

# ADVANCED DATA ANALYSIS FOR PSYCHOLOGICAL SCIENCE

## Part 2. Introduction to multivariate modeling

**Luca Menghini Ph.D.**

luca.menghini@unipd.it

\*\*\*





Master degree in Developmental and Educational Psychology

University of Padova

2023-2024



## Outline of Part 2

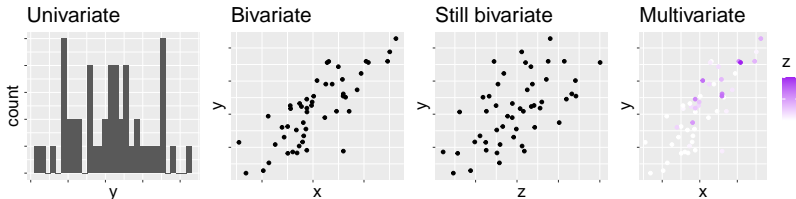
- **sem() intro:** Gentle introduction to the world of structural equation modeling (SEM)
- **Path analysis:** Introduction to path analysis (aka SEM with observed variables) and focus on *mediation models*
- **Data structure:** How to approach a multivariate data structure, how to manipulate and pre-process multivariate data 
- **Model fit & evaluation:** How to fit a path analysis in R, to evaluate model fit, compare multiple models, and interpret model results 
- **cfa():** How to conduct a confirmatory factor analysis (CFA) and to interpret its results 
- **Related topics:** In-depth topics related to multivariate modeling (e.g., cross-lagged panel models, multilevel and Bayesian SEM) 

---

 = not for the exam

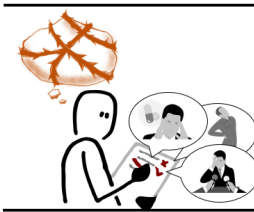
 = exercises with R (bring your laptop!)

# Multivariate analyses for a multivariate reality



- In psychology, we mainly inspect empirical data focusing on **univariate** ( $y$ ) or **bivariate** relationships (either  $y$  by  $x$  or  $y$  by  $z$ )
- But reality (particularly psychosocial reality) is complex, it is **multivariate** i.e., more than two variables covarying at the same time
- It is *reductionist* to separately analyze our variables without considering their overall interactions → **biased effect estimates**
- **Structural equation modeling (SEM)** allow to analyze the relationships of interest by accounting for the multivariate reality of psychosocial phenomena (e.g.,  $y$  by  $x$  covarying with  $z$ ;  $x$  affects  $y$  through  $z$ )

## Observed indicators & latent variables



- In psychology, we are mainly interested in **latent variables** = phenomena that we cannot directly observe, but we can estimate from 1+ **observed indicators** (e.g., 10-item scale measuring anxiety)
- Are we allowed to do that? Yes (let's say yes), provided that we trust the indicator **construct validity** = their relationship with the latent variable they claim to measure
- **SEM** allow to evaluate that by *quantifying the latent variables* and their relationships with observed indicators

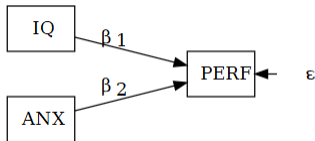
# Structural what!?

Structural equation modeling (SEM)

= multivariate *linear* models formalized by **systems of equations**

**Linear models** (LM): determining the link between a dependent and 1+ independent variables through a **single equation** like:

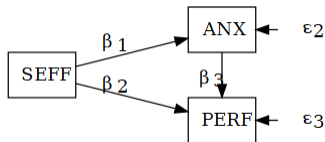
$$PERF = \beta_1 IQ + \beta_2 ANX + \epsilon$$



LM can only predict **one dependent variable at a time**, being either *univariate* (without predictors, i.e., intercept-only) or *bivariate* (with predictors).

**SEM** allow to simultaneously model multiple ~~dependent~~ *endogenous* variables with a **system of equations** like:

$$\begin{cases} ANX = \beta_1 SEFF + \epsilon_2 \\ PERF = \beta_2 SEFF + \beta_3 ANX + \epsilon_3 \end{cases}$$

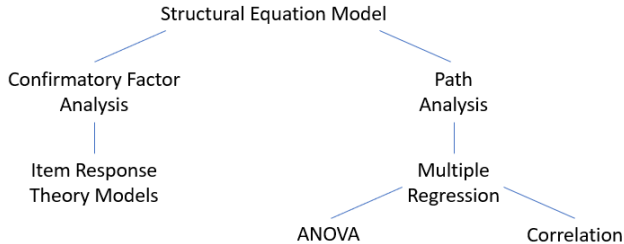


# The SEM family

SEM = broad family of statistical models within which LM, ANOVA, and even correlation can be included.

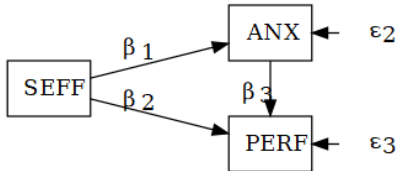
Particularly, 2 main sub-families can be distinguished based on whether **latent variables** are included in the model or not:

- **Path analysis:** multivariate linear models with observed variables only
- **Confirmatory factor analysis (CFA):** multivariate linear models with both observed and latent variables



## Path models & path analysis

**Path models/diagrams** = multivariate models with observed variables only  
= pictorial representations (*diagrams*) of a theory of variable relationships

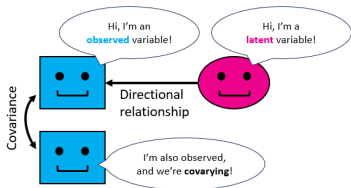


**Paths** = arrows (*edges*) linking the variables (*nodes*) in a model

**Path analysis** = analysis of multivariate relationships between observed variables  
(‘*quantification of the paths accounting for all other paths and errors*’)

# Latent factors & CFA

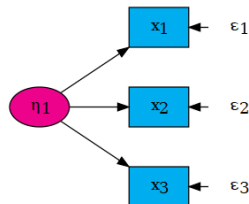
- **Observed/Manifest variable (OV)**  
variable that is directly observable (e.g., height, heart rate, item responses)
- **Latent variable/factor (LV)**  
variable that is *not* directly observable (e.g., anxiety, intelligence), but can be indexed by one or more observed variables
- In SEM, **OVs** are represented by **squares/rectangles** and indexed with **lower case letters** (e.g.,  $x$ ), whereas **LVs** are represented by **circles/ellipses** and indexed by the **Greek letter  $\eta$**



## Confirmatory factor analysis (CFA)

= analysis of the relationships (*factor loadings*) between a set of OVs and one or more LVs

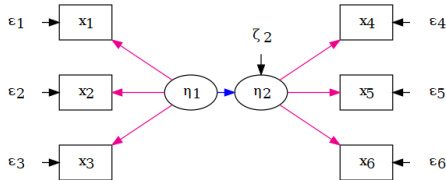
CFA uses **latent variable models** to *form* or *quantify* LVs and their relationships with OVs (evaluation of **construct validity**)





# SEM: Measurement & Structural model

To properly talk about ‘full SEM’ (or just SEM), we need both OV and LVs



A SEM consists of two parts:

1. **Structural model**: Regression-like relationships among the variables, working similar to *path analysis*
2. **Measurement model (or latent variable model)**: Relationships between OVs and LVs, working a little differently

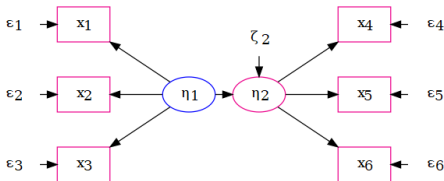
Notes:

In this sense, we may say that a CFA model is a ‘full SEM’ whereas a path model is not

A CFA is a SEM with just the measurement part (without the structural model)

# A new classification: From in/dependent to exo/endogenous variables

In both SEM (e.g., CFA) and path models, the classic independent vs. dependent classification is replaced with a more meaningful one:



- **Exogenous variables:** variables (both OVs and LVs) without a direct ‘cause’ from inside the model (predictors), without error estimate
- **Endogenous variables:** variables (both OVs and LVs) directly ‘caused’ from inside the model (predictors & outcomes), with error estimate  $\epsilon$  (OV) or  $\zeta$  (LV)

# A new starting point: From dataset columns to covariance matrices

The starting point of LM(ER) is a vector (or a set of vectors) of variable values, usually corresponding to one or more columns from a dataset.

```
head(df,4)
```

	MAT	QI	WM	STM
1	57	21	15	18
2	77	22	19	17
3	51	13	13	16
4	58	24	6	21

The starting point of SEM and path models is the **covariance matrix of the observed variables**.

$$\text{cov}(x, y) = \sum (x_i - \bar{x})(y_i - \bar{y})/N$$

```
cov(df[,c("MAT", "QI", "WM", "STM")])
```

	MAT	QI	WM	STM
MAT	100.70	24.89	17.21	7.99
QI	24.89	19.43	6.69	4.04
WM	17.21	6.69	17.33	2.23
STM	7.99	4.04	2.23	5.34

SEM estimate a number of parameters  $\theta$  so that the **implied covariance matrix**  $\sum(\theta)$  (i.e., the covariance matrix predicted by the model based on the parameter estimates) is as close as possible to the **sample covariance matrix**  $S$

 Note: even the model parameters are estimated within **matrices of parameters** 

## Covariance & correlation

- **Variance** = Expected value of the **squared deviation from the mean** of a random variable, or degree to which it deviates from its expected value

$$\text{var}(x) = \sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

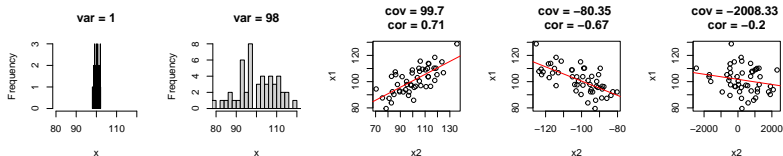
- **Covariance** = Measure of the **joint variability** of two random variables, or Degree to which they tend to deviate from their expected values in similar ways, either directly (positive cov) or inversely (negative cov), whose value depends on the variable scales of measurement (from  $-\infty$  to  $+\infty$ )

$$\text{cov}(x_1, x_2) = \frac{\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{N}$$

- **Correlation** = standardized covariance of two random variables

Correlation ranges from -1 (perfectly negative) to +1 (perfectly positive)

$$\text{cor}(x_1, x_2) = \frac{\text{cov}(x_1, x_2)}{\sigma_{x_1} \sigma_{x_2}}$$



## Covariance matrix ( $S$ )

Given a set of  $p$  variables, we can define the covariance matrix:

$$S = \begin{bmatrix} s_{11} & \dots & s_{1j} & \dots & s_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ s_{i1} & \dots & s_{ij} & \dots & s_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ s_{p1} & \dots & s_{pj} & \dots & s_{pp} \end{bmatrix}$$

Properties of the covariance matrix:

1. **Symmetrical:**  $s_{ij} = s_{ji}$
2. The **main diagonal** shows the **variances** (= covariance between each variable and itself)

```
cov(df[,c("MAT", "QI", "WM", "STM")])
```

	MAT	QI	WM	STM
MAT	100.70	24.89	17.21	7.99
QI	24.89	19.43	6.69	4.04
WM	17.21	6.69	17.33	2.23
STM	7.99	4.04	2.23	5.34

SEM estimate a number of parameters  $\theta$  so that the **implied covariance matrix**  $\hat{\Sigma}(\theta)$  (i.e., the covariance matrix predicted by the model based on the parameter estimates) is as close as possible to the **sample covariance matrix**  $S$

# That's all for now!

## Questions?

### Homework (optional):

- read the slides presented today  
and write in the Moodle forum if you have any doubts
- exeRcises 12-13 from exeRcises.pdf

---

For each exercise, the solution (or one of the possible solutions) can be found in dedicated chunk of commented code within the `exeRcises.Rmd` file

## In the last episode...

### The problem

Psychosocial reality is complex:  
it's **multivariate** (3+ variables  
interacting at the same time) and  
involves **latent variables** (not  
directly measurable)

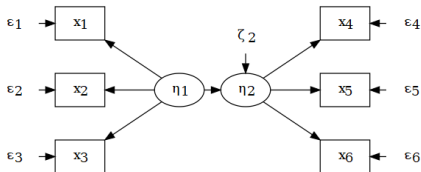
### The solution

SEM allows to analyze the  
multivariate relationships among  
observed and latent variables  
through **systems of equations**:

$$\begin{cases} ANX = \beta_{21}SEFF + \epsilon_2 \\ PERF = \beta_{31}SEFF + \beta_{32}ANX + \epsilon_3 \end{cases}$$

### SEM basics

- **Observed ( $x$ ) vs latent variables ( $\eta$ )** depending on whether can be directly measured or not
- **Exogenous vs endogenous variables** depending on whether directly caused inside the model or not
- **Structural vs measurement model** depending on whether focusing on structural relationships or construct validity of the observed indicators
- **Path model**: SEM with observed variables only
- **CFA** = SEM with measurement model only
- Starting point of any SEM = **covariance matrix**



# Path models: SEM with observed variables

A path model is a pictorial representation (*diagram*) of a theory of variable relationships. Path analysis is widely used to model complex multivariate relationships (e.g., *mediation models*).

- Path analysis tests models of **causal relationships**\* among observed variables
- All variables in path analysis are **observed**
- Path analysis uses **systems of regression equations**

\*Note: Within path analysis (and SEM) we assume that the relationships are *causal*, but this is not necessarily true (e.g., observational studies) → causation requires experimental manipulation, control group, etc.

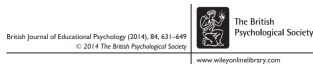


# Case study: Early mathematical abilities

A sample of 120 first-grade children (58 females; mean age: 6 years, 3 months) was assessed over the following variables:

- **MAT**: early mathematical abilities (e.g., comparison, classification) measured with the Early Numeracy Test
- **QI**: intelligence level measured with the Wechsler Intelligence Scale for Children (WISC-III)
- **WM**: working memory capacity measured with the Backward word recall task
- **STM**: short-term memory capacity measured with the Forward word recall task
- **ANS**: approximate number system = innate system for approximate quantity manipulation (e.g., approximate computations, comparing 2+ sets of elements without counting), measured with several tasks

**RQ: How much can MAT abilities be attributed to memory & ANS?**



## **The contribution of general cognitive abilities and approximate number system to early mathematics**

Maria Chiara Passolunghi<sup>1\*†</sup>, Elisa Cargnelutti<sup>1</sup> and Massimiliano Pastore<sup>2</sup>

<sup>1</sup>Department of Life Sciences, University of Trieste, Italy

<sup>2</sup>Department of Developmental and Social Psychology, University of Padua, Italy

# Data exploration

First, let's explore the data:

```
library(devtools); install_github("https://github.com/masspastore/ADati") # install ADati pkg
```

```
data( earlymath, package = "ADati" ) # loading earlymath dataset from ADati pkg
```

```
head(earlymath,3) # showing first 3 rows
```

	gender	MAT	QI	WM	STM	ANS
147	m	57	21	15	18	80
144	m	77	22	19	17	76
155	f	51	13	13	16	79

```
summary(earlymath[,c(2,4:ncol(earlymath))]) # summarizing variables (not showing QI due to space limits)
```

MAT		WM		STM		ANS	
Min.	:36.00	Min.	: 1.00	Min.	:13.00	Min.	:45.00
1st Qu.:	:61.75	1st Qu.:	:12.00	1st Qu.:	:17.00	1st Qu.:	:74.00
Median	:68.00	Median	:14.50	Median	:18.00	Median	:80.00
Mean	:68.56	Mean	:14.55	Mean	:18.43	Mean	:79.34
3rd Qu.:	:75.00	3rd Qu.:	:17.00	3rd Qu.:	:20.00	3rd Qu.:	:85.00
Max.	:91.00	Max.	:28.00	Max.	:26.00	Max.	:94.00

```
round( cor(earlymath[,2:ncol(earlymath)]), 2) # correlations
```

	MAT	QI	WM	STM	ANS
MAT	1.00	0.56	0.41	0.34	0.26
QI	0.56	1.00	0.36	0.40	0.23
WM	0.41	0.36	1.00	0.23	0.12
STM	0.34	0.40	0.23	1.00	0.19
ANS	0.26	0.23	0.12	0.19	1.00

# Linear model as a path diagram

Let's fit a multiple linear model:  $MAT = \beta_0 + \beta_1 WM + \beta_2 STM + \beta_3 ASN + \epsilon$

```
lm.fit <- lm(MAT ~ WM + STM + ANS, data = earlymath) # fitting LM
```

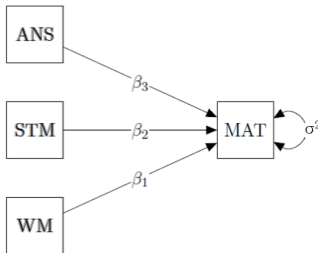
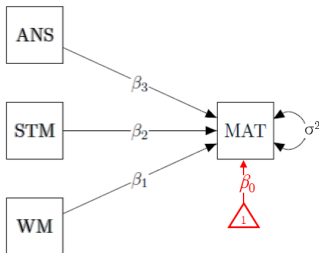
```
summary(lm.fit)$coefficients # LM regression table
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	20.03	9.61	2.09	0.04
WM	0.81	0.20	4.10	0.00
STM	1.01	0.36	2.81	0.01
ANS	0.23	0.11	2.16	0.03

Residual variance  $\sigma^2$ :

```
summary(lm.fit)$sigma^2  
[1] 75.94542
```

This model can be graphically represented as a path diagram and further simplified by removing the **intercept**  $\beta_0$  (note: **triangles represent constants**)



How many parameters?

# Linear model as a path diagram

Let's fit a multiple linear model:  $MAT = \beta_0 + \beta_1 WM + \beta_2 STM + \beta_3 ASN + \epsilon$

```
lm.fit <- lm(MAT ~ WM + STM + ANS, data = earlymath) # fitting LM
```

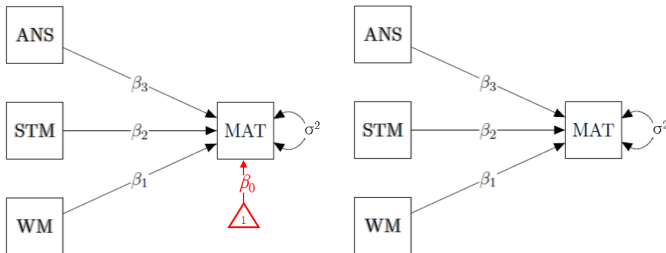
```
summary(lm.fit)$coefficients # LM regression table
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	20.03	9.61	2.09	0.04
WM	0.81	0.20	4.10	0.00
STM	1.01	0.36	2.81	0.01
ANS	0.23	0.11	2.16	0.03

Residual variance  $\sigma^2$ :

```
summary(lm.fit)$sigma^2  
[1] 75.94542
```

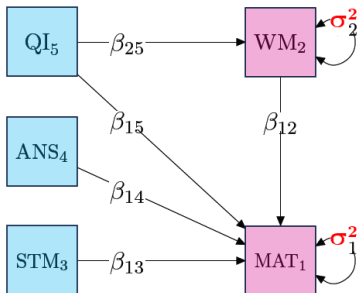
This model can be graphically represented as a path diagram and further simplified by removing the **intercept**  $\beta_0$  (note: **triangles represent constants**)



How many parameters? **Five**: Intercept, 3 slopes, residual variance

## Multivariate path models

In the previous example, we only considered **bivariate relationships** (i.e., 2 variables at a time, controlling for other variables). But what if we include IQ as a common predictor of both WM and MAT? We would have 3 variables interacting at the same time.



Both **MAT** and **WM** are **endogenous variables** because they receive 1+ arrow(s) and have error variance  $\sigma^2$ .

In contrast, **STM**, **ANS**, and **QI** are **exogenous variables** because they do not receive any arrow and have no errors.

A single LM equation is insufficient to describe this model. We need 2 separated equations: one for each variable that depends upon another variable

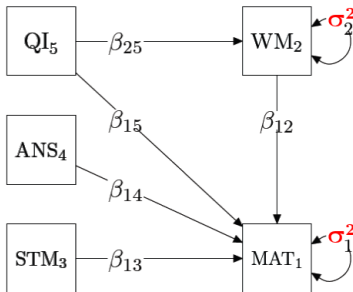
Path analysis (and SEM) uses **one equation per endogenous variable**:

$$\begin{cases} MAT_1 = \beta_{12}WM_2 + \beta_{13}STM_3 + \beta_{14}ANS_4 + \beta_{15}QI_5 + \epsilon_1 \\ WM_2 = \beta_{25}QI_5 + \epsilon_2 \end{cases}$$

# Graphical notation (1/3): Error terms

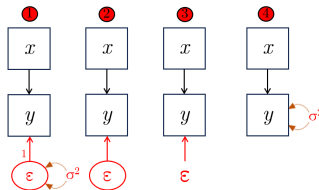
$$\begin{cases} MAT_1 = \beta_{12}WM_2 + \beta_{13}STM_3 + \beta_{14}ANS_4 + \beta_{15}QI_5 + \epsilon_1 \\ WM_2 = \beta_{25}QI_5 + \epsilon_2 \end{cases}$$

**Errors** = *residuals* or *disturbances* = discrepancy between observed and predicted values (as in LM!), they represent something *unexplained* = **exogenous** and *not directly observable* = **latent**



$\sigma^2$  = variance of a variable error (residual var.)

**b** Alternative ways to represent errors: some highlight their latent nature (#1 and #2), some highlight their variance (#1 and #4), and some highlight both (#1).

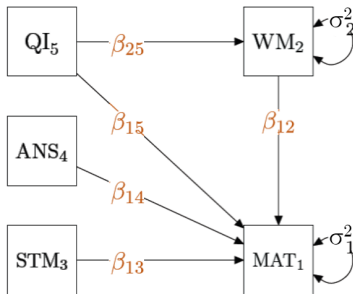


In this course, we use notation #4.

## Graphical notation (2/3): Arrows & coefficients

$$\begin{cases} MAT_1 = \beta_{12}WM_2 + \beta_{13}STM_3 + \beta_{14}ANS_4 + \beta_{15}QI_5 + \epsilon_1 \\ WM_2 = \beta_{25}QI_5 + \epsilon_2 \end{cases}$$

**Arrows** = *relationships* between 2 variables (*paths* or *slopes*) or between a variable and itself (*residual variance*), such that we do not include an arrow when a relationship is not expected (e.g., between QI and ASN) → path models are *complete*



**How to index variables and paths:**

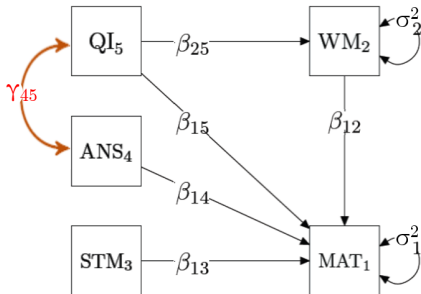
- **Variables** are indexed from the one receiving most arrows ( $MAT_1$ ) to the last exogenous variable ( $QI_5$ )
- **Path coefficients**  $\beta$  are indexed by firstly reporting the index of the endogenous variable and then that of the exogenous variable

**From plot to equations:** endogenous v.  $\sim$  sum of all linked exogenous v. + error

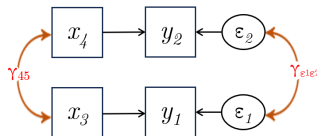
## Graphical notation (3/3): Covariances

$$\begin{cases} MAT_1 = \beta_{12}WM_2 + \beta_{13}STM_3 + \beta_{14}ANS_4 + \beta_{15}QI_5 + \epsilon_1 \\ WM_2 = \beta_{25}QI_5 + \epsilon_2 \\ Cov(ANS_4, QI_5) = \gamma_{ANS_4, QI_5} \end{cases}$$

**Covariances** = *non-directional (symmetric)* relationships between 2 *exogenous* v.



- Covariances are usually *not* reported in the system of equations, but they are graphically represented with (rounded) **double-headed arrows**
- Endogenous variables cannot covary but their errors  $\epsilon$  can**



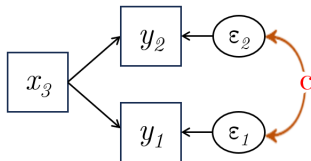


# Regression, partial correlation, and path coefficients

**Path coefficients** (single-headed arrows) are **partial regression coefficients** (*slopes*): as in LM, they index the *effect* of  $x$  on  $y$  by controlling for (i.e., after removing the effect of) other predictors, which are fixed to zero

**Covariances** between two exogenous variables (double-headed arrows), or between the errors of two endogenous variables, are **partial correlation coefficients**: they express the relationship between two variables by controlling for (i.e., after removing the effect of) all other correlated variables, which are fixed to zero

For instance, the figure below (source: Beaujeau, 2014) shows a path model of a partial correlation. Variables  $y_1$  and  $y_2$  are not allowed to covary since they are endogenous, but their errors are allowed to do so. Thus, the  $c$  coefficient is the relationship between  $y_1$  and  $y_2$  after removing the effect of  $x_1$  from both variables.



## Graphical notation: Recap



Directional (asymmetric) relationship



Non-directional (symmetric) relationship  
(covariance/correlation)



**Endogenous** observed variable  
with associated **variance  $\sigma^2$  of errors  $\epsilon$**



**Exogenous** observed variable  
without associated error



**Covarying** exogenous variables

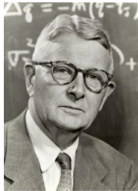


Endogenous variable  
with **covarying errors**



Constant (intercept)

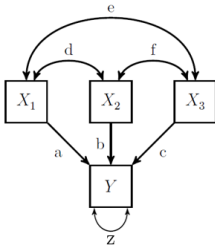
# Tracing rules & path coefficients



Sewall Wright (1889–1988): US geneticist that firstly developed rules for how to estimate values for a path model's coefficients by tracing the paths within it (i.e., path analysis).

**Tracing rules** = rules to estimate the covariance between 2 variables by *summing* the appropriate connecting paths:

1. Trace all paths between 2 variables multiplying all coefficients
2. Start by going backwards along single-headed arrows, no loops
3. Once you start going forward, you cannot no longer go back
4. Each path can only include one double-headed arrow



Starting from *observed covariances* (or correlations), we can compute the value of path coefficients. For instance, to compute path *a* starting from the observed correlations between  $X_1$  and  $Y$  (e.g.,  $r = .70$ ), between  $X_1$  and  $X_2$  (e.g.,  $d = .24$ ), and between  $X_1$  and  $X_3$  (e.g.,  $e = .20$ ):

$$r_{X_1, Y} = a + db + ec \rightarrow .70 = a + .24c + .20b \rightarrow a = .70 - .24c - .20b$$

## Standardized vs. *un*standardized solution

Path coefficients are **partial regression coefficients** (relationship between an exogenous  $x$  and an endogenous variable  $y$ , controlling for all other exogenous variable affecting  $y$ ). Similar to LM, they can be either *un*standardized or standardized:

- **Unstandardized coefficients** are obtained when the model is fitted on the variables expressed in their *natural metrics* (raw score units of measurement) → useful when raw score units are meaningful (e.g., age, meters, bpm) and when comparing the same variable relationship across samples
- **Standardized coefficients (ranging from -1 to 1)** are obtained when the model is fitted on standardized variables (i.e., variables transformed into ***z-scores***:  $z_{x_i} = (x_i - \bar{x})/s_x$ ) → useful to compare coefficients within the same model and/or the same sample

[b](#) To standardize an unstandardized coefficient:  $b^* = b \times (s_Y/s_X)$

[b](#) To unstandardize a standardized coefficient:  $b = b^* \times (s_X/s_Y)$

# That's all for now!

**Questions?**

**Homework** (optional):

- read the slides presented today  
and write in the Moodle forum if you have any doubts
- exeRcises 14-15 from exeRcises.pdf

---

For each exercise, the solution (or one of the possible solutions) can be found in dedicated chunk of commented code within the `exeRcises.Rmd` file

# Credits

The present slides are partially based on:

- Altoè, G. (2023) Corso Modelli lineari generalizzati ad effetti misti - 2023.  
<https://osf.io/b7tkp/>
- Beaujean, A. A. (2014) Latent Variable Modeling Using R. A Step-by-Step Guide. New York: Routledge
- Finch, W. H., Bolin, J. E., Kelley, K. (2014). Multilevel Modeling Using R (2nd edition). Boca Raton: CRC Press
- Pastore, M. (2015). Analisi dei dati in psicologia (e applicazioni in R). Il Mulino.
- Pastore, M. (2021). Analisi dei dati in ambito di comunità

# Achronyms & Greek letters

- CFA: confirmatory factor analysis
- LM: linear models/modeling
- LV: latent variable
- OV: observed variable
- SEM: structural equation models/modeling
- SS: sum of squares
- $\beta = \textit{beta}$ , indexing path coefficients (or regression coefficients)
- $\epsilon = \textit{epsilon}$ , indexing the error of an observed variable
- $\sigma = \textit{sigma}$ , indexing the variance  $\sigma^2$  of the errors  $\epsilon$
- $\eta = \textit{eta}$ , indexing latent variables
- $\theta = \textit{theta}$ , indexing overall model parameters

# Achronyms & Greek letters

- CFA: confirmatory factor analysis
- LM: linear models/modeling
- LV: latent variable
- OV: observed variable
- SEM: structural equation models/modeling
- SS: sum of squares
- $\beta = \textit{beta}$ , indexing path coefficients (or regression coefficients)
- $\epsilon = \textit{epsilon}$ , indexing the error of an observed variable
- $\sigma = \textit{sigma}$ , indexing the variance  $\sigma^2$  of the errors  $\epsilon$
- $\eta = \textit{eta}$ , indexing latent variables
- $\theta = \textit{theta}$ , indexing overall model parameters
- ciao