Lecture 5: SEM (Path models)

Dr Nemanja Vaci

University of Sheffield

2025-03-13

Press record

Intended learning outcomes

Motivate utilisation of path and CFA models; Argue how they connect to other models that we covered at the course.

Calculate number of free parameters and degrees of freedom of the proposed model.

Build a model in R statistical environment, estimate, and interpret the coefficients.

Criticise, modify, compare, and evaluate the fit of the proposed models.

Structural equation modelling (SEM)

General framework that uses various models to test relationships among variables

Other terms: covariance structure analysis, covariance structure modelling, **causal modelling**

Sewell Wright - "mathematical tool for drawing **causal** conclusions from a combination of of observational data and **theoretical assumptions**"

Waves:

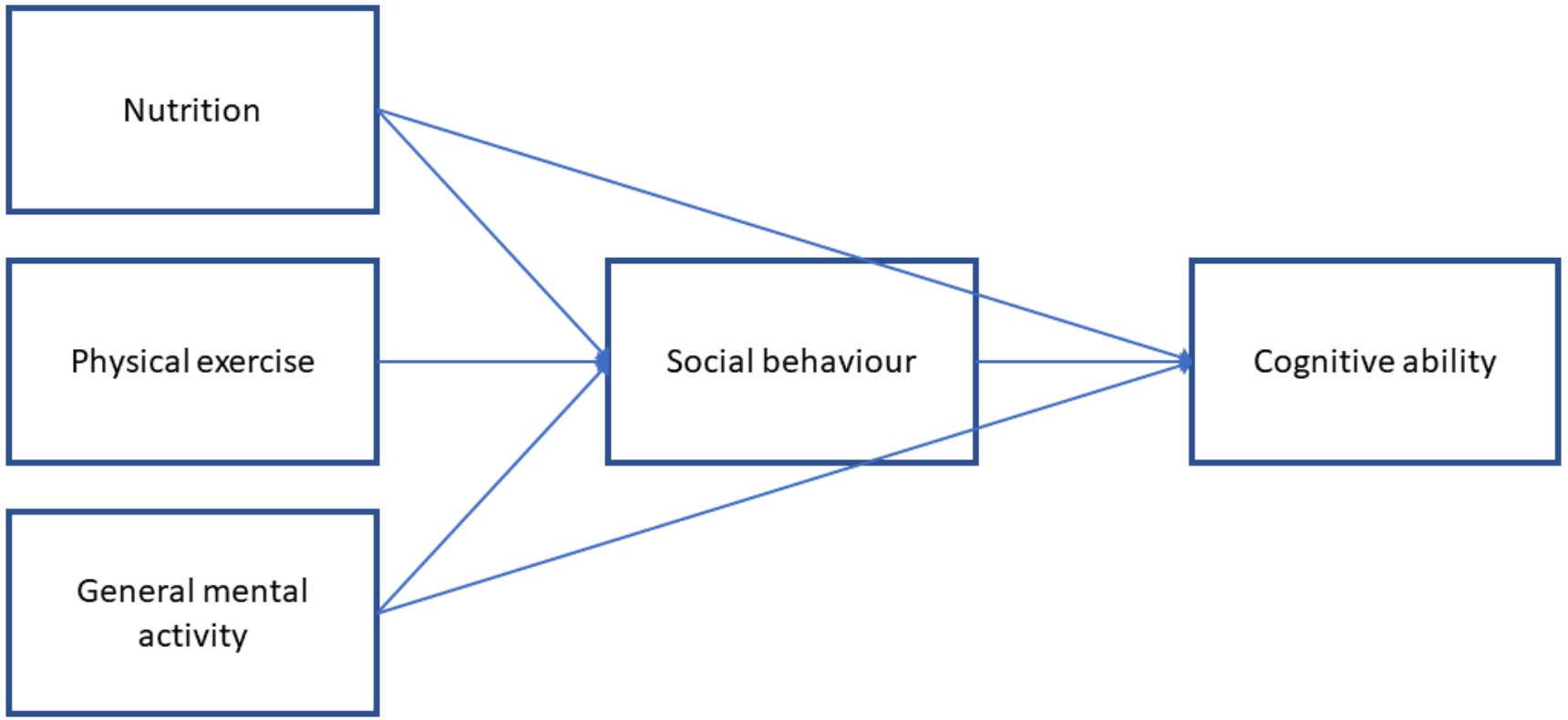
1. Causal modelling through path models
2. Latent structures - factor analysis
3. Structural causal models

SEM is a general modelling framework that is composed of measurement model and the structural model.

Structural part of the model (path analysis)

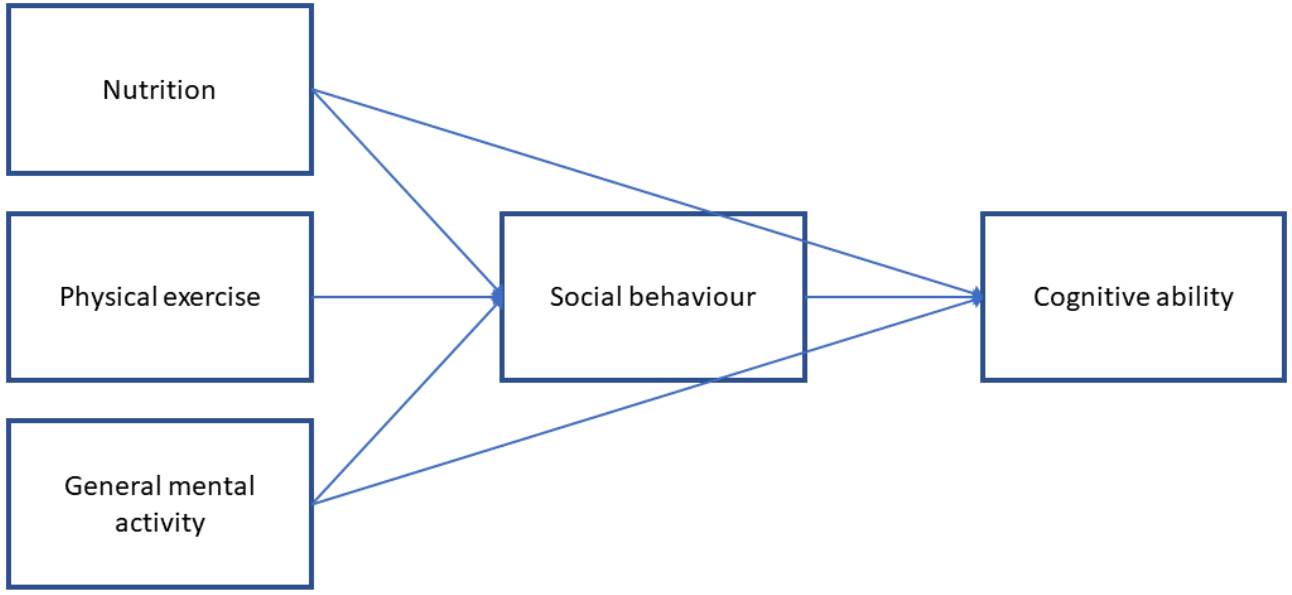
Model that test relationship between set of variables, often arranged in some sort of structural form.

A common focus of the path model is the estimation of mediation between X and Y.



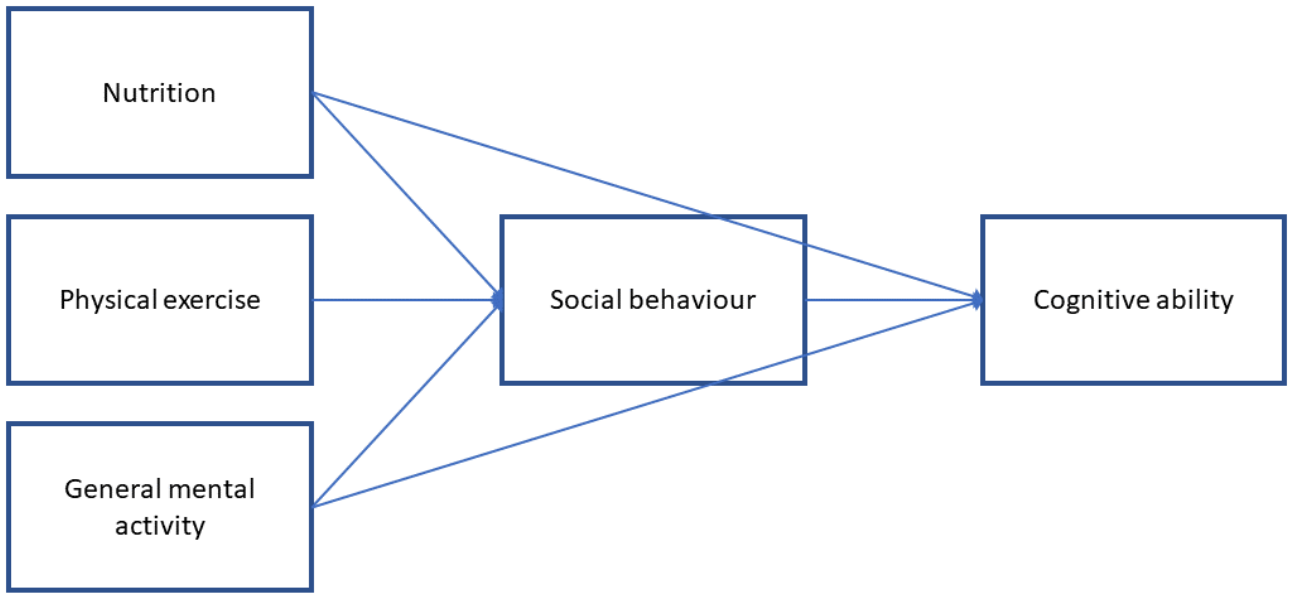
First step: Specification of the model

Previous findings show that development of cognitive abilities in people depends on a range of factors in infancy and early childhood. General mental/cognitive abilities (e.g. reading or drawing), varied nutrition, physical exercises, and social engagement have shown to influence the level of cognitive abilities. Based on some of these studies, researchers postulate that social engagement is mediating factor between the behavioural factors and development of cognitive abilities.



Can model be estimated?

Total Number of the parameters that we can estimate: $\frac{variables*(variables+1)}{2}$



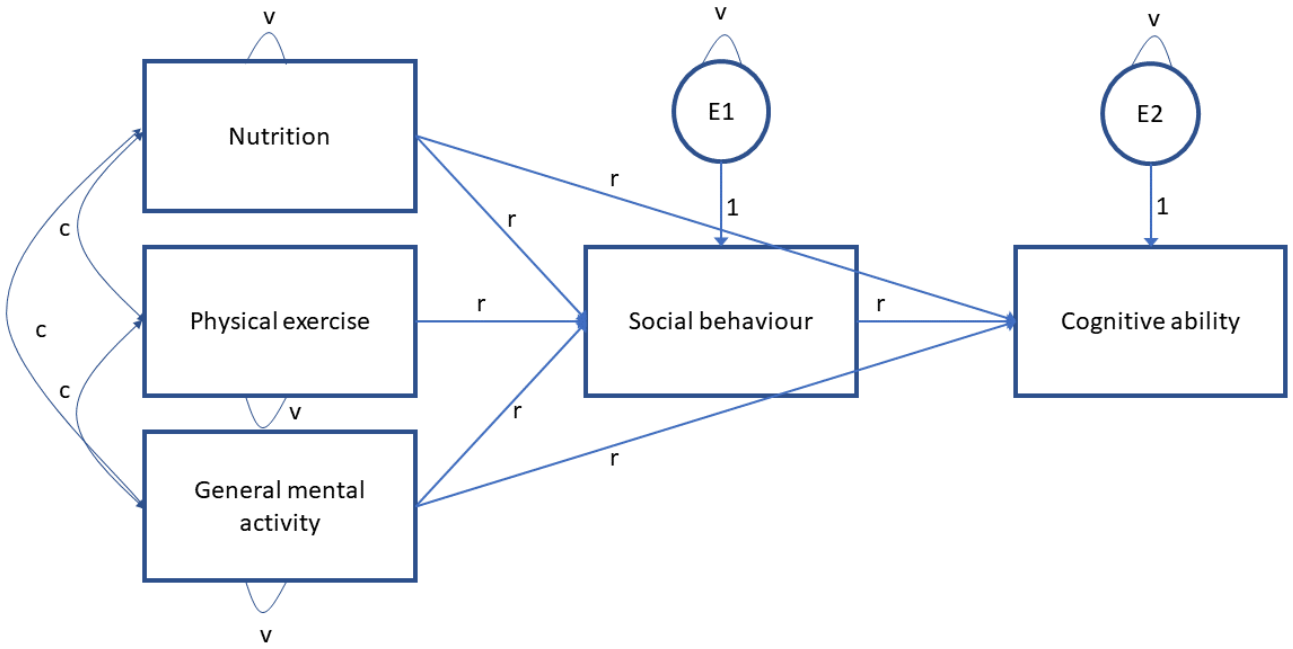
Number of observations

```
Matrix←cov(Babies[,c('Nutrition','PhyExer','GMA','SocialBeh','CognitiveAb')])
Matrix[upper.tri(Matrix)]←NA
knitr::kable(Matrix, format = 'html')
```

	Nutrition	PhyExer	GMA	SocialBeh	CognitiveAb
Nutrition	45.6689837	NA	NA	NA	NA
PhyExer	-10.1006752	2652.9074	NA	NA	NA
GMA	0.5641485	-249.3049	2478.2889	NA	NA
SocialBeh	-11.6168733	3417.8681	-506.1066	9988.898	NA
CognitiveAb	210.6731970	48916.6339	1254.2100	94358.621	1125746

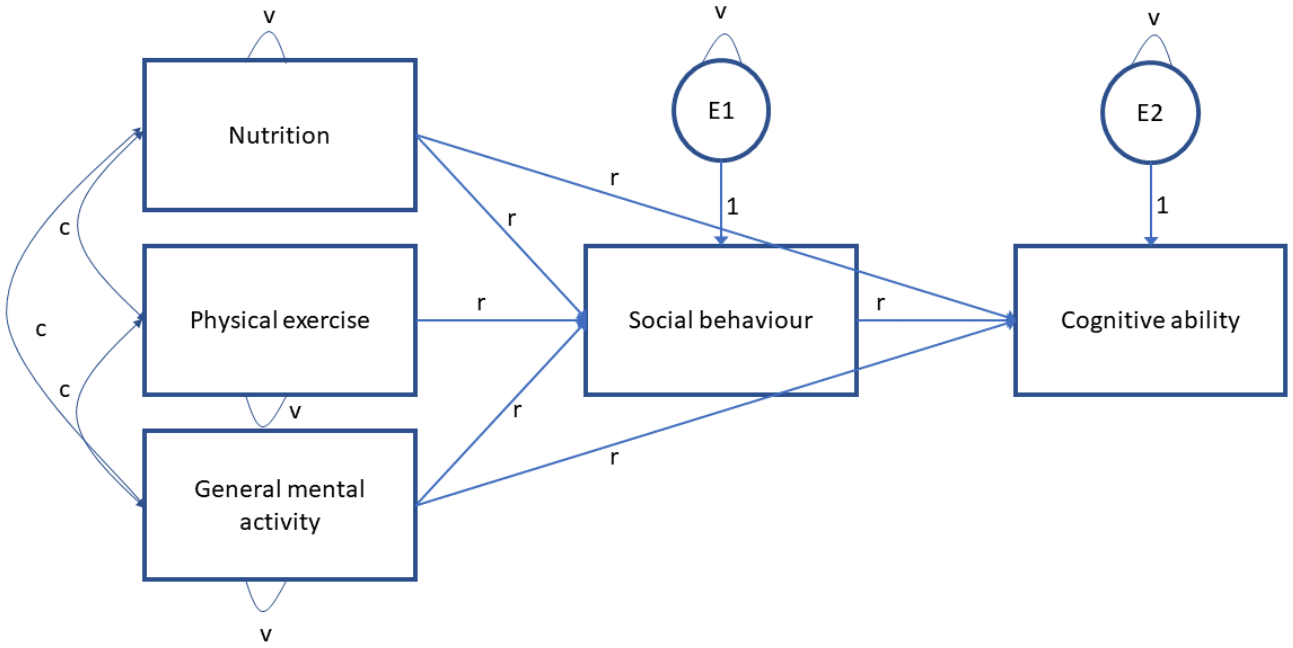
How many parameters are we estimating (path model)?

How many degrees of freedom do we have without the model?



Number of observations (total number of parameters) = 15
Empty model = variances and covariances
Degrees of freedom (df) **= 15 - 8 = 7**

How many parameters (our model)?



Free parameters = variances + covariances + regression pathways = 14

Second step: model identification

1. Under-indentified: more free parameters than total possible parameters
2. Just-identified: equal number of free parameters and total possible parameters
3. Over-identified: fewer free parameters than total possible parameters

Parameters can either be: free, fixed or constrained

Third step: estimation of the model

```
modelAbility←'
SocialBeh~Nutrition+PhyExer+GMA
CognitiveAb~SocialBeh+Nutrition+GMA
'
```

```
fit1←sem(modelAbility, data=Babies)
summary(fit1)
```

```
## Length  Class   Mode
##      1 lavaan     S4
```

Step four: model evaluation

Chi-square test: measure of how well model-implied covariance matrix fits data covariance

We would prefer not to reject the null hypothesis in this case

Assumptions:
Multivariate normality
N is sufficiently large (150+)
Parameters are not at boundary or invalid (e.g. variance of zero)

With the large samples it is sensitive to small misfits
Nonormality induces bias

Other fit indices

```
summary(fit1, fit.measures=TRUE)
```

```
## Length  Class   Mode
##      1 lavaan     S4
```

Other fit indices

TABLE 13.1. Fit Indices for Covariance Structure Models

Equation No.	Fit index		Reference	Goodness- or badness-of-fit index	Theoretical range	Cutoff criterion	Sensitive to *N*	Penalty for model complexity?
T1	$\chi^2 = (N-1)f$		Jöreskog (1969)	Badness	≥ 0	$p < .05$	Yes	No
T2	χ^2 / df	(8)	Jöreskog (1969)	Badness	≥ 0	< 5[d]	Yes	Yes
T3	$\mathrm{GFI} = 1 - \frac{\mathbf{e'We}}{\mathbf{s'Ws}}$	(10)	Jöreskog & Sörbom (1981)	Goodness	0–1[a]	$> .95$[d]	Yes	No
T4	$\mathrm{AGFI} = 1 - \frac{p^*}{df}(1-\mathrm{GFI})$	(6)	Jöreskog & Sörbom (1981)	Goodness	0–1[a]	N/A[d,e]	Yes	Yes
T5	$\mathrm{GFI}^* = \frac{p}{p+2\left(\frac{\chi^2-df}{N-1}\right)}$	(0)	Maiti & Mukherjee (1990); Steiger (1989)	Goodness	0–1[a]	$> .95$	No	No
T6	$\mathrm{AGFI}^* = 1 - \frac{p^*}{df}(1-\mathrm{GFI}^*)$	(0)	Maiti & Mukherjee (1990); Steiger (1989)	Goodness	0–1[a]	N/A[e]	No	Yes
T7	$\mathrm{RMR} = [p^{*-1}(\mathbf{e'Ie})]^{1/2}$	(4)	Jöreskog & Sörbom (1981)	Badness	> 0	N/A[e,f]	Yes	No
T8	$\mathrm{SRMR} = [p^{*-1}(\mathbf{e'W_s e})]^{1/2}$	(13)	Bentler (1995)	Badness	> 0	$< .08$	Yes	No
T9	$\mathrm{RMSEA} = \sqrt{\frac{\hat{\lambda}_N}{df}} = \sqrt{\frac{\max(\chi^2-df,\,0)}{df(N-1)}}$	(42)	Steiger & Lind (1980)	Badness	> 0	$< .06$	Yes to small *N*	Yes
T10	$\mathrm{TLI}^c = \frac{\chi_0^2/df_0 - \chi_k^2/df_k}{\chi_0^2/df_0 - 1}$	(22)	Tucker & Lewis (1973)	Goodness	0–1[a, b]	$> .95$	No	Yes
T11	$\mathrm{NFI} = \frac{f_0 - f_k}{f_0} = \frac{\chi_0^2 - \chi_k^2}{\chi_0^2}$	(7)	Bentler & Bonett (1980)	Goodness	0–1	$> .95$[d]	Yes	No
T12	$\mathrm{IFI} = \frac{\chi_0^2 - \chi_k^2}{\chi_0^2 - df_k}$	(3)	Bollen (1989); Marsh et al. (1988)	Goodness	> 0[b]	$> .95$	Yes to small *N*	Yes
T13	$\mathrm{RNI} = \frac{(\chi_0^2 - df_0) - (\chi_k^2 - df_k)}{(\chi_0^2 - df_0)}$	(3)	Bentler (1990); McDonald & Marsh (1990)	Goodness	> 0[b]	$>.95$	No	Yes
T14	$\mathrm{CFI} = \frac{\max(\chi_0^2 - df_0, 0) - \max(\chi_k^2 - df_k, 0)}{\max(\chi_0^2 - df_0, 0)}$	(42)	Bentler (1990)	Goodness	0–1	$> .95$	No	Yes

Note. χ^2, chi-square test statistic; GFI = goodness-of-fit index; AGFI, adjusted goodness-of-fit index. GFI*, revised GFI; AGFI*, revised AGFI; RMR, root mean square residual; SRMR, standardized root mean square residual; RMSEA, root mean square error of approximation; TLI, Tucker–Lewis index; NFI, normed fit index; IFI, incremental fit index; RNI, relative noncentrality index; CFI, comparative fit index; *f*, minimized discrepancy function; *o*, baseline model; *k*, tested or hypothesized model; *df*, degrees of freedom; *N*, sample size; *p**, the number of nonduplicated elements in the covariance matrix; **e**, a vector of residuals from a covariance matrix; **s**, a vector of the *p** nonredundant elements in the observed covariance matrix; **I**, an identify matrix; **W**, a weight matrix; $\mathbf{W}_s$, a diagonal weight matrix used to standardize the elements in a sample covariance matrix; λ_N, noncentrality parameter, normed so that it is not negative. The numbers in parentheses in the "Fit indices" column represent the number out of 55 articles on structural equation models in substantive American Psychological Association journals in 2004 that reported each of the practical fit indices described here (see Taylor, 2008). No other practical fit indices were reported.

[a]Can be negative. Negative value indicates an extremely misspecified model.

Model modification

Add/take out theoretical pathways:

```
modelAbility2←'
SocialBeh~Nutrition+PhyExer+GMA
CognitiveAb~SocialBeh+Nutrition+GMA+PhyExer
'
fit2←sem(modelAbility2, data=Babies)
summary(fit2, fit.measures=TRUE)
```

```
## Length  Class   Mode
##      1 lavaan     S4
```

We can compare the models

```
lavTestLRT(fit1,fit2)
```

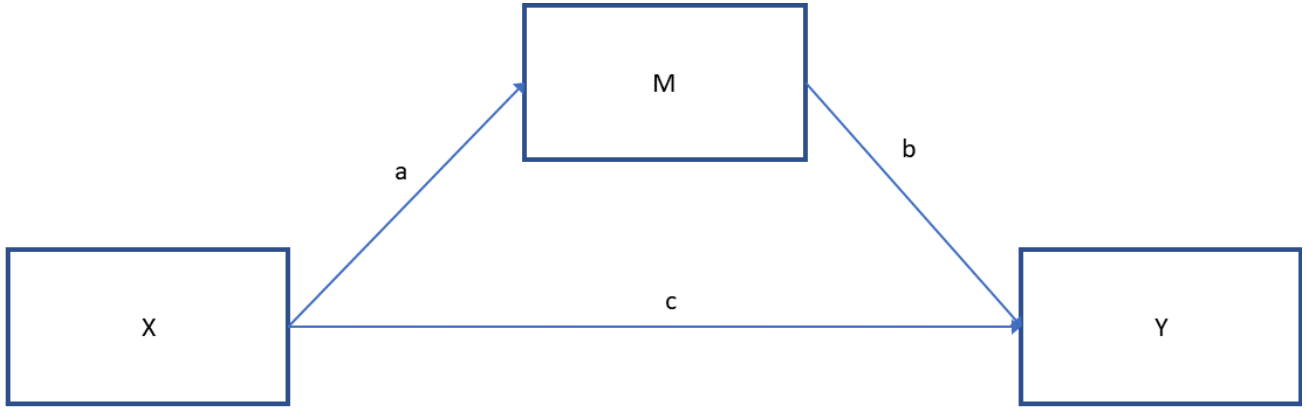
```
##
## Chi-Squared Difference Test
##
##      Df    AIC    BIC  Chisq Chisq diff  RMSEA Df diff Pr(>Chisq)
## fit2  0 2459.8 2483.2   0.00
## fit1  1 2673.0 2693.8 215.24     215.24 1.4637       1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Or check modification indices

```
modindices(fit1, sort=TRUE)
```

```
##             lhs op         rhs     mi    epc sepc.lv sepc.all sepc.nox
## 15   SocialBeh  ~ CognitiveAb 88.379 -0.228  -0.228   -2.420   -2.420
## 16 CognitiveAb  ~     PhyExer 88.379 11.390  11.390    0.553    0.011
## 22     PhyExer  ~ CognitiveAb 82.143  0.128   0.128    2.635    2.635
## 26         GMA  ~ CognitiveAb  1.601  0.025   0.025    0.529    0.529
## 18   Nutrition  ~ CognitiveAb  1.002  0.007   0.007    1.114    1.114
## 21     PhyExer  ~   SocialBeh  0.000  0.000   0.000    0.000    0.000
## 20   Nutrition  ~         GMA  0.000  0.000   0.000    0.000    0.000
## 19   Nutrition  ~     PhyExer  0.000  0.000   0.000    0.000    0.000
```

Direct and indirect



Direct effect (c): subgroups/cases that differ by one unit on X, but are equal on M are estimated to differ by **c** units on Y.

Indirect effect:
a) X -> M: cases that differ by one unit in X are estimated to differ by **a** units on M
b) M -> Y: cases that differ by one unit in M, but are equal on X, are estimated to differ by **b** units on Y

The indirect effect of X on Y through M is a product of **a** and **b**. The two cases that differ by one unit on X are estimated to differ by **ab** units on Y as a result of the effect of X on M which affects Y.

Direct and indirect

```
modelAbilityPath←'
SocialBeh~Nutrition+a*PhyExer+GMA
CognitiveAb~b*SocialBeh+c*PhyExer+GMA

indirect := a*b
direct := c
total := indirect + direct
'
fitPath←sem(modelAbilityPath, data=Babies)
```

```
## Length  Class   Mode
##      1 lavaan     S4
```

Prerequisites

Theory: Strong theoretical assumptions that could be used to draw causal assumptions that could be tested using the data and specification of the model

Data: large samples, N:p rule - 20:1, more data usually better estimates.

- We are not that interested in significance:

  a) Overall behaviour of the model more interesting

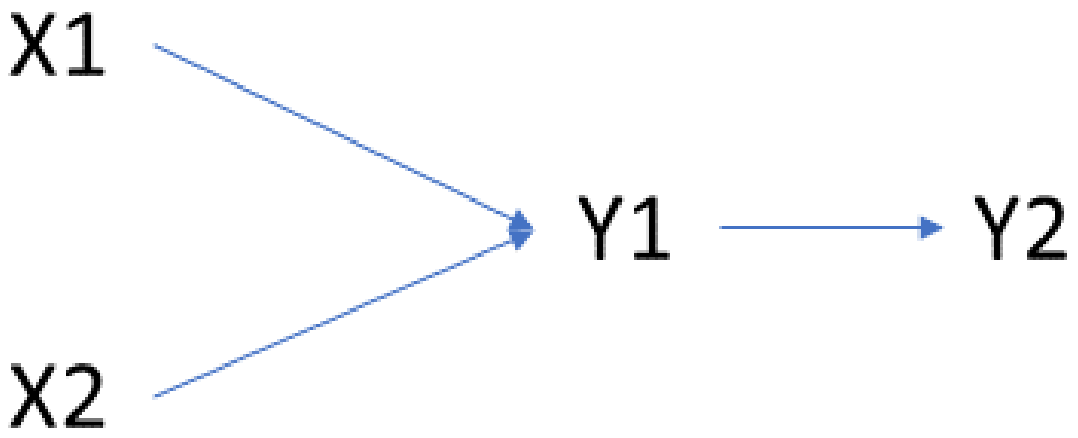
  b) More data higher probability of significant results (weak effects)

  c) Latent models are estimated by anchoring on indicator variables, different estimation can result in different patterns

Problems with SEM and alternatives

1. Variables derived from the normal distribution
2. Observations independent
3. Large sample size

PiecewiseSEM



Variables are causally dependent if there is an arrow between them
They are causally independent if there are no arrows between them

X1 is causally independent from Y2 *conditional* on Y1

PiecewiseSEM performs a test of directional separation (d-sep) and asks whether causally independent paths are significant when controlling for variables on which causal process is conditional.

PiecewiseSEM

```
#install.packages('piecewiseSEM)
require(piecewiseSEM)
model1←psem(lm(SocialBeh~Nutrition+PhyExer+GMA, data=Babies),
            lm(CognitiveAb~SocialBeh+Nutrition+GMA, data=Babies))
summary(model1, .progressBar=FALSE)
```

```
##
## Structural Equation Model of model1
##
## Call:
##   SocialBeh ~ Nutrition + PhyExer + GMA
##   CognitiveAb ~ SocialBeh + Nutrition + GMA
##
##     AIC      BIC
##  229.364   255.416
```

Important aspects: theory

- Difference between moderation and mediation
- Interpretation of the predictors
- Calculation of free parameters and total parameters
- Model identification: three-types of identifications
- Overall fit of the model

Important aspects: practice

- Building path model: both continous and categorical exogenous variables
- Calculation of the direct and indirect pathways for predictors of interest
- Adding an interaction to path model
- Interpretation of the coefficients
- Getting fit indices of the model

Literature

Chapters 1 to 5 of Principles and Practice of Structural Equation Modeling by Rex B. Kline

Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach by Andrew F. Hayes

Latent Variable Modeling Using R: A Step-by-Step Guide by A. Alexander Beaujean

Thank you for your attention