Stochastic Methods for Material Science

Programming Project – Summer term 2023

(Deadline for submission: September 3rd 2023 at 11:59 pm)

Notice: You have to submit an R script including comments and the commands for solving the subsequent tasks. The evaluation and interpretation of the results and data analysis in R can be done via comments within the R script.

1. Task

In an investigation of the Cd contamination of trout in a river, ten trout were caught at each of two locations and their Cd content (in mg/g fresh weight) was determined. The measured values were as follows:

Location A	76.8	72.3	74.0	73.2	46.1	76.5	61.9	62.4	65.9	62.4
Location B	64.4	60.0	59.4	61.2	52.0	58.1	62.0	57.8	57.2	

a) Draw parallel box plots. What do you observe?

b) Test for the level $\alpha = 0.05$ whether the Cd contents measured at **both** locations can be regarded as realizations of a normally distributed random variable.

c) Test for the level $\alpha = 0.05$ whether the variances of the Cd contents are equal or significantly different from each other.

d) Test for the level $\alpha = 0.05$ whether the expected Cd content at location A is significantly greater than that at location B.

2. Task

The measured temperatures (in ^oC) before and 3 hours after taking a drug can be found for ten different patients in the following table:

before	38.4	39.6	39.4	40.1	39.2	38.5	39.3	39.1	38.4	39.5
after	37.6	37.9	39.1	39.4	38.6	38.9	38.7	38.7	38.9	38.7

a) Describe which sample situation (one, two, paired, multiple, ...) we have here.

b) Generate a scatter plot of the data. What do you observe?

c) Estimate the correlation between the temperature values before and after taking the drug.

d) Test for level $\alpha = 0.05$ if the correlation is significant (you can decide in one- or two-sided alternative). To this end, research (in literature or online) which test is suitable for this task and verify that its assumptions are satisfied for the data at hand.

3. Task

In the text file *coal_data.txt* you find the time intervals in days between disasters in British coal mines between 1850 and 1965.

a) Load the data into R and visualize the empirical distribution via a histogram. Also draw a kernel density estimte for the underlying distribution. What do you observe regarding the (shape of the) distribution?

b) Compute a confidence interval for the mean value of days between two disasters for confidence level of 95%.

4. Task

The *cherry.csv* dataset contains measurements of circumference (*Girth*), height (*Height*), and volume (*Volume*) of $n = 30$ of all black cherry trees. Circumference and height are given in feet and volume in cubic feet. The circumference was determined at a height of 4.5 feet above the ground.

a) We want to investigate whether the wood volume of black cherry trees behaves like the volume of a cylinder. Remember the volume formula for cylinders with height h and circumference u:

$$V = \frac{h \cdot u^2}{4\pi}.$$

Fit the following linear model to the data in *cherry.csv* accordingly:

$$volume = \theta_0 + \theta_1 Height \cdot Girth^2$$

(*Hint:* Use the function `I(...)` for the description of the model in the call `lm`).

Answer the following questions in the process:

- Are there any anomalies in the residual analysis?
- Does the model fit the data (goodness-of-fit, R^2)?
- Is θ_0 significantly different from zero?
- Is θ_1 significantly different from $\frac{1}{4\pi}$? If so, is the wood volume of black cherry trees significantly larger or smaller than that of a cylinder of the same height and circumference?

b) We want to fit a quadratic function as an alternative to a):

$$volume = \theta_0 + \theta_1 Height + \theta_2 Girth + \theta_3 Height \cdot Girth + \theta_4 Height^2 + \theta_5 Girth^2.$$

Proceed as follows:

- First fit only a linear function

  $$volume = \theta_0 + \theta_1 Height + \theta_2 Girth$$

  Again, check the residuals as well as the goodness of fit of the model.
- Now, step by step, add the three remaining basis functions to the quadratic function.
- Always check the residuals of the larger model again for anomalies, the goodness of fit of the model and whether the goodness of fit of the model has indeed improved significantly by adding the new function.
- If the larger model is not significantly better, then stay with the smaller model and try the next of the three estimators (this is a *forward model selection*).

Which submodel of the quadratic function above is the final one you choose?

c) Now choose either the adjusted model from a) or the final model from b). Give reasons for your decision.

5. Task

The degradation of a (very high) concentration of pollutants was measured each 30 minutes over the course of 10 hours. The corresponding data can be found in the file *saet.csv*. To model the time decay of the concentration, a Rodbard function

$$c(t) = a + \frac{b}{1 + \left(\frac{t}{c}\right)^d}$$

can be fitted to the data. Here a describes the final concentration for $t \to \infty$, $a + b$ describes the initial concentration at $t = 0$, and c and d describe the rate of degradation.

a) Fit the above model to the data. Plot the data (black circles) and the fitted Rodbard function (red line) in a graph.

b) We are now interested in the time $t_{0.5}$ at which the pollutant concentration has fallen exactly to the mean of the initial and final concentrations. In particular, we want to estimate an upper bound for this time. Give a corresponding upper bound "for $t_{0.5}$ at 95% confidence.

(*Hint:* For b) consider how $t_{0.5}$ can be calculated from the four parameters).