Chapter 4 (Non-parametric Estimation for Right Censored Data)

Variable of interest: survival time T

Observed data: (tᵢ, δᵢ), i=1,2, ..., n.
• tᵢ = time on study (value of T) for the i-th subject
• δᵢ = indicator of whether the tᵢ is event time (δᵢ = 1) or censoring time (δᵢ = 0).

Assumptions:
• Independent subjects.
• Common population.
• Non-informative censoring.

Kaplan – Meier Estimator (KME) of survival function S(t) = P(T > t)

Example: Suppose a group of 10 patients joined a clinical study on Jan 1, 88. 6 of these patients died during that year. On Jan 1, 89 another group of 20 patients joined the study. In 89, 3 of group 1 patients died and 15 of group 2 patients died. The study terminated on Dec 31, 89. The variable of interest (T) is survival time in years. Estimate S(2), the chance of surviving more than 2 years.

Here: n = 30; t = 1 for 21 patients; t = 1+ for 5 patients; t = 2 for 3 patients & t = 2+ for 1 patient.

• Deaths are observed at ordered times t(1) = 1 < t(2) = 2.
• nᵢ = # patients who are at risk of death at time tᵢ, i.e., # patients with T ≥ tᵢ
• dᵢ = # deaths at time tᵢ, i=1,2.

Here, n₁ = ______, n₂ = ______, d₁ = ______, d₂ = ________.

Write S(2) = P(T > 2) =

• Estimate of P(T > 1) =

• Estimate P(T > 2 | T > 1) by the estimate of P(T > 2 | T ≥ 2) =

• KME of S(2) =

• No censoring case:

KME of S(t) in a general setup:

Suppose the deaths in the data are observed at m distinct times.

• Ordered death times: t(1) < t(2) < ... < t(m)
• m __ n.
• dᵢ = # deaths at time tᵢ
• nᵢ = # patients who are at risk of death at tᵢ
• Estimate of S(t) will be a step function with jumps only at the times tᵢ.

First consider estimating S(tᵢ).
\[ S(t_{(i)}) = P(T > t_{(i)}) = \]

- Estimate \( P(T > t_{(j)} \mid T > t_{(j-1)}) \) by the estimate \( P(T > t_{(j)} \mid T \geq t_{(j)}) = \)

- Estimate \( P(T > t_{(1)}) \) by

- \( \hat{S}(t_{(i)}) = \)

**KME:**

\[
\hat{S}(t) = \begin{cases} 
1, & \text{if } t < t_{(1)} \\
\prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i}, & \text{if } t \geq t_{(1)}
\end{cases}
\]

- Also known as the *Product – Limit estimator*.
- Very popular.
- Not well defined for \( t > t_{\text{max}} \) — the largest time on study.
- \( t_{\text{max}} = \text{death time: } \hat{S}(t_{\text{max}}) = \hat{P}(T > t_{\text{max}}) = \). So \( \hat{S}(t) = \) for \( t > t_{\text{max}} \).
- \( t_{\text{max}} = \text{censoring time: } \hat{S}(t_{\text{max}}) = \hat{P}(T > t_{\text{max}}) \).
- Alternatives:
  - \( \hat{S}(t) = \hat{S}(t_{\text{max}}) \) for \( t > t_{\text{max}} \) — assumes that the patient dies at \( t = \) ____. It is ____________ biased.
  - \( \hat{S}(t) = 0 \) for \( t > t_{\text{max}} \) — assumes that the patient dies ____________ time \( t_{\text{max}} \). It is ____________ biased.
  - For large \( n \), both roughly the same.
  - For small \( n \), the first is usually recommended.

**Properties of KME:**

- Under certain regularity conditions, when \( t \) is fixed and \( n \) is large, \( \hat{S}(t) \) approximately follows a Normal \( (S(t), \sqrt{V[\hat{S}(t)]}) \) distribution.
- Greenwood’s formula for \( V[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)} \)
- Non-parametric MLE.
- Reduces to the empirical CDF when there is no censoring.
Example: [HMO-HIV+ study] Suppose that the data for the first 7 subjects in this study are:

<table>
<thead>
<tr>
<th>ID</th>
<th>Time (months)</th>
<th>Censor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

Estimate the KME of survival function and its variance. Here, \( m = \) # distinct death times = _____

\[
\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i} \\
\hat{V}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}
\]

To summarize:

<table>
<thead>
<tr>
<th>( t )</th>
<th>( \hat{S}(t) )</th>
<th>( \hat{V}(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 \leq t &lt; 3 )</td>
<td>( \bullet )</td>
<td>( \bullet )</td>
</tr>
<tr>
<td>( 3 \leq t &lt; 5 )</td>
<td>( \bullet )</td>
<td>( \bullet )</td>
</tr>
<tr>
<td>( 5 \leq t &lt; 22 )</td>
<td>( \bullet )</td>
<td>( \bullet )</td>
</tr>
<tr>
<td>( 22 \leq t )</td>
<td>( \bullet )</td>
<td>( \bullet )</td>
</tr>
</tbody>
</table>

Let us now learn how to use R to get the KME and nice plots.

#### Learning R ####

# load the survival package
> library(survival)

# see the help for available functions #
> library(help=survival)

# enter the data for example 1 #
> stime <- c(5, 6, 8, 3, 22, 22, 8)
> censor <- c(1, 0, 1, 1, 1, 0, 1)

# see help for various functions #
> ?Surv
> ?survfit

# put the data in a format that R understands #
> ex1 <- Surv(stime, censor)
> ex1
[1] 5 6+ 8 3 22 22+ 8

# get the estimate, its SE and the plot of the estimated survival function#
> ex1.out <- survfit(ex1, conf.type = "none")
> plot(ex1.out, xlab = "time (months)", ylab = "KME of S(t)"")
> summary(ex1.out)
Call: survfit(formula = ex1, conf.type = "none")

<table>
<thead>
<tr>
<th>time</th>
<th>n.risk</th>
<th>n.event</th>
<th>survival</th>
<th>std.err</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>7</td>
<td>1</td>
<td>0.857</td>
<td>0.132</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>1</td>
<td>0.714</td>
<td>0.171</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>2</td>
<td>0.357</td>
<td>0.198</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>1</td>
<td>0.179</td>
<td>0.160</td>
</tr>
</tbody>
</table>

Do these results match with the ones we had obtained? Plot of the estimated survival function:

![Survival Function Graph]

What do you notice?

- _____ function. The curve decreases at an observed _____________ time.
- Minimum ____ zero.
- Marks indicate censored observations.
• The curve __________ change at a censored survival time – its value remains the same as the value of the previous observed uncensored death time.

Let us now look the estimated S(t) for the real HMO-HIV+ data:

# read the data #
> hmohiv <- read.table(file="hmohiv.dat",header=TRUE)
# look at first row of the data #
> hmohiv[1,]

ID ENTDATE ENDDATE TIME AGE DRUG CENSOR
1 1 15may90 14oct90 5 46 0 1

# get KME and plot it#
> hmohiv.kme <- survfit(Surv(hmohiv$TIME,hmohiv$CENSOR))
> plot.survfit(hmohiv.kme, xlab="time (months)",
               ylab="survival probability",
               main="KME of survival probability for HMO-HIV+ data",
               conf.int=0, mark.time=FALSE)

Interpretation:
• Curve descends sharply and tails off gradually.
• The minimum is not zero.
Shape of a KME:

- Depends on % of censored and non-censored observations and their actual values.
- Also depends on the pattern of enrollment in a follow-up study.
- No “typical” shape.

Ex: HMO-HIV+ study

- Suppose that many subjects had long survival times with the same % of censored observations. The curve would descend __________ at first and then more rapidly in the end.
- Suppose that the survival times remained same but the censoring % was higher. The curve would look more ____________.

Ex: Consider a 5-year follow-up study with 2-year enrollment period. If it has many late entries, then it is likely to have more censored observations and thus a different looking estimated survival function than the same study with many early entries.

Confidence interval for S(t)

100(1-α)% CI for S(t) at a fixed time point t = t₀:

- Find [L, U] such that

  Recall: \( \hat{S}(t₀) \) is asymptotically normal with mean \( S(t₀) \) and variance \( V(\hat{S}(t₀)) \) given by the Greenwood’s formula.

- No extrapolation: Valid only over the interval \((0, t_{max})\).

Linear CI for S(t₀):

- Drawback:

  Can get better interval for \( S(t₀) \) by first constructing CI for \( \ln(S(t₀)) \) or \( \ln(-\ln(S(t₀))) \) and then transforming back. Recall that \( -\ln[S(t₀)] \) is the cumulative hazard function at time \( t₀ \).

- By default, R uses the \( \ln \) version. Our book recommends the \( \ln-\ln \) version – also available in R.

ln-ln CI for S(t₀):

- Delta-method: \( \ln(-\ln[\hat{S}(t₀)]) \) has asymptotic normal distribution with mean \( \ln(-\ln[S(t₀)]) \) and approximate variance:

\[
\hat{V}[\ln(-\ln\hat{S}(t₀))] = \frac{\hat{V}[\hat{S}(t₀)]}{\hat{S}^2(t₀)\ln[\hat{S}(t₀)]^2}
\]

- CI for \( \ln(-\ln[S(t₀)]) \):
• CI for $S(t_0)$:

• We will use the ln-ln version unless specified otherwise.

**Pointwise CI for $S(t)$:**

• Construct $[L(t), U(t)]$ – an *individual* $100(1-\alpha)\%$ CI for $S(t)$ at each time $t \leq t_{\text{max}}$.

• It satisfies: $P[L(t) \leq S(t) \leq U(t)] \approx 1 - \alpha$ for each $t \leq t_{\text{max}}$ — a pointwise CI.

• Cannot claim that $P[L(t) \leq S(t) \leq U(t) \text{ for each } t \leq t_{\text{max}}] \approx 1 - \alpha$ — a simultaneous CI.

• See book for simultaneous CI.

**Example:** HMO-HIV+ study.

```r
# get the 95% pointwise “ln-ln” CI and plot it #

> hmohiv.loglog <- survfit(Surv(hmohiv$TIME,hmohiv$CENSOR), conf.type="log-log", conf.int=0.95)

> plot.survfit(hmohiv.loglog,xlab="time (months)", ylab="survival probability", main="Estimate of survival probability \nand its 95% pointwise CI for HMO-HIV+ data", mark.time=FALSE)
```

Clearly, the CI also supports our earlier observation that there are many early deaths with a few deaths near the maximum of 5 years of follow-up.

**General properties of a pointwise ln-ln based CI:**

• Skewed for large and small values of the probability and fairly symmetric in the middle.

• Shorter in the beginning.

• Lie within $[0,1]$.

• *Not simultaneous* CI’s.

• valid as long as $n \geq 25$ and right-censoring $\leq 50\%$. 

Estimation of percentiles of survival time $T$

- Focus on median. Other percentiles similar.
- Median $t_{0.5}$ is such that
- More precisely, median $t_{0.5}$ is such that $\inf\{t: S(t) \leq 0.5\}$.

**Graphical approach**: (used in R)

**Example**: HMO-HIV+ study.

```r
> print.survfit(hmohiv.loglog)
Call: survfit(formula = Surv(hmohiv$TIME, hmohiv$CENSOR), conf.type = "log-log", conf.int = 0.95)

         n    events     rmean      se(rmean)    median  0.95LCL  0.95UCL
100.00   80.00     14.67       1.97      7.00      5.00      9.00

# verify the method #

> plot.survfit(hmohiv.loglog, xlab="time (months)", ylab="survival probability", main="Estimate of survival probability and its 95% pointwise CI for HMO-HIV+ data", mark.time=FALSE)

> abline(h=0.5,v=c(5,7,9),lty=3)
```
**Brookmeyer-Crowley approach:**

Consider testing $H_0: S(t) = 0.5$ versus $H_1: S(t) \neq 0.5$ for a specified $t$.

- **ln-ln based test statistic** = $LL(t) = \frac{\ln(-\ln(S(t))) - \ln(-\ln(0.5))}{\hat{SE}[\hat{S}(t)][\hat{S}(t)\ln\hat{S}(t)]}$
- Reject if $LL(t) > z_{\alpha}$
- $\{t: H_0$ is not rejected$\} = \ldots$
- Recommended when there are no ties. Read Example 4.2 of the textbook for an illustration.

**Estimation of mean survival time:** Read Section 4.5 of the text.

**HMO-HIV+ DATA**


DISCLAIMER: This text has been downloaded from the following Wiley's FTP site: ftp://ftp.wiley.com/public/sci_tech_med/survival

Data are in the file hmothiv.dat
n =100

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Codes / Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Subject ID Code</td>
<td>1-100</td>
</tr>
<tr>
<td>ENTDATE</td>
<td>Entry date</td>
<td>ddmmyr</td>
</tr>
<tr>
<td>ENDDATE</td>
<td>Entry date</td>
<td>ddmmyr</td>
</tr>
<tr>
<td>TIME</td>
<td>Survival Time</td>
<td>days between Entry date and End date</td>
</tr>
<tr>
<td>AGE</td>
<td>Age</td>
<td>years</td>
</tr>
<tr>
<td>DRUG</td>
<td>History of IV Drug Use</td>
<td>0 = No, 1 = Yes</td>
</tr>
<tr>
<td>CENSOR</td>
<td>Follow-Up Status</td>
<td>1 = Death due to AIDS or AIDS related factors, 0 = Alive at study end or lost to follow-up</td>
</tr>
</tbody>
</table>