UNIVERSITEIT GENT

FACULTY OF **PSYCHOLOGY AND EDUCATIONAL SCIENCES**

Flexible mediation analysis in the presence of non-linear relations: beyond the mediation formula.

Modern Modeling Methods (M^3) Conference

Beatrijs Moerkerke
Ghent University, Belgium
Beatrijs.Moerkerke@UGent.be

Tom Loeys, Olivia De Smet, Ann Buysse, Johan Steen & Stijn Vansteelandt

May 21, 2013

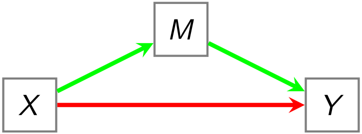
Mediation analysis: Goal?

To unravel causal pathways between exposure X and outcome Y

- What is the effect of X on Y? Total Effect

Mediation analysis: Goal?

To unravel causal pathways between exposure X and outcome Y

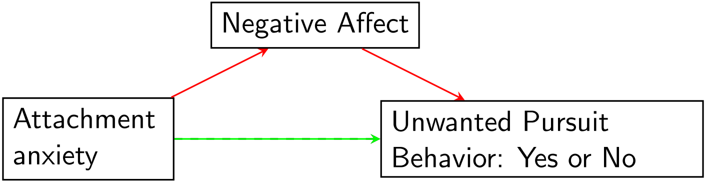


- What part of the effect is mediated by M? Indirect Effect
- What is the remaining causal effect of X on Y? Direct Effect

Example: observational study

Subset from IPOS: big Flemish survey in separating individuals:
(Interdisciplinary Project on the Optimization of Separation Trajectories)

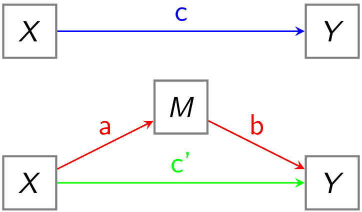
(De Smet, Loeys, & Buysse, 2012)



The effect of attachment anxiety on unwanted pursuit behavior mediated by negative affect? (indirect effect)

The effect of attachment anxiety on unwanted pursuit behavior not mediated by negative affect? (direct effect)

The Baron and Kenny approach



$$E[Y_i \mid X_i] = i_0 + cX_i$$
$$E[M_i \mid X_i] = i_1 + aX_i$$
$$E[Y_i \mid X_i, M_i] = i_2 + c'X_i + bM_i$$

$$Total\ Effect = Direct\ Effect + Indirect\ Effect$$
$$c = c' + a \times b$$

Assumptions

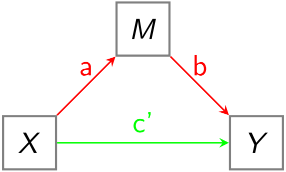
(A1) no unmeasured confounding of the X-M relationship

(A2) no unmeasured confounding of the X-Y relationship

(A3) no unmeasured confounding of the M-Y relationship

(A4) no confounders of the M-Y relationship that are affected by X

Non-linear settings: binary exposure



Example

Logistic regression:

$$\begin{aligned} E[M_i \mid X_i] &= i_1 + aX_i \\ \mathrm{logit}\{E[Y_i \mid X_i, M_i]\} &= i_2 + c'X_i + bM_i \end{aligned}$$

Analysis using the Baron & Kenny approach

$$\begin{aligned}
\text{logit}\{E[Y_i \mid X_i, C_i]\} &= i_0 + cX_i + dC_i \\
E[M_i \mid X_i, C_i] &= i_1 + aX_i + eC_i \\
\text{logit}\{E[Y_i \mid X_i, M_i, C_i]\} &= i_2 + c'X_i + bM_i + fC_i
\end{aligned}$$

- $Y_i = 1$ if showing UPB, else 0
- M_i: (standardized) negative affect
- X_i: (standardized) attachment anxiety
- C_i: baseline covariates: age, gender and education level

	Estimate	standard error	OR (95% CI)
c	0.497	0.112	$1.64(1.32, 2.05)$
c'	0.316	0.121	$1.37(1.08, 1.74)$
$c - c'$	0.181	0.045	
a	0.340	0.048	
b	0.618	0.123	$1.86(1.45, 2.36)$
$a \times b$	0.210	0.054	

$$c - c' \neq a \times b$$

Outline

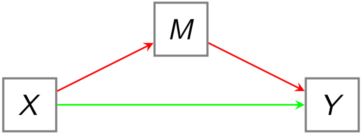
1. Counterfactual framework: natural direct and indirect effects
2. Mediation formula and R package *mediation* (Imai, Keele, & Tingley, D., 2010)
3. Some limitations
4. Natural effects models
5. Case study: IPOS
6. Discussion

Counterfactual framework

- Define measures of direct and indirect effects using *counterfactual outcomes* (Rubin, 2004)
- $M(x)$ and $Y(x)$ denote the mediator and outcome that would have been observed for a subject if X is set to x through some intervention
- $Y(x, M(x^*))$ denotes the outcome that would have been observed if X is set to x and M to the value it would have taken if X is set to x^*

Natural direct and indirect effect

Assume randomized exposure X (0/1)



- the total effect: $E[Y(1)] - E[Y(0)]$
- the natural direct effect: $E[Y(1, M(0))] - E[Y(0, M(0))]$
- the natural indirect effect: $E[Y(0, M(1))] - E[Y(0, M(0))]$

Natural direct and indirect effect

- Non-randomized exposure: conditional on baseline covariates C (think of assumptions (A1), (A2) and (A3) ...)
- More general for continuous exposure:
  The (conditional) natural direct effect

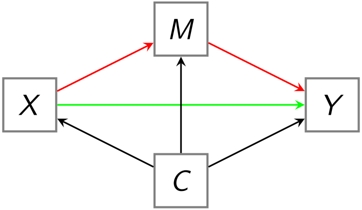
  $$E[Y_i(x, M_i(x^*)) \mid C_i] - E[Y_i(x^*, M_i(x^*)) \mid C_i]$$

  The (conditional) natural indirect effect

  $$E[Y_i(x^*, M_i(x)) \mid C_i] - E[Y_i(x^*, M_i(x^*)) \mid C_i]$$

The mediation formula

Under assumptions (A1) to (A4), i.e.



The *mediation formula* states (Pearl, 2001) that

$$E\{Y(x, M(x^*)) \mid C\} = \sum_m E(Y \mid X = x, M = m, C = c)P(M = m \mid X = x^*, C)$$

⇒ the mediation formula can be used to calculate to obtain analytic expressions for the direct and indirect effect.

The mediation formula

- The mediation formula can be used to obtain analytical expressions for natural direct and indirect effects

  e.g. for linear and logistic models with exposure-mediator interaction (VanderWeele and Vansteelandt, 2009, 2010)
- Only in limited cases closed form expressions available:
  $\rightarrow$ causally defined direct and indirect effects in Mplus (Muthén,2011)
  $\rightarrow$ SAS and SPSS-macros (Valeri & Vanderweele,2013)
- Monte-Carlo approximations add flexibility:
  $\rightarrow$ R-package *mediation* (Imai, Keele, & Tingley, D., 2010)
- But a limitation is that it easily entails complicated results

R-package *mediation*

Monte Carlo draws of potential outcomes:

- Specify 2 models: mediator model and outcome model
- Sample $M(x^*)$ from mediator model
- Given that draw, sample $Y(x, M(x^*))$ from outcome model

Analysis using *mediation* package

```
model1.m<-lm(negaffect~attachment+gender+educ+age)
model1.y<-glm(stalk1~attachment+negaffect+gender+educ+age,
              family=binomial(link="logit"),data=dat)
out1<-mediate(model1.m,model1.y,treat="attachment",
              mediator="negaffect",sims=500,boot=TRUE)
```

direct	$E[Y(1, M(0)) - Y(0, M(0))]$	0.071	(0.023,0.122)
mediation	$E[Y(0, M(1)) - Y(0, M(0))]$	0.047	(0.028,0.069)

Limitations

- Effects expressed on the linear scale
- Default values for ‘control’ and ‘treatment’
- Marginalized over observed covariate distribution
- Testing for moderated mediation?

Limitations - example 1

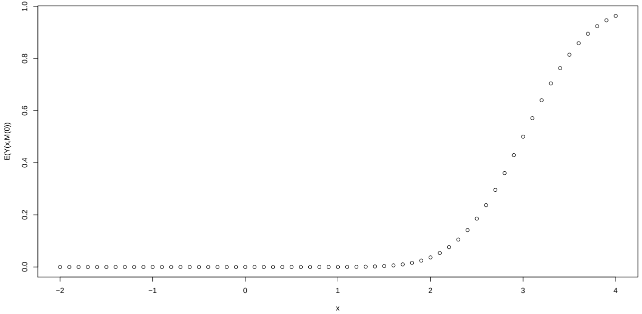
Assume:

- $X \sim N(0, 1)$
- $M \mid X \sim N(\alpha_0 + \alpha_1 X, \sigma_M^2)$
- $\Pr(Y = 1 \mid X, M) = \Phi(\beta_0 + \beta_1 X + \beta_2 M)$ with Φ cumulative normal distribution

$$\Rightarrow E[Y(x, M(x^*)] = \Phi\left(((\beta_0 + \beta_2\alpha_0) + \beta_1 x + \alpha_1\beta_2 x^*)/\sqrt{1 + \beta_2^2\sigma_M^2}\right)$$

Limitations - example 1

With $\alpha_0 = 0$, $\alpha_1 = 1$, $\beta_0 = -6$, $\beta_1 = 2$, $\beta_2 = 0.5$ and $\sigma_M = 1$, the value of $E[Y(x, M(0)]$ as a function of x



$$E[Y(1, M(0))] - E[Y(0, M(0))] \approx 0.0001,$$
$$E[Y(3, M(2))] - E[Y(2, M(2))] \approx 0.642,$$
$$E[Y(3, M(0))] - E[Y(2, M(0))] \approx 0.477$$

Limitations - example 2

Assume:

- 2 correlated baseline confounders $C_{1,i}$ and $C_{2,i}$:
  $C_{1,i} \sim Bern(0.5)$, $C_{2,i} \sim Bern(0.3)$ if $C_{1,i} = 0$ and
  $C_{2,i} \sim Bern(0.9)$ if $C_1 = 1$
- $X_i \sim N(0, 1)$
- $M_i \mid X_i, C_{1,i}, C_{2,i} \sim N(\alpha_0 + \alpha_1 X_i + \alpha_2 C_{1,i} + \alpha_3 C_{2,i}, \sigma_M)$
- $\Pr(Y_i = 1 \mid X_i, M_i, C_{1,i}, C_{2,i}) =$
  $\Phi(\beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 C_{1,i} + \beta_4 C_{2,i})$

Limitations - example 2

with $\alpha_0 = 0$, $\alpha_1 = 1$, $\alpha_2 = 2$, $\alpha_3 = -2$, $\beta_0 = -1$, $\beta_1 = 2$, $\beta_2 = 0.5$, $\beta_3 = 1.5$, $\beta4 = -1.5$, and $\sigma = 1$

mediation package: C_1 stratum specific direct effects:

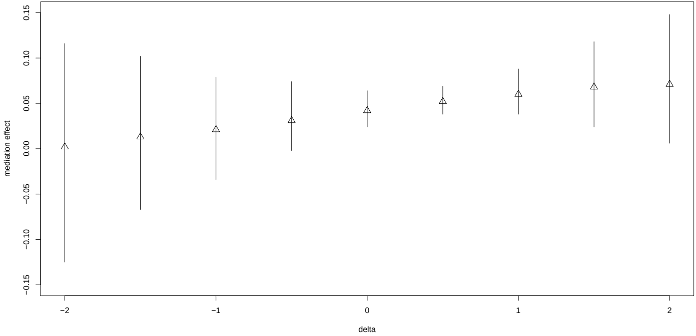
$$\begin{aligned}\sum_{i=1}^{n} E[Y_i(1, M_i(0)) - Y_i(0, M_i(0)) \mid C_{1,i} = 0, C_{2,i}]/n &= 0.2975 \\ \sum_{i=1}^{n} E[Y_i(1, M_i(0)) - Y(0, M_i(0)) \mid C_{1,i} = 1, C_{2,i}]/n &= 0.4107\end{aligned}$$

C_1 stratum specific direct effects, averaged over the observed C_2-distribution in the C_1-specific subsample (size n_0 and n_1):

$$\begin{aligned}\sum_{C_{1,i}=0} E[Y_i(1, M(0)) - Y_i(0, M(0)) \mid C_{1,i} = 0, C_{2,i}]/n_0 &= 0.4554 \\ \sum_{C_{1,i}=1} E[Y_i(1, M(0)) - Y_i(0, M(0)) \mid C_{1,i} = 1, C_{2,i}]/n_1 &= 0.5658\end{aligned}$$

Limitations - example 3

Independent variable - by - mediator interaction?



Estimated mediation effect of negative affect (on a linear scale) with 95% confidence interval at various levels δ of attachment anxiety, i.e. $E[Y(\delta, M(1))] - E[Y(\delta, M(0))]$ versus δ.

Natural effects models

Natural effect models are models for nested counterfactuals

$$g\left[E\{Y_i(x, M_i(x^*)) \mid C\}\right] = \theta' W_i(x, x^*, C_i)$$

(Vansteelandt, S., Bekaert, M. & Lange, T. (2012) Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, 1, 131-158.)

- Linear natural effect model

$$E\{Y_i(x, M_i(x^*)) \mid C_i\} = \theta_0 + \theta_1 x + \theta_2 x^* + \theta_3 C_i$$

Natural direct effect

$$E\left\{Y_i(x+1, M(x)) - Y_i(x, M(x))|C_i\right\} = \theta_1$$

Natural indirect effect

$$E\left\{Y_i(x, M(x+1)) - Y_i(x, M(x))|C_i\right\} = \theta_2$$

- Linear natural effect model with moderation

$$E\{Y_i(x, M_i(x^*)) \mid C_i\} = \theta_0 + \theta_1 x + \theta_2 x^* + \theta_3 C_i + \theta_4 x x^* + \theta_5 x C_i$$

Natural direct effect

$$E\{Y_i(x+1, M_i(x)) - Y_i(x, M_i(x)) | C_i\} = \theta_1 + \theta_4 x + \theta_5 C_i$$

Natural indirect effect

$$E\{Y_i(x, M_i(x+1)) - Y_i(x, M_i(x)) | C_i\} = \theta_2 + \theta_4 x$$

- Logistic natural effect model

$$\mathrm{logit}\{E\{Y_i(x, M_i(x^*)) \mid C_i\}\} = \theta_0 + \theta_1 x + \theta_2 x^* + \theta_3 C_i$$

Natural direct effect

$$\frac{\text{odds}\{Y_i(x+1, M_i(x)) = 1 | C_i\}}{\text{odds}\{Y_i(x, M_i(x)) = 1 | C_i\}} = \exp(\theta_1)$$

Natural indirect effect

$$\frac{\text{odds}\{Y_i(x, M_i(x+1)) = 1 | C_i\}}{\text{odds}\{Y_i(x, M_i(x)) = 1 | C_i\}} = \exp(\theta_2)$$

- Poisson natural effect model

$$\log\{E\{Y_i(x, M_i(x^*)) \mid C_i\}\} = \theta_0 + \theta_1 x + \theta_2 x^* + \theta_3 C_i$$

Natural direct effect

$$\frac{E[Y_i(x+1, M_i(x))]}{E[Y_i(x, M_i(x))]} = \exp(\theta_1)$$

Natural indirect effect

$$\frac{E[Y_i(x, M_i(x+1))]}{E[Y_i(x, M_i(x))]} = \exp(\theta_2)$$

How to fit natural effect models?

- $Y_i(x, M_i(x^*))$ only observed when $x^* = x$, and x^* observed level of X_i
- When $x^* \neq x$, $Y_i(x, M_i(x^*))$ can be predicted from $E(Y_i \mid X_i = x, M_i, C_i)$

  $M_i(x^*) = M_i$ among subjects with x^* observed value of X

Suppose X is dichotomous (0/1)

- for an untreated subject ($X = 0$), we observe

$$Y(0, M(0)) = Y$$

but not

$$Y(1, M(0)) = Y(1, M).$$

- For a treated subject ($X = 1$), we observe

$$Y(1, M(1)) = Y$$

but not

$$Y(0, M(1)) = Y(0, M).$$

Observed data

id	X	x	x^*	$Y(x, M(x^*))$	C
1	0	0	0	Y_1	C_1
1	0	1	0	**?**	C_1
2	1	1	1	Y_2	C_2
2	1	0	1	**?**	C_2
⋮	⋮	⋮	⋮	⋮	⋮

Imputation for the untreated ($X = 0$)

The missing $Y_i(1, M)$ can be predicted

Example

Under model

$$\text{logit}P\left(Y_i = 1 | X_i, M_i, C_i\right) = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 C_i$$

we predict as $Y_i(1, M_i)$ as

$$\text{expit}\left(\beta_0 + \beta_1 + \beta_2 M_i + \beta_3 C_i\right)$$

Imputation for the treated ($X = 1$)

The missing $Y(0, M)$ can be predicted

Example

Under model

$$\text{logit} P(Y_i = 1 | X_i, M_i, C_i) = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 C_i$$

we predict as $Y_i(0, M_i)$ as

$$\text{expit}(\beta_0 + \beta_2 M_i + \beta_3 C_i)$$

Imputed data

id	X	x	x^*	$Y(x, M(x^*))$	C
1	0	0	0	Y_1	C_1
1	0	1	0	$\hat{Y}_1(1, M)$	C_1
2	1	1	1	Y_2	C_2
2	1	0	1	$\hat{Y}_2(0, M)$	C_2
⋮	⋮	⋮	⋮	⋮	⋮

Fit the natural effects model

$$\text{logit}P\left\{Y_i(x, M_i(x^*)) = 1|C_i\right\} = \theta_0 + \theta_1 x + \theta_2 x^* + \theta_3 C_i$$

on the imputed data using standard software

Standard errors based on bootstrap or sandwich estimator

Estimation strategy

(1) Build the outcome model $E(Y_i \mid X_i = x, M_i, C_i)$

(2) Create a new data set by repeating the observed data K times and adding 2 variables:

  (i) x

    First replication: original level of X
    $K - 1$ remaining replications: random draw from $X \mid C$

  (ii) x^* which equals the original level of X

(3) $Y(x, M(x^*))$ predicted by observed Y when $x = x^*$ and by $E(Y \mid X = x, M, C)$ when $x \neq x^*$

(4) Fit natural effects model using predicted $Y(x, M(x^*))$

Analysis using natural effects models

Outcome model:

$$\text{logit}\left[E\{Y_i \mid X_i, M_i, C_i\}\right] = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 C_i$$

Natural effects model:

$$\text{logit}\left[E\{Y_i(x, M_i(x^*)) \mid C_i\}\right] = \theta_0 + \theta_1 x + \theta_2 x^* + \theta_3 C_i$$

θ_1	0.298	(0.077,0.520)
θ_2	0.202	(0.102,0.302)

Analysis using natural effects models - 2

'independent variable-by-mediator' interaction

Outcome model:

$$\text{logit}\left[E\{Y_i \mid X_i, M_i, C_i\}\right] = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 C_i + \beta_4 X_i M_i$$

Natural effects model:

$$\text{logit}\left[E\{Y_i(x, M_i(x^*)) \mid C_i\}\right] = \theta'_0 + \theta'_1 x + \theta'_2 x^* + \theta'_3 C + \theta'_4 x x^*$$

θ'_1	0.285	(0.067,0.502)
θ'_2	0.197	(0.100,0.295)
θ'_4	0.047	(-0.041,0.136)

Discussion

Natural effects models:

- Natural direct and indirect effect each captured by single parameter
- Model direct and indirect effect on the most natural scale
- Moderated mediation easily tested and quantified
- Software in preparation

- As with most imputation methods, a concern may be that the imputation and analysis models are incompatible.

  e.g. if the imputation model excludes $X - C$ interaction,
  it would be inappropriate to investigate if the direct effect is modified by C.

- Number of imputations K