A brief solution to the written exam in STK9900/STK4900 Statistical methods and applications. Thursday 6. June 2013.

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

Exercise 1

a) To test if the expression of gene 1 is a significant explanatory variable for bone density, we use the t-test statistic

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

where $\hat{\beta}_1$ is the estimated effect of gene 1 and $se(\hat{\beta}_1)$ is the corresponding standard error. Using the output in (I), the test statistic takes the value

$$t = \frac{-1.1297}{0.3609} = -3.13.$$

The value -3.13 is to be compared with the t-distribution with 82 degrees of freedom. Using the table for t-distributions, we find that the P-value is smaller than $2 \cdot 0.005 = 0.01$ (two-sided, using 60 degrees of freedom from the table). Hence β_1 is significantly different from 0 and gene 1 is a significant explanatory variable for bone density.

b) In the results (II) we see that gene 1 is no longer significant when gene 2 is included. Gene 2 i highly significant. Hence if we have information on gene 2, gene 1 does not contain enough additional information on bone density to be significant. This can happen when the covariates are (highly) correlated. We see that the correlation between gene 1 and gene 2 is 0.73.

c) The second model has only significant covariates and a slightly higher Adjusted R^2, hence should be considered better that the first (the first model has a higher R^2, but has also one more covariate, so we use Adjusted R^2 for comparison here). For the second model, we interpret the estimated effects as: A high gene expression for gene 2 tends to reduce expected bone density, when the other covariates (genes) are kept unchanged. A high gene expression for gene 3 tends to increase expected bone density, when the other genes are unchanged. The same for gene 4. We also see that gene 3 has the largest estimated effect on bone density.

d) [9900] Short answer: Interaction between covariates is when the effect of one covariate on the response depends on the level of another covariate. Some more discussion appreciated, explaining how this can be modelled etc., see lecture notes Lecture 4. In the setting here, interaction between genes would mean f.ex. that the effect of a high expression of gene r could depend on the expression level of another gene s.

Exercise 2

a) We have data (x_1, x_2, y) for $n = 35$ subjects (surgeries), where x_1 and x_2 are explanatory vaiables and the outcome y is binary (0 and 1). We are interested in modelling the probability of getting a sore throat as a function of the explanatory variables, hence we model $p(x_1, x_2) = P(y = 1|x_1, x_2)$ with a logistic regression model

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

This ensures that $p(x_1 + x_2)$ is a proper probability, and the log odds will be linear in the covariates.

b) With only one explanatory variable x_1, we find

$$P(y = 1|x_1 = 30) = \frac{\exp(\beta_0 + \beta_1 \cdot 30)}{1 + \exp(\beta_0 + \beta_1 \cdot 30)}$$

estimated as

$$\frac{\exp(-2.21358 + 0.07038 \cdot 30)}{1 + \exp(-2.21358 + 0.07038 \cdot 30)} = 0.4745.$$

c) The odds ratio for 40 minutes wrt. 30 minutes is $\exp(10 \cdot \hat{\beta}_1) = \exp(0.7038) = 2.02$. A confidence interval for $10 \cdot \beta_1$ is found through

$$10 \cdot \hat{\beta}_1 \pm 1.96 se(10 \cdot \hat{\beta}_1),$$

where $se(10 \cdot \hat{\beta}_1) = 10 \cdot se(\hat{\beta}_1)$. With the estimates from R we get $0.7038 \pm 1.96 \cdot 0.2667$, giving the 95% confidence interval (0.181068, 1.226532). This gives a 95% confidence interval for the odds ratio as

$$(\exp(0.181068), \exp(1.226532))$$

that is, (1.199, 3.409). The OR tells us that the odds for a sore throat are doubled if the surgery lasts 10 minutes more. From the confidence interval we can conclude that there is some uncertainty about this doubling, but at least we can say that the odds are increased (starting above 1).

d) Wald test:

$$H_0 : \beta_2 = 0 \quad \text{versus} \quad H_a : \beta_2 \neq 0$$

From R we find the z-value -1.798, and a P-value$= 2P(Z > 1.798) = 2 \cdot 0.036 = 0.072$ where Z is standard normal. This P-value is not small enough to reject H_0 at level 0.05, so the type (x_2) is not significantly improving the model.

Deviance test:

$$G = D_0 - D = 33.651 - 30.138 = 3.513$$

G is χ^2 with 1 df under H_0. From the χ^2 table we find that the P-value must be larger than 0.05, and conclude again that type (x_2) is not significantly improving the model.

Exercise 3

a) The estimated rate ratio is $\exp(1.2016) = 3.325$, hence the rate of wave caused damages is more than three times larger for the ships built in 70-74 compared to 60-64.

b) [9900] For an explanation of over dispersion, see lecture notes for Lecture 8. Should write down the overdispersed Poisson regression model, introducing $Var(Y_i) = \phi\mu_i$, and explain that using the quasi family in the glm allows to estimate ϕ and corrects the standard errors to

$$se^* = se\sqrt{\hat{\phi}}.$$

For the ship damage data, $\hat{\phi} = 3.19643$ and the standard errors of the estimated coefficients are scaled up with a factor

$$\sqrt{\hat{\phi}} = 1.788,$$

and the test statistics and p-values corrected correspondingly. The p-values become bigger, but the effects are still significant.

c) This is a Poisson model with observations that are aggregated counts. If Y_i is the number of counts in group i, the model can be written as

$$E(Y_i) = \exp(\log(\text{months}_i) + \beta_0 + \beta_1 x_{Bi} + \beta_2 x_{Ci} + \beta_3 x_{Di} + \beta_4 x_{Ei}$$

$$+\beta_5 x_{65-69,i} + \beta_6 x_{70-74,i} + \beta_7 x_{75-79,i} + \beta_8 x_{op75-79,i})$$

where $x_{Bi} = 1$ if ship type in group i is B and 0 otherwise, etc. We test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_8 = 0$$

using a deviance test:

$$G = D_0 - D = 73.225 - 38.695 = 34.53.$$

G is χ^2 with 5 df under H_0. $P(\chi^2_5 > 34.53) < 0.005$ and we reject H_0 at any reasonable level of significance. Hence we can conclude that the new variables significantly improve the model.

d) Using estimated rate ratios, we find that ship type B, C and D all are safer than ship type A, when all other covariates are unchanged. Ship type E is more dangerous.

Specifically we find RR=exp(-0.6874)=0.5029 for ship type C compared to ship type A, when all other covariates are unchanged. Similarly we find RR=exp(-0.5433)=0.5808 for ship type B compared to A. Ship type C has therefore the lowest rate of wave caused damage, but the standard error of $\hat{\beta}_2$ is much larger than that of $\hat{\beta}_1$, so it is also reasonable to argue that ship type B should be considered the safest, given that the difference in RR is very small.

The RR for operation in the last period wrt the first is RR=exp(0.38447)=1.4688. This means that there was an increased risk of damage for operations in the period 1975-79 compared to the previous period 1960-74, when all other covariates as ship type and year of construction are the same.