Questions 1 and 2 consider the dataset LifeExpectancy.csv.

1. a) Calculate the sample correlation between variables 'Alcohol' and 'LifeExpectancy'. How do you interpret the correlation. What does the correlation mean? (3p)

   b) Assume that life expectancy is estimated by using linear regression model with independent variables 'Schooling' and 'AdultMortality'. Explain what kind of variable you would use to test the significance of that model and how you would calculate the p-value for the test. (3p)

2. a) Create a linear regression model which estimates the life expectancy y by using independent variables $x_1,...,x_{18}$. You can find the data descriptions of the variables in their own files. (1p)

   b) Use backwards elimination to eliminate variables from the model so that every independent variable is significant on significance level $\alpha = 0{,}01$. (2p)

   c) Ireland can be found in the data on row 69. Calculate the estimate for the life expectancy in Ireland. Also calculate the 95 percent prediction interval for the estimate. (3p)

Questions 3 and 4 consider the dataset HeartAttack.csv.

3. a) Use two way analysis of variance to see if the number of cigarettes consumed daily is dependent on the gender inside the group of smokers (in other words: see if the means of the 'cigsPerDay' differ in the groups formed by the variables 'male' and 'currentSmoker'). (3p)

   b) Based purely on the data what would you answer to the question "Do men between the ages 32 to 70 years smoke more than women in the same age range?". Justify your answer.

   c) Why is it wrong to make generalization of the question in part b on population level? (1p)

4. a) Fit a logistic regression model to dataset HeartAttack so you get to calculate the probability for a patient to have a high risk of CHD (coronary heart disease). (1p)

   b) Use backwards elimination to eliminate independent variables from your model so that every independent variable is significant on the significance level $\alpha = 0{,}05$. (1p)

   c) One of the independent variables left in the model is 'male'. Consider male and female patients who have the same values in every other independent variable. How many times greater is the odds ratio

$$\frac{P(Y = 1)}{P(Y = 0)}$$

   of the male patient compared to female patient? (2p).

d) Let’s set 0,5 as a limit for patient to have higher risk for CHD. If the probablity estimated by a model in part b is lower than 0,5 patients risk for CHD is not considered to be high. Otherwise the risk is considered to be high. Examine the goodness of your model in following way:

   1. Create a variable/vector *p*, in which you save the probabilities for the risk of CHD for every patient in the data.
   2. By using variable *p*, create a variable *yhat*, which gets a value 0, if patients risk for CHD is not high and value 1 if the risk is high. You might want to use for-loop and if-else structure here.
   3. Compare the values of the variables *yhat* and *TenY earCHD*. How many percent of the estimates created in steps 1 and 2 correspond the risk in original data? (2p).