STK4900/9900 - Lecture 9

Program

1. Survival data and censoring
2. Survival function and hazard rate
3. Kaplan-Meier estimator
4. Logrank test
5. Proportional hazards and Cox regression

- Section 3.5
- Sections 6.1 and 6.2

Survival data and censoring

The data in this lecture have a different form from what we have seen earlier

The response is the time (from a well defined starting point) until a specific event (end point) occurs, or until observation of the subject stops

Examples:

- Time from birth to the onset of a disease
- Time from onset of a disease to death
- Duration of unemployment
- Time from starting a PhD-study to graduation

We will often call the time until the event a survival time, also when the end point of interest is something else than death

A new aspect for survival data is censoring:

The event of interest does not necessarily occur in the observation period.

Then we only know that the survival time is longer than the observation period, but not exactly how long.

This is referred to as censoring.

Also these censored survival times contain important information and must be included in the analysis.

Example: clinical trial

Assume that we want to study the time from disease onset until death

- New patients are diagnosed and included in the study

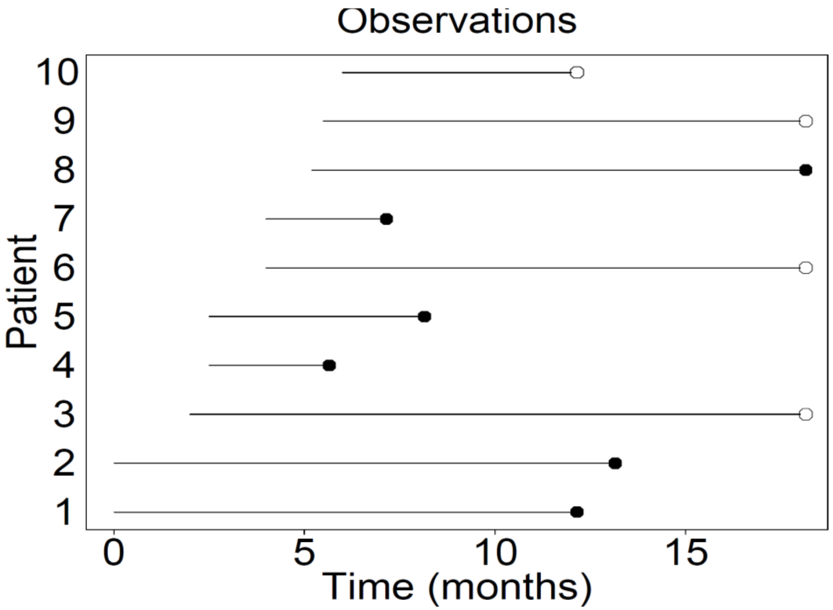
- The patients are then followed until:
    - death
    - they no longer want to participate
    - the study is concluded

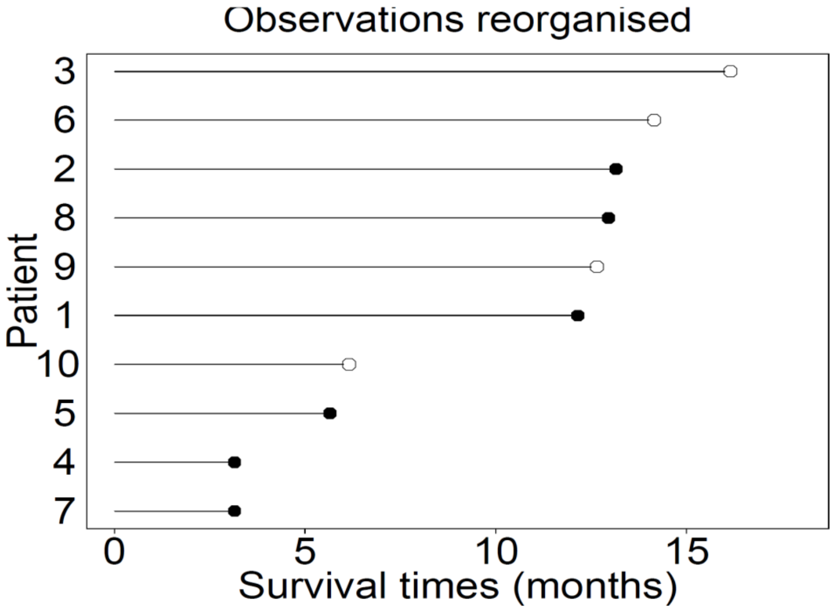
In the second and third case the survival times are censored.

Illustration for a hypothetical clinical trial with 10 patients:

Follow-up of patients on the calendar time scale:

Follow-up of patients on the study time scale:





Death: ● and censoring: ○

Notation for censored survival times

$$T_i^0 = \text{survival time for individual no } i$$

$$C_i = \text{censoring time for individual no } i$$

We do not observe all T_i^0 (or C_i), but only:

$$T_i = \min(T_i^0, C_i) = \text{censored survival time}$$

$$D_i = \begin{cases} 1 & \text{if } T_i^0 \le C_i \quad \text{so the survival time is observed} \\ 0 & \text{if } T_i^0 > C_i \quad \text{so the censoring time is observed} \end{cases}$$

The response for subject i is (T_i, D_i), i.e. a combination of a numerical response T_i and a binary response D_i

Using T_i as response without taking D_i into account does not make sense. We need statistical methods that use data on all subjects, whether their survival times are observed or we only observe time until censoring.

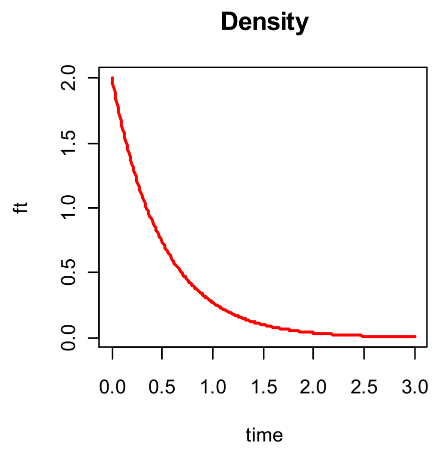
Concepts describing the distribution of survival times

The following concepts may all be used to describe the distribution of a survival time T^0 :

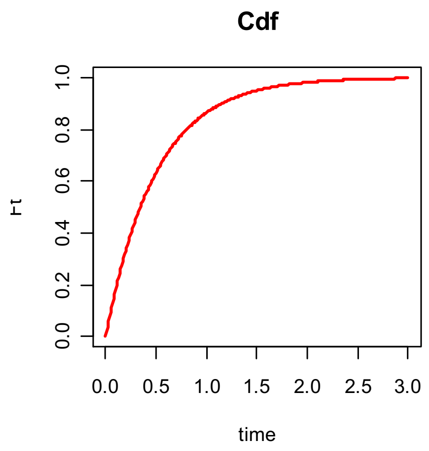
- Density $f(t)$: $P(t \le T^0 < t + \Delta) \approx f(t)\Delta$
- Cumulative distribution function: $F(t) = P(T^0 \le t)$
- Survival function: $S(t) = 1 - F(t) = P(T^0 > t)$
- Hazard function: $h(t)$: $P(t \le T^0 < t + \Delta | T^0 \ge t) \approx h(t)\Delta$
- Cumulative hazard function: $H(t) = \int_0^t h(s)ds$

Example: exponential distribution

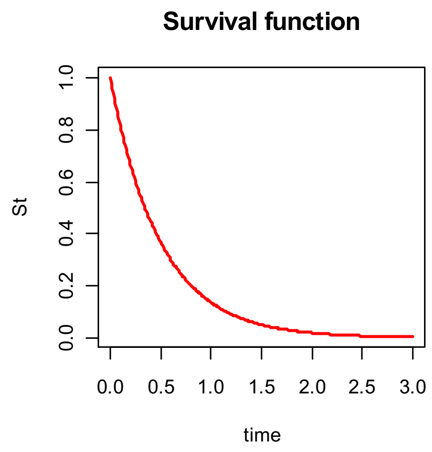
- $f(t) = \lambda e^{-\lambda t}$



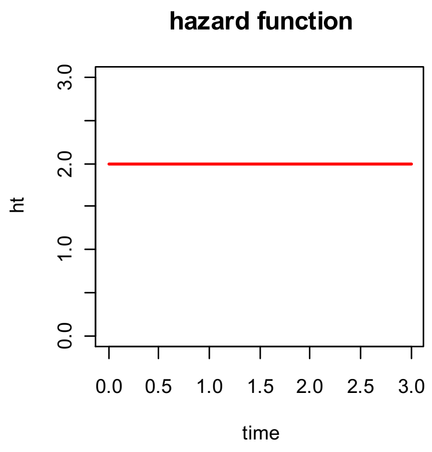
- $F(t) = 1 - e^{-\lambda t}$



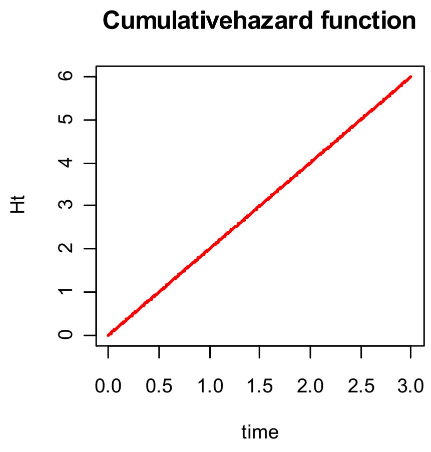
- $S(t) = e^{-\lambda t}$



- $h(t) = \lambda$



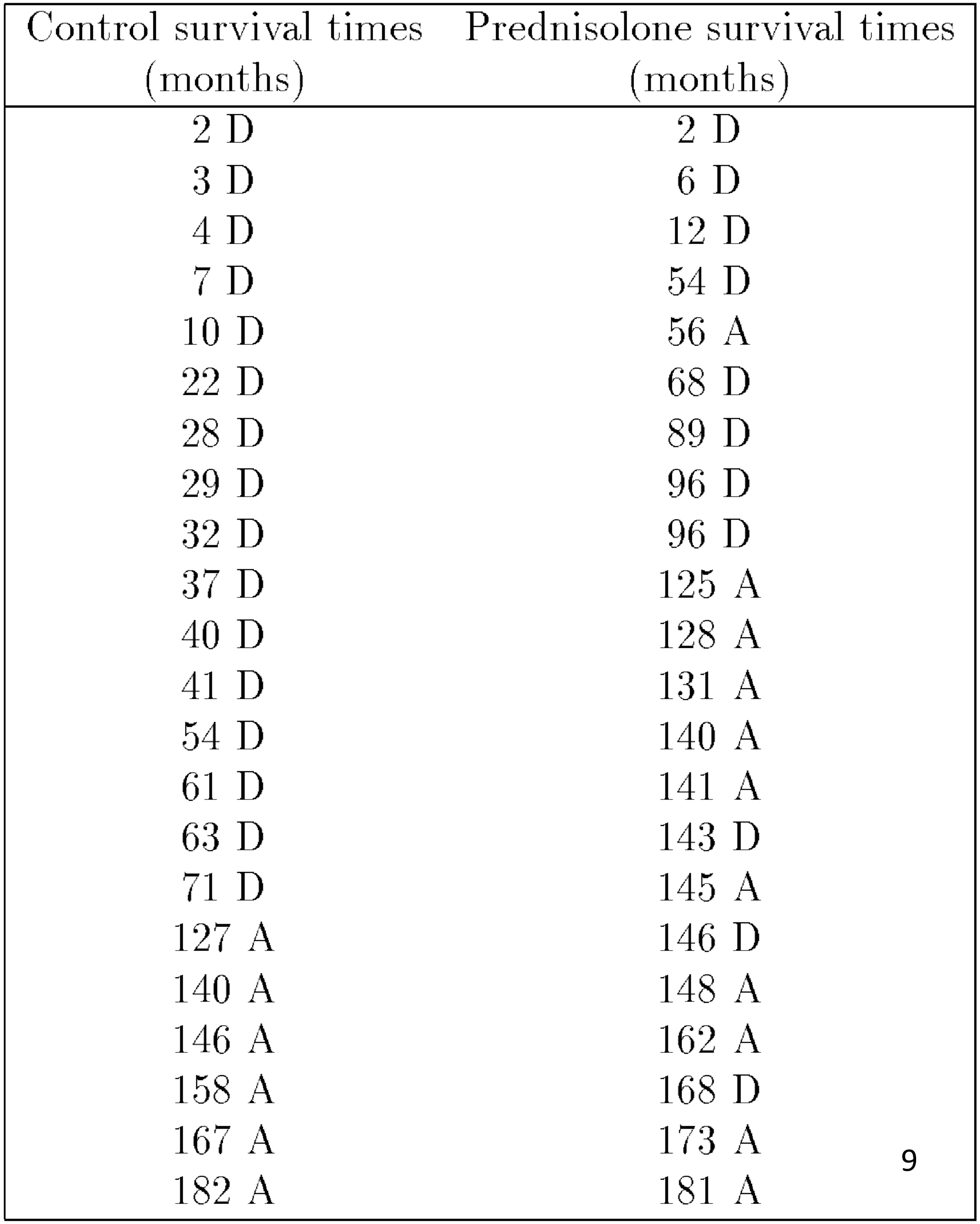
- $H(t) = \lambda t$



Example:

In a clinical trial 44 patients with chronic active hepatitis were randomized either to treatment with prednisolone or to an untreated control group

The table shows the censored survival times and whether a patient died (D) or were still alive (A)



Control survival times (months)	Prednisolone survival times (months)
2 D	2 D
3 D	6 D
4 D	12 D
7 D	54 D
10 D	56 A
22 D	68 D
28 D	89 D
29 D	96 D
32 D	96 D
37 D	125 A
40 D	128 A
41 D	131 A
54 D	140 A
61 D	141 A
63 D	143 D
71 D	145 A
127 A	146 D
140 A	148 A
146 A	162 A
158 A	168 D
167 A	173 A
182 A	181 A

Estimation of the survival function

We want to estimate the survival function $S(t)$ *without* assuming that it belongs to a specific parametric class of distributions (like exponential or gamma).

For illustration we look at the prednisolone group.

19 of the 22 patients live more than 50 months.

Therefore:

$$\hat{S}(50) = \frac{19}{22} = 0.864$$

But how do we find $\hat{S}(100)$?

This can not be found as a simple proportion, since we do not know whether the patient censored at 56 months would live longer than 100 months or not

Kaplan-Meier estimator

Introduce:

- Distinct times of events: $t_1 < t_2 < \cdots$
- $m_j =$ number of events observed at t_j
- $Y(t_j) =$ number "at risk" at t_j

For $t_k \leq t < t_{k+1}$ the survival function is estimated by the product

$$\widehat{S}(t) = \left(1 - \frac{m_1}{Y(t_1)}\right) \cdot \left(1 - \frac{m_2}{Y(t_2)}\right) \cdots \left(1 - \frac{m_k}{Y(t_k)}\right)$$

This is the Kaplan-Meier estimator

More compactly we may write:

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{m_j}{Y(t_j)}\right)$$

Example: prednisolone group

t_j	$Y(t_j)$	m_j	$\frac{m_j}{Y(t_j)}$	$1-\frac{m_j}{Y(t_j)}$	$\widehat{S}(t)$
2	22	1	$\frac{1}{22}$	$\frac{21}{22}$	$\frac{21}{22}$
6	21	1	$\frac{1}{21}$	$\frac{20}{21}$	$\frac{21}{22}\cdot\frac{20}{21}=\frac{20}{22}$
12	20	1	$\frac{1}{20}$	$\frac{19}{20}$	$\frac{19}{20}\cdot\frac{20}{22}=\frac{19}{22}$
54	19	1	$\frac{1}{19}$	$\frac{18}{19}$	$\frac{18}{19}\cdot\frac{19}{22}=\frac{18}{22}$
68	17	1	$\frac{1}{17}$	$\frac{16}{17}$	$\frac{16}{17}\cdot\frac{18}{22}=0.770$
89	16	1	$\frac{1}{16}$	$\frac{15}{16}$	$\frac{15}{16}\cdot 0.770=0.722$
96	15	2	$\frac{2}{15}$	$\frac{13}{15}$	$\frac{13}{15}\cdot 0.722=0.626$
143	8	1	$\frac{1}{8}$	$\frac{7}{8}$	$\frac{7}{8}\cdot 0.626=0.547$
146	6	1	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{5}{6}\cdot 0.547=0.456$
168	3	1	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}\cdot 0.456=0.304$

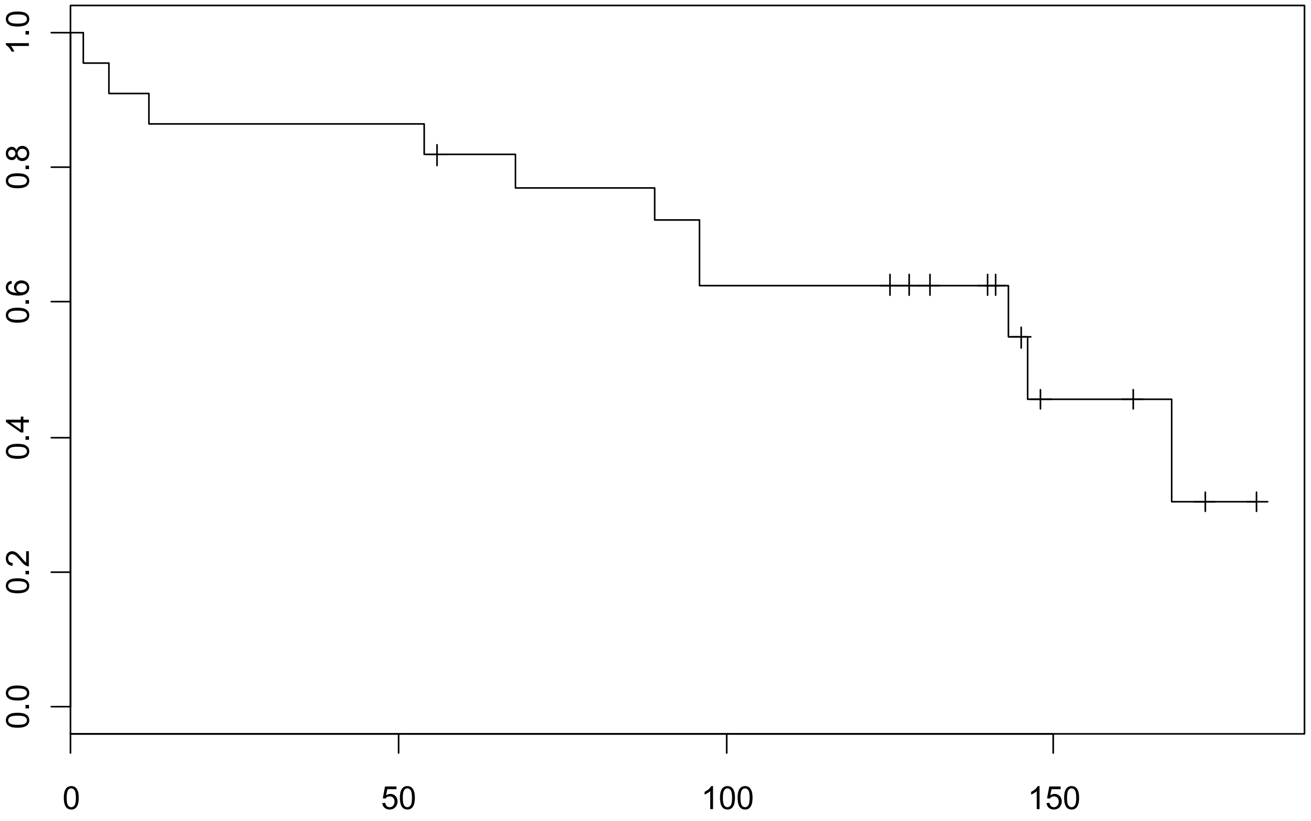
R commands:

```
time=c(2,6,12,54,56,68,89,96,96,125,128,131,140,141,143,145,146,148,162,168,173,181)
cens=c(1,1,1,1,0,1,1,1,1,0,0,0,0,0,1,0,1,0,0,1,0,0)
library(survival)
survpred=survfit(Surv(time,cens)~1, conf.type="none")
summary(survpred)
```

R output :

```
 time   n.risk   n.event   survival    std.err
   2    22         1        0.955     0.0444
   6    21         1        0.909     0.0613
  12    20         1        0.864     0.0732
  54    19         1        0.818     0.0822
  68    17         1        0.770     0.0904
  89    16         1        0.722     0.0967
  96    15         2        0.626     0.1051
 143     8         1        0.547     0.1175
 146     6         1        0.456     0.1285
 168     3         1        0.304     0.1509
```

Plot of Kaplan-Meier estimate



R command: plot(survpred)

Standard error and confidence intervals

The standard error of the Kaplan-Meier estimator is estimated by Greenwood's formula:

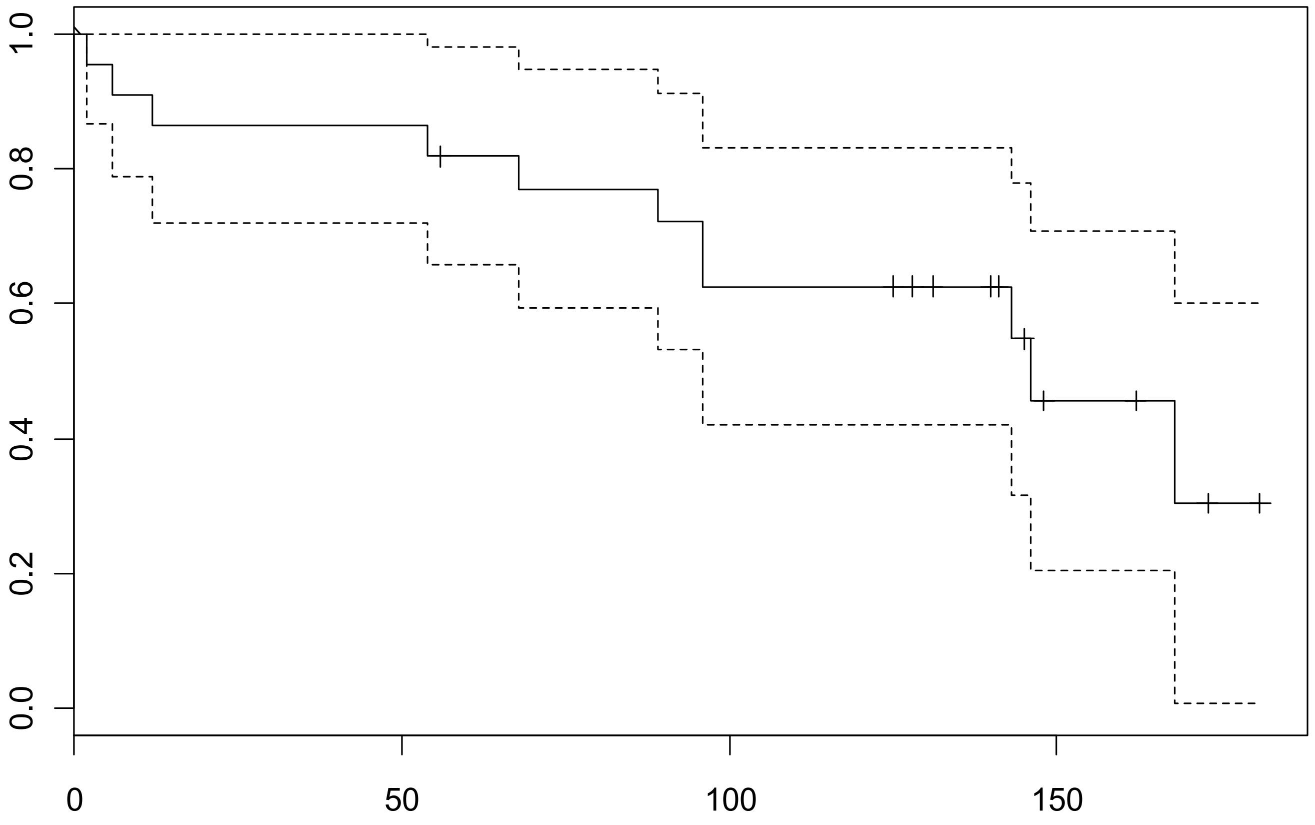
$$\widehat{\mathsf{se}}(\widehat{S}(t)) = \widehat{S}(t)\sqrt{\sum_{t_j \leq t} \frac{m_j}{Y(t_j)(Y(t_j) - m_j)}}$$

A 95% confidence interval for *S*(*t*) is given by:

$$\widehat{S}(t) \pm 1.96 \times \widehat{\mathsf{se}}(\widehat{S}(t))$$

Other options for confidence intervals are available
(but note that R use a silly default)

Kaplan-Meier estimate with confidence limits:



R commands:

```
survpred2=survfit(Surv(time,cens)~1, conf.type="plain")
plot(survpred2)
```

Median survival time

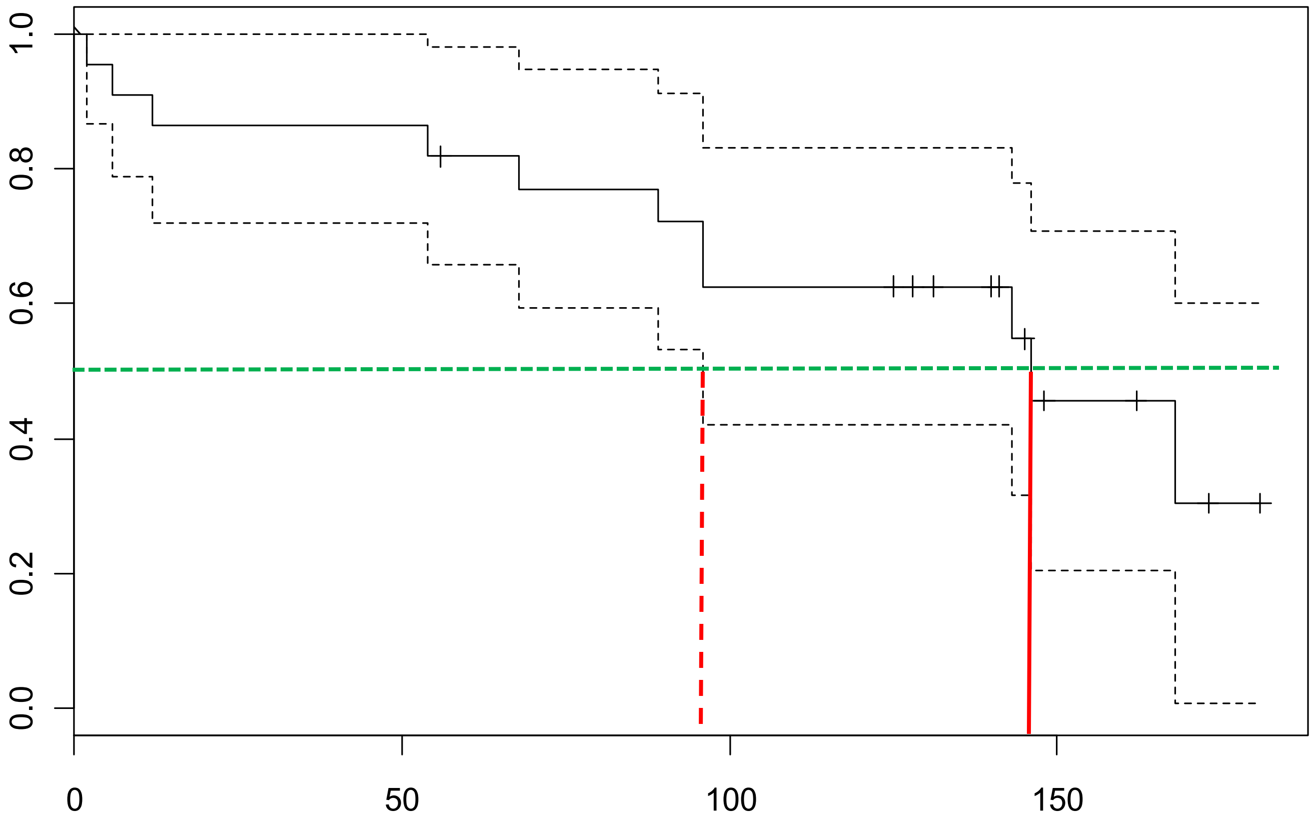
is defined as the time when the Kaplan-Meier estimator $\hat{S}(t) = 0.5$ (or crosses the value 0.5).

This time can be read of graphically from a Kaplan-Meier plot (see next slide).

Other percentiles are defined similarly, for instance the lower and upper quartiles are defined as solving $\hat{S}(t) = 0.75$ and $\hat{S}(t) = 0.25$

Furthermore confidence intervals for the median and percentiles are also found graphically from Kaplan-Meier plots with confidence limits included.

Median survival time:



R commands:

```
print(survpred2)
```

R output:

```
records    n.max    n.start    events    median    0.95LCL    0.95UCL
   22        22        22        11        146        96          NA
```

Comparing two groups

We want to compare the survival in two groups (e.g. treatment and control):

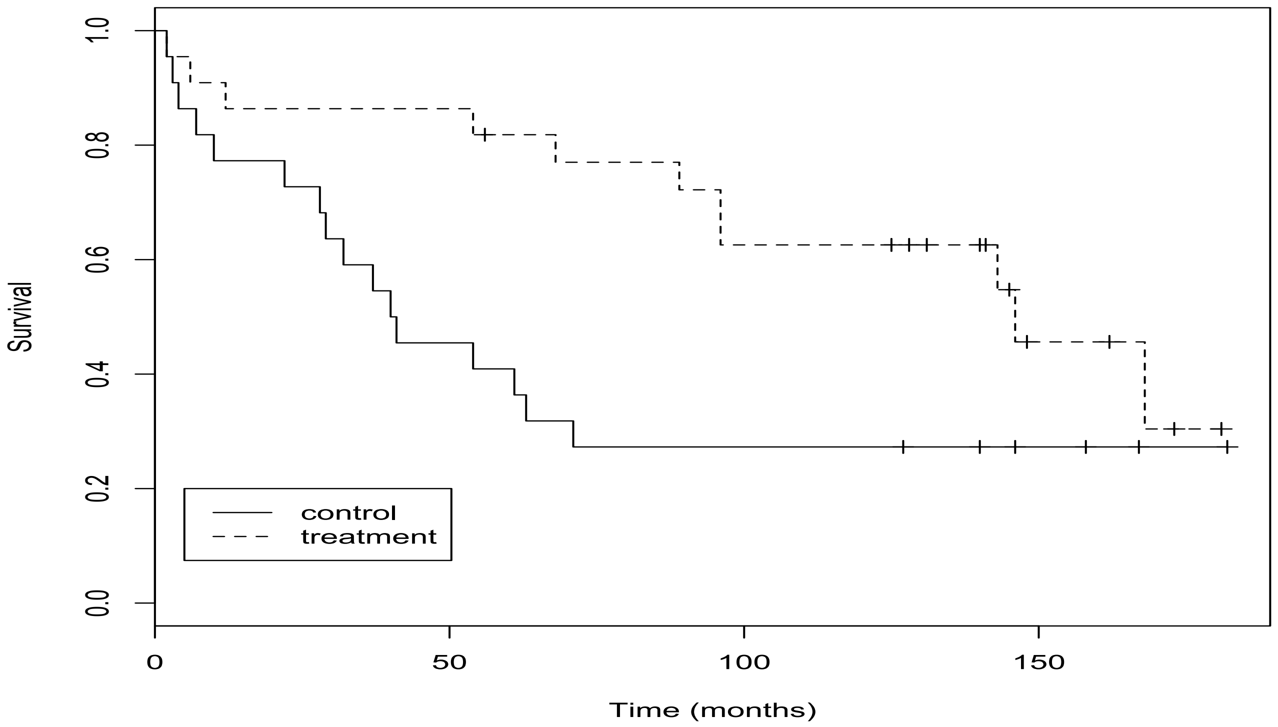
- Group 1: (T_{i1}, D_{i1}) $i = 1, \ldots\ldots, n_1$
- Group 2: (T_{i2}, D_{i2}) $i = 1, \ldots\ldots, n_2$

$\widehat{S}_k(t)$: Kaplan-Meier in group k $(k = 1, 2)$

Comparison:

- Graphically : Plot $\widehat{S}_1(t)$ and $\widehat{S}_2(t)$
- Testing : Log rank-test

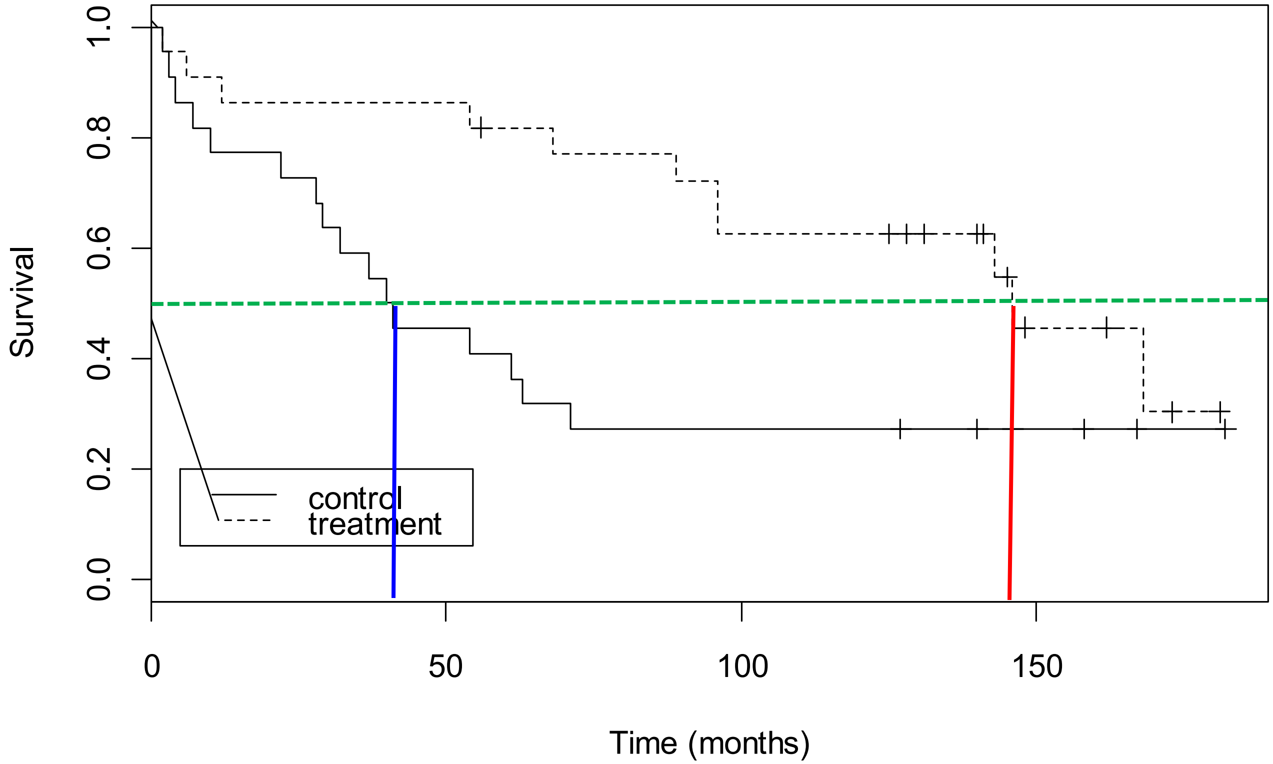
Graphical comparison:



R commands:

```
time=c(2,3,4,7,10,22,28,29,32,37,40,41,54,61,63,71,127,140,146,158,167,182,2,
        6,12,54,56,68,89,96,96,125,128,131,140,141,143,145,146,148,162,168,173,181)
cens=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,
        1,1,1,1,0,1,1,1,1,0,0,0,0,0,1,0,1,0,0,1,0,0)
group=c(rep(1,22),rep(2,22))
survboth=survfit(Surv(time,cens)~group, conf.type="plain")
plot(survboth,lty=1:2,xlab="Time (months)",ylab="Survival")
legend(5,0.2,c("control","treatment"),lty=1:2)
```

Median survival times:



R commands:

```
print(survpred2)
```

R output:

```
             records    n.max    n.start    events    median    0.95LCL    0.95UCL
group=1        22         22        22        16        40         28         71
group=2        22         22        22        11       146         96         NA
```

Logrank test

We will test the null hypothesis that the survival function is the same in both groups:

$$H_0 : S_1(t) = S_2(t) \quad \text{for all } t$$

The test is based on a comparison of the observed and expected number of events (under H$_0$) in the two groups:

O_1 : number of events in group 1

O_2 : number of events in group 2

E_1 and E_2 : “expected” number of events in the two groups
if the survival functions are the same

Define for *both groups combined*:

Times of observed events: $t_1 < t_2 < \cdots < t_d$

m_j : number of events at t_j

$Y(t_j)$: number "at risk" at t_j

Define also:

$Y_k(t_j)$ number "at risk" in group k at t_j

Then:

$$E_k = \sum_{j=1}^{d} m_j \frac{Y_k(t_j)}{Y(t_j)}$$

The test statistic

$$Z = \frac{O_2 - E_2}{\widehat{\mathsf{se}}(O_2 - E_2)}$$

is approximately N(0,1)-distributed under the null hypothesis that the survival functions are the same in the two groups (H_0)

Equivalently:

$$\chi^2 = \frac{(O_2 - E_2)^2}{\widehat{\mathsf{se}}(O_2 - E_2)^2}$$

is approximately chi-squared distributed with 1 df under H_0

The test is called the logrank test

R commands:

```
survdiff(Surv(time,cens)~group)
```

R output :

```
              N      Observed     Expected     (O-E)^2/E    (O-E)^2/V
group=1      22         16          10.6          2.73         4.66
group=2      22         11           16.4         1.77          4.66

 Chisq= 4.7  on 1 degrees of freedom, p= 0.0309
```

The logrank test may be extended to more than two groups

When we compare *K* groups, we get a test statistic with *K* – 1 df

Introducing covariates to explain time to event

We have compared survival curves for two (or more) groups.
Often, we have several covariates for each individual which could influence the time to event.

This is modeled through the hazard function h(t), which with covariates $x_1, x_2 \ldots. , x_p$ is expressed as

$$h(t \mid x_1, x_2 \ldots. , x_p).$$

This means that the hazard for each individual is modeled according to the individual´s covariates.

Example: Time from start of PhD to graduation – suggestions for possible covariates?

Proportional hazards

Consider first the situation with only one covariate

Hazard function for an individual with covariate x

$$h(t \mid x) = h_0(t) \exp(\beta x)$$

The baseline hazard $h_0(t)$ is the hazard for a subject with $x=0$

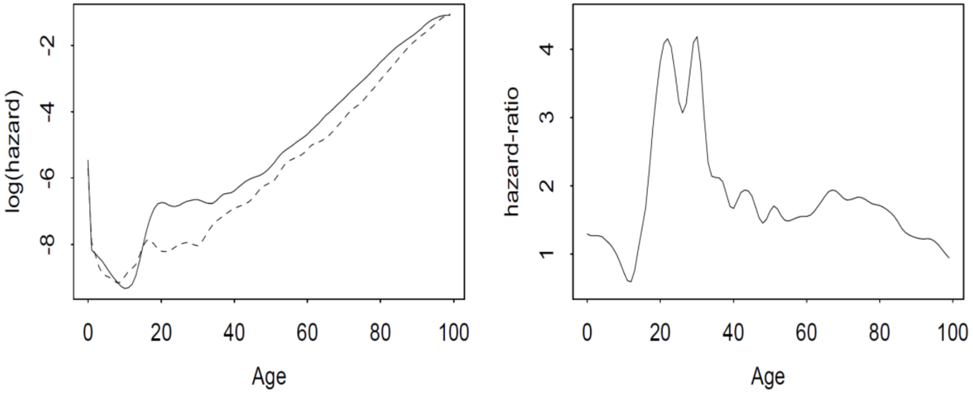
If we consider two subjects with covariate values $x + \Delta$ and x, respectively, their hazard ratio (HR) becomes

$$\frac{h(t \mid x+\Delta)}{h(t \mid x)} = \frac{h_0(t)\exp\big(\beta(x+\Delta)\big)}{h_0(t)\exp(\beta x)} = \exp(\beta\Delta)$$

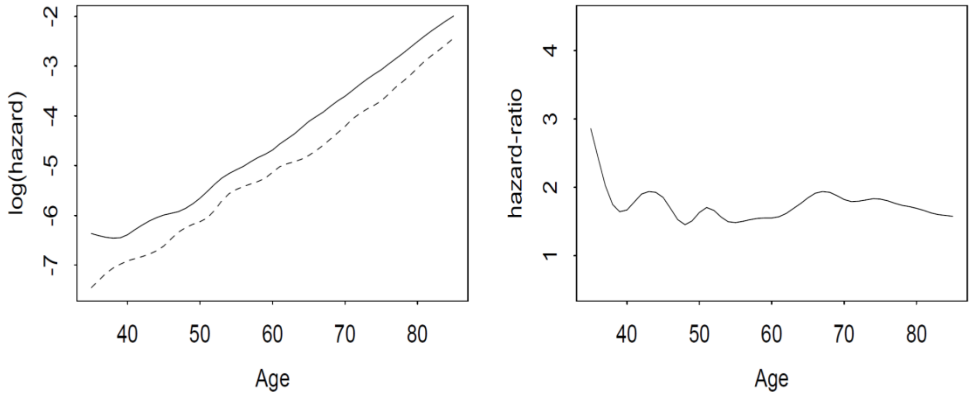
In particular e^{β} is the hazard ratio corresponding to one unit's increase in the value of the covariate

Example: Mortality rates for men and women (from SSB)

Binary covariate x (0=female, 1=male)



A proportional hazards model is *not* valid for 0-100 years



Proportional hazards is a reasonable model for 40-85 years with $HR \approx 1.8$

Example: Melanoma data

205 patients with malignant melanoma were operated during a 15 years period. A number of covariates were recorded at operation
The patients were followed until death or censoring

One covariate of interest was sex ($x=0$ for females; $x=1$ for males)

We fit a proportional hazards model:

$$h(t \mid x) = h_0(t)\exp(\beta x)$$

Estimate

$$\hat{\beta} = 0.662$$

The hazard ratio for males vs females becomes

$$HR = e^{0.662} = 1.94$$

Proportional hazards with several predictors

Consider the situation with several predictors, and assume that the hazard rate for an individual with covariates $x_1, x_2, ..., x_p$ takes the form:

$$h(t \mid x_1, x_2, ..., x_p) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$$

The baseline hazard $h_0(t)$ is the hazard for a subject with all covariates equal to zero

If we consider two subjects with values $x_1 + \Delta$ and x_1, for the first covariate and the same values for all the others, their hazard ratio (HR) becomes

$$\frac{h(t \mid x_1 + \Delta, x_2, ..., x_p)}{h(t \mid x_1, x_2, ..., x_p)} = \exp(\beta_1 \Delta)$$

In particular e^{β_1} is the hazard ratio corresponding to one unit's increase in the value of the first covariate *holding all other covariates constant*

Example: Melanoma data

Consider the covariates:

- $x_1 = 0$ for females; $x_1 = 1$ for males
- $x_2 =$ tumor thickness (mm)

We fit a proportional hazards model:

$$h(t \mid x_1, x_2) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2)$$

Estimates:

$$\hat{\beta}_1 = 0.574 \qquad \text{and} \qquad \hat{\beta}_2 = 0.159$$

Hazard ratios:

$$HR_1 = e^{0.574} = 1.78 \qquad \text{and} \qquad HR_2 = e^{0.159} = 1.17$$

Cox regression

For Cox's regression model the baseline hazard $h_0(t)$ is an *arbitrary* non-negative function

Estimation in Cox's model is based on a *partial likelihood* of the form

$$L(\beta) = \prod_{j=1}^{d} L_j(\beta)$$

where $t_1 < t_2 < .... < t_d$ are the times when events are observed, and the factors $L_j(\beta)$ only depend on the regression parameters (and not on the baseline hazard)

The partial likelihood has similar properties as an ordinary likelihood, and similar methods as for logistic regression and Poisson regression may be used. E.g. confidence intervals, Wald tests and tests based on the difference in deviance (i.e. twice the difference in log likelihoods)

R commands:

```
melanom=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/data/melanoma.dat", header=T)
fit.sex.thickn=coxph(Surv(lifetime,status==1)~factor(sex)+thickn ,data=melanom)
summary(fit.sex.thickn)
```

R output (edited):

```
                coef     exp(coef)        se(coef)        z        Pr(>|z|)
factor(sex)2    0.574      1.776            0.265        2.164      0.0304
thickn          0.159      1.172            0.0327       4.869     1.12e-06

                exp(coef)          exp(-coef)          lower .95          upper .95
factor(sex)2    1.776              0.5632              1.056              2.986
thickn          1.172              0.8529              1.100              1.250

Likelihood ratio test  = 23.82  on 2 df,   p=6.711e-06
Wald test              = 28.77  on 2 df,   p=5.662e-07
Score (logrank) test  = 32.2    on 2 df,   p=1.020e-07
```

(Here the "likelihood ratio test" corresponds to a G-statistic, i.e. the difference between the null and residual deviances in generalized linear models)

```
> melanom
   status      lifetime   ulcer   thickn   sex    age   grthick   logthick
1       4    0.02739726     1       6.76     2     76      3        1.91102300
2       4    0.08219178     2       0.65     2     56      1       -0.43078290
3       2    0.09589041     2       1.34     2     41      1        0.29266960
4       4    0.27123290     2       2.90     1     71      2        1.06471100
5       1    0.50684930     1      12.08     2     52      3        2.49155100
6       1    0.55890410     1       4.84     2     28      2        1.57691500
7       1    0.57534250     1       5.16     2     77      3        1.64093700
8       4    0.63561640     1       3.22     1     60      2        1.16938100
9       1    0.63561640     1      12.88     2     49      3        2.55567600
10      1    0.76438360     1       7.41     1     68      3        2.00283000
11      1    0.80821920     1       4.19     1     53      2        1.43270100
12      4    0.97260270     1       0.16     1     64      1       -1.83258100
13      1    1.05753400     1       3.87     1     68      2        1.35325500
14      1    1.16712300     1       4.84     2     63      2        1.57691500
15      1    1.28493200     1       2.42     1     14      2        0.88376750
16      4    1.35068500     1      12.56     2     72      3        2.53051700
17      1    1.44931500     1       5.80     2     46      3        1.75785800
18      1    1.70137000     1       7.06     2     72      3        1.95444500
19      1    1.72328800     1       5.48     2     95      3        1.70110500
```

The anova-command may be used for Cox regression in the same way as for generalized linear models

R commands:

```
fit.sex=coxph(Surv(lifetime,status==1)~factor(sex) ,data=melanom)
anova(fit.sex,fit.sex.thickn,test="Chisq")
```

R output (edited):

```
Analysis of Deviance Table
Cox model: response is  Surv(lifetime, status == 1)
Model 1: ~ factor(sex)
Model 2: ~ factor(sex) + thickn

        loglik     Chisq          Df       P(>|Chi|)
1      -280.12
2      -271.29    17.673           1       2.623e-05
```

Model fit

When fitting a Cox regression model one should (as for all models!) check that the model fits reasonably well

Checking the fit of a Cox model is, however, somewhat involved and time does not allow us to address this here

A discussion of model fit is given in Section 7.4 in the text book