Series 3

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

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

where $\sigma^2 > 0$, 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 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? Which are the two countries with the highest Cook's distance?

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, leverage, cook's distance) with the commands `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 the previous subtask.

3. In this exercise we study the bias-variance trade-off by adapting the last part of Rcode3.R from the lecture homepage. We will use the notation $\hat{f}_\alpha$ for the function which maps $x \in \mathbb{R}$ to the fitted value of the loess smoother with span parameter α at x. Note that this function implicitly depends on the training data.

a) Create 25 data sets from the model and plot the function `f` and the corresponding loess smoothers for different levels of smoothing. Comment on the variance and bias of the loess estimator at the point `xtest`=2, in relation to the smoothing parameter alpha. Describe the bias-variance trade-off.

b) Recreate the plots from part a) for different values of `sigma` and `n`. Describe how they change.

c) Conduct a small simulation study with 1000 simulations and plot histograms of the loess smoother at `xtest`=2 for the various values of the smoothing parameter alpha. Comment on the histograms and refer to the bias-variance trade-off.

d) Approximate the expected test MSE at `xtest`=2 by averaging the MSE over all simulations. Similarly, estimate the variance and squared bias of $\hat{f}_\alpha(2)$. Visualize the bias-variance trade-off at `xtest`.

4. Let $Y_i = f(x_i) + \varepsilon_i$, $i = 1, \ldots, n$, where the $x_1, \ldots, x_n$ are fixed and $\varepsilon_1, \ldots, \varepsilon_n$ are iid with $E(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Let $\hat{f}(\cdot)$ be an estimator of $f(\cdot)$ based on $(x_1, Y_1), \ldots, (x_n, Y_n)$. Let (x_0, Y_0) be a new data point, where $Y_0 = f(x_0) + \varepsilon$. We saw in class that the expected test MSE can be decomposed as follows:

$$E\left[\left(Y_0 - \hat{f}(x_0)\right)^2\right] = \left[f(x_0) - E(\hat{f}(x_0))\right]^2 + E\left[\left(E(\hat{f}(x_0)) - \hat{f}(x_0)\right)^2\right] + E(\varepsilon^2). \tag{1}$$

Reproduce this derivation, arguing carefuly why the cross-products disappear.

Hint: argue that $\hat{f}(x_0)$ and ε are independent.

Preliminary discussion: Friday, March 15.

Deadline: Friday, March 22.