STK4900 Mandatory Assignment 2

Olivia Beyer Bruvik

Due 22/4/2021

Problem 1a

I have chosen to use a logistic regression model to study how the probability of presence of satellites depends on the explanatory variable width. Logistic regression is a suitable model for this data, because it models the probabilities within a range of {y | 0, 1}. In contrast, a linear model, such as an additive risk model, would give unreasonable values for probabilities, such as a negative probability or probabilities over 1.

To perform logistic regression, I will regress the explanatory variable the female crab's carapace width in cm onto the binary indicator y, which indicates whether one or more satellites were present.

The estimates for the intercept ($\hat{\beta}_0$) and width ($\hat{\beta}_1$) are -12.351 and 0.497, respectively. The p-values are 2.62e-06 and 1.02e-06 for the intercept and width coefficients, suggesting that these results are significant. The full model output is shown below.

This gives the fitted model below, from which we can estimate the probability that one or more satellites are present given a female crab's carapace width in centimeters.

$$\hat{p}(w) = \frac{e^{-12.351+0.497*w}}{1+e^{-12.351+0.497*w}} \tag{1}$$

$$\tag{2}$$

$$\text{where: } w = \text{the female crab's carapace width (cm)} \tag{3}$$

Problem 1b

Odds of presences of satellites

The odds of presences of satellites is the probability of presence of satellites divided by the probability of absence of satellites. The odds for a female crab with a carapace width w is given by:

$$\frac{p(w)}{1-p(w)} = e^{-12.351+0.497*w} \tag{4}$$

$$\tag{5}$$

$$\text{where: } w = \text{the female crab's carapace width (cm)} \tag{6}$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.3508177	2.6287179	-4.698419	2.6e-06
width	0.4972306	0.1017355	4.887482	1.0e-06

Odds ratio calculation

From this formula, we can calculate the odds ratio of presences of satellites between crabs that differ one cm in width as follows:

$$\frac{\frac{p(w+1)}{1-p(w+1)}}{\frac{p(w)}{1-p(w+1)}} = \frac{e^{\hat{\beta}_0+0.497(w+1)}}{e^{\hat{\beta}_0+0.497w}} = e^{0.497} \tag{7}$$

$$\text{where: } w = \text{the female crab's carapace width (cm)} \tag{8}$$

The odds ratio $e^{0.497}$ indicates that the odds of presences of satellites increases by a factor of $e^{0.497}$ per unit increase in the width of a female crab's carapace. In other words, a wider carapace is associated with higher odds of satellite presence. For example, the odds of presences of satellites for a crab with a carapace width of 29cm will be $e^{0.497}$ times higher than that of a crab with a carapace width of 28cm.

Odds ratio and relative risk

The odds ratio can be considered as an approximation to a relative risk when the probabilites of presence of satellites are small for both widths considered: $p(w_1) << 1$ and $p(w_2) << 1$. Therefore, this approximation is only valid for crabs with a width below a certain number of centimeters, for example 22 cm.

For crabs with carapace widths of 22 and 21 cm, the odds ratio is $e^{0.497} = 1.64378$, whereas the relative risk is $RR = \frac{p(22)}{p(21)} = 1.5179323$. This is a reasonable approximation.

In contrast, for crabs with carapace widths of 31 and 30 cm, the odds ratio is $e^{0.497} = 1.64378$, whereas the relative risk is $RR = \frac{p(31)}{p(30)} = 1.0287589$. This is not a reasonable approximation, and odds ratio can no longer be approximated to relative risk.

Confidence intervals for the odds ratio

The confidence intervals for the odds ratio are calculated as follows:

95% confidence intervals for the width coefficient, $\hat{\beta}_1 = 0.4972$:
$CI(\hat{\beta}_1) = \hat{\beta}_1 \pm 1.96 * se(\hat{\beta}_1)$
$CI(\hat{\beta}_1) = 0.4972 \pm 1.96 * 0.1017$
$CI(\hat{\beta}_1) : 0.299 - 0.697$

95% confidence intervals for the odds ratio (OR = 0.4972): $CI(OR) : e^{0.297868} - e^{0.696532}$
$CI(OR) : 1.3469 - 2.007$

The odds ratio with confidence intervals is therefore $1.3469 < 1.644 < 2.007$, and can be concluded to be significant as even the lower end of the confidence interval is more than one.

Problem 1c

Next, I looked at each covariate individually. The results for each logistic regression model are summarized in the table below, where covariates are given in the column labeled "Cov".

Categorical and numerical covariates

The "Cat" column gives information about whether the covariate was included as a categorical or numerical value. As can be seen in the table, the width and weight covariates were included as numerical covariates because they are continuous (ie. measured in cm and kg). Color and spine were included as categorical covariates.

	Cov	Cat	Estimate	SE	z.value	p.value	OR	Lower	Upper
(Intercept)	Width	No	-12.351	2.629	-4.698	0.000	0.000	0.000	0.000
width	Width	No	0.497	0.102	4.887	0.000	1.644	1.346	2.008
(Intercept)1	Weight	No	-3.695	0.880	-4.198	0.000	0.025	0.004	0.140
weight	Weight	No	1.815	0.377	4.819	0.000	6.141	2.933	12.857
(Intercept)2	Color	Yes	1.099	0.667	1.648	0.099	3.001	0.812	11.092
factor(color)2	Color	Yes	-0.123	0.705	-0.174	0.862	0.884	0.222	3.520
factor(color)3	Color	Yes	-0.731	0.734	-0.996	0.319	0.481	0.114	2.027
factor(color)4	Color	Yes	-1.861	0.809	-2.301	0.021	0.156	0.032	0.762
(Intercept)3	Spine	Yes	0.860	0.360	2.392	0.017	2.363	1.167	4.785
factor(spine)2	Spine	Yes	-0.994	0.630	-1.577	0.115	0.370	0.108	1.272
factor(spine)3	Spine	Yes	-0.265	0.407	-0.651	0.515	0.767	0.345	1.703

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.3547261	3.5280465	-2.651531	0.0080128
weight	0.8337917	0.6716445	1.241418	0.2144513
width	0.3067892	0.1819473	1.686143	0.0917682

Selection of variables

The weight covariate is statistically significant ($p < 0.05$) with an intercept of -3.695, a $\hat{\beta}_1$ of 1.815, and an odds ratio of $2.933 < 6.141 < 12.857$. This suggests that weight has a significant influence on the presence of satellites.

The color variable is only significant for color 4, with a p-value of 0.021, $\hat{\beta}_1$ of -1.861, and an odds ratio of $0.032 < 0.156 < 0.762$. The odds ratio is below one, suggesting that a change from color 1 to color 4 will decrease the odds of presence of satellites. This suggests that while changing the color from color 1 to 2 or 3 has no influence on the presence of satellites, changing the color to color 4 does have a slight albeit limited influence.

From the table, we can see that changing from spine 1 (both good) to spine 2 or spine 3 has no significant influence on the presence of satellites ($p > 0.05$).

Problem 1d)

I have chosen to include weight and width as covariates in my model. The spine variable had no significant influence on the presence of satellites. Further, the color variable had a slight influence on the presence of satellites (color 1 vs 4); however, I will not include this variable for simplicity.

To perform logistic regression, I will regress the explanatory variables, the female crab's weight (kg) carapace width (cm) onto the binary indicator y, which indicates satellite presence. The binomial family is used.

As can be seen in the summary above, the neither weight (p=0.214) not width (p=0.092) are significant covariates in this model. This may be due to correlation between width and weight of female crabs in this dataset. The Pearson's correlation coefficient between the two variables are 0.8868715). This is a significant correlation and has, as can be seen in the p-values, impacted the model as variables should be independent.

Problem 1e

Here, interactions between covariates will be explored. As can be seen in the summary below, there does not appear to be any significant interactions between the variables ($p > 0.05$). Individual interactions between different covariates were also analyzed independenlty, but no significant interactions were detected.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.6158915	12.6735204	0.2853107	0.7754061
weight	-3.7570156	5.5786389	-0.6734646	0.5006518
width	-0.1997760	0.4972949	-0.4017254	0.6878861
factor(spine)2	17.4498934	2774.8082334	0.0062887	0.9949824
factor(spine)3	-18.2710337	3956.1804467	-0.0046184	0.9963151
factor(color)2	-0.1888964	0.9812751	-0.1925009	0.8473498
factor(color)3	16.4232299	2206.1657976	0.0074442	0.9940604
factor(color)4	-18.3126604	3956.1804273	-0.0046289	0.9963067
weight:width	0.1790628	0.2077877	0.8617584	0.3888205
factor(spine)2:factor(color)2	-17.8098241	2774.8083652	-0.0064184	0.9948789
factor(spine)3:factor(color)2	18.9586343	3956.1804903	0.0047922	0.9961764
factor(spine)2:factor(color)3	-34.6151202	3544.9582293	-0.0097646	0.9922091
factor(spine)3:factor(color)3	1.6844825	4529.7384864	0.0003719	0.9997033
factor(spine)2:factor(color)4	-16.1300093	6245.1810615	-0.0025828	0.9979392
factor(spine)3:factor(color)4	35.5543688	5594.8840151	0.0063548	0.9949296

Problem 2a

Poisson regression

Poisson regression is a generalized linear model: $Y_i \sim P_o(\lambda_i)$, where $\lambda_i = \lambda(x_{1i}, x_{2i}, ..., x_{pi}) = e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+...+\beta_p x_{pi}}$

A Poisson regression model is suitable to model the olympic data because winning an olympic gold medal is an event that happens randomly over time. In this model, the outcome variable is the total number of gold medals per country in the 2000 Olympics and the predictors are the log-transformed population per country, GDP per capita per country and total number of gold medals won in the 19996 Olympics per country.

Offset terms

Log.athletes is included in the model as an offset term, a covariate where the regression coefficient is known to equal 1. Log.athletes is a sensible choice for such an offset because it is highly correlated with Total2000 with a pearson correlation coefficient of 0.682. The Log.athletes coefficient for a Total2000~Log.athletes regression model is 1.236, which is close to 1.0. By including Log.athletes on our model, we are modelling the following: $Y_i \sim P_o(a_i\lambda_i)$, where a_i is the number of athletes representing country i, given by exp(Log.athletes).

Model variables

The model, summarized below, shows that Total1996 (p = 1.79e-13), the total gold medals won in the 1996 Olympics, and GDP.per.cap (p = 3.29e-06), the GDP per capita, are significant predictors of the total gold medals won in the 2000 Olympics

Rate ratios

The rate ratio (RR) for a specific coefficient β_x in a Poisson regression model is given by $RR = e^{\beta_x \Delta}$, where β_x is a coefficient and Δ is the difference between the values for two countries for coefficient β_x.

The rate ratio for the Total1996 coefficient in this model is $RR = e^{0.0118\Delta=1.01187\Delta}$, holding all else constant. This ratio is above 1, suggesting that the number of medals won at the 1996 is positively associated with the number of medals won at the 2000 Olympics.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.8622989	0.3190759	-8.9705885	0.0000000
Total1996	0.0118319	0.0016067	7.3639547	0.0000000
Log.population	0.0275103	0.0315391	0.8722609	0.3830661
GDP.per.cap	-0.0149242	0.0032083	-4.6517395	0.0000033

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.1109261	0.1305242	-31.495520	0.0000000
Total1996	0.0090753	0.0027766	3.268459	0.0010813
GDP.per.cap	-0.0181694	0.0070697	-2.570041	0.0101687

Problem 2b)

Choosing variables

I used an analysis of variance table to select the most reasonable model for the data. I started with this model: log_medals_per_athlete ~ offset(Log.athletes) + Total1996. I then added various variables. As can be seen below, the addition of the GDP.per.cap variable to the model was significant (p = 0.009).

```
## Analysis of Deviance Table
##
## Model 1: log_medals_per_athlete ~ offset(Log.athletes) + Total1996
## Model 2: log_medals_per_athlete ~ offset(Log.athletes) + Total1996 + GDP.per.cap
## Model 3: log_medals_per_athlete ~ offset(Log.athletes) + Total1996 + Log.population +
##     GDP.per.cap
## Model 4: log_medals_per_athlete ~ offset(Log.athletes) + Total1996 + GDP.per.cap +
##     GDP.per.cap * Log.population
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        64     42.526
## 2        63     35.691  1   6.8348  0.00894 **
## 3        62     35.686  1   0.0052  0.94255
## 4        61     35.170  1   0.5155  0.47277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final model

This model indicates that the rate ratio corresponding to one unit's increase in the number of medals won at the 1996 Olympics is $RR = e^{0.009\Delta} = 1.009\Delta$, holding all else constant. The rate ratio is greater than 1, suggesting that the number of medals won at the 1996 is positively associated with the number of medals won per log athlete at the 2000 Olympics.

The rate ratio for gdp per capita is $RR = e^{-0.018\Delta} = 0.982161\Delta$, holding all else constant. This rate ratio is slightly less than one, suggesting that the gdp per capita is negatively associated with the number of medals won per log athlete at the 2000 Olympics.

Problem 3a

The Kaplan-Meier plots below shows the survival functions for each level of the covariates treatment, sex, ascites and age group. The levels for each covariate are indicated in the legend on each plot.

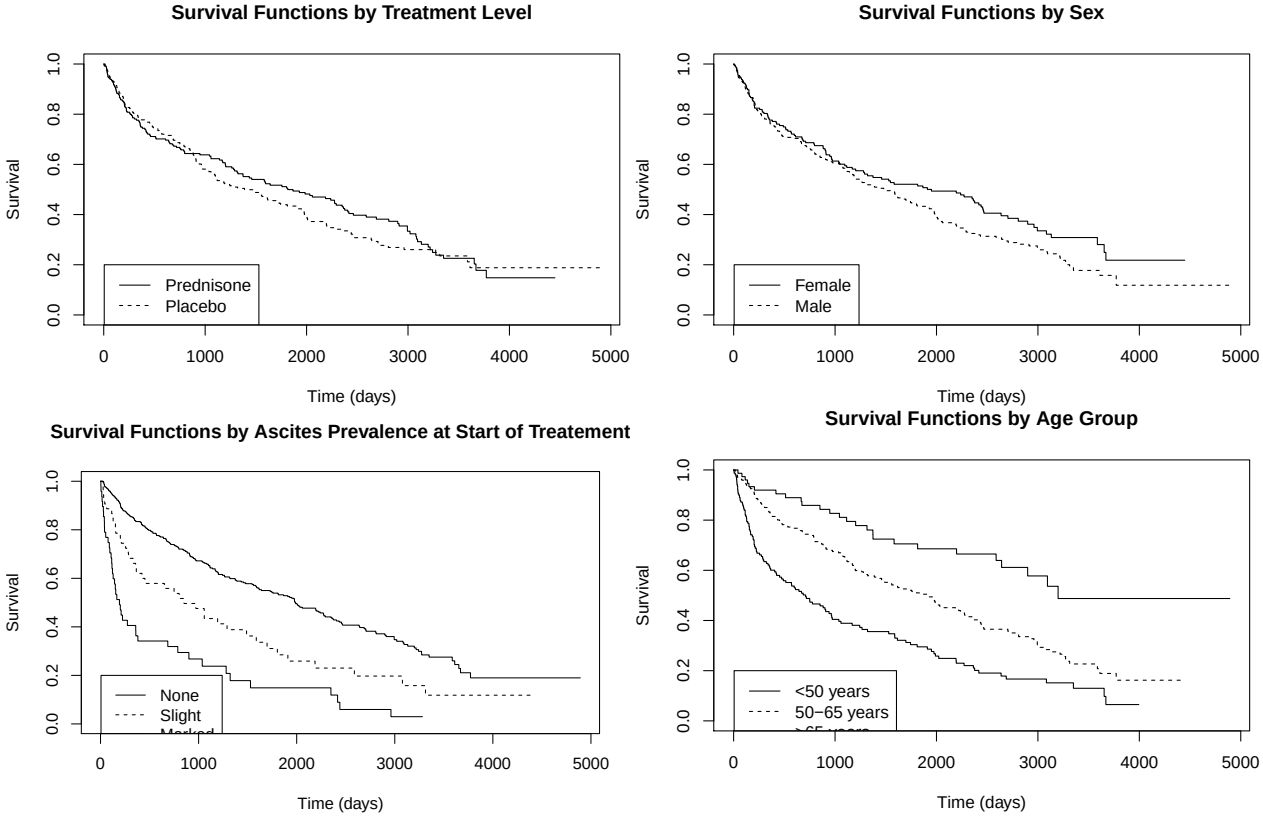
The first plot shows the survival functions for two different treatment groups: the group that received Prednisone and the the group that received a placebo. The plot indicates that the group on prednisone

treatment has a similar survival time as compared with the group on placebo. The survival for the prednisone has a more steeper slope than the placebo for t $<$ 250, but intersects the placebo survival function at around t = 950. The survival decreases with approximately the same rate for the two groups between 1000 $<$ t $<$ 3000, but intersects again at t = 3300. This suggests that while prednisone treatment does not increase survival, it may increase the median time to the event death, albeit slightly.

The second Kaplan-Meier plot shows the survival functions for female and male subjects. The plot indicates that the survival function is very similar for both sexes for t $<$ 1000 days; however, the survival function for female subjects is less steep than the survival function for male subjects. This suggests that liver cirrhosis may be more deadly for male than female patients. An interesting point to note is the drop in survival for females at t = 3500 days. The plot suggests that a significant proportion of the female study participant either left the study at this day or died.

The third figure plots the survival function by the prevalence of ascites at the start of treatment, grouped by none, slight and marked. The survival functions become steeper as the prevalence of ascites increases. As such, the median time to death is significantly longer for the group in the none level than the two other: $\hat{S}_{marked}(t_m) < \hat{S}_{slight}(t_m) < \hat{S}_{none}(t_m)$, where t_m is the median time to death.

The fourth plot gives the survival functions by age group. As can be seen in the plot, while all group start with a survival of 1, the survival decreases markedly faster as the age range increases: $\hat{S}_{<50y}(t_m) < \hat{S}_{50-65y}(t_m) < \hat{S}_{>65y}(t_m)$, where t_m is the median time to death. This suggests that the hazard is greater for older patients.



Problem 3b

We can test the null hypothesis that the survival function is the same for two groups, $H_0 : S_1(t) = S_2(t)$ for all t, with the Logrank test.

Logrank test for the treatment covariate

The Logrank test indicates that there is no significant difference ($P < 0.05$) in the survival functions of the treatment covariate levels of Prednisone and Placebo. The null hypothesis remains and we have not detected a signficant benefit of using Prednisone on survival.

```
## Call:
## survdiff(formula = Surv(time, status) ~ treat, data = cir_data)
##
##           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
```

Logrank test for the sex covariate

The Logrank test indicates that sex does not have a significant impact on the survival function; however, the p-value is close to the level of significance (p=0.06), suggesting that this may be an interesting point for future studies.

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex, data = cir_data)
##
##         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
```

Logrank test for the ascites covariate

The Logrank test rejects the null hypothesis that the survival function is the same for groups across levels of the ascites covariate (p = 7e-16). This confirms the graphical interpretation that the prevalence of ascites at the start of treatment significantly influences the survival function.

```
## Call:
## survdiff(formula = Surv(time, status) ~ asc, data = cir_data)
##
##         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
```

Logrank test for the age group covariate

The Logrank test rejects the null hypothesis that the survival function is the same across age groups (p = 1e-11). This, like that of the ascites covariate, confirms the graphical interpretation that age group significantly influences the survival function.

```
## Call:
## survdiff(formula = Surv(time, status) ~ agegr, data = cir_data)
##
##           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
```

Problem 3c

I next fitted a Cox regression with the following covariates: treatment, sex, ascites and age.

The treatment covariate was not signficant in the model (p = 0.704), suggesting that receiving Prednisone vs place treatment does not signficantly alter the time to event.

The sex covariate (0=female, 1=male) significantly altered the time to event with a coefficient of 0.462287 (p=0.000228). This coefficient is positive and therefore indicates a worse prognosis over time, consistent with the graphical interpretation above and the hazard ratio, discussed below.

The ascites covariate has a significant coefficient of 0.595150 (p=6.86e-13). This indicates that the prognosis worsens as we move up the levels of the covariate, from no ascites to slight and marked ascite prevalences at the start of treatement.

The age covariate is also significant (p = 8.34e-13) with a coefficient of 0.0489. This suggests that prognosis increases as age increases with a hazard ratio of $e^{0.0489} = 1.050$.

```
## Call:
## coxph(formula = Surv(time, status == 1) ~ treat + sex + asc +
##     age, data = cir_data)
##
##   n= 488, number of events= 292
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## treat 0.044637  1.045648 0.117610 0.380 0.704293
## sex   0.462287  1.587702 0.125406 3.686 0.000228 ***
## 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
## 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
##
## 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
```

Hazard ratio for the sex covariate

The hazard ratio for men vs women when all other variables are constant is given by $HR = e^{\hat{\beta}}$, where $\hat{\beta}$ can be found with the proportional hazard function: $h(t|x) = h_0(t)exp(\beta x)$, where x = 0 for female and x = 1 for male).

The hazard ratio for men vs women when all other variables are constant is 1.588 with 95% confidence interval limits of (1.2417, 2.030). This hazard ratio of 1.255 corresponds to the baseline hazard for men over the baseline hazard for women, suggesting that cirrhosis poses a greater hazard for men.

Conclusion on the effect of prednisone in this trial

From the information above, it can be concluded that the treatment choice of prednisone or placebo did not significantly alter the survival function in this trial. This conclusion is based on the failure to detect a significant influence of prednisone on proportional hazard in the cox regression model (p = 0.704), alongth with the failure to reject the null hypothesis that there is a difference between the two treatment groups (p = 0.4).

This does not prove that there is no effect of prednisone; it is merely an indication that no significant effect of prednisone was detected in this trial.