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

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

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. Similar to lecture, 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 $(\hat{\Psi}_L, \hat{\Psi}_U)$. After this, we can apply inverse function of g() to the former CI interval to get log[-log] based CI for S(t) which is $(\hat{S_L}(t), \hat{S_U}(t))$.

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

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

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

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

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 $Var\hat{S}(t)$

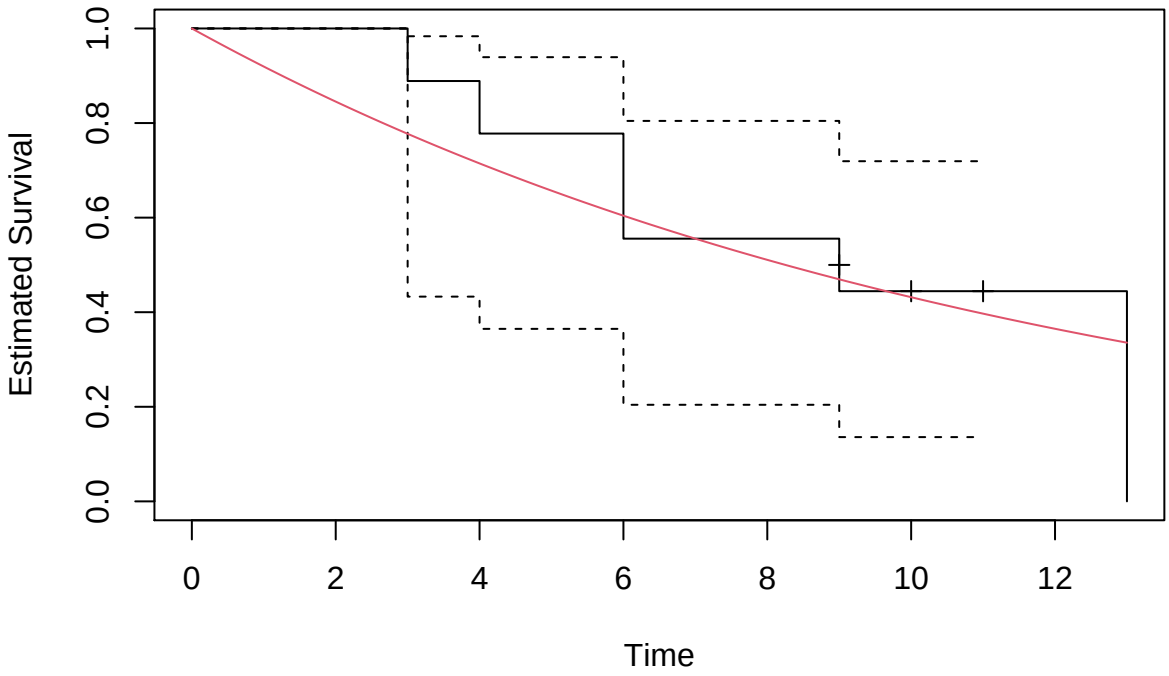
j	a_j	n_j	d_j	$\hat{S}(t)$	$Var\hat{S}(t)$
1	3	10	1	$1-\frac{1}{10}=0.9$	$0.9^2\frac{1}{10(10-1)}=0.009$
2	4	9	1	$0.9(1-\frac{1}{9})=0.8$	$0.8^2[\frac{1}{10(10-1)}+\frac{1}{9(9-1)}]=0.016$
3	6	8	2	$0.8(1-\frac{2}{8})=0.6$	$0.6^2[\frac{1}{10(10-1)}+\frac{1}{9(9-1)}+\frac{2}{8(8-2)}]=0.024$
4	9	6	1	$0.6(1-\frac{1}{6})=0.5$	$0.5^2[\frac{1}{10(10-1)}+\frac{1}{9(9-1)}+\frac{2}{8(8-2)}+\frac{1}{6(6-1)}]=0.025$
5	13	1	1	$0.5(1-1)=0$	$0^2[\frac{1}{10(10-1)}+\frac{1}{9(9-1)}+\frac{2}{8(8-2)}+\frac{1}{6(6-1)}+\frac{1}{1-1}]=0$

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

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

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

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)
```