STK4900 Obligatory 2

Jonas Thoen Faber

April 20, 2021

All coding is done using jupyter lab and can be found at https://github.com/JonasTFab/STK4900_Statistical_Methods_and_Applications/tree/main/obligatory_2. A compressed version of the code are located in the appendix.

Problem 1

a)

We know that the width variable (predictor) is continuous but the satellite variable (response) is binary. Because of the latter, I have chosen to implement logistic regression to solve this problem. It is in general preferable to avoid a linear regression when we have a binary problem. The logistic regression model with one predictor is then given as shown below:

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

where the predictor x can be a continuous numeric value. We then have an expression of the probability of one or more satellites present when the width of the carapace is known (one predictor).

b)

From the lecture notes, we have that the odds ratio is given as following:

$$OR = \frac{p(x + \Delta x)/[1 - p(x + \Delta x)]}{p(x)/[1 - p(x)]} \tag{2}$$

with x and Δx as covariates. It can be shown given the probability from Equation 1 from a logistic regression model, that this expression may be shorten into the following:

$$OR = \exp(\beta_1 \Delta) \tag{3}$$

If we want to consider the odds ratio between crabs that differ one centimeter in width, we get that $\Delta = 1$ cm thus:

$$OR = \exp(\beta_1) \tag{4}$$

This also means that the unit of β_1 must be in cm^{-1} since the dimension in the exponential term should be unitless. The odds ratio is defined as a measure between the relation of the odds of two particular outcomes. The two odds represents the outcome with and without a given exposure. In our case, the odds ratio is the measure of outcome with one unit increase in the exposure.
If we have that the probability is small for both covariates, we may approximate the odds ratio to the relative risk. We observe this from Equation 2 as both of the denominator would be ≈ 1 and the equation becomes the definition of relative risk $RR = p(x + \Delta x)/p(x)$. We are most likely not getting a low probability which means that the odds ratio is not a good approximation for the relative risk.
To find the CI for the odds ratio, we need to fit an logistic regression model in R to get an approximation of β_1. This is done by the `glm()` functionality provided in R. We then get the 95 % CI for odds ratio by transforming the lower and upper CI limits for β_1. From R, we get $\hat{\beta}_1 = 0.4972$ and $se(\hat{\beta}_1) = 0.1017$ which results in a lower and upper CI of 0.2979 and 0.6965. We thus have the estimated $OR = 1.6441$ with upper an lower odds ratio CI of 1.3470 and 2.0068. We note that if $OR = 1$, the width of the carapace would not have any influence on the presence of satellites. The lower CI is clearly greater than 1 meaning that we may assume that the width of the carapace does affect the presence of satellites significantly.

c)

Looking at the data set, we observe that the covariates color and spine already are grouped into categories meaning that we can not consider these as numerical. This is not the case for the weight covariate. We observe that this covariate is continuous and may therefore include it as numerical. Another solution is to group the weight covariate. We will not do so when solving the problem.
We will use the same approach as in b) but with other covariates to fit a logistic regression model in R. The difference is that we need to categorize the covariates color and spine before fitting the model. This can be done by using the `factor()` function in R. The color and spine categorised as 1 will be considered as the initial factor. We then get the results shown in Table 1.

Table 1: Outputs from the fitted logistic regression model using each of the covariates separately.

	Estimate	Standard error	z	Pr(> |z|)
(Intercept)	-3.6947	0.3767	-4.198	2.70e-05
Weight	1.8151	0.3767	4.819	1.45e-06
(Intercept)	1.0986	0.6667	1.648	0.0994
Color 2	-0.1226	0.7053	-0.174	0.8620
Color 3	-0.7309	0.7338	-0.996	0.3192
Color 4	-1.8608	0.8087	-2.301	0.0214
(Intercept)	0.8602	0.3597	2.392	0.0168
Spine 2	-0.9937	0.6303	-1.577	0.1149
Spine 3	-0.2647	0.4068	-0.651	0.5152

First we take a look at the p-value of the weigh covariate which is quite small meaning that this covariate is considered as significant. As the estimate is positive tells us that the increase of weight results in a greater possibility of one or more satellites.
Now looking at the model that include the color of the female crab, we note that color 4 has the lowest p-value that is also the only factor with a p-value lower than 5 %. Color 4 is coded as the dark colored crabs. As the estimate is negative, the probability of having one or more satellites are greater for color 1 than color 4.
Finally, the p-value with spine as predictors is lowest for the intercept which is the only p-value lower than 5 %. The intercept represents both spines as good.

d)

Next we create a logistic regression model using all the covariates. We then get the results as shown in Table 2.

Table 2: Outputs from the fitted logistic regression model using all covariates.

	Estimate	Standard error	z	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
Color 2	-0.10290	0.78259	-0.131	0.8954
Color 3	-0.48886	0.85312	-0.573	0.5666
Color 4	-1.60867	0.93553	-1.720	0.0855
Spine 2	-0.09598	0.70337	-0.136	0.8915
Spine 3	0.40029	0.50270	0.796	0.4259

We note from Table 2 that the intercept has the lowest p-value. The intercept represents color 1 (medium light) and spine 1 (both good). The second to best

covariate is female crabs with color 4 (dark). The p-values we get are still far from what we got with when fitting a logistic model with weight as a predictor. From this model, we got a p-value of 1.45e-06. From b), using width as predictor resulted in a p-value of 1.02e-06. Now we will try to use only width and weight as predictors.

Table 3: Outputs from the fitted logistic regression model with width and weight as covariate.

	Estimate	Standard error	z	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

From Table 3, neither the width nor the weight has sufficient p-value. If we think of it, width and weight are two factors that can be consider as proportional to some degree. Hence we should believe that there are a correlation between these two factors. We can check the correlation between the covariates by using the function `cor()` in R.

Table 4: Correlation between the covariates.

	Width	Weight	Color	Spine
Width	1.0000	0.8869	-0.2644	-0.1219
Weight	0.8869	1.0000	-0.2508	-0.1665
Color	-0.2644	-0.2508	1.0000	0.3785
Spine	-0.1219	-0.1665	0.3785	1.0000

We observe from Table 4 that both width and weight are highly correlated. This imply that we should use only one of these when we fit a model.

e)

Next we want to check if there are any interactions between the covariates. As width and weight are highly correlated, they will not interact and we need only to use one of them as covariate. We will use width. To check of any interactions, we need to add an additional covariate as a product between two variables. We use the same approach as above and we get the best results as shown in Table 5.

Table 5: Outputs from the fitted logistic regression model considered as the best model with interactions between covariates.

	Estimate	Standard error	z	Pr(> |z|)
(Intercept)	-9.06912	7.67811	-1.181	0.238
Width	0.36908	0.29231	1.263	0.207
Spine	-1.37855	3.04961	-0.452	0.651
Width:Spine	0.05389	0.11646	0.463	0.644

We note the insignificance between the covariates. The results tells us that the is no interaction between the covariates that is of any significance.

Problem 2

Before I could extract the data from `olympic.txt`, I had to download the file and replace white space with underscore for country with double names or more. The text file used can be found on my Github page introduced above.

a)

A Poisson regression model is a model that assumes that a random variable Y with a parameter λ follows the distribution defined in Equation 5.

$$P(Y = y) = \frac{\lambda^y}{y!}e^{-\lambda}, \qquad y = 0, 1, 2, ... \tag{5}$$

This equation is better known as the Poisson distribution and can be written as $Y \sim \text{Po}(\lambda)$. Equation 5 produce the probability of getting an integer value y based on the parameter λ. Note that y is allways positive or zero. In this distribution, λ is also allways positive. Poisson regression will estimate λ as a function of covariates. This function should give positive output which means that this regression model can be written as following:

$$\lambda_i = \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}) \tag{6}$$

where the subscript i corresponds to the count of subject y_i. β_j is the regression coefficient while x_{ji} is the covariate number j for subject i.
In Poisson regression models, it is often considered to include an offset term in Equation 6 as $\log(w_i)$ within the exponential. w_i is the counted weight of the number of subjects in group i. Hence, this term is used when the data is on a aggregated form. In our case, we use the data given in `olympic.txt` which is indeed aggregated such that the number of athletes is defined by other covariates (country, population). Therefore, `Log.athletes` fits as an offset term.

b)

We then fit a Poisson regression model on the data by using the function `glm(Total2000 ~ Total1996 + Log.population + offset(Log.athletes) + GDP.per.cap, data=data, family=poisson)` in R with all covariates included. Note that `data` is the data extracted from the `olympic.txt` file. We then get the results shown in Table 6.

Table 6: Fitted Poisson regression model using data from the olympic athletes text file. All covariates are used. Dispersion parameter taken to be 1.

	Estimate	Standard error	z	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
GPD.per.cap	-0.014924	0.003208	-4.652	3.29e-06

Before we investigate the table above, we need to consider if the data is in fact Poisson distributed. For a Poisson distribution, we have that $\overline{y} = s^2$ which are both estimates of λ. We then check for overdispersion by calculating $CD = s^2/\overline{y}$. If $CD >> 1$, we have a overdispersion meaning that the data cannot be considered as Poisson distributed. Luckily for us, R computes CD for us. From the data used, we got an coefficient of dispersion which is taken to be 1. The data is therefore consider as Poisson distributed.
Looking at Table 6, we observe an high P-value of `Log.population` meaning that this covariate is associated with `Total2000` in a insignificant manner, hence is neglected. We will then try to fit a model without this parameter. The results is shown in Table 7

Table 7: Fitted Poisson regression model using data from the olympic athletes text file. Population is not included as covariate. Dispersion parameter taken to be 1.

	Estimate	Standard error	z	Pr(> |z|)
(Intercept)	-2.589318	0.057648	-44.916	<2e-16
Total1996	0.012825	0.001140	11.248	<2e-16
GPD.per.cap	-0.015800	0.003059	-5.164	2.41e-07

The P-values of both covariates is very low implying significance. Even thought both `Total1996` and `GPD.per.cap` show sign of high significance, we observe from the estimate that these values are very low with an even lower standard errors. This means that the total number of medals in 2000 is not highly related to the medals of previous games nor the wealth of the country, but to certain degree. In fact, the negative sign on the estimate of wealth tells us that the wealth in a country has a negative effect on the medals gained in 2000. Therefore based on the results from these statistic procedures, we do not confirm the statement that larger and wealthier countries win more medals.

Problem 3

a)

To make a Kaplan-Meier plot, we have to import the `survival` library provided in R. This library enables the `survfit()` function which is known as a estimator of the survival function. The function requires a time and status parameter which is crucial to know if the total number of patient is still alive after a certain period of time. We may then plot these two parameters as a function of covariates such as treatment, sex, ascites and grouped age to observe if they have any effect. This is shown in Figure 1.

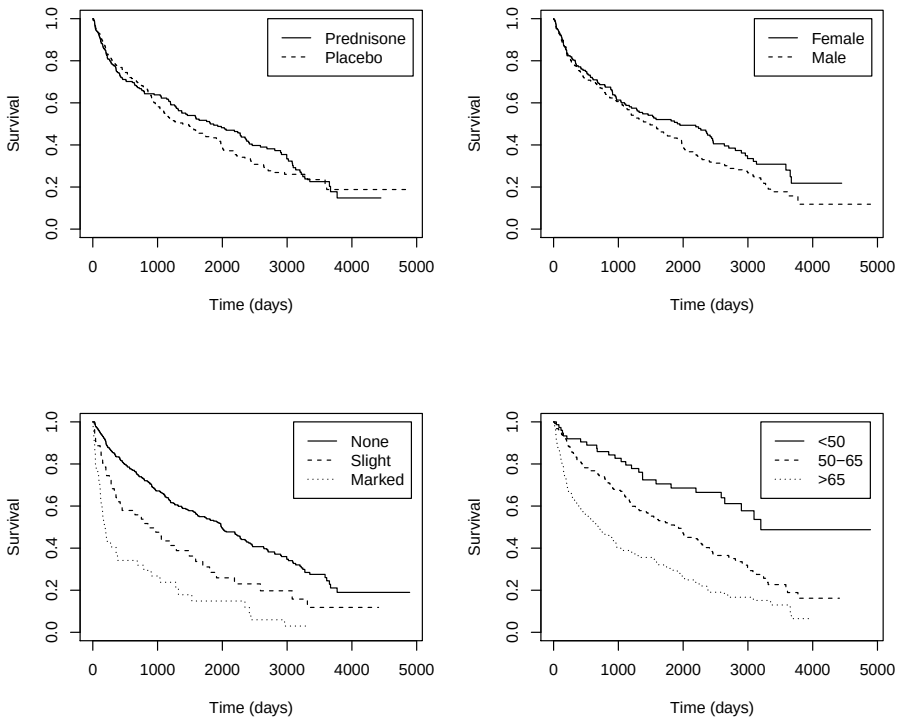

Figure 1: Kaplan-Meier plots as function of four seperate covariates. Upper left: treatment, upper right: sex, lower left: ascites, lower right: age group.

From the Kaplan-Meier plot including treatment as covariate, we note that within the first 1000 days or so that prednisone and placebo has equal results. Prednisone is slightly lower but that may be caused by uncertainties. Between 1000 and 3000 days we observe that having the hormone might have an effect

on the patients. After 3000 days, it looks like the hormone has a negative effect. In the upper left plot, we observe that female are more likely to survive the treatment. Note that the treatment in this case does not specify of prednisone or placebo. From the lower plots, it is quite plausible that ascites and age group has an effect on the survival rate.

b)

We want to test the null hypothesis that the survival function is the same in all groups. We can do so by checking if

$$\chi^2 = \frac{(O_2 - E_2)^2}{\hat{se}(O_2 - E_2)^2} \tag{7}$$

is approximately chi-squared distributed with $K - 1$ degrees of freedom where K is the number of groups. The test is called the logrank test and is a function provided from the `survival` library. The function is called `surdiff()` and require survival time and status as input as well as one or more covariate. We will use a single covariate at a time, e.g. `survdiff(Surv(time, status)~treat)`. The results are given in the following table.

Table 8: Logrank test on patient data with treatment, sex, ascites and age group as covariates. Note that the logrank test is used with each of the covarieates seperately. Treatment has $\chi^2 = 0.7$ on 1 df and $p = 0.4$. Sex has $\chi^2 = 3.5$ on 1 df and $p = 0.06$. Ascites has $\chi^2 = 69.9$ on 2 df and $p = 7e - 16$. Age group has $\chi^2 = 50.6$ on 2 df and $p = 1e - 11$.

	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
sex=0	198	111	127	2.00	3.55
sex=1	290	181	165	1.54	3.55
ascites=0	386	211	251.9	6.63	48.66
ascites=1	54	39	26.2	6.30	6.94
ascites=2	48	42	14.0	56.17	59.60
age_group=1	80	26	58.7	18.18	22.87
age_group=2	250	148	162.0	1.21	2.72
age_group=3	158	118	71.3	30.51	40.87

We note that the computed χ^2 is quite low regarding treatment as covariate. Hence the survival function is close to what we observe. Still, the p-value is quite large and we consider treatment as insignificant. The p-value on sex is also quite large ($> 5\%$) but might in some cases be considered as significant. Both ascites and age group has a very low p-value which imply a significance. We wanted to test the null hypothesis that the survival function is the same for all groups, but we note that there is a great deviance in group 0 and 2 in ascites

as well as group 1 and 3 in age group. Hence, we reject the null hypothesis for these covariates.

c)

Finally we will perform a multiple Cox regression on the data. The regression is supplied by the library previously used and its function is called `coxph()`. In our case, we want to study the effect using all the covariates as main effects. We also replace the categorised `agegr` covariate with the continuously `age` covariate. The Cox regression function requires the categorised covariates to be factorized, so that the function call becomes `coxph(Surv(time, status==1) ~ factor(treat) + factor(sex) + factor(asc) + age, data=data)`. Note that we perform the Cox regression on the events when patients have died, hence `status==1`. We get the results shown in Table 9.

Table 9: Cox regression on cirrhosis data set. Out of 488 patients, 292 events occured. Concordance=0.682 (se=0.017), likelihood ratio test=109.3 with p<=2e-16, Wald test=115.4 with p<=2e-16, score (logrank) test=123.9 with p<=2e-16. df=5.

	coef	exp(coef)	se(coef)	z	Pr(> |z|)
factor(treat)1	0.044818	1.045837	0.117657	0.381	0.703263
factor(sex)1	0.461877	1.587050	0.125631	3.676	0.000236
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

	exp(coef)	exp(-coef)	lower .95	upper .95
factor(treat)1	1.046	0.9562	0.8305	1.317
factor(sex)1	1.587	0.6301	1.2407	2.030
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

We have included all covariates in Table 9 and we observe the high p-value in `factor(treat)1` implying that have a treatment of prednisone is highly insignificant. We note also that the coefficient is almost zero, implying that one unit change in time have very little effect. All the other covariates has a low p-value, hence is considered as significant.
One observation that is interesting is how the sex does have an effect on the survival rate. We know the this covariate is significant and if we look at the confidense interval, we see that the lower 95 % of `factor(sex)1` is greater than 1. This imply that the death rate is higher for sex=1, i.e. for men. We do not know the reason for this but it might be because women are better at seeking treatment than men.
We observe from Figure 1 that patient with prednisone treatment overall may

have a slight positive effect. Studying the data further using other statistical approaches such as logrank test and Cox regression did not support this theory as the uncertainty is too high. Our conclusion is therefore that the treatment with prednisone can not be consider as effective or not.

1 Appendix

Code

```r
# 1a)
data = read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/v21/obliger/crabs.txt", header=T)



# 1b)
fit.binary.width = glm(y~width, data=data, family=binomial)
summary(fit.binary.width)

beta.1.width = 0.4972               # coefficient from fitted model
se.width = 0.1017                   # standard error from fitted model
upper.width = beta.1.width + 1.96*se.width
lower.width = beta.1.width - 1.96*se.width

OR.width = exp(beta.1.width)        # Odds ratio
upper.OR.width = exp(upper.width)   # upper CI
lower.OR.width = exp(lower.width)   # lower CI

print(OR.width)
print(lower.OR.width)
print(upper.OR.width)



# 1c)
fit.binary.weight = glm(y~weight, data=data, family=binomial)
summary(fit.binary.weight)

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

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



# 1d)
fit.binary.all = glm(y~width+weight+factor(color)+factor(spine), data=data, family=binomial)
summary(fit.binary.all)

fit.binary.ww = glm(y~width + weight, data=data, family=binomial)    # ww = width/weight
summary(fit.binary.simplified)

data.predictors = data.frame(width=data$width, weight=data$weight, color=data$color, spine=data$spine)
cor(data.predictors)



# 1e)
data.interaction = glm(y~width + spine + width*spine, data=data, family=binomial)
summary(data.interaction)



# 2a)
data = read.table("olympic.txt", header=T)



```

```r
# 2b)
fit.medals.all = glm(Total2000 ~ Total1996+Log.population+offset(Log.athletes)+GDP.per.cap,
      data=data, family=poisson)
summary(fit.medals.all)

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



# 3a)
data = read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/v21/obliger/
      cirrhosis.txt", header=T)

library(survival)
treat = survfit(Surv(data$time, data$status)~data$treat, conf.type="none")
sex = survfit(Surv(data$time, data$status)~data$sex, conf.type="none")
ascites = survfit(Surv(data$time, data$status)~data$asc, conf.type="none")
grouped.age = survfit(Surv(data$time, data$status)~data$agegr, conf.type="none")

plot(treat, lty=1:2, xlab="Time (days)",ylab="Survival")
legend(2700, 1, c("Prednisone","Placebo"),lty=1:2)

plot(sex, lty=1:2, xlab="Time (days)",ylab="Survival")
legend(3100, 1, c("Female","Male"),lty=1:2)

plot(ascites, lty=1:3, xlab="Time (days)",ylab="Survival")
legend(3100, 1, c("None","Slight","Marked"),lty=1:3)

plot(grouped.age, lty=1:3, xlab="Time (days)",ylab="Survival")
legend(3200, 1, c("<50","50-65",">65"),lty=1:3)



# 3b)
survdiff(Surv(data$time, data$status)~data$treat)

survdiff(Surv(data$time, data$status)~data$sex)

survdiff(Surv(data$time, data$status)~data$asc)

survdiff(Surv(data$time, data$status)~data$agegr)



# 3c)
fit = coxph(Surv(time, status==1)~factor(treat)+factor(sex)+factor(asc)+age, data=data)
summary(fit)
```