MASM22/FMSN30/FMSN40 Spring 2023
Lab 3: Lead in moss — checking and fine tuning

Examination: Mozquizto Lab 3:A+B

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

3 Problem description

This is a direct continuation of Lab 2 where we validate the model and try to improve it.

3.A Regression diagnostics

Continue with Mozquizto test *Lab 3.A: Regression diagnostics.*

3.A(a). Use the data from all regions (`Pb_all.rda`) and fit the (bad) linear model

`Pb ~ I(year - 1975).`

Also reuse the model `log(Pb) ~ I(year - 1975)` from Lab 2.A(b).

Calculate the leverage values, v_{ii}, for each model. Make one plot for each model with leverage against `year` and add horizontal lines at $1/n$ and $2(p+1)/n$ and make sure the y-axis includes $y = 0$.

Note: Since we have many observations each year the points will end up on top of each other. Use `geom_jitter(width = 1)` instead of `geom_point()` to add a random value to the year.

Is there any difference in the leverage between the two models? Why/why not?

Any years where the observations have a high leverage?

Which year would the leverage be at its minimum?

3.A(b). Now reuse the log-transformed model with all the regions using [*mzq*] as reference, from Lab 2.B(d).

Calculate the leverage values, v_{ii} and plot them against `year` using different colours for the different regions, `aes(..., color = region)`. Add horizontal lines at $1/n$ and $2(p+1)/n$ and make sure the y-axis includes $y = 0$.

Any patterns of large leverage? Why are the regions ordered as they are? Hint; Compare the ordering with the number of observations in each region.

3.A(c). For model 2.B(d), calculate the studentized residuals, r_i^*, and plot them against the linear predictor, $\hat{Y}_i$. Add horizontal lines for 0, ± 2 and ± 3 as well as a moving average using `+geom_smooth()`. Are there any trends in the residuals? Are there any unpleasantly large residuals?

3.A(d). Redo the plot separately for each region using `+facet_wrap(~region)`. Are there any trends in the residuals in any of the regions? What might that indicate?

3.A(e). Plot $\sqrt{|r_i^*|}$ against the linear predictor, $\hat{Y}_i$, separately for each region. Make sure the y-axis contains $y = 0$ and add horizontal lines for $\sqrt{\lambda_{0.25}}$, $\sqrt{2}$ and $\sqrt{3}$. Are there any unpleasant trends in the residual variability in any of the regions?

3.A(f). Calculate Cook's Distance and plot them against `year`, separately for each region. Add horizontal lines at $F_{0.5,p+1,n-(p+1)}$ and $4/n$ and make sure the y-axis includes $y = 0$.

Are there any observations with a really large Cook's distance? If not, redo the plot without the $F_{0.5,p+1,n-(p+1)}$-reference line.

Are there any observations that have had an substantially larger influence on the estimates than the rest? Were they also observations with high leverage? Were they also observations with large residuals?

3.A(g). Calculate DFBETAS and save the ones for the time variable. Plot them against time, separately for each location. Did the influential points in 3.A(f) in Örebro have a large influence on the estimate of the rate of decline? Did the one in Västra Götaland?

3.A(h). Identify the influential point in Västra Götaland, redo the plot in 2.B(a) and highlight the point in red. Does it seem logical that it had a large influence on the estimated rate of decline?

3.B Model selection

Continue with Mozquizto test *Lab 3.B: Model selection*.

3.B(a). Fit a model with interaction between time and location,

```
log(Pb) ~ I(year - 1975)*region
```

and use a suitable test to determine whether this gives a significant improvement. Is the rate of decline the same in all locations? Also redo the residual plot in 3.A(d). Has anything happened to the trends?

3.B(b). Calculate R^2, R^2_{adj}, AIC and BIC for all three models, 2.A(b) (only time), 2.B(d) (time and location) and 3.B(a) (interaction). Do the measures agree on which model is best? Worst?

For the model that is best according to AIC, how much of the variability in log-lead concentration does is explain?

Note: you can now do Part 3 of Project 1.