yc4384_Yangyang_Chen_HW1

Yangyang Chen

2024-09-20

Question 1

a) Using the data above, calculate the maximum likelihood estimator of the parameter λ for time to relapse and time to death assuming an exponential distribution:

$$f(t) = \lambda e^{-\lambda t}$$

Write a brief sentence interpreting this parameter.

Solution:

The likelihood functions are:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda T_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} T_i}$$

$$\log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} T_i$$

Maximizing the likelihood:

$$\frac{d}{d\lambda} \log L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} T_i = 0$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} T_i}$$

Time to relapse:

$$n = 6$$

$$\sum T_i = 180$$

$$\hat{\lambda}_{\mathrm{r}} = \frac{6}{180} \approx 0.033$$

Time to death:

$$n = 3$$

$$\sum T_i = 225$$

$$\hat{\lambda}_{\mathrm{d}} = \frac{3}{225} \approx 0.013$$

Interpretation:

- The MLE for the time to relapse is $\hat{\lambda}_{\mathrm{r}} = 0.033$, indicating that approximately 3.3% of patients relapse each month.
- The MLE for the time to death is $\hat{\lambda}_{\mathrm{d}} = 0.013$, meaning that approximately 1.3% of patients die each month.

b) Now you will see how powerful this single parameter can be! Using this parameter estimate (round to 3 decimal places), estimate the following quantities:

(i) The mean time to relapse and mean survival time after bone marrow transplant.

Solution:

The mean time to relapse is:

$$E(T_{\mathrm{r}}) = \frac{1}{0.033} \approx 30.303 \text{ months}$$

The mean survival time to death is:

$$E(T_{\mathrm{d}}) = \frac{1}{0.013} \approx 76.923 \text{ months}$$

(ii) The median time to relapse and median survival time after bone marrow transplant.

Solution:

$$1 - e^{-\lambda t_{\text{median}}} = 0.5$$

$$t_{\text{median}} = \frac{-\log(0.5)}{\lambda}$$

The median time to relapse is:

$$\text{Median}(T_{\mathrm{r}}) = \frac{-\log(0.5)}{0.033} \approx 21.004 \text{ months}$$

The median survival time to death is:

$$\text{Median}(T_{\mathrm{d}}) = \frac{-\log(0.5)}{0.013} \approx 53.319 \text{ months}$$

(iii) The one-year and two-year probabilities of remaining relapse-free and surviving: $S_R(12)$ and $S_R(24)$ for relapse, and $S_D(12)$ and $S_D(24)$ for death.

Solution:

```
lambda_rl = 6/(5+8+12+24+32+17+16+17+19+30)
sr12 = exp(-lambda_rl*12)
sr24 = exp(-lambda_rl*24)

lambda_d = 3/(10+12+15+33+45+28+16+17+19+30)
sd12 = exp(-lambda_d*12)
sd24 = exp(-lambda_d*24)
```

Given:

$$S(t) = P(T > t) = e^{-\lambda t}$$

The probabilities of remaining relapse-free:

$$S_R(12) = e^{-\hat{\lambda}_{\mathrm{r}} \cdot 12} = e^{-0.033 \cdot 12} = 0.67$$

$$S_R(24) = e^{-\hat{\lambda}_{\mathrm{r}} \cdot 24} = e^{-0.033 \cdot 24} = 0.449$$

The survival probabilities:

$$S_D(12) = e^{-\hat{\lambda}_{\mathrm{d}} \cdot 12} = e^{-0.013 \cdot 12} = 0.852$$

$$S_D(24) = e^{-\hat{\lambda}_{\mathrm{d}} \cdot 24} = e^{-0.013 \cdot 24} = 0.726$$

(iv) The cumulative probabilities of relapse and death by one and two years (based on the CDF, $F(t)$).

Solution:

The cumulative probability of relapse is:

$$F_R(12) = 1 - S_R(12) = 1 - 0.67032 = 0.33$$

$$F_R(24) = 1 - S_R(24) = 1 - 0.449329 = 0.551$$

The cumulative probability of death is:

$$F_D(12) = 1 - S_D(12) = 1 - 0.8521438 = 0.148$$
$$F_D(24) = 1 - S_D(24) = 1 - 0.726149 = 0.274$$

(v) Based on the exponential distribution with $\hat{\lambda}$ as calculated in (a), calculate the conditional probability of being relapse-free after 2 years given that one has remained relapse-free for at least one year. How does this compare with the probability of remaining relapse-free one year after bone marrow transplant calculated in part (iii)?

Solution:

$$P(T > 24 \mid T > 12) = \frac{P(T > 24)}{P(T > 12)} = \frac{S(24)}{S(12)} = \frac{e^{-\lambda \cdot 24}}{e^{-\lambda \cdot 12}} = e^{-\lambda \cdot (24-12)} = e^{-\lambda \cdot 12}$$

Since $\hat{\lambda}_{\mathrm{r}} \approx 0.033$, we obtain:

$$P(T > 24 \mid T > 12) = e^{-0.033 \cdot 12} = 0.67$$

From part (iii), the probability of remaining relapse-free is:

$$S(12) = e^{-0.033 \cdot 12} = P(T > 24 \mid T > 12) = 0.67$$

Therefore, the conditional probability of remaining relapse-free after 2 years, given relapse-free for 1 year, is exactly the same as the probability of remaining relapse-free after 1 year. This is due to the memoryless property of the exponential distribution.

(c) If we decide that an exponential distribution is not appropriate and want to estimate the survival distribution non-parametrically, is it possible to estimate the median time to relapse? Is it possible to estimate the median time to death? If so, provide the appropriate estimates.

Solution:

Yes, it is possible. The Kaplan-Meier estimator is used to estimate survival functions and does not rely on any parametric distribution.

```r
relapse_t = c(5, 8, 12, 24, 32, 17, 16, 17, 19, 30)
relapse_s = c(1, 1, 1, 1, 1, 1, 0, 0, 0, 0)

death_t = c(10, 12, 15, 33, 45, 28, 16, 17, 19, 30)
death_s = c(1, 1, 1, 1, 0, 0, 0, 0, 0, 0)

km_relapse = survfit(Surv(relapse_t, relapse_s) ~ 1)
med_relapse = summary(km_relapse)$table["median"]

km_death = survfit(Surv(death_t, death_s) ~ 1)
med_death = summary(km_death)$table["median"]

med_df = data.frame(
  Outcome = c("Relapse", "Death"),
  Median_Time = c(med_relapse, med_death)
)

knitr::kable(med_df, col.names = c("Outcome", "Median Survival Time"))
```

Outcome	Median Survival Time
Relapse	24
Death	33

Question 2

a) t_j: distinct death or censoring times
d_j: the number of death at t_j
r_j: the number of individuals at risk right before the j-th death time
c_j: the number of censored observations between the j-th and $(j+1)$-st death time

t_j	d_j	c_j	r_j	$1-(d_j/r_j)$	$\hat{S}(t_j)$
2	1	0	17	0.941	0.941
3	1	0	16	0.938	0.882
4	1	0	15	0.933	0.824
12	1	0	14	0.929	0.765
22	1	0	13	0.923	0.706
48	1	0	12	0.917	0.647
51	0	1	11	1	0.647
56	0	1	10	1	0.647
80	1	0	9	0.889	0.575
85	1	0	8	0.875	0.503
90	1	0	7	0.857	0.431
94	0	1	6	1	0.431
160	1	0	5	0.8	0.345
171	1	0	4	0.75	0.259
180	1	1	3	0.667	0.173
238	1	0	1	0	0

b) Repeat the above estimation of $\hat{S}(t)$ using any software you choose. Also calculate pointwise 95% confidence intervals for $\hat{S}(t)$ using the "log-log" approach and the linear approach. Do either of the approaches result in lower or upper confidence bounds outside the $[0,1]$ interval?

Solution:

```r
df2 = data.frame(
  t = c(2,3,4,12,22,48,51,56,80,85,90,94,160,171,180,180,238),
  c = c(1,1,1,1,1,1,0,0,1,1,1,0,1,1,1,0,1) # 1: event, 0: censored
)

surv = Surv(df2$t, df2$c)
km = survfit(surv ~ 1)

l_loglog = km.ci(km, method = "loglog")$lower
u_loglog = km.ci(km, method = "loglog")$upper

l_linear = km.ci(km, method = "linear")$lower
u_linear = km.ci(km, method = "linear")$upper

tb = data.frame(
  t = round(km$time,3), # time
  st = round(km$surv,3), # survival
  l_loglog = round(l_loglog, 3),
  u_loglog = round(u_loglog, 3),
  l_linear = round(l_linear, 3),
  u_linear = round(u_linear, 3)
)
kable(tb, col.names = c("t", "S(t)", "Lower 95%CI (log-log)", "Upper 95%CI (log-log)", "Lower 95%CI (li
```

t	S(t)	Lower 95%CI (log-log)	Upper 95%CI (log-log)	Lower 95%CI (linear)	Upper 95%CI (linear)
2	0.941	0.650	0.991	0.829	1.053
3	0.882	0.606	0.969	0.729	1.036
4	0.824	0.547	0.939	0.642	1.005
12	0.765	0.488	0.904	0.563	0.966
22	0.706	0.431	0.866	0.489	0.922
48	0.647	0.377	0.823	0.420	0.874
51	0.647	0.377	0.823	0.420	0.874
56	0.647	0.377	0.823	0.420	0.874
80	0.575	0.307	0.772	0.333	0.817
85	0.503	0.244	0.716	0.254	0.752
90	0.431	0.187	0.656	0.181	0.682
94	0.431	0.187	0.656	0.181	0.682
160	0.345	0.122	0.584	0.094	0.596
171	0.259	0.069	0.505	0.020	0.497
180	0.173	0.030	0.416	-0.038	0.383
238	0.000	NaN	NaN	NaN	NaN

Conclusion:

Log-log approach returns CIs within $[0, 1]$, but linear approach returns some values outside $[0, 1]$.

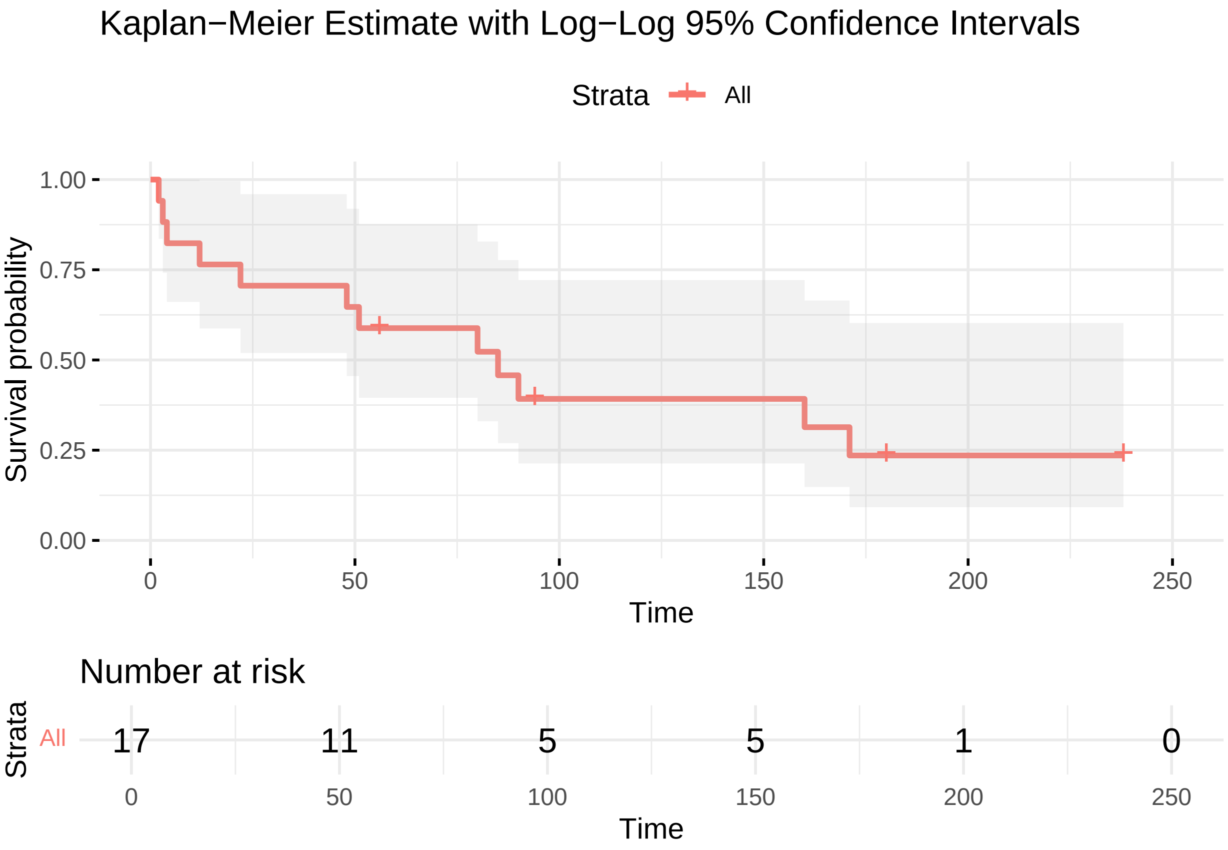
c) Plot the estimated survival function $\hat{S}(t)$ and pointwise 95% confidence intervals, by hand or using any statistical software package.

```
times = c(22, 2, 48, 85, 160, 238, 56, 94, 51, 12, 171, 80, 180, 4, 90, 180, 3)
censoring = c(FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FA

data = data.frame(times = times, censoring = censoring)
surv_obj = Surv(data$times, event = !data$censoring)

km_fit = survfit(surv_obj ~ 1, data = data)

ggsurvplot(
  km_fit,
  data = data,
  conf.int = TRUE,
  conf.int.style = "ribbon",
  conf.int.alpha = 0.2,
  conf.int.type = "log-log",
  ggtheme = theme_minimal(),
  risk.table = TRUE,
  title = "Kaplan-Meier Estimate with Log-Log 95% Confidence Intervals"
)
```



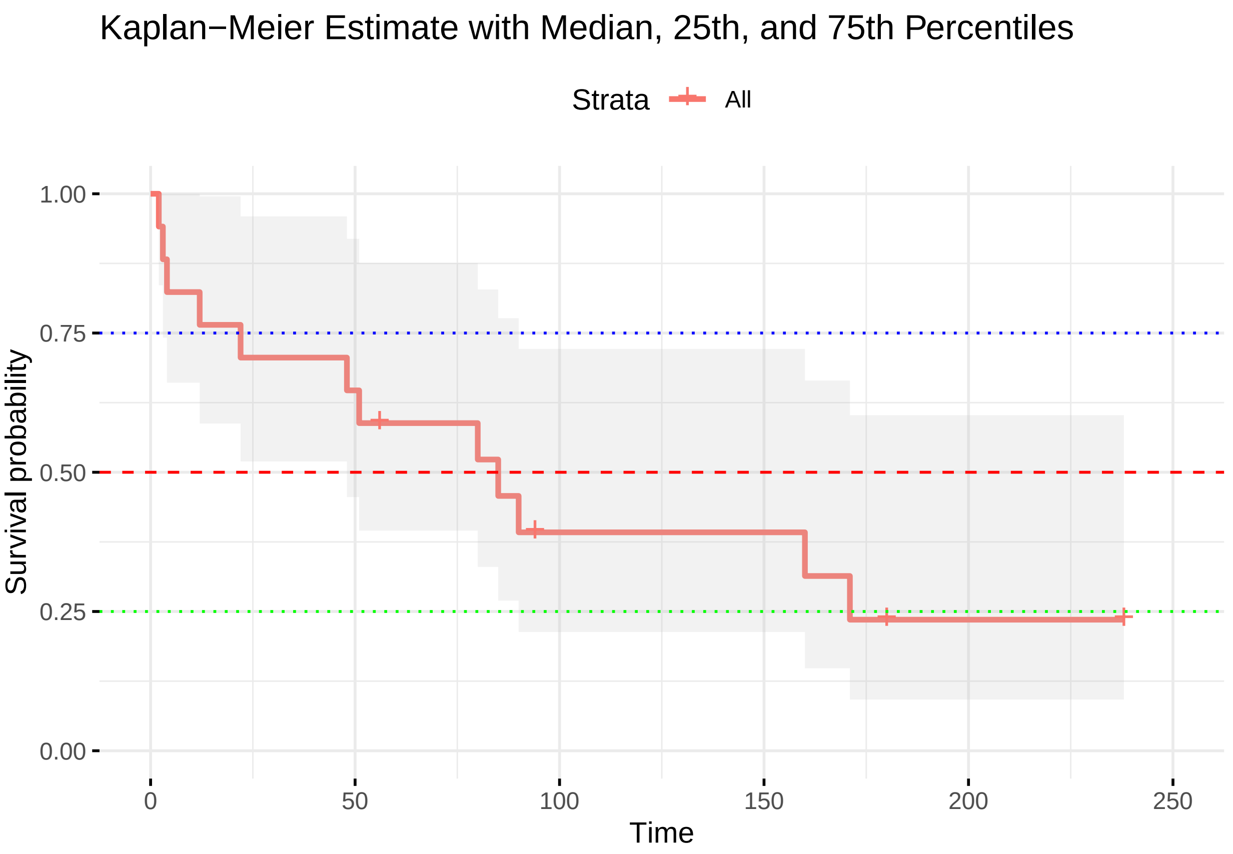
d) Provide the estimated median survival, along with the estimated 25th and 75th percentiles (when possible). Indicate where these percentiles fall on your KM plot from (c) by drawing horizontal lines. What are the actual KM survival estimates corresponding to each of these estimated percentiles?

Solution:

```
median_survival = summary(km_fit)$table["median"]
percentiles = quantile(km_fit, probs = c(0.25, 0.75))
```

```
km_plot = ggsurvplot(
  km_fit,
  data = data,                 # Provide the original data
  conf.int = TRUE,
  conf.int.style = "ribbon",
  conf.int.alpha = 0.2,
  conf.int.type = "log-log",     # Log-log confidence intervals
  ggtheme = theme_minimal(),
  risk.table = TRUE,
  title = "Kaplan-Meier Estimate with Median, 25th, and 75th Percentiles"
)

km_plot$plot +
  geom_hline(yintercept = 0.5, linetype = "dashed", color = "red") +
  geom_hline(yintercept = 0.75, linetype = "dotted", color = "blue") +
  geom_hline(yintercept = 0.25, linetype = "dotted", color = "green")
```



The actual Kaplan-Meier survival estimates for the 25th, 50th, and 75th percentiles are as follows: - 25th percentile: Survival estimate = 0.705 - 50th percentile (median): Survival estimate = 0.431 - 75th percentile: Survival estimate = 0.172 (e)

```
km_summary = summary(km_fit)
survival_prob = km_summary$surv  # S(t)
cumulative_hazard = -log(survival_prob)
result = data.frame(
  time = km_summary$time,
  survival_estimate = survival_prob,
  cumulative_hazard = cumulative_hazard
)
```

t_j	2	3	4	12	22	48	51	56	80	85	90	94	160	171	180	238
$\hat{\Lambda}_{KM}(t)$	0.061	0.125	0.194	0.268	0.348	0.435	0.531	0.531	0.648	0.782	0.936	0.936	1.159	1.447	1.447	1.447

(f)

```
cox_fit = coxph(surv_obj ~ 1, data = data)
na_cumulative_hazard = basehaz(cox_fit, centered = FALSE)
hazard_df = data.frame(
  Time = na_cumulative_hazard$time,
  Hazard = na_cumulative_hazard$hazard |> round(3)
)

knitr::kable(hazard_df, col.names = c("$t_j$", "$\\hat{\\Lambda}_{NA}(t)$"), escape = FALSE)
```

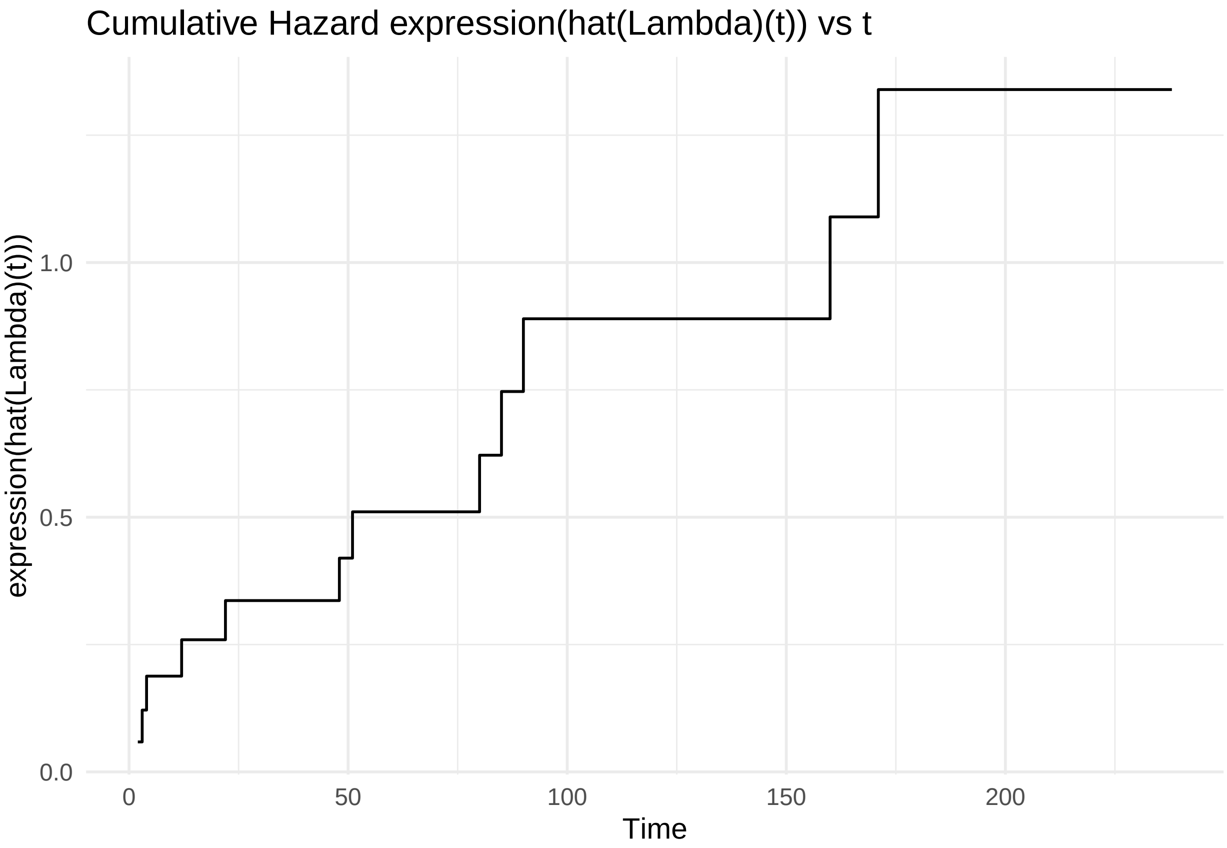
t_j	$\hat{\Lambda}_{NA}(t)$
2	0.059
3	0.121
4	0.188
12	0.259
22	0.336
48	0.420
51	0.511
56	0.511
80	0.622
85	0.747
90	0.890
94	0.890
160	1.090
171	1.340
180	1.340
238	1.340

(g)

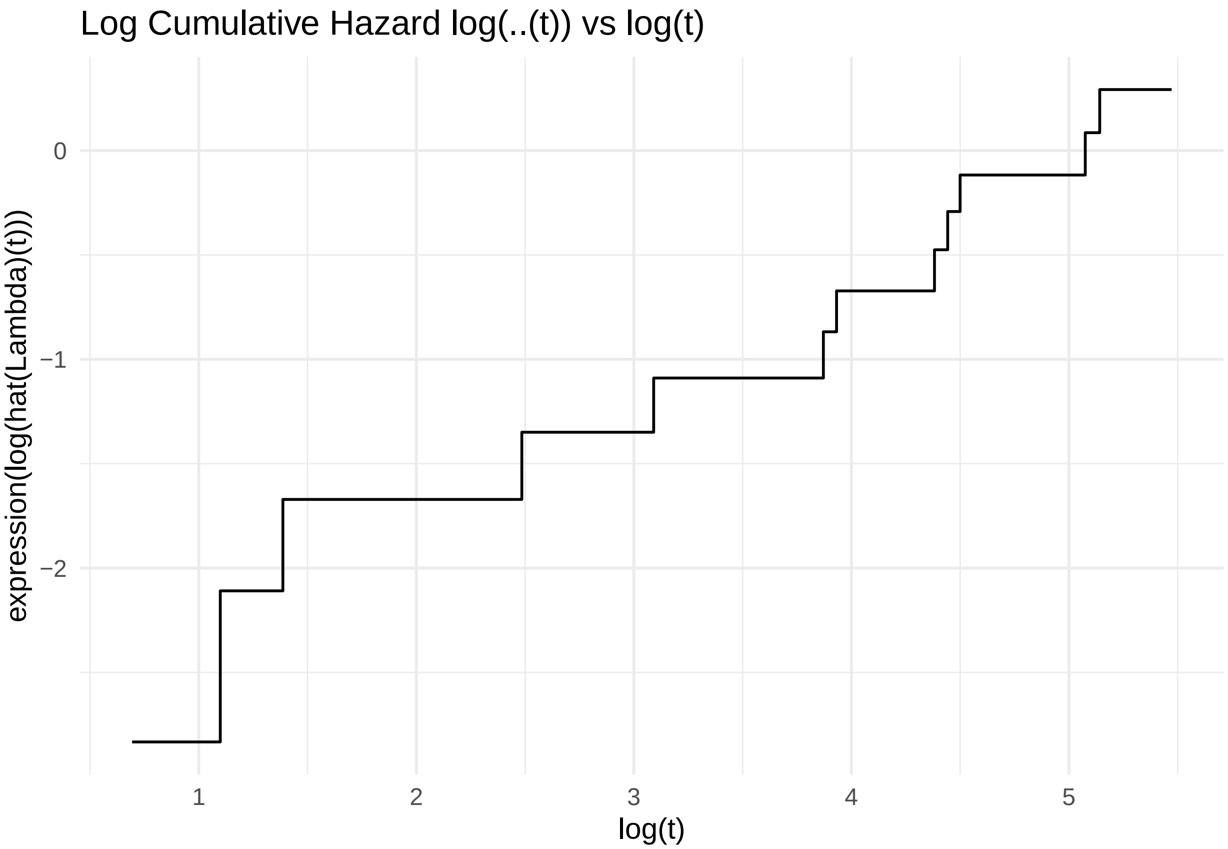
```r
cox_fit = coxph(surv_obj ~ 1, data = data)

na_cumulative_hazard = basehaz(cox_fit, centered = FALSE)

ggplot(na_cumulative_hazard, aes(x = time, y = hazard)) +
  geom_step() +
  labs(title = "Cumulative Hazard expression(hat(Lambda)(t)) vs t", x = "Time", y = "expression(hat(Lam
  theme_minimal()
```



```
ggplot(na_cumulative_hazard, aes(x = log(time), y = log(hazard))) +
  geom_step() +
  labs(title = "Log Cumulative Hazard log(Λ(t)) vs log(t)", x = "log(t)", y = "expression(log(hat(Lambd
  theme_minimal()
```



Plot (i) exhibits deviations from linearity in the cumulative hazard function, whereas plot (ii) displays a more linear trend. This observation suggests that the Weibull model may offer a better fit for the data, as it is more consistent with the linear behavior observed in the cumulative hazard function.

(h)

```
na_cumulative_hazard$survival_FH = exp(-na_cumulative_hazard$hazard)

comparison = data.frame(
  time = km_fit$time,
  NA_survival = na_cumulative_hazard$hazard |> round(3),
  KM_survival = km_fit$surv |> round(3),
  FH_survival = na_cumulative_hazard$survival_FH[match(km_fit$time,
                                                 na_cumulative_hazard$time)] |> round(3)
)

knitr::kable(comparison, col.names = c("$t_j$", "$\\hat{\\Lambda}_{NA}(t)$", "$\\hat{S}(t_j)$", "$\\hat
```

t_j	$\hat{\Lambda}_{NA}(t)$	$\hat{S}(t_j)$	$\hat{S}_{FH}(t_j)$
2	0.059	0.941	0.943
3	0.121	0.882	0.886
4	0.188	0.824	0.829
12	0.259	0.765	0.771
22	0.336	0.706	0.714
48	0.420	0.647	0.657

t_j	$\hat{\Lambda}_{NA}(t)$	$\hat{S}(t_j)$	$\hat{S}_{FH}(t_j)$
51	0.511	0.588	0.600
56	0.511	0.588	0.600
80	0.622	0.523	0.537
85	0.747	0.458	0.474
90	0.890	0.392	0.411
94	0.890	0.392	0.411
160	1.090	0.314	0.336
171	1.340	0.235	0.262
180	1.340	0.235	0.262
238	1.340	0.235	0.262

Overall, the agreement between the Kaplan-Meier and Fleming-Harrington survival estimates is very good, with only small differences observed. These differences are expected due to the distinct methodologies underlying the two estimators, but they do not substantially impact the overall survival trend in this dataset. Both estimators provide reliable estimates of the survival function.

Question 3.

(a)

```
intervals = cut(data$times, breaks = seq(0, max(data$times) + 30, by = 30), right = FALSE) |> levels()

rj = c(17, 12, 9, 7, 5, 5, 3, 1)|> round(3)
dj = c(5, 1, 2, 1, 0, 2, 1, 1) |> round(3)

data.frame(
  time_interval = intervals,
  rj,
  cj = c(0, 2, 0, 1, 0, 0, 1, 0)|> round(3),
  dj,
  sr = cumprod(1 - dj/rj)
) |>
  knitr::kable(col.names = c("Time Intervals", "$r_j$","$c_j$", "$d_j$","$\\hat{S}(t)$"), escape = FALS
```

Time Intervals	r_j	c_j	d_j	$\hat{S}(t)$
[0,30)	17	0	5	0.7058824
[30,60)	12	2	1	0.6470588
[60,90)	9	0	2	0.5032680
[90,120)	7	1	1	0.4313725
[120,150)	5	0	0	0.4313725
[150,180)	5	0	2	0.2588235
[180,210)	3	1	1	0.1725490
[210,240)	1	0	1	0.0000000

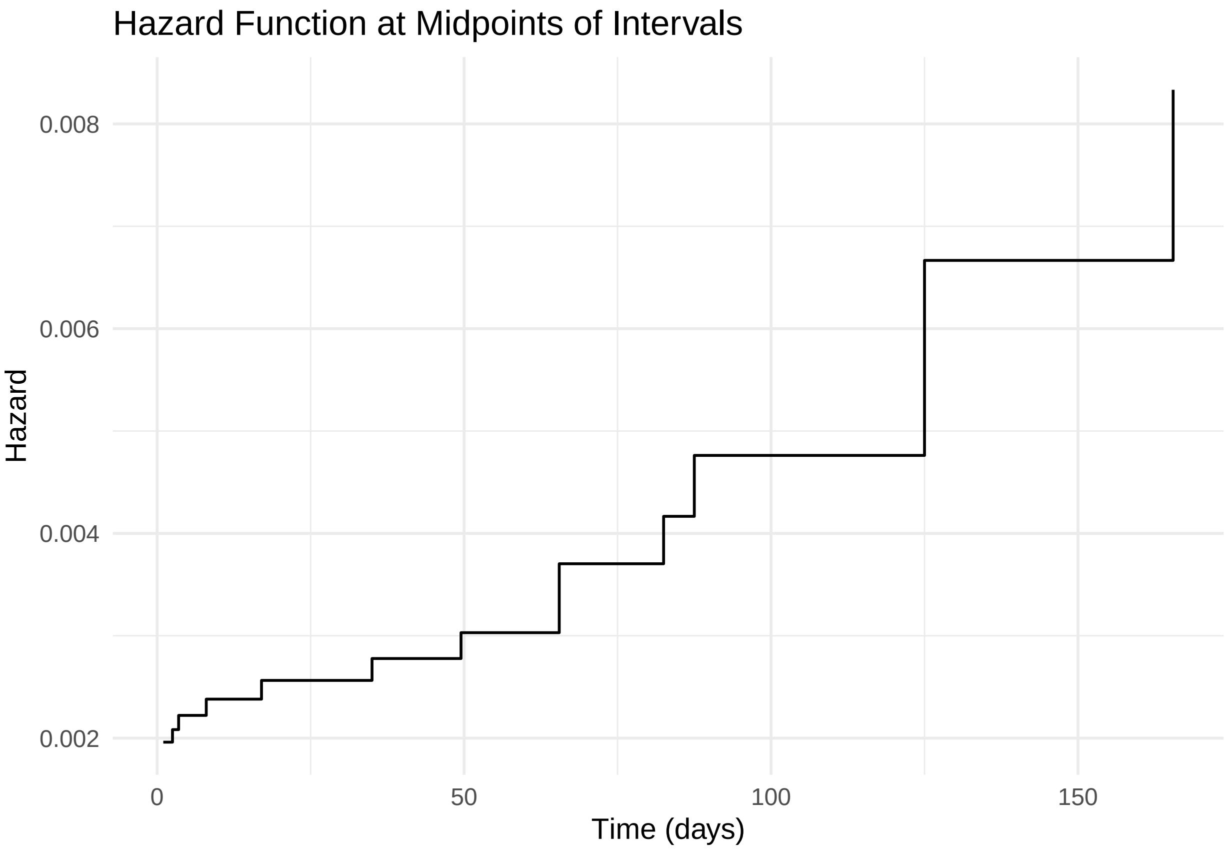
(b)

```
lifetable_fit = survfit(surv_obj ~ 1, data = data, type = "fleming-harrington", timefix = FALSE)

lifetable_summary = summary(lifetable_fit)
interval_midpoints = (lifetable_summary$time + c(0, lifetable_summary$time[-length(lifetable_summary$ti

hazard_function = lifetable_summary$n.event / (lifetable_summary$n.risk * 30)
hazard_data = data.frame(
  interval_midpoint = interval_midpoints,
  hazard = hazard_function
)

ggplot(hazard_data, aes(x = interval_midpoint, y = hazard)) +
  geom_step() +
  labs(title = "Hazard Function at Midpoints of Intervals", x = "Time (days)", y = "Hazard") +
  theme_minimal()
```



(c)

From the plot, since the hazard function is increasing and not constant, an exponential model does not seem appropriate for this data. Instead, a model that allows for a time-varying hazard, such as a Weibull or Cox proportional hazards model, might be more suitable for capturing the behavior of the hazard over time.

Question 4.

1) Brinkhof et al. (2010):

Main Concern: The authors worried that non-informative censoring, which assumes equal mortality risk for those lost to follow-up and those in care, underestimated mortality, as higher rates were found in those lost to follow-up.

Steps Taken: They used multiple imputation and hazard ratios to estimate survival for patients lost to follow-up, conducting sensitivity analyses with varying assumed hazard ratios to account for higher mortality.

2) Braitstein et al. (2011):

Main Concern: The authors feared that mortality was underestimated in HIV-positive and HIV-exposed children lost to follow-up due to stigma and fear of discrimination.

Steps Taken: Braitstein et al. conducted a prospective evaluation, tracing over 80% of lost children with community health workers, revealing that many had died, confirming underestimated mortality. Stigma and disclosure issues were key factors in non-informative censoring.