

STK4900 Oblig 2

Daniel Heinesen, daniehei

13. april 2019

1 Problem 1

1.1 a)

We have a data set where the outcome is whether the female crab has one or more satellites $y = 1$, or none $y = 0$. We are looking for a regression that can, given the covariates, give us a probability that the female crab have satellites. This means that we are looking for a regression model that gives us a probability for a binary outcome. The best choice for such a model is a logistic regression model

$$p(x_1, \dots, x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}, \quad (1)$$

where p is the desired probability, x_i the covariates and β_i their fitted coefficients.

Is that the best way to describe β_i ?

1.2 b)

We want to find the odds ratio between crabs that differ with one centimetre in width. We know that with width as the only covariate, we define the odds as

$$\frac{p(x)}{1 - p(x)} = \exp(\beta_0 + \beta_1 x). \quad (2)$$

From this we get the odds ratio for a difference in one centimetre

$$OR = \frac{p(x+1)/[1 - p(x+1)]}{p(x)/[1 - p(x)]} = \frac{\exp(\beta_0 + \beta_1 \cdot (x+1))}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1 \cdot 1) = \exp(\beta_1). \quad (3)$$

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-12.3508	2.6287	-4.70	0.0000
width	0.4972	0.1017	4.89	0.0000

Tabell 1: Summary of the logical regression done the satellite crabs, with width as the only covariate.

We get the β 's from the logical regression found in table 1. From this we see that

$$OR = \exp(0.4972) = 1.64. \quad (4)$$

This means that odds of a female crab having satellite males increase by 64% if the width of female increases by 1 cm.

If $p(1) = p(x+1)$ and $p(0) = p(x)$ are small, we can approximate the odds ratio with the relative risk $RR = p(1)/p(0)$. In this case we have that $p(1) = 6.8 \cdot 10^{-6}$ and $p(0) = 4.1 \cdot 10^{-6}$, which means that both are very small. This means that we can assume that $RR \approx OR$. And comparing them $OR = 1.6367$ and $RR = 1.6487$ we see that this is correct.

Since we know that

$$t = \frac{\beta}{se(\beta)} \quad (5)$$

is close to normally distributed, we can use this to find a confidence interval for the odds ratio. We simply do this by calculating the confidence interval for β_1 and then taking exp of this interval.

From tab. 2 we see that we get a confidence interval $CI = (1.35, 2.01)$. Since 1 is outside of this interval, we can say that width gives an significant increase in the probability of satellites.

	expcoef	lower	upper
(Intercept)	4.33e-06	2.50-08	0.00075
width	1.64	1.35	2.01

Tabell 2: Confidence interval for the odds ratio for crabs differing by one cm in width.

1.3 c)

1.4 d)

1.5 e)

2 Problem 2

2.1 a)

We have an outcome, medals, which is a count outcome, and we therefore assume that the count is distributed with a Poisson distribution $Y_i \sim Po(\lambda_i)$. Such a distribution is parametrized with a parameter λ , the rate. In our data we will have that this rate is dependent on several covariates, so we need a way to determine this dependence on the covariates. It is here we use Poisson regression. Given n independent subjects, we have that

- y_i is the count of the i^{th} subject
- x_{ij} is the j^{th} covariate for the i^{th} subject

We then define our model as

$$\lambda_i = \lambda(x_{1,i}, \dots, x_{p,i}) = \exp(\beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i}). \quad (6)$$

It is this model we want to fit to the medal counts. But there is something we have to be careful with: A country with a higher number of athletes will most likely have more medals than a country with fewer athletes. So instead of the medal count following the distribution $Y_i \sim Po(\lambda_i)$ we instead say that they follow the distribution $Y_i \sim Po(w_i \lambda_i)$, where w_i is the number of athletes representing the country. But how do we use this in our model? We can see this from taking the expected value

$$E[Y_i] = w_i \lambda_i = w_i \exp(\beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i}) = \exp(\log(w_i) + \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i}). \quad (7)$$

This means that to compensate for this imbalance of athletes, we can fit our model with $\log(w_i)$ as a covariate. This is what we call an offset. This we already *Log.athletes* in our data, we can just use this as the offset in our regression.

2.2 b)

To find a fit for our model we are going to use two methods. The first is to fit a model with all the covariate and see which of them are significant. The other method is to add one covariate after another and use a two-way ANOVA to see if the addition is significant.

From table3 we see the summary of the fit with all covariate. We see that with p-values close to 0

3 Problem 3

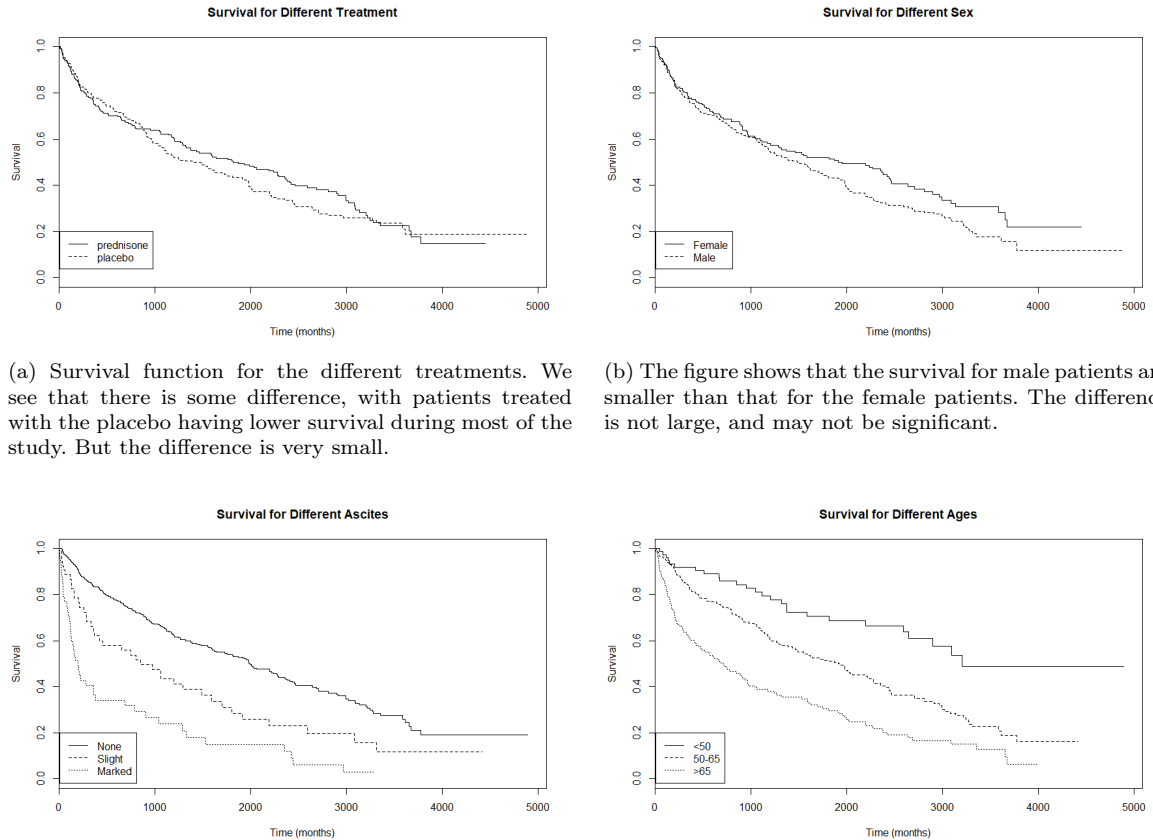
3.1 a)

We are here looking at the survival of patients with cirrhosis treated with prednisone. We are first going to look at the Kaplan-Meier plots, which are plots of the estimated survival function.

	Estimate	Std. Error	z value	$\Pr(> z)$
(Intercept)	-2.8623	0.3191	-8.97	0.0000
Log.population	0.0275	0.0315	0.87	0.3831
GDP.per.cap	-0.0149	0.0032	-4.65	0.0000
Total1996	0.0118	0.0016	7.36	0.0000

Tabell 3: Summary of a Poisson regression with all the covariates. *Log.population* is the logarithm of the nation's population size per 1000, *GDP.per.cap* is the GDP per capita and *Total1996* is the medal count for the previous Olympic Games.

We are going to look at four different plots corresponding to treatment(whether the patient got prednisone or a placebo), the sex of the patient, the severity of fluid build up in the abdomen (ascites) at the start of the observation, and their age group.



(a) Survival function for the different treatments. We see that there is some difference, with patients treated with the placebo having lower survival during most of the study. But the difference is very small.

(b) The figure shows that the survival for male patients are smaller than that for the female patients. The difference is not large, and may not be significant.

(c) The figure shows that there is a large difference in the survival based on the severity of the ascites of the patient at the start of the observation. The more fluid the patient has, the lower the survival.

(d) The figure shows that the age plays a important role in the survival, with older patients surviving shorter than the younger ones.

Figure 1: Four figure showing the estimated survival function for different groups.

If we look at figure 1 we see the different Kaplan-Meier plots for the different groups. Fig. 1a shows the difference in survival based on the treatment. We see that, (especially in the middle of the observation) patients with placebo the survival seems to be lower than for the patients with prednisone. This may indicate that the treatment helps, but we see that the difference is so small that the difference may be due to other factors.

Looking at fig. 1b we can clearly see that male patients has a lower survival for the entire observational period than female period. But again the difference is so small that we ought to be sceptical about the significance.

For the ascites, fig. 1c, the difference in the severity of the fluid buildup seems to have a large impact on the survival of the patients. There is a clear decrease in survival for a patient with slight ascites compared with one with none, and an even worse decrease if the patient has marked ascites.

The same clear difference can be seen in the estimated survival of the groups with different ages, fig. 1d. The survival of the decreases drastically with the increase in age.

3.2 b)

We have seen above that there is some difference in all the K different groups. Now we want to check if these differences are significant. To check this we are going to use *logrank tests*. In this test we test the null hypothesis that a set with groups have the same survival function. We do this by finding the observed number of events in each group, giving us O_i . Then we find the expected number of events E_i , which is the number of observations we would expect if all the groups were the same. With these we can compute the χ^2 test statistic

$$\chi_i^2 = \frac{(O_i - E_i)^2}{se(O_i - E_i)}. \quad (8)$$

This follows a χ^2 distribution with $K - 1$ degrees of freedom.

Type	N	Observed	Expected	$(O - E)^2/E$	$(O - E)^2/V$
treat=0	251	142	149	0.355	0.728
treat=1	237	150	143	0.371	0.728
Treatment	Chisq= 0.7 on 1 degrees of freedom, p= 0.4				
sex=0	198	111	127	2.00	3.55
sex=1	290	181	165	1.54	3.55
Sex	Chisq= 3.5 on 1 degrees of freedom, p= 0.06				
asc=0	386	211	251.9	6.63	48.66
asc=1	54	39	26.2	6.30	6.94
asc=2	48	42	14.0	56.7	59.6
Ascites	Chisq= 69.9 on 2 degrees of freedom, p= 7e-16				
agegr=0	80	26	58.7	18.18	22.87
agegr=1	250	148	162.0	1.21	2.72
agegr=2	158	118	71.3	30.51	40.87
Ascites	Chisq= 50.6 on 2 degrees of freedom, p= 1e-11				

Tabell 4: Table showing logrank tests for the different groups.

Table 4 shows us the logrank tests for the different covariates with their respective groups. Looking at the difference between placebo and prednisone we see that we have a small $\chi^2 = 0.7$ and large $p = 0.4$. This means that we cannot conclude that prednisone increase survival – as we saw in fig. 1a.

In the difference between male and female patients, we see in fig. 1b there female patients seemed to have a bit longer than male patients. But from tab. 4 we see that we get a statistic $\chi^2 = 3.5$ and $p = 0.006$. This means the p is a bit too large for us to say that it is significant. Thus we cannot conclude that there is a difference between the sexes.

Fro both ascites and age group we saw in fig. 1c and 1d that there seems to be a large difference in the survival between the groups. From tab. 4 we see for the difference severities of ascites we get $\chi^2 = 69.9$ and $p = 7 \cdot 10^{-16}$, and for different age groups we get $\chi^2 = 50.6$ and $p = 1 \cdot 10^{-11}$. So we conclude that there is a significant difference between the different severities in ascites, and there is a significant difference between the age groups.

3.3 c)

We now want to estimate the hazard function given out covariates. This is done with a Cox regression, where we estimate the hazard function as

$$h(t|x_1, \dots, x_p) = h_0(t) \exp(\beta_1 x_1 + \dots + \beta_p x_p), \quad (9)$$

where $h_0(t)$ is the baseline hazard function, given as the hazard when all covariates are zero.

Here we want to use Cox regression with ascites, treatment and sex as categorical covariates and age as a numerical covariate.

	coef	exp(coef)	se(coef)	z	p
factor(sex)1	0.46	1.59	0.13	3.68	0.000236
factor(treat)1	0.04	1.05	0.12	0.38	0.703263
factor(asc)1	0.60	1.83	0.18	3.45	0.000564
factor(asc)2	1.19	3.28	0.18	6.78	1.24e-11
age	0.05	1.05	0.01	7.14	9.26e-13

Tabell 5: Table showing a part of the summary of the Cox regression, with the exponent of the coefficients included.

Table 5 shows the result of the Cox regression. We that once again the severity of ascites and age have significant effects on the hazard, with both increase in severity and age leading to an increase in hazard. Contrary to our logrank test, where sex did not have a significant effect on survival, we see that sex have a significant effect on the hazard, with male patients having a larger hazard.

We can find the hazard ratio for men versus women by observing that

$$\frac{h(t|x_1, \dots, x_{p-1}, x_{sex} = 1 = \text{male})}{h(t|x_1, \dots, x_{p-1}, x_{sex} = 0 = \text{female})} = \exp(\beta_{sex}) = \exp(0.46) = 1.59. \quad (10)$$

This means that the hazard increases with 59% if the patient is male. Since

$$t = \frac{\beta_i}{se(\beta_i)} \quad (11)$$

is close to normally distributed , we can find the 95% confidence interval for this hazard ratio as $\exp(\beta_i) \pm \exp(1.96 \cdot se(\beta_i))$, which gives us the interval

$$CI = (1.241, 2.03). \quad (12)$$

This means that 1 is not in the interval, meaning that the increase in hazard is significant.

From 4 and 5 we can conclude that while ascites and age have a significant effect on the hazard and survival function and age a significant effect on the hazard effect, we see that whether the patient receives prednisone or a placebo have no significant effect on neither the survival function nor the hazard function.

is it?