ADVANCED DATA ANALySIS FOR PSYCHOLOGICAL SCIENCE

Part 1. Introduction to multilevel modeling

Luca Menghini Ph.D.

luca.menghini@unipd.it

***

Master degree in Developmental and Educational Psychology

University of Padova

2023-2024

Outline of Part 1

- **`lm()` recap**: Short recap of linear regression modeling
- **`lmer()`**: Introduction to multilevel modeling (aka *linear mixed-effects regression*, LMER)
- **Data structure**: How to approach a multilevel data structure, how to manipulate and pre-process multilevel data
- **Model fit**: How to fit a multilevel model in R, to evaluate model diagnostics, to interpret model results
- **Model evaluation**: How to evaluate a model, compare multiple models, and select the best model
- **Related topics**: In-depth topics related to multilevel modeling (e.g., generalized and Bayesian LMER, power analysis)

= not for the exam

= exercises with R (bring your laptop!)

Linear regression models

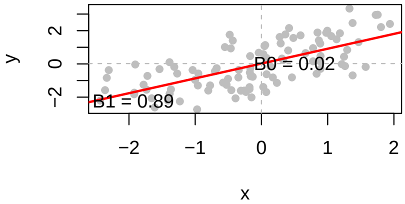
Linear regression models allow to determinate the link between two variables as expressed by a linear function: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

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 is the **residual variance** (distance from the regression line)



Notes: __________

x_i and y_i are the values of individual i for the **casual variables** x and y

β_0, β_1, and ϵ are called "**model parameters**" or "**coefficients**"

Fitting linear models in R

R uses the `lm()` function to fit linear models with the arguments `formula` (`y ~ x1 + x2 + ...`) and `data` (identifying the dataframe with the model variables).

```
data("children", package = "npregfast") # loading children dataset from npregfast pkg
```

Null model

Children' `height` is only predicted by the model **intercept** b_0 = expected (i.e., mean) value of `height` in the sample.

```
m0 <- lm(formula = height ~ 1,
         data = children)
coefficients(m0) # model coefficients
```

```
## (Intercept)
##    153.4013
```

Simple regression model

`height` is now predicted by the **intercept** b_0 (mean value when `age` is 0) and the **slope** b_1 (expected change for 1-unit increase in `age`)

```
m1 <- lm(formula = height ~ age,
         data = children)
coefficients(m1) # model coefficients
```

```
## (Intercept)         age
##   94.904099    4.388803
```

Notes: __________

`~` = tilde, on Windows: `Alt + 126`.

Multiple regression & interactions

LM also allow to include **multiple predictors** and the **interactions** among them. This is done by estimating a separate slope (thus, a separate line) for each predictor by *holding constant* the value of the other predictors, which are fixed to zero.

Multiple regression model

b_0 = expected value in girls with `age` $= 0$

b_1 = `age` effect within the same `sex`

b_2 = `sex` difference when `age` $= 0$

```
m2 <- lm(formula = height ~ age + sex,
         data = children)
coefficients(m2)
```

```
## (Intercept)         age     sexmale
##  95.0075706   4.3887983  -0.2001025
```

Interactive model

b_1 = `age` effect in `girls`

b_2 = `sex` difference in `height` when `age` $= 0$

b_3 = `sex` difference in `age` effect (**interaction**)

```
m3 <- lm(formula = height ~ age * sex,
         data = children)
round(coefficients(m3),2)
```

```
## (Intercept)         age     sexmale age:sexmale
##      104.25        3.70      -19.04        1.41
```

Notes: __________

- In this context, “effect” is used as a synonym of “relationship” (not a *causal* effect).

- The interaction (used in moderation analysis) is computed as the product of x_1 and x_2.

Model comparison & model selection

Likelihood ratio test

Testing the ratio of the log-*likelihoods* of two nested models (one model includes all predictors of the other model and the Y variable is the same)

```
library(lmtest)
lrtest(m0,m1,m2,m3)
```

#Df	LogLik	Df	Chisq	Pr(>Chisq)
2	-10417.84	NA	NA	NA
3	-8582.42	1	3670.84	0.0
4	-8582.19	1	0.45	0.5
5	-8468.86	1	226.67	0.0

Notes: __________

Likelihood = probability of observing your data given your set of parameters, sometimes reffered as the *evidence* of a model.

Information criteria

The Akaike (AIC) and the Bayesian Information Criterion (BIC) account for both likelihood and *parsimony* (the lower number of parameters the better)

```
# AIC: the lower the better
AIC(m0,m1,m2,m3)
```

```
## [1] 20839.68 17170.83 17172.39 16947.72
```

```
# Akaike weights: from 0 (-) to 1 (+)
library(MuMIn)
Weights(AIC(m0,m1,m2,m3)) # Aw
```

```
##  model weights
## [1] 0 0 0 1
```

Parameter estimation in linear regression models

b_0 and b_1 must be **estimated** using sample data taken from a population.

There are several ways to estimate unknown parameters (e.g., maximum likelihood, Bayesian approach), including the widely popular **ordinary least squares** (OLS), which aims at minimizing the sum of the squared residuals.

Linear model:

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

Predicted values:

$\hat{y}_i = \beta_0 + \beta_1 x_i$

Observed values:

$y_i = \hat{y}_i + \epsilon_i$

Residuals = observed - predicted

$\epsilon_i = y_i - \hat{y}_i$

But what are **residuals**?

```
head(data.frame(observed = children$height,
                predicted = fitted(m3),
                residuals = residuals(m3) ))
```

```
##   observed predicted  residuals
## 1   150.77  152.9026 -2.1326167
## 2   170.59  156.6139 13.9760532
## 3   167.31  160.3095  7.0005026
## 4   165.72  165.5202  0.1997761
## 5   171.67  160.3095 11.3605026
## 6   143.74  151.0706 -7.3306208
```

Statistical inference on regression coefficients

Based on NHST, it is possible to test the **statistical significance** of each regression coefficient (*two-tail t-test*), which is automatically done by R in the summary of the model.

```
summary(m3) # model results
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 104.247994 0.88181122 118.22031 0.000000e+00
## age           3.695551 0.06432249  57.45348 0.000000e+00
## sexmale     -19.043493 1.25746134 -15.14440 1.237494e-49
## age:sexmale   1.413741 0.09185516  15.39098 3.897810e-51
```

Effect size:

Coefficient of determination

$R^2 = 1$ - SS residuals/SS total

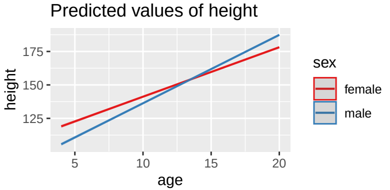
```
summary(m3)$r.squared
```

```
## [1] 0.79
```

The model explains 79% of the variance in height.

Plotting effects:

```
sjPlot::plot_model(m3,type="pred",terms=c("age","sex"))
```



Hands on R

1. Download & read the dataset from the *Pregnancy during the COVID-19 pandemics* study

`depr` = postnatal depression, `age` = mother's age, `NICU` = intensive care, `threat` = fear of COVID

```
library(osfr) # package to interact with the Open Science Framework platform
proj <- "https://osf.io/ha5dp/" # link to the OSF project (see protocol paper & data dictionary)
osf_download(osf_ls_files(osf_retrieve_node(proj))[2, ],conflicts="overwrite") # download
preg <- na.omit(read.csv("OSFData_Upload_2023_Mar30.csv",stringsAsFactors=TRUE)) # read dataset
colnames(preg)[c(2,5,12,14)] <- c("age","depr","NICU","threat") # shortening variable names
```

2. Explore the the variables `depr`, `threat`, `NICU`, and `age` (descr., corr., & plots)
3. Fit a null model `m0` of `depr`
4. Fit a simple regression model `m1` with `depr` being predicted by `threat`
5. Fit a multiple regression model `m2` also controlling for `NICU` and `age`
6. Fit an interactive model `m3` to check whether `age` moderates the relationship between `threat` and `depr`.
7. Compare the models with AIC and likelihood ratio test: which is the best model?
8. Print & interpret the coefficients estimated by the selected model
9. Print & interpret the statistical significance of the estimated coefficients
10. Plot the effects of the selected model
11. Compute the determination coefficient of the selected model

One step back: LM assumptions

Core assumptions:

1. Linearity: x_i and y_i are linearly associated $\rightarrow$ the expected (mean) value of ϵ_i is zero

2. Normality: ϵ_i are normally distributed $\rightarrow \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

3. Homoscedasticity: ϵ_i variance is constant over the levels of x_i (homogeneity of variance)

4. Independence of predictors & errors: x_i is unrelated to ϵ_i

5. Independence of observations: for any two observations i and j with $i \neq j$,
the residual terms ϵ_i and ϵ_j are independent

Additional assumptions:

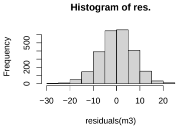
6. Absence of influential observations (multivariate outliers)

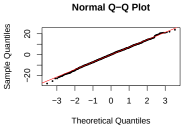
7. Absence of collinearity (for multiple regression):
lack of linear relationship between x_1 and x_2

LM diagnostics: Assessing LM assumptions

Normality & linearity ☺

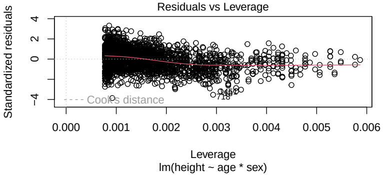
```
hist(residuals(m3))
qqnorm(residuals(m3)); qqline(residuals(m3))
```





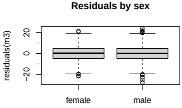
Absence of influential cases ☺

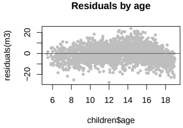
```
plot(m3,which=5)
```



Homoscedasticity & independence x, ϵ ☺

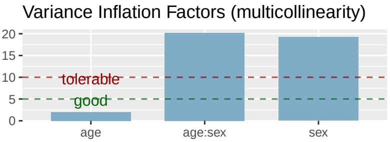
```
plot(residuals(m3) ~ children$sex)
plot(residuals(m3) ~ children$age)
```





Absence of collinearity (multiple regr.) ☹

```
sjPlot::plot_model(m3,"diag")[[1]]
```



Independence of observations ?

Are the unmeasured factors influencing y unrelated from one individual to another?

Cluster variables & nested data

In many cases, the *sampling method* creates **clusters** of *individual observations*

- students → schools
- children → families → neighborhoods → cities → regions → states → planets

Nested data structure (~ *multilevel* or *hierarchical* data structure)
= when data points at the **individual level** appear *in only one group*
of the **cluster level** variable

→ individual observations are ***nested*** within clusters

vs. 'crossed data structure' = individuals can appear in multiple clusters
e.g., after-school activities: a student can be enrolled in multiple activities

Notes: __________

Individual observation = statistical unit = individual entity within a sample or population that is the subject of data collection & analysis (not necessarily a person)

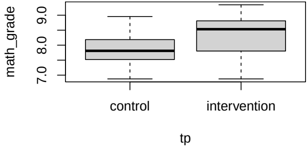
Case study: Innovative math teaching program

We're hired by a school principal to assess whether an innovative teaching program can improve in first-year high-school students' achievement in math.

```
table(itp[,c("classID","tp")])
```

	control	intervention
A	30	0
B	22	0
C	0	27
D	0	11

```
boxplot(math_grade ~ tp, data=itp)
```



The teaching program `tp` was delivered over the first semester to 2 out of 4 `classID` and we got the students' end-of-semester `math_grade` (1-10).

Nested data structure: students are *nested* within classes, with each student only belonging to one class.

Note: The **cluster variable** is related to both x (program delivered at the class level) and y (grades will be more similar between students belonging to the same class).

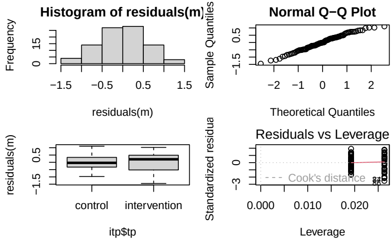
Non-independence of observations with nested data

Let's try with a **linear regression model**:

```
m <- lm(math_grade ~ tp, data=itp)
summary(m)$coefficients[,1:3]
```

```
##                Estimate Std. Error t value
## (Intercept)        7.85       0.08   97.54
## tpintervention     0.48       0.12    3.86
```

```
hist(residuals(m)); qqnorm(residuals(m))
boxplot(residuals(m)~itp$tp); plot(m,5)
```



- Coefficient meaning?
- Linear model assumptions?
- **Independent observations**?

Are ϵ_i and ϵ_j independent for any $i \neq j$? Are the unmeasured factors influencing y unrelated from one individual to another?

NO: students are nested within classes and such cluster variable is likely to explain differences in the y variable as well as in the relationship between x and y

Thus, we cannot rely on linear models to analyze these data.

Local dependencies

Mixed-effects models

Multilevel models are part of the largest mixed-effects family

E.g., when a subject changes group over time, it is still a mixed-effects model but not a multilevel model

Nested data & Multilevel data structure

Case study: Adolescent insomnia

Fitting a multilevel model (in R)

Case study: Adolescent insomnia

LMER assumptions

Diagnostics

pacchetti performance e sjPlot

Model comparison

AIC e BIC, weights

likelihood ratio test

Some topics related to multilevel modeling

- *Power analysis* of multilevel models
- *Generalized* linear mixed-effects regression (GLMER)
- *Bayesian* linear mixed-effects regression (BLMER)

Power analysis of multilevel models

glmer(): Generalized multilevel modeling (1/3)
Rationale

Generalized linear mixed-effects models (GLMER) are a **generalization** of LMER:

In addition to modeling normally distributed quantitative dependent variables (like classic LMER), they can also manage **non-normally distributed variables** such as:

- quantitative variables that only take positive values ← **Gamma**
- count variables ← **Poisson**
- binary/dichotomic variables ← **Binomial**

How is that possible?

glmer(): Generalized multilevel modeling (2/3)
Components of a GLMER model

GLMER models allow to model multiple types of dependent variables thanks to their three components:

- A **probability distribution** for the expected value of the y variable (e.g., normal, Gamma, Poisson, binomial)
- A **linear model** of the model predictors, including both *fixed* and *random* effects (LMER)
- A **link function** that translates the expected values of the y variable into the values predicted by the linear model

glmer(): Generalized multilevel modeling (3/3)
Example with Logistic regression

Logistic regression . . .

stan_glmer(): Bayesian multilevel modeling

Dire solo che esiste, fare un esempio con il pacchetto rstanarm,

Dire che convergono meglio xk lmer ha problemi di convergenza

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. Child development, 85(3), 842-860.

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 psicologie (e applicazioni in R). Il Mulino.

Useful resources

- Bates, D. (2022). lme4: Mixed-effects modeling with R. https://stat.ethz.ch/~maechler/MEMo-pages/lMMwR.pdf
- 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.