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

Some instructions to solve the exercises

The present document includes some optional homework exercises on the contents presented during lectures. Similar to the course slides, exercises will be progressively updated as the course progresses.

To check the **exercise solutions**, just look at the `exeRcises.Rmd` file on either Github or Moodle: each exercise text is followed by a chunk of R code showing point-by-point solutions (or some among the many possible solutions).

If you have any doubts on how to solve the exercises, feel free to write me an e-mail or (even better) try writing in the **Moodle forum**, so that all students can see your question and try to reply it. We can also solve some exercise during lectures, but let me know! ;)

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.

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")`

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?

4. Multilevel data structure

1. Download and read the "*Innovative teaching program*" dataset `studentData.csv` (Moodle/Github, "data" folder)
2. Explore the student-level variables `studId` (identification code of each student), `math_grade` (student grade in math) and `anxiety` (anxiety level). What is the total number of students? How many rows per students? What is the range of `math_grade` and `anxiety`?
3. How many students per class? How many students per level of the `tp` variable?
4. How many classes per level of the `tp` variable? To answer that, you can create the **wide-form dataset** by taking only one row per class (e.g., try using the `duplicated()` function preceded by the `!` symbol to remove duplicated values of `classID`)
5. Compute the mean `math_grade` and `anxiety` value for each class and join them to the wide-form dataset: which is the class with the maximum `math_grade`? Which class has the maximum `anxiety` level?
6. Fit a simple linear regression model predicting `math_grade` by `anxiety` both on the long-form and on the wide-form dataset; inspect and interpret the estimated coefficients and their statistical significance.
7. Which model has the highest standard errors? Why?

5. Data centering

Consider the long- and wide-form datasets from exercise #4:

1. Compute the **grand-mean-centered** `anxiety` values from the wide-form dataset
2. Fit a simple linear model predicting class-level `math_grade` by grand-mean-centered `anxiety` using the wide-form dataset. Inspect and interpret the estimated coefficients, and compare them with those estimated in the previous exercise
3. Use the `join()` function from the `plyr` package to join the cluster-level mean `anxiety` values to the long-form dataset
4. Compute the **cluster-mean-centered** `anxiety` values by subtracting mean class `anxiety` from student-level `anxiety`
5. Considering class `A`, how many students have an `anxiety` level below the class average? How many have a higher value than the average?
6. Fit a simple linear model predicting student-level `math_grade` by cluster-mean-centered `anxiety` values using the long-form dataset. Inspect and intepret the estimated coefficients, and compare them with those estimated in the previous exercise

6. Data centering & level-specific correlations

Do left- and right-side infant pupil sizes correlate more at the within-subject or at the between-subject level?

1. Download and read the "*Infant pupil*" dataset `infantPupil.csv`
2. Subset columns 15, 10, 11, 12, and 13, and rename them as `ID` (subject identification code), `pupil.left` (left-side pupil size in mm), `pupil.left_valid` (validity of left-sized pupil size measurement), `pupil.right` (right-side pupil size in mm), and `pupil.right_valid` (validity of the right-side pupil size measurement)
3. How many valid cases for each eye? (note: 1 = valid, 0 = invalid)
4. Remove all cases with invalid pupil size in either one or the other eye
5. Compute the cluster means and the cluster-mean-centered values for `pupil.left` and `pupil.right`
6. Compute the between-subject and the within-subject correlations between the two variables: Do left- and right-side infant pupil sizes correlate more at the within-subject or at the between-subject level?

7. Intraclass correlation coefficient

Using data from exercise #6, compute the intraclass correlation coefficient (ICC) for both pupil size measures.

1. Do they variate more at the within-subject (lv1) or at the between-subject (lv2) level?
2. What is the percentage of within-subject variability over the total variability?
3. Does one eye variate more within-subject than the other?