Statistics with R

Survival Analysis

Scott Hetzel
University of Wisconsin – Madison
Summer Institute for Training in Biostatistics (2009)

Derived from: “Introductory Statistics with R” by: Peter Dalgaard

and from previous notes by Deepayan Sarkar, Ph.D

Survivial Analysis in R

- Tools are available in the package `survival`
- This is a recommended package, which means it should already be installed
- It has to be loaded using
  ```
  > library(survival)
  ```
- Survival Analysis is covered in Chapter 12 of the text

Functions of Interest

- Create a survival object: `Surv`
- Kaplan-Meier Estimates: `survfit`
- The log-rank test: `survdiff`
- The Cox proportional hazards model: `coxph`
  (we won't be discussing this)

Survival Objects

- Created by the `Surv` function
- Needs two arguments:
  - `time`: follow-up time
  - `event`: status indicator
- `event=TRUE` means event occured
- `event=FALSE` indicates censoring
- Other values possible (see `help(Surv)`)

Example: melanom

We will use the example from the text:

```
> library(ISwR)
> str(melanom)

'data.frame':  205 obs.  of 6 variables:
$ no        :  int 789 13 97 16 21 469 685 7 932 944 ...
$ status    :  int 3 3 2 3 1 1 1 1 3 1 ...
$ days      :  int 10 30 35 99 185 204 210 232 232 279 ...
$ ulc       :  int 1 2 2 2 1 1 1 1 1 1 ...
$ thick     :  int 676 65 134 290 1208 484 516 1288 322 741 ...
$ sex       :  int 2 2 2 1 2 2 2 2 1 1 ...
```

We are interested in:

- `days`: time on study after operation for malignant melanoma
- `status`: the patient's status at the end of study

Censoring Indicator

- The possible values of `status` are
  - 1: dead from malignant melanoma
  - 2: alive at the end of the study
  - 3: dead from other causes

- `Surv` needs a logical status indicator
  (`TRUE` if event occurred, `FALSE` if censored)

- Let's consider "dead from other causes" as censored

- Thus, status vector should be `status == 1`

Creating the Survival Object

```
> msurv <- with(melanom, Surv(days, status==1))
> msurv
  [1]    10+    30+    35+    99+   185   204   210   232   232+   279   295
 [12]   355+   386   426   469   493+   529   621   629   659   667   718
 [23]   752   779   793   817   826+   833   858   869   872   967   977
 [34]   982  1041  1055  1062  1075  1156  1228  1252  1271  1312  1427+
 [45]  1435  1499+  1506  1508+  1510+  1512+  1516  1525+  1542+  1548  1557+
 [56]  1560  1563+  1584  1605+  1621  1627+  1634+  1641+  1641+  1648+  1652+
 [67]  1654+  1654+  1667  1678+  1685+  1690  1710+  1710+  1726  1745+  1762+
 [78]  1779+  1787+  1787+  1793+  1804+  1812+  1836+  1839+  1839+  1854+  1856+
 [89]  1860+  1864+  1899+  1914+  1919+  1920+  1927+  1933  1942+  1955+  1956+
[100]  1958+  1963+  1970+  2005+  2007+  2011+  2024+  2028+  2038+  2056+  2059+
[111]  2061  2062  2075+  2085+  2102+  2103  2104+  2108  2112+  2150+  2156+
[122]  2165+  2209+  2227+  2227+  2256  2264+  2339+  2361+  2387+  2388  2403+
[133]  2426+  2426+  2431+  2460+  2467  2492+  2493+  2521+  2542+  2559+  2565
[144]  2570+  2660+  2666+  2676+  2738+  2782  2787+  2984+  3032+  3040+  3042
[155]  3067+  3079+  3101+  3144+  3152+  3154+  3180+  3182+  3185+  3199+  3228+
[166]  3229+  3278+  3297+  3328+  3330+  3338  3383+  3384+  3385+  3388+  3402+
[177]  3441+  3458+  3459+  3459+  3476+  3523+  3667+  3695+  3695+  3776+  3776+
[188]  3830+  3856+  3872+  3909+  3968+  4001+  4103+  4119+  4124+  4207+  4310+
[199]  4390+  4479+  4492+  4668+  4688+  4926+  5565+
```

The print method for `Surv` objects marks censored observations with a '+' sign after the time. For example 10+ means the patient did not die from melanoma within ten days and was then unavailable for further study. Whereas 185 means that the patient died from melanoma 185 days after the operation.

Operations on the Survival Object

- Not very useful in isolation

- Typically used in other functions

- Caution if trying to find the mean of a survival object. The survival object is saved as a matrix with two columns: one for time and one for status. Trying `mean(msurv)` will give the mean of the whole matrix not just the times which is probably what you really want.

```
> mean(msurv)
[1] 1076.539
```

Use indexing to get the correct mean.

```
> mean(msurv[,1])
[1] 2152.8
```

Check `summary(msurv)` to verify.

The Kaplan-Meier Estimator

- Computed by the function `survfit`
- Simplest case: just needs the survival object
- Note the use of the `data` argument below

```
> mfit <- survfit(Surv(days, status == 1), data = melanom)
> mfit
Call:  survfit(formula = Surv(days, status == 1), data = melanom)
   n  events  median  0.95LCL  0.95UCL
 205      57     Inf      Inf      Inf
```

Notice how the simple print of the `survfit` object does not give much information. In this case the estimate of the median survival is infinite because the survival curve does not reach the 50% line before the end of the study.

The Kaplan-Meier Estimator (Cont)

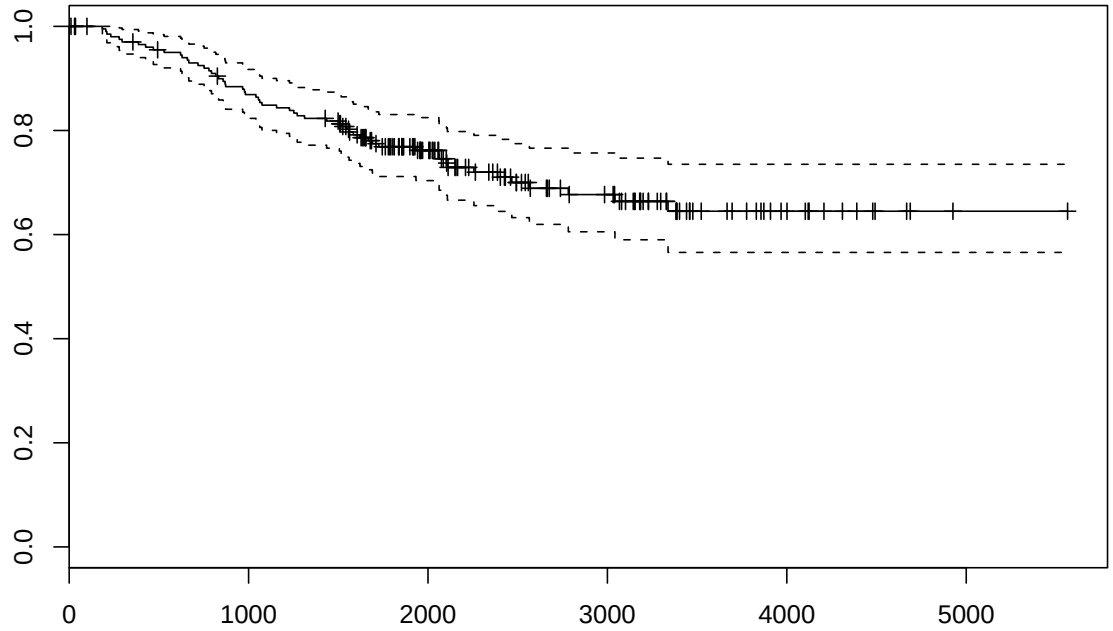
- The `summary` method actually produces the values of *S*
- By default, values of *S* at all event times are listed

```
> summary(mfit, times=seq(185, 3000, 400))
Call:  survfit(formula = Surv(days, status == 1), data = melanom)
 time  n.risk  n.event  survival  std.err  lower 95% CI  upper 95% CI
  185     201        1     0.995  0.00496         0.985         1.000
  585     188        9     0.950  0.01542         0.920         0.981
  985     171       16     0.869  0.02397         0.823         0.917
 1385     162        9     0.823  0.02713         0.772         0.878
 1785     127       10     0.769  0.03033         0.712         0.831
 2185      83        5     0.729  0.03358         0.666         0.798
 2585      61        4     0.689  0.03729         0.620         0.766
 2985      54        1     0.677  0.03854         0.605         0.757
```

The Kaplan-Meier Estimator (Cont)

Naturally the `plot` function plots the estimated survival curve.

```
> plot(mfit)
```



Comparing Survival Curves

- Things get intesting when there are two or more groups to compare

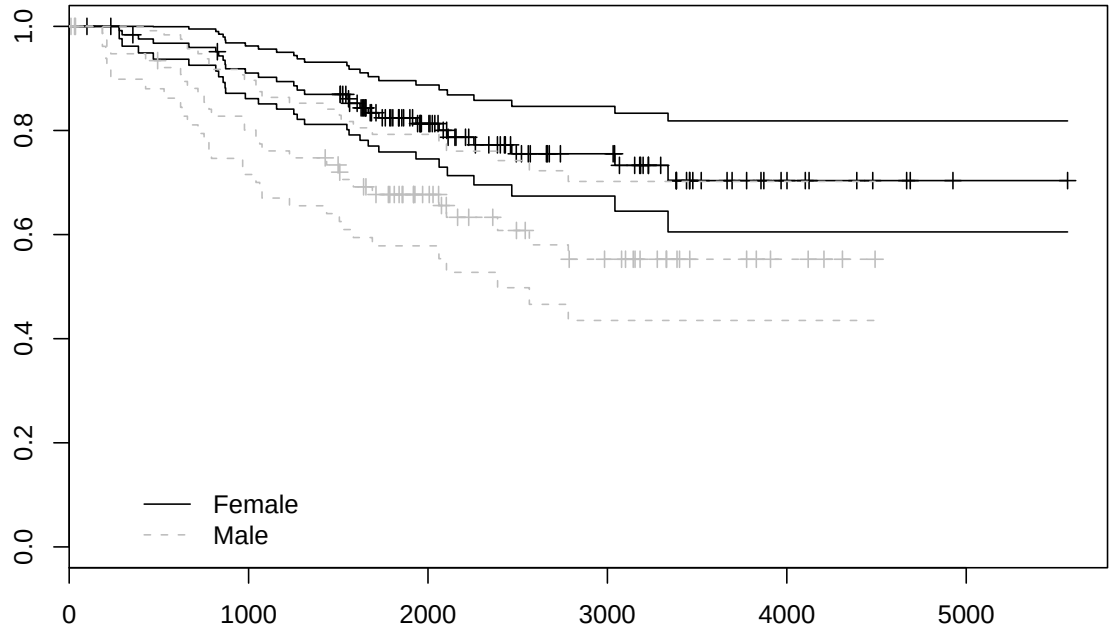
- For example, does survival differ in men and women?

```
> fit.bysex <- survfit(Surv(days, status == 1) ~ sex, data=melanom)
> fit.bysex
Call:  survfit(formula = Surv(days, status == 1) ~  sex, data
= melanom)

         n  events  median  0.95LCL  0.95UCL
 sex=1  126      28     Inf      Inf      Inf
 sex=2   79      29     Inf     2388      Inf
```

Comparing Survival Curves (Cont)

```
> plot(fit.bysex,conf.int=TRUE, col=c("black","grey"),
lty=1:2,
legend.text=c("Female","Male"))
```



Comparing Survival Curves (Cont)

The function `survdiff`, formally tests for differences between groups.

```
> survdiff(Surv(days, status==1) ~ sex, data=melanom)
Call:
survdiff(formula = Surv(days, status == 1) ~ sex, data=melanom)

          N  Observed  Expected  (O-E)^2/E  (O-E)^2/V
 sex=1  126        28      37.1       2.25       6.47
 sex=2   79        29      19.9       4.21       6.47

Chisq= 6.5 on 1 degrees of freedom, p= 0.011
```

Exercises in Using R

An intestigator collected data on survival of patients with lung cancer at Mayo Clinic. The investigator would like you, the statistician, to answer the following questions and provide some graphs. The data is located in the survival package under the name: `cancer`.

1. What is the probability that someone will survive past 300 days?

2. Provide a graph, including 95% confidence limits, of the Kaplan-Meier estimate for the entire study.

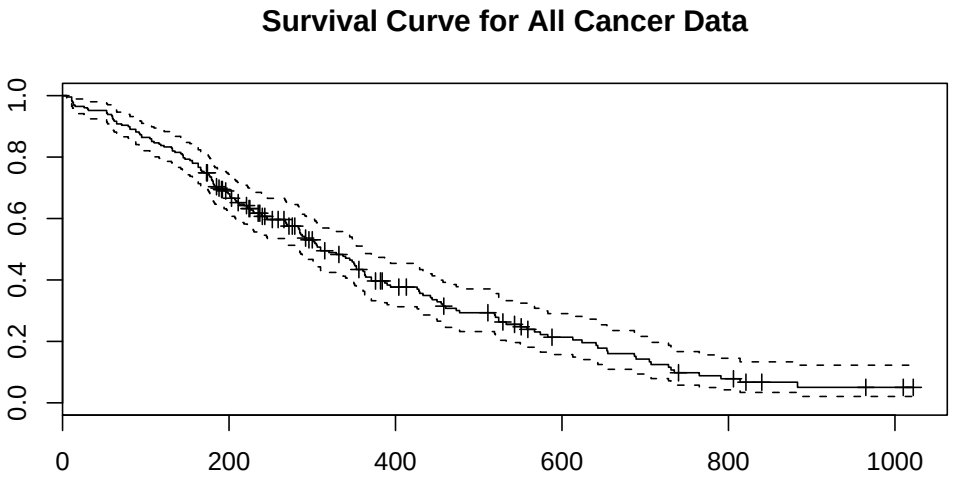
3. Is there a difference in the survival rates between males and females? Provide a formal statistical test with p-value and visual evidence.

4. Is there a difference in the survival rates for the older half of the group versus the younger half? Provide a formal statistical test with p-value and visual evidence.

Exercises in Using R Answers

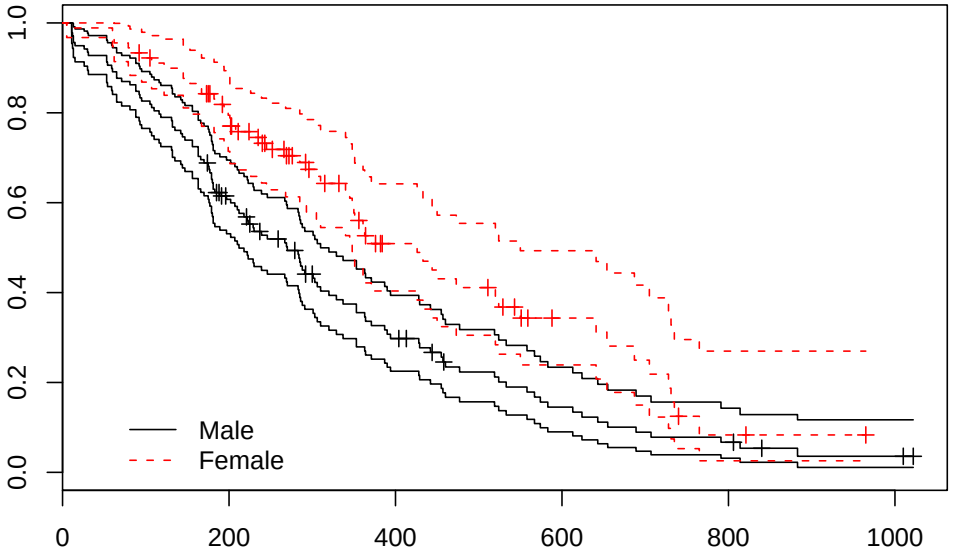
1.
```
> attach(cancer)
> surv.can <- Surv(time, status == 2)
> fit.can <- survfit(surv.can)
> summary(fit.can, time=300)$surv
[1] 0.5306081
```

2.
```
> plot(fit.can, main="Survival Curve for All Cancer Data")
```



Exercises in Using R Answers

3.
```
> can.bysex <- survfit(surv.can ~ sex)
> survdiff(surv.can ~ sex) # See in output a p-value of 0.00131
> plot(can.bysex, conf.int=TRUE, col=c("black", "red"),
lty=1:2, legend.text=c("Male", "Female"))
```



Exercises in Using R Answers

4.
```
> median(age)
[1] 63
> can.byage <- survfit(surv.can ~ age>63)
> survdiff(surv.can ~ age>63) # See in output a p-value of 0.17
> plot(can.byage, conf.int=TRUE, col=c("orangered", "blue"),
lty=c(4,5), legend.text=c("Age <= 63", "Age > 63"))
```