Survival analysis

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

Survival analysis

- In many biomedical studies, the outcome variable is a survival time, or more generally a time to an event. We will describe some of the standard tools for analyzing survival data.
- Most studies of survival last a few years, and at completion many subjects may still be alive. For those individuals, the actual survival time is not known – all we know is how long they survived from their entry in the study. Also, some individuals may drop out from the study early.
- Each of these cases is said to be censored; all's we know is that the event hadn't happened yet the last time we saw them.

HPA staining and breast cancer

- We consider data from a retrospective study of 45 women who had surgery for breast cancer. Tumor cells, surgically removed from each woman, were classified according to the results of staining on a marker taken from the Roman snail, the Helix pomatia agglutinin (HPA).
- The survival times in months t_i and staining results ($x_i = 0$ for negative and $x_i = 1$ for positive) for the 45 women are given. Also included is a censoring indicator d_i.
- Contrary to the normal definition of an indicator variable, the censoring indicator is zero if the observation is right-censored, and one if the observation is uncensored. So it's really a non-censoring indicator!

HPA staining and breast cancer

- A woman's survival time was right censored if the woman was alive at the end of the study or if the woman died of causes unrelated to breast cancer.
- A first step in survival analysis is often to estimate the survival curve, or survival time distribution.
- Where $t > 0$, the survival function is $S(t) = \Pr\{T > t\}$, the probability that a randomly selected individual survives at least until time t. This is also the proportion of population that survives until time t or later.
- $S(t) = \Pr\{T > t\}$ is called the **survival function**.

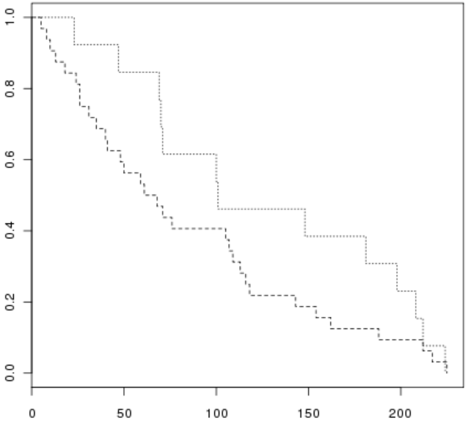
The survival time data

- $x = 0$ (negative) staining:
  23; 47; 69; 70$*$; 71$*$; 100$*$; 101$*$; 148; 181; 198$*$; 208$*$; 212$*$; 224$*$
- $x = 1$ (positive) staining:
  5; 8; 10; 13; 18; 24; 26; 26; 31; 35; 40; 41; 48; 50; 59; 61; 68;
  71; 76$*$; 105$*$; 107$*$; 109$*$; 113; 116$*$; 118; 143; 154$*$; 162$*$; 188$*$;
  212$*$; 217$*$; 225$*$
- * indicates right-censoring.
- What is the estimate of survival function for each group?

Censoring and survival

- **Case I: No censoring**
  If we have a random sample from the population, we can use the empirical survival function; this is the sample proportion that survive at least until time t – very easy to compute.
- But if there is censoring then this is a bad estimate.

Empirical survival function



Short-dashed is negative; long-dashed is positive stained.

Kaplan-Meier estimator

- **Case II: Right censoring**
  We can estimate the survival function using the Kaplan-Meier estimator

  $$\hat{S}(t) = \frac{n_j - d_j}{n_j} \times \frac{n_{j-1} - d_{j-1}}{n_{j-1}} \times \ldots \times \frac{n_1 - d_1}{n_1}$$

  where we group the data into intervals $t_{j-1} < t < t_j$, where n_j is the number at risk of dying at the beginning of the interval, and d_j is the number that die in the interval.
- Note $\frac{n_j - d_j}{n_j}$ is the estimated probability of surviving past t_j given you have survived past t_{j-1}.

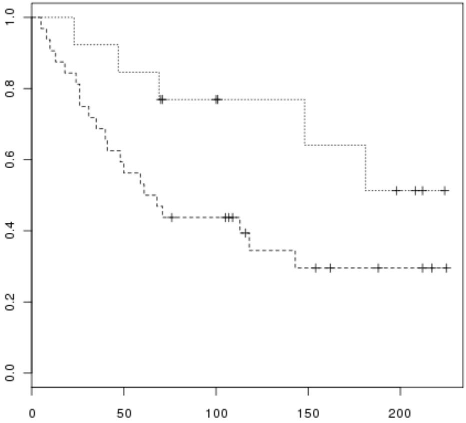
In R

```
> library(survival)
> timeneg = c(23,47,69,70,71,100,101,148,181,198,208,212,224)
> censneg = c(1,1,1,0,0,0,0,1,1,0,0,0,0)
> timepos = c(5, 8, 10, 13, 18, 24, 26, 26, 31, 35, 40, 41, 48, 50, 59, 61, 68, 71, 76, 105,
+ 107, 109, 113, 116, 118, 143, 154, 162, 188, 212, 217, 225)
> censpos = c( rep(1,times=18),0,0,0,0,1,0,1,1,0,0,0,0,0,0)
> group = as.factor( c( rep("neg",times=13), rep("pos",times=32) ) )
> time = c(timeneg,timepos)
> cens = c(censneg,censpos)
> plot( survfit( Surv(time) ~ group ) , lty=3:2)
> fit = survfit( Surv(time,cens) ~ group )
> plot(fit,lty=3:2)
> survdiff( Surv(time,cens) ~ group )
Call:
survdiff(formula = Surv(time, cens) ~ group)

           N Observed Expected (O-E)^2/E (O-E)^2/V
group=neg 13        5     9.57      2.18      3.51
group=pos 32       21    16.43      1.27      3.51

 Chisq= 3.5  on 1 degrees of freedom, p= 0.0608
```

Kaplan-Meier estimates



Short-dashed is negative; long-dashed is positive stained.

Cox's proportional hazards

- We define the hazard function $h(t)$ such that for small enough Δ,
$$\Pr\{t < T < t + \Delta | t \leq T\} = h(t)\Delta.$$
- Cox's proportional hazards model states that the hazard in one group is $h(t)$ and the hazard in the other group is $h(t)e^{\beta}$.
- We want to test $H_0 : \beta = 0$
- We will use the HPA staining example.

Cox PH

```
> coxph( Surv(time,cens) ~ group )
Call:
coxph(formula = Surv(time, cens) ~ group)

          coef exp(coef) se(coef)    z     p
grouppos 0.909      2.48    0.501 1.82 0.069
```

Interpretation

- The estimated coefficient is positive, so the staining result $x = 1$ increases the hazard.
- Note that e^{β} is the relative risk of failing in the next instant for the group denoted by $x = 1$ versus $x = 0$.
- The relative risk in the two groups is $e^{0.909} = 2.48$
- The effect is (not quite) significant, we do not reject $H_0 : \beta = 0$ at the 5% level because $0.069 > 0.05$.

Confidence interval

- A 95% confidence interval for the log relative risk is

$$0.909 \pm 1.96 SE_{b_1} = 0.909 \pm 1.96(0.501) = (-0.073, 1.891).$$

- Exponentiating gives the 95% confidence interval for the relative risk: $(e^{-0.073}, e^{1.891}) = (0.930, 6.626)$.