Lead concentration Project description

In the 1970s several studies focused on the environmental and health risks of lead smelters. The dataset considered here invo lves children who lived near a lead smelter in El Paso, Texas, USA. Children who were exposed to the lead smelter were included in this study, as well as a control group of children. outcomes of interest are:

- finger-wrist tapping score (variable name: *MAXFWT*). It is the number of finger taps in a period of 10 second. It is considered a proxy for the integrity of the neuromuscular system. The higher the score the better. It can be measured for the left and the right hand. We will use the maximum of the two as outcome variable.

- full-scale IQ score (variable name: *Iqf*). The higher the score the better.

One of the research questions was how the MAXWT is related to the lead concentrations in the blood µg/100 ml (*Ld72* and *Ld73* for measurements in 1972 and 1973) and to the total number of years the child lived in the proxy of the lead smelter (*Totyrs*). However, we should perhaps account for other variables in the dataset, such as the age (*Age*) and the gender (*Sex*) of the children. The data can be read as shown in the next chunck of R code: load("lead.RData")

0. Data reading

Read in the dataset and analyze the skim through the data variables.

1. Descriptive statistics

- Summarize your data and calculate the following: mean, median, minimum, maximum, first and third quartile (for each variable).
- For the categorical variable existing, calculate a frequency table
- Calculate the correlation coefficient (MAXWT and *Ld72)* and (MAXWT and Ld73)

2. Graphics

- Generate a bar chart of a categorical variable for the gender (*Sex* parameter).
- Generate a bar chart graph with mean MAXWT inmales and females
- Make a histogram of a continuous variable: "*age*" as well as "MAXWT".
- Make a scatterplot of 2 continuous variables *Ld72* and MAXWT, and add the regression lines for each gender
- Make a boxplot of *age* and a separate boxplots per *Ld72 and per Ld73* (as.factors).

3. Outlier detection

- Explore the data for any existing outliers, identify them (do NOT remove them if found).
- What do you think?

4. Testing for normality/ homoscedasticity (for all features mentioned above)

- Check the normality using two methods
- Check the homoscedasticity using two methods.
- What do you think?

5. Statistical Inference

- Calculate the 90%, 95%, 99% confidence interval for the means of MAXWT per each *gender*.
- How would you describe those inferences and what do you observe in terms of the interval width when request higher confidence (i.e. 99% C.I.)?

6. Hypothesis testing

- We hypothesis that MAXWT is different between male vs female. Assuming normality and homoscedasticity, can you test this hypothesis using statistical hypothesis framework
- Assess whether the previous test assumptions have been meet for the test.
- We hypothesis that MAXWT is "lower" in the group receiving *Ld72 > 40* compared to the control *Ld72 =< 40*. Can you test this hypothesis assuming heteroscedasiticy
- Assess the previous test assumption
- We hypothesis that MAXWT is different between the different Lead types with the different genders (i.e. 4 groups male_leadtype1, male_leadtype2, female_leadtype1, female_leadtype2). Can you perform comparison between the different groups, after assessing the assumptions and performing post-hoc testing (assuming normality and homoscedasticity).

7. Linear model

- Fit a linear regression to the data and interpret the regression coefficient (for the one of the hypotheses mentioned above)
- Calculate and interpret a 95% confidence interval of the regression slope , (bonus)

Estimating the average MAXWT reduction for with increasing the lead concentration (Lead73) to 100 µg /100 ml (bonus)