Statistics, Simulation & Optimization: Assignment 2

In Exercises 2.1 and 2.2 related to the Statistics part: tests should be performed using level $\alpha = 0.05$*, unless stated otherwise; check the model assumptions whenever appropriate.*

Exercise 2.1 Scholastic Aptitude Test

The data in the file `sat.txt` concerns data on the Scholastic Aptitude Test (SAT) for pupils in the United States in 1994/1995. The column `expend` contains the mean expenses per pupil (in $ per pupil), `ratio` is the pupil/teacher ratio, `salary` is the mean salary of teachers, `takers` is the percentage of pupils that takes the SAT. Variables `verbal` and `math` are partial scores of the total SAT score in `total` and not used in the analysis. All data are per state.

a) (1.0) Fit a multiple linear regression model to these data, taking the column `total` as response variable and `expend`, `ratio`, `salary`, `takers` as explanatory variables. Use both the step-up and the step-down methods to find the best model. In case step-up and step-down yield two different models, choose the best one and motivate your choice.

b) (0.5) Investigate whether the square of `takers` is useful as additional explanatory variable. Expand the data set in the following way: `sat$takers2=(sat$takers)^2`. Next, find again the best model, using step-up and step-down, now using `expend`, `ratio`, `salary`, `takers` and `takers2` as possible explanatory variables.

c) (0.5) Which model do you prefer, the model you found in a) or in b)? Motivate your answer. Comment on the estimated parameters in your final model.

d) (0.5) Determine 95% confidence and prediction intervals for the total SAT score using the model you preferred in c) for a state with `expend`=5, `ratio`=20, `salary`=36.000, `takers`=25.

Exercise 2.2 Trees

The 'Amsterdamsche Bos' Forestry wishes to estimate the total wood volume of the trees on its domain. To this end the Forestry has cut a sample of 59 trees of their most prevalent type 'Beech' and 'Oak'. The volume of these trees alongside with their height and trunk diameter have been measured. The latter two characteristics can be measured in the field without sacrificing the tree. The Forestry hypothesizes that these are predictive of the tree's volume. The data (in the file `treeVolume.txt`) from the cut trees is to uncover this relationship, taking into account the tree type.

a) (0.7) Investigate whether tree type influences volume by performing ANOVA, without taking the diameter or height into account. Is an oak more voluminous than a beech? What are the estimated volumes for the two tree types?

b) (1.1) Investigate whether the tree type influences the volume, now including diameter and height (but no interactions) as predictors into the analysis, which results in an ANCOVA analysis. How does tree type influence volume? What are the estimated volumes for the two tree types with average diameter and height? (Hint: use function `predict`.) How does diameter influence volume? Graphically (or propose and perform a relevant test) investigate whether this dependence is similar for both tree types.

c) (0.7) Propose a transformation of (a subset of) the explanatory variables that possibly yields a better model (verify this). (Hint: think of a natural link between the response and explanatory variables.)

Exercise 2.3 Optimal Product Mix

Consider the prices and nutrients of the following foods.

Foods	Price ($) per serving	Calories per serving	Fat (g) per serving	Protein (g) per serving	Carbohydrate (g) per serving
Raw carrots	0.14	23	0.1	0.6	6
Baked potatoes	0.12	171	0.2	3.7	30
Wheat bread	0.2	65	0	2.2	13
Cheddar cheese	0.75	112	9.3	7	0
Peanut butter	0.15	188	16	7.7	2

You are interested in finding the cheapest diet that satisfies the following (you can take fractions of servings):

- Calories must be at least 2000,
- Fat must be at least 50g,
- Protein must be at least 100g,
- Carbohydrates must be at least 250g.

a. (1.25) Solve the problem using Excel.

b. (0.5) Only the first 5 units of peanut butter cost \$0.15, above 5 units the price is \$0.25 per unit. For example, 6 units cost $5 \times \$0.15 + 1 \times \$0.25 = \$1.0$. Model this in a linear way and solve it in Excel. Does it change the answer?

c. (0.25) Solve the problem from a. when fractions of servings are not allowed anymore (using Excel). Explain why this cheapest diet is less/more expensive than that of a.

Exercise 2.4 Transportation Problem (will be discussed in Lecture 9)

Consider a transportation problem in which goods have to be transported from 3 sources labeled as S_1, S_2 and S_3 to 4 destinations labeled as D_1, D_2, D_3 and D_4. The costs in euros per unit transported from a source to a destination are given in the following table. It also contains the demand at the destinations and the available supply at the sources:

Source\Destination	D_1	D_2	D_3	D_4	**Supply at source:**
S_1	10	0	20	11	20
S_2	12	7	9	20	25
S_3	0	14	16	18	15
Demand at destination:	10	15	15	20	

For example, it costs 9 euro per unit transported from S_2 to D_3, S_2 has 25 units of supply available and D_3 requires 15 units.

a. (0.75) Find the cheapest transportation plan using Excel that fulfills all demand while respecting the available supplies using linear optimization.

b. (0.75) Now each source-destination edge that is being used to transport goods costs an additional fixed 100 euro. For example, transporting 10 goods from S_3 to D_2 costs $100 + 10 \cdot 14 = 240$ euro. Formulate this as an (integer) linear optimization problem and solve it using Excel. Shortly reflect on the difference with the transport solution from a.

Exercise 2.5 Call Center Staffing (will be discussed in Lecture 9)
Based on predictive analytics and some service level requirement, the required staffing in a call center from 09:00 to 21:00 is given Table 1. There are 2 types of shifts available:

(i) *8-hour working shifts* with 30 minutes unpaid break after 4 working hours (note: this shift takes 8.5 hours in total). These shifts can start every half an hour from 09:00 to 12:30. The wage is 20 euro/hour.

(ii) *4-hour working shifts* without a break that can start every half an hour from 09:00 to 17:00. The wage is 24 euro/hour.

A staffing schedule indicates the number of each shift type and the starting time for each shift.

a. (0.75) Formulate this as a covering problem and use Excel to find the staffing schedule that minimizes costs and satisfies the required staffing.

The required staffing from a is now "demanded", meaning that it not necessarily needs to be fulfilled anymore. Furthermore, only the 8-hour working shifts are available.

b. (0.75) Using Excel, find the schedule that minimizes the sum of absolute differences between the demanded and scheduled number of workers per time interval by solving an integer linear optimization model.

Time interval	**Required number of workers**
09:00 - 09:30	10
09:30 - 10:00	11
10:00 - 10:30	13
10:30 - 11:00	16
11:00 - 11:30	16
11:30 - 12:00	13
12:00 - 12:30	11
12:30 - 13:00	10
13:00 - 13:30	10
13:30 - 14:00	11
14:00 - 14:30	12
14:30 - 15:00	13
15:00 - 15:30	14
15:30 - 16:00	14
16:00 - 16:30	13
16:30 - 17:00	11
17:00 - 17:30	10
17:30 - 18:00	9
18:00 - 18:30	9
18:30 - 19:00	10
19:00 - 19:30	9
19:30 - 20:00	8
20:00 - 20:30	8
20:30 - 21:00	8

Table 1: Data for Exercise 2.5