

Department of Medicine and Surgery

PhD program in Public Health

Cycle XXXVII

Curriculum in Biostatistics and Epidemiology

CAUSAL MEDIATION ANALYSIS WITH HIDDEN MEDIATOR

Candidate: Laura Montelisciani

Registration number: 891538

Tutor: Laura Antolini

Co-tutor: Eric J. Tchetgen Tchetgen

Coordinator: Luigi Badano

Academic Year: 2024/2025

Contents

1	Introduction	4
2	Mediation analysis with observable mediator in the absence of measurement error	6
2.1	Y continuous (Identity link)	8
2.2	Y count, binary or polytomous (Log link)	11
3	Mediation analysis with hidden mediator or in presence of measurement error	13
3.1	Y continuous (Identity link)	14
3.1.1	Identification through regression coefficients	14
3.1.2	Identification through cross-world counterfactual means	17
3.2	Y count, binary or polytomous (Log link)	20
4	Simulations	23
4.1	Y continuous (Identity link)	24
4.2	Y count, binary or polytomous (Log link)	28
5	Measurement error arising from the use of a proxy in place of the mediator	31
5.1	Measurement error in linear models and attenuation	32
5.2	Bias quantification in NDE ^a in linear model without exposure-mediator interaction . . .	34
6	Application: Brader, Valentino, and Suhay's Immigration Experiment	39
7	Discussion	42
	Appendix	47
A1	Derivation of NDEs and NIEs with observable mediator	47
A1.1	NDE and NIE with continuous outcome Y (Identity link) with observable mediator without exposure-mediator interaction	47
A1.2	NDE and NIE with continuous outcome Y (Identity link) with observable mediator with exposure-mediator interaction	49
A1.3	NDE and NIE with count, binary or polytomous outcome Y (Log link) with observable mediator without exposure-mediator interaction	51
A1.4	NDE and NIE with count, binary or polytomous outcome Y (Log link) with observable mediator with exposure-mediator interaction	53

A2 Derivation of NDEs and NIEs with hidden mediator: proximal framework	55
A2.1 NDE and NIE with continuous outcome Y (Identity link) with hidden mediator without exposure-mediator interaction	55
A2.2 NDE and NIE with continuous outcome Y (Identity link) with hidden mediator with exposure-mediator interaction	56
A2.3 NDE and NIE with continuous outcome Y (Identity link) with hidden mediator with and without exposure-mediator interaction (Identification of cross-world counterfactual means)	58
A2.4 NDE and NIE with count, binary or polytomous outcome Y (Log link) with hidden mediator without exposure-mediator interaction	60
A2.5 NDE and NIE with count, binary or polytomous outcome Y (Log link) with hidden mediator with exposure-mediator interaction	62
A3 The empirical sandwich variance estimator	64
A4 Measurement error	65
A4.1 Classical measurement error model with W (mismeasured variable) unbiased	66
A4.2 Classical measurement error model with W (mismeasured variable) biased	67
A4.3 Classical multiple measurement error model with W (mismeasured variable) unbiased	69
A4.4 Classical multiple measurement error model with W (mismeasured variable) biased	71
A4.5 Bias quantification in NDE ^a in linear model without exposure-mediator interaction	74

1 Introduction

Mediation analysis investigates the role of the mediator in the relationship between an exposure (equivalently, treatment) and an outcome. Rather than solely examining the total effect between the exposure and the outcome, it aims to determine whether part of the effect passes through a mediator; in other words, whether the exposure variable causes the mediator variable, which in turn causes the outcome variable.

The classical approach to mediation analysis, introduced by Baron and Kenny (1986), focuses on linear models with three continuous variables: the exposure A , the mediator M , and the outcome Y . The direct effect from the exposure to the mediator is referred to as path a , from the mediator to the outcome as path b , and from the exposure to the outcome as path c . According to this approach, the direct effect is represented by path c while the indirect effect is given by the product of paths $a * b$. The total effect, denoted by c' , is given by the sum of the direct effect, path c , and the indirect effect, path $a * b$. The indirect effect can also be calculated as the difference between the total effect c' and the direct effect c .

In the classical approach the estimate of the direct effect c is the coefficient of the exposure A in the linear regression model of the outcome Y on both the exposure A and the mediator M . The estimate of the indirect effect $a * b$ is the product of two regression coefficients: the coefficient of the exposure A in the linear regression model of the mediator M on the exposure A , and the coefficient of the mediator M in the linear regression model of the outcome Y on both the exposure A and the mediator M .

This approach is subject to two important limitations: it excludes exposure-mediator interaction and non-linearity. More recent literature (Robins and Greenland, 1992; Pearl, 2001) clarifies the assumptions required for mediation analysis in the counterfactual framework and provides definitions of direct and indirect effects based on the counterfactual means that could be used even in the presence of interaction and nonlinearities.

Identifying causal effects is challenging in observational studies. In mediation analysis, one of the major hurdles is the assumption that the mediator is observable and measured without error. However, these assumptions result particularly strong in observational settings.

For the mismeasured mediator, Le Cessie et al. (2012) proposed a correction for causal effects for different types of measurement error in the mediator when the outcome is modeled using either linear or logistic regression, but in the absence of interaction between exposure and mediator. Valeri et al. (2014) extending this by proposing a correction for measurement error even in the presence of inter-

action in various statistical models; however, specific assumptions regarding the error are necessary, some of these impossible to verify.

In this thesis, we extended the proximal causal inference framework (Miao et al., 2018; Tchetgen Tchetgen et al., 2020) to address both issues: the case where the mediator is unobservable and the case where it has been measured with measurement error, while providing a unique method that does not require assumptions about the type of measurement error. This approach is initially designed to enable the identification of causal effects in the presence of unobservable confounders, contingent upon having data on two proxy variables for the confounder that satisfy certain conditional independence conditions. Specifically, we extended the proximal causal inference framework to the field of causal mediation analysis for cases where the mediator is either unobservable or cannot be measured accurately. This allows for the identification of natural direct and indirect effects of the exposure on the outcome, even with a hidden mediator. Similar to the case of an unobservable confounder, this approach requires data on two proxy variables for the hidden mediator.

This thesis introduces an innovative approach for identifying the causal effects with a hidden or mis-measured mediator, using generalized linear models (GLM). The primary contribution lies in offering a straightforward and readily applicable method that facilitates the identification and the estimation of causal effects. In the case of measurement error in the mediator, its versatility is noteworthy, as it does not require specific assumptions about the nature of the measurement error. With access to two mediator proxies, the approach ensures unbiased estimates of causal effects, regardless of the type of measurement error present.

We developed a methodology for cases where the treatment is dichotomous, the mediator is continuous, and the outcome variable can be continuous (with a linear link), count, binary, or polytomous (with a logarithmic link). Additionally, we explore scenarios involving interaction between exposure and mediator.

Furthermore, we quantify the bias that would arise when using one of the two proxies instead of the unmeasured mediator in the analysis.

The thesis is structured as follows: Section 2 introduces the notation, classic assumptions, and reviews the direct and indirect causal effects in mediation analysis when the mediator is observed without measurement error, for the linear and the logarithmic cases. Section 3 presents the GLM-based proximal approach for identifying causal effects in the presence of mediator measurement error or a hidden mediator. In Section 4, we show the performances of the method through a simulation study. In Section 5, we quantify the bias that results from using one proxy in place of the mediator. Section 6

applies the proposed method to real data, followed by a discussion in Section 7.

2 Mediation analysis with observable mediator in the absence of measurement error

A denotes the dichotomous treatment variable, where $A = a = 1$ signifies treatment received (treated group), and $A = a' = 0$ otherwise (untreated group). Y denotes the outcome of interest, which can be continuous, count, binary, or polytomous. M denotes the continuous mediator and X denotes one or more possible observable confounders.

To express the possible causal effects among the introduced variables, the graphical tool of DAGs (Directed Acyclic Graphs) is utilized. The absence of arrows in the DAG is used to indicate the absence of a direct causal effect from one variable to another.

The pertinent scenarios are illustrated in Figure 1.

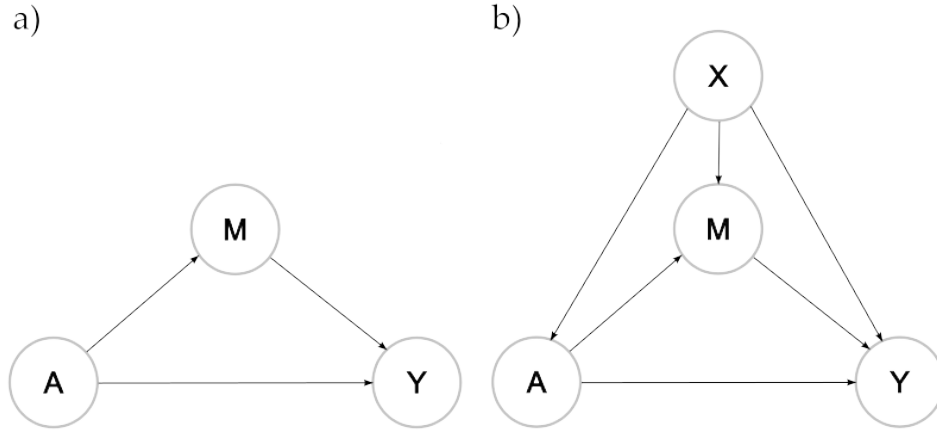


Figure 1: Mediation analysis with observable mediator. The graphs represent the case with an observable mediator without measurement error, without a) and with b) observable confounder respectively.

The exposure has an effect on the outcome both through the mediator, which in turn is causally related to the outcome, and by directly influencing the level of the outcome.

$Y(a, M(a))$ denotes the potential outcome (equivalently, counterfactual outcome) when the treatment A is set to the value $A = a$ and the mediator assumed the value $M(a)$ it has under treatment $A = a$. Obviously, not all the considered counterfactuals have corresponding observed values; for instance, the value of the counterfactual variable $M(a)$ differs from the value of the observed variable M if the patient did not receive the treatment $A = a$.

We posit the following hypotheses, which are standard in causal mediation analysis and essential for a causal interpretation of the effects.

Assumption 1: Positivity (Rosenbaum and Rubin 1983)

$p(A | X) > 0$ and $p(M | A, X) > 0$ for all possible values of A , M and X ;

This assumption ensures that, within each level of the confounder covariate X , no units have an impossible probability of receiving both treatment versions. Similarly, all units have a nonzero probability of attaining any possible value of the mediator, given the confounder level and under each treatment condition.

Assumption 2: Consistency (Robins 1986)

$M(a) = M$ if $A = a$; $Y(a, M(a)) = Y$ if $A = a$ and $M = m$;

This assumption states that if a patient has actually received a certain version of the treatment $A = a$ the counterfactual mediator $M(a)$ is equal to the observed mediator M . Furthermore, the counterfactual outcome $Y(a, M(a))$ is equal to the observed outcome Y only if the patient receives the treatment $A = a$ and has the value $M = M(a) = m$.

Furthermore, Imai, Keele and Yamamoto (2010) introduced an additional assumption specifically for the causal mediation analysis.

Assumption 3: Sequential Ignorability:

1. $Y(a, M(a)) \perp\!\!\!\perp A | X$;
2. $M(a) \perp\!\!\!\perp A | X$;
3. $Y(a, M(a)) \perp\!\!\!\perp M(a') | X$ (Cross-world independence assumption)

With the first two points of assumption 3, the aim is to express the independence of the potential outcome $Y(a, M(a))$ and the potential mediator $M(a)$ from the treatment assignment A , conditioning on the observable pre-treatment covariates. In other words, this part excludes the presence of unobservable confounders in the relationship between potential outcome or mediator and the treatment assignment.

The third part of Assumption 3 (the cross-world independence assumption) expresses the independence between two counterfactuals under different versions of the treatment. Specifically, it assumes independence between the potential outcome when the treatment is set to $A = a$, i.e., $Y(a, M(a))$, and the potential mediator when the treatment is set to $A = a'$, i.e., $M(a')$. Thus, it does not merely posit conditional independence between the potential outcome and the potential mediator, but rather assumes this independence across two distinct treatment regimes. In other words, we assume the absence of a mediator-outcome confounder affected by exposure. The presence of this "intermediate confounder" would prevent the identification of causal effects.

Now, we introduced the two main estimands in the causal mediation framework: the Natural Indirect Effect (NIEs) and the Natural Direct Effect (NDEs).

The NIEs (Pearl, 2001) expresses the change that would occur in the outcome if the mediator value changes, with a fixed treatment. It represents the indirect effect of the treatment on the outcome through the mediator, when the treatment is constant.

The NDEs (Pearl 2001) expresses the change that would occur to the outcome if the treatment value changes, with a fixed mediator. It represents the direct effect of the treatment on the outcome when the mediator is constant.

2.1 Y continuous (Identity link)

Assuming no confounder X , for simplicity, we can define two types of NIEs: one evaluated the indirect effect in the treated group, when $A = a = 1$, and the other in the control group, when $A = a' = 0$.

$$NIE^a = E[Y(a, M(a))] - E[Y(a, M(a'))] \quad (1)$$

$$NIE^{a'} = E[Y(a', M(a))] - E[Y(a', M(a'))] \quad (2)$$

Imai (2010) called this effects Average Causal Mediation Effect (ACME), while Robins (2003) used total indirect effect for NIE^a .

Even the NDEs effect can be considered in the two groups:

$$NDE^a = E[Y(a, M(a))] - E[Y(a', M(a))] \quad (3)$$

$$NDE^{a'} = E[Y(a, M(a'))] - E[Y(a', M(a'))] \quad (4)$$

Imai (2010) called this effects Average Direct Effect (ADE), while Robins (2003) used pure direct effect for $NDE^{a'}$.

The Average Causal Effect (ATE) (equivalently, total effect) of the treatment in terms of potential outcomes can be defined as:

$$ATE = E[Y(a) - Y(a')] = E[Y(a, M(a))] - E[Y(a', M(a'))] \quad (5)$$

When treatment $A = a$ and the mediator M takes the value it would have under $A = a$, the counterfactual outcome $Y(a, M(a))$ simplifies to $Y(a)$. This implies that the counterfactual outcome can be expressed without explicitly referencing the mediator.

The ATE can be partitioned in the two different sums of two versions of NIEs and NDEs (Pearl, 2001) as follow:

$$\text{ATE} = \text{E}[Y(a, M(a))] - \text{E}[Y(a, M(a'))] + \text{E}[Y(a, M(a'))] - \text{E}[Y(a', M(a'))] = \text{NIE}^a + \text{NDE}^{a'} \quad (6)$$

$$\text{ATE} = \text{E}[Y(a, M(a))] - \text{E}[Y(a', M(a))] + \text{E}[Y(a', M(a))] - \text{E}[Y(a', M(a'))] = \text{NDE}^a + \text{NIE}^{a'} \quad (7)$$

Under the assumptions 1-3 it is possible to identify the four effects as follow (Imai, Keele, and Yamamoto (2010)):

$$\text{NDE}^a = \int_m [\text{E}(Y | A = a, M) - \text{E}(Y | A = a', M)] p(M | A = a) \partial m \quad (8)$$

$$\text{NDE}^{a'} = \int_m [\text{E}(Y | A = a, M) - \text{E}(Y | A = a', M)] p(M | A = a') \partial m \quad (9)$$

$$\text{NIE}^a = \int_m \text{E}(Y | A = a, M) [p(M | A = a) - p(M | A = a')] \partial m \quad (10)$$

$$\text{NIE}^{a'} = \int_m \text{E}(Y | A = a', M) [p(M | A = a) - p(M | A = a')] \partial m \quad (11)$$

Assuming no treatment-mediator interaction, we have the following data generation equations:

$$\text{E}[Y | A, M] = \theta_0 + \theta_a A + \theta_m M$$

$$\text{E}[M | A] = \beta_0 + \beta_a A$$

By using the formulas from 8 to 11, it is possible to identify the four causal effects just introduced, expressed in terms of regression coefficients (Appendix A1.1).

$$\text{NDE}^a = \theta_a \quad (12)$$

$$\text{NDE}^{a'} = \theta_a \quad (13)$$

$$\text{NIE}^a = \theta_m \beta_a \quad (14)$$

$$\text{NIE}^{a'} = \theta_m \beta_a \quad (15)$$

As evidenced, in scenarios devoid of interaction between exposure and mediator, the identification of direct and indirect effects across both groups aligns.

Therefore, the ATE can be rewritten using the four causal effects obtained:

$$\text{ATE} = \text{NIE}^a + \text{NDE}^{a'} = \theta_m \beta_a + \theta_a$$

$$\text{ATE} = \text{NDE}^a + \text{NIE}^{a'} = \theta_a + \theta_m \beta_a$$

Considering instead the scenario with interaction between the exposure and the mediator, the data generation equations for Y become:

$$E[Y \mid A, M] = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM$$

Even in the presence of exposure-mediator interaction, it is possible to identify the four effects of interest using formulas 8-11 (Appendix A1.2):

$$NDE^a = \theta_a + \theta_{am}(\beta_0 + \beta_a) \quad (16)$$

$$NDE^{a'} = \theta_a + \theta_{am}\beta_0 \quad (17)$$

$$NIE^a = \beta_a(\theta_m + \theta_{am}) \quad (18)$$

$$NIE^{a'} = \theta_m\beta_a \quad (19)$$

It is noteworthy that in the presence of interaction, the direct and indirect effects diverge between the treated and untreated groups, however, the ATE expressed in formulas 6-7 remained unchanged:

$$ATE = NIE^a + NDE^{a'} = \beta_a(\theta_m + \theta_{am}) + \theta_a + \theta_{am}\beta_0 = \theta_a + \theta_{am}(\beta_0 + \beta_a) + \theta_m\beta_a$$

$$ATE = NDE^a + NIE^{a'} = \theta_a + \theta_{am}(\beta_0 + \beta_a) + \theta_m\beta_a$$

Thus, through formulas 8-11 and under the specified assumptions 1-3, it is feasible to identify the four effects defined in the formulas 1-4, by expressing them in terms of regression coefficients.

As highlighted in definition 5, $E[Y(a, M(a))]$ and $E[Y(a', M(a'))]$ can be expressed as $E[Y(a)]$ and $E[Y(a')]$. Under assumptions 1-3, it is possible to identify $E[Y(a)]$ and $E[Y(a')]$ using the observed conditional means $E(Y \mid A = a)$ and $E(Y \mid A = a')$, without necessarily observing the mediator and without relying on the regression coefficients.

However, identifying the mean counterfactual outcomes $E[Y(a, M(a'))]$ and $E[Y(a', M(a))]$, below indicated as the cross-world counterfactual means, cannot be achieved simply as an expected value over the observed data, but it is necessary to go through the calculation of the following integrals (Pearl, 2001; Imai et al., 2010):

$$E[Y(a, M(a'))] = \int_m E[Y \mid A = a, M] p(M \mid A = a') \partial m \quad (20)$$

$$E[Y(a', M(a))] = \int_m E[Y \mid A = a', M] p(M \mid A = a) \partial m \quad (21)$$

2.2 Y count, binary or polytomous (Log link)

When the outcome is a count, binary, or polytomous variable, the ATE can be expressed as the ratio of counterfactual means rather than their difference.

$$\text{ATE} = \frac{E[Y(a)]}{E[Y(a')]} = \frac{E[Y(a, M(a))]}{E[Y(a', M(a'))]} \quad (22)$$

In this case, the structure of direct and indirect effects will also change.

$$\text{NIE}^a = \frac{E[Y(a, M(a))]}{E[Y(a, M(a'))]} \quad (23)$$

$$\text{NIE}^{a'} = \frac{E[Y(a', M(a'))]}{E[Y(a', M(a))]} \quad (24)$$

$$\text{NDE}^a = \frac{E[Y(a, M(a))]}{E[Y(a', M(a))]} \quad (25)$$

$$\text{NDE}^{a'} = \frac{E[Y(a', M(a'))]}{E[Y(a, M(a'))]} \quad (26)$$

The ATE can be rewritten using the counterfactual ratio, in two different versions:

$$\text{ATE} = \frac{\frac{E[Y(a, M(a))]}{E[Y(a, M(a'))]}}{\frac{E[Y(a', M(a'))]}{E[Y(a, M(a'))]}} = \frac{\text{NIE}^a}{\text{NDE}^{a'}} \quad (27)$$

$$\text{ATE} = \frac{\frac{E[Y(a, M(a))]}{E[Y(a', M(a))]} }{\frac{E[Y(a', M(a'))]}{E[Y(a', M(a))]} } = \frac{\text{NDE}^a}{\text{NIE}^{a'}} \quad (28)$$

We consider the case of a binary, count, or polytomous outcome, assuming no treatment-mediator interaction and no confounder X for simplicity. The data generation process is described by the following equations:

$$\log(E[Y \mid A, M]) = \theta_0 + \theta_a A + \theta_m M \quad (29)$$

$$E[M \mid A] = \beta_0 + \beta_a A \quad (30)$$

$$M = E[M \mid A] + \epsilon_M; \quad E[\epsilon_M] = 0; \quad \epsilon_M \perp A \quad (31)$$

Again, under the assumptions 1-3, it is possible to identify both versions of NDEs and NIEs using formulas 8-11 (Appendix A1.3):

$$\text{NDE}^a = \exp(\theta_a) \quad (32)$$

$$\text{NDE}^{a'} = \frac{1}{\exp(\theta_a)} \quad (33)$$

$$\text{NIE}^a = \exp(\theta_m \beta_a) \quad (34)$$

$$\text{NIE}^{a'} = \frac{1}{\exp(\theta_m \beta_a)} \quad (35)$$

Therefore, the ATE can be rewritten using the four causal effects obtained.

$$\text{ATE} = \frac{\text{NIE}^a}{\text{NDE}^{a'}} = \exp(\theta_m \beta_a) \cdot \exp(\theta_a)$$

$$\text{ATE} = \frac{\text{NDE}^a}{\text{NIE}^{a'}} = \exp(\theta_a) \cdot \exp(\theta_m \beta_a)$$

Note that, in the absence of interaction, the NDEs and NIEs depend on the same combination of regression coefficients. Thus, although they are derived from the same coefficients, one is the inverse of the other for both treated and untreated groups.

Considering the other scenario with interaction, the data generation equations for Y become:

$$\log(\text{E}[Y \mid A, M]) = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM$$

Applying the formulas 8-11, under the assumptions 1-3, it is possible to identify the four effects of interest (Appendix A1.4):

$$\text{NDE}^a = \exp(\theta_a + \theta_{am}(\beta_0 + \beta_a)) \frac{M_{\epsilon_M}(\theta_m + \theta_{am})}{M_{\epsilon_M}(\theta_m)} \quad (36)$$

$$\text{NDE}^{a'} = \frac{1}{\exp(\theta_a + \theta_{am}\beta_0)} \frac{M_{\epsilon_M}(\theta_m)}{M_{\epsilon_M}(\theta_m + \theta_{am})} \quad (37)$$

$$\text{NIE}^a = \exp(\beta_a(\theta_m + \theta_{am})) \quad (38)$$

$$\text{NIE}^{a'} = \frac{1}{\exp(\theta_m \beta_a)} \quad (39)$$

Therefore, the ATE can be rewritten using the four causal effects obtained:

$$\begin{aligned} \text{ATE} &= \frac{\text{NIE}^a}{\text{NDE}^{a'}} = \exp(\beta_a(\theta_m + \theta_{am})) \exp(\theta_a + \theta_{am}\beta_0) \frac{M_{\epsilon_M}(\theta_m + \theta_{am})}{M_{\epsilon_M}(\theta_m)} = \\ &= \exp(\beta_a(\theta_m + \theta_{am}) + \theta_a + \theta_{am}\beta_0) \frac{M_{\epsilon_M}(\theta_m + \theta_{am})}{M_{\epsilon_M}(\theta_m)} \\ \text{ATE} &= \frac{\text{NDE}^a}{\text{NIE}^{a'}} = \exp(\theta_a + \theta_{am}(\beta_0 + \beta_a)) \frac{M_{\epsilon_M}(\theta_m + \theta_{am})}{M_{\epsilon_M}(\theta_m)} \exp(\theta_m \beta_a) = \\ &= \exp(\beta_a(\theta_m + \theta_{am}) + \theta_a + \theta_{am}\beta_0) \frac{M_{\epsilon_M}(\theta_m + \theta_{am})}{M_{\epsilon_M}(\theta_m)} \end{aligned}$$

In this scenario with interaction and a logarithmic link, it can be noted that not only do NIEs and NDEs differ between the treated and untreated groups, but also the direct effects are not expressed solely in terms of regression coefficients: the moment generating function (MGF) $M_{\epsilon_M}(\cdot)$ are also present.

Thus, through formulas 8-11 and under the specified assumptions 1-3, it is feasible to identify the four effects defined in formulas 23-26 by expressing them in terms of regression coefficients, and in the case of exposure-mediator interaction, also through the identification of MGFs.

3 Mediation analysis with hidden mediator or in presence of measurement error

In practice, we may encounter an unmeasured mediator or a mediator subject to measurement error. We demonstrate that it is still possible to identify direct and indirect natural effects as long as we have information on two proxies of the mediator with specific characteristics.

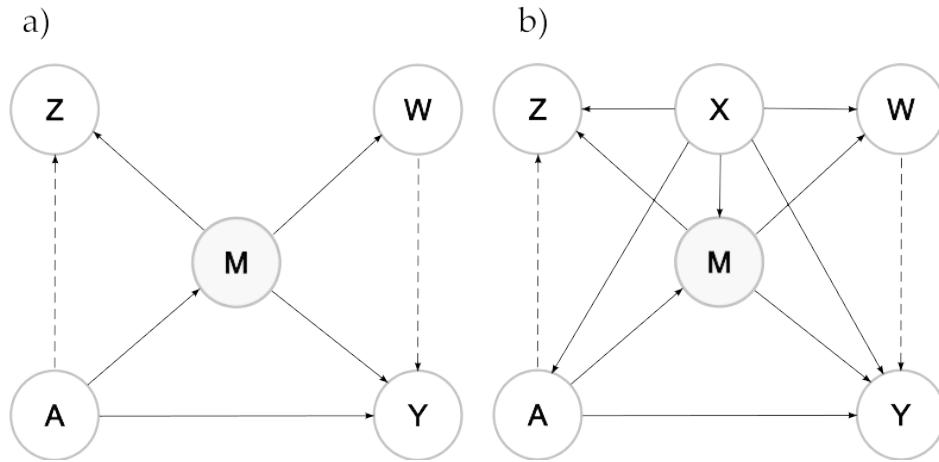


Figure 2: Mediation analysis with hidden mediator. The graphs illustrate the cases with a hidden or mismeasured mediator (node in grey), without a) and with b) confounder respectively. The dashed arrows can be present or absent in the DAG.

The necessary proxies in the proximal approach are two: Z and W. Z is caused by both treatment A and mediator M. Thus, it is associated with outcome Y through mediator M and treatment A, which are common causes of both Z and Y. W is caused only by mediator M and directly affects outcome Y. Its association with treatment A is only through the mediator. These assumptions are graphically represented in the DAGs in Figure 2, which considers both scenarios with and without the observable confounder X. In this figure, the dashed edges may also be absent. More precisely, the existing literature on the proximal causal framework (Tchetgen Tchetgen, Ying, Cui, 2020) formalizes the concepts shown in the DAG in Figure 2b, adaptable even to the case without confounder X shown in Figure 2a, by proposing the following assumption:

Assumption 4: (Conditional independence):

$$Z \perp Y \mid A, M, X;$$

$$W \perp (A, Z) \mid M, X.$$

Figures 2a and 2b illustrates graphical models that satisfy the proxy variable conditional conditions just introduced. We see in the following paragraphs that, under assumptions 1-4, even in the case of a hidden or mismeasured mediator, it is still possible to identify the direct and indirect natural effects in the two groups.

3.1 Y continuous (Identity link)

In the case of a continuous outcome, the identification of the four causal effects of interest can be achieved in two distinct ways, both based on the proximal framework.

In the scenario presented in Section 3.1.1, the necessary regression coefficients are estimated to reconstruct the various effects of interest. In the case described in Section 3.1.2, the cross-world counterfactual means, $E[Y(a, M(a'))]$ and $E[Y(a', M(a))]$, are identified, and the various causal effects are derived as differences or sums of the other two counterfactual means, $E[Y(a, M(a))]$ and $E[Y(a', M(a'))]$, which are always estimable from the observed data.

3.1.1 Identification through regression coefficients

The causal relationship between the variables considered is represented in the DAG in Figure 2a, where, for simplicity, the observable confounder X is excluded. We also assume no interaction between the exposure A and the hidden mediator M . All variables are considered continuous except for the treatment, which is dichotomous. The assumed data generation models for Y and W are as follows:

$$E[Y \mid A, Z, M] = \theta_0 + \theta_a A + \theta_m M$$

$$E[W \mid A, Z, M] = \alpha_0 + \alpha_m M$$

If we could observe the mediator M without measurement error, the NDE^a and $NDE^{a'}$ would be represented by the regression coefficient θ_a , as shown in the formulas 12-13 from paragraph 2.1. However, due to the inability to measure the variable M , the proximal framework based on a GLM approach is employed to identify the NDEs.

To address this, we take the expectation of both sides of the data generation equations for M , conditional on (A, Z) , and obtain the following two regression equations:

$$E[Y \mid A, Z] = \theta_0 + \theta_a A + \theta_m E[M \mid A, Z]$$

$$E[W \mid A, Z] = \alpha_0 + \alpha_m E[M \mid A, Z]$$

Note that $E[M \mid A, Z]$ is a linear transformation of $E[W \mid A, Z]$ and constants:

$$E[M \mid A, Z] = \frac{E[W \mid A, Z] - \alpha_0}{\alpha_m}$$

RESULT 1: Under the assumed data generation models one obtains:

$$E[Y \mid A, Z] = \theta_0^* + \theta_a^* A + \theta_m^* E[W \mid A, Z]$$

It is evident that the NDEs can be estimated by fitting two linear models: the first one is employed to obtain the conditional expectation of W , while the second one, based on the results obtained in the first stage, is utilized to derive the unbiased estimator for θ_a , more precisely:

$$1^{\text{st}} \text{ stage: } E[W \mid A, Z] = \omega_0 + \omega_a A + \omega_z Z$$

$$2^{\text{nd}} \text{ stage: } E[Y \mid A, Z] = \theta_0^* + \theta_a^* A + \theta_m^* E[W \mid A, Z]$$

$$\text{Where } \theta_0^* = \theta_0 - \frac{\theta_m}{\alpha_m} \alpha_0; \quad \theta_a^* = \theta_a; \quad \theta_m^* = \frac{\theta_m}{\alpha_m}.$$

It is worth noting that the regression coefficient of the treatment, θ_a , remains the same as would have been obtained if the mediator M could have been observed.

It is possible now formulate the proximal procedure to estimate the NDEs:

- Specify the linear function for $E[W \mid A, Z]$;
- Run the first-stage linear regression to obtain the regression coefficients $\hat{E}[W \mid A, Z] = \hat{\omega}_0 + \hat{\omega}_a A + \hat{\omega}_z Z$;
- Calculate $\hat{E}[W \mid A, Z]$ using the estimated coefficients from the first-stage regression;
- Run the second-stage linear regression $\hat{E}[Y \mid A, Z] = \hat{\theta}_0^* + \hat{\theta}_a^* A + \hat{\theta}_m^* \hat{E}[W \mid A, Z]$ to obtain $\hat{\theta}_a^*$.

In the context of the NIEs, they can be straightforwardly obtained by simply subtracting the estimated NDEs just derived from the ATE. Indeed, an estimate of the ATE can be obtained using the observed conditional means of the outcome Y in the two groups of treated and untreated individuals, even in presence of hidden mediator.

Alternatively, the NIEs, which from the formulas 14-15 in paragraph 2.1 we know to be $\theta_m \beta_a$, can be estimated by applying the proximal approach to estimate the regression coefficients.

In the following sections, we demonstrate that it is possible to identify the NDEs and NIEs even in the presence of interaction between the exposure A and the hidden mediator M .

In the data generation model for the outcome Y , the interaction term is now included; furthermore, an additional model for the mediator M is required.

$$E[Y \mid A, Z, M] = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM$$

$$E[M \mid A, Z] = \beta_0 + \beta_a A$$

$$M = E[M \mid A, Z] + \epsilon_M; \quad E[\epsilon_M] = 0; \quad \epsilon_M \perp (A, Z)$$

Based on the formula 16 from the previous paragraph 2.1, if we could observe the mediator M without measurement error, the NDE^a would be expressed by the follow regression coefficients $\theta_a + \theta_{am}(\beta_0 + \beta_a)$.

However, the mediator M is unobservable in our case. Once again, it is necessary to use the proximal approach.

Analogously to the scenario without interaction, we take the conditional expectation on both sides of the data generation equations for Y and W .

$$E[Y \mid A, Z] = \theta_0 + \theta_a A + \theta_m E[M \mid A, Z] + \theta_{am} A E[M \mid A, Z]$$

$$E[W \mid A, Z] = \alpha_0 + \alpha_m E[M \mid A, Z]$$

Even in this case, $E[M \mid A, Z]$ is a linear transformation of $E[W \mid A, Z]$ and the constants α_0 and α_m .

RESULT 2: Under the assumed data generation models one obtains:

$$E[Y \mid A, Z] = \theta_0^* + \theta_a^* A + \theta_m^* E[W \mid A, Z] + \theta_{am}^* A E[W \mid A, Z]$$

$$E[E[W \mid A, Z] \mid A, Z] = \beta_0^* + \beta_a^* A$$

Even within this context, estimating the NDE^a remains feasible by fitting various linear models. In the first stage, we derive the conditional expectation of W . Subsequently, in the second stage, based on the coefficients obtained from the first stage, we perform one regression on Y to derive the biased estimators for θ_a and θ_{am} , and another regression of the treatment A on $E[W \mid A, Z]$ from the first stage to derive the biased estimators for β_0 and β_a , more precisely:

$$1^{\text{st}} \text{ stage: } E[W \mid A, Z] = \omega_0 + \omega_a A + \omega_z Z$$

$$2^{\text{nd}} \text{ stage: } E[Y \mid A, Z] = \theta_0^* + \theta_a^* A + \theta_m^* E[W \mid A, Z] + \theta_{am}^* A E[W \mid A, Z]$$

$$E[E[W \mid A, Z] \mid A, Z] = \beta_0^* + \beta_a^* A$$

Where $\theta_0^* = \theta_0 - \frac{\theta_m}{\alpha_m} \alpha_0$; $\theta_a^* = \theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0$; $\theta_m^* = \frac{\theta_m}{\alpha_m}$; $\theta_{am}^* = \frac{\theta_{am}}{\alpha_m}$; $\beta_0^* = \beta_0 \alpha_m + \alpha_0$; $\beta_a^* = \beta_a \alpha_m$.

The combination of these biased estimator results in an unbiased estimator for the NDE^a .

$$NDE^a = \theta_a^* + \theta_{am}^*(\beta_0^* + \beta_a^*) = \left(\theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0 \right) + \left(\frac{\theta_{am}}{\alpha_m} \right) (\beta_0 \alpha_m + \alpha_0 + \beta_a \alpha_m) = \theta_a + \theta_{am}(\beta_0 + \beta_a)$$

It is possible now formulate the proximal procedure to estimate the NDE^a in presence of treatment-mediator interaction:

- Specify the linear function for $E[W \mid A, Z]$;
- Run the first-stage linear regression to obtain the regression coefficients $\hat{E}[W \mid A, Z] = \hat{\omega}_0 + \hat{\omega}_a A + \hat{\omega}_z Z$;
- Calculate $\hat{E}[W \mid A, Z]$ using the estimated coefficients from the first-stage regression;
- Run the second-stage linear regression: one on the outcome, $\hat{E}[Y \mid A, Z] = \hat{\theta}_0^* + \hat{\theta}_a^* A + \hat{\theta}_m^* \hat{E}[W \mid A, Z] + \hat{\theta}_{am}^* A \hat{E}[W \mid A, Z]$, and the other one on $\hat{E}[\hat{E}[W \mid A, Z] \mid A] = \hat{\beta}_0^* + \hat{\beta}_a^* A$ to obtain $\hat{\theta}_a^*$, $\hat{\theta}_{am}^*$, $\hat{\beta}_0^*$ and $\hat{\beta}_a^*$.
- Calculate NDE^a as $\theta_a^* + \theta_{am}^*(\beta_0^* + \beta_a^*)$

To derive $NIE^{a'}$, as in the scenario without interaction, we calculate the difference between the ATE, estimated using the observed conditional means, and the NDE^a obtained through the proximal approach just described. Even in this case, estimating the $NIE^{a'}$ is achievable alternatively by applying the proximal approach to estimate the various regression coefficients that constitute the effect of interest.

The same procedure is applicable for estimating the regression coefficients part of the causal effects $NDE^{a'}$ and NIE^a .

3.1.2 Identification through cross-world counterfactual means

In the previous paragraph 3.1.1, we obtained the NDEs and NIEs by estimating the individual biased regression coefficients that comprise the causal effects of interest. In this section, we identify the same causal effects, still based on the proximal approach, but by estimating the four counterfactual means members of the causal effects of interest (as shown in equations 1-4).

It is feasible to estimate $E[Y(a, M(a))]$ and $E[Y(a', M(a'))]$ utilizing the observed conditional means $E(Y \mid A = a)$ and $E(Y \mid A = a')$, even in the presence of the hidden mediator.

Identification of the cross-world counterfactual means, $E[Y(a, M(a'))]$ and $E[Y(a', M(a))]$, is also possible with the available data, even when the mediator is unobservable, using the proximal approach.

Consider the case where all variables are continuous except for the binary treatment, with an exposure-mediator interaction and no observable confounder X .

To identify the cross-world counterfactual mean $E[Y(a, M(a'))]$, which represents the average outcome for individuals who receive the treatment $A = a = 1$, but with the mediator value corresponding to not receiving the treatment, that is, when $A = a' = 0$, we consider the two data generation equations for Y and W in the treated group (Ghassami et al. 2024):

$$E[Y \mid A = 1, Z, M] = \theta_0 + \theta_a + \theta_m M + \theta_{am} M$$

$$E[W \mid A = 1, Z, M] = \alpha_0 + \alpha_m M$$

Taking the conditional expectation on both sides of the data generation equations we have:

$$E[Y \mid A = 1, Z] = (\theta_0 + \theta_a) + (\theta_m + \theta_{am}) E[M \mid A = 1, Z]$$

$$E[W \mid A = 1, Z] = \alpha_0 + \alpha_m E[M \mid A = 1, Z]$$

Note that, even in this case, $E[M \mid A = 1, Z]$ is a linear transformation of $E[W \mid A = 1, Z]$ and the constants α_0 and α_m .

RESULT 3: Under the assumed data generation models one obtains:

$$E[Y \mid A = 1, Z] = \theta_0^* + \theta_m^* E[W \mid A = 1, Z]$$

It is evident that it can be estimated by fitting two linear models: the first one to obtain the conditional expectation of W in the treated group, and the second one, based on the results from the first stage, to derive the regression coefficients θ_0^* and θ_m^* , more precisely:

$$1^{\text{st}} \text{ stage: } E[W \mid A = 1, Z] = \omega_0 + \omega_z Z$$

$$2^{\text{nd}} \text{ stage: } E[Y \mid A = 1, Z] = \theta_0^* + \theta_m^* E[W \mid A = 1, Z]$$

$$\text{Where } \theta_0^* = \theta_0 + \theta_a - \frac{\theta_m + \theta_{am}}{\alpha_m} \alpha_0 \text{ and } \theta_m^* = \frac{\theta_m + \theta_{am}}{\alpha_m}$$

Once the coefficients of the regression of $E[W \mid A = 1, Z]$ on Y in the treated group are estimated, they can be used, along with the observed expected value of W in the untreated group $E[W \mid A = 0]$, to estimate the desired conditional expected value (Ghassami et al. 2024):

$$E[Y(a, M(a'))] = \theta_0^* + \theta_m^* E[W \mid A = 0]$$

It is possible now formulate the procedure to estimate $E[Y(a, M(a'))]$ and consequently $NDE^{a'}$:

- Specify the linear function for $E[W | A = 1, Z]$;
- Run the first-stage linear regression to obtain the regression coefficients $\hat{E}[W | A = 1, Z] = \hat{\omega}_0 + \hat{\omega}_z Z$;
- Calculate $\hat{E}[W | A = 1, Z]$ using the estimated coefficients from the first-stage regression;
- Run the second-stage linear regression to obtain the regression coefficients $\hat{E}[Y | A = 1, Z] = \hat{\theta}_0^* + \hat{\theta}_m^* \hat{E}[W | A = 1, Z]$;
- Estimate $\hat{E}[Y(a, M(a'))]$ as $\hat{\theta}_0^* + \hat{\theta}_m^* \hat{E}[W | A = 0]$;
- Estimate $\hat{E}[Y(a', M(a'))]$ as $E[Y | A = 0]$;
- Calculate $NDE^{a'}$ as the difference $\hat{E}[Y(a, M(a'))] - \hat{E}[Y(a', M(a'))]$.

Once $E[Y(a, M(a'))]$ is estimated, it is possible to estimate the NIE^a as the difference with $E[Y(a, M(a))]$.

Obviously, this approach is also applicable for estimating $E[Y(a', M(a))]$ and, consequently, NDE^a and $NIE^{a'}$. In this case, the two data generation equations for Y and W are considered in the untreated group, that is $A = a' = 0$, and the coefficients $\hat{\theta}_0^*$ and $\hat{\theta}_m^*$ obtained are used to estimate the cross-world counterfactual mean along with the expected value of W in the treated group $\hat{E}[W | A = 1]$.

Finally, it should be noted that this procedure correctly identify the desired causal effects whether or not there is an interaction between the exposure and the mediator. Simply, if an interaction is present, its effect will be contained in the coefficients of the regression on Y . More precisely:

$$\theta_0^* = \theta_0 + \theta_a - \frac{(\theta_m + \theta_{am})}{\alpha_m} \alpha_0; \quad \theta_m^* = \frac{(\theta_m + \theta_{am})}{\alpha_m} \quad (40)$$

Thus, the advantage of utilizing this approach is that it does not require assuming a priori the presence or absence of the exposure-mediator interaction, as this interaction is appropriately accounted for in the estimation of effects in both cases.

Furthermore, an additional advantage is that the distributional form of M does not affect the method, allowing it to be employed regardless of the distributional characteristics of the unobserved mediator.

One drawback of this approach might be that it is "less parametric" than the previous one based only on the regression coefficients, as it considers the two observed means $E[Y | A = 1]$ and $E[Y | A = 0]$ in the data as estimates of the counterfactual means. Non-parametric methods may sometimes exhibit

lower efficiency compared to parametric methods, especially when the latter are well-specified and the underlying data assumptions are satisfied, as in the simulation cases.

3.2 Y count, binary or polytomous (Log link)

Now we consider the case in which the outcome is binary, count or polytomous, with logarithmic link function, and the other variables in the model are continuous, except for the binary treatment. For simplicity, we exclude a potential observable confounder X and the interaction between exposure and mediator. Again, the causal relationships among the variables are those represented in the DAG in Figure 2a.

The assumed data generation models for Y and W are as follows:

$$\log(E[Y \mid A, Z, M]) = \theta_0 + \theta_a A + \theta_m M$$

$$E[W \mid A, Z, M] = \alpha_0 + \alpha_m M$$

If we could observe the mediator M without measurement error, the NDE^a would be expressed as $\exp(\theta_a)$, as shown in formula 32 of Chapter 2.2. However, because observing the mediator M is not feasible, we use the proximal approach to identify the NDE^a .

Taking the conditional expectation on both sides of the data generation equations we have:

$$\log(E[Y \mid A, Z]) = \theta_0 + \theta_a A + \theta_m E[M \mid A, Z]$$

$$E[W \mid A, Z] = \alpha_0 + \alpha_m E[M \mid A, Z]$$

Again, $E[M \mid A, Z]$ is a linear transformation of $E[W \mid A, Z]$ and constants α_0 and α_m .

RESULT 4: Under the assumed data generation models one obtains:

$$\log(E[Y \mid A, Z]) = \theta_0^* + \theta_a^* A + \theta_m^* E[W \mid A, Z]$$

Even in this context, estimating the NDE^a is possible by fitting two regression models. In the first stage, we derive the conditional expectation of W . In the second stage, based on the coefficients obtained from the first stage, we perform one Poisson regression model on the outcome to derive the estimator for θ_a , more precisely:

$$1^{\text{st}} \text{ stage: } E[W \mid A, Z] = \omega_0 + \omega_a A + \omega_z Z$$

$$2^{\text{nd}} \text{ stage: } \log(E[Y \mid A, Z]) = \theta_0^* + \theta_a^* A + \theta_m^* E[W \mid A, Z]$$

Where $\theta_0^* = \theta_0' - \frac{\theta_m}{\alpha_m} \alpha_0$; $\theta_a^* = \theta_a$; $\theta_m^* = \frac{\theta_m}{\alpha_m}$.

θ_0' is defined as $\theta_0 + \log(M_{\varepsilon_M}(\theta_m))$ as described in the Appendix A2.4.

The regression coefficient of the treatment, θ_a , remains the same as would have been obtained if the mediator M could have been measured accurately.

It is now possible to formulate the proximal procedure to estimate the NDE^a:

- Specify the linear function for $E[W | A, Z]$;
- Run the first-stage linear regression model to obtain the regression coefficients $\hat{E}[W | A, Z] = \hat{\omega}_0 + \hat{\omega}_a A + \hat{\omega}_z Z$;
- Calculate $\hat{E}[W | A, Z]$ using the estimated coefficients from the first-stage regression;
- Run the second-stage Poisson regression model on the outcome Y, $\log(\hat{E}[Y | A, Z]) = \hat{\theta}_0^* + \hat{\theta}_a^* A + \hat{\theta}_m^* \hat{E}[W | A, Z]$ to obtain $\hat{\theta}_a^*$.

The NIE^{a'} can be estimated simply by dividing the NDE^a by the ATE, as derived from formula 28 in Chapter 2.2. Recall that the ATE can be estimated from the data using the observed conditional means, even when the mediator is hidden. Alternatively, the NIE^{a'} can also be estimated by applying the proximal procedure just described.

It is possible to identify the NDEs and NIEs even in the presence of interaction between the exposure and the hidden mediator. In the data generation model for the outcome Y, the interaction term is now included; furthermore, an additional model for the mediator M is required:

$$\log(E[Y | A, Z, M]) = \theta_0 + \theta_a A + \theta_m M + \theta_{am} MA$$

$$E[M | A, Z] = \beta_0 + \beta_a A$$

$$M = E[M | A, Z] + \epsilon_M; \quad E[\epsilon_M] = 0; \quad \epsilon_M \perp (A, Z)$$

According to formulas 36 and 37, both NDEs are not only functions of regression coefficients but also of moment-generating functions, which cannot be identified by the proximal approach. However, this does not prevent us from identifying the direct effects, which can be obtained by dividing or multiplying the indirect effects by the ATE as follows:

$$\text{NDE}^{a'} = \frac{\text{NIE}^a}{\text{ATE}}$$

$$\text{NDE}^a = \text{ATE} \cdot \text{NIE}^{a'}$$

Indeed, both NIEs are estimable using the proximal approach, as they depend solely on regression coefficients and do not require the estimation of moment-generating functions, as shown in formulas 38 and 39.

If we could observe the mediator M without measurement error, the $\text{NIE}^{a'}$ would be expressed as $\frac{1}{\exp(\theta_m \beta_a)}$, as derived from formula 39 in Chapter 2.2. However, due to the impossibility of observing the mediator M , the proximal approach is used.

We take the conditional expectation on both sides of the data generation equations for Y and W .

$$\log(\mathbb{E}[Y \mid A, Z]) = \theta_0 + \theta_a A + \theta_m \mathbb{E}[M \mid A, Z] + \theta_{am} A \mathbb{E}[M \mid A, Z]$$

$$\mathbb{E}[W \mid A, Z] = \alpha_0 + \alpha_m \mathbb{E}[M \mid A, Z]$$

As in the other cases $\mathbb{E}[M \mid A, Z]$ is a linear transformation of $\mathbb{E}[W \mid A, Z]$ and the two constants α_0 and α_m .

RESULT 5: Under the assumed data generation equations one obtains:

$$\log(\mathbb{E}[Y \mid A, Z]) = \theta_0^* + \theta_a^* A + \theta_m^* \mathbb{E}[W \mid A, Z] + \theta_{am}^* A \mathbb{E}[W \mid A, Z]$$

$$\mathbb{E}[\mathbb{E}[W \mid A, Z] \mid A, Z] = \beta_0^* + \beta_a^* A$$

Estimating the $\text{NIE}^{a'}$ is possible by fitting various models. In the first stage, we derive the conditional expectation of W . Subsequently, in the second stage, based on the coefficients obtained from the first stage, we perform a Poisson regression model on Y to derive the biased estimators for θ_m , and another linear regression model of the treatment A on $\mathbb{E}[W \mid A, Z]$ from the first stage to derive the biased estimators for β_a , more precisely:

$$1^{\text{st}} \text{ stage: } \mathbb{E}[W \mid A, Z] = \omega_0 + \omega_a A + \omega_z Z$$

$$2^{\text{nd}} \text{ stage: } \log(\mathbb{E}[Y \mid A, Z]) = \theta_0^* + \theta_a^* A + \theta_m^* \mathbb{E}[W \mid A, Z] + \theta_{am}^* A \mathbb{E}[W \mid A, Z]$$

$$\mathbb{E}[\mathbb{E}[W \mid A, Z] \mid A, Z] = \beta_0^* + \beta_a^* A$$

Where $\theta_0^* = \theta_0'' - \frac{\theta_m}{\alpha_m} \alpha_0$, $\theta_a^* = \theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0$, $\theta_m^* = \frac{\theta_m}{\alpha_m}$, $\theta_{am}^* = \frac{\theta_{am}}{\alpha_m}$, $\beta_0^* = \beta_0 \alpha_m + \alpha_0$, and $\beta_a^* = \beta_a \alpha_m$.

θ_0'' is defined as $\theta_0 + \log(M_{\epsilon_M}(\theta_m + \theta_{am}))$ as described in the Appendix A2.5.

The combination of these biased estimator results in an unbiased estimator for the $\text{NIE}^{a'}$.

$$\text{NIE}^{a'} = \theta_m^* \beta_a^* = \left[\frac{\theta_m}{\alpha_m} \right] [\beta_a \alpha_m] = \theta_m \beta_a$$

It is now possible to formulate the proximal procedure to estimate the $\text{NIE}^{a'}$:

- Specify the linear function for $E[W \mid A, Z]$;
- Run the first-stage linear regression to obtain the regression coefficients $\hat{E}[W \mid A, Z] = \hat{\omega}_0 + \hat{\omega}_a A + \hat{\omega}_z Z$;
- Calculate $\hat{E}[W \mid A, Z]$ using the estimated coefficients from the first-stage regression;
- Run the second-stage regression models: one Poisson model for the outcome Y , $\log(\hat{E}[Y \mid A, Z]) = \hat{\theta}_0^* + \hat{\theta}_a^* A + \hat{\theta}_m^* \hat{E}[W \mid A, Z] + \hat{\theta}_{am}^* A \hat{E}[W \mid A, Z]$, and one linear model on $\hat{E}[\hat{E}[W \mid A, Z] \mid A, Z] = \hat{\beta}_0^* + \hat{\beta}_a^* A$ to obtain $\hat{\theta}_m^*$ and $\hat{\beta}_a^*$.
- Calculate $\text{NIE}^{a'}$ as $\hat{\theta}_m^* \hat{\beta}_a^*$

The NDE^a can be estimated simply by multiplying the ATE by the $\text{NIE}^{a'}$ as shown in formula 28, recalling that the ATE can be estimated from the data.

Alternatively, using the same proximal approach, it is possible to estimate the NIE^a and consequently the $\text{NDE}^{a'}$ by dividing the NIE^a by the ATE.

4 Simulations

To thoroughly evaluate the performance of the estimators proposed in the previous chapters, we conducted extensive simulation studies. These simulations were designed to examine both linear and logarithmic links, allowing us to assess the performance of the estimators across various scenarios. We also considered cases with and without exposure-mediator interactions, providing a comprehensive understanding of how these factors influence the estimation process.

For each scenario, we generated N i.i.d. observations of (Y, A, M, Z, W) , where the sample size N took values from (250, 500, 1000, 1500, 2000, 2500). For each value of N , we evaluated the performance of each estimator through 1000 iterations.

To assess the estimators, we used graphical tools such as boxplots and tables showing the bias (calculated as the difference between the true parameter used in data generation and the estimated value), standard errors (both empirical and those obtained via the Monte Carlo method), and the empirical coverage of the true parameter by the empirical confidence intervals.

4.1 Y continuous (Identity link)

Using the simulated data, we evaluate four estimators:

- The oracle estimator: we run the oracle linear regression model $E[Y | A, M]$ when the mediator M is observed, and we use the regression coefficients to obtain an estimator for NDE^a .
- The proximal estimator: an estimator based on the proximal approach presented in Results 1 and 2 (Section 3.1.1) of this text.
- The naive estimator: we consider the proxy W as the mediator, disregarding M . We run a linear regression model $E[Y | A, W]$ and use the regression coefficients to obtain an estimator for NDE^a .
- The cross-world estimator: an estimator based on the estimation of cross-world counterfactual means presented in Result 3 (Section 3.1.2) of this text.

For the linear case without exposure-mediator interaction we have the following data generating models:

$$E[Y | A, Z, M] = \theta_0 + \theta_a A + \theta_m M$$

$$E[W | A, Z, M] = \alpha_0 + \alpha_m M$$

N	Bias	Monte Carlo s.e.	Empirical s.e.	Coverage (%)
250	0.01	0.357	0.270	86.2
500	0.01	0.245	0.190	88.4
1000	0.00	0.175	0.135	87.0
1500	0.01	0.137	0.110	88.3
2000	0.00	0.124	0.095	86.5
2500	0.00	0.109	0.085	88.3

Table 1: Summary statistics for the cross-world estimator of NDE^a in the linear case without exposure-mediator interaction. The Bias column shows the average empirical bias of 1000 estimates for the estimator. The Monte Carlo s.e. column presents the standard deviations of 1000 estimates for each estimator. The Empirical s.e. column displays the average standard deviation of the 1000 estimates for the estimator. The Coverage column indicates the empirical coverage rates of 95% confidence intervals based on the empirical standard errors.

We consider the following data generation process with predetermined parameters:

- Generate the dichotomous treatment A independently from a Bernoulli distribution with a probability of success equal to $p = 0.6$;
- Generate the continuous mediator $M | A$, where $E[M | A] = 1.5 + 2.5 A$, and the error term follows a Normal distribution with $\mu = 0$ and $\sigma = 1$;
- Generate the continuous proxy $Z | A, M$, where $E[Z | A, M] = 1 + 3.59 A + 2.55 M$, and the error term follows a Normal distribution with $\mu = 0$ and $\sigma = 2$;

- Generate the Bivariate Normal $Y, W \mid A, M$: First, we generate two standard normal variables z_1 and z_2 ; then, we generate $E[W \mid M] = z_1 + 0.48 + 0.6 M$ and $E[Y \mid A, M, W] = \rho z_1 + \sqrt{1 - \rho^2} z_2 + 0.63 + 0.86 A + 1.54 M$, where ρ is the correlation coefficient between Y and W , equal to 0.6.

N	Bias	Monte Carlo s.e.	Empirical s.e.	Coverage (%)
250	0.01	0.328	0.318	94.3
500	0.01	0.222	0.225	94.6
1000	0.00	0.162	0.159	94.9
1500	0.01	0.126	0.129	96.1
2000	0.00	0.116	0.112	94.9
2500	0.01	0.101	0.100	93.8

Table 2: Summary statistics for the proximal estimator of NDE^a in the linear case without exposure-mediator interaction. The Bias column shows the average empirical bias of 1000 estimates for the estimator. The Monte Carlo s.e. column presents the standard deviations of 1000 estimates for each estimator. The Empirical s.e. column displays the average standard deviation of the 1000 estimates for the estimator. The Coverage column indicates the empirical coverage rates of 95% confidence intervals based on the empirical standard errors.

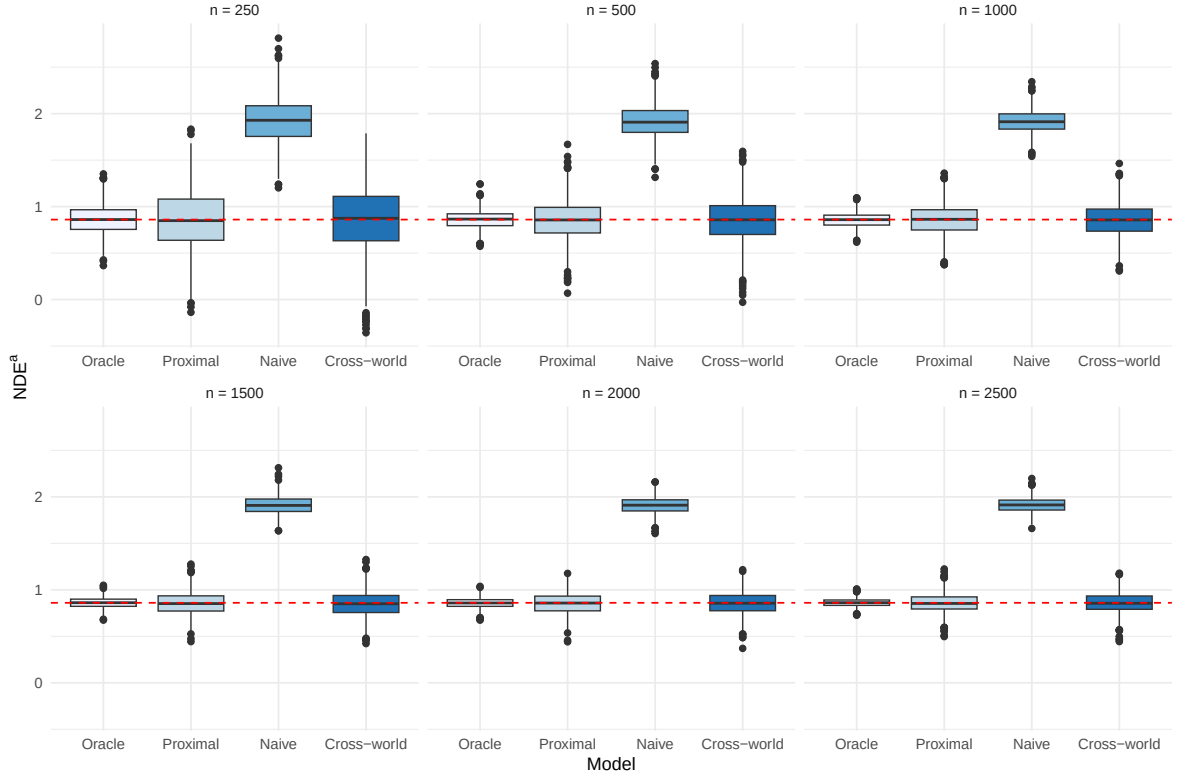


Figure 3: Simulation result for the linear case without exposure-mediator interaction. Each boxplot shows the estimates of NDE^a of the four estimators for $N = 250, 500, 1000, 1500, 2000, 2500$ respectively.

The simulation results are illustrated in Figure 3. It shows that the naive estimator produces a biased estimate of the effect of interest compared to the oracle estimator, across all sample sizes. In contrast, our two estimators, proximal and cross-world, remain unbiased, although they exhibit higher variability.

ity than the oracle estimator. This increased variability is due to the greater uncertainty stemming from the fact that we do not use the mediator M in the analysis but rather estimates based on the two proxies.

Tables 1 and 2 present numerical summaries for both estimators. Inference for these estimators is performed by considering both empirical standard errors and those obtained through the Monte Carlo method, where the standard deviation of 1000 estimates is calculated for each sample size. As the sample size N increases, both the empirical and Monte Carlo standard errors decrease. Additionally, we evaluate the empirical coverage of the 95% confidence intervals, to determine the percentage of intervals that contain the true value of NDE^a used in the data generation process.

The coverage of both estimators is satisfactory, although the coverage of the proximal estimator is better than that of the cross-world estimator.

N	Bias	Monte Carlo s.e.	Empirical s.e.	Coverage (%)
250	0.004	0.489	0.140	42.1
500	0.005	0.350	0.100	43.3
1000	-0.002	0.246	0.070	42.5
1500	0.003	0.196	0.057	41.7
2000	0.002	0.172	0.049	41.7
2500	0.006	0.156	0.044	43.1

Table 3: Summary statistics for the cross-world estimator of NDE^a in the linear case with exposure-mediator interaction. The Bias column shows the average empirical bias of 1000 estimates for the estimator. The Monte Carlo s.e. column presents the standard deviations of 1000 estimates for each estimator. The Empirical s.e. column displays the average standard deviation of the 1000 estimates for the estimator. The Coverage column indicates the empirical coverage rates of 95% confidence intervals based on the empirical standard errors.

N	Bias	Monte Carlo s.e.	Empirical s.e.	Coverage (%)
250	0.004	0.529	0.516	94.7
500	0.001	0.364	0.366	94.7
1000	0.004	0.253	0.258	95.4
1500	0.003	0.206	0.211	95.6
2000	0.004	0.181	0.183	95.5
2500	0.003	0.167	0.163	93.8

Table 4: Summary statistics for the proximal estimator of NDE^a in the linear case with exposure-mediator interaction. The Bias column shows the average empirical bias of 1000 estimates for the estimator. The Monte Carlo s.e. column presents the standard deviations of 1000 estimates for each estimator. The Empirical s.e. column displays the average standard deviation of the 1000 estimates for the estimator. The Coverage column indicates the empirical coverage rates of 95% confidence intervals based on the empirical standard errors.

A similar procedure was also applied in the case with exposure-mediator interaction. Only the as-

sumption regarding $E[Y | A, Z, M]$ is different; it actually includes the coefficient for the interaction:

$$E[Y | A, Z, M] = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM$$

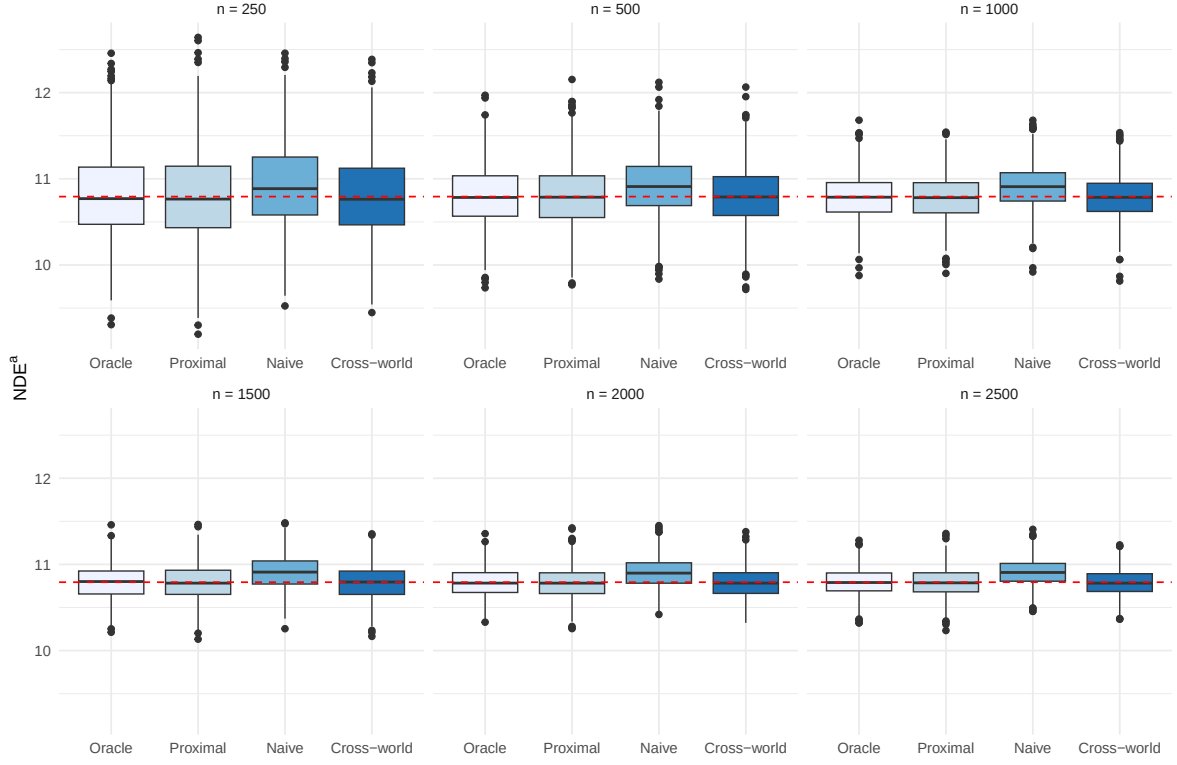


Figure 4: Simulation result for the linear case with exposure-mediator interaction. Each boxplot shows the estimates of NDE^a of the four estimators for $N = 250, 500, 1000, 1500, 2000, 2500$ respectively.

We consider the following data generation process:

- Generate the dichotomous treatment A independently from a Bernoulli distribution with a probability of success equal to $p = 0.6$;
- Generate the continuous mediator $M | A$, where $E[M | A] = 1.66 + 2.25 A$, and the error term follows a Normal distribution with $\mu = 0$ and $\sigma = 1$;
- Generate the continuous proxy $Z | A, M$, where $E[Z | A, M] = 1 + 2 A + 1.5 M$, and the error term follows a Normal distribution with $\mu = 0$ and $\sigma = 2$;
- Generate the Bivariate Normal $Y, W | A, M$: First, we generate two standard normal variables z_1 and z_2 ; then, we generate $E[W | M] = z_1 + 1.98 + 1.5 M$ and $E[Y | A, M, W] = \rho z_1 + \sqrt{1 - \rho^2} z_2 + 0.63 + 0.86 A + 1.54 M + 2.54 A M$, where ρ is the correlation coefficient between Y and W , equal to 0.6.

The results are presented in Tables 3 and 4, as well as in Figure 4. In particular, Figure 4 displays the estimates of NDE^a when the interaction term is present: the naive estimator is biased for every sample size, further illustrating that using a proxy instead of the mediator is not an effective solution. The proximal and cross-world estimators are not biased, but they exhibit greater uncertainty compared to the oracle estimator, as they are derived without utilizing the mediator M .

From Tables 3 and 4, it is clear that even in the presence of exposure-mediator interaction, the standard errors naturally decrease as the sample size increases. The standard errors of the proximal estimator are larger than those of the cross-world estimator; however, its coverage is significantly more satisfactory.

As anticipated at the end of section 3.1.2, it was expected that, in cases of correct model specification (as is the case in these simulations), the performance of the "less parametric" estimator, the cross-world estimator, might be less satisfactory compared to that of the estimator fully based on GLMs.

4.2 Y count, binary or polytomous (Log link)

For the case with a binary Y and logarithmic link, both with and without exposure-mediator interaction, simulations were conducted to evaluate the performance of the three estimators: Oracle, Proximal, and Naive. For the logarithmic case the Cross-world estimator is not applicable.

We have the following data generating models:

$$\log(E[Y | A, Z, M]) = \theta_0 + \theta_a A + \theta_m M$$

$$E[W | A, Z, M] = \alpha_0 + \alpha_m M$$

We consider the following data generation process with predetermined parameters:

- Generate the dichotomous treatment A independently from a Bernoulli distribution with a probability of success equal to $p = 0.6$.
- Generate the continuous mediator $M | A$, where $E[M | A] = 2.8 + 0.8 A$ and the error term has a uniform distribution in the interval $[-1, 1]$.
- Generate the continuous proxy $Z | A, M$, where $E[Z | A, M] = 0.9 + 3.59 A + 2.55 M$ and the error term has a normal distribution with $\mu = 0$ and $\sigma = 2.1$.
- Generate the continuous proxy $W | M$, where $E[W | M] = -0.49 - 0.9 W$ and the error term has a uniform distribution in the interval $[-1, 1]$.
- Generate the dichotomous outcome $Y | A, M, W$, where $\log(E[Y | A, M, W]) = -0.55 + 0.6 A - 0.65 M - 0.3 W$.

The simulation results are illustrated in Figure 5 and summarized in Table 5. It can be observed that the coverage rate is satisfactory for all sample sizes. Additionally, the estimates of NDE^a are biased when using the proxy instead of the mediator, while the proximal estimator is unbiased, despite exhibiting greater variability compared to the oracle estimator.

Sample size	Bias	MC s.e.	Empirical s.e.	Coverage (%)
250	-0.044	0.560	0.522	93.9
500	-0.006	0.359	0.346	94.0
1000	-0.003	0.243	0.238	92.8
1500	0.007	0.194	0.194	92.9
2000	-0.000	0.170	0.164	91.4
2500	-0.001	0.153	0.147	92.1

Table 5: Summary statistics for the proximal estimator of NDE^a in the logarithmic case without exposure-mediator interaction. The Bias column shows the average empirical bias of 1000 estimates for the estimator. The Monte Carlo s.e. column presents the standard deviations of 1000 estimates for each estimator. The Empirical s.e. column displays the average standard deviation of the 1000 estimates for the estimator. The Coverage column indicates the empirical coverage rates of 95% confidence intervals based on the empirical standard errors.

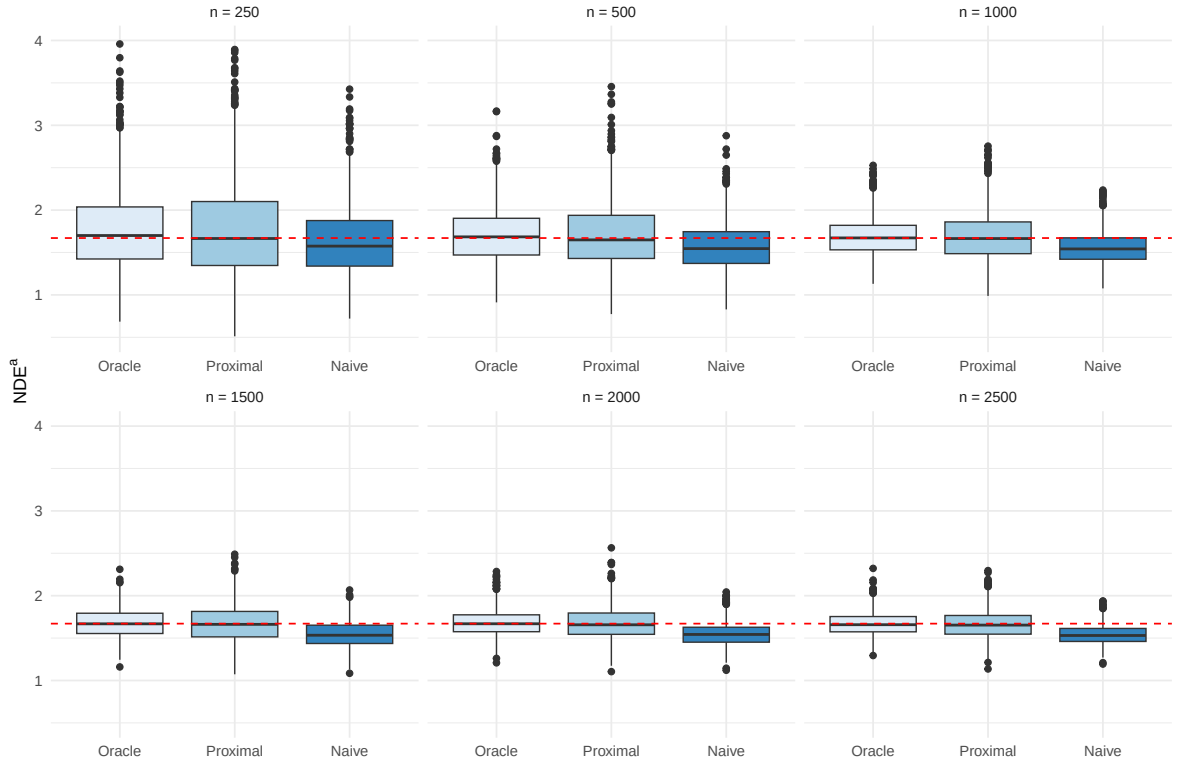


Figure 5: Simulation result for the log case without exposure-mediator interaction. Each boxplot shows the estimates of NDE^a of the three estimators for $N = 250, 500, 1000, 1500, 2000, 2500$ respectively.

In the logarithmic case with exposure-mediator interaction, the assumption regarding $\log(E[Y \mid A, Z, M])$ differs as it includes the coefficient for the interaction:

$$\log(E[Y \mid A, Z, M]) = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM$$

Sample size	Bias	MC s.e.	Empirical s.e.	Coverage (%)
250	-0.138	0.311	0.233	87.5
500	-0.004	0.212	0.156	90.9
1000	0.013	0.131	0.104	90.1
1500	0.001	0.106	0.085	89.2
2000	-0.005	0.088	0.072	90.9
2500	-0.003	0.080	0.065	90.1

Table 6: Summary statistics for the proximal estimator of NDE^a in the logarithmic case with exposure-mediator interaction. The Bias column shows the average empirical bias of 1000 estimates for the estimator. The Monte Carlo s.e. column presents the standard deviations of 1000 estimates for each estimator. The Empirical s.e. column displays the average standard deviation of the 1000 estimates for the estimator. The Coverage column indicates the empirical coverage rates of 95% confidence intervals based on the empirical standard errors.

We consider the following data generation process:

- Generate the dichotomous treatment A independently from a Bernoulli distribution with a probability of success equal to $p = 0.6$.
- Generate the continuous mediator $M \mid A$, where $E[M \mid A] = 2.8 + 0.8 A$ and the error term has a uniform distribution in the interval $[-1, 1]$.
- Generate the continuous proxy $Z \mid A, M$, where $E[Z \mid A, M] = 0.9 + 3.59 A + 2.55 M$ and the error term has a normal distribution with $\mu = 0$ and $s = 2.1$.
- Generate the continuous proxy $W \mid M$, where $E[W \mid M] = -0.49 - 0.9 W$ and the error term has a uniform distribution in the interval $[-1, 1]$.
- Generate the dichotomous outcome $Y \mid A, M, W$, where $\log(E[Y \mid A, M, W]) = -0.1 + 0.7 A - 0.6 M - 0.3 A M - 0.3 W$.

The simulation results are illustrated in Figure 6 and summarized in Table 6. In the logarithmic case with interaction, it is not possible to directly estimate the NDEs due to the moment generating functions ratio. Therefore, only the NIEs can be estimated directly using the regression coefficients, and the NDEs are derived by multiplying the ATE.

In this scenario, the estimates obtained for NDE^a using the proximal estimator remain unbiased, in contrast to those obtained by replacing the mediator with the proxy, that is the naive estimator. While

the standard errors decrease as the sample size increases, the coverage rates remains satisfactory for all sample sizes.

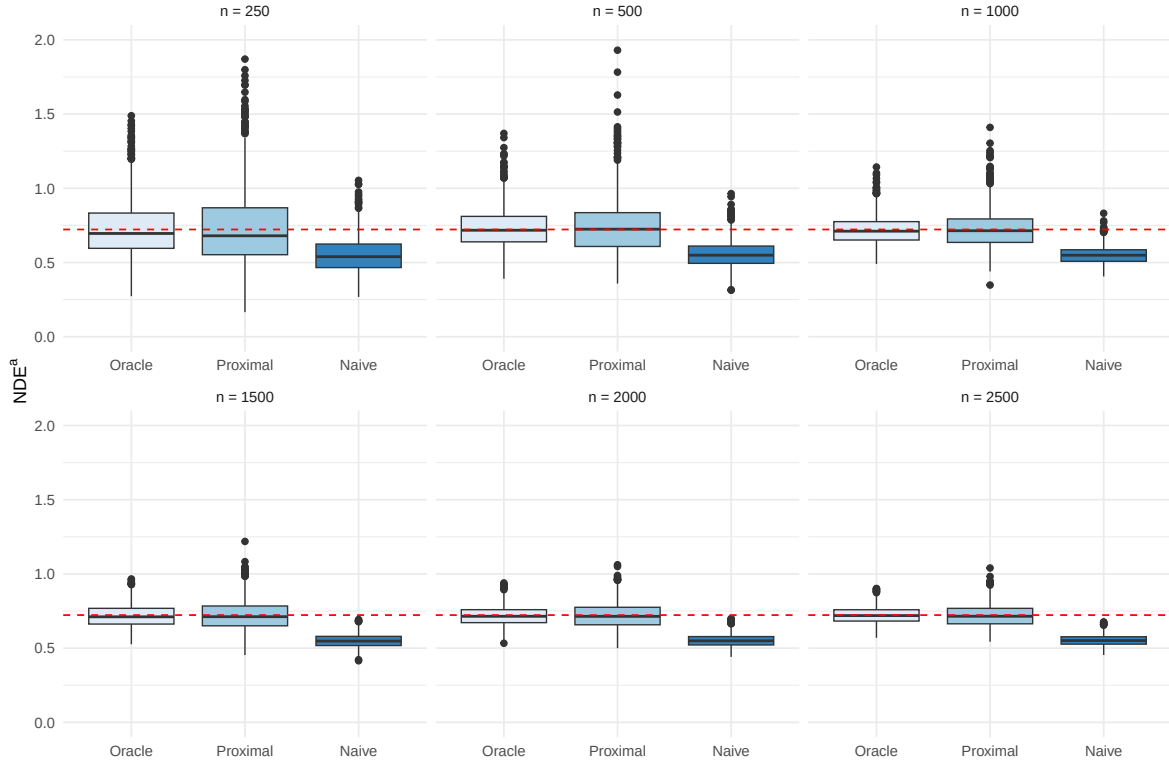


Figure 6: Simulation result for the log case with exposure-mediator interaction. Each boxplot shows the estimates of NDE^a of the three estimators for $N = 250, 500, 1000, 1500, 2000, 2500$ respectively.

5 Measurement error arising from the use of a proxy in place of the mediator

As observed in the previous chapter, the bias resulting from using a proxy instead of the mediator does not appear to follow a specific direction; in fact, it may either underestimate (e.g., Figure 5) or overestimate (e.g., Figure 3) the causal effects that would be obtained using the mediator M . For this reason, it is important to identify the components that contribute to the bias arising from the use of the proxy.

We explore in detail the scenario where, instead of applying methods based on the proximal framework, we opt to use one proxy variable, in this case W , in place of the unobserved mediator M .

In this chapter, the proxy W will be considered as a mismeasured version of the unobservable mediator M . It is particularly relevant to investigate the implications of measurement error in situations where the methods presented in this thesis are not applied. It is expected that the use of W , a mediator

measured with error, would result in biased estimates.

Notably, the magnitude of the bias is expected to diminish as the characteristics of the proxy variable W more closely resemble those of the mediator M .

5.1 Measurement error in linear models and attenuation

The simplest case of measurement error is represented by the classical additive nondifferential measurement error model (Carroll et. al, 2006). We have the following outcome generating equation:

$$Y = \theta_0 + \theta_m M + \epsilon_y^m$$

but instead of observing M , we observe W :

$$W = M + U$$

It is measured with additive error U with constant variance σ_u^2 :

$$U \sim N(0, \sigma_u^2), \quad U \perp M$$

So, we have:

$$Y = \theta_0^* + \theta_m^* W + \epsilon_y^w$$

Furthermore, in this context, the measurement error is non-differential, meaning that $\text{Cor}(\epsilon_y^m, U) = 0$. Additionally, we consider W as a surrogate for M . In other words, the conditional distribution of Y given (M, W) depends only on M ; W provides no additional information about Y beyond what M already provides.

In this simple case the biased coefficient obtained using W instead M is θ_m^* :

$$\theta_m^* = \theta_m \frac{\sigma_m^2}{\sigma_m^2 + \sigma_u^2}$$

where σ_m^2 is the variance of the mediator variable.

Given that

$$\frac{\sigma_m^2}{\sigma_m^2 + \sigma_u^2} < 1$$

the magnitude of the regression coefficient is biased towards zero:

$$|\theta_m^*| < |\theta_m|$$

and $\theta_m^* \rightarrow \theta_m$ as $\sigma_u^2 \rightarrow 0$.

This kind of bias is commonly referred to as "attenuation" or "attenuation to the null". θ_m^* underestimates the "true" effect of the variable M. The measurement error tends to make it more difficult to reject the null hypothesis of no effect (i.e., $H_0 : \theta_m = 0$).

In the case where the aim is to test for the presence of a non null effect, one might also consider to ignore the bias: in fact, even without adjusting for measurement error, if the test yields a significant result, it strongly suggests that there is an effect of the variable M. A significant result indicates that the effect does exist, but it is likely underestimated.

This concept is not applicable if there are deviations from the classical additive non-differential error model.

First, we have the following outcome generating equation, with another variable measured accurately, the treatment A:

$$Y = \theta_0 + \theta_a A + \theta_m M + \epsilon_y^m$$

Second, W is biased for M:

$$W = \gamma_0 + \gamma_m M + U$$

$$U \sim N(0, \sigma_u^2), \quad U \perp (A, M)$$

So, we have:

$$Y = \theta_0^* + \theta_a^* A + \theta_m^* W + \epsilon_y^w$$

Third, the error is differential, meaning there is some correlation between the outcome error term, ϵ_y^m and the error term for the proxy, U. In other words, the bias shows a structured pattern. Here, W does not serve as a simple surrogate for M; rather, it acts as an independent proxy variable. This implies that the conditional distribution of Y given (M, W) differs from the conditional distribution of Y given M alone. Consequently, W is not merely a noisy version of M, but a distinct variable that provides additional information about Y beyond what M offers

In this case we have:

$$\text{Cor}(\epsilon_y^m, U) = \rho_{\epsilon_y^m U} \neq 0$$

Suppose we are interested in estimating θ_a^* . If we use W in place of the mediator M, the coefficient θ_a can be shown to be biased, even if A is measured correctly. The difference ($\theta_a^* - \theta_a$) represents that bias.

Based on Carroll (2006) results, we have:

$$\theta_a^* = \theta_a + \theta_m \left(1 - \frac{\sigma_{m|a}^2}{\gamma_m^2 \sigma_{m|a}^2 + \sigma_u^2} \right) \delta_a$$

where $\sigma_{m|a}^2$ is the residual variances of the regression of M on A and δ_a is the coefficient of A in the regression $E(M | W, A) = \delta_0 + \delta_w W + \delta_a A$.

It should be noted that, as shown in Appendix A4.3, deriving this bias required a stronger assumption compare to the previous case: the joint normality distribution of M, W and A.

$$M | (W, A) \sim \mathcal{N}(\mu_{M|W,A}, \sigma_{M|W,A}^2)$$

Specifically, this assumption implies that the three variables involved, M, W, and A, must be continuous and normally distributed and the joint distribution of M given W and A, is normally distributed, ensuring that the relationship between these variables adheres to the properties of normality.

We can observe that if $\gamma_m \neq 1$, meaning that W is a biased proxy for M, and both the effect of the treatment on the mediator, δ_a , conditioning on W, and the effect of the mediator on the outcome, θ_m , conditioning on the treatment, are positive, it is possible to have:

$$|\theta_a^*| > |\theta_a|$$

In this context, the biased estimate θ_a^* cannot be constrained within the range $[0, \theta_a]$; rather, it may exceed it or even assume negative values. It is not feasible to define a restricted interval within which it may reside. Furthermore, ignoring the measurement error is inadmissible, even testing the null hypothesis of no effect.

5.2 Bias quantification in NDE^a in linear model without exposure-mediator interaction

The linear case without treatment-mediator interaction considered in the first case of the chapters 2.1, 3.1.1 and 3.1.2 of this thesis, has two differences compared to the last linear case previously examined.

The first difference is that the treatment considered is a binary variable, not a continuous one, with a

Bernoulli distribution.

$$A \sim \text{Bernoulli}(p)$$

The second difference is that the variables W and Y are assumed to follow a bivariate normal distribution:

$$\begin{pmatrix} W \\ Y \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mu_w \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_w^2 & \sigma_{wy} \\ \sigma_{wy} & \sigma_y^2 \end{pmatrix} \right)$$

where μ_w and μ_y are the means, and σ_w^2 and σ_y^2 are the variances of W and Y , respectively, while σ_{wy} represents the non-zero covariance between the two variables.

The outcome data generating equation is:

$$Y = \theta_0 + \theta_m M + \theta_a A + \epsilon_y^m$$

Since we are unable to observe the variable M or measure it without error, we use the biased proxy W instead of M .

$$W = \gamma_0 + \gamma_m M + U$$

$$U \sim N(0, \sigma_u^2), \quad U \perp (A, M)$$

Thus, we have:

$$Y = \theta_0^* + \theta_m^* W + \theta_a^* A + \epsilon_y^w$$

In this specific scenario, an effort has been made to quantify the bias introduced in the NDE^a, that is θ_a , by using the mediator proxy W instead of the mediator itself. To achieve this, we need to introduce the treatment data generation equation:

$$A = \alpha_0 + \alpha_w W + \Delta$$

$$E[\Delta] = 0; \quad \Delta \perp W$$

The two data generating models for A and Y can be rewritten as the moment conditions (Appendix A4.5 Proof 1):

$$E \left[\begin{pmatrix} 1 \\ W \\ A \end{pmatrix} (Y - \theta_0^* - \theta_m^* W - \theta_a^* A) \right] = 0 \quad (41)$$

$$E \left[\begin{pmatrix} 1 \\ W \end{pmatrix} (A - \alpha_0 - \alpha_w W) \right] = 0 \quad (42)$$

The one for Y can be rewrite as follow (Appendix A4.5 Proof 2):

$$E[\Delta(Y - \theta_a^* \Delta)] = 0$$

$$E[\Delta Y] = \theta_a^* E[\Delta^2]$$

$$\theta_a^* = E[\Delta Y] \cdot E[\Delta^2]^{-1}$$

In the previous expression we replaced the outcome with $Y = \theta_0 + \theta_a A + \theta_m M + \epsilon_y^m$:

$$\begin{aligned} \theta_a^* &= E[(\theta_0 + \theta_a A + \theta_m M + \epsilon_y^m) \Delta] E[\Delta^2]^{-1} \\ &= (\theta_0 E[\Delta] + \theta_a E[A \Delta] + \theta_m E[M \Delta] + E[\epsilon_y^m \Delta]) E[\Delta^2]^{-1} \end{aligned}$$

We replace $E[A \Delta]$ with $E[(A - \alpha_0 - \alpha_w W) \Delta]$ as proven in Appendix A4.5 Proof 3.

$$\begin{aligned} \theta_a^* &= (\theta_a E[(A - \alpha_0 - \alpha_w W) \Delta] + \theta_m E[M \Delta] + E[\epsilon_y^m \Delta]) E[\Delta^2]^{-1} \\ &= \theta_a E[\Delta^2] E[\Delta^2]^{-1} + \theta_m E[M \Delta] E[\Delta^2]^{-1} + E[\epsilon_y^m \Delta] E[\Delta^2]^{-1} \\ &= \theta_a + \theta_m (\text{Cov}[M, \Delta] + E[M] E[\Delta]) E[\Delta^2]^{-1} + (\text{Cov}[\Delta, \epsilon_y^m] + E[\Delta] E[\epsilon_y^m]) E[\Delta^2]^{-1} \\ &= \theta_a + \theta_m \text{Cov}[M, \Delta] E[\Delta^2]^{-1} + \text{Cov}[\Delta, \epsilon_y^m] E[\Delta^2]^{-1} \end{aligned}$$

The bias $(\theta_a^* - \theta_a)$ can be quantified as follows:

$$\theta_a^* - \theta_a = \theta_m \frac{\text{Cov}[M, \Delta]}{E[\Delta^2]} + \frac{\text{Cov}[\Delta, \epsilon_y^m]}{E[\Delta^2]} \quad (43)$$

First, we wrote the covariance $\text{Cov}[M, \Delta]$ in term of A and M as follows:

$$\begin{aligned} \theta_m \text{Cov}[M, \Delta] &= \theta_m \text{Cov}[M, (A - \alpha_0 - \alpha_w W)] \\ &= \theta_m \text{Cov}[M, (A - \alpha_0 - \alpha_w(\gamma_0 + \gamma_m M + U))] \\ &= \theta_m \text{Cov}[M, (A - \alpha_0 - \alpha_w \gamma_0 - \alpha_w \gamma_m M - \alpha_w U)] \\ &= \theta_m (\text{Cov}[A, M] - \alpha_w \gamma_m V(M)) \end{aligned}$$

Second, we write the covariance $\text{Cov}[\Delta, \epsilon_y^m]$ as follows.

$$\begin{aligned} \text{Cov}[\Delta, \epsilon_y^m] &= \text{Cov}[(A - \alpha_0 - \alpha_w \gamma_0 - \alpha_w \gamma_m M - \alpha_w U), (Y - \theta_0 - \theta_a A - \theta_m M)] \\ &= \text{Cov}[Y, A] - \alpha_w \gamma_m \text{Cov}[Y, M] - \alpha_w \text{Cov}[Y, U] - \theta_a V(A) + \theta_a \alpha_w \gamma_m \text{Cov}[A, M] + \theta_a \alpha_w \text{Cov}[A, U] - \\ &\quad \theta_m \text{Cov}[M, A] + \theta_m \alpha_w \gamma_m V(M) + \theta_m \alpha_w \text{Cov}[M, U] \\ &= \text{Cov}[Y, A] - \alpha_w \gamma_m \text{Cov}[Y, M] - \theta_a V(A) + \theta_a \alpha_w \gamma_m \text{Cov}[A, M] - \theta_m \text{Cov}[M, A] + \theta_m \alpha_w \gamma_m V(M) \\ &= \text{Cov}[Y, A] - \alpha_w \gamma_m \text{Cov}[Y, M] - \theta_a V(A) + (\theta_a \alpha_w \gamma_m - \theta_m) \text{Cov}[M, A] + \theta_m \alpha_w \gamma_m V(M) \end{aligned}$$

Finally, the expected value $E[\Delta^2]$ can be written as:

$$\begin{aligned}
E[\Delta^2] &= V(\Delta) + E[\Delta]^2 \\
&= V(A - \alpha_0 - \alpha_w W) \\
&= V(A) + \alpha_w^2 V(W) - 2\alpha_w \text{Cov}(A, W) \\
&= V(A) + \alpha_w^2 V(W) - 2\alpha_w \text{Cov}(\alpha_0 + \alpha_w W + \Delta, W) \\
&= V(A) + \alpha_w^2 V(W) - 2\alpha_w^2 V(W) \\
&= V(A) - \alpha_w^2 V(W)
\end{aligned}$$

Now it is possible to substitute the obtained results into the bias formula 43:

$$\begin{aligned}
\theta_a^* - \theta_a &= \theta_m \frac{\text{Cov}[M, \Delta] + \text{Cov}[\Delta, \epsilon_y^m]}{E[\Delta^2]} \\
&= \frac{\theta_m \text{Cov}[A, M] - \theta_m \alpha_w \gamma_m V(M) + \text{Cov}[Y, A] - \alpha_w \gamma_m \text{Cov}[Y, M] - \theta_a V(A) + (\theta_a \alpha_w \gamma_m - \theta_m) \text{Cov}[M, A] + \theta_m \alpha_w \gamma_m V(M)}{V(A) - \alpha_w^2 V(W)} \\
&= \frac{\text{Cov}[Y, A] - \alpha_w \gamma_m \text{Cov}[Y, M] - \theta_a V(A) + \theta_a \alpha_w \gamma_m \text{Cov}[M, A]}{V(A) - \alpha_w^2 V(W)}
\end{aligned}$$

Given that:

$$\text{Cov}[Y, A] = \theta_a V(A) + \theta_m \text{Cov}[A, M];$$

$$(-\alpha_w \gamma_m) \text{Cov}[Y, M] = (-\alpha_w \gamma_m) \theta_a \text{Cov}[A, M] + (-\alpha_w \gamma_m) \theta_m V(M).$$

We have:

$$\begin{aligned}
\theta_a^* - \theta_a &= \frac{\theta_a V(A) + \theta_m \text{Cov}[A, M] - \theta_a \alpha_w \gamma_m \text{Cov}[A, M] - \alpha_w \gamma_m \theta_m V(M) - \theta_a V(A) + \theta_a \alpha_w \gamma_m \text{Cov}[M, A]}{V(A) - \alpha_w^2 V(W)} \\
&= \frac{\theta_m (\text{Cov}[A, M] - \alpha_w \gamma_m V(M))}{V(A) - \alpha_w^2 V(W)}
\end{aligned}$$

Given that :

$$\alpha_w = \frac{\text{Cov}[A, W]}{V(W)}$$

$$\gamma_m = \frac{\text{Cov}[W, M]}{V(M)}$$

We have:

$$\theta_a^* - \theta_a = \theta_m \frac{\text{Cov}[A, M] - \left(\frac{\text{Cov}[A, W]}{V(W)} \right) \left(\frac{\text{Cov}[W, M]}{V(M)} \right) V(M)}{V(A) - \left(\frac{\text{Cov}[A, W]}{V(W)} \right)^2 V(W)} \quad (44)$$

Therefore, the extent of the bias in θ_a depends mainly on the covariances between the treatment A, the mediator M and the proxy W.

Thus, the bias does not only depend on the relationship between the mediator M and the proxy W, but also on the association of the treatment A with these two variables.

From the formula 44, it can be observed that it is not sufficient for the mediator M to be independent of the treatment A in order to eliminate the bias in θ_a . Indeed, even if $\text{Cov}[A, M] = 0$, the bias remains non-zero. Furthermore, even if the association between the proxy and the treatment or between the proxy and the mediator is absent (i.e., $\text{Cov}[W, A] = 0$ or $\text{Cov}[W, M] = 0$), the bias is still non-zero if an association between the treatment A and the mediator M exists, that is $\text{Cov}[A, M] \neq 0$.

In conclusion, for the bias to be null, one of the following two conditions must hold: either W and M are identical variables, or the proxy W is independent of both A and M , and M is also independent of A .

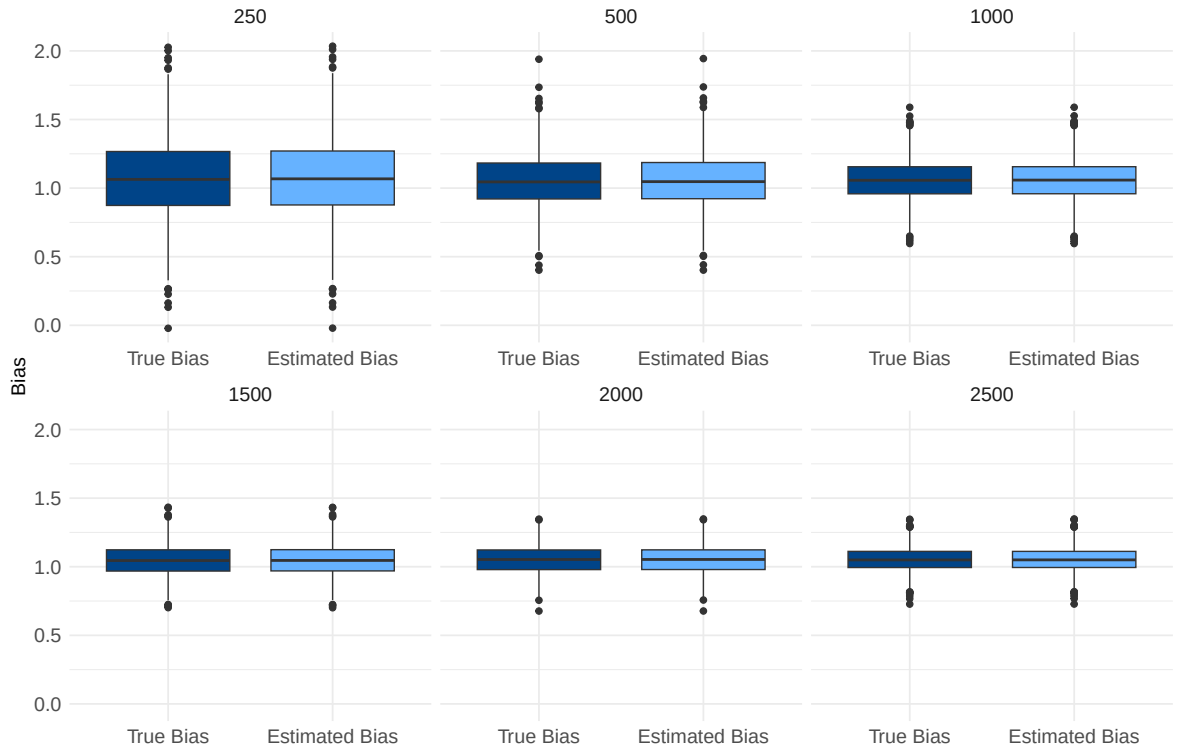


Figure 7: Simulation result for the quantification of bias in the linear case without exposure-mediator interaction. Each boxplot shows the estimates of $\text{bias} = \theta_a^* - \theta_a$ of the four estimators for $N = 250, 500, 1000, 1500, 2000, 2500$ respectively.

The first case, the one in which the proxy W simulates the hidden mediator M perfectly, the bias is null. In this case $M = W$, then $\text{Cov}[A, W] = \text{Cov}[A, M]$, $\text{Cov}[W, M] = \text{Cov}[M, M] = V(M)$ and $V(W) = V(M)$.

$$\theta_a^* - \theta_a = \theta_m \frac{\text{Cov}[A, M] - \left(\frac{\text{Cov}[A, W]}{V(W)} \right) \left(\frac{\text{Cov}[W, M]}{V(M)} \right) V(M)}{V(A) - \left(\frac{\text{Cov}[A, W]}{V(W)} \right)^2 V(W)}$$

$$\begin{aligned}
&= \theta_m \frac{\text{Cov}[A, M] - \left(\frac{\text{Cov}[A, M]}{V(M)} \right) \left(\frac{\text{Cov}[M, M]}{V(M)} \right) V(M)}{V(A) - \left(\frac{\text{Cov}[A, M]}{V(M)} \right)^2 V(M)} \\
&= \theta_m \frac{\text{Cov}[A, M] - \text{Cov}[A, M]}{V(A) - \frac{\text{Cov}[A, M]^2}{V(M)}} = 0
\end{aligned}$$

To show the accuracy of the quantification obtained in formula 44, a simulation was performed. We created 1000 independent and identically distributed samples of the variables (Y, A, M, Z, W), with the sample size N varying among the values 250, 500, 1000, 1500, 2000, 2500. In Figure 7, boxplots are shown depicting the distribution of the true bias and the estimated bias, as previously specified.

It can be observed that the bias can be perfectly expressed as a function of the quantities specified in the formula 44.

Similar results are obtained using Z instead of W:

$$\theta_a^* - \theta_a = \theta_m \frac{\text{Cov}[M, (A - \tau_0 - \tau_z Z)]}{E[(A - \tau_0 - \tau_z Z)^2]} + \frac{\text{Cov}[(A - \tau_0 - \tau_z Z), \epsilon_y^m]}{E[(A - \tau_0 - \tau_z Z)^2]}$$

Where τ_0 and τ_z are the regression coefficients of the regression of Z on A: $E[A | Z] = \tau_0 + \tau_z Z$.

6 Application: Brader, Valentino, and Suhay's Immigration Experiment

To illustrate the proposed estimators in practice, we use the data from the immigration experiment conducted by Brader, Valentino, and Suhay (2008).

Brader et al. hypothesized that U.S. public opinion on immigration is influenced by news content, so they conducted an experiment with a nationally representative sample of 265 white, non-Latino adults.

Participants were told they would read one of several recent newspaper articles, randomly selected for them. The study manipulated news content to emphasize either the benefits or the costs of immigration and varied the immigrant's ethnicity, depicting either a white Russian immigrant or a Hispanic Mexican immigrant. Thus, the treatment is a four-category variable: "News emphasizing benefits of immigration, Russian immigrant", "News emphasizing benefits of immigration, Hispanic immigrant", "News emphasizing costs of immigration, Russian immigrant" and "News emphasizing costs of immigration, Hispanic immigrant". The treatment is randomly assigned to the participants. All stories reported that immigration to the United States is increasing.

The outcome of interest is 'Prefer to reduce immigration,' where higher values indicate greater opposition to immigration. This outcome was considered both as a continuous and a binary variable by

Brader and by us.

Brader et al. posited that respondents' anxiety might be a mediator, altering their attitudes toward immigration, independently of their beliefs about the severity of the issue. They also considered perceived harm as another potential mediator of news influence on immigration opinions.

After reading the stories, participants answered questions about the potential negative financial and social impacts of immigration, which were used to create a continuous perceived harm scale.

To measure anxiety, subjects were asked to complete a post-test consisting of six questions about their feelings towards increased immigration—anxious, proud, angry, hopeful, worried, excited. Responses related to anxiety, worry, and anger were combined into a single continuous scale of anxiety.

Brader et al. found that only one version of treatment, a negative immigration news story featuring a Hispanic immigrant, significantly increased anxiety and reduced support for immigration. Based on the original results, we use the recoded data from Imai et al. (2011), which transformed the four-category treatment into a binary variable. In this recoding, the treatment condition consists of the negative news story combined with the picture of the Hispanic immigrant, while the control condition includes subjects from the other three conditions.

They also controlled for education, age, income, and gender.

In the first analysis, we considered the continuous outcome variable "Prefer to reduce immigration," with values ranging from 1 to 4, where opposition to immigration increases as the score increases.

In the Naive estimator, we used the reported level of anxiety as the mediator. In the proximal and cross-world estimators we used the two mediators from Brader analysis, "perceived harm" and the "measured/reported level of anxiety through the survey", as proxies for what we considered the unobserved underlying true mediator: "true level of anxiety toward immigration."

Our adjusted analyses also accounted for the same four confounding variables: education, age, income, and gender. The results are presented in Table 7 and 8.

It can be observed that for the naive estimator, which uses self-reported anxiety (the proxy) as true level of anxiety (the hidden mediator), both NDEs and NIEs of the article's tone on the inclination to reduce immigration are present (although the NDEs are not significant).

In contrast, using the proximal and cross-world estimators increases the NIEs (both adjusted and unadjusted), while the NDEs are nearly null.

	Naive				Proximal			
	Without Interaction Unadj.	Interaction Adj.	Without Interaction Unadj.	Interaction Adj.	Without Interaction Unadj.	Interaction Adj.	Without Interaction Unadj.	Interaction Adj.
NDE ^a	0.16	0.18	0.14	0.17	0.01	0.04	-0.01	0.02
NDE ^{a'}	0.16	0.18	0.21	0.24*	0.01	0.04	0.07	0.10
NIE ^a	0.28***	0.23***	0.23***	0.18***	0.43***	0.38***	0.37***	0.32***
NIE ^{a'}	0.28***	0.23***	0.30***	0.25***	0.43***	0.38***	0.45***	0.40***
ATE	0.44	0.41	0.44	0.42	0.44	0.42	0.44	0.42

Table 7: Naive and Proximal estimators of causal effects for the linear case with and without exposure-mediator interaction. (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$)

	Cross-world	
	Unadj.	Adj.
NDE ^a	-0.01	0.16
NDE ^{a'}	0.07	-0.05
NIE ^a	0.37***	0.47***
NIE ^{a'}	0.45***	0.25*
ATE	0.44	0.42

Table 8: Cross-world estimators of causal effects for the linear case with and without exposure-mediator interaction. (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$)

This demonstrates that the entire effect of reading news about the benefits or costs of immigration is fully mediated by the presence of anxiety toward immigration in the reader.

The cross-world estimator does not differentiate between cases with or without exposure-mediator interaction because it inherently accounts for the interaction coefficient if present. Given that the results with the proximal estimator including interaction are so similar to those with the cross-world estimator (in the unadjusted case they are exactly the same), it suggests the presence of interaction between exposure and mediator.

In the case of a binary outcome, we used the same observable confounders as in the continuous case, but we considered the outcome to be 0 if the propensity to reduce immigration was 1 or 2, and 1 if the propensity was 3 or 4. The results of binary case are shown in Table 9.

The results differ when considering the binary outcome. In this case, both NDEs and NIEs are statistically significant using both the naive and proximal estimators.

However, with the naive estimator, NDE^{a'} and NIE^{a'} for the untreated group, participants who read an article highlighting either the benefits of immigration or the costs with a Russian immigrant, suggest reduced opposition to immigration. In contrast, NDE^a and NIE^a for the treated group, participants who read an article emphasizing the costs of immigration with a Hispanic protagonist, appear to increase opposition to immigration.

	Naïve				Proximal			
	Without Interaction Unadj.	Interaction Adj.	Without Interaction Unadj.	Interaction Adj.	Without Interaction Unadj.	Interaction Adj.	Without Interaction Unadj.	Interaction Adj.
NDE ^a	1.10***	1.12***	1.07***	1.09***	0.96***	0.98***	0.91***	0.93***
NDE ^{a'}	0.91***	0.89***	0.85***	0.84***	1.05***	1.02***	0.93***	0.93***
NIE ^a	1.17***	1.15***	1.09***	1.07***	1.34***	1.31***	1.19***	1.19***
NIE ^{a'}	0.86***	0.87***	0.84***	0.85***	0.75***	0.76***	0.71***	0.73***
ATE	1.28	1.29	1.27	1.27	1.28	1.28	1.28	1.28

Table 9: Naive and proximal estimators of causal effects considered for the logarithmic case with and without exposure-mediator interaction. (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$)

With the proximal estimator, the results indicate that only NIE^a for the treated group, participants exposed to negative news about a Hispanic immigrant, increases the propensity to oppose immigration, while the other causal effects show a slight reduction or no change in the inclination to oppose it.

Thus, using self-reported anxiety as a mediator suggests that reading news highlighting the costs of Hispanic immigration directly and indirectly increases opposition to immigration. However, applying the proximal estimator, which accounts for true hidden anxiety as mediator, reveals that opposition to immigration is primarily driven by anxiety itself. Specifically, individuals who receive the treatment, reading the negative news on Hispanic immigrant, oppose immigration only if they are subject to an anxiety mechanism. The direct effect of the treatment remains negligible or may even slightly reduce opposition in individuals without anxiety-related issues.

7 Discussion

The innovative proximal approach presented in this thesis allows for the derivation of NDEs and NIEs in the context of mediation analysis, when the mediator is either hidden or mismeasured, and whether or not there is an interaction between the exposure and the mediator, regardless of whether the outcome is continuous, count, binary or polytomous.

We build upon the work of Tchetgen Tchetgen and Ghassami. Specifically, we have used the key concepts from Tchetgen Tchetgen et al. (2023), where they propose the proximal G-computation formula via a two-stage least-squares regression approach to address the presence of unobservable confounder. Their method enables the identification of the causal effect of the treatment on the outcome when hidden confounder is present, provided that two proxies for the unobserved confounder are available.

We extended this framework, initially developed for unobserved confounder, to causal mediation analysis when the mediator is hidden. A solution to identify NIEs and NDEs in the presence of a hidden mediator was previously introduced by Ghassami et al. (2020). However, their method for handling

hidden mediators with exposure-mediator interaction is not based on generalized linear models (GLM) and involves solving certain linear integral equations, which can be computationally complex.

The advantage of the proximal methods introduced in this thesis is twofold: first, they provide unbiased estimators for both the direct and indirect natural effects in the treated and untreated groups, in the presence of a hidden mediator; second, these methods are straightforward to apply, relying on commonly used GLMs. This makes them accessible for epidemiologists, who can apply these methods using widely adopted models. The only requirement is the availability of data on two proxies for the mediator, meeting specific conditional independence assumptions.

Moreover, these methods are not limited to cases where the mediator is unmeasurable but can also be applied when the mediator is measured with error. While various techniques exist to account for measurement error, they often rely on assumptions about the nature of the error. Our approach, however, addresses measurement error without making any assumptions about its characteristics. In fact, the proximal approach does not simply adjust the causal effects obtained with a mismeasured mediator, but instead directly provides unbiased causal effects.

There are several avenues for future research: the next step is to extend this proximal approach in cases with a binary outcome and a logit link. Subsequently, I aim to explore the case of multiple or time-dependent mediators. Finally, I would like to apply this approach to identify the ATE of the exposure on the outcome, in the presence of an unobserved confounder U between the exposure and the outcome, within the context of the Front-Door Criterion. This scenario involves a mediator in the path from the exposure to the outcome, where the mediator is not caused by the unobservable confounder U .

References

- [1] Andrews Ryan M., Vanessa Didelez. “Insights into the Cross-world independence assumption of causal mediation analysis.” *Epidemiology*, vol. 32, no. 2, 30 Dec. 2020, pp. 209–219, <https://doi.org/10.1097/ede.0000000000001313>.
- [2] Baron Reuben M., David A. Kenny. “The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.” *Journal of Personality and Social Psychology*, vol. 51, no. 6, 1986, pp. 1173–1182, <https://doi.org/10.1037/0022-3514.51.6.1173>.
- [3] Brader Ted, Valentino Nicholas A., Suhay Elizabeth. “What triggers public opposition to immigration? Anxiety, group cues, and immigration threat.” *American Journal of Political Science*, vol. 52, no. 4, 29 Sept. 2008, pp. 959–978, <https://doi.org/10.1111/j.1540-5907.2008.00353.x>.
- [4] Carroll Raymond J. “Measurement error in Epidemiologic Studies.” *Encyclopedia of Biostatistics*, 15 Feb. 2005, <https://doi.org/10.1002/0470011815.b2a03082>.
- [5] Carroll Raymond J. “Measurement Error in Nonlinear Models: A Modern Perspective.” Chapman & Hall/CRC, 2006.
- [6] Dukes Oliver, Shpitser Ilya, Tchetgen Tchetgen Eric J. “Proximal mediation analysis.” *Biometrika*, vol. 110, no. 4, 3 Mar. 2023, pp. 973–987, <https://doi.org/10.1093/biomet/asad015>.
- [7] Ghassami Amiremad, Yang Alan, Shpitser Ilya, Tchetgen Tchetgen Eric J. “Causal inference with hidden mediators.” *Biometrika*, 15 July 2024, <https://doi.org/10.1093/biomet/asae037>.
- [8] Greenland Sander, Robins James. “Identifiability, exchangeability, and epidemiological confounding.” *International Journal of Epidemiology*, vol. 15, no. 3, 1986, pp. 413–419, <https://doi.org/10.1093/ije/15.3.413>.
- [9] Hernan Miguel A., James M. Robins. “Causal Inference: What If”. Taylor and Francis, 2024.
- [10] Imai Kosuke, Yamamoto Teppei. “Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments.” *Political Analysis*, vol. 21, no. 2, 2013, pp. 141–171, <https://doi.org/10.1093/pan/mps040>.
- [11] Imai Kosuke, Keele Luke, Tingley Dustin. “A general approach to causal mediation analysis.” *Psychological Methods*, vol. 15, no. 4, 2010, pp. 309–334, <https://doi.org/10.1037/a0020761>.
- [12] Imai Kosuke, Keele Luke, Yamamoto Teppei. “Identification, inference and sensitivity analysis for causal mediation effects.” *Statistical Science*, vol. 25, no. 1, 1 Feb. 2010, <https://doi.org/10.1214/10-sts321>.

- [13] Imai Kosuke, Keele Luke, Tingley Dustin, Yamamoto Teppei. “Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies.” *American Political Science Review*, vol. 105, no. 4, Nov. 2011, pp. 765–789, <https://doi.org/10.1017/s0003055411000414>.
- [14] Le Cessie Saskia, Debeij Jan, Rosendaal Frits R., Cannegieter Suzanne C., Vandenbroucke Jan P. “Quantification of bias in direct effects estimates due to different types of measurement error in the mediator.” *Epidemiology*, vol. 23, no. 4, July 2012, pp. 551–560, <https://doi.org/10.1097/ede.0b013e318254f5de>.
- [15] Lipsitch Marc, Tchetgen Tchetgen Eric J., Cohen Ted. “Negative controls: a tool for detecting confounding and bias in observational studies.” *Epidemiology*, vol. 21, no. 3, May 2010, pp. 383–388, <https://doi.org/10.1097/ede.0b013e3181d61eeb>.
- [16] Miao Wang, Shi Xu, Li Yilin, Tchetgen Tchetgen Eric J. “A confounding bridge approach for double negative control inference on causal effects.” *Statistical Theory and Related Fields*, 30 Aug. 2024, pp. 1–12, <https://doi.org/10.1080/24754269.2024.2390748>.
- [17] Miao Wang, Zhi Geng, Tchetgen Tchetgen Eric J. “Identifying causal effects with proxy variables of an unmeasured confounder.” *Biometrika*, vol. 105, no. 4, 13 Aug. 2018, pp. 987–993, <https://doi.org/10.1093/biomet/asy038>.
- [18] Pearl Judea. “Direct and Indirect Effects.” *Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, 2001, pp. 411–420, https://ftp.cs.ucla.edu/pub/stat_ser/R273-U.pdf.
- [19] Pearl Judea. “The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms.” *Prevention Science*, vol. 13, 2012, pp. 426–436.
- [20] R core team. “R: A Language and Environment for Statistical Computing”, [sdl6x2/documentazione/r/fullrefman.pdf](https://www.R-project.org/doc/manuals/r-sdl6x2/documentazione/r/fullrefman.pdf). Accessed 2008.
- [21] Robins James M., Greenland Sander. “Identifiability and exchangeability for direct and indirect effects.” *Epidemiology*, vol. 3, no. 2, Mar. 1992, pp. 143–155, <https://doi.org/10.1097/00001648-199203000-00013>.
- [22] Saul Bradley C., Hudgens Michael G. “The calculus of M-estimation in r with geex.” *Journal of Statistical Software*, vol. 92, no. 2, 2020, <https://doi.org/10.18637/jss.v092.i02>.
- [23] Shi Xu, Miao Wang, Tchetgen Tchetgen Eric J. “A selective review of negative control methods in epidemiology.” *Current Epidemiology Reports*, vol. 7, no. 4, 15 Oct. 2020, pp. 190–202, <https://doi.org/10.1007/s40471-020-00243-4>.

- [24] Stefanski Leonard, Boos Dennis. “The Calculus of M-Estimation.” *The American Statistician*, 1st ed., vol. 56, 2002, pp. 29–38, <http://www.jstor.org/stable/3087324>.
- [25] Tchetgen Tchetgen Eric J., Ying Andrew, Cui Yifan, Shi Xu, Miao Wang. ”An Introduction to Proximal Causal Learning”, 23 Sept. 2020, <https://doi.org/10.1101/2020.09.21.20198762>.
- [26] Valeri Linda, VanderWeele Tyler J. “Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros.” *Psychological Methods*, vol. 18, no. 2, June 2013, pp. 137–150, <https://doi.org/10.1037/a0031034>.
- [27] Valeri Linda, VanderWeele Tyler J. “The estimation of direct and indirect causal effects in the presence of misclassified binary mediator.” *Biostatistics*, vol. 15, no. 3, 26 Mar. 2014, pp. 498–512, <https://doi.org/10.1093/biostatistics/kxu007>.
- [28] Valeri Linda, Lin Xihong, VanderWeele Tyler J. “Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model.” *Statistics in Medicine*, vol. 33, no. 28, 14 Sept. 2014, pp. 4875–4890, <https://doi.org/10.1002/sim.6295>.
- [29] VanderWeele Tyler J., Valeri Linda, Ogburn Elizabeth L. “The role of measurement error and misclassification in mediation analysis: mediation and measurement error.” *Epidemiology*, vol. 23, no. 4, July 2012, pp. 561–564, <https://doi.org/10.1097/ede.0b013e318258f5e4>.
- [30] Zivich Paul N., Cole Stephen R., Edwards Jessie K., Mulholland Grace E., Shook-Sa Bonnie E., Tchetgen Tchetgen Eric J. “Introducing proximal causal inference for epidemiologists.” *American Journal of Epidemiology*, vol. 192, no. 7, 31 Mar. 2023, pp. 1224–1227, <https://doi.org/10.1093/aje/kwad077>.

Appendix

In Sections A1.1-A1.4, we provide the derivation of NDEs and NIEs when the mediator is observable, using the identification results from Pearl 2001. We examined both dichotomous and continuous outcomes, applying linear and logarithmic links respectively, without and with exposure-mediator interaction, while adjusting for observable confounders X .

In Sections A2.1-A2.5, we present details on the proximal approach for identifying NDEs and NIEs when the mediator is hidden. This analysis includes both continuous and dichotomous outcomes, without and with interaction, and accounts for observable confounders X .

Section A3 addresses statistical inference based on standard M-estimator theory.

Finally, in Section A4.1 and A4.2, we assess the bias introduced by using a proxy for the mediator instead of the mediator itself.

A1 Derivation of NDEs and NIEs with observable mediator

In this chapter, we show how the four causal effects considered are obtained in the case where the mediator is observable. Specifically, we will derive the effects of interest in the four scenarios examined thus far, both with and without exposure-mediator interaction, and for continuous and binary outcomes.

A1.1 NDE and NIE with continuous outcome Y (Identity link) with observable mediator without exposure-mediator interaction

Assuming data are generated from the following model:

Assumption 1A

$$E[Y \mid A, M] = \theta_0 + \theta_a A + \theta_m M + \theta_x X$$

$$E[M \mid A] = \beta_0 + \beta_a A + \beta_x X$$

Now we prove the results in Section 2.1 under Assumption 1A, starting from formulas 8-11 (considering the observable confounder X) to express the four counterfactual means $E[Y(a, M(a))]$, $E[Y(a', M(a))]$, $E[Y(a, M(a'))]$, and $E[Y(a', M(a'))]$, in terms of regression coefficients.

$$\begin{aligned} \text{NDE}^a &= \int_{m,x} [E(Y \mid A = a, M, X) - E(Y \mid A = a', M, X)] p[M \mid A = a, X] p[X] \partial m \partial x \\ &= \int_{m,x} E[Y \mid A = a, M, X] p[M \mid A = a, X] p[X] \partial m \partial x - \int_{m,x} E[Y \mid A = a', M, X] p[M \mid A = a, X] p[X] \partial m \partial x \\ &= E[Y(a, M(a))] - E[Y(a', M(a))] \end{aligned}$$

$$\begin{aligned}
\text{NDE}^{a'} &= \int_{m,x} [\text{E}(Y \mid A = a, M, X) - \text{E}(Y \mid A = a', M, X)] \text{p}[M \mid A = a', X] \text{p}[X] \partial m \partial x \\
&= \int_{m,x} \text{E}[Y \mid A = a, M, X] \text{p}[M \mid A = a', X] \text{p}[X] \partial m \partial x - \int_{m,x} \text{E}[Y \mid A = a', M, X] \text{p}[M \mid A = a', X] \text{p}[X] \partial m \partial x \\
&= \text{E}[Y(a, M(a'))] - \text{E}[Y(a', M(a'))]
\end{aligned}$$

$$\begin{aligned}
\text{NIE}^a &= \int_{m,x} \text{E}[Y \mid A = a, M, X] [\text{p}[M \mid A = a, X] - \text{p}[M \mid A = a', X]] \text{p}[X] \partial m \partial x \\
&= \int_{m,x} \text{E}[Y \mid A = a, M, X] \text{p}[M \mid A = a, X] \text{p}[X] \partial m \partial x - \int_{m,x} \text{E}[Y \mid A = a, M, X] \text{p}[M \mid A = a', X] \text{p}[X] \partial m \partial x \\
&= \text{E}[Y(a, M(a))] - \text{E}[Y(a, M(a'))]
\end{aligned}$$

$$\begin{aligned}
\text{NIE}^{a'} &= \int_{m,x} \text{E}[Y \mid A = a', M, X] [\text{p}[M \mid A = a, X] - \text{p}[M \mid A = a', X]] \text{p}[X] \partial m \partial x \\
&= \int_{m,x} \text{E}[Y \mid A = a', M, X] \text{p}[M \mid A = a, X] \text{p}[X] \partial m \partial x - \int_{m,x} \text{E}[Y \mid A = a', M, X] \text{p}[M \mid A = a', X] \text{p}[X] \partial m \partial x \\
&= \text{E}[Y(a', M(a))] - \text{E}[Y(a', M(a'))]
\end{aligned}$$

$$\begin{aligned}
\text{E}[Y(a, M(a))] &= \int_{m,x} \text{E}[Y \mid A = a, M, X] \text{p}[M \mid A = a, X] \text{p}[X] \partial m \partial x = \\
&\int_{m,x} (\theta_0 + \theta_a + \theta_m M + \theta_x X) \text{p}[M \mid A = a, X] \text{p}[X] \partial m \partial x = \\
&\theta_0 + \theta_a + \theta_m \int_{m,x} M \text{p}[M \mid A = a, X] \text{p}[X] \partial m \partial x + \theta_x \int_x X \text{p}[X] \partial x = \\
&\theta_0 + \theta_a + \theta_m \int_x \text{E}[M \mid A = a, X] \text{p}[X] \partial x + \theta_x \text{E}[X] = \\
&\theta_0 + \theta_a + \theta_m (\beta_0 + \beta_a + \beta_x \int_x X \text{p}[X] \partial x) + \theta_x \text{E}[X] = \\
&\theta_0 + \theta_a + \theta_m (\beta_0 + \beta_a + \beta_x \text{E}[X]) + \theta_x \text{E}[X] = \\
&\theta_0 + \theta_a + \theta_m \beta_0 + \theta_m \beta_a + (\theta_m \beta_x + \theta_x) \text{E}[X]
\end{aligned}$$

$$\begin{aligned}
\text{E}[Y(a', M(a))] &= \int_{m,x} \text{E}[Y \mid A = a', M, X] \text{p}[M \mid A = a, X] \text{p}[X] \partial m \partial x = \\
&\int_{m,x} (\theta_0 + \theta_m M + \theta_x X) \text{p}[M \mid A = a, X] \text{p}[X] \partial m \partial x = \\
&\theta_0 + \theta_m \int_{m,x} M \text{p}[M \mid A = a, X] \text{p}[X] \partial m \partial x + \theta_x \int_x X \text{p}[X] \partial x = \\
&\theta_0 + \theta_m \int_x \text{E}[M \mid A = a, X] \text{p}[X] \partial x + \theta_x \int_x X \text{p}[X] \partial x = \\
&\theta_0 + \theta_m (\beta_0 + \beta_a + \beta_x \int_x X \text{p}[X] \partial x) + \theta_x \text{E}[X] = \\
&\theta_0 + \theta_m \beta_0 + \theta_m \beta_a + (\theta_m \beta_x + \theta_x) \text{E}[X]
\end{aligned}$$

$$\begin{aligned}
\text{E}[Y(a, M(a'))] &= \int_{m,x} \text{E}[Y \mid A = a, M, X] \text{p}[M \mid A = a', X] \text{p}[X] \partial m \partial x = \\
&\int_{m,x} (\theta_0 + \theta_a + \theta_m M + \theta_x X) \text{p}[M \mid A = a', X] \text{p}[X] \partial m \partial x = \\
&\theta_0 + \theta_a + \theta_m \int_{m,x} \text{p}[M \mid A = a', X] \text{p}[X] \partial m \partial x + \theta_x \int_x X \text{p}[X] \partial x = \\
&\theta_0 + \theta_a + \theta_m \int_x \text{E}[M \mid A = a', X] \text{p}[X] \partial x + \theta_x \text{E}[X] = \\
&\theta_0 + \theta_a + \theta_m (\beta_0 + \beta_x \int_x X \text{p}[X] \partial x) + \theta_x \text{E}[X] = \\
&\theta_0 + \theta_a + \theta_m \beta_0 + \theta_m \beta_x \text{E}[X] + \theta_x \text{E}[X] = \\
&\theta_0 + \theta_a + \theta_m \beta_0 + (\theta_m \beta_x + \theta_x) \text{E}[X]
\end{aligned}$$

$$\begin{aligned}
\text{E}[Y(a', M(a'))] &= \int_{m,x} \text{E}[Y \mid A = a', M, X] \text{p}[M \mid A = a', X] \text{p}[X] \partial m \partial x = \\
&\int_{m,x} (\theta_0 + \theta_m M + \theta_x X) \text{p}[M \mid A = a', X] \text{p}[X] \partial m \partial x =
\end{aligned}$$

$$\begin{aligned}
& \theta_0 + \theta_m \int_{m,x} M p[M | A = a', X] p[X] \partial m \partial x + \theta_x \int_x X p[X] \partial x = \\
& \theta_0 + \theta_m \int_x E[M | A = a', X] p[X] \partial x + \theta_x \int_x X p[X] \partial x = \\
& \theta_0 + \theta_m (\beta_0 + \beta_x \int_x X p[X] \partial x) + \theta_x E[X] = \\
& \theta_0 + \theta_m \beta_0 + (\theta_m \beta_x + \theta_x) E[X]
\end{aligned}$$

Utilizing the counterfactual means just identified, we can determine the four causal effects:

$$\begin{aligned}
\text{NDE}^a &= E[Y(a, M(a)) - Y(a', M(a))] = \\
& (\theta_0 + \theta_a + \theta_m \beta_0 + \theta_m \beta_a + (\theta_m \beta_x + \theta_x) E[X]) - (\theta_0 + \theta_m \beta_0 + \theta_m \beta_a + (\theta_m \beta_x + \theta_x) E[X]) = \theta_a
\end{aligned}$$

$$\begin{aligned}
\text{NDE}^{a'} &= E[Y(a, M(a')) - Y(a', M(a'))] = \\
& (\theta_0 + \theta_a + \theta_m \beta_0 + (\theta_m \beta_x + \theta_x) E[X]) - (\theta_0 + \theta_m \beta_0 + (\theta_m \beta_x + \theta_x) E[X]) = \theta_a
\end{aligned}$$

$$\begin{aligned}
\text{NIE}^a &= E[Y(a, M(a)) - Y(a, M(a'))] = \\
& (\theta_0 + \theta_a + \theta_m \beta_0 + \theta_m \beta_a + (\theta_m \beta_x + \theta_x) E[X]) - (\theta_0 + \theta_a + \theta_m \beta_0 + (\theta_m \beta_x + \theta_x) E[X]) = \theta_m \beta_a
\end{aligned}$$

$$\begin{aligned}
\text{NIE}^{a'} &= E[Y(a', M(a)) - Y(a', M(a'))] = \\
& (\theta_0 + \theta_m \beta_0 + \theta_m \beta_a + (\theta_m \beta_x + \theta_x) E[X]) - (\theta_0 + \theta_m \beta_0 + (\theta_m \beta_x + \theta_x) E[X]) = \theta_m \beta_a
\end{aligned}$$

Note that, in the absence of a treatment-mediator interaction, the NDEs and NIEs for treated and untreated groups are the same.

A1.2 NDE and NIE with continuous outcome Y (Identity link) with observable mediator with exposure-mediator interaction

Assuming data are generated from the following model:

Assumption 2A

$$E[Y|A, M] = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM + \theta_x X$$

$$E[M|A] = \beta_0 + \beta_a A + \beta_x X$$

Now we prove the results in Section 2.1 under Assumption 2A, starting from formulas 8-11 (considering the observable confounder X) to express the four counterfactual means $E[Y(a, M(a))]$, $E[Y(a', M(a))]$, $E[Y(a, M(a'))]$, and $E[Y(a', M(a'))]$, in terms of regression coefficients.

$$\begin{aligned}
E[Y(a, M(a))] &= \int_{m,x} E[Y|A = a, M, X] p[M|A = a, X] p[X] \partial m \partial x = \\
& \int_{m,x} (\theta_0 + \theta_a + \theta_m M + \theta_{am} M + \theta_x X) p[M|A = a, X] p[X] \partial m \partial x = \\
& \theta_0 + \theta_a + (\theta_m + \theta_{am}) \int_{m,x} M p[M|A = a, X] p[X] \partial m \partial x + \theta_x \int_x X p[X] \partial x = \\
& \theta_0 + \theta_a + (\theta_m + \theta_{am}) \int_x E[M|A = a, X] p[X] \partial x + \theta_x E[X] = \\
& \theta_0 + \theta_a + (\theta_m + \theta_{am}) (\beta_0 + \beta_a + \beta_x \int_x X p[X] \partial x) + \theta_x E[X] =
\end{aligned}$$

$$\theta_0 + \theta_a + \theta_m\beta_0 + \theta_{am}\beta_0 + \theta_m\beta_a + \theta_{am}\beta_a + (\theta_m\beta_x + \theta_{am}\beta_x + \theta_x) E[X]$$

$$\begin{aligned} E[Y(a', M(a))] &= \int_{m,x} E[Y|A = a', M, X] p[M|A = a, X] p[X] \partial m \partial x = \\ &= \int_{m,x} (\theta_0 + \theta_m M + \theta_x X) p[M|A = a, X] p[X] \partial m \partial x = \\ &= \theta_0 + \theta_m \int_{m,x} M p[M|A = a, X] p[X] \partial m \partial x + \theta_x \int_x X p[X] \partial x = \\ &= \theta_0 + \theta_m + \int_x E[M|A = a, X] p[X] \partial x + \theta_x E[X] = \\ &= \theta_0 + \theta_m + (\beta_0 + \beta_a + \beta_x \int_x X p[X] \partial x) + \theta_x E[X] = \\ &= \theta_0 + \theta_m\beta_0 + \theta_m\beta_a + \theta_m\beta_x E[X] + \theta_x E[X] = \\ &= \theta_0 + \theta_m\beta_0 + \theta_m\beta_a + (\theta_m\beta_x + \theta_x) E[X] \end{aligned}$$

$$\begin{aligned} E[Y(a, M(a'))] &= \int_{m,x} E[Y|A = a, M, X] p[M|A = a', X] p[X] \partial m \partial x = \\ &= \int_{m,x} (\theta_0 + \theta_a + \theta_m M + \theta_{am} M + \theta_x X) p[M|A = a', X] p[X] \partial m \partial x = \\ &= \theta_0 + \theta_a + (\theta_m + \theta_{am}) \int_{m,x} M p[M|A = a', X] p[X] \partial m \partial x + \theta_x \int_x X p[X] \partial x = \\ &= \theta_0 + \theta_a + (\theta_m + \theta_{am}) \int_x E[M|A = a', X] p[X] \partial x + \theta_x E[X] = \\ &= \theta_0 + \theta_a + (\theta_m + \theta_{am}) (\beta_0 + \beta_x \int_x X p[X] \partial x) + \theta_x E[X] = \\ &= \theta_0 + \theta_a + (\theta_m + \theta_{am}) (\beta_0 + \beta_x E[X]) + \theta_x E[X] = \\ &= \theta_0 + \theta_a + \theta_m\beta_0 + \theta_{am}\beta_0 + \theta_m\beta_x E[X] + \theta_{am}\beta_x E[X] + \theta_x E[X] = \\ &= \theta_0 + \theta_a + \theta_m\beta_0 + \theta_{am}\beta_0 + (\theta_m\beta_x + \theta_{am}\beta_x + \theta_x) E[X] \end{aligned}$$

$$\begin{aligned} E[Y(a', M(a'))] &= \int_{m,x} E[Y|A = a', M, X] p[M|A = a', X] p[X] \partial m \partial x = \\ &= \int_{m,x} (\theta_0 + \theta_m M + \theta_x X) p[M|A = a', X] p[X] \partial m \partial x = \\ &= \theta_0 + \theta_m \int_{m,x} M p[M|A = a', X] p[X] \partial m \partial x + \theta_x \int_m X p[X] \partial x = \\ &= \theta_0 + \theta_m \int_x E[M|A = a', X] p[X] \partial x + \theta_x E[X] = \\ &= \theta_0 + \theta_m (\beta_0 + \beta_x \int_m X p[X] \partial x) + \theta_x E[X] = \\ &= \theta_0 + \theta_m\beta_0 + \theta_m\beta_x E[X] + \theta_x E[X] = \\ &= \theta_0 + \theta_m\beta_0 + (\theta_m\beta_x + \theta_x) E[X] \end{aligned}$$

Utilizing the counterfactual means just identified, we can determine the four causal effects:

$$\begin{aligned} NDE^a &= E[Y(a, M(a)) - Y(a', M(a))] = \\ &= (\theta_0 + \theta_a + \theta_m\beta_0 + \theta_{am}\beta_0 + \theta_m\beta_a + \theta_{am}\beta_a + (\theta_m\beta_x + \theta_{am}\beta_x + \theta_x) E[X]) - (\theta_0 + \theta_m\beta_0 + \theta_m\beta_a + (\theta_m\beta_x + \theta_x) E[X]) = \\ &= \theta_a + \theta_{am} (\beta_0 + \beta_a + \beta_x E[X]) \end{aligned}$$

$$\begin{aligned} NDE^{a'} &= E[Y(a, M(a')) - Y(a', M(a'))] = \\ &= (\theta_0 + \theta_a + \theta_m\beta_0 + \theta_{am}\beta_0 + (\theta_m\beta_x + \theta_{am}\beta_x + \theta_x) E[X]) - (\theta_0 + (\theta_m\beta_0 + (\theta_m\beta_x + \theta_x) E[X])) = \\ &= \theta_a + \theta_{am} (\beta_0 + \beta_x E[X]) \end{aligned}$$

$$NIE^a = E[Y(a, M(a)) - Y(a, M(a'))] =$$

$$(\theta_0 + \theta_a + \theta_m \beta_0 + \theta_{am} \beta_0 + \theta_m \beta_a + \theta_{am} \beta_a + (\theta_m \beta_x + \theta_{am} \beta_x + \theta_x) E[X]) - (\theta_0 + \theta_a + (\theta_m \beta_0 + \theta_{am} \beta_0 + (\theta_m \beta_x + \theta_{am} \beta_x + \theta_x) E[X])) = \beta_a (\theta_m + \theta_{am})$$

$$\begin{aligned} \text{NIE}^{(a')} &= E[Y(a', M(a)) - Y(a', M(a'))] = \\ &(\theta_0 + \theta_m \beta_0 + \theta_m \beta_a + (\theta_m \beta_x + \theta_x) E[X]) - (\theta_0 + (\theta_m \beta_0 + (\theta_m \beta_x + \theta_x) E[X])) = \theta_m \beta_a \end{aligned}$$

Note that, in this case with treatment-mediator interaction, the NDEs and NIEs for treated and untreated groups are different.

A1.3 NDE and NIE with count, binary or polytomous outcome Y (Log link) with observable mediator without exposure-mediator interaction

Assuming data are generated from the following model:

Assumption 3A

$$\log(E[Y \mid A, M, X]) = \theta_0 + \theta_a A + \theta_m M + \theta_x X$$

$$E[M \mid A, X] = \beta_0 + \beta_a A + \beta_x X$$

$$M = E[M \mid A, X] + \varepsilon_M; \quad E[\varepsilon_M] = 0; \quad \varepsilon_M \perp A, X$$

$$M_{\varepsilon_M}(\cdot) \text{ exists for any real value } \cdot.$$

Now we prove the results in Section 2.2 under Assumption 3A, starting from formulas 8-11 (considering the observable confounder X) to express the four counterfactual means $E[Y(a, M(a))]$, $E[Y(a', M(a))]$, $E[Y(a, M(a'))]$, and $E[Y(a', M(a'))]$, in terms of regression coefficients.

We use $M_{\varepsilon_M}(\cdot)$ to denote the moment generating function (MGF) of ε_M with input $\cdot$. If $\varepsilon_M \sim N(0, \sigma_\varepsilon^2)$ and t is a constant, then $M_{\varepsilon_M}(t) = e^{-\frac{1}{2}\sigma_\varepsilon^2 t^2}$.

$$\begin{aligned} E[Y(a, M(a))] &= \int_{m,x} E[Y \mid A = a, M, X] p[M \mid A = a, X] p[X] \partial m \partial x \\ &= \int_{m,x} \exp(\theta_0 + \theta_a + \theta_m M + \theta_x X) p[M \mid A = a, X] p[X] \partial m \partial x \\ &= \exp(\theta_0 + \theta_a + \theta_x \int_x X p[X] \partial x) \int_{m,x} \exp(\theta_m M) p[M \mid A = a, X] p[X] \partial m \partial x \\ &= \exp(\theta_0 + \theta_a + \theta_x E[X]) \int_x E[\exp(\theta_m M) \mid A = a, X] p[X] \partial x \\ &= \exp(\theta_0 + \theta_a + \theta_x E[X]) E[\exp(\theta_m (\beta_0 + \beta_a + \beta_x \int_x X p[X] \partial x + \varepsilon_M))] \\ &= \exp(\theta_0 + \theta_a + \theta_x E[X]) \exp(\theta_m \beta_0 + \theta_m \beta_a + \theta_m \beta_x E[X]) E[\exp(\theta_m \varepsilon_M)] \\ &= \exp(\theta_0 + \theta_a + \theta_m \beta_0 + \theta_m \beta_a + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m) \end{aligned}$$

$$E[Y(a', M(a))] = \int_{m,x} E[Y \mid A = a', M, X] p[M \mid A = a, X] p[X] \partial m \partial x$$

$$\begin{aligned}
&= \int_{m,x} \exp(\theta_0 + \theta_m M + \theta_x X) p[M|A = a, X] p[X] \partial m \partial x \\
&= \exp(\theta_0 + \theta_x \int_x X p[X] \partial x) \int_{m,x} \exp(\theta_m M) p[M|A = a, X] p[X] \partial m \partial x \\
&= \exp(\theta_0 + \theta_x E[X]) \int_x E[\exp(\theta_m M)|A = a, X] p[X] \partial x \\
&= \exp(\theta_0 + \theta_x E[X]) E[\exp(\theta_m(\beta_0 + \beta_a + \beta_x \int_x X p[X] \partial x + \varepsilon_M))] \\
&= \exp(\theta_0 + \theta_x E[X]) \exp(\theta_m \beta_0 + \theta_m \beta_a + \theta_m \beta_x E[X]) E[\exp(\theta_m \varepsilon_M)] \\
&= \exp(\theta_0 + \theta_m \beta_0 + \theta_m \beta_a + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)
\end{aligned}$$

$$\begin{aligned}
E[Y(a, M(a'))] &= \int_{m,x} E[Y|A = a, M, X] p[M|A = a', X] p[X] \partial m \partial x \\
&= \int_{m,x} \exp(\theta_0 + \theta_a + \theta_m M + \theta_x X) p[M|A = a', X] p[X] \partial m \partial x \\
&= \exp(\theta_0 + \theta_a + \theta_x \int_x X p[X] \partial x) \int_{m,x} \exp(\theta_m M) p[M|A = a', X] p[X] \partial m \partial x \\
&= \exp(\theta_0 + \theta_a + \theta_x E[X]) \int_x E[\exp(\theta_m M)|A = a', X] p[X] \partial x \\
&= \exp(\theta_0 + \theta_a + \theta_x E[X]) E[\exp(\theta_m(\beta_0 + \beta_x \int_x X p[X] \partial x + \varepsilon_M))] \\
&= \exp(\theta_0 + \theta_a + \theta_x E[X]) \exp(\theta_m \beta_0 + \theta_m \beta_x E[X]) E[\exp(\theta_m \varepsilon_M)] \\
&= \exp(\theta_0 + \theta_a + \theta_m \beta_0 + (\theta_m \beta_x + \theta_x) E[X]) M_{\varepsilon_M}(\theta_m)
\end{aligned}$$

$$\begin{aligned}
E[Y(a', M(a'))] &= \int_{m,x} E[Y|A = a', M, X] p[M|A = a', X] p[X] \partial m \partial x \\
&= \int_{m,x} \exp(\theta_0 + \theta_m M + \theta_x X) p[M|A = a', X] p[X] \partial m \partial x \\
&= \exp(\theta_0 + \theta_x \int_x X p[X] \partial x) \int_{m,x} \exp(\theta_m M) p[M|A = a', X] p[X] \partial m \partial x \\
&= \exp(\theta_0 + \theta_x E[X]) \int_x E[\exp(\theta_m M)|A = a', X] p[X] \partial x \\
&= \exp(\theta_0 + \theta_x E[X]) E[\exp(\theta_m(\beta_0 + \beta_x \int_x X p[X] \partial x + \varepsilon_M))] \\
&= \exp(\theta_0 + \theta_x E[X]) \exp(\theta_m \beta_0 + \theta_m \beta_x E[X]) E[\exp(\theta_m \varepsilon_M)] \\
&= \exp(\theta_0 + \theta_m \beta_0 + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)
\end{aligned}$$

Utilizing the counterfactual means just identified, we can determine the four causal effects:

$$\begin{aligned}
\text{NDE}^a &= \frac{E[Y(a, M(a))]}{E[Y(a', M(a))]} = \frac{\exp(\theta_0 + \theta_a + \theta_m \beta_0 + \theta_m \beta_a + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)}{\exp(\theta_0 + \theta_m \beta_0 + \theta_m \beta_a + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)} = \exp(\theta_a) \\
\text{NDE}^{(a')} &= \frac{E[Y(a', M(a'))]}{E[Y(a, M(a'))]} = \frac{\exp(\theta_0 + \theta_m \beta_0 + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)}{\exp(\theta_0 + \theta_a + \theta_m \beta_0 + (\theta_m \beta_x + \theta_x) E[X]) M_{\varepsilon_M}(\theta_m)} = \frac{1}{\exp(\theta_a)} \\
\text{NIE}^a &= \frac{E[Y(a, M(a))]}{E[Y(a, M(a'))]} = \frac{\exp(\theta_0 + \theta_a + \theta_m \beta_0 + \theta_m \beta_a + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)}{\exp(\theta_0 + \theta_a + \theta_m \beta_0 + (\theta_m \beta_x + \theta_x) E[X]) M_{\varepsilon_M}(\theta_m)} = \exp(\theta_m \beta_a) \\
\text{NIE}^{(a')} &= \frac{E[Y(a', M(a'))]}{E[Y(a', M(a))]} = \frac{\exp(\theta_0 + \theta_m \beta_0 + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)}{\exp(\theta_0 + \theta_m \beta_0 + \theta_m \beta_a + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)} = \frac{1}{\exp(\theta_m \beta_a)}
\end{aligned}$$

Note that, in the absence of a treatment-mediator interaction, the NDEs and NIEs for treated and untreated groups are inverses of each other. Furthermore, although the MGF appears in the expected values, it simplifies when taking the ratio of the expected values and is ultimately absent in the final formula for NDEs and NIEs.

A1.4 NDE and NIE with count, binary or polytomous outcome Y (Log link) with observable mediator with exposure-mediator interaction

Assuming data are generated from the following model:

Assumption 4A

$$\log(E[Y | A, M, X]) = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM + \theta_x X$$

$$E[M | A, X] = \beta_0 + \beta_a A + \beta_x X$$

$$M = E[M | A, X] + \varepsilon_M; \quad E[\varepsilon_M] = 0; \quad \varepsilon_M \perp A, X$$

Now we prove the results in Section 2.2 under Assumption 4A, starting from formulas 8-11 (considering the observable confounder X) to express the four counterfactual means $E[Y(a, M(a))]$, $E[Y(a', M(a))]$, $E[Y(a, M(a'))]$, and $E[Y(a', M(a'))]$, in terms of regression coefficients.

$$\begin{aligned} E[Y(a, M(a))] &= \int_{m,x} E[Y|A=a, M, X] p[M|A=a, X] p[X] \partial m \partial x \\ &= \int_{m,x} \exp(\theta_0 + \theta_a + \theta_m M + \theta_{am} M + \theta_x X) p[M|A=a, X] p[X] \partial m \partial x \\ &= \exp(\theta_0 + \theta_a + \theta_x \int_x X p[X] \partial x) \int_{m,x} \exp((\theta_m + \theta_{am})M) p[M|A=a, X] p[X] \partial m \partial x \\ &= \exp(\theta_0 + \theta_a + \theta_x E[X]) \int_x E[\exp(\theta_m + \theta_{am})M|A=a, X] p[X] \partial x \\ &= \exp(\theta_0 + \theta_a + \theta_x E[X]) E[\exp((\theta_m + \theta_{am})(\beta_0 + \beta_a + \beta_x \int_x X p[X] \partial x + \varepsilon_M))] \\ &= \exp(\theta_0 + \theta_a + \theta_x E[X]) \exp(\theta_m \beta_0 + \theta_m \beta_a + \theta_m \beta_x E[X] + \theta_{am} \beta_0 + \theta_{am} \beta_a + \theta_{am} \beta_x E[X]) E[\exp((\theta_m + \theta_{am})\varepsilon_M)] \\ &= \exp(\theta_0 + \theta_a + \theta_m \beta_0 + \theta_m \beta_a + \theta_{am} \beta_0 + \theta_{am} \beta_a + (\theta_x + \theta_m \beta_x + \theta_{am} \beta_x) E[X]) M_{\varepsilon_M}(\theta_m + \theta_{am}). \end{aligned}$$

$$\begin{aligned} E[Y(a', M(a))] &= \int_{m,x} E[Y|A=a', M, X] p[M|A=a, X] p[X] \partial m \partial x \\ &= \int_{m,x} \exp(\theta_0 + \theta_m M + \theta_x X) p[M|A=a, X] p[X] \partial m \partial x \\ &= \exp(\theta_0 + \theta_x \int_x X p[X] \partial x) \int_{m,x} \exp(\theta_m M) p[M|A=a, X] p[X] \partial m \partial x \\ &= \exp(\theta_0 + \theta_x E[X]) \int_x E[\exp(\theta_m)M|A=a, X] p[X] \partial x \\ &= \exp(\theta_0 + \theta_x E[X]) E[\exp((\theta_m)(\beta_0 + \beta_a + \beta_x \int_x X p[X] \partial x + \varepsilon_M))] \\ &= \exp(\theta_0 + \theta_x E[X]) \exp(\theta_m \beta_0 + \theta_m \beta_a + \theta_m \beta_x E[X]) E[\exp(\theta_m \varepsilon_M)] \\ &= \exp(\theta_0 + \theta_m \beta_0 + \theta_m \beta_a + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m). \end{aligned}$$

$$\begin{aligned} E[Y(a, M(a'))] &= \int_{m,x} E[Y|A=a, M, X] p[M|A=a', X] p[X] \partial m \partial x \\ &= \int_{m,x} \exp(\theta_0 + \theta_a + \theta_m M + \theta_{am} M + \theta_x X) p[M|A=a', X] p[X] \partial m \partial x \\ &= \exp(\theta_0 + \theta_a + \theta_x \int_x X p[X] \partial x) \int_{m,x} \exp((\theta_m + \theta_{am})M) p[M|A=a', X] p[X] \partial m \partial x \\ &= \exp(\theta_0 + \theta_a + \theta_x E[X]) \int_x E[\exp(\theta_m + \theta_{am})M|A=a', X] p[X] \partial x \\ &= \exp(\theta_0 + \theta_a + \theta_x E[X]) E[\exp((\theta_m + \theta_{am})(\beta_0 + \beta_x \int_x X p[X] \partial x + \varepsilon_M))] \\ &= \exp(\theta_0 + \theta_a + \theta_x E[X]) \exp(\theta_m \beta_0 + \theta_{am} \beta_0 + \theta_m \beta_x E[X] + \theta_{am} \beta_x E[X]) E[\exp((\theta_m + \theta_{am})\varepsilon_M)] \end{aligned}$$

$$= \exp(\theta_0 + \theta_a + \theta_m \beta_0 + \theta_{am} \beta_0 + (\theta_x + \theta_m \beta_x + \theta_{am} \beta_x) E[X]) M_{\varepsilon_M}(\theta_m + \theta_{am}).$$

$$\begin{aligned} E[Y(a', M(a'))] &= \int_{m,x} E[Y|A = a', M, X] p[M|A = a', X] p[X] \partial m \partial x \\ &= \int_{m,x} \exp(\theta_0 + \theta_m M + \theta_x X) p[M|A = a', X] p[X] \partial m \partial x \\ &= \exp(\theta_0 + \theta_x \int_x X p[X] \partial x) \int_{m,x} \exp(\theta_m M) p[M|A = a', X] p[X] \partial m \partial x \\ &= \exp(\theta_0 + \theta_x E[X]) \int_x E[\exp(\theta_m M)|A = a', X] p[X] \partial x \\ &= \exp(\theta_0 + \theta_x E[X]) E[\exp((\theta_m)(\beta_0 + \beta_x \int_x X p[X] \partial x + \varepsilon_M))] \\ &= \exp(\theta_0 + \theta_x E[X]) \exp(\theta_m \beta_0 + \theta_m \beta_x E[X]) E[\exp(\theta_m \varepsilon_M)] \\ &= \exp(\theta_0 + \theta_m \beta_0 + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m). \end{aligned}$$

Utilizing the counterfactual means just identified, we can determine the four causal effects:

$$\begin{aligned} \text{NDE}^a &= \frac{E[Y(a, M(a))]}{E[Y(a', M(a))]} = \\ &= \frac{\exp(\theta_0 + \theta_a + \theta_m \beta_0 + \theta_m \beta_a + \theta_{am} \beta_0 + \theta_{am} \beta_a + (\theta_x + \theta_m \beta_x + \theta_{am} \beta_x) E[X]) M_{\varepsilon_M}(\theta_m + \theta_{am})}{\exp(\theta_0 + \theta_m \beta_0 + \theta_m \beta_a + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)} \\ &= \exp(\theta_a + \theta_{am}(\beta_0 + \beta_a + \beta_x E[X])) \frac{M_{\varepsilon_M}(\theta_m + \theta_{am})}{M_{\varepsilon_M}(\theta_m)}. \end{aligned}$$

$$\begin{aligned} \text{NDE}^{a'} &= \frac{E[Y(a', M(a'))]}{E[Y(a, M(a))]} = \\ &= \frac{\exp(\theta_0 + \theta_m \beta_0 + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)}{\exp(\theta_0 + \theta_a + \theta_m \beta_0 + \theta_{am} \beta_0 + (\theta_x + \theta_m \beta_x + \theta_{am} \beta_x) E[X]) M_{\varepsilon_M}(\theta_m + \theta_{am})} \\ &= \frac{1}{\exp(\theta_a + \theta_{am}(\beta_0 + \beta_x E[X]))} \frac{M_{\varepsilon_M}(\theta_m)}{M_{\varepsilon_M}(\theta_m + \theta_{am})}. \end{aligned}$$

$$\begin{aligned} \text{NIE}^a &= \frac{E[Y(a, M(a))]}{E[Y(a, M(a'))]} = \\ &= \frac{\exp(\theta_0 + \theta_a + \theta_m \beta_0 + \theta_m \beta_a + \theta_{am} \beta_0 + \theta_{am} \beta_a + (\theta_x + \theta_m \beta_x + \theta_{am} \beta_x) E[X]) M_{\varepsilon_M}(\theta_m + \theta_{am})}{\exp(\theta_0 + \theta_a + \theta_m \beta_0 + \theta_{am} \beta_0 + (\theta_x + \theta_m \beta_x + \theta_{am} \beta_x) E[X]) M_{\varepsilon_M}(\theta_m + \theta_{am})} \\ &= \exp(\beta_a(\theta_m + \theta_{am})). \end{aligned}$$

$$\begin{aligned} \text{NIE}^{a'} &= \frac{E[Y(a', M(a'))]}{E[Y(a', M(a))]} = \\ &= \frac{\exp(\theta_0 + \theta_m \beta_0 + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)}{\exp(\theta_0 + \theta_m \beta_0 + \theta_m \beta_a + (\theta_x + \theta_m \beta_x) E[X]) M_{\varepsilon_M}(\theta_m)} = \frac{1}{\exp(\theta_m \beta_a)}. \end{aligned}$$

Note that, in the presence of a treatment-mediator interaction, the NDEs and NIEs for treated and untreated groups are different. Furthermore, the MGF ratio is present only in the NDEs.

A2 Derivation of NDEs and NIEs with hidden mediator: proximal framework

A2.1 NDE and NIE with continuous outcome Y (Identity link) with hidden mediator without exposure-mediator interaction

We investigate the case with continuous Y with an Identity link and without exposure-mediator interaction. We assume that the data are generated from the following models:

Assumption 5A:

$$E[Y \mid A, Z, M, X] = \theta_0 + \theta_a A + \theta_m M + \theta_x X$$

$$E[W \mid A, Z, M, X] = \alpha_0 + \alpha_m M + \alpha_x X$$

$$E[M \mid A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

We now prove Result 1 in section 3.1 under Assumption 5A.

Initially, we derive the conditional expected value of W:

$$E[W \mid A, Z, X] = \omega_0 + \omega_a A + \omega_z Z + \omega_x X$$

Then, we take the conditional expectation on both sides of the first two data generation equations for M conditional on (A, Z, X):

$$E[Y \mid A, Z, M, X] = \theta_0 + \theta_a A + \theta_m M + \theta_x X$$

$$E[Y \mid A, Z, X] = \theta_0 + \theta_a A + \theta_m E[M \mid A, Z, X] + \theta_x X$$

$$E[W \mid A, Z, M, X] = \alpha_0 + \alpha_m M + \alpha_x X$$

$$E[W \mid A, Z, X] = \alpha_0 + \alpha_m E[M \mid A, Z, X] + \alpha_x X$$

Thus, we rewrite the conditional mean of M in terms of the conditional mean of W:

$$E[M \mid A, Z, X] = \frac{E[W \mid A, Z, X] - \alpha_0 - \alpha_x X}{\alpha_m}$$

Using the law of iterated expectations, we can rewrite the third data generation equation as follows:

$$E[E[M \mid A, Z, X] \mid A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

We substitute the conditional expectation of M with its version as a function of E[W | A, Z, X]:

$$E[Y \mid A, Z, X] = \theta_0 + \theta_a A + \theta_m \frac{E[W \mid A, Z, X] - \alpha_0 - \alpha_x X}{\alpha_m} + \theta_x X$$

$$E[Y \mid A, Z, X] = \left[\theta_0 - \frac{\theta_m}{\alpha_m} \alpha_0 \right] + \theta_a A + \left[\frac{\theta_m}{\alpha_m} \right] E[W \mid A, Z, X] + \left[\theta_x - \frac{\theta_m}{\alpha_m} \alpha_x \right] X$$

$$E[Y \mid A, Z, X] = \theta_0^* + \theta_a^* A + \theta_m^* E[W \mid A, Z, X] + \theta_x^* X$$

$$E\left(\frac{E[W|A, Z, X] - \alpha_0 - \alpha_x X}{\alpha_m} \mid A, Z, X\right) = \beta_0 + \beta_a A + \beta_x X$$

$$E(E[W \mid A, Z, X] \mid A, Z, X) = [\beta_0 \alpha_m - \alpha_0] + [\beta_a \alpha_m] A + [\beta_x \alpha_m - \alpha_x] X$$

$$E[W \mid A, Z, X] = \beta_0^* + \beta_a^* A + \beta_x^* X$$

Finally, we use the obtained regression coefficients to identify the desired causal effects:

$$NDE^a = NDE^{a'} = \theta_a^* = \theta_a$$

$$NIE^a = NIE^{a'} = \theta_m^* \beta_a^* = \left(\frac{\theta_m}{\alpha_m}\right) (\beta_a \alpha_m) = \theta_m \beta_a$$

A2.2 NDE and NIE with continuous outcome Y (Identity link) with hidden mediator with exposure-mediator interaction

We investigate the case with continuous Y with an Identity link and exposure-mediator interaction.

We assume that the data are generated from the following models:

Assumption 6A:

$$E[Y \mid A, Z, M, X] = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM + \theta_x X$$

$$E[W \mid A, Z, M, X] = \alpha_0 + \alpha_m M + \alpha_x X$$

$$E[M \mid A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

At this point another hypothesis is needed, namely that there is no interaction between treatment A and covariates.

Assumption 6B:

$$A \perp X$$

We now prove Result 2 in the section 3.1 under Assumption 6A and 6B.

Initially, we derive the conditional expected value of W:

$$E[W \mid A, Z, X] = \omega_0 + \omega_a A + \omega_z Z + \omega_x X$$

Then, we take the conditional expectation on both sides of the first two data generation equations for M conditional on (A, Z, X):

$$E[Y \mid A, Z, M, X] = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM + \theta_x X$$

$$E[Y \mid A, Z, X] = E[\theta_0 + \theta_a A + \theta_m M + \theta_{am} AM + \theta_x X \mid A, Z, X]$$

$$E[Y \mid A, Z, X] = \theta_0 + \theta_a A + (\theta_m + \theta_{am} A) E[M \mid A, Z, X] + \theta_x X$$

$$E[W \mid A, Z, M, X] = \alpha_0 + \alpha_m M + \alpha_x X$$

$$E[W \mid A, Z, X] = \alpha_0 + \alpha_m E[M \mid A, Z, X] + \alpha_x X$$

Thus, we rewrite the conditional mean of M in terms of the conditional mean of W:

$$E[M \mid A, Z, X] = \frac{E[W \mid A, Z, X] - \alpha_0 - \alpha_x X}{\alpha_m}$$

Using the law of iterated expectations, we can rewrite the third data generation equation as follows:

$$E[M \mid A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

$$E[E[M \mid A, Z, X] \mid A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

We substitute the conditional expectation of M with its version as a function of $E[W \mid A, Z, X]$:

$$\begin{aligned} E[Y \mid A, Z, X] &= \theta_0 + \theta_a A + \left(\frac{\theta_m}{\alpha_m} + \frac{\theta_{am}}{\alpha_m} A \right) (E[W \mid A, Z, X] - \alpha_0 - \alpha_x X) + \theta_x X \\ E[Y \mid A, Z, X] &= \left[\theta_0 - \frac{\theta_m}{\alpha_m} \alpha_0 \right] + \left[\theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0 \right] A + \left[\frac{\theta_m}{\alpha_m} \right] E[W \mid A, Z, X] + \left[\frac{\theta_{am}}{\alpha_m} \right] A E[W \mid A, Z, X] \\ &+ \left[\theta_x - \frac{\theta_m}{\alpha_m} \alpha_x \right] X + \left[-\frac{\theta_{am}}{\alpha_m} \alpha_x \right] X A \end{aligned}$$

Because of the orthogonalization between A and X we have:

$$E[Y \mid A, Z, X] = \theta_0^* + \theta_a^* A + \theta_m^* E[W \mid A, Z, X] + \theta_{am}^* A E[W \mid A, Z, X] + \theta_x^* X$$

$$E[E[M \mid A, Z, X] \mid A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

$$E\left(\frac{E[W \mid A, Z, X] - \alpha_0 - \alpha_x X}{\alpha_m} \mid A, Z, X\right) = \beta_0 + \beta_a A + \beta_x X$$

$$E[W \mid A, Z, X] = [\beta_0 \alpha_m + \alpha_0] + [\beta_a \alpha_m] A + [\alpha_m \beta_x + \alpha_x] X$$

$$E[W \mid A, Z, X] = \beta_0^* + \beta_a^* A + \beta_x^* X$$

Finally, we use the obtained regression coefficients to identify the desired causal effects:

$$\begin{aligned} \text{NDE}^a &= \theta_a^* + \theta_{am}^* (\beta_0^* + \beta_a^* + \beta_x^* E[X]) \\ &= \left[\theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0 \right] + \left[\frac{\theta_{am}}{\alpha_m} \right] ([\beta_0 \alpha_m + \alpha_0] + [\beta_a \alpha_m] + [\alpha_m \beta_x + \alpha_x] E[X]) \\ &= \theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0 + \left[\frac{\theta_{am}}{\alpha_m} \beta_0 \alpha_m + \frac{\theta_{am}}{\alpha_m} \alpha_0 \right] + \left[\frac{\theta_{am}}{\alpha_m} \beta_a \alpha_m \right] + \left[\frac{\theta_{am}}{\alpha_m} \alpha_m \beta_x + \frac{\theta_{am}}{\alpha_m} \alpha_x \right] E[X] \\ &= \theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0 + [\theta_{am} \beta_0 + \frac{\theta_{am}}{\alpha_m} \alpha_0] + [\theta_{am} \beta_a] + [\theta_{am} \beta_x + \frac{\theta_{am}}{\alpha_m} \alpha_x] E[X] \\ &= \theta_a + \theta_{am} (\beta_0 + \beta_a + \beta_x X + \frac{\alpha_x}{\alpha_m} E[X]) \\ &\neq \theta_a + \theta_{am} (\beta_0 + \beta_a + \beta_x E[X]) \end{aligned}$$

$$\begin{aligned} \text{NDE}^{a'} &= \theta_a^* + \theta_{am}^* (\beta_0^* + \beta_x^* E[X]) \\ &= \left[\theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0 \right] + \left[\frac{\theta_{am}}{\alpha_m} \beta_0 \alpha_m + \frac{\theta_{am}}{\alpha_m} \alpha_0 \right] + \left[\frac{\theta_{am}}{\alpha_m} \alpha_m \beta_x + \frac{\theta_{am}}{\alpha_m} \alpha_x \right] E[X] \\ &= \theta_a + [\theta_{am} \beta_0] + [\theta_{am} (\beta_x + \frac{\alpha_x}{\alpha_m}) E[X]] \end{aligned}$$

$$\neq \theta_a + \theta_{am}(\beta_0 + \beta_x E[X])$$

$$\text{NIE}^a = \beta_a^*(\theta_m^* + \theta_{am}^*) = [\beta_a \alpha_m] \left[\frac{\theta_m}{\alpha_m} + \frac{\theta_{am}}{\alpha_m} \right] = \beta_a(\theta_m + \theta_{am})$$

$$\text{NIE}^{a'} = \theta_m^* \beta_a^* = \left[\frac{\theta_m}{\alpha_m} \right] [\beta_a \alpha_m] = \theta_m \beta_a$$

Note that in the case with exposure-mediator interaction and measurable confounders X , the NDEs are not identified by the regression coefficients obtained using the proximal approach, because of the $\frac{\alpha_x}{\alpha_m}$ term. However, we can still identify them as the difference between the ATE and the NIEs:

$$\text{NDE}^a = \text{ATE} - \text{NIE}^{a'}$$

$$\text{NDE}^{a'} = \text{ATE} - \text{NIE}^a$$

A2.3 NDE and NIE with continuous outcome Y (Identity link) with hidden mediator with and without exposure-mediator interaction (Identification of cross-world counterfactual means)

We investigate the case with continuous Y with an identity link and exposure-mediator interaction.

We assume that the data are generated from the following models:

Assumption 7A:

$$E[Y \mid A, Z, M, X] = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM + \theta_x X$$

$$E[W \mid A, Z, M, X] = \alpha_0 + \alpha_a A + \alpha_m M + \alpha_x X$$

$$E[W \mid A, X] = \lambda_0 + \lambda_a A + \lambda_x X$$

Assumption 8A:

$$E[h_a(w, x) \mid A = a, Z, X] = E[Y \mid A = a, Z, X]$$

Under Assumptions 4 and 8A, the counterfactual mean of the outcome can be written (Ghassami et al., 2023) as follows:

$$E[Y(a, M(a'))] = \int_{w,x} h_a(w, x) p(w \mid A = a', X) p(x) \partial w \partial x$$

We now prove Result 3 in Section 3.1 under Assumptions 4, 7A, and 8A. Initially, we derive the conditional expected value of W :

$$E[W \mid A, Z, X] = \omega_0 + \omega_a A + \omega_z Z + \omega_x X$$

Then, we take the conditional expectation on both sides of the first two data generation equations for

M conditional on (A, Z, X):

$$E[Y | A, Z, M, X] = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM + \theta_x X$$

$$E[Y | A, Z, X] = \theta_0 + \theta_a A + \theta_m E[M | A, Z, X] + \theta_{am} A E[M | A, Z, X] + \theta_x X$$

$$E[W | A, Z, M, X] = \alpha_0 + \alpha_a A + \alpha_m M + \alpha_x X$$

$$E[W | A, Z, X] = \alpha_0 + \alpha_a A + \alpha_m E[M | A, Z, X] + \alpha_x X$$

Thus, we rewrite the conditional mean of M in terms of the conditional mean of W:

$$E[M | A, Z, X] = \frac{E[W | A, Z, X] - \alpha_0 - \alpha_a A - \alpha_x X}{\alpha_m}$$

We substitute the conditional expectation of M with its version as a function of E[W | A, Z, X]:

$$E[Y | A, Z, X] = \theta_0 + \theta_a A + (\theta_m + \theta_{am} A) \frac{E[W | A, Z, X] - \alpha_0 - \alpha_a A - \alpha_x X}{\alpha_m} + \theta_x X$$

If A = a = 1, the conditional expected value of Y and W are:

$$E[Y | A = 1, Z, X] = (\theta_0 + \theta_a) + \frac{(\theta_m + \theta_{am})}{\alpha_m} (E[W | A = 1, Z, X] - \alpha_0 - \alpha_x X) + \theta_x X$$

$$E[Y | A = 1, Z, X] = (\theta_0 + \theta_a) + \frac{(\theta_m + \theta_{am})}{\alpha_m} E[W | A = 1, Z, X] - \frac{(\theta_m + \theta_{am})}{\alpha_m} \alpha_0 - \frac{(\theta_m + \theta_{am})}{\alpha_m} \alpha_x X + \theta_x X$$

$$E[Y | A = 1, Z, X] = \left[\theta_0 + \theta_a - \frac{(\theta_m + \theta_{am})}{\alpha_m} \alpha_0 \right] + \left[\frac{(\theta_m + \theta_{am})}{\alpha_m} \right] E[W | A = 1, Z, X] + \left[\theta_x - \frac{(\theta_m + \theta_{am})}{\alpha_m} \alpha_x \right] X$$

$$E[Y | A = 1, Z, X] = \theta_0^1 + \theta_w^1 E[W | A = 1, Z, X] + \theta_x^1 X$$

$$E[W | A = 1, X] = \lambda_0^1 + \lambda_a^1 A + \lambda_x^1 X$$

If A = a' = 0, the conditional expected value of Y and W are:

$$E[Y | A = 0, Z, X] = \theta_0 + \frac{\theta_m}{\alpha_m} (E[W | A = 0, Z, X] - \alpha_0 - \alpha_x X) + \theta_x X$$

$$E[Y | A = 0, Z, X] = \theta_0 + \frac{\theta_m}{\alpha_m} E[W | A = 0, Z, X] - \frac{\theta_m}{\alpha_m} \alpha_0 - \frac{\theta_m}{\alpha_m} \alpha_x X + \theta_x X$$

$$E[Y | A = 0, Z, X] = \left[\theta_0 - \frac{\theta_m}{\alpha_m} \alpha_0 \right] + \frac{\theta_m}{\alpha_m} E[W | A = 0, Z, X] + \left[\theta_x - \frac{\theta_m}{\alpha_m} \alpha_x \right] X$$

$$E[Y | A = 0, Z, X] = \theta_0^0 + \theta_w^0 E[W | A = 0, Z, X] + \theta_x^0 X$$

$$E[W | A = 0, X] = \lambda_0 + \lambda_a A + \lambda_x X$$

$$E[W | A = 0, X] = (\lambda_0 + \lambda_a) + \lambda_x X$$

$$E[W | A = 0, X] = \lambda_0^0 + \lambda_x^0 X$$

For Assumption 8A, the outcome counterfactual mean can be rewritten in terms of regression coefficients:

$$\begin{aligned}
& E[Y(a, M(a'))] \\
&= \int_{w,x} h_a(w, x) p(w \mid A = a', X) p(x) \partial w \partial x = \\
&= \int_{w,x} (\theta_0^a + \theta_w^a w + \theta_x^a x) p(w \mid A = a', X) p(x) \partial w \partial x = \\
&= \theta_0^a + \theta_w^a \int_{w,x} w p(w \mid A = a', X) p(x) \partial w \partial x + \theta_x^a \int_x x p(x) \partial x = \\
&= \theta_0^a + \theta_w^a \int_x E[W \mid A = a', X] p(x) \partial x + \theta_x^a E[X] = \\
&= \theta_0^a + \theta_w^a \left(\lambda_0^{a'} + \lambda_a^{a'} a' + \lambda_x^{a'} \int_x x p(x) \partial x \right) + \theta_x^a E[X] = \\
&= \theta_0^a + \theta_w^a \left(\lambda_0^{a'} + \lambda_a^{a'} a' + \lambda_x^{a'} E[X] \right) + \theta_x^a E[X]
\end{aligned}$$

Finally, we use the obtained regression coefficients to identify the four outcome counterfactual means:

$$\begin{aligned}
E[Y(1, M(0))] &= \theta_0^1 + \theta_w^1 (\lambda_0^0 + \lambda_x^0 E[X]) \\
E[Y(0, M(1))] &= \theta_0^0 + \theta_w^0 (\lambda_0^1 + \lambda_a^1 + \lambda_x^1 E[X]) + \theta_x^0 E[X] \\
E[Y(0, M(0))] &= \theta_0^0 + \theta_w^0 (\lambda_0^0 + \lambda_x^0 E[X]) + \theta_x^0 E[X] \\
E[Y(1, M(1))] &= \theta_0^1 + \theta_w^1 (\lambda_0^1 + \lambda_a^1 + \lambda_x^1 E[X]) + \theta_x^1 E[X]
\end{aligned}$$

A2.4 NDE and NIE with count, binary or polytomous outcome Y (Log link) with hidden mediator without exposure-mediator interaction

We investigate the case with continuous Y with Logarithm link and without exposure-mediator interaction. We assume that the data are generated from the following models:

Assumption 9A:

$$\log(E[Y|A, Z, M, X]) = \theta_0 + \theta_a A + \theta_m M + \theta_x X$$

$$E[W|A, Z, M, X] = \alpha_0 + \alpha_m M + \alpha_x X$$

$$E[M|A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

$$M = E[M|A, Z, X] + \varepsilon_M; \quad E[\varepsilon_M] = 0; \quad \varepsilon_M \perp A, Z, X;$$

$$M_{\varepsilon_M}(\cdot) \text{ exists for any real value } \cdot$$

We now prove Result 4 in the section 3.2 under Assumption 9A.

Initially, we derive the conditional expected value of W:

$$E[W|A, Z, X] = \omega_0 + \omega_a A + \omega_z Z + \omega_x X$$

Then, we take the conditional expectation on both sides of the Y data generation equation for M

conditional on (A, Z, X) .

$$\log(E[Y|A, Z, M, X]) = \theta_0 + \theta_a A + \theta_m M + \theta_x X$$

$$E[Y|A, Z, M, X] = \exp(\theta_0 + \theta_a A + \theta_m M + \theta_x X)$$

$$E[Y|A, Z, X] = \exp(\theta_0 + \theta_a A + \theta_x X) E[\exp(\theta_m M)|A, Z]$$

$$E[Y|A, Z, X] = \exp(\theta_0 + \theta_a A + \theta_x X) \exp(\theta_m [\beta_0 + \beta_a A + \beta_x X]) E[\exp(\theta_m \varepsilon_M)]$$

$$E[Y|A, Z, X] = \exp(\theta_0 + \theta_a A + \theta_x X + \theta_m E[M|A, Z, X] + \log(M_{\varepsilon_M}(\theta_m)))$$

$$E[Y|A, Z, X] = \exp(\theta'_0 + \theta_a A + \theta_m E[M|A, Z, X] + \theta_x X)$$

where θ'_0 is defined as $\theta_0 + \log(M_{\varepsilon_M}(\theta_m))$.

Similarly, we take the conditional expectation on both sides of the W data generation equation for M conditional on (A, Z, X) .

$$E[W|A, Z, M, X] = \alpha_0 + \alpha_m M + \alpha_x X$$

$$E[W|A, Z, X] = \alpha_0 + \alpha_m E[M|A, Z, X] + \alpha_x X$$

Thus, we rewrite the conditional mean of M in terms of the conditional mean of W.

$$E[M|A, Z, X] = \frac{E[W|A, Z, X] - \alpha_0 - \alpha_x X}{\alpha_m}$$

Using the law of iterated expectations, we can rewrite the third data generation equations as follows:

$$E[M|A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

$$E[E[M|A, Z, X]|A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

We substitute the conditional expectation of M with its version as a function of $E[W | A, Z, X]$:

$$\log(E[Y|A, Z, X]) = \theta'_0 + \theta_a A + \theta_m E[M|A, Z, X] + \theta_x X$$

$$\log(E[Y | A, Z, X]) = \theta'_0 + \theta_a A + \theta_m \frac{E[W|A, Z, X] - \alpha_0 - \alpha_x X}{\alpha_m} + \theta_x X$$

$$\log(E[Y|A, Z, X]) = \theta'_0 + \theta_a A + \frac{\theta_m}{\alpha_m} (E[W|A, Z, X] - \alpha_0 - \alpha_x X) + \theta_x X$$

$$\log(E[Y|A, Z, X]) = \left[\theta'_0 - \frac{\theta_m}{\alpha_m} \alpha_0 \right] + \theta_a A + \left[\frac{\theta_m}{\alpha_m} \right] E[W|A, Z, X] + \left[\theta_x - \frac{\theta_m}{\alpha_m} \alpha_x \right] \theta_x X$$

$$E \left[(E[W | A, Z, X] - \alpha_0 - \alpha_x X) \frac{1}{\alpha_m} | A, Z, X \right] = \beta_0 + \beta_a A + \beta_x X$$

$$E(E[W | A, Z, X] | A, Z, X) = [\beta_0 \alpha_m - \alpha_0] + [\beta_a \alpha_m] A + [\beta_x \alpha_m - \alpha_x] X$$

$$E[W | A, Z, X] = \beta_0^* + \beta_a^* A + \beta_x^* X$$

Finally, we use the obtained regression coefficients to identify the NIEs :

$$\begin{aligned} \text{NIE}^a &= \exp(\beta_a^* \theta_m^*) = \exp\left([\beta_a \alpha_m] \left[\frac{\theta_m}{\alpha_m}\right]\right) = \exp(\beta_a \theta_m) \\ \text{NIE}^{a'} &= \frac{1}{\exp(\beta_a^* \theta_m^*)} = \frac{1}{\exp\left([\beta_a \alpha_m] \left[\frac{\theta_m}{\alpha_m}\right]\right)} = \frac{1}{\exp(\beta_a \theta_m)} \\ \text{NDE}^a &= \exp(\theta_a^*) = \exp(\theta_a) \\ \text{NDE}^{a'} &= \frac{1}{\exp(\theta_a^*)} = \frac{1}{\exp(\theta_a)} \end{aligned}$$

A2.5 NDE and NIE with count, binary or polytomous outcome Y (Log link) with hidden mediator with exposure-mediator interaction

We investigate the case with continuous Y with logarithm link and with exposure-mediator interaction.

We assume that the data are generated from the following models:

Assumption 10A:

$$\log(E[Y|A, Z, M, X]) = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM + \theta_x X$$

$$E[W|A, Z, M, X] = \alpha_0 + \alpha_m M + \alpha_x X$$

$$E[M|A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

$$M = E[M|A, Z, X] + \varepsilon_M; \quad E[\varepsilon_M] = 0; \quad \varepsilon_M \perp A, Z, X;$$

$$M_{\varepsilon_M}(\cdot) \text{ exists for any real value } \cdot$$

We now prove Result 5 in the section 3.2 under Assumption 10A.

Initially, we derive the conditional expected value of W:

$$E[W|A, Z, X] = \omega_0 + \omega_a A + \omega_z Z + \omega_x X$$

Then, we take the conditional expectation on both sides of the Y data generation equation for M conditional on (A, Z, X).

$$\log(E[Y|A, Z, M, X]) = \theta_0 + \theta_a A + \theta_m M + \theta_{am} AM + \theta_x X$$

$$E[Y|A, Z, M, X] = \exp(\theta_0 + \theta_a A + \theta_m M + \theta_{am} AM + \theta_x X)$$

$$E[Y|A, Z, X] = \exp(\theta_0 + \theta_a A + \theta_x X) E[\exp(\theta_m + \theta_{am} A) M | A, Z, X]$$

$$E[Y|A, Z, X] = \exp(\theta_0 + \theta_a A + \theta_x X) \exp((\theta_m + \theta_{am} A) [\beta_0 + \beta_a A + \beta_x X]) E[\exp((\theta_m + \theta_{am} A) \varepsilon_M) | A, Z, X]$$

$$E[Y|A, Z, X] = \exp(\theta_0 + \theta_a A + \theta_x X) \exp((\theta_m + \theta_{am} A) E[M|A, Z, X]) M_{\epsilon_M}((\theta_m + \theta_{am}))$$

$$E[Y|A, Z, X] = \exp(\theta_0 + \theta_a A + \theta_x X + (\theta_m + \theta_{am} A) E[M|A, Z, X]) + \log(M_{\epsilon_M}(\theta_m + \theta_{am}))$$

$$E[Y|A, Z, X] = \exp(\theta_0'' + \theta_a A + (\theta_m + \theta_{am} A) E[M|A, Z, X] + \theta_x X)$$

$$\log(E[Y|A, Z, X]) = \theta_0'' + \theta_a A + (\theta_m + \theta_{am} A) E[M|A, Z, X] + \theta_x X$$

where θ_0'' is defined as $\theta_0 + \log(M_{\epsilon_M}(\theta_m + \theta_{am}))$.

Similarly, we take the conditional expectation on both sides of the W data generation equation for M conditional on (A, Z, X).

$$E[W|A, Z, M, X] = \alpha_0 + \alpha_m M + \alpha_x X$$

$$E[W|A, Z, X] = \alpha_0 + \alpha_m E[M|A, Z, X] + \alpha_x X$$

Thus, we rewrite the conditional mean of M in terms of the conditional mean of W.

$$E[M|A, Z, X] = \frac{E[W|A, Z, X] - \alpha_0 - \alpha_x X}{\alpha_m}$$

Using the law of iterated expectations, we can rewrite the third data generation equations as follows:

$$E[M|A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

$$E[E[M|A, Z, X]|A, Z, X] = \beta_0 + \beta_a A + \beta_x X$$

We substitute the conditional expectation of M with its version as a function of $E[W | A, Z, X]$:

$$\log(E[Y|A, Z, X]) = \theta_0'' + \theta_a A + (\theta_m + \theta_{am} A) E[M|A, Z, X] + \theta_x X$$

$$\log(E[Y | A, Z, X]) = \theta_0'' + \theta_a A + (\theta_m + \theta_{am} A) \frac{E[W|A, Z, X] - \alpha_0 - \alpha_x X}{\alpha_m} + \theta_x X$$

$$\log(E[Y|A, Z, X]) = \theta_0'' + \theta_a A + \frac{\theta_m}{\alpha_m} (E[W|A, Z, X] - \alpha_0 - \alpha_x X) + \frac{\theta_{am}}{\alpha_m} A (E[W|A, Z, X] - \alpha_0 - \alpha_x X) + \theta_x X$$

$$\log(E[Y|A, Z, X]) = \left[\theta_0'' - \frac{\theta_m}{\alpha_m} \alpha_0 \right] + \theta_a A + \left[\frac{\theta_m}{\alpha_m} \right] E[W|A, Z, X] + \frac{\theta_{am}}{\alpha_m} A E[W|A, Z, X] + \left[\theta_x - \frac{\theta_m}{\alpha_m} \alpha_x \right] \theta_x X$$

$$\log(E[Y|A, Z, X]) = \theta_0^* + \theta_a^* A + \theta_m^* E[W|A, Z, X] + \theta_{am}^* A E[W|A, Z, X] + \theta_x^* X$$

$$E \left[(E[W | A, Z, X] - \alpha_0 - \alpha_x X) \frac{1}{\alpha_m} \mid A, Z, X \right] = \beta_0 + \beta_a A + \beta_x X$$

$$E(E[W | A, Z, X] | A, Z, X) = [\beta_0 \alpha_m - \alpha_0] + [\beta_a \alpha_m] A + [\beta_x \alpha_m - \alpha_x] X$$

$$E[W | A, Z, X] = \beta_0^* + \beta_a^* A + \beta_x^* X$$

Finally, we use the obtained regression coefficients to identify the NIEs :

$$NIE^a = \exp(\beta_a^* (\theta_m^* + \theta_{am}^*)) = \exp \left([\beta_a \alpha_m] \left[\frac{\theta_m}{\alpha_m} + \frac{\theta_{am}}{\alpha_m} \right] \right) = \exp(\beta_a (\theta_m + \theta_{am}))$$

$$\text{NIE}^{a'} = \frac{1}{\exp(\theta_m^* \beta_a^*)} = \frac{1}{\exp\left(\frac{\theta_m}{\alpha_m} [\beta_a \alpha_m]\right)} = \frac{1}{\exp(\theta_m \beta_a)}$$

Although the NDEs are not directly identified due to the MGF ratio, their estimates can still be obtained using the NIEs and ATE

$$\begin{aligned} \text{NDE}^a &= \exp(\theta_a^* + \theta_{am}^* (\beta_0^* + \beta_a^* + \beta_x^* \text{E}[X])) \frac{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)}{M_{\epsilon_M}(\theta_m^*)} \\ &= \exp\left([\theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0] + \left[\frac{\theta_{am}}{\alpha_m}\right] [\beta_0 \alpha_m + \alpha_0 + \beta_a \alpha_m + (\alpha_m \beta_x + \alpha_x) \text{E}[X]]\right) \frac{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)}{M_{\epsilon_M}(\theta_m^*)} \\ &= \exp\left(\theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0 + \left[\frac{\theta_{am}}{\alpha_m} \beta_0 \alpha_m + \frac{\theta_{am}}{\alpha_m} \alpha_0 + \frac{\theta_{am}}{\alpha_m} \beta_a \alpha_m + \frac{\theta_{am}}{\alpha_m} (\alpha_m \beta_x + \alpha_x) \text{E}[X]\right]\right) \frac{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)}{M_{\epsilon_M}(\theta_m^*)} \\ &= \exp\left(\theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0 + \left[\theta_{am} \beta_0 + \frac{\theta_{am}}{\alpha_m} \alpha_0 + \theta_{am} \beta_a + \theta_{am} (\beta_x + \frac{\alpha_x}{\alpha_m}) \text{E}[X]\right]\right) \frac{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)}{M_{\epsilon_M}(\theta_m^*)} \\ &= \exp\left(\theta_a + \left[\theta_{am} \beta_0 + \theta_{am} \beta_a + \theta_{am} (\beta_x + \frac{\alpha_x}{\alpha_m}) X\right]\right) \frac{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)}{M_{\epsilon_M}(\theta_m^*)} \\ &= \exp\left(\theta_a + \theta_{am} (\beta_0 + \beta_a + \beta_x \text{E}[X] + \frac{\alpha_x}{\alpha_m} X)\right) \frac{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)}{M_{\epsilon_M}(\theta_m^*)} \\ &\neq \exp(\theta_a + \theta_{am} (\beta_0 + \beta_a + \beta_x \text{E}[X])) \frac{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)}{M_{\epsilon_M}(\theta_m^*)} \\ \\ \text{NDE}^{a'} &= \frac{1}{\exp(\theta_a + \theta_{am}^* (\beta_0^* + \beta_a^* + \beta_x^* \text{E}[X]))} \frac{M_{\epsilon_M}(\theta_m^*)}{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)} \\ &= \frac{1}{\exp\left(\theta_a - \frac{\theta_{am}}{\alpha_m} \alpha_0 + \left[\frac{\theta_{am}}{\alpha_m} \beta_0 \alpha_m + \frac{\theta_{am}}{\alpha_m} \alpha_0 + \frac{\theta_{am}}{\alpha_m} \alpha_m \beta_x + \frac{\theta_{am}}{\alpha_m} \alpha_x\right] \text{E}[X]\right)} \frac{M_{\epsilon_M}(\theta_m^*)}{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)} \\ &= \frac{1}{\exp\left(\theta_a + [\theta_{am} \beta_0] + \left[\theta_{am} (\beta_x + \frac{\alpha_x}{\alpha_m})\right] \text{E}[X]\right)} \frac{M_{\epsilon_M}(\theta_m^*)}{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)} \\ &= \frac{1}{\exp\left(\theta_a + \theta_{am} (\beta_0 + \beta_x \text{E}[X] + \frac{\alpha_x}{\alpha_m} \text{E}[X])\right)} \frac{M_{\epsilon_M}(\theta_m^*)}{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)} \\ &\neq \frac{1}{\exp(\theta_a + \theta_{am} (\beta_0 + \beta_x \text{E}[X]))} \frac{M_{\epsilon_M}(\theta_m^*)}{M_{\epsilon_M}(\theta_m^* + \theta_{am}^*)} \end{aligned}$$

For NDEs, it is not possible to apply a procedure similar to that used for NIEs by combining the obtained regression coefficients, because the $\frac{\alpha_x}{\alpha_m}$ term and the ratio of the MGFs remain unidentified. However, the NDEs can be identified as a ratio or product of NIEs and the ATE.

$$\text{NDE}^a = \text{ATE} \cdot \text{NIE}^{a'}$$

$$\text{NDE}^{a'} = \frac{\text{NIE}^a}{\text{ATE}}$$

A3 The empirical sandwich variance estimator

To estimate the empirical sandwich variance of the four estimators considered, the R package **geex** was used (R Core Team, 2008). This package allows for asymptotic inference based on the estimating function methods or M-estimation methods.

Let N be the number of observations and $O_i = (Y_i, A_i, W_i, Z_i, X_i)$ be the observed data of unit i .

The M-estimation theory states that if an estimator can be obtained as a solution to an estimating equation of the form:

$$\sum_{i=1}^n \psi(O_i, \phi) = 0$$

the estimator, under suitable regularity assumptions, is consistent and asymptotically normal, and its variance can be consistently estimated by the empirical sandwich variance estimator. ϕ , the root of the estimating equation, is the M-estimator.

The models with exposure-mediator interaction terms can be written as follows:

$$g[E(Y | A, Z, X)] = \theta_0^* + \theta_a^* A + \theta_m^* E(W | A, Z, X) + \theta_{am}^* A E(W | A, Z, X) + \theta_x^* X$$

$$E(W | A, Z, X) = \omega_0 + \omega_a A + \omega_z Z + \omega_x X = \hat{W}$$

where g can be logarithmic or linear link functions, according to the nature of Y .

The estimating functions associated with the two regressions are the following:

$$\psi_1(O; \omega) = \begin{bmatrix} 1 \\ A \\ Z \\ X \end{bmatrix} (W - (\omega_0 + \omega_a A + \omega_z Z + \omega_x X)) \in R^4$$

$$\psi_2(O; \omega, \theta) = \begin{bmatrix} 1 \\ A \\ \hat{W} \\ X \end{bmatrix} (Y - g(\theta_0^* + \theta_a^* A + \theta_m^* \hat{W} + \theta_{am}^* A \hat{W} + \theta_x^* X)) \in R^5$$

These two estimating functions satisfy $E[\psi_1(O; \omega) | A, Z, X] = 0$ and $E[\psi_2(O; \omega, \theta) | A, Z, X] = 0$.

In addition to estimating the individual regression coefficients, we used the package to obtain estimates of the four causal effects, derived either as a combination of the regression coefficients or in comparison to the ATE, along with their sandwich variance.

A4 Measurement error

In this section, we will present the results discussed in Chapter 5 of this thesis. Specifically, we will show the biased coefficients for all the cases illustrated by Carroll, based on the different assumptions made about the measurement error in each case (Chapter 5.1). Additionally, we will provide some useful proofs for quantifying the bias in Chapter 5.2, focusing on the case of measurement error of a

linear model without interaction, as considered in the first case of the chapters 2.1, 3.1.1 and 3.1.2 in this thesis.

A4.1 Classical measurement error model with W (mismeasured variable) unbiased

We investigate the univariate continuous case with one correctly measured variable, M, and assume that the data are generated by the following model:

$$Y = \theta_0 + \theta_m M + \epsilon_y^m$$

$$E[\epsilon_y^m | M] = 0 \quad ; \quad \epsilon_y^m \perp M$$

We begin by considering a model where, instead of observing M, we observe W and it is unbiased for M, that is $E[W] = E[M]$.

$$W = M + U$$

$$E[U | M] = 0 \quad ; \quad U \perp M$$

The estimates we obtain are biased, as they derive from a different model:

$$Y = \theta_0^* + \theta_m^* W + \epsilon_y^w$$

Based on this model specification, we get the following identification of θ_m^* :

$$\theta_m^* = \frac{\text{Cov}(Y, W)}{V(W)} = \frac{\text{Cov}(\theta_0 + \theta_m M + \epsilon_y^m, M + U)}{V(M + U)} =$$

$$\frac{\text{Cov}(\theta_0, M) + \text{Cov}(\theta_0, U) + \text{Cov}(\theta_m M, M) + \text{Cov}(\theta_m M, U) + \text{Cov}(\epsilon_y^m, M) + \text{Cov}(\epsilon_y^m, U)}{V(M) + V(U) + 2 \text{Cov}(M, U)}$$

$$= \frac{\theta_m V(M) + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{V(M) + V(U)} = \frac{\theta_m \sigma_m^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\sigma_m^2 + \sigma_U^2}$$

When the error is non-differential, that is $\rho_{\epsilon_y^m U} = 0$, we have:

$$\theta_m^* = \theta_m \frac{\sigma_m^2}{\sigma_m^2 + \sigma_U^2}$$

When the error is differential, that is $\rho_{\epsilon_y^m U} \neq 0$, we have:

$$\theta_m^* = \frac{\theta_m \sigma_m^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\sigma_m^2 + \sigma_U^2}$$

We can write the model for Y as follows:

$$E[Y] = \theta_0 + \theta_x E[M]$$

and we can substitute that expression in the θ_0^* expression:

$$\begin{aligned}
\theta_0^* &= E[Y] - \theta_m^* E[W] \\
&= \theta_0 + \theta_m E[M] - \theta_m^* E[W] \\
&= \theta_0 + \theta_m E[M] - \theta_m^* E[M + U] \\
&= \theta_0 + \theta_m E[M] - \theta_m^* E[M] - \theta_m^* E[U] \\
&= \theta_0 + \theta_m E[M] - \theta_m^* E[M]
\end{aligned}$$

When the error is non-differential, that is $\rho_{\epsilon_y^m U} = 0$, we have:

$$\theta_0^* = \theta_0 + \theta_m E[M] - \theta_m \frac{\sigma_m^2}{\sigma_m^2 + \sigma_U^2} E[M]$$

When the error is differential, that is $\rho_{\epsilon_y^m U} \neq 0$, we have:

$$\begin{aligned}
\theta_0^* &= \theta_0 + \theta_m E[M] - \frac{\theta_m \sigma_m^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\sigma_m^2 + \sigma_U^2} E[M] \\
\theta_0^* &= \theta_0 + E[M] \left(\theta_m - \frac{\theta_m \sigma_m^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\sigma_m^2 + \sigma_U^2} \right)
\end{aligned}$$

A4.2 Classical measurement error model with W (mismeasured variable) biased

We investigate the univariate continuous case with one correctly measured variable, M, and assume the same data generating model as in the previous chapter:

$$\begin{aligned}
Y &= \theta_0 + \theta_m M + \epsilon_y^m \\
E[\epsilon_y^m | M] &= 0 \quad ; \quad \epsilon_y^m \perp M
\end{aligned}$$

We consider a model where, instead of observing M, we observe W and it is biased for M, that is $E[W] \neq E[M]$.

$$\begin{aligned}
W &= \gamma_0 + \gamma_m M + U \\
E[U | M] &= 0 \quad ; \quad U \perp M
\end{aligned}$$

The estimates we obtain are biased, as they derive from a different model:

$$Y = \theta_0^* + \theta_m^* W + \epsilon_y^w$$

Based on this model specification, we get the following identification of θ_m^* :

$$\begin{aligned}
\theta_m^* &= \frac{\text{Cov}(Y, W)}{V(W)} = \frac{\text{Cov}(\theta_0 + \theta_m M + \epsilon_y^m, \gamma_0 + \gamma_m M + U)}{V(\gamma_0 + \gamma_m M + U)} \\
&= \frac{\text{Cov}(\theta_m M, \gamma_0) + \text{Cov}(\theta_m M, \gamma_m M) + \text{Cov}(\theta_m M, U) + \text{Cov}(\epsilon_y^m, \gamma_0) + \text{Cov}(\epsilon_y^m, \gamma_m M) + \text{Cov}(\epsilon_y^m, U)}{V(\gamma_0) + \gamma_x^2 V(M) + V(U)} \\
&= \frac{\text{Cov}(\theta_m M, \gamma_m M) + \text{Cov}(\epsilon_y^m, U)}{V(\gamma_0) + \gamma_x^2 V(M) + V(U)} \\
&= \frac{\theta_m \gamma_m \sigma_m^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\gamma_m^2 \sigma_m^2 + \sigma_U^2}
\end{aligned}$$

When the error is non-differential, that is $\rho_{\epsilon_y^m U} = 0$, we have:

$$\theta_m^* = \theta_m \frac{\gamma_m \sigma_m^2}{\gamma_m^2 \sigma_m^2 + \sigma_U^2}$$

When the error is differential, that is $\rho_{\epsilon_y^m U} \neq 0$, we have:

$$\theta_m^* = \frac{\theta_m \gamma_m \sigma_m^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\gamma_m^2 \sigma_m^2 + \sigma_U^2}$$

We can write the model for Y as follows:

$$E[Y] = \theta_0 + \theta_m E[M]$$

and we can substitute that expression in the θ_0^* expression:

$$\begin{aligned}
\theta_0^* &= E[Y] - \theta_m^* E[W] \\
&= \theta_0 + \theta_m E[M] - \theta_m^* E[W] \\
&= \theta_0 + \theta_m E[M] - \theta_m^* E[\gamma_0 + \gamma_m M + U] \\
&= \theta_0 + \theta_m E[M] - \theta_m^* \gamma_0 - \theta_m^* \gamma_m E[M] - \theta_m^* E[U] \\
&= \theta_0 + \theta_m E[M] - \theta_m^* \gamma_0 - \theta_m^* \gamma_m E[M] \\
&= \theta_0 + \theta_m E[M] - \theta_m^* (\gamma_0 + \gamma_m E[M])
\end{aligned}$$

When the error is non-differential, that is $\rho_{\epsilon_y^m U} = 0$, we have:

$$\theta_0^* = \theta_0 + \theta_m E[M] - \theta_m \frac{\gamma_m \sigma_m^2}{\gamma_m^2 \sigma_m^2 + \sigma_U^2} (\gamma_0 + \gamma_m E[M])$$

When the error is differential, that is $\rho_{\epsilon_y^m U} \neq 0$, we have:

$$\theta_0^* = \theta_0 + \theta_m E[M] - \frac{(\theta_m \gamma_m \sigma_m^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U)}{\gamma_m^2 \sigma_m^2 + \sigma_U^2} (\gamma_0 + \gamma_m E[M])$$

A4.3 Classical multiple measurement error model with W (mismeasured variable) unbiased

We investigate the multivariate continuous case with two correctly measured variables, M and A, and assume that the data are generated by the following model:

$$Y = \theta_0 + \theta_m M + \theta_a A + \epsilon_y^m$$

$$E[\epsilon_y^m \mid M, A] = 0 \quad ; \quad \epsilon_y^m \perp M, A$$

We consider a model where, instead of observing M, we observe W and it is unbiased for M, that is $E[W] = E[M]$.

$$W = M + U$$

$$E[U \mid M, A] = 0 \quad ; \quad U \perp M, A$$

The estimates we obtain are biased, as they derive from a different model:

$$Y = \theta_0^* + \theta_x^* W + \theta_a^* A + \epsilon_y^w$$

The biased coefficient θ_m^* can be written as follows:

$$\theta_m^* = \frac{\text{Cov}(Y, W \mid A)}{V(W \mid A)} = \frac{\text{Cov}(\theta_0 + \theta_m M + \theta_a A + \epsilon_y^m, M + U \mid A)}{V(M + U \mid A)} =$$

$$\frac{\text{Cov}(\theta_m M, M \mid A) + \text{Cov}(\theta_m M, U \mid A) + \text{Cov}(\theta_a A, M \mid A) + \text{Cov}(\theta_a A, U \mid A) + \text{Cov}(\epsilon_y^m, M \mid A) + \text{Cov}(\epsilon_y^m, U \mid A)}{V(M \mid A) + V(U \mid A) + 2 \text{Cov}(M, U \mid A)}$$

$$= \frac{\theta_m V(M \mid A) + \text{Cov}(\epsilon_y^m, U \mid A)}{V(M \mid A) + V(U \mid A)}$$

$$= \frac{\theta_m \sigma_{m|a}^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\sigma_{m|a}^2 + \sigma_U^2}$$

When the error is non-differential, that is $\rho_{\epsilon_y^m U} = 0$, we have:

$$\theta_m^* = \theta_m \frac{\sigma_{m|a}^2}{\sigma_{m|a}^2 + \sigma_U^2}$$

When the error is differential, that is $\rho_{\epsilon_y^m U} \neq 0$, we have:

$$\theta_m^* = \frac{\theta_m \sigma_{m|a}^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\sigma_{m|a}^2 + \sigma_U^2}$$

We can write the model for Y as follows:

$$E[Y] = \theta_0 + \theta_m E[M] + \theta_a E[A]$$

and we can substitute that expression in the θ_0^* expression:

$$\begin{aligned}
\theta_0^* &= E[Y] - \theta_m^* E[W] - \theta_a^* E[A] \\
&= \theta_0 + \theta_m E[M] + \theta_a E[A] - \theta_m^* E[W] - \theta_a^* E[A] \\
&= \theta_0 + \theta_m E[M] + \theta_a E[A] - \theta_m^* E[M + U] - \theta_a^* E[A] \\
&= \theta_0 + \theta_m E[M] + \theta_a E[A] - \theta_m^* E[M] - \theta_m^* E[U] - \theta_a^* E[A] \\
&= \theta_0 + \theta_m E[M] - \theta_m^* E[M] + E[A](\theta_a - \theta_a^*)
\end{aligned}$$

When the error is non-differential, that is $\rho_{\epsilon_y^m U} = 0$, we have:

$$\theta_0^* = \theta_0 + \theta_m E[M] - \theta_m \frac{\sigma_{m|a}^2}{\sigma_{m|a}^2 + \sigma_U^2} E[M] + E[A](\theta_a - \theta_a^*) = \theta_0 + \theta_m E[M] \left(1 - \frac{\sigma_{m|a}^2}{\sigma_{m|a}^2 + \sigma_U^2} \right) + E[A](\theta_a - \theta_a^*)$$

When the error is differential, that is $\rho_{\epsilon_y^m U} \neq 0$, we have:

$$\begin{aligned}
\theta_0^* &= \theta_0 + \theta_m E[M] - \frac{\theta_m \sigma_{m|a}^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\sigma_{m|a}^2 + \sigma_U^2} E[M] + E[A](\theta_a - \theta_a^*) \\
&= \theta_0 + E[M] \left(\theta_m - \frac{\theta_m \sigma_{m|a}^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\sigma_{m|a}^2 + \sigma_U^2} \right) + E[A](\theta_a - \theta_a^*)
\end{aligned}$$

We can write the model of Y as follows:

$$E[Y | W, A] = \theta_0 + \theta_m E[M | W, A] + \theta_a A$$

To proceed with this multivariate case, we need a new data generating model and some strong assumptions:

$$E[M | W, A] = \delta_0 + \delta_w W + \delta_a A$$

Furthermore, under joint normality, the regression of M on (W, A) is linear.

To facilitate the derivation, we rewrite it by centering the variables A and W around their expected values; this leaves the regression coefficients unchanged, except for the intercept.

$$E[M | W, A] = \delta_{0*} + \delta_w (W - E(W | A)) + \delta_a (A - E(A))$$

Since $W - E(W | A)$ represents the part of W not explained by A, this component is obviously uncorrelated with $A - E(A)$. This independence ensures that the two terms are orthogonal, and the conditional relationship between M and A is not influenced by the part of W not explained by A. This implies that $E(M | A) = E(W | A)$. It follows:

$$\begin{aligned}
\delta_w &= \frac{E(E[M(W - E(W | A)) | A])}{E(E[(W - E(W | A))^2 | A])} \\
&= \frac{E(E(MW | A) - E(M | A) E(W | A))}{V(W | A)}
\end{aligned}$$

$$\begin{aligned}
&= \frac{\mathbb{E}[\mathbb{E}(M^2 | A) - \mathbb{E}(M | A)\mathbb{E}(M | A)]}{\mathbb{V}(M + U | A)} \\
&= \frac{\mathbb{E}[\mathbb{E}(M^2 | A) - \mathbb{E}(M | A)^2]}{\mathbb{V}(M | A) + \mathbb{V}(U | A)} \\
&= \frac{\sigma_{m|a}^2}{\sigma_{m|a}^2 + \sigma_U^2}
\end{aligned}$$

Suppose now that $\mathbb{E}(M | A) = \Gamma_0 + \Gamma_a A$, therefore $\mathbb{E}(W | A) = \Gamma_0 + \Gamma_a A$ as well.

Again, because of the orthogonalization, it is immediate that $\Gamma_a = \delta_a$.

We rewrite the $\mathbb{E}[M | W, A]$ and replace it in the $\mathbb{E}[Y | W, A]$ we have:

$$\begin{aligned}
\mathbb{E}[M | W, A] &= \delta_{0*} + \delta_w \{W - \Gamma_0 - \Gamma_a A\} + \delta_a \{A - \mathbb{E}(A)\} \\
&= \delta_{0*} + \delta_w W - \delta_w \Gamma_0 - \delta_w \Gamma_a A + \delta_a A - \delta_a \mathbb{E}(A) \\
\mathbb{E}[Y | W, A] &= \theta_0 + \theta_m \mathbb{E}[M | W, A] + \theta_a A \\
\mathbb{E}[Y | W, A] &= \theta_0 + \theta_m (\delta_{0*} + \delta_w W - \delta_w \Gamma_0 - \delta_w \Gamma_a A + \delta_a A - \gamma_a \mathbb{E}(A)) + \theta_a A \\
\mathbb{E}[Y | W, A] &= \theta_0 + \theta_m \delta_{0*} + \theta_m \delta_w W - \theta_m \delta_w \Gamma_0 - \theta_m \delta_w \Gamma_a A + \theta_m \delta_a A - \theta_m \delta_a \mathbb{E}(A) + \theta_a A \\
\mathbb{E}[Y | W, A] &= (\theta_0 + \theta_m \delta_{0*} - \theta_m \delta_w \Gamma_0) + (\theta_m \delta_w) W + (\theta_m \delta_a - \theta_m \delta_w \Gamma_a + \theta_a) A - \theta_m \delta_a \mathbb{E}(A) \\
\mathbb{E}[Y | W, A] &= \theta_0^* + \theta_m^* W + \theta_a^* A
\end{aligned}$$

It follows that:

$$\begin{aligned}
\theta_a^* &= \theta_a - \theta_m \delta_w \Gamma_a + \theta_m \delta_a \\
&= \theta_a + \theta_m (\delta_a - \delta_w \Gamma_a) \\
&= \theta_a + \theta_m \left(1 - \frac{\sigma_{m|a}^2}{\gamma_m^2 \sigma_{m|a}^2 + \sigma_U^2}\right) \delta_a
\end{aligned}$$

Even if the error is differential, the correlation between U and ϵ_y^m does not directly influence the relationship between of A and U since they are independent. Therefore, the expression for θ_a^* remains unchanged even in the case of differential error.

A4.4 Classical multiple measurement error model with W (mismeasured variable) biased

We investigate the multivariate continuous case with two correctly measured variables, M and A, and assume that the data are generated by the following model:

$$Y = \theta_0 + \theta_m M + \theta_a A + \epsilon_y^m$$

$$E[\epsilon_y^m \mid M, A] = 0 \quad ; \quad \epsilon_y^m \perp M, A$$

We consider a model where, instead of observing M , we observe W and it is biased for M , that is $E[W] \neq E[M]$.

$$W = \gamma_0 + \gamma_m M + U$$

$$E[U \mid M, A] = 0 \quad ; \quad U \perp M, A$$

The estimates we obtain are biased, as they derive from a different model:

$$Y = \theta_0^* + \theta_x^* W + \theta_a^* A + \epsilon_y^w$$

The biased coefficient θ_m^* can be written as follows:

$$\begin{aligned} \theta_m^* &= \frac{\text{Cov}(Y, W \mid A)}{V(W \mid A)} = \frac{\text{Cov}(\theta_0 + \theta_m M + \theta_a A + \epsilon_y^m, \gamma_0 + \gamma_m M + U \mid A)}{V(\gamma_0 + \gamma_m M + U \mid A)} \\ &= \frac{\text{Cov}(\theta_m M, \gamma_m M \mid A) + \text{Cov}(\theta_m M, U \mid A) + \text{Cov}(\theta_a A, \gamma_m M \mid A) + \text{Cov}(\theta_a A, U \mid A) + \text{Cov}(\epsilon_y^m, \gamma_m M \mid A) + \text{Cov}(\epsilon_y^m, U \mid A)}{\gamma_m^2 V(M \mid A) + V(U \mid A)} \\ &= \frac{\text{Cov}(\theta_m M, \gamma_m M \mid A) + \text{Cov}(\epsilon_y^m, U \mid A)}{\gamma_m^2 V(M \mid A) + V(U \mid A)} \\ &= \frac{\theta_m \gamma_m \sigma_{m|a}^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\gamma_m^2 \sigma_{m|a}^2 + \sigma_U^2} \end{aligned}$$

When the error is non-differential, that is $\rho_{\epsilon_y^m U} = 0$, we have:

$$\theta_m^* = \theta_m \frac{\gamma_m \sigma_{m|a}^2}{\gamma_m^2 \sigma_{m|a}^2 + \sigma_U^2}$$

When the error is differential, that is $\rho_{\epsilon_y^m U} \neq 0$, we have:

$$\theta_m^* = \frac{\theta_m \gamma_m \sigma_{m|a}^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\gamma_m^2 \sigma_{m|a}^2 + \sigma_U^2}$$

We can write the model for Y as follows:

$$E[Y] = \theta_0 + \theta_m E[M] + \theta_a E[A]$$

and we can substitute that expression in the θ_0^* expression:

$$\begin{aligned} \theta_0^* &= E[Y] - \theta_m^* E[W] - \theta_a E[A] \\ &= \theta_0 + \theta_m E[M] + \theta_a E[A] - \theta_m^* E[W] - \theta_a E[A] \\ &= \theta_0 + \theta_m E[M] + \theta_a E[A] - \theta_m^* E[\gamma_0 + \gamma_m M + U] - \theta_a E[A] \\ &= \theta_0 + \theta_m E[M] + \theta_a E[A] - \theta_m^* \gamma_0 - \theta_m^* \gamma_m E[M] - \theta_m^* E[U] - \theta_a E[A] \end{aligned}$$

$$= \theta_0 + \theta_m E[M] - \theta_m^* \gamma_0 - \theta_m^* \gamma_m E[M] + E[A](\theta_a - \theta_a^*)$$

$$= \theta_0 + \theta_m E[M] - \theta_m^* (\gamma_0 + \gamma_m E[M]) + E[A](\theta_a - \theta_a^*)$$

When the error is non-differential, that is $\rho_{\epsilon_y^m U} = 0$, we have:

$$\theta_0^* = \theta_0 + \theta_x E[M] - \theta_m \frac{\gamma_m \sigma_{m|a}^2}{\gamma_m^2 \sigma_{m|a}^2 + \sigma_U^2} (\gamma_0 + \gamma_m E[M]) + E[A](\theta_a - \theta_a^*)$$

When the error is differential, that is $\rho_{\epsilon_y^m U} \neq 0$, we have:

$$\theta_0^* = \theta_0 + \theta_m E[M] - \frac{\theta_m \gamma_m \sigma_{m|a}^2 + \rho_{\epsilon_y^m U} \sigma_{\epsilon_y^m} \sigma_U}{\gamma_m^2 \sigma_{m|a}^2 + \sigma_U^2} (\gamma_0 + \gamma_m E[M]) + E[A](\theta_a - \theta_a^*)$$

We can write the model for Y as follows:

$$E[Y | W, A] = \theta_0 + \theta_m E[M | W, A] + \theta_a A$$

To proceed with this multivariate case, we need a new data generating model and some strong assumptions: $E[X | W, A] = \delta_0 + \delta_w W + \delta_a A$

Furthermore, under joint normality, the regression of M on (W, A) is linear.

To facilitate the derivation, we rewrite it by centering the variables A and W around their expected values; this leaves the regression coefficients unchanged, except for the intercept.

$$E[X | W, A] = \delta_{0*} + \delta_w (W - E(W | A)) + \delta_a (A - E(A))$$

Because of the orthogonalization $E(M | A) = E(W | A)$ and it is immediate that:

$$\begin{aligned} \delta_w &= \frac{E(E[M(W - E(W | A)) | A])}{E(E[(W - E(W | A))^2 | A])} \\ &= \frac{E[E(MW | A) - E(M | A)E(W | A)]}{V(W | A)} \\ &= \frac{E[E(M^2 | A) - E(M | A)E(M | A)]}{V(\gamma_0 + \gamma_m M + U | A)} \\ &= \frac{E[E(M^2 | A) - E(M | A)^2]}{V(\gamma_0 | A) + \gamma_m^2 V(M | A) + V(U | A)} \\ &= \frac{E[E(M - E(M | A))^2 | A]}{\gamma_m^2 V(M | A) + V(U | A)} \\ &= \frac{\sigma_{m|a}^2}{\gamma_m^2 \sigma_{m|a}^2 + \sigma_U^2} \end{aligned}$$

Suppose now that $E(W | A) = \Gamma_0 + \Gamma_a A$. Again, because of the orthogonalization, it is immediate that $\Gamma_a = \delta_a$.

We rewrite the $E[M \mid W, A]$ and replace it in the $E[Y \mid W, A]$ we have:

$$\begin{aligned}
E[M \mid W, A] &= \delta_{0*} + \delta_w \{W - \Gamma_0 - \Gamma_a A\} + \delta_a \{A - E(A)\} \\
&= \delta_{0*} + \delta_w W - \delta_w \Gamma_0 - \delta_w \Gamma_a A + \delta_a A - \delta_a E(A) \\
E[Y \mid W, A] &= \theta_0 + \theta_m E[M \mid W, A] + \theta_a A \\
&= \theta_0 + \theta_m (\delta_{0*} + \delta_w W - \delta_w \Gamma_0 - \delta_w \Gamma_a A + \delta_a A - \delta_a E(A)) + \theta_a A \\
&= \theta_0 + \theta_m \delta_{0*} + \theta_m \delta_w W - \theta_m \delta_w \Gamma_0 - \theta_m \delta_w \Gamma_a A + \theta_m \delta_a A - \theta_m \delta_a E(A) + \theta_a A \\
&= (\theta_0 + \theta_m \delta_{0*} - \theta_m \delta_w \Gamma_0) + (\theta_m \delta_w) W + (\theta_m \delta_a - \theta_m \delta_w \Gamma_a + \theta_a) A - \theta_m \delta_a E(A) \\
&= \theta_0^* + \theta_m^* W + \theta_a^* A
\end{aligned}$$

It follows that

$$\begin{aligned}
\theta_a^* &= \theta_a - \theta_x \delta_w \Gamma_a + \theta_x \delta_a \\
&= \theta_a + \theta_x (\delta_a - \delta_w \Gamma_a) \\
&= \theta_a + \theta_m \left(1 - \frac{\sigma_{m|a}^2}{\gamma_m^2 \sigma_{m|a}^2 + \sigma_U^2} \right) \delta_a
\end{aligned}$$

Even in this case, if the error is differential, the correlation between U and ϵ_y^m does not directly influence the relationship of A and U and the expression for θ_a^* remains unchanged even in both cases.

A4.5 Bias quantification in NDE^a in linear model without exposure-mediator interaction

Proof. 1: The data-generating model $A = \alpha_0 + \alpha_w W + \Delta$ can be written using the moment condition as:

$$E \left[\begin{pmatrix} 1 \\ W \end{pmatrix} (A - \alpha_0 - \alpha_w W) \right] = 0$$

This data generating model $A = \alpha_0 + \alpha_w W + \Delta$ can also be written in matrix notation as $\underline{A} = \mathbf{W}\underline{\alpha} + \underline{\Delta}$.

Where $\underline{A}$, $\underline{\alpha}$ and $\underline{\Delta}$ are vectors and $\mathbf{W}$ can be a matrix or a vector.

The sum of squared residual of the model is $S(\alpha)$:

$$S(\alpha) = (\underline{A} - \mathbf{W}\underline{\alpha})'(\underline{A} - \mathbf{W}\underline{\alpha})$$

If we want to find the coefficients in α that 'best' fits the data generating equation, we have to solve the quadratic minimization problem:

$$\hat{\underline{\alpha}} = \arg \min S(\alpha) = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\underline{\mathbf{A}}$$

We can rewrite $S(\alpha)$ as follows:

$$S(\alpha) = \underline{\mathbf{A}}'\underline{\mathbf{A}} - \underline{\mathbf{A}}'\mathbf{W}\underline{\alpha} - \underline{\alpha}'\mathbf{W}'\underline{\mathbf{A}} + \underline{\alpha}'\mathbf{W}'\mathbf{W}\underline{\alpha}$$

$$= \underline{\mathbf{A}}'\underline{\mathbf{A}} - 2\underline{\alpha}'\mathbf{W}'\underline{\mathbf{A}} + \underline{\alpha}'\mathbf{W}'\mathbf{W}\underline{\alpha}$$

To find the minimum, the derivative of $S(\alpha)$ with respect to α is set equal to zero.

$$\frac{\partial S(\alpha)}{\partial \alpha} = -2\mathbf{W}'\underline{\mathbf{A}} + 2\mathbf{W}'\mathbf{W}\underline{\alpha}$$

$$-2\mathbf{W}'\underline{\mathbf{A}} + 2\mathbf{W}'\mathbf{W}\underline{\alpha} = 0$$

$$2\mathbf{W}'\underline{\mathbf{A}} = 2\mathbf{W}'\mathbf{W}\underline{\alpha}$$

$$\mathbf{W}'\underline{\mathbf{A}} = \mathbf{W}'\mathbf{W}\underline{\alpha}$$

$$\mathbf{W}'(\underline{\mathbf{A}} - \mathbf{W}\underline{\alpha}) = 0$$

Taking the expected value, the moment condition is:

$$\mathbb{E}[\mathbf{W}'(\underline{\mathbf{A}} - \mathbf{W}\underline{\alpha})] = 0$$

The moment condition arises from the assumption that the residuals $\underline{\Delta} = \underline{\mathbf{A}} - \mathbf{W}\underline{\alpha}$ are uncorrelated with the explanatory variable $\mathbf{W}$.

Returning to the scalar representation, the moment condition is:

$$\mathbb{E} \left[\begin{pmatrix} 1 \\ W \end{pmatrix} (A - \alpha_0 - \alpha_w W) \right] = 0$$

The same process also explains why the data generating model for the outcome, $Y = \theta_0^* + \theta_a^* A + \theta_m^* W + \epsilon_y^w$, can be expressed using the moment condition.

Proof. 2 $\mathbb{E}[A\Delta] = \mathbb{E}[(A - \alpha_0 - \alpha_w W)\Delta]$

We consider the data generating model $A = \alpha_0 + \alpha_w W + \Delta$, and we assume the error term Δ has zero mean $\mathbb{E}[\Delta] = 0$ and is independent of W , that is $\mathbb{E}[W\Delta] = 0$. Therefore, we have:

$$E[A\Delta] = E[(A - \alpha_0 - \alpha_W W)\Delta]$$

$$\text{Cov}(A, \Delta) + E[A]E[\Delta] = E[A\Delta - \alpha_0\Delta - \alpha_W W\Delta]$$

$$\text{Cov}(A, \Delta) = E[A\Delta] - \alpha_0 E[\Delta] - \alpha_W E[W\Delta]$$

$$\text{Cov}(A, \Delta) = \text{Cov}(A, \Delta) + E[A]E[\Delta]$$

$$\text{Cov}(A, \Delta) = \text{Cov}(A, \Delta)$$

$$\mathbf{Proof. 3} \quad E[\Delta(Y - \theta_a^* \Delta)] = E \left[\begin{pmatrix} 1 \\ A \\ W \end{pmatrix} (Y - \theta_0^* - \theta_a^* A - \theta_m^* W) \right]$$

We can rewrite the right-hand side of the equation as the three separate expected values shown below:

$$E[Y - \theta_0^* - \theta_a^* A - \theta_m^* W] = 0$$

$$E[A(Y - \theta_0^* - \theta_a^* A - \theta_m^* W)] = 0$$

$$E[W(Y - \theta_0^* - \theta_a^* A - \theta_m^* W)] = 0$$

Considering only the second equality:

$$E[(\alpha_0 + \alpha_w W + \Delta)(Y - \theta_0^* - \theta_a^* A - \theta_m^* W)] = 0$$

$$E[\alpha_0(Y - \theta_0^* - \theta_a^* A - \theta_m^* W) + \alpha_w W(Y - \theta_0^* - \theta_a^* A - \theta_m^* W) + \Delta(Y - \theta_0^* - \theta_a^* A - \theta_m^* W)] = 0$$

$$\alpha_0 E[(Y - \theta_0^* - \theta_a^* A - \theta_m^* W)] + \alpha_w E[W(Y - \theta_0^* - \theta_a^* A - \theta_m^* W)] + E[\Delta(Y - \theta_0^* - \theta_a^* A - \theta_m^* W)] = 0$$

From the first and the third equality, we have:

$$E[\Delta(Y - \theta_0^* - \theta_a^* A - \theta_m^* W)] = 0$$

$$E[\Delta(Y - \theta_0^* - \theta_a^*(\alpha_0 + \alpha_w W + \Delta) - \theta_m^* W)] = 0$$

$$E[\Delta(Y - \theta_0^* - \theta_a^* \alpha_0 - \theta_a^* \alpha_w W - \theta_a^* \Delta - \theta_m^* W)] = 0$$

$$E[\Delta(Y - \theta_a^* \Delta)] + \theta_0^* E[\Delta] - \theta_a^* \alpha_0 E[\Delta] - E[(\theta_a^* \alpha_w - \theta_m^*) W \Delta] = 0$$

Even here, $E[\Delta] = 0$ and the model errors Δ are independent of the variable W , that is $E[W\Delta] = 0$

$$E[\Delta(Y - \theta_a^* \Delta)] - (\theta_a^* \alpha_w - \theta_m^*) E[W\Delta] = 0$$

$$E[\Delta(Y - \theta_a^* \Delta)] = 0$$