Lab 9 Part 1 - Survival analysis

Isabel Sassoon

Survival Analysis

This notebook walks through the analysis of the lung data from the survival library.

```
library(survival)
```

The lung data has many columns

```
help("lung")
```

In order to focus I subset the data into a smaller set of columns:

```
lung.analysis<-subset(lung, select = c("time", "status", "sex"))
```

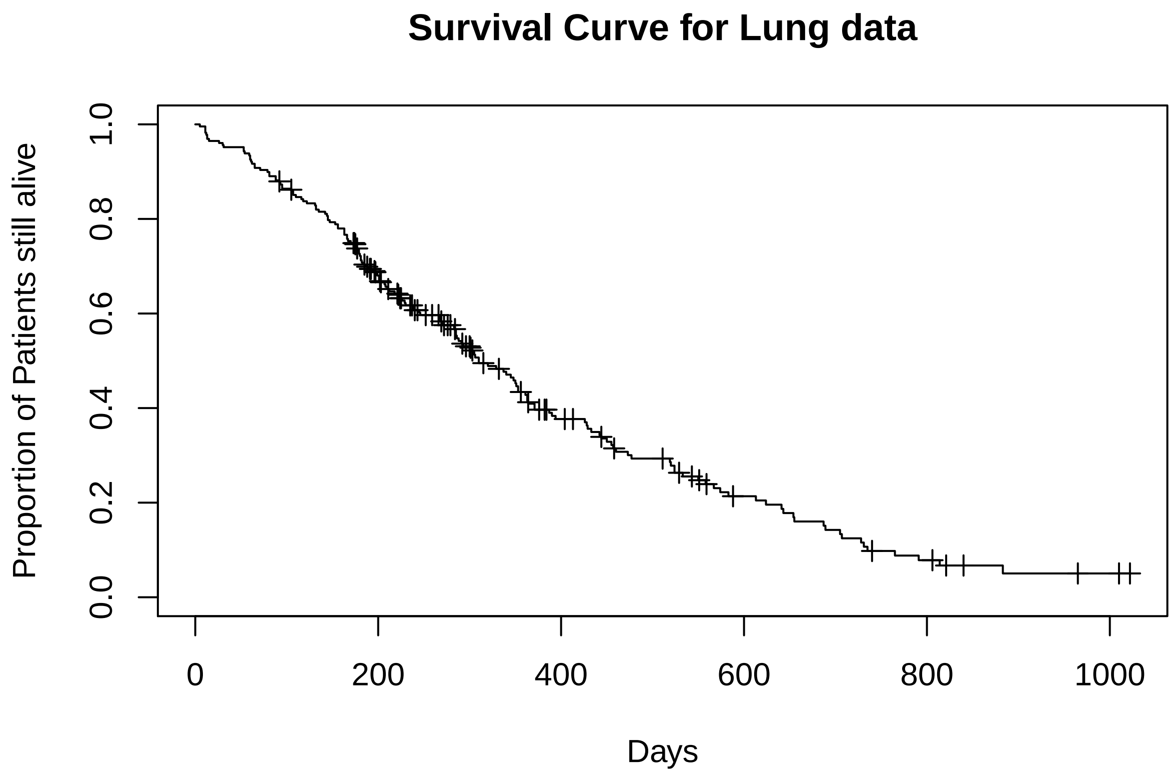
To estimate the survival curve using KM

```
s.survfit <- survfit(Surv(lung.analysis$time, lung.analysis$status)~ 1, data=lung.analysis )
s.survfit
```

```
## Call: survfit(formula = Surv(lung.analysis$time, lung.analysis$status) ~
##     1, data = lung.analysis)
##
##       n  events  median 0.95LCL 0.95UCL
##     228     165     310     285     363
```

and to plot it:

```
plot(s.survfit, mark.time = TRUE, conf.int = FALSE)
title(main="Survival Curve for Lung data", xlab="Days", ylab = "Proportion of Patients still alive")
```



Research Question: Does Gender affect survival?

Here we use survdiff to see if there is a difference.

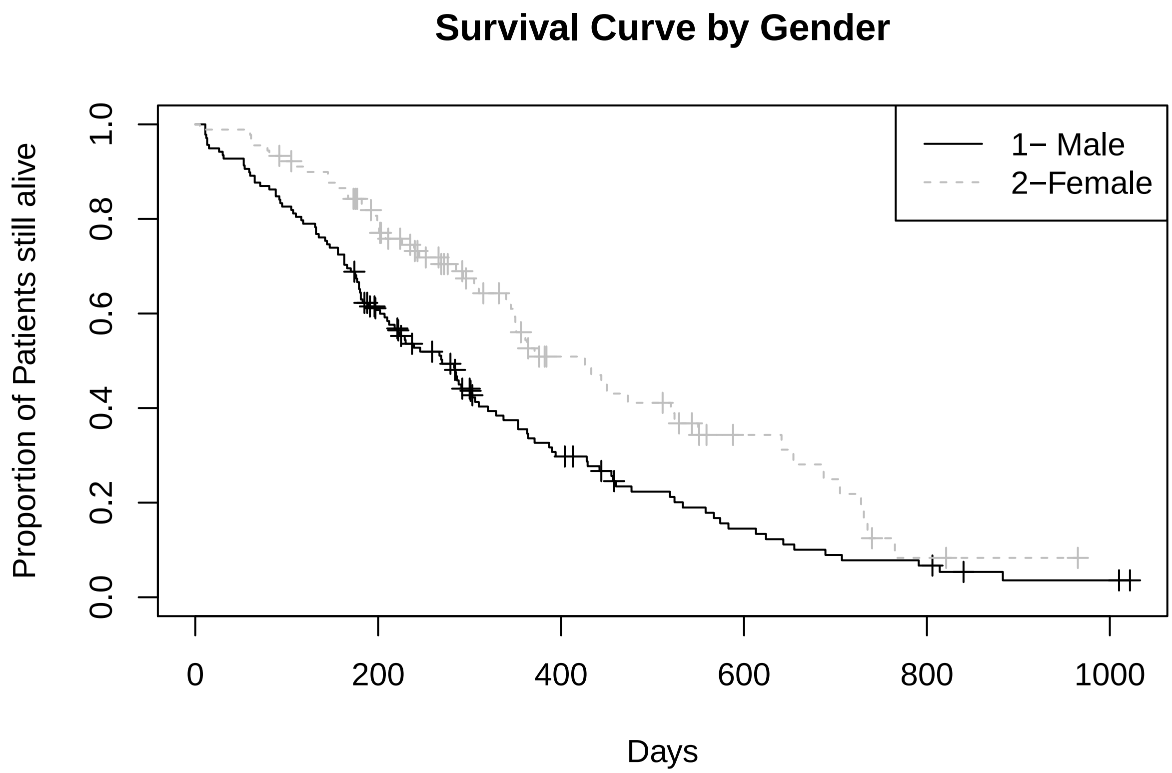
```
lung.gender.survfit<-survfit(Surv(lung.analysis$time, lung.analysis$status)~ lung.analysis$sex, data=lun

lung.gender.survdiff<-survdiff(Surv(lung.analysis$time, lung.analysis$status)~ lung.analysis$sex, data=l

lung.gender.survdiff
```

```
## Call:
## survdiff(formula = Surv(lung.analysis$time, lung.analysis$status) ~
##     lung.analysis$sex, data = lung.analysis)
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## lung.analysis$sex=1 138      112     91.6      4.55      10.3
## lung.analysis$sex=2  90       53     73.4      5.68      10.3
##
##  Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

```
plot(lung.gender.survfit, mark.time = TRUE, col=c("black", "grey75"), lty=1:2)
title(main="Survival Curve by Gender", xlab="Days", ylab = "Proportion of Patients still alive")
legend("topright", c("1- Male", "2-Female"), lty=1:2, col=c("black", "grey75"))
```



The p-value is significant and this supports what we see in the survival curves. There is a difference in survival curves between Male and Female patients.

Another possible research question

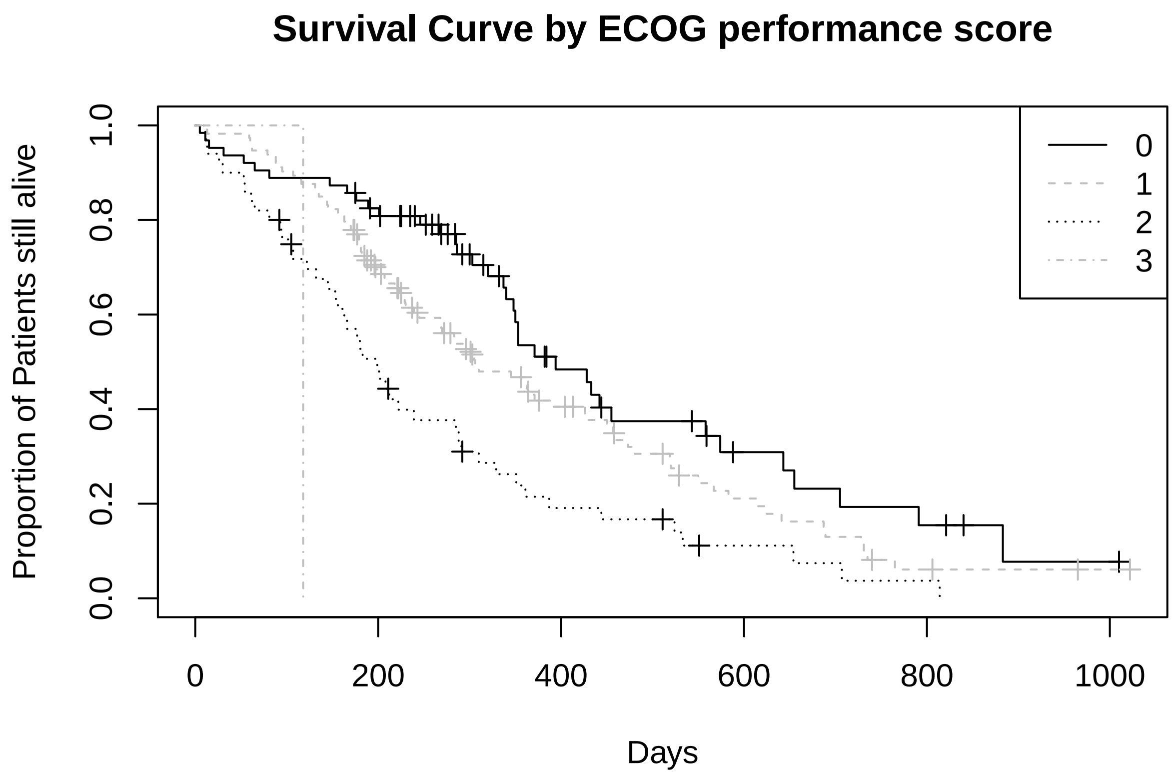
```
lung.perf.survfit<-survfit(Surv(lung$time, lung$status)~ lung$ph.ecog, data=lung)

lung.perf.survdiff<-survdiff(Surv(lung$time, lung$status)~ lung$ph.ecog, data=lung)

lung.perf.survdiff
```

```
## Call:
## survdiff(formula = Surv(lung$time, lung$status) ~ lung$ph.ecog,
##     data = lung)
##
## n=227, 1 observation deleted due to missingness.
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## lung$ph.ecog=0  63       37   54.153    5.4331    8.2119
## lung$ph.ecog=1 113       82   83.528    0.0279    0.0573
## lung$ph.ecog=2  50       44   26.147   12.1893   14.6491
## lung$ph.ecog=3   1        1    0.172    3.9733    4.0040
##
##  Chisq= 22  on 3 degrees of freedom, p= 7e-05
```

```
plot(lung.perf.survfit, mark.time = TRUE, col=c("black", "grey75"), lty=1:4)
title(main="Survival Curve by ECOG performance score", xlab="Days", ylab = "Proportion of Patients stil
legend("topright", c("0", "1", "2", "3"), lty=1:4, col=c("black", "grey75"))
```



We can see that there is one category with only one case. Perhaps this can be removed or grouped with group ecog=2?

Cox Proportional Hazards (OPTIONAL)

```
lung.gender.ph<-coxph(Surv(lung.analysis$time, lung.analysis$status)~ lung.analysis$sex, data=lung.anal
summary(lung.gender.ph)
```

```
## Call:
## coxph(formula = Surv(lung.analysis$time, lung.analysis$status) ~
##     lung.analysis$sex, data = lung.analysis)
##
##   n= 228, number of events= 165
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## lung.analysis$sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##                    exp(coef) exp(-coef) lower .95 upper .95
## lung.analysis$sex     0.588      1.701    0.4237     0.816
##
## Concordance= 0.579  (se = 0.021 )
## Likelihood ratio test= 10.63  on 1 df,   p=0.001
## Wald test            = 10.09  on 1 df,   p=0.001
## Score (logrank) test = 10.33  on 1 df,   p=0.001
```

The exp(coef) column contains $exp(\beta_1)$ This is the hazard ratio – the multiplicative effect of that variable on the hazard rate (for each unit increase in that variable). So, for a categorical variable like gender (in this case), going from male (baseline) to female results in approximately ~40% reduction in hazard. Recall that the CoxPH model is a linear model of the natural log of the hazard at time t, denoted $h(t)$, as a function of the baseline hazard ($h_0(t)$)

$log(h(t)) = log(h_0(t)) + \beta_1 x_1 + \cdots + \beta_p x_p$

if both sides are exponentitated:

$h_1(t) = h_0(t) \times exp(\beta_1 x_1)$

Rearranging makes it possible to estimate the hazard ratio:

$HR(t) = \frac{h_1(t)}{h_0(t)} = exp^{\beta_1}$