Survival Models: Week 5

CIs for Kaplan-Meier Estimate

Last week: $\hat{S}(t) \pm 1.96 \text{ s.d.}(\hat{S}(t))$

An approximate 95% CI for the KM estimator can be found using the "standard" approach: estimate $\pm$ two times the standard error given by Greenwood's formula. With this approach values greater than 1 or less than 0 are replaced with 1 and 0, respectively. In R this "standard CI" corresponds to *conf.type="plain"*. However, the default value is *conf.type="log"*. This default option computes the CI based on the log of the survival function.

CIs for Kaplan-Meier Estimate

Basing the 95% CI for the estimated survival function on the log of the survival function proceeds as follows:

1. Compute a 95% CI for $log[S(t)]$ as:

$$log[\hat{S}(t)] \pm 2 \times \sqrt{\sum_{t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}} = [C_l, C_u],$$

   using the variance computed in week 4.

2. A 95% CI for $S(t)$ can then be obtained by exponentiating the end points C_l and C_u: $[\exp(C_l), \min(1, \exp(C_u))]$. Note: this "log" type CI means that the C_l is guaranteed to be greater than zero.

R Example

```
#Looking at different CI types
library(survival)
data(ovarian)
km.est1<-survfit(Surv(ovarian$futime,ovarian$fustat)~1,conf.type="plain")
km.est2<-survfit(Surv(ovarian$futime,ovarian$fustat)~1,conf.type="log")
km.est3<-survfit(Surv(ovarian$futime,ovarian$fustat)~1,conf.type="log-log")
lower<-cbind(summary(km.est1)$lower,summary(km.est2)$lower,summary(km.est3)$lower)
upper<-cbind(summary(km.est1)$upper,summary(km.est2)$upper,summary(km.est3)$upper)
bounds<-round(cbind(lower,upper),2)
colnames(bounds)<-rep(c("plain","log","log-log"),2)
bounds[1:4,]
     plain  log log-log plain log log-log
[1,]  0.89 0.89    0.76  1.00   1    0.99
[2,]  0.82 0.83    0.73  1.00   1    0.98
[3,]  0.76 0.77    0.68  1.00   1    0.96
[4,]  0.71 0.72    0.64  0.98   1    0.94
```

Comparing survival Curves

It is often of interest to test whether two (or more) survival curves are equivalent. One popular method of conducting this test is the log-rank test. The log-rank test is a large sample chi-square test with the null hypothesis: *There is no difference between the two survival curves.* Briefly, the log-rank test involves ordering the times of death for the entire (combined dataset) and then noting the number of deaths observed from each group, at each of these times. The expected number of deaths from each group, at each time, are also computed. The log-rank test is then based on the observed and expected values from each group. The log-rank statistic has a chi-square distribution with (K-1) degrees of freedom. K is the number of curves being compared. The log-rank test is easily computed in R.

no need to compute, just need to identify

R Example

```
#Example of log-rank test
#kidney data: times to infection once catheter is inserted.
data(kidney)
km.est<-survfit(Surv(kidney$time,kidney$status)~kidney$disease)
plot(km.est,col=c("red","blue","green","black"))
legend(x="topright",c("other","GN","AN","PKD"),lty=1,col=c("red","blue","green","black"))
survdiff(Surv(kidney$time,kidney$status)~kidney$disease,rho=0)
Call:
survdiff(formula = Surv(kidney$time, kidney$status) ~ kidney$disease,
    rho = 0)

                       N Observed Expected (O-E)^2/E (O-E)^2/V
kidney$disease=Other  26       20    23.20     0.442     0.779
kidney$disease=GN     18       14    11.62     0.488     0.631
kidney$disease=AN     24       18    14.70     0.740     1.070
kidney$disease=PKD     8        6     8.48     0.724     0.989

 Chisq= 2.7  on 3 degrees of freedom, p= 0.446
```

talk about them later in week 10.

4 curves

dof = 4-1 = 3

test statistics

decision rule: $p > 0.05$
do not reject null hypothesis: no difference b/w 4 curves

R Example

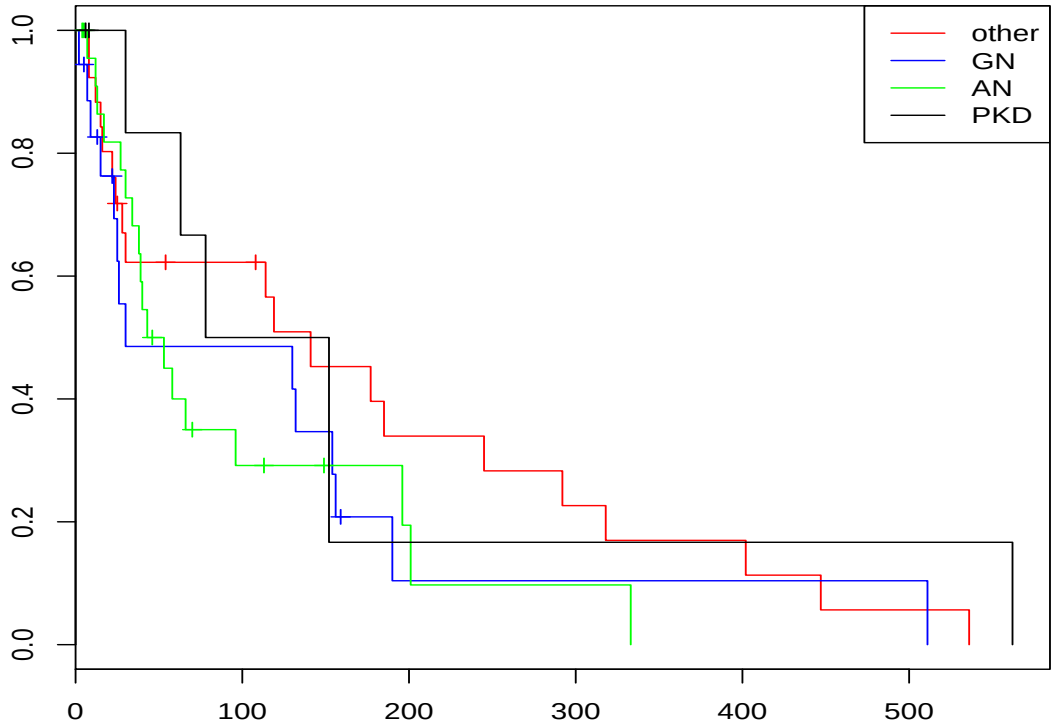


Figure 1: Estimated survival curves for kidney data.

Nelson-Aalen Estimation of Survival Function

The Nelson-Aalen (NA) estimator takes a different approach to estimating the survival function compared to the KM approach. In particular the NA estimator computes an estimate of the integrated hazard $\Lambda(t) = \int_0^t \mu_{x+t} dt$ and then obtains $\hat{S}(t)$ via the relationship $S(t) = \exp(-\Lambda(t))$. The NA estimate of the integrated hazard is computed as:

$$\hat{\Lambda}(t) = \sum_{t_j \leq t} \frac{d_j}{r_j},$$

the NA estimator is then obtained as $\hat{S}(t) = \exp(-\hat{\Lambda}(t))$.

Summary of KM & NA

KM estimate

one way to find the survival function:

$$\hat{S}(t) = \prod_{t_j \le t} \hat{p}_j \quad , \quad \hat{p}_j = \frac{r_j - d_j}{r_j}$$

another way:

$$S(t) = e^{-\int_0^t \mu(s)\,ds}$$

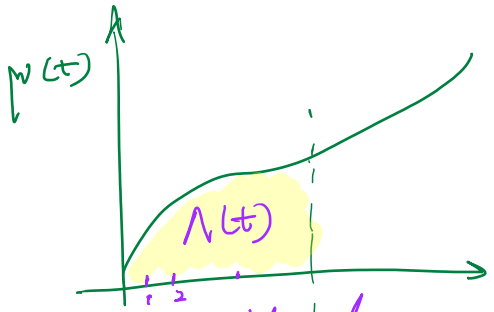
NA estimate

$$\Lambda(t) = \int_0^t \mu(s)\,ds$$

use this to approximate

hazard

$$S(t) = e^{-\Lambda(t)} \Rightarrow \hat{\Lambda}(t) = ?$$ find this



area is integral

if have enough # of deaths $\Rightarrow$

At each time point,
when a death does not occur $\rightarrow 0$
when a death occurs $\rightarrow \frac{d_j}{r_j} = \hat{q}_j$
} to estimate the mortality rate

$$\Rightarrow \hat{\Lambda}(t) = \sum_{t_j \le t} \hat{q}_j \Rightarrow \hat{S}(t) = \exp\left(-\sum_{t_j \le t} \frac{d_j}{r_j}\right)$$

Comparison

KM	t_j	r_j	d_j	$\frac{r_j - d_j}{r_j}$	$\prod_{t_j \le t} \frac{r_j - d_j}{r_j}$
NA	t_j	r_j	d_j	$\frac{d_j}{r_j}$	$\exp\left(-\sum_{t_j \le t} \frac{d_j}{r_j}\right)$

Variance of the Nelson-Aalen Estimator

$$\hat{S}(t) = \exp(-\sum_{t_j \leq t} \frac{d_j}{r_j})$$

Using the same assumptions as for our derivation of Greenwood's formula can we derive its standard error?

$$\hat{S}(t) = \exp\left(-\sum_{t_j \le t} \frac{d_j}{r_j}\right) = \exp\left(\hat{\Lambda}(t)\right)$$

Variance of NA $\quad Var(\hat{S}(t))$

① find $Var[\hat{\Lambda}(t)]$ first

$$\hat{q}_j = \frac{d_j}{r_j} \qquad d_j \sim Bin(r_j, q_j)$$

$$Var(\hat{q}_j) = \frac{1}{r_j^2} \cdot r_j \cdot q_j(1-q_j) \approx \frac{d_j(r_j - d_j)}{r_j^3}$$

$$\Rightarrow Var(\hat{\Lambda}(t)) = Var\left(\sum_{t_j \le t} \hat{q}_j\right) = \sum_{t_j \le t} Var(\hat{q}_j) \qquad (i\,i\,d)$$

$$= \sum_{t_j \le t} \frac{d_j(r_j - d_j)}{r_j^3}$$

Delta Method

② $\hat{S}(t) = e^{-\hat{\Lambda}(t)}$

$g(\mu) = e^{-\mu}$

$g'(\mu) = -e^{-\mu} = -\hat{S}(t)$

$Var(gx) = (g'(\mu))^2 \cdot Var(x)$

$$E(\hat{\Lambda}(t)) = \sum_{t_j \le t} E\left(\frac{d_j}{r_j}\right) = \sum_{t_j \le t} q_j \approx \sum_{t_j \le t} \hat{q}_j = \hat{\Lambda}(t)$$

$$\Rightarrow Var(\hat{S}(t)) = (-\hat{S}(t))^2 \cdot \sum_{t_j \le t} \frac{d_j(r_j - d_j)}{r_j^3}$$

Nelson-Aalen - Example

The observed survival times (or times of censoring) of $N = 8$ individuals after a particular operation (in months) were: $8, 12, 12*, 17, 17, 22, 27*, 30$. The "$*$" correspond to censored observations. What is the NA estimate of the survival function at time $t = 24$?

t_j	r_j	d_j	$\frac{d_j}{r_j}$	$e^{-\sum \frac{d_j}{r_j}}$
8	8	1	$\frac{1}{8}$	$e^{-\frac{1}{8}}$
12	7	1	$\frac{1}{7}$	$e^{-\frac{1}{8}-\frac{1}{7}}$
17	5	2	$\frac{2}{5}$	$e^{-\frac{1}{8}-\frac{1}{7}-\frac{2}{5}}$
22	3	1	$\frac{1}{3}$	...
30	1	1	1	$e^{-\frac{1}{8}-\frac{1}{7}-\cdots-1}$

① $\hat{S}(24) = e^{-(\frac{1}{8}+\frac{1}{7}+\frac{2}{5}+\frac{1}{3})}$ NA estimator

② $\hat{S}(24) = e^{-(\frac{1}{8}+\frac{1}{7}+\frac{1}{5}+\frac{1}{4}+\frac{1}{3})}$ Fleming-Harrington estimator (No tied death) is also one type of NA estimator.

"survfit" function in R uses F.H.

R - Example

```
#Example of NA estimator
library(survival)
data(ovarian)
NA.est<-survfit(Surv(ovarian$futime,ovarian$fustat)~1,type="fleming-harrington")
KM.est<-survfit(Surv(ovarian$futime,ovarian$fustat)~1,type="kaplan-meier")
summary(NA.est)
> summary(NA.est)
Call: survfit(formula = Surv(ovarian$futime, ovarian$fustat) ~ 1, type = "fleming-harrington")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.962  0.0377        0.891        1.000
  115     25       1    0.925  0.0523        0.827        1.000
  156     24       1    0.887  0.0628        0.772        1.000
  268     23       1    0.849  0.0710        0.721        1.000
  329     22       1    0.811  0.0776        0.673        0.979
  353     21       1    0.774  0.0831        0.627        0.955
.
.
.
plot(NA.est)
lines(KM.est,col="red")
```

R Example

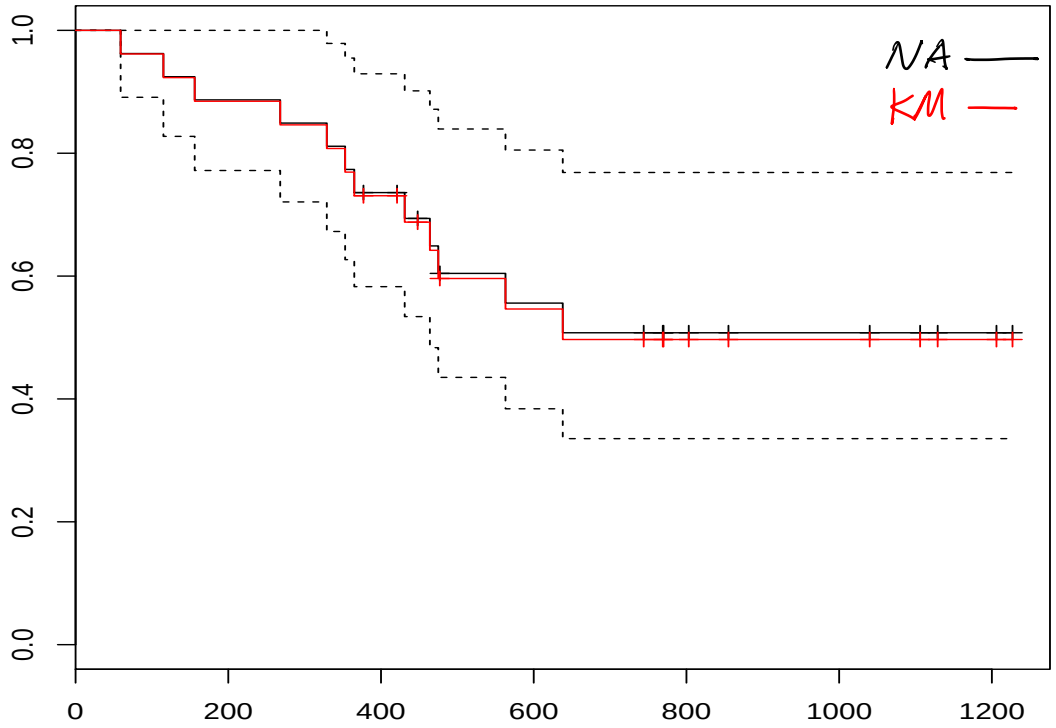


Figure 2: Comparison of NA and KM estimators.

Addict Example

We will look at data from a 1991 Australian study by Capelhorn and Bell (1991). This dataset has information on how long heroin addicts remained in treatment. There are two clinics in the dataset. There is also information on whether an individual had spent time in prison. An example of the data is provided below. In total there are 238 individuals in the dataset.

	id	clinic	status	time	prison	dose
1	1	1	1	428	0	50
2	2	1	1	275	1	55
3	3	1	1	262	0	55
4	4	1	1	183	0	30
5	5	1	1	259	1	65

Note: status= 1 and 0 correspond to death and censoring, respectively.

Addict Example

Does time in treatment depend on clinic? To answer this question we can compute survival curves based on clinic and then use a log-rank test.

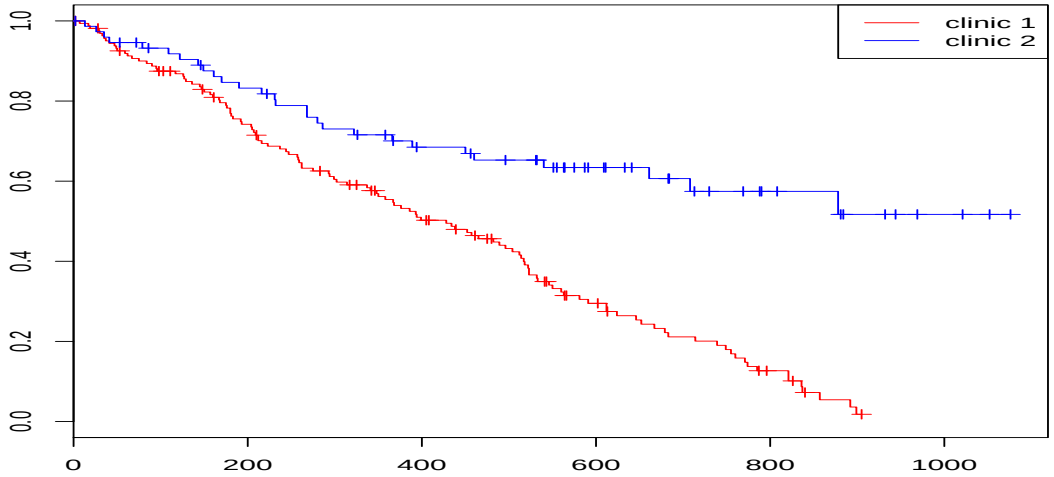


Figure 3: Estimated survival curves for addict data.

Addict Example

A log-rank test for this data gives the following result:

```
Chisq= 27.9  on 1 degrees of freedom, p= 1.28e-07
```

Clinic appears to be highly important. What about whether the individual has a prison record?

```
KM.est<-survfit(Surv(addict$time,addict$status)~addict$clinic+addict$prison)
plot(KM.est,col=c("red","blue","green","black"))
legend(x="topright",c("clinic=1, prison=0","clinic=1, prison=1","clinic=2,
prison=0","clinic=2, prison=1"),lty=1,col=c("red","blue","green","black"))
survdiff(Surv(addict$time,addict$status)~addict$clinic+addict$prison,rho=0)

Chisq= 32.6  on 3 degrees of freedom, p= 3.88e-07
```

small ⇒ at least 2 of the curves are different from each other.

Addict Example

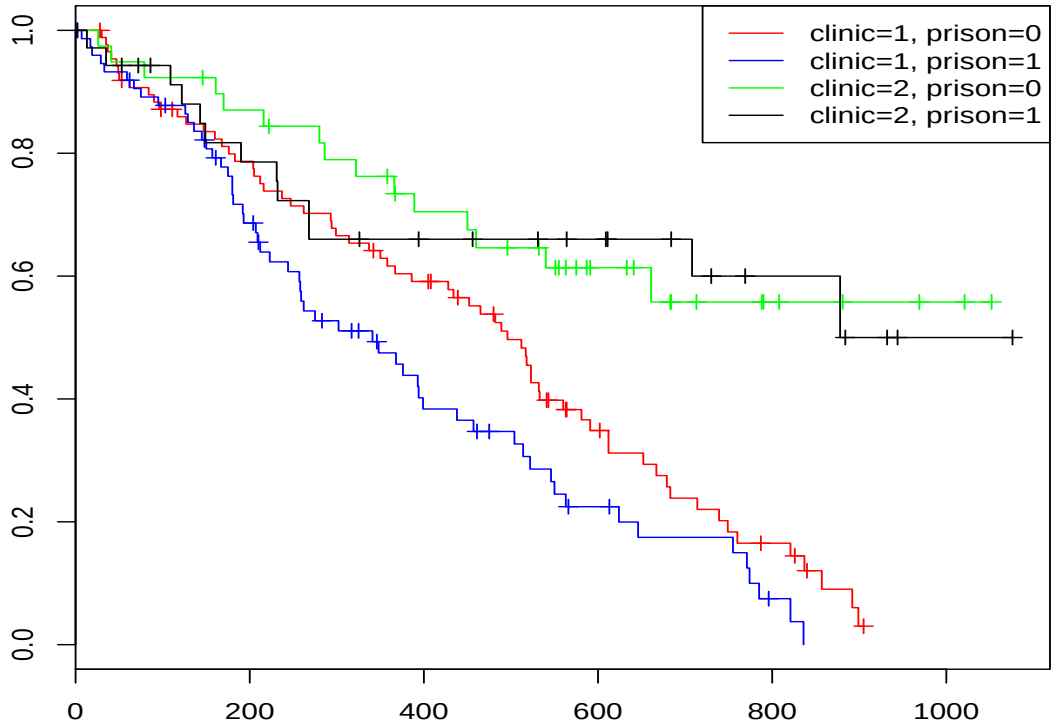


Figure 4: Estimated survival curves for addict data.

Addict Example

Using the estimated survival curves we can also compute the median time in treatment. As an example we will look at the entire addict data as one group.

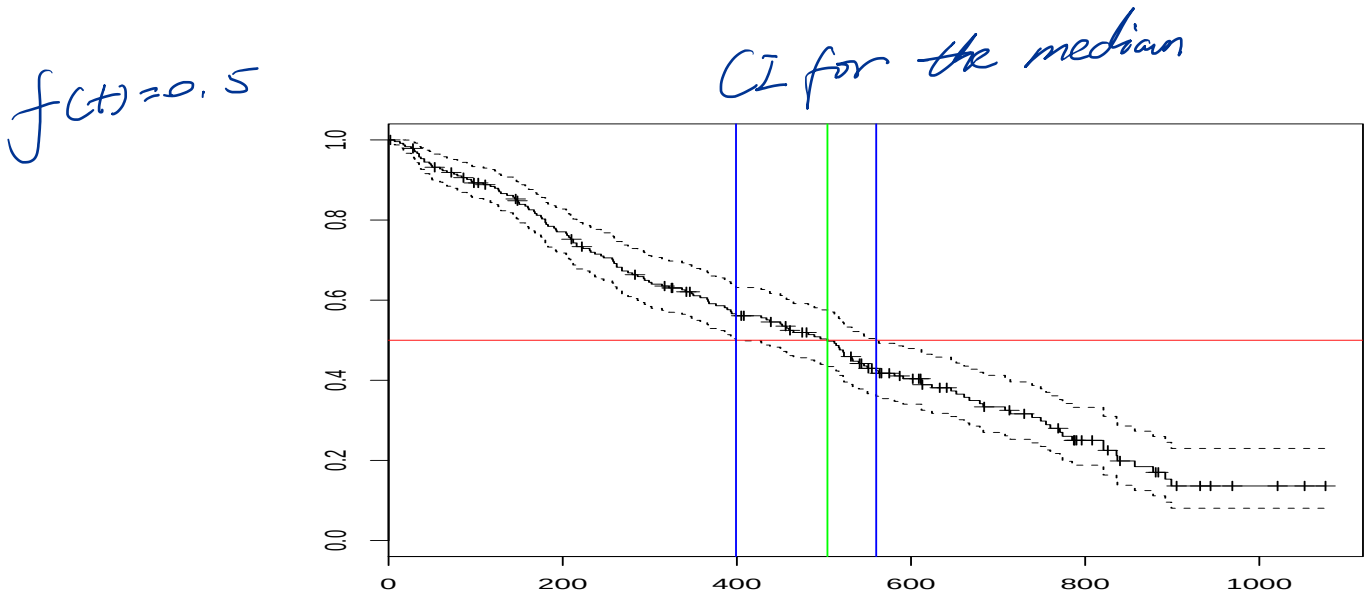


Figure 5: Estimated survival curves for addict data. Blue lines give endpoints of 95% CI for median survival time and green line is our estimate of the median.

Likelihood Approach

We can also use likelihood based approaches when we have censored data. In an example where we have non-informative right censoring the likelihood can be written as:

$$\prod_{complete} f(t_i) \prod_{censored} S(c_i).$$

differ from completed $f(t_i)$

The likelihood can then be maximized to obtain the parameter estimates.

$$L(\lambda) = \begin{cases} f(t) & \text{complete} \\ S(c) & \text{censoring} \end{cases}$$

$S(c_i) = P(T_i \geq c_i)$

i^{th} individual

Likelihood Approach

The following survival times after a particular operation were observed 2,2.5*,3,24,1*,0.5. The density ($f(t) = \lambda \exp(-\lambda t)$ is believed to be suitable for modelling the survival times of this population. Find the MLE estimate of λ.

$$f(t) = \lambda \exp(-\lambda t)$$

$$L(\lambda) = \underbrace{\prod_{i=1}^{4} \lambda e^{-\lambda t_i}}_{\text{pdf}} \cdot \underbrace{\prod_{j=1}^{2} e^{-\lambda c_j}}_{P(T_i \geq c_j) \text{ cdf}}$$

$$l(\lambda) = \sum_{i=1}^{4} \log(\lambda e^{-\lambda t_i}) + \sum_{j=1}^{2} \log(\lambda e^{-\lambda c_j})$$

$$= 4 \log \lambda + \sum_{i=1}^{4} (-\lambda t_i) + \sum_{j=1}^{2} (-\lambda c_j)$$

$$l'(\lambda) = \frac{4}{\lambda} - \sum_{i=1}^{4} t_i - \sum_{j=1}^{2} c_j \Rightarrow \hat{\lambda} = \frac{4}{\sum \cdots} = \frac{4}{33}$$

Cox Proportional Hazards Regression

- In many situations we are interested in the impact of a number of covariates on the survival distribution. Some of these covariates will typically be continuous. KM won't work for continuous variables
- Examples of possible covariates include: age, weight, blood pressure.
- Using KM estimation with continuous variables presents a number of difficulties.
- Cox regression can be used in these situations.

Cox Proportional Hazards Regression

In basic Cox regression the hazard at time t for an individual with covariates given by the p-vector x can be expressed as:

$$\lambda(t; x) = \lambda_0(t) \exp(\beta^T x),$$

where β is a p-vector of unknown parameters. The quantity $\lambda_0(t)$ is the baseline hazard. It represents the hazard for a life who has all covariates equal to zero at time t. For a given individual, the covariates scale the baseline hazard by an amount equal to $\exp(\beta^T x)$. Importantly the baseline hazard is arbitrary. For this reason the form of $\lambda(t; x)$ specified above is known as a semi-parametric model.

Cox Proportional Hazards Regression

For example, we may be interested in the hazard for a life who underwent an operation at age x, who is female at time t years after the operation.

For this life we define two covariates x_1 and x_2. The first covariate is a continuous variable and represents the age of the life at the time of the operation. The second covariate is concerned with the gender of the individual. This clearly requires an indicator variable and we will assume $x_2 = 1$ for males and $x_2 = 0$ for females.

Using the expression for the hazard on the previous slide, the hazard for a male aged 25 who underwent an operation 5 years ago is:

$$\lambda(t; x) = \lambda_0(5) \exp(20\beta_1 + \beta_2).$$

Similarly the hazard for a male aged 30 who underwent an operation 5 years ago is

$$\lambda(t; x) = \lambda_0(5) \exp(25\beta_1 + \beta_2).$$

The ratio of these two hazards is:

$$\frac{\exp(20\beta_1 + \beta_2)}{\exp(25\beta_1 + \beta_2)}.$$

ratio is time-invariant

Note: the baseline hazard cancels out in this ratio. This hazard will NOT change with time. As the two males considered above get older the ratio of their predicted hazards will remain constant. We therefore can say that the hazards of two individuals are proportional to each other.

Cox Proportional Hazards Regression

$$S(t) = S_0(t)^{\exp(\beta^T x)}$$

where $S_0(t) = \exp(-\int_0^t \lambda_0(s)ds)$ is the survival function for an individual whose covariates values are all zero. The above relationship implies that the survival functions for different covariate values cannot cross. Note: this restriction does not apply for the KM estimator. How to derive this.

$$\beta = \begin{pmatrix} \beta_1 \\ \\ \beta_p \end{pmatrix} \quad ?$$

$$x_i = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$$

$$\beta^T x = \sum_{i=1}^{p} \beta_i \cdot x_i$$

estimate β first

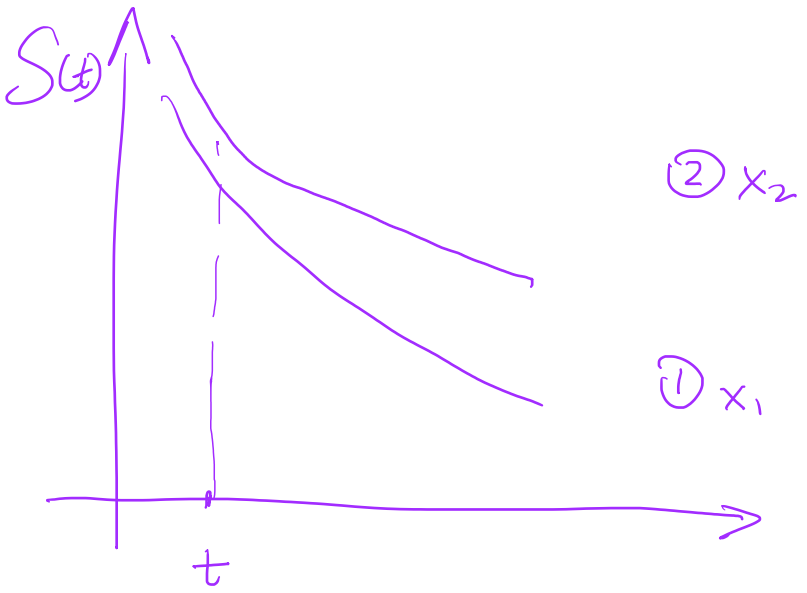
$$S(t) = \exp\left(-\int_0^t \lambda(s; x) \cdot ds\right)$$

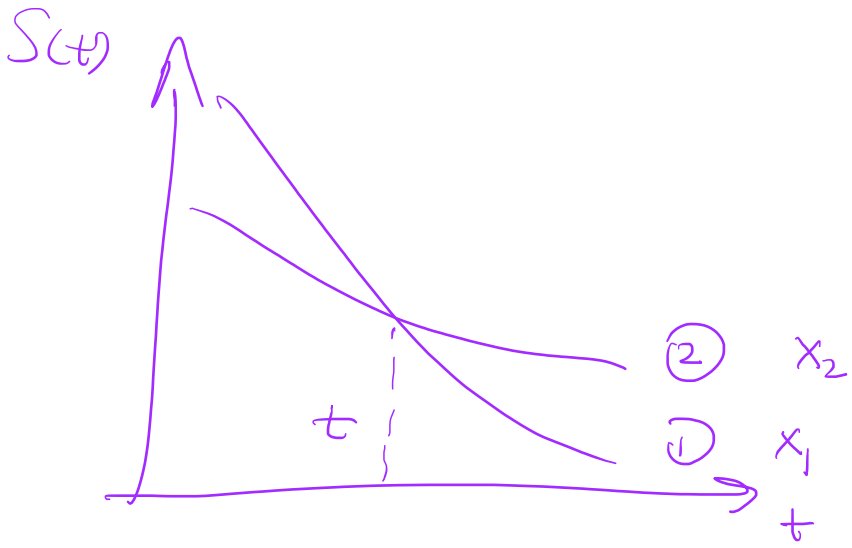
$$= \exp\left(-\int_0^t \lambda_0(s) \cdot \exp(\beta^T x)\, ds\right)$$

$$= \exp\left(-\int_0^t \lambda_0(s)\, ds\right) \exp(\beta^T x)$$

$$= \left[\exp\left(-\int_0^t \lambda_0(s)\, ds\right)\right]^{\exp(\beta^T x)}$$

$$= S_0^{\exp(\beta^T x)}$$





Cannot use Cox-proportional hazard model.

$$S_1(t) = S_2(t)$$

$$\Rightarrow S_0(t)^{\exp(\beta^T x_1)} = S_0(t)^{\exp(\beta^T x_2)}$$

$$\Rightarrow x_1 = x_2 \quad \text{it's not true!}$$

But we say:

when $x_1 \neq x_2 \Rightarrow S_1(t) \neq S_2(t)$

The survival functions can never cross.

Partial Likelihood

The parameters of the Cox regression model are estimated as the values that maximise the partial likelihood (PL). We will use the same notation that was established in our discussion of the KM estimator. In addition, we define the following:

- $x_{(i)}$: the covariates for the person that was observed to die at time t_i.
- $R(t_i)$: the set of individuals still under study at time t_i.

Partial Likelihood

(calculation won't be in exam)

The partial likelihood (assuming no tied deaths) can be expressed as:

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(\beta^T x_{(i)})}{\sum_{j \in R(t_i)} \exp(\beta^T x_j)},$$

where, m is the number of deaths. The PL can be treated as a typical likelihood.

generally, don't have closed form solution.

Partial Likelihood

The PL is motivated by asking the following: "what is the probability that a person with covariates given by $x_{(s)}$ dies at t_s given that one of the persons in $R(t_s)$ dies at this time".

if this prob ↑ ⇒ person with $x_{(s)}$ is more likely to die than other persons in this time.

$$P(\text{person with covariates } x_{(s)} \text{ dies in time } (t_s, t_s+\delta) \mid \text{give one death in } (t_s, t_s+\delta))$$

$$= \frac{P(\text{person with covariates } x_{(s)} \text{ dies in time } (t_s, t_s+\delta) \mid \text{survival to } t_s)}{P(\text{one death in } (t_s, t_s+\delta) \mid \text{survival to } t_s)}$$

$$= \frac{\exp(\beta^T x_{(s)})}{\sum_{j \in R(t_s)} \exp(\beta^T x_j)}$$

Note: here we are using the fact that:
$P(\text{person with covariates } x_i \text{ dies at time } t_s \mid \text{survival to } t_s) \approx$
$\lambda_0(t_s) \exp(\beta^T x_i)\delta$.

$$\mu_h \approx \frac{{}_h q_x}{h} \qquad {}_h q_x \approx \mu_h \cdot h$$

$$\frac{P(\text{person with } x_{(s)} \text{ dies within } \dots)}{P(\text{one death within } \dots)}$$

$$= \frac{\delta \cdot \lambda(x_{(s)}, t_s)}{\sum_{j \in R(t_s)} \delta \cdot \lambda(x_{(j)}, t_s)}$$

$$= \frac{\exp(\beta^T x_{(s)}) \cdot \lambda_0}{\sum_{j \in R(t_s)} \exp(\beta^T x_j) \cdot \lambda_0}$$

"partial likelihood"

↓

try to maximize this prob.

Partial Likelihood

When there are tied deaths the PL can be expressed as:

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(\beta^T s_{(i)})}{[\sum_{j \in R(t_i)} \exp(\beta^T x_j)]^{d_i}}.$$

where $s_{(i)} = x_{(i),1} + \ldots + x_{(i),d_i}$, is the sum of the covariates for the persons observed to die at t_i. This is approximation is due to Breslow.

tied death: example.
2 covariates ① age & blood pressure
3 deaths at time t_i
$s_{(i)} = \begin{bmatrix} age \\ bp \end{bmatrix} = \begin{bmatrix} age_1 + age_2 + age_3 \\ bp_1 + bp_2 + bp_3 \end{bmatrix}$

R Example - Recidivism

```
library(RcmdrPlugin.survival)
data(Rossi) #see J fox notes for more details
Rossi[1:3,1:10]
  week arrest fin age  race wexp          mar paro prio educ
1   20      1  no  27 black   no not married  yes    3    3
2   17      1  no  18 black   no not married  yes    8    4
3   25      1  no  19 other  yes not married  yes   13    3
cox.fit<-coxph(Surv(week,arrest)~fin+age+race+wexp+mar+
prio,data=Rossi)
summary(cox.fit)
```

Output is found on next slide. Note: in this table the baseline values of the categorical covariates are "no" for financial aid, "black" for race, "no" for work experience, and "married" for married.

R Example - Recidivism

```
Call:
coxph(formula = Surv(week, arrest) ~ fin + age + race + wexp +
    mar + prio, data = Rossi)

  n= 432, number of events= 114

                   coef exp(coef) se(coef)      z Pr(>|z|)
finyes          -0.37352   0.68831  0.19082 -1.957 0.050295 .
age             -0.05640   0.94516  0.02184 -2.583 0.009796 **
raceother       -0.30983   0.73357  0.30780 -1.007 0.314133
wexpyes         -0.15331   0.85786  0.21218 -0.723 0.469957
marnot married   0.44339   1.55799  0.38136  1.163 0.244958
prio             0.09336   1.09785  0.02832  3.296 0.000981 ***
```