MATH6143 coursework

Etienne Latif

October 21, 2023

1 Question 1

1.1 1(a)

The survivor (or survival) function gives the probability of an individual surviving to a given time — a lower value (close to 0) implies a very low probability of survival while a high value (close to 1) conveys the opposite. This function necessarily (and intuitively) is non-increasing, meaning it always remains the same or decreases as time progresses.

We have used the Kaplan-Meier estimates of the survivor function to compare the survival probabilities between the asymptomatic and symptomatic groups in the lymphoma data set. The results are displayed in Figure 1. Due to page limit constraints, the exact values and standard errors of each estimate can not be displayed explicitly in this report. The relevant rows needed to answer the questions in this project are displayed in Tables 1.1 and 1.2 for asymptomatic and symptomatic groups respectively. The estimated survival function decreases much more rapidly for the symptomatic group, implying the probability of survival as time passes gets lower more quickly for patients showing symptoms of the disease. This seems intuitive as we would expect Non-Hodgkin's lymphoma (a type of cancer) patients to die earlier than individuals without the disease.

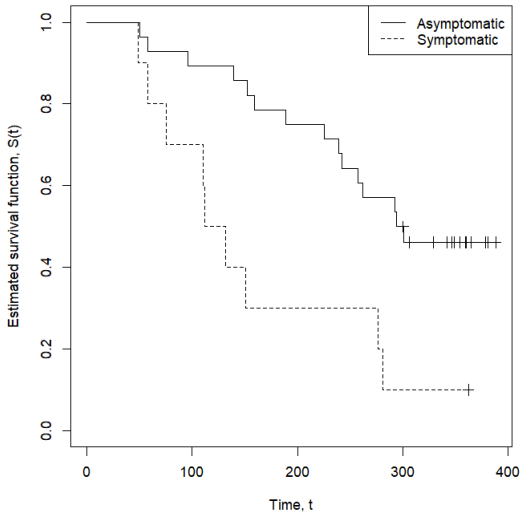

Figure 1: Estimated survival distribution by symptom group on the lymphoma data set using the Kaplan-Meier estimator. Crosses denote censored values.

1.2 1(b)

A 95% confidence interval is a range of values between which we would expect 95% of our estimates to lie between if we repeatedly perform an experiment. In this context, a 95% confidence interval for the survival function at a given time is a range of values for which 95% of our estimated survival functions would lie between for different sample populations.

Here we present 95% confidence intervals for the probability of 6-year (312 week) survival for the two groups. The last rows of the survival function estimate tables for the asymptomatic group and symptomatic group in Tables

1.1 and 1.2 respectively. As our estimated survival functions are step-wise functions and the time of interest (312 weeks) lies beyond the last time in the domain of both our functions, we take the 95% confidence intervals given in the last row of each table as our 95% confidence intervals for the probability of 6-year survival. Therefore the confidence interval for the asymptomatic group is (0.345, 0.724) and for the symptomatic group it is (0.0156, 0.642).

(1.1)

Time	**Estimated survival function**	**Standard error**	95% Confidence interval
...	...	...	...
301	0.462	0.0947	(0.309, 0.690)

(1.2)

Time	**Estimated survival function**	**Standard error**	95% Confidence interval
...	...	...	...
132	0.4	0.1549	(0.1872, 0.855)
151	0.3	0.1449	(0.1164, 0.773)
276	0.2	0.1265	(0.0579, 0.691)
281	0.1	0.0949	(0.0156, 0.642)

Table 1: (1.1) Last row of the survival function estimate tables for the asymptomatic group with the Kaplan-Meier estimator. (1.2) Last four rows of the survival function estimate tables for the symptomatic group with the Kaplan-Meier estimator.

1.3 1(c)

From Table 1.2 we have an estimate for the time that probability of survival falls below 30% as 276 weeks.

1.4 1(d)

The Nelson-Aalen estimates of the survival function are shown in Figure 2. Graphically it is clear that the estimates with both estimators are similar, but not identical; For example, for the symptomatic group, the survival function falls as low as 0.1 using the Kaplan-Meier estimator, but only to 0.145 using the Nelson-Aalen estimator. The Nelson-Aalen estimator estimates the cumulative hazard function and take the exponential function of the additive inverse of these estimates. The exponential function of the additive inverse of the cumulative hazard gives the survival function, which is also estimated by the Kaplan-Meier estimator, hence the two estimates are similar.

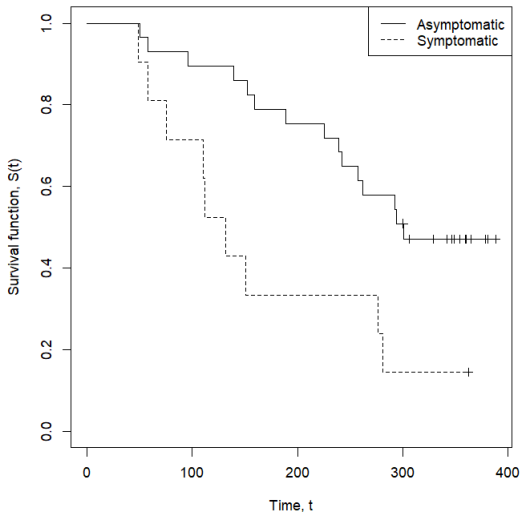

Figure 2: Estimated survival distribution by symptom group on the lymphoma data set using the Nelson-Aalen estimator. Crosses denote censored values.

1.5 1(e)

Using the exact same logic as 1(b) and drawing values from Table 2, we obtain the 95% confidence interval for probability of 6-year survival under the Nelson-Aalen estimator as (0.319, 0.697) for the asymptomatic group and

(0.0340, 0.621) for the symptomatic group.

(2.1)

Time	Estimated survival function	Standard error	95% Confidence interval
...	...	...	...
301	0.471	0.0940	(0.319, 0.697)

(2.2)

Time	Estimated survival function	Standard error	95% Confidence interval
...	...	...	...
281	0.145	0.1077	(0.0340, 0.621)

Table 2: (2.1) Last row of the survival function estimate tables for the asymptomatic group with the Nelson-Aalen estimator. (2.2) Last row of the survival function estimate tables for the symptomatic group with the Nelson-Aalen estimator.

2 Question 2

2.1 2(a)

A Weibull distribution is a statistical distribution that we may use to model the survival time of individuals. Such a model may depend on some explanatory variables, meaning that by entering some values of these variables to the model we are able to draw some predictions about the expected survival time of the individual. An intercept is also included in the model, giving a baseline value when the values of all explanatory variables are zero.

Starting from an intercept-only Weibull model, we can try to add main effects of the explanatory variables in a forward-selection process, using the 5% significance level. We find no statistically significant evidence to add any of the variable main effects to our model, including under a logarithmic transformation. As a result we conclude that survival does not depend on any of our potential explanatory variables in a Weibull model, and conclude that our preferred model is the intercept only model:

$$T_i \sim \text{Weibull}(1.212, \exp(-4.593)) \tag{1}$$

2.2 2(b)

Looking at Figure 3, the assumption of a Weibull distribution does not appear valid as the line is not straight and does not pass through the origin.

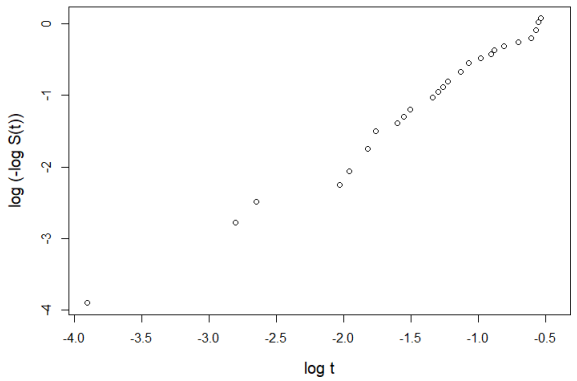

Figure 3: Plot of $\log(-\log S(t))$ against $\log t$ for the preferred model from 2(a).

2.3 2(c)

As there is no clear interpretation of a null ('intercept only') Cox proportional hazard model, we will begin by trying to find the single main effect model which best describes the data. Experimenting with different variables

shows that the one-parameter model with the lowest p-value for the likelihood ratio test and Wald test value is the model containing only log(Weight). Trying to add more terms to this model does not turn up any parameters worth adding, and so we settle on the model containing just log (Weight). The estimated value of the coefficient is -3.557, so a unit increase in log(Weight) is estimated to increase the hazard by a factor of $\exp(-3.557) = 0.029$, or in other words an increase in Weight results in a decrease in hazard.

2.4 2(d)

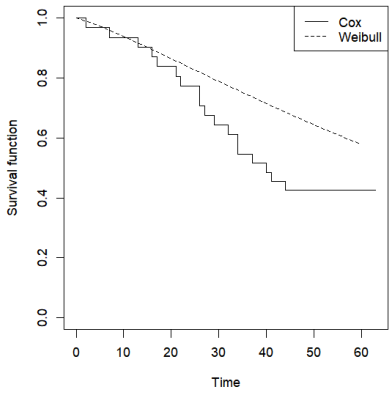

Figure 4: Estimated survival function for duck with age 1, weight 1000 and length 250 under the preferred models. Smooth curve shows the estimated survival function under the Weibull model and the stepwise function shows the estimated survival function under the Cox proportional hazards model.

3 Question 3

3.1 3(a)

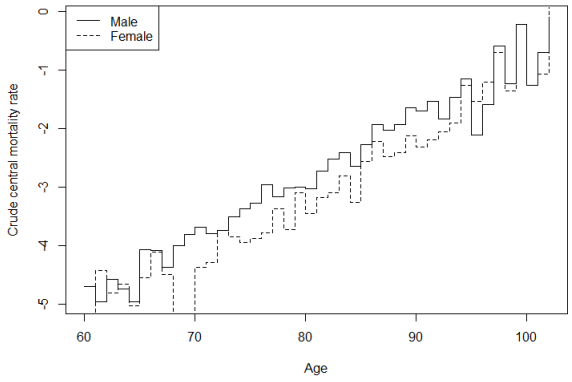

Figure 5: The crude force of mortality rates for male and female pensioners.

3.2 3(b)

	Constant force of mortality				Uniform distribution			
x	**Male** q_x	**Female** q_x	**Male** l_x	**Female** l_x	**Male** q_x	**Female** q_x	**Male** l_x	**Female** l_x
60	0.009	0.001	100000	100000	0.009	0.001	100000	100000
65	0.017	0.011	95860	96317	0.017	0.011	95860	96317
70	0.025	0.013	87849	92020	0.025	0.013	87849	92020
75	0.037	0.021	76681	83981	0.037	0.021	76680	83981
80	0.047	0.031	60816	72406	0.047	0.031	60813	72404
85	0.099	0.075	42577	58160	0.099	0.075	42567	58156
90	0.168	0.095	20673	35940	0.168	0.095	20639	35923
95	0.115	0.194	6807	16559	0.115	0.195	6759	16509
100	0.249	0	932	3336	0.25	0	866	3265

Table 3: The life table for males and females from radix $l_{60} = 100000$ in intervals of 5 years.

3.3 3(c)

Making the assumption of uniform distribution of deaths within each year of age we have curtate life expectancy 21.594 years and complete life expectancy 22.094 years for males and curtate life expectancy 24.927 years and complete life expectancy 25.427 years for females.

3.4 3(d)

For the male and female populations we use a chi-squared test to compare the mortality rates in the study population to standard mortality rates from the ELT17 data set:

$$H_0 : m_x = m_x^S \quad \text{against} \quad H_1 : m_x \neq m_x^S$$

For the male population we calculate a test statistic of 35.218, giving a p-value of 0.795 for the chi-squared distribution with 43 degrees of freedom. For the female population we calculate a test statistic of 57.727, giving a p-value of 0.066 for the chi-squared distribution with 43 degrees of freedom. Hence for both male and female we do not reject the null hypothesis at the 5% significance level, and conclude that the standard mortality rates provide a reasonable model of the study population rates.