STAT886 A2

Zhiwen Tan

2/9/2022

Question 1

From lecture, we know by greenwood's formula, we have

$$\hat{Var}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{j:a_j<t} \frac{d_j}{n_j(n_j - d_j)}$$

Now, let $Z = \hat{S}(t), g(Z) = log[-log\hat{S}(t)]$, then we have

$$\begin{aligned}
g'(Z) &= \frac{d}{dZ} log[-logZ] \\
&= \frac{-1}{log(Z)} \frac{d}{dZ}[-log(Z)] && \text{(by chain rule)} \\
&= \frac{-1}{log(Z)} \frac{-1}{Z} \\
&= \frac{1}{Zlog(Z)} \\
&= \frac{1}{\hat{S}(t)log(\hat{S}(t))} && \text{(since } Z = \hat{S}(t)\text{)}
\end{aligned}$$

By the Δ-method, we have $Var[g(Z)] \approx [g'(Z)]^2 Var(Z)$, the we have

$$\begin{aligned}
Var[g(Z)] &= [\frac{1}{\hat{S}(t)log(\hat{S}(t))}]^2 * [\hat{S}(t)]^2 \sum_{j:a_j<t} \frac{d_j}{n_j(n_j - d_j)} \\
&= [\frac{1}{log(\hat{S}(t))}]^2 * \sum_{j:a_j<t} \frac{d_j}{n_j(n_j - d_j)} \\
&= [\frac{1}{log(\prod_{j:a_j<t}(1 - \frac{d_j}{n_j}))}]^2 * \sum_{j:a_j<t} \frac{d_j}{n_j(n_j - d_j)} && \text{(KM estimate)} \\
&= [\frac{1}{\sum_{j:a_j<t} log(1 - \frac{d_j}{n_j})}]^2 * \sum_{j:a_j<t} \frac{d_j}{n_j(n_j - d_j)} \\
&= \frac{\sum_{j:a_j<t} \frac{d_j}{n_j(n_j - d_j)}}{[\sum_{j:a_j<t} log(1 - \frac{d_j}{n_j})]^2} \\
&= \frac{\sum_{j:a_j<t} \{d_j/n_j(n_j - d_j)\}}{[\sum_{j:a_j<t} log(1 - d_j/n_j)]^2}
\end{aligned}$$

Now, we have proved

$$\hat{Var}\{log[-log\hat{S}(t)]\} = \frac{\sum_{j:a_j<t}\{d_j/n_j(n_j-d_j)\}}{[\sum_{j:a_j<t} log(1-d_j/n_j)]^2}$$

Now, we have proved

$$\hat{Var}\{log[-log\hat{S}(t)]\} = \frac{\sum_{j:a_j<t}\{d_j/n_j(n_j-d_j)\}}{[\sum_{j:a_j<t} log(1-d_j/n_j)]^2}$$

For confidence interval, we can use transformation method because we already have log(-log) transformation. Then let $\Psi = log[-log\hat{S}(t)]$, so we have $\hat{Var}(\Psi) = \hat{Var}\{log[-log\hat{S}(t)]\}$ Since we already have $\hat{Var}\{log[-log\hat{S}(t)]\}$, so we can just use this. By the property of MLE's, we know $Z = \frac{\hat{\Psi}-\Psi}{\hat{Var}(\Psi)^{\frac{1}{2}}} \approx N(0,1)$. follow the same procedure from lecture, we have a CI for Ψ, $\hat{\Psi}_L \leq \Psi \leq \hat{\Psi}_U$ or $(\hat{\Psi}_L, \hat{\Psi}_U)$.

$$\begin{aligned}
\hat{\Psi}_L &= \hat{\Psi} - Z_{1-\alpha/2}[\hat{Var}(\hat{\Psi})]^{\frac{1}{2}} \\
&= log[-log(\prod_{j:a_j<t}(1-\frac{d_j}{n_j}))] - Z_{1-\alpha/2}[\frac{\sum_{j:a_j<t}\{d_j/n_j(n_j-d_j)\}}{[\sum_{j:a_j<t} log(1-d_j/n_j)]^2}]^{\frac{1}{2}} \\
\hat{\Psi}_U &= \hat{\Psi} + Z_{1-\alpha/2}[\hat{Var}(\hat{\Psi})]^{\frac{1}{2}} \\
&= log[-log(\prod_{j:a_j<t}(1-\frac{d_j}{n_j}))] + Z_{1-\alpha/2}[\frac{\sum_{j:a_j<t}\{d_j/n_j(n_j-d_j)\}}{[\sum_{j:a_j<t} log(1-d_j/n_j)]^2}]^{\frac{1}{2}} \\
P(\hat{\Psi}_L &\leq \Psi \leq \hat{\Psi}_U) = 1-\alpha \\
P(\hat{\Psi}_L &\leq log[-logS(t)] \leq \hat{\Psi}_U) = 1-\alpha
\end{aligned}$$

After this, we can apply inverse function of g() to the former CI interval to get CI for S(t) which is $(\hat{S_L}(t), \hat{S_U}(t))$.

$$\begin{aligned}
P(e^{\hat{\Psi_L}} &\leq -logS(t) \leq e^{\hat{\Psi_U}}) = 1-\alpha \\
P(-e^{\hat{\Psi_L}} &\leq logS(t) \leq -e^{\hat{\Psi_U}}) = 1-\alpha \\
P(e^{-e^{\hat{\Psi_L}}} &\leq S(t) \leq e^{-e^{\hat{\Psi_U}}}) = 1-\alpha
\end{aligned}$$

Question 2

a) Since $logT = \mu + \sigma W$, so $W = \frac{logT-\mu}{\sigma}$. Then we have

$$\begin{aligned} F_T(t) &= P(T \leq t) \\ &= P(e^{\mu+\sigma W} \leq t) \\ &= P(W \leq \frac{logt-\mu}{\sigma}) \\ &= F_W(\frac{logt-\mu}{\sigma}) \\ &= 1 - e^{-e^{\frac{logt-\mu}{\sigma}}} \end{aligned}$$

Now, let $\mu = -\sigma log(\lambda), \sigma = 1/\beta$, then we have $\lambda = e^{-\mu/\sigma}, \beta = 1/\sigma$. plug this back in we have

$$\begin{aligned} F_T(t) &= 1 - e^{-e^{(logt+\frac{1}{\beta}log\lambda)\beta}} \\ &= 1 - e^{-e^{(logt^\beta+log\lambda)}} \\ &= 1 - e^{-e^{(log\lambda t^\beta)}} \\ &= 1 - e^{-\lambda t^\beta} \\ S_T(t) &= 1 - F_T(t) \\ &= e^{-\lambda t^\beta} \qquad (t > 0, \lambda > 0, \beta > 0) \end{aligned}$$

Now we can see this is the survival function of Weibull distribution, so we can conclude T has a Weibull distribution. The relationships are $\mu = -\sigma log(\lambda), \sigma = 1/\beta$.

b) Since $\beta = 1/\sigma$, so we have $\beta = 1/1 = 1$, Plug this back to the survival function for T, Then

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

This matches the survival function for exponential distribution, so we know T has an exponential distribution when $\sigma = 1$

Question 3

a) because $S_Y(y) = e^{-e^{\frac{y-\mu}{\sigma}}}$, so we have

$$f(y) = -\frac{d}{dy} S_Y(y)$$
$$= \frac{1}{\sigma} e^{\frac{y-\mu}{\sigma}} \cdot e^{-e^{\frac{y-\mu}{\sigma}}}$$

Since $y_i = \log x_i$, and the data consist of n pair of $(y_1, \delta_1) \cdots (y_n, \delta_n)$

Then we have

$$L(\mu, \sigma) = \prod_{i=1}^{n} [f(y_i)]^{\delta_i} [S(y_i)]^{1-\delta_i}$$
$$= \prod_{i=1}^{n} \left[\frac{1}{\sigma} e^{\frac{y_i-\mu}{\sigma}} \cdot e^{-e^{\frac{y_i-\mu}{\sigma}}}\right]^{\delta_i} \left[e^{-e^{\frac{y_i-\mu}{\sigma}}}\right]^{1-\delta_i}$$
$$= \frac{1}{\sigma^{\sum_{i=1}^{n}\delta_i}} \cdot e^{\sum_{i=1}^{n} \frac{y_i-\mu}{\sigma}\delta_i} \cdot e^{-\sum_{i=1}^{n} e^{\frac{y_i-\mu}{\sigma}}}$$

Since $\phi = \log \sigma$, therefore, $\sigma = e^{\phi}$. Then we have

$$L(\mu, \phi) = \frac{1}{e^{\phi \sum_{i=1}^{n}\delta_i}} \cdot e^{\sum_{i=1}^{n} \frac{(y_i-\mu)}{e^{\phi}}\delta_i} \cdot e^{-\sum_{i=1}^{n} e^{\frac{(y_i-\mu)}{e^{\phi}}}}$$

b) From $L(\mu, \phi)$, we can get $\ell(\mu, \phi)$.

$$\ell(\mu, \phi) = \log[L(\mu, \phi)]$$
$$= -\phi \sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n} \frac{y_i-\mu}{e^{\phi}}\delta_i - \sum_{i=1}^{n} e^{\frac{y_i-\mu}{e^{\phi}}}$$

now, calculate score function for μ and ϕ

Score function for μ

$$\frac{d\ell}{d\mu} = 0 - \sum_{i=1}^{n} \delta_i \cdot \frac{1}{e^{\phi}} + \sum_{i=1}^{n} e^{(y_i-\mu)e^{-\phi}} \cdot \frac{1}{e^{\phi}}$$
$$= \sum_{i=1}^{n} e^{(y_i-\mu)e^{-\phi}} e^{-\phi} - \sum_{i=1}^{n} \delta_i e^{-\phi}$$

Score function for ϕ

$$\frac{d\ell}{d\phi} = -\sum_{i=1}^{n}\delta_i - \sum_{i=1}^{n}(y_i-\mu)e^{-\phi}\delta_i + \sum_{i=1}^{n}(y_i-\mu)e^{-\phi} e^{(y_i-\mu)e^{-\phi}}$$
$$= -\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}(\mu-y_i)e^{-\phi}\delta_i + \sum_{i=1}^{n}(y_i-\mu)e^{-\phi} e^{(y_i-\mu)e^{-\phi}}$$

Then we have $U(\mu, \phi) = \begin{bmatrix} \frac{d\ell}{d\mu} \\ \frac{d\ell}{d\phi} \end{bmatrix}$

C) Let $\frac{\partial \ell}{\partial \mu} = 0$, then we have

$$0 = \sum_{i=1}^{n} e^{(y_i-\mu)e^{-\phi}} e^{-\phi} - \sum_{i=1}^{n} \delta_i e^{-\phi}$$

$$\sum_{i=1}^{n} \delta_i e^{-\phi} = \sum_{i=1}^{n} e^{(y_i-\mu)e^{-\phi}} e^{-\phi}$$

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} e^{(y_i-\mu)e^{-\phi}}$$

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} e^{\hat{w}_i}$$

$$\hat{\alpha} = \sum_{i=1}^{n} e^{\hat{w}_i}$$

Let $\frac{\partial \ell}{\partial \phi} = 0$, then we have

$$0 = -\sum_{i=1}^{n} \delta_i + \sum_{i=1}^{n} (\mu - y_i) e^{-\phi} \delta_i + \sum_{i=1}^{n} (y_i - \mu) e^{-\phi} e^{(y_i-\mu)e^{-\phi}}$$

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} (\mu - y_i) e^{-\phi} \delta_i + \sum_{i=1}^{n} (y_i - \mu) e^{-\phi} e^{(y_i-\mu)e^{-\phi}}$$

$$\sum_{i=1}^{n} \delta_i = -\sum_{i=1}^{n} \hat{w}_i \delta_i + \sum_{i=1}^{n} \hat{w}_i e^{\hat{w}_i}$$

$$\hat{\alpha} = -\sum_{i=1}^{n} \hat{w}_i \delta_i + \sum_{i=1}^{n} \hat{w}_i e^{\hat{w}_i}$$

$$I(\hat{\mu}, \hat{\phi}) = \begin{bmatrix} -\frac{\partial^2 \ell}{\partial \hat{\mu}^2} & -\frac{\partial^2 \ell}{\partial \hat{\mu} \partial \hat{\phi}} \\ -\frac{\partial^2 \ell}{\partial \hat{\phi} \partial \hat{\mu}} & -\frac{\partial^2 \ell}{\partial \hat{\phi}} \end{bmatrix}$$

$$\frac{\partial^2 \ell}{\partial \mu^2} = \sum_{i=1}^{n} -e^{-\phi} \cdot e^{-\phi} \cdot e^{(y_i-\mu)e^{-\phi}}$$

$$\frac{\partial^2 \ell}{\partial \hat{\mu}^2} = \sum_{i=1}^{n} -e^{-2\hat{\phi}} \cdot e^{\hat{w}_i}$$

$$= -\sum_{i=1}^{n} e^{-2\hat{\phi}} \cdot e^{\hat{w}_i}$$

$$\frac{\partial^2 \ell}{\partial \mu \partial \phi} = \sum_{i=1}^{n} \delta_i e^{-\phi} + \sum_{i=1}^{n} \left(-e^{-\phi} e^{(y_i-\mu)e^{-\phi}} - (y_i-\mu)e^{-\phi} e^{(y_i-\mu)e^{-\phi}} \cdot e^{-\phi}\right)$$

$$\frac{\partial^2 \ell}{\partial \hat{\mu} \partial \hat{\phi}} = \frac{d}{e^{\hat{\phi}}} + \sum_{i=1}^{n} \left(-e^{-\hat{\phi}} e^{\hat{w}_i} - \hat{w}_i e^{\hat{w}_i} \cdot e^{-\hat{\phi}}\right)$$

$$= \sum_{i=1}^{n} -e^{-\hat{\phi}} \hat{w}_i e^{\hat{w}_i}$$

$$= -\sum_{i=1}^{n} e^{-\hat{\phi}} \hat{w}_i e^{\hat{w}_i}$$

$$\frac{\partial^2 \ell}{\partial \phi^2} = \sum_{i=1}^{n} (y_i-\mu)e^{-\phi} \delta_i - \sum_{i=1}^{n} (y_i-\mu)e^{-\phi} e^{(y_i-\mu)e^{-\phi}} - \sum_{i=1}^{n} (y_i-\mu)^2 e^{-2\phi} e^{(y_i-\mu)e^{-\phi}}$$

$$\frac{\partial^2 \ell}{\partial \hat{\phi}^2} = \sum_{i=1}^{n} \hat{w}_i \delta_i - \sum_{i=1}^{n} \hat{w}_i e^{\hat{w}_i} - \sum_{i=1}^{n} \hat{w}_i^2 e^{\hat{w}_i}$$

$$= -d - \sum_{i=1}^{n} \hat{w}_i^2 e^{\hat{w}_i}$$

Then we can plug those numbers in $I(\mu, \phi)$

$$\underline{I}(\hat{\mu}, \hat{\phi}) = \begin{bmatrix} \sum_{i=1}^{n} e^{-2\hat{\phi}} e^{\hat{w}_i} & \sum_{i=1}^{n} e^{-\hat{\phi}} \hat{w}_i e^{\hat{w}_i} \\ \sum_{i=1}^{n} e^{-\hat{\phi}} \hat{w}_i e^{\hat{w}_i} & d + \sum_{i=1}^{n} \hat{w}_i^2 e^{\hat{w}_i} \end{bmatrix}$$

Since $\hat{Var}\begin{pmatrix} \hat{\mu} \\ \hat{\phi} \end{pmatrix} = I^{-1}(\hat{\mu}, \hat{\phi})$, then we have

$$\hat{Var}(\hat{\mu}) = [I^{-1}(\hat{\mu}, \hat{\phi})]_{1,1}$$

$$\hat{Cov}(\hat{\mu}, \hat{\phi}) = [I^{-1}(\hat{\mu}, \hat{\phi})]_{1,2} = [I^{-1}(\hat{\mu}, \hat{\phi})]_{2,1}$$

$$\hat{Var}(\hat{\phi}) = [I^{-1}(\hat{\mu}, \hat{\phi})]_{2,2}$$

Then we can use those to estimate the variance matrix

$$\hat{Var}\begin{pmatrix} \hat{\mu} \\ \hat{\phi} \end{pmatrix} = \begin{bmatrix} [I^{-1}(\hat{\mu}, \hat{\phi})]_{1,1} & [I^{-1}(\hat{\mu}, \hat{\phi})]_{1,2} \\ [I^{-1}(\hat{\mu}, \hat{\phi})]_{2,1} & [I^{-1}(\hat{\mu}, \hat{\phi})]_{2,2} \end{bmatrix}$$

Question 4

a) Since we have an exponential distribution with hazard rate λ, Then

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

. From here, we have

$$\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} [f(t)]^{\delta_i} [S(t)]^{1-\delta_i} \\
&= \prod_{i=1}^{n} [\lambda e^{-\lambda x_i}]^{\delta_i} [e^{-\lambda x_i}]^{1-\delta_i} && \text{(let } x_i \text{ be t at certain time point)} \\
&= \prod_{i=1}^{n} [\lambda]^{\delta_i} e^{-\lambda x_i} \\
&= [\lambda]^{\sum_{i=1}^{n} \delta_i} e^{-\lambda \sum_{i=1}^{n} x_i} \\
\ell(\lambda) &= log([\lambda]^{\sum_{i=1}^{n} \delta_i} e^{-\lambda \sum_{i=1}^{n} x_i}) && \text{(log likelihood)} \\
&= \sum_{i=1}^{n} \delta_i log\lambda - \lambda \sum_{i=1}^{n} x_i
\end{aligned}$$

Now, we have the log-likelihood function for λ, and next step is to get the mle

$$\begin{aligned}
\ell'(\lambda) &= \frac{d}{d\lambda}(\sum_{i=1}^{n} \delta_i log\lambda - \lambda \sum_{i=1}^{n} x_i) \\
&= \sum_{i=1}^{n} \delta_i \frac{1}{\lambda} - \sum_{i=1}^{n} x_i \\
0 &= \sum_{i=1}^{n} \delta_i \frac{1}{\lambda} - \sum_{i=1}^{n} x_i && \text{(score equation)} \\
\hat{\lambda} &= \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i} \\
&= \frac{6}{6+4+6+3+9+9+10+13+11} && \text{(take in values)} \\
&= 0.084
\end{aligned}$$

Now, for mean survival time

$$\begin{aligned}
\theta &= E(T) = \frac{1}{\lambda} \\
\hat{\theta} &= \frac{1}{\hat{\lambda}} && \text{(invariance property)} \\
&= \frac{1}{\frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i}} \\
&= \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} \delta_i} \\
&= 1/0.084 \\
&= 11.9
\end{aligned}$$

Now we have the mle for both hazard rate and mean survival time.

b) We can just use the KM estimate formula and greenwood's formula to get $\hat{S}(t)$, and $Var\hat{S}(t)$

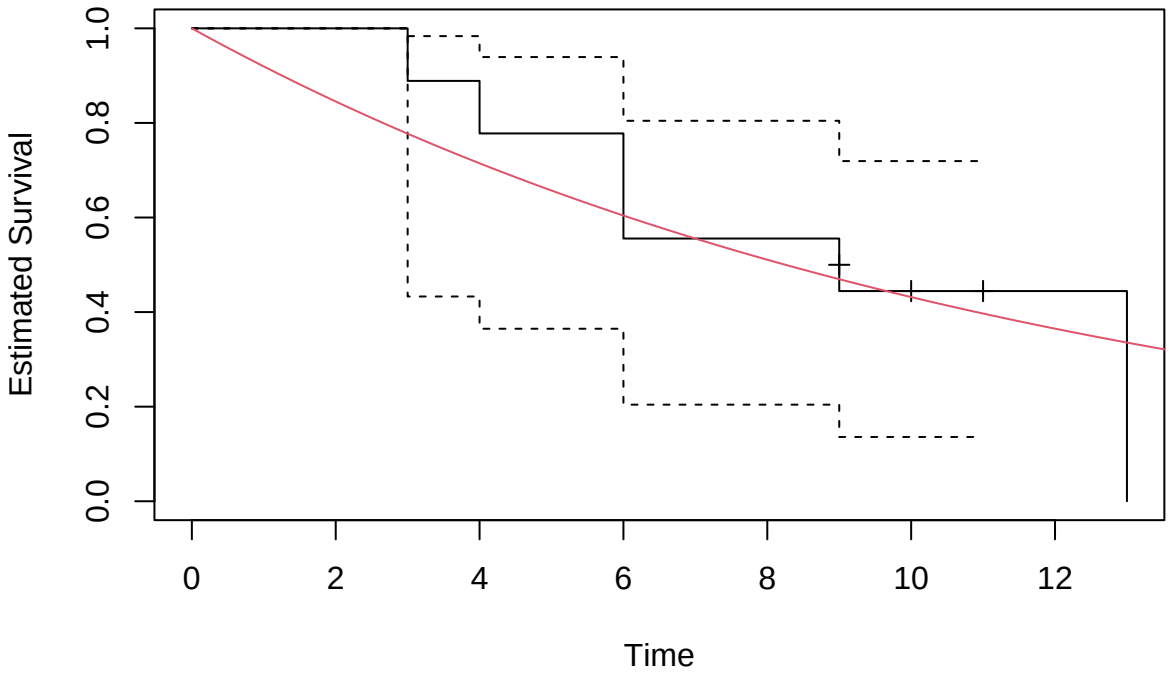
j	a_j	n_j	d_j	$\hat{S}(t)$	$Var\hat{S}(t)$
1	3	9	1	$1-\frac{1}{9}=\frac{8}{9}$	$\frac{8}{9}^2\frac{1}{9(9-1)}=0.011$
2	4	8	1	$\frac{8}{9}(1-\frac{1}{8})=\frac{7}{9}$	$\frac{7}{9}^2[\frac{1}{9(9-1)}+\frac{1}{8(8-1)}]=0.019$
3	6	7	2	$\frac{7}{9}(1-\frac{2}{7})=\frac{5}{9}$	$\frac{5}{9}^2[\frac{1}{9(9-1)}+\frac{1}{8(8-1)}+\frac{2}{7(7-2)}]=0.027$
4	9	5	1	$\frac{5}{9}(1-\frac{1}{5})=\frac{4}{9}$	$\frac{4}{9}^2[\frac{1}{9(9-1)}+\frac{1}{8(8-1)}+\frac{2}{7(7-2)}+\frac{1}{5(5-1)}]=0.027$
5	13	1	1	$\frac{4}{9}(1-1)=0$	$0^2[\frac{1}{9(9-1)}+\frac{1}{8(8-1)}+\frac{2}{7(7-2)}+\frac{1}{5(5-1)}+\frac{1}{1-1}]=0$

c) we know $\hat{\mu}$ = area under $\hat{S}(t)$, so we can use the data from part b to calculate the mean

$$\begin{aligned}\hat{\mu} &= a_1+\hat{S}(a_1t)(a_2-a_1)+...+\hat{S}(a_{k-1}t)(a_k-a_{k-1})\\ &= 3+0.9*1+0.8*2+0.6*3+0.5*4+0\\ &= 8.7\end{aligned}$$

we can see the result are not exactly the same, this may because we are assuming the survival time has exponetial distribution, but we don't make any assumption here. Also, we can see most of the censoring happends to the end of the study, so this will cause the mean estimation in the mle method be greater. in addition, the mle method will calculate integral of S(t) from 0 to ∞, this can also the cause a greater estimation. Overall, this nonparametric estimate seems agree with the corresponding parametric estimate.

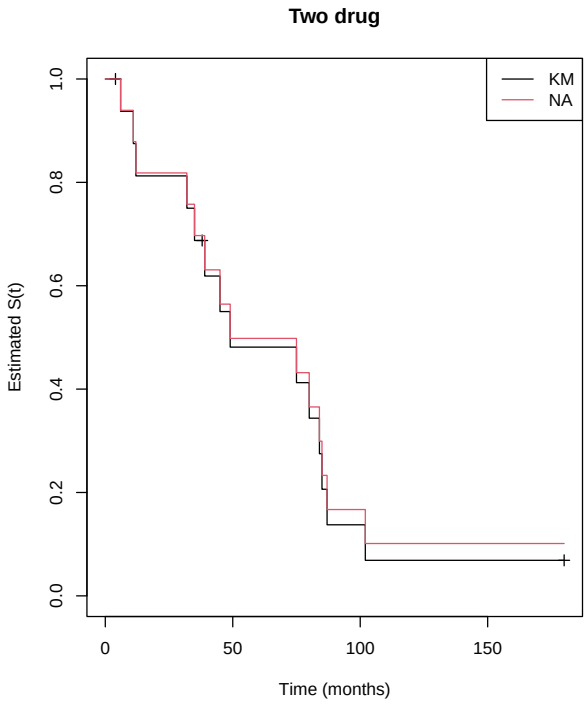
d) To plot the KM curve, we need to first set one vector for time and another vactor for failure status. then we can use survfit function to fit a KM curve on the two vectors, then we can plot it out. In here, I used the log-log parameter for CI, this ensure that CI will not exceed 1.

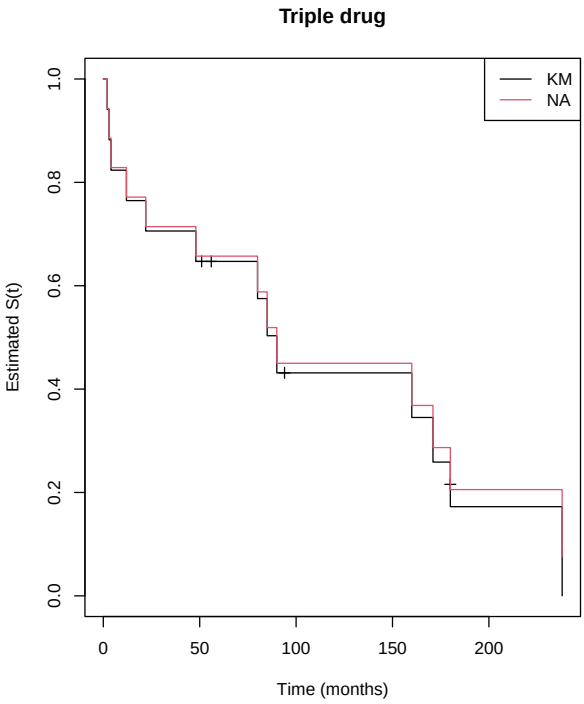


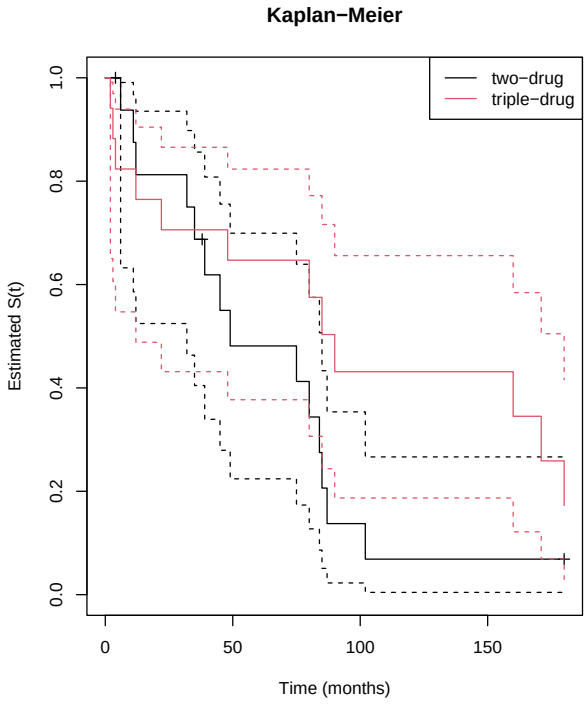
The exponential model in a) fits the data pretty well, we can see the exponential curve is in between the point-wise CI from 3 to 13, therefore, between this interval, the exponential model can give a good estimation.

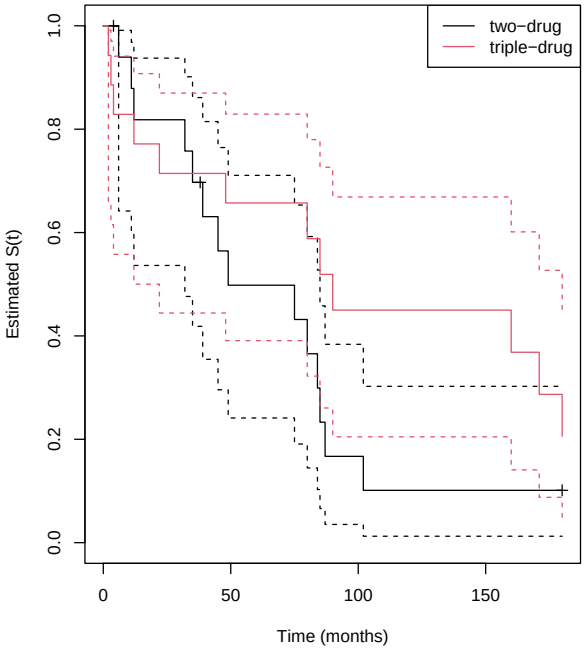
Question 5

a) we will first split the hiv data by two treatment, then use survfit function to fit both KM and MA estimate for thew two groups. After this. we ca plot them in a single plot to see the difference. because we are compare the survival function, not the cumulative hazard, so I used the default parameter for ctype in survfit function. confidence interval was setted to log-log base









From the top two plots, we can see both Kaplan-Meier estimate and the Nelson-Aalen estimate gives similar results, the overall pattern are very similar. in general, the NA estimate is higher than the KM estimate. For the two plot below, we can clearly see the overall survival function has a big difference between two drug group and triple drug group. However, when we look at the confidence interval, we can see the confidence interval is overlapping at any time point between two group, therefore, there is a chance that the two method may have similar effect. Thus, we can not simply conclude the triple drug group has higher survival probability than the two drug group. we can just say the participants in triple drug group may performs batter than the two drug group.

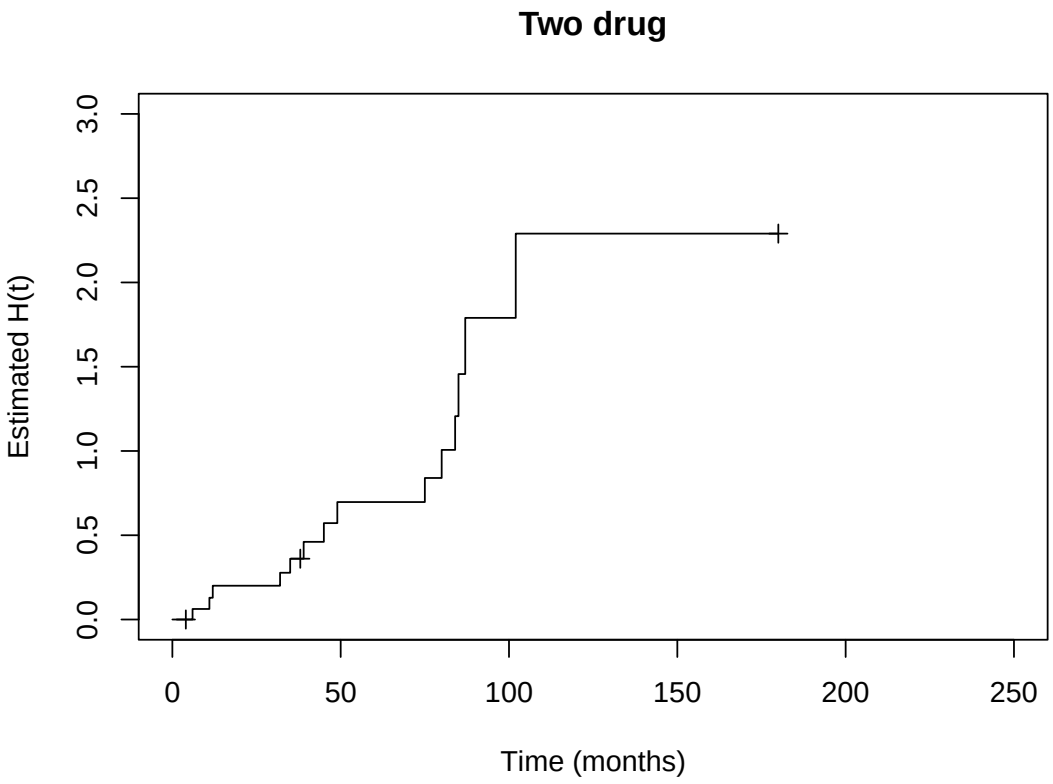
b) we can use summary function to check the median for each group, then we can use survfit function to git model with conf.type set to log-log and plain two get the two confidence interval at time 60. since we already used log-log confidence interval in part b) so we can just use summary function on the model we have in part b).

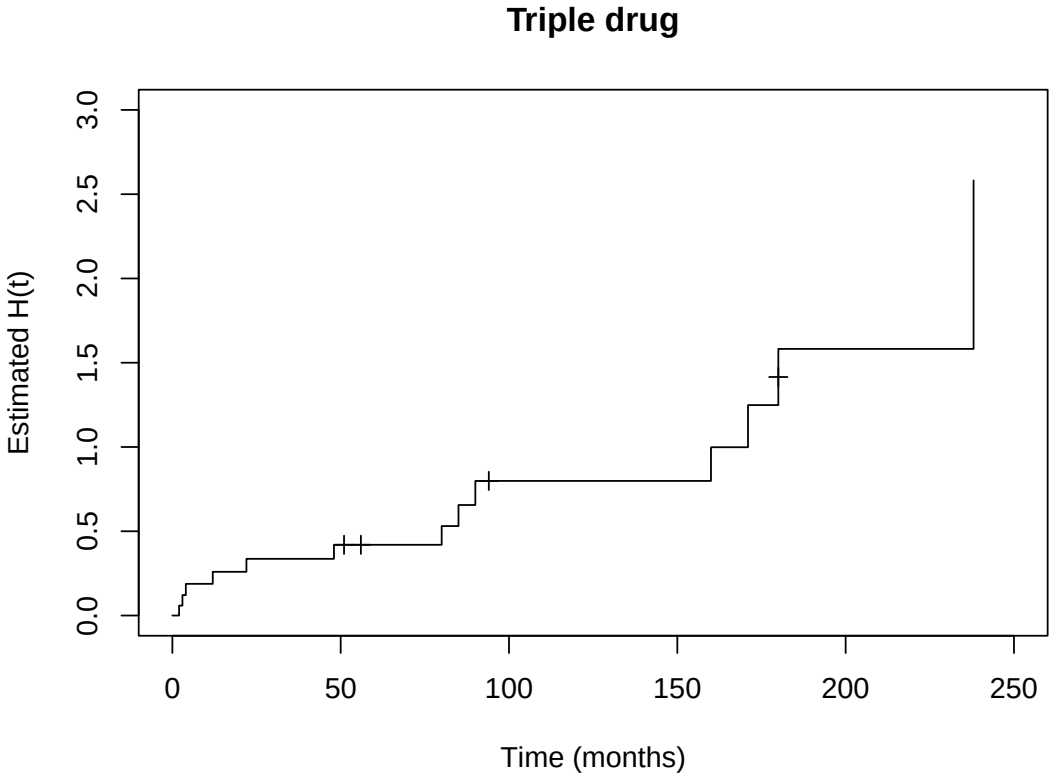
```
## Call: survfit(formula = Surv(x1, delta1) ~ 1, stype = 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     16       1   0.9375  0.0605      0.63235        0.991
##    11     15       1   0.8750  0.0827      0.58598        0.967
##    12     14       1   0.8125  0.0976      0.52460        0.935
##    32     13       1   0.7500  0.1083      0.46343        0.898
##    35     12       1   0.6875  0.1159      0.40460        0.856
##    39     10       1   0.6188  0.1230      0.33929        0.808
##    45      9       1   0.5500  0.1271      0.27933        0.756
##    49      8       1   0.4813  0.1285      0.22410        0.699
##    75      7       1   0.4125  0.1272      0.17339        0.639
##    80      6       1   0.3438  0.1232      0.12728        0.575
##    84      5       1   0.2750  0.1162      0.08617        0.507
##    85      4       1   0.2063  0.1055      0.05082        0.433
##    87      3       1   0.1375  0.0900      0.02265        0.354
##   102      2       1   0.0688  0.0662      0.00443        0.267

## Call: survfit(formula = Surv(x1, delta1) ~ 1, stype = 1, conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     16       1   0.9375  0.0605       0.8189        1.000
##    11     15       1   0.8750  0.0827       0.7130        1.000
##    12     14       1   0.8125  0.0976       0.6213        1.000
##    32     13       1   0.7500  0.1083       0.5378        0.962
##    35     12       1   0.6875  0.1159       0.4604        0.915
##    39     10       1   0.6188  0.1230       0.3777        0.860
##    45      9       1   0.5500  0.1271       0.3009        0.799
##    49      8       1   0.4813  0.1285       0.2294        0.733
##    75      7       1   0.4125  0.1272       0.1632        0.662
##    80      6       1   0.3438  0.1232       0.1023        0.585
##    84      5       1   0.2750  0.1162       0.0473        0.503
##    85      4       1   0.2063  0.1055       0.0000        0.413
##    87      3       1   0.1375  0.0900       0.0000        0.314
##   102      2       1   0.0688  0.0662       0.0000        0.199

## Call: survfit(formula = Surv(x2, delta2) ~ 1, stype = 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
```

```
##     2     17       1    0.941  0.0571       0.6502        0.991
##     3     16       1    0.882  0.0781       0.6060        0.969
##     4     15       1    0.824  0.0925       0.5471        0.939
##    12     14       1    0.765  0.1029       0.4883        0.904
##    22     13       1    0.706  0.1105       0.4315        0.866
##    48     12       1    0.647  0.1159       0.3771        0.823
##    80      9       1    0.575  0.1233       0.3065        0.772
##    85      8       1    0.503  0.1272       0.2436        0.716
##    90      7       1    0.431  0.1277       0.1870        0.656
##   160      5       1    0.345  0.1280       0.1216        0.584
##   171      4       1    0.259  0.1217       0.0691        0.505
##   180      3       1    0.173  0.1074       0.0296        0.416
##   238      1       1    0.000     NaN           NA           NA

## Call: survfit(formula = Surv(x2, delta2) ~ 1, stype = 1, conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2     17       1    0.941  0.0571       0.8293        1.000
##     3     16       1    0.882  0.0781       0.7292        1.000
##     4     15       1    0.824  0.0925       0.6423        1.000
##    12     14       1    0.765  0.1029       0.5631        0.966
##    22     13       1    0.706  0.1105       0.4893        0.922
##    48     12       1    0.647  0.1159       0.4199        0.874
##    80      9       1    0.575  0.1233       0.3335        0.817
##    85      8       1    0.503  0.1272       0.2541        0.752
##    90      7       1    0.431  0.1277       0.1811        0.682
##   160      5       1    0.345  0.1280       0.0942        0.596
##   171      4       1    0.259  0.1217       0.0204        0.497
##   180      3       1    0.173  0.1074       0.0000        0.383
##   238      1       1    0.000     NaN          NaN          NaN
```

From the first summary, we can see the estimate median is between T=45 and T=49, so the approximated median T would be 49 because the survival rate at T=49 is the greatest number that smaller than 0.5. For two drug group, we can see from the plot in b) that at time 60, the survival probability and confidence interval should be the same as T=49, so we have approximate plain CI: (0.2294, 0.733), and log-log CI:(0.22410, 0.699). For triple drug group, T = 60 is in between T=48 and T=80, so simile as above, we have CI at T=60 = CI at T=60, therefore, we have approximate plain CI: (0.4199, 0.874) and log-log CI: (0.3771, 0.823)

c) To plot cumulative hazard function, we just need to specify the parameter cumhaz=T in plot function. because the default ctype in survfit function is NA estimate, so we don't need to change the code too much.


Two drug
Estimated H(t)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
50
100
150
200
250
Time (months)

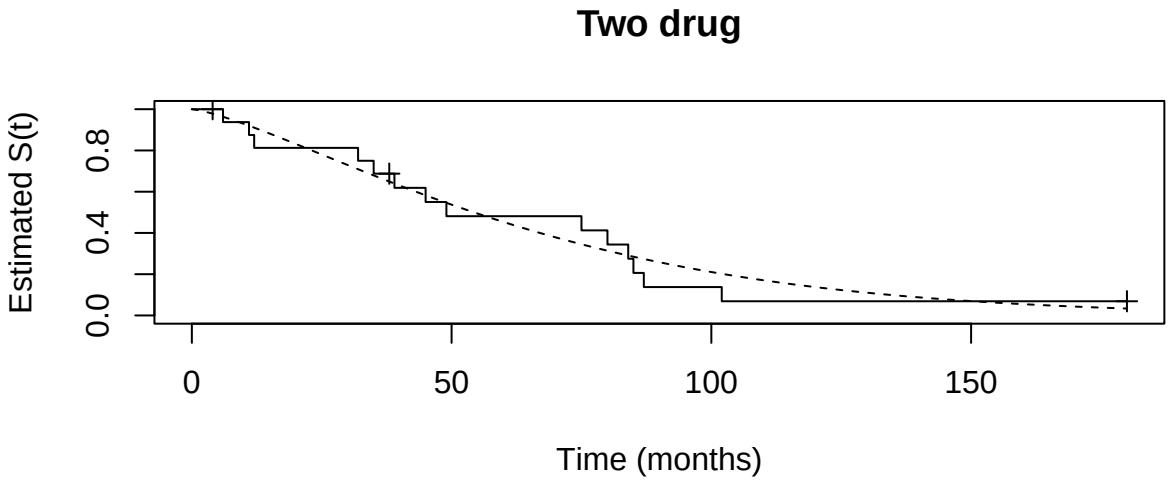

Triple drug
Estimated H(t)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
50
100
150
200
250
Time (months)

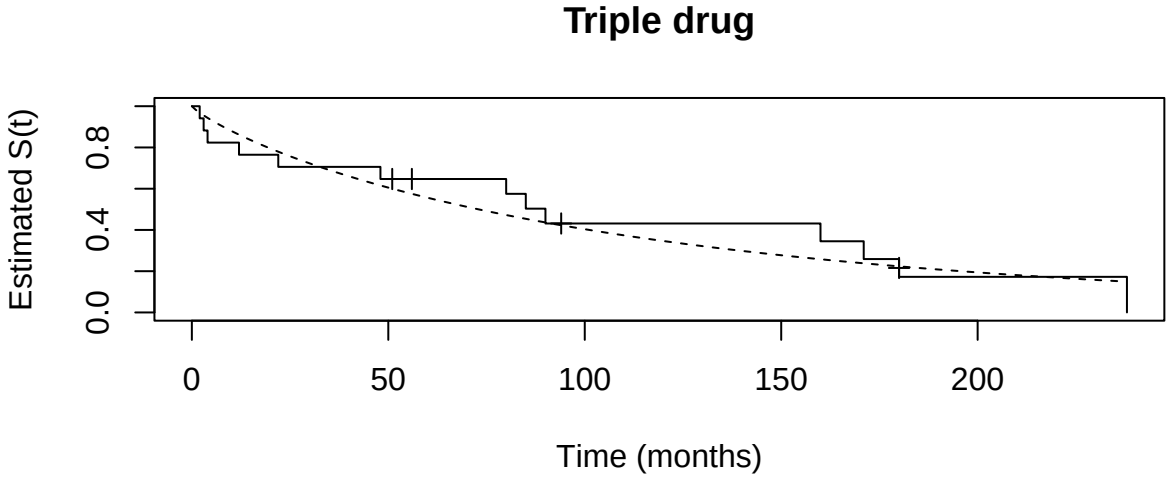
For two-drug group, we can see the the H(t) stop increases at the end period, this is because the last data point in the file is censored. if we ignore the last part, then the H(t) seems have curve shape. the slop increase as time increases.

For triple drug group, because the last person is a failure, so the data has an increasing at the end. The interest thing about this plot is the slope of this plot decreases as the time increases at the beginning, and then the slope increases as the time increases.

In both plot, when there is censoring happens, the slope is decreasing in both graph. Also, it means no failure happens in this period of time. this is because the time between two time points increases and stretch the curve to the right.

Weibull distribution would be reasonable to model T, because form the H(t) plot, we can see there is a increase trend happens in both group and the increasing rate is getting bigger as time pass, this fits the h(t) function for weibull distribution. ALso, we can see the survival plot form b) has similar shape as a weibull distribution. So Weibull distribution is reasonable to model T. To check this, I will plot the Weibull distribution survival curve in to the KM curve





We can see the Weibull distribution gives similar estimate as KM curve, so we can conclude that Weibull distribution is reasonable to model T.

Appendix

Question 4

KM plots

```
library(survival)
times <- c(6,4,6,3,9,9,10,13,11)
delta <- c(1,1,1,1,0,1,0,1,0)
eq <- function(x) exp(-0.084*x)
fit <- survfit(Surv(times, delta) ~ 1, conf.type = "log-log")
plot(fit, xlab = "Time", ylab = "Estimated Survival", mark.time = T, conf.int = T)
curve(eq, from = 0, to = 13, col=2, add = T)
```

Question 5

part a)

```
df <- read.table("hiv.txt", header = T)
g1 <- df[which(df$grp == 1),]
g2 <- df[which(df$grp == 2),]
x1 <- g1$time; x2 <- g2$time
delta1 <- g1$status; delta2 <- g2$status
fit_km1 <- survfit(Surv(x1, delta1) ~ 1, stype = 1)
fit_na1 <- survfit(Surv(x1, delta1) ~ 1, stype = 2)
fit_km2 <- survfit(Surv(x2, delta2) ~ 1, stype = 1)
fit_na2 <- survfit(Surv(x2, delta2) ~ 1, stype = 2)
plot(fit_km1, xlab="Time (months)", ylab="Estimated S(t)", mark.time=T,conf.int=F)
lines(fit_na1, conf.int=F,lty=2)
lines(fit_km2, conf.int=F, col = 2)
lines(fit_na2, conf.int=F,lty=2, col=2)
legend("topright",0.95,c("Kaplan-Meier", "Nelson-Aalen", "two-drug", "triple-drug"),
       lty=c(1,2,1,1), col = c(1,1,1,2))
```

part b)

```
fit_km3 <- survfit(Surv(x1, delta1) ~ 1, stype = 1, conf.type = "plain")
fit_km4 <- survfit(Surv(x2, delta2) ~ 1, stype = 1, conf.type = "plain")
summary(fit_km1)
summary(fit_km3)
summary(fit_km2)
summary(fit_km4)
```

part c)

```
par(mfrow = c(1,2))
plot(fit_na1, xlab="Time (months)", ylab="Estimated H(t)",
     mark.time=T,conf.int=F, main = "Two drug", cumhaz = T,
     xlim = c(0,250), ylim = c(0,3))
plot(fit_na2, xlab="Time (months)", ylab="Estimated H(t)",
     mark.time=T,conf.int=F, main = "Triple drug", cumhaz = T,
     xlim = c(0,250), ylim = c(0,3))
```

part c) check

```
fit.weib1 <- survreg(Surv(x1, delta1) ~ 1)
sigma1 <- fit.weib1$scale
mu1 <- fit.weib1$coefficients
t1 <- 0:180
sf.weib1 <- exp(-exp((log(t1)- mu1)/sigma1))

fit.weib2 <- survreg(Surv(x2, delta2) ~ 1)
sigma2 <- fit.weib2$scale
mu2 <- fit.weib2$coefficients
t2 <- 0:238
sf.weib2 <- exp(-exp((log(t2)- mu2)/sigma2))

par(mfrow = c(2,1))

plot(fit_km1, xlab="Time (months)", ylab="Estimated S(t)",
     mark.time=T,conf.int=F, main = "Two drug")
lines(t1,sf.weib1, lty = 2)

plot(fit_km2, xlab="Time (months)", ylab="Estimated S(t)",
     mark.time=T,conf.int=F, main = "Triple drug")
lines(t2,sf.weib2, lty = 2)
```