Series 2

1. In this exercise, we analyze a dataset about fruitflies, see [1] and [2]". This dataset contains observations on five groups of male fruitflies – 25 fruitflies in each group – from an experiment designed to test if increased reproduction reduces longevity for male fruitflies. (Such a cost has already been established for females.) The five groups are: males forced to live alone, males assigned to live with one or eight interested females, and males assigned to live with one or eight non-receptive females. The observations on each fly were longevity, thorax length, and the percentage of each day spent sleeping." Note that the fruitflies were assigned randomly to the five groups.

a) Read in the dataset and remove the variables `id` and `sleep`. Then create a pairs plot and comment on it.
R-hints:

```
> url <- "https://ww2.amstat.org/publications/jse/datasets/fruitfly.dat.txt"
> data <- read.table(url)
> data <- data[,c(-1,-6)] # remove id and sleep
> names(data) <- c("partners","type","longevity","thorax")
```

b) Make a scatterplot of longevity versus thorax, using colors for the number of females and different plotting symbols for the different types of females. Comment on the plot.

c) Make three separate plots of longevity versus thorax, one for the flies with 0 females, one for the flies with 1 female and one for the flies with 8 females. Use the same plotting colors and symbols as above. Comment on the plot. Do you see evidence for an interaction between the number of females and type of females in their effect on longevity?

d) Create dummy variables for the different groups. Make a boxplot of thorax for the five different groups. What test can we use to test whether thorax length is significantly different between at least two of the groups? Verify that the test indicates no significant difference. Argue why this was to be expected.

e) Given the result above (thorax is not a so-called confounding variable), and the fact that we are not interested in thorax, we could argue to omit thorax from the model. Test the effect of type of female on longevity for the two groups with 1 female. Conduct this test one time without thorax, and one time including thorax. Comment on the results. How can you explain this?

f) We want to test for interaction between type of female and number of females. First try to model with `as.factor(partners)`, `as.factor(type)` and the product. What goes wrong?

g) We still want to test for interaction between type of female and number of females. The full model is

$$y = \beta_0 + \beta_1 l_t + \gamma_{1,0} p_1 t_0 + \gamma_{1,1} p_1 t_1 + \gamma_{8,0} p_8 t_0 + \gamma_{8,1} p_8 t_1 + \epsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 > 0$, l_t is thorax length and t_0, t_1, p_1, p_8 are dummy variables that are 1 if and only if `type`=0, `type`=1, `partners`=1 respectively `partners`=8. This means that we use "no females" (`type`=9 and `partners`=0) as baseline. Show that $\gamma_{1,0} - \gamma_{1,1} = \gamma_{8,0} - \gamma_{8,1}$ if there is no interaction. Plug this into the full model to obtain the reduced model under the assumption of no interaction. Fit the reduced model and conduct a partial F-test. What do you conclude?
R-Hint: If you have two predictors `q1` and `q2` you can use `I(q1+q2)` in the model formula in R to treat the sum as a predictor on its own.

2. In this exercise we will fit multiple linear regression to a dataset about the life expectancy in different countries, the average number of people per doctor and the average number of people per TV. The data can be loaded as follows.

[1] `https://ww2.amstat.org/publications/jse/v2n1/datasets.hanley.html`
[2] `https://ww2.amstat.org/publications/jse/datasets/fruitfly.txt`

```
> url <- "https://raw.githubusercontent.com/jawj/coffeestats/master/lifeexp.dat"
> data <- read.table(url, sep="\t", header=T, row.names=1)
> data <- data[,c("LifeExp","People.per.TV","People.per.Dr")]
```

a) At first, have a closer look at the data. Plot a histogram for each of the three variables and a pairs plot. Which are the three countries with the highest life expectancy, which are the three countries with the highest number of people per TV and which are three countries with the highest number of people per doctor? **R-hint:** `?order`

b) Exclude the two countries with missing values. Then fit a linear model for `LifeExp` against `log2(People.per.TV)` and `log2(People.per.Dr)`. Interpret the regression coefficients on the original, untransformed scale.
R-hint: `complete.cases(data)`

c) Can we conclude that more TVs imply a higher life expectancy? Why or why not? Can we use the number of people per TV to predict life expectancy?

d) Assume the model assumptions are sufficiently met. Construct a 95% confidence interval for `LifeExp` in a country with 50 people per TV and 3000 people per doctor? What would be a 95% prediction interval for this country?

e) Consider the model diagnostics plots (Tukey-Anscombe, Q-Q, scale-location, Cook's distance, residuals vs. leverage) by `plot(fit, which=1)` to `plot(fit, which=5)`, and comment on them.

f) Exclude the two observations which have the highest Cook's distance, and recompute the confidence and prediction intervals from subtask d).

3. In a study on the contribution of air pollution to mortality, General Motors collected data from 60 US Standard Metropolitan Statistical Areas (SMSAs). The dependent variable is the age adjusted mortality. The data includes variables measuring demographic characteristics of the cities, variables measuring climate characteristics and variables recording the pollution potential.

Mortality	Age Adjusted mortality
JanTemp	Mean January temperature (degrees Fahrenheit)
JulyTemp	Mean July temperature (degrees Fahrenheit)
RelHum	Relative Humidity
Rain	Annual rainfall (inches)
Educ	Median education
Dens	Population density
NonWhite	Percentage of non whites
WhiteCollar	Percentage of white collar workers
Pop	Population
House	Average population per household
Income	Median income
HC	Hydrocarbon pollution potential
NOx	Nitrous Oxide pollution potential
SO2	Sulfur Dioxide pollution potential

Try to find a good linear model that represents the data. Therefore, take the following steps:

1. Inspect the data using scatterplots of the response **Mortality** versus the predictors, and produce diagnostic plots such as, for instance, residual plots. Which (if any) of these plots are useful in order to assess the validity of a model? Why? Should we perform any transformation?

Remark: It is good practice to log-transform data that are very right-skewed even though none of the model assumptions are violated. This and other techniques had been developed by Tukey and are known as "first aid" transformations. The reasons behind these transformations are a little bit elaborated and involve the diagonal entries of the matrix P of the subsection "Geometrical Interpretation" of the script as well as the "Residual vs Leverage" plot produced by the R-function "lm".

2. Perform the transformations suggested in the R-hints. Now, perform a stepwise variable selection. Do forward selection and backward elimination lead to the same model? Which variables have a significant influence on the **Mortality**-level and in which direction? Compare these results to an all-subsets regression using Mallows-C_p.

R-hints:

```
> ## Reading the dataset
> url <- "https://stat.ethz.ch/Teaching/Datasets/mortality.csv"
> mortality <- read.csv(url, header = TRUE)
> mortality <- mortality[,-1]
> ## Create pairs plot using the splom() function of the 'lattice' package
> library(lattice)
> splom( ~mortality, pscales=0, cex=0.5)
> ## Note that the plots of the response Mortality versus the predictors
> ## are shown in the last row.

> ## Transform the data
> mortality_old <- mortality ## It is always a good idea to keep the original data.
> mortality[,"logPop"] <- log(mortality[,"Pop"])
> mortality[,"logHC"]  <- log(mortality[,"HC"])
> mortality[,"logNOx"] <- log(mortality[,"NOx"])
> mortality[,"logSO2"] <- log(mortality[,"SO2"])
> col_num <- which(names(mortality) %in% c("Pop","HC","NOx","SO2"))
> mortality <- mortality[,-col_num]

> ## Fit the full model
> mortal.full <- lm(Mortality ~ . , data=mortality)

> ## Fit the empty model. This is not very useful in itself, but is required
> ## as a starting model for stepwise forward variable selection
> mortal.empty <- lm(Mortality ~ 1, data = mortality)

> ## Backward elimination, starting from the full model
> mortal.bw <- step(mortal.full, direction = "backward")
> ## Forward selection, starting from the empty model
> mortal.fw <- step(mortal.empty, direction = "forward",
                    scope = list(upper=mortal.full, lower=mortal.empty))

> ## Loading the package for all-subsets regression
> library(leaps) #-> function regsubsets()

## All subsets model choice, compare to the stepwise methods
mortal.alls  <- regsubsets(...)

## Load function to produce a nice figure of C_p versus p
## (taken from a "Computational Statistics" course in some previous year)
source("https://stat.ethz.ch/Teaching/maechler/CompStat/cp-plot.R")
p.regsubsets(mortal.alls)
```

Preliminary discussion: Friday, March 11.

Deadline: Friday, March 18.