Statistics for Data Science Exercise 11
Wintersemester 2023/24
Please complete these problems before the exercise session on Tuesday 16 January, 2024, 8:30. Please be prepared to present your solutions to any problems that you completed successfully.

1. A factory produces warm, woollen slippers. The number of weekly slipper orders y is believed to be linearly dependent on last week's median temperature x (in Celsius degrees). Consider the following sample of the variables:

x	-10	-17	-4	-7	-5	-6	-11
y	105	163	43	69	48	56	115

   (a) Assume that the residuals of the model follow a normal distribution. Calculate the 90% confidence intervals for the intercept and the slope.
   (b) Assume that this week's median temperature is -15 degrees Celsius. Predict the number of next week's orders.
   (c) What can you say about next week's orders if this week's median temperature is $+30$ degrees Celsius?

2. Ice cream consumption (liters/day) and blood sausage consumption (kg/day) are thought to be linearly dependent on the maximum temperature of the day (in Celsius degrees) and on the number of hours the local radio station plays accordion music (hours/day). Assume that the maximum temperature and the number of hours of accordion music played on the radio are independent. Consider the following sample of the variables:

maximum temperature	hours of accordion music	ice cream consumption	blood sausage consumption
15	4	272	74
24	7	371	123
17	3	286	54
29	8	432	140
18	6	311	111
16	11	314	216
30	2	415	19
25	9	390	165

   (a) Formulate the corresponding linear model.
   (b) Calculate the (generalized) least squares estimates of the parameters of the model.
   (c) Estimate the consumption of ice cream and the consumption of blood sausages when the temperature is 22 degrees Celsius and 5 hours of accordion music is played on the radio.

The exercises continue on the next page!

3. It is not known what proportion p of purchases of a certain brand of breakfast cereal are made by women and what proportion are made by men. In a random sample of 70 purchases of this cereal, it was found that 58 were made by women and 12 were made by men. Find the maximum likelihood (ML) estimator of p.

4. Consider the problem

$$y_j = F(x) + \eta_j, \quad j \in \{1, \dots, n\}, \tag{1}$$

where the unknown quantity $x \in \mathbb{R}^d$ is assumed to remain static, $F\colon \mathbb{R}^d \to \mathbb{R}^k$ is a function, we assume Gaussian observational noise $\eta_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \gamma^2 I)$, $\gamma > 0$, and we assume to have n independent and identically distributed observations $y_1, \dots, y_n \in \mathbb{R}^k$.

   (a) What is the likelihood function of parameter x?

   (b) Show that

$$\sum_{j=1}^{n} \|y_j - F(x)\|^2 = n\|\overline{y} - F(x)\|^2 + n\left(\frac{1}{n}\sum_{j=1}^{n} y_j^{\mathrm{T}} y_j - \overline{y}^{\mathrm{T}}\overline{y}\right),$$

   where $\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$.

   (c) Use parts (a) and (b) to deduce that the problem (1) has the same maximum likelihood (ML) estimator as the problem

$$\overline{y} = F(x) + \eta, \quad \eta \sim \mathcal{N}\left(0, \frac{\gamma^2}{n} I\right).$$

   Interpretation: averaging a number of independent measurements of a static target results in variance reduction of the noise.