ADVANCED DATA ANALYSIS
FOR PSYCHOLOGICAL SCIENCE

Introduction and general course information

Luca Menghini Ph.D.

luca.menghini@unipd.it

***

Master degree in Developmental and Educational Psychology

University of Padova

2023-2024

The prof
●○○○

The course
○○○○○○○○○○

useR!
○○○○○○

The exam
○○○

IntRoduction
○○○○○○○○○

Resources
○○○

My career path

Luca Menghini Ph.D.
Work & Org. Psychologist
Postdoc in Applied psychology &
Quantitative research methods
@uniTN

- 2014: Bsc in Work & Social Psych Sciences @uniPD
  Biofeedback training for work stress management

- 2016: Msc in Social, Work, & Communication Psych @uniPD
  Psychophysiological workplace stress assessment protocol

- 2017: Psychology consultance internship @InsidePerformance
  Stress managament & biofedback training in professional sport drivers
  + Research internship @uniPD *Actigraphy in sleep research* ← ®️ ☺️

- 2020: Ph.D. in Psychological Sciences @uniPD
  Ecological momentary assessment of workplace stress ← 🎲 ☺️

- 2020: Visiting scholar @SRI International (CA, USA)
  Wearable sleep trackers, sleep and stress in adolescent insomnia

- 2021: Postdoc @uniBO
  Workaholism and daily fluctuations in blood pressure, emotional exhaustion, and sleep quality

- 2022: Postdoc @uniTN *Youth, transitions, challenges, & opportunities*

Some of my studies related to the course content

Intensive longitudinal designs
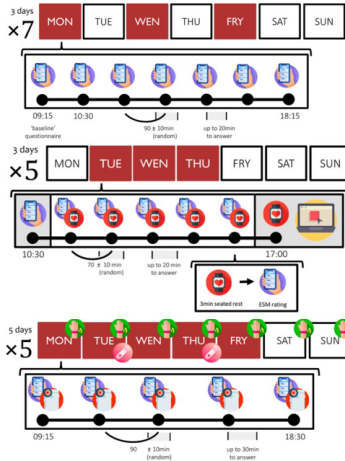
- **Multilevel**: Menghini et al. (2023). Wearable and mobile technology to characterize daily patterns of sleep, stress, pre-sleep worry, and mood in adolescent insomnia. *Sleep Health 9*(1), 108-116. [FULL-TEXT] [R CODE]

- **Multivariate**: Menghini et al. (2022). Italian adaptation of the Warr's Job-related Affective Wellbeing Scale: Factorial structure and relationships with the HSE Management Standards Indicator Tool. *TPM, 29*(3), 309-325. [FULL-TEXT] [R CODE]

- **Multilevel & Multivariate**: Menghini et al. (2022). Workplace Stress in Real Time: Three Parsimonious Scales for the Experience Sampling Measurement of Stressors and Strain at Work. *European Journal of Psych Assessment.* [FULL-TEXT] [R CODE]

Warnings

- I'm not a statistician
- I'm not a mathematical psychologist
- I'm not a programmer

I'm an Applied psychologist passionate about modeling and psychometrics.

Plus, this is my first time with this course: suggestions and critiques are welcomed!

_____

Ethical code for psychology research: Explicit aknowledgement of limitations

Contact, office hours, & master theses

Contact: Moodle or mail to: luca.menghini@unipd.it

Office hours: Wednesday 3.30 PM - 5.30 PM

Where: Psico 1 pink building, ground floor, between the computer rooms

We can also schedule Zoom meetings

Theses: As a contracted professor, I can only accept up to 2 master students

Ongoing projects: Ambulatory: Workaholism daily diary;

In-lab: ECG smart t-shirt, rumination & response inhibition

Advanced data analysis for psychological science

This course aims at providing basic notions of **multi-level & multi-variate** linear regression modeling, focusing on applications in developmental, educational, and applied psychology.

The course aims at transmitting **basic knowledge** on *linear mixed-effects regression* (LMER) and two common examples of multivariate techniques within the structural equation modeling (SEM) framework, namely *path analysis & confirmatory factor analysis* (CFA).

The course also aims at providing **practical competences** on advanced data analysis, with a particular emphasis on data preparation and pre-processing, model fit and evaluation, coefficient interpretation, and data visualization. Also, we'll try to **understand the results reported in published papers**.

The course is characterized by an **applied approach** that prioritizes real case studies and includes **practical exercises** (optional) using R.

The prof
0000

The course
0●00000000

useR!
000000

The exam
000

IntRoduction
000000000

Resources
000

Prerequisites

Students should have good knowledge about basic concepts linked to probability theory and associated topics (e.g., random variables, probability distributions, hypothesis testing), including **linear regression modeling** (but we will review that).

The prof
○○○○

The course
○○●○○○○○○○

useR!
○○○○○○

The exam
○○○

IntRoduction
○○○○○○○○○

Resources
○○○

Course contents

1. Intro and course info 📍

Multi-level

2. Introduction: From `lm()` to `lmer()`

3. Data preparation ®

4. Model fit & random effects ®

5. Coefficient interpretation ®

6. Model evaluation & selection ®

7. Generalized models `glmer()`,
   Bayesian LMER, power analysis 🔬

Multi-variate

8. Introduction: From `lm()` to `sem()`

9. Observed variables & path analysis

10. Data preparation ®

11. Model evaluation & selection ®

12. Coefficient interpretation ®

13. Latent variables & CFA ®

14. Multilevel SEM, Bayesian SEM,
    power analysis 🔬

The course program will end where we arrive.

———

® = Practical exercise sessions with R (bring your PC!)

🔬 = In-depth topics (not for the exam!)

The prof
○○○○

The course
○○○●○○○○○○

useR!
○○○○○○

The exam
○○○

IntRoduction
○○○○○○○○○

Resources
○○○

When & where

The course will last 42 hours (6 ECTS).

All lectures will be delivered in the Psico 2 gray building, room 3F - via Venezia 12.

Day	Date	Time	Room
1	10-4 (wed)	12:30-14:30	3F
2	10-5 (thu)	08:30-10:30	3F
3	10-11 (wed)	12:30-14:30	3F
4	10-12 (thu)	08:30-10:30	3F
5	10-18 (wed)	12:30-14:30	3F
6	10-19 (thu)	08:30-10:30	3F
7	10-25 (wed)	12:30-14:30	3F
8	10-26 (thu)	08:30-10:30	3F
9	11-2 (thu)	08:30-10:30	3F
10	11-8 (wed)	12:30-14:30	3F

Day	Date	Time	Room
11	11-9 (thu)	08:30-10:30	3F
12	11-15 (wed)	12:30-14:30	3F
13	11-16 (thu)	08:30-10:30	3F
14	11-22 (wed)	12:30-14:30	3F
15	11-23 (thu)	08:30-10:30	3F
16	11-29 (wed)	12:30-14:30	3F
17	11-30 (thu)	08:30-10:30	3F
18	12-6 (wed)	12:30-14:30	3F
19	12-7 (thu)	08:30-10:30	3F
20	12-13 (wed)	12:30-14:30	3F
21	12-14 (thu)	08:30-10:30	3F

Course materials

All course materials can be accessed from the Moodle page of the course
and from https://github.com/Luca-Menghini/advancedDataAnalysis-course

The contents required by the exam are exhaustively covered in the **main course slides**:
1. **Intro & course info** (the present slides)
2. **Multilevel modeling**
3. **Multivariate modeling**

Suggested **textbooks to deepen the topics of the course:**

- Finch, W. H., Bolin, J. E., Kelley, K., Multilevel Modeling Using R (2nd edition). Boca Raton: CRC Press, 2014

- Beaujean, A. A., Latent Variable Modeling Using R. A Step-by-Step Guide. New York: Routledge, 2014

The prof
oooo

The course
ooooo●ooooo

useR!
oooooo

The exam
ooo

IntRoduction
ooooooooo

Resources
ooo
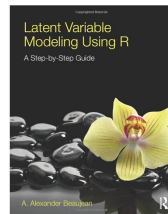
Course slides

The course slides are structured by intermixing theory, R code, plots, examples, and exercises. The slides are *dense* but we'll stay on each slide as long as needed ;)

The R code used in any exercise/example is also provided.

Slides and sentences with the microscope 🔬 icon cover in-depth but still useful topics that will be possibly presented but are not part of the core course topics and **not required for the exam**!

All course materials can be accessed from Moodle ***and*** from
https://github.com/Luca-Menghini/advancedDataAnalysis-course

The prof
○○○○

The course
○○○○○○●○○○

useR!
○○○○○○

The exam
○○○

IntRoduction
○○○○○○○○○

Resources
○○○

Additional resources & extra slides

Additional resources that are not presented during classes will be also available from Moodle and Github. These will include published papers and online resources, R code and exercises, extra slides, and other.

For instance, you can already find the `R-intro.pdf` extra slides (introduction to R), and you can already give a first look at the "Latent Variable Modeling using R" book website (e.g., "R syntax" section):
https://blogs.baylor.edu/rlatentvariable/

PSICOSTAT meetings & workshops:

Interdisciplinary research group on quantitative psychology, psychometrics, psychological testing, & statistics - monthly online meetings + weekly in person workshops
https://psicostat.dpss.psy.unipd.it/index.html

Teaching modalities

📽 Frontal theoretical sessions on the rationale of the analytical techniques focused by the course

💻 Practical sessions with individual and group exercises

Practical sessions will be based on the freely-available ® software.
Students are encouraged to bring their **laptops**, if possible.

The course will emphasize practical examples and **cases studies**
in developmental, educational, and applied psychology.

The prof
○○○○

The course
○○○○○○○○○●○○

useR!
○○○○○○

The exam
○○○

IntRoduction
○○○○○○○○○

Resources
○○○

Case studies:
What are your research and/or applied interests?

The prof
oooo

The course
ooooooooo●

useR!
oooooo

The exam
ooo

IntRoduction
ooooooooo

Resources
ooo

Attending the course: "PACATE"

Participation: You are expected to contribute to the class by participating in class discussion and working with each other during practical sessions. If you find something unclear or discordant with other information, please tell it to the professor. If you find it uncomfortable to speak up in class, feel free to contact the professor and work on this skill.

Attendance: Class attendance is not mandatory but encouraged. It is recommended to gradually but constantly familiarize with the content of the course.

Collaboration: Please, help each other, for instance, by working together on practical sessions and assignments, and/or by exchanging notes and useful materials for the exam.

Assignments: Over the course, several exercises/homework will be *suggested* to consolidate the course contents. While some of these will be discussed in class, feel free to contact the professor if you find any issue with the assignments.

Timekeeping: You are expected to be on time. You should be in your seat and ready to begin class when the class starts.

Exe®cise!

The prof
○○○○

The course
○○○○○○○○○○

useR!
●○○○○○

The exam
○○○

IntRoduction
○○○○○○○○○

Resources
○○○

An inseparable companion

R is a programming language and a programming environment for **statistical computing** and **graphics**.

It is based on the S language (Becker & Chambers, 1984), subsequently used to develop the S-Plus sofware and then R, originally created in 1996 by Ross Ihaka and Robert Gentleman.

Today, it is supported by an international research group (R Core Team and R Foundation for Statistical Computing) that periodically update (each year) the base sofware (***Base R***).

Progressive and exponential inclusion of new **packages** that extend its capabilities.

https://www.r-project.org/

A bottom-up resource

R provides a wide range of statistical and graphical techniques. It is designed to be **user-friendly** but at the same time to generate **high-quality outputs** (graphics, tables, and reports with equations, mathematical symbols, etc.).

Optimized default functions
+ dedicated packages
+ possibility to fully control.

Differently from other statistical software that implement multilevel and multivariate analyses (e.g., Mplus), R is a **free sofware** (GNU General Public Licence) that can be used anywhere worldwide, it is an **open-source software** (all functions are documented and can be inspected in detail), and **works on all main OS**: Windows, MacOS, and UNIX (e.g. Linux)

Moreover, there is a massive community of useRs: For any issue, just Google it! 😊

The prof
○○○○

The course
○○○○○○○○○○

useR!
○○●○○○

The exam
○○○

IntRoduction
○○○○○○○○○○

Resources
○○○

Even better than googling

Try https://chat.openai.com/ or https://rtutor.ai/

> **L** How can I compute the mean of 3, 5, and 8 in one line of R code?

> You can compute the mean of the numbers 3, 5, and 8 in one line of R code like this:

```R
mean_value <- mean(c(3, 5, 8))
```

RTutor   Home   Data   Report   EDA   Ask   About

Dataset: mpg   Reset

Example requests:

...

Specify a multilevel model and print the fixed effects.

Submit   Settings   ☐ Python

AI generated code: ☐ Continue from

Chunk #4

```R
library(lme4)
model <- lmer(dis ~ city + highway + (1 | model), data = df)
summary(model)$coefficients[,"Estimate"]
```

Results:

```
(Intercept)        city      highway
5.80940320 -0.16226330  0.02046682
```

Introduction to ®: Additional materials

Extra slides on Moodle/GitHub

- `extra/R-intro.pdf`: How to install and get started with `R` and `RStudio`, elementary commands, `R` objects, functions, and workspace, how to read and export datasets, `R` graphics, and linear models. Link to extra slides

Free tutorials

- Navarro, D. Learning statistics with R: A tutorial for psychology students and other beginners - https://learningstatisticswithr.com/

- `learnr`: an R package for learning how to use R https://rstudio.github.io/learnr/

- excellent STAT545: Data wrangling, exploration, and analysis with R https://stat545.com/

The prof ○○○○
The course ○○○○○○○○○○
useR! ○○○○○●○
The exam ○○○
IntRoduction ○○○○○○○○○
Resources ○○○

Key ℝ packages used in the course

The course uses several packages with customized and optimized functions. Here are the main packages used in the course (and the code to install all of them):

```r
pckg <- c("lme4","lavaan", # model fit
          "MuMIn","sjPlot","knitr" # model outputs & visualization
          "ggplot2","gridExtra" # visualization
          "plyr", # data preparation
          "lmtest","influence.ME","lattice","osfr") # other
install.packages(pckg)
```

To install a single package, just run:

```r
install.packages("package_name")
```

To load an installed package:

```r
library("package_name")
```

The prof
○○○○

The course
○○○○○○○○○○

useR!
○○○○○○●

The exam
○○○

IntRoduction
○○○○○○○○○

Resources
○○○

Some key ® functions used in the course

```r
# Aggregating scores by group
aggregate(x = sleep$extra,
          by = list(sleep$group), FUN = mean)
```

```
  Group.1    x
1       1 0.75
2       2 2.33
```

```r
# Merging wide- and long-form datasets
plyr::join(long, wide, by = "ID", type = "left")
```

```r
# Fitting LMER models and printing fixed effects
fit <- lmer(extra ~ group + (1|ID), data = sleep)
fixef(fit)
```

```
(Intercept)        group2
       0.75          1.58
```

```r
# Fitting SEM and printing coefficients
fit <- sem("visual =~ x1 + x2 + x3

            textual =~ x4 + x5 + x6

            visual ~ textual",

           data=HolzingerSwineford1939)
standardizedsolution(fit)[1:7,1:4]
```

lhs	op	rhs	est.std
visual	=~	x1	0.78
visual	=~	x2	0.43
visual	=~	x3	0.57
textual	=~	x4	0.85
textual	=~	x5	0.85
textual	=~	x6	0.84
visual	~	textual	0.46

When & where (to be updated!)

Note: These are the current dates and places for the exam, but they will probably change as I asked moving to a room with computers (the exam will be done on Moodle)

→ **Check the annoucements on Moodle for keeping updated**

Session	Date	Time	Room
1. Jan	2024-01-17	14:30	3L
2. Feb	2024-02-14	14:30	3L
3. Jun	2024-06-10	14:30	3L
4. Jul	2024-07-10	14:30	3L
5. Sept	2024-09-10	14:30	3L

The prof
○○○○

The course
○○○○○○○○○○

useR!
○○○○○○

The exam
○●○

IntRoduction
○○○○○○○○○

Resources
○○○

Exam structure & contents

The final exam will be **written**, computerized (on Moodle), and will last **40 min**. The exam will consist of **31 closed-ended (multi-choice) questions** on:

- theoretical topics covered by the course
- interpretation of data analysis outputs
- interpretation of paper results

Exam contents: The contents required by the exam are exhaustively covered in the main course slides. Note that the required contents will depend on where we arrive at the end of the course

Exam scores: The exam score will be computed as the sum of the scores obtained to the 31 questions (right answer = 1; wrong/missing answer = 0; 18+ right answers = exam passed; 31 right answers = 30 cum laude).

———

More information about the final exam will be provided later, along with an exam simulation.

Example questions

- **Theoretical topics**:

  What does "ME" mean in the "LMER" achronym?

  Mean Effect | Mixed Effects | Multilevel Effects | Multivariate Errors

  Which of the following is a fixed effect?

  Intercept | Residual variance | Cluster variance | None

- **Output interpretation**:

  Which is the variance of the random slope in the following table?

  0.75 | 0.76 | 2.85 | Not shown

  Which of the following random effects are shown in the table?

  Random slope | R. Intercept & residuals | Residuals & R. slope | None

Predictors	b (SE)	CI	p
(Intercept	409.50 (5.40)	$398.93 - 420.08$	**<0.001**
Stress	-7.19 (2.29)	$-11.68 - -2.70$	**0.002**
Group [Insomnia]	2.76 (7.57)	$-12.09 - 17.60$	0.716
Stress × Group [Insomnia]	2.92 (3.19)	$-3.33 - 9.17$	0.359
Random Effects			
σ^2		5071.75	
$\tau_{00 \ ID}$		1196.32	

Multi-LEVEL & Multi-VARIATE

Advanced statistical techniques to deal with **large & complex data structures**:

Multilevel regression model
To be used with **hierarchical data** where statistical units are *nested within* higher-level variables (clusters):

- students → classes → schools
- experiences → days → individuals
- trials → items & individuals

Linear mixed-effects regression (LMER) allows to estimate *fixed effects* that are constant across all clusters + *random effects* varying from cluster to cluster.

Multivariate regression model
To be used to account for modeling **mutliple variables interacting at the same time** (e.g., multiple outcomes, mediation).

Path model = Pictorial representation of a theory of variable relationships

Measurement model = relationships that form a *latent variable* from multiple *observed variables*

Structural equation model (SEM)
= measurement + structural model

Linear models as the common root

Linear regression models allow to determinate the link between two variables as expressed by a linear function: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Such a function can be graphically represented as a **straight line** where

β_0 is the **intercept** (value assumed by y when x = 0)

β_1 is the **slope** (predicted change in y when x increases by 1 unit)

ϵ_i is the **error** (distance between observation i and the regression line)
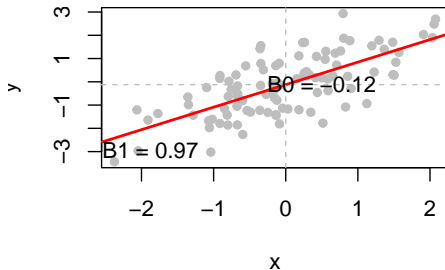
The only three formulas to keep in mind

Linear model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Mixed-effects model:

$$y_{ij} = (\beta_{00} + \lambda_{0j}) + (\beta_{10} + \lambda_{1j})x + \epsilon_{ij}$$

For each observation i and each cluster j, the intercept and the slope are decomposed into the **fixed** components β_0 and β_1 referred to the whole sample, and the **random** components λ_{0j} and λ_{1j} randomly varying between clusters

Structural equation model:

$$\begin{cases} y = \Lambda\eta + \epsilon \ (measurement) \\ \\ \eta = B\eta^* + \zeta \ (structural) \end{cases}$$

The **measurement model** clarifies the relationships Λ between the *observed variables* y and the corresponding *latent variables* η, whereas the **structural model** clarifies the relationship B between two or more latent variables η^*.
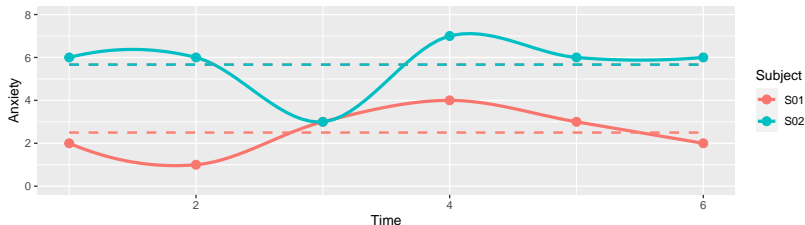
Multilevel models: Let the visuals talk

Visual introduction to multilevel modeling:

http://mfviz.com/hierarchical-models

Between & Within

When a random variable y is measured from multiple data points per subject (or group), multilevel modeling **partition the y variance** into the *within-subject* (level 1) & *between-subject* (level 2) components.

The same applyes when multiple individuals (e.g., students) are nested within a number of groups (e.g., schools) $\rightarrow$ *within-group* vs. *between-group*

The prof
oooo

The course
oooooooooo

useR!
oooooo

The exam
ooo

IntRoduction
ooooo●oooo

Resources
ooo

Multilevel models: Fixed vs. Random effects

In the literature, multilevel modeling is sometimes called with different terms, e.g., *hierarchical linear modeling*, *random slope models*, *variance component models*, …

All these models are part of the broader ***mixed-effects models*** family, identifying models with both fixed and random effects:

- **Fixed effects**: effects that remains constant across all clusters, whose *levels* are exhaustively considered (e.g., gender, levels of a Likert scale) and generally controlled by the researcher (e.g., experimental conditions)

- **Random effects**: effects that vary from cluster to cluster, whose *levels* are randomly sampled from a population (e.g., schools, participants, days, experimental stimuli)

The prof
oooo

The course
oooooooooo

useR!
oooooo

The exam
ooo

IntRoduction
oooooo●oooo

Resources
ooo

Multivariate models: Let the visuals talk

Linear regression: determining the link between a dependent and an independent variables through linear functions like:
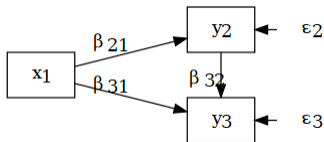
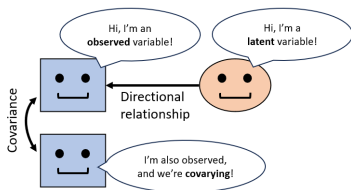$y = \beta_1 x_1 + \beta_2 X_2 + \epsilon$
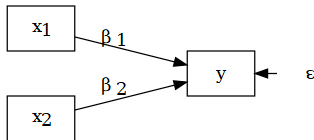
As a limitation, linear models can only predict **one dependent variable at time** with a single equation. They can be *univariate* (without predictors) or *bivariate* (with predictors).

Structural equation models (SEM) are *multivariate* models that allow simultaneously modeling multiple ~~dependent~~ *endogenous* variables with a **system of equations**:

$$\begin{cases} y_2 = \beta_{21} x_1 + \epsilon_2 \\ \\ y_3 = \beta_{31} x_1 + \beta_{32} Y_2 + \epsilon_3 \end{cases}$$

The prof
oooo

The course
oooooooooo

useR!
oooooo

The exam
ooo

IntRoduction
ooooooo●oo

Resources
ooo

# Multivariate models: observed vs. latent

Observed variables: directly observable and measurable (e.g., heart rate), represented by *squares* and the *lowercase letters* x (exogenous) and y (endogenous)

Latent variables: hypothetical and not directly measurable but **indexed** by one or multiple observed variables (e.g., happiness), represented by *circles* and the *greek letter* η

When including **observed variables only**, SEM are called **path analysis**
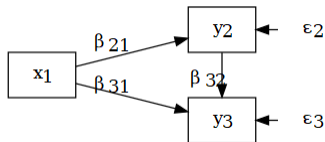
When **both observed and latent** variables are included, we can talk of **'full SEM'**
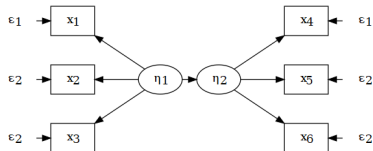
Multilevel & multivariate models:
It's a matter of theory!

While *any model is a formal representation of a theory* (Bollen, 1989), multilevel and multivariate models are particularly directly linked to the underlying theoretical model.

Multilevel modeling:
Theories determinate whether a clustering variable is meaningful or not, the number of levels to be considered (e.g., individuals, days, weeks, schools), and whether a given construct can be meaningfully attributed to a given level (e.g., happy people, happy days, happy schools).

SEM:
Theory determines how a latent variable is reflected by a set of observed variables (*measurement model*) and what are the regression-like relationships among the variables (*structural model*).

The prof
0000

The course
0000000000

useR!
000000

The exam
000

IntRoduction
00000000●

Resources
000

That's all for now!

- Any question?

- Next lecture: LM recap, ok?

- Needed a recap on how to use R?

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.

Useful resources: Multilevel

- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59*(4), 390-412.

- Bliese, P. (2022). Multilevel modeling in R (2.7).
  https://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf

- McElreath, R. (2020). Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.

- Pinheiro, J., & Bates, D. (2006). Mixed-effects models in S and S-PLUS. Springer science & business media.
  see also Bates, D. (2022). lme4: Mixed-effects modeling with R.
  https://stat.ethz.ch/~maechler/MEMo-pages/lMMwR.pdf

The prof
OOOO

The course
OOOOOOOOOO

useR!
OOOOOO

The exam
OOO

IntRoduction
OOOOOOOOO

Resources
OO●

Useful resources: Multivariate

- Kline, R.B. (2005). Principles and Practice of Structural Equation Modeling. Guilford Press, NY.

- Lin, J. Introduction to structural equation modeling (SEM) in R with lavaan. https://stats.oarc.ucla.edu/r/seminars/rsem/

- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*, 1-36.