Nonparametric Survival Analysis in R

We discuss how to implement nonparametric methods used in the analysis of survival data in the R software package. Survival data are given by \( D = \{(y_i, x_i, \delta_i) : i = 1, 2, 3, \ldots, n\} \), where \( y_i = \min(T_i, C_i) \) is the minimum of the failure time \( T_i \) and censoring time \( C_i \) for subject \( i \), \( x_i \) denotes a collection of covariate information obtained on subject \( i \), and \( \delta_i \) is the event indicator for subject \( i \). Here \( \delta_i = 1 \) if the event was observed and \( \delta_i = 0 \) otherwise. Typically the goal is to compare two populations, for example exposed and unexposed populations with respect to survival prospects, having adjusted for other variables where appropriate. In this computing handout, we present R code for estimating the survival functions, \( S_E(t) \) and \( \bar{S}_E(t) \), for the exposed and unexposed groups using Kaplan-Meier curves. We also discuss log-rank tests for comparing the distribution of survival times for two populations. The methods considered in this handout are nonparametric in that we do not specify a form for the distribution of survival times for the \( E \) and \( \bar{E} \) populations.

In this handout we present the R code that was used to obtain the inferences in Survival Analysis Notes 2. The R functions used for survival analysis are stored in the survival library. You need to use

\[
\text{library(survival)}
\]

to have access to the functions that we will need.

1 The Burn Data

We consider data from a study designed to assess the effectiveness of a new wound cleansing treatment on the time to infection of burn patients. The analyses presented are for illustrative purposes only. The covariates included in the data set that we will consider include (i) treatment (1 = new, 0 = standard), (ii) female (1 = female, 0 = male), (iii) white (1 = white, 0 = other), (iv) surface area burned, and (v) burn type (1 = chemical, 2 = scald, 3 = electric, 4 = flame).

To read in the burn data I used the code

1
burndata <- read.table(file="C:\Documents and Settings\abranscum\My Documents\CPH630\Fall2005\SurvivalAnalysisNotes\BurnData\subsetoftheburndata.txt",header=T, sep="\t")
attach(burndata)
options(contrasts=c("contr.treatment","contr.poly"))
library(survival)

2 Estimates of the Survivor Function

Kaplan and Meier developed a nonparametric method for estimating \( S(t) = \Pr(T > t) \). We will denote the Kaplan-Meier estimate of \( S(t) \) by \( \hat{S}_{KM}(t) \). Given samples of survival data from the \( E \) and \( \bar{E} \) populations, we can use \( \hat{S}_{KM,E}(t) \) and \( \hat{S}_{KM,\bar{E}}(t) \) to compare the survival prospects for the \( E \) and \( \bar{E} \) populations. In the absence of censoring, the Kaplan-Meier estimate reduces to the empirical survival function, which is discussed in the next subsection.

2.1 Empirical survival function

Consider the following survival times in months for 11 male patients with malignant bone tumors

\[ 11, 13, 13, 13, 13, 14, 14, 15, 15, 17 \]

The graph of \( \hat{S}(t) \) is a step-function with steps at each observed death time. For values of \( t \) that occur between death times, \( \hat{S}(t) \) is constant. For all \( t \leq 11 \), \( \hat{S}(t) = 1 \) and for \( t > 17 \), \( \hat{S}(t) = 0 \). Figure 1 presents a plot of the empirical survival function for the bone tumor data, which was produces using the following code.

\[ x <- c(11,13,13,13,14,14,15,15,17) \]
\[ plot(stepfun(c(11,13,14,15,17),c(1,10/11,5/11,3/11,1/11,0)), \]
\[ main="",xlab="t(Months)",ylab="Empirical Survival Beyond t") \]

2.2 Kaplan-Meier estimate of \( S(t) \)

Kaplan and Meier developed a nonparametric method for estimating \( S(t) \). As for the empirical survival function, \( \hat{S}_{KM}(t) \) is a step-function that has steps
at each observed death time and is constant in between death times. But unlike the empirical survival function, censored observations are incorporated into the computing of $S_{KM}(t)$.

2.2.1 Kaplan-Meier curves comparing the new and routine treatment

The Kaplan-Meier estimates of the survivor functions for each of the two treatment groups are plotted in Figure 2. The new treatment appears stronger than the routine treatment with respect to infection time. Figure 2 was produced using the following code.

```r
KM.fit <- survfit(Surv(time, status) ~ newtrt)
plot(KM.fit, conf.int=F, lty = c(1,2), xlab="Time to Infection", ylab="Survival Probability", main="Kaplan-Meier Curves for New and Standard Treatments")
legend(10, .1, c("Routine Treatment", "New Treatment"), lty=c(1,2),bty="n")
```

2.2.2 Kaplan-Meier curves comparing the new and routine treatment adjusted for another variable

We can obtain Kaplan-Meier curves within any strata. For example, we can compare the new and routine treatments for males and for females separately. The R code for males is

```r
KM.fitM <- survfit(Surv(time, status) ~ newtrt, subset=female==0)
plot(KM.fitM, conf.int=F, lty = c(1,2), xlab="Time to Infection", ylab="Surv Prob", main="Kaplan-Meier Curves \n for Males")
legend(8, .1, c("Routine", "New"), lty = c(1,2), bty="n")
```

which produced Figure 3.

Similarly the following R code produced Figure 4.
KM.fitF <- survfit(Surv(time,status)~newtrt, subset=female==1)
plot(KM.fitF, conf.int=F, lty = c(1,2), xlab="Time to Infection", ylab="Surv Prob", main="Kaplan-Meier Curves \n for Females")
legend(8, .1, c("Routine", "New"), lty = c(1,2), bty="n")

We can present the Kaplan-Meier estimates for the four groups (Routine, Male), (New, Male), (Routine, Female), and (New, Female) using

KM.fit0 <- survfit(Surv(time, status)~newtrt +female)

plot(KM.fit0, conf.int=F, lty = c(1,2,3,4), xlab="Time to Infection", ylab="Surv Prob", main="Kaplan-Meier Curves \n for Males and Females")
legend(7, .3, c("Routine, Male", "New, Male","Routine, Female", "New, Female" ), lty = c(1,2,3,4), bty="n")

but the plot (Figure 5) is somewhat busy and hence difficult to read.

3 Log-rank Tests for Two Populations

Log-rank tests can be used to test the hypothesis of equality of two survival curves, for example

\[ H_0 : S_{New}(t) = S_{Std}(t) \text{ for all } t. \]

The R code used to test this null hypothesis for the burn data is

survdiff(Surv(time, status)~newtrt, rho=0)
The output was

```
Call: survdiff(formula = Surv(time, status) ~ newtrt, rho = 0)

 N Observed Expected (O-E)^2/E (O-E)^2/V
newtrt=0 70 28 21.4 2.07 3.79
newtrt=1 84 20 26.6 1.66 3.79

Chisq= 3.8 on 1 degrees of freedom, p= 0.0515
```

so the p-value is 0.0515 which suggests that the two treatment groups differ with respect to the distributions of times to infection.

We can test whether the distributions of infection times are different for patients with different burn types, and obtain Kaplan-Meier estimates for each burn type using the following code.

```
survdiff(Surv(time, status) ~ factor(burntype))

KM.fit1 <- survfit(Surv(time, status)~ factor(burntype))
```

```
plot(KM.fit1, conf.int=F, lty = 1:4, xlab="Months", ylab="Surv Prob", main="Kaplan-Meier curves for the 4 Burn Types")
```

```
legend(5, 0.2, c("Chemical", "Scald", "Electric", "Flame"), lty=1:4, bty="n")
```

The output of the log-rank test was

```
Call: survdiff(formula = Surv(time, status) ~ factor(burntype))

 N Observed Expected (O-E)^2/E (O-E)^2/V
factor(burntype)=1 9 1 2.97 1.310 1.429
factor(burntype)=2 18 6 5.11 0.155 0.177
factor(burntype)=3 11 6 2.04 7.677 8.261
factor(burntype)=4 116 35 37.88 0.218 1.065

Chisq= 9.6 on 3 degrees of freedom, p= 0.0222
```

so the p-value was 0.02 so that the distributions of infection times are significantly different for at least one burn type.
3.1 Stratified Log-rank Tests

Stratified log-rank tests can be used to compare the survival distributions of 
$E$ and $\bar{E}$ populations adjusted for another variable.

For the burn data, for example, we can test whether the distribution of 
infection times are the same for the two treatment groups within each burn 
type:

$$H_0 : S_{New,e}(t) = S_{Std,e}(t),$$
$$S_{New,a}(t) = S_{Std,a}(t),$$
$$S_{New,e}(t) = S_{Std,e}(t),$$
$$S_{New,f}(t) = S_{Std,f}(t), \text{ for all } t.$$  

The R code used to test $H_0$ is

```R
survdiff(Surv(time, status) ~ newtrt + strata(factor(burntype)))
```

which resulted in the following output.

Call: survdiff(formula = Surv(time, status) ~ newtrt + 
strata(factor(burntype)))

<table>
<thead>
<tr>
<th></th>
<th>N Observed Expected (O-E)^2/E (O-E)^2/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>newtrt=0</td>
<td>70  28  21  2.35  4.35</td>
</tr>
<tr>
<td>newtrt=1</td>
<td>84  20  27  1.82  4.35</td>
</tr>
</tbody>
</table>

Chisq= 4.3 on 1 degrees of freedom, p= 0.037

The p-value here is 0.037. Therefore the distributions of infection times for 
the two treatment groups are significantly different for at least one burn type.
Figure 1: Empirical survival function for the bone tumor data.
Figure 2: Kaplan-Meier estimates of the survival functions for the new and standard treatment groups.
Figure 3: Kaplan-Meier estimates of the survival functions for the new and standard treatment groups for males.
Figure 4: Kaplan-Meier estimates of the survival functions for the new and standard treatment groups for females.
Figure 5: Kaplan-Meier estimates of the survival functions for the new and standard treatment groups for males and females.
Figure 6: Kaplan-Meier estimates of the survival functions for patients with different burn types.