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

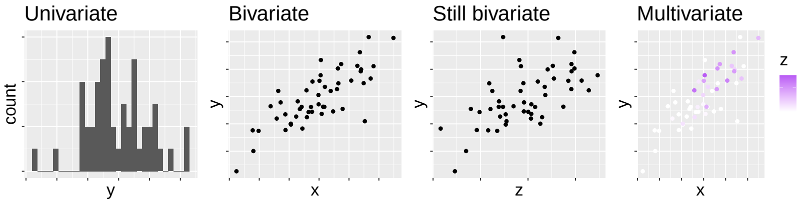
Outline of Part 2

- **`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

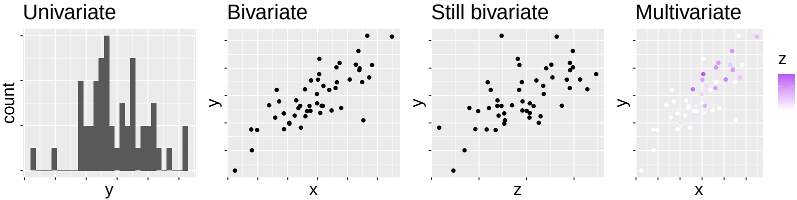
= exercises with R (bring your laptop!)

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)

Multivariate analyses for a multivariate reality

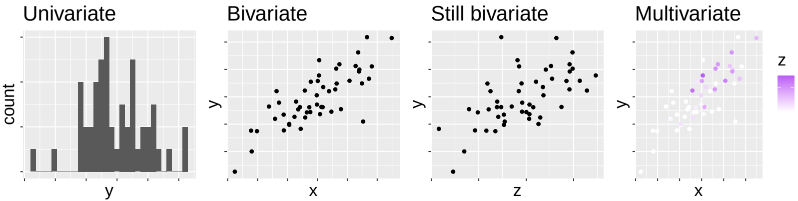


- In psychology, we mainly inspect empirical data focusing on **univariate** (y) or **bivariate** relationships (either y by x or y by z)
- But reality (particularly psychosocial reality) is complex, it is **multivariate** i.e., more than two variables covarying at the same time

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**

Multivariate analyses for a multivariate reality



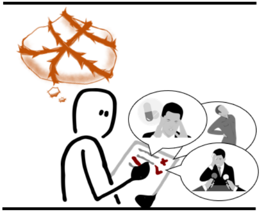
- In psychology, we mainly inspect empirical data focusing on **univariate** (y) or **bivariate** relationships (either y by x or y by z)

- 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**

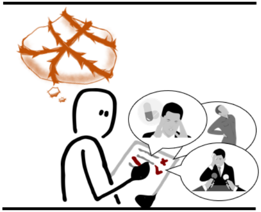
- **Structural equation modeling (SEM)** allow to analyze the relationships of interest by accounting for the multivariate reality of psychosocial phenomena (e.g., y by x covarying with z; x affects y through z)

Observed indicators & latent variables



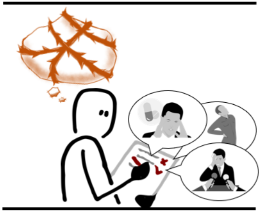
- In psychology, we are mainly interested in **latent variables** = phenomena that we cannot directly observe, but we can estimate from 1+ **observed indicators** (e.g., 10-item scale measuring anxiety)

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

Observed indicators & latent variables

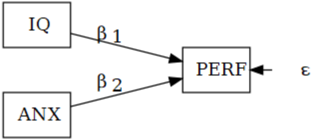


- In psychology, we are mainly interested in **latent variables** = phenomena that we cannot directly observe, but we can estimate from 1+ **observed indicators** (e.g., 10-item scale measuring anxiety)
- Are we allowed to do that? Yes (let's say yes), provided that we trust the indicator **construct 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:
$PERF = \beta_1 IQ + \beta_2 ANX + \epsilon$

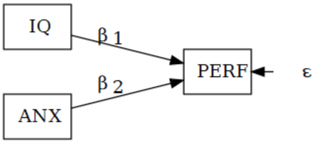


LM can only predict **one dependent variable at a time**, being either *univariate* (without predictors, i.e., intercept-only) or *bivariate* (with predictors).

Structural what!?

Structural equation modeling (SEM)
= multivariate *linear* models formalized by **systems of equations**

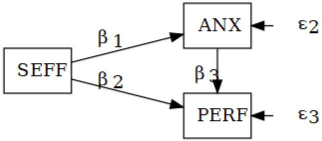
Linear models (LM): determining the link between a dependent and 1+ independent variables through a **single equation** like:
$PERF = \beta_1 IQ + \beta_2 ANX + \epsilon$



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:

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

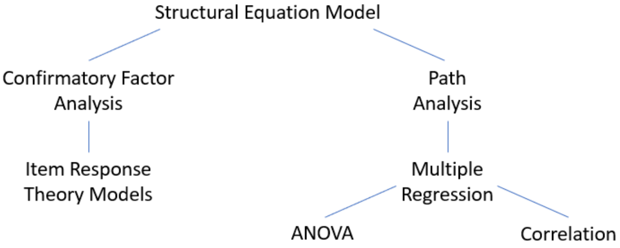


The SEM family

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:

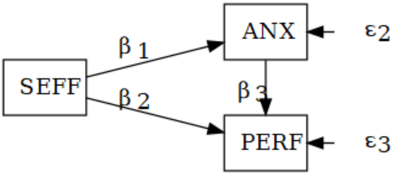
- **Path analysis**: multivariate linear models with observed variables only
- **Confirmatory factor analysis (CFA)**: multivariate linear models with both observed and latent variables



Source: Beaujean (2014)

Path models & path analysis

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

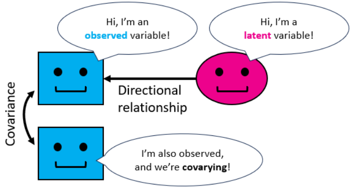


Paths = arrows (*edges*) linking the variables (*nodes*) in a model

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

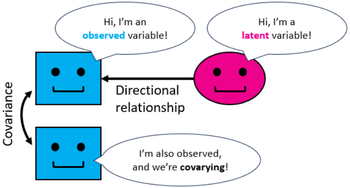
Latent factors & CFA

- **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, **OV**s are represented by squares/rectangles and indexed with lower case letters (e.g., x), whereas **LV**s are represented by circles/ellipses and indexed by the Greek letter η



Latent factors & CFA

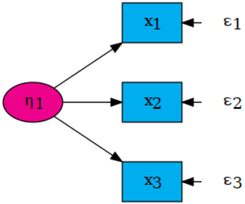
- **Observed/Manifest variable (OV)**
  variable that is directly observable (e.g., height, heart rate, item responses)
- **Latent variable/factor (LV)**
  variable that is *not* directly observable (e.g., anxiety, intelligence), but can be indexed by one or more observed variables
- In SEM, **OV**s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



Confirmatory factor analysis (CFA)

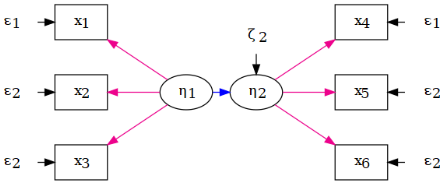
= analysis of the relationships (*factor loadings*) between a set of OVs and one or more LVs

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 OVs 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

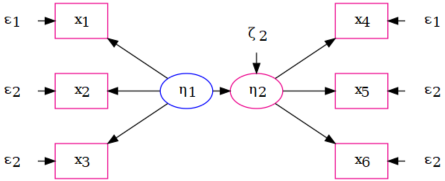
Notes:

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 ϵ (OV) or ζ (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**.

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

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

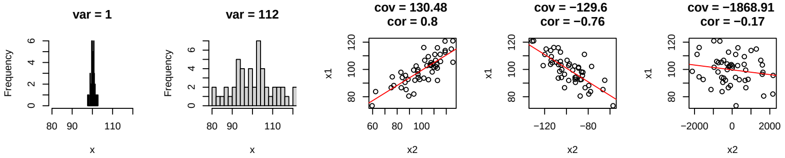
```
       MAT    QI    WM  STM
MAT 100.70 24.89 17.21 7.99
QI   24.89 19.43  6.69 4.04
WM   17.21  6.69 17.33 2.23
STM   7.99  4.04  2.23 5.34
```

SEM estimate a number of parameters θ so that the **implied covariance matrix** $\hat{\sum}(\theta)$ (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**

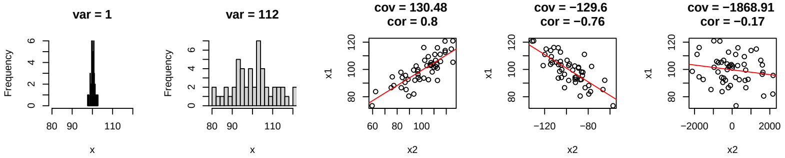
Covariance & correlation

- **Variance** = Expected value of the **squared deviation from the mean** of a random variable, or degree to which it deviates from its expected value
  $var(x) = \sigma_x^2 = \frac{\sum (x_i - \overline{x})^2}{N}$



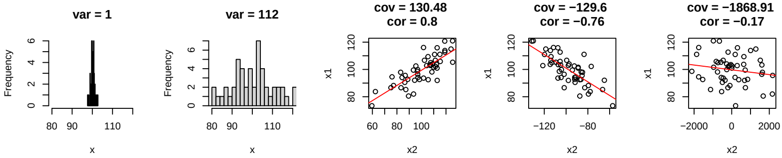
Covariance & correlation

- **Variance** = Expected value of the **squared deviation from the mean** of a random variable, or degree to which it deviates from its expected value
  $var(x) = \sigma_x^2 = \frac{\sum (x_i - \overline{x})^2}{N}$
- **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 $+\infty$)
  $cov(x_1, x_2) = \frac{\sum (x_{1i} - \overline{x_1})(x_{2i} - \overline{x_2})}{N}$



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
  $var(x) = \sigma_x^2 = \frac{\sum (x_i - \overline{x})^2}{N}$
- **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 $+\infty$)
  $cov(x_1, x_2) = \frac{\sum (x_{1i} - \overline{x_1})(x_{2i} - \overline{x_2})}{N}$
- **Correlation** = standardized covariance of two random variables
  Correlation ranges from -1 (perfectly negative) to +1 (perfectly positive)
  $cor(x_1, x_2) = \frac{cov(x_1, x_2)}{\sigma_{x_1}^2 \sigma_{x_2}^2}$



Covariance matrix (S)

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

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

```r
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
```

Properties of the covariance matrix:

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

SEM estimate a number of parameters θ so that the **implied covariance matrix** $\hat{\sum}(\theta)$ (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

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

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

- β = *beta*, indexing path coefficients (or regression coefficients)
- ϵ = *epsilon*, indexing the error of an observed variable
- σ = *sigma*, indexing the variance σ^2 of the errors ϵ
- η = *eta*, indexing latent variables
- θ = *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

- β = *beta*, indexing path coefficients (or regression coefficients)
- ϵ = *epsilon*, indexing the error of an observed variable
- σ = *sigma*, indexing the variance σ^2 of the errors ϵ
- η = *eta*, indexing latent variables
- θ = *theta*, indexing overall model parameters
- ciao