Multivariate Statistics – Homework 1

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 13th of April.

This exercise deals with *linear regression*. We take a data set on cars from the **R** package `gamair` to build some linear models with **R** ...

1. (2 points) Install the `gamair` package, load it and type

```
data(mpg)
```

   to get access to the `mpg` data set on the fuel efficiency of a number of cars. Convince yourself (also by using the help page of this data set) that the object `mpg` is a data frame consisting of $n = 205$ observations on $p = 26$ mainly technical parameters of these cars (from the mid 1980s). The data set contains both categorical as well as numerical variables. A quick way to get an overview of the class of each column is by typing

```
unlist(lapply(mpg, is.factor))
```

   which will result in a vector with a `TRUE` entry for categorical variables and `FALSE` if it's a numeric variable/column. How can you use the result of the above expression to determine the number of numerical variables in the data set?

2. (2 points) In this exercise we will build (multiple) linear regression models to predict the fuel efficiency from other technical parameters. At first, restrict the data set `mpg` to the variables (columns) `hw.mpg` (fuel consumption on highway in miles per gallon), `wb` (wheel base in inches, in German: *Radstand*), `length`, `width`, `height`, `weight`, `eng.cc` (capacity of the engine, German: *Hubraum*), `bore`, `stroke` (diameter and height of cylinders) and `hp` (horsepower) and eliminate rows with `NA`s. How many observations are remaining in the data set (which we now call `mpg2`)?

3. (1 point) Replace the `hw.mpg` variable with a fuel efficiency variable `lphk` (in liters per 100 kilometer).

```
mpg2$lphk <- 100 / ((mpg2$hw.mpg / 0.621371) * 0.264172)
```

4. (2 points) To start with a simple model, first create a scatterplot of the `lphk` variable on the vertical versus the `weight` variable on the horizontal axis. Do the same for the pair `lphk` and `height`. How would you describe the relationship between `lphk` and the two variables?

5. (3 points) Now use the `lm()` function to build the two simple linear regression models[1] and the model with both predictor variables. Which of the 3 models would you choose and why? Did you expect this result?

6. (2 points) Choose one of the two simple regression models above, create a scatterplot of the involved variables and add the regression line (experts may try different line widths, plotting characters, colors, ...).

[1] The term *simple* in this context just means that the model contains a single x variable contrary to a multiple regression model containing $x_1, x_2, \ldots$.

7. (2 points) Let's assume we have another car with a weight of 2750 lbs. Using the model containing only `weight` as independent variable, which fuel consumption would you predict for it? Further assume, that we also know the height of the car (55 in). Using the model containing both `weight` and `height` as predictors, which fuel consumption would you estimate now?

8. (4 points) Now, create a linear model for the response (dependent) variable `lphk` with all other (i.e. 9) variables in the data frame `mpg2`. Would you prefer this model over the above bivariate model? Which method can be used? What is the interpretation of the regression coefficients?

9. (4 points) Finally use function `regsubsets` (`method = "exhaustive"`) from **R** package `leaps` to find the optimal set of parameters for a linear model for the response (dependent) variable `lphk` starting with all other (i.e. 9) variables in the data frame `mpg2`. Plot the BIC versus the number of predictors. Which model would you choose? Now set up your optimal model. Would you prefer this model over the above bivariate model? What is the interpretation of the regression coefficients?

10. (3 points) Contrary to a simple linear regression model, where we have a single x and a single y variable, a multiple regression model is a little more difficult to visualize. It is often done by plotting the *true* response values (which are the `lphk` values in our case) versus the predicted values (which we can obtain with the `predict()` function) in a scatter plot. Do this for the optimal model from the previous task. Add a 45 degree line to this plot with `abline(a = 0, b = 1)`.