**Solutions to exam problems**

# STK4900 and STK9900 June 12th, 2012

The exam problems for STK4900 and STK9900 have substantial overlap, but are not the same. The solutions below cover both courses, and it is commented when the questions differ for the two courses.

## Problem 1

**a)** The model can be written

$$Y_{ij} = \mu_j + \epsilon_{ij} = \mu_1 + \gamma_j + \epsilon_{ij},$$

where $Y_{ij}$ is response no. $i = 1, ..., 6$ from variety no. $j = 1, ..., 10$, $\mu_j$ the expected value for variety no. $j$ and $\epsilon_{ij}$ are independent error terms with normal distribution, expectation zero and common variance $\sigma^2$. Also $\gamma_j = \mu_j - \mu_1$ is the difference in expectation between variety $j$ and variety 1 for $j = 2, 3, ..., 10$.

The varieties are significantly different as the p-value (for the F-value) is below 0.05. The degrees of freedom for variety equals 9 (= no. of varieties $-1$). The F-value equals $F = 454.34/95.13 = 4.78$.

**b)** The model can be written

$$Y_{ij} = \mu_j + \beta x_{ij} + \epsilon_{ij} = \mu_1 + \gamma_j + \beta x_{ij} + \epsilon_{ij},$$

where $x_{ij}$ is the moisture for variety $j$ and plot $i$ and $\beta$ the regression coefficient for this covariate. The remaining quantities are defined as in question a).

When moisture increases one unit the expected yield will increase by $\hat{\beta} = 0.671$. Furthermore the expected yield for variety one with zero moisture equals 31.99 (may be an extrapolated value) and variety 2 has an expectation 2.88 less than variety 1 (etc.).

**c)** $R^2$ is the explained fraction of the variation. With $\hat{Y}_{ij}$ the predicted yield and $\bar{Y}$ the average yield we can compute

$$R^2 = 1 - \frac{\sum(Y_{ij} - \hat{Y}_{ij})^2}{\sum(Y_{ij} - \bar{Y})^2}$$

For the model in question a) we get

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{MSS}{MSS + RSS} = \frac{4089.1}{4089.1 + 4756.3} = 0.46.$$

Thus the model in question b) with $R^2 = 0.995$ has considerably better ability to predict expected yield.

**Problem 2**

**a)** We consider a situation where the outcome for a fish is 0 or 1, with 0 corresponding to no infection by the parasite and 1 corresponding to infection. For such a situation it is appropriate to use a logistic regression model, which specifies the probability $p$ that a fish is infected as a function of parameters and the covariates. When `year` is the only covariate, the logistic regression model takes the form

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}, \tag{1}$$

where $x_1 = 1$ if the fish is caught in year 2000 ($x_1 = 0$ otherwise), while $x_2 = 1$ if the fish is caught in 2001 ($x_2 = 0$ otherwise).

**b)** To test the null hypothesis that `year` has no effect on the probability that a fish is infected by the parasite, we look at

$$G = D^* - \hat{D}.$$

Here $D^*$ is the null deviance (i.e. the deviance for the model with no covariates) and $\hat{D}$ is the residual deviance (i.e. the deviance for the model with `year` as the only covariate). If there is no effect of `year`, $G$ will be approximately chi-square distributed with 2 degrees of freedom. From output 3 we find that $G = 467.82 - 445.80 = 22.02$. Using the table for the chi square distribution with 2 degrees of freedom, this gives a P-value of less than 0.5%, so `year` has a significant effect.

To find estimates for the probability that a fish is infected, we use formula (1) and the estimates from output 3. This gives:

- For year 1999:

$$\hat{p} = \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = \frac{e^{-0.991}}{1+e^{-0.991}} = 0.271$$

- For year 2000:

$$\hat{p} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1}} = \frac{e^{-0.991+1.287}}{1+e^{-0.991+1.287}} = 0.573$$

- For year 2001:

$$\hat{p} = \frac{e^{\hat{\beta}_0+\hat{\beta}_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_2}} = \frac{e^{-0.991+0.079}}{1+e^{-0.991+0.079}} = 0.287$$

Thus for the years 1999 and 2001 the probability that a fish is infected is less than 30%, while it is almost 60% in year 2000.

**c)** We here consider a model with the covariates `year` and `weight`. The logistic regression model then takes the form

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}.$$

Here $x_1$ and $x_2$ are given as in question a), while $x_3 = \text{weight} - 1.75$.

Let $p_1$ and $p_2$ denote the probabilities of infection for two fishes, labeled 1 and 2, that were caught in the same year (so they have the same values of $x_1$ and $x_2$). We assume that fish 2 weighs 1 kg more than fish 1 (so their weights are $x_3 + 1$ and $x_3$). Then the odds ratio for these fishes becomes:

$$OR = \frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3(x_3+1)}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}} = e^{\beta_3}$$

This is the odds ratio corresponding to 1 kg increase in the weight of a fish (the year they were caught being the same).

Using output 4, we get the estimated odds ratio

$$\widehat{OR} = e^{\hat{\beta}_3} = e^{-0.224} = 0.799.$$

Thus the odds of infection is reduced by 20% when the weight is increased by 1 kg.

A 95% confidence interval for the odds ratio is given by (with $\widehat{se}_3$ the standard error corresponding to $\hat{\beta}_3$):

$$e^{\hat{\beta}_3 \pm 1.96 \cdot \widehat{se}_3} = e^{-0.224 \pm 1.96 \cdot 0.100} = e^{-0.224 \pm 0.196}$$

Thus we are 95% confident that the odds ratio is between $e^{-0.224-0.196} = e^{-0.420} = 0.657$ and $e^{-0.224+0.196} = e^{-0.028} = 0.972$.

**d)** In output 5 we consider the model with covariates `year`, `weight`, and `age`. For this model $e^{\beta_3}$ is the odds ratio for 1 kg increase in weight *keeping age constant*. From output 5 we find the estimated odds ratio $e^{-0.782} = 0.457$. This is much smaller than the estimated odds ratio $e^{-0.224} = 0.799$ found from output 4.

The reason why the odds ratios for the two models differ, is that the effect of weight for the model in output 4 is confounded by age. When age increases, weight will increase as well.

The model of output 5 takes the form

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4}},$$

with $x_1$, $x_2$, and $x_3$ are given as in question c), while $x_4 = \text{age} - 4.4$. The intercept applies to a fish with $x_1 = x_2 = x_3 = x_4 = 0$, i.e. a fish caught in 1999 with weight 1.75 kg and age 4.4 years. The estimate of the probability that such a fish is infected, is given by

$$\hat{p} = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} = \frac{e^{-1.141}}{1 + e^{-1.141}} = 0.319.$$

**Problem 3**

**a)** A discussion of (right) censored survival data is given in Lecture 9. The response is a combination of the censored survival time $T_i = \min(T_i^0, C_i)$, where $T_i^0$ is the true survival time and $C_i$ the censoring time, and the indicator $D_i$ that $T_i = T_i^0$. Doing linear regression

directly on $T_i$ would ignore that many survival times are larger, and doing logistic regression on $D_i$ would ignore that the potential follow-up times are different.

**b)** We see that the survival function for men fall below that of women, thus men tend to live shorter than women. Furthermore, the survival function of non-smokers is higher than those in any of the smoker groups, thus it appears that non-smokers live longer than smokers. Also the high smoking groups seems to have the highest mortality.

**c)** The Cox-model and the hazard ratios are described in Lecture 9, slide 25-28. We see that women have lower mortality than men and smokers appears to have higher mortality than non-smokers, but the difference is not significant.

**d)** This question is only for the STK4900-students.

In the model with only smoking categories (not sex) the smoking groups have significantly higher mortality than non-smokers. Also the hazard ratios are higher than in the model in question c). The reason for the difference between the models is that there is a dependency between sex and smoking. Men aged 65-75 in 1966-71 tended to smoke more than women. The observed difference between smoking groups (figure and only smoking Cox-regression) can partly be attributed to this sex difference.

### Problem 4

This problem is only for the STK9900-students.

**a)** The random effects model assumes that the distance measurement number $j$ for the $i$-th child can be written

$$Y_{ij} = \beta_0 + B_i + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \epsilon_{ij}. \tag{2}$$

for $j = 1, 2, 3, 4$ and $i = 1, 2, \ldots, 27$. Here $x_{1ij}$ is the age of the $i$-th child at the $j$-th measurement, while $x_{2ij} = 0$ if the $i$-th child is a boy and $x_{2ij} = 1$ it it is a girl. Further the random effects $B_1, B_2, \ldots, B_{27}$ are independent and $N(0, \sigma^2_{subj})$-distributed, and independent of the random errors, which are independent and $N(0, \sigma^2_\epsilon)$-distributed.

The corresponding linear regression model is similar to (2), but without the random effect $B_i$. The random effects induce a correlation between the four measurements for a child, and this makes the model more appropriate than the linear regression model (which assumes that the measurements are independent).

**b)** The estimated effect of age is 0.66, so on average the measured distance increases by 0.66 mm per year. The estimated sex effect is $-2.32$, which means that on average the measured distance for girls is 2.32 mm less than for boys.

The estimate of the standard deviation of the random effects takes the value $\hat{\sigma}_{subj} = 1.81$, while the estimated residual standard deviation becomes $\hat{\sigma}^2_\epsilon = 1.43$. From these we obtain the estimate $1.81^2/(1.81^2 + 1.43^2) = 0.62$ for the correlation of two measurements for a child.