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 (*linear mixed-effects regression*)
- **Data processing**: How to approach a multilevel data structure? How to manipulate and pre-process multilevel data?
- **Descriptives**: Which descriptive stats should be reported from a multilevel dataset? How to compute and interpret them?
- **Model fit**: How to fit a multilevel model in R? How to inspect, report, visualize, and interpret the results of a multilevel model?
- **Model evaluation**: Which are the assumptions of multilevel models? How to evaluate them? How to compare multiple models and select the best model?
- **Related**: Summaries & in-depth topics related to multilevel modeling (e.g., generalized and Bayesian LMER, power analysis)

= not for the exam

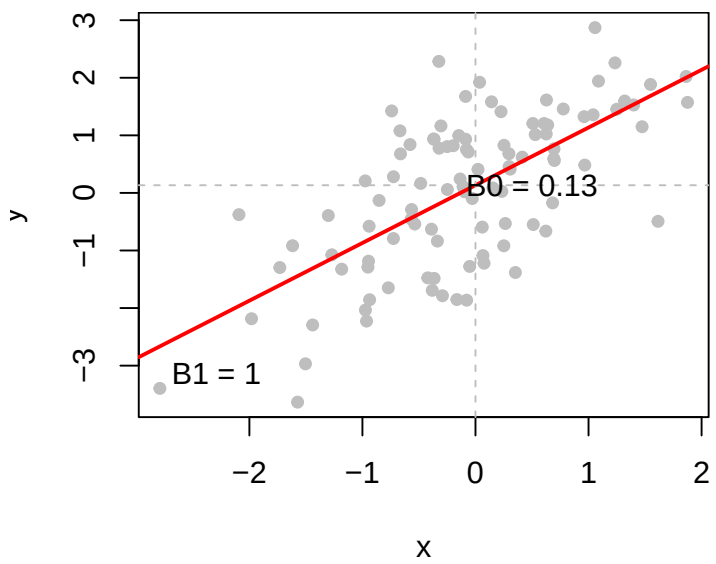
= exercises with R (bring your laptop!)

LM recap: Linear regression models

Linear models (LM) 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 are the **errors** (distance between observation i and the regression line)



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

β_0, β_1, and ϵ_i are called "**parameters**", or "**coefficients**". They are *estimated* from the sampled data and *generalized* to the whole population.

Fitting linear models in R

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

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).

Null model

Children' `height` is only predicted by the model **intercept** β_0 = expected (i.e., mean) value of `height` in the sample. σ^2 is the **variance of the residuals** ϵ_i (deviations from the intercept).

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

```
(Intercept)
   153.4013
```

```r
summary(m0)$sigma^2 # residual variance
```

```
[1] 243.9085
```

Simple regression model

`height` is now predicted by the **intercept** β_0 (mean value when `age` is 0), the **slope** β_1 (expected change for 1-unit increase in `age`), and the **residual variance** σ^2.

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

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

```r
summary(m1)$sigma^2 # residual variance
```

```
[1] 56.19656
```

Multiple regression & interactions

LM also allow to include **multiple predictors** and the **interactions**[1] 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

β_0 = expected value in girls with `age` = 0

β_1 = `age` effect[2] within the same `sex`

β_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

β_1 = `age` effect in girls

β_2 = `sex` difference in `height` when `age` = 0

β_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
```

---

[1] The **interaction** between x_1 and x_2 is computed as the **product of** x_1 **and** x_2.

[2] In this context, "effect" is used as a synonym of "relationship" (not a *causal* effect).

Model comparison & model selection

Likelihood ratio test

Compares the *fit* of two *nested* models (i.e., predicting the same y variable, with the more complex model including all predictors included in the simpler model).

```
library(lmtest)

lrtest(m0,m1,m2,m3) # returns Chisq statistic
```

```
  #Df    LogLik Df   Chisq   Pr(>Chisq)
1   2 -10417.84 NA      NA           NA
2   3  -8582.42  1 3670.84 0.000000e+00
3   4  -8582.19  1    0.45 5.046155e-01
4   5  -8468.86  1  226.67 3.176229e-51
```

Information criteria

The Akaike (AIC) and the Bayesian Information Criterion (BIC) compare multiple models in terms of *fit & parsimony* (the lower number of parameters the better)

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

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

```
# Akaike weights: from 0 (-) to 1 (+)

MuMIn::Weights(AIC(m0,m1,m2,m3))
```

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

Here, *model fit to the data* is expressed by its **likelihood** = probability of observing the sampled data given the parameters estimated by the model, sometimes referred as the *evidence* of a model, or its *ability to predict/forecast* new data that are similar to the sampled data (see interactive visualization by Kristoffer Magnusson).

Parameter estimation in linear regression models

β_0 , β_1 , and ϵ must be **estimated** based on data sampled from a population:
$\hat{\beta}_0 = b_0$; $\hat{\beta}_1 = b_1$; $\hat{\epsilon} = e$).

There are several methods to estimate unknown parameters, such as:

- **Ordinary least squares (OLS)**: finds the *parameter values* that *minimize the sum of the squared residuals* (default LM estimator)
- **Maximum likelihood estimator (MLE)**: finds the *parameter values* that *maximize the model likelihood*, making the observed data the most probable under that model
- **Bayesian estimator**: finds the *parameter posterior distributions* based on prior knowledge/beliefs (*prior*) and observed data (*likelihood*)

Regardless of the used method, parameters values (or distributions) are always accompanied with a measure of the **uncertainty/precision** associated with their estimate:
Standard errors (SE) = predicted *variability* in the parameter estimate if the data were collected from different random samples from the same population.
SE are used for computing *test statistics* (Est/SE) & *confidence intervals* ($Est \pm 1.96 \times SE$)

In LM, under the assumption of normally distributed residuals, OLS = MLE

What are residuals?

Residuals are the model-based estimates of the population errors.

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:

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

Residuals = observed - predicted

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

```r
head(data.frame(observed = children$height,
                predicted = fitted(m3),
                residuals = residuals(m3)
                squared = residuals(m3)^2 ))
```

```
  observed predicted residuals squared
1   150.77    152.90     -2.13    4.55
2   170.59    156.61     13.98  195.33
3   167.31    160.31      7.00   49.01
4   165.72    165.52      0.20    0.04
5   171.67    160.31     11.36  129.06
6   143.74    151.07     -7.33   53.74
```

```r
sum(residuals(m3)^2) # sum of squared (SS) residuals
```

```
## [1] 128188.3
```

```r
var(residuals(m3)) # residual variance SIGMA2
```

```
## [1] 51.29585
```

In LM, **model parameters** include:

(1) intercept, (2) slope(s), and (3) **residual variance** σ^2

→ *How many parameters in the previous models? (= **No. predictors + 2**)*

Statistical inference on regression coefficients

In the NHST approach, we can **test the statistical significance** of regression coefficients (*two-tail t-test*). This is automatically done by R in the model summary.

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

```
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)   104.25       0.88  118.22 0.000000e+00
age             3.70       0.06   57.45 0.000000e+00
sexmale       -19.04       1.26  -15.14 1.237494e-49
age:sexmale     1.41       0.09   15.39 3.897810e-51
```

- `Estimate` = estimated parameter
- `Std. Error` = parameter standard error
- `t value` = test statistic computed as $t = Estimate/Std.Error$
- `p-value` = p corresponding to the t-value with $No.\ Obs. - No.\ Coeff. - 1$ degrees of freedom

Effect size:

Coefficient of determination $R^2 = 1 - SS_{residuals}/SS_{total}$

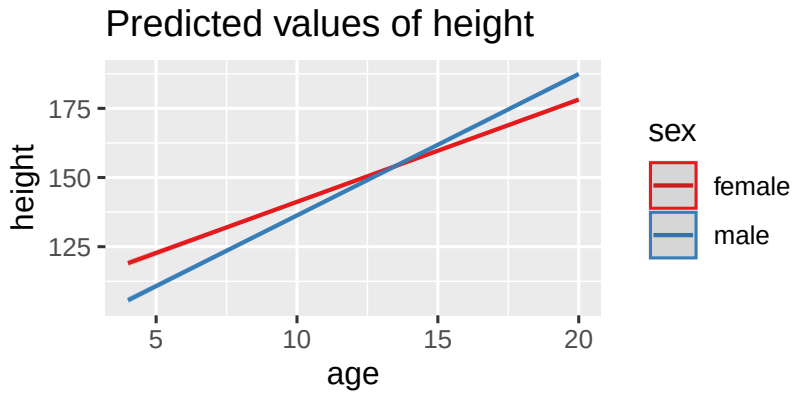
```
summary(m3)$r.squared
```

```
[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 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
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 data
colnames(preg)[c(2,5,12,14)] <- c("age","depr","NICU","threat") # set 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: Linear model 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**: residuals ϵ_i are normally distributed with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
3. **Homoscedasticity**: ϵ_i variance is constant over the levels of x_i (homogeneity of variance)
4. **Independence of predictors & errors**: predictors x_i are unrelated to residuals 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 (no common disturbance factors)

Additional assumptions:

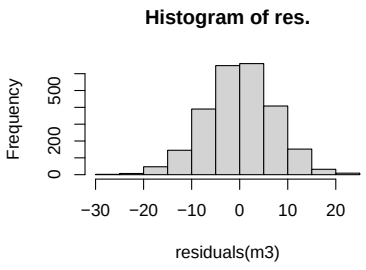
6. **Absence of influential observations** (multivariate outliers)
7. **Absence of multicollinearity (for multiple regression)**:

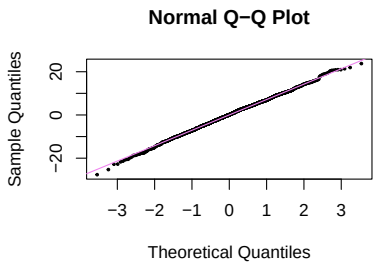
lack of linear relationship between x_1 and x_2

Model diagnostics: Assessing LM assumptions

Normality & linearity ☺

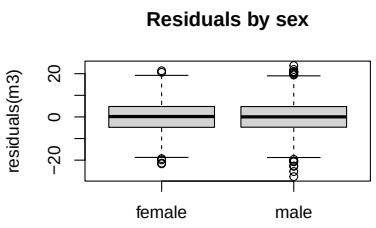
```
hist(residuals(m3))
qqnorm(residuals(m3)); qqline(residuals(m3))
```





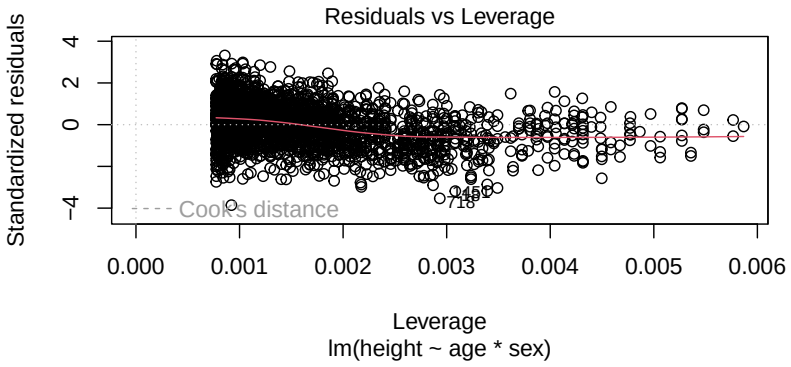
Homoscedasticity & independence x, ϵ ☺

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



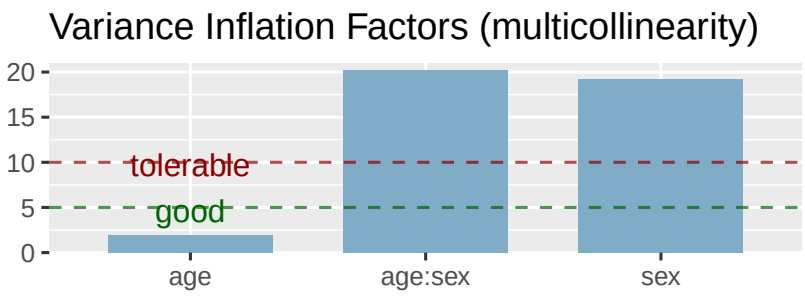
Absence of influential cases ☺

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



Absence of multicollinearity ☹

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



Independence of observations ?

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

That's all for now!

Questions?

Homework (optional):

- read the slides presented today
  and write in the Moodle forum if you have any doubts
- refresh your familiarity with R: `R-intro.pdf`
- **exeRcises 1-3** 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

Achronyms & Greek letters

- AIC = Akaike Information Criterion
- BIC = Bayesian Information Criterion
- LM = linear models
- CI = confidence intervals
- MLE = maximum likelihood estimator
- OLS = ordinary least squares
- NHST = null hypothesis significance testing
- SE = standard error
- SS = sum of squares

- β = *beta*, used to index population-level intercept (β_0) and slope (β_1, β_2, etc.) parameters
- ϵ = *epsilon*, used to index population-level errors to be estimated based on model residuals
- σ = *sigma*, used to index the variance σ^2 of population-level errors (or model residual)
- $\mathcal{N}$ = capital *nu*, used to index that a variable is normally distributed

Achronyms & Greek letters

- AIC = Akaike Information Criterion
- BIC = Bayesian Information Criterion
- LM = linear models
- CI = confidence intervals
- MLE = maximum likelihood estimator
- OLS = ordinary least squares
- NHST = null hypothesis significance testing
- SE = standard error
- SS = sum of squares

- β = *beta*, used to index population-level intercept (β_0) and slope (β_1, β_2, etc.) parameters
- ϵ = *epsilon*, used to index population-level errors to be estimated based on model residuals
- σ = *sigma*, used to index the variance σ^2 of population-level errors (or model residual)
- $\mathcal{N}$ = capital *nu*, used to index that a variable is normally distributed
- ciao