STATISTICS FOR EES — EXERCISE SHEET 9

1. The following table shows how many semesters several students studied. All of them studied Mathematics or Computer Science in the same cohort at the Yule–Simpson University. The table also shows the salary (in k €) of the graduated in their first year in the job.

Semester	12	14	16	12	15	14	13	14	11	13	10	12	14	13	14	15
Salary	39.4	38.2	37.4	39.5	32.8	35.3	39.1	35.2	37.9	35.7	41	40.9	34.2	38.4	36.2	38.4
Semester	9	11	9	9	12	13	11	10	10	10	9	10	12	10		
Salary	33.7	35.9	36.1	34.2	29.9	31.9	33.3	36.2	33.8	32.9	33.3	35.1	34.2	35.3		

a) Do students who studied longer tend to earn more money? Determine the regression line to predict the starting salary from the number of semesters studied.

b) The top two lines of the table refer to Computer Scientists and the two bottom lines refer to Mathematicians. Do you get different results if you analyse the data of these two groups separately?

c) Visualize the data in a way that supports your conclusions.

2. Sixty crop plants have been grown with thee different fertilizers (A,B,C) and different amounts of water (in ml per day). The yield from each plant has been measured in gram.

```
> str(plantyield)
'data.frame': 60 obs. of  3 variables:
 $ yield     : num  147 151 148 149 159 ...
 $ water     : num  51 52 53 54 55 56 57 58 59 60 ...
 $ fertilizer: Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
```

(a) The data has been analyzed with a linear model:

```
> model1 <- lm(yield~water+fertilizer,plantyield)
> summary(model1)

Call:
lm(formula = yield ~ water + fertilizer, data = plantyield)

Residuals:
     Min       1Q   Median       3Q      Max
-22.4072  -5.2699   0.4253   5.9653  15.9066

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.8502    12.0978  -3.542 0.000809 ***
water         3.5518     0.1973  18.003  < 2e-16 ***
fertilizerB  66.9414     2.7867  24.022  < 2e-16 ***
fertilizerC   5.2006     2.7867   1.866 0.067246 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.812 on 56 degrees of freedom
Multiple R-squared:  0.9488,Adjusted R-squared:  0.9461
F-statistic: 346.1 on 3 and 56 DF,  p-value: < 2.2e-16
```

What yield does this model predict for a plant that obtained fertilizer A and 60 ml of water per day?

(b) What do the p-values in this table tell you? (using a 5 % significance level; one and only one answer is correct)

i. Fertilizer B has a significant effect and the effect of fertilizer C is not significant (both compared to using no fertilizer, only water).

ii. The effect of fertilizer B is significantly different than that of A, and the effect of fertilizer C is not significantly different than the effect of fertilizer B.

iii. The effect of fertilizer B is significantly different than that of A, and the effect of fertilizer C is not significantly different than the effect of fertilizer A.

iv. The effect of fertilizer B is significantly different than the average effects of fertilizers A and C, and and the effect of fertilizer C is not significantly differentthan the effect of fertilizers A and B.

v. The effect of of fertilizer A is not significant, the effect of fertilizer B is significant, and for the effect of fertilizer C we observe a trend.

(c) The data has been analyzed with another linear model:

```
> model2 <- lm(yield~water*fertilizer,plantyield)
> summary(model2)

Call:
lm(formula = yield ~ water * fertilizer, data = plantyield)

Residuals:
    Min      1Q  Median      3Q     Max
-18.419  -5.763   1.057   4.846  13.157

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -13.94804   19.09222  -0.731  0.46820
water               3.07411    0.31415   9.785 1.47e-13 ***
fertilizerB       -16.26299   27.00047  -0.602  0.54948
fertilizerC         1.69853   27.00047   0.063  0.95007
water:fertilizerB   1.37528    0.44428   3.096  0.00311 **
water:fertilizerC   0.05789    0.44428   0.130  0.89682
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.101 on 54 degrees of freedom
Multiple R-squared:  0.9583,Adjusted R-squared:  0.9544
F-statistic: 248.2 on 5 and 54 DF,  p-value: < 2.2e-16
```

Predict, based on this model, the yield of a plant that obtained fertilizer B and 60 ml of water per day (with three decimal places).

(d) What conclusion can you draw from the following R output? (One and only one answer is correct.)

```
> drop1(model2,test="F")
Single term deletions

Model:
yield ~ water * fertilizer
                Df Sum of Sq    RSS    AIC F value   Pr(>F)
<none>                       3544.0 256.72
```

```
water:fertilizer  2    804.71 4348.7 265.00  6.1307 0.003986 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i. Model 2 explains the data significantly better than model 1.

ii. The ratio of the amount of water and the amount of fertilizer has a significant effect on the yield.

iii. The yield is modeled as the product of the amount of water and the amout of fertilizer.

iv. The AIC values indicate overfitting issues for model 2.

v. Model 2 explains the data significantly better than assuming that neither water nor fertilizer has an effect, but we cannot conclude that model 2 fits better than model 1.

3. A certain gene in a certain bacteria species is known to be highly expressed if in the environment of the bacterium the potassium (K) concentration is higher than the sodium (Na) concentration. In 40 bacteria cultures in Petri dishes with different ratios of K and Na concentrations you measure expression values for this gene. Your measure for gene expression is proportional to the number of mRNA molecules per bacterium. The file `RNASeq_K_Na.csv` contains the gene expression values and the ratios of K- and Na-concentrations. Fit a linear model that explains how the gene expression value depends on the ratio of concentrations. Use visualizations to check the assumptions of linear regression and apply transformtions to the data as necessary to fulfill these assumptions.

4. Habanero Chili plants of five different varieties were watered with different amounts of water (in ml) per day from two weeks after germination on. The data in file `chili.csv` shows the height of the chili plants (in cm) after four months.

   (a) Use R to fit a linear model that explains the height as a linear function of the amount of water, with a intercept that depends on the variety. What height (after 4 months) does this model predict for a plant of variety B that obtained 100 ml of water per day?

   (b) Now fit a model in which the dependence of height on watering can differ among the varieties. Give for this model a mathematical formula that predicts the height of the chili plant after four months, given the amount of water w per day and the variety V of the plant. Further, specify the precise assumptions underlying the linear model.

   (c) Inspect the residuals of the model from (b). Do they look normally distributed? Do they still seem to depend on the amount of water, on the variety or on the predicted height in some way? Based on your observation, propose a better model and check it too by visually inspecting the residuals. Again, give a mathematical formula that predicts the height of the chili plant after four months, given the amount of water w per day and the variety V of the plant.

   (d) It turned out that the conditions were not identical for the different plants. Sensors that had been placed near the plants have measured the average number s of sunshine hours per day and the average temperature T in °C during the four months. These data are given in file `chilis.csv` in colums `light` and `temp`. Use some model fitting strategy to find a linear model (which may of course include non-linear scaling and interactions among the variables) to predicts the expected height of the plant as a function of w, V, s and T. What height does this model predict for a plant of variety C that obtains 100 ml water and 8 hours of sun each day and grows in an average temperature of 25 °C?