Nonparametric Survival Analysis

David M. Rocke

October 8, 2024

Bone Marrow Transplant Data

- Copelan et al. (1991) study of allogeneic (from a donor) bone marrow transplant therapy for acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).
- Possible intermediate events are graft vs. host disease (GVHD), an immunological rejection response to the transplant, and platelet recovery, a return of platelet count to normal levels. One or the other, both in either order, or neither may occur.
- End point events are relapse of the disease or death.
- Any or all of these events may be censored.

KMsurv bmt data

```
The bmt data frame has 137 rows and 22 columns.

This data frame contains the following columns:

group   Disease Group 1-ALL, 2-AML Low Risk, 3-AML High Risk
t1      Time To Death Or On Study Time
t2      Disease Free Survival Time (Time To Relapse, Death, Or End Of Study)
d1      Death Indicator 1-Dead 0-Alive
d2      Relapse Indicator 1-Relapsed, 0-Disease Free
d3      Disease Free Survival Indicator 1-Dead Or Relapsed, 0-Alive Disease Free)
ta      Time To Acute Graft-Versus-Host Disease
da      Acute GVHD Indicator 1-Developed Acute GVHD 0-Never Developed Acute GVHD)
tc      Time To Chronic Graft-Versus-Host Disease
dc      Chronic GVHD Indicator 1-Developed Chronic GVHD
          0-Never Developed Chronic GVHD
tp      Time To Platelet Recovery
dp      Platelet Recovery Indicator 1-Platelets Returned To Normal,
          0-Platelets Never Returned to Normal
```

KMsurv bmt data

```
z1       Patient Age In Years
z2       Donor Age In Years
z3       Patient Sex: 1-Male, 0-Female
z4       Donor Sex: 1-Male, 0-Female
z5       Patient CMV Status: 1-CMV Positive, 0-CMV Negative
z6       Donor CMV Status: 1-CMV Positive, 0-CMV Negative
z7       Waiting Time to Transplant In Days
z8       FAB: 1-FAB Grade 4 Or 5 and AML, 0-Otherwise
z9       Hospital: 1-The Ohio State University, 2-Alferd , 3-St. Vincent,
           4-Hahnemann
z10      MTX Used as a Graft-Versus-Host- Prophylactic: 1-Yes 0-No
```

Bone Marrow Transplant Example

- We concentrate for now on disease-free survival (`t2` and `d3`) for the three risk groups, ALL, AML Low Risk, and AML High Risk.
- We will construct the Kaplan-Meier survival curves, compare them, and test for differences.
- We will construct the cumulative hazard curves and compare them.
- We will estimate the hazard functions, interpret, and compare them.
- Then we will introduce the Cox proportional hazards model.

Survival Function

$$\hat{S}(t) = \prod_{t_i < t} [1 - d_i / Y_i]$$

where Y_i is the group at risk at time t_i.
The estimated variance of $\hat{S}(t)$ is (Greenwood's formula)

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i < t} \frac{d_i}{Y_i(Y_i - d_i)}$$

which we can use for confidence intervals for a survival function or a difference of survival functions.

To see where Greenwood's formula comes from, let $x_i = Y_i - d_i$. We approximate the solution treating each time as independent, with Y_i fixed and ignore randomness in times of failure and we treat x_i as independent binomials $Bin(Y_i, p_i)$. Letting $S(t)$ be the "true" survival function

$$
\begin{aligned}
\hat{S}(t) &= \prod_{t_i < t} x_i / Y_i \\
S(t) &= \prod_{t_i < t} p_i
\end{aligned}
$$

$$\begin{aligned}\frac{\hat{S}(t)}{S(t)} &= \prod_{t_i<t} \frac{x_i}{p_i Y_i} = \prod_{t_i<t} \frac{\hat{p}_i}{p_i} \\ &= \prod_{t_i<t} \left(1 + \frac{\hat{p}_i - p_i}{p_i}\right) \\ &\approx 1 + \sum_{t_i<t} \frac{\hat{p}_i - p_i}{p_i}\end{aligned}$$

because $(\hat{p}_i - p_i)/p_i$ is small and any term with more than one such factor will be negligible.

$$
\begin{aligned}
\mathrm{Var}\left(\frac{\hat{S}(t)}{S(t)}\right) &\approx \mathrm{Var}\left(1 + \sum_{t_i<t} \frac{\hat{p}_i - p_i}{p_i}\right) \\
&= \sum_{t_i<t} \frac{1}{p_i^2} \frac{p_i(1-p_i)}{Y_i} \\
&= \sum_{t_i<t} \frac{(1-p_i)}{p_i Y_i} \approx \sum_{t_i<t} \frac{(1 - x_i/Y_i)}{x_i} \\
&= \sum_{t_i<t} \frac{Y_i - x_i}{x_i Y_i} = \sum_{t_i<t} \frac{d_i}{Y_i(Y_i - d_i)} \\
\mathrm{Var}(\hat{S}(t)) &\approx \hat{S}(t)^2 \sum_{t_i<t} \frac{d_i}{Y_i(Y_i - d_i)}
\end{aligned}
$$

Cumulative Hazard

$$h(t) = -\frac{d \ln S(t)}{dt}$$

The cumulative hazard function is

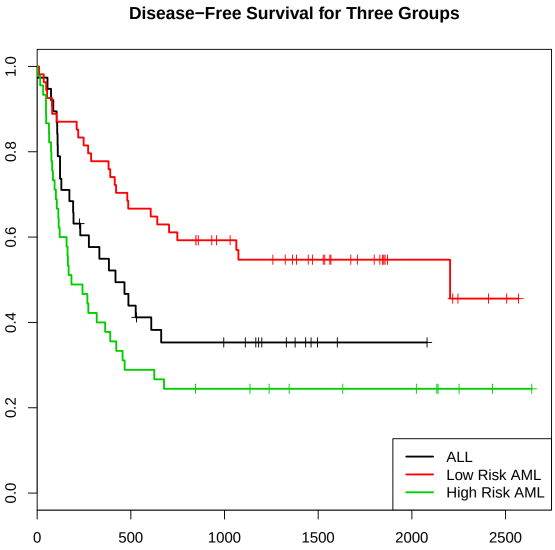
$$\begin{aligned} H(t) &= \int_0^t h(t)dt \\ &= -\ln S(t) \\ \hat{H}(t) &= -\ln \hat{S}(t) \end{aligned}$$

```
> library(KMsurv)
> library(survival)
> data(bmt)
> dfsurv <- Surv(bmt$t2,bmt$d3)
```

The last command creates a survival object from the time variable `t2` (disease-free survival) and the associated status variable `d3`. This is usually the first step in computer analysis of survival data.

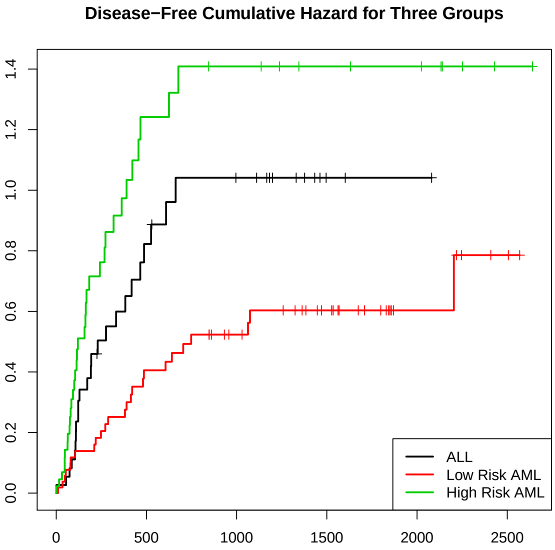
```
> plot(survfit(dfsurv~group,data=bmt),col=1:3,lwd=2)
> title("Disease-Free Survival for Three Groups")
> legend("bottomright",c("ALL","Low Risk AML","High Risk AML"),col=1:3,lwd=2)
```

This plots the estimated survival curves for the three groups on the same graph in three colors with associated legend.


Disease−Free Survival for Three Groups
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
ALL
Low Risk AML
High Risk AML

```
> plot(survfit(dfsurv~group,data=bmt),col=1:3,lwd=2,fun="cumhaz")
> title("Disease-Free Cumulative Hazard for Three Groups")
> legend("bottomright",c("ALL","Low Risk AML","High Risk AML"),col=1:3,lwd=2)
```

This plots the cumulative hazards for the three groups.



Disease−Free Cumulative Hazard for Three Groups
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
500
1000
1500
2000
2500
ALL
Low Risk AML
High Risk AML

```
> survdiff(dfsurv~group,data=bmt)
         N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 38       24     21.9     0.211     0.289
group=2 54       25     40.0     5.604    11.012
group=3 45       34     21.2     7.756    10.529

 Chisq= 13.8  on 2 degrees of freedom, p= 0.00101
```

This tests whether the three groups could have a common survival function. Note that group is treated as a factor even though it is numeric. This is the Mantel-Haenszel test.

Nelson-Aalen Survival Function Estimate

The point hazard at time t_i can be estimated by d_i/Y_i which leads to the estimate of the cumulative hazard

$$\hat{H}(t) = \sum_{t_i<t} d_i/Y_i$$

which has approximate variance

$$\hat{V}[\hat{H}(t)] = \sum_{t_i<t} \frac{(d_i/Y_i)(1-d_i/Y_i)}{Y_i} \approx \sum_{t_i<t} \frac{d_i}{Y_i^2}$$

giving an alternate estimate of the survival function

$$\hat{S}_{NA}(t) = \exp[-\hat{H}(t)]$$

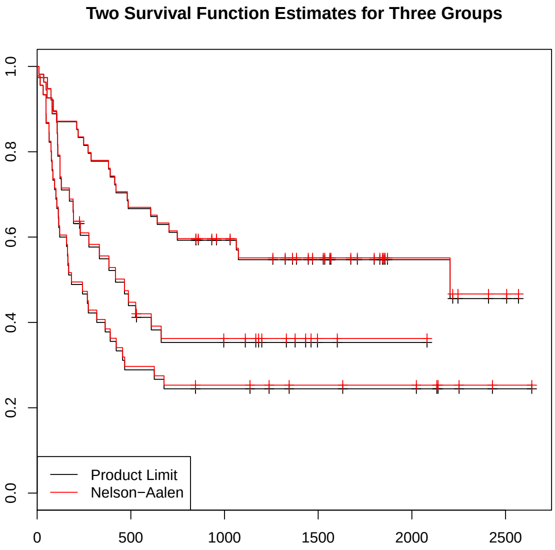
KM and NA Survival Function Estimates

$$
\begin{aligned}
\hat{S}_{KM}(t) &= \prod_{t_i<t}[1-d_i/Y_i] \\
\hat{V}[\hat{S}_{KM}(t)] &= \hat{S}(t)^2 \sum_{t_i<t} \frac{d_i}{Y_i(Y_i-d_i)} \\
\hat{S}_{NA}(t) &= \exp[-\sum_{t_i<t} d_i/Y_i] \\
&= \prod_{t_i<t} \exp(-d_i/Y_i) \\
&\approx \prod_{t_i<t}[1-d_i/Y_i]
\end{aligned}
$$

The product limit estimate and the Nelson-Aalen estimate often do not differ by much. The latter is considered more accurate in small samples and also directly estimates the cumulative hazard. The "fleming-harrington" method reduces to Nelson-Aalen when the data are unweighted. We can also estimate the cumulative hazard as the negative log of the KM survival function estimate.

```
nafit <- survfit(dfsurv~group,type="fleming-harrington",data=bmt)

plot(survfit(dfsurv~group,data=bmt))
lines(nafit,col=2)
legend("bottomleft",c("Product Limit","Nelson-Aalen"),col=1:2,lwd=1)
title("Two Survival Function Estimates for Three Groups")
```


Two Survival Function Estimates for Three Groups
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Product Limit
Nelson−Aalen

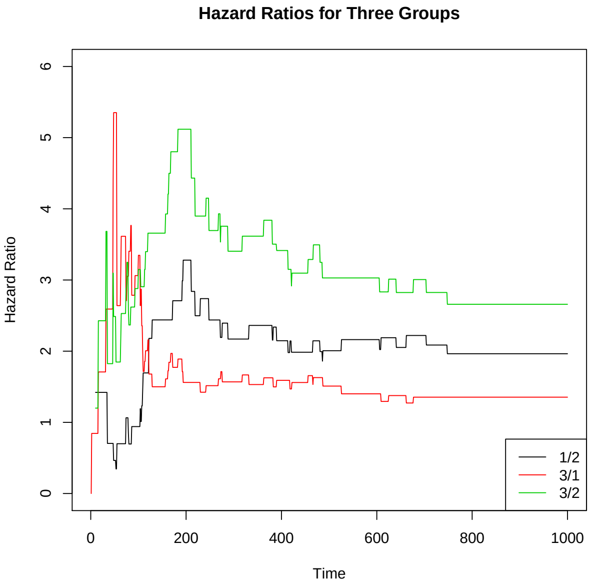
Nelson-Aalen Survival Function Estimate

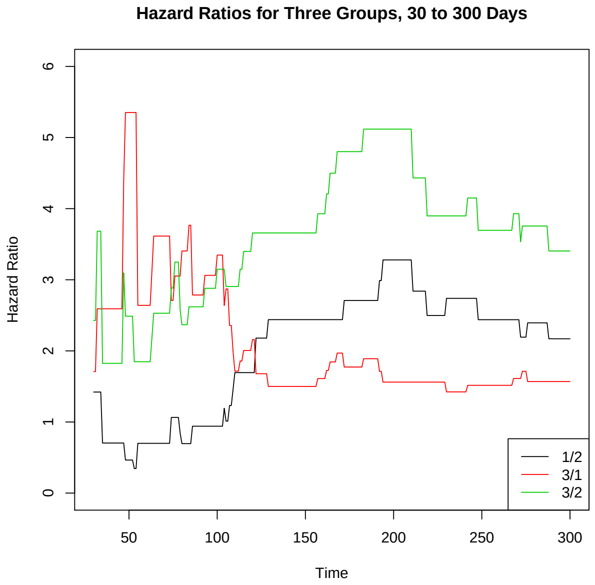
The Nelson-Aalen estimate of the cumulative hazard is usually used for estimates of the hazard and often the cumulative hazard.

If the hazards of the three groups are proportional, that means that the ratio of the hazards is constant over t. We can test this using the ratios of the estimated cumulative hazards, which also would be proportional.

```
nafit <- survfit(dfsurv˜group,type="fleming-harrington",data=bmt)
timevec <- 1:1000
sf1 <- stepfun(nafit[1]$time,c(1,nafit[1]$surv))
sf2 <- stepfun(nafit[2]$time,c(1,nafit[2]$surv))
sf3 <- stepfun(nafit[3]$time,c(1,nafit[3]$surv))
cumhaz1 <- -log(sf1(timevec))
cumhaz2 <- -log(sf2(timevec))
cumhaz3 <- -log(sf3(timevec))

plot(timevec,cumhaz1/cumhaz2,type="l",ylab="Hazard Ratio",xlab="Time",ylim=c(0,6))
lines(timevec,cumhaz3/cumhaz1,ylab="Hazard Ratio",xlab="Time",col=2)
lines(timevec,cumhaz3/cumhaz2,ylab="Hazard Ratio",xlab="Time",col=3)
legend("bottomright",c("1/2","3/1","3/2"),col=1:3,lwd=1)
title("Hazard Ratios for Three Groups")
```


Hazard Ratios for Three Groups
Hazard Ratio
0
1
2
3
4
5
6
Time
0
200
400
600
800
1000
1/2
3/1
3/2


Hazard Ratios for Three Groups, 30 to 300 Days
Hazard Ratio
0
1
2
3
4
5
6
Time
50
100
150
200
250
300
1/2
3/1
3/2

The Nelson-Aalen estimate of the cumulative hazard is usually used for estimates of the hazard. Since the hazard is the derivative of the cumulative hazard, we need a smooth estimate of the cumulative hazard, which is provided by smoothing the step-function cumulative hazard.

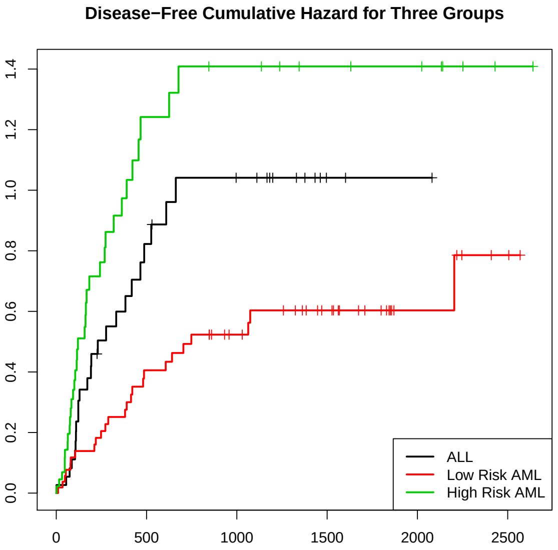
The R package `muhaz` handles this for us. What we are looking for is whether the hazard function is more or less the same shape, increasing, decreasing, constant, etc. Are the hazards “proportional”?

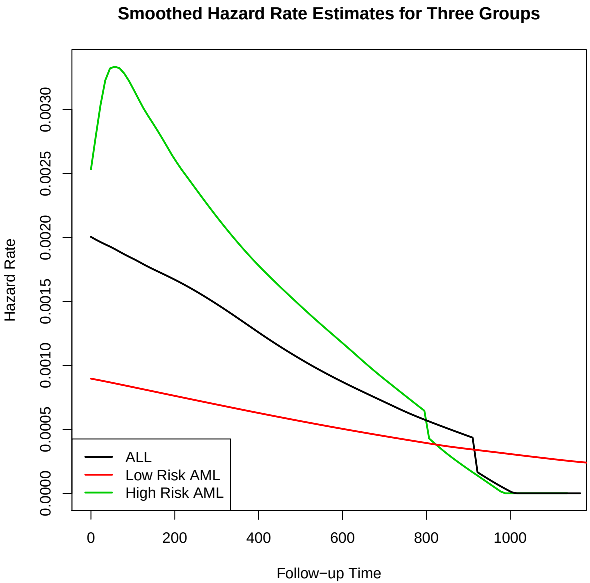
```
> library(muhaz)

> plot(muhaz(bmt$t2,bmt$d3,bmt$group==3),lwd=2,col=3)
> lines(muhaz(bmt$t2,bmt$d3,bmt$group==1),lwd=2,col=1)
> lines(muhaz(bmt$t2,bmt$d3,bmt$group==2),lwd=2,col=2)
> legend("bottomleft",c("ALL","Low Risk AML","High Risk AML"),col=1:3,lwd=2)
> title("Smoothed Hazard Rate Estimates for Three Groups")
```

Group 3 was plotted first because it has the highest hazard. We could also have set the ylim value in plot.

We will see that except for an initial blip in the high risk AML group, the hazards look roughly proportional . They are all strongly decreasing.


Disease−Free Cumulative Hazard for Three Groups
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
500
1000
1500
2000
2500
ALL
Low Risk AML
High Risk AML


Smoothed Hazard Rate Estimates for Three Groups
Hazard Rate
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0
200
400
600
800
1000
Follow−up Time
ALL
Low Risk AML
High Risk AML

Background on the Proportional Hazards Model

The exponential distribution has constant hazard

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t} \\ S(t) &= e^{-\lambda t} \\ h(t) &= \lambda \end{aligned}$$

Let's make two generalizations. First, let the hazard depend on covariates $x_1, x_2, \ldots x_p$. Second, let the base hazard depend on t but not on the covariates.

The Cox Model

The generalization is that the hazard function is

$$\begin{aligned} \eta &= \beta_1 x_1 + \cdots + \beta_p x_p \\ h(t|\text{covariates}) &= h_0(t)e^{\eta} \end{aligned}$$

This has a log link as in a generalized linear model. It is semi-parametric because the linear predictor depends on estimated parameters but the base hazard function is unspecified. There is no constant term because it is absorbed in the base hazard. Note that for two different individuals with possibly different covariates, the ratio of the hazard functions is $\exp(\eta_1)/\exp(\eta_2) = \exp(\eta_1 - \eta_2)$ which does not depend on t.

The Cox Model

How do we fit this model? We need to estimate the coefficients of the covariates, and we need to estimate the base hazard $h_0(t)$. For the covariates, supposing for simplicity that there are no tied event times, let the event times for the whole data set be $t_1, t_2, \ldots, t_D$. Let the risk set at time t_i be $R(t_i)$ and

$$\begin{aligned} \eta_j &= \beta_1 x_{j1} + \cdots + \beta_p x_{jp} \\ \theta_j &= e^{\eta_j} \\ h(t|\text{covariates}) &= h_0(t)e^{\eta} = \theta h_0(t) \end{aligned}$$

The Cox Model

Conditional on a single failure at time t_i, the probability that the event is due to subject $f \in R(t_i)$ is approximately

$$\begin{aligned}\Pr(f \text{ fails}|1 \text{ failure at } t_i) &= \frac{h_0(t_i)e^{\eta_f}}{\sum_{k\in R(t_i)} h_0(t_i)e^{\eta_k}} \\ &= \frac{\theta_f}{\sum_{k\in R(t_i)} \theta_k}\end{aligned}$$

The Cox Model

If subject $f(i)$ is the one who fails at time t_i, then the *partial likelihood* is

$$L(\beta|T) = \prod_i \frac{\theta_{f(i)}}{\sum_{k \in R(t_i)} \theta_k}$$

and we can numerically maximize this with respect to the coefficients β_j. When there are tied event times adjustments need to be made, but the likelihood is still similar. Note that we don't need to know the base hazard to solve for the coefficients.

The Cox Model

If subject $f(i)$ is the one who fails at time t_i, then the *partial likelihood* is

$$L(\beta|T) = \prod_i \frac{\theta_{f(i)}}{\sum_{k \in R(t_i)} \theta_k}$$

From the data, the covariate values x_{ji}, failure times, and the subject who fails are known. We vary the coefficients β_j which determine the

$$\hat{\theta}_k = \hat{\beta}_1 x_{k1} + \cdots + \hat{\beta}_p x_{kp}$$

and that in turn determines the likelihood.