Homework 2

Xinyu Zhang

Due date: Thursday, October 11

1. Show that (algebraically) in the absence of censoring $\hat{S}_{\text{KM}}(t) = \hat{S}_{\text{e}}(t)$.
   Assume $m = max\{i : t_{(i)} \leq t\}$, *then* :

   $\hat{S}_{\text{KM}}(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i} = \frac{n-d_1}{n} \cdot \frac{n-d_1-d_2}{n-d_1} \cdot \cdots \cdot \frac{n-d_1-d_2-\cdots-d_m}{n-d_1-d_2-\cdots-d_{m-1}} = \frac{n-d_1-d_2-\cdots-d_m}{n}$
   $= \frac{\#individuals\ with\ survival\ times \geq t}{\#individuals\ in\ the\ data\ set} = \hat{S}_{\text{e}}(t)$

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.
   $Var\{\widehat{S}_{KM}(t)\} \approx \widehat{S}^2_{KM}(t) \cdot \sum_{t_{(i)} \leq t} \frac{d_i}{n_i \cdot (n_i - d_i)}$
   $= (\frac{n-d_1-d_2-\cdots-d_m}{n})^2 \cdot \sum_{t_{(i)} \leq t} \frac{d_i}{n_i \cdot (n_i - d_i)}$ *assume no censoring, follow from Q*1
   $= (\frac{n-t_{(m)}}{n})^2 \cdot \sum_{t_{(i)} \leq t} \frac{1}{n_i \cdot (n_i - 1)}$ *assume no ties*
   $= (\frac{n-t_{(m)}}{n})^2 \cdot (\frac{1}{n(n-1)} + \frac{1}{(n-1)(n-2)} + \cdots + \frac{1}{(n-(t_{(m)}-1))(n-t_{(m)})})$ *assume no censoring*
   $= (\frac{n-t_{(m)}}{n})^2 \cdot ((-\frac{1}{n} + \frac{1}{n-1}) + (-\frac{1}{n-1} + \frac{1}{n-2}) + \cdots + (-\frac{1}{n-(t_{(m)}-1)} + \frac{1}{n-t_{(m)}}))$
   $= (\frac{n-t_{(m)}}{n})^2 \cdot (\frac{1}{n-t_{(m)}} - \frac{1}{n})$
   $= \frac{\frac{n-t_{(m)}}{n} - (\frac{n-t_{(m)}}{n})^2}{n}$
   $= \frac{\widehat{S}_{KM}(t) \times \{1 - \widehat{S}_{KM}(t)\}}{n}$

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

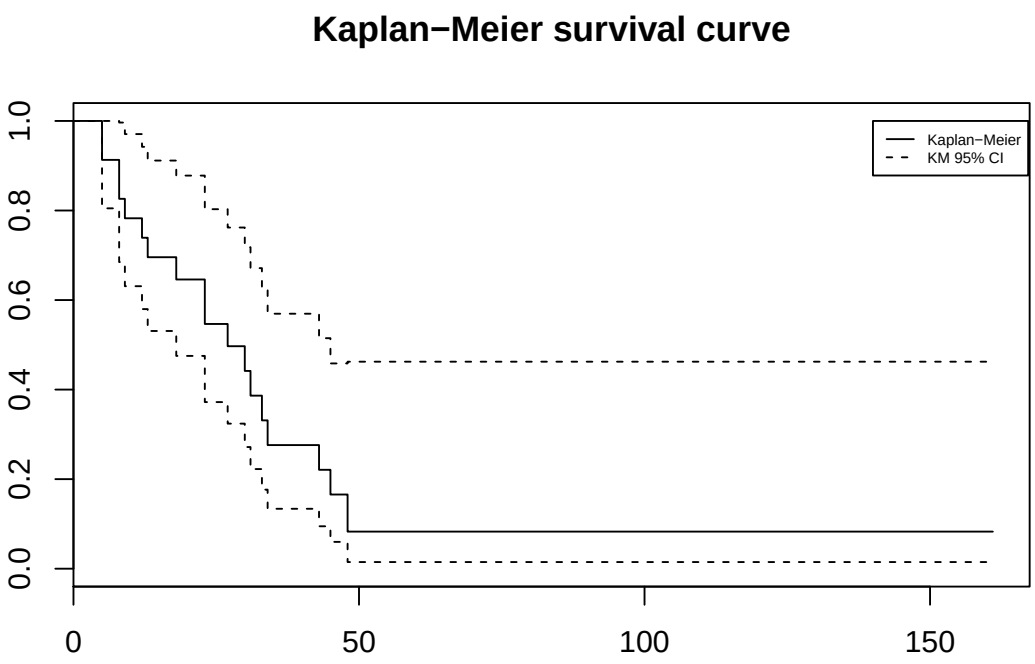
```
> library(survival)
> head(aml)
```

```
  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.

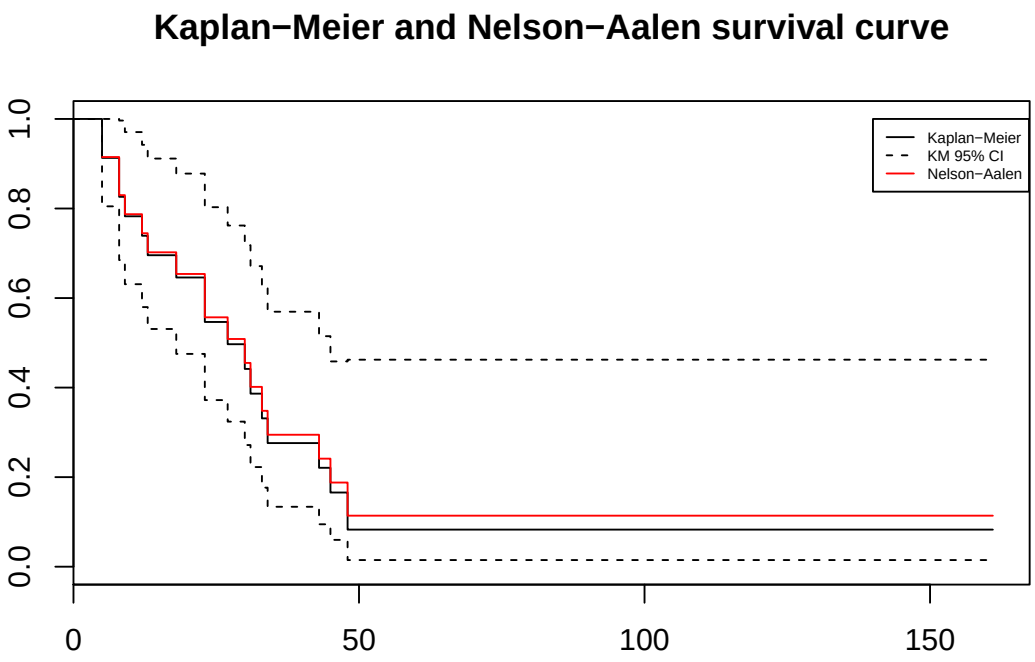
   a. Plot the Kaplan-Meier survival curve for the data.

```
> km <- survfit(Surv(time, status) ~ 1, data = aml)
> plot(km,main="Kaplan-Meier survival curve")
> legend(140, 1, legend=c("Kaplan-Meier","KM 95% CI"),col=c("black","black"),
+ lty=c(1,2), cex=0.5)
```



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

```
> cox <- coxph(Surv(time, status) ~ 1, data = aml)
> H0 <- basehaz(cox)
> plot(km,main="Kaplan-Meier and Nelson-Aalen survival curve")
> lines(H0$time, exp(-H0$hazard), 's', col = 2)
> legend(140, 1, legend=c("Kaplan-Meier","KM 95% CI", "Nelson-Aalen"),
+ col=c("black","black", "red"),lty=c(1,2,1), cex=0.5)
```



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).

```
> matrix(c(km$time,(km$surv-tail(km$surv, n=1))/(1-tail(km$surv, n=1))),nrow=length(km$time),
+ dimnames = list(NULL, c("death_time","km_surv"))) %>%
+ as.data.frame() %>%
+ add_row(death_time = 0, km_surv = 1,.before = 1) %>%
+ arrange(desc(km_surv)) %>%
+ mutate(time_diff=lead(death_time,default = 0)-death_time) %>%
+ filter(death_time<161) %>%
+ summarise(expected_lifetime=sum(time_diff*km_surv))
```

```
  expected_lifetime
1          25.11061
```

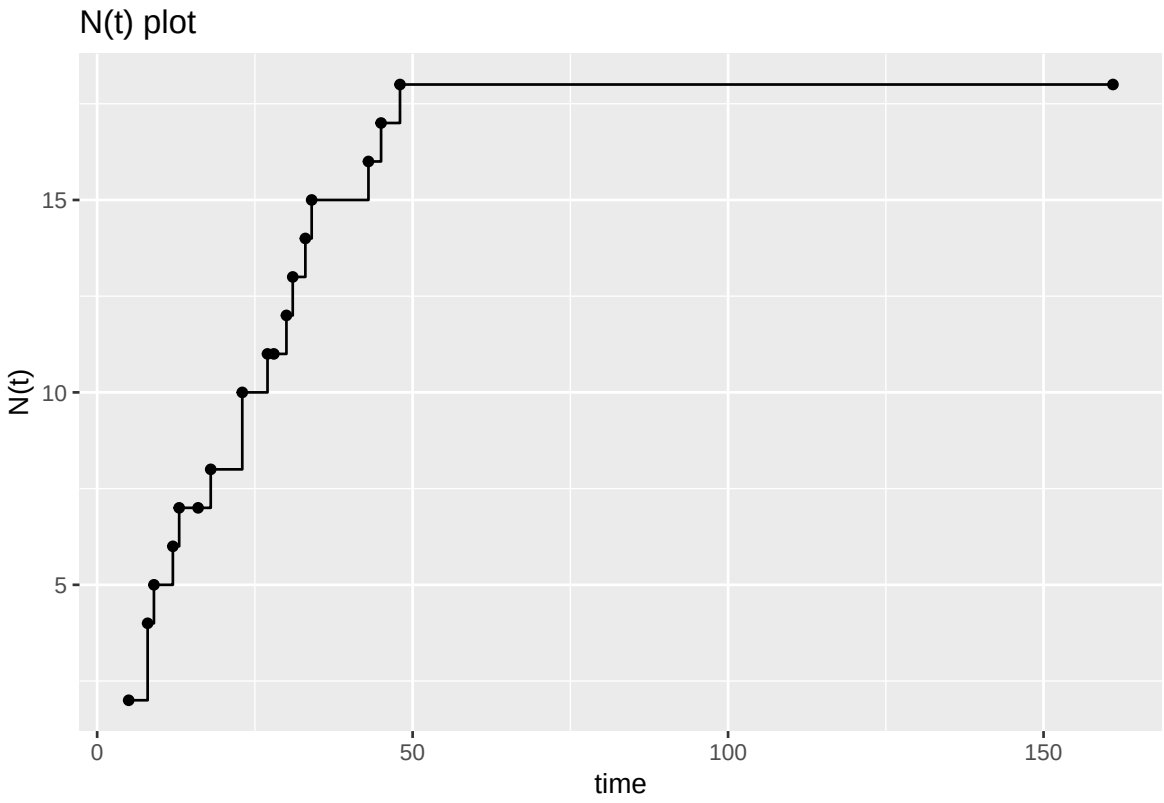
The conditional expected lifetime is 25.11061.

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 $N(t)$.

```
> calculate_Nt=
+ aml %>% arrange(time) %>% select(-x) %>% group_by(time) %>%
+ summarize(di = sum(status)) %>%
```

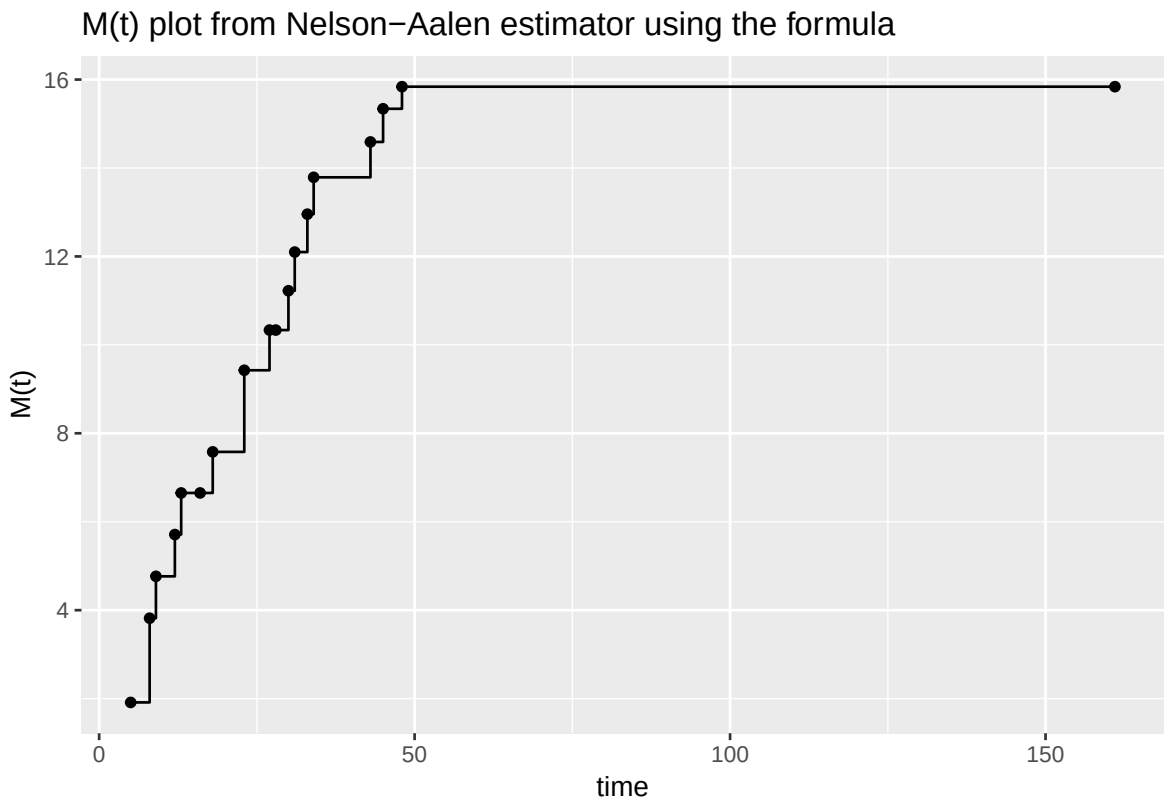
```
+ mutate(Nt = cumsum(di))
>
> qplot(time, Nt, data = calculate_Nt, geom = "step",
+ main="N(t) plot") +
+ ylab("N(t)") + geom_point()
```



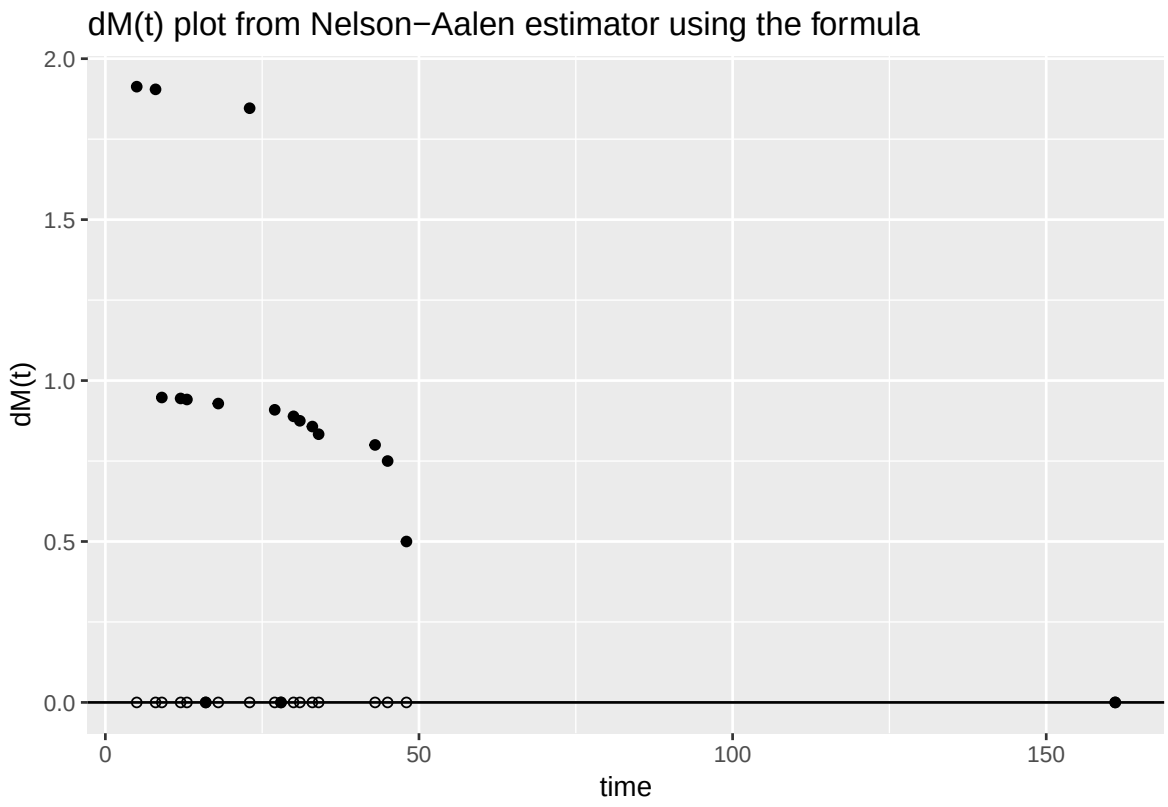
b. Plot $M(t)$, where $M(t) = N(t) - \hat{H}(t)$ and $\hat{H}(t)$ is the Nelson-Aalen estimator for the cumulative hazard function.

- 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.

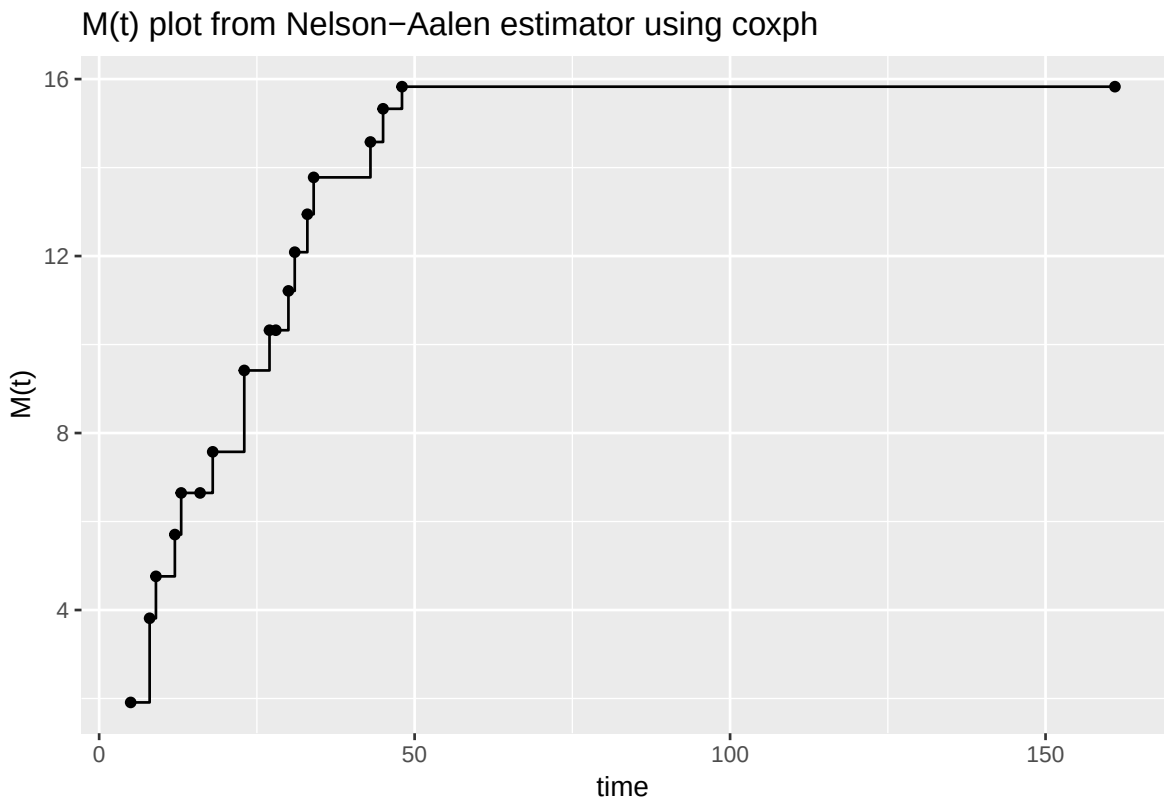
```
> na_surv_by_formula=
+ aml %>% arrange(time) %>% select(-x) %>% group_by(time) %>%
+ summarize(di = sum(status), ni = length(status)) %>%
+ mutate(ni = rev(cumsum(rev(ni))), hi = di/ni,
+ Nt = cumsum(di), Ht = cumsum(hi), Mt = Nt-Ht,
+ dNt = di, htdt = hi, dMt = di-hi)
>
> qplot(time, Mt, data = na_surv_by_formula, geom = "step",
+ main="M(t) plot from Nelson-Aalen estimator using the formula") +
+ ylab("M(t)") + geom_point()
```



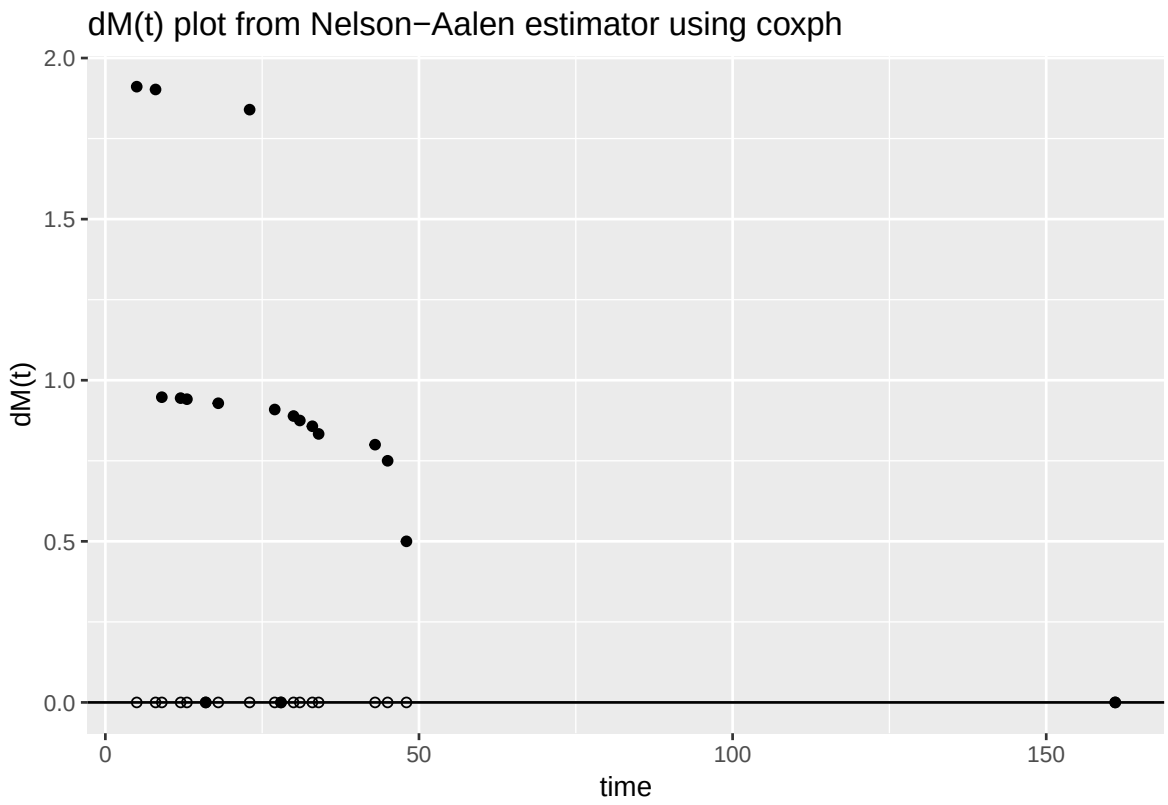

```
> qplot(time, dMt, data = na_surv_by_formula,
+ main="dM(t) plot from Nelson-Aalen estimator using the formula") +
+ geom_abline(intercept = 0, slope = 0) +
+ geom_point(aes(na_surv_by_formula$time, rep(0,18)), pch=1) + ylab("dM(t)")
```




```
> na_surv_by_cosph=
+ aml %>% arrange(time) %>% select(-x) %>% group_by(time) %>%
+ summarize(di = sum(status), ni = length(status)) %>%
+ mutate(ni = rev(cumsum(rev(ni))), hi = di/ni,
+ Nt = cumsum(di), Ht = H0$hazard, Mt = Nt-Ht,
+ dNt = di, htdt = diff(c(0,Ht)), dMt = di-htdt)
>
> qplot(time, Mt, data = na_surv_by_cosph, geom = "step",
+ main="M(t) plot from Nelson-Aalen estimator using coxph") +
+ ylab("M(t)") + geom_point()
```




```
> qplot(time, dMt, data = na_surv_by_cosph,
+ main="dM(t) plot from Nelson-Aalen estimator using coxph") +
+ geom_abline(intercept = 0, slope = 0) +
+ geom_point(aes(na_surv_by_cosph$time, rep(0,18)), pch=1) + ylab("dM(t)")
```



Since the estimators from the formula and the coxph are very close, they plots are almost identical.