Solutions to exam problems

STK4900 and STK9900 June 9th, 2011

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) To test if pressure has a significant influence on the emission of nitrous oxide, we use the t-test statistic

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

where $\widehat{\beta}_1$ is the estimated effect of pressure and $se(\widehat{\beta}_1)$ is the corresponding standard error. Using the output for model 1, the test statistic takes the value $t = 0.018122/0.003494 = 5.19$. The value 5.19 shall be compared with the t-distribution with 18 degrees of freedom. Using the table for t-distributions, we find that the P-value is (much) smaller than $2 \cdot 0.005 = 0.01 = 1\%$, so there is a significant effect of pressure.

b) For model 1 we have $R^2 = 0.599$, while for model 2 we have $R^2 = 0.770$. This means that humidity alone accounts for 77.0% of the variation in the emission of nitrous oxide, while pressure alone accounts for 59.9%. Thus humidity is the best predictor for the emission of nitrous oxide.

c) For model 1 we fit the simple linear regression model:

$$\texttt{N20} = \beta_0 + \beta_1 \, \texttt{PRESS} + \epsilon.$$

Here β_1 is the expected change in emission of nitrous oxide when pressure is increased by 1 mmHg. This expected change is estimated to 0.018 ppm.

For model 3 we fit the multiple linear regression model:

$$\texttt{N20} = \beta_0 + \beta_1 \, \texttt{PRESS} + \beta_2 \, \texttt{HUM} + \epsilon.$$

(For ease of notation we use the same symbols for the two models, even though the βs and ϵ are not the same.) For this model β_1 is the expected change in emission of nitrous oxide when pressure is increased by 1 mmHg *keeping humidity constant*. This expected change is estimated to 0.006 ppm, which is only a third of the estimate from model 1.

The reason why the results for the two models differ, is that the effect of pressure in model 1 is confounded by humidity. When pressure increases, humidity tends to decrease (cf. the matrix scatter plot), and this causes increased emission of nitrous oxide (since the estimated regression coefficient for humidity is negative). Thus parts of the effect attributed to pressure in model 1 is in reality an effect of humidity. When we in model 3 control for this confounding, the effect of pressure is reduced.

Using the output for model 3, the t-test statistic for pressure now takes the value

$$t = \frac{0.0060639}{0.0038901} = 1.56.$$

Comparing this value of the test statistic with the t-distribution with $n-p-1 = 20-2-1 = 17$ degrees of freedom, we get a P-value between $2 \cdot 0.05 = 0.10 = 10\%$ and $2 \cdot 0.10 = 0.20 = 20\%$. Thus the effect of pressure is not significant when humidity is included in the model.

Problem 2

a) The outcome for a child is 0 or 1, with 0 corresponding to no respiratory disease and 1 corresponding to development of respiratory disease during first year of life[1]. For such a situation, it is appropriate to use a regression model that relates the probability p that a child develops respiratory disease to its covariates, and this is achieved by using a logistic regression model.

b) When type of feeding is the only covariate, the logistic regression model takes the form

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

Here $x_1 = 1$ if the child is fed by some breast with supplement ($x_1 = 0$ otherwise), while $x_2 = 1$ if the child is breast fed ($x_2 = 0$ otherwise). Thus the reference is a child fed by bottle for whom $x_1 = x_2 = 0$. The corresponding odds is given by

$$\text{odds}(x_1, x_2) = \frac{p(x_1, x_2)}{1 - p(x_1, x_2)} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}.$$

Hence the odds ratio for a child who is breast fed ($x_1 = 0$, $x_2 = 1$) relative to a child who is fed by bottle ($x_1 = x_2 = 0$) takes the form

$$OR = \frac{\text{odds}(0,1)}{\text{odds}(0,0)} = \frac{e^{\beta_0 + \beta_2}}{e^{\beta_0}} = e^{\beta_2}.$$

Using the output for model 4, the estimated odds ratio becomes

$$\widehat{OR} = e^{\hat{\beta}_2} = e^{-0.67645} = 0.508.$$

Thus the odds for developing respiratory disease for a child who is breast fed is about half the odds for a child who is fed by bottle.

A 95% confidence interval for the odds ratio is given by

$$e^{\hat{\beta}_2 \pm 1.96 \cdot se(\hat{\beta}_2)} = e^{-0.67645 \pm 1.96 \cdot 0.15281} = e^{-0.67645 \pm 0.29951}.$$

[1] The data in the problem are given on aggregated form. For each combination of the levels of the factors (sex, feeding), we have the total number of children and the number of children who develop respiratory disease.

Thus we are 95% confident that the odds ratio is between $e^{-0.67645-0.29951} = e^{-0.97596} = 0.377$ and $e^{-0.67645+0.29951} = e^{-0.37694} = 0.686$.

c) The STK4900-students are just asked to test the null hypothesis that sex has no effect, and then any one of the two tests given below will do. But the STK9900-students are asked to test the null hypothesis in two different ways, so they should provide both tests.

To test the null hypothesis that sex has no effect for the risk of developing respiratory disease (when adjusting for type of feeding), we may either use a Wald test or a test based on the deviances (which is the same as a likelihood ratio test).

Wald test: We use the test statistic:

$$z = \frac{\widehat{\beta}_3}{se(\widehat{\beta}_3)},$$

where $\widehat{\beta}_3$ is the estimated effect of sex and $se(\widehat{\beta}_3)$ is the corresponding standard error. Using the output for model 5, the test statistic takes the value $z = -0.3126/0.1410 = -2.22$. The value -2.22 shall be compared with the standard normal distribution. Using the table for the standard normal distribution, we find the P-value $2 \cdot 0.0132 = 0.026 = 2.6\%$. Thus it is a significant effect of the sex of the child, with girls having lower risk of developing respiratory disease than boys.

Deviance test: Here we use the test statistic

$$G = D_0 - D,$$

where D_0 is the (residual) deviance for the model without sex (model 4) and D is the (residual) deviance for the model with sex included (model 5). If there is no effect of sex, G will be approximately chi-square distributed with 1 degree of freedom. Using the output from models 4 and 5 we find $G = 5.699 - 0.722 = 4.977$. Comparing this value with the table for the chi-square distribution with 1 degree of freedom, we obtain a P-value between 2.5% and 5%. Thus sex has a significant effect on the risk of developing respiratory disease.

d) This question is only for the STK9900-students.

There is an interaction between type of feeding and sex if the effect of type of feeding differs between boys and girls (or equivalently if the effect of sex differs according to the type of feeding). To test if there is an interaction, we my use the deviance test. The test statistic is $G = D_0 - D$, where now D_0 is the (residual) deviance for the model with main effects of type of feeding and sex, but no interaction (model 5), while D is the (residual) deviance for the model where also interaction between type of feeding and sex is included. If there is no interaction, G will be approximately chi-square distributed with 2 degrees of freedom. The model with interaction gives a perfect fit to the (grouped) data, so $D = 0$. Hence $G = 0.722 - 0 = 0.722$. Comparing this value with the table for the chi square distribution with 2 degree of freedom gives a P-value (much) larger than 5%. Thus we do not reject the hypothesis that there is no interaction.

Problem 3

a) We start out by considering one cell (lymphocyte) in experiment number i; $i = 1, 2, \ldots, 27$. If we assume that

- the rate of occurrence of chromosomal abnormalities is constant over time,
- the number of occurrences in disjoint intervals are independent,
- chromosomal abnormalities occur one at a time,

then the process counting the occurrences of chromosomal abnormalities in the cell is a Poisson process. It follows that the number of chromosomal abnormalities that occur in the cell during the whole experiment is Poisson distributed with an expected value λ_i, say. Then λ_i is the rate of occurrence of chromosomal abnormalities per cell in experiment number i. If we let w_i be the number of cells in experiment i, then the total number of chromosomal abnormalities (i.e. for all cells) in the experiment is Poisson distributed with expected value $w_i\lambda_i$.

b) Let y_i be the total number of chromosomal abnormalities observed in experiment number i, and introduce

$$\begin{aligned}
x_{1i} &= \begin{cases} 1 & \text{if dose in experiment } i \text{ is 2.5 Grays} \\ 0 & \text{otherwise} \end{cases} \\
x_{2i} &= \begin{cases} 1 & \text{if dose in experiment } i \text{ is 5.0 Grays} \\ 0 & \text{otherwise} \end{cases} \\
x_{3i} &= \log(\texttt{doserate}) \quad \text{in experiment number } i
\end{aligned}$$

Then model 6 assumes that the y_i are independent and Poisson distributed, $y_i \sim \text{Po}(w_i\lambda_i)$, where the rate of occurrence of chromosomal abnormalities per cell is specified as

$$\lambda_i = e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}}.$$

The rate ratio for dose 2.5 Grays relative to dose 1.0 Gray is e^{β_1}, and the estimate of this rate ratio becomes $e^{\hat{\beta}_1} = e^{1.65299} = 5.22$. Thus we expect to get about five time as many chromosomal abnormalities when the dose is 2.5 Grays than when the dose is 1.0 Gray.

c) We then consider the model with interaction between dose and the logarithm of dose rate. For this model the rate of occurrence of chromosomal abnormalities per cell is assumed to take the form

$$\lambda_i = e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}+\beta_4 x_{1i}x_{3i}+\beta_5 x_{2i}x_{3i}}.$$

The model assumes that the effect of dose rate is different for the three dose levels. Spelled out the model may be written:

$$\lambda_i = \begin{cases} e^{\beta_0+\beta_3 \log(\texttt{doserate})} & \text{if dose in experiment } i \text{ is 1.0 Gray} \\ e^{\beta_0+\beta_1+(\beta_3+\beta_4)\log(\texttt{doserate})} & \text{if dose in experiment } i \text{ is 2.5 Grays} \\ e^{\beta_0+\beta_2+(\beta_3+\beta_5)\log(\texttt{doserate})} & \text{if dose in experiment } i \text{ is 5.0 Grays} \end{cases}$$

To test if there is an interaction, we my use the deviance test. The test statistic is $G = D_0 - D$, where D_0 is the (residual) deviance for the model without interaction (model 6), while D is the (residual) deviance for the model with interaction (model 7). If there is no interaction, G will be approximately chi-square distributed with 2 degrees of freedom (equal to the number of parameters used to model the interaction). From the output for models 6 and 7 we obtain $G = 42.78 - 21.75 = 21.03$. Comparing this value with the table for the chi square distribution with 2 degrees of freedom gives a P-value smaller than 0.5%. Thus we reject the hypothesis that there is no interaction and conclude that there is a significant interaction between dose and the logarithm of dose rate.