MASM22/FMSN30/FMSN40 Spring 2023
Lab 1: Lead in moss — change over time

Examination: Mozquizto Lab 1:A+B+C

Perform the tasks by writing and running appropriate R-code while answering the questions in the accompanying three Mozquizto-tests at `quizms.maths.lth.se`. These tests will also provide you will the information marked [*mzq*] below.

1 Problem description

In order to assess the concentration levels of poisonous metals in the environment, Naturvårdsverket (Swedish Environmental Protection Agency) measures the levels of a number of different metals in moss every five years. Since the largest sources of emission have disappeared (been banned) during the later decades, the levels in nature are declining over time.

We have access to the measurements from a handfull of Swedish regions. Which region you should use is determined by your Mozquizto test.

1.A Linear model

Start the Mozquizto-test *Lab 1.A: Linear model* in order to find out which data set you should use. Download the data and save it in the subfolder `Data` in your RStudio project folder. Start the project and create an R-script file, e.g., `lab1.R`, and save it in the subfolder `R` in your RStudio project folder. Write and run the necessary R-code in the script file.

Load the data and look at the summary and the first few lines (replace [*mzq*] as appropriate):

```
load("Data/Pb_[mzq].rda")
summary(Pb_[mzq])
head(Pb_[mzq])
```

The variable `Pb` is the measured concentration level of lead (Pb) in mg/kg dry weight, that is, the amount of mg lead per kg dried moss. We are interested in how fast the lead concentration decreases.

1.A(a). Plot `Pb` against `year`. Does the relationship look linear? Should it be linear?

1.A(b). The measurements started in 1975 and it would be more resonable to use that as a starting point instead of year = 0. Replot the data using `x = I(year - 1975)` instead of `x = year`. The `I()` is necessary in order to use a function, ”-” = ”subtraction”, that R could confuse with ”-” = ”leave this variable out”. This trick can, and should, be used when fitting the model as well.

1.A(c). Ignore the non-linear appearance and fit a linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $Y =$ `Pb` and $x =$ `I(year-1975)`. Calculate the β-estimates, their standard errors and 95 % confidence intervals.

1.A(d). What is the average concentration of lead in 1975, according to this model?

1.A(e). How does the concentration change over time, on average, according to this model?

1.A(f). Use the model to calculate a 95 % confidence interval for the expected average lead concentration in the year [*mzq*]. Does it seem reasonable?

1.A(g). Use the model to calculate a 95 % prediction interval for observations of the lead concentration in the year [*mzq*]. Does it seem reasonable?

1.A(h). Plot the data together with the fitted line, a 95 % confidence interval for the fitted line and a 95 % prediction interval for new observations. Does it look reasonable?

1.A(i). Calculate the residuals and plot them against the predicted values. As visual aides, add a horizontal reference line at 0 and a moving average, `+geom_smooth()`. Any problems here?

1.A(j). Make a Q-Q-plot and a histogram for the residuals. Problems?

1.B Log-transformation

Continue with Mozquizto-test *Lab 1.B: Log-transformation.*

1.B(a). The concentration of lead, when there are no new additions, should decrease at a constant rate, e.g., 10 % per year. What type of model would this indicate and what type of transformations would be natural for Y and/or x?

1.B(b). Plot `y = log(Pb)` against `x = I(year-1975)`. Does this seem like a linear relationship?

1.B(c). Fit this log-transformed linear model and calculate the β-estimates, their standard errors and 95 % confidence intervals.

1.B(d). What is the average log-concentration of lead in 1975, according to this model?

1.B(e). How does the log-concentration change over time, on average, according to this model?

1.B(f). Use the model to calculate a 95 % confidence interval for the expected average log-lead concentration in the year [*mzq*].

1.B(g). Use the model to calculate a 95 % prediction interval for the logarithm of an observation of the lead concentration in the year [*mzq*].

1.B(h). Plot the log-transformed data together with the fitted line, 95 % confidence interval and prediction interval. How does this look compared to 1.A(h)?

1.B(i). Calculate the residuals and plot them against the predicted log-values. Also make a Q-Q-plot and a histogram. How do these look compared to 1.A(i) and 1.A(j)? Comment?

1.C Back to the original scale

Continue with Mozquizto-test *Lab 1.C: Original scale*.

1.C(a). Write down how the (untransformed) concentration of lead, `Pb`, develops over time, as a function of the β-parameters in the model in 1.B, as well as a function of $a = e^{\beta_0}$ and $b = e^{\beta_1}$.

1.C(b). Calculate estimates of a and b together with 95 % confidence intervals.

1.C(c). How high was the average (or rather, median) concentration of lead in 1975, according to the estimated model?

1.C(d). How fast is the rate of decrease in lead concentration, according to the estimated model?

1.C(e). Use the model to calculate a 95 % confidence interval for the expected average (median) lead concentration in the year [*mzq*].

1.C(f). Use the model to calculate a 95 % prediction interval for the an observation of the lead concentration in the year [*mzq*].

1.C(g). Plot `y = Pb` against `x = I(year-1975)` together with the fitted relationship, 95 % confidence interval and prediction interval. How does this look compared to the model in 1.A(h)?

1.C(h). Redo the plot with `x = year` instead. You only need to change the `x=` in the aesthetics, everything else can be the same (except the label on the x-axis).

Note: you can now do Part 1 of Project 1.