Exact parametric causal mediation analysis for non-rare binary outcomes with binary mediators

Marco Doretti\textsuperscript{1}, Martina Raggi\textsuperscript{2}, and Elena Stanghellini\textsuperscript{3}

\textsuperscript{1}University of Perugia, Department of Political Science
\textsuperscript{2}University of Neuchâtel, Faculty of Economics and Business
\textsuperscript{3}University of Perugia, Department of Economics

Abstract

In this paper, we derive the exact parametric expressions of natural direct and indirect effects, on the odds-ratio scale, in settings with a binary mediator. The effect decomposition we propose does not require the outcome to be rare and generalizes the existing one, allowing for interactions between both the exposure and the mediator and confounding covariates. Further, it outlines a more interpretable relationship between the causal effects and the correspondent pathway-specific logistic regression parameters. Our findings are applied to data coming from a microfinance experiment performed in Bosnia and Herzegovina. A simulation study for a comparison with estimators relying on the rare outcome assumption is also implemented.

1 Introduction

The concepts of direct and indirect effects have been capturing the interest of researchers for a long time, with the first formalizations dating back to the early 1900s (Wright, 1934; Cochran, 1938). In the last decades, these studies have given rise to the so-called mediation analysis framework, that was developed both without a formal causal inference approach (Judd and Kenny, 1981; Baron and Kenny, 1986) and within it (Robins and Greenland, 1992; Pearl, 2001), even in non-counterfactual frameworks (Geneletti, 2007). As is well-known, the general purpose of mediation analysis is to understand whether and to what extent the effect of a treatment/exposure on an outcome is due to the presence of intermediate variables influenced by the treatment and affecting the outcome in turn. Examples abound in epidemiology, where identifying the components of the overall effect of a drug/medical intervention is usually of interest (Lok, 2016; Valeri and VanderWeele, 2013), but also in the social sciences, with applications ranging from psychology (MacKinnon, 2012) to policy evaluation (Keele et al., 2015; Huber et al., 2018).
Figure 1: The simplest mediation scheme in a general (a) and parametric (b) setting.

The simplest graphical conceptualization of mediation is depicted by the direct acyclic graph (Lauritzen, 1996) in Figure 1a, where the overall effect of $X$ on $Y$ is split in two parts: a direct effect, represented by the path $X \rightarrow Y$, and an indirect effect, acting through the intermediate variable $W$ (also called mediator) and represented by the path $X \rightarrow W \rightarrow Y$. In a non-causal framework, a relationship between effects and regression coefficients of basic parametric linear models has been well-known for many years and referred to as Cochran’s formula (Cochran, 1938). Denoting by $\gamma_x$ the effect of $X$ in the model for $W$ given $X$ and by $\beta_x$ ($\beta_w$) the partial effect of $X$ ($W$) in the model for $Y$ given $X$ and $W$, Cochran’s formula states that the total (i.e., marginal) effect of $X$ on $Y$ is given by $\beta_x + \gamma_x \beta_w$, where $\beta_x$ plays the role of the direct effect while $\gamma_x \beta_w$ represents the indirect effect (see Figure 1b). Recently, extensions beyond linear regression, though with a less clear mapping from coefficients to effects, have also been proposed (Cox, 2007; Lupparelli, 2018; Stanghellini and Doretti, 2018).

In the causal inference framework, formal definitions of total, direct and indirect effects were first introduced by Pearl (2001); see also Pearl (2009) for a comprehensive overview. Typically, in Pearl’s approach causal effects are additive, meaning that they are defined as differences between two average outcome levels. As a consequence, effect decompositions also work on an additive scale: in particular, the so-called natural direct and indirect effects are defined in a way such that their sum equals the total causal effect. Under certain assumptions, these causal effects can be identified, and thus estimated, from observational data with a class of methods known as the mediation formula (Pearl, 2010). The mediation formula embeds the aforementioned decomposition for parametric linear models as a special case, since it extends to more general parametric settings (like the presence of interactions) as well as to non-parametric settings.

For binary outcomes, the above approach leads to evaluating causal effects on the risk difference scale. However, ratio scales, when the relationship between the total effect and the natural effects is no longer additive but multiplicative, have also been recently investigated in the literature. In particular, the odds ratio scale has been introduced with the overall aim of establishing links between natural effects and the coefficients of logistic - rather than linear - regression models, in
the spirit of Figure 1b. Specifically, VanderWeele and Vansteelandt (2010) have first defined causal effects on the odds ratio scale, providing their parametric identification expressions for the case of a continuous mediator. Later, Valeri and VanderWeele (2013) have extended this approach to the case of a binary mediator. However, since they rely on approximations of the logit function to the logarithmic function, results contained in these papers are based on the assumption that the outcome is rare, i.e., that \( P(Y = 1) \) is reasonably small. In practice, this corresponds to approximate effects on the odds ratio scale to effects on the risk ratio scale. When the outcome is not rare, such an approximation is no longer valid, thereby representing a serious limitation in many empirical data analyses.

For settings with a continuous mediator, some alternative solutions to overcome the rare outcome assumption were devised, though under different assumptions about the distribution of the mediator (Tchetgen Tchetgen, 2013) or introducing other approximations (Gaynor et al., 2018). In the case of a binary mediator, the approach proposed by Gaynor et al. (2018) substitutes the probabilities involved in the mediation formula with their parametric equivalent deriving from the assumed logistic models, without obtaining a full characterization of causal natural effects in terms of their associate pathway-specific logistic regression parameters. In this paper, we aim at filling this gap and provide an alternative exact parametric expression for the natural direct and indirect effects, on the odds ratio scale, when logistic regression models for both the outcome and the mediator are specified. In this way, the exact relationship between causal natural effects and their correspondent logistic regression coefficients becomes clearer and much easier to interpret. Our results account for the presence of any kind of confounders. Therefore, they can be used to estimate causal effects in the presence of observational data, provided that all relevant confounders are measured without error. Furthermore, our formulas handle all possible interactions in regression models, including those between the exposure (as well as the mediator) and the confounders. These interactions are not modelled in the derivations of Gaynor et al. (2018), nor they are considered within the approaches relying on the rare outcome assumption (see VanderWeele and Vansteelandt (2010) and Valeri and VanderWeele (2013) for a related discussion). Approximate standard errors of our estimators are also derived and compact formulas for their computation reported. Also these formulas generalize the existing ones in a number of ways; see Section 2 for details.

We apply our derivations to data gathered from an experiment aimed at assessing the impact of a microcredit programme developed in Bosnia and Herzegovina during the years 2009-2010 (Augsburg et al., 2015). This study is part of a wider programme involving other countries; see Banerjee et al. (2015) and references therein for a more thorough explanation of the whole project. Evidence from Augsburg et al. (2015) shows that microcredit interventions at an household or individual level impact on a number of factors including access to new loans, the development of business activities and self-employment. However, no mention of any mediation analysis is contained in the study, though subject specific considerations suggest that a number of mediation schemes might be plausible in this context. Specifically, the most pertinent research question appears to be whether the
The effect of microcredit financing on client’s access to new credit lines is mediated by a binary variable measuring whether or not after 14 months the recipient owns an active business. Indeed, a business can be started or maintained thanks to the initial microcredit financing, and business owners are more likely to access the loan market than others. From a technical standpoint, the three random variables composing the mediation scheme are all binary. Moreover, the rare outcome assumption is violated since the proportion of units who have received new loans (that is, \( P(Y = 1) \)) is equal to 0.57. These facts clearly motivate the application of the derived theoretical results to the dataset at hand.

The paper is structured as follows. In Section 2, we outline the general theory leading to the natural effect decomposition, reporting the parametric formulas for the natural direct and indirect effects on the odds ratio scale. Relationships with another decomposition, developed in a non-causal framework, are also illustrated. In Section 3, we discuss the application to the Bosnian microcredit data, showing the results obtained under the aforementioned mediation scheme. In Section 4, we present evidence from a simulation study we have conducted to compare - for different levels of outcome rareness - the estimators deriving from our exact formulas to those of Valeri and VanderWeele (2013), which rely on the rare outcome assumption. Finally, in Section 5 some concluding remarks are offered.

## 2 Parametric effect decomposition

### 2.1 Notation and assumptions

We denote the binary outcome by \( Y \), the binary mediator by \( W \) and the treatment/exposure, which can be of any nature, by \( X \). We take a potential outcome approach (Rubin, 1974) and let \( Y_x \) and \( W_x \) be, respectively, the random variables representing the outcome and the mediator had the exposure been set, possibly contrary to the fact, to level \( x \). Further, \( Y_{xw} \) indicates the value of the outcome if \( X \) had been set to \( x \) and \( W \) to \( w \).

In line with the classical causal mediation framework, we make the standard assumptions needed to identify causal natural direct and indirect effects. These are the so-called consistency and composition assumptions. Consistency states that, in the subgroup of units with \( X = x \), the observed variables \( Y \) and \( W \) equal the potential outcome variables \( Y_x \) and \( W_x \) respectively (VanderWeele, 2009). In the mediation framework, consistency also requires that for units with \( X = x \) and \( W = w \), \( Y \) is equal to the potential outcome \( Y_{xw} \) (VanderWeele and Vansteelandt, 2009). On the other hand, composition requires that \( Y_x = Y_{xW_x} \), i.e., that the potential outcome associated to the intervention \( X = x \) be equal to the potential outcome associated to setting \( X \) to \( x \) and the mediator to \( W_x \), which is the value it would have naturally attained under \( X = x \) (VanderWeele and Vansteelandt, 2009).

A number of assumptions concerning confounding are also required which are graphically sum-
In the language of conditional independence (Dawid, 1979), this corresponds to the conditional independence statements $Y_{xw} \perp \perp X \mid T$, $Y_{xw} \perp \perp W \mid (X, S)$ and $W_{x} \perp \perp X \mid V$, which have to hold for every level $x$ and $w$. Another necessary assumption, sometimes termed cross-world independence (Steen and Vansteelandt, 2018), is encoded by the conditional independence statement $Y_{xw} \perp \perp W_{x^*} \mid C$ (for all $x$, $x^*$ and $w$), meaning in practice that none of the variables tackling the mediator-outcome confounding can be affected by the treatment (which would correspond to an arrow from $X$ to $S$ in Figure 2). Such an effect would compromise the identification of natural effects (Valeri and VanderWeele, 2013; VanderWeele and Vansteelandt, 2010). In what follows, to simplify formulas we will use the condensed notation $Z = (T, S)$. Without loss of generality, we can think of $Z$ and $V$ as of univariate random variables. This is equivalent to assume that one covariate addresses both the exposure-outcome and the mediator-outcome confounding and another covariate manages the exposure-mediator confounding. Results for the case of multiple $Z$ and $V$ follow in a straightforward way; see Appendix A.

In this framework, with reference to two levels of the exposure (say $x$ and $x^*$), VanderWeele and Vansteelandt (2010) introduced the definitions of natural effects on the odds-ratio scale. Specifically, the natural direct effect is defined as

$$
\text{OR}^{\text{NDE}}_{x,x^*|c} = \frac{P(Y_{xW_{x^*}} = 1 \mid C = c)/P(Y_{xW_{x^*}} = 0 \mid C = c)}{P(Y_{x^*W_{x^*}} = 1 \mid C = c)/P(Y_{x^*W_{x^*}} = 0 \mid C = c)},
$$

whereas the natural indirect effect is

$$
\text{OR}^{\text{NIE}}_{x,x^*|c} = \frac{P(Y_{xW_{x}} = 1 \mid C = c)/P(Y_{xW_{x}} = 0 \mid C = c)}{P(Y_{x^*W_{x^*}} = 1 \mid C = c)/P(Y_{x^*W_{x^*}} = 0 \mid C = c)}.
$$
so that the causal total effect
\[
\text{OR}_{x,x^*|c}^{\text{TE}} = \frac{P(Y_x = 1 \mid C = c)/P(Y_{x^*} = 1 \mid C = c)}{P(Y_x = 0 \mid C = c)/P(Y_{x^*} = 0 \mid C = c)}
\]
can be decomposed in the multiplicative fashion
\[
\text{OR}_{x,x^*|c}^{\text{TE}} = \text{OR}_{x,x^*|c}^{\text{NDE}} \times \text{OR}_{x,x^*|c}^{\text{NIE}}
\]
or, as it is often presented, in an additive fashion on the logarithmic scale. A subset of the assumptions listed above allows to identify the causal total effect on the odds-ratio scale by the conditional associational odds-ratio
\[
\text{OR}_{x,x^*|c}^{\text{TE}} = \frac{P(Y = 1 \mid X = x, C = c)/P(Y = 0 \mid X = x, C = c)}{P(Y = 1 \mid X = x^*, C = c)/P(Y = 0 \mid X = x^*, C = c)}
\]
(VanderWeele and Vansteelandt, 2010).

2.2 Parametric formulas for causal natural effects

As mentioned in the Introduction, the natural effects \(\text{OR}_{x,x^*|c}^{\text{NDE}}\) and \(\text{OR}_{x,x^*|c}^{\text{NIE}}\) can be explicitly related to their pathway-associate coefficients of the logistic regression models for \(Y\) and \(W\), that can be easily estimated from observational data. To better fix the key idea, we first present results for the case of absence of covariates, i.e. for \(C = \emptyset\). Then, the more general case will be addressed. Specifically, in the absence of covariates we can formulate the two models as
\[
\log \frac{P(Y = 1 \mid X = x, W = w)}{P(Y = 0 \mid X = x, W = w)} = \beta_0 + \beta_x x + \beta_w w + \beta_{xw} x w
\]
(1)
and
\[
\log \frac{P(W = 1 \mid X = x)}{P(W = 0 \mid X = x)} = \gamma_0 + \gamma_x x.
\]
(2)
Combining (1) and (2) with the identification assumptions above, it is possible to show that
\[
\log \text{OR}_{x,x^*|c}^{\text{NDE}} = \beta_x (x - x^*) + \log A_{x,x^*}/A_{x^*,x^*}
\]
(3)
and
\[
\log \text{OR}_{x,x^*|c}^{\text{NIE}} = \log A_{x,x}/A_{x^*,x^*},
\]
(4)
where
\[
A_{x,x^*} = \frac{\exp(\beta_w + \beta_{xw} x) \exp(\gamma_0 + \gamma_{x^*} x^*) \{1 + \exp(\beta_0 + \beta_x x)\} + 1 + \exp(\beta_0 + \beta_w + (\beta_x + \beta_{xw}) x)}{\exp(\gamma_0 + \gamma_{x^*} x^*) \{1 + \exp(\beta_0 + \beta_x x)\} + 1 + \exp(\beta_0 + \beta_w + (\beta_x + \beta_{xw}) x)}.
\]
(5)
The mathematical derivations leading to Equations (3) and (4) are given in Appendix A.

Although the general expression of the term $A_{x,x^*}$ involves the whole set of parameters, it is not difficult to observe that $\log(A_{x,x^*}/A_{x^*,x})$ is a quantity driven by $\beta_x$ and $\beta_{xw}$, in line with its being part of the natural direct effect. Similarly, $\log(A_{x,x}/A_{x,x^*})$ is driven by $\gamma_x$, which makes sense since $\log(A_{x,x}/A_{x,x^*})$ is the natural indirect effect itself. More formally, if $\beta_x = 0 = \beta_{xw}$ there is no direct causal effect and $OR_{TE}^{x,x^*} = OR_{NIE}^{x,x^*}$, while if $\gamma_x = 0$ there is no indirect effect and $OR_{TE}^{x,x^*} = OR_{NIE}^{x,x^*}$. Furthermore, if $\beta_w = 0 = \beta_{xw}$, then $A_{x,x^*}$ is equal to 1 for every combination of $x$ and $x^*$. Clearly, also in this case the natural indirect effect vanishes as it is meant to do since there is no influence of $W$ on $Y$ conditional on $X$. It is useful to bear in mind that such a clear correspondence between model parameters and direct/indirect effects does not hold with associational (non-causal) effects. In particular, for non-linear models it is known that the condition $\gamma_x = 0$ does not guarantee the marginal and conditional effects of $X$ on $Y$ be equal (Cox and Wermuth, 2003). For logistic regression, this is related to the well-known fact that odds-ratios are non-collapsible association measures (Greenland et al., 1999).

The extension of the above setting to the parametric inclusion of covariates $C$ is immediate. In detail, if Equations (1) and (2) are modified to account for these additional covariates and for all their possible interactions, i.e., to

$$
\log \frac{P(Y = 1 \mid X = x, W = w, Z = z)}{P(Y = 0 \mid X = x, W = w, Z = z)} = \beta_0 + \beta_x x + \beta_w w + \beta_z z + \beta_{xw} xw + \beta_{xz} xz + \beta_{uw} wz + \beta_{xz} xwz
$$

and

$$
\log \frac{P(W = 1 \mid X = x, V = v)}{P(W = 0 \mid X = x, V = v)} = \gamma_0 + \gamma_x x + \gamma_v v + \gamma_{xv} xv,
$$

then natural effects become conditional on the covariate configuration $C = c$ and can be written as

$$
\log OR_{NIE}^{x,x^*|c}(x - x^*) = (\beta_x + \beta_{xz})(x - x^*) + \log \frac{A_{x,x^*|c}}{A_{x^*,x^*|c}}
$$

and

$$
\log OR_{NIE}^{x|x^*|c} = \log \frac{A_{x|x^*|c}}{A_{x^*,x^*|c}},
$$

where, denoting the exponentiated linear predictors by the compact forms

$$
e_y(x, w, z) = \exp(\beta_0 + \beta_x x + \beta_w w + \beta_z z + \beta_{xw} xw + \beta_{xz} xz + \beta_{uw} wz + \beta_{xz} xwz)
$$

$$
e_w(x, v) = \exp(\gamma_0 + \gamma_x x + \gamma_v v + \gamma_{xv} xv),
$$

the conditional version of (5) is given by

$$
A_{x,x^*|c} = \frac{\exp(\beta_w + \beta_{xw} x + \beta_{uw} z + \beta_{xuw} wz)e_w(x^*, v)}{e_w(x^*, v)\{1 + e_y(x, 0, z)\}} + 1 + e_y(x, 1, z)
$$

and

$$
A_{x^*,x^*|c} = \frac{\exp(\beta_w + \beta_{xw} x + \beta_{uw} z + \beta_{xuw} wz)e_w(x^*, v)}{e_w(x^*, v)\{1 + e_y(x, 0, z)\}} + 1 + e_y(x, 1, z).
$$
The proof follows immediately from the one of the previous case and is also reported in Appendix A.

Clearly, all these decompositions allow the consistent estimation of causal natural effects by simply plugging-in the parameter estimates of logistic regression models in the formulas above. Approximate standard errors for the estimates \( \hat{e} = (\hat{\text{OR}}_{x,x^*|c}^{\text{NDE}}, \hat{\text{OR}}_{x,x^*|c}^{\text{NIE}})' \) can be obtained via the delta method (Oehlert, 1992). Explicit formulas for the first-order approximate variance-covariance matrix \( V(\hat{e}) \), obtained with such a method, are reported in Appendix B. These expressions extend those contained in the R code provided by Gaynor et al. (2018), which, for the case of a binary mediator, do not account for the exposure/mediator interaction or for any interaction involving covariates and appear to be limited to changes between \( x^* = 0 \) and \( x = 1 \).

### 2.3 Links with other effect decompositions

Another interesting decomposition of associational effects exists which has links with the decomposition of causal natural effects introduced above. Indeed, denoting the conditional risk ratio of \( W \) and \( Y \) given \((X,C)\) by

\[
\text{RR}_{W|Y,X=X^*,C=C} = \frac{P(W = 1 | Y = 1, X = x, C = c)}{P(W = 1 | Y = 0, X = x, C = c)}
\]

and letting \( \bar{W} = 1 - W \), the marginal log-odds ratio \( \log \text{OR}_{x,x^*|c} \) can be written as

\[
\log \text{OR}_{x,x^*|c} = (\beta_x + \beta_xz)(x - x^*) + \log \text{RR}_{W|Y,X=X^*,C=C} - \text{RR}_{W|Y,X=X,C=C}
\]

\[
= (\beta_x + \beta_xz)(x - x^*) + \log \text{RR}_{W|Y,X=X^*,C=C} - \text{RR}_{W|Y,X=X,C=C}.
\]  

(11)

Equation (11) is a generalization of the result proposed in Stanghellini and Doretti (2018), that applies to two consecutive levels of a discrete treatment. Since the conditional total causal effect \( \text{OR}_{x,x^*|c}^{\text{TE}} \) is identified by \( \text{OR}_{x,x^*|c}^{\text{NDE}} \), it is possible to relate natural causal effects to (11) to obtain an alternative decomposition of them. Specifically, we have

\[
\log \text{OR}_{x,x^*|c}^{\text{NDE}} = (\beta_x + \beta_xz)(x - x^*) + \log \text{RR}_{W|Y,X=X^*,C=C} + \log A_{x,x^*|c}
\]  

(12)

and

\[
\log \text{OR}_{x,x^*|c}^{\text{NIE}} = -\log \text{RR}_{W|Y,X=X,C=C} - \log A_{x,x^*|c}
\]  

(13)

(see the final part of Appendix A for the proof). Equations (12) and (13) do not have a well-established causal interpretation, since neither the risk ratio term nor \( A_{x,x^*|c} \) are conceptually ascribable to any of the causal paths from \( X \) to \( Y \). Nevertheless, we think it is worth to acknowledge their existence, also in the light of possible extensions; see Section 5.
3 Application to Bosnian microcredit data

Microcredit, as the main tool of microfinance, makes credit accessible even to people who are normally excluded from the institutional financial system. Normally, this kind of people, often termed “unbankables”, are considered too risky and financially unreliable to access regular loans granted by financial institutions. Therefore, the main purpose of microcredit is the enhancement of the overall “bankability” - that is, the capability to attract loans from banks or other microfinance institutions (MFIs) - of the financially disadvantaged individuals. As mentioned in Section 1, we here offer an empirical application of our analytical results to the microcredit experiment implemented in Bosnia and Herzegovina by Augsburg et al. (2015); see also Banerjee et al. (2015) for details about the more general project involving similar experiments in other countries. This study was performed during the period 2009-2010 and was addressed to a particular segment of unbankable people formed by the potential clients of a well-established MFI of the country.

The main goal of the experiment was to evaluate the impact of randomly allocated microcredit loans not only on client’s bankability but also on a number of other socioeconomic outcomes including self-employment, business ownership, income, time worked, consumption and savings. At baseline, clients were selected to take part to the experiment and enrolled in a pre-intervention survey in order to collect main information concerning them and their household. Then, they were randomly assigned to the exposure (access to the microloan) or control group. After 14 months, the research team conducted a follow-up survey on the same respondents recruited at baseline. In total, 995 respondents were interviewed at the two waves. The average microloan amount was equal to 1’653 Bosnian marks (BAM, with an exchange rate at baseline of US$1 to BAM 1.634) with an average maturity of 57 weeks.

Since individuals were free to use money from the loan for business activities as well as for household consumption, a positive effect of the financing policy on many of the above-mentioned socioeconomic indicators was found (Augsburg et al., 2015). However, some of these measures can be reasonably thought of not only as final outcomes, but also as determinants of client’s future credit attractiveness, lending themselves to the role of possible mediators of the overall effect of microcredit on bankability. In particular, Banerjee et al. (2015) acknowledge, though without any formal analysis, business ownership as the main candidate as a mediator variable. In line with their hypothesis, we apply the derived decomposition of causal natural effects in order to try and validate such a mediation scheme. Specifically, we make use of the notation of Section 2 and denote by $X$ the binary exposure taking value 1 if the client gets the microcredit financing at baseline and 0 otherwise, by $Y$ the binary outcome taking value 1 for clients who have access to at least one new credit line from an MFI at follow-up, and by $W$ the binary mediator with value 1 for units owning a personal business and 0 otherwise, which is also measured at follow-up. Notice that individuals with $Y = 1$ might have received loans also from other traditional institutions like banks. A graphical representation of the setting under investigation is shown in Figure 3.
Figure 3: Causal mediation setting for the Bosnian microcredit study.

\[ Y \sim \beta_0 + \beta_x X + \beta_w W + \beta_{xw} X W + \beta_a A + \beta_u U + \beta_l L \]

| Estimate | Std. Error | 95% Conf. Interval | p-value |
|----------|------------|--------------------|---------|
| \( \beta_0 \) | -1.542 | 0.290 | -2.118 -0.981 | 0.000 |
| \( \beta_x \) | 1.903 | 0.213 | 1.492 2.327 | 0.000 |
| \( \beta_w \) | 0.758 | 0.211 | 0.349 1.175 | 0.000 |
| \( \beta_{xw} \) | 0.137 | 0.296 | -0.444 0.718 | 0.643 |
| \( \beta_a \) | 0.008 | 0.006 | -0.004 0.020 | 0.214 |
| \( \beta_u \) | -1.001 | 0.363 | -1.729 -0.299 | 0.006 |
| \( \beta_l \) | 0.185 | 0.085 | 0.020 0.355 | 0.029 |

\[ W \sim \gamma_0 + \gamma_x X \]

| Estimate | Std. Error | 95% Conf. Interval | p-value |
|----------|------------|--------------------|---------|
| \( \gamma_0 \) | 0.027 | 0.095 | -0.159 0.213 | 0.776 |
| \( \gamma_x \) | 0.262 | 0.128 | 0.011 0.513 | 0.041 |

Table 1: Results from the fitted logistic models for the outcome and the mediator.

All the sample marginal probabilities for \( X \), \( W \) and \( Y \) are close to 0.5. In detail, we have \( P(X = 1) = 0.55 \), \( P(W = 1) = 0.54 \) and \( P(Y = 1) = 0.57 \). Moreover, from Figure 3 it is possible to note that no exposure-outcome and exposure-mediator confounders are included in the analysis since the exposure assignment is randomized. On the contrary, a set of possible mediator-outcome confounders \( S \) needs to be determined. Some preliminary research combined with subject matter considerations led to include in \( S \) client’s age (\( A \)), educational level (\( U \)) and number of active loans (\( L \)). In particular, age is measured in years while educational level is coded as a binary variable taking value 1 for individuals with at least a university degree and 0 otherwise. All these covariates are measured at baseline and can be considered as pre-treatment variables. This should ensure that the cross-world independence assumption (see Section 2) holds, since none of
Table 2: Estimates, standard errors (SEs), 95% confidence intervals (CIs) and p-values of the causal odds ratios for the mediation scheme of Figure 3.

| A = 37, U = 0, L = 0 | A = 37, U = 1, L = 0 |
|----------------------|----------------------|
| **OR**<sub>NDE</sub><<sub>1,0</sub> | **OR**<sub>NDE</sub><<sub>1,0</sub> |
| 6.652 | 6.796 |
| 0.953 | 0.988 |
| 5.024 | 5.112 |
| 8.809 | 9.036 |
| 0.000 | 0.000 |
| **OR**<sub>NIE</sub><<sub>1,0</sub> | **OR**<sub>NIE</sub><<sub>1,0</sub> |
| 1.059 | 1.059 |
| 0.033 | 0.033 |
| 0.997 | 0.997 |
| 1.125 | 1.125 |
| 0.063 | 0.063 |
| **OR**<sub>TE</sub><<sub>1,0</sub> | **OR**<sub>TE</sub><<sub>1,0</sub> |
| 7.046 | 7.197 |
| 1.022 | 1.071 |
| 5.302 | 5.376 |
| 9.364 | 9.635 |
| 0.000 | 0.000 |

| A = 37, U = 0, L = 1 | A = 37, U = 1, L = 1 |
|----------------------|----------------------|
| **OR**<sub>NDE</sub><<sub>1,0</sub> | **OR**<sub>NDE</sub><<sub>1,0</sub> |
| 6.646 | 6.757 |
| 0.954 | 0.967 |
| 5.017 | 5.072 |
| 8.806 | 8.913 |
| 0.000 | 0.000 |
| **OR**<sub>NIE</sub><<sub>1,0</sub> | **OR**<sub>NIE</sub><<sub>1,0</sub> |
| 1.059 | 1.059 |
| 0.033 | 0.033 |
| 0.997 | 0.997 |
| 1.125 | 1.125 |
| 0.062 | 0.063 |
| **OR**<sub>TE</sub><<sub>1,0</sub> | **OR**<sub>TE</sub><<sub>1,0</sub> |
| 7.039 | 7.157 |
| 1.022 | 1.057 |
| 5.296 | 5.358 |
| 9.356 | 9.559 |
| 0.000 | 0.000 |

| A = 37, U = 0, L = 2 | A = 37, U = 1, L = 2 |
|----------------------|----------------------|
| **OR**<sub>NDE</sub><<sub>1,0</sub> | **OR**<sub>NDE</sub><<sub>1,0</sub> |
| 6.647 | 6.723 |
| 0.957 | 0.967 |
| 5.012 | 5.072 |
| 8.815 | 8.913 |
| 0.000 | 0.000 |
| **OR**<sub>NIE</sub><<sub>1,0</sub> | **OR**<sub>NIE</sub><<sub>1,0</sub> |
| 1.059 | 1.059 |
| 0.033 | 0.033 |
| 0.997 | 0.997 |
| 1.125 | 1.125 |
| 0.062 | 0.063 |
| **OR**<sub>TE</sub><<sub>1,0</sub> | **OR**<sub>TE</sub><<sub>1,0</sub> |
| 7.040 | 7.121 |
| 1.024 | 1.045 |
| 5.294 | 5.341 |
| 9.361 | 9.494 |
| 0.000 | 0.000 |

the variables in S is causally affected by the exposure (VanderWeele and Vansteelandt, 2009; Steen and Vansteelandt, 2018). In the sample, the age ranges from 17 to 70 years, with an average of 37.81 years and a median of 37 years. The first and third quartiles are 28 and 47 years respectively. Further, only 5% of sample units own a university degree, while only 4% have three or more active loans at baseline.

Table 1 contains the output of the fitted logistic regression models for the outcome and the mediator, whereas Table 2 shows the estimates, together with their uncertainty measures, of the causal effects obtained from these model parameters. The effects refer to individuals with median age and all the most relevant patterns of the other covariates. The asymptotic standard errors and confidence intervals are constructed using the delta method as illustrated in Appendix B. As in standard analyses on odds ratios, the 95% confidence intervals are first built on the logarithmic scale and then exponentiated. Also, the p-values refer to tests where the null hypotheses are formulated on the logarithmic scale, that is, that log-odds ratios are equal to zero. In the outcome model, the presence of interaction terms involving the confounders was explored, but none of these effects
resulted statistically significant or, to the best of our judgment, worth to be added to the model.

Table 1 shows that all the estimated coefficients related to the mediation pathways \( X \rightarrow W \rightarrow Y \) and \( X \rightarrow Y \) are positive and statistically significant, with the exception of the interaction \( \hat{\beta}_{xw} \), which is positive but not significant (p-value 0.643). However, it is possible to notice a relevant difference in the coefficient magnitudes. Indeed, \( \hat{\beta}_w \) and \( \hat{\gamma}_x \) are much smaller than \( \hat{\beta}_x \), suggesting that the natural direct effect is the dominant component of the total effect. This is confirmed by the results in Table 2 which are rather stable across the covariate patterns examined. Specifically, the estimated natural effects always lie between 6.646 and 6.796, whereas all the estimates of natural indirect effects are very close to 1.059 (digits from the fourth onwards are not reported in the table). Other unreported effects computed for some additional covariate patterns are also in line with the values in Table 2. As a consequence, we can conclude that the estimated causal odds ratios for the total effects lie around 7.10, a value slightly greater but essentially in line with the marginal outcome/exposure odds ratio (6.885, s.e. 0.985 with the delta method), which also has a causal interpretation - though not in a mediation setting; see Section 5 - due to randomization.

All the 95% confidence intervals for the indirect effects barely contain 1, corresponding to p-values around 6%. The low magnitude of the natural indirect effects might be due to the relatively limited temporal distance occurring between the baseline and the follow-up measurement occasions. Indeed, also from the original study by Augsburg et al. (2015) it seems that a 14-month period may be not long enough to register any relevant effect of microcredit on business ownership. We also replicated these results starting from an alternative outcome model where the \( XW \) interaction is removed. In this model, the main effects modify to \( \hat{\beta}_x = 1.974 \) (s.e. 0.149) and \( \hat{\beta}_w = 0.828 \) (s.e. 0.149), while the other parameters do not sensibly change. The causal odds ratios resulting from this model are substantially equivalent to the previous one.

### 4 Simulation study

In this section, we present the results of a small simulation study conducted to compare the estimators proposed here to those in Valeri and VanderWeele (2013), which are based on the rare outcome assumption. For simplicity, we will henceforth refer to the former as to the exact estimators and to the latter as to the approximate estimators. The behavior of the estimated standard errors is also analyzed. We consider a simplified context with no covariates and a binary treatment generated with probability \( P(X = 1) = 0.5 \). Because of the absence of covariates, the data generating process for \( Y \) and \( W \) can be represented by Equations (1) and (2).

Two settings characterized by different magnitudes of the regression coefficients are built. For each setting, we adopt a scheme similar to the one of Gaynor et al. (2018) and govern the outcome rareness by varying the intercept \( \beta_0 \). Specifically, we set three different values of \( \beta_0 \) in order to obtain, in combination with other parameters, a marginal probability \( P(Y = 1) \) approximately equal to 20%, 40% and 60%. In what follows, we borrow from the epidemiological terminology and
Table 3: Simulation study: true parameter values (up to the fourth digit) for the two settings.

| Setting | Prevalence | $\beta_0$ | $\beta_x$ | $\beta_w$ | $\beta_{xw}$ | $\gamma_0$ | $\gamma_x$ | OR$^{\text{NDE}}_{1,0}$ | OR$^{\text{NIE}}_{1,0}$ | OR$^{\text{TE}}_{1,0}$ |
|---------|------------|-----------|-----------|-----------|-------------|-----------|--------|----------------|----------------|----------------|
| 1       | 20%        | -1.30     | 0.10      | -0.08     | -0.02       | 1         | 0.60   | 1.0893         | 0.9898         | 1.0782         |
|         | 40%        | -0.45     |           |           |             |           |        | 1.0892         | 0.9999         | 1.0782         |
|         | 60%        | 0.45      |           |           |             |           |        | 1.0890         | 0.9900         | 1.0781         |
| 2       | 20%        | -2.10     | 0.80      | 0.70      | -0.10       | -0.50     | 0.60   | 2.0971         | 1.0926         | 2.2914         |
|         | 40%        | -1.10     |           |           |             |           |        | 2.0988         | 1.0918         | 2.2915         |
|         | 60%        | -0.20     |           |           |             |           |        | 2.1145         | 1.0903         | 2.3055         |

refer to $P(Y = 1)$ as to the outcome prevalence. Finally, three sample sizes ($n = 250, 500, 1000$) are considered, so that an overall total of $2 \times 3 \times 3 = 18$ simulation scenarios is obtained. For each scenario, 1000 datasets are generated and the exact and approximate estimates of the three causal parameters OR$^{\text{NDE}}_{1,0}$, OR$^{\text{NIE}}_{1,0}$ and OR$^{\text{TE}}_{1,0}$, together with their standard errors, are computed. Formulas for the standard errors of the approximate estimators are obtained with the delta method as well, as illustrated in the web appendix of Valeri and VanderWeele (2013). The true parameter values for the two settings, together with the true casual effects associated to them, are reported in Table 3.

Figure 4 contains a summary of the simulation results for the second setting. It depicts a $3 \times 3$ panel where rows vary with the sample size and columns with the causal estimand. Each plot in the panel compares the relative root mean squared error (RRMSE) of the exact and approximate estimator for the three prevalence levels. Filled dots represent the RRMSEs of the exact estimators, while void squares denote those of the approximate estimators. For the natural direct effect and the total effect, our exact method always (slightly) outperforms the approximate one in terms of RRMSE. Differences are more pronounced for lower sample sizes and increase with the prevalence as expected, though gains are modest. Instead, for the natural indirect effects the performances are essentially equivalent. The same plot for the results obtained under the first setting (not shown) is almost identical.

The better finite-sample behavior of our estimation method was somehow expected, since the approximate identification formulas the alternative estimator relies on are supposed to reflect on its bias. In any case, exact estimators are also found to have a smaller simulation variance almost in every case. Specifically, the OR$^{\text{NDE}}_{1,0}$ and OR$^{\text{TE}}_{1,0}$ exact estimators present a smaller empirical variance in all the 18 scenarios, with the maximum difference equal to 0.0620 for the total effect in the scenario with a 60% prevalence and $n = 500$. Conversely, the exact OR$^{\text{NIE}}_{1,0}$ estimator has a variance smaller than that of the approximate estimator in 11 out of 18 scenarios, though absolute differences between these variances are always almost negligible. Notice that the performance of the approximate estimator does not monotonically worsen as the prevalence increases. As a matter
Figure 4: Simulation results for setting 2: relative root mean squared error (RRMSE) as a function of the outcome prevalence for the three estimands and the three sample sizes. Filled dots represent exact estimators, whereas void squares represent approximate ones.
Figure 5: Simulation results for setting 2: boxplots of the standard errors of approximate and exact estimators for the three estimands and the three sample sizes (prevalence=20%).
of fact, we observe a performance improvement moving from a 20% to a 40% prevalence, and a worsening from 40% to 60%. In a setting with a continuous mediator, an analogous behavior was spotted, for relatively small magnitudes of the regression coefficients, in the simulation study implemented by VanderWeele and Vansteelandt (2010). However, this might not be the case for higher parameter magnitudes.

A comparison involving variance estimators is also made. Figure 5 contains the same panel as in Figure 4, but with each quadrant reporting the boxplots of the standard errors of the exact and approximate estimators computed at every iteration. The figure refers to the second setting and to the 20% prevalence case. It is possible to notice that the standard errors of the exact estimators tend to be slightly smaller than the competitor standard errors for the natural direct effect and the total effect, while they appear to be slightly higher (though the difference in the scales must be noticed) for the natural indirect effect. Again, results for every other prevalence/setting combination turn to be similar, with a more marked difference in favor of exact estimators at higher prevalences. Overall, evidence from this simulation study seem to suggest that the exact causal effect estimators we propose are worth-using.

5 Conclusions

In this paper, we have focussed on causal mediation for binary outcomes, deriving a novel parametric decomposition of natural direct and indirect effects, on the odds ratio scale, for settings where the mediator is a binary random variable. Our formulas are exact in the sense that they do not require the rare outcome assumption, and distinctly mark the link between the natural effects and their pathway-correspondent coefficients of the logistic regression models assumed to govern the data generating process of the outcome and the mediator. We have also formalized the expressions of the approximate standard errors of the causal effect estimators, obtained via the delta method. The formulas for both the causal effects and the standard errors generalize the existing ones, especially with regard to the presence of parametric interactions between the exposure (and the mediator) and the confounders in the logistic models. Furthermore, in a simulation study our exact estimators were found to overcome, even if only slightly, the approximate estimators relying on the rare outcome assumption proposed by Valeri and VanderWeele (2013).

Our theoretical findings were applied to a dataset coming from a microcredit experiment performed in Bosnia and Herzegovina (Augsburg et al., 2015), where it is plausible to think that the effect of randomly allocated microloans on client’s bankability (i.e., the capability to obtain loans from financial institutions) at a 14-month follow up might be mediated by whether or not the client owns an active business. It is worth to remark that, although in the dataset at hand the exposure was randomly assigned, our formulas are designed to address any kind of observed confounding, so they can be used also in the presence of observational data. In the experiment description contained in Augsburg et al. (2015), non-compliance issues appear to be absent, but a 17% loss to follow-
up was registered. However, the reasons leading to client dropout, together with formal analyses reported in the paper, do not induce to think that a non-ignorable missigness mechanism (Molenberghs et al., 2008) occurred.

The causal effects estimated for the microcredit data are conditional on client’s age, educational level and number of active loans at baseline, that is, on the patterns of the mediator-outcome confounder variables we have identified by combining empirical analyses and subject matter knowledge. Nevertheless, the results we have obtained are pretty stable across these patterns. Specifically, all the total causal effect odds ratios lie around 7.10, but only a small part of such total effects appears to be mediated by business ownerships. As mentioned in Section 3, since access to microloans is randomly assigned one could estimate the unconditional total causal effect through the marginal outcome/exposure odds ratio, regardless of any mediation setting. In the examined dataset, this is equal to 6.885. A sensible comparison between these results has to account for the fact that the two approaches rely on different parametric assumptions since logistic models are generally not collapsible but in the trivial case of absence of covariates and binary exposure (Lin et al., 1998).

As a sensitivity analysis, we have also computed the effect estimates after widening the set of observed confounders with some other potentially relevant variables excluded in the first place like income, value of family assets, gender and marital status. Results do not substantially change. Though this tends to confirm the validity of our causal estimates, like in every empirical study the absence of unobserved confounding cannot be guaranteed with certainty, and results have to be interpreted with caution. In this sense, it would be nice to adapt the existing methods of sensitivity analysis with respect to unobserved confounding to the present context. In particular, the interval identification method introduced by (Lindmark et al., 2018) for probit regression could be adapted to logistic models. Other promising extensions involve multiple mediators, for which a decomposition like the one of Equations (12) and (13) has already been proposed (Stanghellini and Doretti, 2018).

References

Wright S. The method of path coefficients. The Annals of Mathematical Statistics 1934;5(3):161—215.

Cochran WG. The omission or addition of an independent variate in multiple linear regression. Supplement to J R Statist Soc 1938;5(2):171–176.

Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. Evaluation review 1981;5(5):602–619.

Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of personality and social psychology 1986;51(6):1173–1182.
Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology 1992;3(2):143–155.

Pearl J. Direct and Indirect Effects. In: Proceedings of the 17th International Conference on Uncertainty in Artificial Intelligence UAI’01, San Francisco: Morgan Kaufmann Publishers Inc.; 2001. p. 411–420.

Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. J R Stat Soc Ser B Stat Methodol 2007;69(2):199–215.

Lok JJ. Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible. Statistics in Medicine 2016;35(22):4008–4020.

Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. Psychol Methods 2013;18(2):137–150.

MacKinnon D. Introduction to statistical mediation analysis. Routledge; 2012.

Keele L, Tingley D, Yamamoto T. Identifying mechanisms behind policy interventions via causal mediation analysis. Journal of Policy Analysis and Management 2015;34(4):937–963.

Huber M, Lechner M, Strittmatter A. Direct and indirect effects of training vouchers for the unemployed. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2018;181(2):441–463.

Lauritzen SL. Graphical Models. Oxford University Press; 1996.

Cox DR. On a generalization of a result of W.G. Cochran. Biometrika 2007;94(3):755–759.

Lupparelli M. Conditional and marginal relative risk parameters for a class of recursive regression graph models. Statistical Methods in Medical Research 2018;https://doi.org/10.1177/0962280218802603, pMID: 30352537.

Stanghellini E, Doretti M. On marginal and conditional parameters in logistic regression models. ArXiv preprint. Available at https://arxiv.org/abs/1804.03062; 2018.

Pearl J. Causality: Models, Reasoning, and Inference. 2nd ed. New York, NY, USA: Cambridge University Press; 2009.

Pearl J. The mediation formula: A guide to the assessment of causal pathways in non-linear models. University of California, Los Angeles; 2010.
VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. American Journal of Epidemiology 2010;172(12):1339–1348.

Tchetgen Tchetgen E. A note on formulae for causal mediation analysis in an odds ratio context. Epidemiological Methods 2013;2(1):21–31.

Gaynor SM, Schwartz J, Lin X. Mediation analysis for common binary outcomes. Statistics in Medicine 2018;https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7945.

Augsburg B, De Haas R, Harmgart H, Meghir C. The impacts of microcredit: Evidence from Bosnia and Herzegovina. American Economic Journal: Applied Economics 2015;7(1):183–203.

Banerjee A, Karlan D, Zinman J. Six randomized evaluations of microcredit: Introduction and further steps. American Economic Journal: Applied Economics 2015;7(1):1–21.

Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 1974;66(5):688–701.

VanderWeele TJ. Concerning the Consistency Assumption in Causal Inference. Epidemiology 2009;20(6):880–883.

VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, intervention and composition. Statistics and Its Interface 2009;2(4):457–468.

Dawid AP. Conditional independence in statistical theory (with discussion). J R Stat Soc Ser B Stat Methodol 1979;41(1):1–31.

Steen J, Vansteelandt S. Graphical models for mediation analysis. arXiv preprint arXiv:180106069 2018:.

Cox DR, Wermuth N. A general condition for avoiding effect reversal after marginalization. J R Statist Soc B 2003;65(4):937–941.

Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Statistical Science 1999;14(1):29–46.

Oehlert GW. A note on the delta method. The American Statistician 1992;46:27–29.

Molenberghs G, Beunckens C, Sotto C, Kenward MG. Every missingness not at random model has a missingness at random counterpart with equal fit. J R Stat Soc Ser B Stat Methodol 2008;70(2):371–388.
Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 1998;54(3):948–963.

Lindmark A, de Luna X, Eriksson M. Sensitivity analysis for unobserved confounding of direct and indirect effects using uncertainty intervals. Statistics in Medicine 2018;37(10):1744–1762.

A Mathematical derivation of \( \text{OR}_{x,x^*}^{\text{NDE}} \) and \( \text{OR}_{x,x^*}^{\text{NIE}} \)

Given the standard causal inference assumptions of Section 2, the two natural effects can be non-parametrically identified by using Pearl’s mediation formula (Pearl, 2001, 2010). For a binary mediator, the expression identifying the natural direct effect is

\[
\text{OR}_{x,x^*}^{\text{NDE}} = \frac{\sum_{w} P(Y = 1 \mid X = x, W = w) P(W = w \mid X = x^*) / \sum_{w} P(Y = 0 \mid X = x, W = w) P(W = w \mid X = x^*)}{\sum_{w} P(Y = 1 \mid X = x^*, W = w) P(W = w \mid X = x^*) / \sum_{w} P(Y = 0 \mid X = x^*, W = w) P(W = w \mid X = x^*)}
\]

Given the parametric models assumed, the numerator of the expression above can be written as

\[
Q_1 = P(Y = 1 \mid X = x, W = 1)P(W = 1 \mid X = x^*) + P(P(Y = 1 \mid X = x, W = 0)P(W = 0 \mid X = x^*)
\]

\[
= \frac{\sum_{w} \exp[\beta_0 + \beta_w + (\beta_x + \beta_{wx}x)] P(W = w \mid x) \exp(\gamma_0 + \gamma_x x^*)}{\sum_{w} \exp(\gamma_0 + \gamma_x x^*) P(W = w \mid x) + \sum_{w} \exp(\beta_0 + \beta_x x) \exp(\gamma_0 + \gamma_x x^*) + \sum_{w} \exp(\beta_0 + \beta_w x) \exp(\gamma_0 + \gamma_x x^*) + 1}
\]

\[
= \frac{\exp[\beta_0 + \beta_w + (\beta_x + \beta_{wx}x)] \exp(\gamma_0 + \gamma_x x^*)}{\exp(\gamma_0 + \gamma_x x^*) + 1 + \exp(\beta_0 + \beta_x x) + \exp(\beta_0 + \beta_w x)}
\]

For the denominator, an analogous calculation leads to \( Q_2 = \exp(\beta_0 + \beta_x x^*) A_{x^*}^{x^*} \) and therefore to \( \log \text{OR}_{x,x^*}^{\text{NDE}} = \log Q_1 - \log Q_2 = \beta_x(x - x^*) + \log(A_{x^*}^{x^*}/A_{x^*}^{x^*}) \), which proves Equation (3).

Derivations for the natural indirect effect are similar since we have

\[
\text{OR}_{x,x^*}^{\text{NIE}} = \frac{\sum_{w} P(Y = 1 \mid X = x, W = w) P(W = w \mid X = x^*) / \sum_{w} P(Y = 0 \mid X = x, W = w) P(W = w \mid X = x^*)}{\sum_{w} P(Y = 1 \mid X = x^*, W = w) P(W = w \mid X = x^*) / \sum_{w} P(Y = 0 \mid X = x^*, W = w) P(W = w \mid X = x^*)}
\]

with \( Q_3 = \exp(\beta_0 + \beta_x x) A_{x,x} \), leading to \( \log \text{OR}_{x,x^*}^{\text{NIE}} = \log Q_3 - \log Q_1 \), that is, Equation (4).
The algebraic developments above remain unchanged once confounding-removing covariates $C$ are added, provided that linear predictors are suitably modified. Specifically, the conditional versions of $Q_1$, $Q_2$ and $Q_3$ become

\[ Q_{1|c} = e_y(x, 0, z)A_{x|x*|c} \]
\[ Q_{2|c} = e_y(x^*, 0, z)A_{x^*x^*|c} \]
\[ Q_{3|c} = e_y(x, 0, z)A_{x|x|c}, \]

where $e_y(x, w, z)$ and $A_{x,x^*|c}$ are as in Section 2. The derivation of Equations (8) and (9) is then immediate. Notice that this approach can be immediately generalized to account for multiple confounders; it suffices to replace $z$ with $z = (z_1, \ldots, z_p)'$ and $v$ with $v = (v_1, \ldots, v_q)'$ in the formulas above, substituting every product involving $z$ and $v$ with the corresponding row column product (for instance, $\beta z$ is replaced by $\beta' z$ with $\beta = (\beta_{z1}, \ldots, \beta_{zp})'$ and so on).

To prove (12) and (13), it is necessary to rewrite $Q_{2|c}$ and $Q_{3|c}$ in an alternative form. First, one has to notice that for any triplet of events $(E_1, E_2, E_3)$

\[
P(E_1 \mid E_2) = \frac{P(E_1 \mid E_2, E_3)P(E_2, E_3)}{P(E_3 \mid E_1, E_2)P(E_2)}
\]

(14)

holds. Therefore, we have

\[
Q_{2|c} = \frac{\sum_w P(Y = 1 \mid X = x^*, W = w, C = c)P(W = w \mid X = x^*, C = c)}{\sum_w P(Y = 0 \mid X = x^*, W = w, C = c)P(W = w \mid X = x^*, C = c)}
\]
\[
\quad = \frac{P(Y = 1 \mid X = x^*, C = c)}{P(Y = 0 \mid X = x^*, C = c)} \times \frac{P(W = 0 \mid Y = 0, X = x^*, C = c)}{P(W = 0 \mid Y = 1, X = x^*, C = c)}
\]
\[
\quad = e_y(x^*, 0, z) \left( RR_{W|Y,X=x^*,C=c} \right)^{-1},
\]

where the second-to-last equality is obtained applying (14) to both the numerator and the denominator with $E_1 = \{Y = y\}$, $E_2 = \{X = x^*, C = c\}$ and $E_3 = \{W = 0\}$. As a consequence,

\[
\log \frac{Q_{1|c}}{Q_{2|c}} = (\beta_x + \beta_{xz}z)(x - x^*) + \log A_{x,x^*|c} + \log RR_{W|Y,X=x^*,C=c},
\]

which proves (12). Interestingly, the equality $A_{x,x^*|c} = \left( RR_{W|Y,X=x^*,C=c} \right)^{-1}$ holds. The same arguments apply to $Q_{3|c} = e_y(x, 0, z) \left( RR_{W|Y,X=x^*,C=c} \right)^{-1}$, which leads to (13) by developing $\log(Q_{3|c}/Q_{1|c})$.
B Variance-covariance matrix of estimated causal natural effects

Denoting by $\beta = (\beta_0, \beta_x, \beta_z, \beta_{xz}, \beta_{w}, \beta_{wx}, \beta_{wz}, \beta_{wxz})'$ and $\gamma = (\gamma_0, \gamma_x, \gamma_v, \gamma_{xv})'$ the two vectors of model parameters and by $\Sigma_\beta$ and $\Sigma_\gamma$ the variance-covariance matrices of their estimators $\hat{\beta}$ and $\hat{\gamma}$, the first-order approximate variance-covariance matrix of $\hat{e} = (\hat{\text{OR}}_{x,x^*|c}, \hat{\text{OR}}_{x,x^*|c})'$ can be obtained as $V(\hat{e}) = E \Sigma D' E'$, where $E = \text{diag}(e)$.

$$\Sigma = \begin{pmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\gamma \end{pmatrix}$$

and $D$ is the matrix of derivatives $D = \partial \log e / \partial \theta'$, with $\theta = (\beta', \gamma')'$ denoting the whole parameter vector. To obtain $D$, it is convenient to compute the row vector $d_{x,x^*|c} = \partial A_{x,x^*|c} / \partial \theta'$ first. To this end, it is worth to write $A_{x,x^*|c}$ as

$$A_{x,x^*|c} = \frac{p_1 p_2 p_3 + p_4}{p_2 p_3 + p_4},$$

with $p_1 = \exp(\beta_w + \beta_{wx} x + \beta_{wz} z + \beta_{wxz} x z)$, $p_2 = e_w(x^*, w)$, $p_3 = 1 + e_y(x, 0, z)$ and $p_4 = 1 + e_y(x, 1, z)$. Under this notation, the three key derivatives to compute are

$$d_{\beta_0}(x, x^* | c) = \frac{\partial A_{x,x^*|c}}{\partial \beta_0} = \frac{\{p_1 p_2 (p_3 - 1) + p_4 - 1\}(p_2 p_3 + p_4) - \{p_1 p_2 p_3 + p_4\}(p_2 (p_3 - 1) + p_4 - 1)}{(p_2 p_3 + p_4)^2},$$

$$d_{\beta_w}(x, x^* | c) = \frac{\partial A_{x,x^*|c}}{\partial \beta_w} = \frac{(p_1 p_2 p_3 + p_4 - 1)(p_2 p_3 + p_4) - (p_1 p_2 p_3 + p_4)(p_2 p_3 + p_4)}{(p_2 p_3 + p_4)^2},$$

$$d_{\gamma_0}(x, x^* | c) = \frac{\partial A_{x,x^*|c}}{\partial \gamma_0} = \frac{(p_1 p_2 p_3)(p_2 p_3 + p_4) - (p_1 p_2 p_3 + p_4)(p_2 p_3)}{(p_2 p_3 + p_4)^2},$$

while the others can be written as functions thereof. Specifically, a compact form for $d_{x,x^*|c}$ is given by

$$d_{x,x^*|c} = [(d_{\beta_0}(x, x^* | c), d_{\beta_w}(x, x^* | c)) \otimes d(x, z), d_{\gamma_0}(x, x^* | c)d(x^*, v)],$$

where $\otimes$ denotes the Kronecker product and, letting $I_2$ be a diagonal matrix of order 2, $d(a, b)$ is the row vector returned by the vector-matrix multiplication $d(a, b) = (1, a)[(1, b) \otimes I_2]$. The vectors $d_{x,x^*|c}$ and $d_{x^*,x^*|c}$ can be calculated applying the same formulas above to $A_{x,x|c}$ and $A_{x^*,x^*|c}$ respectively. Then, the matrix $D$ can be obtained as

$$D = \begin{pmatrix} d_1 + d_2 \\ d_3 \end{pmatrix},$$
where

\[
d_2 = \frac{d_{x,x^*}}{A_{x,x^*}} - \frac{d_{x^*,x^*}}{A_{x^*,x^*}},
\]

\[
d_3 = \frac{d_{x,x}}{A_{x,x}} - \frac{d_{x,x^*}}{A_{x,x^*}},
\]

while \(d_1\) is a row vector of the same length of \(\theta\) with all its components set to zero but the ones in the positions of \(\beta_x\) and \(\beta_{xz}\), worth \(x - x^*\) and \(z(x - x^*)\) respectively. Again, extension to multiple confounders is immediate provided that \(\beta\) and \(\gamma\) are extended as follows:

\[
\beta = (\beta_0, \beta_x, \beta_1, \ldots, \beta_p, \beta_{x^*}, \ldots, \beta_{xz_1}, \ldots, \beta_{xz_p}, \beta_{xw}, \beta_{xw_1}, \ldots, \beta_{xw_p}, \beta_{xwz_1}, \ldots, \beta_{xwz_p})',
\]

\[
\gamma = (\gamma_0, \gamma_x, \gamma_{v_1}, \ldots, \gamma_{v_q}, \gamma_{xv_1}, \ldots, \gamma_{xv_p})'.
\]

Given the multiplicative nature of the effect decomposition on the odds-ratio scale, the delta method can be further applied to compute the approximate variance of the estimated causal total effect as

\[
V(\hat{\text{OR}}_{x,x^*|c}^{\text{TE}}) = \bar{e}'V(\hat{e})\bar{e},
\]

where \(\bar{e} = (\text{OR}_{x,x^*|c}^{\text{NIE}}, \text{OR}_{x,x^*|c}^{\text{NDE}})'\), which is \(e\) with its two components exchanged. Clearly, in finite-sample analyses one has plug in the estimates \(\hat{\beta}\) and \(\hat{\gamma}\) in the formulas above to obtain the estimated variances/covariances.