COMPULSORY ASSIGNMENT 2
STK4900: STATISTICAL METHODS AND APPLICATIONS

ELISABETH CHRISTENSEN

Problem 1:

a) Since we have a binary dependent variable y then logistic regression will be our best approach to describe a model to fit the data. The logistic model regression can also be used to describe one or more predictors, as is in our case. The model will then follow a logistic distribution as described by the logistic function

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)} \tag{1}$$

By applying the logistic regression model to our data with only width as a covariate. In doing so we end up with the table 1. Here, we see that the value width is significant for the binary data with a p-value of $p = 1.02e{-}06$. Thus, we can reject the null-hypothesis where we assume the predictor width will not have an impact on the variance of y. We can also plot the distribution between width and y, where we fit a logistic regression model to the data. This is shown in figure 1, where we see our regression model clearly follows that of a logistic distribution.

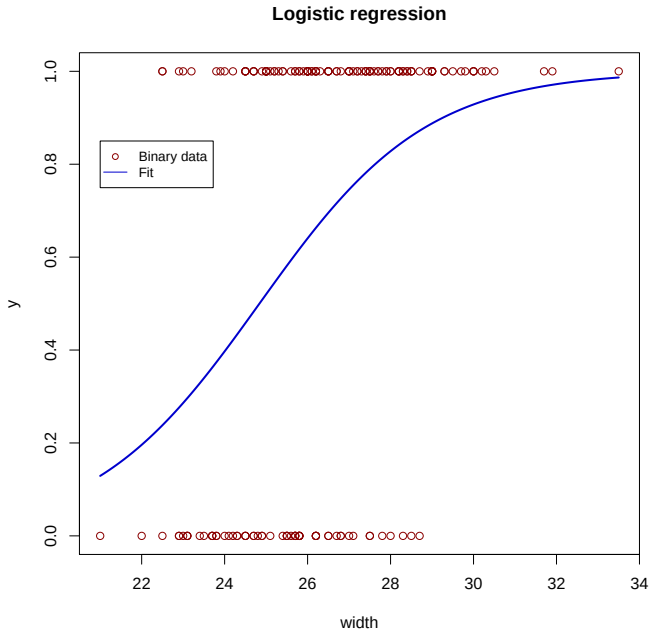

FIGURE 1. Logistic regression model of binary data y with width as only predictor

TABLE 1. Summary of the logistic regression model when only considering the numerical covariate 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

TABLE 2. Odds ratio of the presences of satellites between crabs that differ one cm in width with a 95% confidence interval

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

TABLE 3. y∼weight

	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

b) The odds ratio (OR) describes how much more probable the outcome of a variable is depending on the change of one predictor. OR can be calculated between two subjects with covariate values $x + \Delta$ and x as

$$\begin{aligned} OR &= \frac{p(x_1)/[1-p(x_1)]}{p(0)/[1-p(x_0)]} \\ &= \frac{\exp(\beta_0 + \beta_1(x+\Delta))}{1+\exp(\beta_0+\beta_1 x)} \\ &= e^{\beta_1 \Delta} \end{aligned} \tag{2}$$

Using eq. 2 we can calculate the OR with a 95% confidence interval for the presences of satellites between crabs that differ one cm in width. This is shown in table 2, where we see that the $OR = 1.644$. I.e. the likelihood of one or more satellites is 1.6 times larger if the width of the carapace of a female crab increases by one cm. Meaning the exposure of the predictor width is associated with a higher odds of outcome for one or more satellites. Our confidence interval for this value is $CI \in [1.347, 2.007]$. Compared to the relative risk (RR), we know that since the odds ratio is greater than 1, then $1 < RR < OR$, such that RR is also greater than 1, but lower than OR. Thus, since our odds ratio is only 1.6 this must mean that $RR \in (1, 1.6)$, and we can/cannot thus consider the odds ratio as an approximation to the relative risk.

c) Since weight is only measured by one category (weight in kg), and not separated into different weight groups, it can be included as a numerical variable, just as with width. The two other explanatory variables, color and spine, are on the other hand separated into different groups dependent on the color of a female crab and the conditions of the spine, should thus be included as categorical.

Running the same tests as before, we can fit a logistic regression model to each of the covariates to check the dependency of the data compared to single predictors. In doing so, we get the values as shown in table 4.

As seen here, each predictor seems to have a significant influence on the variance of the data set, however not all groups of the categorical variables color and spine are significant. In this case, we see that only the female

TABLE 4. Summary of each predictor evaluated as a logistic regression model individually

		Estimate	Std. Error	z-value	Pr(>|z|)
$y \sim$ width	(Intercept)	-3.6947	0.8802	-4.198	2.70e-05
	weight	1.8151	0.3767	4.819	1.45e-06
		Estimate	Std. Error	z-value	Pr(>|z|)
$y \sim$ factor(color)	(Intercept)	1.0986	0.6667	1.648	0.0994
	color_group2	-0.1226	0.7053	-0.174	0.8620
	color_group3	-0.7309	0.7338	-0.996	0.3192
	color_group4	-1.8608	0.8087	-2.301	0.0214
		Estimate	Std. Error	z-value	Pr(>|z|)
$y \sim$factor(spine)	(Intercept)	0.8602	0.3597	2.392	0.0168
	spine_group2	-0.9937	0.6303	-1.577	0.1149
	spine_group3	-0.2647	0.4068	-0.651	0.5152

TABLE 5. Treating all variables in the same logistic regression model, where $y \sim$weight+width+factor(color)+factor(spine)

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-8.06501	3.92855	-2.053	0.0401
weight	0.82578	0.70383	1.173	0.2407
width	0.26313	0.19530	1.347	0.1779
color_group2	-0.10290	0.78259	-0.131	0.8954
color_group3	-0.48886	0.85312	-0.573	0.5666
color_group4	-1.60867	0.93553	-1.720	0.0855
spine_group2	-0.09598	0.70337	-0.136	0.8915
spine_group3	0.40029	0.50270	0.796	0.4259

crabs with the darkest color and those with a condition on the spine where both are good are considered as the predictors which allows us to neglect the null-hypothesis H_0, due to their p-values $p = 0.0214$ and $p = 0.0168$ respectively.

d) Considering all predictors in the same regression model, we get the values as shown in table 5. Here, we clearly see the significance of the intercept, i.e. of color group 1 and spine group 1. Even though each variable has an estimate different from zero, they cannot be considered significant enough to have an effect on the outcome of the variable y.

However, we can try and run an analysis of variance to find the most compatible regression model with the data. From table 5, we see that the variable width has a lower p-value than the variable weight, and we can thus try to first test the variable width and expand the logistic regression models with the predictors with the lowest p-values for each run. An example of this is shown in table **??**. From this, we see that the only significant regression model is model 2 with a p-value of $p = 7.041e - 07$.

e) Continuing with the best fit model from above we can extend with interactions between the covariates, and run an analysis of variance to once more check the best model, as seen in table 7. Here, we clearly see the significance of model 2 with a p-value $p < 2e-16$ where we have an interaction between the weight and width. This would also make sense seen as when the weight of a horseshoe crab increases so too will the width of the carapace. The best logistic regression fit to our data set is thus model 2.

TABLE 6.
Model 1: $y \sim$factor(color)
Model 2: $y \sim$ width + factor(color)
Model 3: $y \sim$ width + factor(color) + factor(spine)
Model 4: $y \sim$ width + factor(color) + factor(spine) + weight

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	169	212.06			
2	168	187.46	1	24.6038	7.041e-07
3	166	186.61	2	0.8451	0.6554
4	165	185.20	1	1.4099	0.2351

TABLE 7.
Model 1: $y \sim$ factor(color)+width
Model 2: $y \sim$ factor(color)+width+weight:width
Model 3: $y \sim$ factor(color)+width+weight:width+width:factor(color)
Model 4: $y \sim$ factor(color)+width+weight:width+width:factor(color)+ weight:factor(color)
Model 5: $y \sim$ factor(color)+width+width:weight+width:factor(color)+ weight:factor(color)+weight:factor(color):factor(spine)

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	168	187.457			
2	167	31.797	1	155.660	<2e-16
3	164	181.702	3	-149.906	
4	161	176.923	3	4.779	0.1887
5	153	165.633	8	11.291	0.1858

Problem 2:

a) A Poisson regression model can be thought of as a generalized linear model and is often used when dealing with *count data* happening at random over time, when given one or more independent variables. The assumptions for a Poisson process is that the rate[1] of events λ is constant, that the events are independent of one another and that they occur separately. We can see the Poisson distribution from the count data in the Olympics from both 1996 and 2000 in figure 2.

When the observations are aggregated counts, i.e. grouping of observations, we can express the regression model for a Poisson process as

$$\begin{aligned} E(Y_i) &= w_i\lambda_i \\ &= w_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_n x_{ni}) \\ &= \exp(\log(w_i) + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_n x_{ni}) \end{aligned} \tag{3}$$

where y_i is observation no. i, w_i is the number of subjects in group i (also known as the weight), and $\log(w_i)$ is the so-called offset where the regression coefficient is known to equal 1.
When trying to simplify the model in this case we can turn to the log.athletes variable as our offset. This is due to the fact that this is the only variable which allows a country to have a chance at getting at medal, our dependent variable, meaning without any athletes to compete for a country, there is no probability for that country to get a medal. The two other variables

[1] expected number of events per unit time

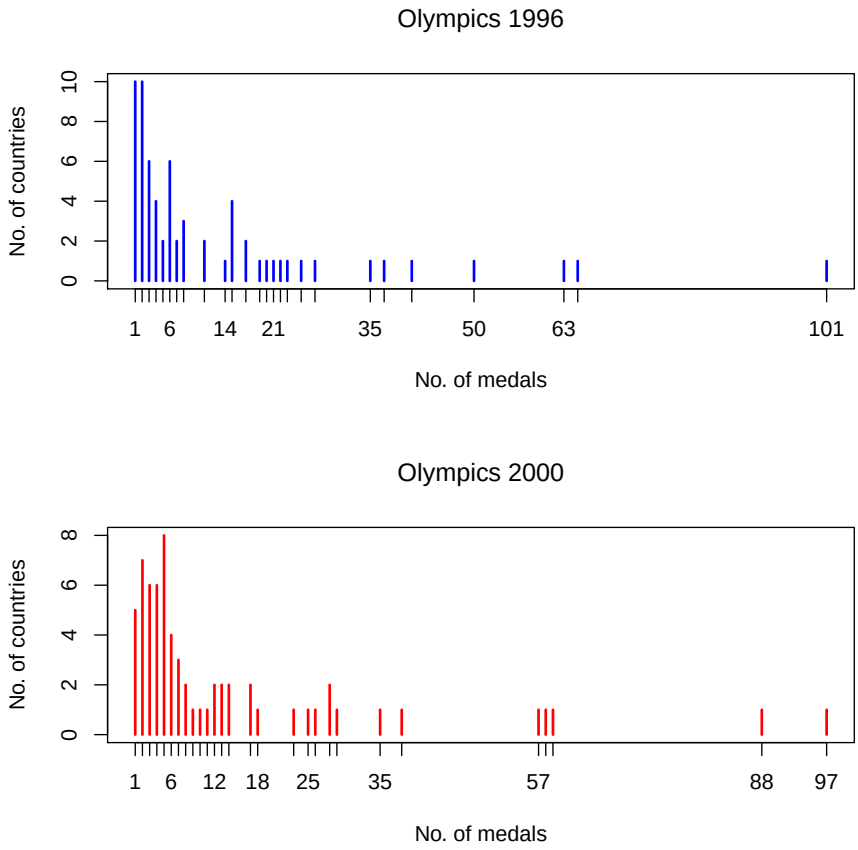

FIGURE 2. Poisson distribution of the amount of medals gained by a country in the Olympics from years 2000 and 1996.

TABLE 8. medals2000 $\sim$ offset(log.ath) + log.pop + GDP

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-4.255144	0.250782	-16.968	<2e-16
log.pop	0.179605	0.022466	7.995	1.3e-15
GDP	-0.004340	0.002726	-1.592	0.111

GDP and log.pop will affect the number of athletes competing, but only log.athletes have a direct impact on the amount of medals gained.

b) With the amount of athletes as the intercept, we see from table 8 that it is significant on all significant levels with a value of $p < 2e-16$. We also see that the logarithmic size of the populations included is also highly significant with $p = 1.3e-15$, allowing the probability of a nation getting a medal 0.18 times higher if the logarithm of a population size increases by 1. While for GDP we see that it is not significant with a p-value of $p = 0.111$.

We can simplify the regression model through the maximum likelihood estimation. I.e. by running an analysis of variance we can find the model which best fits the distribution by looking at the p-value for each model. This can be seen in table 9, where we have compared three different models with medals2000 $\sim$ offset(log.ath) as our null-model. Here, we see that

TABLE 9.
Model 1: medals2000 ∼ offset(log.ath)
Model 2: medals2000 ∼ offset(log.ath) + log.pop
Model 3: medals2000 ∼ offset(log.ath) + log.pop + GDP

	Resid. Df	Resid. Dev	Deviance	Df	Pr(>Chi)
1	65	254.11			
2	64	190.35	1	63.766	1.401e-15
3	63	187.80	1	2.552	0.1101

TABLE 10.
Model 1: medals2000 ∼ offset(log.ath) + log.pop Model 2: medals2000 ∼ offset(log.ath) + log.pop + GDP + log.pop:log.ath
Model 3: medals2000 ∼ offset(log.ath) + log.pop + GDP + log.pop:log.ath + log.pop:GDP

	Resid. Df	Resid. Dev	Deviance	Df	Pr(>Chi)
1	64	190.35			
2	62	173.13	2	17.213	0.0001829
3	61	173.09	1	0.0497	0.8236118

model 2 gives the best fit as it has a p-value of p = 1.401e-15. This is as we expected as seen from above where only the amount of athletes and the population size were significant for the count data, while GDP could be ruled out. We can further improve our model by including interactions between the different covariates, as seen in table 10, with the null model as the previous best fit. Here it is clear that model 2 from table 10 is the most significant model, ith $p = 0.000183$ with an interaction between the population size and the amount of athletes. This makes sense as more athletes are available for a country with a large population size, giving them a higher probability to get a medal, than a country with a smaller population size. So in conclusion larger countries will have a significantly greater chance at earning a medal at the Olympics, but the gross domestic product of the nation is not significant enough to be taken into account when finding the best possible regression model. Thus, only the population size has a great enough significance to account for some of the variances in the data set.

Problem 3:

a) When analyzing survival data when can visualize the data set using Kaplan-Meier plot. This can be seen in figures 3, 4, 5 and 6. In figure 3 we see that there is no real distinction between the two treatments of prednisone vs inactive placebo. Meaning, the prednisone treatment has no significant impact on the lifetime of a patient affected by cirrhosis. This can also be seen in table 11, showing the number of events during the trial period, and the median survival time for a patient during the two treatments. Here, we see that the number of events are roughly the same,although there is a difference in the median survival time of about 426 days. When comparing between the genders, we see that there is neither a real significant difference in the survival time between males and females. This can also be seen in table 11 where the median survival time for men and women are 1950 days and 1480 days respectively.

TABLE 11. Number of events and the median lifetime of patients in days within a certain category.

	n	Events	Median	0.95LCL	0.95UCL
treat=0	251	142	1814	1322	2376
treat=1	237	150	1388	1078	1910
sex=0	198	111	1950	1281	2461
sex=1	290	181	1480	1173	1909
agegr=1	80	26	3199	2642	NA
agegr=2	250	148	1909	1459	2277
agegr=3	158	118	710	453	972
asc=0	386	211	1985	1641	2351
asc=1	54	39	851	365	1698
asc=2	48	42	193	122	686

While in the Kaplan-Meier plots in figure 5 shows a significant difference in the median survival times of the different age groups. Here, patients less than 50 years seem to be the least vulnerable to ascites, with a median survival time of 3199 days, while patients between the age of 50-65 years have a median survival time of 1480 days. The most exposed patients are those above the age of 65, with a median survival time of 710 days.

We can also see the significance of the degree of ascites at the beginning of a treatment. If a patient has no sign of ascites then his/her survival time is about 1985 days. While for a patient diagnosed with a slight degree of ascites have a median survival time of about 851 days, meaning their time is less than half the survival time of those without any ascites. While those who are marked only have a survival time of 193 days, less than 10% compared to those without any sign of ascites.

b) Using the logrank test we can find the most significant covariates. An overview of the tests can be found in table 12. Here we see that the only significant covariates are the age groups and the degree of ascites with p-values $p = 7\text{e-}16$ and $p = 1\text{e-}11$ respectively, as expected from our Kaplan-Meier plots. Neither treatments prednisone and placebo nor the gender of the patients has a significant impact on the survival time of the patients, each with a p-value $p = 0.4$ and $p = 0.06$ respectively.

c) Furthermore, performing a multiple Cox regression allows us to find the Hazard ratio, i.e. the risk of failure, which can be expressed as

$$HR = \frac{h(t|x+\Delta)}{h(t|x)} = \frac{h_0 \exp(\beta(x+\Delta))}{h_0 \exp(\beta x)} = e^{\beta\Delta} \tag{4}$$

In doing so, we can find the hazard ratio where the effects of all covariates are studied simultaneously with a 95% confidence interval, as seen in table 13. Here, we see that the hazard ratio for men and women are $HR = 1.587$ and $HR = 0.6301$. I.e. the risk of failure is greater among men than women, as can also be seen from figure 4, where men have a lower median survival time than women.

In conclusion, the effects of prednisone on cirrhosis is not significant enough to have a greater impact on the survival time than an inactive placebo treatment. The only covariates significant enough to make a difference in the survival time is the age and degree of ascites of each patient during the start of the treatment.

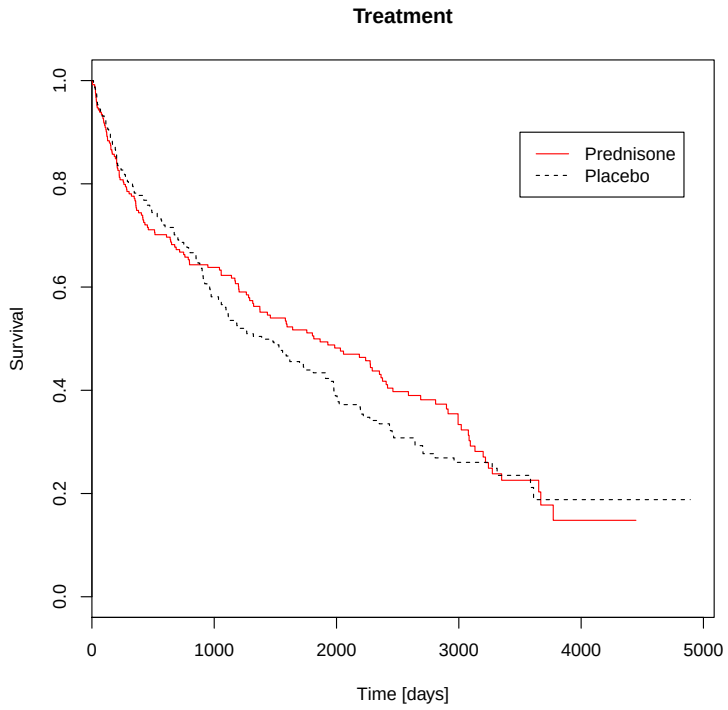

FIGURE 3. Comparison between patients treated with the hormone prednisone and patients treated with an inactive placebo.

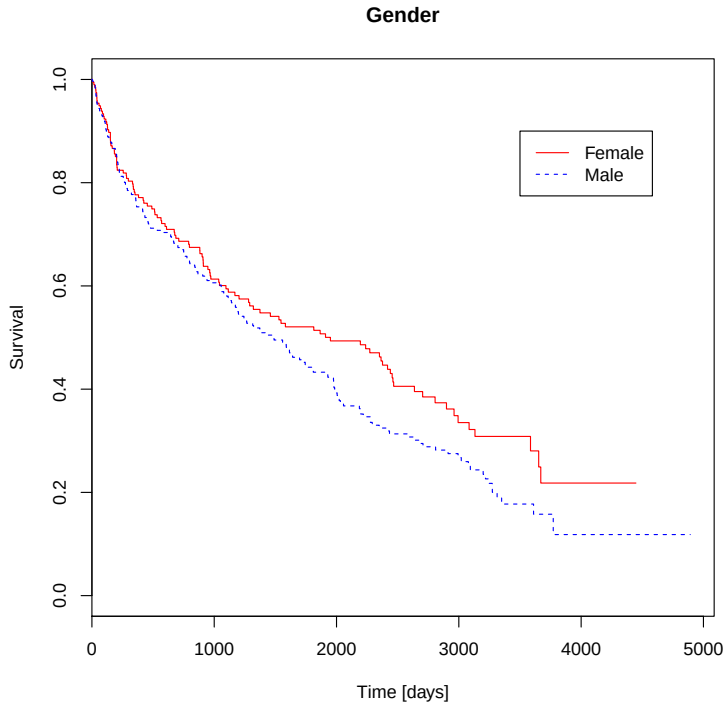

FIGURE 4. Comparison between females and males.

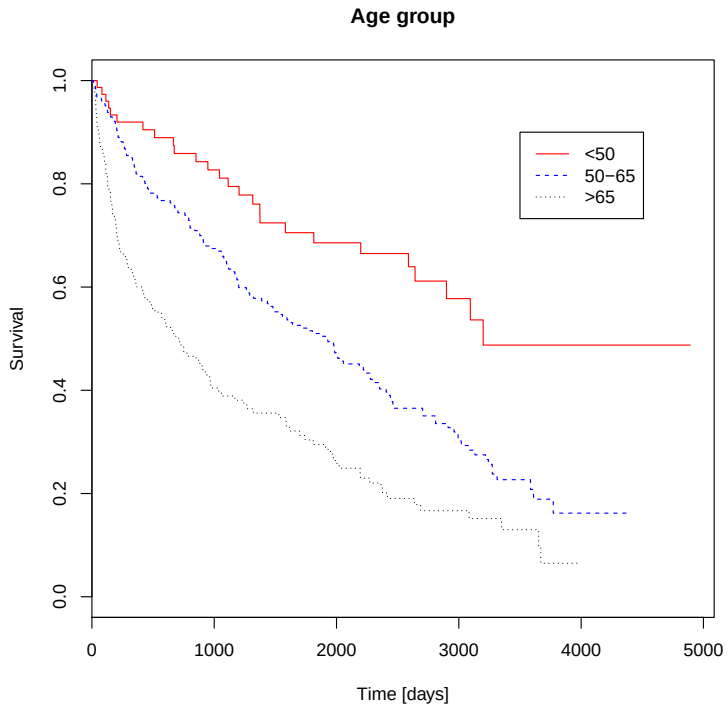

FIGURE 5. Kaplan-Meier plots of the age groups 1<=50, 2=50-65 and 3=>65.

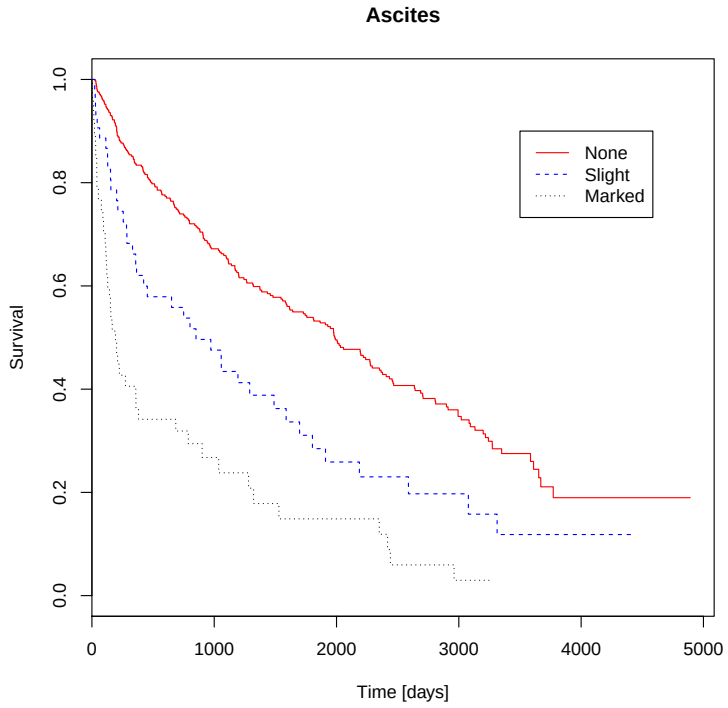

FIGURE 6. Kaplan-Meier plots visualizing the different degrees of ascites at the start of the treatment.

TABLE 12. Logrank test of each covariate

	χ^2	p-value
treat	0.7	0.4
sex	3.5	0.06
agegr	69.9	7e-16
asc	50.6	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
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
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

TABLE 13. Multiple Cox regression considering the cases of men and women separately

Likelihood ratio test= 109.3 on 5 df, p=<2e-16
Wald test = 115.4 on 5 df, p=<2e-16
Score (logrank) test = 123.9 on 5 df, p=<2e-16

	exp(coef)	exp(-coef)	lower .95	upper .95
factor(sex == 0)	0.6301	1.5871	0.4926	0.806
factor(treat)1	1.0458	0.9562	0.8305	1.317
age	1.0501	0.9523	1.0361	1.064
factor(asc)1	1.8285	0.5469	1.2975	2.577
factor(asc)2	3.2781	0.3051	2.3252	4.621

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