# Exercises day 3

Basic Statistics for health researchers 2025

March 3rd, 2025

## Warming up

Before starting the exercise below, learn from the R-demo of Lecture 3 (available from the course webpage):

1. Read and run the code.
2. Check that the output matches the results presented on the slides.
3. Do not hesitate to add your own comments into the script.

## Exercise A (linear regression)

For this exercise we will work with the Sickle Cell Disease (SCD) data, as in Day 1. The data and their description are available from the course webpage.

### Question 1

(Similar to some questions of Exercise Day 1)

1. Read the description of the data and load the data into R. Visualize the first lines of the data and a summary of the data.
2. Create (and add to the data) the new variable `MAP` to define the Mean Arterial Pressure as the sum of the diastolic pressure (`Pdias`) plus one third of the difference between the systolic pressure (`Psys`) and the diastolic pressure (`Pdias`). In short,

$$\texttt{MAP} = \texttt{Pdias} + \frac{1}{3}\big(\texttt{Psys} - \texttt{Pdias}\big).$$

**Question 2**

1. Make a scatter plot to visualize `MAP` versus age.
2. Fit a linear model for the Mean Arterial Pressure as a function of age.
   **Note**: here and in the rest of the exercise, we assume that there is no specific reason to pre-specify a log-transformation the outcome or any other transformation of the outcome (unlike in the lecture of this morning).
3. Add the regression line to the plot in red.

**Question 3**

Interpret the results of the linear regression:

1. How do you interpret the estimate of the slope?
2. Do the data show a significant association between age and MAP?
3. Compute a 95% confidence interval for the slope.
4. Does the intercept have a meaningful interpretation?

**Question 4**

1. Create and add to the data the variable `Zage` defined as

$$\texttt{Zage} = \frac{\texttt{age} - 30}{10} \ .$$

   Note: this almost corresponds to a standardized version of `age` (the only difference being that we have rounded `mean(age)` and `sd(age)` down).
2. Fit a linear model for the Mean Arterial Pressure as a function of this new variable.
3. Can we interpret the new estimate of the slope? How? Does the new estimate of intercept have a meaningful interpretation? Do the p-values of the two estimated slopes match? Is it surprising?

**Question 5**

Make a new scatter plot to visualize `MAP` versus age for subjects with and without SCD separately:

- first plot only the observations of the subjects with no SCD in blue,
- then add those from the subjects with SCD in red,
- finally add a legend.

**Hint:** you can use the function `points` for adding the red dots. You can use the options `xlim` and `ylim` in the `plot` function to make sure that we can see all values (not only those from subjects with no SCD).

**Question 6**

1. Fit two linear models: one for those with SCD, one for the those without.
2. Add the two regression lines to your scatter plot with the matching blue/red color. For which of the two groups the association seems the strongest?

**Question 7**

Produce the QQ-plots and residual plots to check the model assumptions for each of the two models. Do the assumptions seem reasonable?

**Question 8**

Compute a 95% confidence interval for the slope of each model. From these two confidence intervals, can you easily conclude whether there is a significantly different association between age and MAP in the two groups? Why?

**Question 9 (advanced and can be skipped)**

1. Compute the estimate of the difference in slope betwen the two groups.
2. Compute the estimate of the standard error of the difference in slope betwen the two groups. To do so, you can use this formula

$$\text{s.e.}(\hat{\beta}_{\text{SCD}} - \hat{\beta}_{\text{no SCD}}) = \sqrt{\left[\text{s.e.}(\hat{\beta}_{\text{SCD}})\right]^2 + \left[\text{s.e.}(\hat{\beta}_{\text{no SCD}})\right]^2} \, .$$

3. Compute the ratio of the estimated difference divided by its standard error, i.e.

$$z = \frac{\hat{\beta}_{\text{SCD}} - \hat{\beta}_{\text{no SCD}}}{\text{s.e.}(\hat{\beta}_{\text{SCD}} - \hat{\beta}_{\text{no SCD}})} \, .$$

4. Because the number of observations in each group is large ($n = 88$), it is reasonable to approximate the distribution of $z$ by that of a standard normal distribution under the null hypothesis that the two slopes are equal, i.e. $\mathcal{H}_0 : \beta_{\text{SCD}} - \beta_{\text{no SCD}} = 0$. Consequently, can we reject the null hypothesis that the two slopes are equal?

**Question 10**

1. Compute the 95% prediction intervals of the mean MAP given age in the two populations of subjects with and without SCD.
2. Add the the 95% prediction intervals to the plot produced at Questions 5 and 6.
3. Describe the distribution of the age of subjects with and without SCD. Use e.g. median, first and third quartiles.
4. How do the age distributions impact your interpretation of the prediction intervals? i.e. for which ages do you trust the prediction intervals most and least?

# Exercise B (regression to the mean)

1. Read the following Statistics notes from Bland and Altman. "Statistics notes: some examples of regression towards the mean: " Bmj 309.6957 (1994): 780. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2541039/pdf/bmj00458-0034.pdf

2. Does any of the examples of regression to the mean relates to what you might encounter in your own research field? (in your own studies or in those you read about in the literature).

3. If that is the case, present and discuss the case with other students seating next to you.