R Worksheet 4: Bivariate Statistics

Please document your code for answering the following questions in an R script and check that your code compiles.

For all of the following exercises, use the dataset mgus2.xlsx
As you already did for Worksheet 1, import the dataset into R and make sure factors and numerical variables are appropriately defined.

Exercise 1: Correlational analysis

a. You hypothesize that the variables "age" and "futime" (time until death or last contact) are correlated. To better understand your data, you first plot the two variables in a scatterplot.
b. Create a plot that shows the histogram and qq-plot for each variable.
c. With the information you have from b., calculate the correlation between the two variables with the appropriate method.
d. You now want to test the correlation for significance. After looking at the scatterplot, you hypothesize that the correlation is less than zero. Write down your hypotheses and use an appropriate function to test the correlation. What is your conclusion?
e. You want to know more about the correlation between Haemoglobin, Creatinine and age. Calculate a correlation table that includes these three variables. Use a function that also calculates the p-value for a correlation test. Which correlations would you explore further?

Exercise 2: Linear regression

a. You want to test the hypothesis that haemoglobin can be predicted by age. Draw a scatterplot with the two variables where age is the predictor. Add a line modelling a linear regression.
b. Calculate the statistical values of the linear regression model for the two variables. How much of the variance in haemoglobin can be explained by age?
c. To split your linear regression model for the categorical variable fu_cat you previously created (WS1 /2d.), use the scatterplot() function and interpret the result.

Exercise 3: Multiple linear regression

a. From your previous analyses, you conclude that age and creatinine are closely connected. Use a multiple linear regression model to determine how much additional variance creatinine explains if your first model is the linear model predicting haemoglobin with age.