Mediation analysis with the mediator and outcome missing not at random

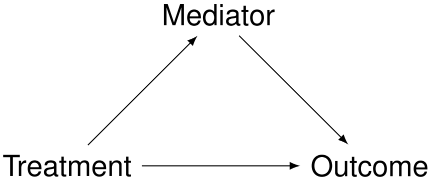
Fan Yang

Yau Mathematical Sciences Center, Tsinghua University

Joint work with
Shuozhi Zuo and Debashis Ghosh (University of Colorado)
Peng Ding (University of California, Berkeley)

Motivation

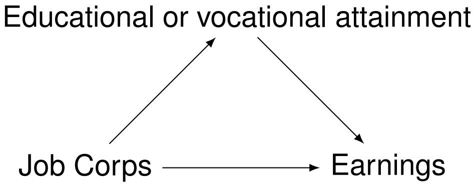
- Mediation analysis: a useful and widely adopted approach for investigating the direct and indirect causal pathways through which an effect arises.



- However, many mediation studies are challenged by missingness in the mediator and/or the outcome.

A motivating study: National Job Corps Study (NJCS)

- Job Corps: the largest education and training program for 16-24 year old disadvantaged youths administered by the U.S. Department of Labor.
- A research question:



A motivating study: National Job Corps Study (NJCS)

Table 1: Missingness patterns in the mediator and the outcome

Mediator	Outcome	Treatment N (%)	Control N (%)
Missing	Observed	$545\ (10.72\%)$	$361\ (9.96\%)$
Observed	Missing	$538\ (10.58\%)$	$400\ (11.04\%)$
Missing	Missing	$497\ (9.78\%)$	$426\ (11.76\%)$
Observed	Observed	$3504\ (68.92\%)$	$2436\ (67.24\%)$
Total Number of Subjects		5084	3623

- Concern: the missingness may be missing not at random (MNAR).
- Example: the missingness in the mediator may depend on whether or not the subject received certificate.

Literature on mediation analysis with missing data

- Challenge: the underlying data distribution can not be identified in general without further assumptions if MNAR.
- Most of the previous literature assume either missingness completely at random or missingness at random. (Enders et al., 2013; Zhang and Wang, 2013; Wu and Jia, 2013; Qin et al., 2021)
- Consider missingness in outcomes only, Li and Zhou (2017) utilized an instrumental variable type of covariate to identify the direct and indirect effect when the missingness in the outcome is MNAR.

Outline

Notation

- T: treatment assignment, $t = 1$ if assigned to the experimental group; $t = 0$ otherwise.
- $M(t)$: potential mediator value under treatment condition t.
- $Y(t, M(t))$, or equivalently, $Y(t)$: potential outcome value under treatment condition t.
- X : vector of measured covariates values.
- M and Y: observed value of the mediator and the outcome.

Mediation analysis without missing data

$$\text{ATE} = \text{NIE} + \text{NDE}$$

$$\text{ATE} = \text{E}[Y(1) - Y(0)] \equiv \text{E}[Y(1, M(1)) - Y(0, M(0))]$$

$$\text{NIE} = \text{E}[Y(1, M(1)) - Y(1, M(0))]$$

$$\text{NDE} = \text{E}[Y(1, M(0)) - Y(0, M(0))]$$

- Sequential Ignorability Assumption (Imai et al., 2010a,b):

  For $t, t' \in \{0, 1\}$,

$$\{Y(t', m), M(t)\} \perp\!\!\!\perp T \mid X = x$$

$$Y(t', m) \perp\!\!\!\perp M(t) \mid T = t, X = x$$

Mediation analysis without missing data

- Nonparametric identification result under sequential ignorability assumption when there exists no missing data:

$$\mathrm{E}[Y(t, M(t'))|X = x] = \int_{\mathcal{M}} \mathrm{E}[Y \mid T = t, M = m, X = x] \, dF(m \mid T = t', X = x)$$

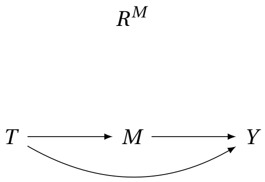
- When there exists missing data, the key would be to identify:
  1. $P(Y = y \mid T = t, M = m, X = x)$
  2. $P(M = m \mid T = t, X = x)$

  Or equivalently, $P(Y = y, M = m \mid T = t, X = x)$

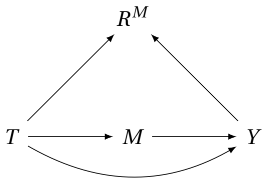
Outline

Missingness mechanisms: MCAR and MAR

R^M: missingness indicator of M, 1 if observed and 0 otherwise.



MCAR



MAR

- All graphs condition on X.
- If $R^M \perp\!\!\!\perp M, Y, T, X$, the missingness is MCAR.
- If $R^M \perp\!\!\!\perp M \mid Y, T, X$, the missingness is MAR.

However, in the Job Corps study, we are concerned that people who failed to obtain an educational or vocational certificate were less likely to report compared to people who successfully obtained an educational or vocational certificate.

Proposed MNAR mechanism

$$
\begin{tikzcd}
 & R^M & \\
T \arrow[r] \arrow[ur] \arrow[rr, bend right] & M \arrow[u] \arrow[r] & Y
\end{tikzcd}
$$

- MNAR Assumption I: $R^M \perp\!\!\!\perp Y \mid (M, T, X)$.
- Allows the missingness R^M to depend on the missing value M.
- Since the outcome Y occurs later, it is plausible in many studies to assume that the missingness of the mediator is conditionally independent of the outcome.

Identifiability under MNAR Assumption I

1. $P(Y = y \mid M = m, T = t, X = x) = P(Y = y \mid M = m, T = t, R^M = 1, X = x)$.
2. Define

$$P_{my1|t,x} = P(M = m, Y = y, R^M = 1 \mid T = t, X = x),$$

$$P_{+y0|t,x} = P(Y = y, R^M = 0 \mid T = t, X = x).$$

Then

$$\begin{aligned} P_{+y0|t,x} &= \sum_{m \in \mathcal{M}} P(M = m, Y = y, R^M = 0 \mid T = t, X = x) \\ &= \sum_{m \in \mathcal{M}} P_{my1|t,x} \frac{P(R^M = 0 \mid M = m, T = t, X = x)}{P(R^M = 1 \mid M = m, T = t, X = x)}. \end{aligned}$$

Note that if the ratios are identifiable, then, $P(M = m \mid T = t, X = x)$ can be identified by

$$P(M = m \mid T = t, X = x) = \frac{P(M = m, R^M = 1 \mid T = t, X = x)}{P(R^M = 1 \mid M = m, T = t, X = x)}.$$

Identifiability under MNAR Assumption I

When will the ratios be identifiable?

$$P_{+y0|t,x} = \sum_{m \in \mathcal{M}} P_{my1|t,x} \frac{P(R^M = 0 \mid M = m, T = t, X = x)}{P(R^M = 1 \mid M = m, T = t, X = x)}.$$

- The ratios are identifiable if the above system of linear equations has full rank, which essentially requires that
  1. The number of elements in the support of Y is not smaller than the number of elements in the support of M.
  2. $M \not\perp\!\!\!\perp Y \mid T, X$.

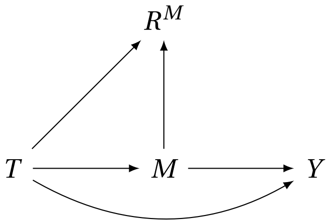
Completeness condition

- Completeness: Define a function $f(A,B)$ to be complete in B if $\int g(A)f(A,B)d\nu(A) = 0$ implies $g(A) = 0$ almost surely for any square-integrable function g. Here, $\nu(\cdot)$ denotes a generic measure.
- The assumption of completeness is routinely made in nonparametric identification problems.
- The completeness condition holds under some frequently used parametric models, such as exponential families of distributions (Newey and Powell, 2003) and a class of location-scale distribution families (Hu and Shiu, 2018).

Identifiability under MNAR Assumption I

Theorem 1

Under sequential ignorability and MNAR Assumption I, if $P(R^M = 1 \mid M = m, T = t, X = x) > 0$ *for all* m, t, x, *and if* $P(Y, M, R^M = 1 \mid T = t, X = x)$ *is complete in* Y *for all* t, x, $P(Y, M \mid T, X)$ *is identifiable, and therefore, the* NIE *and* NDE *are identifiable.*



- Since $P(Y \mid M, T, X)$ can be identified without completeness, when $M \perp\!\!\!\perp Y \mid T, X$, we have $P(Y \mid M, T, X) = P(Y \mid T, X)$, $\mathrm{NIE} = 0$ and $\mathrm{NDE} = \mathrm{ATE} = \int_{\mathscr{X}} [\mathbb{E}(Y \mid T = 1, X = x) - \mathbb{E}(Y \mid T = 0, X = x)]\, dF(x)$.

Simulation study under MNAR Assumption I

Four setups with different relationships in the supports of M and Y:
(A) binary M and binary Y, (B) binary M and continuous Y,
(C) continuous M and continuous Y, (D) continuous M and binary Y.

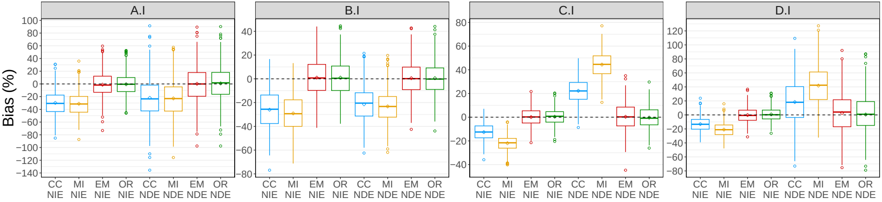
- $X \sim \mathrm{N}(0,1),\ T \sim \mathrm{Bernoulli}(0.5)$
- $M:\mathrm{logit}\ P(M=1 \mid T,X) = \alpha_0 + \alpha_t T + \alpha_x X,\ M \sim \mathrm{N}(\alpha_0 + \alpha_t T + \alpha_x X, 1)$
- $Y:\mathrm{logit}\ P(Y=1 \mid M,T,X) = \beta_0 + \beta_m M + \beta_t T + \beta_{mt} M \cdot T + \beta_x X,$
  $Y \sim \mathrm{N}(\beta_0 + \beta_m M + \beta_t T + \beta_{mt} M \cdot T + \beta_x X, 1)$
- $R^M:\mathrm{logit}\ P(R^M=1 \mid M,T,X) = \lambda_0 + \lambda_m M + \lambda_t T + \lambda_x X$
- Missing rates in M: $20 \sim 25\%$ (with $\lambda_m \neq 0$)
- 500 simulated data sets with sample size 1000

Methods we compare:

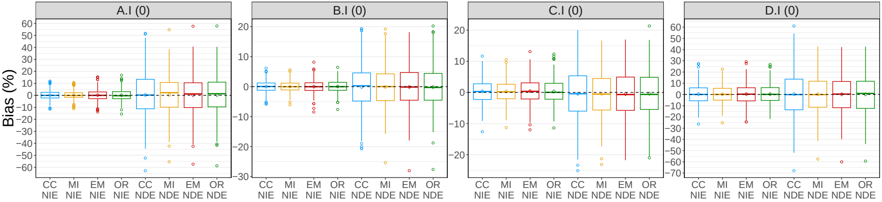
1. Complete Case: using subjects without missing data
2. Multiple Imputation: by chained equations assuming MAR
3. Our EM Algorithm: incorporating the MNAR mechanism
4. Oracle: with true values of the missing data

Simulation results under MNAR Assumption I

- $M \not\!\perp\!\!\!\perp Y \mid T, X$

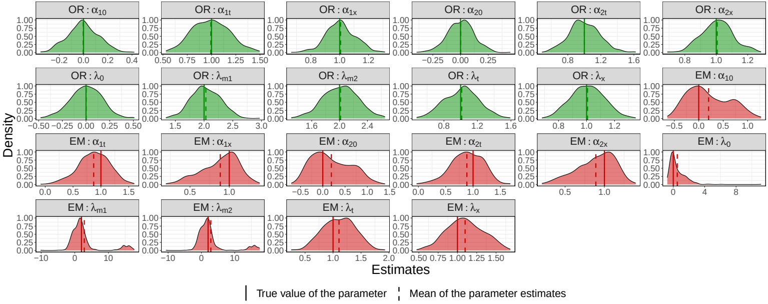


- $M \perp\!\!\!\perp Y \mid T, X$



Simulation results under MNAR Assumption I

- The distribution of parameters may display multimodality and other irregular patterns when the nonparametric identification can not be achieved.
- M has more categories than Y where M is generated by a multinomial logistic regression model and Y is generated by a logistic regression model.



Simulation results on parameters in the M and R^M models.

Outline

Missingness mechanisms: MCAR and MAR

R^Y: the missingness indicator of Y, 1 if observed and 0 otherwise.

$$\begin{tikzcd} & R^M \arrow[r] & R^Y \\ T \arrow[r] \arrow[rr, bend right] & M \arrow[r] & Y \end{tikzcd}$$

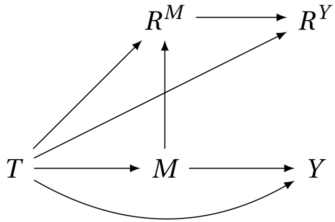
MCAR

$$\begin{tikzcd} & R^M \arrow[r] & R^Y \\ T \arrow[ur] \arrow[urr] \arrow[r] \arrow[rr, bend right] & M \arrow[r] & Y \end{tikzcd}$$

MAR

- If $(R^M, R^Y) \perp\!\!\!\perp (M, Y, T, X)$, the missingness is MCAR.
- If $(R^M, R^Y) \perp\!\!\!\perp (M, Y) \mid T, X$, the missingness is MAR.

MNAR Assumption II



- MNAR Assumption II: $(R^Y, R^M) \perp\!\!\!\perp Y \mid M, T, X$ and $R^Y \perp\!\!\!\perp M \mid R^M, T, X$.
- MAR is a special case of the proposed MNAR Assumption II.

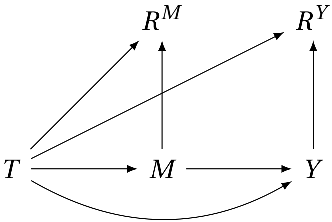
Identification Result under MNAR Assumption II

Theorem 2

Under sequential ignorability and MNAR Assumption II, if $P(R^M = 1, R^Y = 1 \mid M = m, T = t, X = x) > 0$ *and* $P(R^M = 0, R^Y = 1 \mid M = m, T = t, X = x) > 0$ *for all* m, t, x, *and if the* $P(Y, M, R^M = 1, R^Y = 1 \mid T = t, X = x)$ *is complete in* Y *for all* t, x, *the* $P(Y, M \mid T, X)$ *is identifiable, and therefore, the* NIE *and* NDE *are identifiable.*

- The completeness is only used to identify $P(M|T,X)$, and therefore, when $M \perp\!\!\!\perp Y \mid T, X$, we have $P(Y \mid M, T, X) = P(Y \mid T, X)$, $\mathrm{NIE} = 0$ and $\mathrm{NDE} = \mathrm{ATE}$.

MNAR Assumption III



- MNAR Assumption III: $R^Y \perp\!\!\!\perp (R^M, M) \mid Y, T, X$ and $R^M \perp\!\!\!\perp (R^Y, Y) \mid M, T, X$.
- It allows both the missingness of M and the missingness of Y to depend on the missing value itself.

Identification Result under MNAR Assumption III

Theorem 3

Under sequential ignorability and MNAR Assumption III, if
$P(R^M = 1, R^Y = 1 \mid Y = y, M = m, T = t, X = x) > 0$,
$P(R^M = 0, R^Y = 1 \mid Y = y, M = m, T = t, X = x) > 0$ *and*
$P(R^M = 1, R^Y = 0 \mid Y = y, M = m, T = t, X = x) > 0$ *for all* y, m, t, x, *and if*
$P(Y, M, R^M = 1, R^Y = 1 \mid T = t, X = x)$ *is complete in* Y *for all* t, x, *and*
$P(Y, M, R^M = 1, R^Y = 1 \mid T = t, X = x)$ *is also complete in* M *for all* t, x,
$P(Y, M \mid T, X)$ *is identifiable, and therefore, the* NIE *and* NDE *are identifiable.*

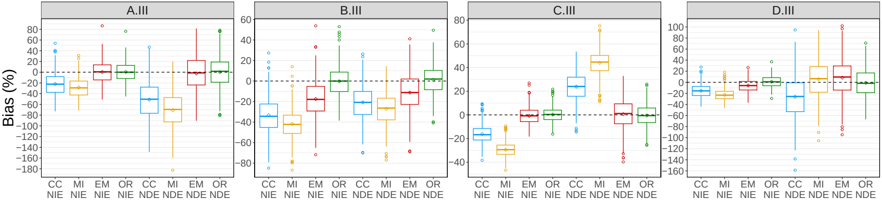
- Different from before, the identification of both $P(Y|M, T, X)$ and $P(M|T, X)$ rely on the above completeness conditions, and therefore, when $M \perp\!\!\!\perp Y \mid T, X$, the NIE and NDE are no longer identifiable.

Simulation study under MNAR Assumption III

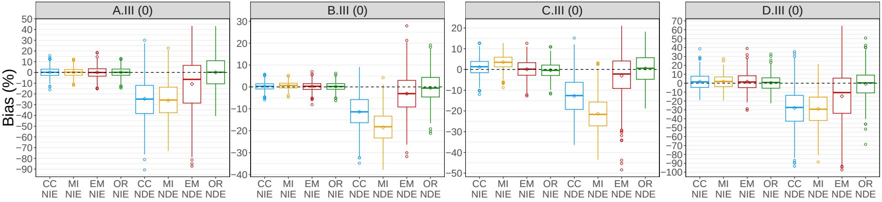
- $R^M : \text{logit } P(R^M = 1 \mid M, T, X) = \lambda_0 + \lambda_m M + \lambda_t T + \lambda_x X$
- $R^Y : \text{logit } P(R^Y = 1 \mid Y, T, X) = \gamma_0 + \gamma_y Y + \gamma_t T + \gamma_x X$
- Missing rates in M and Y: $20 \sim 25\%$ (with $\lambda_m \neq 0, \gamma_y \neq 0$)

Simulation results under MNAR Assumption III

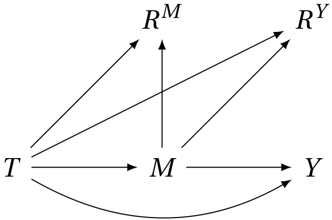
- $M \not\perp\!\!\!\perp Y \mid T, X$



- $M \perp\!\!\!\perp Y \mid T, X$



MNAR Assumption IV



- MNAR Assumption IV: R^Y, R^M and Y are mutually independent given M, T, X.
- The mediator M drives the missingness in both M and Y given T, X.

Identification Result under MNAR Assumption IV

Theorem 4

Under sequential ignorability and MNAR Assumptions IV, if either of the following two conditions holds, the joint distribution $P(Y, M \mid T, X)$ is identifiable, and therefore, the NIE *and* NDE *are identifiable:*

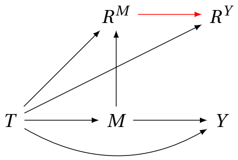
(i) $P(R^M = 1, R^Y = 1 \mid M = m, T = t, X = x) > 0$,
$P(R^M = 0, R^Y = 1 \mid M = m, T = t, X = x) > 0$ *and*
$P(R^M = 1, R^Y = 0 \mid M = m, T = t, X = x) > 0$ *for all* m, t, x, *and*
$P(Y, M, R^M = 1, R^Y = 1 \mid T = t, X = x)$ *is complete in* Y *for all* t, x;

(ii) $P(R^M = 1 \mid M = m, T = t, X = x) > 0$ *for all* m, t, x *and*
$P(M, R^M = 1, R^Y \mid T = t, X = x)$ *is complete in* R^Y *for all* t, x.

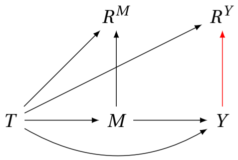
- The completeness is only used to identify $P(M|T, X)$, and therefore, when $M \perp\!\!\!\perp Y \mid T, X$, we have $P(Y \mid M, T, X) = P(Y \mid T, X)$, $\mathrm{NIE} = 0$ and $\mathrm{NDE} = \mathrm{ATE}$.

MNAR mechanisms: a summary of identification

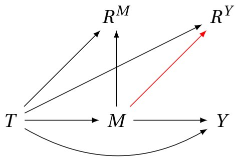
- Assuming Y is not affecting R^M, and allowing M to have an impact on R^M, we have shown that identification of NIE and NDE can be achieved under some completeness assumptions when R^Y only depend on one of (R^M, Y, M) given T, X.



MNAR Assumption II



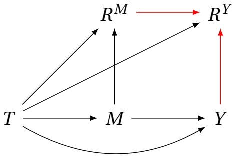
MNAR Assumption III



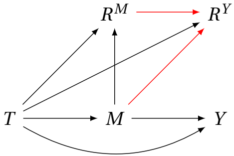
MNAR Assumption IV

MNAR mechanisms: a summary of identification

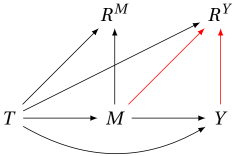
- When R^Y depend on more than one of (R^M, Y, M) given T, X, the identification of NIE and NDE cannot be achieved without further assumptions.



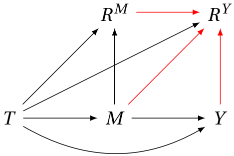
(i) unidentifiable case



(ii) unidentifiable case



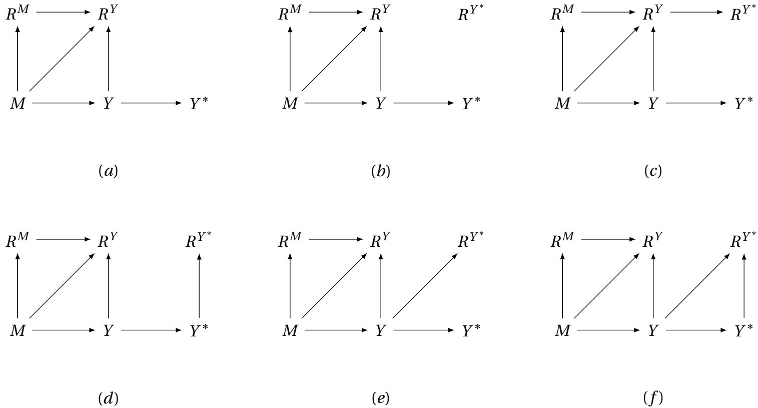
(iii) unidentifiable case



(iv) unidentifiable case

Improved identifiability with a future outcome

- We provide some scenarios where the identification is plausible under the unidentifiable cases by exploiting the information on a future outcome (Y^*).

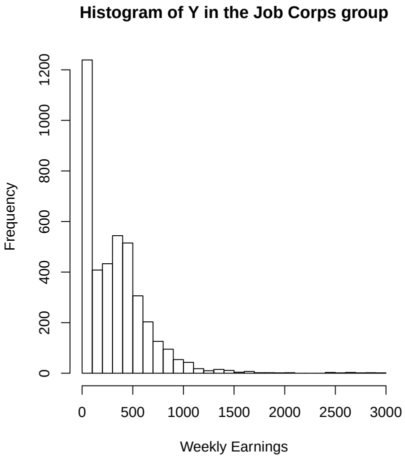


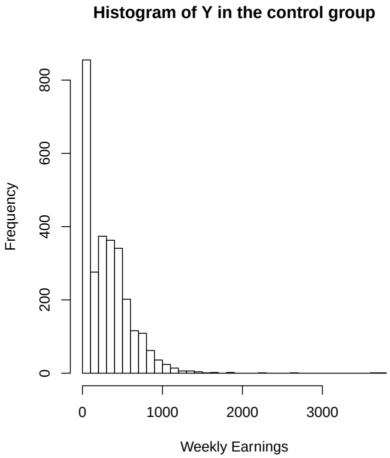
Outline

National Job Corps Study (NJCS)

- The data describes $8,707$ subjects who were randomized to either Job Corps group ($T = 1$) or control group ($T = 0$).
- M: whether subject obtained an educational/vocational certificate or not (collected at 30-month followup); 1 if obtained a certificate, and 0 otherwise.
- Y: weekly earnings four years after randomization.
- Missingness both in M and in Y.
- X: gender, age, race, education level, earnings in the year before participating in the study, whether the subject had a child or not, and whether the subject had ever been arrested or not.

Outcome distribution





- Use two-part models to address the excessive zero values and skewed positive values of earnings.

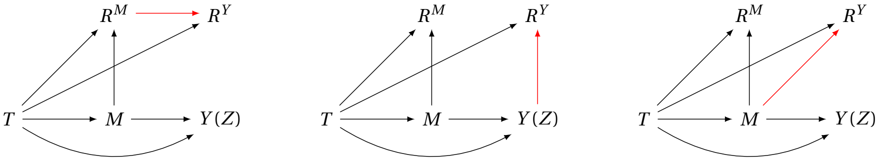
Models

- Model for M: $\text{logit } P(M_i = 1|T_i, X_i) = \alpha_0 + \alpha_t T_i + \alpha_x^T X_i$
- Model for Y: define $Z_i = 0$ if $Y_i = 0$, and $Z_i = 1$ if $Y_i > 0$. Two-part Gamma model for Y with log link:
  $\text{logit } P(Z_i = 1 \mid M_i, T_i, X_i) = \delta_0 + \delta_m M_i + \delta_t T_i + \delta_{mt} M_i \cdot T_i + \delta_x^T X_i$,
  $Y_i \mid Z_i = 1, M_i, T_i, X_i \sim \text{Gamma}(\nu, \nu/\mu_i(M_i, T_i, X_i))$, where
  $\mu_i(M_i, T_i, X_i) = \exp(\beta_0 + \beta_m M_i + \beta_t T_i + \beta_{mt} M_i \cdot T_i + \beta_x^T X_i)$
- Model for R^M: $\text{logit } P(R_i^M = 1|M_i, T_i, X_i) = \lambda_0 + \lambda_m M_i + \lambda_t T_i + \lambda_x^T X_i$
- Model for R^Y:
  Under Assumption II:
  $\text{logit } P(R_i^Y = 1|R_i^M, T_i, X_i) = \gamma_0 + \gamma_{r^M} R_i^M + \gamma_t T_i + \gamma_x^T X_i$;
  Under Assumption III:
  $\text{logit } P(R_i^Y = 1|Z_i, T_i, X_i) = \gamma_0 + \gamma_z Z_i + \gamma_t T_i + \gamma_x^T X_i$;
  Under Assumption IV:
  $\text{logit } P(R_i^Y = 1|M_i, T_i, X_i) = \gamma_0 + \gamma_m M_i + \gamma_t T_i + \gamma_x^T X_i$.

Model comparison

Table 2: Model comparison among models under MNAR Assumptions II, III and IV using two-part Gamma and log-normal models. The log-likelihoods are evaluated at the corresponding MLEs.

Assumption	Model	Log-likelihood	NIE	NDE
II	Gamma	$-53131.35\checkmark$	$10.94\ (7.94,\ 14.29)$	$12.93\ (-1.95,\ 27.64)$
III	Gamma	-53488.54	$14.87\ (11.59,\ 18.35)$	$9.99\ (-2.99,\ 22.73)$
IV	Gamma	-53475.01	$10.14\ (7.26,\ 13.25)$	$11.29\ (-3.85,\ 26.21)$
II	Log-normal	-53799.79	$15.50\ (11.06,\ 20.16)$	$4.14\ (-18.14,\ 26.42)$
III	Log-normal	-54159.23	$19.23\ (14.79,\ 23.81)$	$3.21\ (-15.37,\ 21.74)$
IV	Log-normal	-54145.92	$14.36\ (10.16,\ 18.71)$	$1.68\ (-20.82,\ 24.21)$



MNAR Assumption II　　MNAR Assumption III　　MNAR Assumption IV

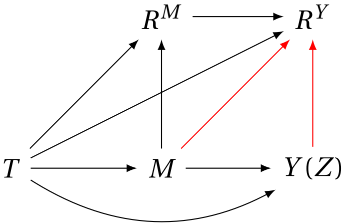
Results: Gamma under MNAR Assumption II

Table 3: CI based on 500 bootstrap samples; λ_m, coefficient of M in the R^M model; γ_{r^M}, coefficient of R^M in the R^Y model.

	Complete Case		Multiple Imputation		EM Algorithm	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
λ_m	NA	NA	NA	NA	1.73	(0.34, 3.33)
γ_{r^M}	NA	NA	NA	NA	1.87	(1.76, 2.00)
NIE	12.00	(8.65, 15.57)	12.04	(8.25, 14.60)	10.94	(7.94, 14.29)
NDE	14.75	(−0.05, 29.50)	9.22	(−5.85, 23.23)	12.93	(−1.95, 27.64)

- The causal conclusions regarding the NIE and NDE are the same among the three methods, in spite of the significant effect of $M \rightarrow R^M$ (λ_m) and the significant effect of $R^M \rightarrow R^Y$ (γ_{r^M}).

Sensitivity analysis



Missing data mechanism for the sensitivity analysis.

- Revised model for R^Y: $\text{logit } P(R^Y = 1 \mid R^M = r^M, Z = z, M = m, T = t, X = x) = \gamma_0 + \gamma_{r^M} r^M + \gamma_z z + \gamma_m m + \gamma_t t + \gamma_x^T x$, where γ_z and γ_m are sensitivity parameters.
- When $\gamma_z = \gamma_m = 0$, it is the same as the previous analysis under MNAR Assumption II.

Sensitivity analysis

Table 4: Est, estimate; CI, confidence interval based on 500 bootstrap samples; γ_z (sensitivity parameter), coefficient of Z in the R^Y model; γ_m (sensitivity parameter), coefficient of M in the R^Y model.

		$\gamma_z = -2$		$\gamma_z = 0$		$\gamma_z = 2$	
	γ_m	Est	95% CI	Est	95% CI	Est	95% CI
NIE	-2	11.15	(7.97, 14.49)	11.49	(8.24, 14.83)	14.33	(11.02, 17.89)
	0	11.30	(8.12, 14.58)	10.94	(7.94, 14.29)	13.40	(10.22, 16.78)
	2	11.39	(8.19, 14.63)	10.83	(7.98, 14.25)	10.48	(7.15, 14.81)
NDE	-2	13.18	(−1.53, 27.88)	13.90	(−1.00, 28.57)	11.72	(−2.33, 26.34)
	0	12.82	(−1.90, 27.52)	12.93	(−1.95, 27.64)	11.27	(−3.12, 25.57)
	2	12.50	(−2.24, 27.16)	12.25	(−2.38, 27.33)	15.43	(−0.18, 29.47)

- The causal conclusions on the NIE and NDE are not sensitive to a strong impact of $Z \to R^Y$ and/or $M \to R^Y$ in addition to the impact of R^M on R^Y.

Outline

Summary

- Show some positive results on nonparametric identification of NIE and NDE when mediator and/or outcome are MNAR.
- One of our favorite statistics quotes:

  "If an issue can be addressed nonparametrically then it will often be better to tackle it parametrically; however, if it cannot be resolved nonparametrically then it is usually dangerous to resolve it parametrically."

  "Principles of Applied Statistics," Cox and Donnelly (2011)

References

Enders, C. K., Fairchild, A. J., and MacKinnon, D. P. (2013). A bayesian approach for estimating mediation effects with missing data. *Multivariate Behavioral Research*, 48(3):340–369.

Hu, Y. and Shiu, J.-L. (2018). Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory*, 34(3):659–693.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71.

Li, W. and Zhou, X.-H. (2017). Identifiability and estimation of causal mediation effects with missing data. *Statistics in Medicine*, 36(25):3948–3965.

Newey, W. and Powell, J. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Qin, X., Deutsch, J., and Hong, G. (2021). Unpacking complex mediation mechanisms and their heterogeneity between sites in a job corps evaluation. *Journal of Policy Analysis and Management*, 40(1):158–190.

Wu, W. and Jia, F. (2013). A new procedure to test mediation with missing data through nonparametric bootstrapping and multiple imputation. *Multivariate Behavioral Research*, 48(5):663–691.

Zhang, Z. and Wang, L. (2013). Methods for mediation analysis with missing data. *Psychometrika*, 78(1):154–184.