可靠性数据与生存分析作业

计 91 刘程华 20180116897
清华大学计算机系

日期：2021 年 6 月 10 日

1 第二次作业

1.1 Work with built-in ovarian data set from the survival library. This dataset comprises a cohort of ovarian cancer patients and respective clinical information, including the time patients were tracked until they either died or were lost to follow-up (futime), whether patients were censored or not (fustat).

(a) Plot the KM curve $\hat{S}_{KM}$ (i.e., non-parametric estimate) of this censored survival data

(b) Fit a Weibull model to this censored survival data and add the Weibull estimate to the curve you made in (a).

(c) Perform Wald test to test whether or not the survival data are from exponential distribution.

(d) Suggest ways to check the Weibull model assumptions and conduct the diagnostic.

(a)and(b) figure 1

```
library(survival)
data(ovarian)
ovarian_surv <- with(ovarian, Surv(futime, fustat))
t.vals <- 1:max(ovarian$futime)
# Kaplan-Meier
fit.KM <- survfit(ovarian_surv ~ 1)
# Weibull
fit.Wei <- survreg(ovarian_surv ~ 1, dist = "weibull")
mu.hat <- fit.Wei$coefficients
sigma.hat <- fit.Wei$scale
lambda.hat <- exp(-mu.hat)
alpha.hat <- 1 / sigma.hat
prob.hat.Wei <- 1 - pweibull(t.vals, shape = alpha.hat, scale = 1/lambda.hat)
# Plot
plot(fit.KM, conf.int = F, xlab = "patient time", ylab = "survival probability"
    )
lines(prob.hat.Wei ~ t.vals, col="blue")
```

(c)

```
summary(fit.Wei)
```

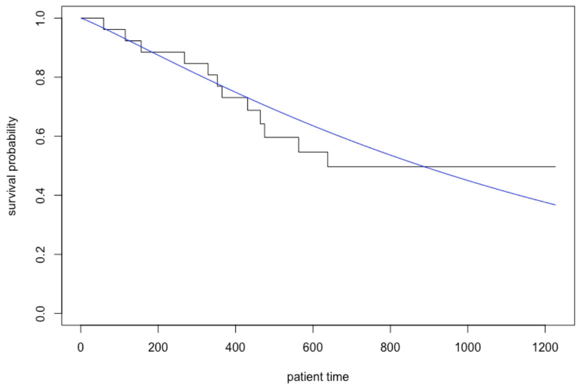


图 1

```
Call:
survreg(formula = ovarian_surv ~ 1, dist = "weibull")
              Value Std. Error    z       p
(Intercept)   7.111      0.293 24.3 <2e-16
Log(scale)   -0.103      0.254 -0.4   0.69

Scale= 0.902

Weibull distribution
Loglik(model)= -98   Loglik(intercept only)= -98
Number of Newton-Raphson Iterations: 5
n= 26
```

We can see that p-value for $\log(scale)$ is 0.69 .So we can not reject $H_0 : \alpha = 1$. We can't reject that data come from an exponential distribution.

(d)

Plot $\log[-\log\{S(t)\}]$ vs. log(t), is it a straight line?

```
plot(log(fit.KM$time), log(-log(fit.KM$surv)),xlab = "log(t)", ylab = "KM␣
    estimate␣of␣S(t)")
```

Alternative, plot the Weibull estimate of $S(t)$ vs. KM estimate, is it a straight line?

```
plot(1 - pweibull(fit.KM$time, shape = alpha.hat, scale = 1/lambda.hat), fit.
    KM$surv, xlab = "Weibull␣estimate␣of␣S(t)", ylab = "KM␣estimate␣of␣S(t)")
```

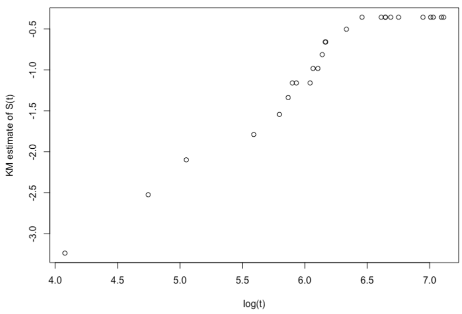
According to the figures 2(a)2(b), we can say that the Weibull assumption is plausible.

1.2 For the following data $(n = 10)$: $6, 9+, 10, 10+, 11, 13+, 16+, 17, 19+, 20$

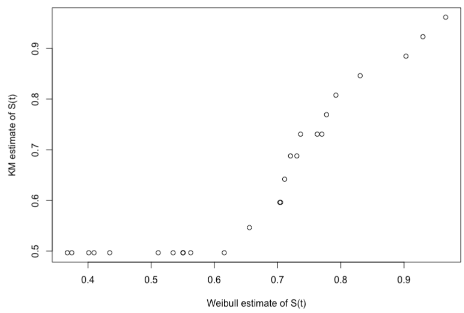
(a) Find the $K-M$ estimate of the survival function and an approximately 95%Cl for S(t) when $t = 10$

(b) Use the above KM estimate to get an estimate of the cumulative hazard at $t = 10$.

(c) Find the Nelson-Aalen estimate of the cumulative hazard function and its variance at t = 10



(a) 1



(b) 2

(d) Find the estimate and its variance of the survival function using the Nelson-Aalen estimate you got in (c) at $t = 10$.

(a) The KM estimate of the survival function is

$$\hat{S}_{\text{KM}}(10) = \frac{9}{10} \times \frac{7}{8} = \frac{63}{80} = 0.7875$$

According to log-log approach, an approximately 95% CI for $S(10)$ is

$$\left([\hat{S}(10)]^{\exp\left(z_{0.975}\,\text{se}(\hat{L}(10))\right)}, [\hat{S}(10)]^{-\exp\left(z_{0.975}\,\text{se}(\hat{L}(10))\right)} \right)$$

where

$$\text{se}(\hat{L}(10)) = \sqrt{\frac{1}{[\log \hat{S}(t)]^2} \sum_{j:\tau_j \le t} \frac{d_j}{(r_j - d_j)\, r_j}} = \sqrt{\frac{1}{[\log \hat{S}(10)]^2} \left(\frac{1}{(10-1)10} + \frac{1}{(8-1)8} \right)} = 0.7124587$$

an approximately 95% CI for $S(10)$ is $(0.3808, 0.9426)$ **(b)** An estimate of the cumulative hazard at $t = 10$ is

$$\hat{\Lambda}_{\text{KM}}(10) = -\log \hat{S}_{\text{KM}}(10) = 0.2388919$$

(c) The Nelson-Aalen estimate of the cumulative hazard function is

$$\hat{\Lambda}_{\text{NA}}(10) = \sum_{j:\tau_j \le 10} \frac{d_j}{r_j} = \frac{1}{10} + \frac{1}{8} = 0.225$$

and its variance is

$$\widehat{\text{Var}}(\hat{\Lambda}(10)) = \sum_{j:\tau_j \le 10} \frac{d_j}{r_j^2} = \frac{1}{10^2} + \frac{1}{8^2} = 0.025625$$

(d) The estimate using the Nelson-Aalen estimate at $t = 10$ is

$$\hat{S}(10) = \exp\left(-\hat{\Lambda}_{\text{NA}}(10)\right) = 0.7985162$$

and its variance of the survival function using the Nelson-Aalen estimate at $t = 10$ is

$$\begin{aligned} \widehat{\text{var}}\left(\hat{S}(10)\right) &= \widehat{\text{var}}\left(\exp\left(-\hat{\Lambda}_{\text{NA}}(10)\right)\right) = \left[\hat{S}(10)\right]^2 \widehat{\text{var}}\left(\hat{\Lambda}_{\text{NA}}(10)\right) \\ &= (0.7985162)^2 \times \left(\frac{1}{10^2} + \frac{1}{8^2}\right) = 0.01633922 \end{aligned}$$

1.3 Here is a follow-up data for male patients with heart diseases. Every patient was visited at the end of each year after being diagnosed. The follow-up continued for 15 years for every patient, or ended earlier due to death or censoring. (Data on Page 3)

(a) Find the life-table estimate of survival function of the time to death at years 6,8 and 10

(b) Find the variance of the estimate you got in (a) at years 6,8 and 10

(c) Repeat the above using R.

Aassuming censoring during the intervals

(a)

$$\hat{S}(6) = \left(1 - \frac{456}{2418}\right)\left(1 - \frac{226}{1962 - 39/2}\right)\left(1 - \frac{156}{1697 - 22/2}\right)\left(1 - \frac{171}{1523 - 23/2}\right)\left(1 - \frac{135}{1329 - 24/2}\right)$$
$$\cdot \left(1 - \frac{125}{1170 - 107/2}\right) = 0.4611239$$
$$\hat{S}(8) = \hat{S}(6)\left(1 - \frac{83}{938 - 133/2}\right)\left(1 - \frac{74}{722 - 102/2}\right) = 0.3711964$$
$$\hat{S}(10) = \hat{S}(8)\left(1 - \frac{51}{546 - 68/2}\right)\left(1 - \frac{42}{427 - 64/2}\right) = 0.2986843$$

(b)

$$\widehat{\text{Var}}(\hat{S}(6)) = (\hat{S}(6))^2\left(\frac{456}{2418 \times (2418 - 456)} + \cdots + \frac{125}{1116.5 \times (1116.5 - 125)}\right) = 0.0001077433$$
$$\widehat{\text{Var}}(\hat{S}(8)) = 0.0001119129$$
$$\widehat{\text{Var}}(\hat{S}(10)) = 0.0001186082$$

(c)

```
library(KMsurv)
tis <- seq(from=0, to=11, by=1)
ninit <- 2418
nlost <- c(0,39,22,23,24,107,133,102,68,64,45)
nevent <- c(456,226,152,171,135,125,83,74,51,42,43)
lifetab(tis, ninit, nlost, nevent)
```

This is consistent with the previous calculation.

1.4 Show that the CI for K-M estimator using log-log approach is $\left([\hat{S}(t)]^{e^A}, [\hat{S}(t)]^{e^{-A}}\right)$ where $L(t) = \log(-\log(S(t))), S(t) = \exp(-\exp(L(t))$, and $A = 1.96 * \text{se}(\hat{L}(t))$, and $\text{se}(\hat{L}(t)) = \text{sqrt}\left(\frac{1}{[\log \hat{S}(t)]^2} \sum_{j:\pi_j \le t} \frac{d_j}{(r_j - d_j) r_j}\right)$

First, let's talk about Delta method, which often appears in class of survival analysis.While the delta method generalizes easily to a multivariate setting, careful motivation of the technique is more easily demonstrated in univariate terms. Roughly, if there is a sequence of random variables X_n satisfying

$$\sqrt{n}\,[X_n - \theta] \xrightarrow{\mathcal{V}} \mathcal{N}\left(0, \sigma^2\right)$$

where θ and σ^2 are finite valued constants and $\xrightarrow{\mathcal{V}}$ denotes convergence in distribution, then

$$\sqrt{n}\,[g(X_n) - g(\theta)] \xrightarrow{\mathcal{V}} \mathcal{N}\left(0, \sigma^2 \cdot [g'(\theta)]^2\right)$$

for any function g satisfying the property that $g'(\theta)$ exists and is non-zero valued.

Proof:

Demonstration of this result is fairly straightforward under the assumption that $g'(\theta)$ is continuous. To begin, we use the mean value theorem (i.e.: the first order approximation of a Taylor series using Taylor's theorem):

$$g\left(X_n\right) = g(\theta) + g'(\tilde{\theta})\left(X_n - \theta\right),$$

$$g'(\tilde{\theta}) \xrightarrow{H} g'(\theta)$$

where $\xrightarrow{P}$ denotes convergence in probability. Rearranging the terms and multiplying by $\sqrt{n}$ gives

$$\sqrt{n}\left[g\left(X_n\right) - g(\theta)\right] = g'(\tilde{\theta})\sqrt{n}\left[X_n - \theta\right]$$

Since

$$\sqrt{n}\left[X_n - \theta\right] \xrightarrow{\nu} \mathcal{N}\left(0, \sigma^2\right)$$

by assumption, it follows immediately from appeal to Slutsky's theorem that

$$\sqrt{n}\left[g\left(X_n\right) - g(\theta)\right] \xrightarrow{\nu} \mathcal{N}\left(0, \sigma^2\left[g'(\theta)\right]^2\right)$$

This concludes the proof. Besides, one more step to obtain the order of approximation:

$$\begin{aligned}
\sqrt{n}\left[g\left(X_n\right) - g(\theta)\right] &= g'(\tilde{\theta})\sqrt{n}\left[X_n - \theta\right] = \sqrt{n}\left[X_n - \theta\right]\left[g'(\tilde{\theta}) + g'(\theta) - g'(\theta)\right] \\
&= \sqrt{n}\left[X_n - \theta\right]\left[g'(\theta)\right] + \sqrt{n}\left[X_n - \theta\right]\left[g'(\tilde{\theta}) - g'(\theta)\right] \\
&= \sqrt{n}\left[X_n - \theta\right]\left[g'(\theta)\right] + O_p(1) \cdot o_p(1) \\
&= \sqrt{n}\left[X_n - \theta\right]\left[g'(\theta)\right] + o_p(1)
\end{aligned}$$

This suggests that the error in the approximation converges to 0 in probability.

Q.E.D.

We have defined that $L(t) = \log(-\log(S(t)))$. Form a 95% confidence interval for $L(t)$ based on $\hat{L}(t)$, yielding $[\hat{L}(t) - A, \hat{L}(t) + A]$ .Since $S(t) = \exp(-\exp(L(t))$, the confidence bounds for the 95% CI of $S(t)$ are:

$$\left[\exp\left\{-e^{\hat{L}(t)+A}\right\}, \exp\left\{-e^{\hat{L}(t)-A}\right\}\right]$$

Substituting $\hat{L}(t) = \log(-\log(\hat{S}(t)))$ back into the above bounds, we get confidence bounds of

$$\left([\hat{S}(t)]^{e^{A}}, [\hat{S}(t)]^{e^{-A}}\right)$$

where A is $1.96 \cdot \mathrm{se}(\hat{L}(t))$. To calculate this, we need to calculate

$$\operatorname{Var}(\hat{L}(t)) = \operatorname{Var}[\log(-\log(\hat{S}(t)))]$$

From previous calculations in class, we know

$$\widehat{\operatorname{Var}}(\log[\hat{S}(t)]) = \sum_{j:\tau_j \le t} \frac{d_j}{\left(r_j - d_j\right) r_j}$$

Applying the delta method, we get:

$$\begin{aligned}
\widehat{\operatorname{Var}}(\hat{L}(t)) &= \widehat{\operatorname{Var}}(\log(-\log[\hat{S}(t)])) \\
&= \frac{1}{[\log \hat{S}(t)]^2} \sum_{j:\tau_j \le t} \frac{d_j}{\left(r_j - d_j\right) r_j}
\end{aligned}$$

1.5 The following table shows data on time to HIV development for a sample of 100 individuals with STD but free of HIV at time 0 : Use the data in this table to do the following (here we assume that censoring occurred in the middle of the interval):

Year intervals	of HIV positive	lost to follow-up
0 – 2	1	1
2 – 4	2	1
4 – 6	8	4
6 – 8	5	8
8 – 10	5	18
10 – 12	3	20
12 – 14	8	16

(a) Find the life-table estimate of the survival function of the time to HIV at years 6,8 , and 10 for the individuals with STD.

(b) Find the variance of the estimate you got in (a) at year 6,8, and 10 .

(c) Repeat the above using R.

Aassuming censoring during the intervals

(a)

$$\hat{S}(6) = \left(1 - \frac{1}{99.5}\right)\left(1 - \frac{2}{97.5}\right)\left(1 - \frac{8}{93}\right) = 0.8862329$$

$$\hat{S}(8) = \hat{S}(6)\left(1 - \frac{5}{79}\right) = 0.8301422$$

$$\hat{S}(10) = \hat{S}(8)\left(1 - \frac{5}{61}\right) = 0.7620978$$

(b)

$$\widehat{\text{Var}}(\hat{S}(6)) = (\hat{S}(6))^2\left(\frac{1}{99.5 \times 98.5} + \frac{2}{97.5 \times 95.5} + \frac{8}{93 \times 85}\right) = 0.001043686$$

$$\widehat{\text{Var}}(\hat{S}(8)) = 0.001505163$$

$$\widehat{\text{Var}}(\hat{S}(10)) = 0.002118635$$

(c)

```
library(KMsurv)
tis <- seq(from=0, to=14, by=2)
ninit <- 100
nlost <- c(1,1,4,8,18,20,16)
nevent <- c(1,2,8,5,5,3,8)
lifetab(tis, ninit, nlost, nevent)
```

This is consistent with the previous calculation.

1.6 For the following small data set of survival time: 13, 14, 15+, 16, 16+, 18, 21+, 24+, 25, 26+, **where "+" means a right censored survival time, do the following:**

(a) Find the Kaplan-Meier estimate of the survival function and its variance at each failure time.

(b) Use the above Kaplan-Meier estimate to get an estimate and its variance of the cumulative hazard function at each failure time.

(c) Find the Nelson-Aalen estimate of the cumulative hazard function and its variance at each failure time.

(d) Find an estimate and its variance of the survival function using the Nelson-Aalen estimate you got in (c) at each failure time.

(a) The KM estimate of the survival function is

$$\hat{S}_{KM}(t) = \prod_{j:\tau_j \le t} \frac{r_j - d_j}{r_j} = \prod_{j:\tau_j \le t} \left(1 - \frac{d_j}{r_j}\right)$$

its variance at each failure time t is

$$\widehat{\mathrm{Var}}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:\tau_j \le t} \frac{d_j}{(r_j - d_j)\, r_j}$$

So we get value as following

t	S_{KM}	$Var(S)$
13	0.9	0.009
14	0.8	0.016
16	0.6857143	0.02295044
18	0.5485714	0.02973481
25	0.2742857	0.04505002

(b) An estimate of the cumulative hazard at each failure time t is

$$\hat{\Lambda}_{\mathrm{KM}}(t) = -\log \hat{S}_{\mathrm{KM}}(t)$$

and its variance

$$\widehat{\mathrm{Var}}(\hat{\Lambda}(t)) = \sum_{i:t_i \le t} \frac{d_i}{n_i\,(n_i - d_i)}$$

So we get value as following

t	$\hat{\Lambda}_{\mathrm{KM}}(t)$	$\widehat{\mathrm{Var}}(\hat{\Lambda}(t))$
13	0.1053605	0.11
14	0.2231436	0.249
16	0.3772942	0.487
18	0.6004378	0.0987
25	1.293585	0.05987

(c) The Nelson-Aalen estimate of the cumulative hazard function is

$$\hat{\Lambda}_{\mathrm{NA}}(t) = \sum_{j:\tau_j \le t} \frac{d_j}{r_j}$$

and its variance is

$$\widehat{\mathrm{Var}}(\hat{\Lambda}(t)) = \sum_{j:\tau_j \le t} \frac{d_j}{r_j^2}$$

So we get value as following

t	$\Lambda_{\text{NA}}(t)$	$Var(\Lambda(t))$
13	0.1	0.001
14	0.2111111	0.02234568
16	0.3539683	0.04275384
18	0.5539683	0.08275384
25	1.053968	0.3327538

(d) The estimate using the Nelson-Aalen estimate at $t =$ is

$$\hat{S}(t) = \exp\left(-\hat{\Lambda}_{\text{NA}}(t)\right)$$

and its variance of the survival function using the Nelson-Aalen estimate at t is

$$\widehat{\text{var}}\left(\hat{S}(t)\right) = \widehat{\text{var}}\left(\exp\left(-\hat{\Lambda}_{\text{NA}}(t)\right)\right) = \left[\hat{S}(t)\right]^2 \widehat{\text{var}}\left(\hat{\Lambda}_{\text{NA}}(t)\right)$$

So we get value as following

t	$S(t)$	$Var(S(t))$
13	0.9048374	0.0008187307
14	0.8096841	0.01464957
16	0.7018972	0.02106309
18	0.5746648	0.0273286
25	0.3485519	0.04042574