MASM22/FMSN30/FMSN40 SPRING 2024

LAB 2: LEAD IN MOSS — DIFFERENT REGIONS

Uses the techniques from Lecture 3 and 4

Deadline: Friday 19 April 2024

Examination: Mozquizto Lab 2: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.

2 Lab 2: Problem description

This is a direct continuation of Lab 1, where we will take the locations, `region`, of the observations into account. The data is available in the file `Pb_all.rda`.

```
#Lab2####
library(ggplot2)
library(dplyr)
load("Data/Pb_all.rda")
summary(Pb_all)
```

2.A Does the lead level change over time?

Continue with Mozquizto-test *Lab 2.A: Change over time*

2.A(a). Plot `Pb` against `year` using all the data, ignoring `region`. Does it look like an exponential decline might work?

2.A(b). Fit a linear model with `y = log(Pb)` and `x = I(year - 1975)` using all the data.

2.A(c). Use a suitable test to test whether there is a significant linear relationship between `log(Pb)` and time, i.e. H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$ on significance level $\alpha = 5\,\%$. Report the test statistic, its degrees of freedom, the P-value, and the conclusion.

2.A(d). Use the model to calculate a 95 % confidence interval and a 95 % prediction interval for the lead concentration (mg/kg) in the year [*mzq*].

2.A(e). Calculate the residuals and plot them against the predicted log values. Also make a Q-Q-plot. Problems?

Note: if `geom_smooth()` gives an error message, use `geom_smooth(method = "loess")` instead.

2.A(f). Redo the residual plots separately for each region by adding `+ facet_wrap(~ region)`. Does this reveal any problems with the residuals?

2.B Add region to the model

Continue with Mozquizto test *Lab 2.B: Regional differences*.

We have measurements from five different Swedish regions and it is likely that they have different concentration levels of lead to start with. This would explain the behaviour of the residuals in 2.A(f). We will fit a model for decrease in (log) lead concentration over time assuming that the rate of deterioration is the same in all regions, but the starting levels may be different.

The `region` variable is already a factor variable. This means that R will automatically create internal dummy variables for each non-reference category. However, we might want to change the reference category.

2.B(a). Plot the data again, as in 2.A(a), but separately for each region by adding
`+ facet_wrap(~ region)`

Does it look like the same type of relationship (exponential decline) for all regions?

2.B(b). Redo the plot with `log(Pb)` on the y-axis. Does this seem like a good idea for all the regions?

2.B(c). Fit the linear model `log(Pb) ~ I(year - 1975) + region`.

Which region was used as reference? Is this a good idea?

2.B(d). The reference category should always be a large category. Change the region variable so that [*mzq*] will be used as reference instead:

```
###2.B(d)####
Pb_all <- mutate(Pb_all, region = relevel(region, "[mzq]"))
```

Then refit the model in 2.B(c).

How high was the average concentration of lead in region [*mzq*] in 1975, according to the model? How fast is the rate of decrease in lead concentration, according to this model? Compare with 1.C(c) and 1.C(d) in Lab 1.

2.B(e). Use the model to calculate a 95 % confidence interval and a 95 % prediction interval for the lead concentration (mg/kg) in region [*mzq*] in the year [*mzq*].

2.B(f). Estimate the ratio between the expected Pb-level in Örebro and the level in [*mzq*], for any given year.

2.B(g). Örebro does not have any observations for 1975. Use the model to estimate the expected Pb-level in Örebro in 1975, together with a 95 % confidence interval.

2.B(h). Use the model to estimate the expected Pb-level in Örebro in the year [*mzq*], together with a 95 % confidence interval.

2.C Tests and residuals

Continue with Mozquizto test *Lab 2.C: Tests and residuals*.

2.C(a). Is there a significant linear relationship between `log(Pb)` and time, now that we have taken the different levels in the regions into account? Test H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$, using a suitable test and report the test-statistic, its degrees of freedom, the P-value and the conclusion.

2.C(b). Are there any significant differences in the starting levels (1975) between the different regions? Test this using a suitable test, comparing the model from 2.A(b) with 2.B(d). Report the test statistic, its degrees of freedom, the P-value, and the conclusion.

2.C(c). Add fitted lines, confidence intervals and prediction intervals to the plot(s) in 2.B(b) and in 2.B(a). Do they seem reasonable?

2.C(d). Calculate the residuals and plot them against the predicted log values. Also make a Q-Q-plot. Problems? Compare with 2.A(e).

2.C(e). Redo the residual plots separately for each region by adding `+ facet_wrap(~ region)`. Did adding the regions to the model solve the problems revealed in 2.A(f)?

In Lab 3 we will look closer at the one or two large residuals and find out if they are a problem.

Note: you can now do Part 2 of Project 1.