Lecture: Survival

Peter Ström

April 8, 2015

Table of Contents

Intro to survival data

Kaplan-Meier

- The survivor function
- Kaplan-Meier: An estimator of the survivor function
- Example data HIV
- The logrank test

Cox PH regression

- The hazard function
- Cox proportional hazards model
- The proportional hazards assumtion

What is survival data?

Is it the same as binomial data?

Censoring

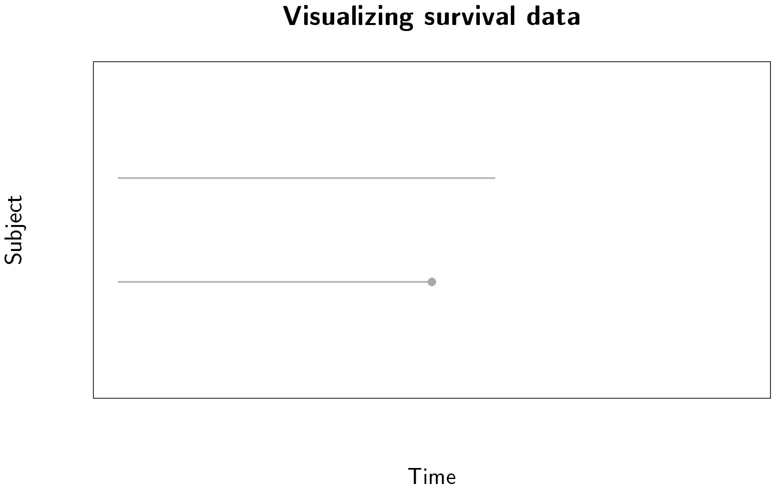


Table of Contents

Intro to survival data

Kaplan-Meier

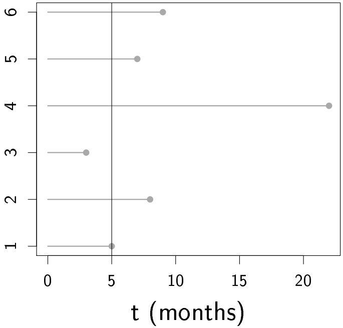
- The survivor function
- Kaplan-Meier: An estimator of the survivor function
- Example data HIV
- The logrank test

Cox PH regression

- The hazard function
- Cox proportional hazards model
- The proportional hazards assumtion

Survivor function, S(t)

The survivor function gives the probability of surviving beyond t.

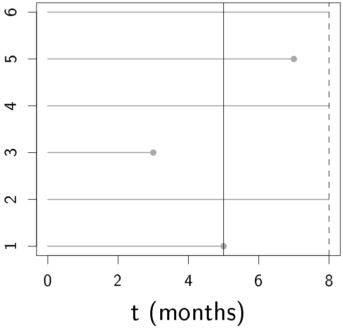


Example: No Censoring

How to estimate the probability of surviving beyond 5 months, S(t=5), when there is no censoring, i.e. we know the time of the event for all subjects?

Survivor function, S(t)

The survivor function gives the probability of surviving beyond t.

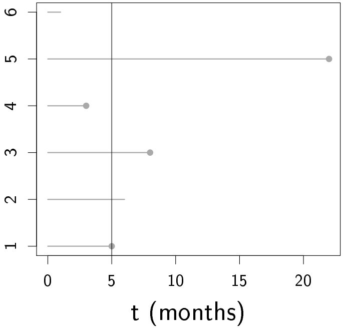


Example: A single right censoring time

How to estimate the probability of surviving beyond 5 months, S(t=5), when there is only a single censoring time at 8 months?

Survivor function, S(t)

The survivor function gives the probability of surviving beyond t.



Warning: Censoring during follow-up

The main issue of survival analysis is how to deal with censoring! Has subject 6 died before or after 5 months?

Kaplan-Meier: An estimator of S(t)

If a subject is censored before time t, then estimating S(t) simply as the observed proportion with event times greater than t can be biased - the censored subject may have died before time t without our knowledge.

The solution is to look at each event time $t_1 < t_2 < ... < t_k$. Let d_j and n_j be the number who die and are *at risk* of dying, respectively, at time t_j.

The Kaplan-Meier (KM) estimator

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

At risk

At risk means thay have not (yet) died nor have been censored. If one already died she is no longer at risk. And if one has been censored she is not considered at risk anymore since even if she will die, we can't observe it.

HIV data

- Question: A Health Maintenance Organization (HMO) wants to evaluate the survival time of HIV+ members using a follow-up study.
- Enter: Members diagnosed with HIV from Jan 1, 1989 to Dec 31, 1991 were enrolled into the study.
- Exit: Follow-up until death due to AIDS or AIDS-related complications, until end of study (Dec 31, 1995), or lost to follow-up.
- Baseline measures: Age and drug use.

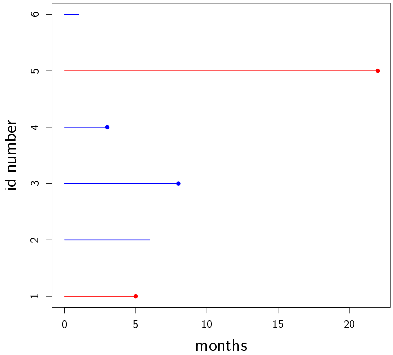
HIV data

OK, let's look at the structure of the data:

```
hiv <- read.table(
  "http://www.ats.ucla.edu/stat/R/examples/asa/hmohiv.csv",
  sep=",", header = TRUE)

head(hiv)

  ID time age drug censor    entdate     enddate
1  1    5  46    0      1 5/15/1990  10/14/1990
2  2    6  35    1      0 9/19/1989   3/20/1990
3  3    8  30    1      1 4/21/1991  12/20/1991
4  4    3  30    1      1  1/3/1991    4/4/1991
5  5   22  36    0      1 9/18/1989   7/19/1991
6  6    1  32    1      0 3/18/1991   4/17/1991
```

Variables

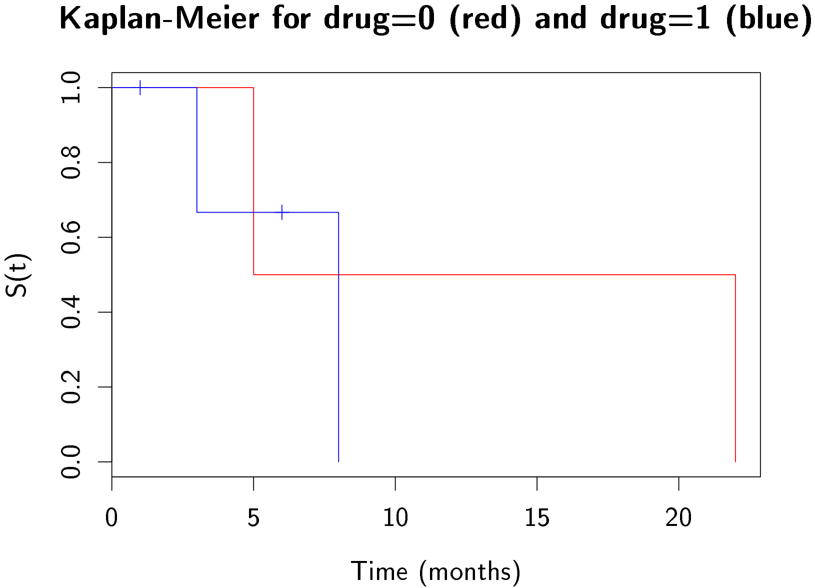
- time: follow-up time (months)
- censor: 1 = dead, 0 = censored

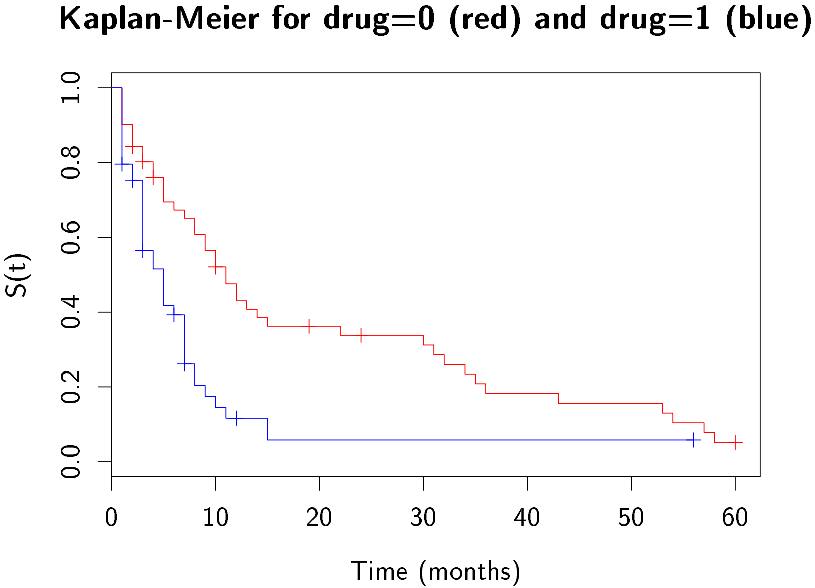


```
ID time drug censor
 6    1    1      0
 5   22    0      1
 4    3    1      1
 3    8    1      1
 2    6    1      0
 1    5    0      1
```

Variables

- time: follow-up time (months)
- censor: 1 = dead, 0 = censored


Kaplan-Meier for drug=0 (red) and drug=1 (blue)
S(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Time (months)


Kaplan-Meier for drug=0 (red) and drug=1 (blue)
S(t)
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
Time (months)

Test for difference in survivor functions

Proceed as we always do:

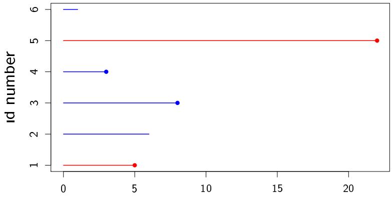
1. **The null**. Assume there is no difference; we call this the null hypothesis (or simply the null).
2. **Find statistic**. Find a statistic (i.e. a function of the data) for which we know the distribution under the null (usually a chisq-distribution).
3. **Test**. See if the value for your statistic is unusually large for what could be expected under the null.

The **logrank** statistic

The test statistic is the sum of $(O - E)^2$/E for each group, where O and E are the totals of the observed and expected events.

For each event time we calculate the expected death in each group as the proportion of the subjects at risk times the number of deaths. Then we sum up these expected deaths to get E for each group.

Test for difference in survivor functions - Example 1



Example: 6 subjects

- Drug 0:

$$E = \frac{2}{5}1 + \frac{2}{4}1 + \frac{1}{2}1 + \frac{1}{1}1 = 2.4$$

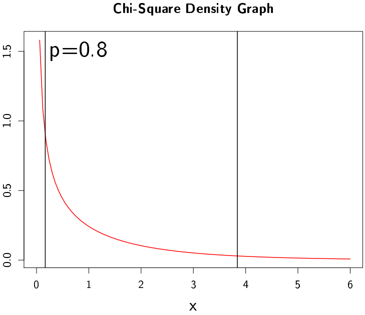
- Drug 1:

$$E = \frac{3}{5}1 + \frac{2}{4}1 + \frac{1}{2}1 + \frac{0}{1}1 = 1.6$$

The logrank test statistic:

$$\frac{(2-2.4)^2}{2.4} + \frac{(2-1.6)^2}{1.6} = 0.17$$

Is 0.17 an extreme value under the null?



Test for difference in survivor functions - Example 2

```
hiv$agecat <- cut(hiv$age, c(min(hiv$age), 29, 34, 39,
                             max(hiv$age)), include.lowest=T)
survdiff(Surv(time=time, event=censor) ~ agecat, data=hiv)

Call:
survdiff(formula = Surv(time = time, event = censor) ~ agecat,
    data = hiv)


                N Observed Expected (O-E)^2/E (O-E)^2/V
agecat=[20,29] 12        8     19.9   7.10608   12.4419
agecat=(29,34] 34       29     29.4   0.00641    0.0117
agecat=(34,39] 25       20     17.8   0.26894    0.3834
agecat=(39,54] 29       23     12.9   7.98170   11.1799

 Chisq= 19.9  on 3 degrees of freedom, p= 0.000178
```

Table of Contents

Intro to survival data

Kaplan-Meier

- The survivor function
- Kaplan-Meier: An estimator of the survivor function
- Example data HIV
- The logrank test

Cox PH regression

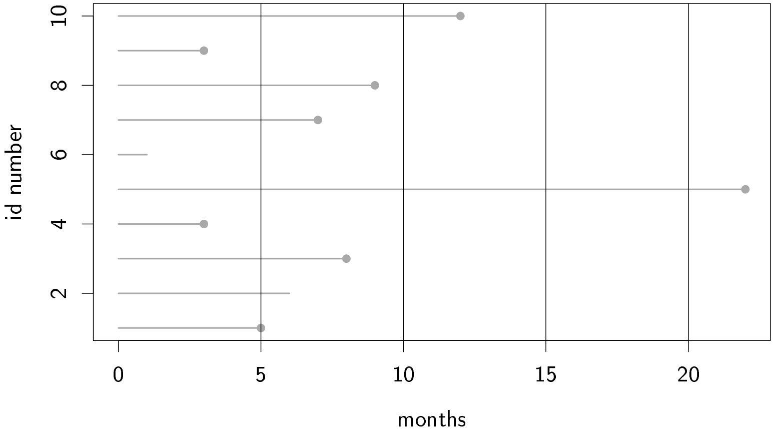
- The hazard function
- Cox proportional hazards model
- The proportional hazards assumtion

The hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

The hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$



Model the hazard function - Cox regression

$$\log\ h(t) = \log\ h_0(t) + \beta x$$

$$\Leftrightarrow$$

$$h(t) = h_0(t)\mathrm{e}^{\beta x}$$

Hazard ratio (or risk ratio) from Cox regression

$$X_1 = \begin{cases} 1 & (\mathit{Male}) \\ 0 & (\mathit{Female}) \end{cases} \qquad X_2 = \begin{cases} 1 & (\mathit{Blue\ eyes}) \\ 0 & (\mathit{Not\ blue\ eyes}) \end{cases}$$

$$\begin{aligned} h(t|X_1 = 0, X_2 = 0) &= h_0(t) \\ h(t|X_1 = 1, X_2 = 0) &= h_0(t)exp(\beta_1) \\ h(t|X_1 = 0, X_2 = 1) &= h_0(t)exp(\beta_2) \\ h(t|X_1 = 1, X_2 = 1) &= h_0(t)exp(\beta_1 + \beta_2) \end{aligned}$$

Now we can obtain the Hazard ratio (risk ratio) for any combination of groups, e.g.:

$$\text{HR(Male vs Female)}=\frac{h_0(t)exp(\beta_1)}{h_0(t)} = exp(\beta_1)$$

Cox regression in R

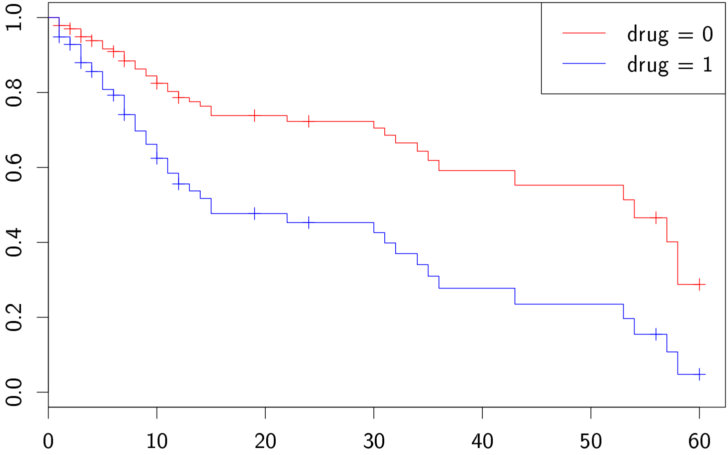
```
table(hiv$agecat)


[20,29] (29,34] (34,39] (39,54]
     12      34      25      29

coxph(Surv(time=time, event=censor) ~ agecat, data=hiv)

Call:
coxph(formula = Surv(time = time, event = censor) ~ agecat, data = hiv)


              coef exp(coef) se(coef)    z       p
agecat(29,34] 1.20      3.33    0.450 2.67 7.5e-03
agecat(34,39] 1.33      3.80    0.458 2.91 3.6e-03
agecat(39,54] 1.91      6.78    0.468 4.09 4.3e-05

Likelihood ratio test=20.9  on 3 df, p=0.000109  n= 100, number of events= 80
```

Cox regression in R

```
hiv$drug <- as.factor(hiv$drug)
coxph(Surv(time=time, event=censor) ~ drug + age, data=hiv)

Call:
coxph(formula = Surv(time = time, event = censor) ~ drug + age,
    data = hiv)


        coef exp(coef) se(coef)    z       p
drug1 1.0167      2.76   0.2562 3.97 7.2e-05
age   0.0971      1.10   0.0186 5.21 1.9e-07

Likelihood ratio test=39.1  on 2 df, p=3.18e-09  n= 100, number of events= 80
```

Cox regression in R

```
cox <- coxph(Surv(time=time, event=censor) ~ agecat + drug, data=hiv)
predict <- data.frame(drug=c(0,1), agecat=rep(levels(hiv$agecat)[1], 2))
```



Comparison of nested models

Make a likelihood ratio (LR) test to see if there is an significant overall effect of agecat:

```
model1 <- coxph(Surv(time=time, event=censor) ~ drug, data=hiv)
model2 <- coxph(Surv(time=time, event=censor) ~ agecat + drug, data=hiv)
anova(model1, model2)

Analysis of Deviance Table
 Cox model: response is  Surv(time = time, event = censor)
 Model 1: ~ drug
 Model 2: ~ agecat + drug
   loglik  Chisq Df P(>|Chi|)
1 -290.12
2 -279.25 21.745  3 7.369e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The proportional hazards assumtion

Cox regression a.k.a Cox Proportional Hazards regression

$$\log\ h(t,x) = \log\ h_0(t) + \beta x$$

$$\Leftrightarrow$$

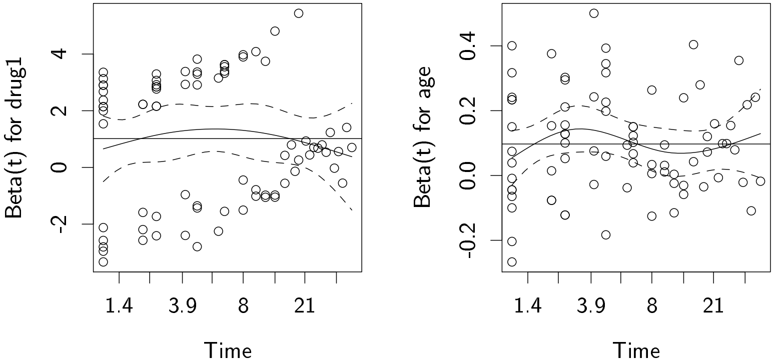
$$h(t,x) = h_0(t)\mathrm{e}^{\beta x}$$

$$\mathrm{HR} = \frac{h(t,x^*)}{h(t,x)} = \frac{h_0(t)\mathrm{e}^{\beta x^*}}{h_0(t)\mathrm{e}^{\beta x}} = \mathrm{e}^{\beta(x^*-x)}$$

The proportional hazards assumtion - Graphically

Graphs of the Schoenfeld residuals help us detect if the parameters vary over the follow-up (i.e. non-PH).

```
cox <- coxph(Surv(time=time, event=censor) ~ drug + age, data=hiv)
plot(cox.zph(cox))
```



The proportional hazards assumtion - Test

Test if there is a correlation between Schoenfeld residuals and time:

```
cox.zph(cox)

           rho    chisq     p
drug1  0.00188 0.000276 0.987
age    0.01626 0.018958 0.890
GLOBAL      NA 0.019077 0.991
```

Summary

- Start with a sample who are at risk of some event. Follow them until they either get the event or are censored.
- **The Survivor function** is the probability of surviving beyond time t, and is estimated with **Kaplan-Meier**.
- Do two or more groups have different Survivor function? Use the **Logrank test**!
- The hazard can be thought of as the probability that an event will occur at time t.
- Often we don't care about the hazard but only the **extra** proportion hazard in one group (e.g. males) compared to another group (e.g. females). Use **Cox regression**!