PROJECT 1: LINEAR REGRESSION

MASM22/FMSN30/FMSN40: LINEAR AND LOGISTIC REGRESSION (WITH DATA GATHERING), 2019

Peer assessment: **Monday 15 April (13^{15}–14)**

Final deadline: **Wednesday 17 April, 16.00**

---

1 Instructions

You must write a report, in English. Work in groups of two. Discussion between groups is permitted (and encouraged), as long as your report reflects your own work. Write a clear report, presenting your approach to the assignment, discussing methods and results. Results discussion and interpretation **is important**. Just reporting results is not enough! It should be noted that for some questions there isn't a unique "right" answer and there are a myriad of different issues that you could discuss, so use your imagination. In addition to the text, use as many figures and tables as necessary, with explanatory captions.

The report should be *readable*, not a random disorganized collection of thoughts, plots and tables (see also the Peer Review Guidelines at the end of this document). For example, it should be possible for the reader to understand what you are doing without having access to your code. Also, key information may be better summarized in tables than by including the R printouts (e.g. it may be enough to give regression coefficients and p-values without all the accompanying information provided by R). There is no need to include your R code in the report, but you can include some of the R output.

1.1 Peer review

Bring a **printed version** of your report to the peer assessment (15 April, 13^{15}), or email the report to anna.lindgren@matstat.lu.se at least 1 hr in advance so I can print a copy.

1.2 Final submission

E-mail the final version (a single PDF document) to one of (depending on your course) the following addresses by the deadline **16.00 at Wednesday 17 April**. Also attach to the same message your R-files (or implementation in other language), in a file named `proj1.R` that can be used to run your analyses.

- MASM22/FMSN30 students: email to fmsn30@matstat.lu.se
- FMSN40 students: email to fmsn40@matstat.lu.se

Subject field of the email: write "Project1 by `studid1` and `studid2`" where `studid1` and `studid2` are the id numbers for two students in a given group (forgot your id? Go to the link in the "Form groups" section).

Example: Project1 by d08xhj and d08fjh

2 Atmospheric particles in Olso

The data is a random subsample of 500 observations from a data set that originates in a study where air pollution at a road is related to traffic volume and meteorological variables, collected by the Norwegian Public Roads Administration. The data is hourly measurements at Alnabru in Oslo, Norway, between October 2001 and August 2003. In order to get rid of the strong correlation between successive measurements a random sample of the original, larger, data has been taken. The objective is to model the concentration ($\mu g/m^3$) of atmospheric particles[1] with a diameter between 2.5 and 10 μm, PM_{10}.

The data file `olso.txt` can be downloaded from the course home page. Save it to your R project directory and then read it into R and put it in a data frame called `oslo` with

```
oslo <- read.delim("oslo.txt")
```

pm10	concentration of PM_{10} particles ($\mu g/m^3$)
cars	number of cars
windspeed	wind speed (m/s)
direction	wind direction (0–360 degrees, 0 = North)
temp2m	temperature 2 meters above ground (°C)
tempdiff	temperature difference between 25 and 2 meters above ground (°C)

3 Concentration of PM_{10} and the number of cars

We want to model how the concentration of PM_{10} varies with the number of cars. Plot them against each other and determine, visually, whether a linear relationship might be appropriate.

We will consider three different models:

$$\begin{aligned} \text{linlin}: &\quad PM_{10} = \beta_0 + \beta_1 \cdot \text{cars} + \varepsilon, \\ \text{loglin}: &\quad \ln(PM_{10}) = \beta_0' + \beta_1' \cdot \text{cars} + \varepsilon', \\ \text{loglog}: &\quad \ln(PM_{10}) = \beta_0'' + \beta_1'' \cdot \ln(\text{cars}) + \varepsilon'' \end{aligned}$$

where the errors are assumed to be independent, unbiased, normally distributed with constant variances.

Write down how the, un-transformed, PM_{10} concentration depends on the, un-transformed, number of cars in each of the three models. If we suspect that a relative change in the number of cars would give the same relative change in the PM_{10} concentration, regardless of the number of cars, which model should we use?

Fit the three models and investigate their residuals. Do they fulfil the model assumptions? If not, what seems to be the problem?

Ignoring any problems with the residuals, calculate a 95 % confidence interval for the average (ln) PM_{10}-concentration and a 95 % prediction interval for the observed (ln) PM_{10}-concentration as a function of the (ln) number of cars and plot them together with the data and the estimated relationship (one plot for each model). Also plot the data and intervals on the original, un-transformed, scale. Do the problems with the residuals have any obvious impact on any of the intervals? Explain why/why not.

Use the three models to calculate 95 % prediction intervals for the observed PM_{10}-concentrations when there are 1000, 2000 and 4000 cars. Comment on any interesting differens between the three models.

Compare the two models "loglin" and "loglog" and try to determine whether one can be said to be better than the other. The comparison should also include, but not be limited to, their ability to explain the variability and their sensitivity to problematic observations, e.g., outliers, if any.

[1] For more on atmospheric particles, see `https://en.wikipedia.org/wiki/Particulates` and `https://www.naturvardsverket.se/Stod-i-miljoarbetet/Vagledningar/Luft-och-klimat/Miljokvalitetsnormer-for-utomhusluft/Gransvarden-malvarden-utvarderingstrosklar/`

4 Concentration of PM_{10} and cars, wind and temperature

The amount of pollution also depends on the weather. Start by checking for possible colinearity problems by plotting all the variables against each other. Are there any problematic variable combinations?

The effect of the temperature difference might be non-linear so that negative and positive differences have very different effects on the amount of pollution. A large difference can create a lid keeping the pollution in place. Create an additional variable, `diffcategory`, by dividing the temperature difference into three groups: "negative" = "difference less than -1", "zero" = "difference between -1 and $+1$", "positive" = "difference larger than $+1$". Determine which of the three categories is most suited as reference category.

Using $\ln(PM_{10})$ as dependent variable, fit linear models using the following as explanatory variables:

(a). number of cars (the loglin model),

(b). number of cars and wind speed,

(c). number of cars, wind speed and temperature 2 meters above ground,

(d). number of cars, wind speed, temperature 2 meters above ground and temperature difference between 25 and 2 meters above ground (`tempdiff`),

(e). number of cars, wind speed, temperature 2 meters above ground and categorized temperature difference (`diffcategory`).

For each model, test whether the last added variable(s) is a significant improvement on the previous smaller model (state clearly which model that is). Also report the amount of variation in $\ln(PM_{10})$ that the different models explain, adjusting for the number of covariates. Which model is best, in this sense? Are all parameters in this model necessary, or should we remove some? If so, do that. For the final best model, report the parameter estimates, with 95 % confidence intervals. Do the parameters have the expected signs?

Use the final model to calculate 95 % prediction intervals for the observed concentration of PM_{10} particles during an hour for the six different combinations where the number of cars is 1000, 2000 and 4000, while the temperature difference is either 0 or $-2°C$. Keep wind speed fixed at 2 m/s and the temperature 2 m above ground fixed at 0°C. Compare these intervals to the ones for the loglin model, using only cars as covariate.

Investigate the residuals of the final model and determine whether they follow the model assumptions. Plot Cook's distance against each of the different explanatory variables.

Find the observation that has the highest Cook's distance and redo all of the analyses, noting any interesting differences. Did this observation have a large influence on the final model?

5 Wind direction

Since Alnabru in Oslo is situated at the end of the Oslo fjord and surrounded by mountains[2] one might expect that the wind direction might matter. Since the direction is circular (360 = 0) it might be best to divide the directions into categories using, e.g., 0, 90, 180, 270 and 360 as cutpoints. Try to improve your previous model, taking the wind direction into account.

[2] `https://www.google.se/maps/place/Alnabru,+0668+Oslo,+Norway`

6 Peer review guidelines

The following guidelines for peer-review should be followed.

6.1 Questions regarding the content

		Yes	No
1.	Have all the tasks in the assignment been completed?	☐	☐
2.	Does the report contain relevant figures and tables?	☐	☐
3.	Has all notation been properly introduced and/or explained?	☐	☐
4.	Has the model been properly introduced?	☐	☐
5.	Are the results properly presented and discussed?	☐	☐

6.2 Questions regarding the report presentation

		Yes	No
1.	Does the report have:		
	• Title, authors, and date?	☐	☐
	• Page numbers?	☐	☐
	• Introduction?	☐	☐
	• Results and/or conclusions?	☐	☐
2.	Has the report been proofread? Have language and spelling mistakes been corrected?	☐	☐
3.	Are figures and tables:		
	• Numbered?	☐	☐
	• Equipped with suitable captions?	☐	☐
	• Referred to in the text?	☐	☐
4.	Is the text divided into paragraphs and well structured with clear and suitable section headings?	☐	☐
5.	Is the report easy to read, and understandable without access to the project description?	☐	☐