STK4900, ASSIGMENT 2

Katarina Andersen

February 27, 2022

Problem 1

In this problem we worked with data from a study of horseshoe crabs on an island in the Gulf of Mexico. During spawning season, a female migrates to the shore to breed. With a male attached to her posterior spine, she burrows into the sand and lays clusters of eggs. The eggs are fertilized externally, in the sand beneath the pair. During spawning, other male crabs may cluster around the pair and may also fertilize the eggs. These male crabs are called satellites.

The response outcome for each of the $n = 173$ female crabs is a binary indicator y of whether one or more satellites were present. Explanatory variables are the female crab's color, spine condition, weight and carapace width. The data set was available in the file **crabs.txt** on the course web page. The file consisted of the variables

- **y**, indicator for one or more satellites (0=no, 1=yes)
- **width**, width of carapace of female crab (in cm)
- **weight**, weight of female crab (in kg)
- **color**, color of female crab (1=medium light, 2=medium, 3=medium dark, 4=dark)
- **spine**, conditions of spine (1=both good, 2=one worn or broke, 3=both broken)

1.1 Choosing a suitable regression model for studying how the probability of presence of satellites depends on the explanatory variable width

We are looking at the probability of how much the presence of satellites depends on the width of the carapace of female crabs. y is the indicator for one or more satellites and is a binary outcome, and the width x_i is a predictor for the subject. The probability may thus be given by the logistic regression model:

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

where β_0 and β_1 are the regression coefficients. Plugging this into R using the glm-function, we get

Table 1: Summary of the logistic regression model with width as predictor

	Estimate	**Std. Error**	**z value**	**Pr($>|z|$)**
(Intercept)	-12.3508	2.6287	-4.698	$2.62e-06$
width	0.4972	0.1017	4.887	$1.02e-06$

1.2 Finding the odds ratio of presences of satellites between crabs that differ one cm in width, and the confidence interval for the odds ratio

From equation (1) we can obtain the expression for the odds

$$\begin{aligned} p(x) &= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \\ p(x)\big(1 + \exp(\beta_0 + \beta_1 x)\big) &= \exp(\beta_0 + \beta_1 x) \\ p(x) &= \exp(\beta_0 + \beta_1 x) - p(x)\exp(\beta_0 + \beta_1 x) \\ &= \exp(\beta_0 + \beta_1 x)\big(1 - p(x)\big) \end{aligned}$$

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

If we consider two female crabs, one with a carapace of width x and another with a width of $x + \Delta$, we can find the odds ratio using (2)

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

which is the odds ratio corresponding to one unit's increase in the value of the covariate (in this case, the width). In our case, we are looking to find the odds ratio of presences of satellites between crabs that differ one cm in width, thus $\Delta = 1\ cm$. Plugging this and the value for β_1 obtained from table 1 in equation (3), we get

$$\exp(\beta_1 \Delta) = \exp(0.4972) = 1.644 \tag{4}$$

Meaning that if the carapace of the female crab increases with 1 cm, the odds of having satellites increases by 64%. The odds ratio can be considered as an approximation of the relative risk if $p(1)$ and $p(0)$ are small. Then

$$OR = \frac{p(1)/[1 - p(1)]}{p(0)/[1 - p(0)]} \approx \frac{p(1)}{p(0)} = RR \tag{5}$$

Using equation (1) we find that the value of $p(0)$ is $4.33e-06$ and the value of $p(1)$ is $7.11e-06$, and as they are very small we can conclude that the odds ratio can be a good approximation of the relative risk. To confirm this, we calculate the RR to be 1.6420 which is fairly close to the OR.

We use R to find the confidence interval to be [1.34,2.01]. As 1 is outside of our interval, we can determine that the width increases the probability of presence of satellites significantly.

1.3 Considering the other explanatory variables

Weight and width are both numerical covariates as they are continuous and only measured by one category. Color and spine are separated into four and three categories respectively, where color tells us the color of the crab and the spine variable tells us the condition of the spine, and thus color and spine are categorical covariates. Performing the same analysis as in **1.1**, we get

Table 2: Summary of the logistic regression model for each covariate separately

	Estimate	Std. Error	z value	Pr($>\|z\|$)
Weight				
(Intercept)	-3.6947	0.8802	-4.198	$2.70e-05$
weight	1.8151	0.3767	4.819	$1.45e-06$
Width				
(Intercept)	-12.3508	2.6287	-4.698	$2.62e-06$
width	0.4972	0.1017	4.887	$1.02e-06$
Color				
(Intercept)	1.09865	0.6667	1.648	0.0994
factor(color)2	-0.1226	0.7053	-0.174	0.8620
factor(color)3	-0.7309	0.7338	-0.996	0.3192
factor(color)4	-1.8608	0.8087	-2.301	0.0214
Spine				
(Intercept)	0.8602	0.3597	2.392	0.0168
factor(spine)2	-0.9937	0.6303	-1.577	0.1149
factor(spine)3	-0.2647	0.4068	-0.651	0.5152

From table 2 we see that both weight and width have a $Pr(>|z|)$ value of less than 0.05. This suggests that these covariances have statistically significance on the presence of satellites. Looking at the $Pr(>|z|)$ value for the spine, we see that none of the spine categories show any significance, whilst looking at the color categories we find that the category 4 - dark colored - have some significance. Calculation of the odds ratio, $exp(-1.8608) = 0.155$, shows that the probability of presence of satellites on a dark crab is 15% that of a dark medium crab.

1.4 Using all variables in the regression (as main effects)

In this task we are going to look at a multiple regression model where we use all covariates. The result is given by table 3

Table 3: Summary of multi binary linear regression using all covariates

	Estimate	**Std. Erros**	**z value**	**Pr(>|z|)**
(Intercept)	-8.06501	3.92855	-2.053	0.0401
width	0.26313	0.19530	1.347	0.1779
weight	0.82578	0.7083	1.173	0.2407
Factor(color)2	-0.10290	0.78259	-0.131	0.8954
Factor(color)3	-0.48886	0.85312	-0.573	0.5666
Factor(color)4	-1.60867	0.93553	-1.720	0.0855
Factor(spine)2	-0.09598	0.70337	-0.136	0.8915
Factor(spine)3	0.40029	0.50270	0.796	0.796

In table 2 where we looked at the covariates one by one, we found that weight, width and the color dark were significant factors when looking at the probability of presence of satellites. However, looking at our model where we take all covariates into account we find that none of the variables have a p-value that suggests significance, meaning we cannot determine any variables to being insignificant on the outcome of y.

Table 4: Summary of multi binary linear regression model using weight and width as covariates

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-9.3547	3.5280	-2.652	0.00801
width	0.3068	0.1819	1.686	0.09177
weight	0.8338	0.6716	1.241	0.21445

To look into this further, we can look at a multiple linear regression model with only width and weight as factors. In table 4 we see that neither width nor weight has any significance when it comes to the presence of satellites. In order to figure out why this occurs, we plot these to covariates against each other.

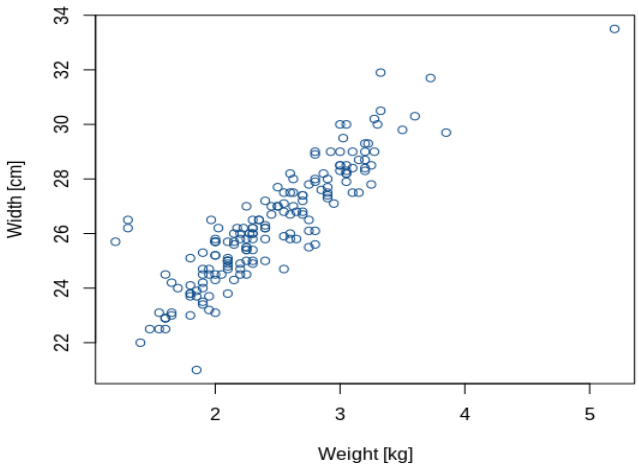


Figure 1: The correlation between width and weight

From figure 1 we have plotted the correlation between width and wight, and we can clearly see that the width can be describe as a function of the weight. Calculations show that the Pearson coefficient is 0.887, suggesting a strong linear correlation, meaning that the width follows the same tendencies as weight or vice versa. This implies that we only need to include one of these factors in our model.

1.5 Investigating whether there are interactions between the covariates

We are now checking to see whether or not there are any interaction between the covariates. In order to do this, we have performed an ANOVA test on different models. The result in 5 shows no indication of significance, thus there is no interactions between the covariates.

Table 5: ANOVA analysis to study interactions between covariates

	Df	**Deviance**	**Resid. Df**	**Resid. Dev**	**Pr(>Chi)**
Null			172	225.76	
Weight	1	30.0214	171	195.74	4.273e-08
Width	1	2.8453	170	192.89	0.09164
factor(spine)	2	0.0941	168	192.80	0.95403
factor(color)	3	7.5958	165	185.20	0.05515
weight:width	1	0.8159	164	184.39	0.36639
width:factor(spine)	2	0.4080	162	183.98	0.81546
width:factor(color)	3	6.6865	159	177.29	0.08259
weight:factor(spine)	2	3.8095	157	173.48	0.14886
weight:factor(color)	3	5.9718	154	167.51	0.11299
factor(spine):factor(color)	6	8.2569	148	159.25	0.21988
weight:width:factor(color)	3	0.1563	145	159.10	0.98431
weight:width:factor(spine)	2	1.5349	143	157.56	0.46419

Problem 2

It is often claimed that participants from larger and wealthier nations are more likely to win medals in competitions like the Olympic games. The file **olympics.txt** on the course web page contains the total number of medals each nation won under the Olympic games in Sydney in the year 2000. Only the 66 nations that won at least one medal in both 2000 and 1996 (in Atlanta) are considered in the file. In addition to the number of medals for each nation in 2000, the file contains information on

- **Total1996**, Number of medals won by the nation in the previous game
- **Log.population**, Logarithm of the nation's population size per 1000@
- **Log.athletes**, Logarithm of the number of athletes representing the nation
- **GDP.per.cap** ,The per capita Gross Domestic Product of the nation

2.1 Develop a Poisson Regression model, including offset term

In this task, we are going to use Poisson regression in order to analyze the outcome of medals in 2000. The outcome of medals in 2000 consists of n independent data where y_i (where $i = 1, 2, \ldots, n$) is the count for subject number i, and x_{ji} is the covariate number j for subject i. Integrating over all y_i, we get the distribution given by

$$Y_i \sim Po(\lambda_i) \tag{6}$$

where the function of covariates, λ_i, is given by

$$\lambda_i = \lambda(x_i, x_{2i}, ..., x_{pi}) = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} \tag{7}$$

But we need to take into account that it is more likely for a larger country with a larger population to win medals than for a smaller country with smaller population. We may thus have data on aggregated form where we record counts for groups of individuals who share the same covariate value. Our observations are aggragated counts, and integrating over our observations we get the distribution to be on the form

$$Y_i \sim Po(w_i \lambda_i) \tag{8}$$

where the weight w_i is the number of subjects in i. Combining (8) with (6), we get

$$\begin{aligned} E(Y_i) &= w_i \lambda_i \\ &= w_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}) \\ &= \exp(\log(w_i) + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}) \end{aligned}$$

$\log(w_i)$ is the offset, where the regression coefficient is equal 1. As we already have $\log .athletes$ in our dataset, we will us this as the offset in our regression model.

2.2 Fitting a model for the rate of medals won per athlete, using (possibly only some of) the predictors above

In this task, we will attempt to find the best fitted model for our data. We will start by doing a Poisson regression analysis with all the covariates. The result is given in table 6

Table 6

	Estimate	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	-2.862299	0.319076	-8.971	$< 2e-16$
Total1996	0.011832	0.001607	7.364	$1.79e-13$
Log.population	0.027510	0.031539	0.872	0.383
GDP.per.cap	-0.014924	0.003208	-4.652	$3.29e-06$

From our first model shown in table 6, we observe that the covariate **Total1996** is the most significant with a p-value of $< 1.79e-13$, followed by **GDP.per.cap** with a p-value of $3.29e-06$. **Total1996** is the number of medals won by the nation in the previous game and **GDP.per.cap** is the per capita Gross Domestic product. It makes sense that these covariates are significant. Many of the athletes competing in the 1996 Olympics will also compete in the 2000 Olympics, and usually the countries with higher GDP have higher budget thus better chances of winning.

To see what else may affect the rate of medals won by athletes, we will now try a regression model without the **Total1996** covariate. Even though it shows high significance, it is not very statistically important in our case. Our new Poisson regression analysis is given by

Table 7

	Estimate	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	-4.255144	0.250782	-16.968	<2e-16
Log.population	0.179605	0.022466	7.995	1.3e-15
GDP.per.cap	-0.004340	0.002726	-1.592	0.111

In table 7 we now see that **Log.population** is more significant than **GDP.per.cap**. We thus remove the **GDP.per.cap** and end up with:

Table 8

	Estimate	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	-4.34619	0.24585	-17.678	<2e-16
Log.population	0.18212	0.02256	8.073	6.84e-16

Looking at table 8 we now see that the significance of **Log.population** has increased even more,

which is no surprise. If we calculate the risk ratio

$$RR = \exp(\beta_i) = \exp(0.18212) = 1.199 \tag{9}$$

We find that by increasing **Log.population** by 1 gives an increase in medals won per athletes of about 20%. Thus countries with large population have a higher chance of achieving medals. Though, we may wonder if this is the best model as countries such as India have a high population, but doesn't necessarily win more medals. We have to consider that the likelihoods of winning more medals also depend on the wealth of the country and how much the country is willing to invest in sports.

Problem 3

488 patients with liver cirrhosis at various hospitals in Copenhagen were included in a randomized clinical trial with several years of follow up. The purpose of the study was to investigate whether patients treated with the hormone prednisone had better survival than patients who got an inactive placebo treatment. 251 of the patients received prednisone while 237 received placebo. The file **cirrhosis.txt** contained the data for the study. The data were organised with one line for each of the 488 patients who took part in the study, and with the following variables in the seven columns:

- **status**, Indicator for death/censoring (1=dead; 0=censored)
- **time**, Time in days from start of treatment to death/censoring
- **treat**, Treatment (0=prednisone; 1=placebo)
- **sex**, Gender (0=female; 1=male)
- **asc**, Ascites at start of treatment (0=none; 1=slight; 2=marked)
- **age**, Age in years at start of treatment
- **agegr**, Age group (1=<50; 2=50-65; 3=>65)

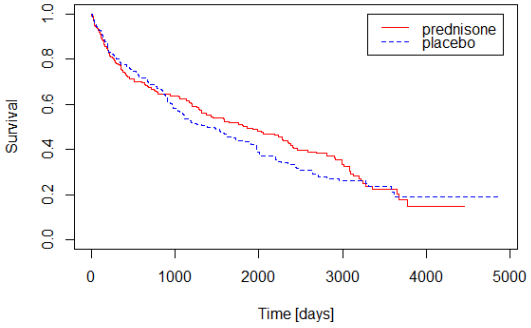
In this problem, we studied the effect of treatment with prednisone, sex, age, and ascites

3.1 Making Kaplan-Meier plots for the survival function for each level of the covariates treatment, sex, ascites, and grouped age

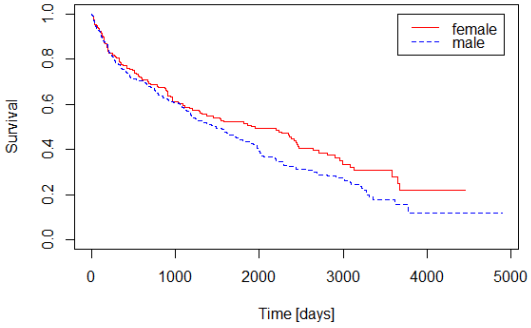
In this task we are going to study the effect of treatment with prednisone, sex, age, and ascites. We start of by making Kaplan-Meier plots for the survival function for each level of the covariates treatment, sex, ascites, and grouped age.

For figure 2a we see that the estimated survival function for people being treated with prednisone and placebo are somewhat the same, with prednisone being more effective in the earlier stages. However, the stabilization of the survival function for people being treated with placebo is slightly higher than that of people being treated with prednisone.

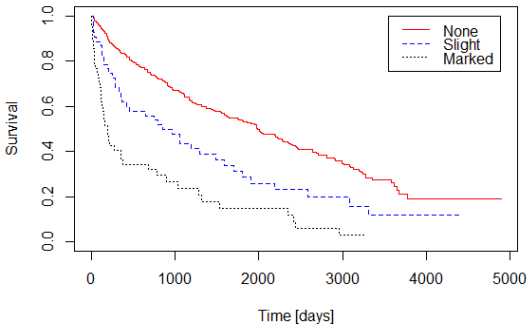
Looking at 2b we see that the estimated survival of male and female are approximately the same in the earlier stages of the function, but as time goes there develops a distinction where females have a higher survival probability.



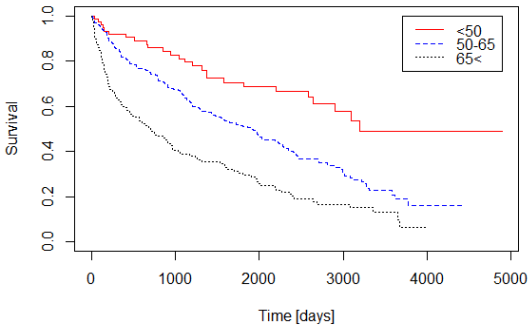
(a) Survival of different treatments



(b) Survival for male and female patients



(c) Survival for different ascites



(d) Survival for different ages

Figure 2: Four figures showing the estimated survival function for different groups.

The figure for the survival for different ascites, 2c, shows that the excess fluid in the abdomen seems to have a large impact on the survival of the patients. We see clear distinctions early on in the case study, where having no ascites has higher survival than that of slight and marked ascites. This trend is carried out throughout the case study.

Figure 2d shows the survival for different ages. Here we clearly see that the survival of patients that are under the age of 50 is much higher than that of the two other age groups. The survival of patients at ages between 50-65 has higher survival than that of patients over 65.

3.2 Using the logrank test for each of the covariates to investigate if the covariate has a significant effect on survival

In this part of the task, we are going to study the differences plotted above using the logrank test. We start the logrank test by making a null hypothesis, stating that the survival function is same for all treatment groups.

$$H_0 : S_1(t) = S_2(t) \ \big(= S_3(t) \big), \ \text{for all } t \tag{10}$$

Table 9

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Treat					
treat=0	251	142	149	0.355	0.728
treat=1	237	150	143	0.371	0.728
Chisq= 0.7 on 1 degrees of freedom, p= 0.4					
Sex					
sex=0	198	111	127	2.00	3.55
sex=1	290	181	165	1.54	3.55
Chisq= 3.5 on 1 degrees of freedom, p= 0.06					
Ascites					
acs=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	56.17	59.60
Chisq= 69.9 on 2 degrees of freedom, p= 7e-16					
Age					
agegr=1	80	26	58.7	18.18	22.87
agegr=2	250	148	163.0	1.21	2.72
agegr=3	158	118	71.3	30.51	40.87
Chisq= 50.6 on 2 degrees of freedom, p= 1e-11					

We then test this null hypothesis by comparing the observed (O_i) and expected (E_i) number of events where H_0 is fulfilled in the two groups. Proceeding by using these values in the test statistics χ^2 which is given by

$$\chi_i^2 = \frac{(O_i - E_i)^2}{se(O_i - E_i)^2} \tag{11}$$

By comparing K groups, we get a test statistics with $K - 1$ degree of freedom. The result we get from performing this test is shown in table 9.

Firstly, we look at the difference in survival of different treatments, placebo and prednisone. The $\chi^2 = 0.7$ and $p = 0.4$. With a fairly low value of χ^2 and a large value of p, we can not conclude whether or not the prednisone increases or decreases the survival function. This comes as no surprise as it was we concluded just about the same thing while looking at our figure.

Recall that the figure 2b showing the survival for male and female patients showed a slightly higher survival function for female. The logrank test shows a χ^2-value of 3.5 and a p-value of 0.06, which means that the p-value is too large for us to conclude if the sex is significant or not for survival.

The figure for survival of different ascites, 2c, and the figure for different ages, 2d, showed a bigger difference between the groups. The logrank test for the ascites gives us $\chi^2 = 69.9$ and $p = 7e - 16$, and the logrank test for the ages gives us $\chi^2 = 50.6$ and $p = 1e - 11$. Both these logrank tests gives us a high value of χ^2 and a low value of p, thus we conclude that there is a significant difference between the ascites groups and the age groups.

3.3 Making a multiple Cox regression where the effects of all the covariates are studied simultaneously

Doing a multiple hazard regression in the form of a multiple Cox regression in R, we obtain the result given in table 10

Table 10

	coef	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**
factor(sex)1	0.461877	1.587050	0.125631	3.676	0.000236
factor(treat)1	0.044818	1.045837	0.117657	0.381	0.703263
factor(asc)1	0.603507	1.828520	0.175019	3.448	0.000564
factor(asc)2	1.187254	3.278068	0.175224	6.776	1.24e-11
age	0.048877	1.050091	0.006844	7.141	9.26e-13

In contrast to the logrank test, the multiple Cox regression shows that the difference in sex is significant with a p-value of 0.000236. From the table we see again that both ascites and ages have significant differences between the groups, which was the same conclusion we obtained from studying the figures and performing the logrank test.

We can also find a 95 % confidence interval for the hazard ratio for men versus women. We start

by finding the hazard ratio

$$HR = \frac{h(t|x_1 + \Delta, x_2, \ldots, x_p)}{h(tx_1, x_2, \ldots, x_p)} = \exp \beta_{sex} = \exp(0.461877) \approx 1.59 \tag{12}$$

This means that the hazard ratio increases by 59% if the sex of the patient is male. We can then find the 95% confidence interval by plugging this into $\exp(\beta_i \pm 1.96 \cdot se(\beta_i))$, or by reading of table 11.

Table 11

	exp(coef)	**exp(-coef)**	**lower .95**	**upper .95**
factor(sex)1	1.587	0.6301	1.2407	2.030
factor(treat)1	1.046	0.9562	0.8305	1.317
factor(asc)1	1.829	0.5469	1.2975	2.577
factor(asc)2	3.278	0.3051	2.3252	4.621
age	1.050	0.9523	1.0361	1.064

The 95% confidence interval for the hazard ratio is thus given by

$$CI = [1.2407, 2.030] \tag{13}$$

As 1 is not within our confidence interval, we conclude that the hazard is significant.

We clearly see that different methods of analyzing the same data gives different results. By comparing the results of the result from the different analyzing methods, we can maybe get a clearer image of the data.

Code for problem 1

```
# Reading in the data from the website:

crabcake = read.table('https://www.uio.no/studier/emner/matnat/math/STK4900/data/crabs.txt',header = TRUE)
summary(crabcake)

# Setting a binary logistic regression model

fit.width = glm(y~width, data = crabcake, family = binomial)
summary(fit.width)

# Odds ratio

delta = 1
OR = exp(fit.width[["coefficients"]]["width"]*delta)

# Finding confidence interval
expcoef=function(glmobj)
{
  regtab=summary(glmobj)$coef
  expcoef=exp(regtab[,1])
  lower=expcoef*exp(-1.96*regtab[,2])
  upper=expcoef*exp(1.96*regtab[,2])
  cbind(expcoef,lower,upper)
}
expcoef(fit.width)


fit.multi = glm(y~width+weight+factor(color)+factor(spine), data = crabcake, family =binomial)
summary(fit.multi)

# Grouped logistic fit for width and weight as the significant covariates
fit.multiww = glm(y~width+weight, data = crabcake, family = binomial)
summary(fit.multisig)

# Checking the correlation between width and weight
plot(crabcake$weight, crabcake$width, xlab = "Weight [kg]", ylab = "Width [cm]", col = 'dodgerblue4')
cor(crabcake$weight, crabcake$width)

# Checking interaction between the covariates

crabcake.fit.interaction = glm(y~weight + width + width:weight + factor(spine) + factor(color) + width:factor(spine) + width:factor(color) + weight:factor(spine) + weight:factor(color) + factor(spine):factor(color) + weight:width:factor(color)+weight:width:factor(spine), data = crabcake, family = binomial)
anova(crabcake.fit.interaction,test = "Chisq")
```

Code for problem 2

```
# Reading in the data
olympic=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/data/olympic.txt",sep="\t",header=TRUE)

# Making a fit for all the covariates
fit.1 = glm(Total2000~offset(Log.athletes) + Total1996 + Log.population + GDP.per.cap,data=olympic,family=poisson)
summary(fit.1)

# Making a fit without 1996
fit.2 = glm(Total2000~offset(Log.athletes)+Log.population + GDP.per.cap,data=olympic,family=poisson)
summary(fit.2)

# Making a fit without GDP and 1996
fit.3 = glm(Total2000~offset(Log.athletes)+Log.population,data=olympic,family=poisson)
summary(fit.3)

# Computing rate ratio
exp(fit.3$coefficients)
```

Code for problem 3

```
## PROBLEM 3 ###

### A ###

#read data
df3 = read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/data/cirrhosis.txt", header=TRUE)
#load library
library(survival)

surv_treat = survfit(Surv(time,status)~treat, data = df3, conf.type="none")
summary(surv_treat)
plot(surv_treat,lty=1:2, xlab="Time [days]", ylab = "Survival", col=c('red','blue'))
legend(3300,1,legend=c('prednisone', 'placebo'), col=c('red','blue'),lty=1:2)


surv_sex = survfit(Surv(time, status)~sex, data=df3, conf.type="none")
summary(surv_sex)
plot(surv_sex, lty=1:2, xlab="Time [days]", ylab = "Survival", col=c('red','blue'))
legend(3650,1,legend=c('female', 'male'), col=c('red','blue'),lty=1:2)

surv_ascites = survfit(Surv(time, status)~asc, data=df3, conf.type="none")
summary(surv_ascites)
plot(surv_ascites, lty=1:3, xlab="Time [days]", ylab = "Survival", col=c('red','blue','black'))
legend(3550,1,legend=c('None', 'Slight', 'Marked'), col=c('red', 'blue','black'),lty=1:3)

surv_agegr = survfit(Surv(time, status)~agegr, data=df3, conf.type="none")
summary(surv_agegr)
plot(surv_agegr, lty=1:3, xlab="Time [days]", ylab = "Survival",col=c('red','blue','black'))
legend(3700,1,legend=c('<50', '50-65', '65<'), col=c('red','blue','black'), lty=1:3)

### B ###

survdiff(Surv(time, status)~treat, data=df3)
survdiff(Surv(time, status)~sex, data = df3)
survdiff(Surv(time, status)~asc, data = df3)
survdiff(Surv(time, status)~agegr, data=df3)
```