SKETCH of the SOLUTIONS STK4900/9900 - 2022

Problem 1

a

A box plot is a graphical tool useful to describe some important characteristics of the data distribution. It consists of a box that ranges from the first to the third quartile, displaying the interquartile range. In the middle of the box, a line represents the value of the median. Starting from the box, two whiskers display the distance from the first interquartile and the minimum and the distance between the third quartile and the maximum. If the minimum and the maximum are very far from the quartiles, the extreme points are denoted by circles and the whiskers stop before the extremes.

One assumption of the ANOVA is the Gaussian distribution of the observations within the groups. It is clear that, in this case, the distribution is not symmetric due to the high values of some observations and it is useful to apply a logarithmic transformation. The box-plots on the right are, indeed, compatible with a Gaussian distribution.

b

The null hypothesis and the alternative hypothesis for the test mentioned at point (a) are:

H_0 $\mu_{\text{spring}} = \mu_{\text{summer}} = \mu_{\text{autumn}} = \mu_{\text{winter}}$;

H_A at least one mean is different from the other;

where μ_{spring}, μ_{summer}, μ_{autumn} and μ_{winter} are the mean area burned by the fire in spring, summer, autumn, and winter, respectively.

Since we have $K = 4$ seasons, $n = 270$ observations, the model sum of squares (MSS) equal to 8.7 and the residual sum of squares (RSS) equal to 618.3, then

$$F = \frac{MSS/(K-1)}{RSS/(n-K)} = \frac{8.7/3}{618.3/266} \approx 1.24.$$

Therefore, we have 3 degrees of freedom at the numerator, 266 at the denominator, and the observed test statistics is approximately 1.24. Comparing with the F table, we note that the p-value is larger than 0.25, so the test is not significant.

Therefore, there is no evidence in the data against the equality of the means, and in average it does not make any difference in terms of (log) area burned in which season of the year the fire happens.

c

Forward selection is a procedure to select a model in which only relevant variables are included. The idea is to start from the null model, only consisting of the intercept, and add one by one variables that improve the model. This can be done by adding variables whose regression coefficients, when the variables are added into the model, are statistically different from 0. Alternatives include looking at the AIC or the adjusted R^2. When no further variables are significant (once added to the model), or another stopping criterion is met (e.g., AIC starts to increase), the procedure stops.

In this case, we would add to the model the variable `DMC`, as the test evaluating if its inclusion improves the model is significant at level 0.05.

d

The intercept $\beta_0 = 1.89$ shows the expected logarithm area burned in spring when `DMC` is equal to 0. $\beta_1 = -0.22$ means that in average we expect -0.22 less in the logarithm of the area burned when the fire is in summer compared to a similar (i.e., in this case, same value for DMC) fire that happens in spring. Same for autumn ($\beta_2 = -0.65$) and winter ($\beta_3 = -0.54$). The last regression coefficient, instead, denoted the expected increase in the logarithm of the area burned when DMC increases of 1, keeping fixed the season. This means that the expected logarithm of the size of the area burned increases of about 0.04 when the DMC increases of 10.

Finally, the expected size of a burned are for a fire that happens in summer with DMC equal to 50 is

$$\log(area) = 1.885746 - 0.215159 + 0.003958 * 50 \approx 1.89,$$

so the expected burned area is $\exp\{1.89\} \approx 6.62$ ha.

e

The model at the previous point has a very low R^2, and when tested against the null model, it does not show a significant difference. In the current model, no regression coefficient is significantly different from 0, which may let us suspect that the additional variables do not help in explaining the variability of the response. More importantly, the adjusted R^2 of this model is smaller than the one in the previous point.

We need to test two nested models, so we can use the likelihood-ratio test

$$G = D_0 - D.$$

Since for the linear regression the deviance is the sum of squares divided by $\hat{\sigma^2}$, and considering that the estimation of the standard error is very similar in the two cases, we can approximate

$$G \approx (609.24 - 599.35)/1.52^2 \approx 4.28$$

that can be evaluated on the table of a χ^2 with 7 degrees of freedom. The probability that $\chi^2_7 > 4.28$ is larger than 0.5, so we do not reject the null hypothesis, i.e., the model at point (d) is preferable to the last one.

Problem 2

a

To test whether the proportion of people with arrhythmia is the same in the male and female populations (assuming, as per the exercise text, that this sample is representative of the population), we can perform a test to compare the two proportions $p_{\text{male}} = 117/202 \approx 0.58$ and $p_{\text{female}} = 88/248 \approx 0.35$. We need to compute

$$\begin{aligned} \hat{p}_{\text{male}} &= 117/202 \approx 0.58 \\ \hat{p}_{\text{female}} &= 88/248 \approx 0.35 \\ \hat{p} &= 205/450 \approx 0.46 \\ se_0(\hat{p}_{\text{male}} - \hat{p}_{\text{female}}) &= \sqrt{\frac{0.46(1-0.46)}{202} + \frac{0.46(1-0.46)}{248}} \approx 0.047 \end{aligned}$$

We are now in the position to compute the test statistic

$$z = \frac{0.58 - 0.35}{0.047} \approx 4.89.$$

This value is larger than 1.96, so we reject the null hypothesis (the proportion of people with arrhythmia is the same in the male and female population) at level 0.05. Looking at the tables (e.g., Student's t with infinite degrees of freedom), the p-value is smaller than 0.0001, so we are quite confident in rejecting the null hypothesis.

Alternatively, one could have performed a χ^2 test,

$$\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E},$$

for which we would have needed to compute the expected numbers,

	normal	arrythmia			normal	arrythmia
male	202*245/450	202*205/450	≈	male	110	92
female	248*245/450	248*205/450		female	135	113

which would have led to

$$\chi^2 = \frac{(85-110)^2}{110} + \frac{(117-92)^2}{92} + \frac{(160-135)^2}{135} + \frac{(88-113)^2}{113} \approx 22.64,$$

Since it is a 2×2 table, i.e., number of columns - 1 times number of rows - 1 is equal 1, we know that the test statistic is distributed as a χ^2 with 1 degree of freedom. Since

$$Pr(\chi^2_1 > 22.64) \approx 0$$

we would have rejected the null hypothesis (i.e., reject the hypothesis that the proportion of people with arrhythmia is the same in the male and female population).

For the following points:

Excess risk: it measures the effect of "exposure" by subtracting from the probability of experiencing the event in the exposed population to that of the control one, $ER = p(y = 1|x = 1) - p(y = 1|x = 0)$;

Relative risk: it also measures the effect of "exposure", but in this case showing how many times more probable is the probability of experiencing the event in the exposed population than that of the control one, $RR = p(y = 1|x = 1)/p(y = 1|x = 0)$;

Odds ratio: it is similar to the relative risk, but instead of using the probabilities it uses the odds, where the “odds” measure how many times is more probable that an event occurs than it does not. In formula, $odds = p/(1-p)$, and $OR = \frac{p(y=1|x=1)/(1-p(y=1|x=1))}{p(y=1|x=0)/(1-p(y=1|x=0))}$.

Here:

- $ER = 117/202 - 88/248 \approx 0.22$;
- $RR = (117/202)/(88/248) \approx 1.63$;
- $OR = (117/202)/(85/202)/((88/248)/(160/248)) \approx 2.50$.

b

The regression coefficient related to the variable `Age` is the log-odds ratio for one year increase in age. When considering the exponential of its value, we get the odd ratio for one year increase in age.

The z-value is $0.003369/0.005853 \approx 0.58$ and the p-value is therefore around 50%. The confidence interval at 95% is

$$0.003369 \pm 1.96 \times 0.005853$$

that leads to a confidence interval approximately equal to $(-0.008; 0.014)$. In terms of odds-ratio, it would be $(0.992; 1.015)$.

c

From this model with all the available variables included, we can see that gender and height are related to the occurrence of arrhythmia. In particular, the odds-ratio between female and male is about 0.26, strongly significant, showing that being a male is a risk factor for arrhythmia. This confirms the result of point (a), and we can say that the gender is significant even when we control for age, height and weight. The fact that we control for other variables is also the reason for which the regression coefficient changes a bit for the variable `Age`. It can be expected, for example, a correlation between age and height and weight. This as an influence on the estimate of the regression coefficients.

Finally, this model seems an improvement with respect to that at the previous point. Roughly speaking, we expect this from the fact that part of the variables added are significant to a very small level. More formally, we can use a likelihood-ratio test,

$$G = D_0 - D = (619.94 - 585.36) \approx 34.58,$$

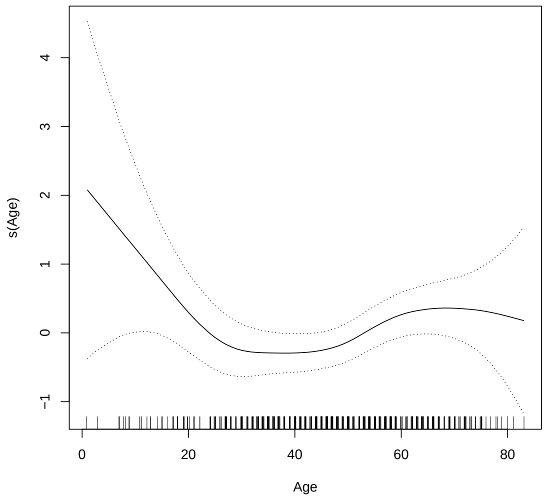
that should be compared to a χ^2 distribution with $448 - 445$ degrees of freedom. The p-value,

$$Pr(\chi^2_3 > 34.58) \approx 0,$$

indicates that there is strong evidence in the data against the null hypothesis that the models are equivalent.

d

It is clear that age has not a linear effect, but it has a quadratic effect on the response (through the link function). The figure below shows the real effect,



In terms of solutions for the exam, a correct plot should have had these characteristics:

- show a non-linear effect, as the quadratic term of `Age` is significant;
- have a quadratic behaviour, as the quadratic term of `Age` is significant;
- be more or less symmetric, as the linear effect is not significant, so a least squares line should be parallel to the x axis.

Problem 3

a

The null hypothesis is that the survival function is the same for the four groups, while the alternative is that at least one is different. To test if there is support in the data against it, we can implement a log-rank test.

In particular, we contrast the value of the test statistic, 2.4 with a χ^2 distribution with 3 degrees of freedom, as in this case "number of groups - 1" is equal to 3. Looking at the tables, we note that $Pr(\chi^2_3 > 2.4) \approx 0.5$, so there seems no support in the data against the null hypothesis.

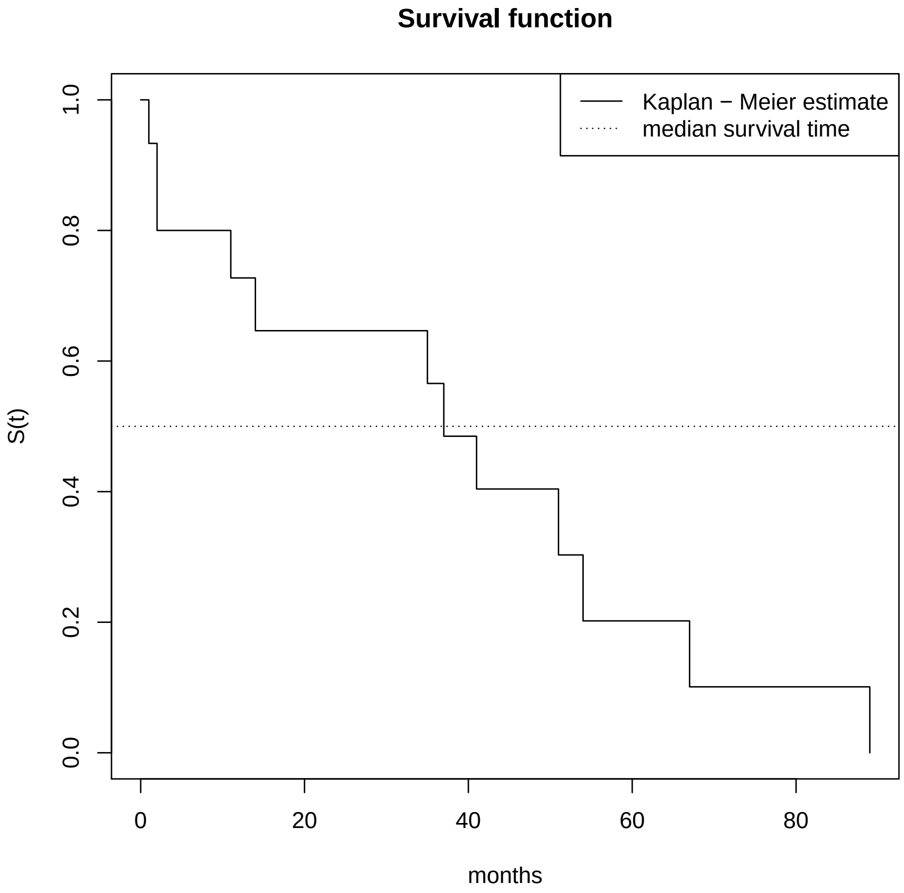
b

Using the same notation of slide 12, lecture 9,

t_j	$Y(t_j)$	m_j	$\frac{m_j}{Y(t_j)}$	$1 - \frac{m_j}{Y(t_j)}$	$\hat{S}(t)$
1	15	1	1/15	14/15	$14/15 \approx 0.93$
2	14	2	2/14	12/14	$14/15 \times 12/14 = 0.8$
11	11	1	1/11	10/11	$0.8 \times 10/11 \approx 0.73$
14	9	1	1/9	8/9	$0.73 \times 8/9 \approx 0.65$
35	8	1	1/8	7/8	$0.65 \times 7/8 \approx 0.57$
37	7	1	1/7	6/7	$0.57 \times 6/7 \approx 0.48$
41	6	1	1/6	5/6	$0.48 \times 5/6 \approx 0.40$
51	4	1	1/4	3/4	$0.40 \times 3/4 \approx 0.30$
54	3	1	1/3	2/3	$0.30 \times 2/3 \approx 0.20$
67	2	1	1/2	1/2	$0.20 \times 1/2 \approx 0.10$
89	1	1	1/1	0	0

so that the median survival time is 37 months.

The plot is displayed in the following page. To identify the median survival time, we need to draw a horizontal line at 0.5 and check at which time it crosses the estimated survival function.



c

This is a Cox model that describes the relationship of four variables, `logBUN`, `age`, `sex`, and `BJprotein`, with the survival time. In a Cox model, we model the hazard, and the exponential of the regression coefficient estimates are called hazard ratios. In particular, 0.9306 is the hazard ratio for females versus males, and 1.9380 that of a subject without the Bence Jone protein present in the urine at diagnosis versus a subject that has it. In both cases, considering all the other covariates constant. For the continuous variables, instead, 8.6028 is the hazard ratio corresponding to one unit's increase in the logarithm of the amount of urea nitrogen in the blood given all the other covariates constant., and, finally, 0.9789 is the hazard ratio corresponding to one year increase in the age of the subject at the beginning of the study, given all the other covariates constant.

From the R output, we note that only `logBUN` and `BJprotein` are significant, with related 95% confidence intervals equal to $(2.4736; 29.919)$ and

$(1.0084; 3.724)$, respectively. It is clear from these intervals that the two variables are significant at 5% because 1 is not included in the intervals.

Finally, the p-value of the likelihood ratio test is 0.01, as the probability of a χ^2 with 4 degrees of freedom (in a Cox model we do not estimate a parameter for the intercept), $Pr(\chi^2_4) \leqslant 13.23$, is 0.99. As $1 - Pr(\chi^2_4) \leqslant 13.23$ it is smaller than 0.05, we can say that, at level 5%, we reject the null hypothesis of the equivalence between this and the null model, and therefore prefer this model to the null one. In other words, there is enough information in the covariates to justify their inclusion.