Compulsory assignment number 1 (of 2) spring 2021

STK4900/9900: Statistical methods and applications

This is the first compulsory assignment (out of two) for 2021. This assignment must be handed in electronically using the Canvas system **no later than 14.30 pm Thursday March 11th 2021**.

You are allowed to collaborate and discuss the problems with other students, but each student has to formulate her or his own answers. You should give the names of the students you collaborate with, so that it is possible to compare the written solutions.

It is not sufficient to present the numerical results and the plots, you should also discuss what you can learn from them. In an appendix you may include computer printouts and other technical material that do not fit nicely into the main part (you should only include the final code, not all trial and errors).

You may use a software package of your choice, but whether you use R or not, you must be able to answer all questions. We recommend that you use R.

The data files are provided on the course web page. If you have problems reading the data, please send an email to osamuels@math.uio.no.

Problem 1

You are asked to study how air pollution at a measuring station at Alnabru in Oslo is related to explanatory variables such as traffic volume and meteorological conditions in the same place. Air pollution is measured by the concentration of NO_2 particles.

The data set consists of 500 observations collected by the Norwegian Public Roads Administration (Vegvesenet). The variables in the data set are

log.no2	The logarithm of the concentration of NO_2
log.cars	The logarithm of the number of cars per hour
temp	Temperature 2 meters above the ground (degrees C)
wind.speed	Wind speed (meters/second)
hour.of.day	Hour of the day the measurements were collected (1-24)

You find the dataset **no2.txt** on the course web page. There is one line for each of the 500 measurements of the 5 variables.

a) Report the main features of the variables **log.no2** and **log.cars** by numerical summaries and plots. Make a scatterplot with **log.cars** on the x-axis and **log.no2** on the y-axis. What do you see?

b) Fit a simple linear model where the log concentration of NO_2 is explained by the amount of traffic, measured by log(number of cars per hour). Give an interpretation

of the estimated coefficients and construct a plot with the observations and the fitted line. Explain what the R^2 measure tells you.

c) Check various residual plots to judge if the model assumptions for the model in b) are reasonable.

d) Then use multiple regression to study the simultaneous effect of the various covariates on the log concentration of NO_2. Use an appropriate measure in order to find the 'best' model for prediction of NO_2 concentration at Alnabru. You need not include interactions, but check if you should transform some of the variables in addition to the number of cars, which is already log-transformed.

e) For the model you have chosen in d), write an interpretation of the model coefficients and check if the model assumptions seem reasonable through various plots. Remember to comment all plots you include in your report.

Problem 2

In the table below, you find measurements of the blood pressure of random samples of 12 men in each of three age groups.

30–45 years	46 – 59 years	60 –75 years
128, 104, 132, 112	120, 136, 174, 166	214, 146, 138, 148
136, 124, 112, 118	138, 124, 160, 157	156, 110, 188, 158
116, 108, 160, 116	108, 110, 154, 122	182, 148, 138, 136

The data are available in the file `blood.txt` on the course web page. Blood pressure is in column 1 and age group in column 2. Age group is coded with values 1, 2, and 3 and you have to remember to specify that they should be considered as categorical (factors).

a) Describe the data using boxplots and numerical summaries. From these, does it seem that blood pressure is varying across age groups?

b) Use one-way ANOVA to answer the question above. Specify assumptions and the hypotheses you are testing. Write a summary of your findings.

c) Formulate this problem using a regression model with age group as categorical predictor variable. Use treatment-contrast and the youngest group as reference. Run the analysis, interpret the results and write a conclusion. Compare with the analysis in b).