PROJECT 1: DATA HANDLING, VISUALIZATION AND LINEAR REGRESSION

FMSF90F: STATISTICAL LEARNING AND VISUALIZATION 2023

Hand in: **17.00 on Monday January 30**

Intro

Aim

This project consists of 2 main parts:

1. to read, manipulate and visualize data from a real data-set and to fit a linear regression model, and visualize and interpret the results.
2. to perform a principal component analysis (PCA) and interpret the results.

On a **voluntary** basis you can also try to

- use PCA to fill in missing data

We encourage students to work in groups of 2, and that as many groups as possible work with data from (one of) the students. We will try to match students with relevant data from their own research with students that lack such data.

Data prerequisites

You can either use data from your own research, or use the datasets with plasma levels of beta-carotene described below.

If you want to analyze your own research data, send a short description of the data and what variables you plan to focus your attention on, plus the data or a screenshot showing what the data looks like. The prerequisites to fit the methods intended for this project are:

- Reasonable size of data: ≥ 3 continuous variables and > 20 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 2 (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. A short description of the dataset you analyse, including how it was collected

2. Well selected, publication ready graphs (In total 2-4) to illustrate some important features of your data. (Could be scatter-plots, histograms, boxplots, etc.)

3. Relevant results of simple linear regression analyses of `y` on each of the (selected) `x` variables[1], presented eg as a table (with 1 row per analysis = 1 row per x-variable). (Be careful to use a continuous outcome variable (`y`)!).

4. For one simple linear regression[2], make a deeper dive in the analysis by presenting

   - a clearly stated model (with parameter estimates including confidence intervals)
   - graphs that illustrate the findings, including
     - a graph of the data together with the linear fit including both confidence and prediction intervals,
     - an assessment of model fit (i.e. residual analysis).
   - results of an attempt of a polynomial regression (of order 2 or 3), and your conclusions on whether you prefer a linear or polynomial regression

   Be sure to clearly state **your** interpretation about the relation between `y` and `x`, and to highlight any concerns with model assumptions not fulfilled.

5. Fit a multiple regression model of X on y, and interpret the results. **For this model just select 2-5 interesting/relevant x variables.** You do not have to perform model selection, we will do that in Project 2!
   Focus on answering questions 1-3 in Section 3.2.2 in ISLR2[3].

   For the last question on model fit you should create your own plots of residuals vs predicted values (by saving the calculated residuals and predicted values and plot them in a scatterplot) and look for patterns, and outliers. If you are worried about collinearty in the data, investigate it by calculating the Variance inflation factor for each predictor. Remember to clearly state your interpretations also for the investigation of model fit.

6. Perform a PCA of the x-variables in your data material, and present it graphically and with your interpretation of the result in the text. (Use all continuous X-variables you have available in your data.)

7. Finally **if you have time**, emulate the sitation with missing data by making a copy of your data and in this copy you randomly select 10 % of your rows of data, and for each such row select a variable of your dataset randomly and make that value missing (`NA`). Now, use Algorithm 12.1 in ISLR2 to fill in the missing values using PCA. Plot the computed values vs the original data to assess the quality of the fill

[1] preferably 5-10 variables, not more!
[2] I suggest the one with the smallest p-value,and a continuous `x`
[3] Introduction to Statistical Learning (2nd edition)

General requirements for the report

Report should have:

- 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.

Introduction

The purpose of this project is to visualize relations between variables (personal characteristics and dietary variables) in a dataset from a study of plasma concentrations of beta-carotene.

> Observational studies have suggested that low dietary intake or low plasma concentrations of beta-carotene or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients. We designed a cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of beta-carotene and other carotenoids. Study subjects ($N = 315$) were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous. We display the data for only one of the analytes.
>
> We conclude that there is wide variability in plasma concentrations of these micronutrients in humans, and that much of this variability is associated with dietary habits and personal characteristics. A better understanding of the physiological relationship between some personal characteristics and plasma concentrations of these micronutrients will require further study.

Data file

The datafile `plasma.txt` is slightly simplified compared to published data and contains 314 observations on 13 variables and can be downloaded from the course home page. Save it to your R data directory and then read it into R, put it in a data frame called `plasma` and look at it with

```
plasma <- read.delim("Data/plasma.txt")
head(plasma)
summary(plasma)
```

Variable description

age	Age (years)
sex	Sex (1 = Male, 2 = Female).
smokstat	Smoking status (1 = Never, 2 = Former, 3 = Current Smoker)
quetelet	Quetelet (weight/height^2 kg/m^2) a.k.a. BMI
bmicat	BMI category (1 = Underweight, 2 = Normal, 3 = Overweight, 4 = Obese)
vituse	Vitamin use (1 = Yes, fairly often, 2 = Yes, not often, 3 = No)
calories	Number of calories consumed per day.
fat	Grams of fat consumed per day.
fiber	Grams of fiber consumed per day.
alcohol	Number of alcoholic drinks consumed per week.
cholesterol	Cholesterol consumed (mg per day).
betadiet	Dietary beta-carotene consumed (μg per day).
betaplasma	Plasma beta-carotene (ng/ml)

Suggestions for the analyses:

- In the linear regression `log(betaplasma)` is a relevant outcome variable (`y`), and any of the (continuous) other variables are possible explanatory variables (`x`). (You do not have to try them all, but you may if you like.)

End of Project 1