HW # 4 : Nonparametric Survival Analysis

Smita Sukhadeve

November 2016

Introduction

The purpose of this paper is to demonstrate the use of co-variates for the prediction of survival time from the diagnosis. We aim to use non-parametric survival analysis to estimate the survival and hazard function from the survival data. The Scotland and Newcastle Lymphoma Group (SNLG) has collected the data used during this study. The sample data constitutes information for 784 patients diagnosed with Non-Hodgkin's lymphoma, their survival times from diagnosis and other explanatory variables as described in the table-1. For the scope of this study, we intend to apply Kaplan-Meier Survival estimation and Nelson-Aalen Hazard estimation functions to determine and analyse the survival and hazard functions for this data. We also wish to compare the survival functions for the patients having different clinical stages of the Non-Hodgkin's lymphoma.

Table-1 Dataset Summary

Variable	Description
ref	SNLG reference number of the patient
status	$0/1$ indicating censored$/$lost-to-follow-up and dead respectively
t	time in months from diagnosis to death or censoring
Sex	sex (1 male and 2 :female)
Age	Age at time of diagnosis in years
stage	clinical stage at the time of diagnosis, possible values 1, 2, 3, 4
ecog	fitness grades , possible values 1, 2, 3, 4, 5, 6
albumin	Serum albumin concentration. If within range then 0 else $X - L/L$
ldh	serum lactate dehydrogenase (ldh) concentration. If within range then 0 else $X - U/U$
alk	serum alkaline phosphatase concentration. If within range then 0 else $X - U/U$
urea	blood urea concentration. If within range then 0 else X-(L+U)/2
hb	blood haemoglobin level (X-125)/25 where X is haemoglobin level (g/l)
wbc	White blood cell count (X-8)/5 where X is the white blood cell count
bulk	This is 0 if there is no "bulk disease" and 1 if there is "bulk disease".
marrow	This is 0 if there is no lymphoma and 1 if there is lymphoma found by a marrow trephine
extranod	the number of "extra nodal" sites with evidence of disease

Method and Results

1.1 Kaplan-Meier Estimation of Survival Function

First, we use Kaplan-Meier or Product Limit estimator to predict the survival time from the diagnosis of Non-Hodgkin's lymphoma to the death of the subject. In this non-parametric approach, we divide the observed time-span into definite intervals so that it considers all the deaths or censoring events. Using the formula for the conditional probability, we find the survival function S(t), which represents the probability of survival of the patient after the diagnosis of certain disease/event after a specific time t. The conditional probability of surviving the subject in the j^{th} interval is $(1 - \frac{d_j}{n})$ and $S(t) = 1$, when no deaths occur during that interval. d_j represents death's count in the j^{th} interval. Finally, we take product of all the probabilities for intervals including t or preceding time t to get the survival time. Mathematically, it can be represented as follows:

$\hat{S}(t) = P[T > t] = \prod_{i=1}^{k}$ P[Survive jth interval I_j | survive to start of I_j]

We used survfit() function from the R-Survival package to find the Kaplan-Meier Estimator to predict the survival time. The beginning of each interval is determined by the death of the patient. Following table represents survival probability $\hat{S}(t)$ for first few and last intervals. We observe the cumulative probability of surviving at least a half month is 0.991.

Table-2: Survival Probability using Kaplan-Meier Estimation

t_i	d_i	n_i	S(t)	Range of t
0	-	-	1	$0 <= t < 0.03$
0.03	784	2	$1 - \frac{2}{784} = 0.997$	$0.03 <= t < 0.26$
0.26	780	1	$(1 - \frac{1}{780} = 0.998) * 0.997 = 0.996$	$0.26 <= t < 0.36$
0.36	779	1	$(1 - \frac{1}{779} = 0.998) * 0.996 = 0.995$	$0.36 <= t < 0.43$
0.43	778	1	0.994	$0.43 <= t < 0.46$
0.46	777	1	0.992	$0.46 <= t < 0.49$
0.49	776	1	0.991	$0.49 <= t < 0.59$
..	...	...	...	....
153.10	4	1	0.225	$153.10 <= t < NA$

Figure-1, shows the Kaplan-Meier(KM) estimation curve for the survival function along with confidence interval and simultaneous confidence band. KM curve represents the point estimates of the survival function. In the plot, X-axis represents the time-to-death in months and vertical axis indicates the survival probability ranging from 0 to 1. Each interval is terminated by the occurrences of death after the diagnosis of the Non-Hodgkin's lymphoma. The vertical marks on the curve represent the censored events. Step height represents the change in cumulative probability as the curve progresses. The initial part of curve is steep and probability of surviving the first 50 months after the diagnosis of lymphoma is 50%. Similarly, step size in the first 50 months is smaller than the later intervals since many patients censored after the first half of the study. The middle part of curve is smooth and cumulative probabilities of survival between 50 to 100 months varies from 0.5 to 0.4.

Width of confidence interval band increases as the curve progresses due to the increased point-wise standard errors in the estimation of survival probabilities. Figure-1 also shows the simultaneous confidence band that is wider than point-wise confidence interval for KM survival function. We used Hall-Wellner approach to construct the simultaneous confidence band. We construct the point-wise confidence interval using the

simple linear method. By the end of study, the cumulative probability of surviving at most 153 months is 0.225 and confidence interval [0.0862,0.364] have 95% chance to include this value.
We use delta method to calculate the standard error in the survival function estimates. The formula derived using this technique is known as Greenwood Formula and is as follows:

$$SE(\hat{S}(t)) = \hat{S}(t)\sqrt{\sum \frac{d_i}{N_i(N_i - d_i)}}$$

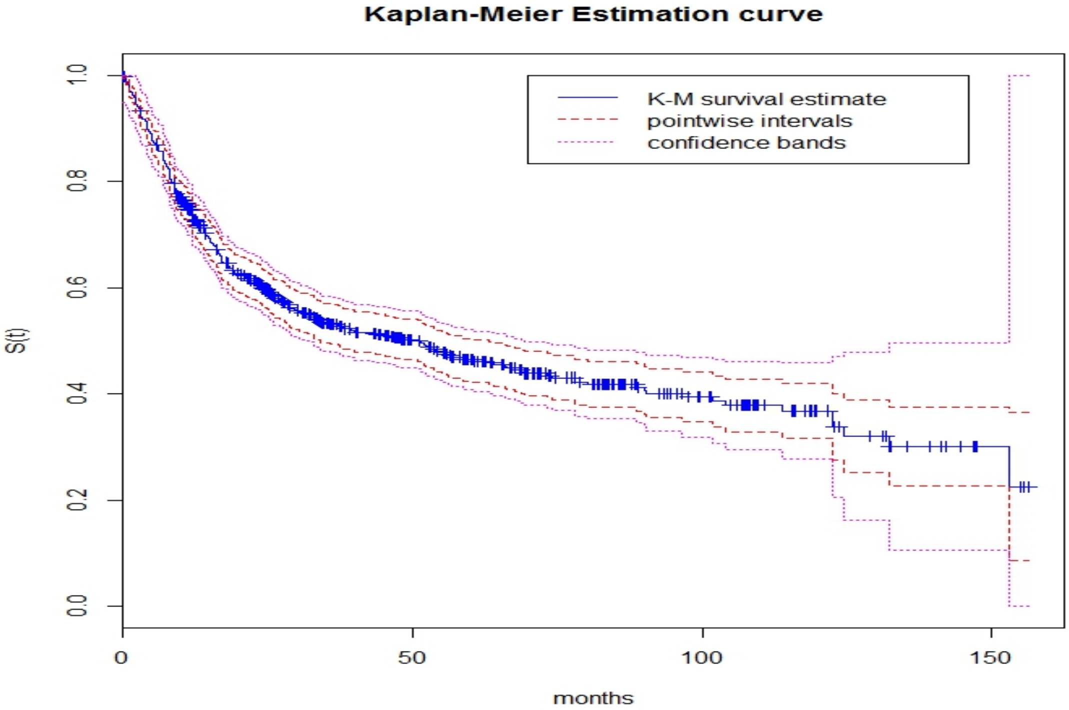


Figure 1: KM estimation plot with 95% confidence interval and simultaneous confidence band

1.2 Nelson and Aalen Estimation of Cumulative hazard function

Hazard function tells the probability that event will happen at time t given the subject is at risk at time t. We use Nelson-Aalen estimator to find the cumulative hazard rate function. Nelson-Aalen Estimator is as follows:

$$\Lambda = \sum_{x<t} \frac{dN(x)}{Y(x)}$$

The variance of $\hat{\Lambda}(t)$ is calculated as follows:

$$\sum_{x<t} \frac{\frac{dN(x}{Y(x)}[1 - \frac{dN(x)}{Y(x)}]}{Y(x) - 1}$$

After some calculation, we can show that, $\hat{V}(\hat{\Lambda}(t)) \approx dN(x)/Y(x)^2$
Where, $dN(x) =$ number of deaths in the interval $[x, x + \Delta x)$
$Y(x) =$ number of subject at risk at point x

Following table represents cumulative hazard rates $\hat{H}(t)$ and the cumulative variance of the hazard function for the first few and last intervals. We calculated 95% confidence interval of the hazard function as follows:
C.I. of hazard function = $[H(t) \pm z_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\Lambda})}]$

Table-3 Hazard function estimation using Nelson-Aalen estimator

t_i	d_i	n_i	H(t)	σ_H^2
0	-	-	0	0
0.03	784	2	$\frac{2}{784} = 0.00255$	$\frac{2}{784^2} = 3.25e^{-6}$
.				.
0.26	780	1	$\frac{1}{780} = 0.00128 + 0.00255 = 0.0038$	$3.253e^{-06} + \frac{1}{780^2} = 4.89e{-}6$
.				.
0.36	779	1	0.00511	$6.545^e{-}6$
0.43	778	1	0.00640	$8.197^e{-}6$
0.46	777	1	0.00768	$9.853^e{-}6$
...	...	..	....	....
153.10	4	1	1.443	7.77

Figure-2, shows the Nelson-Aalen curve for the Hazard function along with 95% confidence interval band. From the plot, we can observe that death rate increases with the time. The hazard rate for the patient dieing after the 50 months of the diagnosis is 0.6. We observed hazard rate of patient dieing after 150 months is 1.44 with 95% chance that this value will be in confidence interval [0.89 1.99].

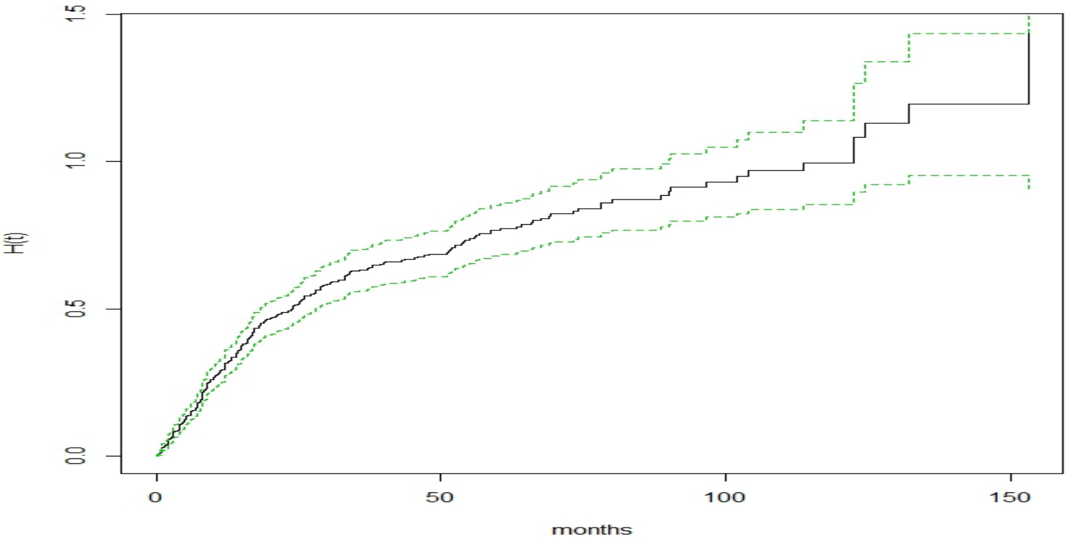


Figure 2: Nelson-Aalen hazard estimation plot with 95% confidence interval

2.1 Estimation of Mean Survival time and it's confidence interval

We calculate the point estimate of the mean survival time as follow:

$\mu_\tau = \int_0^\tau \hat{S}(t)dt$

Where, $\hat{S}(t) =$ Kaplan-Meier Estimator and $\tau =$ Pre-specified largest possible survival time

Confidence Interval for the mean survival time is given as:

$\mu_\tau \pm z_{1-\frac{\alpha}{2}} sqrt(\hat{V}(\hat{\mu}_\tau))$

Where,
$\hat{V}(\hat{\mu}_\tau) = \sum_{i=1}^{D} [\int_{t_i}^{\tau} \hat{S}(t)dt]^2 \frac{d_i}{n_i(n_i-d_i)}$
d_i =number of deaths in the i^th interval
n_i =number of subject at risk

Using above approach, we got mean survival time 72.58 with standard error of 2.85. 95% confidence interval for mean survival time is [66.994, 78.166]. It means mean survival time of the patient after the diagnosis of Non-Hodgkin's lymphoma is approximately 73 months and confidence interval [66.994, 78.166] has 95% chance to include this value.

2.2 Advantages and Disadvantages of various interval Estimates

Figure-3 represents the KM curve along with different types of confidence intervals. For the survival data used during this study, these confidence interval bands roughly coincides with each other.

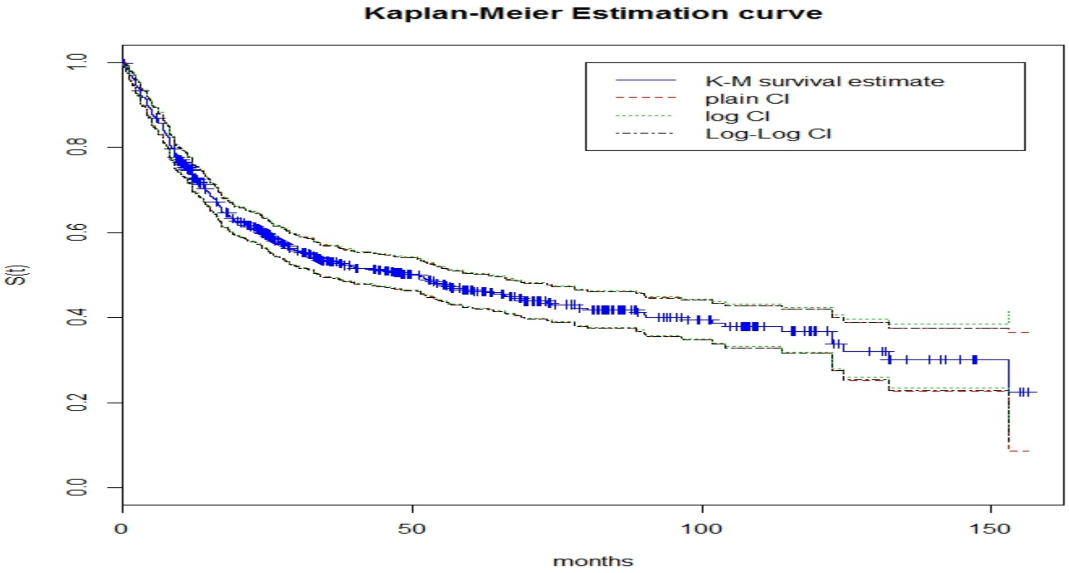


Figure 3: KM curve with different point-wise confidence interval bands

The primary advantage of using standard approach is we don't need any transformation to interpret confidence interval. However, using this approach the range for confidence interval may exceed the normal range of survival function i.e.[0, 1]. To resolve this issue, we take log transformation of the survival function and find the variance/standard error for the log($\hat{S}(t)$) using the delta approach and get the confidence interval(CI) for log$\hat{S}(t)$. Then, we transform it to the original scale to interpret the result. We use same technique to get the log-log confidence interval for survival function. We take log of log$\hat{S}(t)$ and find its variance to construct CI. Main advantage of using log-log transformation is to get approximately normal distribution for the confidence intervals which may be useful to evaluate the results in real scenario.

3. Kaplan-Meier estimation based on patient clinical stages

In this section, we demonstrated use of Kaplan-Meier estimator to find the survival functions based on the clinical stages of the disease. Patient health is categorised into four stages based on the growth of lymph nodes in the body. The KM-estimation plot based on patient's clinical stage is as shown in Figure-4:

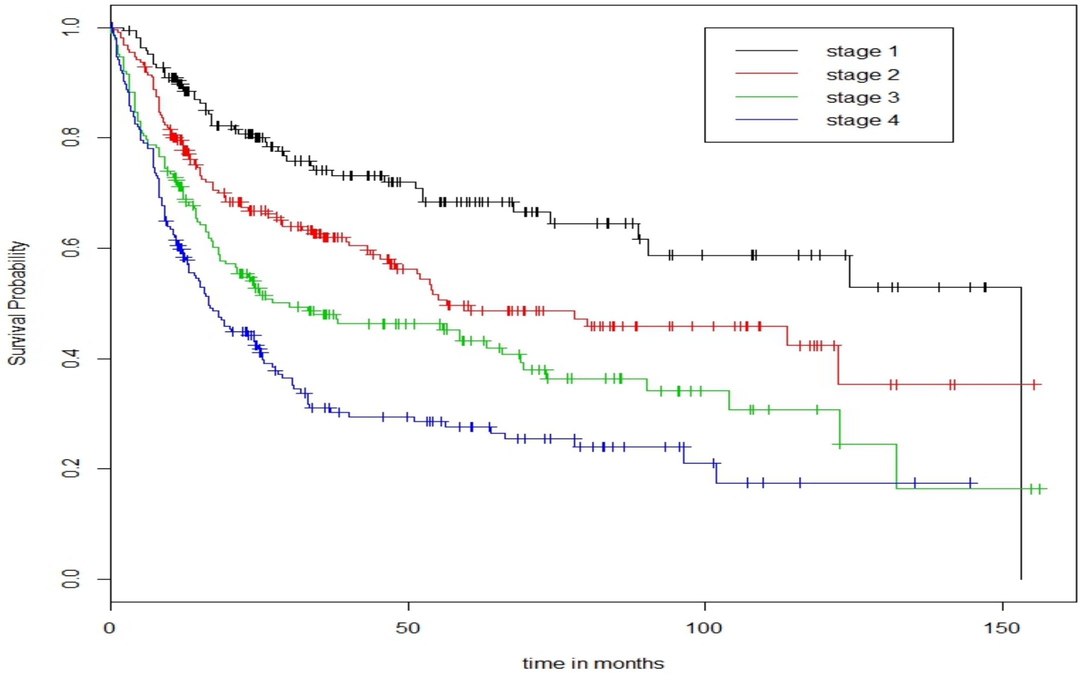


Figure 4: Kaplan-Meier estimation plot stratified by patient's clinical stage

From Figure-4, we can clearly see the survival curves based on clinical stages are very different from each other. KM estimation plot for stage 4 patient's is steeper. Probability of patient surviving in stage 4 after 30 months is 0.30. However, the survival probability of patient at stage 3, 2, 1 is approximately 0.50, 0.70 and 0.80 respectively after this duration. Patient with stage 1 lymphoma died towards the end of study i.e. after 150 months. For stage 2 lymphoma patients, probability of surviving at most 120 month is 0.3. Whereas, stage 3 patients have only 16% chance of surviving after 132 months. Last death for patient with stage 4 lymphoma occurred after 100 months. We can also analyse mean and median survival time for these different groups. Mean survival time is the area under the survival curve whereas median time represents the time at which survival probability is 0.5. Following table shows the mean and median survival times for different stages of Non-Hodgkin's lymphoma.

Table-5 Mean and Median estimates of survival time for patients with distinct clinical stages

Stage	mean S(t)	95%CI of mean	Median S(t)	95%CI of median
1	103.0	[91.3,114.7]	153.1	[90.3, NA]
2	80.8	[69.3,90.6]	56.0	[48.0, NA]
3	62.2	[51.6, 72.7]	30.0	[21.1, 65.7]
4	46.0	[36.6, 55.3]	16.5	[13.0, 24.0]

4. Log-Rank test on the difference in the survival distribution among clinic stage

We used Log-Rank test to compare the survival functions of these different groups. The log-rank test is a chi-squared test. During this test, we compare the death rates between the different groups at distinct time points. We construct k by k table for analysing the difference between k groups. We compare observed number of deaths/event with the expected number of deaths. Expected number of deaths in group j at death time t_i is calculated as : $e_{ij} = d_i \frac{n_{ij}}{n_i}$. then simple χ^2 test statistics is calculated as follows:

$$\chi^2 = \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j}$$

which is χ^2 with $K - 1$ degrees of freedom. K is number of groups in the study.

For the null hypothesis, we assume that survival functions for the groups are same. We used survdiff() function from R 'survival' package to perform this test. The value of χ^2 is 74.8 with 3 degrees of freedom. p-value is $4.44e^{-6}$ which significant at 0.05 significance level. Therefore, we will reject null hypothesis and confirm that there is significance difference between survival functions for the patient with different stages.

Discussion

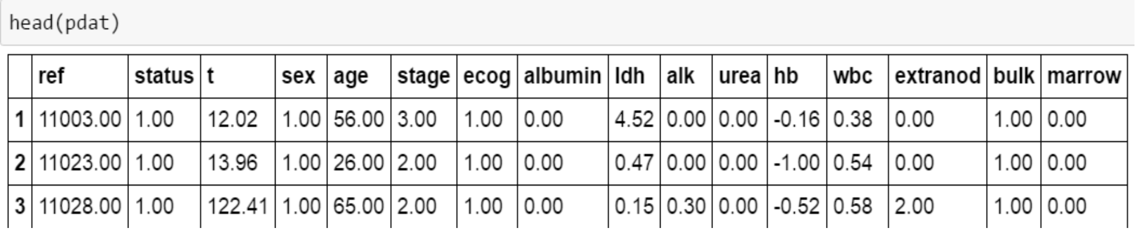
To summarize, we used Kaplan-Meier and Nelson-Aalen estimators to find survival and hazard function respectively. We analysed survival and hazard estimation plots to compare probability of surviving and death rates among the patient diagnosed with Non-Hodgkin's lymphoma. At this point, we can only evaluate the survival probability at different time points and analyse how mean, median and confidence intervals varies with the different groups. Moreover, we observed half of the patients censored by the end of this study. This could get misleading results. Therefore, I think, we need other methods to validate our results.

References

1. Dr. Cheng Peng (Fall 2016 - STA 588) Notes 1-6, Survival Analysis
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3932959/
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3227332/

Appendix

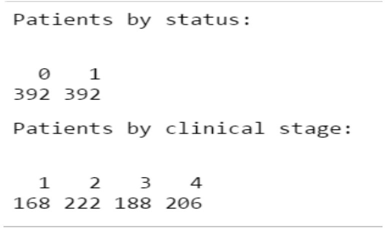
I.Dataset structure



```
head(pdat)
```

	ref	status	t	sex	age	stage	ecog	albumin	ldh	alk	urea	hb	wbc	extranod	bulk	marrow
1	11003.00	1.00	12.02	1.00	56.00	3.00	1.00	0.00	4.52	0.00	0.00	-0.16	0.38	0.00	1.00	0.00
2	11023.00	1.00	13.96	1.00	26.00	2.00	1.00	0.00	0.47	0.00	0.00	-1.00	0.54	0.00	1.00	0.00
3	11028.00	1.00	122.41	1.00	65.00	2.00	1.00	0.00	0.15	0.30	0.00	-0.52	0.58	2.00	1.00	0.00

II.Frequency by patients status and stages



```
Patients by status:


  0   1 
392 392 
Patients by clinical stage:


  1   2   3   4 
168 222 188 206 
```

III. R code to reproduce the result