Compulsory assignment number 2 spring 2020 STK4900/9900: Statistical methods and applications

Victoria Jensen

Department of Mathematics

UNIVERSITETET I OSLO

April 2020

Problem 1

The exercise studies the horseshoe crabs on an island in the Gulf of Mexico. The motivation of the study is to investigate how different features of the female horseshoe crab affects the number of the present satellites. The satellites are male crabs from outside that may fertilized the female crabs' eggs during spawning. There are the following variables considered in the study:

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)

a) Choose a suitable regression model for studying how the probability of presence of satellites depends on the explanatory variable width. Give the reasons for your choice of regression model.

```
rm(list=ls(all=TRUE))

# Reading datafiles

crabs <-
read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/data/crabs.txt",heade
r=T)

summary(crabs)

# Choose a binary logistic regression model

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

summary(fit.width)
```

The goal is to find the effect of the covariate width. The most suitable model for analysis here is binary logistic model. Response variable with value 0 and 1 supports this decision as far as concern the choice of the model. Generally if the data set has a form $(x_1, y_1),\ldots,(x_n, y_n)$ with y as a binary response variable with value (0 or 1) for subject i, and if x_1 is a numerical or factor variable, the model of the logistic regression can be written in the following form:

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

where p(x) is the probability of response, the β_0 and β_1 are the regression coefficients or effect variables and x is a covariate of interest. The model shows ”S-shaped” relationships between p(x) and x. After reading the data and producing the summary of the data, R gives the following fit summarized in table 1.

Table 1. Results from the binary logistics regression.

```
Coefficients:
            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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

b) Find the odds ratio of presences of satellites between crabs that differ one cm in width and explain what this odds ratio means. Can the odds ratio be considered as an approximation to a relative risk in this situation? Also, find a confidence interval for the odds ratio and determine whether width influences presence of satellites significantly.

```
# Computing the odd 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)
```

The odds ratio of presences of satellites between crabs that differ with 1 cm in width can be calculated by the following formula:

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

The odds ratio, OR, between two subjects with covariate values x and x + Δ, is calculated by the following formula:

$$OR = \frac{\exp(\beta_0 + \beta_1(x+\Delta)}{1 + \exp(\beta_0 + \beta_1 x)} = \exp(\beta_1 \Delta).$$

The odds ratio exp(β1) corresponds to one unit increase in the value of the covariate. In our case, we want the odds ratio for a Δ = 1 cm increase in width. This is then given as $\exp(\beta_1\Delta) = \exp(\beta_1) = \exp(0.4972) = 1.644$, which corresponds to exactly one unit increase in width. This means that the odds of a female crab having satellites, increases by ≈ 64% if the width of the female increases by 1 cm. The table 2 shows the results from the running the regression.

Table 2. Results for variable width.

	expcoef	lower	upper
(Intercept)	4.326214e-06	2.503215e-08	0.0007476835
width	1.644162e+00	1.346931e+00	2.0069822897

The odds ratio can be approximated to the relative risk RR by the following way if p(1) and p(0) are very small: $RR = \frac{p(1)}{p(0)}$,

For our fit, we find that for x = 1, we get $p(1) = 7.12\times10^{-6}$, and for x = 0 we get $p(0) = 4.326\times10^{-6}$. These are very small numbers, suggesting that the approximation should hold. Computing the relative risk, RR = 1.644, shows that the approximation is exact for the model. The 95 % CI for the odds ratio via R gives the interval: [1.346,2.006].

c) Then consider the other explanatory variables weight, colour and spine as covariates, one at a time. Discuss whether these covariates should be included as categorical or numerical. Determine which variables has a significant influence on the presence of satellites.

```
# c. Binary logistic regression model for the other covariates one by one
fit.weight=glm(y~weight , data=crabs ,family=binomial)

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

fit.weight=glm(y~weight , data=crabs ,family=binomial)

fit.color=glm(y~factor(color), data=crabs ,family=binomial)

fit.spine=glm(y~factor(spine), data=crabs ,family=binomial)

summary(fit.width)

summary(fit.weight)

summary(fit.color)

summary(fit.spine)

# Grouped logistic fit for all significant covariates

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

summary(fit.multi)
```

Now it needs to undertake the further analysis using other covariates. The predictors width and weight are numerical covariates as they are continuous. Both are statistically significant and that's why can be used to explain variation in indicator (variable y). The color and spine, however, are discrete covariates as they can only take finite values. Thus, they are categorical variables, and we need to include those as factor variables in regression analysis. The covariate responsible for spine, needs to check whether both spines are functioning or if both are broken, or if one of them are working. Therefore, they take three definite values. The color variable is hard to quantify as color is a continuous spectrum. Therefore, the color variable will be grouped into 4 groups to see whether color is a significant factor for the regression. The color variable also could be numerical normalized to, for instance, the brightest and darkest crabs observed. Grouping covariates can bring the model better fit – what we are going to test via a deviance test.

Table 3. Summary of the binary logical regression models for each covariate separately. Width.

```
            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 ***

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1977  -0.9424   0.4849   0.8491   2.1198
> summary(fit.weight)
```

Table 4.Summary of the binary logical regression models for each covariate separately. Weight.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
weight        1.8151     0.3767   4.819 1.45e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 195.74  on 171  degrees of freedom
AIC: 199.74

Number of Fisher Scoring iterations: 4
```

Table 5. Summary of the binary logical regression models for each covariate separately. Color.

```
> summary(fit.color)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.0986     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 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 212.06  on 169  degrees of freedom
AIC: 220.06

Number of Fisher Scoring iterations: 4
```

Table 6 Summary of the binary logical regression models for each covariate separately. Spine.

```
> summary(fit.spine)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
```

```
Residual deviance: 223.23  on 170  degrees of freedom
AIC: 229.23

Number of Fisher Scoring iterations: 4
```

Table 8. Grouped logistic fit for all significant covariates.

```
Coefficients:
               Estimate Std. Error 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    0.70383   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   0.4259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

d) Next use all variables in the regression (as main effects) and describe your findings. Try to simplify the model only using the significant covariates. Discuss the covariates weight and width.

```
# d. Grouped logistic fit for all significant covariates

fit.multisig=glm(y~width+weight , data = crabs , family = binomial)

summary(fit.multisig)
```

Now we will fit a multiple logistic regression model where we use all covariates instead of each of them separately. The resulting model can be seen in tables 3-6.

From table 7 it is observed that such model with all covariates included, results in none of the covariates being significant anymore.

Table 7. The model with all covariates included

```
Coefficients:
               Estimate Std. Error 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    0.70383   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   0.4259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 185.20  on 165  degrees of freedom
AIC: 201.2
```

This is perhaps a result of overfitting as we include covariates which is known mostly to be non-significant. Therefore, another model can be constructed using the multiple logistic regression but including only two statistically significant covariates covariates responsible for width and weight. The result from the regression is summarized in the table 8.

Table 8. The model with the grouped logistic fit for all significant covariates

```
Coefficients:
            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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 192.89  on 170  degrees of freedom
AIC: 198.89

Number of Fisher Scoring iterations: 4
```

In the new model the width and weight are not significant. In order to investigate why the high significance levels of the two covariates width and weight are lost when combined, we need to consider relations between them. This can be done by plotting the width of the female horseshoe crabs as a function of their weight. The figure 1 shows the results from the plotting.

```
# Checking correlation between width and weight

plot(crabs$weight , crabs$width , xlab = "Weight [kg]", ylab = "Width [cm]")

cor(crabs$weight , crabs$width)
```

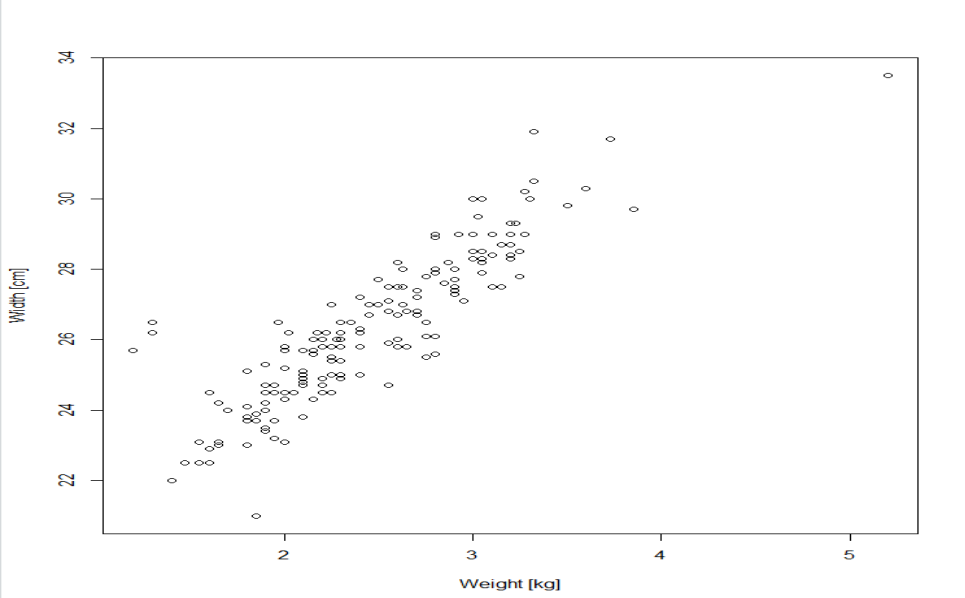


Figure 1. Plotting the relations between weight and width.

The figure 1 shows a strong linear correlation between the two covariates. A heavier crab means a wider crab, as expected. The Pearson correlation coefficient between the two predictors are 0.887 meaning that the correlation is highly linear.

```
# Checking correlation between width and weight
plot(crabs$weight , crabs$width , xlab = "Weight [kg]", ylab = "Width [cm]")
cor(crabs$weight , crabs$width)
```

Therefore, we can conclude that the covariate width is a confounder for weight, meaning that we only need to include one of these variables in our model. The final model is then constructed using either the width or weight. We choose the width, as it might be the" visual" covariate out of the two which the horseshoe crabs might base their decisions of. The model is summarized in table 8, in which we see that the significance of the covariate is back.

Table 8. The new model with only one covariate responsible for the width.

```
Coefficients:
            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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
```

```
AIC: 198.45
```

e) Finally investigate whether there are interactions between covariates.

```
# Fitting the final model
fit.multisig=glm(y~width , data = crabs , family = binomial)
summary(fit.multisig)
```

This is also useful to check whether there are any significant interactions between the chosen covariates. This can be done via finding if two binary predictors have the causal effect on the outcome. Table 9 shows results from the regression with interactions between covariates. The weight and width are confounding covariates as was shown before. By fitting a model which includes interactions for the pair combination of spine and color with width, the results shows that there are in fact no significant interactions between the covariates in the analysis.

Table 9. Results from the regression with interactions between covariates

```
Coefficients:

                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                    1.2528     0.8018   1.562    0.118
factor(color)2                -0.5596     0.9112  -0.614    0.539
factor(color)3                15.3133  1385.3780   0.011    0.991
factor(color)4               -17.8188  2399.5449  -0.007    0.994
factor(spine)2                15.3133  1696.7345   0.009    0.993
factor(spine)3               -17.8188  2399.5449  -0.007    0.994
factor(color)2:factor(spine)2 -16.5173  1696.7347  -0.010    0.992
factor(color)3:factor(spine)2 -31.8794  2190.4750  -0.015    0.988
factor(color)4:factor(spine)2 -15.3133  3794.0134  -0.004    0.997
factor(color)2:factor(spine)3  18.4728  2399.5449   0.008    0.994
factor(color)3:factor(spine)3   1.5247  2770.7557   0.001    1.000
factor(color)4:factor(spine)3  33.7659  3393.4688   0.010    0.992

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 196.87  on 161  degrees of freedom
AIC: 220.87
```

Problem 2

The exercise analyses the data from the 1996 and 2000 Olympics games. The motivation is to analyse if participants from larger and wealthier nations have a larger probability to win any

medals. The data set consists of 66 nations that get at least one medal in OL. There are following variables to be considered:

Total2000: Number of medals won by the nation in the Olympics of 2000

Total1996: Number of medals won by the nation in the Olympics of 1996

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 nation

a) Since the outcome total medals in 2000 is a count variable it might be reasonable to analyze the data by Poisson regression. Present and explain such a model. Often Poisson regression models include offset terms. Explain why Log.athletes is a sensible choice for an offset. In the following analyses you should include Log.athletes as offset.

It needs to build the model for data analysing. The outcome of the study is a number of medals earned by a nation, which makes the outcome a count variable. When working with count variables Poisson regression is best to use for analysis. Thus, the count is distributed as Y ~ Po(λ), where the parameter λ is the rate. It is a multicovariate model, meaning that the parameter λ is not necessarily the same for all subjects. The model then becomes $Y_i \sim P_0(\lambda_i)$ with a rate in the form

$$\lambda_1 = \boldsymbol{\lambda}\,(x_{1i}, x_{2i}, \ldots, x_{pi}) = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}).$$

If one works with aggregated counts, an observation y_i is a realization of $Y_i \sim P_0\ (w_i\lambda_i)$ with weight w_i as a number of subjects in group i. The weight covariate must be included into the model because we know that countries with higher athletes' amount have a higher probability to win. Then the model that calculate the count expectation will be as follows:

$$\mathrm{E}(Y_i) = w_i\lambda_i = w_i\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}) = \exp\big(\log(w_i) + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\big)$$

where the term $\log(w_i)$ is a covariate where the regression coefficient is equal to 1. This term is also called the offset and adding it to the model will compensate for the imbalance if athletes' number. A natural choice for such covariate would be the Log.atheletes.

```
# a. Reading in 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 covariates

fit.1 = glm(Total2000~offset(Log.athletes) + Total1996 + Log.population + GDP.per.cap ,data=olympic ,family=poisson)

summary(fit.1)

# Checking correlation between covariates

plot(olympic)

cor(olympic$Total2000 , olympic$Total1996)
```

Table 10 shows the summary from the Poisson regression model using all covariates with Log athletes as the offset. The covariates Total1996 and GDP per cap are strongly statistically significant and can be used to explain variation in y.

Table 10. Summary of the Poisson regression model, using all covariates with Log.athletes as the offset.

```
Coefficients:
                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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 254.11  on 65  degrees of freedom
Residual deviance: 131.63  on 62  degrees of freedom
AIC: 392.31

Number of Fisher Scoring iterations: 4
```

```
# Checking correlation between covariates

plot(olympic)

cor(olympic$Total2000 , olympic$Total1996)

# 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)
```

Table 11 shows the summary from the Poisson regression model, using a model without the covariate Total1996. Without the covariate Total1996 the Log.population covariates becomes statistically significant and GDP.per.cap not statistically significant.

Table 11. Summary of the Poisson regression model, using a model without the covariate Total1996.

```
Coefficients:
                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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 254.11  on 65  degrees of freedom
Residual deviance: 187.80  on 63  degrees of freedom
AIC: 446.48

Number of Fisher Scoring iterations: 4
```

Table 12 shows summary of the Poisson regression model using covariate Log.population. With only one covariate this covariate fully explains variation in y.

Table 12. Summary of the Poisson regression model, using only Log. Population as covariate.

```
Coefficients:
               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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 254.11  on 65  degrees of freedom
Residual deviance: 190.35  on 64  degrees of freedom
AIC: 447.03

Number of Fisher Scoring iterations: 4
```

b) Fit a model for the rate of medals won per athlete, using (possibly only some of) the predictors above. Explain how you arrived at your final model. Give an interpretation of the results and write a summary of your findings. Do you confirm the statement above?

Now we will start to fit the Poisson regression model with the covariates Total 1996, Log.population and GDP.per.cap. with Log.athletes as the offset for better model fitting using the available data set (Table 10). In the output the Total1996 and GDP. per cap both statistically significant and Total1996 is even more significant covariate than GDP.per.cap. The fact that the results of the previous Olympics appears to be the most important factor comes as no big surprise. Many of the previous contenders which won, are likely to compete again. The correlation between Total1996 and Total2000 is 0.967, which means that the chosen model is highly multicollinear. This can cause problems when trying to interpret the results. In addition, using medal winners from previous years as the dominating covariate will not help to learn about statistical importance. Therefore, we will try a model without Total1996 to see what the significance of the other covariates are in this case. The table 11 shows that the Log.population is the most significant covariate, while GDP.per.cap is no longer significant. By removing the insignificant GDP.per.cap, we end up with the final model seen in the table 12. The rate ratio, RR, for this model is RR = $\exp(\beta_j)$ = exp(0.182) = 1.997, meaning that the rate of medals won by a nation increases by ≈ 18,2% for one unit increase in Log.population.

The findings confirm that if the motivation is to predict the nation which would gather most medals the next Olympics, the best source would be to check the last year's winners. However, if the goal is to check out the winners of an arbitrary Olympics without the knowledge of the previous winners, the best bet would be to look at the population. The country data such as India with one of the highest Log.pop values, won only1 medal both in 2000 and 1996. This means that the model prediction considered only on the population can be very wrong for some scenarios. There are probably some covariates which are also significant such as the public information and the levels of education in a nation. To conclude, it seems that the wealth of the country does not matter much when it comes to their medal count. Large countries, however, are more likely to bring home more medals.

Problem 3

This report a randomized clinical trial is considered to analyse. The dataset consists of 488 patients with liver cirrhosis at various hospitals in Copenhagen who have been included in the test. The motivation is to find out what effects the provided hormone prednisone had vs a placebo treatment. This can be done via comparation of survival rate for treatment groups (251 patients) and the control group who received a placebo treatment (237 patients). The explanatory variables of pasients are sex, age, and ascites (excess fluid in the abdomen).

The variables to be considered are:

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)

a) Make Kaplan-Meier plots for the survival function for each level of the covariate's treatment, sex, ascites, and grouped age (so 4 plots in total). Discuss what the plots tell you.

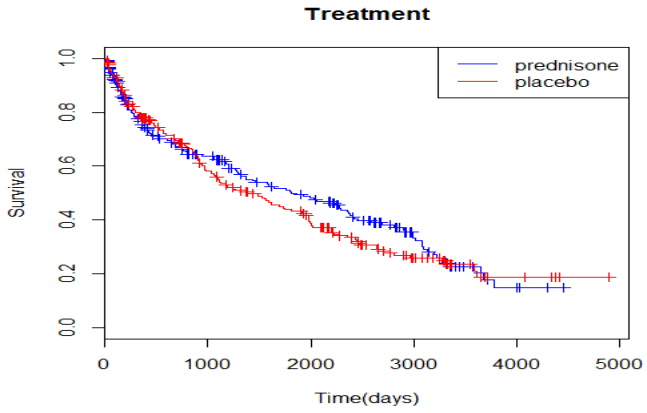
```
#a. Reading in data 2

cirrhosis <-
read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/data/cirrhosis.txt",hea
der=TRUE)

library(survival)

survpred <- survfit(Surv(time,status)~1, conf.type="none")

summary(survpred,xlab="Time(days)",ylab="Survival")

plot(survpred)

survpred1 <- survfit(Surv(time,status)~treat, conf.type="plain")

summary(survpred1)

plot(survpred1,col=c('blue', 'red'),xlab="Time(days)",ylab="Survival")

legend('topright', c("prednisone","placebo"), col=c('blue', 'red'), lty=1)

title("Treatment")

survpred2 <- survfit(Surv(time,status)~sex, conf.type="plain")

summary(survpred2)

plot(survpred2,col=c('blue', 'red'),xlab="Time(days)",ylab="Survival")

legend('topright', c("Male","Female"), col=c('blue', 'red'), lty=1)

title("Sex")

survpred3 <- survfit(Surv(time,status)~asc, conf.type="plain")

summary(survpred3)

plot(survpred3,col=c('blue', 'red', 'green'),xlab="Time(days)",ylab="Survival")

legend('topright', c("none","slight", "marked"), col=c('blue', 'red', 'green'), lty=1)

title("Ascites")

survpred4 <- survfit(Surv(time,status)~agegr, conf.type="plain")

summary(survpred4)

plot(survpred4,col=c('blue', 'red', 'green'),xlab="Time(days)",ylab="Survival")

legend('topright', c("<50","50-65", ">65"), col=c('blue', 'red', 'green'), lty=1)

title("Age group")
```

We start by looking at the survival of the studied patients by plotting the covariate treat (prednisone or placebo), sex (female or male), asc (the severity of built up abdomen fluid) and agegr (age group) into the Kaplan-Meier plots of the the survival function for the variables treat,

sex, asc and agegr. They are useful for analysis of the differences by the levels of each covariate. They are normalized on the y-axis, such that they start at a survival of 1.0, e.g. all patients' life. The x-axis is the time from the beginning of treatment until death/censoring. The figure 2 shows the plot result. The curves show the survival of patients under treatment of prednisone and placebo. They end in almost the same manner with more deaths by patients treated with prednisone.

```
# Plotting the Kaplan -Meier estimate for each fit
plot(fit.treat ,lty=1:2,xlab="Time [days]",ylab="Survival")
legend(3000,1,c("prednisone","placebo"),lty=1:2)
plot(fit.sex,lty=1:2,xlab="Time [days]",ylab="Survival")
legend(3300,1,c("Female","Male"),lty=1:2)
plot(fit.asc,lty=1:3,xlab="Time [days]",ylab="Survival")
legend(3300,1,c("None","Slight","Marked"),lty=1:3)
plot(fit.agegr ,lty=1:3,xlab="Time [days]",ylab="Survival")
legend(3500,1,c("< 50","50-65",">65"),lty=1:3)
```

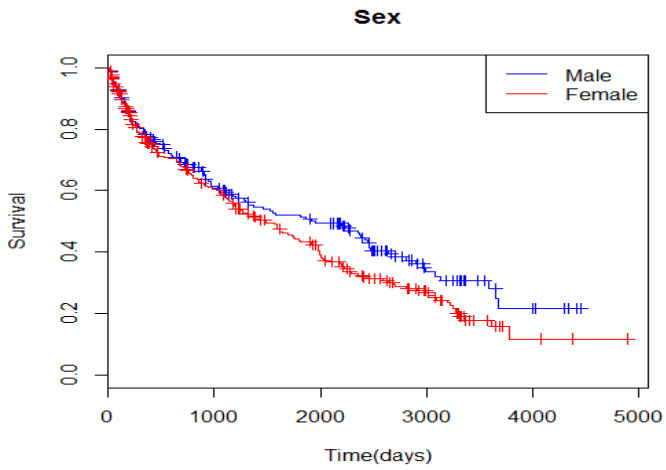
The figure 2a shows the treatment survival function where it is observed that the patients with the prednisone treatment have slightly lower survival during the first 1000 days compared to those with the placebo treatment. For most of the remaining time, however, the prednisone treatment seems to increase the survival of the patients.



(a) Survival function for the two treatments.

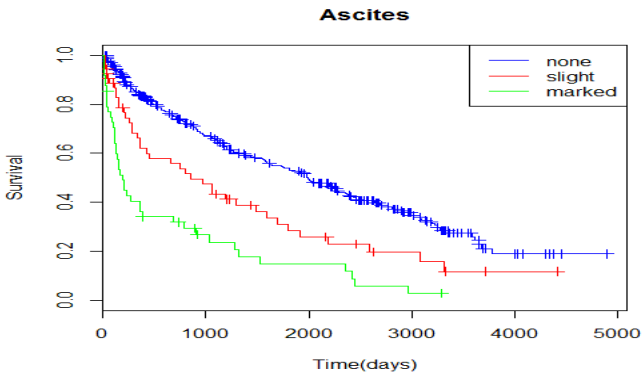
The figure 2b shows the survival function between men and female where it is observed a marginal difference and where the male patients have a higher survival rate compared to the

female. After a couple of years, the male patient curve starts to flat out a bit more than the female patient curve.



(b) Survival function for the genders.

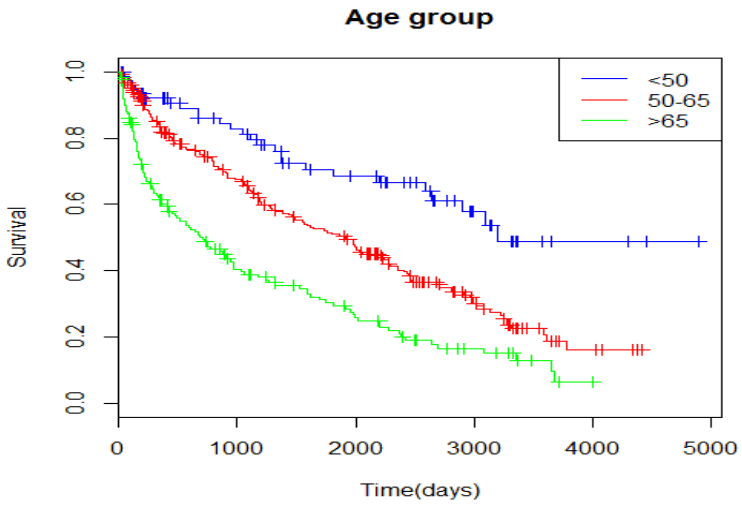
The figure 2c shows the survival function of the patients started with different levels of ascites. Ascites are the excess fluid in the abdomen of patients before treatment. This is a significant cause of lower survival rate. Its observed that the patients with no ascites have the highest survival compared to those with the most ascites have the lowest. the marked patients do not survive long, and the curve drops down very low at the first year. The none case however have a gentler slope.



(c) Survival function for the initial level of ascites in the patients.

The figure 2d shows the survival function of the different age groups. The blue curve, representing the youngest group below 50 years of age, seems to be the bulk of surviving patients. The older groups are more exposed. We observe that the predictors do not satisfy

the proportional hazard assumption, for which the curves would be basically the same, and the separation between the curves would remain proportional through time.



(d) Survival function for the different age groups.

Figure 2. Survival functions for the four covariates of interest: treat, sex, asc and agegr.

b) For each of the covariates, use the logrank test to investigate if the covariate has a significant effect on survival.

```
#b) # Logrank tests
survdiff(Surv(time,status)~treat, data=cirrhosis)
survdiff(Surv(time,status)~sex, data=cirrhosis)
survdiff(Surv(time,status)~asc, data=cirrhosis)
survdiff(Surv(time,status)~agegr, data=cirrhosis)
```

From the Kaplan-Meier plots the differences in the different groups are observed. Logrank tests helps to check if these differences are significant by testing the null hypothesis that the survival function is the same in both treatment groups. It is an efficient way to check how our variables affect status over the time. i.e. $H_0 : S_1(t) = S_2(t)$, for any t. The H_0 will be checked with test statistic $\chi 2$, i.e. $X_i^2 = = \frac{(O_i - E_i)^2}{\widehat{SE}(O_i - E_i)^2}$, where O_i is the observed number of events and E_i is the expected number of events in group i. The results from logrank test can be found in the table 13. By utilizing R we find that ascites and age are highly significant, with X_i^2 of 69.9 and 50.6, vs age and treatment with X_i^2 of 3.5 and 0.7. This is in coherent with our observations in the

plots above. We make use of Cox regression where the effects of the covariates are studied simultaneously.

Table 13. Results from the logrank test.

```
survdiff(formula = Surv(time, status) ~ treat)

          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
Chisq= 0.7  on 1 degrees of freedom, p= 0.4
```

```
> survdiff(Surv(time,status)~sex)
Call:
survdiff(formula = Surv(time, status) ~ sex, data = cirrhosis)

        N Observed Expected (O-E)^2/E (O-E)^2/V
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
```

```
> survdiff(Surv(time,status)~asc)
Call:
survdiff(formula = Surv(time, status) ~ asc)

        N Observed Expected (O-E)^2/E (O-E)^2/V
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     56.17     59.60

 Chisq= 69.9  on 2 degrees of freedom, p= 7e-16
```

```
> survdiff(Surv(time,status)~agegr)
Call:
survdiff(formula = Surv(time, status) ~ agegr)

          N Observed Expected (O-E)^2/E (O-E)^2/V
agegr=1  80       26     58.7     18.18     22.87
agegr=2 250      148    162.0      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
```

In the case of the asc and agegr survival functions it is visually clear that the null hypothesis in both cases are wrong. This is strongly confirmed by the log rank tests in table 13 which result in huge $\chi 2$ values and very small p-values.

c) Then do multiple Cox regression where the effects of all the covariates are studied simultaneously. Use age in years (not grouped). Summarize (and interpret) your findings. For this 'full' model with all covariates as main effects, find a 95% confidence interval for the hazard ratio for men versus women when all other covariates are constant. Write a conclusion about the effect of prednisone in this trial.

```
#c) Cox regression fitting all covariates
fit.all=coxph(Surv(time,status==1)~factor(sex)+factor(treat)+factor(asc)+age ,data=cirrhosis)
summary(fit.all)
```

Another way to study survival data is via proportional hazard models. In general, a multicovariate hazard function is given as h(t|x_1, x_2, …, x_p) = h_o(t)exp($\beta_1 x_1$ + $\beta_2 x_2$ + … + $\beta_p x_p$), where h_o(t) is the baseline hazard for a subject with all covariates equal to zero. We will now estimate the hazard function via a multiple Cox regression with all the covariates studied simultaneously. Table 14 shows the results from the Cox regression.

Table 14. Results from the Cox regression

```
coxph(formula = Surv(time, status) ~ sex + treat + asc + age)

  n= 488, number of events= 292

          coef exp(coef) se(coef)     z Pr(>|z|)
sex   0.462287  1.587702 0.125406 3.686 0.000228 ***
treat 0.044637  1.045648 0.117610 0.380 0.704293
asc   0.595150  1.813304 0.082864 7.182 6.86e-13 ***
age   0.048851  1.050064 0.006827 7.155 8.34e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
sex       1.588     0.6298    1.2417     2.030
treat     1.046     0.9563    0.8304     1.317
asc       1.813     0.5515    1.5415     2.133
age       1.050     0.9523    1.0361     1.064

Concordance= 0.682  (se = 0.017 )
Likelihood ratio test= 109.3  on 4 df,   p=<2e-16
Wald test            = 115.5  on 4 df,   p=<2e-16
Score (logrank) test = 122  on 4 df,     p=<2e-16
```

The column marked with z is the ratio of each regression coefficient to its standard error. Observe that ascites and age have highly statistical significance, while treatment is marginally significant. The exponential of the coefficient can be interpreted as multiplicative on the hazard.

Here we again see that age and asc are both highly significant covariates. In contrast to the logrank test however, sex is now also a significant covariant with a p-value of p = 0.0002.

We can also find the hazard ratio, HR, which is given as $\mathrm{HR} = \frac{\mathrm{h}(\mathrm{t}|x_1+\Delta,x_2,\ldots,x_p)}{\mathrm{h}(\mathrm{t}|x_1,x_2,\ldots,x_p)} = \exp(\beta_1\Delta)$,

where exp (β_1) is the hazard ratio corresponding to one unit increase in the value of the covariate x_1 while all other covariates are held constant. For our model, if its considered the hazard ratio for males and females, exp (β_{sex}) = exp (0,461) = 1,587, meaning that the hazard increases by a whole ≈ 58.7% if the subject is male. The 95% CI can be found for the above ratio from the usual exp (β_i)±exp (1.96·se (β_i)). Results from R code shows the CI to be in the interval [1.24,2.03] for the sex hazard ratio. Since 1 is not in the confidence interval, then the hazard is significant.

It is important to use multiple models for analysis of the survival data, as the logrank method and Cox regression yield different results on the significance of the sex covariant. However, it is still cannot be conclude that the effects of the prednisone treatment has any significant effect on the survival of the patients tested.

An alternative is the use a stratified version of the Cox regression to look at male and female patients independently. Below there is the summary of such operation.

Table 15. Stratified version of the Cox regression

```
Call:
coxph(formula = Surv(time, status) ~ strata(sex) + treat + asc +
    age)

  n= 488, number of events= 292

          coef exp(coef) se(coef)     z Pr(>|z|)
treat 0.033296  1.033857 0.118024 0.282    0.778
asc   0.595872  1.814613 0.083205 7.162 7.98e-13 ***
age   0.048932  1.050149 0.006855 7.138 9.46e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
treat     1.034     0.9673    0.8203     1.303
asc       1.815     0.5511    1.5416     2.136
age       1.050     0.9522    1.0361     1.064

Concordance= 0.688  (se = 0.018 )
Likelihood ratio test= 105.2  on 3 df,   p=<2e-16
Wald test            = 112.9  on 3 df,   p=<2e-16
Score (logrank) test = 120  on 3 df,   p=<2e-16
```

Here the 95% CI is listed in the second table, from 0.8203 - 1.303. The p-value indicates the significance of each variable. The stars after the last column can be interpreted as the significance code, as noted below. The more stars the more important the variable is. Treatment contributed almost nothing, and it can be safely removed from the model. Thus, the summary becomes (table 16):

Table 16. Summary from the Cox regression without treatment

```
> summary(model2)
Call:
coxph(formula = Surv(time, status) ~ asc + age + sex)

  n= 488, number of events= 292

        coef exp(coef) se(coef)     z Pr(>|z|)
asc 0.597047  1.816746 0.082785 7.212 5.51e-13 ***
age 0.048783  1.049993 0.006825 7.148 8.80e-13 ***
sex 0.462537  1.588098 0.125395 3.689 0.000225 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .95 upper .95
asc     1.817     0.5504     1.545     2.137
age     1.050     0.9524     1.036     1.064
sex     1.588     0.6297     1.242     2.031

Concordance= 0.682  (se = 0.017 )
Likelihood ratio test= 109.1  on 3 df,   p=<2e-16
Wald test            = 115.7  on 3 df,   p=<2e-16
Score (logrank) test = 122  on 3 df,   p=<2e-16
```

Notice that the sex variable has three stars as well, although the p-value is much larger than that of age and ascites. The analysis will be proceeded by removing the least significant variables from the model. Earlier we saw that the most significant covariates are ascites and age. Sex had a slight effect. Let us see how removing the sex variable will affect our results as well.

Table 17. Summary from the Cox regression without sex variable

```
> summary(model3)
Call:
coxph(formula = Surv(time, status) ~ asc + age)

  n= 488, number of events= 292

        coef exp(coef) se(coef)     z Pr(>|z|)
asc 0.571241  1.770463 0.082173 6.952 3.61e-12 ***
age 0.042720  1.043645 0.006504 6.569 5.08e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
    exp(coef) exp(-coef) lower .95 upper .95
asc     1.770     0.5648     1.507     2.080
age     1.044     0.9582     1.030     1.057

Concordance= 0.678  (se = 0.018 )
Likelihood ratio test= 95.12  on 2 df,   p=<2e-16
Wald test            = 107  on 2 df,   p=<2e-16
Score (logrank) test = 112.1  on 2 df,   p=<2e-16
```

The change is slight, but still significant based on our previous plots and analysis. Thus, the last model will include ascites, age and sex, while the treatment is excluded. Ascites and age are the most important factors, and of the two ascites are most significant.

The purpose of this study was to find the effects of the provided hormone prednisone vs a placebo treatment. It is observed that there was almost no effect of the prednisone treatment on liver cirrhosis, thus we can conclude that its necessary to continue research for another cure to treat this condition.

Appendix. Code

```
rm(list=ls(all=TRUE))

# a. Reading datafiles

crabs <- read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/data/crabs.txt",header=T)

summary(crabs)

# Choose a binary logistic regression model

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

summary(fit.width)

# b. Computing the odd 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)
# c. Binary logistic regression model for the other covariates one by one
fit.width=glm(y~width , data=crabs ,family=binomial)
fit.weight=glm(y~weight , data=crabs ,family=binomial)
fit.color=glm(y~factor(color), data=crabs ,family=binomial)
fit.spine=glm(y~factor(spine), data=crabs ,family=binomial)
summary(fit.width)
summary(fit.weight)
summary(fit.color)
summary(fit.spine)

# Grouped logistic fit for all significant covariates
fit.multi=glm(y~width+weight+factor(color)+factor(spine), data = crabs , family = binomial)
summary(fit.multi)
# d. Grouped logistic fit for all significant covariates
```

```
fit.multisig=glm(y~width+weight , data = crabs , family = binomial)

summary(fit.multisig)

# Checking correlation between width and weight

plot(crabs$weight , crabs$width , xlab = "Weight [kg]", ylab = "Width [cm]")

cor(crabs$weight , crabs$width)

# Fitting the final model

fit.multisig=glm(y~width , data = crabs , family = binomial)

summary(fit.multisig)

# e. Checking interaction

fit.multiint=glm(y~width+ factor(color) + width:factor(color), data = crabs , family = binomial)

fit.multiint=glm(y~width+ factor(spine) + width:factor(spine), data = crabs , family = binomial)

fit.multiint=glm(y~factor(color)+ factor(spine) + factor(color):factor(spine), data = crabs , family = binomial)

summary(fit.multiint)

#Exercise 2.

# a. Reading in 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 covariates

fit.1 = glm(Total2000~offset(Log.athletes) + Total1996 + Log.population + GDP.per.cap ,data=olympic ,family=poisson)

summary(fit.1)

# Checking correlation between covariates

plot(olympic)
```

```
cor(olympic$Total2000 , olympic$Total1996)

# 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)

#Exercise 3.

#a. Reading in data 2

cirrhosis <- read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/data/cirrhosis.txt",header=TRUE)

library(survival)

survpred <- survfit(Surv(time,status)~1, conf.type="none")

summary(survpred,xlab="Time(days)",ylab="Survival")

# Plotting the Kaplan-Meier estimate for each fit

plot(survpred)

survpred1 <- survfit(Surv(time,status)~treat, conf.type="plain")

summary(survpred1)

plot(survpred1,col=c('blue', 'red'),xlab="Time(days)",ylab="Survival")

legend('topright', c("prednisone","placebo"), col=c('blue', 'red'), lty=1)
```

```
title("Treatment")
survpred2 <- survfit(Surv(time,status)~sex, conf.type="plain")
summary(survpred2)
plot(survpred2,col=c('blue', 'red'),xlab="Time(days)",ylab="Survival")
legend('topright', c("Male","Female"), col=c('blue', 'red'), lty=1)
title("Sex")
survpred3 <- survfit(Surv(time,status)~asc, conf.type="plain")
summary(survpred3)
plot(survpred3,col=c('blue', 'red', 'green'),xlab="Time(days)",ylab="Survival")
legend('topright', c("none","slight", "marked"), col=c('blue', 'red', 'green'), lty=1)
title("Ascites")
survpred4 <- survfit(Surv(time,status)~agegr, conf.type="plain")
summary(survpred4)
plot(survpred4,col=c('blue', 'red', 'green'),xlab="Time(days)",ylab="Survival")
legend('topright', c("<50","50-65", ">65"), col=c('blue', 'red', 'green'), lty=1)
title("Age group")
#b) # Logrank tests
survdiff(Surv(time,status)~sex + age)
survdiff(Surv(time,status)~treat)
survdiff(Surv(time,status)~sex)
survdiff(Surv(time,status)~asc)
survdiff(Surv(time,status)~agegr)
cox <- coxph(Surv(time,status)~sex+treat+ asc+ age)
```

```
cox

summary(cox)

cox.strata <- coxph(Surv(time,status)~strata(sex)+treat+ asc+ age)

cox.strata

summary(cox.strata)

model1 <- coxph(Surv(time,status)~sex+ asc+ age)

model1

summary(model1)

model2 <- coxph(Surv(time,status)~asc+ age + sex)

model2

summary(model2)

model3 <- coxph(Surv(time,status)~asc+ age)

model3

summary(model3)

model4 <- coxph(Surv(time,status)~asc)

model4

summary(model4)
```