STAT6390_HW2_CongZhang

Cong Zhang

October 6, 2018

1. Show that (algebraically) in the absence of censoring $\hat{S}_{\text{KM}}(t) = \hat{S}_{\text{e}}(t)$.

Answer

$$\hat{S}_{\text{KM}}(t) = P(T > t) = P(T > t_{(0)}) \cdot P(T > t_{(1)} | T > t_{(0)}) \cdot ... \cdot P(T > t | T > t_{(i)})$$

for a series of time intervals $0 = t_0 < t_1 < ... < t_i < t$ for some $i \leq n$.

When there is no censoring, we have

$$\hat{S}_{\text{KM}}(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i} = 1 \cdot \frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdot ... \cdot \frac{n_i - \sum_{j=1}^{i} d_j}{n_i} = \frac{\sum_{i=1}^{n} I(T_i > t)}{n} = \hat{S}_e(t)$$

2. In the absence of censoring, show that the Greenwood Formula (page 30 on note 2) can be reduced to

$$\frac{\hat{S}_{\text{KM}}(t) \times \{1 - \hat{S}_{\text{KM}}(t)\}}{n}.$$

You might assume there are no ties among the observations.

Answer

$$\begin{aligned}
&\hat{S}^2_{\text{KM}}(t) \cdot \sum_{t_{(i)} \leq t} \frac{d_i}{n_i \cdot (n_i - d_i)} \\
&= \hat{S}^2_{\text{KM}}(t) \cdot (\frac{1}{n_i - d_i} - \frac{1}{n_i} + \frac{1}{n_{i-1} - d_{i-1}} - \frac{1}{n_{i-1}} + ... + \frac{1}{n_1 - d_1} - \frac{1}{n_1}) \\
&= \hat{S}^2_{\text{KM}}(t) \cdot (\frac{1}{n_1 - \sum_{j=1}^{i} d_j} - \frac{1}{n_1 - \sum_{j=1}^{i-1} d_j} + \frac{1}{n_1 - \sum_{j=1-1}^{i} d_j} - \frac{1}{n_1 - \sum_{j=1}^{i-2} d_j} + ... + \frac{1}{n_1 - d_1} - \frac{1}{n_1}) \\
&= \hat{S}^2_{\text{KM}}(t) \cdot (\frac{1}{n_1 - \sum_{j=1}^{i} d_j} - \frac{1}{n_1}) \\
&= \hat{S}^2_{\text{KM}}(t) \cdot \frac{\sum_{j=1}^{i} d_j}{n_1(n_1 - \sum_{j=1}^{i} d_j)} \\
&= \hat{S}_{\text{KM}}(t) \cdot \frac{n_1 - \sum_{j=1}^{i} d_j}{n_1} \cdot \frac{\sum_{j=1}^{i} d_j}{n_1} \cdot \frac{1}{n_1 - \sum_{j=1}^{i} d_j} \\
&= \hat{S}_{\text{KM}}(t) \cdot (1 - \frac{n_1 - \sum_{j=1}^{i} d_j}{n_1}) \cdot \frac{1}{n_1} \\
&= \frac{\hat{S}_{\text{KM}}(t) \times \{1 - \hat{S}_{\text{KM}}(t)\}}{n}
\end{aligned}$$

3. Consider the Leukemia data from the `survival` package:

```
library(survival)
head(aml)
```

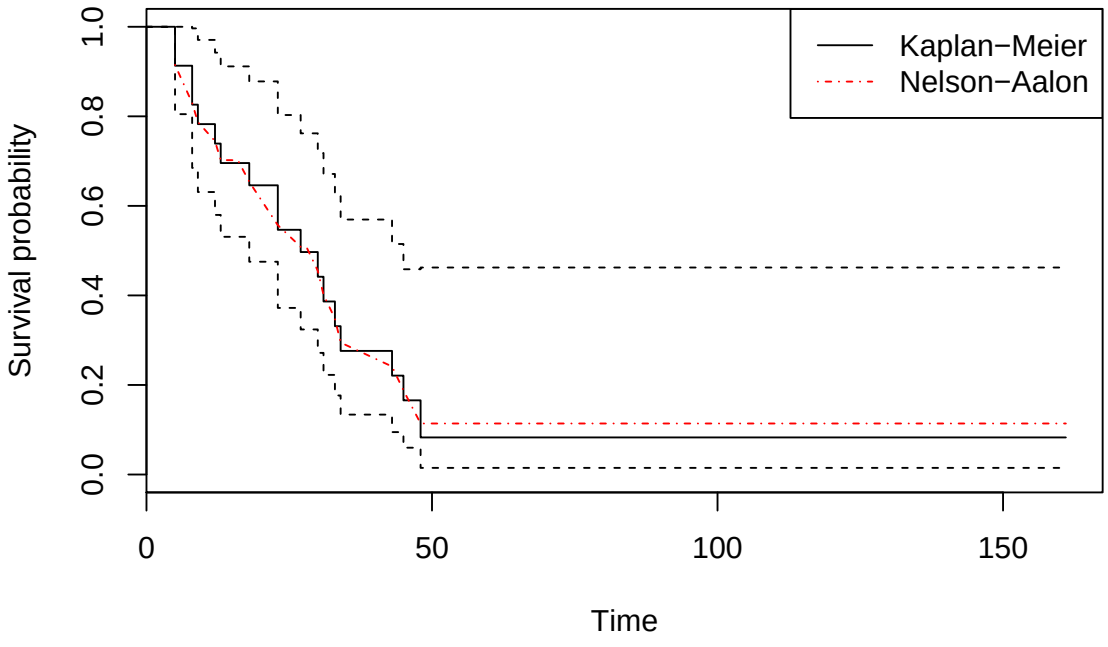
```
##   time status          x
## 1    9      1 Maintained
## 2   13      1 Maintained
## 3   13      0 Maintained
## 4   18      1 Maintained
## 5   23      1 Maintained
## 6   28      0 Maintained
```

In here, each row represent one patient. `aml` is the observed survival time, `status` is the censoring indicator (1 = event, 0 = censored), and `x` is the treatment indicator. We will ignore the treatment indicator for now.

a. Plot the Kaplan-Meier survival curve for the data.
b. Add the Nelson-Aalen survival curve to the Kaplan-Meier plot from (3a).

Answer

```
library(survival)
library(survMisc)
library(knitr)
km <- survfit(Surv(time, status) ~ 1, data = aml)
cox <- coxph(Surv(time, status) ~ 1, data = aml)
H0 <- basehaz(cox)
plot(km, xlab = "Time", ylab = "Survival probability")
lines(H0$time, exp(-H0$hazard), lty = 4, col = 2)
legend("topright" , c("Kaplan-Meier", "Nelson-Aalon"), col = c(1, 2), lty = c(1, 4) )
```



4. The expected survival time for the Leukemia data in #(3) does not exist because the last observation is a censored event. An alternative is to lookInstead of looking at the expected survival time, an alternative is to look at the restricted mean survival time. Compute $E(T|T < 161)$ based on the survival

curve in (3a).

Answer $E(T|T < 161) = 36.36$

```
survival:::survmean(km, rmean=161)
```

```
## $matrix
##    records      n.max    n.start     events     *rmean *se(rmean)
##  23.000000  23.000000  23.000000  18.000000  36.364389   9.854101
##     median    0.95LCL    0.95UCL
##  27.000000  18.000000  45.000000
##
## $end.time
## [1] 161
```

5. Let $N_i(t)$ be the number of events over time interval $(0, t]$ for the ith patient in #(3). Let $N(t) = \sum_{i=1}^{n} N_i(t)$ be the aggregated counting process.
   a. Plot $N(t)$.
   b. Plot $M(t)$, where $M(t) = N(t) - \hat{H}(t)$ and $\hat{H}(t)$ is the Nelson-Aalen estimator for the cumulative hazard function.

   - Note on 5b: After giving some thought, I think it is more meaningful to plot $dM(t) = dN(t) - \hat{h}(t)dt$. Both plots will receive full credit for 5b.

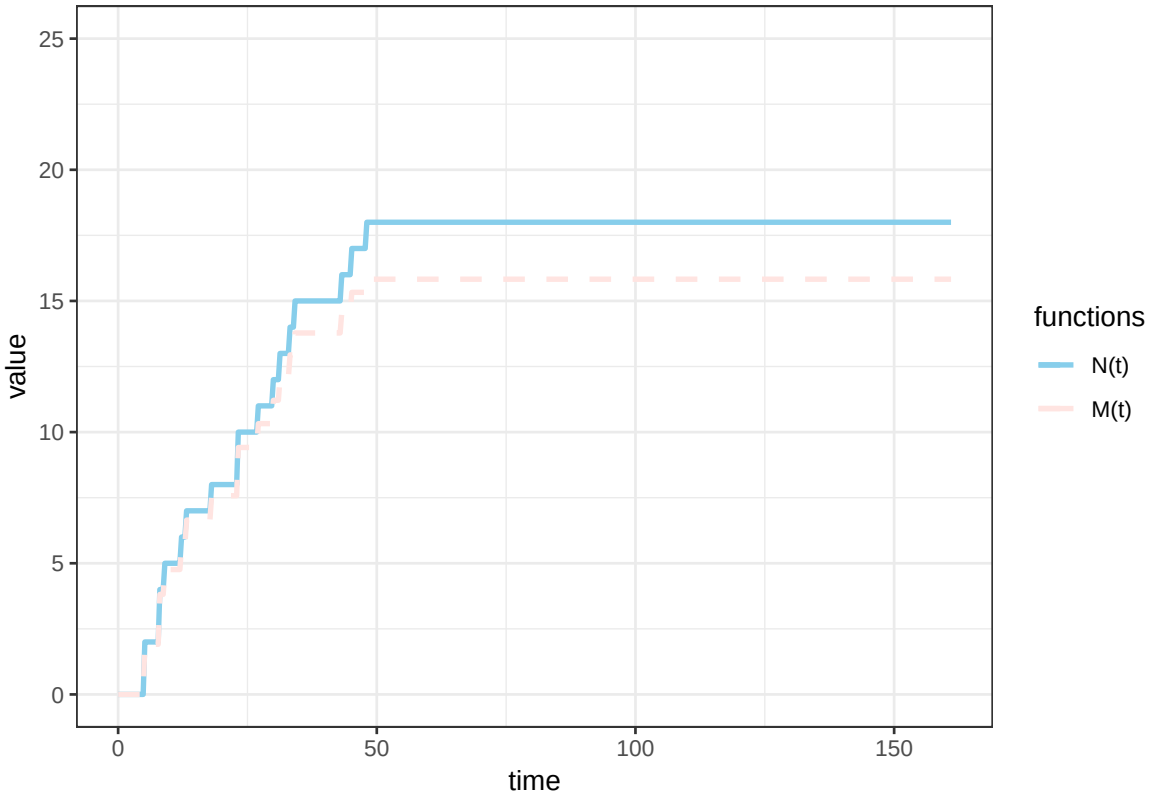
Answer

```
library(ggplot2)
library(dplyr)
N <- Vectorize(function(t)
{
    sum(aml[aml$time <= t, "status"])
})

HNA <- Vectorize(function(t)
{
    if (H0$time[1] > t) 0 else H0[last(which(H0$time <= t)), "hazard"]
})

M <- function(t)
{
    N(t) - HNA(t)
}

xlim <- c(0, max(aml$time))
ylim <- c(0, 25)
ggplot(data = data.frame(x = xlim), aes(x)) +
    stat_function(fun = N, aes(color = "N"), size = 1, n = 500) +
    stat_function(fun = M, aes(color = "M"), size = 1, n = 500, linetype = "dashed") +
    scale_x_continuous(name = "time",  limit = xlim) +
    scale_y_continuous(name = "value", limit = ylim,
                       breaks =seq(ylim[1], ylim[2], length.out = 6)) +
    scale_color_manual(name = "functions",
                       values = c("N" = "skyblue", "M" = "mistyrose"),
                       breaks = c("N","M"),
                       labels = c("N(t)","M(t)")) +
    theme_bw()
```


25
20
15
10
5
0
value
0
50
100
150
time
functions
N(t)
M(t)