A brief solution to the written exam in STK9900/STK4900 Statistical methods and applications. Tuesday 7. June 2016.

The exams in STK4900 and STK9900 have substantial overlap, but are not the same. This is a very brief solution to the STK9900 questions, which includes the questions for STK4900.

Exercise 1

a) Due to the presence of some species with large (heavy) brains (Figure 1a), we see that using a log transformed variable makes it reasonable to assume a linear association with TotalSleep (Figure 1b), hence we can use a linear regression model. From Figure 1c) we see that log(BrainWT) and log(BodyWt) are highly correlated (actually, r=0.95). Using both together as covariates will lead to large standard errors for the estimated regression coefficients, see Lecture 3. In practice, this can typically lead to non significant effects of both variables (while they would be significant alone).

b) We have a standard simple linear regression model with regression parameters β_0 and β_1. To test if log(BrainWt) is a significant explanatory variable for TotalSleep, that is, testing

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$$

we use the test statistic

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

where $\hat{\beta}_1$ is the estimated effect of log(BrainWt) and $se(\hat{\beta}_1)$ is the corresponding standard error. Using the output in (I), the test statistic takes the value

$$t = \frac{-1.1688}{0.2325} = -5.028.$$

The value -5.028 is to be compared with the t-distribution with 40 degrees of freedom (40 comes from $42 - 2$). Using the table for t-distributions, we find that the P-value is smaller than $2 \cdot 0.005 = 0.01$ (two-sided, using 40 degrees of freedom from the table). Hence β_1 is significantly different from 0 and log(BrainWt) is a significant explanatory variable for TotalSleep.

c) Since the model is linear in log(BrainWt), and not in BrainWt, the only thing we can say directly from fit (I) is that the negative sign in $\hat{\beta}_1$ tells us that the expected amount of total sleep will be reduced when brain weight increases. The estimated reduction for an increase of 100 g will vary for varying brain weight B as $1.1688 \cdot \log((B + 100)/B)$. For f.ex. $B = 1000$, the expected total sleep is reduced with 0.11 hours, while for $B = 100$, it is reduced with 0.81 hours for an increase of 100 g. In (I) we find Multiple R-squared

=0.3873. For simple linear regression this means that 38.73 % of the variation in total sleep can be explained by the linear relation with log(BrainWt).
d) Danger is an index for the danger category a species is in, and hence a categorical variable. Though ordered, it is better to code Danger as a factor, otherwise the index numbers 1-5 will be interpreted as numerical variables, and we will fit a linear model to those. With "treatment contrast", the model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i$$

where $x_{ki} = 1$ if species i is in danger category $k+1$ and $x_{ki} = 0$ otherwise, $k = 1, 2, 3, 4$ and $i = 1, \ldots, 42$. The ϵ_i are independent $N(0, \sigma^2)$. The β_js are the differences between the expected outcome in index group $j+1$ and the first group.
e) The fit of the model above in (II) says that species tend to sleep less the more in danger they are. Species that are least in danger (index 1) sleep on the average 13.86 hours a day. Danger index 2 gives a reduction in expected total sleep of 1.82 hours, but the difference between index 1 and 2 is not significant. When it comes to the most endangered animals (index 5), they have a reduction compared to the least endangered of 9.3 hours. This reduction is highly significant. With a significance level of 0.1, all coefficients except β_1 are significant.
f) In the fitted model (III), all coefficients are significant at a significance level of 0.1. Also, the Adjusted R-squared has been considerably improved, hence this is a better model than (II). We see that increased brain weight gives reduced total sleep, and that the danger index has almost the same type of effect as described above. One difference is that now being in danger category 4 gives less reduction (with respect to category 1) than being in danger category 3, possibly explained by some special dependencies between brain weight and danger index. Assumptions: 1) linearity, 2) uncorrelated, 3) normally distributed errors with 4) constant variance. The three plots shed light on linearity, normality and variance. Should discuss well.
g) See page 14 Lecture 3. $n = 42, p = 5$.

Exercise 2

a) Deviance test:

$$G = D_0 - D = 19.176 - 10.059 = 9.117.$$

G is χ^2 with 1 df under H_0 (model (I)). From the χ^2 table we find that the P-value must be smaller than 0.005, and conclude that the inclusion of horsepower (hp) is significantly improving the model. (Exact P-value is 0.002532).
b) Model (II) is a logistic regression model. We have data (x_1, x_2, y) for $n = 32$ cars, where x_1 and x_2 are explanatory variables and the outcome y is binary (0 and 1). We are interested in modelling the probability of having a manual transmitter hence we model

$p(x_1, x_2) = P(y = 1|x_1, x_2)$ with a logistic regression model

$$p(x_1, x_2) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}.$$

With $x_1 = 120$ (horsepower) and $x_2 = 2.8$ (weight in 1000 lbs) and the estimated coefficients from (II) we get $P(y = 1|x_1 = 120, x_2 = 2.8)$ estimated as

$$\frac{\exp(18.8663 + 0.03626 \cdot 120 - 8.08348 \cdot 2.8)}{1 + \exp(18.8663 + 0.03626 \cdot 120 - 8.08348 \cdot 2.8)} = 0.64$$

We find the odds ratio for an increase of 10 units in horsepower, while keeping weight fixed, as the ratio between the odds $p(x_1+10, x_2)/(1-p(x_1+10, x_2))$ and $p(x_1, x_2)/(1-p(x_1, x_2))$, which is estimated as $\exp(10 \cdot \hat{\beta}_1) = \exp(0.3626) = 1.44$. Hence the odds of a manual transmitter is increased with 44% for an increased horsepower of 10.

c) We want to check if the effect of weight is linear on the log odds scale. By creating suitable age groups (with approx. the same number of observations in each) we could fit a logistic regression model with the age group as factor (categorical variable), and compare with a model where we use f.ex. the mean of each group as numerical covariate, using a deviance test. If there is no improvement in the fit using a more flexible categorical model, we conclude that the linearity assumption is ok.

Exercise 3

a) We find only approximations from the figure, median survival is approx. 8 weeks for the control group and 23 weeks for the 6MP group. The log rank test tests if the survival function is the same in both groups:

$$H_0 : S_{\text{control}}(t) = S_{\text{6MP}}(t) \text{ for all } t$$

The test statistic for this χ^2-test is 16.8, giving a very small P-value. Hence we reject H_0 and conclude that the two survival functions are different.

b) The model for the hazard rate is

$$h(t|x_1, x_2, x_3) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$

where x_1 is the treatment, $x_1 = 1$ if the patient received 6MP and 0 otherwise, x_2 is gender, $x_2 = 1$ if female, 0 otherwise, and x_3 is the logged white blood cells counts. From R we find $\hat{\beta}_1 = -1.5036$, $\hat{\beta}_2 = 0.3147$ and $\hat{\beta}_3 = 1.6819$.

c) The hazard rate for treatment versus placebo is $\exp(\beta_1)$, estimated as 0.2223. A 95% confidence interval is (0.08998, 0.5493) from R. This means that the hazard is substantially and significantly reduced for the treatment group, and we can conclude that 6MP has a positive effect on the relapse free survival for leukemia patients.