Survival Analysis by Example

Hands on approach using R

Faye Anderson

Survival Analysis by Example
Hands on approach using R

First Edition

Faye Anderson, MS, PhD

Contents

Veritas

Preface

Throughout my work as a statistician, the most challenging task has been to relay what the analyses were saying to non-statisticians. This book offers help in this area but focusing only on survival analyses.

This book explains survival analysis using examples and plain English. There are many good books on survival analysis but most require the reader to go through the theory first. This book skips the formulas and dives right in the applied substance of the matter.

R was selected because of its free accessibility but the examples can easily be replicated using other statistical software. Last but not least, this book assumes basic knowledge of R. Code on how to import data, install a package, or save results can easily be obtained from the multitude of public sources in the internet.

Enjoy!

Chapter 1: Survival Analysis Terminology

Survival analysis is about analyzing data where the outcome is the time to the occurrence of an event. This even can be death, onset of disease, or machine failure. Examples include the number of years until an economic downturn happens or the number of years until a person develops a disease. This time period is called survival time and when it is unknown, the observation is called censored. Survival time is greater or equal to zero. If a subject drops from the study before its end then his/her survival time is censored or missing, which is included in the dataset in order to avoid bias.

Censoring can be left, right, or interval (refer Example 1). Dealing with censored/missing data is beyond the scope of this book. Replacing the missing data with zeroes, averages, or other values can significantly skew the results because the analysis will be based on a **different** sample that might not be representative of the population of interest.

Each subject/machine survival prospects remain constant throughout the study period. Observations are independent and each is only included once.

Example 1: Censored Observations

A clinical trial studies patients with risk of heart attack for four months. If a patient does not have an attack throughout the duration of the study then his/her record is called right censored and the survival time for this subject is four months. If another patient had a heart attack before entering the study, or before the study ends then his/her survival time is left censored. Interval censoring happens if a subject had a heart attack during the four months of the study but the exact time of the attack was not recorded for some reason. Censoring status does not affect the survival prospect of a patient/machine.

Why not use regression?

Since survival time is usually a continuous number, why not use ordinary regression analyses where survival time is the dependent variable (outcome)? Because survival time cannot be negative, linear regression models would be skewed. Moreover, because the dependent variable in survival analysis has two aspects: time to event and status, ordinary regression models cannot answer two important questions:
Q1) what’s the probability of surviving past a point in time (survival function)?
Q2) what’s the failure rate to a certain point in time (also known as hazard function)? E.g.; how many will die by the age of 75? How many years will the machine work properly before we need to buy a new one?

Parametric, Non-parametric and Semi-parametric Survival Analysis

The table below summarizes the differences between the three approaches. Detailed examples are presented in the following chapters.

Parametric	Non-parametric	Semi-parametric
Assume knowledge of the statistical distribution of survival times	Make no assumptions on the distribution of survival times like Kaplan Meier estimator	Has parametric and non-parametric components like the Cox regression model

Example 2: Exploratory Analyses

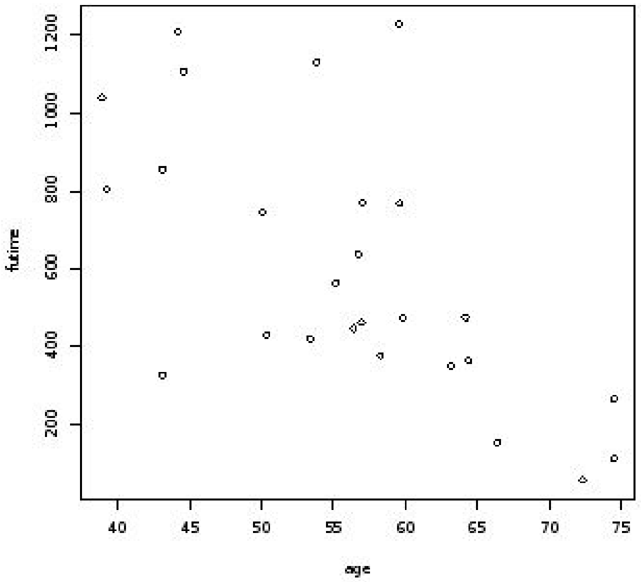
The ovarian cancer data comes with the survival package in R. The following few commands explore the data.

```
install.packages("survival",repos="http://cran.r-project.org") #install survival library
library(survival) # load survival library
> data(ovarian)
> dim(ovarian) # Ovarian data has 26 rows and 6 columns
[1] 26  6
> help(ovarian)
starting httpd help server ... done
```

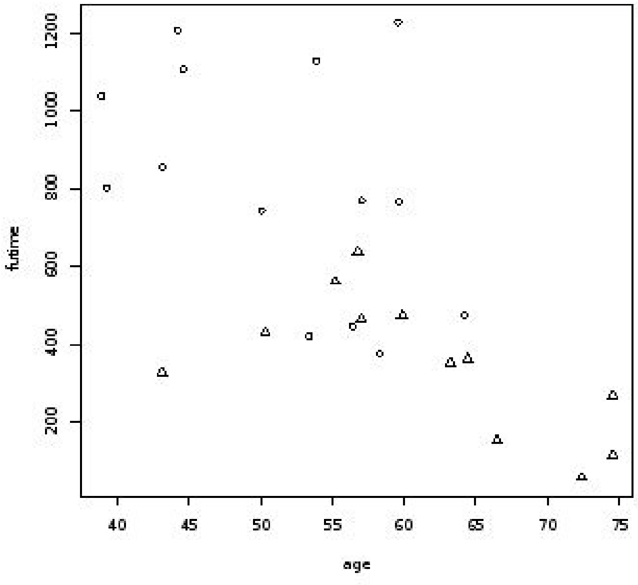
Description
Survival in a randomised trial comparing two treatments for ovarian cancer
Usage
ovarian
Format
futime: survival or censoring time
fustat: censoring status
age: in years
resid.ds: residual disease present (1=no,2=yes)
rx: treatment
groupecog.ps: ECOG performance status (1 is better, see reference)

```
> ovarian
   futime fustat     age resid.ds rx ecog.ps
1      59      1 72.3315        2  1       1
2     115      1 74.4932        2  1       1
3     156      1 66.4658        2  1       2
4     421      0 53.3644        2  2       1
5     431      1 50.3397        2  1       1
6     448      0 56.4301        1  1       2
7     464      1 56.9370        2  2       2
8     475      1 59.8548        2  2       2
9     477      0 64.1753        2  1       1
10    563      1 55.1781        1  2       2
11    638      1 56.7562        1  1       2
12    744      0 50.1096        1  2       1
13    769      0 59.6301        2  2       2
14    770      0 57.0521        2  2       1
15    803      0 39.2712        1  1       1
16    855      0 43.1233        1  1       2
17   1040      0 38.8932        2  1       2
18   1106      0 44.6000        1  1       1
```

```
19   1129      0 53.9068        1  2        1
20   1206      0 44.2055        2  2        1
21   1227      0 59.5890        1  2        2
22    268      1 74.5041        2  1        2
23    329      1 43.1370        2  1        1
24    353      1 63.2192        1  2        2
25    365      1 64.4247        2  2        1
26    377      0 58.3096        1  2        1
> attach(ovarian) # remember to detach it at the end
> summary(futime) # survival time
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   59.0   368.0   476.0   599.5   794.8  1227.0
> summary(age) # subjects age
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  38.89   50.17   56.85   56.17   62.38   74.50
> cor(futime, age) # pair-wise correlation
[1] -0.6483612

> psymbol<-fustat+1
> table(psymbol) # 2 = censored
psymbol
1  2
14 12
> plot(age, futime)
```



```
plot(age, futime, pch=(psymbol))
```



```
> detach(ovarian)# so the variables do not overlap
```

Interpretation: This dataset had 26 subjects (patients), and 12 censored observations (fustat). Survival time ranged from 59 to 1227 weeks, with an average of 599.8 weeks. The patients' ages averaged at 56 years and ranged from 38.9 to 74.5 years. The first plot contrasts survival time against age regardless of censored status. Notice that as age increases survival time decreases. This is also manifested in the negative strong pairwise correlation between the two (-0.65). The second plot differentiates the subjects with censored status (triangle = censored). Censored ones are presented with triangles. They show relatively lower survival time.

Chapter 2: Parametric Survival Analysis

This approach assumes that survival time data follows a certain distribution like exponential, Weibull, lognormal, log logistic, or generalized gamma. It rarely, if ever follows a normal distribution. Not all R functions support all of these distributions, so you will need to read the documentation of the function in order to find out which distribution it supports. You can do this by typing *help(function-name)* in R environment, which takes you to an online description of whatever function or dataset that is between the parentheses.

Example 3: Fitting a Parametric Model

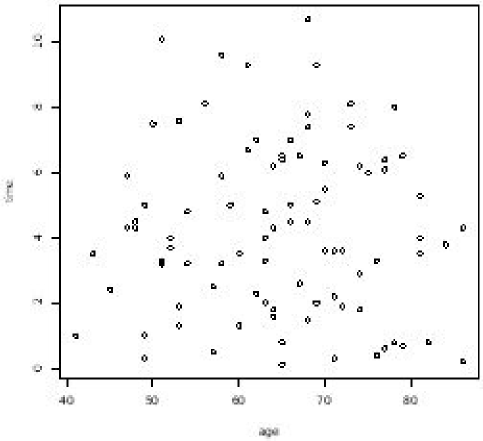
This comprehensive example explores the larynx cancer data which is available from the KMsurv package.

```
>install.packages("KMsurv",repos="http://cran.r-project.org")
> library(KMsurv)
Warning message:
package ‘KMsurv’ was built under R version 3.1.3
> data(larynx)
help(larynx)
```

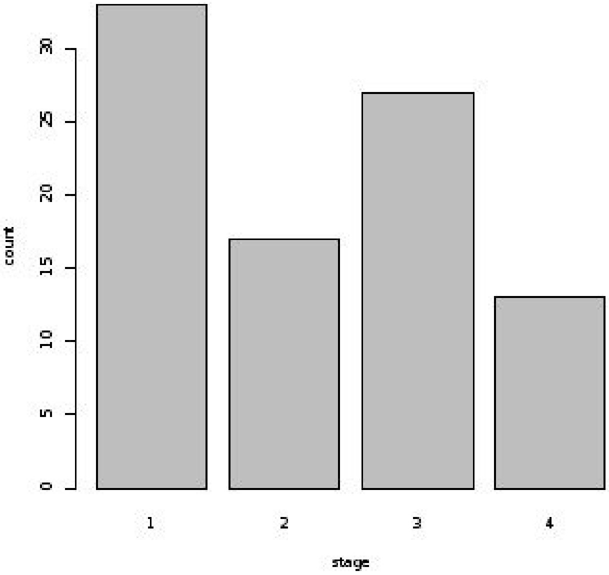
Description
The larynx data frame has 90 rows and 5 columns.
Format
This data frame contains the following columns:
stage
Stage of disease (1=stage 1, 2=stage2, 3=stage 3, 4=stage 4)
time
Time to death or on-study time, months
age
Age at diagnosis of larynx cancer
diagyr
Year of diagnosis of larynx cancer
delta
Death indicator (0=alive, 1=dead)

```
> colnames(larynx)
[1] "stage"  "time"   "age"    "diagyr" "delta"
> dim(larynx)
[1] 90  5

> attach(larynx)
> summary(time)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.100   2.000   4.000   4.198   6.200  10.700

> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  41.00   57.00   65.00   64.61   72.00   86.00

> summary(diagyr)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  70.00   72.25   74.00   74.24   76.00   78.00

> sum(delta) # how many died, 1=dead
[1] 50
```

```
> plot(age,time)
```



```
> table(stage) # count subjects in each stage
stage
1  2  3  4
33 17 27 13
> barplot(table(stage), xlab ="stage", ylab="count") # bar chart
```



```
> library(survival)
> WeibullFit <- survreg(Surv(time, delta) ~ as.factor(stage) + age, dist="weibull") # fit a parametric survival regression model where survival time follows Weibull distribution
> WeibullFit
Call:
survreg(formula = Surv(time, delta) ~ as.factor(stage) + age,
    dist = "weibull")

Coefficients:
      (Intercept) as.factor(stage)2 as.factor(stage)3 as.factor(stage)4
       3.52875996       -0.14770417       -0.58655845       -1.54407608
              age
      -0.01746367
```

```
Scale= 0.8848432

Loglik(model)= -141.4   Loglik(intercept only)= -151.1
        Chisq= 19.37 on 4 degrees of freedom, p= 0.00066
n= 90

> summary(WeibullFit)

Call:
survreg(formula = Surv(time, delta) ~ as.factor(stage) + age,
    dist = "weibull")
                    Value Std. Error      z        p
(Intercept)        3.5288     0.9041  3.903 9.50e-05
as.factor(stage)2 -0.1477     0.4076 -0.362 7.17e-01
as.factor(stage)3 -0.5866     0.3199 -1.833 6.68e-02
as.factor(stage)4 -1.5441     0.3633 -4.251 2.13e-05
age               -0.0175     0.0128 -1.367 1.72e-01
Log(scale)        -0.1223     0.1225 -0.999 3.18e-01

Scale= 0.885

Weibull distribution
Loglik(model)= -141.4   Loglik(intercept only)= -151.1
        Chisq= 19.37 on 4 degrees of freedom, p= 0.00066
Number of Newton-Raphson Iterations: 5
n= 90

> WeibullFit$coeff # model coeffficients
      (Intercept) as.factor(stage)2 as.factor(stage)3 as.factor(stage)4               age
       3.52875996       -0.14770417       -0.58655845       -1.54407608       -0.01746367

> WeibullFit$icoef  # intercept and scale coefficients
  Intercept  Log(scale)
2.01689111 -0.01479215

> WeibullFit$var  # variance-covariance matrix
                  (Intercept) as.factor(stage)2 as.factor(stage)3 as.factor(stage)4           age    Log(scale)
(Intercept)        0.81743869     -0.0904873892     -0.0847916997     -0.0444384841 -0.0111447001  0.0259112033
as.factor(stage)2 -0.09048739      0.1661119777      0.0531882202      0.0506790554  0.0005697726  0.0001585792
as.factor(stage)3 -0.08479170      0.0531882202      0.1023678205      0.0566836066  0.0004230051 -0.0073114005
as.factor(stage)4 -0.04443848      0.0506790554      0.0566836066      0.1319623242 -0.0002043104 -0.0107031439
age               -0.01114470      0.0005697726      0.0004230051     -0.0002043104  0.0001632873 -0.0002596823
Log(scale)         0.02591120      0.0001585792     -0.0073114005     -0.0107031439 -0.0002596823  0.0150075201

> WeibullFit$scale  # scale parameter
[1] 0.8848432
```

```
detach(larynx)
```

Interpretation: This dataset had 90 subjects (patients), of average age of 64.6 years and survival time of 4.2 years. Fifty of them died of larynx cancer. There is no censored information in this dataset. The command of *table(stage)* counts the number of subjects in each of the stages of larynx cancer. A parametric survival model was fit (WeibullFit) where the independent variable is the stage of the disease and the dependent variables are both survival time and death indicator. It was assumed that survival time followed Weibull distribution. The p-value of the model is less than the default significance level of 0.05 (*chi-square p= 0.00066)* indicates that the model is significant. As for the model coefficients (summary command), their p-values were all less than the default significance level of 0.05 and they were all negative indicating strong negative association. By default stage 1 of the disease was considered as the reference stage so the coefficients of the remaining three stages are relative to stage 1 having a coefficient of 1. You can try the model again with different distributions by replacing “weibull” by "exponential", "gaussian", "logistic","lognormal" or "loglogistic".

Chapter 3: Non-parametric Survival Analysis

One of the most famous non-parametric models is called the Kaplan-Meier estimator because it does not call for any assumptions.

Example 4: Fitting a Non-parametric Model

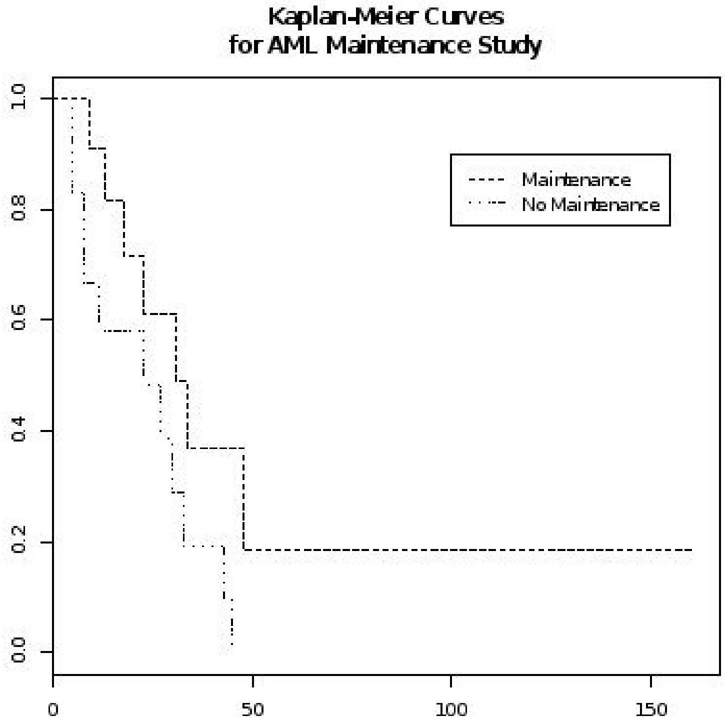
This is an example using the Acute Myelogenous Leukemia (aml) survival data, and is extracted from the survival library documentation. Some R versions recognize it with the name “leukemia”.

```
library(survival)
data(aml) # or data(leukemia)
help(aml)
starting httpd help server ... done
```

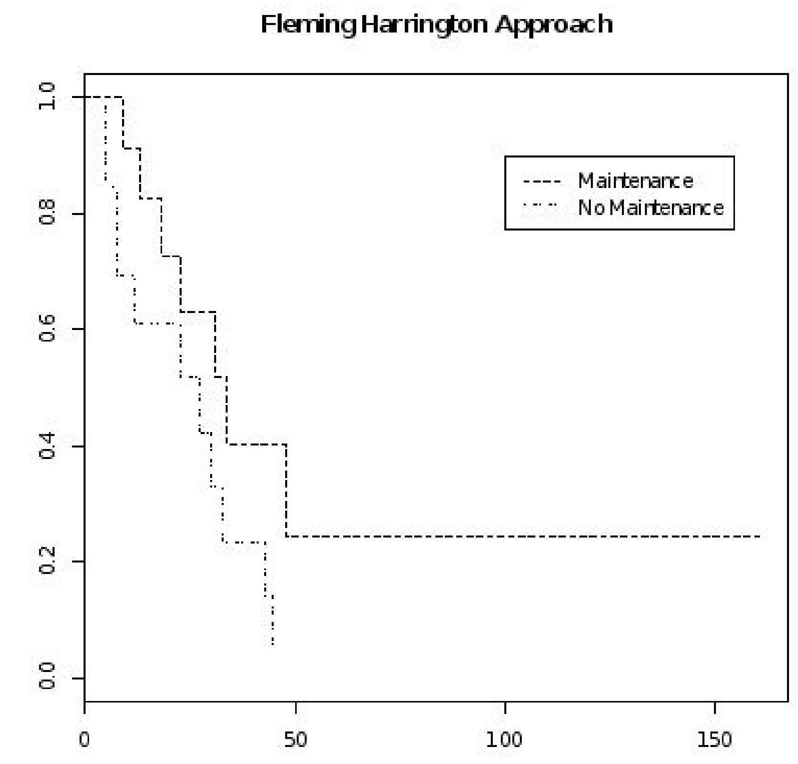
time: survival or censoring time
status: censoring status
x: maintenance chemotherapy given? (factor)

```
> head(aml)# status 0 = censored
  time status          x
1    9      1 Maintained
2   13      1 Maintained
3   13      0 Maintained
4   18      1 Maintained
5   23      1 Maintained
6   28      0 Maintained
```

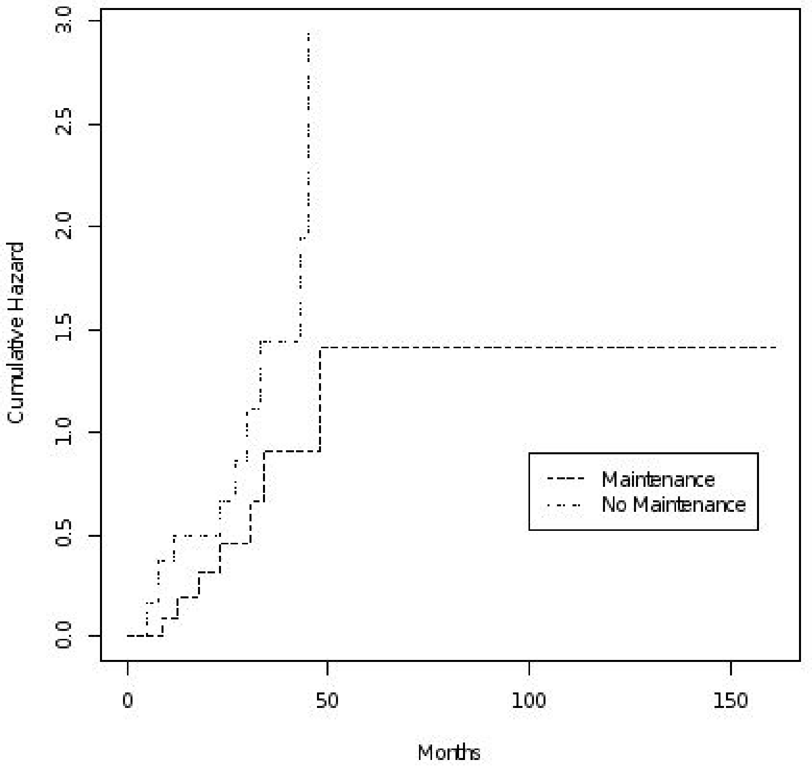
```
> leukemia.surv <- survfit(Surv(time, status) ~ x, data = aml)
> leukemia.surv
Call: survfit(formula = Surv(time, status) ~ x, data = aml)

                 n events median 0.95LCL 0.95UCL
x=Maintained    11      7     31      18      NA
x=Nonmaintained 12     11     23       8      NA

plot(leukemia.surv, lty = 2:3)
legend(100, .9, c("Maintenance", "No Maintenance"), lty = 2:3)
title("Kaplan-Meier Curves\nfor AML Maintenance Study")
```



```
> lsurv2 <- survfit(Surv(time, status) ~ x, aml, type='fleming')
> lsurv2
Call: survfit(formula = Surv(time, status) ~ x, data = aml, type = "fleming")

                 n events median 0.95LCL 0.95UCL
x=Maintained    11      7     34      18      NA
x=Nonmaintained 12     11     27       8      NA

lsurv2 <- survfit(Surv(time, status) ~ x, aml, type='fleming')
> plot(lsurv2, lty=2:3)
> legend(100, .9, c("Maintenance", "No Maintenance"), lty = 2:3)
> title("Fleming Harrington Approach")
```



```
plot(lsurv2, lty=2:3, fun="cumhaz", xlab="Months", ylab="Cumulative Hazard") # hazard function
legend(100, .9, c("Maintenance", "No Maintenance"), lty = 2:3)
```



Interpretation: In the survival first two curves above, each drop in the curve represents an event. And between events the estimated survival remains the same or constant. Absence of events means that the hazard does not exist or is zero. In both models, patients' survival time is being estimated based on their maintenance or chemotherapy status (taken or not taken). The first model leukemia.surv is fit using Kaplan-Meier approach (default method), whereas the second model or lsurv2 was fit using the Fleming-Harrington method. The two survival plots did not present substantial differences. They both indicate the same conclusion: survival rate was lowest at 20% and it increases with maintenance.

Note that survfit function fits each group (maintenance vs non-maintenance) to its own survival curve. Looking at the median survival time for each group (lsurv2), median is the preferred descriptive measure of typical survival time because it is more resistant to presence of extreme values.

The last plot was for the cumulative hazard function of the second model. Hazard is higher for patients with no maintenance, hazard increases with time, hazard plateaus for those under maintenance but goes to its highest right before the first 50 months.

Example 5: Another Non-parametric Model

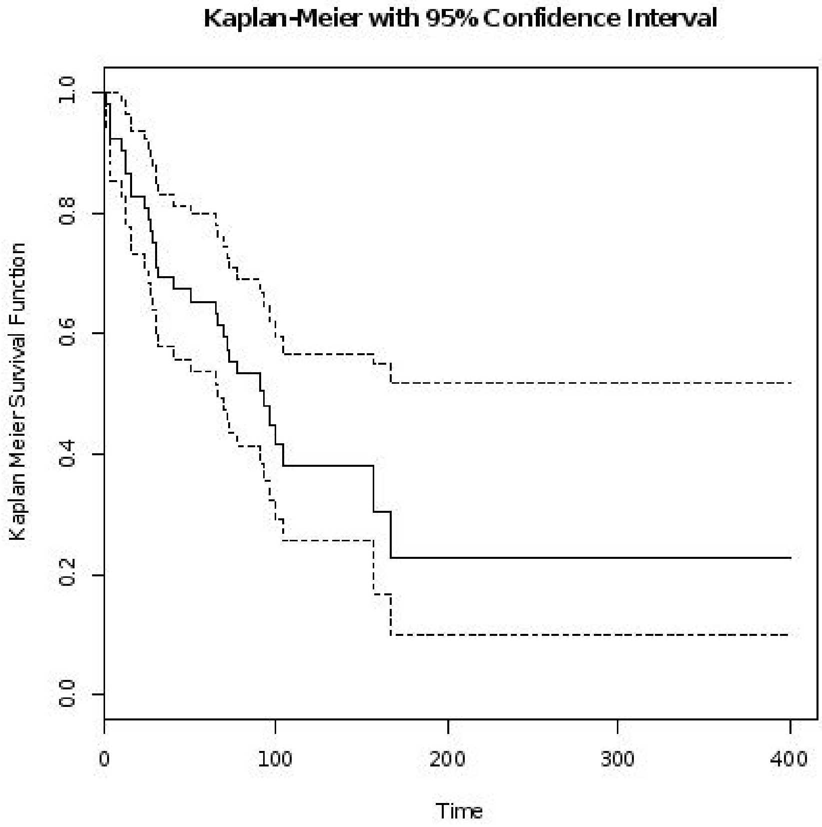
```
install.packages("OIsurv ",repos="http://cran.r-project.org") #install survival library
library(OIsurv)
data(tongue)
attach(tongue)
help(tongue)
```

Description
The tongue data frame has 80 rows and 3 columns.
Format
This data frame contains the following columns:
type
Tumor DNA profile (1=Aneuploid Tumor, 2=Diploid Tumor)
time
Time to death or on-study time, weeks
delta
Death indicator (0=alive, 1=dead)

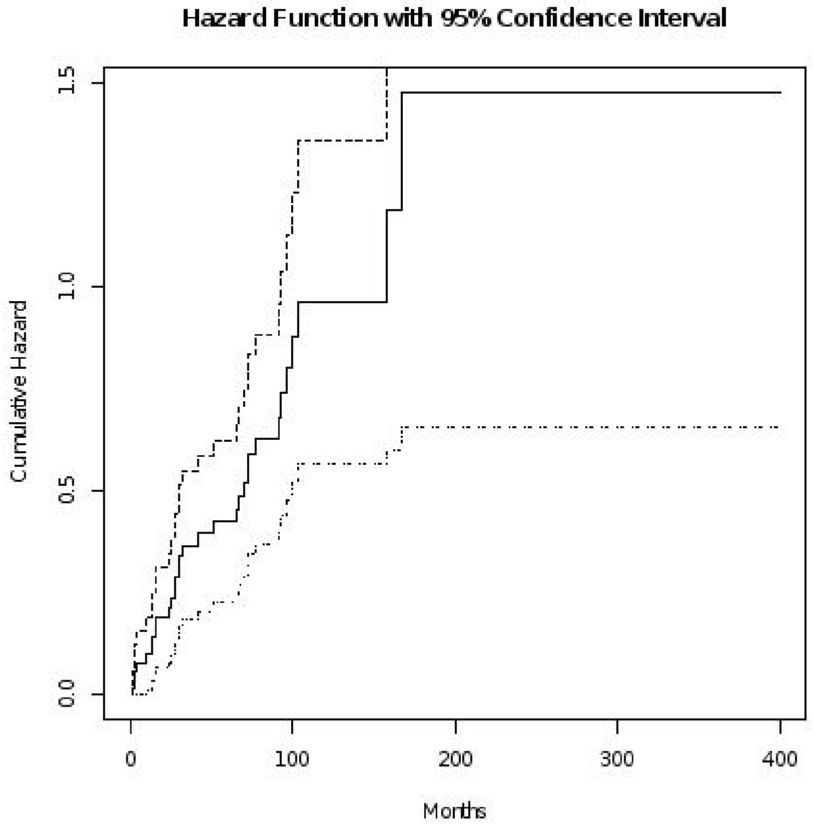
```
> sum(delta)# count dead subjects
[1] 53
```

```
mysurv <- Surv(time[type==1], delta[type==1])# Create a survival object for Aneuploid Tumor
```

```
KMfit = survfit(mysurv ~ 1) #create survival curve from mysurv
plot(KMfit, main="Kaplan-Meier with 95% Confidence Interval", xlab="Time", ylab="Kaplan Meier
Survival Function")
```



```
> KMfit$surv #  Kaplan-Meier estimates for each time
[1] 0.9807692 0.9423077 0.9230769
[4] 0.9038462 0.8653846 0.8269231
[7] 0.8076923 0.7884615 0.7692308
[10] 0.7500000 0.7115385 0.6923077
[13] 0.6730769 0.6538462 0.6538462
[16] 0.6340326 0.6142191 0.5944056
[19] 0.5745921 0.5547786 0.5547786
[22] 0.5342312 0.5342312 0.5342312
[25] 0.5342312 0.5342312 0.5342312
[28] 0.5342312 0.5061138 0.4779963
[31] 0.4481216 0.4481216 0.4161129
[34] 0.4161129 0.3814368 0.3814368
[37] 0.3814368 0.3814368 0.3814368
[40] 0.3814368 0.3051494 0.2288621
[43] 0.2288621 0.2288621 0.2288621
>

#### now let’s estimate the hazard function
plot(KMfit, lty=1:3, fun="cumhaz", xlab="Months", ylab="Cumulative Hazard") # hazard function
title("Hazard Function with 95% Confidence Interval")
```



detach(tongue)

Interpretation: The Kaplan-Meier survival plot shows that survival goes down to about 20% after about 180 weeks. The hazard function goes to about 150% at that same time. No assumptions were made the distribution of survival time variable.

Example 6: Test Survival Curve Differences

This example tests the difference between survival curves of the groups included within the data. Let’s revisit the ovarian cancer data one more time to compare the survival times based on treatment group (rx).

```
> library(survival)
Warning message:
package ‘survival’ was built under R version 3.1.3
> data(ovarian)
> head(ovarian)
  futime fustat     age resid.ds rx ecog.ps
1     59      1 72.3315        2  1       1
2    115      1 74.4932        2  1       1
3    156      1 66.4658        2  1       2
4    421      0 53.3644        2  2       1
5    431      1 50.3397        2  1       1
6    448      0 56.4301        1  1       2

> help(ovarian)
starting httpd help server ... done
```

1. **futime: survival or censoring time**
2. **fustat: censoring status**
3. **age: in years**
4. **resid.ds: residual disease present (1=no,2=yes)**
5. **rx: treatment**

```
groupecog.ps: ECOG performance status (1 is better, see reference)

> survdiff(Surv(futime, fustat) ~ rx,data=ovarian)
Call:
survdiff(formula = Surv(futime, fustat) ~ rx, data = ovarian)

       N Observed Expected (O-E)^2/E (O-E)^2/V
rx=1 13        7     5.23     0.596      1.06
rx=2 13        5     6.77     0.461      1.06

Chisq= 1.1  on 1 degrees of freedom, p= 0.303
```

Interpretation: The log-rank test (also known as Mantel-Haenzel Test) is a non-parametric test and it can be used to test the difference between two or more groups, if any. For the survdiff function, the argument rho = 0 returns the log-rank or Mantel-Haenszel test, whereas rho = 1 returns the Peto and Peto modification of the Gehan-Wilcoxon test. There are many other tests but this one does the job.

The p-value (p=0.303) of the test is greater than the significance level of 0.05. This means that the survival times between the two treatment groups are not significantly different.

Example 7: Significant Log-rank Test

This example compares the survival of lung cancer patients based on a score value (Karnofsky performance score as rated by patient).

```
library(survival)
data(lung)
head(lung)
```

```
inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
  1    3  306     2  74   1       1       90       100     1175      NA
   2    3  455     2  68   1       0       90        90     1225      15
   3    3 1010     1  56   1       0       90        90       NA      15
   4    5  210     2  57   1       1       90        60     1150      11
   5    1  883     2  60   1       0      100        90       NA       0
   6   12 1022      1  74   1       1       50        80      513       0
```

```
> help(lung)
```

Description

Survival in patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities.

Usage

lung

cancer

Format

inst:	Institution code
time:	Survival time in days
status:	censoring status 1=censored, 2=dead
age:	Age in years
sex:	Male=1 Female=2
ph.ecog:	ECOG performance score (0=good 5=dead)
ph.karno:	Karnofsky performance score (bad=0-good=100) rated by physician
pat.karno:	Karnofsky performance score as rated by patient
meal.cal:	Calories consumed at meals
wt.loss:	Weight loss in last six months

```
> survdiff(Surv(time, status) ~ pat.karno, data=lung)
Call:
survdiff(formula = Surv(time, status) ~ pat.karno, data = lung)
n=225, 3 observations deleted due to missingness.
                N Observed Expected (O-E)^2/E (O-E)^2/V
pat.karno=30   2        1    0.658    0.1774     0.179
pat.karno=40   2        1    1.337    0.0847     0.086
pat.karno=50   4        4    1.079    7.9088     8.013
pat.karno=60  30       27   15.237    9.0808    10.148
pat.karno=70  41       31   26.264    0.8540     1.027
pat.karno=80  51       39   40.881    0.0865     0.117
pat.karno=90  60       38   49.411    2.6354     3.853
pat.karno=100 35       21   27.133    1.3863     1.684

Chisq= 22.6  on 7 degrees of freedom, p= 0.00202
```

Interpretation: The p-value (p=0.00202) of the test is less than the significance level of 0.05. This means that the difference in survival times between the groups is statistically significant. Looking at the observed and expected survival times, looks like the groups with Karnofsky performance score of 80 and 90 has the highest observed survival time and highest expected survival time, respectively. The group with highest expected estimate is the one with highest survival (49.411, pat.karno=90).

Example 8: Survival by Group

This example differentiates between gender groups by plotting as well as conducting the log-rank test, using the melanoma data from the MASS library.

```
library(survival); library(MASS)
data(Melanoma)
head(Melanoma)
```

```
  time status sex age year thickness ulcer
1   10      3   1  76 1972      6.76     1
2   30      3   1  56 1968      0.65     0
3   35      2   1  41 1977      1.34     0
4   99      3   0  71 1968      2.90     0
5  185      1   1  52 1965     12.08     1
6  204      1   1  28 1971      4.84     1
```

```
help(Melanoma)
```

Description

The Melanoma data frame has data on 205 patients in Denmark with malignant melanoma.

Usage

Melanoma

Format

This data frame contains the following columns:

time: survival time in days, possibly censored.

status: 1 died from melanoma, 2 alive, 3 dead from other causes.

sex: 1 = male, 0 = female.

age: age in years.

year: of operation.

thickness: tumor thickness in mm.

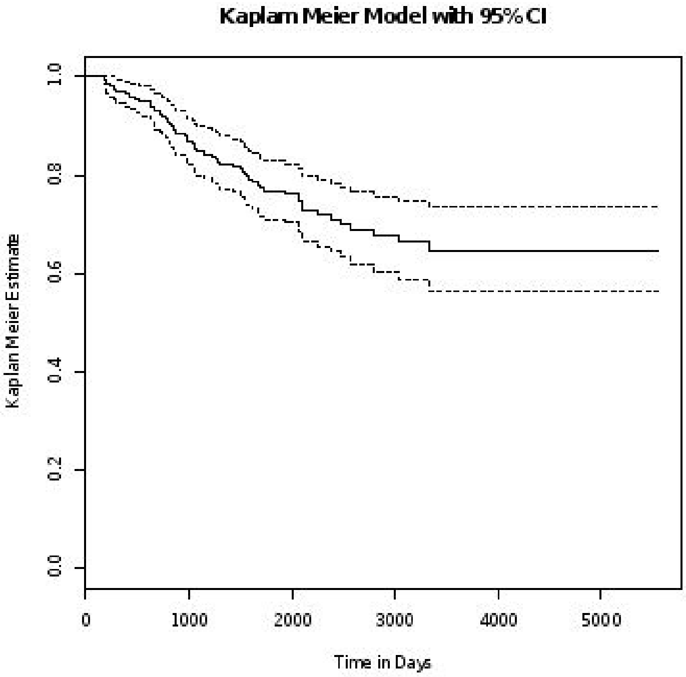
ulcer: 1 = presence, 0 = absence.

```
summary(Melanoma$time) # summarize days of survival
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     10    1525    2005    2153    3042    5565 
```

```
KaplanMfit <- survfit(Surv(time, status == 1) ~ 1, data = Melanoma) # fit Kaplan Meier model
summary(KaplanMfit,  times = seq(0, 6000, 1000)) # summarize the model with respect to time in 1000-day intervals
```

```
Call: survfit(formula = Surv(time, status == 1) ~ 1, data = Melanoma)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    0    205       0    1.000  0.0000        0.000        0.000
 1000    171      26    0.869  0.0240        0.823        0.917
 2000    103      20    0.762  0.0308        0.704        0.825
 3000     54       9    0.677  0.0385        0.605        0.757
 4000     13       2    0.645  0.0431        0.566        0.735
 5000      1       0    0.645  0.0431        0.566        0.735
```

```
> plot(KaplanMfit, ylab="Kaplan Meier Estimate", xlab="Time in Days")
> title ("Kaplam Meier Model with 95% CI") #CI = confidence interval
```



```
KM.sex <- survfit(Surv(time, status == 1) ~ sex, data = Melanoma)
summary(KM.sex, times = seq(0, 6000, 1000))
```

```
Call: survfit(formula = Surv(time, status == 1) ~ sex, data = Melanoma)

                sex=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    0    126       0    1.000  0.0000        0.000        0.000
 1000    111      11    0.910  0.0258        0.861        0.962
 2000     68      11    0.813  0.0362        0.745        0.887
 3000     36       4    0.755  0.0439        0.674        0.846
 4000      9       2    0.704  0.0542        0.605        0.818
 5000      1       0    0.704  0.0542        0.605        0.818

                sex=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    0    126       0    1.000  0.0000        0.000        0.000
 1000     60      15    0.801  0.0461        0.715        0.896
 2000     35       9    0.677  0.0544        0.578        0.792
 3000     18       5    0.553  0.0675        0.435        0.702
 4000      4       0    0.553  0.0675        0.435        0.702
```

```
> survdiff(Surv(time, status == 1) ~ sex, data = Melanoma)
Call:
survdiff(formula = Surv(time, status == 1) ~ sex, data = Melanoma)
        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=0 126       28     37.1      2.25      6.47
sex=1  79       29     19.9      4.21      6.47
```

Interpretation: The plot shows that after 3,500 days, little over 60% of the patients survived the disease. The second model (KM.sex) is for those who died from melanoma and uses sex as the independent variable or predictor. Looking at the estimators from summary(KM.sex), it can be seen that females had higher survival rate (sex = 0). This is also shown in the log-rank test result where females had 37.1 expected survival rate as opposed to 19.9 for males.

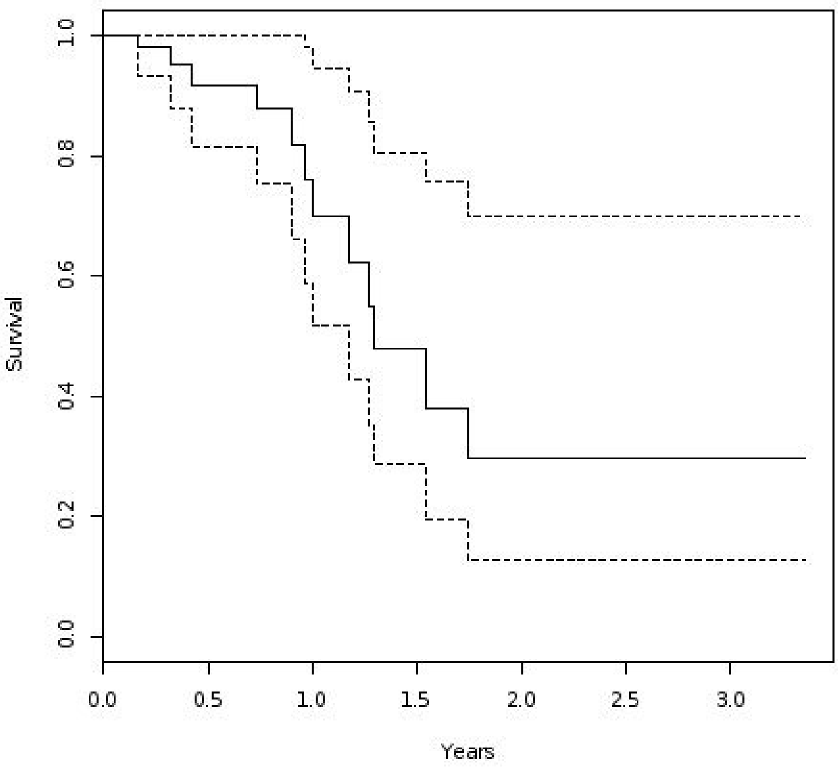
Chapter 4: Semi-parametric Approach

Non-parametric and semi-parametric methods are more flexible than parametric ones because they allow us to estimate a model without knowing the distribution of survival times. The most famous semi-parametric survival model is Cox proportional hazard regression model. It is semi-parametric because it has parametric and non-parametric parts. The parametric part is the assumption that the survival time follows a certain distribution but there is no assumption on the shape of the hazard function.

Example 9: Cox Proportional Hazard Model

The following Cox model example is from the survival package documentation. It fits a Cox proportional hazards model for those who were 60 years old.

```
Library(survival)
fit <- coxph(Surv(futime, fustat) ~ age, data = ovarian)
plot(survfit(fit, newdata=data.frame(age=60)),
    xscale=365.25, xlab = "Years", ylab="Survival")
```



```
> fit$var # variance-covariance matrix
           [,1]
[1,] 0.00247408
> sqrt(diag(fit$var))
[1] 0.04974012
```

```
> summary(fit)
Call:
coxph(formula = Surv(futime, fustat) ~ age, data = ovarian)

  n= 26, number of events= 12

       coef exp(coef) se(coef)     z Pr(>|z|)
age 0.16162   1.17541  0.04974 3.249  0.00116 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .95 upper .95
age     1.175     0.8508     1.066     1.296

Concordance= 0.784  (se = 0.091 )
Rsquare= 0.423   (max possible= 0.932 )
Likelihood ratio test= 14.29  on 1 df,   p=0.0001564
Wald test            = 10.56  on 1 df,   p=0.001157
Score (logrank) test = 12.26  on 1 df,   p=0.0004629
```

Interpretation: The Cox Proportional Hazard Model named fit gives the hazard ratio for a one unit change in the predictor (age). The survival plot shows that survival goes down to about 30% after about 1.75 years. The model indicates that age is a significant predictor of a patient’s survival (p-value 0.00116 < 0.05).

Example 10: Stratified Cox Model

Cox model can be stratified in order to study the hazard function across the different levels of the stratification variable.

```
> library(survival)
> help(ovarian)
starting httpd help server ... done
```

Description
Survival in a randomised trial comparing two treatments for ovarian cancer
Usage
ovarian
Format
futime: survival or censoring time
fustat: censoring status
age: in years
resid.ds: residual disease present (1=no,2=yes)
rx: treatment
groupecog.ps: ECOG performance status (1 is better, see reference)

```
> coxph(Surv(futime, fustat)~rx+strata(age>60), ovarian) # first model using stratified Cox
Call:
coxph(formula = Surv(futime, fustat) ~ rx + strata(age > 60),
    data = ovarian)

    coef exp(coef) se(coef)     z    p
rx -0.43      0.65     0.60 -0.72 0.47

Likelihood ratio test=0.52  on 1 df, p=0.471
n= 26, number of events= 12
```

```
> coxph(Surv(futime, fustat)~rx+(age>60), ovarian) # second model, no stratification
Call:
coxph(formula = Surv(futime, fustat) ~ rx + (age > 60), data = ovarian)

                coef exp(coef) se(coef)     z       p
rx            -0.273     0.761    0.613 -0.44 0.65688
age > 60TRUE  2.258     9.560    0.685  3.29 0.00099

Likelihood ratio test=11.8  on 2 df, p=0.00281
n= 26, number of events= 12
```

<u>Interpretation:</u> The first model fits a Cox Proportional Hazard Model where survival time is estimated using treatment (rx) for a sample that is stratified/split into two strata: women over the age of 60 and women who are 60 or younger. The second model uses two predictors: treatment and age above 60. That is it excludes women who are 60 or younger.

Chapter 5: Model Assessment

Goodness of fit for statistical models investigates how well your model fits the used data. There are many ways to measure the goodness of fit for statistical models including but not limited to hypothesis testing (e.g. test the residuals follow a normal distribution), analysis of variance (anova), or using the Akaike **I**nformation **C**riterion (AIC).

Example 11: AIC

```
> model1 = coxph(Surv(futime, fustat) ~ age, data = ovarian)
> model2 = coxph(Surv(futime, fustat) ~ rx, data = ovarian)
> model3 = coxph(Surv(futime, fustat) ~ rx + age, data = ovarian)
> extractAIC(model1) # gives degrees of freedom & AIC
[1]  1.00000 57.67629

> extractAIC(model2)
[1]  1.00000 70.91843

> extractAIC(model3)
[1]  2.0000 58.0838
```

Interpretation: Here we fit three models for the ovarian cancer data. The one with the least AIC is the best model, which is model2 in this case. This indicates that treatment is a better predictor of the subjects survival than their age or having both variables. AIC is nonparametric (i.e.; no assumption on the distribution of the data).

Example 12: ANOVA

```
library(survival) # load survival library
> data(ovarian)
> dim(ovarian) # Ovarian data has 26 rows and 6 columns
[1] 26  6
> help(ovarian)
starting httpd help server ... done
```

Description

Survival in a randomised trial comparing two treatments for ovarian cancer

Usage

ovarian

Format

futime: survival or censoring time

fustat: censoring status

age: in years

resid.ds: residual disease present (1=no,2=yes)

rx: treatment

groupecog.ps: ECOG performance status (1 is better, see reference)

```
> fit1 <- coxph(Surv(futime, fustat) ~ rx, data = ovarian)

> anova(fit1)
Analysis of Deviance Table
Cox model: response is Surv(futime, fustat)
Terms added sequentially (first to last)

      loglik  Chisq Df Pr(>|Chi|)
NULL -34.985
rx   -34.459 1.0515  1     0.3052

fit2 <- coxph(Surv(futime, fustat) ~ rx+age,data=ovarian)

> anova(fit1,fit2)
Analysis of Deviance Table
Cox model: response is  Surv(futime, fustat)
Model 1: ~ rx
Model 2: ~ rx + age
   loglik  Chisq Df P(>|Chi|)
1 -34.459
2 -27.042 14.835  1 0.0001174 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation: The first anova (anova(fit1)) tests the significance of having treatment (rx) as a predictor, whereas the second one (anova(fit2,fit1)) compares the two models against each other in the order specified. The p-value of this comparison which is an F-test is 0.0001174, which is less than the significance level of 0.05 which means that it is worth it to add age as a predictor (as well as treatment).

The End