

To infer causality from observed data requires the standard causal inference assumptions: consistency, positivity, and exchangeability. Let $Y(x,m)$ denote the potential outcome given the exposure $X = x$ and the mediator $M = m$, and let $M(x)$ denote the potential mediator given the exposure $X = x$. The consistency assumption is that for an individual who actually has exposure $X = x$, the observed Y

and M equal the potential outcomes $Y(x)$ and $M(x)$, respectively. Furthermore, we assume for an individual with exposure $X = x$ and mediator $M = m$, the observed Y equals the potential outcome $Y(x, m)$. The positivity assumption is that everyone has a non-zero probability of having a particular exposure and mediator value: $0 < \mathrm{P}(X = x|C = c)$ and $0 < \mathrm{P}(M = m|X = x, C = c)$ for all $x \in X$, $c \in C$, and $m \in M$.

The exhangeability (no unmeasured confounding) assumptions depicted in Figure 2.2 can be written as i) $Y(x, m) \perp X|C$; ii) $Y(x, m) \perp M|\{X, C\}$; iii) $M(x) \perp X|C$; iv) $Y(x, m) \perp M(x_\circ)|C$ (VanderWeele 2015). When the exposure is randomized, the assumptions of no unmeasured exposure-outcome or exposure-mediator confounders (assumptions i and iii) are considered reasonable, but the possibility of unmeasured mediator-outcome confounders still require consideration. Under assumptions i) and ii), the average total effect and average controlled direct effects are identified. All four assumptions are required to identify natural direct and indirect effects. VanderWeele (2015) and Imai, Keele and Yamamoto (2010) provide sensitivity analysis techniques to quantify how robust conclusions are to the potential violation of the exchangeability assumption.

	Total Effect (TE)	Controlled Direct Effect (CDE)	Portion Eliminated (PE)
	$g(\mathrm{E}[Y\|x]) - g(\mathrm{E}[Y\|x_\circ])$ $= g(\eta_R(x)) - g(\eta_R(x_\circ))$	$g(\mathrm{E}[Y\|x,m]) - g(\mathrm{E}[Y\|x_\circ,m])$ $= g(\eta_F(x,m)) - g(\eta_F(x_\circ,m))$	$g(\mathrm{E}[Y\|x]) - g(\mathrm{E}[Y\|x_\circ])$ $-(g(\mathrm{E}[Y\|x,m]) - g(\mathrm{E}[Y\|x_\circ,m]))$ $= g(\eta_R(x)) - g(\eta_R(x_\circ))$ $-(g(\eta_F(x,m)) - g(\eta_F(x_\circ,m)))$
g=identity	$\mathrm{E}[Y\|x] - \mathrm{E}[Y\|x_\circ]$ $= \beta_X^*(x - x_\circ)$	$\mathrm{E}[Y\|x,m] - \mathrm{E}[Y\|x_\circ,m]$ $= \beta_X(x - x_\circ)$	$\mathrm{E}[Y\|x] - \mathrm{E}[Y\|x_\circ] - (\mathrm{E}[Y\|x.m] - \mathrm{E}[Y\|x_\circ,m])$ $= (\beta_X^* - \beta_X)(x - x_\circ)$
g=log	$\mathrm{RR}^{\mathrm{TE}} = \dfrac{\mathrm{E}[Y\|x]}{\mathrm{E}[Y\|x_\circ]}$ $= e^{\beta_X^*(x-x_\circ)}$	$\mathrm{RR}^{\mathrm{CDE}} = \dfrac{\mathrm{E}[Y\|x,m]}{\mathrm{E}[Y\|x_\circ,m]}$ $= e^{\beta_X(x-x_\circ)}$	$\mathrm{RR}^{\mathrm{PE}} = \dfrac{\mathrm{RR}^{\mathrm{TE}}}{\mathrm{RR}^{\mathrm{CDE}}}$ $= e^{(\beta_X^*-\beta_X)(x-x_\circ)}$
g=logit	$\mathrm{OR}^{\mathrm{TE}} = \dfrac{\mathrm{E}[Y\|x]/(1-\mathrm{E}[Y\|x])}{\mathrm{E}[Y\|x_\circ]/(1-\mathrm{E}[Y\|x_\circ])}$ $= e^{\beta_X^*(x-x_\circ)}$	$\mathrm{OR}^{\mathrm{CDE}} = \dfrac{\mathrm{E}[Y\|x,m]/(1-\mathrm{E}[Y\|x,m])}{\mathrm{E}[Y\|x_\circ,m]/(1-\mathrm{E}[Y\|x_\circ,m])}$ $= e^{\beta_X(x-x_\circ)}$	$\mathrm{OR}^{\mathrm{PE}} = \dfrac{\mathrm{OR}^{\mathrm{TE}}}{\mathrm{OR}^{\mathrm{CDE}}}$ $= e^{(\beta_X^*-\beta_X)(x-x_\circ)}$

Table 4.1: Average causal mediation effects from equations (4.1) and (4.2) defined on the link function scale for commonly used link functions (identity, log, logit). Note that in the setting of no exposure-mediator interaction, the controlled direct effect (CDE) and the natural direct effect (NDE) are eqivalent, so the portion eliminated (PE) and the natural indirect effect (NIE) are also equivalent. For notational simplicity, this table excludes confounders C (without loss of generality).

	Total Effect (TE)	Controlled Direct Effect (CDE)	Portion Eliminated (PE)
	$\mathrm{E}[Y\|x] - \mathrm{E}[Y\|x_\circ]$ $= g^{-1}(\eta_R(x)) - g^{-1}(\eta_R(x_\circ))$	$\mathrm{E}[Y\|x,m] - \mathrm{E}[Y\|x_\circ,m]$ $= g^{-1}(\eta_F(x,m)) - g^{-1}(\eta_F(x_\circ,m))$	$\mathrm{E}[Y\|x] - \mathrm{E}[Y\|x_\circ] - (\mathrm{E}[Y\|x,m] - \mathrm{E}[Y\|x_\circ,m])$ $= g^{-1}(\eta_R(x)) - g^{-1}(\eta_R(x_\circ))$ $-(g^{-1}(\eta_F(x,m)) - g^{-1}(\eta_F(x_\circ,m)))$
g=identity	$\eta_R(x) - \eta_R(x_\circ) = \beta^*_X(x - x_\circ)$	$\eta_F(x,m) - \eta_F(x_\circ,m) = \beta_X(x - x_\circ)$	$\eta_R(x) - \eta_R(x_\circ) - (\eta_F(x,m) - \eta_F(x_\circ,m))$ $= (\beta^*_X - \beta_X)(x - x_\circ)$
g=log	$e^{\eta_R(x)} - e^{\eta_R(x_\circ)}$	$e^{\eta_F(x,m)} - e^{\eta_F(x_\circ,m)}$	$e^{\eta_R(x)} - e^{\eta_R(x_\circ)} - (e^{\eta_F(x,m)} - e^{\eta_F(x_\circ,m)})$
g=logit	$\dfrac{e^{\eta_R(x,c)}}{1+e^{\eta_R(x)}} - \dfrac{e^{\eta_R(x_\circ)}}{1+e^{\eta_R(x_\circ)}}$	$\dfrac{e^{\eta_F(x,m)}}{1+e^{\eta_F(x,m)}} - \dfrac{e^{\eta_F(x_\circ,m)}}{1+e^{\eta_F(x_\circ,m)}}$	$\dfrac{e^{\eta_R(x)}}{1+e^{\eta_R(x)}} - \dfrac{e^{\eta_R(x_\circ)}}{1+e^{\eta_R(x_\circ)}}$ $- \left(\dfrac{e^{\eta_F(x,m)}}{1+e^{\eta_F(x,m)}} - \dfrac{e^{\eta_F(x_\circ,m)}}{1+e^{\eta_F(x_\circ,m)}} \right)$

Table 4.2: Average causal mediation effects from equations (4.1) and (4.2) defined on the expected value scale for commonly used link functions (identity, log, logit). Note that in the setting of no exposure-mediator interaction, the controlled direct effect (CDE) and the natural direct effect (NDE) are eqivalent, so the portion eliminated (PE) and the natural indirect effect (NIE) are also equivalent. For notational simplicity, this table excludes confounders C (without loss of generality).

4.2.3 Estimating the portion eliminated from a single regression equation

Recall that for linear models with normal errors, the *exact* distribution of the maximum likelihood estimates (MLEs) $\hat{\boldsymbol{\beta}}$ is multivariate normal: $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, (D^{\mathrm{T}}D)^{-1}\sigma^2)$ where D is the $n \times p$ design matrix and σ^2 is the error variance. We used properties of the normal distribution to obtain an estimator of the change in the exposure coefficients $\Delta = \beta_X^* - \beta_X$ (the "essential mediation components" or EMCs). Our estimator of the EMCs is $\hat{\Delta} = -\hat{V}_{XM}\hat{V}_M^{-1}\hat{\beta}_M$, where $\hat{V}_{XM}$ is the estimated covariance matrix between $\hat{\beta}_X$ and $\hat{\beta}_M$, and $\hat{V}_M$ is the estimated covariance of the mediator coefficient(s) $\hat{\beta}_M$. The product $-\hat{V}_{XM}\hat{V}_M^{-1}$ actually equals the vector of estimated exposure coefficients $\hat{\boldsymbol{\alpha}}_X$ from the linear mediator model $\mathrm{E}[M|X, C] = \alpha_0 + h(X)\alpha_X + \alpha_C C$. Importantly, our formula for the EMCs automatically incorporates a mediator model that is as flexible as the full model; that is, the form of $h(X)$ specified in the full model is used to capture the exposure-mediator relationship. We showed that $[h(x) - h(x_\circ)]\hat{\Delta}$ is an estimator of the portion eliminated that requires fitting only the full model.

Generalized linear models are fit using iteratively re-weighted least squares, an algorithm which is equivalent to Fisher scoring and leads to maximum likelihood estimates (McCullagh and Nelder 1989). The *large sample* distribution of the MLEs from a GLM is multivariate normal: $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, \phi(D^{\mathrm{T}}WD)^{-1})$, where W is an $n \times n$ weight matrix and ϕ is a dispersion parameter. For logistic regression, the estimated coefficients $\hat{\boldsymbol{\beta}} = (D^{\mathrm{T}}WD)^{-1}D^{\mathrm{T}}WZ$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (D^{\mathrm{T}}WD)^{-1}$, where Z is the fitted log odds, $\log(\hat{\pi}/(1-\hat{\pi}))$. The weight matrix W has diagonal elements $\hat{\pi}(1-\hat{\pi})$ inversely proportional to the variance of the log-odds. For GLMs, the product $-\hat{V}_{XM}\hat{V}_M^{-1}$ is now the weighted least squares estimator $\hat{\boldsymbol{\alpha}}_{\mathrm{WLS}}$ weighted by W, so $-\hat{V}_{XM}\hat{V}_M^{-1}\hat{\beta}_M = \hat{\alpha}_{X.WLS}\hat{\beta}_M$ (Marshall et al. 2002). In other words, we are actually using the product $\hat{\alpha}_{X.WLS}\hat{\beta}_M$ to approximate the EMCs $\hat{\beta}_X^* - \hat{\beta}_X$. Thus, $-\hat{V}_{XM}\hat{V}_M^{-1}\hat{\beta}_M \approx \hat{\Delta}$ is an approximation to the EMCs from generalized linear models, and an approximate single-model estimator of the portion eliminated is $[h(x) - h(x_\circ)][-\hat{V}_{XM}\hat{V}_M^{-1}\hat{\beta}_M]$. Simply exponentiate to obtain the odds ratio $\mathrm{OR}^{\mathrm{PE}}$ or risk ratio $\mathrm{RR}^{\mathrm{PE}}$.

4.2.4 Visualizing mediation effects from GLMs

For generalized linear models (4.1) and (4.2) with link function $g(\cdot)$ and $h(X) = X$, the portion eliminated on the link scale is $[h(x) - h(x_\circ)]\Delta = (\beta_X^* - \beta_X)(x - x_\circ)$. This can be visualized as the distance between the function $h(X)\Delta$ evaluated at x and $x_\circ$. Figure 4.1 panel A shows the portion eliminated from a logistic regression on the log-odds scale (which is a simple function of the EMCs Δ).

The portion eliminated on the expected value scale is $\text{PE}(x, x_\circ, m) = \text{E}[Y|x] - \text{E}[Y|x_\circ] - (\text{E}[Y|x, m] - \text{E}[Y|x_\circ, m]) = g^{-1}(\eta_R(x)) - g^{-1}(\eta_R(x_\circ)) - (g^{-1}(\eta_F(x, m)) - g^{-1}(\eta_F(x_\circ, m))$. This function of predicted probabilities η_F and η_R can't be written strictly in terms of Δ, as it depends on the values of x and m. If we define a function $S(X, M) \equiv \text{E}[Y|X] - \text{E}[Y|X, M] = g^{-1}(\eta_R(X)) - g^{-1}(\eta_F(X, M))$, then $\text{PE}(x, x_\circ, m) = S(x, m) - S(x_\circ, m)$. That is, the portion eliminated comparing exposure levels x and $x_\circ$ for a fixed mediator value m can be visualized as is the difference between the function $S(X, M)$ evaluated at x, $x_\circ$, and m. Figure 4.1 panel B shows the PE from a logistic regression on the probability scale. Figure 4.8 in Appendix 4.8 shows an example of the portion eliminated on the probability scale for several quantiles of the mediator.

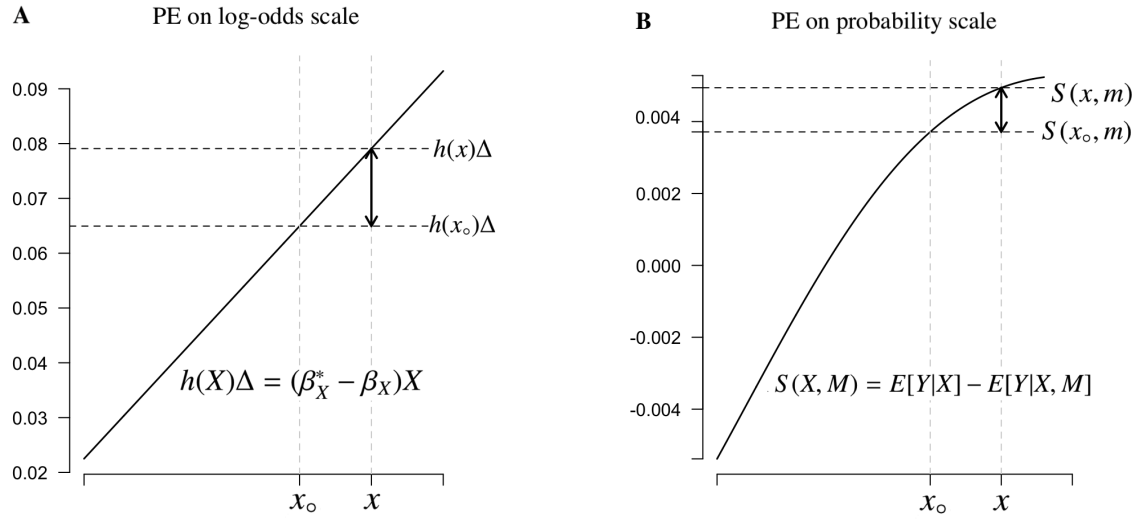


Figure 4.1: Visualizing the portion eliminated as a function of the exposure values $(x, x_\circ)$. Panel A shows the portion eliminated on the log-odds (link) scale as the distance (arrow $\leftrightarrow$) between the line $h(X)\Delta$ evaluated at x and $x_\circ$. Panel B shows the portion eliminated on the probability scale as the distance between the function $S(X, M)$ evaluated at x and $x_\circ$ for a specified value m of the mediator.

4.2.5 What is the true marginal model?

Using the full and reduced outcome models (4.1) and (4.2) to estimate mediation effects requires that both be properly specified. This model specification assumes the mediator is distributed such that the reduced model with mean $\text{E}[Y|X, C]$ has the same distribution as the full model with mean $\text{E}[Y|X, M, C]$. In other words, we

assume a distribution $f(M|X,C)$ such that for link function $g(\cdot)$:

$$\begin{aligned} g^{-1}(\eta_R) &= \mathrm{E}_{M|X,C}\mathrm{E}[Y|X,M,C] \\ &= \int_m \int_y Y f(Y|X,M,C) dy dF(m|x,c) \\ &= \int_m g^{-1}(\eta_F(X,M,C)) dF(m|x,c) \end{aligned}$$

Marginalizing a conditional logistic distribution may lead to a distribution that is no longer logistic.

$$\begin{aligned} &\mathrm{E}_{M|X,C}\mathrm{E}[Y|X,M,C] \\ &= \int_m \int_y Y f(Y|X,M,C) dy dF(m|X,C) \\ &= \int_m \frac{\exp\left(\beta_0 + h(X)\beta_X + \beta_M M + \beta_C C\right)}{1+\exp\left(\beta_0 + h(X)\beta_X + \beta_M M + \beta_C C\right)} dF(m|X,C) \\ &= \exp(\beta_0 + h(X)\beta_X + \beta_C C) \int_m \frac{\exp(\beta_M M)}{1+\exp(\beta_0 + h(X)\beta_X + \beta_M M + \beta_C C)} dF(m|X,C) \\ &= \frac{\exp\left(\beta_0 + h(X)\beta_X + \beta_C C + \lambda(X,C)\right)}{1+\exp\left(\beta_0 + h(X)\beta_X + \beta_C C + \lambda(X,C)\right)} \end{aligned} \tag{4.3}$$

where $\lambda(X,C)$ is a function of X and C. However, analytical and numerical studies by Lin et al. (1998) have shown that for rare events, the reduced logistic model for the mean $\mathrm{E}[Y|X,C]$ written as (4.2) is a reasonable approximation to the marginal model for $\mathrm{E}_{M|X,C}\mathrm{E}[Y|X,M,C]$ given in (4.3). Additionally, they suggest that for a binary mediator, the reduced model will provide an adequate approximation even for non-rare events. The forms of $\lambda(X,C)$ for continuous and binary mediators are given in Appendix 4.8.1 and 4.8.2, respectively.

4.3 Existing approaches

We briefly describe existing approaches to estimating mediation effects from generalized linear models. We focus on the setting in which logistic regression is used to model a binary outcome and a linear model is used for a continuous mediator. We describe methods by Karlson et al. (2012) (the so-called KHB method), the mediation formula, the difference of coefficients, and an approach using the bridge distribution for a logistic model.

4.3.1 The KHB method

We present a latent variable formulation of the mediation system to introduce the KHB method (Karlson et al. 2012). Suppose there is a latent continuous outcome variable Y' and the mediation system is

$$Y' = \gamma_0 + \gamma_X X + \gamma_M M + \gamma_C C + \varepsilon_F \tag{4.4}$$

$$Y' = \gamma_0^* + \gamma_X^* X + \gamma_C^* C + \varepsilon_R \tag{4.5}$$

To derive the logit models, assume $\varepsilon_F = \sigma_F u$ and $\varepsilon_R = \sigma_R \nu$, where u and ν are standard logistic random variables with mean zero and standard deviation $\pi/\sqrt{3}$. Then $\text{sd}(\varepsilon_F) = \sigma_F \pi/\sqrt{3}$ and $\text{sd}(\varepsilon_R) = \sigma_R \pi/\sqrt{3}$. The observed coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ from the logit models are equal to the underlying latent variable model coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ divided by the scale parameters σ_F and σ_R, respectively (Winship and Mare 1983):

$$\text{logit P}[Y = 1|X, M, C] = \beta_0 + \beta_X X + \beta_M M + \beta_C C = \frac{\gamma_0}{\sigma_F} + \frac{\gamma_X}{\sigma_F} X + \frac{\gamma_M}{\sigma_F} M + \frac{\gamma_C}{\sigma_F} C$$

$$\text{logit P}[Y = 1|X, C] = \beta_0^* + \beta_X^* X + \beta_C^* C = \frac{\gamma_0^*}{\sigma_R} + \frac{\gamma_X^*}{\sigma_R} X + \frac{\gamma_C^*}{\sigma_R} C$$

The observed coefficients are influenced by the magnitude of residual variance, so the amount of mediation measured by the change $\beta_X^* - \beta_X$ is masked by the different scale parameters: $\beta_X^* - \beta_X = \frac{\gamma_X^*}{\sigma_R} - \frac{\gamma_X}{\sigma_F} \neq \gamma_X^* - \gamma_X$. Thus, the change in exposure coefficients across logit models with and without the mediator can be due to mediation or to rescaling. Even if the posited mediator is orthogonal to the exposure (such that there is no mediation), if M explains variation in the outcome then the reduced model will have a larger residual error variance ($\sigma_R \geq \sigma_F$) and the coefficient for X may change simply due to rescaling.

Karlson et al. (2012) proposed a solution to this "cross-model coefficient comparability problem" for nonlinear models that are linear in their parameters. The KHB method reparameterizes the full latent variable model (4.4) as

$$Y' = \tilde{\gamma}_0 + \tilde{\gamma}_X X + \tilde{\gamma}_M \varepsilon_{M|X,C} + \tilde{\gamma}_C C + \tilde{\varepsilon}_F \tag{4.6}$$

where $\tilde{\varepsilon}_F = \tilde{\sigma}_F k$, $k \sim$ logistic so that $\text{sd}(\tilde{\varepsilon}_F) = \tilde{\sigma}_F \pi/\sqrt{3}$, and $\varepsilon_{M|X,C}$ is the residual vector from the linear regression $\text{E}[M|X, C] = \alpha_0 + \alpha_X X + \alpha_C C$ with mean zero. Note that $\varepsilon_{M|X,C}$ is orthogonal to X and the exposure coefficients and scale parameters from equations (4.4) and (4.6) are equal (since the fitted values are the same, the residuals

are the same), i.e., $\tilde{\gamma}_X = \gamma^*_X$ and $\varepsilon_F = \tilde{\varepsilon}_F$. Thus, fitting

$$\begin{aligned}\text{logit P}[Y = 1|X, M, C] &= \tilde{\beta}_0 + \tilde{\beta}_X X + \tilde{\beta}_M \varepsilon_{M|X,C} + \tilde{\beta}_C C \\ &= \frac{\tilde{\gamma}_0}{\tilde{\sigma}_F} + \frac{\tilde{\gamma}_X}{\tilde{\sigma}_F} X + \frac{\tilde{\gamma}_M}{\tilde{\sigma}_F} \varepsilon_{M|X,C} + \frac{\tilde{\gamma}_C}{\tilde{\sigma}_F} C\end{aligned}$$

allows one to measure a change in the exposure coefficients that holds the scale and the fit of the error to the assumed distribution constant: $\tilde{\beta}_X - \beta_X = \frac{\tilde{\gamma}_x}{\tilde{\sigma}_F} - \frac{\gamma_X}{\sigma_F} = \frac{\gamma^*_X - \gamma_X}{\sigma_F}$. This approach allows the total, direct, and indirect effects to be identified *relative to the scale* σ_F. A scale-free measure of the *proportion* eliminated is $(\tilde{\beta}_x - \beta_X)/\tilde{\beta}_X = \frac{(\gamma^*_X - \gamma_X)/\sigma_F}{\gamma^*_X/\sigma_F} = (\gamma^*_X - \gamma_X)/\gamma^*_X$.

4.3.2 The mediation formula

A generalization of the product of coefficients approach, the mediation formula (Pearl 2001) can be used for estimating causal mediation effects from any type of model and could be considered the gold standard. The mediation formula requires specification of an outcome model and a mediator model:

$$\begin{aligned}g_1(\text{E}[Y|X, M, C]) &= \beta_0 + \beta_X X + \beta_M M + \beta_C C \\ g_2(\text{E}[M|X, C]) &= \alpha_0 + \alpha_X X + \alpha_c C\end{aligned}$$

VanderWeele (2015) used the mediation formula to derive regression-based solutions to mediation effects from several different combinations of link functions $g_1(\cdot)$ and $g_2(\cdot)$, and Imai, Keele and Tingley (2010) provide R software for their simulation-based approach. For binary outcome Y and continuous mediator M, VanderWeele (2015) provides regression-based solutions from the system

$$\begin{aligned}\text{logit P}[Y = 1|X, M, C] &= \beta_0 + \beta_X X + \beta_M M + \beta_C C \\ \text{E}[M|X, C] &= \alpha_0 + \alpha_X X + \alpha_C C\end{aligned}$$

He defines the average controlled direct effect as $\text{OR}^{\text{CDE}} = \exp\{\beta_X(x - x_\circ)\}$, the average natural direct effect $\text{OR}^{\text{NDE}} = \exp\{\beta_X(x - x_\circ)\}$, and the average natural indirect effect as $\text{OR}^{\text{NIE}} = \exp\{\alpha_X \beta_M (x - x_\circ)\}$. He notes that these definitions hold provided the outcome is rare; if logistic regression is used to model a common outcome, these estimators will be biased.

4.3.3 The difference of coefficients

The difference of coefficients approach fits the "full" and "reduced" outcome models with the same link function. For logistic regression, the mediation system is

$$\text{logit P}[Y = 1|X, M, C] = \beta_0 + \beta_X X + \beta_M M + \beta_C C \tag{4.7}$$

$$\text{logit P}[Y = 1|X, C] = \beta_0^* + \beta_X^* X + \beta_C^* C \tag{4.8}$$

The total effect of X on the log odds of Y is $\beta_X^*(x - x_\circ)$; the controlled direct effect of X on the log odds of Y is given by $\beta_X(x - x_\circ)$; the portion eliminated on the log odds of Y is $(\beta_X^* - \beta_X)(x - x_\circ)$. Exponentiating gives $\text{OR}^{\text{TE}} = \exp\{\beta_X^*(x - x_\circ)\}$, $\text{OR}^{\text{CDE}} = \exp\{\beta_X(x - x_\circ)\}$, and $\text{OR}^{\text{PE}} = \exp\{(\beta_X^* - \beta_X)(x - x_\circ)\}$. We include some properties of the difference of coefficients approach and the circumstances under which it provides a valid measure of mediation in the generalized linear model setting.

Non-collapsibility of the odds ratio: Because odds ratios are not collapsible, it is possible for the exposure effect to differ between the full and reduced models but for there to be no mediation (Greenland et al. 1999). As discussed in Section 4.3.1, the exposure coefficient can change due to both rescaling and mediation.

Rare outcome assumption: Under the rare outcome assumption, the odds ratio approximates a risk ratio (which is collapsible) so one can use the difference of coefficients from a logistic regression model to estimate the portion eliminated. When one uses logistic regression to model a common binary outcome, the odds ratio does not approximate a risk ratio and the aforementioned collapsibility issue arises. However, the rare outcome assumption can be relaxed by using a log-linear regression instead, and the formulas for mediation effects above will have a risk ratio interpretation (VanderWeele 2015).

Conservative estimate of mediation effects: The difference method is conservative for estimating the natural indirect effect, a result proven by Jiang and VanderWeele (2015). Because of the noncollapsibility of odds ratios, including additional covariates in a logistic model tends to increase the magnitude of the exposure coefficient (so $\hat{\beta}_X$ will *overestimate* the magnitude of the natural direct effect). Since $\hat{\beta}_X^*$ is a consistent estimator of the total effect, the difference $\hat{\beta}_X^* - \hat{\beta}_X$ underestimates the NIE. If the direct effect $\hat{\beta}_X$ and difference $\hat{\beta}_X^* - \hat{\beta}_X$ are both positive, then the NIE must be positive and $\text{OR}^{\text{NIE}} > e^{\hat{\beta}_X^* - \hat{\beta}_X}$ (if $\hat{\beta}_X^* - \hat{\beta}_X \leq 0$, the sign of the NIE is inconclusive). Similarly, if $\hat{\beta}_X$ and $\hat{\beta}_X^* - \hat{\beta}_X$ are both negative, we can conclude the NIE must be negative (but if $\hat{\beta}_X^* - \hat{\beta}_X \geq 0$, we cannot draw any conclusions).

4.3.4 The bridge distribution

To obtain closed-form expressions for mediation effects when the outcome is not rare, Tchetgen Tchetgen (2014) replaces the assumption of a normally distributed mediator and a rare outcome with the assumption that the mediator follows a bridge distribution. Using the bridge distribution to marginalize over the full model with a given link function yields a marginal model with the same link function and regression coefficients scaled by ϕ and offset by k. Wang and Louis (2003) showed that for binary outcomes, the necessary bridge density $f(M)$ and constants k and ϕ are unknown but identified as the solutions to the equation

$$\mathrm{E}[Y|X,C] = g^{-1}(\phi\eta_R(X,C)+k) = \int g^{-1}\left(\eta_R(X,M,C)\right) f(M)dm$$

For binary outcomes, there is a *unique* bridge distribution for a given link function (Molenberghs et al. 2013). When g=logit, the bridge density is the unique mediator distribution under which marginalization of the full logistic model with respect to M produces another logistic model, with coefficients scaled by an amount determined by ϕ:

$$B_{\text{logit}}(0,\phi) : f(d) = \frac{sin(\pi\phi)}{cos(\pi\phi)+cosh(\pi d)}, -\infty < d < \infty, 0 < \phi < 1$$

where $cosh(x) = \frac{1}{\exp(x)+\exp(-x)}$. $B_{\text{logit}}(0,\phi)$ is symmetric with variance of $\frac{\pi^2}{3}(\phi^{-2}-1)$. Figure 4.2 shows the normal, logistic, and $B_{\text{logit}}(0,\phi)$ bridge distributions all scaled to have unit variance. Using the result by Wang and Louis (2003), Tchetgen Tchetgen (2014) shows that if $[\varepsilon_M|X,C] \sim B_{\text{logit}}(0,\phi)$, where $\varepsilon_M = M - \mathrm{E}[M|X,C]$, then $\mathrm{E}[Y|X,C] = \text{expit}\left(\tilde{\phi}\left[\beta_0+\beta_X X+\beta_M(\alpha_0+\alpha_X X+\alpha_C C)+\beta_C C\right]\right)$, where $\tilde{\phi} = \left(\beta_M^2(\phi^{-2}-1)+1\right)^{-1/2}$ is a rescaling parameter. It follows that logit $\mathrm{P}[Y=1|X,C] = \beta_0^* + \beta_X^* X + \beta_C^* C$ where $\beta_X^* = \tilde{\phi}(\beta_X+\beta_M\alpha_X)$.

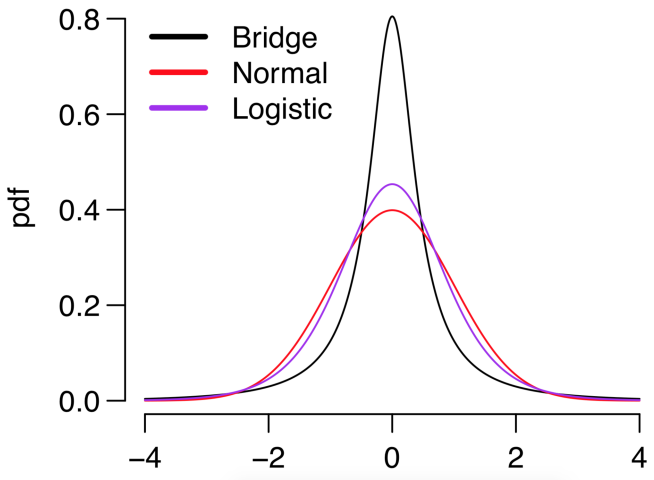


Figure 4.2: Comparing the bridge, normal, and logistic probability density functions, all scaled to have unit variance.

4.4 Comparison of approaches

We discuss the similarities and differences between the aforementioned approaches and our single-model formula, particularly when the outcome is binary and the mediator is continuous. The key difference between using the mediation formula and the difference of coefficients approach is whether or not the total effect is implied or estimated explicitly. The mediation formula approach fits the full model and the mediator model, and uses the sum of the estimated natural direct and indirect effects as the estimated total effect. The difference of coefficients approach, in contrast, uses the maximum likelihood estimate of the exposure effect from the reduced model as the estimated total effect. Note that the rare outcome assumption is needed for both the regression-based solutions given by VanderWeele and the difference of coefficients approach.

For certain models, the difference of coefficients approach is known to coincide with the mediation formula. For correctly-specified models without an exposure-mediator interaction, the difference of coefficients and the mediation formula will agree for linear and log-linear outcome models. For logistic outcome models (4.7) and (4.8), $\beta_X^* - \beta_X \approx \alpha_X \beta_M$ provided the outcome is rare. Under the rare outcome assumption, logit $\mathrm{P}[Y = 1|X, M, C] \approx \log \mathrm{P}[Y = 1|X, M, C]$ and logit $\mathrm{P}[Y = 1|X, C] \approx \log \mathrm{P}[Y = 1|X, C]$.

Thus, $\beta_X^* \approx (\beta_X + \alpha_X\beta_M)$ so $\beta_X^* - \beta_X \approx \alpha_X\beta_M$:

$$
\begin{aligned}
&\exp\{\beta_0^* + \beta_X^* X + \beta_C^* C\} \\
&\approx \mathrm{P}(Y=1|X,C) \\
&= \mathrm{E}_M[\mathrm{P}(Y=1|X,C,M)|X,C] \\
&\approx \mathrm{E}_M[\exp\{\beta_0 + \beta_X X + \beta_M M + \beta_C C\}|X,C] \\
&= \exp\{\beta_0 + \beta_X X + \beta_C C\}\,\mathrm{E}_M[\exp\{\beta_M M\}|X,C] \\
&= \exp\{\beta_0 + \beta_X X + \beta_C C\}\exp\left\{\beta_M(\alpha_0 + \alpha_X X + \alpha_C C) + 0.5\beta_M^2\sigma^2\right\} \\
&= \exp\left\{\beta_0 + 0.5\beta_M^2\sigma^2 + \alpha_0\beta_M + (\beta_X + \alpha_X\beta_M)X + (\beta_C C + \alpha_C\beta_M)C\right\}
\end{aligned}
$$

For the logistic outcome model with no exposure-mediator interaction, the KHB difference of coefficients solution is equivalent to the product of coefficients, $\tilde{\beta}_x - \beta_X = \frac{\tilde{\gamma}_x - \gamma_X}{\sigma_F} = \frac{\alpha_X\gamma_M}{\sigma_F} = \alpha_X\beta_M$, and an equivalent scale-free measure of the proportion eliminated is $\frac{\alpha_X\beta_M}{\beta_X + \alpha\beta_M}$. The difference between the KHB difference $\tilde{\beta}_X - \beta_X = \frac{\gamma_X^*}{\sigma_F} - \frac{\gamma_X}{\sigma_F}$ and the EMCs $\Delta = \beta_X^* - \beta_X = \frac{\gamma_X^*}{\sigma_R} - \frac{\gamma_X}{\sigma_F}$ is $(\tilde{\beta}_X - \beta_X) - (\beta_X^* - \beta_X) = \frac{\gamma_X^*}{\sigma_F} - \frac{\gamma_X^*}{\sigma_R}$. The EMCs can be decomposed into the amount due to rescaling of the coefficient of X and the amount due to mediation: $\beta_X^* - \beta_X = (\beta_X^* - \tilde{\beta}_x) + (\tilde{\beta}_x - \beta_X)$. The ratio $\frac{\tilde{\beta}_x}{\beta_X^*} = \frac{\tilde{\gamma}_x/\tilde{\sigma}_F}{\gamma_X^*/\sigma_R} = \frac{\gamma_X^*/\sigma_F}{\gamma_X^*/\sigma_R} = \sigma_R/\sigma_F$ measures the impact of rescaling, net of mediation (Karlson et al. 2012).

Table 4.3 shows the models that are fit to estimate the total effect, controlled direct effect, and the portion eliminated using the difference of coefficients, the KHB method, the mediation formula, and our formula. The portion eliminated from the KHB method, the mediation formula, and our formula can each be written as some product of coefficients. Our approximation to the EMCs $\beta_X^* - \beta_X$ is $-V_{XM}V_M^{-1}\beta_M = \alpha_{X.WLS}\beta_M$, where $\alpha_{X.WLS}$ is the estimated effect of the exposure on the mediator from a flexible weighted least squares regression of M on X. The mediation formula uses the product $(\mathrm{E}[M|x,c] - \mathrm{E}[M|x_\circ,c])\,\beta_M$, which equals $\alpha_X\beta_M(x - x_\circ)$ for a continuous mediator. If the mediator were binary and modeled with logistic regression, the mediation formula would give $(\mathrm{expit}(\alpha_0 + \alpha_X x + \alpha_C c) - \mathrm{expit}(\alpha_0 + \alpha_X x_\circ + \alpha_C c))\,\beta_M$. The KHB method would still use a linear mediator model to obtain the residuals $\varepsilon_{M|X,C}$ even if the mediator were binary, so the KHB product $\alpha_X\beta_M$ would no longer agree with the mediation formula. For a continuous mediator, the degree to which $\alpha_{X.WLS}$ differs from α_X will determine the departure of our formula from the mediation formula and the KHB method.

	Fitted Models	TE	CDE	PE=TE-CDE
Diff. of coefs	$\text{logit } \mathrm{E}[Y\|X,M] = \beta_0 + \beta_X X + \beta_M M$ $\text{logit } \mathrm{E}[Y\|X] = \beta_0^* + \beta_X^* X$	β_X^*	β_X	$\beta_X^* - \beta_X$
KHB method	$\text{logit } \mathrm{E}[Y\|X,M] = \beta_0 + \beta_X X + \beta_M M$ $\mathrm{E}[M\|X] = \alpha_0 + \alpha_X X \rightarrow \varepsilon_{M\|X}$ $\text{logit } \mathrm{E}[Y\|X,M] = \tilde{\beta}_0 + \tilde{\beta}_X X + \tilde{\beta}_M \varepsilon_{M\|X}$	$\tilde{\beta}_X$	β_X	$\tilde{\beta}_X - \beta_X$ $= \alpha_X \beta_M$
Mediation formula	$\text{logit } \mathrm{E}[Y\|X,M] = \beta_0 + \beta_X X + \beta_M M$ $\mathrm{E}[M\|X] = \alpha_0 + \alpha_X X$	$\beta_X + \alpha_X \beta_M$	β_X	$\alpha_X \beta_M$
Our formula	$\text{logit } \mathrm{E}[Y\|X,M] = \beta_0 + \beta_X X + \beta_M M$	$\beta_X - V_{XM} V_M^{-1} \beta_M$ $= \beta_X + \alpha_{X.WLS} \beta_M$	β_X	$-V_{XM} V_M^{-1} \beta_M$ $= \alpha_{X.WLS} \beta_M$

Table 4.3: Comparing mediation estimators for a binary outcome Y, continuous mediator M, and a unit change in the exposure X (so that $x - x_\circ = 0$) using the difference of coefficients, the KHB method, the mediation formula, and our formula. The columns show the fitted models, the total effect (TE), controlled direct effect (CDE), and the portion eliminated (TE-CDE). For notational simplicity, this table excludes confounders C (without loss of generality).

4.5 Application to genetic epidemiology

Background: Genetic variants on chromosome 15q25.1 have been associated with lung cancer as well as smoking behavior in genome-wide association studies (Amos et al. 2008). There has been discussion regarding whether the effect of genetic variants on lung cancer is direct or if it operates through pathways related to smoking. To investigate whether smoking mediates the relationship between genetics and risk of lung cancer in African Americans, we analyze genotyping data on chromosome 15q25 from 4253 subjects in the African American Lung Cancer Consortium. Table 4.4 provides summary statistics of the dataset.

	N	No cancer $N = 2843$	Cancer $N = 1410$
Sex	4253		
Male		42% (1202)	50% (698)
Female		58% (1641)	50% (712)
Age	4253	49 57 66 (57 ±13)	54 61 69 (61 ±11)
Patient status (case/control)	4253		
Control		100% (2843)	0% (0)
Case		0% (0)	100% (1410)
Lung cancer histology	1410		
Adenocarcinoma			45% (632)
Large cell			2% (34)
Other			22% (306)
Small Cell			6% (81)
Squamous Cell			24% (337)
Unknown/Missing			1% (20)
Smoking category (never/former/current)	4209		
Never		43% (1208)	8% (110)
Former		30% (838)	37% (520)
Current		27% (755)	55% (778)
Smoking category (never/ever)	4209		
Never Smoker		43% (1208)	8% (110)
Ever Smoker (current or former)		57% (1593)	92% (1298)
Pack years	4051	0.0 0.4 15.5 (10.3 ±17.4)	3.5 22.0 42.0 (28.4 ±29.3)
Heavy smoker (yes/no)	4051		
Heavy smoker		19% (520)	53% (729)
Non-heavy smoker		81% (2149)	47% (653)
Study site	4253		
MDA		35% (1001)	26% (373)
NCI_UMD		13% (375)	15% (208)
SCCS		12% (338)	12% (168)
UCSF		23% (661)	23% (325)
WayneState		16% (468)	24% (336)
Estimated European global ancestry	4253	0.11 0.17 0.25 (0.19 ±0.12)	0.10 0.16 0.25 (0.19 ±0.13)
Estimated African global ancestry	4253	0.75 0.83 0.89 (0.81 ±0.12)	0.75 0.84 0.90 (0.81 ±0.13)

$_a\, b\, _c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
$x \pm s$ represents $\bar{X} \pm 1$ SD.N is the number of non–missing values. Numbers after percents are frequencies.

Table 4.4: Summary statistics of variables in the genetic epidemiology example

Assumptions: The exposure X is the number of copies of the minor allele at a particular SNP on gene region 15q25, the mediator M is smoking measured in pack-years, and the outcome Y is the presence or absence of lung cancer. We include

subjects' age, sex, study site, and estimated global African ancestry as confounding variables $\boldsymbol{C}$. The directed acyclic graph in Figure 4.3 presents the assumed causal relationships. We assume that conditional on $\boldsymbol{C}$, there is no unmeasured confounding of the exposure-outcome, mediator-outcome, exposure-mediator relationships, and that no effect of the exposure is itself a mediator-outcome confounder. Since the exposure is a genetic variant among African Americans, assumptions i) and iii) are believable. Global African ancestry can be associated with both the genetic marker and disease frequency and has also been shown to be correlated with environmental and biomedical covariates, so we consider it a surrogate for unmeasured environmental exposures (Ziv and Burchard 2003). We assume adjusting for study site and global African ancestry are sufficient surrogates for SES, neighborhood of residence, or other smoking-lung cancer confounders, so that assumption ii) is met. We also assume the genetic variants do not affect mediator-outcome confounders (like neighborhood of residence). We use the following logistic regression model (age is splined and we use the log(1+x) transformation for pack-years):

$$\begin{aligned}\text{logit P}(\texttt{lung cancer}|X, M, C) =& \beta_0 + \beta_X \texttt{SNP} + \beta_M \texttt{pack-years} + \beta_{C1} \texttt{age} + \\ & \beta_{C2} \texttt{sex} + \beta_{C3} \texttt{site} + \beta_{C4} \texttt{Afr}\end{aligned}$$

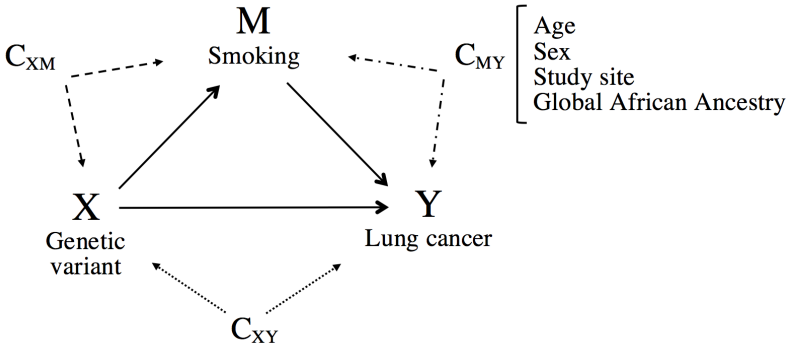


Figure 4.3: A directed acyclic graph for the genetic epidemiology example looking at whether the effect of a genetic variant (X) on lung cancer (Y) is mediated by smoking (M). We assume that conditional on $\boldsymbol{C}$ = {age, sex, study site, and global African ancestry}, we have controlled for i) exposure-outcome confounding (C_{XY}), ii) mediator-outcome confounding (C_{MY}), iii) exposure-mediator confounding (C_{XM}), and iv) that no effect of the exposure is itself a mediator-outcome confounder.

Interpretation at a particular SNP: Consider the effect of increasing copies of the minor allele, such that $x - x_\circ = 1$. The total effect $\text{OR}^{\text{TE}} = \frac{\text{odds E}[Y(1)|c]}{\text{odds E}[Y(1)|c]}$ is the odds ratio for lung cancer comparing those with and without an additional copy of the

minor allele. The controlled direct effect $\text{OR}^{\text{CDE}} = \frac{\text{odds E}[Y(1,m)|c]}{\text{odds E}[Y(0,m)|c]}$ is the odds ratio for lung cancer comparing those with and without an additional copy of the minor allele if pack-years were fixed at a certain level m. The portion eliminated $\text{OR}^{\text{PE}} = \text{OR}^{\text{TE}}/\text{OR}^{\text{CDE}}$ is the portion of the total effect of increasing copies of the minor allele on lung cancer that would be eliminated if we were to fix pack-years to the same value m for all persons. The natural direct effect $\text{OR}^{\text{NDE}} = \frac{\text{odds E}[Y(1,M(0))|c]}{\text{odds E}[Y(0,M(0))|c]}$ is the odds ratio for lung cancer comparing those with and without an additional copy of the minor allele *if pack-years were what it would have been without an additional copy of the minor allele*. The natural indirect effect $\text{OR}^{\text{NIE}} = \text{OR}^{\text{TE}}/\text{OR}^{\text{NDE}} = \frac{\text{odds E}[Y(1,M(1))|c]}{\text{odds E}[Y(1,M(0))|c]}$ is the odds ratio for lung cancer for those with an additional copy of the minor allele *comparing what would happen if pack-years were set to what it would have been with and without the additional copy of the minor allele.*

The natural indirect effect measures the reduction in the total effect of increasing copies of the minor allele on risk of lung cancer that would occur if we were to disable the pathway from SNP to smoking (i.e., setting pack-years to the value it would have had without the additional copy of the minor allele). Although the natural indirect effect has an interesting metaphysical interpretation, in this example it does not correspond to an intervention that could actually be carried out in practice. The portion eliminated, on the other hand, measures the reduction in the total effect that would be obtained by intervening to fix smoking to $M = m$ pack-years for all subjects. The portion eliminated tells us how much of the exposure's effect on lung cancer we could block by intervening on smoking.

Results: We estimated the odds ratios for the controlled direct effect, the total effect, and the portion eliminated at each SNP using the difference of coefficients, the KHB method, the mediation formula, our formula, and a "simple" product of coefficients. The simple product of coefficients estimates the exposure-mediator effect α_X from a mediator model $\text{E}[\texttt{pack-years}|X] = \alpha_0 + \alpha_X\texttt{SNP}$ that excludes the confounders. The KHB method, the mediation formula, and our formula estimate α_X from the mediator model $\text{E}[\texttt{pack-years}|X, \boldsymbol{C}] = \alpha_0 + \alpha_X\texttt{SNP} + \boldsymbol{\alpha_C}\boldsymbol{C}$; however, while the KHB method and the mediation formula use the ordinary least squares estimate of α_X, our formula uses the weighted least squares estimate as described in Section 4.2.3.

The estimated controlled direct effect odds ratios shown in Figure 4.4 agree across methods because they use the same full model. On the other hand, the portion eliminated odds ratios shown in Figure 4.5 differ across methods. We do not include the mediation formula results because they agree perfectly with the KHB results in

this example. Figure 4.6 provides pairwise comparisons of the estimated OR^{PE}.

To assess how each method performs compared to the mediation formula (which could be considered the gold standard), it is equivalent to compare the results to the KHB method. We present the estimated mediation effect odds ratios and the estimated α_X for four SNPs in Figure 4.7, and we interpret some of the mediation effects with varying magnitudes and directions in Appendix 4.8.3. Notice that our sweep formula appears to coincide with the difference of coefficients, which is not surprising since this is what our formula is approximating. The simple product performs similarly to the KHB method, which suggests that excluding confounders from the mediator model did not significantly change the estimate of α_X. Since the exposure is increasing copies of the minor allele (which could be considered randomly assigned), perhaps it is unlikely for there to be influential confounders of the genetic-smoking relationship.

4.6 Future directions

It is of interest to further develop our single-model framework so that it provides an appropriate estimator of the portion eliminated from generalized linear models. For mediation analyses involving large datasets with multiple exposure variables of interest, implementing the simulation-based approach by Imai, Keele and Yamamoto (2010) can be computationally intensive, so a single-model framework is desirable. Directions for future research include using simulations to assess the degree to which our estimator $-\hat{V}_{XM}\hat{V}_M^{-1}\hat{\beta}_M$ approximates the EMCs $\beta_X^* - \beta_X$. We'd also like to understand how the bias of the difference of coefficients estimator of the portion eliminated compares to that of the mediation formula. Finally, we are exploring how to extend our double-sweep approach to generalized linear models.

4.7 Acknowledgements

Example data was obtained from the African American Lung Cancer Consortium, supported by grants CA060691, CA87895, PC35145 and CA22453, CA68578, CA140388, ES007784, ES06717, CA55769, CA127219, CA1116460S1, CA1116460, CA121197, CA141716, CA121197S2, CPRIT RP100443, CA148127, University Cancer Foundation, Duncan Family institute, Center for Community, Implementation, and Dissemination Research Core, and by grants to the component studies: PLCO and the Maryland Studies (NCI Intramural Research Program), LAAPC (Cancer Research Fund 99-00524V-10258), and both MEC and LAAPC (PC35139, DP000807).

4.8 Appendix

4.8.1 Form for $\lambda(X, C)$ when M is normal

When the mediator M is normally distributed with conditional mean $\mu_{M|X,C}$,

$$\lambda(X,C) = \beta_M \mu_{M|X,C} + 0.5\beta_M^2 + \log \frac{\int \left(1+\exp(\beta_0+\beta_X X+\beta_M m+\beta_C C+\beta_M^2)\right)^{-1} \exp(-(m-\mu_{M|X,C})^2/2)dm}{\int \left(1+\exp(\beta_0+\beta_X X+\beta_C C+\beta_M m)\right)^{-1} \exp(-(m-\mu_{M|X,C})^2/2)dm}$$

For a normal mediator, if the event is rare or if the magnitude of β_M is small, $\lambda(X, C) \approx \beta_M \mu_{M|X,C} + 0.5\beta_M^2$ and $\beta_X^* \approx \beta_X + \alpha_X \beta_M$:

$$\begin{aligned}
\text{logit E}[Y|X,C] &= \beta_0 + h(X)\beta_X + \beta_C C + \lambda(X,C) \\
&\approx \beta_0 + h(X)\beta_X + \beta_C C + \beta_M \mu_{M|X.C} + 0.5\beta_M^2 \\
&= \beta_0 + h(X)\beta_X + \beta_C C + \beta_M(\alpha_0 + h(X)\alpha_X + \alpha_C C) + 0.5\beta_M^2 \\
&= (\beta_0 + \beta_M\alpha_0 + 0.5\beta_M^2) + h(X)(\beta_X + \beta_M\alpha_X) + (\beta_C C + \beta_M \alpha_C)C \\
&= \beta_0^* + h(X)\beta_X^* + \beta_C^* C
\end{aligned}$$

4.8.2 Form for $\lambda(X, C)$ when M is binary

When the mediator M has a binomial distribution with conditional mean $P_{M|X,C}$,

$$\lambda(X,C) = \log \frac{e^{\beta_M}\left(1+e^{\beta_0+\beta_X X+\beta_C C}\right)P_{M|X,C} + \left(1+e^{\beta_0+\beta_X X+\beta_M+\beta_C C}\right)(1-P_{M|X,C})}{\left(1+e^{\beta_0+\beta_X X+\beta_C C}\right)P_{M|X,C} + \left(1+e^{\beta_0+\beta_X X+\beta_M+\beta_C C}\right)(1-P_{M|X,C})}$$

If the event is rare or if the magnitude of β_M is small, $\lambda(X, C)$ is approximately equal to $\log\left(e^{\beta_M} P_{M|X,C} + (1 - P_{M|X,C})\right)$ and

$$\begin{aligned}
\text{logitE}[Y|X,C] &= \beta_0 + h(X)\beta_X + \beta_C C + \lambda(X,C) \\
&\approx \beta_0 + h(X)\beta_X + \beta_C C + \log\left(e^{\beta_M} P_{M|X,C} + (1-P_{M|X,C})\right) \\
&= \beta_0 + h(X)\beta_X + \beta_C C + \log\left(1 + \frac{e^{\alpha_0+h(X)\alpha_X+\alpha_C C}}{1+e^{\alpha_0+h(X)\alpha_X+\alpha_C C}}(e^{\beta_M}-1)\right)
\end{aligned}$$

4.8.3 Interpreting mediation effects from logistic regression

We interpret some of the results shown in Figure 4.7. These are not meant to be interpreted scientifically, but rather to demonstrate how one would interpret mediation effect odds ratios with varying magnitudes and directions. The following

inferences assume we've correctly specified our models and that the 95% confidence intervals for these effects exclude the null OR=1.

For rs17184851, the estimated total effect odds ratios are greater than 1, and the estimated controlled direct effects are greater than the estimated total effects. As a result, the estimated portion eliminated odds ratios are less than 1. This implies that increasing copies of the minor allele at rs17184851 directly increases risk of lung cancer, but that increasing copies of the minor allele has a negative indirect effect through smoking. Since we know that smoking is positively associated with lung cancer, this suggests that increasing copies of the minor allele at rs17184851 is associated with reduced pack-years. This is confirmed by the estimated odds ratios for the exposure-mediator effect being less than 1.

For rs1509557, the directions of the results are not consistent across methods. The difference of coefficients and our sweep formula give estimated TE and CDE odds ratios less than 1, and the magnitudes of the estimated CDE odds ratios are larger than the TE odds ratios. Thus, the estimated PE odds ratios are greater than 1. This suggests that increasing copies of the minor allele at rs1509557 directly reduces risk of lung cancer, but that increasing copies of the minor allele has a positive indirect effect through smoking. Since smoking is positively associated with lung cancer, this suggests that increasing copies of the minor allele at rs1509557 is associated with increased pack-years. This is supported by the sweep estimate of the odds ratio for the exposure-mediator effect being greater than 1. However, the magnitudes of the negative CDEs are greater than the magnitudes of the positive PEs, leading to estimated total effects of increasing copies of the minor allele that appear protective for lung cancer.

Now we interpret the results at rs1509557 from the KHB method and the simple product approach. The estimated TE and CDE odds ratios are still less than 1, but now the magnitudes of the estimated TE odds ratios are larger than the CDE odds ratios. As a result, the portion eliminated estimates are also less than 1, suggesting that increasing copies of the minor allele are associated with reduced pack-years, which is supported by the estimated odds ratios for the exposure-mediator effect being less than 1. These results imply that both the direct and indirect effects of increasing copies of the minor allele at rs1509557 are protective. Results for the other two SNPs could be interpreted in a similar manner.

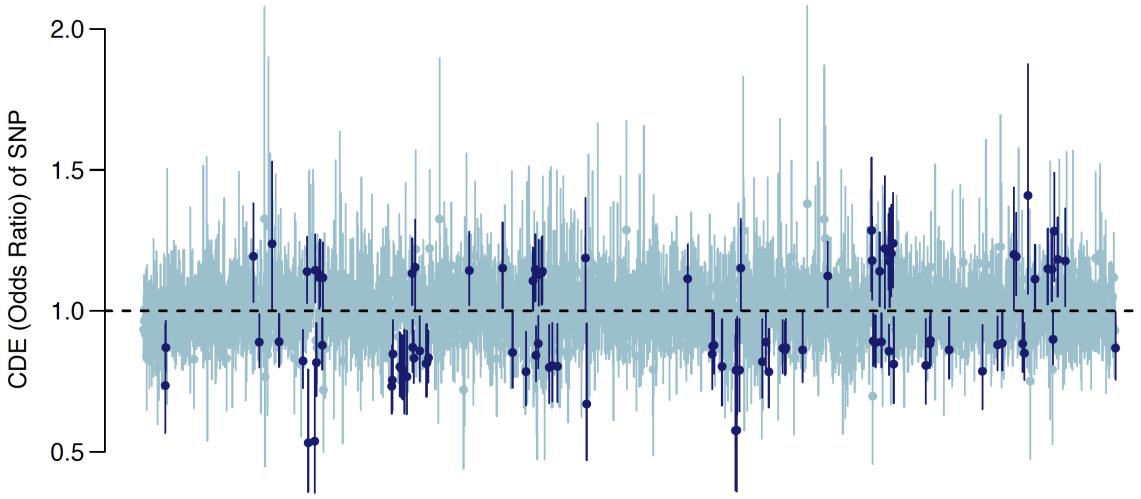


Figure 4.4: The controlled direct effect odds ratios of SNPs on chromosome 15q25.

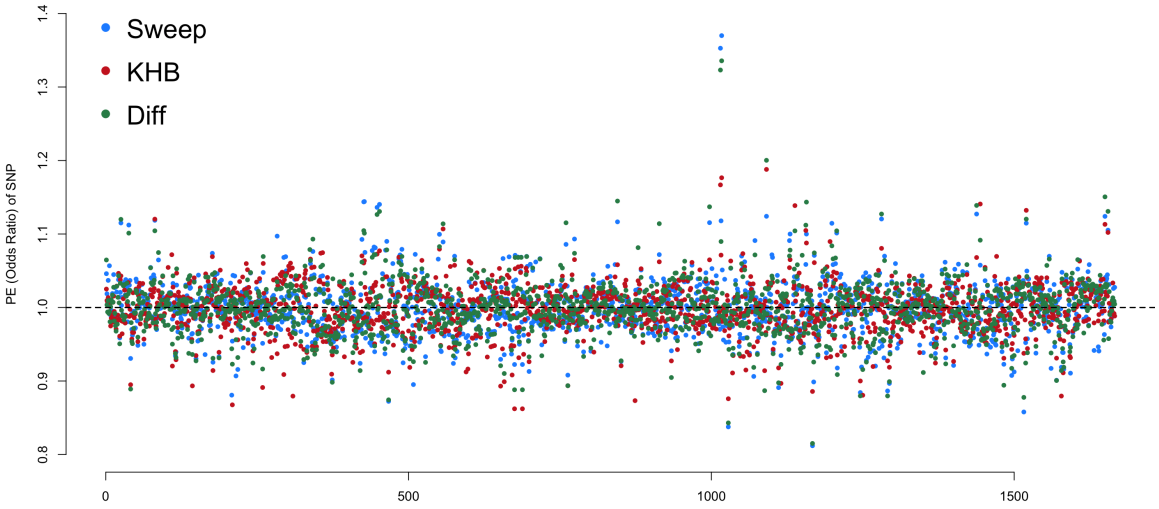


Figure 4.5: The portion eliminated odds ratios of SNPs on chromosome 15q25: comparing results obtained using the sweep formula, the KHB method, and the difference of coefficients.

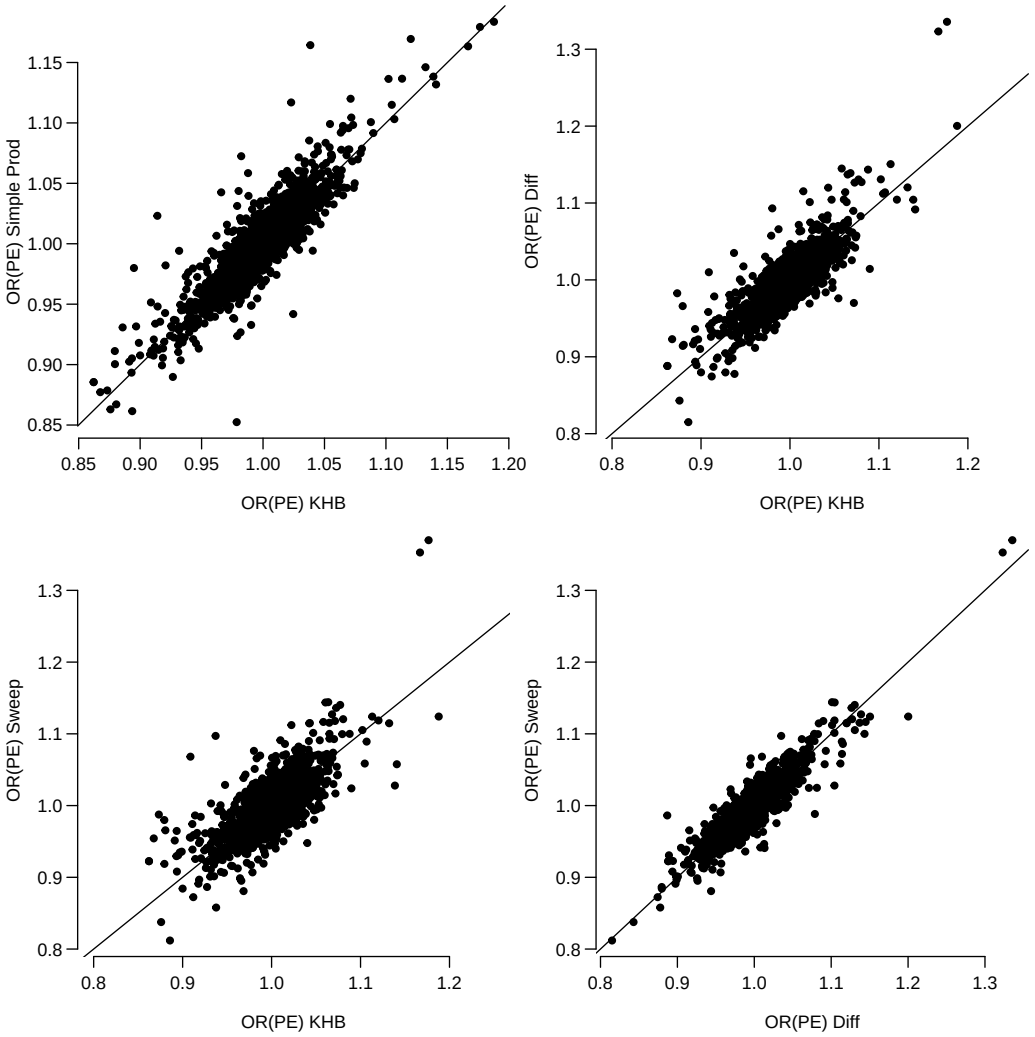


Figure 4.6: Pairwise comparisons of the portion eliminated obtained using the KHB method, the product of coefficients with a simple mediator model (excluding confounders), the difference of coefficients, and our sweep formula.

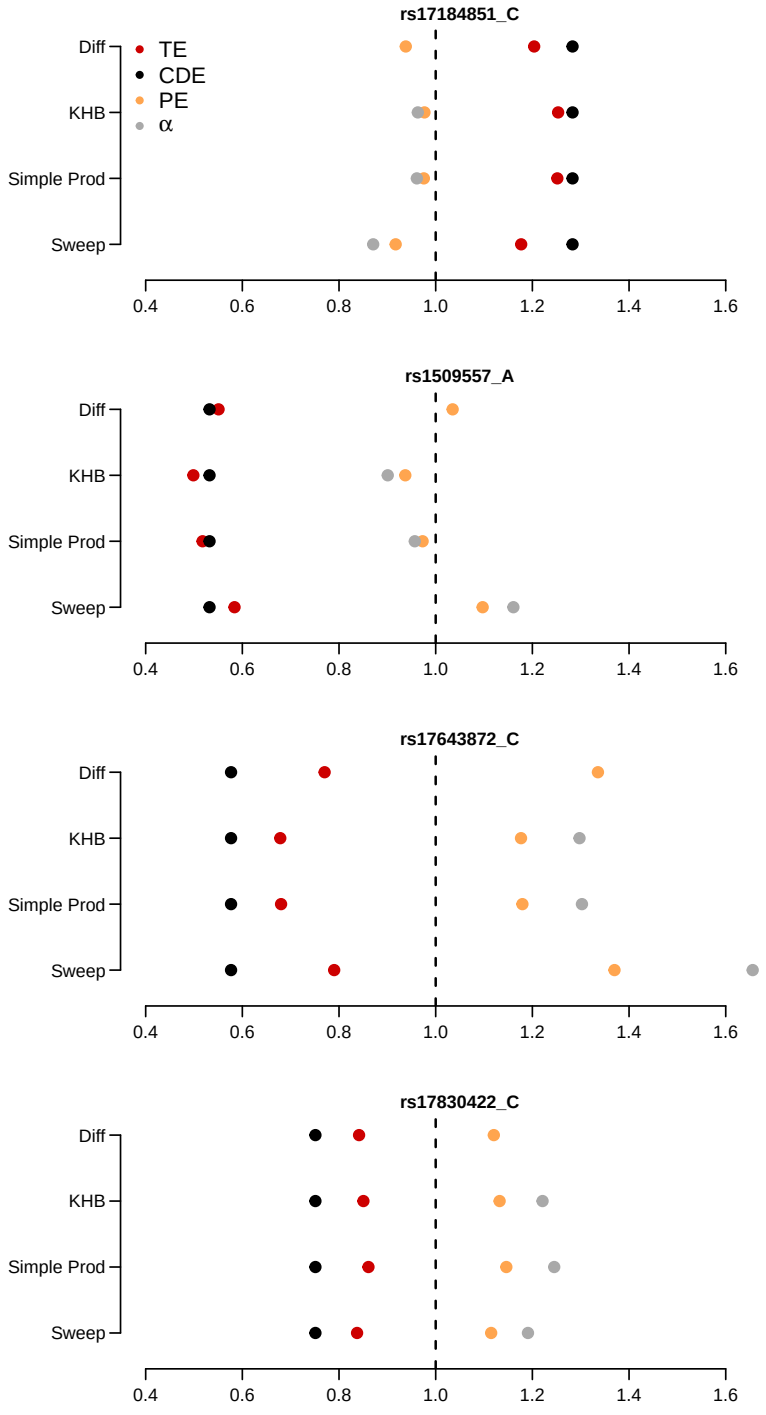


Figure 4.7: Comparison of mediation effect odds ratios (total effect, controlled direct effect, portion eliminated). We compare the estimated odds ratios at four SNPs obtained using the difference of coefficients, the KHB method, the simple product of coefficients, and our sweep formula. The estimated odds ratios for the exposure-mediator effect are represented by α.

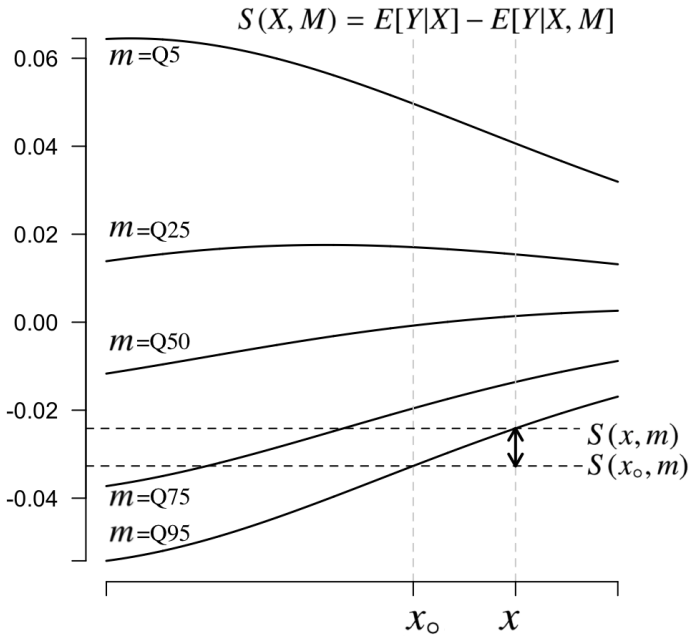


Figure 4.8: The portion eliminated on the probability scale can be visualized as the distance between the function $S(X, M)$ evaluated at x and $x_\circ$. We show results for various values of the mediator (the 5th, 25th, 50th, 75th, and 95th percentiles of M).

CHAPTER 5

CONCLUSION

The statistical literature abounds with methods for measuring mediation in the simple setting of one exposure, one mediator, and one outcome. Our proposed single-model framework is straightforward to implement for simple and sophisticated mediation models that may involve nonlinearities, interactions, splines, and multiple mediators. Rather than fitting several separate regression equations, one can use our formula for the EMCs to estimate mediation effects and their variance from the fit of the full regression equation. This approach allows for mediation analysis with a straightforward application of regression modeling tools - e.g., penalization procedures, multiple imputation, and cross-validation. One simply applies these techniques to the full model and their impact is automatically incorporated in the mediation functionals.

In addition to providing a useful formula for estimating the portion eliminated and its variance, we drew connections between existing approaches, noting their similarities and the settings in which they agree and diverge. The standard approach in the causal inference literature is to use the implied total effect that results from fitting the full outcome model and the mediator model. We say “implied" here because the marginal outcome model is not actually fit to the data. Instead, the sum of the estimated natural direct and indirect effects is used as an estimate of the total effect. In contrast, our approach estimates the total effect that would be directly obtained from *fitting* the marginal model. Importantly, the sum of the estimated natural direct and indirect effects does not necessarily equal the estimated total effect from the fitted marginal model, an unexpected finding. One explanation is that several different systems of equations will yield the same reduced model, but a specific reduced model is implied once the outcome model and mediator model are fit. This is an interesting finding that merits further study.

We also showed how the proposed framework extends to generalized linear models. We presented mediation effects defined on the link scale, using the essential mediation components, and on the expected value scale, using a transformation of the linear predictors. We discussed the distinguishing features of the difference of coefficients, the KHB method, the mediation formula, and our formula. We compared results of a mediation analysis applying each of these methods to genotyping data, noted the ways in which the results are similar and dissimilar, and provided interpretations of

the mediation effect odds ratios.

The proposed framework can be viewed as having two key steps: first, estimation of one well-specified full outcome model and second, estimation of mediation functionals from that model. For mediation analysis with large datasets, efficiency gains from using our estimator of the portion eliminated and its variance can be substantial. The single-model regression framework for mediation analysis imparts conceptual clarity and provides practical and computational advantages. Our directions for future research include further development of the single-model framework so that it provides an appropriate estimator of the portion eliminated from generalized linear models.

REFERENCES

Alwin, D. F. and Hauser, R. M. (1975), The Decomposition of Effects in Path Analysis, *American Sociological Review 1975* **40**, 37–47.

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., Sullivan, K., Matakidou, A., Wang, Y., Mills, G., Doheny, K., Tsai, Y.-Y., Chen, W. V., Shete, S., Spitz, M. R. and Houlston, R. S. (2008), Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1, *Nature Genetics* **40**(5), 616–622.

Baron, R. M. and Kenny, D. A. (1986), The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations, *Journal of Personality and Social Psychology* **51**, 1173–1182.

Bollen, K. A. (1987), Total, Direct, and Indirect Effects in Structural Equation Models, *Sociological Methodology* **17**, 37–69.

Bollen, K. A. and Pearl, J. (2013), Eight Myths About Causality and Structural Equation Models, *in* S. L. Morgan, ed., 'Handbook of Causal Analysis for Social Research', Springer, 301–328.

Cole, D. A. and Maxwell, S. E. (2003), Testing Mediational Models With Longitudinal Data: Questions and Tips in the Use of Structural Equation Modeling., *Journal of Abnormal Psychology* **112**, 558–577.

de Heus, P. (2012), R squared effect-size measures and overlap between direct and indirect effect in mediation analysis, *Behavior Research Methods* **44**, 213–221.

Fairchild, A. J., MacKinnon, D. P., Taborga, M. P. and Taylor, A. B. (2009), R squared effect-size measures for mediation analysis, *Behavior Research Methods* **41**(2), 486–498.

Fisher, R. A. (1915), Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika* .

Fritz, M. S. and MacKinnon, D. P. (2007), Required Sample Size to Detect the Mediated Effect, *Psychological Science* **18**(3), 233–239.

Galles, D. and Pearl, J. (1998), An Axiomatic Characterization of Causal Counterfactuals, *Foundations of Science* **1**, 151–182.

Garcia, D. M., Schmitt, M. T., Branscombe, N. R. and Ellemers, N. (2010), Women's reactions to ingroup members who protest discriminatory treatment: The importance of beliefs about inequality and response appropriateness, *European Journal of Social Psychology* **40**(5), 733–745.

Gelfand, L. A., Mensinger, J. L. and Tenhave, T. (2009), Mediation Analysis: A Retrospective Snapshot of Practice and More Recent Directions, *J Gen Psychol* **136**(2), 153–176.

Goodman, L. A. (1960), On the Exact Variance of Products, *Journal of the American Statistical Association* **55**, 708–713.

Goodnight, J. H. (1979), A Tutorial on the SWEEP Operator, *The American Statistician* **33**, 149–158.

Greenland, S., Robins, J. M. and Pearl, J. (1999), Confounding and Collapsibility in Causal Inference, *Statistical Science* **14**, 29–46.

Gunzler, D., Chen, T., Wu, P. and Zhang, H. (2013), Introduction to mediation analysis with structural equation modeling., *Shanghai archives of psychiatry* **25**(6), 390–394.

Harrell, F. E. (2015), *Regression Modeling Strategies*, 2 edn, Springer.

Hayes, A. F. (2009), Beyond Baron and Kenny: Statistical Mediation Analysis in the New Millennium, *Communication Monographs* **76**(4), 408–420.

Hayes, A. F. (2013), *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, Guilford Press.

Hayes, A. F. (2017), *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, 2 edn, Guilford Press.

Holland, P. (1986), Statistics and Causal Inference, *Journal of the American Statistical Association* **81**, 945–960.

Huang, B., Sivaganesan, S., Succop, P. and Goodman, E. (2004), Statistical assessment of mediational effects for logistic mediational models, *Statistics in medicine* **23**(17), 2713–2728.

Imai, K., Keele, L. and Tingley, D. (2010), A General Approach to Causal Mediation Analysis., *Psychological Methods* **15**, 309–334.

Imai, K., Keele, L. and Yamamoto, T. (2010), Identification, Inference and Sensitivity Analysis for Causal Mediation Effects, *Statistical Science* **25**, 51–71.

Jiang, Z. and VanderWeele, T. J. (2015), When Is the Difference Method Conservative for Assessing Mediation?, *American Journal of Epidemiology* **182**(2), 105–108.

Judd, C. M. and Kenny, D. A. (1981), Process Analysis: Estimating Mediation in Treatment Evaluations, *Evaluation Review* **5**, 602–619.

Karlson, K. B., Holm, A. and Breen, R. (2012), Comparing Regression Coefficients Between Same-sample Nested Models Using Logit and Probit, *Sociological Methodology* **42**(1), 286–313.

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

Last, J. M. (1988), *in* 'A Dictionary of Epidemiology, Second Edition', New York: Oxford University Press.

Lin, D. Y., Psaty, B. M. and Kronmal, R. A. (1998), Assessing the sensitivity of regression results to unmeasured confounders in observational studies., *Biometrics* **54**(3), 948–963.

MacKinnon, D. P. (2008), *Introduction to Statistical Mediation Analysis*, Taylor & Francis Group.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G. and Sheets, V. (2002), A Comparison of Methods to Test Mediation and Other Intervening Variable Effects, *Psychological Methods* **7**, 83–104.

MacKinnon, D. P., Lockwood, C. M. and Williams, J. (2004), Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods, *Multivariate Behavioral Research* **39**(1), 99–128.

MacKinnon, D. P., Warsi, G. and Dwyer, J. H. (1995), A Simulation Study of Mediated Effect Measures, *Multivariate Behavioral Research* **30**(1), 41.

Marshall, G., Warner, B., MaWhinney, S. and Hammermeister, K. (2002), Prospective prediction in the presence of missing data., *Statistics in medicine* **21**(4), 561–570.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, 2 edn, Chapman and Hall.

McDonald, R. P. (1997), Haldane's Lungs: A Case Study in Path Analysis, *Multivariate Behavioral Research* **32**(1), 1–38.

Meinert, C. L. (1986), *in* 'Clinical Trials: Design, Conduct, and Analysis', Oxford University Press, 285.

Molenberghs, G., Kenward, M. G., Verbeke, G., Iddi, S. and Efendi, A. (2013), On the Connections Between Bridge Distributions, Marginalized Multilevel Models, and Generalized Linear Mixed Models, *International Journal of Statistics and Probability* **2**.

Naimi, A. I., Kaufman, J. S. and MacLehose, R. F. (2014), Mediation misgivings: ambiguous clinical and public health interpretations of natural direct and indirect effects, *International journal of epidemiology* **43**(5), 1656–1661.

Pandharipande, P. P., Girard, T. D., Jackson, J. C., Morandi, A., Thompson, J. L., Pun, B. T., Brummel, N. E., Hughes, C. G., Vasilevskis, E. E., Shintani, A. K., Moons, K. G., Geevarghese, S. K., Canonico, A., Hopkins, R. O., Bernard, G. R., Dittus, R. S., Ely, E. W. and BRAIN-ICU Study Investigators (2013), Long-term cognitive impairment after critical illness., *The New England Journal of Medicine* **369**(14), 1306–1316.

Pearl, J. (2001), Direct and indirect effects, *In Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, San Francisco, CA: Morgan Kaufmann Publishers Inc.* 411–420.

Pearl, J. (2012*a*), The causal mediation formula–a guide to the assessment of pathways and mechanisms., *Prevention science : the official journal of the Society for Prevention Research* **13**(4), 426–436.

Pearl, J. (2012*b*), The mediation formula: A guide to the assessment of causal pathways in nonlinear models, *in* C. Berzuini, P. Dawid and L. Bernard, eds, 'Causality: Statistical Perspectives and Applications', John Wiley and Sons, 151–179.

Pollack, J. M., Vanepps, E. M. and Hayes, A. F. (2012), The moderating role of social ties on entrepreneurs' depressed affect and withdrawal intentions in response to economic stress, *Journal of Organizational Behavior* **33**(6), 789–810.

Preacher, K. J. (2015), Advances in Mediation Analysis: A Survey and Synthesis of New Developments, *The Annual Review of Psychology* **66**, 4.1–4.28.

Preacher, K. J. and Hayes, A. F. (2008*a*), Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models, *Behavior Research Methods* **40**(3), 879–891.

Preacher, K. J. and Hayes, A. F. (2008*b*), Contemporary Approaches to Assessing Mediation in Communication Research, *in* A. F. Hayes, M. D. Slater and L. B. Snyder, eds, 'Advanced Data Analysis Methods for Communication Research', SAGE Publications, Inc, 13–54.

Preacher, K. J., Rucker, D. D. and Hayes, A. F. (2007), Addressing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions., *Multivariate Behavioral Research* **42**(1), 185–227.

Robins, J. M. and Greenland, S. (1992), Identifiability and Exchangeability for Direct and Indirect Effects, *Epidemiology* **3**, 143–155.

Saunders, C. T. and Blume, J. D. (2017), A classical regression framework for mediation analysis: fitting one model to estimate mediation effects, *Biostatistics* kxx054. **URL:** *+ http://dx.doi.org/10.1093/biostatistics/kxx054*

Sobel, M. E. (1982), Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models, *Sociological Methodology* **13**, 290–312.

Springer, M. D. and Thompson, W. E. (1966), The distribution of products of independent random variables, *SIAM Journal on Applied Mathematics* **14**(3), 511–526.

Stone, C. A. and Sobel, M. E. (1990), The robustness of estimates of total indirect effects in covariance structure models estimated by maximum likelihood, *Psychometrika* **55**(2), 337–352.

Tal-Or, N., Cohen, J., Tsfati, Y. and Gunther, A. (2010), Testing causal direction in the influence of presumed media influence, *Communication Research* **37**(6), 801–824.

Tchetgen Tchetgen, E. (2014), A Note on Formulae for Causal Mediation Analysis in an Odds Ratio Context, *Epidemiologic Methods* **2**(1), 21–31.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L. and Imai, K. (2014), mediation: R Package for Causal Mediation Analysis., *Journal of Statistical Software* **59**(5), 1–38.

VanderWeele, T. J. (2013), Policy-Relevant Proportions for Direct Effects, *Epidemiology* **24**(1), 175–176.

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

VanderWeele, T. and Vansteelandt, S. (2013), Mediation Analysis with Multiple Mediators, *Epidemiologic Methods* **2**(1), 95–115.

Vansteelandt, S., Bekaert, M. and Lange, T. (2012), Imputation Strategies for the Estimation of Natural Direct and Indirect Effects, *Epidemiologic Methods* **1**(1).

Vinokur, A. D. and Schul, Y. (1997), Mastery and inoculation against setbacks as active ingredients in the jobs intervention for the unemployed, *Journal of Consulting and Clinical Psychology* **65**(5), 867–77.

Wang, Z. and Louis, T. A. (2003), Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function, *Biometrika* **4**, 765–775.

Winship, C. and Mare, R. D. (1983), Structural Equations and Path Analysis for Discrete Data, *The American Journal of Sociology* **89**, 54–110.

Woodworth, R. S. (1928), Dynamic Psychology, *in* 'Psychologies of 1925', 111–126.

Zhao, X., Lynch, Jr, J. G. and Chen, Q. (2010), Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis, *Journal of Consumer Research* **37**(2), 197–206.

Ziv, E. and Burchard, E. G. (2003), Human population structure and genetic association studies., *Pharmacogenomics* **4**(4), 431–441.