DSCI 610 Exploring Time-to-Event Data

Required packages

```r
library("tidyverse")
```

```
## -- Attaching packages -------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.3
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ----------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library("survival")
library("ggfortify")
```

Set Working Directory

```r
setwd("~/Box/MyDocs/Teaching/Spring/2021/DSCI 610/LectureMaterials/Week 12/Lecture")
```

Product-Limit (Kaplan-Meier) Estimator of Survival function

Let $t_1, t_2, \ldots, t_k$ be the distinct ordered failure times in the sample. For each $t_i, i = 1, 2, \ldots k$, let D_i be the number of events (deaths) at time t_i, S_i be the number that are known to have survived beyond t_i, and N_i be the number `at risk` of being observed to die at time t_i.

The Product-Limit (Kaplan-Meier) estimator of the survival function is given by the following formula:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{D_i}{N_i}\right) = \prod_{i:t_i \leq t} \frac{S_i}{N_i}$$

The standard error estimator for the product limit estimator is given by the Greenwood's formula as follows:

$$\hat{se}(\hat{S}(t)) = \hat{S}(t)\sqrt{\sum_{i:t_i \leq t} \frac{D_i}{N_i S_i}}.$$

The $(1-\alpha)100\%$ confidence interval (CI) for $S(t)$ is given by:

$$\hat{S}(t) - z_{1-\alpha/2}\hat{se}(\hat{S}(t)), \hat{S}(t) + z_{1-\alpha/2}\hat{se}(\hat{S}(t))$$

Note the above CI for $S(t)$ can contain values < 0 or > 1 at extreme values of t. You can truncate or avoid the problem by applying the normal approximation to a transformation of $S(t)$ for which the range is unrestricted.

The asymptotic variance of $\hat{v}(t) = \log(-\log \hat{S}(t))$ is estimated by:

$$\hat{se}(\hat{v}(t)) = \frac{\sqrt{\sum_{i:t_i \le t} \frac{D_i}{N_i S_i}}}{\log \hat{S}(t)}$$

The 95% CI for $v(t) = \log(-\log S(t))$: $\hat{v}(t) \pm 1.96\hat{se}(\hat{v}(t))$.

The 95% CI for $S(t)$: $\hat{S}(t)^{[\exp\{\pm 1.96 \hat{se}(\hat{v}(t))\}]}$

Example 1: The 6-MP versus placebo clinical trial in acute leukemia (Cox and Oakes(1984))

A randomized, double-blind, placebo controlled sequential study was conducted in which 42 leukaemia patients were paired by remission status at each of the eleven institutions participating in the study, and randomized to 6-MP or placebo within each pair of patients.

The dataset has the following variables:

Time: remission time in weeks

delta: censoring indicator

Group: treatment indicator, 0 for placebo and 1 for 6-MP

```
# Survival functions for time to remission (weeks) of leukaemia patients
temp <- read.csv("6-mp.csv", header=T)
head(temp)
```

```
##   Time delta Group
## 1    1     1     0
## 2    1     1     0
## 3    2     1     0
## 4    2     1     0
## 5    3     1     0
## 6    4     1     0
```

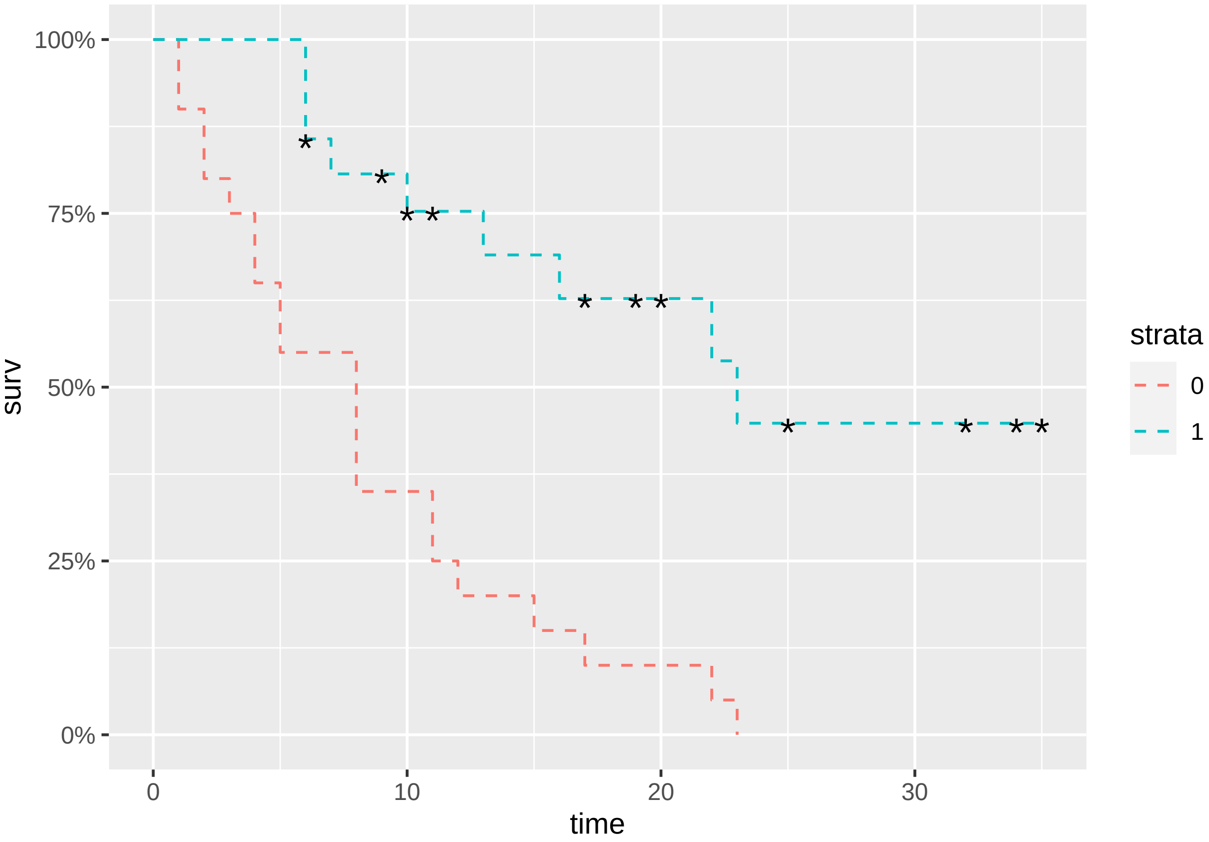
```
time <- temp[,1] # survival time in weeks
delta <- temp[,2] # censoring indicator, 0 = censored, 1 = failed
group <- temp[,3] # treatment groups, 0 = control, 1 = 6-MP drug

fit <- survfit(Surv(time, delta)~group,conf.type="log-log")
```

```r
summary(fit)
```

```
## Call: survfit(formula = Surv(time, delta) ~ group, conf.type = "log-log")
## 
##                 group=0 
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     20       2     0.90  0.0671      0.65603        0.974
##     2     18       2     0.80  0.0894      0.55115        0.920
##     3     16       1     0.75  0.0968      0.49994        0.887
##     4     15       2     0.65  0.1067      0.40300        0.815
##     5     13       2     0.55  0.1112      0.31340        0.735
##     8     11       4     0.35  0.1067      0.15656        0.552
##    11      7       2     0.25  0.0968      0.09099        0.449
##    12      5       1     0.20  0.0894      0.06238        0.393
##    15      4       1     0.15  0.0798      0.03733        0.335
##    17      3       1     0.10  0.0671      0.01698        0.272
##    22      2       1     0.05  0.0487      0.00345        0.205
##    23      1       1     0.00     NaN           NA           NA
## 
##                 group=1 
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     21       3    0.857  0.0764        0.620        0.952
##     7     17       1    0.807  0.0869        0.563        0.923
##    10     15       1    0.753  0.0963        0.503        0.889
##    13     12       1    0.690  0.1068        0.432        0.849
##    16     11       1    0.627  0.1141        0.368        0.805
##    22      7       1    0.538  0.1282        0.268        0.747
##    23      6       1    0.448  0.1346        0.188        0.680
```

```r
autoplot(fit, surv.linetype = 'dashed', censor.shape = '*', conf.int = FALSE, censor.size = 5)
```



Example 2: Comparison of two treatments for prostatic cancer (Andrews and Herzberg, 1985)

A randomized controlled clinical trial to compare treatments for prostatic cancer was begun in 1967 by the Veteran's Administration Cooperative Urological Research Group. The trial was double-blind and two of the treatments used in the study were a placebo and 1.0 mg of diethylstilbestrol (DES).

The treatments were administered daily by mouth. The time origin of the study is the date on which a patient was randomized to a treatment, and the end-point is the death of the patient from prostatic cancer.

The dataset has patients with Stage III cancer, that is patients for whom there was evidence of a local extension of the tumor beyond the prostatic capsule but without elevated serum prostatic acid phosphatase.

In addition to recording the survival time of each patient in the study, information was recorded on a number of other prognostic factors. These included the age of the patient at trial entry, their serum haemoglobin level in gm/100 ml, the size of their primary tumor in cm^2 and the value of a combined index of tumor stage and grade. This index is known as `Gleason index`. The more advanced the tumor, the greater the value of the index.

```
pr.cancer = read.table("prostatic_cancer.dat", header=T)
head(pr.cancer)
```

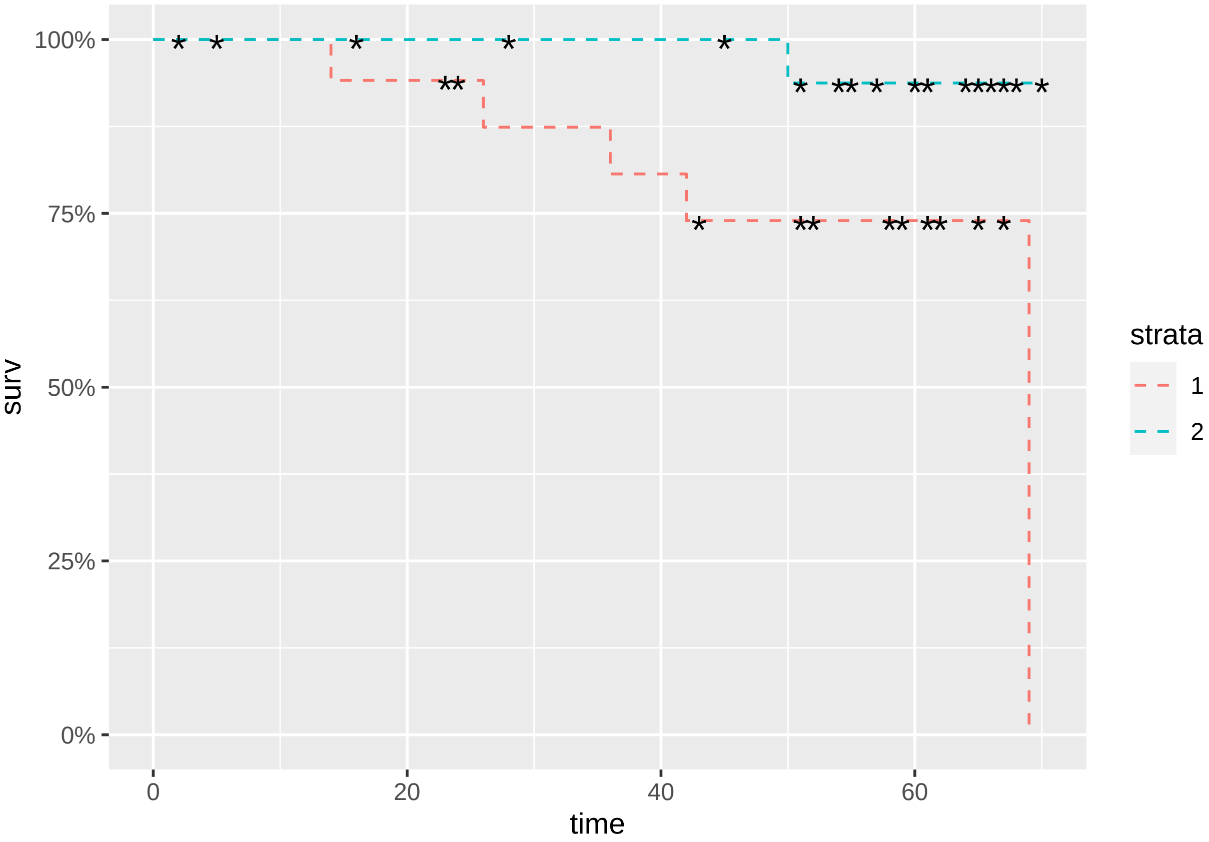
```
##   patient treatment time status age  shb size index
## 1       1         1   65       0   67 13.4   34     8
## 2       2         2   61       0   60 14.6    4    10
## 3       3         2   60       0   77 15.6    3     8
## 4       4         1   58       0   64 16.2    6     9
## 5       5         2   51       0   65 14.1   21     9
## 6       6         1   51       0   61 13.5    8     8
```

```
fit.1 <- survfit(Surv(time, status)~treatment,conf.type="log-log", data = pr.cancer)

summary(fit.1)
```

```
## Call: survfit(formula = Surv(time, status) ~ treatment, data = pr.cancer,
##     conf.type = "log-log")
##
##                 treatment=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    14     17       1    0.941  0.0571        0.650        0.991
##    26     14       1    0.874  0.0837        0.581        0.967
##    36     13       1    0.807  0.1007        0.511        0.934
##    42     12       1    0.739  0.1125        0.445        0.894
##    69      1       1    0.000     NaN           NA           NA
##
##                 treatment=2
##         time       n.risk      n.event     survival      std.err lower 95% CI
##      50.0000      16.0000       1.0000       0.9375       0.0605       0.6323
## upper 95% CI
##       0.9910
```

```
autoplot(fit.1, surv.linetype = 'dashed', censor.shape = '*', conf.int = FALSE, censor.size = 5)
```



The survival function plots for two groups provide a simple approach for comparing survival experience of two groups of patients. In order to compare two groups of survival data, we can consider the non-parametric log-rank test.

The Logrank Test

Given two sample data as follows, we want to compare two survival functions.

$$(X_{1i}, \delta_{1i}), \ \ \text{for} \ \ i = 1, 2, \dots n_1 \ \ iid \ \ T_{1i} \sim S_1$$

$$(X_{2i}, \delta_{2i}), \ \ \text{for} \ \ i = 1, 2, \dots n_2 \ \ iid \ \ T_{2i} \sim S_2$$

The hypothesis of interest is:

$$H_0 : S_1(t) = S_2(t), \quad \text{versus} \quad H_A : S_1(t) \neq S_2(t)$$

Example 3: Comparison of survival functions for the patients on 6-MP versus placebo

```
survdiff(Surv(Time, delta)~Group, rho=0, data=temp)
```

```
## Call:
## survdiff(formula = Surv(Time, delta) ~ Group, data = temp, rho = 0)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## Group=0 20       20     9.99     10.04      16.8
## Group=1 21        9    19.01      5.27      16.8
##
##  Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

Example 2: Comparison of survival functions for the prostatic patients on DES versus placebo

```
survdiff(Surv(time,status)~treatment,rho=0, data = pr.cancer)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ treatment, data = pr.cancer,
##     rho = 0)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## treatment=1 18        5     2.47      2.58      4.42
## treatment=2 20        1     3.53      1.81      4.42
##
##  Chisq= 4.4  on 1 degrees of freedom, p= 0.04
```