Multivariate Statistics – Homework 2

Note: Solve the tasks below using **R** following the **Guidelines to Exercises** and upload your solutions containing R code, R output, R graphics and answers to the questions until Tuesday, the 6th of May.

This exercise deals with *linear regression* and *clustering*.

1. (2 points) We take a data set on black cherry trees `trees` from base **R**. Load the data and have a look at the help page. Do pairwise scatterplots of the dependent variable `Volume` versus the two predictors. What relationships do you observe? Is the variance constant?

```
data (trees)
```

2. (2 points) Try log transformations of the variables and visualize. What relationships do you observe?

3. (2 points) Set up a bivariate regression model on the transformed variables. Interpret the result.

4. (4 points) We take a data set `women` from base **R** on average heights and weights of American women. Load the data and have a look at the help page. Visualize the relationship between the dependent variable `weight` and the independent variable `height`. Fit polynomial models up to order 5 using either `poly`, `ns` or `bs`. Visualize the fits. Which order of the polynomial would you choose? Verify using `anova` or `extractAIC`.

```
data(women)
```

5. (2 points) We take a data set on mammal's milk from the **R** package `flexclust` to group the data with **R**. Install the `flexclust` package, load it and type

```
data(milk)
```

   to get access to the `milk` data set on the mammal's milk of 25 animals. Have a look at the help page. Give a summary of the data and visualize the data.

6. (1 point) We start with agglomerative hierarchical clustering of the animals. First, the data needs to be converted from a `data.frame` to a `matrix`. Next, a distance measure has to be chosen and the distance matrix has to be generated. For a first try, we use Euclidean distance. Also check the class of the objects.

7. (2 points) Now you can generate your fist hierarchical clustering solution named `hc1` using function `hclust`. Plot the corresponding dendrogram. What is your interpretation of the cluster structure? What is the default linkage method?

8. (1 point) In order to get a grouping of the data we can use function `cutree`. How do we get 3 clusters? How many clusters do we get when we cut the tree at height 10?

9. (2 points) Now repeat the procedure using two more distance measures and two more linkage methods. Plot the dendrograms as above. What effects do you see?

10. (2 points) Generate a heatmap of the data including dendrograms on top and on the left of the matrix. Either use the `heatmap` function of base **R** or have a look at function `heatmap.2` in package `gplots`.

11. (3 points) We switch to partitioning clustering using function `kmeans`. Generate a cluster solution `km1` of the `milk` data. Try different numbers of clusters and several restarts to avoid local optima. How many clusters would you suggest and why? Visualize your selected cluster solution using function `pairs` by coloring of the dots by `km1$cluster`.

12. (2 points) Use function `kcca` from package `flexclust` to generate a cluster solution `cl1` with 4 clusters. Visualize the results using `barplot(cl1)`. How do the clusters differ? What is your interpretation of the cluster solution?