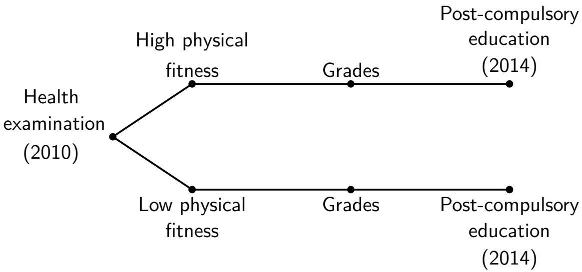
A comparison of 5 software implementations of mediation analysis

Liis Starkopf, Thomas A. Gerds, Theis Lange

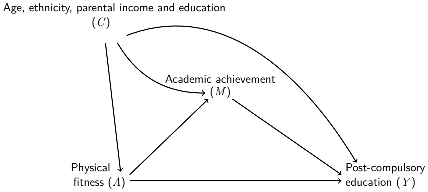
Section of Biostatistics, University of Copenhagen

Illustrative example

- Examine pathways between physical fitness and post-cumpolsory education
- 1084 students from all public elementary schools in Aalborg



Mediation question



- How much of the effect of physical fitness on attendance in post-compulsory education is mediated through academic achievement?

Objectives

- How to do mediation analysis in practice?
  - Implement mediation analysis on the illustrative data using
    - SAS mediation macro
    - R package **medflex**
- What software solutions are available? What can they do? What are the differences?
  - Comparison of five estimation methods and their software solutions
  - Simulation study

Counterfactual framework

General definiton:

$M(a)$	the mediator that would have been observed if, possibly contrary to the fact, the exposure A was set to a
$Y(a, M(a^*))$	the outcome that would have been observed if, possibly contrary to the fact, the exposure A had been set to a and the mediator M was set to the value it would have taken if A was set to a^*.

Counterfactual framework

Example:

$M(0)$	grade point average that would have been observed if, possibly contrary to the fact, the level of physical fitness had been set to low $(A = 0)$
$Y(1, M(0))$	attendance in post-compulsory education that would have been observed if, possibly contrary to the fact, the level of physical fitness had been set to high $(A = 1)$ and grade point average was set to the value it would have taken if the level of physical fitness was set to low $(M(0))$.

Marginal natural effects

$$\underbrace{g\{E[Y(a,M(a))]\} - g\{E[Y(a^*,M(a^*))]\}}_{\text{marginal total effect}}$$
$$= \underbrace{g\{E[Y(a,M(a))]\} - g\{E[Y(a^*,M(a))]\}}_{\text{marginal natural direct effect}}$$
$$+ \underbrace{g\{E[Y(a^*,M(a))]\} - g\{E[Y(a^*,M(a^*))]\}}_{\text{marginal natural indirect effect}}$$

for some link function g.

Identifiability conditions

- No uncontrolled confounding for the exposure-outcome, exposure-mediator or mediator-outcome relations

$$
\begin{aligned}
Y(a,m) \perp\!\!\!\perp A \mid C & \quad \text{for all levels of } a \text{ and } m, \\
M(a) \perp\!\!\!\perp A \mid C & \quad \text{for all levels of } a, \\
Y(a,m) \perp\!\!\!\perp M \mid A = a, C & \quad \text{for all levels of } a \text{ and } m.
\end{aligned}
$$

- No intertwined causal pathways

$$Y(a,m) \perp\!\!\!\perp M(a^*) \mid C \quad \text{for all levels } a, a^* \text{ and } m.$$

- Positivity

$$f(m \mid A, C) > 0 \text{ w.p.1 for each } m.$$

- Consistency

$$
\begin{aligned}
\text{if } A = a, & \quad \text{then } M(a) = M \text{ w.p.1}, \\
\text{if } A = a \text{ and } M = m, & \text{ then } Y(a,m) = Y \text{ w.p.1}.
\end{aligned}
$$

Analytic formulas of natural effects

- Regression model for the outcome Y

$$g_Y\{E[Y \mid A=a, M=m, C=c]\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4^T c$$

- Regression model for the mediator M

$$g_M\{E[M \mid A=a, C=c]\} = \beta_0 + \beta_1 a + \beta_2^T c,$$

- Under identifiability conditions, closed form expressions of natural direct and indirect effects can be derived as a combination of β and θ
- Valeri and VanderWeele [1] implemented the analytic fomulas in SAS/SPSS mediation macros.

[1] Linda Valeri and Tyler J VanderWeele. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with sas and spss macros. *Psychological methods*, 18(2):137, 2013.

SAS/SPSS mediation macro

- Conditional natural effects at a fixed level of C on the scale of linear predictor (g_y)
- Seperate formulas for each combination of the mediator and outcome models
- Binary outcome:
  - For logistic regression model, the formulas hold if the outcome is **rare**
  - Alternatively, log-linear model has to be used

Illustrative example

- Not attending post-compulsory education is a rare event, $P(Y=0)=0.08$
- Logistic regression model for Y

$$\text{logit}\{P(Y=0 \mid A=a, M=m, C=c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_4^T c$$

- Liner regression model for M

$$E[M \mid A=a, C=c] = \beta_0 + \beta_1 a + \beta_2^T c$$

- Resulting formulas

$$\log\{OR_{NDE}(a=1, a^*=0)\} = \theta_1$$
$$\log\{OR_{NIE}(a=1, a^*=0)\} = \beta_1\theta_2$$
$$\log\{OR_{TE}(a=1, a^*=0)\} = \theta_1 + \beta_1\theta_2$$

Implementation

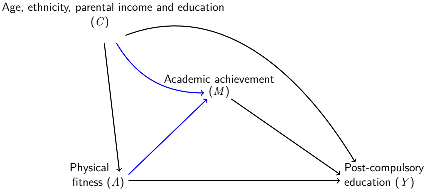
```
%INC "MEDIATION.sas";

PROC IMPORT DATAFILE="d1.csv" OUT=d1 DBMS=csv;
RUN;

%MEDIATION(data=d1,yvar=attend,avar=fitness,mvar=gpa,cvar=ethni age14 age15 income1 income2 income3
            educ1 educ2 educ3,a0=0,a1=1,m=0,nc=4,c=,yreg=logistic,mreg=linear,interaction=false)
RUN;
```

The SAS System

Obs	fitness	gpa	attend	age	income	educ	age14	age15	income1	income2	income3	educ1	ecuc2	educ3
1	0	7.9746053553	0	1	0	2	1	0	0	0	0	0	1	0
2	0	5.4656986857	0	1	1	2	1	0	1	0	0	0	1	0
3	1	7.0326496536	0	1	2	2	1	0	0	1	0	0	1	0
4	0	8.7003049361	0	1	1	3	1	0	1	0	0	0	0	1
5	0	5.4053362093	0	1	1	3	1	0	1	0	0	0	0	1
6	1	7.342120629	0	1	0	2	1	0	0	0	0	0	1	0

Regression for mediator M

```
%MEDIATION(data=d1,
    yvar=attend,
    avar=fitness,
    mvar=gpa,
    cvar=ethni age14 age15
    income1 income2 income3
    educ1 educ2 educ3,
    a0=0,  a1=1,  m=0,  nc=4,   c=,
    yreg=logistic,
    mreg=linear,
    interaction=false)
RUN;
```

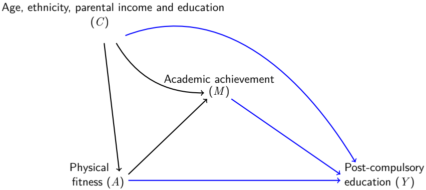


Regression for mediator M

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	5.39284	0.24974	21.59	<.0001
fitness	1	0.74596	0.10639	7.01	<.0001
ethni	1	0.00729	0.16047	0.05	0.9638
age14	1	-0.17335	0.18901	-0.92	0.3595
age15	1	-0.97692	0.16073	-6.08	<.0001
income1	1	0.07487	0.14093	0.53	0.5955
income2	1	0.44570	0.14350	3.11	0.0020
income3	1	0.83715	0.14150	5.92	<.0001
educ1	1	0.86340	0.19138	4.51	<.0001
educ2	1	1.57952	0.19851	7.96	<.0001
educ3	1	2.29918	0.21440	10.72	<.0001

Regression for outcome Y

```
%MEDIATION(data=d1,
    yvar=attend,
    avar=fitness,
    mvar=gpa,
    cvar=ethni age14 age15
    income1 income2 income3
    educ1 educ2 educ3,
    a0=0,  a1=1,  m=0,  nc=4,   c=,
    yreg=logistic,
    mreg=linear,
    interaction=false)
RUN;
```



Regression for outcome Y

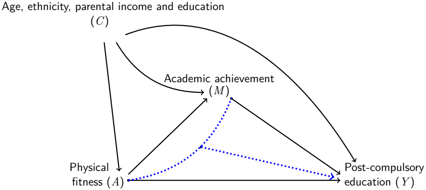
The SAS System

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	3.1458	1.1666	7.2713	0.0070
fitness	1	-0.6067	0.4800	1.5975	0.2063
gpa	1	-0.4354	0.1746	6.2156	0.0127
ethni	1	-1.3247	0.4802	7.6101	0.0058
age14	1	-0.8542	0.8090	1.1151	0.2910
age15	1	0.6492	0.4685	1.9204	0.1658
income1	1	-0.8252	0.4871	2.8706	0.0902
income2	1	-0.4169	0.5119	0.6635	0.4153
income3	1	-0.5031	0.5142	0.9575	0.3278
educ1	1	-1.2325	0.4825	6.5244	0.0106
educ2	1	-1.3308	0.5939	5.0204	0.0250
educ3	1	-1.5868	0.8001	3.9335	0.0473

Exposure-mediator interaction

```
%MEDIATION(data=d1,
    yvar=attend,
    avar=fitness,
    mvar=gpa,
    cvar=ethni age14 age15
    income1 income2 income3
    educ1 educ2 educ3,
    a0=0,  a1=1,  m=0,  nc=4,   c=,
    yreg=logistic,
    mreg=linear,
    interaction=false)
RUN;
```



Other options

```
%MEDIATION(data=d1,
    yvar=attend,
    avar=fitness,
    mvar=gpa,
    cvar=ethni age14 age15
    income1 income2 income3
    educ1 educ2 educ3,
    a0=0,  a1=1,   m=0,  nc=4,   c=,
    yreg=logistic,
    mreg=linear,
    interaction=false)
 RUN;
```

a0 - baseline level of exposure (unexposed)

a1 - new exposure level

c - fixed value for C at which conditional effects are computed

nc - number of baseline covariates

m - fixed value for M at which controlled direct effect is computed

Estimates of natural and controlled direct effects

The SAS System

Obs	Effect	Estimate	p_value	_95__CI_lower	_95__CI_upper
1	cde=nde	0.54517	0.20626	0.21280	1.39669
2	nie	0.72268	0.01882	0.55114	0.94763
3	total effect	0.39399	0.04286	0.15995	0.97050

$$OR_{NDE} = 0.545$$

$$OR_{NIE} = 0.723$$

$$OR_{TE} = 0.394$$

Interpretation

$OR_{NIE} = 0.723$ changing the grade point average from the value that would have been observed at the low level of physical fitness ($M(0)$) to the value that would have been observed at high level of physical fitness ($M(1)$), while actually keeping the physical fitness at the high level ($A = 1$) increases the odds of attending post-compulsory education by $\frac{1}{0.723} = 1.383$ times

Natural effect models

- Lange [2], Vansteelandt[3] suggested using so-called natural effect models

$$g_Y\{E[Y(a, M(a^*)) \mid C = c]\} = \theta_0 + \theta_1 a + \theta_2 a^* + \theta_3^T c$$

- Conditional natural effects at the observed values of C given on the scale of linear predictor g_Y

$\theta_1(a - a^*)$ - natural direct effect

$\theta_2(a - a^*)$ - natural indirect effect

- Implemented in the R package **medflex**

[2] Theis Lange, Stijn Vansteelandt, and Maarten Bekaert. A simple unified approach for estimating natural direct and indirect effects. *American journal of epidemiology*, 176(3):190-195, 2012.

[3] Stijn Vansteelandt, Maarten Bekaert, and Theis Lange. Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, 1(1):13-158, 2012

Estimation of natural effect models

- At first glance, it seemes that fitting natural effect models requires data for nested counterfactuals $Y(a, M(a^*))$

i	A_i	a	a^*	M_i	$M_i(a^*)$	Y_i	$Y_i(a, M_i(a^*))$
1	0	0	0	M_1	M_1	Y_1	Y_1
1	0	1	0	M_1	M_1	Y_1	?
2	1	1	1	M_2	M_2	Y_2	Y_2
2	1	0	1	M_2	M_2	Y_2	?

Estimation of natural effect models

- Vansteelandt et al. suggested imputing the missing counterfactuals $Y(a, M(a^*))$

i	A_i	a	a^*	M_i	$M_i(a^*)$	Y_i	$Y_i(a, M_i(a^*))$
1	0	0	0	M_1	M_1	Y_1	Y_1
1	0	1	0	M_1	M_1	Y_1	$\hat{E}[Y_1 \mid A = a, M_1, C_1]$
2	1	1	1	M_2	M_2	Y_2	Y_2
2	1	0	1	M_2	M_2	Y_2	$\hat{E}[Y_2 \mid A = a, M_2, C_2]$

- Imputation model

$$g_Y\{E[Y \mid A = a, M = m, C = c]\} = \beta_0 + \beta_1 a + \beta_2 m + \beta_3 c.$$

Illustrative example

- Logistic regression model as the natural effects model

$$\text{logit}\{P(Y(a, M(a^*) = 0 \mid C = c)\} = \theta_0 + \theta_1 a + \theta_2 a^* + \theta_4^T c$$

- Logistic regression model for Y

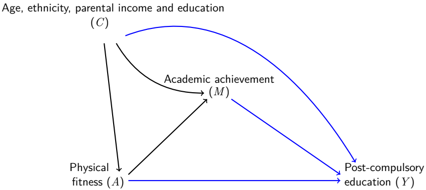
$$\text{logit}\{P(Y = 0, \mid A = a, M = m, C = c)\} = \beta_0 + \beta_1 a + \beta_2 m + \beta_3^T c$$

- Conditional natural effects as odds ratios given the observed level of covariates C

$$\log\{OR_{NDE}(a = 1, a^* = 0)\} = \theta_1$$
$$\log\{OR_{NIE}(a = 1, a^* = 0)\} = \theta_2$$
$$\log\{OR_{TE}(a = 1, a^* = 0)\} = \theta_1 + \theta_2$$

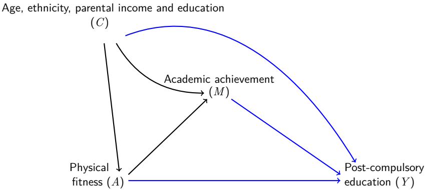
Expanding the data

```
R> library(medflex)
R> d1 <- read.csv("d.csv")
R> Yimp <- glm(attend~fitness+
+           gpa+
+           age+ethni+income+educ,
+           family="binomial")
R> impData <- neImpute(Yimp)
```



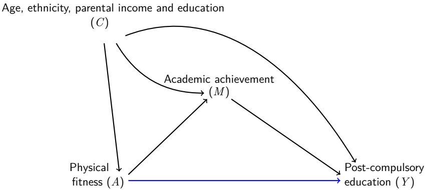
Expanding the data

```
R> library(medflex)
R> d1 <- read.csv("d.csv")
R> impData <- neImpute(attend~fitness+
+            gpa+
+            age+ethni+income+educ,
+            data=d1,
+            family="binomial")
```



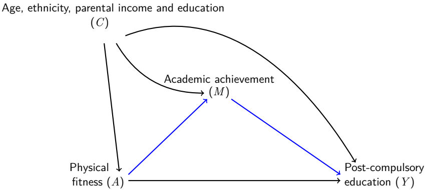
Fitting the natural effect model

```
R>  Yfit <- neModel(formula=attend~
+             fitness0+
+             fitness1+
+             age+ethni+income+educ,
+             expData=impData,
+             se="robust",
+             family="binomial")
```



Fitting the natural effect model

```
R>  Yfit <- neModel(formula=attend~
+            fitness0+
+            fitness1+
+            age+ethni+income+educ,
+            expData=impData,
+            se="robust",
+            family="binomial")
```



Other arguments

```
R>  Yfit <- neModel(formula=attend~
+             fitness0+
+             fitness1+
+             age+ethni+income+educ,
+             expData=impData,
+             se="robust",
+             family="binomial")
```

expData - expanded and imputed data set

se - standard errors (robust - based on Delta method, bootstrap-based on 1000 bootstrap)

Estimates of the natural effects

```
R>  summary(Yfit)

Natural effect model
with robust standard errors based on the sandwich estimator
---
Exposure: fitness
Mediator(s): gpa
---
Parameter estimates:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.40784    0.84595   -1.66   0.0961 .
fitness01   -0.51320    0.46077   -1.11   0.2654
fitness11   -0.43246    0.16205   -2.67   0.0076 **
ethni       -1.00395    0.46373   -2.16   0.0304 *
age1        -0.82623    0.48893   -1.69   0.0911 .
age2        -1.51465    0.88044   -1.72   0.0854 .
income1      0.70755    0.46724    1.51   0.1299
income2      0.09448    0.54964    0.17   0.8635
income3     -0.00342    0.57855   -0.01   0.9953
educ1        1.87121    0.63729    2.94   0.0033 **
educ2        0.43409    0.58175    0.75   0.4556
educ3        0.28027    0.63118    0.44   0.6570
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
      ' 1
```

$$OR_{NDE} = \exp\{-0.513\} = 0.599$$

$$OR_{NIE} = \exp\{-0.432\} = 0.650$$

Estimate of the total effect

```
R>  summary(neEffdecomp(Yfit))

Effect decomposition on the scale of the linear predictor
with standard errors based on the sandwich estimator
---
conditional on: ethni, age, income, educ
with x* = 0, x = 1
---
                        Estimate Std. Error z value Pr(>|z|)
natural direct effect     -0.513      0.461   -1.11   0.2654
natural indirect effect   -0.432      0.162   -2.67   0.0076 **
total effect              -0.946      0.453   -2.09   0.0368 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Univariate p-values reported)
```

$$OR_{TE} = \exp\{-0.946\}$$
$$= 0.388$$

Interpretation

$OR_{NIE} = 0.650$ changing the grade point average from the value that would have been observed at the low level of physical fitness ($M(0)$) to the value that would have been observed at high level of physical fitness ($M(1)$), while actually keeping the physical fitness at the high level ($A = 1$) increases the odds of attending post-compulsory education by $\frac{1}{0.650} = 1.538$ times

Other estimation methods considered in the paper

- Weighting approach in the R package **medflex**
- Approach based on Monte Carlo approximations implemented in the R package **mediation**
- Inverse odds ratio weighted estimation of natural effects with R code examples

Comparison

Variable	SAS/SPSS	medflex (W)	medflex (I)	mediation	IORW
Parameters of interest					
Marginal or conditional	Conditional at a *fixed* level of C	Conditional at the *observed* level of C	Conditional at the *observed* level of C	Marginal	Conditional at the *observed* level of C
Scale	Corresponds to g	Corresponds to g	Corresponds to g	Always difference, i.e. $g = identity$	Corresponds to g
Modelling					
Required models	$M \mid A, C$ $Y \mid A, M, C$	$M \mid A, C$ $Y(a, M(a^*)) \mid C$	$Y \mid A, M, C$ $Y(a, M(a^*)) \mid C$	$M \mid A, C$ $Y \mid A, M, C$	$A \mid M, C$ $Y \mid A, C$
Interactions	$A \times M$	$A \times M$ $A \times C$	$A \times M$ $A \times C$	$A \times M$ $A \times C$	$A \times M$ $A \times C$
Type of variables					
Exposure	Continuous Binary Polytomous	Continuous Binary Polytomous	Continuous Binary Polytomous	Continuous Binary Polytomous	Continuous Binary Polytomous
Mediator	Continuous Binary	Continuous Binary Count Polytomous	Continuous Binary Count Polytomous Multidimensional	Continuous Binary Count Polytomous Failure time Multidimensional	Continuous Binary Count Polytomous Failure time Multidimensional
Outcome	Continuous Binary Count Failure time	Continuous Binary Count	Continuous Binary Count	Continuous Binary Count Polytomous Failure time	Continuous Binary Count Polytomous Failure time

Simulation study

- 2000 samples of data sets with 200 observations
- Set up:

$$
\begin{aligned}
P(C=1) &= 0.7 \\
P(A=1|C=c) &= \Phi(-0.3c) \\
M &= 6.7 + A - 0.7C + \varepsilon \\
P(Y=1|A=a, M=m, C=c) &= \Phi(-0.3 + 0.3a + 0.2m + 0.2c)
\end{aligned}
$$

with

$$\varepsilon \sim t(df = 10)$$

Simulation study

- True natural effects:

  estimated from a simulated data set with 100,000 observations

- Relative bias:

$$\frac{1}{2000}\sum_{i=1}^{2000}\frac{\widehat{NDE}_i - \widehat{NDE}_{true}}{\widehat{NDE}_{true}}$$

- Relative RMSE:

$$\sqrt{\frac{1}{2000}\sum_{i=1}^{2000}\left(\frac{\widehat{NDE}_i - \widehat{NDE}_{true}}{\widehat{NDE}_{true}}\right)^2}$$

Results for direct effect

Method	Rel.Bias	Rel.RMSE	Cov.P
SAS macro	0.0653	1.802	94.4%
medflex (I)	0.309	2.418	94.5%
medflex (W)	0.368	2.587	92.6%
R mediation	0.014	0.866	95.2%
IORW	0.416	2.441	90.1%

Results for indirect effect

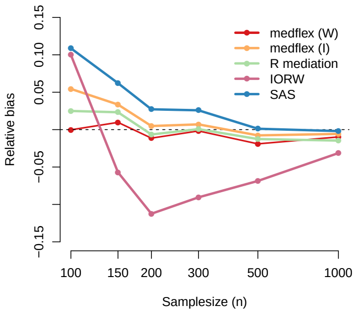
Method	Rel.Bias	Rel.RMSE	Cov.P
SAS macro	0.014	0.511	94.8%
medflex (I)	0.001	0.474	95.2%
medflex (W)	-0.010	0.513	94.0%
R mediation	-0.033	0.479	93.4%
IORW	-0.102	1.110	87.4%

Results for total effect

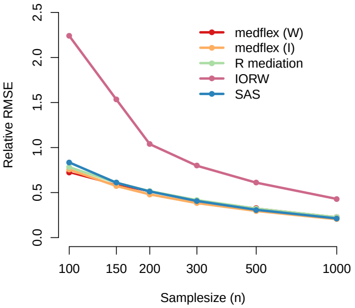
Method	Rel.Bias	Rel.RMSE	Cov.P
SAS macro	0.079	1.081	93.8%
medflex (I)	0.188	1.457	94.0%
medflex (W)	0.212	1.508	93.4%
R mediation	-0.004	0.498	95.7%
IORW	0.188	1.453	94.1%

Results depending on sample size

Relative bias



Relative RMSE



Conclusions

- All estimation methods peform good in this particular setting
- IORW estimation seems to have larger relative bias and relative RMSE, needs further investigation
- Choice of estimation method depends on
  - parameter of interest aimed for
  - software preferance.
- Mediation analysis can be applied fairly easily in most of the standard software packages
- Our paper will give guidance and examples how to apply mediation analysis with 5 different software solutions
- If you run into problems, you are very welcome to contact us!

Thank you!