STK4900 Oblig 2

Daniel Heinesen, daniehei

22. 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

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

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 with. We know that with width as the only covariate, we define the odds as

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

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

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

	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

$$OR = \exp(0.4972) = 1.64. \tag{4}$$

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

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

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 $CI = \exp(\beta \pm 1.96 \cdot se(\beta)) = (1.35, 2.01)$. 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)

We now want to use width, weight, color and spine as covariates. The first two, width and weight, can have a continuous range of values, thus being numerical covariates. Color and spine discrete properties of the crab, and can only take on four and three values respectively. This means that these are categorical covariates. We then try to fit model with each of the covariates.

	Estimate	Std. Error	z value	Pr(> |z|)
Model: Weight				
(Intercept)	-3.6947	0.8802	-4.20	2.70e-05
weight	1.8151	0.3767	4.82	1.45e-06
Model: Width				
(Intercept)	-12.3508	2.6287	-4.70	2.62e-06
width	0.4972	0.1017	4.89	1.02e-06
Model: Color				
(Intercept)	1.0986	0.6667	1.65	0.0994
factor(color)2	-0.1226	0.7053	-0.17	0.8620
factor(color)3	-0.7309	0.7338	-1.00	0.3192
factor(color)4	-1.8608	0.8087	-2.30	0.0214
Model: Spine				
(Intercept)	0.8602	0.3597	2.39	0.0168
factor(spine)2	-0.9937	0.6303	-1.58	0.1149
factor(spine)3	-0.2647	0.4068	-0.65	0.5152

Tabell 3: Summary of the fitted model using each of the covariates.

In table 3 we see one model for each of the covariates. We can see that width and weight have a significance on the presence of satellites with p-values far below 0.05.

For spines there seems to be no difference in the condition of the females spine.

For color, the effects are mostly insignificant, except for color category 4, which corresponds to a dark colouration. It there for seems that if the female crab is dark, contrary to light which is the default color, the odds ratio is $OR = exp(-1.86) = 0.156$, meaning that the odds of a dark female crab having satellites are only 15% of that of a light female crab .

Check if this is interpreted correctly!

So it seems that the most significant covariates are width and weight, and if the female crab is dark or light.

Maybe write more on this?

1.4 d)

We will now try to fit a model with all of covariates used above.

The resulting model can be found in table 4. We see that even though width, weight and one color were significant, none of the covariates are significant anymore. This means that we cant remove any of the covariates with the reason that they are insignificant here. Since we saw that width and weight were significant we can try to use only them – we will ignore the one significant color now, to make the model easier. Table 5 shows the resulting model. But again we see that none of the covariates are significant.

To try to find out why we width and weight are insignificant we try to plot them against eachother. From fig. 1 we see that we can describe the width of the female crabs as a linear function of the weight of the crabs. This means that we have one covariate, weight, that describes

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-8.0650	3.9286	-2.05	0.0401
width	0.2631	0.1953	1.35	0.1779
weight	0.8258	0.7038	1.17	0.2407
factor(color)2	-0.1029	0.7826	-0.13	0.8954
factor(color)3	-0.4889	0.8531	-0.57	0.5666
factor(color)4	-1.6087	0.9355	-1.72	0.0855
factor(spine)2	-0.0960	0.7034	-0.14	0.8915
factor(spine)3	0.4003	0.5027	0.80	0.4259

Tabell 4: Summary of the model fitted with width, weight, color and spine as covariates.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-9.3547	3.5280	-2.65	0.0080
width	0.3068	0.1819	1.69	0.0918
weight	0.8338	0.6716	1.24	0.2145

Tabell 5: Summary of the model fitted with width and weight as covariates.

an other covariate, width, as well as the outcome. This means that we have confounding. This means that we can use only one of these covariates in our model.

Check if all of this is correct, and if we can just use one of them. Also: In the slides confounding can be solved with interaction. Is this the case here?

1.5 e)

Write about Interaction

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

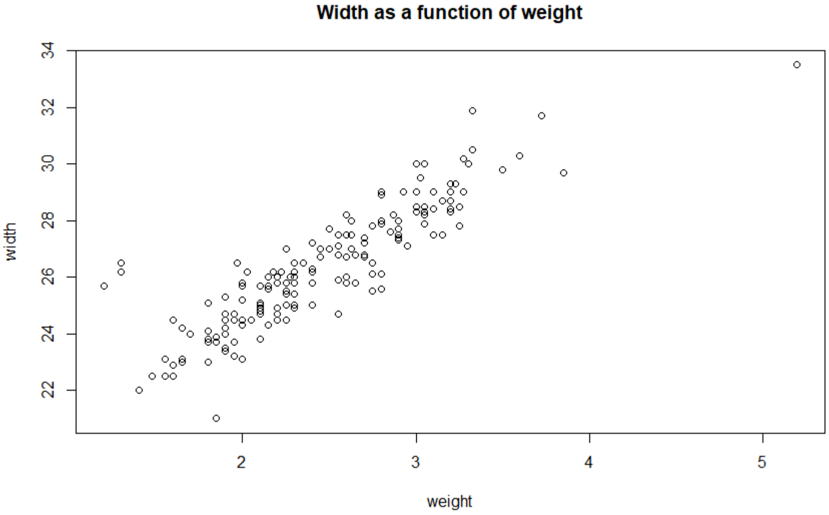
We then define our model as

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

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 $Y_i\ Po(\lambda_i)$ we instead say that they follow the distribution $Y_i\ Po(w_i\lambda_i)$, where w_i is the number of athletes representing the country. But how do use this in our model? We can see this from taking the expected value

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

This means that to compensate for this imbalance of athletes, we can fit out 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.



Figur 1: The figure shows that we can describe the width of the crabs as a linear function of the weight of the crabs.

2.2 b)

We will try to find the best model by first fitting a model with all the variables as covariates.

	Estimate	Std. Error	z value	Pr(>$\|z\|$)
(Intercept)	-2.8623	0.3191	-8.97	< 2e-16
Log.population	0.0275	0.0315	0.87	0.3831
GDP.per.cap	-0.0149	0.0032	-4.65	3.29e-06
Total1996	0.0118	0.0016	7.36	1.79e-13

Tabell 6: Summary of a Poisson regression with all the covariates. $Log.popilation$ 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.

From table6 we see the summary of the fit with all covariate. We see that both GDP per capita and number of medals in 1996 are both significant, with p-values of $p = 3.29 \cdot 10^{-6}$ and $p = 1.79 \cdot 10^{-13}$ respectively, while the logarithm of the population seems no to be significant with $p = 0.38$. If we now fit the model with only the significant covariates we get the model found in table 7. Thus we seem to have found the best model for the amount of models won in the 2000 Olympic games.

	Estimate	Std. Error	z value	Pr(>$\|z\|$)
(Intercept)	-2.5893	0.0576	-44.92	< 2e-16
GDP.per.cap	-0.0158	0.0031	-5.16	2.41e-07
Total1996	0.0128	0.0011	11.25	< 2e-16

Tabell 7: Summary of our first attempt of the best model, with GDP per capita and medals won in the 1996 Olympic games as covariates.

But if we instead try to build this model from the ground up we see something interesting with just athletes and GDP per capita as covariates, we see something else. From tab. 8 we see that without the number of medals from 1996, we see that, contrary to tab. 6, the population is significant, while GDP per capita is not. So something is happening here...

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-4.2551	0.2508	-16.97	< 2e-16
Log.population	0.1796	0.0225	7.99	1.3e-15
GDP.per.cap	-0.0043	0.0027	-1.59	0.111

Tabell 8: Model using logarithm of the population of the nation and GDP per capita as covariates.

I think that this model is more correct than that of tab. 7. This is because the medal count of 1996, have more or less the same dependence of population and GDP per capita as the medal count from 2000. The medal count of 2000 is also very correlated with that of 1996 , which is expected, since a nation has many of the same athletes and causing factors as they had 4 years ago. This means that if we use $Total1996$ as a covariates for the medal count in 2000, the dependency on $Log.population$ is hidden inside $Total1996$, thus making $Log.population$ seem insignificant. Why it enhances GDP per capita, I don't know. So why shouldn't we just $Total1996$? This is like forecasting the weather tomorrow by saying it is the same as today: You will most likely be right, but it doesn't explain anything. So if we ignore $Total1996$, and using tab. 8 we find our final model by noting that GDP per capita is insignificant, and end up tab. 9 as our final model.

Maybe show this dependency?

Show this?

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-4.3462	0.2459	-17.68	< 2e-16
Log.population	0.1821	0.0226	8.07	6.84e-16

Tabell 9: The final model with the logarithm of the nation's population as the only covariate.

This model tells us that given a rise in one $Log.population$ – which is the logarithm of population size per 1000 – we get a *rate ratio*

$$RR = \exp(\beta) = \exp(0.18212) = 1.199758, \tag{8}$$

meaning that the rate of medals won increases by 20%.

This model partly confirm the initial statement from the exercise: Large nations will win more medals, but their wealth is insignificant for the medal count.

3 Problem 3

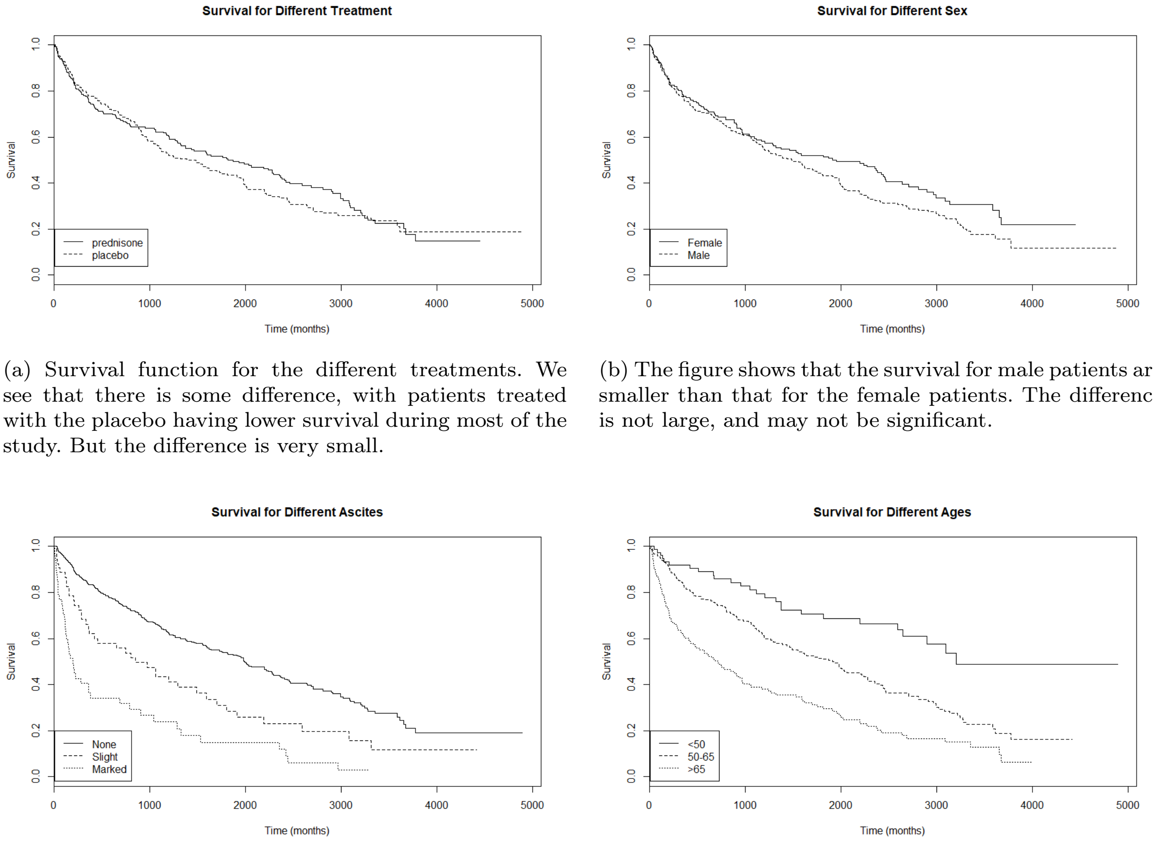
3.1 a)

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. 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.

If we look at figure 2 we see the different Kaplan-Meier plots for the different groups. Fig. 2a 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. 2b 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. 2c, 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 wore decrease if the patient has marked ascites.



(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.

Figur 2: Four figure showing the estimated survival function for different groups.

The same clear difference can be seen in the estimated survival of the groups with different ages, fig. 2d. 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

$$\chi_i^2 = \frac{(O_i - E_i)^2}{se(O_i - E_i)}. \tag{9}$$

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 10: Table showing logrank tests for the different groups.

Table 10 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. 2a.

In the difference between male and female patients, we see in fig. 2b there female patients seemed to have a bit longer than male patients. But from tab. 10 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. 2c and 2d that there seems to be a large difference in the survival between the groups. From tab. 10 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

$$h(t|x_1, ..., x_p) = h_0(t)\exp(\beta_1 x_1 + ... + \beta_p x_p), \tag{10}$$

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

Table 11 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

$$\frac{h(t|x_1, ..., x_{p-1}, x_{sex} = 1 = male)}{h(t|x_1, ..., x_{p-1}, x_{sex} = 0 = female)} = \exp(\beta_{sex}) = \exp(0.46) = 1.59. \tag{11}$$

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

$$t = \frac{\beta_i}{se(\beta_i)} \tag{12}$$

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

is it?

$$CI = (1.241, 2.03). \tag{13}$$

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

From 10 and 11 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.