

Homework1

Yuki Joyama

1. Exponential Density and Survival-related Functions

- (a) Let $\hat{\lambda}$ be the maximum likelihood estimator of the parameter λ .

For relapse time:

$$\hat{\lambda} = \frac{6}{5+8+12+24+32+17+16+17+19+30} \approx 0.033$$

This indicates that the rate of relapse is about 3.33% per month.

For relapse time:

$$\hat{\lambda} = \frac{3}{10+12+15+33+45+28+16+17+19+30} \approx 0.013$$

This indicates that the rate of death is about 1.33% per month.

- (b)

- i. Mean is $\int_0^\infty t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}$ and I will use $\hat{\lambda}$ to derive the following values.

Mean time to relapse:

$$\frac{1}{0.033} \approx 30.303 \text{ months}$$

$$\frac{1}{0.013} \approx 76.923 \text{ months}$$

- ii. Median is $0.5 = e^{-\lambda\tau} \Rightarrow \tau = \frac{-\log(0.5)}{\lambda}$. By $\hat{\lambda}$,

Median time to relapse:

$$\frac{-\log(0.5)}{0.033} \approx 21.004 \text{ months}$$

$$\frac{-\log(0.5)}{0.013} \approx 53.319 \text{ months}$$

- iii. The survival function of exponential distribution is $S(t) = e^{-\lambda t}$. For relapse:

$$S_R(12) = e^{-0.033 \times 12} = 0.67$$

$$S_R(24) = e^{-0.033 \times 24} = 0.449$$

For death:

$$S_D(12) = e^{-0.013 \times 12} = 0.852$$

$$S_D(24) = e^{-0.013 \times 24} = 0.726$$

- iv. The cumulative probabilities can be calculated as: $F(t) = \int_0^t \lambda(u) du = \lambda t$

For relapse:

$$F_R(12) = 0.033 \times 12 = 0.4$$

$$F_R(24) = 0.033 \times 24 = 0.8$$

For death:

$$F_D(12) = 0.013 \times 12 = 0.16$$

$$F_D(24) = 0.013 \times 24 = 0.32$$

- v. The conditional probability can be expressed as $P(T > 24 | T > 12) = \frac{S_R(24)}{S_R(12)} = 0.67$

It is the same as what we observed in (iii) $S_R(12)$. This means that the conditional probability of being relapse-free after 2 years given that one has remained relapse-free for at least 1 year simplifies to the survival function for the remaining time period (memoryless property of the exponential distribution).

- (c) We can use Kaplan-Meier estimator to estimate median time to relapse. I will calculate them using R.

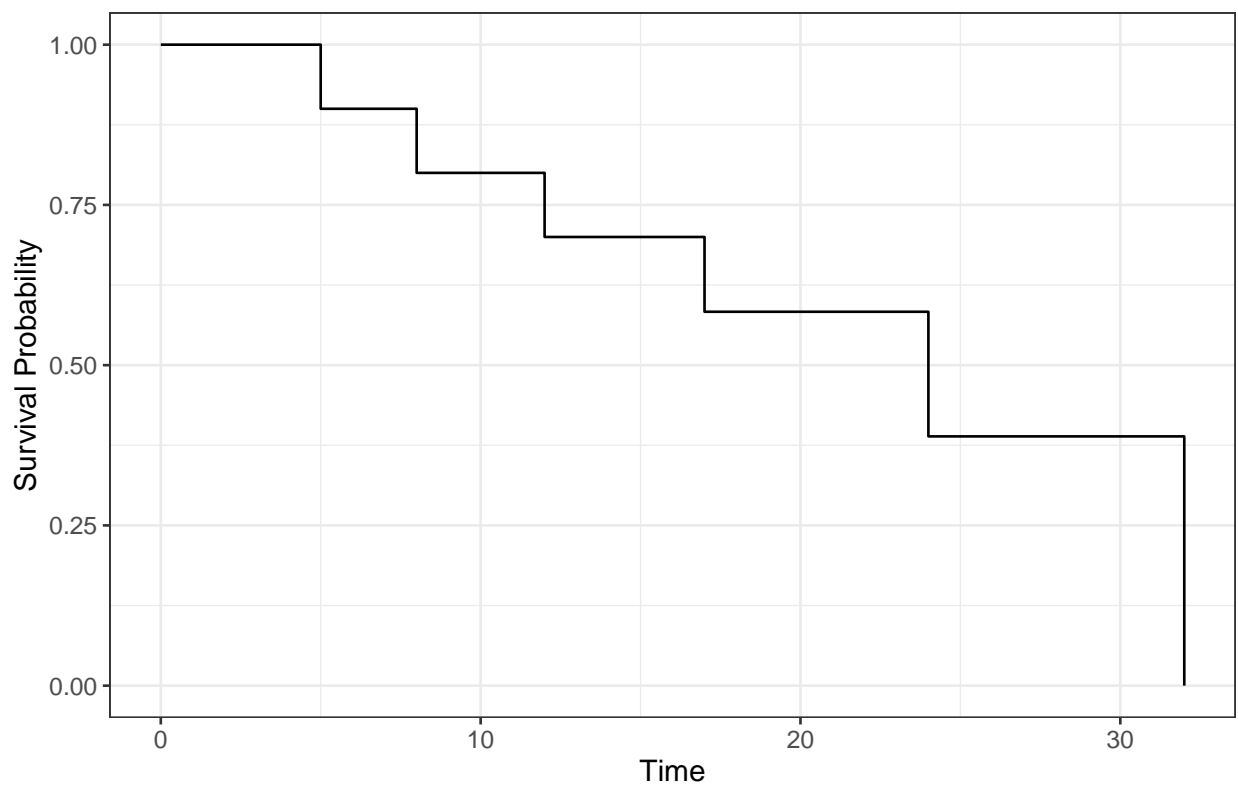
```

library(survival)
library(ggsurvfit)

# set up data
df1 <- data.frame(
  relapse_time = c(5, 8, 12, 24, 32, 17, 16, 17, 19, 30),
  relapse_censored = c(1, 1, 1, 1, 1, 1, 0, 0, 0, 0) # 1: event, 0: censored
)

# fit KM curve
relapse_surv <- Surv(df1$relapse_time, df1$relapse_censored)
relapse_km <- survfit(relapse_surv ~ 1)
relapse_km |>
  ggsurvfit()

```



This tells us that the median time τ s.t. $\hat{S}_R(\tau) \leq 0.50$ is 24 months. Meanwhile, we cannot estimate median time to death because the survival probability does not go below 0.50.

2. Kaplan-Meier Survival Estimate

- (a) I will make a table with a row for every death or censoring time.
 - 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) I will use R to calculate $\hat{S}(t)$ and their pointwise 95% confidence intervals using the “log-log” approach and the linear approach.

```
# set up data
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
)

# fit KM curve
surv <- Surv(df2$t, df2$c)
km <- survfit(surv ~ 1)

# log-log approach to obtain 95%CI
l_loglog <- km.ci(km, method = "loglog")$lower
u_loglog <- km.ci(km, method = "loglog")$upper

# linear approach to obtain 95%CI
l_linear <- km.ci(km, method = "linear")$lower
u_linear <- km.ci(km, method = "linear")$upper

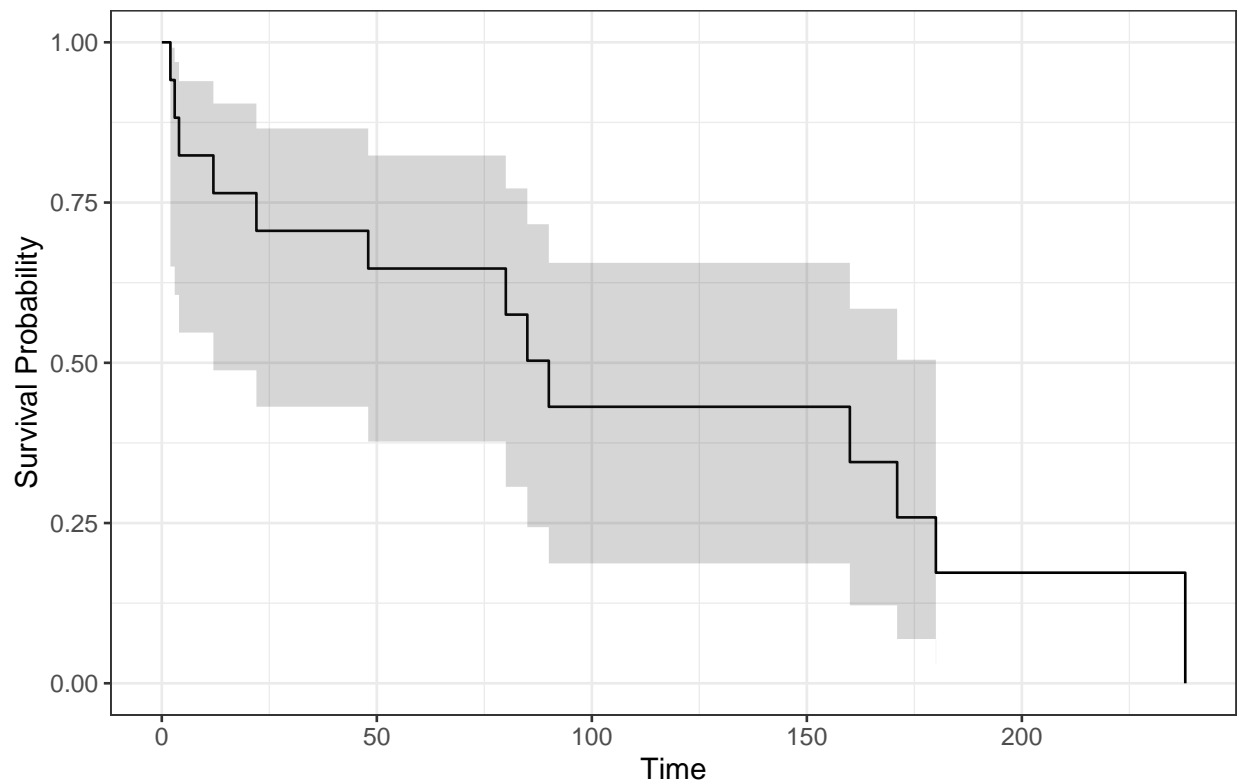
# Create a table
tb <- data.frame(
  t = round(km$time,3), # time
  st = round(km$surv,3), # survival
  l_loglog = round(l_loglog, 3),
  u_loglog = round(u_loglog, 3),
  l_linear = round(l_linear, 3),
  u_linear = round(u_linear, 3)
)
kable(tb, col.names = c("$t_j$", "$\\hat{S}(t_j)$", "Lower 95%CI (log-log)", "Upper 95%CI (log-log)", "Lower 95%CI (linear)", "Upper 95%CI (linear)"))
```

t_j	$\hat{S}(t_j)$	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

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

(c) I will use 95%CI by log-log approach.

```
km.ci(km, method = "loglog") |>
  ggsurvfit() +
  add_confidence_interval()
```



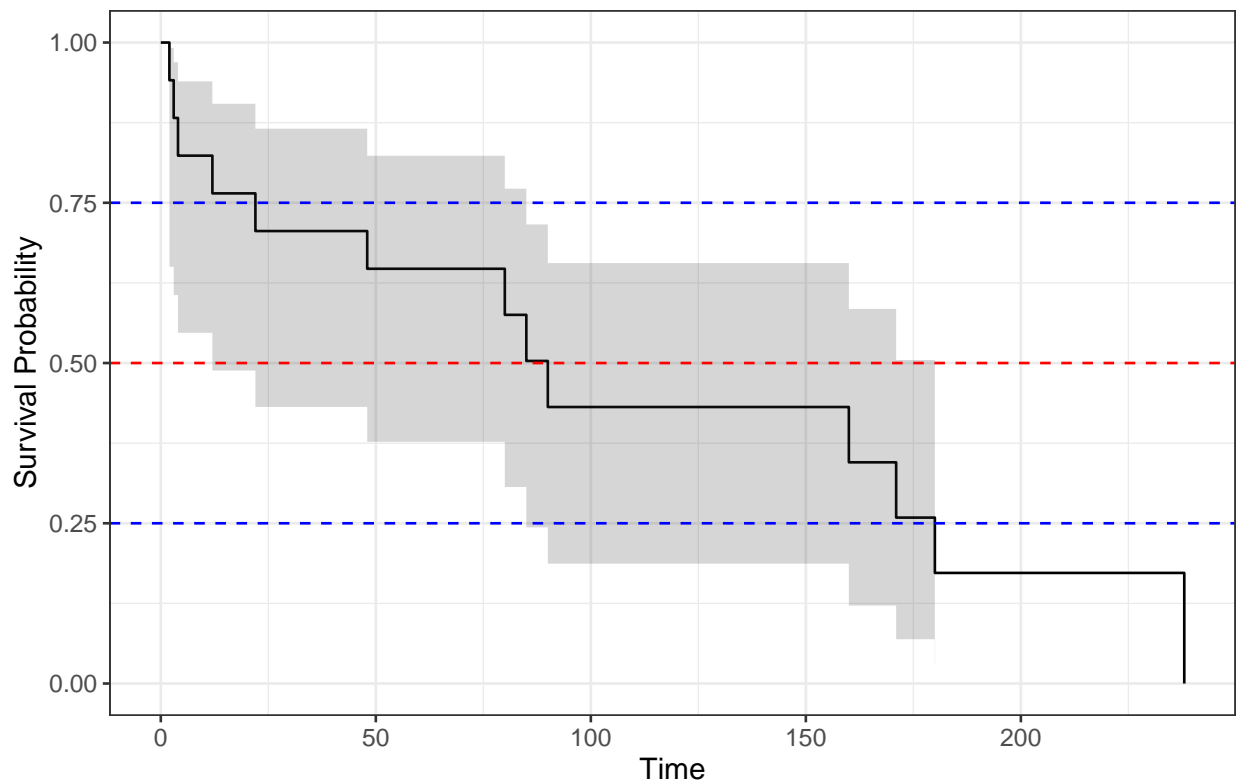
(d)

```
# estimated median survival time
median <- summary(km)$table["median"]

# 25th and 75th percentiles
p25 <- km$time[km$surv <= 0.75][1] # S(t) <= 0.75
p75 <- km$time[km$surv <= 0.25][1] # S(t) <= 0.25
```

The estimated median, 25th and 75th percentile survival time are 90, 22, 180 months, respectively. Let's add this to the KM plot.

```
km.ci(km, method = "loglog") |>
  ggsurvfit() +
  add_confidence_interval() +
  geom_hline(yintercept = 0.5, color = "red", linetype = "dashed") + # median
  geom_hline(yintercept = 0.75, color = "blue", linetype = "dashed") + # 25th
  geom_hline(yintercept = 0.25, color = "blue", linetype = "dashed") # 75th
```



The actual KM survival estimates corresponding to median, 25th and 75th percentile are 0.431, 0.706, 0.173, respectively.

(e) Cumulative hazard rate using the KM survival estimates: $\hat{\Lambda}_{KM}(t) = -\log \hat{S}_{KM}(t)$

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	0.435	0.435	0.553	0.687	0.841	0.841	1.064	1.352	1.757	∞

(f) Cumulative hazard rate using the Nelson-Aalen estimator: $\hat{\Lambda}_{NA}(t) = \sum_{t_j \leq t} \frac{d_j}{r_j}$

```
# Create a table
tb <- data.frame(
  t = round(km$time,3), # time
  ch = round(cumsum((km$n.event)/(km$n.risk)),3) # NA estimator
)
kable(tb, col.names = c("$t_j$", "$\\hat{\\Lambda}_{NA}(t)$"), escape = FALSE)
```

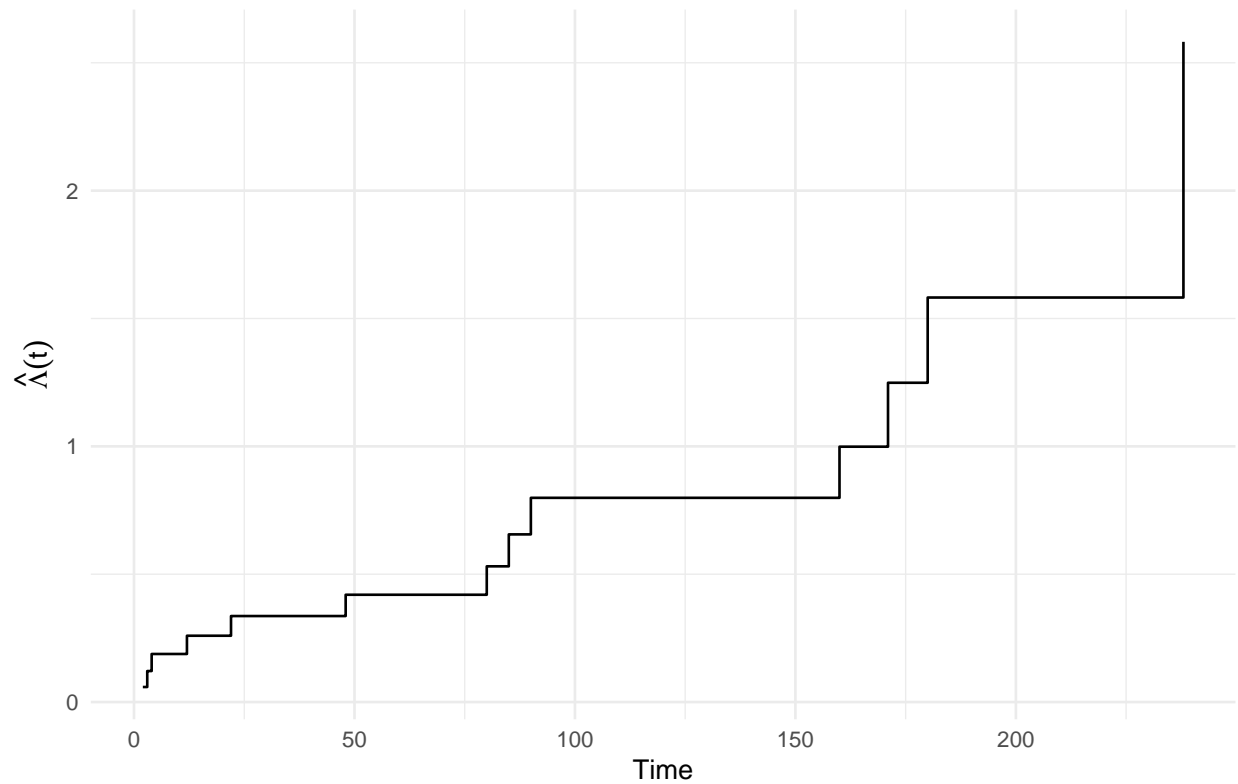
t_j	$\hat{\Lambda}_{NA}(t)$
2	0.059
3	0.121
4	0.188
12	0.259
22	0.336
48	0.420
51	0.420
56	0.420
80	0.531
85	0.656
90	0.799
94	0.799
160	0.999
171	1.249
180	1.582
238	2.582

(g)

(h) Plot the estimated cumulative hazard $\hat{\Lambda}(t)$ vs t

```
# prepare data for plot
tb_ht <- data.frame(
  t = km$time,
  ch = cumsum((km$n.event)/(km$n.risk)) # NA estimator
)

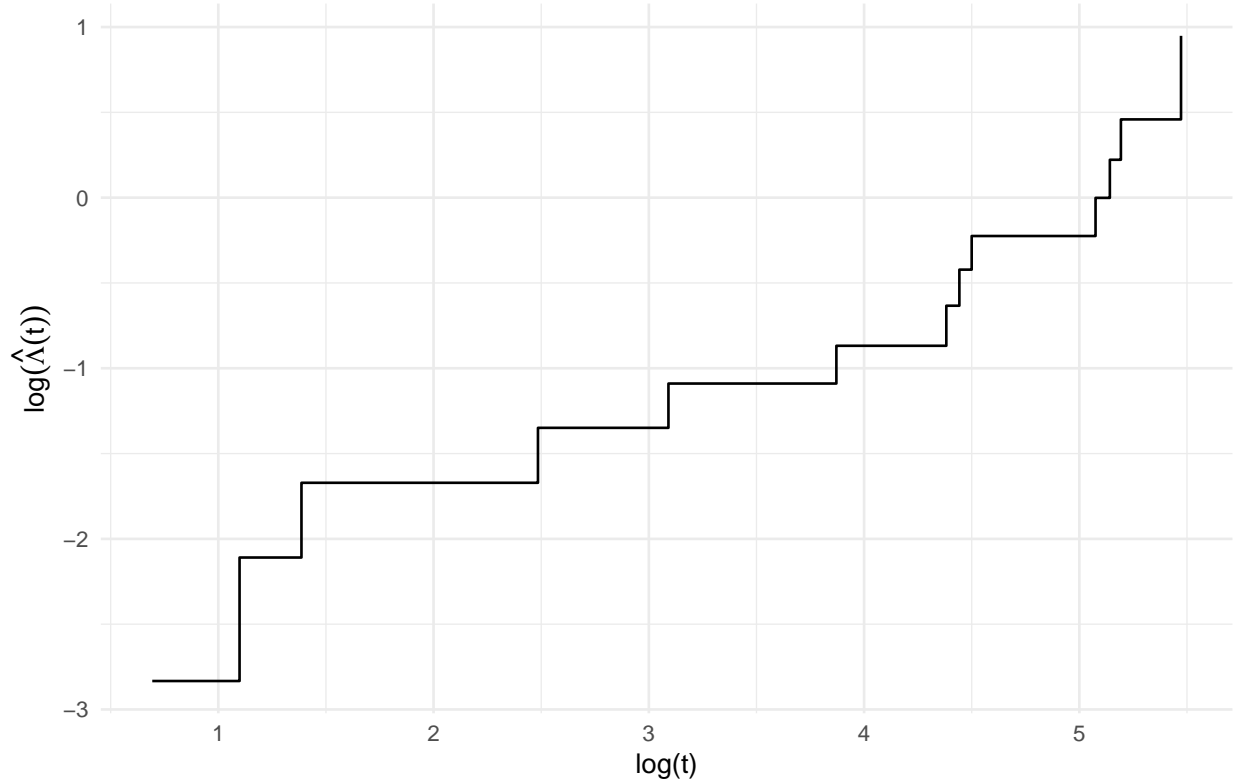
# plot cumulative hazard vs time
ggplot(tb_ht, aes(x = t, y = ch)) +
  geom_step() + # Step plot for cumulative hazard
  labs(
    x = "Time",
    y = expression(hat(Lambda)(t)) +
  theme_minimal()
```



(ii) Plot the estimated log cumulative hazard $\log \hat{\Lambda}(t)$ vs $\log(t)$

```
# prepare data for plot
tb_ht_log <- data.frame(
  t = log(km$time),
  ch = log(cumsum((km$n.event)/(km$n.risk))) # NA estimator
)

# plot cumulative hazard vs time
ggplot(tb_ht_log, aes(x = t, y = ch)) +
  geom_step() + # Step plot for cumulative hazard
  labs(
    x = "log(t)",
    y = expression(log(hat(Lambda)(t))) +
  theme_minimal()
```



Plot (i) shows some deviations from linearity in the cumulative hazard plot, while plot (ii) appears more linear. This suggests that the Weibull model may provide a better fit for the data.

(h) Let's calculate Fleming-Harrington estimator of the survival function using $\hat{S}_{FH}(t) = \exp(-\hat{\Lambda}_{NA}(t))$.

```
tb <- data.frame(
  t = round(km$time,3), # time
  ch = round(cumsum((km$n.event)/(km$n.risk)),3), # NA estimator
  st = round(km$surv,3), # survival
  sfh = round(exp(-cumsum((km$n.event)/(km$n.risk))),3) # FH estimator
)
kable(tb, col.names = c("$t_j$", "$\\hat{\\Lambda}_{NA}(t)$", "$\\hat{S}(t_j)$", "$\\hat{S}_{FH}(t_j)$"))
```

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
51	0.420	0.647	0.657
56	0.420	0.647	0.657
80	0.531	0.575	0.588
85	0.656	0.503	0.519
90	0.799	0.431	0.450
94	0.799	0.431	0.450

t_j	$\hat{\Lambda}_{NA}(t)$	$\hat{S}(t_j)$	$\hat{S}_{FH}(t_j)$
160	0.999	0.345	0.368
171	1.249	0.259	0.287
180	1.582	0.173	0.206
238	2.582	0.000	0.076

The table indicates that both $\hat{S}(t_j)$ and $\hat{S}_{FH}(t_j)$ estimators show close agreement throughout much of the time period, especially for smaller values of t_j .

3. Lifetable (Actuarial) Survival Estimate

1-month intervals: [0, 30), [30, 60), [60, 90), [90, 120), [120, 150), [150, 180), [180, 210), [210, 240)

(a) The actuarial estimate of the survival function can be calculated by $\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$.

```
# calculate variables for lifetable
int = c("[0,30)", "[30,60)", "[60,90)", "[90,120)", "[120, 150)", "[150, 180)", "[180, 210)", "[210, 240)")
j = 1:8
dj = c(5, 1, 2, 1, 0, 2, 1, 1)
cj = c(0, 2, 0, 1, 0, 0, 1, 0)
rj = c(17, 12, 9, 7, 5, 5, 3, 1)
s = cumprod(1 - dj/rj)

# lifetable
tb <- data.frame(
  int, j, dj, cj, rj, round(s, 3)
)
kable(tb, col.names = c("$Days After Treatment$", "$j$", "$d_j$", "$c_j$", "$r_j$", "$\\hat{S}(t)$"), e
```

<i>DaysAfterTreatment</i>	<i>j</i>	<i>d_j</i>	<i>c_j</i>	<i>r_j</i>	$\hat{S}(t)$
[0,30)	1	5	0	17	0.706
[30,60)	2	1	2	12	0.647
[60,90)	3	2	0	9	0.503
[90,120)	4	1	1	7	0.431
[120, 150)	5	0	0	5	0.431
[150, 180)	6	2	0	5	0.259
[180, 210)	7	1	1	3	0.173
[210, 240)	8	1	0	1	0.000

(b) Let the width of the j -th interval be $b_j = t_j - t_{j-1} = 30$ and the conditional probability of dying be $\hat{q}_j = d_j/r'_j$ where $r'_j = r_j - c_j/2$. Also, let t_{mj} be the midpoint of j -th interval ($t_{mj} = \frac{t_j + t_{j-1}}{2}$). The estimated hazard in the j -th interval is $\hat{\lambda}(t_{mj}) = \frac{\hat{q}_j}{b_j(1 - \hat{q}_j/2)}$.

```
bj = 30
rj_dash = rj - cj/2
qj = dj/rj_dash
tmj = c(15, 45, 75, 105, 135, 165, 195, 225)
```

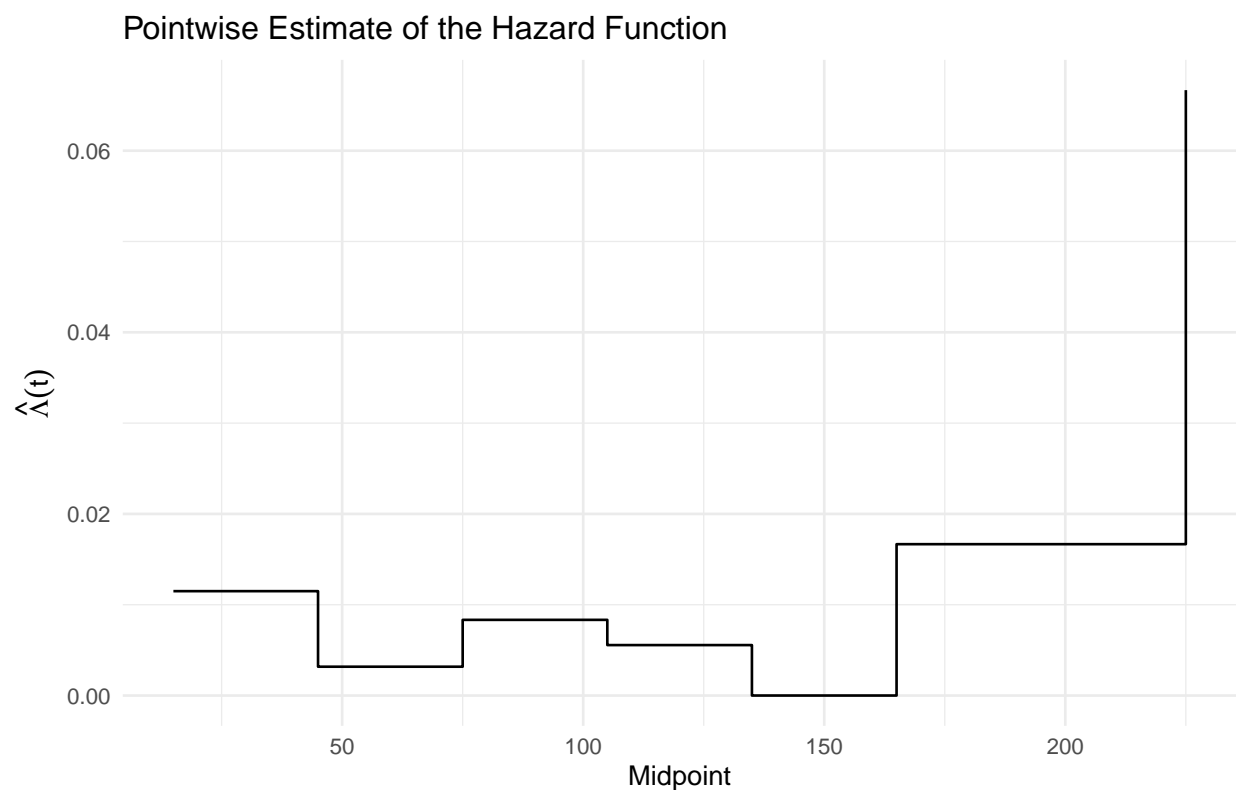
```

hj = qj/(bj*(1-qj/2)) # hazard
hj_c = cumsum(hj)

# prepare data for plot
tb_hj <- data.frame(tmj, hj)
tb_hj_c <- data.frame(tmj, hj_c)

# plot
ggplot(tb_hj, aes(x = tmj, y = hj)) +
  geom_step() +
  labs(title = "Pointwise Estimate of the Hazard Function",
       x = "Midpoint",
       y = expression(hat(Lambda)(t))) +
  theme_minimal()

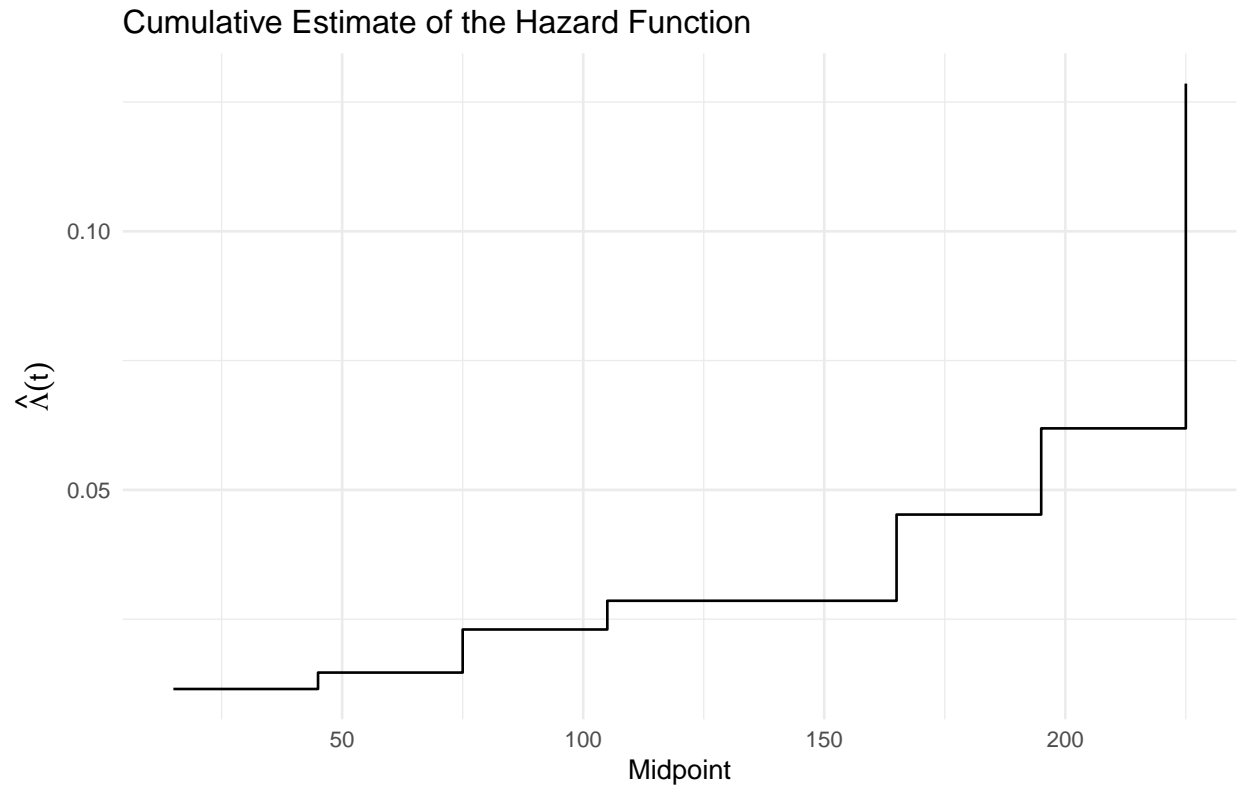
```



```

ggplot(tb_hj, aes(x = tmj, y = hj_c)) +
  geom_step() +
  labs(title = "Cumulative Estimate of the Hazard Function",
       x = "Midpoint",
       y = expression(hat(Lambda)(t))) +
  theme_minimal()

```



- (c) The hazard for treatment response over time is not constant; it increases as time progresses. The plot in (b) does not appear to be linear. Therefore, the exponential model seem inappropriate for this data.

4. Non-Informative and Informative Censoring

Braitstein et al. pointed out that the mortality rate may have been underestimated by non-informative censoring, with previous study indicating higher mortality among children with LTFU. To address this, the authors attempted to locate them and collect information on the reason that had led to LTFU.

Brinkhof et al. also addressed the issue of potential bias caused by non-informative censoring due to LTFU. They adjusted for mortality by estimating the hazard ratio associated with LTFU and accommodated this by imputing missing data based on these estimates.