IMPACT OF MODEL MISSPECIFICATION ON PROPORTION MEDIATED IN CAUSAL MEDIATION ANALYSIS

Yi Li[1], Maya B. Mathur[2], Robert W. Platt[1], Kazuki Yoshida[3]

[1] *McGill University*, [2] *Stanford University*, [3] *Brigham and Women's Hospital & Harvard Medical School*

Background and Aim

Background. In causal mediation analysis, researchers may encounter circumstances where natural direct and indirect effects are in opposite directions, rendering proportion mediated (PM) outside [0, 1] (inconsistent mediation). Though this is possible under the true data generating process (DGP), it can also be caused by model misspecification.

Aim. We conduct a simulation study to show how neglecting EMM impacts the range and the magnitude of bias in PM.

Methods

We use the effect measure modification (EMM)-extended regression-based causal mediation approach[1,2].

Treatment A, Mediator M, Outcome Y, Covariate C.

Mediator model: $g_M[E(M|A,C)] = \beta_0 + \beta_1 a + \beta_2 c + \beta_3 ac;$

Outcome model: $g_Y[E(Y|A,M,C)] = \theta_0 + \theta_1 a + \theta_3 am + \theta_4 c + \theta_5 ac + \theta_6 mc.$

g_M, g_Y are link functions.

Natural direct effect (NDE) = $E[Y_{a,M_a*} \mid C] - E[Y_{a*,M_a*} \mid C]$,
Natural indirect effect (NIE) = $E[Y_{a,M_a} \mid C] - E[Y_{a,M_a*} \mid C]$,
where a^* and a are reference exposure level and the level of interest. Here we assign $a^* = 0$ and $a = 1$.

Data Generating Process

1. **Generation of variables.** A, M, Y, C are generated as below. M_{cont} and M_{bin} represent continuous mediator and binary mediator, respectively. Y_{cont} and Y_{bin} represent continuous outcome and binary outcome, respectively.

$$C \sim N(1, \sigma = 0.5),$$
$$A \sim Ber(0.2C + 0.5C^2),$$
$$M_{cont} = \beta_1 + \beta_2 A + \beta_3 C + \beta_4 A \cdot C + \epsilon,$$
$$M_{bin} \sim Ber(\beta_0 + \beta_1 A + \beta_2 C + \beta_3 A \cdot C + \epsilon),$$
$$Y_{cont} = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 A \cdot M + \theta_4 C + \theta_5 A \cdot C + \theta_6 M \cdot C + \epsilon,$$
$$Y_{bin} \sim Ber(\theta_0 + \theta_1 A + \theta_2 M + \theta_3 A \cdot M + \theta_4 C + \theta_5 A \cdot C + \theta_6 M \cdot C + \epsilon),$$

where $\epsilon \sim N(0, \sigma = 0.05)$.

2. **Sign of coefficients.** We consider only the cases when not all β's and θ's are positive or negative, otherwise PM will always be inside [0, 1] and therefore no need to run simulation. We focus on varying the signs of β_1, θ_1, and θ_2 because they are main effects. Specially, four coefficient settings are:
(1) $\beta_2 > 0, \theta_2 > 0, \theta_3 < 0;$
(2) $\beta_2 > 0, \theta_2 < 0, \theta_3 > 0;$
(3) $\beta_2 < 0, \theta_2 > 0, \theta_3 < 0;$
(4) $\beta_2 < 0, \theta_2 < 0, \theta_3 > 0.$

3. **Sampling of coefficients.** We randomly sample β_1, θ_1, and θ_2 from `seq(-10, 0, 0.2)` or `seq(0, 10, 0.2)`, based on the coefficients sign above. All the other coefficients can be either positive or negative and are sampled from `seq (-10, 10, 0.2)`.

4. **True effects, correct and misspecified models.** First, we obtain the true PM by the closed-form formulas[2]. Then we fit correctly specified and misspecified models, respectively. In misspecified models, we omit one, two or all three EMM terms in mediator and outcome models.

NDE and NIE are conditional on $C = 1$.

Results

For each mediator and outcome model type (mediator and outcome can be either linear or logistic), we iterate the DGP 1000 times, each time with different coefficients in mediator and outcome models. In each DGP, sample size $N = 2000$.

Main parameters

- $\beta_1 > 0, \theta_1 > 0, \theta_2 < 0$
- $\beta_1 > 0, \theta_1 < 0, \theta_2 > 0$
- $\beta_1 < 0, \theta_1 > 0, \theta_2 < 0$
- $\beta_1 < 0, \theta_1 < 0, \theta_2 > 0$

Figure 1. Probability of having consistent true PM but inconsistent estimated PM

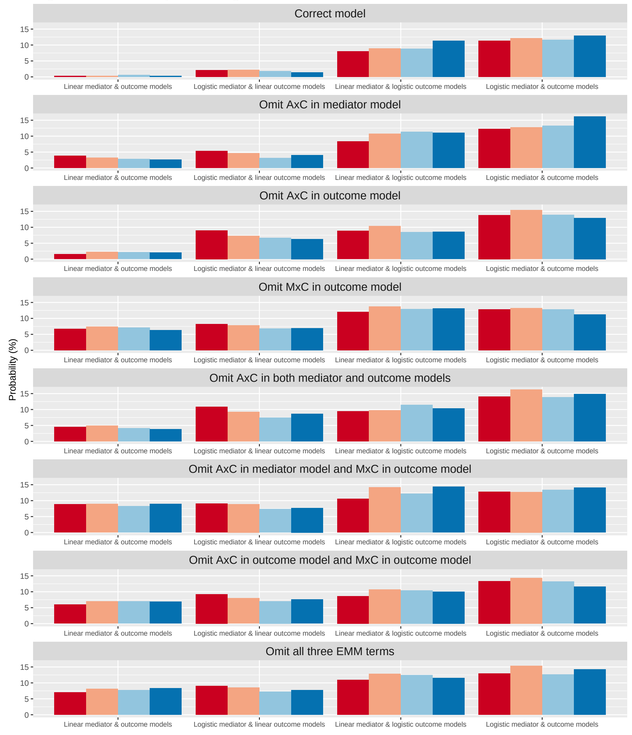

Figure 1 shows the conditional probability of having estimated PM outside $[0, 1]$ when true PM is inside $[0, 1]$, *i.e.* P(estimated PM < 0 or > 1 | $0 \leq$ true PM ≤ 1). For linear outcome model, omitting EMM terms increases the probability of having PM outside $[0, 1]$ by up to 11% (Columns 1 & 2). For logistic outcome model, even the correct model gives up to 12 % inconsistent PM due to random variability, and the change is less noticeable when EMM terms are omitted. Overall, omitting $M \times C$ (Panels 4, 6, 7 & 8) has higher probability of generating inconsistent mediation than other scenarios.

Figure 2 shows the absolute difference between estimated PM and true PM, when both PMs are inside $[0, 1]$. When both true and estimated PMs are inside $[0, 1]$, the error tends to be larger when mediator model is linear (Columns 1 & 3), and the error can span the whole range (from -100% to +100 %) when outcome model is logistic, even when there is no model misspecification (Columns 3 & 4). Again, omitting $M \times C$ (Panels 4, 6, 7 & 8) has larger error than the other scenarios. Across four parameter settings, when $\beta_1 > 0, \theta_1 > 0, \theta_2 < 0$ (red), the error is larger than in the other three scenarios.

Results (cont.)

Figure 2. Absolute error of estimated PM compared to true PM, when both PMs are consistent

◆ represents mean.

Conclusions

1. Omitting EMM terms can lead to high error in proportion mediated, ranging from -100% to 100%, and even inconsistent proportion mediated.
2. Omitting mediator-covariate EMM term performs worse than omitting exposure-covariate EMM terms.
3. Logistic outcome model is more prone to large error and inconsistent proportion mediated than linear outcome model, even when there is no model misspecification.

References

[1] Valeri L. and VanderWeele TJ. "Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros." In: *Psychological methods* 18.2 (2013), p. 137.

[2] Li Y. et al. "Effect Measure Modification by Covariates in Mediation: Extending Regression-Based Causal Mediation Analysis". In: *OSF Preprints* (2022).