STK4900 - Assignment 2

Simen Nyhus Bastnes

6. April 2017

Problem 1

In this problem, we want to take a closer look at data regarding contraceptive use amoungst currently married and fecund women, taken from the "Fiji Fertility Survey".

a) First, we are interested in seeing how the probability of not using contraceptives depends on the desire to have more children. To do this, we need to choose a suitable regression model. The probability of not using contraceptives can be modelled by the fraction of the variables `no.no_use/no.tot`, and since our response is binary (uses contraceptives/does not use contraceptives), we can use a logistical regression model.

```
fit.wants = glm(cbind(no.no_use,no.tot-no.no_use)~wants.more,data=cuse,family=binomial)
```

A summary of this gives us the following output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.23499    0.07677  16.086   <2e-16 ***
wants.more  -1.04863    0.11067  -9.475   <2e-16 ***
```

where the coefficients correspond to β_0 and β_1 in the logistic regression model given by

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

and x corresponds to the covariate `wants.more`. The odds is then given by

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

Now we can find the odds ratio between the different groups of `wants.more`

$$\begin{aligned} OR &= \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))} \\ &= \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp\{\beta_1\} \end{aligned}$$

In our case, the odds ratio is

$$OR = \exp(-1.04863) = 0.35$$

The odds ratio here is the ratio of the ones that don't want more children using contraceptives compared to those who want. An odds ratio < 1 means that in the population where they do not want more children, not using contraceptives is less likely to occur. Meaning that those who wants more children are also less likely to use contraceptives (almost 3 times as likely to not use contraceptives), which makes sense, logically.

b) Using all the variables as main effects in the regression

```
fit.all = glm(cbind(no.no_use,no.tot-no.no_use)~wants.more+factor(agegr)+education, data
    =cuse, family=binomial)
```

A summary of this gives us the following

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.9662     0.1720  11.429  < 2e-16 ***
wants.more      -0.8330     0.1175  -7.091 1.33e-12 ***
factor(agegr)2  -0.3894     0.1759  -2.214  0.02681 *
factor(agegr)3  -0.9086     0.1646  -5.519 3.40e-08 ***
factor(agegr)4  -1.1892     0.2144  -5.546 2.92e-08 ***
education       -0.3250     0.1240  -2.620  0.00879 **
```

From this table, we see that all variables are significant, and that all variables seem to be covariates.

c) Finally, we want to check if adding interaction between age groups and the desire for more children in the future changes the results.

```
fit.int = glm(cbind(no.no_use,no.tot-no.no_use)~wants.more+factor(agegr)+education+
    factor(agegr):wants.more, data=cuse, family=binomial)
```

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 1.80317    0.18018  10.008  < 2e-16 ***
wants.more                 -0.06622    0.33071  -0.200  0.84130
factor(agegr)2             -0.39460    0.20145  -1.959  0.05013 .
factor(agegr)3             -0.54666    0.19842  -2.755  0.00587 **
factor(agegr)4             -0.57952    0.34742  -1.668  0.09530 .
education                  -0.34065    0.12577  -2.709  0.00676 **
wants.more:factor(agegr)2 -0.25918    0.40975  -0.633  0.52704
wants.more:factor(agegr)3 -1.11266    0.37404  -2.975  0.00293 **
wants.more:factor(agegr)4 -1.36167    0.48433  -2.811  0.00493 **
```

Now we see that the variable `wants.more` is no longer significant. We check the accuracy of our new model by checking the deviances

```
anova(fit.all,fit.int, test="Chisq")
```

which gives us the printout

```
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        10     29.917
2         7     12.630  3   17.288 0.0006167 ***
```

We see that the P value is very small, showing that the model with the interaction is significantly better than the model we tried earlier.

Problem 2

We want to look closer at the claim that participants from larger and wealthier nations are more likely to win medals in competitions like the Olympic games.

a) The number of medals won by a nation is clearly related to the number of athletes participating. This makes the rate of medals won per athlete a much more interesting way of measuring the quantities. We can do this by introducing an "offset" covariate

$$E_i = \exp\left(\log(w_i) + \beta_0 + \beta_1 x_{1i} + \beta_{2i} + \cdots + \beta_p x_{pi}\right)$$
$$\frac{E_i}{w_i} = \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_{2i} + \cdots + \beta_p x_{pi}\right)$$

where E_i is medals won per country, and $\log(w_i)$ is a "covariate" where the regression coefficient is known to equal 1. Using this offset as an exposure allows us to get the rate of medals won per athlete.

b) Using Poisson regression, we fit a model for the rate of medals won per athlete, setting `Log.athletes` as an offset, and trying to find a model that works well using some (or all) of the other predictors.

Of the different predictors, one might think that `Total1996` is quite important, as a lot of top athletes are still at the top four years later at the next Olympics. The total population of the country might also be relevant as it allows for a larger pool of potential talents. We test with the model

```
fit.olympic1 = glm(Total2000~offset(Log.athletes)+Total1996+Log.population+GDP.per.cap,
    data=olympic, family=poisson)
```

```
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 ***
```

We see that as suspected, `Total1996` is highly significant, while the population on the other hand is not significant at all. We do however see that the GDP predictor is significant, however the coefficient is negative, seeming to indicate that the poorer a country, the higher amount of medals they win, which seems a bit weird. Removing population from our model yields.

```
fit.olympic2 = glm(Total2000~offset(Log.athletes) +Total1996+ GDP.per.cap, data=olympic,
     family=poisson)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.589318   0.057648 -44.916  < 2e-16 ***
Total1996    0.012825   0.001140  11.248  < 2e-16 ***
GDP.per.cap -0.015800   0.003059  -5.164 2.41e-07 ***
```

where we see the all the predictors are very much significant, and most importantly, that the hypothesis that the GDP negatively affects poorer countries seem to actually be the opposite.

Problem 3

In this problem, we want to look at patients with liver cirrhosis, and how they respond to different forms of treatment based on age, sex, and other variables.

a) We want to make Kaplan-Meier plots for the survival function for each level of covariates treatment, sex, ascites and grouped age.

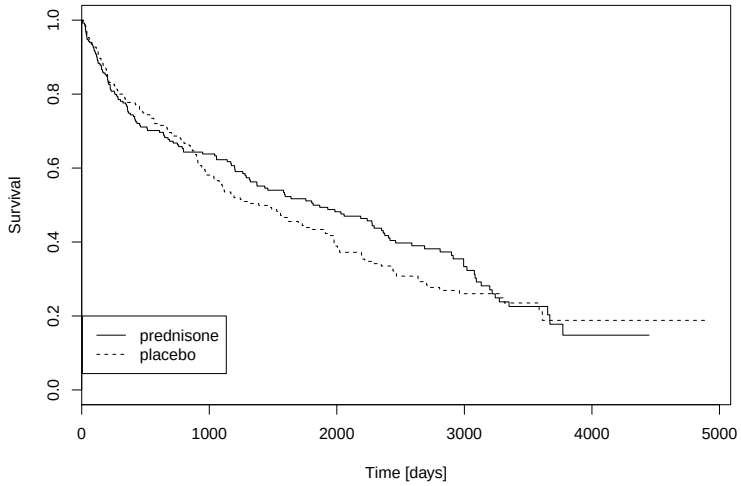

Figure 1: Kaplan-Meier plot for the survival function of patients undergoing prednisone and placebo treatment. We see that the differences are relatively small in the end points, but in a large segment of the middle, the prednisone treatment is a bit better.

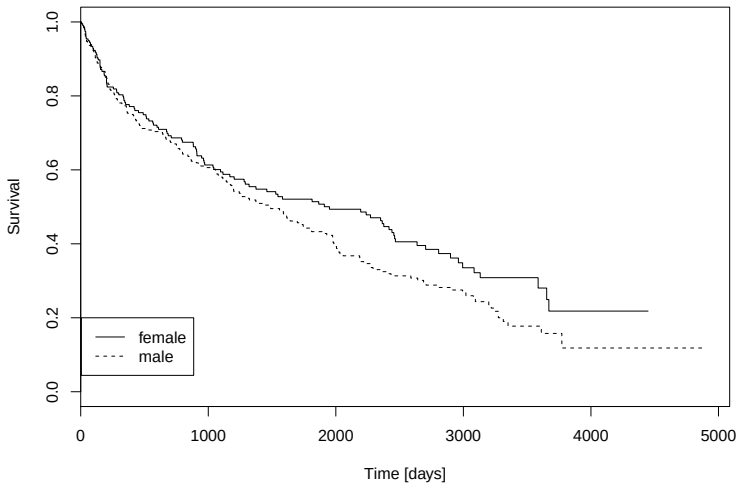

Figure 2: Kaplan-Meier plot for the survival function of patients of both genders. For larger time, we see that the survival is slightly higher for females.

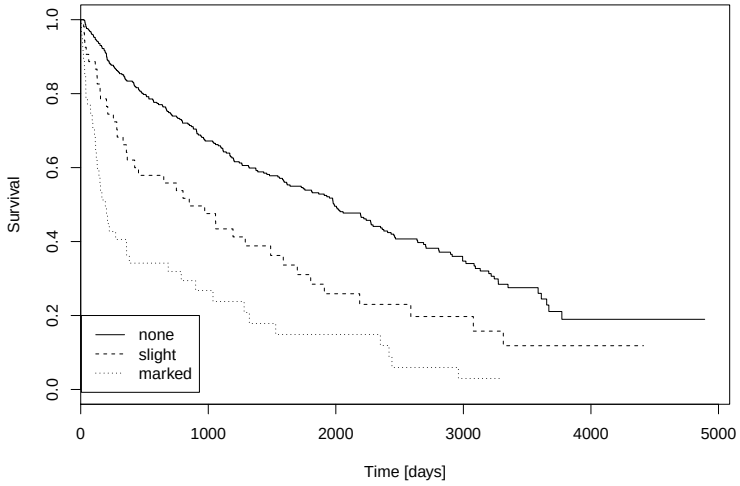

Figure 3: Kaplan-Meier plot for the survival function for ascites at the start of treatment. We see clearly that the level of ascites negatively affect the survival greatly.

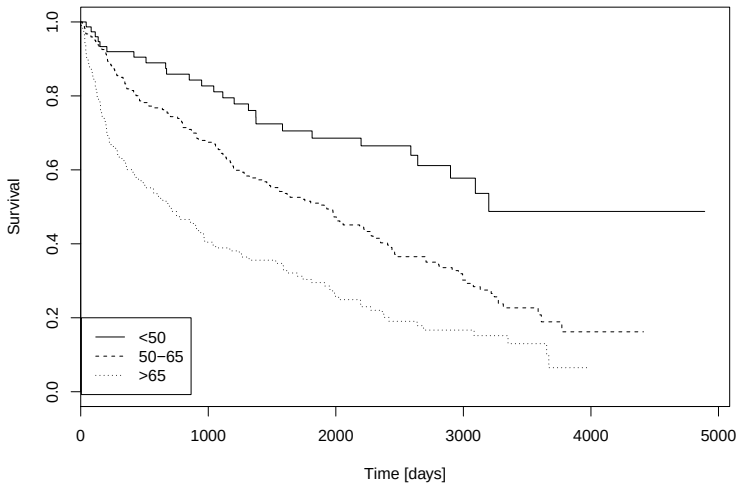

Figure 4: Kaplan-Meier plot for the survival function for the different age groups. The survival drops for increasing level of age group.

Worthy of note, is that for all the plots, the difference between the levels are small for short time periods, which is not surprising, as one would think that at small timescales, the differences don't have time to affect the survival of the patients.

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

```
survdiff(Surv(time,status)~treat, data=cirrhosis)
```

Running the logrank test on all covariates gives us the printout

```
[1] "Logrank of treatment"
Call:
survdiff(formula = Surv(time, status) ~ treat, data = cirrhosis)

          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= 0.394
[1] "Logrank of sex"
Call:
survdiff(formula = Surv(time, status) ~ sex, data = cirrhosis)
```

```
46
47          N Observed Expected (O-E)^2/E (O-E)^2/V
48 sex=0 198      111      127      2.00      3.55
49 sex=1 290      181      165      1.54      3.55
50
51  Chisq= 3.5  on 1 degrees of freedom, p= 0.0596
52 [1] "Logrank of ascites"
53 Call:
54 survdiff(formula = Surv(time, status) ~ asc, data = cirrhosis)
55
56          N Observed Expected (O-E)^2/E (O-E)^2/V
57 asc=0 386      211    251.9      6.63     48.66
58 asc=1  54       39     26.2      6.30      6.94
59 asc=2  48       42     14.0     56.17     59.60
60
61  Chisq= 69.9  on 2 degrees of freedom, p= 6.66e-16
62 [1] "Logrank of age group"
63 Call:
64 survdiff(formula = Surv(time, status) ~ agegr, data = cirrhosis)
65
66            N Observed Expected (O-E)^2/E (O-E)^2/V
67 agegr=1  80       26     58.7     18.18     22.87
68 agegr=2 250      148    162.0      1.21      2.72
69 agegr=3 158      118     71.3     30.51     40.87
70
71  Chisq= 50.6  on 2 degrees of freedom, p= 1.05e-11
```

From the tests, we see that the p-values are high for the covariates `treat` and `sex`, meaning that they do not have a significant effect on the survival. The covariates `asc` and `agegr` however are quite significant.

c) Finally, we do a multiple Cox regression where the effects of all the covariates are studied simultaneously.

```
7 fit.cox = coxph(Surv(time,status)~treat+sex+asc+age, data=cirrhosis)
```

A summary of this gives us

```
72           coef exp(coef) se(coef)     z Pr(>|z|)
73 treat 0.044637  1.045648 0.117610 0.380 0.704293
74 sex   0.462287  1.587702 0.125406 3.686 0.000228 ***
75 asc   0.595150  1.813304 0.082864 7.182 6.86e-13 ***
76 age   0.048851  1.050064 0.006827 7.155 8.34e-13 ***
77 ---
78
```

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

Where we see that the p-value of `treat` is very large and not in any way significant. At the same time, `sex` has three stars, indicating that it is highly significant, but also has a value many magnitudes lower than the remaining covariates.

In the lower half of the table above, the 95% confidence interval for the different covariates are found. The 95% confidence interval for the hazard ratio of men versus women when all other covariates are constant is then [1.2417, 2.030].

In the end, looking at the Kaplan-Meier plots, the logrank tests, and the multiple Cox regression, we have to conclude that the prednisone treatment does not have any significant effect on the survival of cirrhosis patients. Much more importantly is the ascites at the start of treatment, age, and somewhat the gender of the person affected.

Code

Problem 1

```r
# Problem 1
cuse <- read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v17/cuse.txt",sep="\t",header=TRUE)

print("Task a)")
fit.wants = glm(cbind(no.no_use,no.tot-no.no_use)~wants.more, data = cuse, family=binomial)
summary(fit.wants)

print("Odds ratio")
expcoef(fit.wants)

print("Task b)")
fit.all = glm(cbind(no.no_use,no.tot-no.no_use)~wants.more+factor(agegr)+education, data=cuse, family=binomial)
summary(fit.all)

print("Task c)")

fit.int = glm(cbind(no.no_use,no.tot-no.no_use)~wants.more+factor(agegr)+education+factor(agegr):wants.more, data=cuse, family=binomial)

summary(fit.int)
anova(fit.all,fit.int, test="Chisq")
```

Problem 2

```
olympic <- read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v17/olympic.txt",
    sep="\t",header=TRUE)

# Task 2b)
fit.olympic1 = glm(Total2000~offset(Log.athletes)+Total1996+Log.population+GDP.per.cap, data=
    olympic, family=poisson)

fit.olympic2 = glm(Total2000~offset(Log.athletes) +Total1996+ GDP.per.cap, data=olympic,
    family=poisson)

summary(fit.olympic2)
```

Problem 3

```
cirrhosis <- read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v17/cirrhosis.
    txt",sep="\t",header=TRUE)
attach(cirrhosis)
library(survival)

# 3a) Make Kaplan-Meier plots

# Treatment
surv.treat = survfit(Surv(time,status)~treat, conf.type="plain")
pdf('surv_treat.pdf', width=8, height=6)
plot(surv.treat,lty=1:2,xlab="Time [days]",ylab="Survival")
legend(5,0.2,c("prednisone", "placebo"), lty=1:2)
dev.off()

# Sex
surv.sex = survfit(Surv(time,status)~sex, conf.type="plain")
pdf('surv_sex.pdf', width=8, height=6)
plot(surv.sex,lty=1:2,xlab="Time [days]",ylab="Survival")
legend(5,0.2,c("female", "male"), lty=1:2)
dev.off()

# Ascites
surv.asc = survfit(Surv(time,status)~asc, conf.type="plain")
pdf('surv_asc.pdf', width=8, height=6)
plot(surv.asc,lty=1:2:3,xlab="Time [days]",ylab="Survival")
legend(5,0.2,c("none", "slight", "marked"), lty=1:2:3)
dev.off()

# Grouped age
```

```
surv.agegr = survfit(Surv(time,status)~agegr, conf.type="plain")
pdf('surv_agegr.pdf', width=8, height=6)
plot(surv.agegr,lty=1:2:3,xlab="Time [days]",ylab="Survival")
legend(5,0.2,c("<50", "50-65", ">65"), lty=1:2:3)
dev.off()


print("Task b)")
print("Logrank of treatment")
survdiff(Surv(time,status)~treat, data=cirrhosis)
print("Logrank of sex")
survdiff(Surv(time,status)~sex, data=cirrhosis)
print("Logrank of ascites")
survdiff(Surv(time,status)~asc, data=cirrhosis)
print("Logrank of age group")
survdiff(Surv(time,status)~agegr, data=cirrhosis)

print("Task c)")
fit.cox = coxph(Surv(time,status)~treat+sex+asc+age, data=cirrhosis)
summary(fit.cox)
```