ADVANCED DATA ANALYSIS FOR PSYCHOLOGICAL SCIENCE

Homework exercises

Luca Menghini Ph.D.

luca.menghini@unipd.it

***

Master degree in Developmental and Educational Psychology

University of Padova

2023-2024

Exercise 1: correlation & regression

For each couple of variables (x, y) generated as specified below:

a) represent univariate (boxplot) and bivariate distributions (scatter plot)

b) compute their correlation

c) use the `lm()` function to get the slope coefficient β_1 and determinate whether the relationship significantly differs from zero

1. `y <- rnorm(50)` and `x1 <- y`
2. `x2 <- y + 10`
3. `x3 <- rnorm(50)`
4. `x4 <- x3 + 10`
5. Which conclusions can we draw? Which relationship between correlation and regression coefficient?

Source: Pastore, M. (2021). Analisi dei dati in contesti di comunità. Course held at the University of Padova, Academic Year 2021-2022.

Exercise 2: LM assumptions & diagnostics

Using the "*Pregnancy during pandemics*" data* that we saw in class, graphically evaluate the diagnostics of the selected model `m2`:

1. **Linearity**: are model residuals centered on zero?
2. **Normality**: are model residuals normally distributed?
3. **Homoscedasticity**: is residual variance constant over the levels of any predictor?
4. **Independence error-predictor**: are residuals unrelated to any predictor?
5. **Independence of observations**: based on the considered variables (`depr`, `threat`, `NICU`, and `age`), are individual observations independent?
6. **Absence of influential observations**: is there any observation that strongly influence the estimated coefficients?
7. **Absence of multicollinearity**: are predictors mutually unrelated?

---

*To read the dataset, you can either use the code in `2-multilevel.pdf` slide #10 or download the `pregnancy.RData` file from Moodle/Github ("data" folder) and use the command `load("pregnancy.RData")`

Exercise 3: Towards multilevel modeling

1. Download and read the "*Adolescent insomnia*" dataset `INSA.RData` (Moodle/Github, "data" folder)
2. Explore the variables `dayNr` (day of assessment), `stress` (bedtime rating of daily stress), `insomnia` (categorical: insomnia vs. controls), and `TST` (total sleep time, in minutes) → mean, SD, frequencies, plots, and correlations
3. Fit a null model `m0` predicting `TST`
4. Fit a simple regression model `m1` predicting `TST` by `stress`
5. Fit a multiple regression model `m3` predicting `TST` by `stress` and `insomnia`
6. Compare the three models with the AIC and the likelihood ratio test
7. Print and interpret the coefficients (and their statistical significance) of the selected model
8. Now create two subsets of the `insa` dataset: `insa1` only including observations from the participant `s001` and `insa2` with observations from participant `s002`: how many rows in each dataset?
9. Repeat points 3-7 by using the two subsets: Are results consistent with what you found in the full sample?