PROJECT 2: REGRESSION TYPE PROBLEMS
FMSF90F: STATISTICAL LEARNING AND VISUALIZATION 2023
Hand in: **17.00 on Wednesday February 15**

Intro

In this project you should use some popular methods used for regression type problems, i.e. when the outcome that we try to predict is continuous, and interpret the results.

We encourage students to work in groups of 2, and that as many groups as possible work with data from (one of) the students.

Aim

This project consists of 2 main parts:

1. to try different model selection methods, using cross-validation for fine-tuning and assessment
2. to fit a spline-model and assess the optimal flexibility using cross-validation

Data prerequisites

You can either use data from your own research, or use the datasets with plasma levels of beta-carotene described below. We will also use the greenhouse-gas data from Laboration 1 to try models with regression splines.

- Reasonable size of data: $\geq$ 6 continuous variables and $>$ 100 observations, ideally many more. If there are also categorical variables that is a plus.
- Data must be in reasonably good shape so that manipulating the data is not too time consuming
- Identify at least one continuous variable that could serve as an interesting outcome variable (`y`) of a linear regression, and at least 5 (preferably continuous) variables that are interesting to use as explanatory variables (`x`).

Report contents

Using either your own data or the ”plasma”-dataset supplied on the course webpage, the project consists of producing and presenting the following in a written report:

1. **Multivariable regression.**
   The first part of the project concerns model selection for a continuous outcome variable (y) and several potential predictors (x). Using either the supplied `plasma` dataset (with `log(betaplasma)` as outcome), or your own data, present results from model selection:

(a) Fit prediction models using different methods from Chapter 6.
Try (at least):

- Lasso with Cross-validation employed to select the tuning parameters. (Include relevant graphs for the tuning, and cleary state which λ you suggest to use.)
- Principal components regression (OK to just use the 2 first PCs, but if you want you can include more or select based on cross-validation.)
- Random Forest

Compare what variables were selected with the different methods.

(b) Estimate the prediction error of the final model from each method, using K-fold cross-validation, and compare with "the null model" (=a model with no predictors, but only an intercept), and the full Least-squares fit (= using all x-variables in the model).
For Random Forest use the out-of-bag error that is given in the resulting `randomForest`-object.
Present the resulting cross-validated prediction errors in a graph and/or in a table.

State your thoughts about pros and cons of the different type of models (usual least-squares regression models, LASSO, Principal components regression and Random forest). Which of the models you tested would you prefer for final analysis? Does the choice change if you focus on prediction rather than interpretation or vice versa?

2. **Non-linear effects.**
The second part of the project is about modeling non-linear effects. If you have analysed your own dataset you can do so also in this task[1].
If you have analysed the `plasma` data set that is not suitable here. Instead use the Greenhouse-gas data set that you used in Laboration 1.
Your task is [2] to try using regression splines for fitting a more flexible model for the relation between temperature and accumulated Greenhouse gases (`GHGcum`).

- Try first a natural spline model with, say, 4 degrees of freedom. (use function `ns` in `library(splines)` with parameter `df=4`.)
- Calculate predictions (with prediction bands) for all observations and plot the data together with predictions and prediction bands.
- Also calculate the residuals and plot residuals vs fitted values.
- Compare the 2 above plots with your results from simple linear regression (Laboration 1), and comment which model you would prefer to use.
- Finally use 10-fold crossvalidation to assess the predictive ability for different levels of flexibility (use eg df=1:10). Plot the resulting predictions error vs df, and state your conclusions on which model provides the best fit.

General requirements for the report

Report should have:

[1] = If you have a continuous predictor x for which you have reason to believe that a non-linear relation would improve the fit. Otherwise, please use the Green-house data here!

[2] If you use your own data just exchange y=`temperature` and x=`GHGcum` to your y- and x-variable of interest

- Title page, authors, date, page numbers
- Introduction, results and/or conclusions
- been proof read
- Have language and spelling mistakes corrected
- Figures and tables that are
  - Numbered
  - Equipped with suitable captions
  - Referred to in the text
- text divided into paragraphs and well structured with clear and suitable section headings

The R-code and the data should be submitted as separate files.

Plasma concentrations of beta-carotene

A suitable data set if your group does not have access to relevant data of your own. See Project 1 for details about the data.
Suggestions for the analyses:

- In the linear regressions `log(betaplasma)` is a relevant outcome variable (`y`), and any of the (continuous) other variables are possible explanatory variables (`x`).

End of Project 2