Solutions to exam questions

STK9900 June 8th, 2010

Problem 1

a) To test if there is a significant effect of mother's age, we use the t-test statistic

$$t = \frac{\widehat{\beta}_{\text{age}}}{\widehat{se}_{\text{age}}},$$

where $\widehat{\beta}_{\text{age}}$ is the estimated effect of `AGE` and $\widehat{se}_{\text{age}}$ is the corresponding standard error. Using the output for Model 1, the test statistic takes the value $t = 0.00854/0.00399 = 2.14$. The value 2.14 shall be compared with the t-distribution with $n - p - 1 = 500 - 3 - 1 = 496$ degrees of freedom, which is (almost) the same as the standard normal distribution. Using the table for the standard normal distribution, we get the (two-sided) P-value 3.2%, so there is a significant effect of the age of the mother.

b) Using the output for Model 2, the t-test statistic now takes the value $t = 0.00206/0.00425 = 0.48$. Comparing this value of the test statistic with the t-distribution with $500 - 4 - 1 = 495$ degrees of freedom, which is (almost) the same as the standard normal distribution, we now get the P-value 63%. Thus the effect of mother's age is not significant in Model 2.

From Model 2, however, we see that there is a significant effect of `FIRST`, showing that the first child of a woman on average has a lower birth weight than later children. Moreover, since a woman who gets her first child tends to be younger than a woman who has got at least one child, there will be a positive correlation between `AGE` and `FIRST`. This implies that `FIRST` is a confounder in Model 1, and that the effect of `FIRST` is taken up by `AGE` for this model.

c) The estimated effects of the covariates in Model 3 have the following interpretation:

- `SEX`: If we consider newborn boys and girls with the same values of the covariates `WEEKS` and `FIRST`, the girls will on average weigh 0.114 kg less than the boys.
- `WEEKS`: If we consider two groups of newborn babies with the same values of the covariates `SEX` and `FIRST`, but where the pregnancies for one of the groups lasted one week longer than for the other group, then the group with the longest pregnancies will on average weigh 0.160 kg more than the other group. Thus babies on average put on 0.160 kg per week towards the end of the pregnancy.

- `FIRST`: If we consider two groups of newborn babies with the same values of the covariates `SEX` and `WEEKS`, but where one group is firstborn babies and the other groups is not, then the babies who are not firstborn will on average weigh 0.173 kg more than the firstborn babies.

We predict the weight of a newborn girl who is the second child of her mother, and where the length of the pregnancy is 40 weeks, to be

$$\hat{y} = -2.857 - 0.114 + 0.1597 \cdot 40 + 0.173 = 3.590 \text{ kg}$$

Problem 2

a) We consider a situation where the outcome for a worker is 0 or 1, with 0 corresponding to absence of byssinosis and 1 corresponding to presence of the disease[1]. For such a situation it is appropriate to use a regression model that relates the probability p that a worker suffers from byssinosis to the covariates, and this is achieved by using a logistic regression model.

b) When DUST is the only covariate, the logistic regression model takes the form

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

Here $x_1 = 1$ if there is medium dustiness of the workplace ($x_1 = 0$ otherwise), while $x_2 = 1$ if there is low dustiness of the workplace ($x_2 = 0$ otherwise). From the output for Model 1, we find that the (estimated) probability $\hat{p}$ that a worker suffers from byssinosis depends on the dustiness of the workplace in the following way:

- If the workplace has high dustiness we have

$$\hat{p} = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} = \frac{e^{-1.681}}{1 + e^{-1.681}} = 0.157$$

- If the workplace has medium dustiness we have

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}} = \frac{e^{-1.681 - 2.585}}{1 + e^{-1.681 - 2.585}} = 0.014$$

- If the workplace has low dustiness we have

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_2}} = \frac{e^{-1.681 - 2.715}}{1 + e^{-1.681 - 2.715}} = 0.012$$

[1]The data in the problem are given on aggregated form. For each combination of the levels for the factors, we know the total number of workers and the number of workers who suffer from byssinosis.

Thus while the probability of suffering from byssinosis is 15.7% for a worker in a workplace with heavy dustiness, it is only 1.4% if the dustiness is moderate and 1.2% if the dustiness is low.

c) We here consider a model with the factors `DUST` and `EMPLOY`. The logistic regression model then takes the form

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

Here x_1 and x_2 are given as in question b, while $x_3 = 1$ for a worker who has been employed between 10 and 20 years ($x_3 = 0$ otherwise), and $x_4 = 1$ for a worker who has been employed more than 20 years ($x_4 = 0$ otherwise).

Let p_1 and p_2 denote the probabilities of suffering from byssinosis for two workers, labeled 1 and 2, who have a workplace with the same level of dustiness (i.e. the same values of x_1 and x_2). Worker 2 has been employed between 10 and 20 years, while worker 1 has been employed less than 10 years. The odds ratio between worker 2 and worker 1 is given by

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

Using the output from Model 2, the estimated odds ratio becomes

$$\widehat{OR} = e^{\hat{\beta}_3} = e^{0.564} = 1.76.$$

Thus the odds for a worker who has been employed 10 to 20 year is 76% higher than the odds for a worker who has been employed less than 10 years.

Further, let p_3 denote the probability of suffering from byssinosis for a worker, labeled 3, who has a workplace with the same level of dustiness as worker 2, and who has been employed more than 20 years. The odds ratio between worker 3 and worker 2 is given by

$$OR = \frac{\frac{p_3}{1-p_3}}{\frac{p_2}{1-p_2}} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_4}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3}} = e^{\beta_4-\beta_3}.$$

An estimate for the odds ratio becomes

$$\widehat{OR} = e^{\hat{\beta}_4-\hat{\beta}_3} = e^{0.673-0.564} = 1.12.$$

Thus the odds for a worker who has been employed more than 20 years is 12% higher than the odds for a worker who has been employed between 10 to 20 years.

d) To test the null hypothesis that smoking has no effect for the risk of suffering from byssinosis, we look at the difference in deviance for Model 2 and Model 3. More precisely, we look at

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

where D^* is the (residual) deviance for the model without smoking (Model 2) and $\hat{D}$ is the (residual) deviance for the model with smoking (Model 3).

If there is no effect of smoking, G will be approximately chi-square distributed with 1 degree of freedom. Using the output from Models 2 and 3 we find that $G = 23.53 - 12.09 = 11.44$. Using the table for the chi square distribution with 1 degree of freedom this gives a P-value of less than 0.5%, so smoking has a significant effect on the risk of suffering from byssinosis.

Problem 3

a) The Kaplan-Meier curves are estimates of the survival function $S(t)$, i.e. the probability that an individual will be alive t days after the start of the study. From the plot we see that the cirrhosis patients with no ascites have better survival than those with ascites. After 1000 days, for example, the survival probability is about 70% for patients without ascites, while it is about 40% for those with ascites. After 2000 day the corresponding survival probabilities are about 50% and 20%, respectively. By looking at the median survival time (i.e. the time when the survival probability is 50%) we get another illustration of the difference in survival for the two groups. For patients without ascites the median survival time is about 2000 days, while it is only 3-400 days for patients with ascites.

b) For the cirrhosis patients we do not know all the survival times – for some patients we only know that they were still alive at the end of the study. This gives rise to censored observations.

If there had been no censoring, an option would be to use multiple linear regression for the logarithms of the survival times. Due to censoring, this is not feasible, however. Cox regression is an appropriate method for handling censored survival data. For Cox's model one relates the hazard of an individual to its covariates.

In this problem we have two binary covariates: $x_1 = 1$ for a male and $x_1 = 0$ for a female, while $x_2 = 1$ for a patient with ascites and $x_2 = 0$ for a patient without ascites. Then Cox's regression model specifies the hazard for a patient with covariates x_1 and x_2 to be of the form

$$\lambda_{x_1,x_2}(t) = \lambda_0(t)\, e^{\beta_1 x_1 + \beta_2 x_2}.$$

Here $\lambda_0(t)$ is the hazard for a reference individual – a female without ascites.

c) Let $\lambda_1(t)$ and $\lambda_2(t)$ denote the hazards for two cirrhosis patients, labeled 1 and 2, who are of the same gender (i.e. have the same value of x_2). Patient 1 has no ascites, while patient 2 has ascites. Then the hazard ratio between patient 2 and patient 1 is given by:

$$RR = \frac{\lambda_2(t)}{\lambda_1(t)} = \frac{\lambda_0(t)\, e^{\beta_1 + \beta_2 x_2}}{\lambda_0(t)\, e^{\beta_2 x_2}} = e^{\beta_1}.$$

This is the hazard ratio between patients (of the same gender) with and without ascites.

Using the output given in the problem, we get the estimated hazard ratio

$$\widehat{RR} = e^{\hat{\beta}_1} = e^{0.894} = 2.44.$$

Thus the hazard for patient with ascites is more than twice as high as the hazard for a patient without ascites.

A 95% confidence interval for the hazard ratio is given by (with $\widehat{se}_1$ the standard error corresponding to $\hat{\beta}_1$):

$$e^{\hat{\beta}_1 \pm 1.96 \cdot \widehat{se}_1} = e^{0.894 \pm 1.96 \cdot 0.132} = e^{0.894 \pm 0.259}.$$

Thus we are 95% confident that the hazard ratio is between $e^{0.894-0.259} = e^{0.635} = 1.89$ and $e^{0.894+0.259} = e^{1.153} = 3.17$.