UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

Mediation Analysis for Binary Random Variables

Parametric Decomposition of the Total Effect

PhD thesis submitted to the Faculty of Economics and Business

For the PhD degree in Applied Statistics

by

Martina RAGGI

Approved by the dissertation committee:

Prof. Kilian Stoffel, University of Neuchâtel, thesis co-supervisor
Prof. Laurent Donzé, University of Neuchâtel, thesis co-supervisor
Prof. Adrian Holzer, University of Neuchâtel, head of jury
Prof. Rino Bellocco, University of Milano Bicocca, external expert
Prof. Giovanni Marchetti, University of Firenze, external expert
Prof. Elena Stanghellini, University of Perugia, external expert

Defended on September 10, 2020

UNIVERSITÉ DE NEUCHÂTEL

IMPRIMATUR POUR LA THÈSE

Mediation Analysis for Binary Random Variables:
Parametric decomposition of the total effect

Martina RAGGI

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

La Faculté des sciences économiques,
sur le rapport des membres du jury

Prof. Kilian Stoffel (co-directeur de thèse, Université de Neuchâtel)
Prof. Laurent Donzé (co-directeur de thèse, Université de Fribourg)
Prof. Adrian Holzer (président du jury, Université de Neuchâtel)
Prof. Giovanni Maria Marchetti (Université de Florence)
Prof. Rino Bellocco (Université de Milano Bicocca)
Prof. Elena Stanghellini (Université de Pérouse)

Autorise l'impression de la présente thèse.

Neuchâtel, le 22 septembre 2020

Annik Dubied

La doyenne
Annik Dubied

■ Téléphone : +41 32 718 15 00 ■ E-mail : secretariat.seco@unine.ch ■ www.unine.ch

"Per fare tutto...ci vuole un fiore"

(Gianni Rodari, 1974)

Contents

List of Tables

List of Figures

Acknowledgements

Questa tesi è il frutto di infiniti alti e bassi, tira e molla, cambiamenti, ansie, full immersion, traguardi. È il risultato di una Martina più consapevole dei propri limiti e soprattutto delle proprie potenzialità. Non è stato semplice, tante volte avrei voluto mollare, tornare in Italia, trovarmi un lavoro che non mi occupasse mentalmente così tanto. Eppure, eccomi alla fine. Se oggi sono arrivata fin qui lo devo sicuramente a tutte le persone che ho avuto la fortuna di incontrare lungo questo cammino.

Prima di tutto vorrei ringraziare i miei relatori di tesi. Il Prof. Laurent Donzé, che grazie al suo impegno e alla sua professionalità mi ha accolta e sostenuta durante tutto il percorso. È stato un piacere ed un onore aver lavorato al suo fianco. Così, il Prof. Kilian Stoffel, per il suo sostegno formale e l'opportunità concessa.

Vorrei poi ringraziare il Prof. Catalin Starica, nonostante il difficile inizio, mi ha comunque permesso di poter continuare a lavorare al suo fianco.

Un particolare ringraziamento va ai membri esterni della commissione di tesi. Il Prof. Giovanni Marchetti e il Prof. Rino Bellocco per i loro preziosi e stimolanti commenti sulla tesi.

L'IMI, la mia tana. Ringrazio Adrian per l'organizzazione e l'efficienza in qualità di presidente di giuria. Ringrazio Paul per la sua pazienza e disponibilità, per aver avuto sempre una risposta ad ogni mia domanda, anche la più scontata. Grazie ad Eugenia per la sua dolcezza e gentilezza. Grazie a tutti i miei colleghi, vecchi e nuovi. Alessio per le chiacchiere metà francesi metà italiane, Kris per le sue domande esistenziali, Michael per aver condiviso l'ufficio con me quest'ultimo anno, Aditya per la two-way anova. Selena e Simin, che seppur non più colleghe, sono e saranno delle ottime amiche.

Eliane, che dire! La persona che augurerei a tutti di incontrare almeno una volta nella vita. Uno dei pilastri fondamentali della riuscita di questo dottorato... e non solo.

Gli amici, quelli conosciuti in Svizzera. Grazie a Jeff, per essere un fantastico coinquilino, un'ottima spalla, un vero amico. Alex et Aure per le serate insieme, Alicia per la breve ma piacevole convivenza. Marti, Marielle, Giorgia, Klara, ognuna ha un posto speciale nel mio cuore, grazie per aver contribuito a rendere questa esperienza più bella. Silvia, la mia compagna d'avventura, la mia intellettuale preferita, grazie per aver reso viva la vita a Neuchâtel. David un nome una garanzia, grazie per il tuo genio e la tua follia. Un grazie speciale ad

Andrea, seppur non conosciuto in Svizzera, mi ha aiutato moltissimo in un periodo non molto stabile, grazie per il nostro cocktail Cox-Wermuth!

Le amiche di sempre, dove la distanza con loro è solo uno stato mentale. Uno di quei pochi casi dove la lontananza ci ha avvicinate. Francesca, un punto fisso. Chiara, separate alla nascita. Grazie ad entrambe per esserci sempre state.

La famiglia. Grazie papà per avermi sempre sostenuta seppur a modo tuo, per esserti sempre occupato di me nonostante le difficoltà e il mio essere sempre scontrosa con te (è solo una forma strana d'affetto). Grazie mamma per essere quella che sei, per me più di una mamma, una migliore amica. La persona che mi conosce meglio e ahimè la persona che ha dovuto sopportare più di tutti le mie continue lamentele. Sei il mio ossigeno. Sei la mia forza. Grazie Ludo per tutte le "sgridate" che mi hai dato quando ero nelle mie fasi depressive, ma soprattutto grazie per i tuoi, seppur pochi, momenti di affetto (forse meglio così, perché li ricordo tutti con grande emozione). Grazie a tutta la mia big family per essere la famiglia più pazza del mondo.

L'amore. Enrico, che fortunatamente ha vissuto solo l'ultimo anno del mio dottorato, spero abbia ancora riserve di pazienza da usare con me! Grazie per aver sopportato le mie ansie da covid, per avermi capita ed aver accettato tante mie stranezze. Grazie per il sostengo e l'amore che mi dai ogni giorno.

L'esperienza perugina. Grazie Marco, per il lavoro fatto insieme, per i preziosi consigli e per essere stato un super collega. Il Prof. Fabio Santucci, per avermi fatto viaggiare in giro per il mondo grazie ai suoi racconti e le sue avventure.

La Prof. Elena Stanghellini. Il pilastro fondamentale della mia tesi e della mia crescita professionale. Un punto di riferimento importantissimo. Grazie Prof, per tutto il tempo che mi ha concesso, per la sua esperienza, la sua professionalità e le sue conoscenze condivise con me. Ho avuto la fortuna di conoscerla anche al di fuori del lavoro e questo ha consolidato ancora di più il mio pensiero e la mia stima nei suoi confronti.

Abstract

This thesis is centered on the evaluation of direct and indirect effects via mediation analysis. A researcher is usually interested in assessing to what extent an exposure variable affects an outcome. However, identifying the overall effect does not answer questions concerning how and why such an effect arises. Single mediation analysis decomposes the overall effect of the exposure on the outcome into an indirect and a direct effect. The former refers to the to the effect of the exposure on the outcome due to a third variable, the mediator, which is supposed to fall in the pathway. The latter effect is the effect of the exposure on the outcome after keeping the mediator to whatever value might be of interest. Specifically, we derived novel exact parametric decompositions of the total effect into direct and indirect effect for binary random variables, both in the counterfactual and path-analysis frameworks. In the single mediation context, we derive parametric expressions of the counterfactual entities and their relationships with the associational definitions coming from the path analysis context. We apply these methodological results on a dataset coming from a randomly allocated microcredit program in Bosnia-Herzegovina to evaluate the effect on client's bankability. We re-analyse the data, in order to build a mediation scheme that allows a better understanding of the main effect of the study, by assuming business ownership as a possible mediator. We also implement a simulation study to compare the proposed estimator to several competing ones. When multiple mediators are involved, we found alternative definitions for the decomposition of the total effect. These new definitions are more appropriate for variables modelled as a recursive system of univariate logistic regressions. Thus, by making use of graphical models, the overall effect was defined as the sum of the direct, indirect effects and a residual term that is null under certain hypotheses. In general, these expressions are written such that one can maintain the link between effects and their corresponding coefficients in logistic regression models assumed in the system.

Key words: Mediation analysis, Direct effect, Indirect effect, Counterfactual framework, Path analysis, Parametric decomposition, Logistic regression, Multiple binary mediators, Graphical models

Résumé

Cette thèse est centrée sur l'évaluation des effets directs et indirects dans l'analyse de médiation. Habituellement, un chercheur souhaite évaluer dans quelle mesure une variable explicative affecte une variable réponse. Cependant, l'identification du seul effet total ne répond pas aux questions concernant comment et pourquoi un tel effet se produit. L'analyse de médiation simple décompose l'effet total de la variable explicative sur la variable réponse en des effets indirects et directs. Le premier concerne l'effet de la variable explicative sur la variable réponse à travers une troisième variable intermédiaire, le médiateur. Le médiateur est influencé par la variable explicative et à son tour influence la variable réponse. Au contraire, l'effet direct est l'effet qui reste après avoir maintenu le médiateur constant à une certaine valeur. Plus précisément, nous dérivons de nouvelles décompositions paramétriques exactes de l'effet total en effets directs et indirects pour des variables aléatoires binaires, à la fois dans le cadre contrefactuel et dans la ''path analysis". En présence d'un seul médiateur, nous dérivons des expressions paramétriques des quantités d'intérêt dans les cadres contrefactuel et de la ''path analysis". Nous appliquons ces résultats méthodologiques à un ensemble de données provenant d'un programme de microcrédit alloué au hasard en Bosnie-Herzégovine pour évaluer l'effet sur la bancabilité des individus. Nous réanalysons les données, afin de construire un schéma de médiation qui permet une meilleure compréhension de l'effet principal de l'étude, en utilisant la variable indiquant la création/propriété d'une entreprise comme un possible médiateur. Une étude de simulation est également effectuée, en comparant les estimateurs proposés à un nombre d'estimateurs concurrents. Nous avons trouvé des définitions alternatives pour la décomposition de l'effet total lorsque plusieurs médiateurs sont impliqués. Ces définitions sont plus appropriées pour les variables modélisées comme un système récursif de régressions logistiques univariées. Ainsi, en utilisant des modèles graphiques, nous avons défini l'effet total comme la somme des effets directs, indirects et d'un terme résiduel qui est nul sous certaines hypothèses. En général, ces expressions mettent en évidence le lien entre les effets et les coefficients de régression logistique des modèles définis dans le système.

Mot clés : Analyse de médiation, Effet direct, Effet indirect, Contrefactuel, Décomposition paramétrique, Régression logistique, Médiateurs multiples binaires, Modèles graphiques

Introduction

For many decades, applied researchers have been interested in assessing to what extent an exposure/treatment variable affects an outcome. For example, an epidemiologist could be interested in analysing the effect of a therapy on a disease, while a psychologist would be interested in evaluating the effect of a stimulus on a behaviour, or a labour-market analyst could be interested in studying the determinants of hiring discrimination. In general, it is important to understand the consequences of a given phenomenon. Nonetheless, identifying the overall treatment effect does not answer questions concerning how and why such an effect could arise. Recently, among social and medical sciences, the number of research papers has grown rapidly in order to answer those questions. In addition to the treatment effect, i.e. the effect of a cause, typically interesting in policy evaluation, many scientists have been giving more attention to the mechanisms leading to the overall effect of the treatment, i.e. the cause of the effect, (Huber 2019). To do this, we need to identify and measure the underlying pathways that are involved in the relationship under study.

Mediation analysis is one of the most common tools for that purpose. Its aim is to disentangle the total effect of a treatment on the outcome by investigating the role of one or more intermediate variables, also called mediators, which are assumed to lie in the pathway between the treatment and the outcome. In mediation analysis, one can decompose the total effect into an indirect effect, that is the effect of the treatment which arises indirectly by affecting the mediators, which in turn affect the outcome, and a direct effect that can be seen as the effect of the treatment on the outcome which remains after keeping the mediators constant to whatever value might be of interest. As an instance, borrowing the example in Pearl (2001, 2014), in policy evaluation, a labor analyst might be interested not only in evaluating to what extent race or sex discrimination affects hiring decision but mostly by which mechanisms and how to intervene to solve that issue. In the mediation framework, the policy maker, in order to eliminate the gender inequalities, may intervene directly by making hiring decisions gender-blinded (direct effect) or indirectly acting through some intermediated mechanisms like education or job qualification (indirect effect). Another illustrative example that helps to clarify the functionality of mediation analysis is given in Richiardi et al. (2013). Let us consider the effect of walking to work on the coronary heart disease (CHD) and assume there is no (total) effect. This could be explained by the fact that, even if we expect a protective direct effect of walking to work on CHD, it in turn may also produce an indirect harmful effect due to the exposure to air pollution (mediator), which counter balances the protective effect and results into a no significant total effect, as the positive and negative effect cancel out. The phenomenon

of having opposite signs in direct and indirect effects is called *inconsistent mediation* (Valeri and VanderWeele 2013). Mediation analysis is also important when the total effect is totally explained by the mediated effect, i.e. the indirect effect through the intermediate variable. In that case, we say there is a *complete mediation*, in contrast to a *partial mediation* in which both direct and indirect effect are significant and contribute to the explanation of the total effect (MacKinnon 2008). These concepts, however, should be considered carefully; see Hayes (2017, p. 119) for a discussion.

In social sciences, mediation analysis has became popular thanks to the most cited papers of Judd and Kenny (1981) and Baron and Kenny (1986). However, its origins must be traced back to the work of Wright (1921) in the context of path analysis and structural equation models (SEMs). In general, these approaches are developed for linear regression models and they allow to estimate the effects by products and sums of path-specific regression coefficients. Specifically, in order to identify the indirect effect one may use the product of coefficients method, i.e. the path coefficient of the mediator in the conditional model of the outcome (given the treatment) times the path coefficient of the treatment in the mediator model, or equivalently the difference method, i.e. the path coefficient of the treatment in the marginal model of the outcome minus the path coefficient of the treatment in the conditional model of the outcome (given the mediator). These practical and intuitive methods, however, are no longer suitable when non-linearities arise and they produce inconsistent estimates. Furthermore, they lack of specific identifiability conditions that provide causal interpretations. During the last three decades, novel definitions and identifications of direct, indirect and total effect have been developed to overcome these limitations (Robins and Greenland 1992; Pearl 2001). These new approaches have been framed within the counterfactual notation, thus, allowing for a causal interpretation of the estimated effects. The approaches solve the issue of linearity and offer formulas that do not require any particular form of the error distributions, allowing to be applied for parametric as well as non-parametric models, for linear as well as non-linear models. However, also this framework presents some difficulties when binary variables are modelled via logistic regressions. The focus of this thesis is on mediation analysis with recursive systems of logistic regressions. The decomposition of the total effect for random binary variables in both approaches, the path analysis and the counterfactual framework, is investigated. The work presented in this thesis is the outcome of research studies conducted during my PhD. Specifically, they are based on two important works (Doretti et al. 2020; Raggi et al. 2020), both under submission for academic review.

Chapter 1 briefly reviews the origins of mediation analysis and the contributions therein. The introduction of the main notation used in the counterfactual framework, i.e. the potential outcome (Rubin 1974) and an overview of the principal findings obtained in the causal mediation analysis is then offered (Robins and Greenland 1992; Pearl 2001; VanderWeele 2013a, 2014). In this framework, the effects are usually defined on a difference scale, so that the total effect can be decomposed into the sum of a direct and indirect effects. Generally, in causal mediation literature, a distinction is made between two alternative classes of effects, so-called *pure* and *total* direct/indirect effects (Robins and Greenland 1992). These effects sum up to the total effect as follows: *total direct effect* plus *pure indirect effect*, or equivalently *pure direct effect* plus *total indirect effect*. Commonly, they are both referred as *natural effects* in contrast to the *controlled* effects (Pearl 2001). In general, the total and pure effects distinguish

by the way they take into account interactions and non-linearities (Robins 2003). In particular, VanderWeele (2014) offered the highest insight of the total effect decomposition, isolating the component due only to the mediation, the component due to the interaction alone, the component due to both mediation and interaction, and the component due to neither of them. The identifiability of the natural direct and indirect effects is based on strong counterfactual assumptions about the unconfoundness of the relationships between the variables involved in the system (Pearl 2001). Under those assumptions the effects can be (non) parametrically identified through a method known as *mediation formula* (Pearl 2001, 2012). Despite the counterfactual framework overcomes some typical limits of the traditional approach, like for example the presence of interactions and non-linearities, however, it is based on strong assumptions about the counterfactual identification of the effects, which often are difficult to test (VanderWeele and Vansteelandt 2010). Hence, novel contributions to the decomposition of the total effect are presented and are based on specific conditional independence assumptions easier to test. Further, they allow to appreciate the role of the path-specific regression coefficients. In particular, starting from the specific form of the marginal (total) effect a direct, indirect and a residual effects are derived. The direct and indirect effects are defined by zeroing the path-specific regression coefficients, that, graphically, correspond to deleting the arrows in the related graph. Thus, in this sense they are given in a path analysis approach. The residual effect, on the other hand, vanishes under some specific (graphical) conditions. In a setting with a continuous outcome and a continuous mediator, a parametric comparison highlighting the situations where the two methods, counterfactual and path analysis approach, coincide is offered and the situation of absence of interaction where the methods lead back to the traditional approaches of mediation analysis (i.e. the product and difference methods) is explored.

Chapter 2 shows the main findings of mediation analysis when binary variables are modelled via logistic regressions. For binary outcomes, VanderWeele and Vansteelandt (2010) and Valeri and VanderWeele (2013) have recently extended the definition of the counterfactual effects on a ratio scale. In particular, they have defined causal effects on the odds ratio scale which can be identified under the same stringent unconfoundness assumptions given in Pearl (2001). However, their parametric identification is based on the severe assumption that the outcome needs to be rare (VanderWeele and Vansteelandt 2010). Using the logarithmic function in place of the logistic function, they approximate the effects on the odds ratio scale to effects on the risk ratio scale. Nonetheless, their formulations are still appealing since they highlight the role of path-specific regression coefficients in a rather intuitive way. However, in many empirical applications the rare outcome assumption is unrealistic and the approximation is no longer valid, thereby representing a serious limitation. In a setting with a binary outcome and a binary mediator, an exact parametric expression for the counterfactual effects, on the log odds ratio scale, that does not rely on the rare outcome assumption is offered (Doretti et al. 2020). The proposed parametric formulations are more compact and less complex than those already introduced in the literature (Gaynor et al. 2018; Samoilenko et al. 2018). The expressions easily extend to all the components of the total effect, thus allowing to generalize, on the log odds ratio scale, the derivations of the other decompositions of the total effect (VanderWeele 2014). Like the approaches developed under the rare outcome assumption, these formulations allow to appreciate the role pathway-specific coefficients play in natural effects. Also in this chapter, the decomposition of the total effect in the path analysis

approach in a setting with a binary outcome and a binary mediator is derived. Based on the recent result of Stanghellini and Doretti (2019), who explore the link between marginal and conditional logistic models, definitions of direct and indirect effect on the log odds ratio scale are then developed. The approach adopted in this frame, like one used in the previous chapter, is based on zeroing the path-specific logistic regression coefficients. The marginal effect can be written as the sum of the indirect and direct effects plus a residual term that vanishes under some specific conditions (Raggi et al. 2020). The proposed parametric relationship allows to solve the debate on which method should be used to disentangle the total effect (Breen et al. 2018) and to overcome the problem of different variances when fitting two nested non-linear models (Winship and Mare 1983). The expressions proposed in both frameworks handle every possible interaction in regression models, including those between the treatment and the mediator and between other covariates. These interactions were not previously considered by either exact or approximate approaches (Valeri and VanderWeele 2013; Gaynor et al. 2018; Samoilenko et al. 2018). Furthermore, compact formulas for the standard errors of the effect estimators are computed via the delta method. Like in the previous chapter, a comparison between the estimated effects derived in the counterfactual framework and those derived in the path analysis approach is offered, showing the specific conditions where the two methods coincide. The analytical results presented in Chapter 2 are applied on a dataset gathered from a randomized microcredit experiment performed in Bosnia and Herzegovina (Augsburg et al. 2015). A mediation scheme is build in order to facilitate the understanding of the mechanisms arising the effects of microcredit on outstanding loans. The subjects who got the microcredit are assumed to probably increase their capability to collect new funding (Augsburg et al. 2015), but at the same time the microcredit might produce a positive impact on business ownership, which in turn could lead to a better access to liquidity (Banerjee et al. 2015). In line with this scheme, re-analyses of the data are made, allowing to verify which component has a greater influence in explaining the total effect of a microcredit intervention. Finally, a simulation study is implemented. For several scenarios, an analysis comparing the exact natural effect estimators against the approximated ones of Valeri and VanderWeele (2013) is conducted.

Multiple mediation is the most frequently situation in many empirical analysis where the relations between variables are more complex. Clearly, in the traditional setting of structural linear equation modeling, multiple mediation can be assessed by taking the mediators separately and then summing the products between the linear regression coefficients, which form the indirect effects to compute the mediated effect globally (MacKinnon 2000). However, when there are interactions and non-linearities the traditional approach to multiple mediation fails (VanderWeele and Vansteelandt 2014). Chapter 3 moves through a setting which involves multiple intermediate binary random variables. In this chapter, a review of several works that have contributed to the multiple mediation literature is presented. Then, the decomposition of the total effect on a binary outcome is derived, which extends the approach of path analysis used previously. In a setting where all the variables are modelled as a recursive system of univariate logistic regressions, an exact parametric forms of the marginal effect of the treatment on the outcome is derived (Raggi et al. 2020). Starting from this result, definitions on the log odds scale for the direct, indirect effects and a residual term are obtained. Also in this setting with multiple mediators, the residual term vanishes under specific conditions, as shown in the preceding chapters. In this context, further path-specific effects can be identified and other research questions can be investigated, like for example assessing the effect of the treatment

on the outcome when some mediators are kept constant while others are marginalised over, or evaluating the extent to which the treatment affects the outcome when some parameters of the original model are imposed to zero. The derivation of the direct and indirect effects are then generalized to the situation when a marginalization is made over an intermediate or outer mediator. Furthermore, if the recursive system of logistic regressions can be defined as a probabilistic causal model (Pearl 2009b, Ch. 7), a causal interpretation of the proposed derivations might be envisaged.

Chapter 1

Mediation analysis: from path analysis to the counterfactual framework

1.1 An introduction to mediation analysis

Assessing to what extent one variable affects another variable is a central topic in empirical research. A researcher may also be interested in understanding the relationship between two variables. In particular, let us consider a setting where the effect of a treatment X on the outcome Y can be expressed through other variables involved in the system. Let us assume that the overall effect of X on Y can be decomposed into indirect effects, which are transmitted through some intermediate variables, also called mediators, and a direct effect which is the effect of the treatment after keeping the mediators to whatever value might be of interest. In doing so, we refer to the mediation analysis as the principal method for that decomposition. Studying the phenomenon of mediation is useful to highlight the contribution of specific mechanisms that explain how the treatment acts on the outcome. Related to this phenomenon is also the concept of interaction between the treatment and the mediators that explain why the treatment affects the outcome (VanderWeele 2015). The combination of both phenomena allows to assess the extent to which the effect of the treatment on the outcome is due to each phenomenon alone, both or none. In this chapter, we consider a setting with one single mediator W.

To better depict the mediation structure, let us consider the JOBS II study (Vinokur et al. 1995; Vinokur and Schul 1997), frequently used as an illustrative-example in mediation analysis. The main goal of the study was to seek if a well-randomized job training workshop (treatment), among unemployed people, had an effect on several health-related or work-related outcomes, such as depressive symptoms, employment status or the level of subject's job search

self-efficacy.[1] Mediation analysis permits to detect possible mechanisms that may explain the effect of the job program on depressive symptoms, for example, by the mediated effect due to the level of subject's job search self-efficacy. One may expect that receiving the job training workshop increases the level of subject's job search self-efficacy which in turn decreases the depressive symptoms. Disentangle the total effect of the job program on the depressive status leads to a better understanding of how the job training works and through which mechanisms it operates. As a consequence, the policy-maker can intervene to improve the effect of the program acting through the intermediate variable in order to obtain the results desired.

The concept of mediation analysis dates back to the path analysis of Wright (1921, 1934). As well-known, path analysis is a method to evaluate direct and indirect relationships between random variables by using path diagrams combined with well-specified recursive systems of linear regressions, also known as structural equations model (SEM). Specifically, path analysis relates correlation and variance-covariance matrices to path-coefficients in order to test the hypothesized relationships displayed in the path diagrams. An example of path diagram is in Figure 1.1. The nodes represent random variables, while the arrows indicate the path-coefficients that relate the nodes. In Fig. 1.1(c), we assume that the treatment X affects both directly and indirectly, through the mediator W, the outcome Y. This type of graph is also known as Directed Acyclic Graph (DAG) and, in this context, it informs us with the data generating process; see Pearl (1995), Lauritzen (1996), and Elwert (2013), to which we refer for the main definitions. Always in the context of linear models, another notable contribution to the effects decomposition is due to Cochran (1938). As well-known, Cochran's formula decomposes the marginal regression coefficient of Y on X into sums of products of linear regression coefficients. However, it is only from the eighties that mediation analysis increases in popularity among social scientists, varying from psychological (Judd and Kenny 1981; Baron and Kenny 1986; MacKinnon et al. 1991), sociological (Alwin and Hauser 1975; Sobel 1982) to economic sciences (Bollen 1987). We here refer to these works as "traditional approaches" to mediation analysis.

These methods are confined to the context of linear structural equations model and they allow to express the total effect as function of linear regression parameters. In particular, the total effect of X on Y, displayed by the path-coefficient $X \to Y$ in Fig. 1.1(a), can be disentangled into the sum of the conditional regression parameter of X in the model of Y given X and W, and the product between the conditional regression parameter of W in the model of Y given X and W, and the regression parameter of X in the model of W given X. Specifically, the first part of the sum is referred to the *direct effect*, corresponding to the path-coefficient $X \to Y$ as in Fig. 1.1(c), while the second one refers to the *indirect effect*, obtained through the so-called product method, i.e. the path-coefficient of $X \to W$ times the path-coefficient of $W \to Y$, see Fig. 1.1(c). Notice that, in this case, the indirect effect can be equivalently identified by the difference method, that is the difference between the regression parameter of X in the model of Y given X only, which can be referred to the *total effect*, minus the direct effect. Unfortunately, these traditional approaches to mediation analysis hold only for linear models. The one-to-one mapping between effects decomposition and regression coefficients, and the equivalence between the difference and the product method are lost as

[1] For job search self-efficacy we refer to *the belief that one can successfully perform specific job search behaviors and obtain employment*, Saks et al. (2015, p.105).

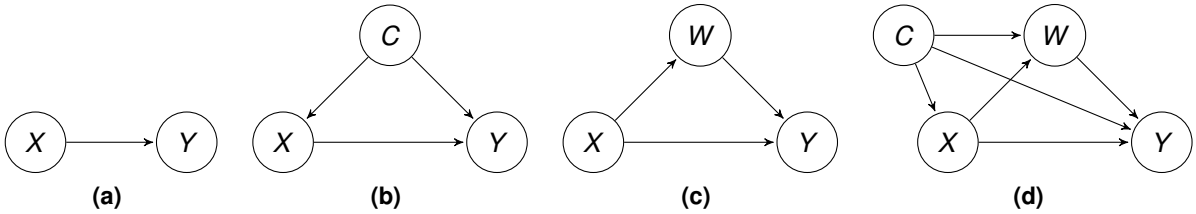


Figure 1.1: DAG (a) for the treatment-outcome relationship; (b) when conditioning on a set of covariates; DAG (c) for the mediation setting; (d) when conditioning on a set of covariates.

soon as non-linearities are introduced in the system, such as an interaction term or when using logistic or probit models. In addition, these approaches lack a clear setting which allows to provide causal interpretations (Cole and Hernán 2002; Glynn 2012; Pearl 2012).

The limits and criticisms of the traditional approach to mediation analysis have opened the way to several alternative works. Relevant contributions have addressed the complications behind linear models. For example, Cox (2007) generalized the Cochran's formula for quantile regression, while Lupparelli (2018) offers a comparison between marginal and conditional parameters for log-linear models. For logistic models similar results have been obtained by Stanghellini and Doretti (2019). Other authors have addressed the case of a single continuous mediator offering modifications to the traditional approach, mainly based on standardized parameters (MacKinnon and Dwyer 1993; MacKinnon et al. 2007; Karlson et al. 2012; Breen et al. 2013).

On the other hand, seminal works to mediation analysis have been offered by Robins and Greenland (1992), Pearl (2001), Imai et al. (2010a,b), VanderWeele and Vansteelandt (2010), and VanderWeele (2013a, 2014), which generalize and formalize the effects decomposition within the causal inference literature. In particular, Pearl (2001) formalized definitions of total, direct and indirect effect based on counterfactuals or potential outcome notation (Rubin 1974). The turning point of Pearl's contribution was to relate graphical models to the notion of causality (Pearl 1995, 2009b). As we will display in Sec.1.2 and 1.3, Pearl (2001) in its work — also known as *mediation formula* — has furnished effects' definitions on difference scale, so that the direct and indirect effect sum up to the total effect. In general, these new approaches allow to identify the effects despite the nature of the models observed, thus working for linear as well as for non linear models and for parametric as well as non-parametric models. They are based on strong assumptions (Pearl 2001; VanderWeele 2015) which are necessary to effects' identification and to give a causal interpretation. Nonetheless, some violations of those assumptions may create misleading conclusions (Richiardi et al. 2013).

The functional implications of mediation analysis make it crucial among applied sciences. Many works on mediation analysis can be found in medical sciences as well as in social sciences, for example, in public health Bellavia et al. (2017) have investigated the association between perinatal HIV infection and delayed sexual maturation mediated by poorer growth in late childhood, while Valeri (2017) has studied the effects of the maternal environmental exposure on perinatal outcomes mediated through genetic and epigenetic factors. Similarly, in

health behaviour Feingold et al. (2019) have assessed the effects of social influences on alcohol outcomes mediated by alcohol expectancies, whereas Mortensen et al. (2009) quantify the role of maternal body mass index and smoking in decomposing the effect of maternal education on birthweight. Applied works on mediation analysis in social sciences, like for example in labour market, have been offered by Huber et al. (2017), who have decomposed the treatment effect of lesser cooperative caseworkers on employment into the indirect effect which acts through the role of the assignment of an active labour market program, while Huber et al. (2018) have investigated employment effects of receiving vouchers driven by the participation to training program.

1.2 The total causal effect

The causal interpretation of a relationship between two phenomena has always been a primary goal in natural and social sciences. The word *cause* requires particular caution and it has been a source of controversies between scientists, in particular statistical researchers. *Correlation is not causation* is a famous statement that we can find in every book of basic statistics. We learned that statistics can tell us only about association. Nonetheless, when a phenomenon is under observation, it is impossible to ignore the consequences that it could cause and the underlying mechanisms of its effects. In this section we aim to recall two basic concepts of causal inference, namely, the definition of a total effect and the related assumptions allowing its causal interpretation (Pearl 2009b; Berzuini et al. 2012; Maathuis et al. 2018).[2]

Answering causal questions is useful to predict consequences and thus encourage or prevent outcomes. For our purpose, we focus on hypothetical causal queries, such as "what would be the value of the outcome if a subject had been exposed to a specific value of a treatment?", which brings out the counterfactual reasoning. The formal language of counterfactual reasoning originates from the works of Neyman (1923) and Fisher (1935), who used randomized experiments to address questions concerning causality. In observational study, the work of Cochran and Chambers (1965) was the first to point out the discussion about causal inference. However, it was during the seventies that the concepts of causality were developed by Rubin (1974, 1980) and they were formalized within the potential outcome framework, also known as Rubin Causal Model (RCM). Rubin's work translated the counterfactual notions within the probabilistic and statistical languages, so that today concepts of causality are largely used also in statistical sciences.

Recently, the counterfactual framework has been broadly developed also by Pearl (2001, 2009b), upon which we will refer throughout this manuscript. Nevertheless, the counterfactual approach to causal inference is not free from critics. In general, the major criticism concerns the strong assumptions needed to answer and interpret causal-effect queries by using real data. An interesting discussion on the use of counterfactuals in causal analysis can be found in Dawid (2000), Pearl (2000), Robins and Greenland (2000), and Rubin (2000).

[2]For a comprehensive overview of causal inference we refer to the following seminal manuscripts Cox (1992), Cox and Wermuth (2004), Pearl (2009a,b, 2010), Berzuini et al. (2012), Imbens and Rubin (2015), Morgan and Winship (2015), Pearl et al. (2016), and Maathuis et al. (2018).

Definition and identification of the total effect

Let X and Y be two random variables, respectively, the treatment and the outcome, which are supposed to be related as in Fig. 1.1(a). Furthermore, let the random variable $Y_i(x)$ be the potential outcome for the $i-th$ unit that would be observed if it had been exposed to the treatment level x. Thus, the potential outcome is the potential response of Y on the $i-th$ unit after intervening on $X = x$. The statement "if it had been exposed to the treatment" can be also formalized with the do-operator, $do(X = x)$.[3] It refers to an external intervention transforming the original probability distribution $P(Y = y \mid X = x)$ into a new distribution $P(Y = y \mid do(X = x))$, or equivalently $P(Y(x) = y)$, defined for all levels x of X.

Using the potential outcome notation, and referring to the situation in which the treatment X can change from a reference level x^* to another level x, respectively, the no-treatment and the treatment condition, we define the *total effect* (TE) of the treatment X on the outcome Y on a difference scale, both at individual and population level, as follows:

Definition 1.2.1. The individual level *Total Effect* describes the difference between the potential outcomes for the unit $i-th$ under two different states of the treatment:

$$TE_{i;x,x^*} = Y_i(x) - Y_i(x^*).$$

Definition 1.2.2. The population average *Total Effect* expresses the difference between the expected value of the potential outcomes under two different states of the treatment:

$$TE_{x,x^*} = E\,[\,Y(x) - Y(x^*)\,].$$

Following the illustrative example on the JOBS II study (Vinokur and Schul 1997), the total effect describes by how much the prevalence of depressive symptoms would change, on population average, if all unemployed subjects were invited to participate in the job training workshop contrary to the fact that all the subjects were not invited (Steen and Vansteelandt 2018).[4]

In order to obtain unbiased estimates and infer causal conclusions from real data, we need several assumptions necessary to identify the effects defined above. This means understanding under which conditions we can move from the hypothetical probability distribution $P(Y(x) = y)$ to the conditional one $P(Y = y \mid X = x)$. Here, we focus on aggregate entities, i.e. at population level, but what follows is true also at individual level. In the following, we use the notation $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B \mid C$ to say that, respectively, A and B are marginal independent, and A is conditional independent of B given C (Dawid 1979).

Temporal ordering Assumption

To define an effect as causal is first necessary to assume that the events follow a temporal sequence. The events thought of cause should precede the events thought of their effects. In

[3] See Pearl (2009b, p. 70) for the "do" notation.

[4] The subjects in the control group received a booklet with job search tips.

general, the treatment precedes the outcome; see Fig. 1.1(a). To achieve this assumption, it would be better to work with designs that measure the variables at two different points in time.

Consistency Assumption

The observed random variable Y is assumed to be related to the potential outcome $Y(x)$, as follows:

$$Y = \begin{cases} Y(x) & \text{if } X = x; \\ Y(x^*) & \text{if } X = x^*. \end{cases} \tag{1.1}$$

In other words, when the population is actually exposed to the treatment $X = x$, then the potential outcome $Y(x)$ equals the observed outcome Y for that population. The same is true also for $X = x^*$. This assumption is known as *Consistency* Assumption (Cole and Frangakis 2009; Pearl 2009b; VanderWeele 2009; VanderWeele and Vansteelandt 2009).

Ignorability Assumption

As well-known, the randomized experiment is the gold standard to evaluate causal effects. If the treatment conditions are well-randomly assigned between subjects, then the two subgroups are comparable, in average, in all pre-treatment characteristics. Therefore, the potential outcome $Y(x^*)$ that would be obtained if X had been set to x^* is independent of the treatment X actually received. In other words:

$$Y(x^*) \perp\!\!\!\perp X \quad \forall x^*. \tag{1.2}$$

The variables X and Y do not share common cause that could confound their relationship. The randomized experiment may be conceived as external intervention since each unit is "forced" to be exposed to the treatment condition or vice versa. This assumption is well-known as *Ignorability* (Rosenbaum and Rubin 1983), or *Exchangeability* (Pearl 2009b), or *No unmeasured confounders* (VanderWeele 2009) assumption.

If Ass. (1.1) and Ass. (1.2) hold then the probability distribution of the potential outcome becomes:

$$P(\ Y(x) = y\ ) = P(\ Y(x) = y \mid X = x\ ) = P(\ Y = y \mid X = x\ ).$$

Hence, we can identify and estimate from real data the total effect by the difference between the expected values of the two subgroups of the outcomes, those who are exposed to the treatment condition and those who are not exposed. From Definition 1.2.2 and the above assumption, it follows:

$$TE_{x,x^*} = E[\ Y \mid X = x\ ] - E[\ Y \mid X = x^*\ ]. \tag{1.3}$$

When the randomization experiment cannot be performed, because ethical or financial reasons, the assignment mechanism of the treatment is unknown. This is typical of observational studies. Under this situation, Ass. (1.2) does not hold any more. In order to identify the

total causal effect, we need to adjust for all common causes that could confound the relationship of interest $X \rightarrow Y$. Let $C = (C_1, ..., C_k)$ a set of observed covariates, which are potentially related with both the treatment and the outcome; see Fig. 1.1(b). These variables, most of the time, are permanent attributes of the units (e.g. sex, age, weight, height, etc.). By conditioning on a set of covariates, ideally, we re-produce the "randomization condition", that is the unconfoundness of the relationship between X and Y. The natural extension of Ass. (1.2) becomes:

$$Y(x^*) \perp\!\!\!\perp X \mid C \quad \forall x^*, c. \tag{1.4}$$

With Ass. (1.4), we assume that the value of the outcome $Y(x^*)$ is independent of the treatment X actually received given a set of covariates C. Within each level of C, the units with different states of the treatment are comparable in every characteristic, regardless the treatment itself. The Ass. (1.4) ensures that all other possible paths from X to Y are blocked[5].

If Ass. (1.1) and Ass. (1.4) hold, then we can identify and estimate from observational data the total effect taking the difference between the expected values in the two subgroups of the outcomes, those who are exposed to the treatment condition and those who are not exposed, both within homogeneous levels of covariates. It follows that:

$$\begin{aligned} TE_{x,x^*|c} &= E[\, Y(x) \mid C = c\,] - E[\, Y(x^*) \mid C = c\,] \\ &= E[\, Y \mid X = x, C = c\,] - E[\, Y \mid X = x^*, C = c\,]. \end{aligned}$$

If one is interested to get the population average total effect, one should weight each stratum of C by its probability distribution $P(C = c)$.

1.3 The decomposition of the total effect: the counterfactual approach

The total effect of X on Y assesses to what extent the variables are causally related. Instead, as explained in Sec. 1.1, if we are interested to better understand the causal mechanisms whereby the treatment affects the outcome, then mediation analysis may be a useful method for this purpose. Mediation analysis assumes that a third intermediate variable W is affected by X and in turn affects Y. Hence, one may decompose the total effect into the sum of an indirect effect of X on Y, which acts through the mediator W, and a direct effect of X on Y while fixing W at whatever value would be of interest. In the counterfactual framework, these effects are traditionally known as *natural* effects (Pearl 2001), or equivalently, *pure* direct effect and *total* indirect effect (Robins and Greenland 1992). As we will see in the next section, these methods provide general counterfactual effects' definitions that implicitly take into account also the effect due to the interaction phenomenon, i.e. the phenomenon whereby one variable modify the effect on the outcome of a different variable. This phenomenon helps to

[5] As defined in Pearl (2009b, p. 16): "Blocking is to be interpreted as stopping the flow of information (or of dependency) between the variables that are connected by such paths."

explain why a particular cause sometimes affects the outcome and sometimes does not (VanderWeele 2015, p. 9). Based on those two seminal works on the counterfactual decomposition of the total effect, recently, VanderWeele (2014) has derived the highest insight of that decomposition. In his work, the total effect is decomposed into four components, separating the mediation effect from the interaction effect. The four-way decomposition explains how and for whom an effect occurs, thus, showing the component of the total effect which is: due to the mediator alone; due to the interaction alone; due neither to the mediator nor the interaction, and due to both mediation and interaction (VanderWeele 2015).

In parallel with the potential outcome notation given in Sec. 1.2 and following the work of Pearl (2001, 2009b) and VanderWeele (2015), let $Y(x)$ and $W(x)$ be, respectively, the potential outcome and the potential value of the mediator that would be obtained if the treatment X had been set to the level x. Let $Y(x, w)$ be the potential outcome that would have been if the treatment X and the mediator W had been set, respectively, to the level x and w. Finally, let $Y(x, W(x^*))$ be the potential outcome if the treatment had been set to the value x and the mediator to the value $W(x^*)$, which is the value that it would have naturally achieved if the treatment had been fixed at level x^*. This latter entity involves a nested-counterfactual and it does not lead to a practical interpretation since it implies a match between different contrasts of the treatment X for the same variable Y (Robins and Greenland 1992; Pearl 2001; VanderWeele 2015; Steen and Vansteelandt 2018). As explained in Pearl (2009b), the nested-counterfactual cannot be written in terms of the *do*-operator because it is impossible to intervene by fixing the value of X to x in the direct path $X \to Y$ and at the same time by fixing the value of X to x^* in the indirect path $X \to W \to Y$. For this reason the identification of the effects involving nested-counterfactuals requires stronger assumptions.

In order to define all components of the total effect, we first take into account the *Composition* assumption (Pearl 2009b; VanderWeele 2009). The assumption states that the potential value of the outcome under the treatment $Y(x)$ is equal to the nested-counterfactual $Y(x, W(x))$, that is the potential value of the outcome if the treatment were fixed to x and the mediator at the value that it would have naturally attained under the same condition, i.e. $X = x$. Thus:

$$Y(x) = Y(x, W(x)) \quad \forall x.$$

By the composition assumption we can define again the total effect as follows:

Definition 1.3.1. The *Total Effect* describes the difference between the expected value of the nested counterfactual outcomes under two different states of the treatment:

$$TE_{x,x^*} = E\,[\; Y(x, W(x)) - Y(x^*, W(x^*))\;]\,.$$

Clearly, since the composition assumption the effect defined above equals the total effect defined in Sec. 1.2.

Causal components of the total effect

The traditional decomposition of the total effect, as given in Pearl (2001) and Robins and Greenland (1992), provides two alternative ways to define the direct and indirect effect. Conceptually, they differ in the way that we include the interaction between the treatment and the mediator, as we explain below. Analytically, this arises since the addition and subtraction to the total effect of different versions of the nested-counterfactuals according to several configurations of the treatment. For a setting in which X takes values x and x^*, we obtain two alternative ways to decompose the total effect (Robins and Greenland 1992; Pearl 2001; Robins 2003):

$$\begin{aligned} E[\, Y(\, x, W(x)\,) - Y(\, x^*, W(x^*)\,)\,] &= E[\, Y(\, x, W(x^*)\,) - Y(\, x^*, W(x^*)\,)\,] \\ &\quad + E[\, Y(\, x, W(x)\,) - Y(\, x, W(x^*)\,)\,] \\ &= E[\, Y(\, x, W(x)\,) - Y(\, x^*, W(x)\,)\,] \\ &\quad + E[\, Y(\, x^*, W(x)\,) - Y(\, x^*, W(x^*)\,)\,]. \end{aligned}$$

The first expected value on the right hand side of the first equality is called *Pure Direct Effect* (*PDE*), while the second one is called *Total Indirect Effect* (*TIE*). On the other hand, the first expected value on the right hand side of the second equality is called *Total Direct Effect* (*TDE*), while the second one is called *Pure Indirect Effect* (*PIE*).

Definition 1.3.2. The *Pure Direct Effect* describes the expected difference between the potential outcomes by moving the treatment from x^* to x, keeping, in both cases, the mediator at the value that it would naturally achieved if it had been not exposed to the treatment, i.e. under $X = x^*$:

$$PDE_{x,x^*} = E[\, Y(\, x, W(x^*)\,) - Y(\, x^*, W(x^*)\,)\,].$$

Notice that the pure direct effect is also known as the *Natural Direct Effect* (Pearl 2001). Following the illustrative example on the JOBS II study, *PDE* expresses by how much the prevalence of depressive symptoms would change if all unemployed workers' treatment assignment status were to be changed (i.e. not participate versus participate to the job training workshop), but their job search self-efficacy were to be fixed to whatever status would be observed if they had not originally assigned to the workshop (Steen and Vansteelandt 2018).

Definition 1.3.3. The *Total Indirect Effect* describes the expected difference between the potential outcomes that would obtained if the treatment had been set to x, but the mediator would be free to vary to its natural level by moving the treatment from the x^* to x:

$$TIE_{x,x^*} = E[\, Y(\, x, W(x)\,) - Y(\, x, W(x^*)\,)\,].$$

Notice that the total indirect effect is also known as the *Natural Indirect Effect* (Pearl 2001). Following the example, *TIE* denotes the expected change of depressive symptoms if all unemployed subjects were to be invited to participate in the job search program, but their job search self-efficacy were to be changed to whatever status would be observed if the subjects had participated in the workshop versus they had not participated (Steen and Vansteelandt 2018).

Definition 1.3.4. The *Total Direct Effect* describes the expected difference between the potential outcomes by moving the treatment from x^* to x, keeping, in both cases, the mediator at the value that it would be naturally achieved if it had been exposed to the treatment, i.e. under $X = x$:

$$TDE_{x,x^*} = E[\, Y(\, x, W(x)\,) - Y(\, x^*, W(x)\,)\,].$$

Following the example, *TDE* expresses by how much the prevalence of depressive symptoms would change if all unemployed workers' treatment assignment status were to be changed, but their job search self-efficacy were to be fixed to whatever status would be observed if they had originally assigned to the workshop.

Definition 1.3.5. The *Pure Indirect Effect* describes the expected difference between the potential outcomes that would obtained if the treatment had been exposed to x^*, but if the mediator would be free to vary to its natural level by moving the treatment from x^* to x:

$$PIE_{x,x^*} = E[\, Y(\, x^*, W(x)\,) - Y(\, x^*, W(x^*)\,)\,].$$

Following the example, *PIE* denotes the expected change of depressive symptoms if all unemployed subjects were not to be invited to participate in the job search workshop, but their job search self-efficacy were to be changed to whatever status would be observed if the subjects had participated in the workshop versus they had not participated.

The choice between pure or total (direct or indirect) effects is related to the different questions which we want to answer when conducting a mediation analysis. When a researcher aims to quantify the effect of the treatment explained through the intermediate variable then the total indirect effect should be considered. On the other hand, if the researcher wants to investigate the effect of the treatment which does not affect the mediator, then the pure direct effect should be taken into account. The pure (natural) direct effect and the total (natural) indirect effect are the most used effects in mediation literature. However, one may attempt to investigate the effect of the treatment that it would have had if it only acts on the mediator, keeping its effect on the outcome to the no-treatment value, that is the pure indirect effect. Finally, if the researcher wants to investigate the effect of the treatment on the outcome setting the mediator to its natural level that would be attained under the treatment, then the total direct effect should be taken into account (Hafeman and Schwartz 2009).

The difference between the two contrasts (total or pure) in the effects is mainly due to the phenomenon of interaction. As well-explained in VanderWeele (2015, p. 193), the term "total" in these effects refers to the fact that they include the interaction between the mediator and the treatment, while the term "pure" indicates that the effects do not consider the interaction. It has been shown that to isolate the effect of the interaction who operates through the direct or indirect effect, one may consider the difference *TDE*-*PDE* or *TIE*-*PIE* which results into a quantity called *Mediated Interaction* (INT^{med}), (VanderWeele 2013a).

Definition 1.3.6. The *Mediated Interaction* describes the expected difference between the total direct (indirect) effect and the pure direct (indirect) effect:

$$\begin{aligned} INT^{med}_{x,x^*} &= TDE_{x,x^*} - PDE_{x,x^*} \\ &= TIE_{x,x^*} - PIE_{x,x^*} \\ &= E[\, Y(\, x, W(x)\,) - Y(\, x^*, W(x)\,) + Y(\, x^*, W(x^*)\,) - Y(\, x, W(x^*)\,)\,]. \end{aligned}$$

As pointed out in VanderWeele (2015, pp. 194, 384), the mediated interaction is a type of mediated effect since it requires that the treatment changes the mediator; but it is also a direct effect as long as the effect of the treatment changes according to the different levels of the mediator. Thus, the mediated interaction solves the ambiguity in the choice of the two different ways to decompose the overall effect.

Combining the pure effects with the mediated interaction effect, we obtain the three-way decomposition of the total effect as derived by VanderWeele (2013a):

$$TE_{x,x^*} = PDE_{x,x^*} + PIE_{x,x^*} + INT^{med}_{x,x^*}.$$

Furthermore, VanderWeele (2014) derived the highest insight to the effect decomposition. He disentangled the pure direct effect into a component related to a direct effect totally explained by the treatment alone, i.e. without interference by mediator, and a residual component who is referred to the interaction effect. In doing so, one should consider the so-called *Controlled Direct Effect* (*CDE*) (Pearl 2001), which explains the effect of X on Y while keeping the mediator W to a reference level w^*.

Definition 1.3.7. The *Controlled Direct Effect* describes the expected difference between the potential outcomes by moving the treatment from x^* to x, while holding the mediator constant to a specific level for all entire population:

$$CDE_{x,x^*;w^*} = E\left[\, Y(x, w^*) - Y(x^*, w^*) \,\right].$$

The controlled direct effect provides a measure of the treatment effect on the outcome while keeping all other variables fixed to a specific value. Following the JOBS II example, the *CDE* expresses the difference in average of depressive symptoms if all the unemployed workers' treatment status were to be changed, but their job search self-efficacy were to be fixed for all subjects to a specific status, e.g. to level zero that is no belief in successfully performing particular job search behaviour and thus obtain employment.

There are various constraints concerning *CDE* (Pearl 2001; VanderWeele and Vansteelandt 2009). First, it is difficult to choose at what value we should have to fix W entirely for all subjects. Second, as in the linear case without interaction, one may be tempted to evaluate the indirect effect by taking the difference between the total effect and the controlled direct effect. However, this will not correspond to the indirect effect since it will be nonzero even if there is not effect transmitted by the treatment on the mediator, i.e. when the path $X \rightarrow W$ is null. The controlled direct effect is still a good measure to identify the direct effect of the treatment on the outcome when the mediator is absent and is widely use in many applied sciences.

Definition 1.3.8. The *Reference Interaction* describes the expected difference between the pure direct and the controlled direct effect:

$$\begin{aligned} INT^{ref}_{x,x^*;w^*} &= PDE_{x,x^*} - CDE_{x,x^*;w^*} \\ &= E\left[\, Y(\, x, W(x^*)\,) - Y(\, x^*, W(x^*) - Y(\, x, w^*\,) + Y(\, x^*, w^*\,) \,\right]. \end{aligned}$$

The reference interaction (INT^{ref}) requires the mediator to operate, but the effect does not come about by the treatment changing the mediator — it simply requires that the mediator itself is present even when the exposure is absent; the effect is "unmediated" in the sense that it does not operate by the treatment changing the mediator, but it requires the presence of the mediator nonetheless, as stated by VanderWeele (2015, p. 384).

Combining the principal components of the total effect we obtain the four-way decomposition (VanderWeele 2014), as follows:

$$TE_{x,x^*} = CDE_{x,x^*;w^*} + PIE_{x,x^*} + INT^{med}_{x,x^*} + INT^{ref}_{x,x^*;w^*}.$$

In Sec. 1.2, we have shown that for the identification and a causal interpretation of the total effect three important assumptions need to hold: temporal ordering, consistency and ignorability assumption. The mediation analysis in the counterfactual framework requires stronger assumptions.

Temporal ordering and Consistency Assumption

Also in the mediation setting, the events thought of cause should precede the events thought of their effects. The treatment precedes the mediator and the mediator precedes the outcome; see Fig. 1.1(c). On the other hand, the consistency assumption is referred to the hypothesis that the value of the potential outcome $Y(x)$ and the potential value of the mediator $W(x)$, corresponds, respectively, to the observed value Y and W when the units actually received the treatment $X = x$, i.e. $Y = Y(x)$ and $W = W(x)$ if $X = x$, for all levels of x. In addition, the value of the potential outcome $Y(x, w)$ corresponds to the observed value Y if the value of the treatment X actually equals to x and the value of the mediator W actually equals to w, i.e. $Y = Y(x, w)$ if $X = x$ and $W = w$, for all levels of x and w. Unfortunately, this assumption is not valid for the nested-counterfactual $Y(x, W(x^*))$. It is impossible to observe Y when actually X takes value x but W should be observed under value x^*.

Ignorability Assumption

To estimate the potential outcomes and the potential values of the mediator it is necessary to understand the procedures that determine when their probability distributions can be obtained from the observed conditional probability of Y and W. As we have shown in Sec. 1.2, the randomization procedure assures the independence between the potential outcome and the treatment. Alternatively, in observational studies this is carried out by conditioning on a set of pre-treatment covariates. However, in the mediation setting, the Ass. (1.2) and (1.4) are valid just to identify the overall effect, but the randomization procedure cannot be run on the mediator. Thus, as pointed out in Pearl (2001), Cole and Hernán (2002), and VanderWeele and Vansteelandt (2009), we need to adjust for all common cause of not only the treatment-outcome relationships but also the treatment-mediator as well as the mediator-outcome relationship. By conditioning on a set of covariates we are assuming to take into account all others

possible causes that could confound the mentioned relationships. To address the issue of unmeasured counfounding variables, we adopt the following stronger ignorability assumptions.

$$Y(x,w) \perp\!\!\!\perp X \mid C \qquad \forall x, w \text{ and } c, \tag{1.5}$$
$$Y(x,w) \perp\!\!\!\perp W \mid \{X, C\} \qquad \forall x, w \text{ and } c, \tag{1.6}$$
$$W(x) \perp\!\!\!\perp X \mid C \qquad \forall x \text{ and } c, \tag{1.7}$$
$$Y(x,w) \perp\!\!\!\perp W(x^*) \mid C \qquad \forall x, x^*, w \text{ and } c. \tag{1.8}$$

Following Pearl (2009b) and VanderWeele and Vansteelandt (2009), the ignorability assumption for the treatment-outcome and the mediator-outcome relationships, when we fix W at level w, are defined, respectively, in assumption (1.5) and (1.6). The assumption (1.7) refers to the treatment-mediator relationship, thus, conditional on C, the potential value of the mediator is independent on the treatment actually assigned. Notice that the assumptions (1.5) and (1.7) can also be ensured by the treatment randomization. Finally, the assumption (1.8) allows to ensure that there is no unmeasured confounders of the relationship mediator-outcome which are affected by the treatment. When the latter assumption is violated, the identification of natural effects (Pearl 2001; VanderWeele 2015; Steen and Vansteelandt 2018) becomes problematic, unless alternative approaches are introduced, like for example the multiple mediation analysis (Daniel et al. 2015; Steen et al. 2017a).

The above assumptions allow to derive from the nested counterfactual probability distribution of the outcome, i.e. $P(Y(x, W(x^*))$, the observed conditional probability distribution of Y given X and W.[6] It follows[7]:

$$P(Y(x, W(x^*)) = y \mid C = c) = \int_{-\infty}^{\infty} P\,(\,Y = y \mid X = x, W = w, C = c\,)\, dP\,(\,W = w \mid X = x^*\,)\,.$$

Therefore, under the above assumptions, we can correctly identify all components of the total effect from observational data.

By Definition 1.3.2 the pure direct effect can be non-parametrically identified as follows:

$$PDE_{x,x^*|c} = \int_{-\infty}^{\infty} \{\, E\,[\,Y \mid X = x, W = w, C = c\,] - E\,[\,Y \mid X = x^*, W = w, C = c\,]\,\} \; dP\,(\,W = w \mid X = x^*\,)\,. \tag{1.9}$$

Following the Definition 1.3.3 the total indirect effect can be non-parametrically obtained as follows:

$$TIE_{x,x^*|c} = \int_{-\infty}^{\infty} E\,[\,Y \mid X = x, W = w, C = c\,]\{\, dP(\,W = w \mid X = x\,) - dP\,(\,W = w \mid X = x^*\,)\,\}. \tag{1.10}$$

[6] See VanderWeele (2015, p. 465) to which we refer for the proof.
[7] For a discrete W one should replace integrals by summation for all level w of W.

Similarly, by Definition 1.3.4 the total direct effect can be non-parametrically obtained as:

$$TDE_{x,x^*|c} = \int_{-\infty}^{\infty} \{ E[\, Y \mid X = x, W = w, C = c\,] - E[\, Y \mid X = x^*, W = w, C = c\,] \} \\ dP(\, W = w \mid X = x\,), \tag{1.11}$$

while by Definition 1.3.5 the pure indirect effect can be non-parametrically identified as:

$$PIE_{x,x^*|c} = \int_{-\infty}^{\infty} E[\, Y \mid X = x^*, W = w, C = c\,] \{ dP(\, W = w \mid X = x\,) \\ - dP(\, W = w \mid X = x^*\,) \}. \tag{1.12}$$

As defined in Definition 1.3.6 the mediated interaction effect can be non-parametrically identified as:

$$INT^{med}_{x,x^*|c} = \int_{-\infty}^{\infty} \{ E[\, Y \mid X = x, W = w, C = c\,] - E[\, Y \mid X = x^*, W = w, C = c\,] \} \\ \{ dP(\, W = w \mid X = x\,) - dP(\, W = w \mid X = x^*\,) \}, \tag{1.13}$$

while the controlled direct effect, following the Definition 1.3.7, can be non-parametrically obtained as:

$$CDE_{x,x^*|w^*,c} = E[\, Y \mid X = x, W = w^*, C = c] - E[\, Y \mid X = x^*, W = w^*, C = c]. \tag{1.14}$$

Finally, as in Definition 1.3.8, the reference interaction can be non-parametrically identified as follows:

$$INT^{ref}_{x,x^*|w^*,c} = \int_{-\infty}^{\infty} \{ E[\, Y \mid X = x, W = w, C = c\,] - E[\, Y \mid X = x^*, W = w, C = c\,] \} \\ dP(\, W = w \mid X = x^*\,) - E[\, Y \mid X = x, W = w^*, C = c\,] \\ + E[\, Y \mid X = x^*, W = w^*, C = c\,] \}. \tag{1.15}$$

1.4 The decomposition of the total effect: the path analysis approach

Similarly to the previous sections, let X, W, Y be three random variables, respectively, the treatment, the mediator, and the outcome, the data generating process of which is displayed in Fig. 1.1(c). Let $M_{Y|X,W}$, $M_{W|X}$ and $M_{Y|X}$ be, respectively, the probabilistic model for Y given X and W, the probabilistic model for W given X, and the probabilistic model for Y given X, which arises as a consequence. Let $\theta_{Y|X,W}$ and $\theta_{W|X}$ be the corresponding vector of parameters of $M_{Y|X,W}$ and $M_{W|X}$. We define $\lambda_{Y|X,W}$ and $\lambda_{W|X}$ features of interest of $M_{Y|X,W}$ and $M_{W|X}$, either a scalar or a vector. Notice that $\lambda_{Y|X,W}$ and $\lambda_{W|X}$, respectively, are functions

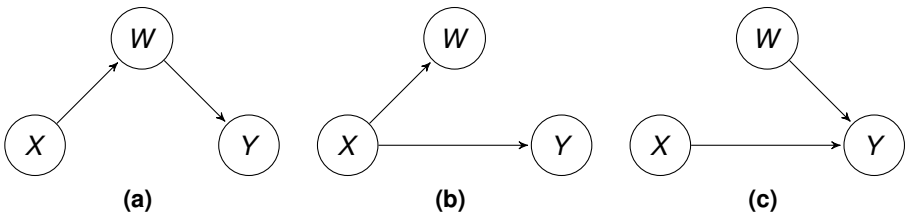


Figure 1.2: DAG when (a) $X \perp\!\!\!\perp Y \mid W$; (b) $W \perp\!\!\!\perp Y \mid X$ and (c) $W \perp\!\!\!\perp X$.

of $\theta_{Y|X,W}$ and $\theta_{W|X}$. Let $\lambda_{Y|X}$ the feature of interest in the marginal model $M_{Y|X}$. $\lambda_{Y|X}$ is a function of both $\theta_{Y|X,W}$ and $\theta_{W|X}$. In many applications λ could be the expected value as well as some dispersion or shape parameters.

In the following, we refer to the marginal and conditional independence notation $A \perp\!\!\!\perp B$ and $C \perp\!\!\!\perp B \mid A$ (Dawid 1979), where relating to path analysis it means that the edges from A to B and from C to B given A are removed, i.e. the related path coefficients are null. Furthermore, we follow the results of Ma et al. (2006) and Xie et al. (2008), who derive sufficient conditions for the collapsibility of an association measure. To be specific, they have considered the dependence between X and Y, taking W as a background variable. Notice that there are several types of collapsibility one might investigate (Whittaker 2009). We here refer to the simple collapsibility of an association measure (Ma et al. 2006; Xie et al. 2008).

In this context, we say that $\theta_{Y|X}$ is collapsible over W if $\theta_{Y|X,W}$ remains unchanged for all levels w of W and it equals to $\theta_{Y|X}$ obtained after marginalizing over W. As demonstrated by Ma et al. (2006, *Theorem 1*), letting θ be a dependence measure, if $W \perp\!\!\!\perp Y \mid X$ then $\theta_{Y|X,W} = \theta_{Y|X}$, and this is a sufficient condition for the collapsibility of $\theta_{Y|X}$. Therefore, in what follows we impose a similar condition on $\lambda_{Y|X}$, i.e. if $W \perp\!\!\!\perp Y \mid X$ then $\lambda_{Y|X,W} = \lambda_{Y|X}$.

Total, direct, indirect and residual effect

In this subsection, we offer new insights on the decomposition of the total effect. We propose a novel decomposition which can be easily related to the path-analysis and conditional independence assumptions. This decomposition does not require counterfactuals entities. However, if the unconfoundness assumptions of the mediation setting hold, the effects derived might be interpreted as causal. In other terms if the recursive systems are structural and represented by specific probabilistic causal models, then one may assert causal interpretations.[8]

We identify three components of the total effect: *i*) one referred to the direct effect alone, *ii*) one referred to the indirect effect alone, which under well-specified models brings back to the traditional product method as described in Sec. 1.1, and *iii*) one referred to a residual effect that vanishes under specific conditions. The approach we propose holds for linear as well as for non-linear models.

[8] See Pearl (2009b, Ch.7).

Definition 1.4.1. The *Total Effect* of the treatment X on the outcome Y describes the difference between $\lambda_{Y|X}$ evaluated under two different conditions of X.

$$TE_{x,x^*} = \lambda_{Y|X=x} - \lambda_{Y|X=x^*}.$$

In path analysis, TE corresponds to the path coefficient $X \to Y$ of Fig. 1.1(a) attributable to the probabilistic model of Y against X, which is obtained after marginalizing over W the conditional probabilistic model of Y against X and W, as displayed in Fig. 1.1(c).

Notice that the definitions of the direct and indirect effect below are based on the following reasoning: by looking at the graphs in Fig. 1.2, we can obtain the indirect path of X on Y through W by removing the directed arrow between X and Y. This is possible only assuming $X \perp\!\!\!\perp Y \mid W$, see Fig. 1.2(a). On the other hand, to obtain the direct path from X to Y not transmitted by W we need to interrupt the path from X to Y via W. This is possible under two different conditions: $W \perp\!\!\!\perp Y \mid X$, see Fig. 1.2(b); or $X \perp\!\!\!\perp W$, see Fig. 1.2(c). Notice that only the first condition $W \perp\!\!\!\perp Y \mid X$, see Fig. 1.2(b), holds for the sufficiency of a collapsible dependence measure. This idea conducts us to propose the following definitions.

Definition 1.4.2. The *Indirect Effect* of the treatment X on the outcome Y is defined as the part of the total effect after assuming that X is conditional independent to Y given W:

$$IE_{x,x^*} = TE_{x,x^*} \mid_{\{X \perp\!\!\!\perp Y|W\}}.$$

IE corresponds to a combination of the path coefficients $X \to W \to Y$, as displayed in Fig. 1.2(a). It is evaluated after interrupting the path $X \to Y$ in the model for Y given X and W. Thus, the effect of X on Y can be transmitted only through the mediator W. The indirect effect is null whenever $X \perp\!\!\!\perp W$ and/or $W \perp\!\!\!\perp Y \mid X$.

Definition 1.4.3. The *Direct Effect* of the treatment X on the outcome Y is defined as the part of the total effect after assuming that W is independent of Y given X:

$$DE_{x,x^*} = TE_{x,x^*} \mid_{\{W \perp\!\!\!\perp Y|X\}}.$$

DE corresponds to the path coefficients $X \to Y$, as displayed in Fig. 1.2(b). It is evaluated after interrupting the path $W \to Y$. Thus, the effect of X on Y cannot be transmitted through the mediator W. The direct effect is null whenever $X \perp\!\!\!\perp Y \mid W$.

Definition 1.4.4. The *Residual Effect* is defined as the difference between the total effect (Def. 1.4.1), direct (Def. 1.4.3) and indirect effect (Def. 1.4.2):

$$RE_{x,x^*} = TE_{x,x^*} - IE_{x,x^*} - DE_{x,x^*}.$$

RE is a remaining term which vanishes as soon as $X \perp\!\!\!\perp Y \mid W$ and/or $W \perp\!\!\!\perp Y \mid X$. Notice that when $X \perp\!\!\!\perp W$ the residual term is nonzero and takes into account the violation of the collapsibility condition.

Finally, given the definitions above we can rewrite the total effect as a sum of its three components, thus:

$$TE_{x,x^*} = IE_{x,x^*} + DE_{x,x^*} + RE_{x,x^*}.$$

On the other hand, one might consider the non-collapsibility effect as a component of the direct effect by assuming the independence $W \perp\!\!\!\perp X$. This means removing the edge between the path $X \to W$ and considering the remaining paths imputable to the direct effect from X to Y. The alternative decomposition of the total effect is given by:

Definition 1.4.5. The *Direct Effect*^{nc} of the treatment X on the outcome Y is defined as the part of the total effect after assuming that W is independent of X:

$$DE^{nc}_{x,x^*} = TE_{x,x^*} \,|_{\{W \perp\!\!\!\perp X\}} .$$

DE^{nc} still involves the path coefficient $X \to Y$, as displayed in Fig. 1.2(b). It is evaluated after interrupting the path $X \to W$. Thus, the effect of X on Y cannot be transmitted through the mediator W but it does not ensure that the sufficient condition to avoid the non-collapsibility of a measure of association is fulfilled. The direct effect still vanishes whenever $X \perp\!\!\!\perp Y \mid W$. As we will see in the next sections, the direct effect defined above can take two different parametric forms depending on which indirect path is disconnected.

Definition 1.4.6. The *Residual Effect*^{nc} is defined as the difference between the total effect (Def. 1.4.1), direct (Def. 1.4.5) and indirect effect (Def. 1.4.2):

$$RE^{nc}_{x,x^*} = TE_{x,x^*} - IE_{x,x^*} - DE^{nc}_{x,x^*}.$$

In this case, RE^{nc} is a remaining term which vanishes as soon as $X \perp\!\!\!\perp Y \mid W$ and/or $W \perp\!\!\!\perp Y \mid X$ and/or $W \perp\!\!\!\perp X$.

Therefore, given the definitions above we can rewrite the total effect as a sum of its two new components:

$$TE_{x,x^*} = IE_{x,x^*} + DE^{nc}_{x,x^*} + RE^{nc}_{x,x^*}.$$

Notice that the above setting easily generalizes to the inclusion of a set of covariates C.

1.5 The parametric form of the total effect and its components

Under a parametric point of view, we aim to identify and compare the quantities presented in the previous sections. For simplicity, we assume the ideal situation where there are no unmeasured confounders among all relationships under analysis. Thus, the data generating process is displayed in Fig. 1.1(c).

In particular, let the outcome Y and the mediator W be two continuous random variables, and let the treatment X be a discrete random variable. Let us consider Y be a function of X, W and a stochastic error term ϵ_y; W be a function of X and a stochastic error term ϵ_w; X be function of a stochastic error term ϵ_x. Let us also assume that $\epsilon_i \perp\!\!\!\perp \epsilon_j \, \forall \, i, j = \{y, x, w\}, \; i \neq j$. To obtain the parametric expression of the total effect and its components we consider, respectively, the conditional expectation of Y given X and W and the conditional expectation of W given X, as follows:

$$E[\, Y \mid X = x, W = w \,] = \beta_0 + \beta_x x + \beta_w w + \beta_{xw} xw \tag{1.16}$$

and

$$E[\, W \mid X = x \,] = \gamma_0 + \gamma_x x. \tag{1.17}$$

By the law of the total expectation we obtain the expected value of Y given X marginalized over W, as follows:

$$\begin{aligned} E[\, Y \mid X = x \,] &= E_{W|X}[\, E[\, Y \mid X = x, W = w \,] \,] \\ &= \beta_0 + \beta_w \gamma_0 + (\, \beta_x + \beta_w \gamma_x + \beta_{xw} \gamma_0 + \beta_{xw} \gamma_x x \,)\, x. \end{aligned} \tag{1.18}$$

One might also consider the second moment of the distribution of Y given X marginalized over W. Thus, by the law of total variance we obtain that:

$$\begin{aligned} V(\, Y \mid X = x \,) &= E_{W|X}[\, V(\, Y \mid X = x, W = w \,) \,] + V_{W|X}(\, E[Y \mid X = x, W = w \,]\,) \\ &= V(\, Y \mid X = x, W = w \,) + (\, \beta_w + \beta_{xw} x \,)^2 \, V(\, W = w \mid X = x \,). \end{aligned} \tag{1.19}$$

Notice that the marginal variance in Eq. (1.19) is also function of the conditional parameters and the treatment X. As a consequence, the variance of the marginal model is not constant.

Starting from the effects definition given in Sec. 1.2 and1.3, and using the conditional models in Eqs. (1.16), (1.17) and (1.18), we can identify the parametric effects based on the counterfactual approach. From Definition 1.2.2 and Equation (1.18), the total effect is parametrically identified as

$$TE_{x,x^*} = (\, \beta_x + \beta_w \gamma_x + \beta_{xw} \gamma_0 + \beta_{xw} \gamma_x x + \beta_{xw} \gamma_x x^* \,)\,(\, x - x^* \,). \tag{1.20}$$

Following Equation (1.9) and models (1.16) and (1.17), the pure direct effect can be obtain as follows

$$PDE_{x,x^*} = (\, \beta_x + \beta_{xw} \gamma_0 + \beta_{xw} \gamma_x x^* \,)\,(\, x - x^* \,), \tag{1.21}$$

whereas by Equation (1.10) and models (1.16) and (1.17), the total indirect effect is

$$TIE_{x,x^*} = (\, \beta_w \gamma_x + \beta_{xw} \gamma_x x \,)\,(\, x - x^* \,). \tag{1.22}$$

Likewise, by Equations (1.11) and models (1.16) and (1.17), the total direct effect is parametrically identified such as

$$TDE_{x,x^*} = (\, \beta_x + \beta_{xw} \gamma_0 + \beta_{xw} \gamma_x x \,)\,(\, x - x^* \,), \tag{1.23}$$

while by Eq. (1.12) and models (1.16) and (1.17), the pure indirect effect is as follows

$$PIE_{x,x^*} = (\beta_w\gamma_x + \beta_{xw}\gamma_x x^*) (x - x^*). \tag{1.24}$$

As obtained in Eq. (1.13) and following models (1.16) and (1.17), the mediated interaction effect is parametrically identified as

$$INT^{med}_{x,x^*} = (\beta_{xw}\gamma_x) (x - x^*)(x - x^*), \tag{1.25}$$

while the controlled direct effect, following the Equation (1.14) and models (1.16) and (1.17), can be identified as

$$CDE_{x,x^*|w^*} = (\beta_x + \beta_{xw} w^*) (x - x^*). \tag{1.26}$$

Finally, as in Eq. (1.15), the reference interaction is parametrically identified as follows

$$INT^{ref}_{x,x^*|w^*} = (\beta_{xw}\gamma_0 + \beta_{xw}\gamma_x x^* - \beta_{xw} w^*) (x - x^*), \tag{1.27}$$

On the other hand, given the effects' definitions in the path analysis approach (Sec. 1.4), and relating $\lambda_{Y|X}$ to the expected value in Eq. (1.18), the total effect can be parametrically identified as follows:

$$\begin{aligned} TE_{x,x^*} &= E[\, Y \mid X = x \,] - E[\, Y \mid X = x^* \,] \\ &= (\beta_x + \beta_w\gamma_x + \beta_{xw}\gamma_0 + \beta_{xw}\gamma_x x + \beta_{xw}\gamma_x x^*) (x - x^*). \end{aligned} \tag{1.28}$$

Clearly, it coincides with the parametric form of the total effect in Eq. (1.20).

The indirect effect, following the Def. 1.4.2, is identified by imposing in the total effect that all parameters of X in the model of Y, given X and W, are zero, since we assume $X \perp\!\!\!\perp Y \mid W$:

$$IE_{x,x^*} = TE_{x,x^*} \,|_{\beta_x=\beta_{xw}=0} = \beta_w\gamma_x(x - x^*). \tag{1.29}$$

Notice that the indirect effect coincides with the traditional product method. It is zero if and only if $\beta_w = 0$ (i.e. $W \perp\!\!\!\perp Y \mid X$) or $\gamma_x = 0$ (i.e. $X \perp\!\!\!\perp W$) or both are zero.

Following the Def. 1.4.3, the direct effect is identified by imposing in the total effect that all parameters of W in the model of Y, given X and W, are zero, since we assume $W \perp\!\!\!\perp Y \mid X$:

$$DE_{x,x^*} = TE_{x,x^*} \,|_{\beta_w=\beta_{xw}=0} = \beta_x(x - x^*). \tag{1.30}$$

Notice that the direct effect in Eq.(1.30) equals the controlled direct effect in Eq.(1.26) when $W = 0$. Clearly, the direct effect is zero if $\beta_x = 0$ (i.e. $Y \perp\!\!\!\perp X \mid W$).

As explained in Sec.1.4, when a measure of dependence is not collapsible, the direct and indirect effect do not sum up to the total effect and a residual term remains. The parametric form of the residual effect, as in Def. 1.4.4, is given as follows:

$$\begin{aligned} RE_{x,x^*} &= TE_{x,x^*} - IE_{x,x^*} - DE_{x,x^*} \\ &= (\beta_{xw}\gamma_0 + \beta_{xw}\gamma_x x + \beta_{xw}\gamma_x x^*) (x - x^*). \end{aligned} \tag{1.31}$$

We immediately observe that the residual effect is governed by the interaction terms between X and W in the model of Y, i.e. it is zero when $\beta_{xw} = 0$, thus bringing back to the traditional approach of mediation analysis. The residual effect tells us that even in the absence of the indirect effect due to the condition $X \perp\!\!\!\perp W$ the total effect and the direct effect of X on Y are different, proving again the results on the non-collapsibility of a dependence measure.

It then follows that unlike the linear case without interaction, the indirect effect cannot be identified by the difference between the total effect and the direct effect (or alternatively the controlled direct effect in $W = 0$). Indeed, we find:

$$TE_{x^*,x} - DE_{x^*,x} = (\beta_w\gamma_x + \beta_{xw}\gamma_0 + \beta_{xw}\gamma_x x + \beta_{xw}\gamma_x x^*)(x - x^*). \tag{1.32}$$

We can easily verify that this quantity cannot be imputable to the indirect effect since it is different from zero when $\gamma_x = 0$ (i.e. if $X \perp\!\!\!\perp W$). In mediation literature this term is also known as *proportion eliminated* (*PE*) (Robins and Greenland 1992; Pearl 2001; VanderWeele 2013b) and is defined as residual effect of the treatment on the outcome if we set the mediator to zero. Notice that $PE = PIE + INT^{ref} + INT^{med}$.

Alternatively, following the Def. 1.4.5, the direct effect is identified by imposing in the total effect that all parameters of X in the model of W are zero, since we assume $W \perp\!\!\!\perp X$:

$$DE^{nc}_{x,x^*} = TE_{x,x^*} \mid_{\gamma_x=0} = (\beta_x + \beta_{xw}\gamma_0)(x - x^*). \tag{1.33}$$

Notice that in this case the direct effect in Eq.(1.33) does not equal the controlled direct effect in Eq.(1.26). Clearly, the direct effect is zero if $\beta_x = \beta_{xw} = 0$ (i.e. $Y \perp\!\!\!\perp X \mid W$). Notice also that under the condition $X \perp\!\!\!\perp W$, i.e. no indirect effect, the direct effect is not equal to the total effect, due to the non-collapsibility effect.

The residual effect (Def. 1.4.6) is given by:

$$\begin{aligned} RE^{nc}_{x,x^*} &= TE_{x,x^*} - IE_{x,x^*} - DE^{nc}_{x,x^*} \\ &= (\beta_{xw}\gamma_x x + \beta_{xw}\gamma_x x^*)(x - x^*). \end{aligned} \tag{1.34}$$

In this case, we can observe that the residual effect is governed by the interaction terms between X and W in the model of Y, i.e. it is zero when $\beta_{xw} = 0$, and by the effect of X on W, i.e. it vanishes when $\gamma_x = 0$.

Similar to MacKinnon et al. (2020), we can highlight the link between the two approaches presented above. Let consider a binary treatment X taking values $x = 1$, $x^* = 0$ and a reference value for the mediator W to $w^* = 0$. Under this setting, it can be easily shown that the controlled direct effect while keeping $W = 0$ equals the total effect after assuming the conditional independence between the mediator and the outcome given the treatment, i.e. if $W \perp\!\!\!\perp Y \mid X$ then $CDE_{1,0|0} = DE_{1,0} = \beta_x$. We can also notice that the pure natural direct effect equals the total effect after assuming the independence between the mediator and the treatment, i.e. if $X \perp\!\!\!\perp W$ then $PDE_{1,0} = DE^{nc}_{1,0} = \beta_x + \beta_{xw}\gamma_0$. This means that $INT^{ref}{}_{1,0|0} = DE^{nc}_{1,0} - DE_{1,0} = \beta_{xw}\gamma_x$. Likewise, the pure natural indirect effect equals the total effect after assuming that the treatment and the outcome are independent given the mediator, i.e. if $X \perp\!\!\!\perp Y \mid W$ then $PIE_{1,0} = IE_{1,0} = \beta_w\gamma_x$. Finally, the interaction mediated equals the

residual term built with the direct effect under $X \perp\!\!\!\perp W$, i.e. $INT^{med}_{1,0} = RE^{nc}_{1,0} = \beta_{xw}\gamma_x$. On the other hand, if we take the sum of the interaction effects from the counterfactual approach thus we obtain the residual term in the path analysis approach, under the assumption $W \perp\!\!\!\perp Y \mid X$, i.e. $INT^{med}_{1,0} + INT^{ref}_{1,0|0} = RE_{1,0} = \beta_{xw}\gamma_0 + \beta_{xw}\gamma_x$.

So far, we have evaluated the effect of X on the expected value of Y. However, as we can notice from Eq. (1.19), since the variance is not constant but varies within the values of X, it seems, therefore, interesting to evaluate the effect of X on the marginal variance of Y by moving X from x^* to x. Thus:

$$V(Y \mid X = x) - V(Y \mid X = x^*) = \left(\beta_w\beta_{xw} + \beta^2_{xw}x^2 + \beta^2_{xw}x^{*2}\right) V(W \mid X = x)(x - x^*). \tag{1.35}$$

1.6 Discussion

In mediation literature there are two different approaches to evaluate the decomposition of the total effect into a direct and indirect effect. On one hand, we found the traditional approach, which allows to identify the effects by simply taking the sum of products between linear regression parameters. On the other hand, crucial works to mediation analysis have been offered by Robins and Greenland (1992) and Pearl (2001), who have given definitions of the effects in a counterfactual framework, which allow for a causal interpretation of the effects. Similarly, novel contributions have been offered by VanderWeele (2013a, 2014), who has provided the highest insight of the total effect's decomposition in the counterfactual framework.

In a setting with a continuous outcome and a continuous mediator, we have proposed a novel decomposition of the total effect by using the method of path analysis. The method proposed overcomes the issues of non-linearity, typical of structural equation modelling, and provides effects' definition for any type of models. Furthermore, a parametric comparison between the traditional, the counterfactual and the path analysis approaches has been offered. We have noticed that under the assumption of no-interaction the approaches bring back to the traditional one. Behind the linear case the traditional approach is no longer consistent to identify direct and indirect effect, thus, one should prefer either the counterfactual or the path analysis approach that we have proposed. In the next chapter we will generalize the previous derivations into a setting with a binary outcome and a binary mediator, modelled via recursive systems of logistic regressions.

Chapter 2

Mediation analysis in recursive systems of logistic regression models

2.1 Mediation analysis for binary outcome: an overview

In linear models, as shown in Chapter 1, the relationship between total, direct and indirect effects is well known. The product and the difference methods are equivalent in identifying the indirect effect (Wright 1921; Cochran 1938; Alwin and Hauser 1975; Baron and Kenny 1986; Bollen 1987). Nonetheless, this equivalence is lost as soon as nonlinearities are introduced in the system. This is due to the fact that the parameters of the original models link in a rather complex way to form the effect of X on Y only.

In non-linear models, the specific case of a binary outcome Y, modeled via logistic regression, has drawn the attention of many researchers. In this situation, the difficulty arises in the way the coefficients of the conditional model of Y given X and W, and W given X must come together to form the marginal model of Y given X. As recently investigated by Stanghellini and Doretti (2019), for X continuous or discrete, the marginal logistic model of Y against X is non-linear and a rather complex formula links the marginal and conditional effect of X on Y. Furthermore, as well-known, the parameters of logistic models are non-collapsible without additional conditional independence assumptions. In a setting with a binary outcome the notion of collapsibility becomes crucial to elucidate practical and conceptual notions about the identification of the effects.[1]

Within the traditional methods of mediation analysis, the decomposition of the total effect for binary outcome has been first addressed by MacKinnon and Dwyer (1993), and more recently by MacKinnon et al. (2007), Karlson et al. (2012), and Breen et al. (2013, 2018). Generally, those works have offered rescaled or standardized parameter solutions derived from latent variable models underlying logit and probit models, in a setting with a continuous

[1] See Section 1.4 for the definition of a collapsible measure of association in this context.

mediator, thus turning to modifications of the product method, to avoid to compare coefficients across different logit or probit models, that are not measured on the same scale.

In the counterfactual framework, the decomposition of the total effect for binary outcome has been tackled by VanderWeele and Vansteelandt (2010), Pearl (2012), and Valeri and VanderWeele (2013). Specifically, VanderWeele and Vansteelandt (2010) have proposed counterfactual definitions of the total, direct and indirect effect on the odds ratio scale (OR). However, their parametric identification are based on the rare outcome assumption, meaning that the outcome should be rare within all the strata formed by the treatment and the mediator, i.e. that $P(Y = 1 \mid X = x, W = w)$ is small for every (x, w) configuration.[2] VanderWeele and Vansteelandt (2010) have derived approximated parametric forms for the causal effects in a setting with a continuous mediator, while Valeri and VanderWeele (2013) have generalized to a setting with a binary mediator, both works assuming the rareness of the outcome. In those situations one can use the logarithmic function instead of the logistic function, corresponding in practice to approximate effects on the odds ratio scale to effects on the relative risk scale. When the outcome is not rare, such approximation is no longer valid, thus, they have suggested to use directly effects on the relative risk scale by fitting log-linear models. Nevertheless, this represents a limitation in many empirical data analyses since logistic regressions are considerably the most used models for a binary response.

To relax the rare outcome assumption, many authors have proposed alternative solutions (Vansteelandt et al. 2012; Tchetgen 2014; Steen et al. 2017b; Gaynor et al. 2018; Samoilenko et al. 2018). As an instance, Vansteelandt et al. (2012) and Steen et al. (2017b) use a class of models named "natural effect models" to estimate causal effects. These methods directly specify a regression model for the potential outcome probabilities, allowing the estimation via imputation or weighting procedures and without the requirement of the rare outcome assumption. However, they do not express the effects as a function of the regression coefficients. Conversely, Tchetgen (2014) replaces the rare outcome assumption with the assumption that a continuous mediator follows a so-called Bridge distribution, obtaining then closed-form expressions for the direct and indirect causal effects. The methods proposed, however, "rely on fairly strong modeling assumptions and can deliver severely biased inferences under modeling error of regression models" - as pointed out by Tchetgen (2014, p. 5), which as alternative solutions suggests to adopt semiparametrics models (Tchetgen and Shpitser 2012). On the other hand, Gaynor et al. (2018) suggest closed-form expressions based on the approximation of the probit to the logit model, for the setting with a continuous mediator, while for a binary mediator, they simply replace the probabilities which identify the effect on the odds ratio scale by the corresponding parametric logistic models. The same approach has been adopted also by Samoilenko et al. (2018), but without offering standard errors estimates for their effects.

In this Chapter, we provide parametric expressions for the counterfactual effects, on the log odds ratio scale – following the definitions in VanderWeele and Vansteelandt (2010) – which are not based on the rare outcome assumption, and thus, obtained in closed-form formulas. Like the approaches developed under the rare outcome assumption, the expressions proposed in this chapter emphasize the role of path-specific coefficients quite intuitively. They

[2] See discussion in Samoilenko et al. (2018), Samoilenko and Lefebvre (2018), and VanderWeele et al. (2018).

are easier to relate with the corresponding counterfactual effects and intuitively generalize to all components of the total effect (VanderWeele 2013a, 2014).

In the path analysis approach, we elaborate a proposal for the definition of the direct and indirect effect which is built on the log odds ratio scale and based on the exact parametric form of the marginal logistic model of Y given X. Starting from this parametric form, which depends on the nature of X, the effects are defined under conditional independence assumptions. This implies the suppression of the corresponding path-coefficients. The decomposition proposed under this approach also avoids fitting two nested models, thus sidestepping the issue of unequal variance (Winship and Mare 1983). We then propose a comparison between the counterfactual and the path analysis approach that allows to better understand the link between the effects in both frameworks. The identification of the effects in both approaches handles for every possible interaction in regression models, including those between the treatment and the mediator as well as the treatment, the mediator and the confounding covariates. These interactions were not previously considered in mediation literature.

2.2 The counterfactual decomposition of the total effect on the log odds scale

Similarly to Chapter 1, let us denote the outcome by the binary random variable Y, the mediator by the binary random variable W, and the treatment by the discrete random variable X. The DAG in Fig. 1.1(c) informs us with the data generating process. From the perspective of a counterfactual framework we briefly recall the potential outcome notation (Rubin 1974; Robins and Greenland 1992; Pearl 2009b; VanderWeele 2015). Let $Y(x)$ and $W(x)$ be, respectively, the potential values of the outcome and the mediator had the treatment been set to level x. Further, $Y(x, w)$ indicates the potential value of the outcome if X had been set to x and W to w. We retain the standard assumptions required to identify causal direct and indirect effects as exposed in Sec. 1.2 and 1.3. First, we assume a temporal ordering between variables, the treatment should precede the mediator which, in turn, should precede the outcome. Moreover, the consistency assumption states that $Y(x) = Y$ and $W(x) = W$ if $X = x$, and also that $Y(x, w) = Y$ if $X = x$ and $W = w$. On the other hand, the composition assumption requires 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 W to $W(x)$, which is the value it would have naturally attained under $X = x$, i.e. $Y(x) = Y(x, W(x))$. In the following we adopt the effects as defined in VanderWeele and Vansteelandt (2010) and Valeri and VanderWeele (2013) and generalize the definitions for the three-way and four-way decompositions on the log odds scale (VanderWeele 2013a, 2014).

Definition 2.2.1. For a treatment X changing from a reference level x^* to another level x, the *Total Effect* on the log odds scale is given by:

$$\begin{aligned}\log \mathrm{OR}^{TE}_{x,x^*} &= \log \frac{P\big(Y(x) = 1\big)}{P\big(Y(x) = 0\big)} - \log \frac{P\big(Y(x^*) = 1\big)}{P\big(Y(x^*) = 0\big)} \\ &= \log \frac{P\big(Y(x, W(x)) = 1\big)}{P\big(Y(x, W(x)) = 0\big)} - \log \frac{P\big(Y(x^*, W(x^*)) = 1\big)}{P\big(Y(x^*, W(x^*)) = 0\big)}.\end{aligned}$$

Notice that the first equality is given by definition, while the second one is derived by composition assumption. $\log \mathrm{OR}^{TE}_{x,x^*}$ describes the difference in the log odds for the outcome Y if the treatment X were set to level x and for the outcome Y if the treatment X were set to level x^*.

One solution to the decomposition of the total effect is to add and subtract from Def. 2.2.1 the log odds of the nested-counterfactual $Y(x, W(x^*))$, i.e. the potential value of the outcome that it would be observed if X had been set to x and the mediator had been set to the value that it would naturally attained if X had been set to x^*, i.e. $W(x^*)$.

Definition 2.2.2. The *Pure (Natural) Direct Effect* of X on Y on the log odds scale is defined by:

$$\log \mathrm{OR}^{PDE}_{x,x^*} = \log \frac{P\big(Y(x, W(x^*)) = 1\big)}{P\big(Y(x, W(x^*)) = 0\big)} - \log \frac{P\big(Y(x^*, W(x^*)) = 1\big)}{P\big(Y(x^*, W(x^*)) = 0\big)}.$$

The pure direct effect compares the log odds of Y if the treatment X had been set to x, against the log odds of Y if the treatment had been set to x^*, but in both cases the mediator were set to the value that it would have naturally observed if the treatment had been set to x^* (VanderWeele and Vansteelandt 2010).

Definition 2.2.3. The *Total (Natural) Indirect Effect* of X on Y on the log odds scale is defined as follows:

$$\log \mathrm{OR}^{TIE}_{x,x^*} = \log \frac{P\big(Y(x, W(x)) = 1\big)}{P\big(Y(x, W(x)) = 0\big)} - \log \frac{P\big(Y(x, W(x^*)) = 1\big)}{P\big(Y(x, W(x^*)) = 0\big)}.$$

The total indirect effect compares the log odds of Y if the treatment were set to x and the mediator to the value that it would have naturally obtained if the treatment had been set to x, against the log odds of Y if the exposure were set to x but the mediator to the value that it would have naturally obtained if the treatment had been set to x^* (VanderWeele and Vansteelandt 2010).

Clearly, the pure direct (Def. 2.2.2) and the total indirect (Def. 2.2.3) effects sum up to the total effect (Def. 2.2.1), such as:

$$\log \mathrm{OR}^{TE}_{x,x^*} = \log \mathrm{OR}^{PDE}_{x,x^*} + \log \mathrm{OR}^{TIE}_{x,x^*}.$$

On the other hand, one might be interested in decomposing the total effect by taking the opposite contrast of the nested-counterfactual $Y(x^*, W(x))$. This contrast is defined as the potential value of the outcome that it would be observed if X had been set to x^* and the mediator had been set to the value that it would naturally get if X had been set to x, i.e. $W(x)$.

Definition 2.2.4. The *Total Direct Effect* of X on Y on the log odds scale is defined as follows:

$$\log \mathrm{OR}^{TDE}_{x,x^*} = \log \frac{P\big(Y(x, W(x)) = 1\big)}{P\big(Y(x, W(x)) = 0\big)} - \log \frac{P\big(Y(x^*, W(x)) = 1\big)}{P\big(Y(x^*, W(x)) = 0\big)}.$$

The total direct effect compares the log odds of Y if the treatment X had been set to x, against the log odds of Y if the treatment had been set to x^*, but in both cases the mediator were set to the value that it would have naturally observed if the treatment had been set to x (VanderWeele and Vansteelandt 2010).

Definition 2.2.5. The *Pure Indirect Effect* of X on Y on the log odds scale is given by:

$$\log OR^{PIE}_{x,x^*} = \log \frac{P\big(Y(x^*, W(x)) = 1\big)}{P\big(Y(x^*, W(x)) = 0\big)} - \log \frac{P\big(Y(x^*, W(x^*)) = 1\big)}{P\big(Y(x^*, W(x^*)) = 0\big)}.$$

The pure indirect effect compares the log odds of Y if the treatment were set to x^* and the mediator to the value that it would have naturally obtained if the treatment had been set to x, against the log odds of Y if the exposure were set to x^* but the mediator to the value that it would have naturally obtained if the treatment had been set to x^* (VanderWeele and Vansteelandt 2010).

Even in this case, the total direct (Def. 2.2.4) and the pure indirect (Def. 2.2.5) effects sum up to the total effect (Def. 2.2.1), such as:

$$\log \mathrm{OR}^{TE}_{x,x^*} = \log \mathrm{OR}^{TDE}_{x,x^*} + \log \mathrm{OR}^{PIE}_{x,x^*}.$$

As already noticed for the difference scale, even in the log odds scale $TDE \neq PDE$ (or $TIE \neq PIE$). This difference results into a remaining term, which we still label as mediated interaction, although, in this context, it would be better to refer to the effect due to the non-linearity, in general terms. It is nonzero in the absence of interaction.

Definition 2.2.6. The *Mediated Interaction* of X on Y on the log odds scale is given by:

$$\begin{aligned}\log \mathrm{OR}^{INT^{med}}_{x,x^*} = &\log \frac{P\big(Y(x, W(x)) = 1\big)}{P\big(Y(x, W(x)) = 0\big)} - \log \frac{P\big(Y(x^*, W(x)) = 1\big)}{P\big(Y(x^*, W(x)) = 0\big)} \\ &- \log \frac{P\big(Y(x, W(x^*)) = 1\big)}{P\big(Y(x, W(x^*)) = 0\big)} + \log \frac{P\big(Y(x^*, W(x^*)) = 1\big)}{P\big(Y(x^*, W(x^*)) = 0\big)}.\end{aligned}$$

The quantity in Def. 2.2.6 brings out the way one might account for the non-linearity phenomenon in the counterfactual setting. As a consequence, we obtain the three-way decomposition on the log odds scale as follows:

$$\log \mathrm{OR}^{TE}_{x,x^*} = \log \mathrm{OR}^{PDE}_{x,x^*} + \log \mathrm{OR}^{PIE}_{x,x^*} + \log \mathrm{OR}^{INT^{med}}_{x,x^*}.$$

Definition 2.2.7. The *Controlled Direct Effect* of X on Y given W equals zero, on the log odds scale is defined as follows:

$$\log \mathrm{OR}^{CDE}_{x,x^*|W=0} = \log \frac{P\big(Y(x, 0) = 1\big)}{P\big(Y(x, 0) = 0\big)} - \log \frac{P\big(Y(x^*, 0) = 1\big)}{P\big(Y(x^*, 0) = 0\big)}.$$

Notice that in mediation literature a more general definition of controlled direct effect exists and it worth for all value w of W. For our purpose we consider only the case in which $W = 0$.

Definition 2.2.8. The *Reference Interaction* effect of X on Y on the log odds scale is defined as the difference between the pure natural direct effect (Def. 2.2.2) and the controlled direct effect (Def. 2.2.7), such as:

$$\log \mathrm{OR}^{INT^{ref}}_{x,x^*|W=0} = \log \frac{P\big(Y(x, W(x^*)) = 1\big)}{P\big(Y(x, W(x^*)) = 0\big)} - \log \frac{P\big(Y(x^*, W(x^*)) = 1\big)}{P\big(Y(x^*, W(x^*)) = 0\big)} - \log \frac{P\big(Y(x, 0) = 1\big)}{P\big(Y(x, 0) = 0\big)} + \log \frac{P\big(Y(x^*, 0) = 1\big)}{P\big(Y(x^*, 0) = 0\big)}.$$

Also in this case, as we will see, the quantity in Def. 2.2.8 would not totally represent the interaction phenomenon.

Finally, on the log odds scale, the four-way decomposition of the total effect of a discrete treatment X on a binary outcome Y with a single binary mediator W is given by:

$$\log \mathrm{OR}^{TE}_{x,x^*} = \log \mathrm{OR}^{CDE}_{x,x^*|W=0} + \log \mathrm{OR}^{INT^{ref}}_{x,x^*|W=0} + \log \mathrm{OR}^{INT^{med}}_{x,x^*} + \log \mathrm{OR}^{PIE}_{x,x^*}. \tag{2.1}$$

2.3 The path analysis decomposition of the total effect on the log odds scale

In the path analysis framework, the decomposition of the total effect on the log odds scale follows the same reasoning and definitions given in Ch. 1, Sec. 1.4. We here briefly recall the main notation. The Fig. 1.1(c) informs us with the data generating process of three random variables. Specifically, let us consider the binary outcome Y, the binary mediator W, and the discrete treatment X which changes from a reference level x^* to another level x. Let $M_{Y|X,W}$, $M_{W|X}$ and $M_{Y|X}$ be, respectively, the probabilistic model for Y given X and W, the probabilistic model for W given X, and the probabilistic model for Y given X, which arises as a consequence. Let $\theta_{Y|X,W}$ and $\theta_{W|X}$ be the corresponding vector of parameters of $M_{Y|X,W}$ and $M_{W|X}$. We define $\lambda_{Y|X,W}$ and $\lambda_{W|X}$ as features of interest of $M_{Y|X,W}$ and $M_{W|X}$, either a scalar or a vector. Notice that $\lambda_{Y|X,W}$ and $\lambda_{W|X}$ depend, respectively, on $\theta_{Y|X,W}$ and $\theta_{W|X}$. Let $\lambda_{Y|X}$ be the feature of interest in the marginal model $M_{Y|X}$. $\lambda_{Y|X}$ is a function of both $\theta_{Y|X,W}$ and $\theta_{W|X}$. In this setting, let $\lambda_{Y|X}$ be the log odds of the outcome Y given the treatment X.

Definition 2.3.1. The *Total Effect* represents the difference in the log odds of Y after moving X from x^* to x. That is:

$$\log \mathrm{OR}^{TE}_{x,x^*} = \lambda_{Y|X=x} - \lambda_{Y|X=x^*} = \log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} - \log \frac{P(Y = 1 \mid X = x^*)}{P(Y = 0 \mid X = x^*)}.$$

Definition 2.3.2. The *Indirect Effect* of X on Y on the log odds scale is defined by assuming in the total effect that $X \perp\!\!\!\perp Y \mid W$, thus:

$$\log \mathrm{OR}^{IE}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} \mid_{\{X \perp\!\!\!\perp Y|W\}}.$$

The indirect effect represents the effect of the treatment on the outcome totally attributable to the indirect path $X \to W \to Y$; see Fig. 1.2(a).

Definition 2.3.3. The *Direct Effect* of X on Y on the log odds scale is defined by assuming in the total effect that $W \perp\!\!\!\perp Y \mid X$, thus:

$$\log \mathrm{OR}^{DE}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} \mid_{\{W \perp\!\!\!\perp Y \mid X\}} .$$

The direct effect represents the effect of the treatment on the outcome totally attributable to the direct path $X \to Y$; see Fig. 1.2(b). As already pointed out in Sec. 1.4 of Ch. 1, the direct effect defined above is based on the sufficient condition to obtain a collapsible measure of association.[3]

Definition 2.3.4. The *Residual Effect* of X on Y on the log odds scale is defined by the difference between the total (Def. 2.3.1), the indirect (Def. 2.3.2) and the direct (Def. 2.3.3) effects, such as:

$$\log \mathrm{OR}^{RE}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} - \log \mathrm{OR}^{IE}_{x,x^*} - \log \mathrm{OR}^{DE}_{x,x^*}.$$

The residual term as defined above takes into account the effect due to the non-collapsibility. It is zero when $X \perp\!\!\!\perp Y \mid W$ and/or under the condition of collapsibility, i.e. $W \perp\!\!\!\perp Y \mid X$, as we will prove in the next sections.

By construction, the decomposition of the total effect on the log odds ratio is given by:

$$\log \mathrm{OR}^{TE}_{x,x^*} = \log \mathrm{OR}^{IE}_{x,x^*} + \log \mathrm{OR}^{DE}_{x,x^*} + \log \mathrm{OR}^{RE}_{x,x^*}. \tag{2.2}$$

Conversely, one might consider an alternative way to define the direct effect.

Definition 2.3.5. The *Direct Effect*^{nc} of X on Y on the log odds scale is determined by assuming in the total effect that $W \perp\!\!\!\perp X$, such as:

$$\log \mathrm{OR}^{DE^{nc}}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} \mid_{\{W \perp\!\!\!\perp X\}},$$

where the superscript *nc* in Def. 2.3.5 stands for *non-collapsibility*. The independence assumption $X \perp\!\!\!\perp W$ removes the path $X \to W$, thus also dismissing the indirect effect; see Fig. 1.2(c). However, as the condition to avoid the non-collapsibility does not hold, the effect due to the non-collapsibility is here absorbed by the direct effect, leading to two different possible values of the direct effect, as shown in the next section.

Definition 2.3.6. The *Residual Effect*^{nc} of X on Y on the log odds scale is defined by the difference between the total (Def. 2.3.1), the indirect (Def. 2.3.2) and the direct (Def. 2.3.5) effects, thus:

$$\log \mathrm{OR}^{RE^{nc}}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} - \log \mathrm{OR}^{IE}_{x,x^*} - \log \mathrm{OR}^{DE^{nc}}_{x,x^*}.$$

As we will see in Sec. 2.4, in this case, the residual effect vanishes as soon as at least one of the paths from X to Y is dismissed.

Finally, by construction, the total effect can be decomposed as follows:

$$\log \mathrm{OR}^{TE}_{x,x^*} = \log \mathrm{OR}^{IE}_{x,x^*} + \log \mathrm{OR}^{DE^{nc}}_{x,x^*} + \log \mathrm{OR}^{RE^{nc}}_{x,x^*}. \tag{2.3}$$

[3] See Ma et al. (2006, *Theorem 1*).

2.4 Exact parametric form of the total effect and its components

We first concentrate on a very simple model with a binary outcome Y, a binary mediator W and a treatment X, that can be either discrete or continuous; see Fig. 1.1(c). Let the postulated models be, respectively, a logistic regression for W given X, and for Y given X and W such as:

$$\log \frac{P(W = 1 \mid X = x)}{P(W = 0 \mid X = x)} = \gamma_0 + \gamma_x x \tag{2.4}$$

and

$$\log \frac{P(Y = 1 \mid X = x, W = w)}{P(Y = 0 \mid X = x, W = w)} = \beta_0 + \beta_x x + \beta_w w + \beta_{xw} xw. \tag{2.5}$$

Notice that we allow for the interaction between X and W in the outcome equation (2.5).

Similarly to Ch. 1, by the law of the total probability, the conditional probability of Y given X (marginal over W) can be written as a function of both the conditional probability of Y given X and W, and the conditional probability of W given X, thus:

$$P(Y = y \mid X = x) = \sum_w P(Y = y \mid X = x, W = w) P(W = w \mid X = x)$$

with $w = \{0, 1\}$. After some algebra, we can obtain the logistic regression model of Y given X by the following exact parametric form:

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = \beta_0 + \beta_x x + \log A_{x,x}, \tag{2.6}$$

with

$$A_{x,x} = \frac{\exp(\beta_w + \beta_{xw} x) \exp(\gamma_0 + \gamma_x x)(1 + \exp(\beta_0 + \beta_x x)) + 1 + \exp(\beta_0 + \beta_x x + \beta_w + \beta_{xw} x)}{\exp(\gamma_0 + \gamma_x x)(1 + \exp(\beta_0 + \beta_x x)) + 1 + \exp(\beta_0 + \beta_x x + \beta_w + \beta_{xw} x)}. \tag{2.7}$$

We here refer to the first subscript of A as the value x of X in the model of Y given X and W, while the second one refers to the value x of X in the model of W given X. The double subscript in the A term will have a useful and specific function especially in the counterfactual framework. We immediately notice that the $A_{x,x}$ vanishes as soon as β_w and β_{xw} are zero, for all configuration of x, i.e. when $W \perp\!\!\!\perp Y \mid X$. Notice that the result in Eq. (2.6) has been previously obtain by Lin et al. (1998). Stanghellini and Doretti (2019) proved that $A_{x,x}$ corresponds to the inverse of the relative risk (RR) of $\bar{W} = W - 1$ for varying Y given $X = x$. We refer to the Appendix A.1 for all details.

The extension of the effects presented above to the parametric inclusion of covariates $C = (C_1, ..., C_k)$ is immediate; see Figure 1.1(d). Let us here consider an example for $k = 1$.

In detail, if Equations (2.5) and (2.4) are modified to account for an additional covariate C and for all their possible interactions, we obtain:

$$\log \frac{P(Y=1 \mid X=x, W=w, C=c)}{P(Y=0 \mid X=x, W=w, C=c)} = \beta_0 + \beta_x x + \beta_w w + \beta_c c + \beta_{xw} xw + \beta_{xc} xc + \beta_{wc} wc + \beta_{xwc} xwc \tag{2.8}$$

and

$$\log \frac{P(W=1 \mid X=x, C=c)}{P(W=0 \mid X=x, C=c)} = \gamma_0 + \gamma_x x + \gamma_c c + \gamma_{xc} xc. \tag{2.9}$$

It follows that the model of Y given X and C after marginalizing over W can be written as

$$\log \frac{P(Y=1 \mid X=x, C=c)}{P(Y=0 \mid X=x, C=c)} = \beta_0 + \beta_x x + \beta_c c + \beta_{xc} xc + \log A_{x,x|c}, \tag{2.10}$$

where the conditional version of (2.7) is given by

$$A_{x,x|c} = \frac{\exp(\beta_w + \beta_{xw} x + \beta_{wc} c + \beta_{xwc} xc) e_w(x,c)\{1 + e_y(x,0,c)\} + 1 + e_y(x,1,c)}{e_w(x,c)\{1 + e_y(x,0,c)\} + 1 + e_y(x,1,c)}, \tag{2.11}$$

with

$$e_y(x,w,c) = \exp(\beta_0 + \beta_x x + \beta_w w + \beta_c c + \beta_{xw} xw + \beta_{xc} xc + \beta_{wc} wc + \beta_{xwc} xwc)$$
$$e_w(x,c) = \exp(\gamma_0 + \gamma_x x + \gamma_c c + \gamma_{xc} xc).$$

The proof follows immediately from the one of the simple case and is shown in Appendix A.2. Clearly, the definition and the parametric formula of the total effect and its components will vary within the level of C.

The identification of the natural effects in the counterfactual approach

In the counterfactual framework, in order to identify the causal effects on the log odds scale as defined in Sec. 2.2, specific assumptions are needed. Let us recall the assumptions given in Sec. 1.3 of Ch. 1. Conditioning on a set of covariates $C = (C_1, ..., C_k)$ we have to ensure that $Y(x,w) \perp\!\!\!\perp X \mid C$ (Ass. 1.5), $Y(x,w) \perp\!\!\!\perp W \mid \{X, C\}$ (Ass. 1.6) and $W(x) \perp\!\!\!\perp X \mid C$ (Ass. 1.7), which have to hold for every level x and w. This guarantees that there are no unmeasured confounders between the exposure-outcome, the mediator-outcome and the exposure-mediator relationships. Similarly, we need to assume that $Y(x,w) \perp\!\!\!\perp W(x^*) \mid C$, for all x, $x^\star$ and w (Ass. 1.8), meaning in practice that none of the variables that guarantees the uncounfoundness of the mediator-outcome relationship may be affected by the treatment. Furthermore, contrary to what VanderWeele and Vansteelandt (2010), Valeri and VanderWeele (2013), and VanderWeele (2015), made, we do not make any assumption about the rareness of the outcome. In the following, under the simple setting without covariates, we derived exact parametric expressions for the counterfactual effects defined in Sec. 2.2, generalizing the approximated ones of Valeri and VanderWeele (2013). In what follows, we study in details the

decomposition of the total effect in the counterfactual framework, for a discrete treatment X and for the simple case without covariates. Results to a set of covariates C can be found in Appendix A.2.

The total effect, following Def. 2.2.1, is parametrically identified on the log odds scale as:

$$\log \mathrm{OR}_{x,x^*}^{TE} = \beta_x(x - x^*) + \log A_{x,x} - \log A_{x^*,x^*}. \tag{2.12}$$

The pure (natural) direct effect and total (natural) indirect effect, as defined in Defs. 2.2.2 and 2.2.3, are parametrically identified on the log odds scale, respectively, as follows:

$$\log \mathrm{OR}_{x,x^*}^{PDE} = \beta_x(x - x^*) + \log A_{x,x^*} - \log A_{x^*,x^*} \tag{2.13}$$

and

$$\log \mathrm{OR}_{x,x^*}^{TIE} = \log A_{x,x} - \log A_{x,x^*}. \tag{2.14}$$

On the other hand, the exact parametric expressions for the total direct and pure indirect effect, as defined in Defs. 2.2.4 and 2.2.5, are parametrically identified on the log odds scale, respectively, such as:

$$\log \mathrm{OR}_{x,x^*}^{TDE} = \beta_x(x - x^*) + \log A_{x,x} - \log A_{x^*,x} \tag{2.15}$$

and

$$\log \mathrm{OR}_{x,x^*}^{PIE} = \log A_{x^*,x} - \log A_{x^*,x^*}. \tag{2.16}$$

Following the Definition 2.2.6, the mediated interaction effect is identified as:

$$\log \mathrm{OR}_{x,x^*}^{INT^{med}} = \log A_{x,x} - \log A_{x^*,x} - \log A_{x,x^*} + \log A_{x^*,x^*}. \tag{2.17}$$

Finally, by Definition 2.2.7, the controlled direct effect in $W = 0$, is parametrically identified as:

$$\log \mathrm{OR}_{x,x^*|W=0}^{CDE} = \beta_x(x - x^*), \tag{2.18}$$

whereas the parametric form of the reference interaction effect, as defined in Def. 2.2.8, is given by:

$$\log \mathrm{OR}_{x,x^*|W=0}^{INT^{ref}} = \log A_{x,x^*} - \log A_{x^*,x^*}. \tag{2.19}$$

We refer to the Appendix A.3 for the proof of the mathematical derivations of the counterfactual effects. The parametric expressions derived above link the counterfactual effects definition to their pathway-specific regression parameters in a clear and exact form. As explained before, the subscripts in the A term help to distinguish the effects under estimation. As an instance, in A_{x,x^*} we known that the first subscript, i.e. x, corresponds to the value of X in the model for Y, while the second one, i.e. x^*, refers to the value of X in the model for

W (the so-called cross-worlds notation,Pearl (2001), VanderWeele and Vansteelandt (2010), and Steen and Vansteelandt (2018)). Therefore, the parametric form of A_{x,x^*} is:

$$A_{x,x^*} = \frac{\exp(\beta_w + \beta_{xw}x)\exp(\gamma_0 + \gamma_x x^*)(1 + \exp(\beta_0 + \beta_x x)) + 1 + \exp(\beta_0 + \beta_x x + \beta_w + \beta_{xw}x)}{\exp(\gamma_0 + \gamma_x x^*)(1 + \exp(\beta_0 + \beta_x x)) + 1 + \exp(\beta_0 + \beta_x x + \beta_w + \beta_{xw}x)}.$$

It is important to underline that A_{x,x^*} does not have the same statistical interpretation of $A_{x,x}$ or A_{x^*,x^*} since derived from the nested-counterfactual definitions. Finally, notice that A_{x,x^*} ($A_{x^*,x}$) equals one as soon as $\beta_w = \beta_{xw} = 0$. As a consequence the pure (2.16) and total (2.14) indirect effect are equal to zero, as well as the mediated interaction (2.17) and the reference interaction (2.19), whereas the pure (2.13), total (2.15) and controlled (2.18) direct effects are all equal to the conditional log odds ratio of X on Y given $W = 0$, i.e β_x. A deeper analysis of the parametric components of the total effect under some case of interest is given in Sec. 2.5.

The identification of the effects in the path analysis approach

In path analysis, the conditional independence assumptions necessary to identify the effect under analysis mean that the edges between nodes must be removed, i.e. the corresponding path-coefficients are zero. In the setting with a binary outcome and a binary mediator, let $\lambda_{Y|X,W}$, $\lambda_{W|X}$ and $\lambda_{Y|X}$ be, respectively, the logit function of Y given X and W (Eq. (2.5)), the logit function of W given X (Eq. (2.4)) and the logit function of Y and X after marginalizing over W (Eq. (2.6)). The results below are obtained for a discrete treatment X taking value x, x^*. Results for a continuous treatment in the path analysis approach are presented in Appendix A.4.

Following the Def. 2.3.1, the total effect of X on Y on the log odds scale is parametrically identified as follows:

$$\log \mathrm{OR}^{TE}_{x,x^*} = \beta_x(x - x^*) + \log \mathrm{A}_{x,x} - \log \mathrm{A}_{x^*,x^*}. \tag{2.20}$$

Clearly it equals the total effect in the counterfactual framework (2.12), meaning that under well-specified assumption the total effect in the path analysis approach can be interpreted as a causal effect.

The indirect effect on the log odds scale, according to the Def. 2.3.2, is defined by imposing in the total effect that all parameters of X in the model of Y given X and W are zero, i.e. $\beta_x = \beta_{xw} = 0$, thus:

$$\log \mathrm{OR}^{IE}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} \mid_{\beta_x=\beta_{xw}=0} = \log \mathrm{A}^*_{x,x} - \log \mathrm{A}^*_{x^*,x^*}, \tag{2.21}$$

where $\mathrm{A}^*_{x,x}$ is Eq. (2.7) evaluated in $\beta_x = \beta_{xw} = 0$, that is:

$$\mathrm{A}^*_{x,x} = \frac{\exp(\beta_w)\exp(\gamma_0 + \gamma_x x)(1 + \exp(\beta_0)) + 1 + \exp(\beta_0 + \beta_w)}{\exp(\gamma_0 + \gamma_x x)(1 + \exp(\beta_0)) + 1 + \exp(\beta_0 + \beta_w)}.$$

Notice that the indirect effect on the log odds scale does not coincide with the traditional product method, i.e. $\beta_w\gamma_x(x - x^*)$. However, it can be shown that when $\beta_w = 0$ then $A^*_{x,x} = 1$ and $A^*_{x^*,x^*} = 1$, while when $\gamma_x = 0$ then $A^*_{x,x} = A^*_{x^*,x^*}$. In both cases the indirect effect vanishes.

According to the Def. 2.3.3, the direct effect on the log odds scale is defined by imposing in the total effect that all parameters of W in the model of Y given X and W are zero, i.e. $\beta_w = \beta_{xw} = 0$. Thus:

$$\log \mathrm{OR}^{DE}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} \mid_{\beta_w=\beta_{xw}=0} = \beta_x(x - x^*). \tag{2.22}$$

As explained above, when $\beta_w = \beta_{xw} = 0$, then $A_{x,x}$ equals one $\forall x$. Notice that the direct effect in Eq. (2.22) equals the controlled direct effect in Eq. (2.18) under $W = 0$.

The parametric form of the residual effect as in Def. 2.3.4 is given by:

$$\log \mathrm{OR}^{RE}_{x,x^*} = \log \mathrm{A}_{x,x} - \log \mathrm{A}_{x^*,x^*} - \log \mathrm{A}^*_{x,x} + \log \mathrm{A}^*_{x^*,x^*}. \tag{2.23}$$

Given the explicit forms of the A terms, we can observe that this effect is driven by β_x, β_w and β_{xw}, meaning that it is zero as soon as $X \perp\!\!\!\perp Y \mid W$ and/or $W \perp\!\!\!\perp Y \mid X$. As pointed out in Sec. 2.3, when $\gamma_x = 0$, i.e. in $X \perp\!\!\!\perp W$ only, the residual effect is nonzero. Thus, the total and the direct effect are different, in line with the results on the non-collapsibility of the odds ratio.

Alternatively, according to the Def. 2.3.5, the direct effect on the log odds scale is defined by imposing in the total effect that all the parameters of X in the model of W are zero, i.e. $\gamma_x = 0$. Thus:

$$\log \mathrm{OR}^{DE^{nc}}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} \mid_{\gamma_x=0} = \beta_x(x - x^*) + \log \mathrm{A}^{nc}_{x,x} - \log \mathrm{A}^{nc}_{x^*,x^*}, \tag{2.24}$$

where $\mathrm{A}^{nc}_{x,x}$ is Eq. (2.7) evaluated in $\gamma_x = 0$, that is:

$$\mathrm{A}^{nc}_{x,x} = \frac{\exp(\beta_w + \beta_{xw}x)\exp(\gamma_0)(1 + \exp(\beta_0 + \beta_x x)) + 1 + \exp(\beta_0 + \beta_x x + \beta_w + \beta_{xw}x)}{\exp(\gamma_0)(1 + \exp(\beta_0 + \beta_x x)) + 1 + \exp(\beta_0 + \beta_x x + \beta_w + \beta_{xw}x)}.$$

Notice that the effect in Eq. (2.24) can be seen as the direct effect in Eq. (2.22) plus a residual term given by the difference between the $\log \mathrm{OR}^{DE^{nc}}_{x,x^*}$ and $\log \mathrm{OR}^{DE}_{x,x^*}$, thus embracing the effect due to the non-collapsibility. This is similar to what is done between the pure natural direct effect and the controlled direct effect, the difference of which gives rise to the reference interaction as in Eq. (2.19).

Finally, by Def. 2.3.6 the residual effect on the log odds scale is given by:

$$\log \mathrm{OR}^{RE^{nc}}_{x,x^*} = \log \mathrm{A}_{x,x} - \log \mathrm{A}_{x^*,x^*} - \log \mathrm{A}^{nc}_{x,x} + \log \mathrm{A}^{nc}_{x^*,x^*} - \log \mathrm{A}^*_{x,x} + \log \mathrm{A}^*_{x^*,x^*}. \tag{2.25}$$

We can observe that it is zero whenever $X \perp\!\!\!\perp Y \mid W$, i.e. $\beta_x = \beta_{xw} = 0$, or $W \perp\!\!\!\perp Y \mid X$, i.e. $\beta_w = \beta_{xw} = 0$ or $X \perp\!\!\!\perp W$, i.e. $\gamma_x = 0$.

Finally, all the decompositions given in Sec. 2.2 and 2.3 allow the consistent estimation of effects by simply plugging the parameter estimates of logistic regression models in the formulas above. Approximate standard errors for the estimates can be obtained via the delta method (Oehlert 1992). Explicit formulas for the first-order approximate variance-covariance matrix, obtained with such a method, are reported in Appendix A.5.

2.5 A comparison between the two approaches

In the following, we study in details the parametric effects presented above, for some cases of interest in order to offer a deeper understanding of the link between the counterfactual and the path analysis approaches. To be specific, we analyze the decompositions of the total effect in Eqs. (2.1), (2.2) and Eq. (2.3) under the following conditional independence assumptions: $X \perp\!\!\!\perp Y \mid W$ (case *i*); $W \perp\!\!\!\perp Y \mid X$ (case *ii*) and $X \perp\!\!\!\perp W$ (case *iii*). Additionally, we study the condition of no-interaction, i.e. $\beta_{xw} = 0$ (case *iv*), and finally we show the link between the two approaches when X is a binary treatment (case *v*).

Case *i*) When the recursive logistic models can be depicted as in Fig. 1.2(a), i.e. $X \perp\!\!\!\perp Y \mid W$, it follows from the definitions above that:

$$\log \mathrm{OR}^{TE}_{x,x^*} \mid_{\beta_x=\beta_{xw}=0} = \log \mathrm{OR}^{IE}_{x,x^*} = \log \mathrm{OR}^{PIE}_{x,x^*} \mid_{\beta_x=\beta_{xw}=0} .$$

Meaning that if $\beta_x = \beta_{xw} = 0$ then $\mathrm{A}_{x,x} = \mathrm{A}_{x^*,x} = \mathrm{A}^*_{x,x}$ and $\mathrm{A}_{x^*,x^*} = \mathrm{A}_{x,x^*} = \mathrm{A}^*_{x^*,x^*}$ and $\mathrm{A}^{nc}_{x,x} = \mathrm{A}^{nc}_{x^*,x^*}$. It follows that the direct effects (2.13, 2.15, 2.18, 2.22, and 2.24), the residual effects (2.23 and 2.25) and the interaction effects (2.17 and 2.19) are all zero.

Case *ii*) When the recursive logistic models can be depicted as in Fig. 1.2(b), i.e. $W \perp\!\!\!\perp Y \mid X$, it follows from the definitions above that:

$$\log \mathrm{OR}^{TE}_{x,x^*} \mid_{\beta_w=\beta_{xw}=0} = \log \mathrm{OR}^{DE}_{x,x^*} = \log \mathrm{OR}^{CDE}_{x,x^*|0},$$

due again to the sufficient condition to guarantee the collapsibility of the log odds ratio. In this case all A terms equal one, then the indirect effects (2.14, 2.16, and 2.21), the mediated and reference interaction effects (2.17 and 2.19) and the residual effects (2.23 and 2.25) are all zero.

Case *iii*) When the recursive logistic models can be depicted as in Fig. 1.2(c), i.e. $W \perp\!\!\!\perp X$, it follows from the definitions above that:

$$\begin{aligned} \log \mathrm{OR}^{TE}_{x,x^*} \mid_{\gamma_x=0} &= \log \mathrm{OR}^{DE^{nc}}_{x,x^*} \mid_{\gamma_x=0} = \log \mathrm{OR}^{DE}_{x,x^*} + \log \mathrm{OR}^{RE}_{x,x^*} \mid_{\gamma_x=0} \\ &= \log \mathrm{OR}^{CDE}_{x,x^*|0} + \log \mathrm{OR}^{INT^{ref}}_{x,x^*|0} \mid_{\gamma_x=0} . \end{aligned} \tag{2.26}$$

In the same way, under this situation, we recover the non-collapsibility of the log odds-ratio. Notice that in this case $\mathrm{A}_{x,x} = \mathrm{A}_{x,x^*} = \mathrm{A}^{nc}_{x,x}$ and $\mathrm{A}_{x^*,x^*} = \mathrm{A}_{x^*,x} = \mathrm{A}^{nc}_{x^*,x^*}$ and $\mathrm{A}^*_{x,x} = \mathrm{A}^*_{x^*,x^*}$, meaning that the indirect effects (2.14, 2.16 and 2.21), the mediated interaction effect (2.17) and the residual effect (2.25) are all zero. After some algebra, it is possible to verify that this assumption is sufficient to avoid effect reversal between the total effect and the (controlled) direct effect, as shown by Cox and Wermuth (2003) in a more general context.

Case *iv*) When we consider the absence of interaction between the treatment and the mediator, i.e. $\beta_{xw} = 0$, we have that

$$\begin{aligned} \log \mathrm{OR}^{TE}_{x,x^*} \mid_{\beta_{xw}=0} &= \log \mathrm{OR}^{DE}_{x,x^*} + \log \mathrm{OR}^{IE}_{x,x^*} + \log \mathrm{OR}^{RE}_{x,x^*} \mid_{\beta_{xw}=0} \\ &= \log \mathrm{OR}^{DE^{nc}}_{x,x^*} \mid_{\beta_{xw}=0} + \log \mathrm{OR}^{IE}_{x,x^*} + \log \mathrm{OR}^{RE^{nc}}_{x,x^*} \mid_{\beta_{xw}=0} \\ &= \log \mathrm{OR}^{CDE}_{x,x^*|0} + \log \mathrm{OR}^{INT^{ref}}_{x,x^*|0} \mid_{\beta_{xw}=0} + \log \mathrm{OR}^{INT^{med}}_{x,x^*} \mid_{\beta_{xw}=0} + \log \mathrm{OR}^{PIE}_{x,x^*} \mid_{\beta_{xw}=0} . \end{aligned}$$

It can be noticed that under this condition no components of the total effect vanishes. Furthermore, if we assume $\gamma_x = 0$ we obtain, from the above definition, that

$$\begin{aligned}\log \mathrm{OR}^{TE}_{x,x^*} \mid_{\beta_{xw}=\gamma_x=0} &= \log \mathrm{OR}^{DE^{nc}}_{x,x^*} \mid_{\beta_{xw}=\gamma_x=0} = \log \mathrm{OR}^{DE}_{x,x^*} + \log \mathrm{OR}^{RE}_{x,x^*} \mid_{\beta_{xw}=\gamma_x=0} \\ &= \log \mathrm{OR}^{CDE}_{x,x^*|0} + \log \mathrm{OR}^{INT^{ref}}_{x,x^*|0} \mid_{\beta_{xw}=\gamma_x=0}\end{aligned}$$

and after some algebra we can verify that

$$|(\log \mathrm{OR}^{TE}_{x,x^*} \mid_{\gamma_x=\beta_{xw}=0})| \leq |\log \mathrm{OR}^{DE}_{x,x^*}| = |\log \mathrm{OR}^{CDE}_{x,x^*|0}|,$$

meaning that, under this situation, $\log \mathrm{OR}^{INT^{ref}}_{x,x^*|0}$, as well as $\log \mathrm{OR}^{RE}_{x,x^*}$ are negative. This result is in line with that obtained by Neuhaus and Jewell (1993) in a more general context.

Case *v*) Similarly to Sec. 1.5, if we consider a binary treatment X taking values $x = 1$ and $x^* = 0$, it follows from the definitions above that $\mathrm{A}_{0,0} = \mathrm{A}^*_{0,0} = \mathrm{A}^{nc}_{0,0}$, $\mathrm{A}_{0,1} = \mathrm{A}^*_{1,1}$ and $\mathrm{A}_{1,0} = \mathrm{A}^{nc}_{1,1}$.

Therefore, the controlled direct effect (2.18) equals the direct effect (2.22). That is:

$$\log \mathrm{OR}^{CDE}_{1,0|0} = \log \mathrm{OR}^{DE}_{1,0} = \beta_x.$$

We can also notice that the pure direct effect (2.13) equals the direct effect (2.24). That is:

$$\begin{aligned}\log \mathrm{OR}^{PDE} = \log \mathrm{OR}^{DE^{nc}}_{1,0} &= \beta_x + \log \mathrm{A}_{1,0} - \log \mathrm{A}_{0,0} \\ &= \beta_x + \log \mathrm{A}^{nc}_{1,1} - \log \mathrm{A}^{nc}_{0,0}.\end{aligned}$$

Therefore, the reference interaction effect (2.19) equals the difference between the two direct effects derived with the path analysis approach. It follow that:

$$\begin{aligned}\log \mathrm{OR}^{INT^{ref}}_{1,0|0} = \log \mathrm{OR}^{DE^{nc}}_{1,0} - \log \mathrm{OR}^{DE}_{1,0} &= \log \mathrm{A}_{1,0} - \log \mathrm{A}_{0,0} \\ &= \log \mathrm{A}^{nc}_{1,1} - \log \mathrm{A}^{nc}_{0,0}.\end{aligned}$$

Likewise, the pure indirect effect (2.16) equals the indirect effect (2.21), then:

$$\begin{aligned}\log \mathrm{OR}^{PIE}_{1,0} = \log \mathrm{OR}^{IE}_{1,0} &= \log \mathrm{A}_{0,1} - \log \mathrm{A}_{0,0} \\ &= \log \mathrm{A}^*_{1,1} - \log \mathrm{A}^*_{0,0}.\end{aligned}$$

Finally, the interaction mediated effect (2.17) equals to the residual term (2.25), that is:

$$\begin{aligned}\log \mathrm{OR}^{INT^{med}}_{1,0} = \log \mathrm{OR}^{RE^{nc}}_{1,0} &= \log \mathrm{A}_{1,1} - \log \mathrm{A}_{0,1} - \log \mathrm{A}_{1,0} + \log \mathrm{A}_{0,0} \\ &= \log \mathrm{A}_{1,1} - \log \mathrm{A}^{nc}_{1,1} - \log \mathrm{A}^*_{1,1} + \log \mathrm{A}_{0,0}.\end{aligned}$$

On the other hand, if we take the sum of the interaction effects from the counterfactual approach we obtain the residual term (2.23), such as:

$$\begin{aligned}\log \mathrm{OR}^{INT^{med}}_{1,0} + \log \mathrm{OR}^{INT^{ref}}_{1,0|0} = \log \mathrm{OR}^{RE}_{1,0} &= \log \mathrm{A}_{1,1} - \log \mathrm{A}_{1,0} \\ &= \log \mathrm{A}_{1,1} - \log \mathrm{A}^*_{1,1}.\end{aligned}$$

Notice that in this setting there is an exact mathematical correspondence, and thus an interchangeability, between the effects derived from the counterfactual framework and those from the path analysis approach. Despite that, the effects derived in this latter approach holds no interpretation for causal inference, unless the recursive system of equations is structural, and the assumptions of no-unmeasured confounders hold. In this situation, they may be given a causal interpretation.[4]

2.6 Direct and indirect effects of a microcredit program: the case of Bosnia and Herzegovina

In the last twenty years, the microfinance has strongly contributed to the financial inclusion of disadvantaged people and of those individuals considered too risky and financially unreliable to have access to the traditional financial services, the so-called "unbankables" individuals. The primary goal of microfincance is to eliminate the poverty, as well as to guarantee a universal primary education, to reduce the gender gap and to increase the empowerment of women. The microcredit, as main tool of the microfinance, aims to make credit accessible to the unbankables individuals and to increase their capability of attracting loans from banks or other microfinance institutions (MFIs).

The randomized controlled trials (RCT) are often used as standard methodologies for the impact evaluation of microcredit programs. Briefly, in such randomized experiments the financial institutions identify the intervention – program or product of microcredit – and a set of potential clients they want to evaluate. A baseline survey of these potential clients is conducted to measure characteristics that might influence the outcomes of interest. Some of the potential clients are randomly assigned to receive the microloan (treatment group) whereas the rest of the clients do not receive the microloan (control group). Finally, after performing the intervention and after a period of time judged necessary for the loans to produce an impact, the outcomes of interest are measured through an endline survey for both the treatment and the control groups.

The strong point of the randomization is that, with a sufficiently large sample, the two groups are similar in both observable and unobservable characteristics. As a result, the treatment in question (microcredit program) is the only systematic difference between the two groups, allowing us to determine its causal effect (i.e. Ass. (1.2) holds). However, the microcredit effect could be confounded by self-selection processes (especially in social science), as well as non-compliance (cross over and no-show), attrition (drop-out) and spillover issues, or quite often because ethics and legal limits, so that the randomization procedure is difficult to carry out. Additionally, even if the randomization allows for a causal interpretation of the microcredit effect, it does not provide explanations and estimations about the process through which the microcredit effect arises. For that reasons, as discussed in the previous chapter, mediation analysis could be an interesting method to identify the extent to which a treatment

[4] See Pearl (2009b, Ch. 7).

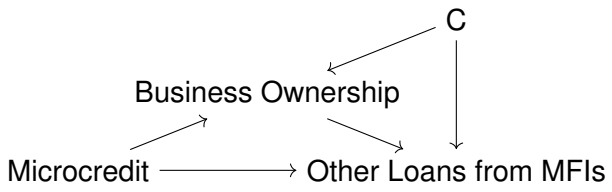


Figure 2.1: Mediation setting for the microcredit study.

affects an outcome through a mediator. In practice, we can decompose the total effect of the microcredit into the sum of its direct and indirect effects.

Many researchers have contributed to the causal evidence of microcredit programs with randomized designs. Some specific examples can be found in a series of articles published in 2015 by the American Economic Journal: Applied Economics. These works have evaluated the impact of randomized microcredit programs in six different developing countries such as Bosnia-Herzegovina, Ethiopia, India, Mexico, Mongolia, and Morocco; see Banerjee et al. (2015) and references therein for a more thorough explanation of the whole project. In general all these studies have estimated the overall effect of a microcredit program, but none of them has implemented a mediation analysis. Hence, we offer an empirical application of our derived analytical results to the microcredit experiment implemented by Augsburg et al. (2015).

The microcredit experiment was performed in Bosnia and Herzegovina during the period 2009-2010. The target population was a subset of clients of a well-established MFIs in the country, who were originally rejected for a regular microloan because lacking of collateral and solid financial history. As stated in Augsburg et al. (2015), given that this particular group of clients was loans applicants of the MFIs, this allowed to increase the effectiveness of the experiment, which was made in (quasi-) perfect compliance. Precisely, only 1.5% of clients selected to receive a microcredit did not take-up it later on. The main goal of the experiment was to emphasize how the impact of a microcredit program could be expanded even to poorer and financially disadvantaged people.

At baseline period, the loan officers selected 1196 marginal clients who would take part to the experiment. Through a survey company, they collected the main information about them and their household. Upon the baseline survey, the clients were randomly allocated to the treatment group (access to the microloan) or the control group (no microloan). After 14 months, at follow-up, the clients were surveyed on a similar questionnaire at baseline. In total, 995 clients were interviewed at the two waves, meaning that an attrition rate of 17% was registered at follow-up (Augsburg et al. 2015). The microloans were individual-liability loans with monthly reimbursement. Their amount ranged between 300 and 3000 BAM (Bosnian marks currency, with an exchange rate at baseline of US$ 1 to BAM 1.634). The average amount was equal to 1'653 BAM, with an average maturity of 57 weeks.

The research team evaluated the microcredit effects on several outcomes, including outstanding loan, business activities, time worked, individual income, savings and household's consumption. Nevertheless, some of these outcomes can be seen as final outcomes as well

	$Y \sim \beta_0 + \beta_x X + \beta_w W + \beta_{xw} XW + \beta_a A + \beta_u U + \beta_l L$				
	Estimate	Std. Error	95% Conf. Interval		p-value
β_0	-1.542	0.290	-2.118	-0.981	0.000
β_x	1.903	0.213	1.492	2.327	0.000
β_w	0.758	0.211	0.349	1.175	0.000
β_{xw}	0.137	0.296	-0.444	0.718	0.643
β_a	0.008	0.006	-0.004	0.020	0.214
β_u	-1.001	0.363	-1.729	-0.299	0.006
β_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
γ_0	0.027	0.095	-0.159	0.213	0.776
γ_x	0.262	0.128	0.011	0.513	0.041

Table 2.1: Results from the fitted logistic models for the outcome and the mediator: the microcredit study.

as determinants of other outcomes, i.e. intermediate outcomes. Let us consider, for example, the excess of outstanding loan which can be seen as a good proxy of a greater access to liquidity. Thus, one may find that the clients who got the microcredit have increased their capability to collect new funding, as effectively evaluated in Augsburg et al. (2015). However, this does not explain how that could arise. One possible explanation is that the microcredit had a positive impact on business ownership which in turn had a better access to liquidity, proving that the microcredit was delivering on its promise of reducing poverty by relaxing credit constraints that inhibit business growth, as pointed out in Banerjee et al. (2015). In line with those hypotheses, we re-analyse the data in order to bring out and verify such a mediation scheme.

Specifically, let us denote by X the binary treatment, taking value 1 if the client got the microcredit at baseline and 0 otherwise. Let Y be the binary outcome measured at follow-up, taking value 1 for clients who have accessed to at least one new loan from any MFIs and 0 otherwise. Let us denote by W the binary mediator measured at follow-up, taking value 1 for clients owning a personal business and 0 otherwise. The graphical representation of such mediation setting is shown in Figure 2.1. From this DAG, it is possible to note that no treatment-outcome and treatment-mediator confounders are included in the analysis since the treatment has been randomly assigned to clients. However, as well known, since the mediator cannot be randomized, a set of possible mediator-outcome confounders C needs to be determined (Ass. 1.7). Specifically, after some preliminary research combined with subject matter considerations, we have included in C: the clients' age (A) measured in years; the clients' educational level (U) taking value 1 for clients with at least a university degree and 0 otherwise; the number of active loans (L), which ranges from 0 to 7. Notice that all the covariates are measured at baseline and can be considered as pre-treatment variables, thus

	$A=37, U=0, L=0$					$A=37, U=1, L=0$				
	Est.	SE	95% CI		p-value	Est.	SE	95% CI		p-value
$OR^{PDE}_{1,0\|c}$	6.652	0.953	5.024	8.809	0.000	6.796	0.988	5.112	9.036	0.000
$OR^{TDE}_{1,0\|c}$	6.717	0.965	5.069	8.901	0.000	6.868	1.019	5.134	9.187	0.000
$OR^{TIE}_{1,0\|c}$	1.059	0.033	0.997	1.125	0.063	1.059	0.033	0.997	1.125	0.063
$OR^{PIE}_{1,0\|c}$	1.049	0.028	0.996	1.105	0.072	1.048	0.027	0.996	1.102	0.070
$OR^{TE}_{1,0\|c}$	7.046	1.022	5.302	9.364	0.000	7.197	1.071	5.376	9.635	0.000
	$A=37, U=0, L=1$					$A=37, U=1, L=1$				
	Est.	SE	95% CI		p-value	Est.	SE	95% CI		p-value
$OR^{PDE}_{1,0\|c}$	6.646	0.954	5.017	8.806	0.000	6.757	0.976	5.091	8.969	0.000
$OR^{TDE}_{1,0\|c}$	6.709	0.962	5.065	8.887	0.000	6.828	1.005	5.118	9.111	0.000
$OR^{TIE}_{1,0\|c}$	1.059	0.033	0.997	1.125	0.062	1.059	0.033	0.997	1.125	0.063
$OR^{PIE}_{1,0\|c}$	1.049	0.028	0.996	1.106	0.073	1.048	0.027	0.996	1.103	0.071
$OR^{TE}_{1,0\|c}$	7.039	1.022	5.296	9.356	0.000	7.157	1.057	5.358	9.559	0.000
	$A=37, U=0, L=2$					$A=37, U=1, L=2$				
	Est.	SE	95% CI		p-value	Est.	SE	95% CI		p-value
$OR^{PDE}_{1,0\|c}$	6.647	0.957	5.012	8.815	0.000	6.723	0.967	5.072	8.913	0.000
$OR^{TDE}_{1,0\|c}$	6.708	0.962	5.065	8.885	0.000	6.793	0.992	5.103	9.044	0.000
$OR^{TIE}_{1,0\|c}$	1.059	0.033	0.997	1.125	0.062	1.059	0.033	0.997	1.125	0.063
$OR^{PIE}_{1,0\|c}$	1.049	0.028	0.995	1.106	0.073	1.048	0.027	0.996	1.103	0.071
$OR^{TE}_{1,0\|c}$	7.040	1.024	5.294	9.361	0.000	7.121	1.045	5.341	9.494	0.000

Table 2.2: Estimates, standard errors (SEs), 95% confidence intervals (CIs) and p-values of the causal odds ratios for the mediation scheme of Figure 2.1.

none of the variables in C can be affected by the treatment (Ass. (1.8) holds).[5]

The sample marginal probabilities for X, W and Y are, respectively, $P(X = 1) = 0.55$, $P(W = 1) = 0.54$ and $P(Y = 1) = 0.57$. Furthermore, we have $P(Y = 1 \mid X = 0, W = 0) = 0.24$, $P(Y = 1 \mid X = 1, W = 0) = 0.67$, $P(Y = 1 \mid X = 0, W = 1) = 0.41$ and $P(Y = 1 \mid X = 1, W = 1) = 0.84$, in contrast with the rare outcome assumption, which is clearly violated also conditionally on the covariates C. The age covariate varies between 17 and 70 years. The first and third quartiles are, respectively, 28 and 47 years, whereas the median is 37 years and the average 37.81 years. Furthermore, 95% of sample units do not own a university degree, while around 96% have less than three active loans at baseline. The estimated causal odds ratios for the total effects range from 7.039 to 7.197, a value slightly greater than, but essentially in line with, the marginal outcome/treatment odds ratio (6.885, SE 0.985 with the delta method). Such a marginal odds ratio also has a causal interpretation, since the microcredit is randomly

[5] See Section 1.3 and VanderWeele and Vansteelandt (2009) and Steen and Vansteelandt (2018).

assigned, but not in a mediation setting. A reasonable comparison between those effects has to account for the fact that the two approaches generating them rely on different parametric assumptions.[6]

The output of the fitted logistic regression models for the outcome and the mediator are shown in Table 2.1. We immediately notice that all the estimated coefficients related to the mediation pathways $X \to W \to Y$ and $X \to Y$ 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, there is a relevant difference in the coefficient magnitudes. Indeed, $\hat{\beta}_w$ and $\hat{\gamma}_x$ are much smaller than $\hat{\beta}_x$, suggesting that all the controlled, pure and total direct effects might be the dominant component of the total effect. Furthermore, it is important to notice that we have explored, in the outcome model, the presence of interaction terms involving the covariates, but none of these coefficients resulted statistically significant or, to the best of our judgement, worth to be added to the model. We have also replicated the above results on the estimated effects 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 (SE 0.149) and $\hat{\beta}_w$ = 0.828 (SE 0.149), while the other parameters do not sensibly change. The causal odds ratios resulting from this model are substantially equivalent to those presented here.

Table 2.2 shows the estimates, together with their variability measures, of the causal effects on the odds ratio scale obtained from the model parameters in Tab. 2.1. The effects refer to clients with median age and all the most frequent configurations of the other covariates. The asymptotic standard errors (SEs) and confidence intervals (CIs) are constructed on the odds ratio scale using the delta method as illustrated in Appendix A.5. 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. The results in Table 2.2 displays that the estimated effects are rather stable across the covariate patterns examined.

The controlled direct effect does not appear in Table 2.2 since it corresponds to the exponentiated regression coefficient $\hat{\beta}_x$ of Table 2.1, which is 6.704 (SE 1.427). In Table 2.2 we can see that the estimated direct effects (*PDE* and *TDE*) always lie between 6.646 and 6.868, with the 95% confidence intervals far away from 1, corresponding to highly significant effects. On the other hand, all the estimates of indirect effects (*PIE* and *TIE*) vary between 1.048 and 1.059, with the 95% confidence intervals barely contain 1, corresponding to p-values around 6% and 7%. The low magnitude of the indirect effects might be due to the relatively limited temporal distance occurring between the baseline and the follow-up measurement occasions. Indeed, as also pointed out 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.

Ultimately, as a sensitivity analysis, we have assessed the effect estimates with alternative relevant covariates like income, value of family assets, gender and marital status. Even under a different setting the results are in line with the values in Table 2.2. Thus, we confirm the validity of our causal estimates, even though that, like in every empirical study, the absence of unobserved confounding cannot be guaranteed with certainty, and results have to be interpreted with caution.

[6] See Lin et al. (1998) and Stanghellini and Doretti (2019).

2.7 Simulation

In this section we report the results of a simulation study. The aim is to compare the exact effect estimators proposed in Sec. 2.4 to the approximate ones of Valeri and VanderWeele (2013), which rely on the rare outcome assumption. Specifically, we compare the relative root mean squared error (RRMSE) of the effect estimators for different scenarios, as well as the coverage probability rate of 95% confidence intervals. Additionally, we also observe the variance behaviour by the empirical cumulative density function (ECDF) of the standard errors of the effect estimators.

The simulation is conducted in a simple framework with all binary variables, including a covariate C, which is supposed to be related to the treatment, the mediator and the outcome. As shown in Sec. 2.5, in the case of a binary treatment, the proposed effects coincide in both the counterfactual and the path analysis framework. For that reason, in the following we show the results only in the counterfactual approach. The data generating process for Y, W, X and C can be represented in Figure 1.1(d), while in Table 2.3 we report the magnitude of the regression coefficients used in such a process.[7] Notice that the covariate C has been generated from a Bernoulli distribution with probability $P(C = 1) = 0.5$. Finally, three sample sizes ($n = 250, 500, 1000$) are considered. For each sample size, 1000 datasets are generated and the estimates of the five causal parameters $\text{OR}^{PDE}_{1,0|c}$, $\text{OR}^{TIE}_{1,0|c}$, $\text{OR}^{TDE}_{1,0|c}$, $\text{OR}^{PIE}_{1,0|c}$ and $\text{OR}^{TE}_{1,0|c}$, together with their standard errors, are computed. The standard errors are obtained with the delta method.[8]

We report results for the scenario with sample size 1000 and for both levels of the covariate. Other scenarios results are reported in Appendix A.11. Overall, they show that exact estimators slightly overcome the approximate ones, as we expected. In particular, for all the sample sizes, all the prevalences and for the setting with the covariate level equal to zero, the exact estimators outperform the approximate estimators in terms of RRMSE, with more pronounced differences for lower sample sizes. Similarly, for the covariate level equal to one, the trend is slightly better for the exact estimators, with some exceptions for the total direct effects and total effects. Regarding the coverage probability, all those for the pure and total direct effects and total effects are very close to the theoretical value of 95%, whereas for the indirect effects the coverage probability is slightly greater than the nominal value (around 97-99%), thus meaning more conservative standard error estimates. Additionally, we can notice a similar trend for all scenarios in both methods, in which the RRMSEs decrease for higher prevalences, specifically from prevalence 20% to prevalence 40%.

Lastly, a graphical comparison involving the behaviour of variance estimators is also reported. To this end, we have built a series of plots shown in Figures A.11.1, A.11.2 and A.11.3. The plots contain a 5×2 panel with rows varying with the five effects estimators and columns with the covariate level. Each quadrant reports ECDF of the standard errors of the five estimators, computed at every iteration. The black line represents the ECDF of the exact estimator

[7] In what follows, we borrow from the epidemiological terminology and refer to $P(Y = 1)$ as to the outcome *prevalence*.

[8] See Appendix A.5 for the exact estimators, and the web appendix of Valeri and VanderWeele 2013 for the approximated ones.

Models	$Y \sim \beta_0 + \beta_x X + \beta_c C + \beta_w W + \beta_{xw} XW$			
	$W \sim \gamma_0 + \gamma_x X + \gamma_c C$			
	$X \sim \theta_0 + \theta_c C$			
Parameters[a]	$\beta_0 = (-1.30, -0.45, 0.45)$		$\gamma_0 = 1$	$\theta_0 = 0.1$
	$\beta_c = 0.3$	$\beta_x = 0.10$	$\gamma_c = 0.2$	$\theta_c = 0.3$
	$\beta_w = -0.08$	$\beta_{xw} = 0.03$	$\gamma_x = 0.6$	

Table 2.3: Simulation study: true parameter values.

[a]We followed a scheme similar to the one of Gaynor et al. (2018), three different values of β_0 are set in order to govern the outcome rareness. In particular, in combination with the other parameters, the marginal probability $P(Y = 1)$ is approximately equal to 20%, 40% and 60%, with all the conditional probabilities $P(Y = 1 \mid X = x, W = w, C = c)$ close to the marginal one.

standard errors, whereas red lines denote the approximate ones. Specifically, each figure shows results for the setting with $n = 1000$ and for each outcome prevalence (20%, 40%, 60%). It is possible to observe that variance estimates for the approximate estimators tend to be higher for pure indirect effects, especially at higher prevalences. The plots referring to the other simulation sample sizes (not reported) show similar trend, which lead to conclude that the behaviour of the variability estimators is slightly more suitable for the exact estimation method.

2.8 Discussion

In this chapter, we have focused on a setting with a binary outcome and a binary mediator, both modelled via logistic regressions. In the counterfactual framework, the novel parametric derivation we have provided does not rely on the rare outcome assumption, thus gaining a closed-form expression. It allows to appreciate the role pathway-specific coefficients play in the causal effects and easily generalize to all type of effects and accounts for every possible interaction in regression models, including those between the treatment (as well as the mediator) and the confounding covariates. These interactions were not previously considered in the mediation literature.

In the path analysis approach, we have proposed a novel decomposition of the total effect which is based on the exact form of the marginal logistic regression model of the outcome against the treatment only. This decomposition is more suitable for non-linear models and, in some cases, reduces to the traditional definitions of linear models. We have showed that the marginal effect can be written as the sum of the indirect and direct effects plus a residual term that vanishes under some specific conditions. This overcomes the debate on which method (i.e. the difference method or the product method) should be used to disentangle the total effect. It also avoids fitting two nested models, thereby sidestepping the issue of unequal variance.

Methods		**Exact**				**Approx.**			
***prev.* 20%**	true	mean	var	RRMSE	CP%	mean	var	RRMSE	CP%
$OR^{PDE}_{1,0\|0}$	1.1294	1.1498	0.0342	0.1648	94.4	1.1505	0.0347	0.1660	94.5
$OR^{TIE}_{1,0\|0}$	0.9949	0.9931	0.0009	0.0297	96.6	0.9921	0.0009	0.0302	96.2
$OR^{TDE}_{1,0\|0}$	1.1329	1.1526	0.0341	0.1640	94.0	1.1529	0.0341	0.1640	94.1
$OR^{PIE}_{1,0\|0}$	0.9919	0.9906	0.0008	0.0292	96.6	0.9897	0.0009	0.0295	96.2
$OR^{TE}_{1,0\|0}$	1.1237	1.1407	0.0321	0.1601	94.2	1.1401	0.0322	0.1604	94.5
***prev.* 40%**									
$OR^{PDE}_{1,0\|0}$	1.1296	1.1431	0.0242	0.1383	94.1	1.1438	0.0251	0.1409	94.3
$OR^{TIE}_{1,0\|0}$	0.9950	0.9940	0.0006	0.0253	97.3	0.9926	0.0007	0.0263	97.6
$OR^{TDE}_{1,0\|0}$	1.1330	1.1460	0.0236	0.1362	94.3	1.1463	0.0239	0.1369	94.4
$OR^{PIE}_{1,0\|0}$	0.9919	0.9912	0.0006	0.0239	97.8	0.9899	0.0006	0.0250	97.3
$OR^{TE}_{1,0\|0}$	1.1239	1.1354	0.0226	0.1341	94.4	1.1342	0.0230	0.1353	94.7
***prev.* 60%**									
$OR^{PDE}_{1,0\|0}$	1.1297	1.1363	0.0270	0.1456	93.6	1.1373	0.0281	0.1486	94.0
$OR^{TIE}_{1,0\|0}$	0.9950	0.9968	0.0007	0.0258	97.9	0.9946	0.0007	0.0268	96.6
$OR^{TDE}_{1,0\|0}$	1.1332	1.1406	0.0267	0.1443	93.0	1.1410	0.0269	0.1448	93.6
$OR^{PIE}_{1,0\|0}$	0.9920	0.9928	0.0006	0.0246	97.6	0.9908	0.0007	0.0263	97.5
$OR^{TE}_{1,0\|0}$	1.1241	1.1319	0.0258	0.1430	93.3	1.1301	0.0263	0.1443	93.2

Table 2.4: Simulation results for the setting with C=0, n=1000. *prev*: outcome prevalence; var: variance; RRMSE: Relative Root Mean Squared Error; CP: Coverage Probability of 95% Confidence Intervals.

Methods		**Exact**				**Approx.**			
***prev.* 20%**	true	mean	var	RRMSE	CP%	mean	var	RRMSE	CP%
$OR^{PDE}_{1,0\|1}$	1.1307	1.1503	0.0336	0.1631	94.0	1.1509	0.0339	0.1638	94.5
$OR^{TIE}_{1,0\|1}$	0.9955	0.9941	0.0007	0.0261	97.0	0.9928	0.0007	0.0269	96.5
$OR^{TDE}_{1,0\|1}$	1.1339	1.1535	0.0350	0.1659	93.8	1.1538	0.0348	0.1655	93.8
$OR^{PIE}_{1,0\|1}$	0.9928	0.9917	0.0006	0.0255	97.1	0.9905	0.0007	0.0261	96.0
$OR^{TE}_{1,0\|1}$	1.1257	1.1427	0.0324	0.1607	94.2	1.1417	0.0324	0.1605	94.0
***prev.* 40%**									
$OR^{PDE}_{1,0\|1}$	1.1309	1.1435	0.0237	0.1366	93.9	1.1440	0.0244	0.1387	94.1
$OR^{TIE}_{1,0\|1}$	0.9955	0.9949	0.0005	0.0224	97.7	0.9933	0.0005	0.0236	97.7
$OR^{TDE}_{1,0\|1}$	1.1340	1.1466	0.0240	0.1371	94.2	1.1467	0.0241	0.1374	93.8
$OR^{PIE}_{1,0\|1}$	0.9928	0.9923	0.0004	0.0213	98.0	0.9908	0.0005	0.0227	97.1
$OR^{TE}_{1,0\|1}$	1.1258	1.1371	0.0227	0.1342	94.4	1.1355	0.0230	0.1349	94.0
***prev.* 60%**									
$OR^{PDE}_{1,0\|1}$	1.1311	1.1377	0.0263	0.1436	93.4	1.1384	0.0270	0.1455	93.7
$OR^{TIE}_{1,0\|1}$	0.9956	0.9973	0.0005	0.0229	98.2	0.9949	0.0006	0.0240	96.5
$OR^{TDE}_{1,0\|1}$	1.1341	1.1419	0.0270	0.1451	93.2	1.1422	0.0270	0.1449	93.2
$OR^{PIE}_{1,0\|1}$	0.9929	0.9937	0.0005	0.0219	97.6	0.9915	0.0006	0.0239	97.9
$OR^{TE}_{1,0\|1}$	1.1260	1.1342	0.0258	0.1427	93.2	1.1319	0.0260	0.1432	93.4

Table 2.5: Simulation results for the setting with C=1, n=1000. *prev*: outcome prevalence; var: variance; RRMSE: Relative Root Mean Squared Error; CP: Coverage Probability of 95% Confidence Intervals.

Chapter 3

Mediation analysis for a binary outcome with multiple binary mediators

3.1 Multiple mediation analysis

In Chapter 1 and 2, we have considered notions and methods for mediation analysis focusing only on a single mediational process. Nonetheless, in many empirical research it is more suitable and realistic to investigate the effects of a treatment on the outcome by including several mediators. As an instance, we consider the example given in the introduction of Ch. 1 on the JOBS II study (Vinokur et al. 1995; Vinokur and Schul 1997). At first, we have assumed a simplified mediation scheme considering the level of subject's job search self-efficacy as the only mediator, which might be affected by the job-program and might affect the depressive symptoms of job seekers in turn. However, as actually performed in the original study, one may think that the job-training program affects both the job search self-efficacy and the reemployment status which in turn might decrease the depressive symptoms of the subjects under analysis. Furthermore, notice that the mediators can also be related each other, e.g. the job search self-efficacy may be a predictor of the reemployment status.

Multiple mediation analysis provides a deeper insight into the complexity of the relations between variables. As for the single mediator case, one of the approaches to the multiple mediation analysis develops from the linear structural equation models and path analysis. Thus, under the narrow assumption of linearity one may use the product method in order to estimate both globally and separately the indirect effect of several mediators. In this situation, the sum of each specific indirect effect equals the overall indirect effect of all mediators (MacKinnon 2000; Preacher and Hayes 2008; Hayes 2017). However, in most situations it is difficult to account for these stringent assumptions considering that even if the outcome variable is normally distributed, it is almost impossible to ensure the absence of interaction between the treatment

and the mediators or the independence between mediators. In path analysis, multiple mediation analysis with recursive systems of non-linear models seems to have not been earlier investigated in mediation literature.

In the counterfactual framework several contributions on mediation analysis have been recently addressed with multiple mediators. In particular, Avin et al. (2005) have presented assumptions for non-parametric identifications of path-specific effects, while Albert and Nelson (2011) have assessed path effects that are parametrically identifiable. However, in both approaches the path effects do not sum up to the total effect. Daniel et al. (2015) have overcame this issue by allowing to obtain a complete decomposition of the total effect. They have offered counterfactual definitions of causal effects, on difference scale, presenting all possible effects' combinations which sum up to the total effect. Concerning the estimation procedures, VanderWeele and Vansteelandt (2014) have used regression based approaches or weighting ones in a setting with a continuous mediators. However, they have offered effects' identification mainly on difference scale, or on ratio scale but handling with the rare outcome assumption. Similarly, Bellavia and Valeri (2018) have offered definitions and identifications criteria, on difference scale, for the decomposition of the total effect which accounts for both the mediation and the interaction phenomena with multiple mediators, thus obtaining, with two mediators, a ten-way decomposition of the total causal effect, generalizing the four-way decomposition of VanderWeele (2014). Despite these relevant contributions, an exact parametric identification for the decomposition of the total effect with binary outcome and binary mediators appears to be lacking in the causal mediation literature. Some results on multiple mediators analysis for binary outcome can be related to the work of Nguyen et al. (2015) who have provided an estimation procedure based on the inverse odds ratio weighting approach to estimate causal direct and indirect effect. This latter approach circumvents some issues of model specifications but it presents higher dispersion of the effects' estimate, as pointed out by the same authors Nguyen et al. (2015, p.351).

In the context of multiple mediation for binary variables, the counterfactual framework presents important limitations and obstacles to the effects' identification and estimation. For that reason, in the following sections, we focus only on the path analysis approach, offering a generalization of the exact parametric identification of the total, direct and indirect effect with several binary intermediate variables. As in the previous chapter, we deal with recursive systems of univariate logistic regression models for the outcome and the mediators. It is also important to clarify that when we consider multiple mediators that are related each other, one may distinguish between the indirect effects which involve each single path from the treatment to the outcome through one or more mediators, namely path-specific indirect effects, and the global indirect effect that considers all the indirect paths simultaneously. These two different classes of indirect effects are equal only under specific conditions, as we will explain in the following sections. Furthermore, with multiple mediators, additional questions can be addressed, such as the decomposition of the total effect when some mediators are marginalized over, as well as when the marginalization arises over the outer mediator instead of the inner one.[1]

[1]For outer mediator we mean the intermediate variable who immediately follows the treatment, whereas for inner mediator we mean the intermediate variable who immediately precedes the outcome.

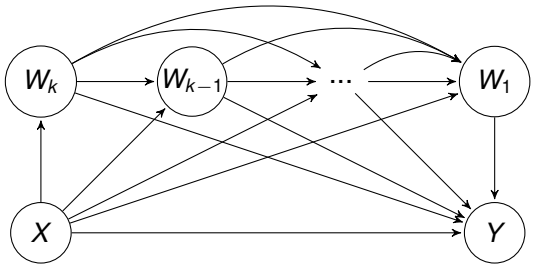


Figure 3.1: DAG with k mediators.

3.2 The decomposition of the total effect with k binary mediators

Let us suppose the general case with k binary mediators. We assume a full ordering among the variables under analysis $(Y, W_1, \dots, W_{k-1}, W_k, X)$, i.e. each variable is a possible response variable for the subsequent ones. This system is displayed via a DAG, as in Figure 3.1. Let us assume that each response model is a hierarchical logistic regression model. This means that if we impose a regression coefficient of one variable to be zero, then all higher order interaction terms involving that covariate are implicitly imposed to zero. Let us indicate by $\text{rhs}(Y \mid X, W_j, W_{>j})$ and $\text{rhs}(W_j \mid X, W_{>j})$, the right hand side (rhs) of the logistic models, respectively, for Y against $(X, W_j, W_{>j})$ and for W_j against X and $W_{>j}$, with $W_{>j}$ the set of all W_r such that $r > j$, $j = \{1, \dots k\}$. As an instance, let $k = 4$ and $j = 2$ then we write $\text{rhs}(Y \mid X, W_2, W_3, W_4)$ and $\text{rhs}(W_2 \mid X, W_3, W_4)$.

Let $\mathrm{A}_{x|w_{>1}}$ be the function which generalizes the one derived in Appendix A.2, after imposing $W = W_1$ and $C = W_{>1}$. Moreover, let $\mathrm{A}^{(w_{<j})}_{x|w_{>j}}$ be the A function obtained recursively after marginalizing over $W_{<j}$ mediators, i.e. $(W_1, \dots, W_{j-1})$ and including $W_{>j}$ as remaining mediators, i.e. $(W_{j+1}, \dots, W_k)$. In analogy with the case of a single mediator in Sec. 2.4, the logistic regression model of Y given X only, obtained after marginalization over k mediators, is given by

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = \beta_0 + \beta_x x + \sum_{j=1}^{k} \log \mathrm{A}^{(w_{<j})}_{x|w_{>j}}, \tag{3.1}$$

where

$$\mathrm{A}^{(w_{<j})}_{x|w_{>j}} = \frac{p_1^{(w_{<j})} p_2^{(w_{<j})} p_3^{(w_{<j})} + p_4^{(w_{<j})}}{p_2^{(w_{<j})} p_3^{(w_{<j})} + p_4^{(w_{<j})}}, \tag{3.2}$$

in which

$$\begin{aligned}
p_1^{(w_{<j})} &= \exp(\text{rhs}(Y \mid W_j = 1, X = x, W_{>j} = w_{>j}) - \text{rhs}(Y \mid W_j = 0, X = x, W_{>j} = w_{>j})),\\
p_2^{(w_{<j})} &= \exp\{\text{rhs}(W_j \mid X = x, W_{>j} = w_{>j})\},\\
p_3^{(w_{<j})} &= 1 + \exp\{\text{rhs}(Y \mid W_j = 0, X = x, W_{>j} = w_{>j})\},\\
p_4^{(w_{<j})} &= 1 + \exp\{\text{rhs}(Y \mid W_j = 1, X = x, W_{>j} = w_{>j})\},
\end{aligned}$$

for all $j = \{1, \dots k\}$. To better understand the above notation, let consider a case with $k = 4$. Equation (3.1) becomes

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = \beta_0 + \beta_x x + \log \mathrm{A}_{x|w_2,w_3,w_4} + \log \mathrm{A}^{(w_1)}_{x|w_3,w_4} + \log \mathrm{A}^{(w_1,w_2)}_{x|w_4} + \log \mathrm{A}^{(w_1,w_2,w_3)}_{x}.$$

The specific form of each A term follows in a straightforward manner given Eq. (3.2). We will refer to Section 3.3 for a complete example with $k = 2$.

Noticing that in this chapter, since we do not offer any formulation in the counterfactual framework, the A term in Eq. 3.2 has only one subscript for x, which refers to the value x of X in each model, i.e. that for Y and for all k mediators. Contrary to what was done in Ch. 2, Sec. 2.4, where the A term (Eq. 2.7) has two subscripts for the value of X, one referred to the value x of X in the model for Y and one referred to the value x of X in the model for W, allowing to an easier identification of the counterfactual effects.

Finally, notice that as for the A term in Eq. (2.7) of Sec. 2.4, also the A term in Eq. (3.2) has a clear statistical interpretation. As already derived by Raggi et al. (2020), which generalize the formulations of Stanghellini and Doretti (2019), we can easily prove that $\mathrm{A}^{(w_{<j})}_{x|w_{>j}}$ is the inverse of the relative risk of $\bar{W}_j = 1 - W_j$ for varying Y after conditioning on $X = x$ and $W_{>j} = w_{>j}$. In Appendix A.6 we show the iterative procedure to obtain Eq. (3.1) and the proof linking Eq. (3.2) to the inverse of the relative risk of $\bar{W}_j$.

Effects' definitions in the path analysis approach

In path analysis approach, similarly to what was done for the case of a single mediator (cf. Sec 2.4), we offer a generalization of the definitions and the parametric identifications for the total, direct, indirect and residual effects under the situation of k mediators and a discrete treatment X taking values x and x^*. Extensions to the case of a continuous and differentiable treatment X follows in a similar way.[2]

Definition 3.2.1. The *Total Effect* of X on Y on the log odds scale is defined as the difference in the log odds of Y (Eq. 3.1) after moving X from x^* to x:

$$\log \mathrm{OR}^{TE}_{x,x^*} = \beta_x(\, x - x^* \,) + \sum_{j=1}^{k} \log \mathrm{A}^{(w_{<j})}_{x|w_{>j}} - \sum_{j=1}^{k} \log \mathrm{A}^{(w_{<j})}_{x^*|w_{>j}},$$

with $\mathrm{A}^{(w_{<j})}_{x|w_{>j}}$ as in Eq.(3.2). Clearly, with $k = 1$ we find the result of Eq. (2.20).

Due to the complexity of the models and the huge number of possible relations between variables, in this chapter we deal with only one definition of the direct effect, the one that ensures adherence to the collapsibility condition.[3]

[2] See Appendix A.7 to which we refer for an example with two mediators and a continuous and differentiable X.
[3] See Sec. 1.4, Ch. 1 and Ma et al. (2006, *Theorem 1*).

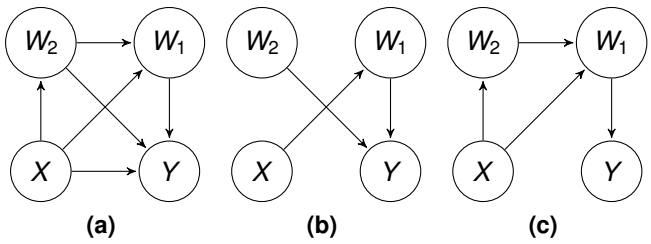


Figure 3.2: DAG with $k = 2$ mediators when (a) no conditional independences hold (b) $X \perp\!\!\!\perp Y \mid \{W_1, W_2\}$, $W_2 \perp\!\!\!\perp W_1 \mid X$ and $X \perp\!\!\!\perp W_2$ (c) $Y \perp\!\!\!\perp \{X, W_2\} \mid W_1$.

Definition 3.2.2. The *Direct Effect* of X on Y on the log odds scale is defined after imposing in the total effect that $\bar{\beta}_W = 0$, thereby $Y \perp\!\!\!\perp \{W_1, \dots, W_k\} \mid X$:

$$\log \mathrm{OR}^{DE}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} \mid_{\bar{\beta}_W=0} = \beta_x(\, x - x^*\,),$$

where $\bar{\beta}_W$ is the set of all β regression coefficients of each mediator in the logit model of Y given X and the all set of mediators. This means interrupting all the indirect paths through the k mediators so that the effect from X to Y acts directly. Notice that the direct effect can be also interpreted as the effect of X on Y keeping all mediators fixed to zero. The definition above is in line with the results on the collapsibility of the odds ratio and with those in the context of a single mediator (See Eq. (2.22) of Sec. 2.4, Ch. 2).

Definition 3.2.3. The *Global Indirect Effect* of X on Y on the log odds scale is defined after imposing in the total effect that $\bar{\beta}_X = 0$ is zero, thereby $Y \perp\!\!\!\perp X \mid \{W_1, \dots, W_k\}$:

$$\log \mathrm{OR}^{GIE}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} \mid_{\bar{\beta}_X=0},$$

where $\bar{\beta}_X$ is the set of all β regression coefficients of X (i.e. β_x and all the interaction coefficients with X) in the logit model of Y given X and the all set of mediators. Notice that GIE describes the effect of X on Y through all the mediators together once the direct path $X \to Y$ is interrupted.

Definition 3.2.4. The *Residual Effect* of X on Y on the log odds scale is defined by the difference between the total (Def. 3.2.1), the direct (Def. 3.2.2) and the global indirect (Def. 3.2.3) effects:

$$\log \mathrm{OR}^{RE}_{x,x^*} = \log \mathrm{OR}^{TE}_{x,x^*} - \log \mathrm{OR}^{DE}_{x,x^*} - \log \mathrm{OR}^{GIE}_{x,x^*}.$$

It can be proven that the residual effect is zero whenever one of the two following graphical conditions holds: (i) there is no direct path from X to Y or (ii) there is a direct path from X to Y and no other arrow is pointing to Y. As we will see in the next section, for example, the model corresponding to the DAG in Fig. 3.2(a) has a non-zero residual effect as there is a direct arrow from X to Y and two other arrows are pointing to Y, while models corresponding to DAGs as in Figs. 3.2(b) and 3.2(c) are such that $\log \mathrm{OR}^{TE}_{x,x^*} = \log \mathrm{OR}^{GIE}_{x,x^*}$ and thus $\log \mathrm{OR}^{DE}_{x,x^*} = \log \mathrm{OR}^{RE}_{x,x^*} = 0$. As shown in the previous chapters, with just one mediator, the residual effect

is nonzero whenever more than one arrow are pointing to Y (see for example Fig. 1.1(c), i.e. no conditional independence assumption holds, and Fig. 1.2(c), i.e. when $X \perp\!\!\!\perp W$).

In the setting with multiple mediators, we can also be interested in deriving the path-specific indirect effects, i.e. the effects that are due to some mediators only, and are null whenever one arrow along the pathway vanishes. First, let V be one of the $2^k - 1$ subsets of $(W_1, \dots, W_k)$, containing at least one element of W. Let i_V be the ordered set of indices j such that $W_j \in V$. Let $\gamma_{j,0}$, $\gamma_{j,x}$, $\gamma_{j,i}$, etc. denote in order, the intercept, the coefficient of X, the coefficient of W_i, etc., in the logistic regression model of W_j given X and the set of all mediators W_r, such that $r > j$ and $j = \{1, \dots k\}$.

Definition 3.2.5. The *Path-Specific Indirect Effect* ($\log \mathrm{OR}_{x,x^*}^{PSIE_V}$) of X on Y on the log odds scale is defined from the total effect after imposing that:

- $\beta_x = 0$;
- $\beta_{w_j} = 0$ with $j = \{1, 2, \dots, k\}, j \neq \min\{i_V\}$ the smallest index in i_V;
- $\gamma_{r,j} = 0$ with $W_r \in V$, $j > r, j \neq \ell_r$, where ℓ_r is the index following r in i_V;
- $\gamma_{r,x} = 0$ with $W_r \in V$, $r \neq \max\{i_V\}$ the largest index in i_V.

Recall that all higher-order interaction coefficients between the involved variables are also zero, as we are imposing that the models are hierarchical. In this way, each PSIE contains only the parameters relating to its path (including the intercepts). It then follows that $\log \mathrm{OR}_{x,x^*}^{PSIE_V}$ is null whenever one of the following conditions holds:

- $\beta_{w_j} = 0$ with $j = \min\{i_V\}$;
- $\gamma_{r,\ell_r} = 0$ with $W_r \in V$;
- $\gamma_{r,x}$ with $r = \max\{i_V\}$.

Notice that each of the above conditions implies deleting one arrow in the corresponding DAG, i.e. imposing a specific conditional independence assumption.

For example, let $k = 4$, $V = \{W_1, W_2, W_4\}$, $i_V = \{1, 2, 4\}$. The path-specific indirect effect is obtained from the total effect after imposing that:

- $\beta_x = 0$;
- $\beta_{w_2} = \beta_{w_3} = \beta_{w_4} = 0$;
- $\gamma_{1,3} = \gamma_{1,4} = \gamma_{2,3} = 0$;
- $\gamma_{1,x} = \gamma_{2,x} = 0$.

The above definition allows for only one path from X to Y, which is $X \to W_4 \to W_2 \to W_1 \to Y$. It is null whenever $\gamma_{4,x}$ or $\gamma_{2,4}$ or $\gamma_{1,2}$ or β_{w_1} are zero.

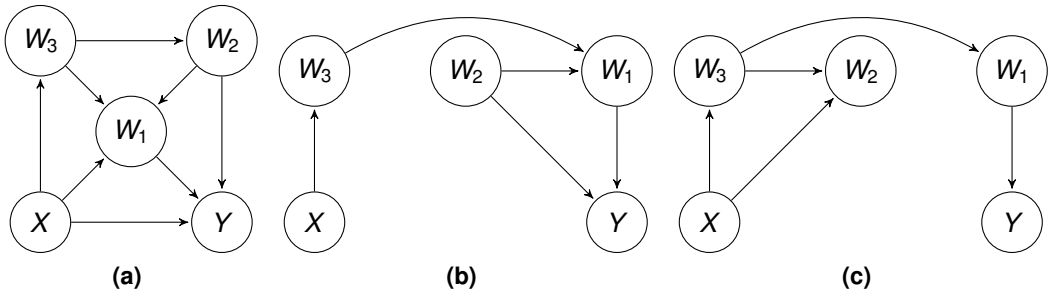


Figure 3.3: DAG with $k = 3$ mediators (a) where W_1 acting as a collider node in the path $X \to W_3 \to W_1 \leftarrow W_2 \to Y$ (b) where $Y \perp\!\!\!\perp \{X, W_3\} \mid \{W_1, W_2\}$, $W_1 \perp\!\!\!\perp X \mid \{W_2, W_3\}$, $W_2 \perp\!\!\!\perp X$ (c) where $Y \perp\!\!\!\perp \{X, W_3\} \mid \{W_1, W_2\}$, $W_1 \perp\!\!\!\perp \{X, W_2\} \mid W_3$, $W_2 \perp\!\!\!\perp X$.

Some considerations on the path-specific indirect effects

The path-specific indirect effects, as defined above, bring out three interesting remarks. First, we can notice that the definition of path-specific indirect effects allows only for arrows pointing in the same direction. Indeed, only ordered subsets of W are admitted to form V because they are the only subsets with a non-zero path specific indirect effect. To clarify this point, see the graph in Fig. 3.3(a). The configuration $X \to W_3 \to W_1 \leftarrow W_2 \to Y$ is not admitted as it gives rise to (W_3, W_1, W_2), not an ordered subset of W. However, notice that as W_1 is a collider node the path between X and Y is blocked by (W_3, W_1, W_2) and the corresponding path-specific effect is zero.[4]

Second, as stated previously, the direct effect above coincides with the effect of X on Y keeping $W_1 = W_2 = ... = W_k = 0$. However, the path specific indirect effect does not in general coincide with the indirect effect after keeping the mediators not in V equal to zero. To see this, notice that in Fig. 3.3(a) the path specific indirect effect for $V = (W_3, W_2)$ is evaluated after imposing that $\beta_x = \beta_{w_1} = 0$. This effect does not coincide with the one obtained after conditioning on $W_1 = 0$.[5]

Third, notice that the sum of all the path-specific indirect effects in general is not equal to the global indirect effect. This is related to the effects which are derived from submodels involving three mediators (W_i, W_j, W_r), $i > j > r$ such that there are two arrows pointing to W_r. In this case, the residual effect of W_i on W_r is non-zero. In particular, the global indirect effect includes all residual effects, whereas path-specific indirect effects do not. This is true even when there is just one path from X to Y. As an instance, consider the models in Figures 3.3(b) and 3.3(c). In both DAGs there is only an indirect path from X to Y, i.e. $X \to W_3 \to W_1 \to Y$, with $V = (W_3, W_1)$. The model in Figure 3.3(b) has a $\log \mathrm{OR}^{GIE}_{x,x^*} \neq \log \mathrm{OR}^{PSIE_V}_{x,x^*}$, while the model in Figure 3.3(c) has a $\log \mathrm{OR}^{GIE}_{x,x^*} = \log \mathrm{OR}^{PSIE_V}_{x,x^*}$. The former situation arises because the presence of the non-zero residual effect in the subgraph formed by W_3, W_2, W_1 in Figure 3.3(b). The difference points out the different interpretation of the

[4] See Pearl (2009b, Ch. 1 and 3).
[5] See Elwert and Winship (2014) to which we refer for further details.

parameters related to the arrow $W_3 \to W_1$ in the two effects: in the $\log \mathrm{OR}^{GIE}_{x,x^*}$ it is the total effect of W_3 on W_1, while in the path-specific effect $\log \mathrm{OR}^{PSIE_V}_{x,x^*}$ it is the direct effect of W_3 on W_1. Regarding to this situation, we may determine the graphical conditions where the sum of all path-specific indirect effects equals to the global indirect effect. That is when (i) there is just one path from X to Y and (ii) all the variables along the path have just one incoming arrow.

3.3 An example with two mediators

In the following we offer an example for a specific case with $k = 2$. In addition to the notation given in Sec. 3.2, let us denote with $\beta_0^{(w_j)}$, $\beta_x^{(w_j)}$, $\beta_j^{(w_j)}$, etc., in order, the intercept, the main effect of X, of W_i, etc., of the logistic model for Y against the X and W_r, $r > j$, which arises after marginalization on W_j. Similarly, we denote with $\beta_0^{(w_j,w_{>j})}$, $\beta_x^{(w_j,w_{>j})}$, etc., in order, the intercept, the main effect of X, etc., of the logistic model for Y against X and the remaining variables, which arises after marginalization on W_j and $W_{>j}$. For all models, higher order coefficients are denoted in a similar way.

Let us assume three logistic regression models, respectively, for the binary outcome Y, and for two binary mediators W_1, W_2, with interaction terms up to the second order:

$$\begin{aligned}\log \frac{P(Y=1 \mid X=x, W_1=w_1, W_2=w_2)}{P(Y=0 \mid X=x, W_1=w_1, W_2=w_2)} &= \beta_0 + \beta_x x + \beta_{w_1} w_1 + \beta_{w_2} w_2 \\ &\quad + \beta_{xw_1} x w_1 + \beta_{xw_2} x w_2 + \beta_{w_1 w_2} w_1 w_2\end{aligned} \tag{3.3}$$

and

$$\log \frac{P(W_1=1 \mid X=x, W_2=w_2)}{P(W_1=0 \mid X=x, W_2=w_2)} = \gamma_{1,0} + \gamma_{1,x} x + \gamma_{1,2} w_2 + \gamma_{1,x2} x w_2, \tag{3.4}$$

$$\log \frac{P(W_2=1 \mid X=x)}{P(W_2=0 \mid X=x)} = \gamma_{2,0} + \gamma_{2,x} x. \tag{3.5}$$

Notice that no conditional independences are assumed. The DAG representing the postulated models is in Fig. 3.2(a).

After the first marginalization over the inner node W_1, the logistic model of Y against X and W_2 becomes:

$$\begin{aligned}\log \frac{P(Y=1 \mid X=x, W_2=w_2)}{P(Y=0 \mid X=x, W_2=w_2)} &= \beta_0^{(w_1)} + \beta_x^{(w_1)} x + \beta_{w_2}^{(w_1)} w_2 + \beta_{xw_2}^{(w_1)} x w_2 \\ &= \beta_0 + \beta_x x + \beta_{w_2} w_2 + \beta_{xw_2} x w_2 + \log \mathrm{A}_{x|w_2},\end{aligned} \tag{3.6}$$

where the first equality arises by definition, while the second one by the marginalization over W_1 (see Appendix A.6). $\mathrm{A}_{x|w_2}$ is given by Eq. (3.2) with $W_j = W_1$ and $W_{>j} = W_2$, that is

$$\mathrm{A}_{x|w_2} = \frac{p_1 p_2 p_3 + p_4}{p_2 p_3 + p_4} \tag{3.7}$$

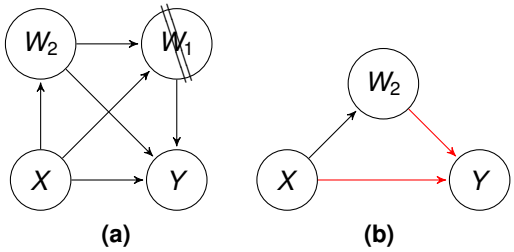


Figure 3.4: (a) Marginalization over the inner mediator W_1 and (b) quantifying the parameters (in red the parameters that change).

in which

$$
\begin{aligned}
p_1 &= \exp(\beta_{w_1} + \beta_{xw_1} x + \beta_{w_1 w_2} w_2),\\
p_2 &= \exp(\gamma_{1,0} + \gamma_{1,x} x + \gamma_{1,w_2} w_2 + \gamma_{1,xw_2} x w_2),\\
p_3 &= 1 + \exp(\beta_0 + \beta_x x + \beta_{w_2} w_2 + \beta_{xw_2} x w_2),\\
p_4 &= 1 + \exp(\beta_0 + \beta_x x + \beta_{w_1} + \beta_{xw_1} x + \beta_{w_2} w_2 + \beta_{xw_2} x w_2 + \beta_{w_1 w_2} w_2).
\end{aligned}
$$

We can notice that the marginalization over the inner mediator induces changes in the parameters of the outcome equation only; see Fig. 3.4(b). We can easily verify that, for a binary X, the regression coefficients of model (3.6) are

$$
\begin{aligned}
\beta_0^{(w_1)} &= \log \frac{P(Y=1 \mid X=0, W_2=0)}{P(Y=0 \mid X=0, W_2=0)}\\
&= \beta_0 + \log \mathrm{A}_{0|W_2=0}
\end{aligned}
$$

for the intercept, while for the main effects

$$
\begin{aligned}
\beta_x^{(w_1)} &= \log \frac{P(Y=1 \mid X=1, W_2=0)}{P(Y=0 \mid X=1, W_2=0)} - \log \frac{P(Y=1 \mid X=0, W_2=0)}{P(Y=0 \mid X=0, W_2=0)}\\
&= \beta_x + \log \mathrm{A}_{1|W_2=0} - \log \mathrm{A}_{0|W_2=0}
\end{aligned}
$$

and

$$
\begin{aligned}
\beta_{w_2}^{(w_1)} &= \log \frac{P(Y=1 \mid X=0, W_2=1)}{P(Y=0 \mid X=0, W_2=1)} - \log \frac{P(Y=1 \mid X=0, W_2=0)}{P(Y=0 \mid X=0, W_2=0)}\\
&= \beta_{w_2} + \log \mathrm{A}_{0|W_2=1} - \log \mathrm{A}_{0|W_2=0},
\end{aligned}
$$

with $\mathrm{A}_{x|w_2}$ as in Eq. (3.7). Instead, the second order interaction coefficient is

$$
\begin{aligned}
\beta_{xw_2}^{(w_1)} &= \log \frac{P(Y=1 \mid X=1, W_2=1)}{P(Y=0 \mid X=1, W_2=1)} - \log \frac{P(Y=1 \mid X=0, W_2=1)}{P(Y=0 \mid X=0, W_2=1)}\\
&\quad - \log \frac{P(Y=1 \mid X=1, W_2=0)}{P(Y=0 \mid X=1, W_2=0)} - \log \frac{P(Y=1 \mid X=0, W_2=0)}{P(Y=0 \mid X=0, W_2=0)}\\
&= \beta_{xw_2} + \log \mathrm{A}_{1|W_2=1} - \log \mathrm{A}_{0|W_2=1} - \log \mathrm{A}_{1|W_2=0} + \log \mathrm{A}_{0|W_2=0},
\end{aligned}
$$

with $\mathrm{A}_{x|w_2}$ as in Eq. (3.7). Notice that even if we did not include any type of interaction in the original model, after the first marginalization, a second order interaction term would appear.

Similarly, after the second marginalization, we obtain that the marginal model (3.1) becomes

$$\begin{aligned}\log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} &= \beta_0^{(w_1,w_2)} + \beta_x^{(w_1,w_2)} x \\ &= \beta_0^{(w_1)} + \beta_x^{(w_1)} x + \log \mathsf{A}_x^{(w_1)} \\ &= \beta_0 + \beta_x x + \log \mathsf{A}_{x|W_2=0} + \log \mathsf{A}_x^{(w_1)},\end{aligned} \tag{3.8}$$

where the first equality is given by definition of the logistic regression model of Y given X, the second one arises by marginalization of (3.6) over W_2 and the last one by simple substitution. Notice that in this case $\mathsf{A}_x^{(w_1)}$ is given by Eq. (3.2) with $W_j = W_2$, that is

$$\mathsf{A}_x^{(w_1)} = \frac{p_1^{(w_1)} p_2^{(w_1)} p_3^{(w_1)} + p_4^{(w_1)}}{p_2^{(w_1)} p_3^{(w_1)} + p_4^{(w_1)}} \tag{3.9}$$

in which

$$\begin{aligned}p_1^{(w_1)} &= \exp(\beta_{w_2} + \beta_{xw_2} x + \log \mathsf{A}_{x|W_2=1} - \log \mathsf{A}_{x|W_2=0}), \\ p_2^{(w_1)} &= \exp(\gamma_{2,0} + \gamma_{2,x} x), \\ p_3^{(w_1)} &= 1 + \exp(\beta_0 + \beta_x x + \log \mathsf{A}_{x|W_2=0}), \\ p_4^{(w_1)} &= 1 + \exp(\beta_0 + \beta_x x + \beta_{w_2} + \beta_{xw_2} x + \log \mathsf{A}_{x|W_2=1})\end{aligned}$$

and $\mathsf{A}_{x|w_2}$ as in Eq. (3.7). Notice that $\mathsf{A}_{x|w_2}$ and $\mathsf{A}_x^{(w_1)}$ are, respectively, the inverse of the relative risk of $\bar{W}_1$ for varying Y given $X = x$ and $W_2 = w_2$, i.e. $RR_{\bar{W}_1|Y,X=x,W_2=w_2}$, and the inverse of the relative risk of $\bar{W}_2$ for varying Y given $X = x$, i.e. $RR_{\bar{W}_2|Y,X=x}$. We refer to Stanghellini and Doretti (2019) and Appendix A.6 for the proof.

Total, direct, indirect and residual effects with two mediators

Given the marginal model (3.8) above and following the definitions in Sec. 3.2, we can parametrically identify the total effect and its components for the case with two binary mediators. For simplicity, we consider the situation with a binary X. We refer to Appendix A.7 for the case of a continuous and differentiable X.

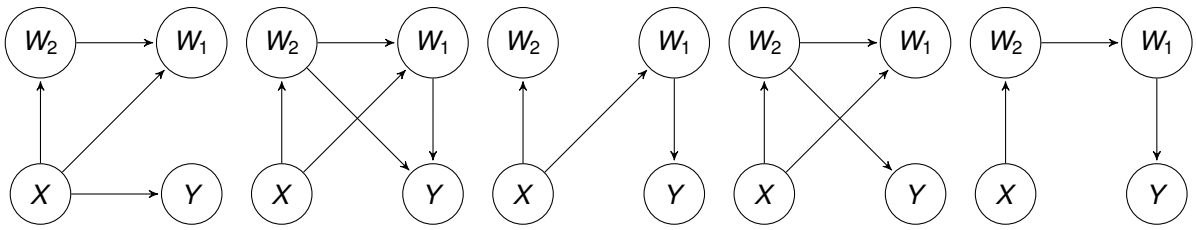
By Definition 3.2.1, the total effect of X on Y, after the marginalization over two mediators, is given by the difference between the marginal logistic model (3.8) evaluated in $x = 1$ and in $x^* = 0$:

$$\log \mathsf{OR}_{1,0}^{TE} = \beta_x + \log \mathsf{A}_{1|W_2=0} - \log \mathsf{A}_{0|W_2=0} + \log \mathsf{A}_1^{(w_1)} - \log \mathsf{A}_0^{(w_1)},$$

with $\mathsf{A}_{x|w_2}$ and $\mathsf{A}_x^{(w_1)}$ as in Eqs. (3.7) and (3.9).

By Definition 3.2.2, the direct effect is obtained after imposing in the total effect that $Y \perp\!\!\!\perp \{W_1, W_2\} \mid X$:

$$\log \mathsf{OR}_{1,0}^{DE} = \log \mathsf{OR}_{1,0}^{TE} \mid_{\beta_{w_1}=\beta_{w_2}=0} = \beta_x,$$



(a) *DE* **(b)** *GIE* **(c)** $PSIE_{W_1}$ **(d)** $PSIE_{W_2}$ **(e)** $PSIE_{W_1,W_2}$

Figure 3.5: Decomposition of the total effect with two binary mediators.

recalling that we also assume the interactions terms related to both mediators be zero. The direct effect can be interpreted as the effect of X on Y not transmitted through the mediators; see Fig. 3.5(a). It can also be seen as the conditional log odds ratio between X and Y keeping $W_1 = W_2 = 0$. Additional, notice that as soon as $\beta_{w_1} = \beta_{w_2} = 0$ then $\mathrm{A}_{x|w_2} = \mathrm{A}_x^{(w_1)} = 1$.

By Definition 3.2.3, the global indirect effect is obtained after imposing that $Y \perp\!\!\!\perp X \mid \{W_1, W_2\}$:

$$\log \mathrm{OR}_{1,0}^{GIE} = \log \mathrm{OR}_{1,0}^{TE} \mid_{\beta_x=0} = \log \mathrm{A}^*_{1|W_2=0} - \log \mathrm{A}^*_{0|W_2=0} + \log \mathrm{A}_1^{*(w_1)} - \log \mathrm{A}_0^{*(w_1)},$$

where $\mathrm{A}^*_{x|w_2}$ and $\mathrm{A}_x^{*(w_1)}$ are $\mathrm{A}_{x|w_2}$ and $\mathrm{A}_x^{(w_1)}$ evaluated with $\beta_x = 0$ (including the interactions). The global indirect effect is depicted in Fig. 3.5(b), and represents the effect of X on Y through the mediators, i.e. interrupting the direct path from X to Y. However, it does not tell us which is the effect contribution of each specific indirect path.

Following the Definition 3.2.4, the residual effect is given by the difference between the total, the direct and the global indirect effect:

$$\begin{aligned}\log \mathrm{OR}_{1,0}^{RE} &= \log \mathrm{OR}_{1,0}^{TE} - \log \mathrm{OR}_{1,0}^{DE} - \log \mathrm{OR}_{1,0}^{GIE} \\ &= \log \mathrm{A}_{1|W_2=0} - \log \mathrm{A}^*_{1|W_2=0} + \log \mathrm{A}_1^{(w_1)} - \log \mathrm{A}_1^{*(w_1)}.\end{aligned}$$

After some algebra we can verify that the residual effect is zero whenever $\beta_x = 0$ or $\beta_{w_1} = \beta_{w_2} = 0$ (including the interactions).

By Definition 3.2.5, for a set of two mediators $W = \{W_1, W_2\}$ we have three subsets $V = \{W_1\}, \{W_2\}$ and $\{W_1, W_2\}$, thus obtaining three different path-specific indirect effects. This means that the indirect effect through each mediator alone, respectively W_1, W_2, is:

$$\begin{aligned}\log \mathrm{OR}_{1,0}^{PSIE_{W_1}} &= \log \mathrm{OR}_{1,0}^{TE} \mid_{\beta_x=\beta_{w_2}=\gamma_{1,2}=0} = \log \mathrm{A}^{**}_{1|W_2=0} - \log \mathrm{A}^{**}_{0|W_2=0} \\ \log \mathrm{OR}_{1,0}^{PSIE_{W_2}} &= \log \mathrm{OR}_{1,0}^{TE} \mid_{\beta_x=\beta_{w_1}=0} = \log \mathrm{A}_1^{**(w_1)} - \log \mathrm{A}_0^{**(w_1)},\end{aligned}$$

where $\mathrm{A}^{**}_{x|W_2=0}$ is $\mathrm{A}_{x|W_2=0}$ evaluated with $\beta_x = \beta_{w_2} = \gamma_{1,2} = 0$, that is:

$$\mathrm{A}^{**}_{x|W_2=0} = \frac{\exp(\beta_{w_1}) \exp(\gamma_{1,0} + \gamma_{1,x} x)(1 + \exp(\beta_0)) + 1 + \exp(\beta_0 + \beta_{w_1})}{\exp(\gamma_{1,0} + \gamma_{1,x} x)(1 + \exp(\beta_0)) + 1 + \exp(\beta_0 + \beta_{w_1})},$$

while $\mathrm{A}_x^{**(w_1)}$ is $\mathrm{A}_x^{(w_1)}$ evaluated with $\beta_x = \beta_{w_1} = 0$, that is:

$$\mathrm{A}_x^{**(w_1)} = \frac{\exp(\beta_{w_2})\exp(\gamma_{2,0}+\gamma_{2,x}x)(1+\exp(\beta_0)+1+\exp(\beta_0+\beta_{w_2})}{\exp(\gamma_{2,0}+\gamma_{2,x}x)(1+\exp(\beta_0)+1+\exp(\beta_0+\beta_{w_2})}.$$

Instead, the indirect effect through both mediators is given by:

$$\log \mathrm{OR}_{1,0}^{PSIE_{w_1,w_2}} = \log \mathrm{OR}_{1,0}^{TE} \mid_{\beta_x=\gamma_{1,x}=\beta_{w_2}=0} = \log \mathrm{A}_1^{***(w_1)} - \log \mathrm{A}_0^{***(w_1)}$$

where $\mathrm{A}_x^{***(w_1)}$ is $\mathrm{A}_x^{(w_1)}$ evaluated with $\beta_x = \gamma_{1,x} = \beta_{w_2} = 0$, that is:

$$\mathrm{A}_x^{***(w_1)} = \frac{p_1^{***}p_2^{***}p_3^{***}+p_4^{***}}{p_2^{***}p_3^{***}+p_4^{***}},$$

with

$$\begin{aligned}
p_1^{***} &= \exp(\log \mathrm{A}_{x|W_2=1}^{***} - \log \mathrm{A}_{x|W_2=0}^{***}),\\
p_2^{***} &= \exp(\gamma_{2,0}+\gamma_{2,x}x),\\
p_3^{***} &= 1+\exp(\beta_0+\log \mathrm{A}_{x|W_2=0}^{***}),\\
p_4^{***} &= 1+\exp(\beta_0+\log \mathrm{A}_{x|W_2=1}^{***}),
\end{aligned}$$

in which $\mathrm{A}_{x|w_2}^{***}$ is $\mathrm{A}_{x|w_2}$ evaluated with $\beta_x = \gamma_{1,x} = \beta_{w_2} = 0$, such as:

$$\mathrm{A}_{x|w_2}^{***} = \frac{\exp(\beta_{w_1})\exp(\gamma_{1,0}+\gamma_{1,w_2}w_2)(1+\exp(\beta_0))+1+\exp(\beta_0+\beta_{w_1})}{\exp(\gamma_{1,0}+\gamma_{1,w_2}w_2)(1+\exp(\beta_0))+1+\exp(\beta_0+\beta_{w_1})}.$$

We can verify that each indirect effect contains only the parameters of its path (including the intercepts); see Figures 3.5(c) and 3.5(d). Therefore, they are null whenever β_{w_j} or $\gamma_{j,x}$ are zero, $\forall j = 1, 2$. Similarly, the indirect effect through both mediators contains only the parameters of its path (including the intercepts); see Figure 3.5(e), and it is null whenever $\gamma_{2,x}$ or $\gamma_{1,2}$ or β_{w_1} are zero.

Finally, we can notice that the Figures 3.5(a)-(c)-(d)-(e) correspond to models with a residual effect equals to zero. Instead, Figures 3.5(c)-(d)-(e) correspond to models with the global indirect effect equal the path specific indirect effect.

3.4 Other cases of particular interest

In the setting with multiple binary mediators, other research questions are of interest. One significant example may involve the analysis of the path-specific effects in a model that arises after marginalization over some mediators while others are maintained in the model. In this situation, one should proceed in evaluating the path-specific indirect effects only after identifying the parameters of the marginal model of interest. As displayed in Fig. 3.6, for example, we might aim to derive the indirect path-specific effects after the marginalization over W_1. First of

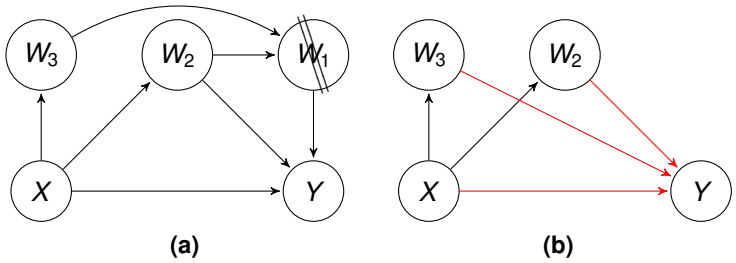


Figure 3.6: (a) Marginalization over the inner mediator W_1 and (b) quantifying the parameters (in red the parameters that change).

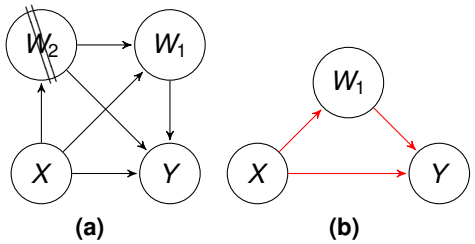


Figure 3.7: (a) Marginalization over the outer mediator W_2 and (b) quantifying the parameters (in red the parameters that change).

all, we can notice that, as expected, the total effect and the global indirect effect in the marginal model (Fig. 3.6(b)) are equal to the total and the global indirect effects in the original model (Fig. 3.6(a)). However, in Fig. 3.6(b) we can notice that parameters of the marginal model of interest has changed after the marginalization over W_1, implying that the only nonzero path-specific indirect effects are for $V = \{W_2\}$ and $V = \{W_3\}$. These effects can be evaluated from the total effect by using the new parameters of the marginal model. The parametric forms of such parameters are given in Appendix A.8.

Another interesting situation may arise when we marginalize over one intermediate/outer node, as displayed in the DAGs in Figure 3.7. In this situation more technicalities are necessary. Quantification of effects in models obtained after marginalization over intermediate or outer nodes involves repeated use of the derivations presented above. We detail the steps to be followed for the case with $k = 2$ mediators. Generalizations follow in a straightforward way. Suppose that we wish to evaluate the coefficients that arise in the model obtained after marginalization over W_2, i.e. with W_1 the only mediator; see Figure 3.7(b). This implies deriving the parametric formulation of

$$\log \frac{P(Y=1 \mid X=x, W_1=w_1)}{P(Y=0 \mid X=x, W_1=w_1)} = \log \frac{P(Y=1 \mid X=x, W_1=w_1, W_2=0)}{P(Y=0 \mid X=x, W_1=w_1, W_2=0)} + \log \frac{P(W_2=0 \mid Y=0, X=x, W_1=w_1)}{P(W_2=0 \mid Y=1, X=x, W_1=w_1)}. \tag{3.10}$$

Notice that the first term of the right hand side of the equation above is known, i.e. the original model (3.3). On the contrary, the second term of the right hand side of the equation needs to

be determined. Let B_{x,w_1} the second term of the right hand side of Eq. (3.10). From repeated use of the derivations in Stanghellini and Doretti (2019), we have that the parametric form of the marginal model over the outer mediator can be rewritten as:

$$\log \frac{P(Y = 1 \mid X = x, W_1 = w_1)}{P(Y = 0 \mid X = x, W_1 = w_1)} = \beta_0 + \beta_x x + \beta_{w_1} w_1 + \beta_{xw_1} x w_1 + \log B_{x,w_1} \tag{3.11}$$

with the parametric expression of B_{x,w_1} in Appendix A.9. Once derived this function, the values of the marginal parameters are straightforward; see for example Section 3.3.

3.5 Discussion

The proposed method generalizes the path-analysis approach commonly used with continuous variables. Behind the linear models, as well known, the most used models to handle with binary variables is the logistic regression. Iterative use of the total effect's decomposition allows to investigate complex research questions like quantifying the total, direct and indirect effects when some mediators are marginalized over. Beyond the mediation context, the method that we have proposed may be useful in longitudinal studies with a binary outcome measured at different points in time.

Although the derivations proposed in this chapter are developed in the path analysis context, if the recursive system of equation is a probabilistic causal model (Pearl 2009b), the total effect and some of its components can be seen as causal effects. The link between the total effect and its causal interpretation is intuitive. Concerning the direct effect, it can be seen as the controlled direct effect, if an external intervention would be made by setting $X = x$ while fixing all mediators to zero. Nonetheless, the causal definitions of indirect effects are more complex. A similar notion of controlled indirect effects does not exist, since it is not clear to what values other variables in the model should be fixed (Avin et al. 2005). The definitions proposed above, however, emphasize the total and controlled direct effect when a marginalization over some intermediate variables is made.

Conclusion

In this thesis we have broadly explored two important methods to decompose the total effect via mediation analysis. In the traditional approach of path analysis, mediation analysis has been developed in the context of linear models. Adopting the same reasoning of path analysis by interrupting the specific pathways between the variables of interest, we have provided novel definitions for the direct and indirect effects, which are more appropriate when non-linearities and interactions arise. On the other hand, in the counterfactual framework, mediation analysis has became very popular in the last twenty years. This framework is based on the causal DAG and no unmeasured confounders. A board range of contributions have covered different setting for the parametric identifications of the natural direct and indirect effects. However, in the setting with binary outcome and binary mediators modelled via logistic regressions some limitations were present. We overcame these issues proposing novel contributions to the exact parametric decomposition of the total effect with recursive system of logistic regressions. We have seen that the information about the parametric structure of the models allows flexible definition and identification of any type of effects. The results that we have proposed can be developed for sensitivity analysis and allow for possible extension to other setting, like for example to ordinal categorical data.

In Chapter 1, we have introduced the concept of the mediation analysis. In this setting, an intermediate variable is supposed to fall in the pathway between the treatment and the outcome and it allows to explain the underlying mechanisms of that relationship. We have seen that in the mediation literature there are two different approaches to evaluate the decomposition of the total effect into a direct and indirect effect. On one hand, we have introduced the traditional approach of Sobel (1982), Baron and Kenny (1986), Bollen (1987), and MacKinnon et al. (1991), who allows to identify the effects by simply sums the products between linear regression parameters. However, this approach lacks of well-specified assumptions for the causal interpretation of the effects and it does not work outside the linear models. In these regards, crucial works to mediation analysis have been offered by Robins and Greenland (1992) and Pearl (2001), who have given definitions of the effects in a counterfactual framework, identifiable from real data regardless the specific forms of the observed models. Furthermore, these definitions have been given in a setting that allows to a causal interpretation of the effects. Successively, novel contributions have been offered by VanderWeele (2013a, 2014), who has provided the highest insight of the total effect decomposition in the counterfactual framework. The total effect can be decomposed into three or four components allowing to isolate the phenomenon of mediation from that of interaction. Similarly, we have proposed

a novel decomposition of the total effect by using the method of path analysis. We have offered specific definitions of the total, direct, indirect and residual effect. The method proposed overcomes the issues of non-linearity, typical in structural equation modeling, and provides the definitions of the effects for any type of models. Furthermore, a parametric comparison between all the approaches has been offered. We have noticed that under the assumption of no-interaction the approaches bring back to the traditional one. Behind the linear case the traditional approach is no longer consistent to identify direct and indirect effect, thus, one should prefer either the counterfactual or the path analysis approach that we propose. In Chapter 1, we have focused in a setting with a countinuous outcome and a continuous mediator, giving definitions on a difference scale.

In Chapter 2, we have focused on mediation analysis for a binary outcome and a binary mediator, in both the counterfactual and the path analysis approaches. On one hand, we have derived an exact parametric decomposition of the total effect and its components, on the log odds ratio scale. We have not based our derivations on the rare outcome assumption, in contrast to what was done by Valeri and VanderWeele (2013). Furthermore, the proposed expressions are written in order to highlight the link between the definition of the effects and their pathway-specific coefficients of the corresponding logistic regression models, which we assumed to govern the data generating process of the outcome and the mediator. On the other hand, from the path analysis approach, we have presented a novel decomposition of the total effect which is based on the exact form of the marginal logistic regression model of Y against X only. We have overcome the issue of unequal variance which arises when fitting two nested models. For both approaches, we have generalized the expressions with regard to the presence of parametric interactions between the treatment, the mediator and the covariates. We have also derived the expressions of the approximate standard errors of the effect estimators, obtained via the delta method. Further, we have explored the links existing between the effects introduced in both frameworks, as for the continuous setting. Studying some cases of interest, we have emphasized the analytical and conceptual correspondence between the two approaches. As an empirical example, we have re-analyzed a dataset coming from a microcredit experiment performed in Bosnia and Herzegovina (Augsburg et al. 2015). We have estimated the decomposition of the overall effect of randomly allocated microloans on individual's capability in obtaining further loans from financial institutions, which we supposed to be mediated by whether or not the individual owns an active business. We have made the analysis in the counterfactual framework. The estimated effects are conditional on individual's age, educational level and number of active loans at baseline, allowing to assure no-unmeasured confounding of the mediator-outcome relationship and to interpret the estimated effects as causal. Finally, in a simulation study, we have examined the behaviour of the exact estimators and of their variance, contrasting them to the approximated ones of Valeri and VanderWeele (2013). Results have shown that exact estimators narrowly exceed the approximate ones, as expected.

In Chapter 3, we have generalized the previous results in a setting which includes multiple binary intermediate variables. Modelling the variables via a series of univariate logistic regressions, we have derived the exact form of the marginal model of Y against X only. In the path analysis approach, we have defined the direct and indirect effects plus a residual term due to the non-linearity. In particular for the indirect effects, we have distinguished between the

global indirect effect which incorporates simultaneously all the indirect paths from the treatment to the outcome, and the path-specific indirect effects which refer to each single indirect path. These latter effects involves only the parameters of their path and allow to appreciate the contribution of each single mediator or combination of them. As we have seen, the sum of all path-specific indirect effects does not reconstruct the global indirect effect. This is due to the multiple use of paths or to the residual effects which arise after the marginalization of non-linear models. The approach that we propose allows to appreciate also other effects of interest, like for example the indirect effect after the marginalization over an intermediate mediator. In the counterfactual framework, parametric identifications of the effects based on ratio scale are still missing. This is probably due to the difficulty of the nested counterfactuals involved in the definitions. Some recent developments have been offered by Steen et al. (2017a) who use particular models, so called natural effects models. However, this approach is not based on parametric assumptions and thus it does not offers effects identification as functions of the regression coefficients. Therefore, we can conceive our formulations in order to infer some causal conclusions. If we assume causal graphical models as explained in Pearl (2009b), the direct effect can be seen as the controlled direct effect of X on Y, i.e. the effect of the intervention performed setting all mediators to zero.

Knowing the exact parametric relationship between the treatment and the outcome allows to flexibly adapt the definitions of the effects for any type of model, and to easily extend our results to other roles of the intermediate variables, e.g. as source nodes; or as sink nodes but with some limitations. Further research may be addressed to the identification of direct and indirect effects with ordinal categorical data, in particular with several simultaneous categorical treatments. Relating to the counterfactual framework, we think that our results open the way for further developments mostly in the multiple mediators setting where parametric results with logistic regression models are still missing. They can be also applied to build bounds and to investigate methods for sensitivity analysis. As an instance, Lindmark et al. (2018) have introduced, in the counterfactual framework, the interval identification method for probit regression. The same method could be generalized to logistic models under our proposed results.

Appendix

A.1 Statistical interpretation of the A term under the case of a single binary mediator

The exact parametric form of the marginal logistic model of Y against X only by marginalizing over W has been recently derived by Stanghellini and Doretti (2019). In this Appendix we summarize the main findings in order to link their results to the formulation of the A term as in Eq. (2.7).

Let us write

$$\log \frac{P(W=1 \mid Y=y, X=x)}{P(W=0 \mid Y=y, X=x)} = \log \frac{P(Y=y \mid W=1, X=x)}{P(Y=y \mid W=0, X=x)} + \log \frac{P(W=1 \mid X=x)}{P(W=0 \mid X=x)}$$

and

$$\log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = -\log \frac{P(W=w \mid Y=1, X=x)}{P(W=w \mid Y=0, X=x)} + \log \frac{P(Y=1 \mid W=w, X=x)}{P(Y=0 \mid W=w, X=x)}.$$

We define $\log \frac{P(W=1|Y=y,X=x)}{P(W=0|Y=y,X=x)}$ by the function $g_y(x)$, that is

$$g_y(x) = y(\beta_w + \beta_{xw}x) + \log \frac{1+\exp(\beta_0+\beta_x x)}{1+\exp(\beta_0+\beta_x x+\beta_w+\beta_{xw}x)} + \gamma_0 + \gamma_x x. \tag{A.1.1}$$

Noticing that $(1+\exp g_y(x))^{-1}$ has the interpretation of $P(W=0 \mid Y=y, X=x)$.

Let $RR_{W|Y,X=x}$ be the relative risk of W for varying Y given $X=x$, that is

$$RR_{W|Y,X=x} = \frac{P(W=1 \mid Y=1, X=x)}{P(W=1 \mid Y=0, X=x)}.$$

It then follows

$$RR_{W|Y,X=x} = \frac{\exp g_1(x)\{1+\exp g_0(x)\}}{\exp g_0(x)\{1+\exp g_1(x)\}}.$$

Analogously, after denoting with $\bar{W} = 1 - W$, we have

$$RR_{\bar{W}|Y,X=x} = \frac{1+\exp g_0(x)}{1+\exp g_1(x)}.$$

Therefore, the marginal logistic model of Y given X only can be written as

$$\log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = \beta_0 + \beta_x x + \beta_w + \beta_{xw} x - \log RR_{W|Y,X=x}, \tag{A.1.2}$$

or alternatively as

$$\log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = \beta_0 + \beta_x x - \log RR_{\bar{W}|Y,X=x}. \tag{A.1.3}$$

After some algebra we can easily verify that the inverse of $RR_{\bar{W}|Y,X=x}$, i.e. $\frac{1+\exp g_1(x)}{1+\exp g_0(x)}$, equals $A_{x,x}$ as derived in Eq. (2.7).

A.2 Generalization to a set of k covariates

The inclusion of a set of covariates $C = (C_1, C_2, \ldots, C_k)$ follows in a straightforward way from the simple case of $C = 1$ shown in Sec. 2.4. Let $c = (c_1, c_2, \ldots, c_k)$ denote a possible value of C. Let $\text{rhs}(Y \mid W = w, X = x, C = c)$ denote the right hand side of the logit model for Y against X, W and C. Notice that this notation easily generalize to the inclusion of the entire set of possible interactions between variables. Furthermore, notice that $\text{rhs}(Y \mid W = 1, X = 0, C = 0)$ is the part of $\text{rhs}(Y \mid W = w, X = x, C = c)$ that contains w terms only (and the intercept). Let $\text{rhs}(W \mid X = x, C = c)$ be the right hand side of the logit model for W against X and C, which may include also the interaction terms.

The logit model of Y against X and C becomes:

$$\log \frac{P(Y=1 \mid X=x, C=c)}{P(Y=0 \mid X=x, C=c)} = \text{rhs}(Y \mid W=0, X=x, C=c) + \log \mathrm{A}_{x,x|c} \tag{A.2.1}$$

with

$$\mathrm{A}_{x,x|c} = \frac{p_1 p_2 p_3 + p_4}{p_2 p_3 + p_4} \tag{A.2.2}$$

in which

$$\begin{aligned}
p_1 &= \exp\{\text{rhs}(Y \mid W=1, X=x, C=c) - \text{rhs}(Y \mid W=0, X=x, C=c)\} \\
p_2 &= \exp\{\text{rhs}(W \mid X=x, C=c)\} \\
p_3 &= 1 + \exp\{\text{rhs}(Y \mid W=0, X=x, C=c)\} \\
p_4 &= 1 + \exp\{\text{rhs}(Y \mid W=1, X=x, C=c)\}
\end{aligned}$$

Clearly, also in the setting with k covariate, the function $\mathrm{A}_{x,x|c}$ corresponds to the inverse of the relative risk of $\bar{W} = W - 1$ for varying Y given $X = x$ and $C = c$. Thus, rewriting the marginal model as

$$\log \frac{P(Y=1 \mid X=x, C=c)}{P(Y=0 \mid X=x, C=c)} = \text{rhs}(Y \mid W=0, X=x, C=c) - \log \mathrm{RR}_{\bar{W}|Y,X=x,C=c} \tag{A.2.3}$$

with

$$\log \mathrm{RR}_{\bar{W}|Y,X=x,C=c} = \log \frac{1+\exp g_0(x,c)}{1+\exp g_1(x,c)},$$

in which

$$\begin{aligned} g_y(x,c) = y\{&\mathrm{rhs}(Y \mid W=1, X=x, C=c) - \mathrm{rhs}(Y \mid W=0, X=x, C=c)\} \\ &+ \log \frac{1+\exp\{\mathrm{rhs}(Y \mid W=0, X=x, C=c)\}}{1+\exp\{\mathrm{rhs}(Y \mid W=1, X=x, C=c)\}} + \mathrm{rhs}(W \mid X=x, C=c). \end{aligned} \tag{A.2.4}$$

The algebraic equivalence between $\mathsf{A}_{x,x|c}$ and inverse of $\mathrm{RR}_{\bar{W}|Y,X=x,C=c}$ is straightforward.

A.3 Mathematical derivations counterfactual effects

Given the assumptions in Sec. 1.3 of Ch. 1 and the definition of the causal effects on the log odds scale in Sec. 2.2 of Ch. 2, by using Pearl's mediation formula (Pearl (2001)) the causal effects can be non-parametrically identified. [6] For a binary mediator, the expression identifying the pure direct effect (Def. 2.2.2) is[7]

$$\mathrm{OR}^{PDE}_{x,x^*} = \frac{\overbrace{\sum_w P(Y=1 \mid X=x, W=w)P(W=w \mid X=x^\star) / \sum_w P(Y=0 \mid X=x, W=w)P(W=w \mid X=x^\star)}^{Q_1}}{\underbrace{\sum_w P(Y=1 \mid X=x^\star, W=w)P(W=w \mid X=x^\star) / \sum_w P(Y=0 \mid X=x^\star, W=w)P(W=w \mid X=x^\star)}_{Q_2}}$$

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

$$\begin{aligned} Q_1 &= \frac{P(Y=1 \mid X=x, W=1)P(W=1 \mid X=x^\star) + P(Y=1 \mid X=x, W=0)P(W=0 \mid X=x^\star)}{P(Y=0 \mid X=x, W=1)P(W=1 \mid X=x^\star) + P(Y=0 \mid X=x, W=0)P(W=0 \mid X=x^\star)} \\ &= \frac{\frac{\exp\{\beta_0+\beta_w+(\beta_x+\beta_{xw})x\}}{1+\exp\{\beta_0+\beta_w+(\beta_x+\beta_{xw})x\}} \times \frac{\exp(\gamma_0+\gamma_x x^\star)}{1+\exp(\gamma_0+\gamma_x x^\star)} + \frac{\exp(\beta_0+\beta_x x)}{1+\exp(\beta_0+\beta_x x)} \times \frac{1}{1+\exp(\gamma_0+\gamma_x x^\star)}}{\frac{1}{1+\exp\{\beta_0+\beta_w+(\beta_x+\beta_{xw})x\}} \times \frac{\exp(\gamma_0+\gamma_x x^\star)}{1+\exp(\gamma_0+\gamma_x x^\star)} + \frac{1}{1+\exp(\beta_0+\beta_x x)} \times \frac{1}{1+\exp(\gamma_0+\gamma_x x^\star)}} \\ &= \frac{\exp\{\beta_0+\beta_w+(\beta_x+\beta_{wx})x\}\exp(\gamma_0+\gamma_x x^\star)\{1+\exp(\beta_0+\beta_x x)\} + \exp(\beta_0+\beta_x x)[1+\exp\{\beta_0+\beta_w+(\beta_x+\beta_{xw})x\}]}{\exp(\gamma_0+\gamma_x x^\star)\{1+\exp(\beta_0+\beta_x x)\} + 1 + \exp\{\beta_0+\beta_w+(\beta_x+\beta_{xw})x\}} \\ &= \exp(\beta_0+\beta_x x)\mathsf{A}_{x,x^\star}. \end{aligned}$$

[6] See also VanderWeele and Vansteelandt (2010), Valeri and VanderWeele (2013), and VanderWeele (2015) and appendix there in, for the proof of the non-parametric identification of the counterfactual effects on the ratio scale.

[7] This appendix is the extension of Appendix A in Doretti et al. (2020).

For the denominator, an analogous calculation leads to $Q_2 = \exp(\beta_0+\beta_x x^*)\mathsf{A}_{x^*,x^*}$ and therefore to $\log \mathrm{OR}^{PDE}_{x,x^*} = \log Q_1 - \log Q_2 = \beta_x(x-x^*)+\log \mathsf{A}_{x,x^*} - \log \mathsf{A}_{x^*,x^*}$, which proves Equation (2.13). Derivations for the total indirect effect (Def. 2.2.3) are similar since we have

$$\mathrm{OR}^{TIE}_{x,x^*} = \frac{\overbrace{\sum_w P(Y=1 \mid X=x, W=w)P(W=w \mid X=x) / \sum_w P(Y=0 \mid X=x, W=w)P(W=w \mid X=x)}^{Q_3}}{\underbrace{\sum_w P(Y=1 \mid X=x, W=w)P(W=w \mid X=x^\star) / \sum_w P(Y=0 \mid X=x, W=w)P(W=w \mid X=x^\star)}_{Q_1}},$$

with $Q_3 = \exp(\beta_0 + \beta_x x)\mathsf{A}_{x,x}$, leading to $\log \mathrm{OR}^{TIE}_{x,x^\star} = \log Q_3 - \log Q_1 = \log \mathsf{A}_{x,x} - \log \mathsf{A}_{x,x^*}$, that is, Equation (2.14).

Similarly, the expression identifying the total direct effect (Def. 2.2.4) is

$$\mathrm{OR}^{TDE}_{x,x^*} = \frac{\overbrace{\sum_w P(Y=1 \mid X=x, W=w)P(W=w \mid X=x) / \sum_w P(Y=0 \mid X=x, W=w)P(W=w \mid X=x)}^{Q_3}}{\underbrace{\sum_w P(Y=1 \mid X=x^*, W=w)P(W=w \mid X=x) / \sum_w P(Y=0 \mid X=x^*, W=w)P(W=w \mid X=x)}_{Q_4}}$$

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

$$\begin{aligned}
Q_4 &= \frac{P(Y=1 \mid X=x^*, W=1)P(W=1 \mid X=x) + P(Y=1 \mid X=x^*, W=0)P(W=0 \mid X=x)}{P(Y=0 \mid X=x^*, W=1)P(W=1 \mid X) + P(Y=0 \mid X=x^*, W=0)P(W=0 \mid X=x)} \\
&= \frac{\frac{\exp\{\beta_0+\beta_w+(\beta_x+\beta_{xw})x^*\}}{1+\exp\{\beta_0+\beta_w+(\beta_x+\beta_{xw})x^*\}} \times \frac{\exp(\gamma_0+\gamma_x x)}{1+\exp(\gamma_0+\gamma_x x)} + \frac{\exp(\beta_0+\beta_x x^*)}{1+\exp(\beta_0+\beta_x x^*)} \times \frac{1}{1+\exp(\gamma_0+\gamma_x x)}}{\frac{1}{1+\exp\{\beta_0+\beta_w+(\beta_x+\beta_{xw})x^*\}} \times \frac{\exp(\gamma_0+\gamma_x x)}{1+\exp(\gamma_0+\gamma_x x)} + \frac{1}{1+\exp(\beta_0+\beta_x x^*)} \times \frac{1}{1+\exp(\gamma_0+\gamma_x x)}} \\
&= \frac{\exp\{\beta_0+\beta_w+(\beta_x+\beta_{wx})x^*\}\exp(\gamma_0+\gamma_x x)\{1+\exp(\beta_0+\beta_x x^*)\} + \exp(\beta_0+\beta_x x^*)[1+\exp\{\beta_0+\beta_w+(\beta_x+\beta_{xw})x^*\}]}{\exp(\gamma_0+\gamma_x x)\{1+\exp(\beta_0+\beta_x x^*)\} + 1 + \exp\{\beta_0+\beta_w+(\beta_x+\beta_{xw})x^*\}} \\
&= \exp(\beta_0+\beta_x x^*)\mathsf{A}_{x^*,x}.
\end{aligned}$$

For the numerator, an analogous calculation leads to $Q_3 = \exp(\beta_0 + \beta_x x)\mathsf{A}_{x,x}$ and therefore to $\log \mathrm{OR}^{TDE}_{x,x^*} = \log Q_3 - \log Q_4 = \beta_x(x-x*) + \log \mathsf{A}_{x,x} - \log \mathsf{A}_{x^*,x}$, which proves Equation (2.15). Derivations for the pure indirect effect (Def. 2.2.5) are similar since we have

$$\mathrm{OR}^{PIE}_{x,x^*} = \frac{\overbrace{\sum_w P(Y=1 \mid X=x^*, W=w)P(W=w \mid X=x) / \sum_w P(Y=0 \mid X=x^*, W=w)P(W=w \mid X=x)}^{Q_4}}{\underbrace{\sum_w P(Y=1 \mid X=x^*, W=w)P(W=w \mid X=x^*) / \sum_w P(Y=0 \mid X=x^*, W=w)P(W=w \mid X=x^*)}_{Q_2}},$$

with $Q_2 = \exp(\beta_0 + \beta_x x^*)\mathsf{A}_{x^*,x^*}$, leading to $\log \mathrm{OR}^{PIE}_{x,x^*} = \log Q_4 - \log Q_2 = \log \mathsf{A}_{x^*,x} - \log A_{x^*,x^*}$, that is, Equation (2.16).

Notice that once derived the parametric expression of Q_1, Q_2, Q_3 and Q_4, the reference interaction and the mediated interaction are straightforward. Furthermore, the mathematical developments above remain unchanged even if we add a set of covariates C. We refer to the Appendix A.2 for the generalization to the A term with k covariates.

A.4 Identification of treatment effects with X continuous

In this appendix, we report and extend results previously derived in Raggi et al. (2020). Let us consider a treatment X continuous and differentiable. Thus, the total effect is defined as the derivative of (2.6) or equivalently (A.1.3) with respect to x. For simplicity, we derive the following results considering the marginal model as expressed in Eq. (A.1.3). Therefore:

$$\begin{aligned}\log \mathrm{OR}^{TE}(x) &= \frac{d}{dx}\left[\beta_0 + \beta_x x + \log\frac{1+\exp g_1(x)}{1+\exp g_0(x)}\right] \\ &= \beta_x + P(W=1 \mid Y=1, X=x)\frac{d}{dx}g_1(x) - P(W=1 \mid Y=0, X=x)\frac{d}{dx}g_0(x).\end{aligned}$$

After some algebra, it is possible to show that:

$$\begin{aligned}\log \mathrm{OR}^{TE}(x) = {} & \beta_x\{1 - \Delta_y(x)\Delta_w(x)\} \\ & + \beta_{xw}\{P(W=1|Y=1, X=x) - \Delta_w(x)P(Y=1 \mid W=1, X=x)\} \\ & + \gamma_x \Delta_w(x)\end{aligned} \tag{A.4.1}$$

where

$$\begin{aligned}\Delta_y(x) &= P(Y=1 \mid W=1, X=x) - P(Y=1 \mid W=0, X=x) \\ &= \frac{\exp(\beta_0 + \beta_x x + \beta_w + \beta_{xw}x)}{1+\exp(\beta_0 + \beta_x x + \beta_w + \beta_{xw}x)} - \frac{\exp(\beta_0 + \beta_x x)}{1+\exp(\beta_0 + \beta_x x)}\end{aligned}$$

and

$$\begin{aligned}\Delta_w(x) &= P(W=1 \mid Y=1, X=x) - P(W=1 \mid Y=0, X=x) \\ &= \frac{\exp g_1(x)}{1+\exp g_1(x)} - \frac{\exp g_0(x)}{1+\exp g_0(x)}\end{aligned}$$

with $g_y(x)$ as in (A.1.1). Eq (A.4.1) confirms the well-known fact that the marginal logistic model is non linear in x, also providing the explicit expression of it. Notice that all terms in curly bracket are bounded between 0 and 1, while $\Delta_w(x)$ is bounded between -1 and 1. Notice further that $\Delta_w(x)$ and $\Delta_y(x)$ share the same sign, and they are both zero whenever $W \perp\!\!\!\perp Y \mid X$. See Stanghellini and Doretti (2019) to which we refer for the proof.

Therefore, we can derive the components of the total effect for the case of a continuous treatment X. The indirect effect of X on Y through W, following the Definition 2.3.2, is obtained after assuming, in the total effect, $\beta_x = \beta_{xw} = 0$, that is

$$\log \mathrm{OR}^{IE}(x) = \log \mathrm{OR}^{TE}(x) \mid_{\beta_x=\beta_{xw}=0} = \gamma_x \Delta_w^*(x) \tag{A.4.2}$$

where $\Delta_w^*(x)$ is $\Delta_w(x)$ evaluated at $\beta_x = \beta_{xw} = 0$. The indirect effect depends on the value of x, and is null if either γ_x or β_w are zero.

The direct effect of X on Y which is not transmitted by W, following the Definition 2.3.3, is obtained after assuming, in the total effect, $\beta_w = \beta_{xw} = 0$, that is

$$\log \mathrm{OR}^{DE}(x) = \log \mathrm{OR}^{TE}(x) \mid_{\beta_w=\beta_{xw}=0} = \beta_x. \tag{A.4.3}$$

Notice that (A.4.3) follows as $\Delta_y(x)$ and $\Delta_w(x)$ are zero when $\beta_w = \beta_{xw} = 0$.

Instead, the residual effect of X on Y, following the Definition 2.3.4, is given by difference as follows

$$\begin{aligned}\log \mathrm{OR}^{RE}(x) &= \log \mathrm{OR}^{TE}(x) - \log \mathrm{OR}^{IE}(x) - \log \mathrm{OR}^{DE}(x) \\ &= -\beta_x \Delta_y(x) \Delta_w(x) \\ &+ \beta_{xw}\{P(W = 1|Y = 1, X = x) - \Delta_w(x) P(Y = 1 \mid W = 1, X = x)\} \\ &+ \gamma_x\{\Delta_w(x) - \Delta_w^*(x)\}.\end{aligned} \tag{A.4.4}$$

Notice that the effect above vanishes whenever $\beta_x = \beta_{xw} = 0$ or $\beta_w = \beta_{xw} = 0$. As a matter of fact, in the former case we have $\Delta_w(x) = \Delta_w^*(x)$, whereas in the latter case $\Delta_w(x) = \Delta_y(x) = 0$. Notice that the latter case coincides with the condition of collapsibility of odds ratio.

On the other hand, as done for the case of a discrete treatment X, we can derive an alternative version of the direct and residual effect. Following the definition 2.3.5, the direct effect is obtained after assuming, in the total effect, $\gamma_x = 0$, that is

$$\begin{aligned}\log \mathrm{OR}^{DE^{nc}}(x) &= \log \mathrm{OR}^{TE}(x) \mid_{\gamma_x = 0} \\ &= \beta_x\{1 - \Delta_y(x)\{\Delta_w(x)\} \mid_{\gamma_x = 0}\} \\ &+ \beta_{xw}\{P(W = 1|Y = 1, X = x) - \Delta_w(x) P(Y = 1 \mid W = 1, X = x)\} \mid_{\gamma_x = 0}\end{aligned} \tag{A.4.5}$$

In this case, there is an effect modification due to conditioning of an additional variable, in line with well-known results on non-collapsibility of parameters of logistic regression models. In addition, we notice that even in this simple case the linearity of X in the marginal model is lost.

Finally, the residual effect, following the definition 2.3.6, is given by difference as follows

$$\begin{aligned}\log \mathrm{OR}^{RE^{nc}}(x) &= \log \mathrm{OR}^{TE}(x) - \log \mathrm{OR}^{IE}(x) - \log \mathrm{OR}^{DE^{nc}}(x) \\ &= \beta_x \Delta_y(x)\{\Delta_w(x) \mid_{\gamma_x = 0} - \Delta_w(x)\} \\ &+ \beta_{xw}\{P(W = 1 \mid Y = 1, X = x) - P(W = 1 \mid Y = 1, X = x) \mid_{\gamma_x = 0} \\ &+ P(Y = 1 \mid W = 1, X = x)\{\Delta_w(x) \mid_{\gamma_x = 0} - \Delta_w(x)\}\} \\ &+ \gamma_x(\Delta_w(x) - \Delta_w^*(x))\end{aligned} \tag{A.4.6}$$

with $\Delta_w^*(x)$ as above.

Reformulating the total effect as follows

$$\begin{aligned}\log \mathrm{OR}^{TE}(x) &= \log \mathrm{OR}^{IE}(x) + \log \mathrm{OR}^{DE} + \log \mathrm{OR}^{RE} \\ &= \log \mathrm{OR}^{IE}(x) + \log \mathrm{OR}^{DE^{nc}} + \log \mathrm{OR}^{RE^{nc}}\end{aligned}$$

one may be interested in studying the total effect under some relevant situation, as we have done for the case of discrete treatment in Sec. 2.5. We refer to Raggi et al. (2020) for the main results.

A.5 Variance-covariance matrix of estimated effects

In this appendix we report the results for the standard errors of the estimated effects, generalizing that in the Appendix B of Doretti et al. (2020).

Let us denote $\boldsymbol{\beta} = (\beta_0, \beta_x, \beta_z, \beta_{xz}, \beta_w, \beta_{xw}, \beta_{wz}, \beta_{xwz})'$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_x, \gamma_v, \gamma_{xv})'$ the two vectors of model parameters and by $\boldsymbol{\Sigma}_{\hat{\beta}}$ and $\boldsymbol{\Sigma}_{\hat{\gamma}}$ the variance-covariance matrices of their estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$. Further, let us denote

$$\boldsymbol{e} = (\mathrm{OR}^{\mathrm{PDE}}_{x,x^\star|c}, \mathrm{OR}^{\mathrm{TIE}}_{x,x^\star|c}, \mathrm{OR}^{\mathrm{TDE}}_{x,x^\star|c}, \mathrm{OR}^{\mathrm{PIE}}_{x,x^\star|c}, \mathrm{OR}^{\mathrm{TE}}_{x,x^\star|c})'$$

the true causal effects vector. The first-order approximate variance-covariance matrix of the estimator

$$\hat{\boldsymbol{e}} = (\widehat{\mathrm{OR}}^{\mathrm{PDE}}_{x,x^\star|c}, \widehat{\mathrm{OR}}^{\mathrm{TIE}}_{x,x^\star|c}, \widehat{\mathrm{OR}}^{\mathrm{TDE}}_{x,x^\star|c}, \widehat{\mathrm{OR}}^{\mathrm{PIE}}_{x,x^\star|c}, \widehat{\mathrm{OR}}^{\mathrm{TE}}_{x,x^\star|c})'$$

can be obtained as $V(\hat{\boldsymbol{e}}) = \boldsymbol{E}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{E}'$, where $\boldsymbol{E} = \mathrm{diag}(\boldsymbol{e})$,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\hat{\beta}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\hat{\gamma}} \end{pmatrix}$$

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

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

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

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

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

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

while the others can be written as functions thereof. Specifically, a compact form for $\boldsymbol{d}_{x,x^\star|c}$ is given by

$$\boldsymbol{d}_{x,x^\star|c} = [(d_{\beta_0}(x, x^\star \mid c), d_{\beta_w}(x, x^\star \mid c)) \otimes \boldsymbol{d}(x, z)\,,\; d_{\gamma_0}(x, x^\star \mid c)\boldsymbol{d}(x^\star, v)],$$

where $\otimes$ denotes the Kronecker product and, letting $\boldsymbol{I}_2$ be a diagonal matrix of order 2, $\boldsymbol{d}(a, b)$ is the row vector returned by the vector-matrix multiplication $\boldsymbol{d}(a, b) = (1\,,\, a)[(1\,,\, b) \otimes \boldsymbol{I}_2]$. The

vectors $\boldsymbol{d}_{x,x|c}$, $\boldsymbol{d}_{x^\star,x^\star|c}$ and $\boldsymbol{d}_{x^\star,x|c}$ can be calculated applying the same formulas above to $A_{x,x|c}$, $A_{x^\star,x^\star|c}$ and $A_{x^\star,x|c}$ respectively. Then, the matrix $\boldsymbol{D}$ can be obtained as

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{d}_1 + \boldsymbol{d}_2 \\ \boldsymbol{d}_3 \\ \boldsymbol{d}_1 + \boldsymbol{d}_4 \\ \boldsymbol{d}_5 \\ \boldsymbol{d}_6 \end{pmatrix},$$

where

$$\begin{aligned}
\boldsymbol{d}_2 &= \boldsymbol{d}_{x,x^\star}/A_{x,x^\star} - \boldsymbol{d}_{x^\star,x^\star}/A_{x^\star,x^\star} \\
\boldsymbol{d}_3 &= \boldsymbol{d}_{x,x}/A_{x,x} - \boldsymbol{d}_{x,x^\star}/A_{x,x^\star} \\
\boldsymbol{d}_4 &= \boldsymbol{d}_{x,x}/A_{x,x} - \boldsymbol{d}_{x^\star,x}/A_{x^\star,x} \\
\boldsymbol{d}_5 &= \boldsymbol{d}_{x^\star,x}/A_{x^\star,x} - \boldsymbol{d}_{x^\star,x^\star}/A_{x^\star,x^\star} \\
\boldsymbol{d}_6 &= \boldsymbol{d}_{x,x}/A_{x,x} - \boldsymbol{d}_{x^\star,x^\star}/A_{x^\star,x^\star}
\end{aligned}$$

while $\boldsymbol{d}_1$ is a row vector of the same length of θ with all its components set to zero but the ones in the positions of β_x and β_{xz}, worth $x - x^\star$ and $z(x - x^\star)$ respectively. Again, extension to multiple confounders is immediate provided that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are extended as follows:

$$\begin{aligned}
\boldsymbol{\beta} &= (\beta_0, \beta_x, \beta_{z_1}, \dots, \beta_{z_p}, \beta_{xz_1}, \dots, \beta_{xz_p}, \beta_w, \beta_{xw}, \beta_{wz_1}, \dots, \beta_{wz_p}, \beta_{xwz_1}, \dots, \beta_{xwz_p})' \\
\boldsymbol{\gamma} &= (\gamma_0, \gamma_x, \gamma_{v_1}, \dots, \gamma_{v_q}, \gamma_{xv_1}, \dots, \gamma_{xv_p})'.
\end{aligned}$$

The same logic applies also in the path analysis approach. Letting

$$\boldsymbol{e} = (\mathrm{OR}^{DE}_{x,x^\star|c}, \mathrm{OR}^{IE}_{x,x^\star|c}, \mathrm{OR}^{DE^{nc}}_{x,x^\star|c}, \mathrm{OR}^{TE}_{x,x^\star|c})'$$

the true effects vector. The first-order approximate variance-covariance matrix of the estimator

$$\hat{\boldsymbol{e}} = (\hat{\mathrm{OR}}^{DE}_{x,x^\star|c}, \hat{\mathrm{OR}}^{IE}_{x,x^\star|c}, \hat{\mathrm{OR}}^{DE^{nc}}_{x,x^\star|c}, \hat{\mathrm{OR}}^{TE}_{x,x^\star|c})'$$

can be obtained as $V(\hat{\boldsymbol{e}}) = \boldsymbol{E}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{E}'$, where $\boldsymbol{E} = \mathrm{diag}(\boldsymbol{e})$,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\hat{\beta}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\hat{\gamma}} \end{pmatrix}$$

To obtain $\boldsymbol{D}$ we need the vectors $\boldsymbol{d}_{x,x|c}$, $\boldsymbol{d}_{x^\star,x^\star|c}$, $\boldsymbol{d}^*_{x,x|c}$, $\boldsymbol{d}^*_{x^\star,x^\star|c}$ $\boldsymbol{d}^{nc}_{x,x|c}$ and $\boldsymbol{d}^{nc}_{x^\star,x^\star|c}$ can be calculated applying the same formulas above to $A_{x,x|c}$, $A_{x^\star,x^\star|c}$, $A^*_{x,x|c}$, $A^*_{x^\star,x^\star|c}$, $A^{nc}_{x,x|c}$ and $A^{nc}_{x^\star,x^\star|c}$ respectively. Then, the matrix $\boldsymbol{D}$ can be obtained as

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \\ \boldsymbol{f}_1 + \boldsymbol{f}_3 \\ \boldsymbol{f}_4 \end{pmatrix},$$

where

$$\begin{aligned}\boldsymbol{f}_2 &= \boldsymbol{d}^*_{x,x^\star}/A^*_{x,x^\star} - \boldsymbol{d}^*_{x^\star,x^\star}/A^*_{x^\star,x^\star}\\ \boldsymbol{f}_3 &= \boldsymbol{d}^{nc}_{x,x}/A^{nc}_{x,x} - \boldsymbol{d}^{nc}_{x,x}/A^{nc}_{x,x}\\ \boldsymbol{f}_4 &= \boldsymbol{d}_{x,x}/A_{x,x} - \boldsymbol{d}_{x^\star,x^\star}/A_{x^\star,x^\star}\end{aligned}$$

while $\boldsymbol{f}_1$ is a row vector of the same length of $\boldsymbol{\theta}$ with all its components set to zero but the ones in the positions of β_x, worth $x - x^\star$.

Clearly, in finite-sample analyses one has plug in the estimates $\hat{\beta}$ and $\hat{\gamma}$ everywhere in the formulae above to obtain the estimated variance/covariance matrix $\hat{V}(\hat{\boldsymbol{e}})$.

A.6 Statistical interpretation of the A term with k mediators

In this appendix we proof the procedure to obtain the marginal model of Y against X in Eq. (3.1), which is obtained after k marginalization and, then, the final formulation of the A term as in Eq. (3.2). The expression of $\mathrm{A}^{(w_{<j})}_{x|w_{>j}}$ can be derived recursively, after noting that $\mathrm{A}_{x|w_{>1}}$ is $\mathrm{A}_{x,x|c}$ in Eq. (A.2.1) after imposing $W = W_1$ and $C = W_{>1}$.

We marginalize iteratively over the inner mediator, that is starting from the marginalization over W_1, we have

$$\mathrm{rhs}(Y \mid X = x, W_2 = w_2, \dots, W_k = w_k) = \mathrm{rhs}(Y \mid X = x, W_1 = 0, W_2 = w_2, \dots, W_k = w_k) + \log \mathrm{A}_{x|w_{>1}}$$

and then over W_2, we have

$$\begin{aligned}\mathrm{rhs}(Y \mid X = x, W_3 = w_3, \dots, W_k = w_k) = \mathrm{rhs}(Y \mid X = x, W_1 = 0, W_2 = 0, W_3 = w_3, \dots, W_k = w_k)\\ + \log \mathrm{A}_{x|W_2=0,W_{>2}=w_{>2}} \mathrm{RR}_{\bar{W}_1|Y,W_2=0,X=x,W_{>2}=w_{>2}} + \log \mathrm{A}^{(w_1)}_{x|w_{>2}}\end{aligned}$$

up to marginalization over W_k, that is

$$\mathrm{rhs}(Y \mid X = x) = \mathrm{rhs}(Y \mid X = x, W_1 = 0, \dots W_k = 0) + \sum_{j=1}^{k} \log \mathrm{A}^{(w_{<j})}_{x|W_{>j}=0}$$

with, at each step, $\mathrm{A}^{(w_{<j})}_{x|w_{>j}}$ is (3.1).

Furthermore, from Appendix A.2 we have shown that $\mathrm{A}_{x,x|c}$ equals the inverse of the relative risk $\mathrm{RR}_{\bar{W}|Y,X=x,C=c}$. Therefore, it follows that if we denote with $g^{(w_{<j})}_y(x, w_{>j})$ the function (A.2.4) evaluated after marginalizing over the $W_{<j}$ mediators and considering $(W_{j+1}, \dots, W_k)$ as covariates. Analogously, the expression of $g^{(w_{<j})}_y(x, w_{>j})$ can be derived recursively, after noting that $g_y(x, w_{>1})$ is (A.2.4) after imposing $W = W_1$ and $C = W_{>1}$. We marginalize iteratively over the inner mediator, that is starting from the marginalization over W_1. We have

$$\begin{aligned}\mathrm{rhs}(Y \mid X = x, W_2 = w_2, \dots, W_k = w_k) = \mathrm{rhs}(Y \mid X = x, W_1 = 0, W_2 = w_2, \dots, W_k = w_k)\\ - \log \mathrm{RR}_{\bar{W}_1|Y,X=x,W_{>1}=w_{>1}}\end{aligned} \tag{A.6.1}$$

and then over W_2

$$\begin{aligned}\text{rhs}(Y \mid X = x, W_3 = w_3, \dots, W_k = w_k) = \text{rhs}(Y \mid X = x, W_1 = 0, W_2 = 0, W_3 = w_3, \dots, W_k = w_k) \\ - \log \text{RR}_{\bar{W}_1 \mid Y, W_2=0, X=x, W_{>2}=w_{>2}} \\ - \log \text{RR}_{\bar{W}_2 \mid Y, X=x, W_{>2}=w_{>2}}\end{aligned} \tag{A.6.2}$$

up to marginalization over W_k, that is

$$\text{rhs}(Y \mid X = x) = \text{rhs}(Y \mid X = x, W_1 = 0, \dots W_k = 0) - \sum_{j=1}^{k} \log \text{RR}_{\bar{W}_j \mid Y, X=x, W_{>j}=0} \tag{A.6.3}$$

where, at each step,

$$\log \text{RR}_{\bar{W}_j \mid Y, X=x, W_{>j}} = \log \frac{1 + \exp g_0^{(w_{<j})}(x, w_{>j})}{1 + \exp g_1^{(w_{<j})}(x, w_{>j})}$$

and

$$\begin{aligned}g_y^{(w_{<j})}(x, w_{>j}) = y\{\text{rhs}(Y \mid W_j = 1, X = x, W_{>j} = w_{>j}) - \text{rhs}(Y \mid W_j = 0, X = x, W_{>j} = w_{>j})\} \\ + \log \frac{1 + \exp\{\text{rhs}(Y \mid W_j = 0, X = x, W_{>j} = w_{>j})\}}{1 + \exp\{\text{rhs}(Y \mid W_j = 1, X = x, W_{>j} = w_{>j})\}} + \text{rhs}(W_j \mid X = x, W_{>j} = w_{>j}).\end{aligned} \tag{A.6.4}$$

The algebraic equivalence between $\text{RR}^{-1}_{\bar{W}_j \mid Y, X=x, W_{>j}}$ and $\text{A}^{(w_{<j})}_{x \mid w_{>j}}$ follows in a straightforward way.

A.7 Marginal model with $k = 2$ and X continuous

In this appendix we offer results fo a setting with a continuous and differentiable treatment X and $k = 2$ mediators. The total effect of X on Y is given by taking the derivative of the marginal model as derived in Eq. (3.8) w.r. to x. For simplicity, we use the alternative formulation with the inverse of the relative risk, i.e. by using the $g_y(x, w_2)$ functions, as expressed in Eq.(A.2.4) after imposing $C = W_2$ (see Appendix A.2). Therefore:

$$\begin{aligned}\log \text{OR}^{TE}(x) &= \frac{d}{dx}\left[\beta_0 + \beta_x x + \log \frac{1 + \exp g_1(x, 0)}{1 + \exp g_0(x, 0)} + \log \frac{1 + \exp g_1^{(w_1)}(x)}{1 + \exp g_0^{(w_1)}(x)}\right] \\ &= \beta_x + P(W_1 = 1 \mid Y = 1, X = x, W_2 = 0)\frac{d}{dx} g_1(x, 0) \\ &- P(W_1 = 1 \mid Y = 0, X = x, W_2 = 0)\frac{d}{dx} g_0(x, 0) \\ &+ P(W_2 = 1 \mid Y = 1, X = x)\frac{d}{dx} g_1^{(w_1)}(x) \\ &- P(W_2 = 1 \mid Y = 1, X = x)\frac{d}{dx} g_0^{(w_1)}(x)\end{aligned} \tag{A.7.1}$$

where the parametric form of the probabilities $P(W_1 = 1 \mid Y = y, X = x, W_2 = 0)$ and $P(W_2 = 1 \mid Y = y, X = x)$ are respectively

$$P(W_1 = 1 \mid Y = y, X, W_2 = 0) = \frac{\exp g_y(x, 0)}{1 + \exp g_y(x, 0)}$$

and

$$P(W_2 = 1 \mid Y = y, X) = \frac{\exp g_y^{(w_1)}(x, 0)}{1 + \exp g_y^{(w_1)}(x, 0)}.$$

On the other hand, the derivatives of $g_y(x)$ and $g_y^{(w_1)}(x)$ are respectively

$$\begin{aligned}\frac{d}{dx} g_y(x, w_2) = y(\beta_{xw_1}) &+ \frac{\exp\{\text{rhs}(Y \mid W_1 = 0, X = x, W_2 = w_2)\}}{1 + \exp\{\text{rhs}(Y \mid W_1 = 0, X = x, W_2 = w_2)\}}(\beta_x + \beta_{xw_2} w_2) \\ &- \frac{\exp\{\text{rhs}(Y \mid W_1 = 1, X = x, W_2 = w_2)\}}{1 + \exp\{\text{rhs}(Y \mid W_1 = 1, X = x, W_2 = w_2)\}}(\beta_x + \beta_{xw_1} + \beta_{xw_2} w_2) \\ &+ \gamma_{1,x} + \gamma_{1,xw_2} w_2\end{aligned}$$

and

$$\begin{aligned}&\frac{d}{dx} g_y^{(w_1)}(x) = y\Big\{\beta_{xw_2} + \frac{\exp g_1(x, 1)}{1 + \exp g_1(x, 1)} \frac{d}{dx} g_1(x, 1) - \frac{\exp g_0(x, 1)}{1 + \exp g_0(x, 1)} \frac{d}{dx} g_0(x, 1) \\ &- \frac{\exp g_1(x, 0)}{1 + \exp g_1(x, 0)} \frac{d}{dx} g_1(x, 0) + \frac{\exp g_0(x, 0)}{1 + \exp g_0(x, 0)} \frac{d}{dx} g_0(x, 0)\Big\} + \frac{\exp\{\text{rhs}(Y \mid X, W_2 = 0)\}}{1 + \exp\{\text{rhs}(Y \mid X, W_2 = 0)\}} \\ &\times \big(\beta_x + P(W_1 = 1 \mid Y = 1, X, W_2 = 0) \frac{d}{dx} g_1(x, 0) \\ &- P(W_1 = 1 \mid Y = 0, X, W_2 = 0) \frac{d}{dx} g_0(x, 0)\big) - \frac{\exp\{\text{rhs}(Y \mid X, W_2 = 1)\}}{1 + \exp\{\text{rhs}(Y \mid X, W_2 = 1)\}} \\ &\times \big(\beta_x + \beta_{xw_2} + P(W_1 = 1 \mid Y = 1, X, W_2 = 1) \frac{d}{dx} g_1(x, 1) \\ &- P(W_1 = 1 \mid Y = 0, X, W_2 = 1) \frac{d}{dx} g_0(x, 1)\big) + \gamma_{2,x}\end{aligned}$$

where

$$\text{rhs}(Y \mid X, W_2) = \beta_0 + \beta_x x + \beta_{w_2} w_2 + \beta_{xw_2} x w_2 + \log \frac{1 + \exp g_1(x, w_2)}{1 + \exp g_0(x, w_2)}.$$

Notice that the total effect in Eq.(A.7.1) is exact. this result is different from what done in Stanghellini and Doretti (2019) where an analogous decomposition is computed for a particular value x_0 and using a linear approximation at that point.

Finally, once obtained the expression for the total effect, its decomposition follows the same rules for the discrete case; see Section 3.3. Thus, the parametric forms for the direct, indirect and residual effects follow in a straightforward way.

A.8 Marginal parameters with $k = 3$

Let us consider the model with $k = 3$ and X binary. No conditional independences are assumed. We report the parametric forms for the parameters changed after the marginalization over W_1. Let us consider four logistic regression models, respectively, for the outcome Y, and for the three mediators W_1, W_2, W_3, with interaction terms up to the second order.

Notice that marginalization over the inner mediator induces changes in the parameters of the outcome equation only. From Eq. (A.2.1), considering W_2 and W_3 as covariates and marginalizing over W_1 we obtain the first marginal model for Y against X, W_2, W_3 such as

$$\begin{aligned}\log \frac{P(Y = 1 \mid X = x, W_2 = w_2, W_3 = w_3)}{P(Y = 0 \mid X = x, W_2 = w_2, W_3 = w_3)} &= \beta_0^{(w_1)} + \beta_x^{(w_1)} x + \sum_{j=2}^{k} \beta_{w_j}^{(w_1)} w_j \\ &+ \sum_{j=2}^{k} \beta_{xw_j}^{(w_1)} x w_j + \beta_{xw_2w_3}^{(w_1)} x w_2 w_3 \\ &= \beta_0 + \beta_x x + \beta_{w_2} w_2 + \beta_{xw_2} x w_2 \\ &+ \log \mathsf{A}_{x|w_2,w_3}\end{aligned}$$

where where $\mathsf{A}_{x|w_2,w_3}$ is obtained from (A.2.2) after imposing $W = W_1$ and $C = (W_2, W_3)$. The intercept, the main effects, the second order interactions and the third order interaction, respectively, are:

$$\beta_0^{(w_1)} = \beta_0 + \log \mathsf{A}_{0|W_2=0,W_3=0},$$

$$\begin{aligned}\beta_x^{(w_1)} &= \log \mathrm{OR}(Y, X \mid W_2 = 0, W_3 = 0) \\ &= \beta_x + \log \mathsf{A}_{1|W_2=0,W_3=0} - \log \mathsf{A}_{0|W_2=0,W_3=0},\end{aligned}$$

$$\begin{aligned}\beta_{w_2}^{(w_1)} &= \log \mathrm{OR}(Y, W_2 \mid X = 0, W_3 = 0) \\ &= \beta_{w_2} + \log \mathsf{A}_{0|W_2=1,W_3=0} - \log \mathsf{A}_{0|W_2=0,W_3=0},\end{aligned}$$

$$\begin{aligned}\beta_{w_3}^{(w_1)} &= \log \mathrm{OR}(Y, W_3 \mid X = 0, W_2 = 0) \\ &= \beta_{w_3} + \log \mathsf{A}_{0|W_2=0,W_3=1} - \log \mathsf{A}_{0|W_2=0,W_3=0}.\end{aligned},$$

$$\begin{aligned}\beta_{xw_2}^{(w_1)} &= \log \mathrm{OR}(Y, X \mid W_2 = 1, W_3 = 0) - \log \mathrm{OR}(Y, X \mid W_2 = 0, W_3 = 0) \\ &= \beta_{xw_2} + \log \mathsf{A}_{1|W_2=1,W_3=0} - \log \mathsf{A}_{0|W_2=1,W_3=0} \\ &- \log \mathsf{A}_{1|W_2=0,W_3=0} + \log \mathsf{A}_{0|W_2=0,W_3=0},\end{aligned}$$

$$\begin{aligned}\beta_{xw_3}^{(w_1)} &= \log \mathrm{OR}(Y, X \mid W_2 = 0, W_3 = 1) - \log \mathrm{OR}(Y, X \mid W_2 = 0, W_3 = 0) \\ &= \beta_{xw_3} + \log \mathsf{A}_{1|W_2=0,W_3=1} - \log \mathsf{A}_{0|W_2=0,W_3=1} \\ &- \log \mathsf{A}_{1|W_2=0,W_3=0} + \log \mathsf{A}_{0|W_2=0,W_3=0},\end{aligned}$$

$$\begin{aligned}\beta^{(w_1)}_{w_2w_3} &= \log \mathrm{OR}(Y, W_2 \mid X = 0, W_3 = 1) - \log \mathrm{OR}(Y, W_2 \mid X = 0, W_3 = 0) \\ &= \beta_{w_2w_3} + \log \mathrm{A}_{0|W_2=1,W_3=1} - \log \mathrm{A}_{0|W_2=0,W_3=1} \\ &\quad - \log \mathrm{A}_{0|W_2=1,W_3=0} + \log \mathrm{A}_{0|W_2=0,W_3=0}\end{aligned}$$

and

$$\begin{aligned}\beta^{(w_1)}_{xw_2w_3} &= [\log \mathrm{OR}(Y, X \mid W_2 = 1, W_3 = 1) - \log \mathrm{OR}(Y, X \mid W_2 = 1, W_3 = 0)] \\ &\quad - [\log \mathrm{OR}(Y, X \mid W_2 = 0, W_3 = 1) - \log \mathrm{OR}(Y, X \mid W_2 = 0, W_3 = 0)] \\ &= \log \mathrm{A}_{1|W_2=1,W_3=1} - \log \mathrm{A}_{0|W_2=1,W_3=1} - \log \mathrm{A}_{1|W_2=1,W_3=0} + \log \mathrm{A}_{0|W_2=1,W_3=0} \\ &\quad - \log \mathrm{A}_{1|W_2=0,W_3=1} + \log \mathrm{A}_{0|W_2=0,W_3=1} + \log \mathrm{A}_{1|W_2=0,W_3=0} - \log \mathrm{A}_{0|W_2=0,W_3=0}.\end{aligned}$$

Theses quantities are generic for models without any restrictions in terms of conditional independence assumptions. Therefore, for the specific example depicted in Fig. 3.6, the values of the marginalized parameters must take into account the conditional independences of the corresponding models.

Notice that even if we did not include any type of interaction in the full model, after the first marginalization, the interactions of second and third order appear.

A.9 Marginal model over the outer node with $k = 2$

We can write

$$\begin{aligned}\log \frac{P(W_2 = 1 \mid Y = y, X = x, W_1 = w_1)}{P(W_2 = 0 \mid Y = y, X = x, W_1 = w_1)} &= \log \frac{P(Y = y \mid X = x, W_1 = w_1, W_2 = 1)}{P(Y = y \mid X = x, W_1 = w_1, W_2 = 0)} \\ &\quad + \log \frac{P(Y = y \mid X = x, W_1 = w_1, W_2 = 1)}{P(Y = y \mid X = x, W_1 = w_1, W_2 = 0)} \\ &\quad + \log \frac{P(W_2 = 1 \mid X = x, W_1 = w_1)}{P(W_2 = 0 \mid X = x, W_1 = w_1)}\end{aligned} \tag{A.9.1}$$

and

$$\begin{aligned}\log \frac{P(Y = 1 \mid X = x, W_1 = w_1)}{P(Y = 0 \mid X = x, W_1 = w_1)} &= \log \frac{P(Y = 1 \mid X = x, W_1 = w_1, W_2 = 0)}{P(Y = 0 \mid X = x, W_1 = w_1, W_2 = 0)} \\ &\quad + \log \frac{P(W_2 = 0 \mid Y = 0, X = x, W_1 = w_1)}{P(W_2 = 0 \mid Y = 1, X = x, W_1 = w_1)}.\end{aligned} \tag{A.9.2}$$

We can write

$$\log \frac{P(W_2 = 1 \mid Y = y, X = x, W_1 = w_1)}{P(W_2 = 0 \mid Y = y, X = x, W_1 = w_1)} = h_{y,w_1}(x)$$

where

$$h_{y,w_1}(x) = \log \frac{P(Y = y \mid X = x, W_1 = w_1, W_2 = 1)}{P(Y = y \mid X = x, W_1 = w_1, W_2 = 0)} + \log \frac{P(W_1 = w_1 \mid X = x, W_2 = 1)}{P(W_1 = w_1 \mid X = x, W_2 = 0)} + \log \frac{P(W_2 = 1 \mid X = x)}{P(W_2 = 0 \mid X = x)} \tag{A.9.3}$$

where all the quantities in (A.9.3) are known from the assumed models in Eqs. (3.3) and (3.4). In particular, following the notation used in Appendix A.1 and A.2, we obtain that

$$\begin{aligned} \log \frac{P(Y = y \mid X = x, W_1 = w_1, W_2 = 1)}{P(Y = y \mid X = x, W_1 = w_1, W_2 = 0)} &= y\{\text{rhs}(Y \mid W_2 = 1, X = x, W_1 = w_1) \\ &\quad - \text{rhs}(Y \mid W_2 = 0, X = x, W_1 = w_1)\} \\ &\quad + \log \frac{1 + \exp\{\text{rhs}(Y \mid W_2 = 0, X = x, W_1 = w_1)\}}{1 + \exp\{\text{rhs}(Y \mid W_2 = 1, X = x, W_1 = w_1)\}}, \end{aligned}$$

$$\begin{aligned} \log \frac{P(W_1 = w_1 \mid X = x, W_2 = 1)}{P(W_1 = w_1 \mid X = x, W_2 = 0)} &= w_1\{\text{rhs}(W_1 \mid W_2 = 1, X = x) - \text{rhs}(W_1 \mid W_2 = 0, X = x)\} \\ &\quad + \log \frac{1 + \exp\{\text{rhs}(W_1 \mid W_2 = 0, X = x)\}}{1 + \exp\{\text{rhs}(W_1 \mid W_2 = 1, X = x)\}} \end{aligned}$$

and

$$\log \frac{P(W_2 = 1 \mid X = x)}{P(W_2 = 0 \mid X = x)} = \text{rhs}(W_2 \mid X = x) = \gamma_{2,0} + \gamma_{2,x} x.$$

Similarly to what done previously, we notice that $\frac{P(W_2=0|Y=0,X=x,W_1=w_1)}{P(W_2=0|Y=1,X=x,W_1=w_1)} = \frac{1+\exp h_{1,w_1}(x)}{1+\exp h_{0,w_1}(x)}$ (i.e. B_{x,w_1} of Eq. (3.11)), that is the inverse of the relative risk of $W_2 = 0$ for varying Y conditional on $X = x$ and $W_1 = w_1$.

A.10 Data availability

All applied and simulated results are generated using R-project (R Core Team 2014), version R-4.0.2. The scripts and the data are available upon request.

A.11 Simulation: other results

Methods		**Exact**				**Approx.**			
***prev.* 20%**	true	mean	var	RRMSE	CP%	mean	var	RRMSE	CP%
$OR_{1,0\|0}^{PDE}$	1.1294	1.1806	0.1383	0.3324	95.8	1.1861	0.1464	0.3425	96.3
$OR_{1,0\|0}^{TIE}$	0.9949	0.9969	0.0049	0.0704	98.6	0.9928	0.0049	0.0704	98.7
$OR_{1,0\|0}^{TDE}$	1.1329	1.1810	0.1317	0.3231	96.2	1.1837	0.1335	0.3256	96.6
$OR_{1,0\|0}^{PIE}$	0.9919	0.9940	0.0039	0.0628	99.1	0.9909	0.0038	0.0620	98.2
$OR_{1,0\|0}^{TE}$	1.1237	1.1706	0.1267	0.3196	96.1	1.1699	0.1290	0.3222	96.4
***prev.* 40%**									
$OR_{1,0\|0}^{PDE}$	1.1296	1.1787	0.1158	0.3044	94.7	1.1840	0.1261	0.3180	95.5
$OR_{1,0\|0}^{TIE}$	0.9950	0.9963	0.0033	0.0580	99.4	0.9907	0.0034	0.0590	98.7
$OR_{1,0\|0}^{TDE}$	1.1330	1.1812	0.1132	0.3000	94.6	1.1832	0.1169	0.3050	94.8
$OR_{1,0\|0}^{PIE}$	0.9919	0.9930	0.0027	0.0528	99.5	0.9885	0.0028	0.0538	98.9
$OR_{1,0\|0}^{TE}$	1.1239	1.1701	0.1082	0.2956	94.4	1.1674	0.1123	0.3006	94.8
***prev.* 60%**									
$OR_{1,0\|0}^{PDE}$	1.1297	1.1832	0.1171	0.3065	94.9	1.2067	0.1618	0.3626	95.8
$OR_{1,0\|0}^{TIE}$	0.9950	0.9946	0.0034	0.0588	99.5	0.9850	0.0043	0.0668	99.7
$OR_{1,0\|0}^{TDE}$	1.1332	1.1789	0.1121	0.2983	95.1	1.1914	0.1275	0.3193	95.4
$OR_{1,0\|0}^{PIE}$	0.9920	0.9964	0.0026	0.0511	99.7	0.9898	0.0028	0.0535	99.5
$OR_{1,0\|0}^{TE}$	1.1241	1.1715	0.1071	0.2941	95.8	1.1771	0.1237	0.3164	96.0

Table A.11.1: Simulation results for the setting with C=0, n=250. *prev*: outcome prevalence; *PDE*: *Pure Direct Effect*; *TIE*: *Total Indirect Effect*; *TDE*: *Total Direct Effect*; *PIE*: *Pure Indirect Effect*; *TE*: *Total Effect*; var: variance; RRMSE: Relative Root Mean Squared Error; CP: Coverage Probability of 95% Confidence Intervals.

Methods		Exact				Approx.			
***prev.* 20%**	true	mean	var	RRMSE	CP%	mean	var	RRMSE	CP%
$OR_{1,0\|1}^{PDE}$	1.1307	1.1802	0.1363	0.3294	95.4	1.1860	0.1434	0.3384	95.9
$OR_{1,0\|1}^{TIE}$	0.9955	0.9982	0.0041	0.0643	98.5	0.9930	0.0041	0.0646	98.5
$OR_{1,0\|1}^{TDE}$	1.1339	1.1831	0.1366	0.3288	95.8	1.1857	0.1373	0.3299	96.3
$OR_{1,0\|1}^{PIE}$	0.9928	0.9952	0.0031	0.0557	99.1	0.9912	0.0030	0.0556	98.5
$OR_{1,0\|1}^{TE}$	1.1257	1.1735	0.1290	0.3219	96.0	1.1718	0.1299	0.3227	96.1
***prev.* 40%**									
$OR_{1,0\|1}^{PDE}$	1.1309	1.1782	0.1115	0.2983	94.9	1.1831	0.1202	0.3101	95.4
$OR_{1,0\|1}^{TIE}$	0.9955	0.9974	0.0028	0.0530	99.7	0.9907	0.0029	0.0545	98.8
$OR_{1,0\|1}^{TDE}$	1.1340	1.1823	0.1139	0.3007	94.8	1.1838	0.1164	0.3040	94.9
$OR_{1,0\|1}^{PIE}$	0.9928	0.9942	0.0022	0.0472	99.7	0.9888	0.0023	0.0489	98.9
$OR_{1,0\|1}^{TE}$	1.1258	1.1722	0.1071	0.2936	94.9	1.1679	0.1100	0.2970	95.0
***prev.* 60%**									
$OR_{1,0\|1}^{PDE}$	1.1311	1.1778	0.1135	0.3008	95.3	1.1979	0.1477	0.3449	95.8
$OR_{1,0\|1}^{TIE}$	0.9956	0.9966	0.0026	0.0512	99.8	0.9859	0.0035	0.0603	99.7
$OR_{1,0\|1}^{TDE}$	1.1341	1.1771	0.1139	0.3001	95.2	1.1871	0.1251	0.3154	95.7
$OR_{1,0\|1}^{PIE}$	0.9929	0.9972	0.0020	0.0450	99.7	0.9899	0.0023	0.0483	99.4
$OR_{1,0\|1}^{TE}$	1.1260	1.1704	0.1076	0.2939	95.7	1.1726	0.1196	0.3099	96.0

Table A.11.2: Simulation results for the setting with C=1, n=250. *prev*: outcome prevalence; *PDE*: *Pure Direct Effect*; *TIE*: *Total Indirect Effect*; *TDE*: *Total Direct Effect*; *PIE*: *Pure Indirect Effect*; *TE*: *Total Effect*; var: variance; RRMSE: Relative Root Mean Squared Error; CP: Coverage Probability of 95% Confidence Intervals.

Methods		Exact				Approx.			
***prev.* 20%**	true	mean	var	RRMSE	CP%	mean	var	RRMSE	CP%
$OR_{1,0\|0}^{PDE}$	1.1294	1.1558	0.0718	0.2385	94.6	1.1573	0.0732	0.2408	94.8
$OR_{1,0\|0}^{TIE}$	0.9949	0.9961	0.0018	0.0425	98.7	0.9942	0.0018	0.0427	97.9
$OR_{1,0\|0}^{TDE}$	1.1329	1.1600	0.0695	0.2340	94.3	1.1607	0.0696	0.2342	94.3
$OR_{1,0\|0}^{PIE}$	0.9919	0.9916	0.0015	0.0393	99.2	0.9901	0.0016	0.0401	98.2
$OR_{1,0\|0}^{TE}$	1.1237	1.1489	0.0672	0.2319	94.6	1.1480	0.0676	0.2323	94.9
***prev.* 40%**									
$OR_{1,0\|0}^{PDE}$	1.1296	1.1434	0.0497	0.1978	94.3	1.1468	0.0513	0.2011	94.3
$OR_{1,0\|0}^{TIE}$	0.9950	0.9952	0.0015	0.0388	98.4	0.9923	0.0015	0.0394	97.7
$OR_{1,0\|0}^{TDE}$	1.1330	1.1443	0.0494	0.1964	94.5	1.1463	0.0495	0.1968	94.7
$OR_{1,0\|0}^{PIE}$	0.9919	0.9940	0.0011	0.0337	99.0	0.9919	0.0011	0.0341	98.2
$OR_{1,0\|0}^{TE}$	1.1239	1.1363	0.0474	0.1939	94.1	1.1360	0.0477	0.1946	94.4
***prev.* 60%**									
$OR_{1,0\|0}^{PDE}$	1.1297	1.1504	0.0546	0.2076	94.6	1.1545	0.0610	0.2198	94.9
$OR_{1,0\|0}^{TIE}$	0.9950	0.9989	0.0015	0.0396	98.6	0.9942	0.0017	0.0412	97.6
$OR_{1,0\|0}^{TDE}$	1.1332	1.1582	0.0540	0.2063	94.1	1.1600	0.0562	0.2105	94.6
$OR_{1,0\|0}^{PIE}$	0.9920	0.9914	0.0012	0.0350	98.8	0.9875	0.0014	0.0380	98.6
$OR_{1,0\|0}^{TE}$	1.1241	1.1470	0.0513	0.2026	94.2	1.1446	0.0541	0.2078	94.8

Table A.11.3: Simulation results for the setting with C=0, n=500. *prev*: outcome prevalence; *PDE*: *Pure Direct Effect*; *TIE*: *Total Indirect Effect*; *TDE*: *Total Direct Effect*; *PIE*: *Pure Indirect Effect*; *TE*: *Total Effect*; var: variance; RRMSE: Relative Root Mean Squared Error; CP: Coverage Probability of 95% Confidence Intervals.

Methods		Exact				Approx.			
***prev.* 20%**	true	mean	var	RRMSE	CP%	mean	var	RRMSE	CP%
$OR_{1,0\|1}^{PDE}$	1.1307	1.1556	0.0691	0.2335	94.6	1.1570	0.0700	0.2352	94.7
$OR_{1,0\|1}^{TIE}$	0.9955	0.9967	0.0014	0.0380	98.6	0.9944	0.0015	0.0384	98.2
$OR_{1,0\|1}^{TDE}$	1.1339	1.1608	0.0698	0.2343	94.9	1.1613	0.0696	0.2340	94.5
$OR_{1,0\|1}^{PIE}$	0.9928	0.9924	0.0012	0.0354	99.2	0.9905	0.0013	0.0366	98.0
$OR_{1,0\|1}^{TE}$	1.1257	1.1503	0.0664	0.2299	94.5	1.1487	0.0664	0.2298	94.5
***prev.* 40%**									
$OR_{1,0\|1}^{PDE}$	1.1309	1.1408	0.0482	0.1944	94.4	1.1441	0.0492	0.1965	94.3
$OR_{1,0\|1}^{TIE}$	0.9955	0.9964	0.0012	0.0349	98.5	0.9930	0.0013	0.0357	97.7
$OR_{1,0\|1}^{TDE}$	1.1340	1.1432	0.0498	0.1969	94.3	1.1449	0.0495	0.1965	94.7
$OR_{1,0\|1}^{PIE}$	0.9928	0.9946	0.0009	0.0300	99.1	0.9921	0.0009	0.0307	98.3
$OR_{1,0\|1}^{TE}$	1.1258	1.1356	0.0470	0.1928	93.8	1.1347	0.0469	0.1926	94.3
***prev.* 60%**									
$OR_{1,0\|1}^{PDE}$	1.1311	1.1537	0.0543	0.2070	94.1	1.1574	0.0598	0.2175	94.6
$OR_{1,0\|1}^{TIE}$	0.9956	0.9995	0.0012	0.0354	98.9	0.9943	0.0014	0.0374	97.8
$OR_{1,0\|1}^{TDE}$	1.1341	1.1614	0.0555	0.2091	94.6	1.1630	0.0572	0.2124	94.8
$OR_{1,0\|1}^{PIE}$	0.9929	0.9928	0.0009	0.0309	99.0	0.9884	0.0011	0.0343	98.6
$OR_{1,0\|1}^{TE}$	1.1260	1.1516	0.0523	0.2044	94.3	1.1484	0.0544	0.2081	94.6

Table A.11.4: Simulation results for the setting with C=1, n=500. *prev*: outcome prevalence; *PDE*: *Pure Direct Effect*; *TIE*: *Total Indirect Effect*; *TDE*: *Total Direct Effect*; *PIE*: *Pure Indirect Effect*; *TE*: *Total Effect*; var: variance; RRMSE: Relative Root Mean Squared Error; CP: Coverage Probability of 95% Confidence Intervals.

A.11. Simulation: other results

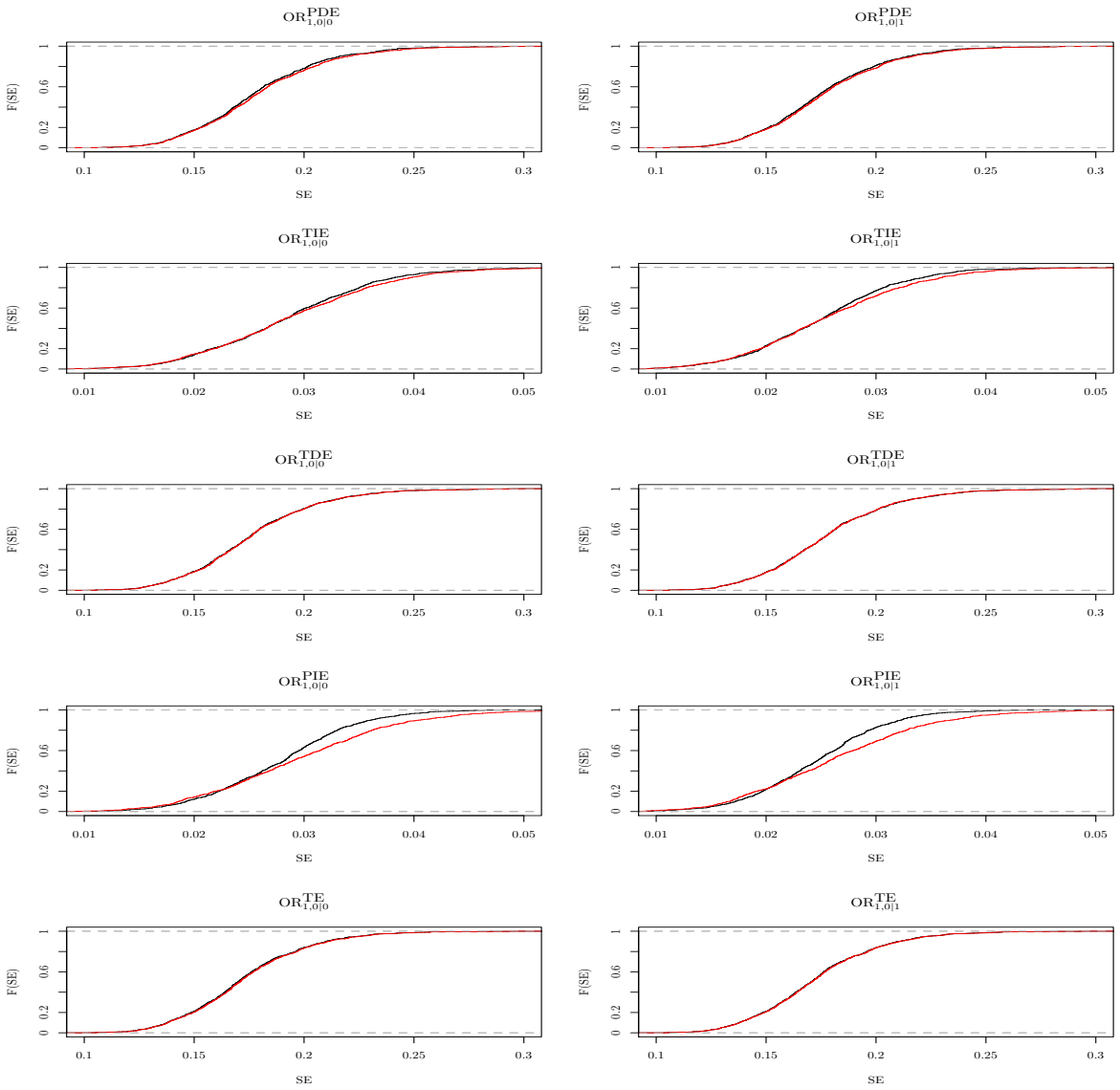


Figure A.11.1: Simulation results for setting with n = 1000 and prevalence 20%: empirical cumulative density function of estimated standard errors (— exact estimators, — approximate estimators).

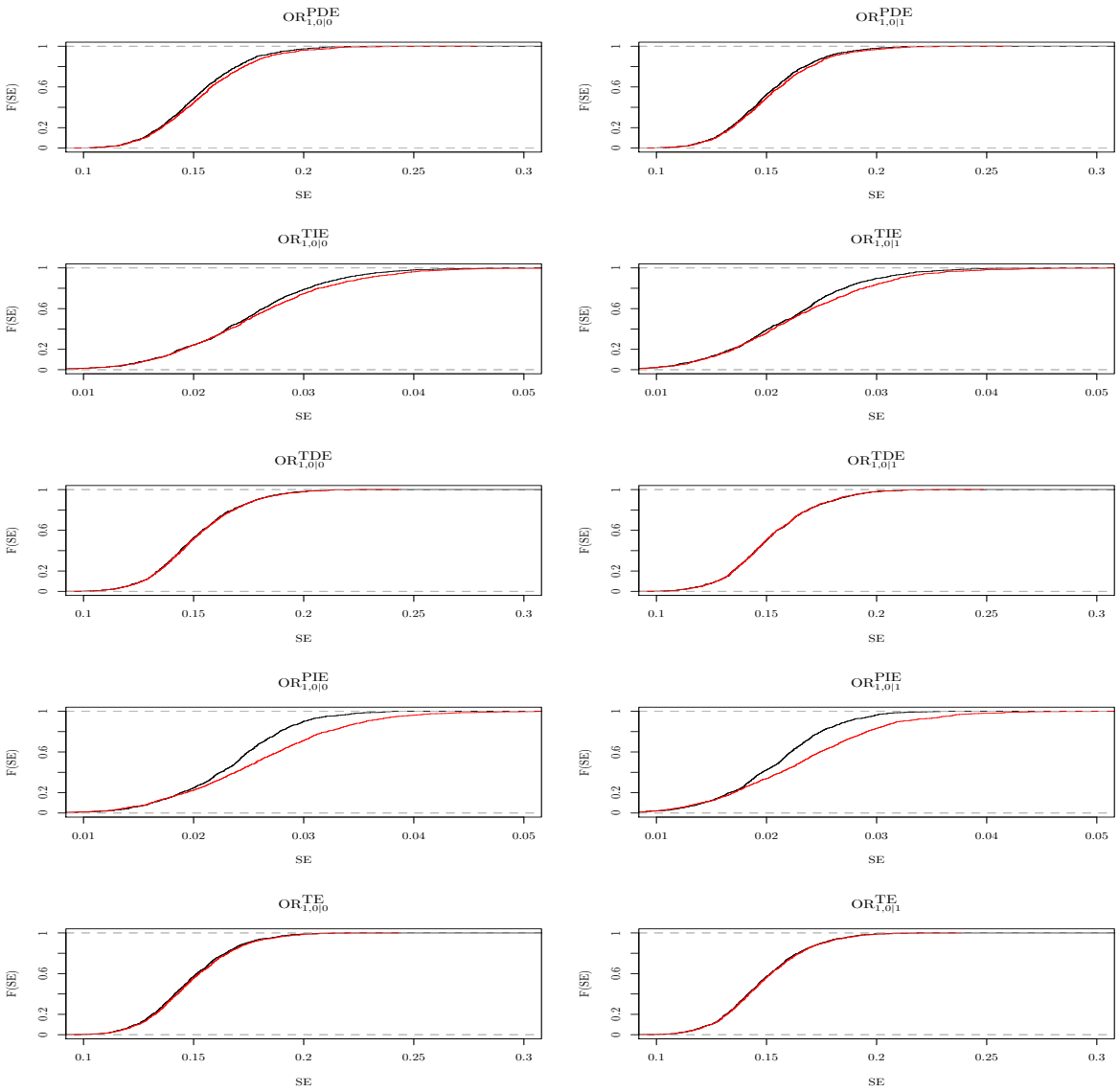


Figure A.11.2: Simulation results for setting with n = 1000 and prevalence 40%: empirical cumulative density function of estimated standard errors (— exact estimators, — approximate estimators).

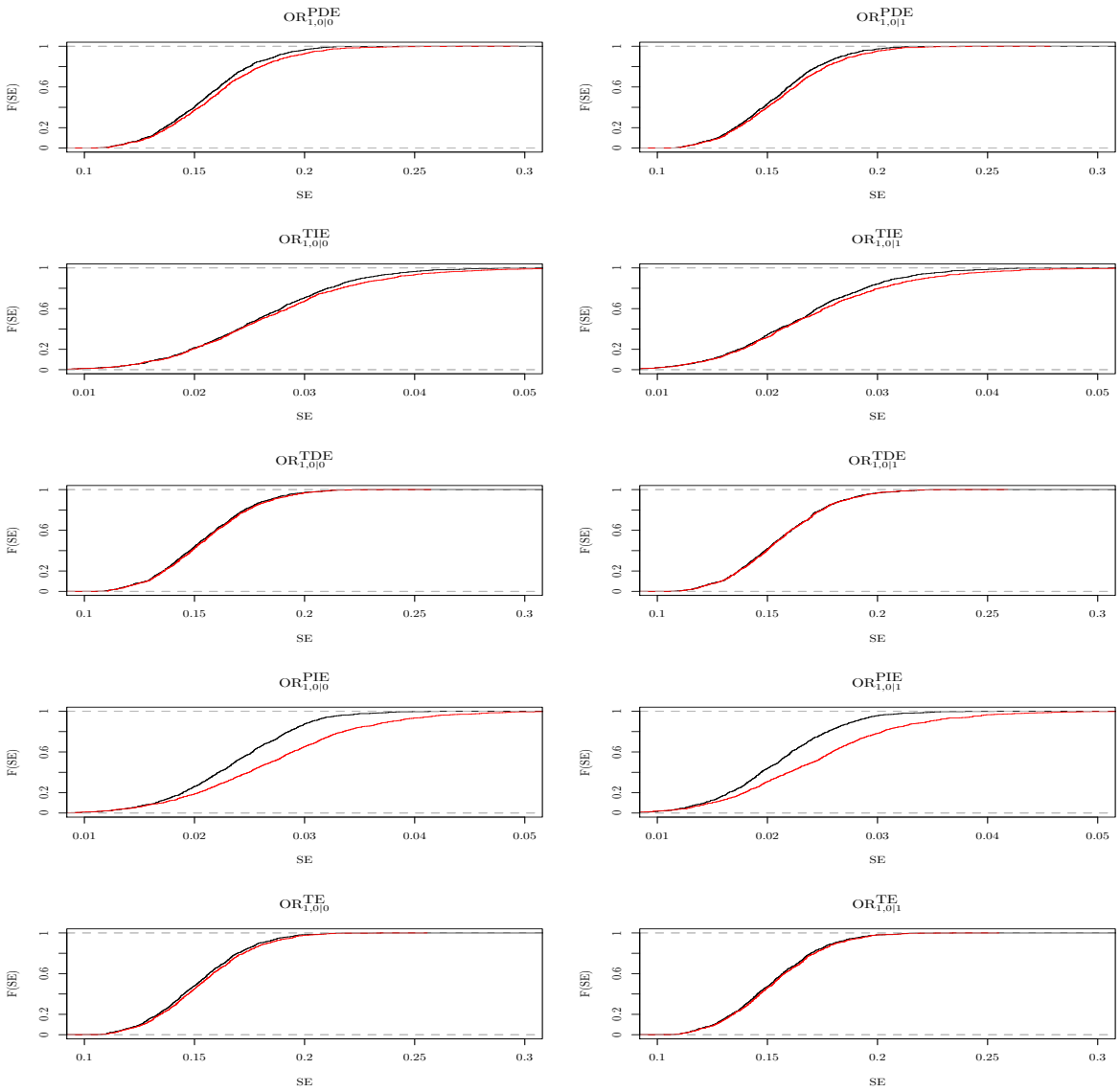


Figure A.11.3: Simulation results for setting with n = 1000 and prevalence 60%: empirical cumulative density function of estimated standard errors (— exact estimators, — approximate estimators).

Acronyms

BAM	Bosnian Marks Currency 44
CDE	Controlled Direct Effect 17
CHD	Coronary Heart Disease 1
CIs	Confidence Intervals 47
DAG	Directed Acyclic Graph 8, 9, 21, 31, 45, 53, 55–58, 63, 65
DE	Direct Effect 22
DE^{nc}	Direct Effect (non-collapsible) 23
ECDF	Empirical Cumulative Density Function 48
GIE	Global Indirect Effect 55
IE	Indirect Effect 22
INT^{med}	Mediated Interaction 16
INT^{ref}	Reference Interaction 18
MFIs	Microfinance Institutions 43–45
OR	Odds Ratio 30

PDE	Pure Direct Effect 15, 16, 33, 47
PE	Proportion Eliminated 26
PIE	Pure Indirect Effect 15, 16, 33, 47
PSIE	Path-specific Indirect Effect 56
RCM	Rubin Causal Model 10
RCT	Randomized Controlled Trials 43
RE	Residual Effect 22
RE^{nc}	Residual Effect (non-collapsible) 23
rhs	right hand side 53
RR	Relative Risk 36
RRMSE	Root Mean Squared Error 48
SEM	Structural Equations Model 2, 8
SEs	Standard Errors 47
TDE	Total Direct Effect 15, 16, 33, 47
TE	Total Effect 11, 22
TIE	Total Indirect Effect 15, 16, 33, 47

Bibliography

Albert, J. M. and S. Nelson (2011): “Generalized causal mediation analysis”. In: *Biometrics* 67.3, pp. 1028–1038.

Alwin, D. F. and R. M. Hauser (1975): “The Decomposition of Effects in Path Analysis”. In: *American Sociological Review* 40.1, pp. 37–47.

Augsburg, B., R. De Haas, H. Harmgart, and C. Meghir (2015): “The impacts of microcredit: Evidence from Bosnia and Herzegovina”. In: *American Economic Journal: Applied Economics* 7.1, pp. 183–203.

Avin, C., I. Shpitser, and J. Pearl (2005): “Identifiability of path-specific effects”. In: *Proceedings of the 19th international joint conference on Artificial intelligence*, pp. 357–363.

Banerjee, A., D. Karlan, and J. Zinman (2015): “Six randomized evaluations of microcredit: Introduction and further steps”. In: *American Economic Journal: Applied Economics* 7.1, pp. 1–21.

Baron, R. M. and D. A. Kenny (1986): “The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations”. In: *Journal of personality and social psychology* 51.6, p. 1173.

Bellavia, A. and L. Valeri (2018): “Decomposition of the total effect in the presence of multiple mediators and interactions”. In: *American journal of epidemiology* 187.6, pp. 1311–1318.

Bellavia, A., P. L. Williams, L. A. DiMeglio, R. Hazra, M. J. Abzug, K. Patel, D. L. Jacobson, R. B. Van Dyke, and M. E. Geffner (2017): “Delay in sexual maturation in perinatally HIV-infected youth is mediated by poor growth”. In: *Aids (London, England)* 31.9, p. 1333.

Berzuini, C., A. P. Dawid, L. Bernardinelli, et al. (2012): “Causality: Statistical Perspectives and Applications”. In: *Wiley Series in Probability and Statistics*.

Bollen, K. A. (1987): “Total, direct, and indirect effects in structural equation models”. In: *Sociological methodology*, pp. 37–69.

Breen, R., K. B. Karlson, and A. Holm (2013): “Total, direct, and indirect effects in logit and probit models”. In: *Sociological Methods & Research* 42.2, pp. 164–191.

— (2018): “A Note on a Reformulation of the KHB Method”. In: *Sociological Methods & Research*.

Cochran, W. G. (1938): “The omission or addition of an independent variate in multiple linear regression”. In: *Supplement to the Journal of the Royal Statistical Society* 5.2, pp. 171–176.

Cochran, W. G. and S. P. Chambers (1965): “The planning of observational studies of human populations”. In: *Journal of the Royal Statistical Society. Series A (General)* 128.2, pp. 234–266.

Cole, S. R. and C. E. Frangakis (2009): “The consistency statement in causal inference: a definition or an assumption?” In: *Epidemiology* 20.1, pp. 3–5.

Cole, S. R. and M. A. Hernán (2002): “Fallibility in estimating direct effects”. In: *International journal of epidemiology* 31.1, pp. 163–165.

Cox, D. R. (1992): “Causality: some statistical aspects”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 155.2, pp. 291–301.

— (2007): “On a generalization of a result of W.G. Cochran”. In: *Biometrika* 94.3, pp. 755–759.

Cox, D. R. and N. Wermuth (2003): “A general condition for avoiding effect reversal after marginalization”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.4, pp. 937–941.

— (2004): “Causality: A statistical view”. In: *International Statistical Review* 72.3, pp. 285–305.

Daniel, R. M., B. L. De Stavola, S. N. Cousens, and S. Vansteelandt (2015): “Causal mediation analysis with multiple mediators”. In: *Biometrics* 71.1, pp. 1–14.

Dawid, A. P. (1979): “Conditional independence in statistical theory (with discussion)”. In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 41.1, pp. 1–31.

— (2000): “Causal inference without counterfactuals”. In: *Journal of the American statistical Association* 95.450, pp. 407–424.

Doretti, M., M. Raggi, and E. Stanghellini (2020): “Exact parametric causal mediation analysis for a binary outcome with a binary mediator”. In: *arXiv preprint arXiv:1811.00439v3*.

Elwert, F. (2013): “Graphical causal models”. In: *Handbook of causal analysis for social research*. Springer, pp. 245–273.

Elwert, F. and C. Winship (2014): “Endogenous selection bias: The problem of conditioning on a collider variable”. In: *Annual review of sociology* 40, pp. 31–53.

Feingold, A., D. P. MacKinnon, and D. M. Capaldi (2019): “Mediation analysis with binary outcomes: Direct and indirect effects of pro-alcohol influences on alcohol use disorders”. In: *Addictive behaviors* 94, pp. 26–35.

Fisher, R. A. (1935): *The design of experiments.* Oliver and Boyd, Edinburgh.

Gaynor, S. M., J. Schwartz, and X. Lin (2018): “Mediation analysis for common binary outcomes”. In: *Statistics in Medicine*.

Glynn, A. N. (2012): “The product and difference fallacies for indirect effects”. In: *American Journal of Political Science* 56.1, pp. 257–269.

Hafeman, D. M. and S. Schwartz (2009): “Opening the Black Box: a motivation for the assessment of mediation”. In: *International journal of epidemiology* 38.3, pp. 838–845.

Hayes, A. F. (2017): *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Huber, M. (2019): *A review of causal mediation analysis for assessing direct and indirect treatment effects*. Tech. rep. Université de Fribourg.

Huber, M., M. Lechner, and G. Mellace (2017): “Why Do Tougher Caseworkers Increase Employment? The Role of Program Assignment as a Causal Mechanism”. In: *The Review of Economics and Statistics* 99.1, pp. 180–183.

Huber, M., M. Lechner, and A. Strittmatter (2018): “Direct and indirect effects of training vouchers for the unemployed”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.2, pp. 441–463.

Imai, K., L. Keele, and D. Tingley (2010a): “A General Approach to Causal Mediation Analysis”. In: *Psychological Methods* 15.4, pp. 309–334.

Imai, K., L. Keele, and T. Yamamoto (2010b): “Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects”. In: *Statistical Science* 25.1, pp. 51–71.

Imbens, G. W. and D. B. Rubin (2015): *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Judd, C. M. and D. A. Kenny (1981): “Process analysis: Estimating mediation in treatment evaluations”. In: *Evaluation review* 5.5, pp. 602–619.

Karlson, K. B., A. Holm, and R. Breen (2012): “Comparing regression coefficients between same-sample nested models using logit and probit: A new method”. In: *Sociological methodology* 42.1, pp. 286–313.

Lauritzen, S. L. (1996): *Graphical models*. Vol. 17. Clarendon Press.

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

Lindmark, A., X. de Luna, and M. Eriksson (2018): “Sensitivity analysis for unobserved confounding of direct and indirect effects using uncertainty intervals”. In: *Statistics in Medicine* 37.10, pp. 1744–1762.

Lupparelli, M. (2018): “Conditional and marginal relative risk parameters for a class of recursive regression graph models”. In: *Statistical Methods in Medical Research*.

Ma, Z., X. Xie, and Z. Geng (2006): “Collapsibility of distribution dependence”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1, pp. 127–133.

Maathuis, M., M. Drton, S. Lauritzen, and M. Wainwright (2018): *Handbook of graphical models*. CRC Press.

MacKinnon, D. P. (2000): “Contrasts in multiple mediator models”. In: *Multivariate applications in substance use research: New methods for new questions*, pp. 141–160.

— (2008): *Introduction to statistical mediation analysis*. Routledge.

MacKinnon, D. P. and J. H. Dwyer (1993): “Estimating mediated effects in prevention studies”. In: *Evaluation review* 17.2, pp. 144–158.

MacKinnon, D. P., C. A. Johnson, M. A. Pentz, J. H. Dwyer, W. B. Hansen, B. R. Flay, and E. Y. Wang (1991): “Mediating mechanisms in a school-based drug prevention program: first-year effects of the Midwestern Prevention Project.” In: *Health Psychology* 10.3, p. 164.

MacKinnon, D. P., C. M. Lockwood, C. H. Brown, W. Wang, and J. M. Hoffman (2007): “The intermediate endpoint effect in logistic and probit regression”. In: *Clinical Trials* 4.5, pp. 499–513.

MacKinnon, D. P., M. J. Valente, and O. Gonzalez (2020): “The correspondence between causal and traditional mediation analysis: the link is the mediator by treatment interaction”. In: *Prevention Science* 21.2, pp. 147–157.

Morgan, S. L. and C. Winship (2015): *Counterfactuals and causal inference*. Cambridge University Press.

Mortensen, L. H., F. Diderichsen, G. D. Smith, and A. M. N. Andersen (2009): “The social gradient in birthweight at term: quantification of the mediating role of maternal smoking and body mass index”. In: *Human Reproduction* 24.10, pp. 2629–2635.

Neuhaus, J. M. and N. P. Jewell (1993): “A geometric approach to assess bias due to omitted covariates in generalized linear models”. In: *Biometrika* 80.4, pp. 807–815.

Neyman, J. (1923): “Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes”. In: *Roczniki Nauk Rolniczych* 10, pp. 1–51.

Nguyen, Q. C., T. L. Osypuk, N. M. Schmidt, M. M. Glymour, and E. J. T. Tchetgen (2015): “Practical guidance for conducting mediation analysis with multiple mediators using inverse odds ratio weighting”. In: *American journal of epidemiology* 181.5, pp. 349–356.

Oehlert, G. W. (1992): “A note on the delta method”. In: *The American Statistician* 46, pp. 27–29.

Pearl, J. (1995): “Causal diagrams for empirical research”. In: *Biometrika* 82.4, pp. 669–688.

— (2000): “Causal inference without counterfactuals: Comment”. In: *Journal of the American Statistical Association* 95.450, pp. 428–431.

— (2001): “Direct and indirect effects”. In: *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 411–420.

— (2009a): “Causal inference in statistics: An overview”. In: *Statistics surveys* 3, pp. 96–146.

— (2009b): *Causality*. Cambridge University Press.

— (2010): “An Introduction to Causal Inference”. In: *The International Journal of Biostatistics* 6.2, pp. 1–62.

— (2012): *The mediation formula: A guide to the assessment of causal pathways in nonlinear models*. Wiley Online Library.

— (2014): "Interpretation and identification of causal mediation." In: *Psychological methods* 19.4, p. 459.

Pearl, J., M. Glymour, and N. P. Jewell (2016): *Causal inference in statistics: A primer*. John Wiley & Sons.

Preacher, K. J. and A. F Hayes (2008): "Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models". In: *Behavior research methods* 40.3, pp. 879–891.

R Core Team (2014): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Raggi, M., E. Stanghellini, and M. Doretti (2020): "Path Analysis for Binary Random Variables".

Richiardi, L., R. Bellocco, and D. Zugna (2013): "Mediation analysis in epidemiology: methods, interpretation and bias". In: *International journal of epidemiology* 42.5, pp. 1511–1519.

Robins, J. M. (2003): "Semantics of causal DAG models and the identification of direct and indirect effects". In: *Oxford Statistical Science Series*, pp. 70–82.

Robins, J. M. and S. Greenland (1992): "Identifiability and exchangeability for direct and indirect effects". In: *Epidemiology*, pp. 143–155.

— (2000): "Causal inference without counterfactuals: comment". In: *Journal of the American Statistical Association* 95.450, pp. 431–435.

Rosenbaum, P. R. and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects". In: *Biometrika*, pp. 41–55.

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of Educational Psychology* 66.5, pp. 688–701.

— (1980): "Randomization analysis of experimental data: The Fisher randomization test comment". In: *Journal of the American Statistical Association* 75.371, pp. 591–593.

— (2000): "Causal inference without counterfactuals: comment". In: *Journal of the American Statistical Association* 95.450, pp. 435–438.

Saks, A. M., J. Zikic, and J. Koen (2015): "Job search self-efficacy: Reconceptualizing the construct and its measurement". In: *Journal of Vocational Behavior* 86, pp. 104–114.

Samoilenko, M., L. Blais, and G. Lefebvre (2018): "Comparing logistic and log-binomial models for causal mediation analyses of binary mediators and rare binary outcomes: evidence to support cross-checking of mediation results in practice". In: *Observational Studies* 4, pp. 193–216.

Samoilenko, M. and G. Lefebvre (2018): "Natural Direct and Indirect Effects' Risk Ratio Expressions in Causal Mediation Analysis of Binary Mediator and Binary Outcome: A Fresh Look at the Formulas". In: *American Journal of Epidemiology*.

Sobel, M. E. (1982): “Asymptotic confidence intervals for indirect effects in structural equation models”. In: *Sociological methodology* 13, pp. 290–312.

Stanghellini, E. and M. Doretti (2019): “On marginal and conditional parameters in logistic regression models”. In: *Biometrika* 106.3, pp. 732–739.

Steen, J., T. Loeys, B. Moerkerke, and S. Vansteelandt (2017a): “Flexible mediation analysis with multiple mediators”. In: *American journal of epidemiology* 186.2, pp. 184–193.

— (2017b): “medflex : an R package for flexible mediation analysis using natural effect models”. In: *Journal of Statistical Software* 76.11.

Steen, J. and S. Vansteelandt (2018): “Graphical models for mediation analysis”. In: *arXiv preprint arXiv:1801.06069*.

Tchetgen, E. J. T. (2014): “A note on formulae for causal mediation analysis in an odds ratio context”. In: *Epidemiologic methods* 2.1, pp. 21–31.

Tchetgen, E. J. T. and I. Shpitser (2012): “Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis”. In: *Annals of statistics* 40.3, p. 1816.

Valeri, L. (2017): “Causal Mediation Analysis in Pregnancy Studies: the Case of Environmental Epigenetics”. In: *Current Epidemiology Reports* 4.2, pp. 117–123.

Valeri, L. and T. J. VanderWeele (2013): “Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros.” In: *Psychological methods* 18.2, pp. 137–150.

VanderWeele, T. J. (2009): “Concerning the consistency assumption in causal inference”. In: *Epidemiology* 20.6, pp. 880–883.

— (2013a): “A three-way decomposition of a total effect into direct, indirect, and interactive effects”. In: *Epidemiology (Cambridge, Mass.)* 24.2, p. 224.

— (2013b): “Policy-relevant proportions for direct effects”. In: *Epidemiology (Cambridge, Mass.)* 24.1, p. 175.

— (2014): “A unification of mediation and interaction: a four-way decomposition”. In: *Epidemiology (Cambridge, Mass.)* 25.5, p. 749.

— (2015): *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. J., L. Valeri, and C. V. Ananth (2018): “Mediation Formulas with Binary Mediators and Outcomes and the “Rare Outcome Assumption””. In: *American Journal of Epidemiology*.

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

— (2010): “Odds ratios for mediation analysis for a dichotomous outcome”. In: *American journal of epidemiology* 172.12, pp. 1339–1348.

— (2014): “Mediation analysis with multiple mediators”. In: *Epidemiologic methods* 2.1, pp. 95–115.

Vansteelandt, S., M. Bekaert, and T. Lange (2012): “Imputation strategies for the estimation of natural direct and indirect effects”. In: *Epidemiologic Methods* 1.1.

Vinokur, A. D., R. H. Price, and Y. Schul (1995): “Impact of the JOBS intervention on unemployed workers varying in risk for depression”. In: *American journal of community psychology* 23.1, pp. 39–74.

Vinokur, A. D. and Y. Schul (1997): “Mastery and inoculation against setbacks as active ingredients in the JOBS intervention for the unemployed.” In: *Journal of consulting and clinical psychology* 65.5, p. 867.

Whittaker, J. (2009): *Graphical models in applied multivariate statistics*. Wiley Publishing.

Winship, C. and R. D. Mare (1983): “Structural equations and path analysis for discrete data”. In: *American Journal of Sociology* 89.1, pp. 54–110.

Wright, S. (1921): “Correlation and causation”. In: *Journal of agricultural research* 20.7, pp. 557–585.

— (1934): “The method of path coefficients”. In: *The annals of mathematical statistics* 5.3, pp. 161–215.

Xie, X., Z. Ma, and Z. Geng (2008): “Some association measures and their collapsibility”. In: *Statistica Sinica*, pp. 1165–1183.