Lecture 9 – Program

1. Survival data and censoring

2. Survival function and hazard rate

3. Kaplan Meier estimator

4. Log rank test

5. Proportional hazards and Cox regression.

Data structure and basic questions

In this lecture the data have a different form from what we have seen earlier:

subject	time	censoring	covariates
1	y_1	yes/no	$x_{11} \cdots x_{1p}$
2	y_2	yes/no	$x_{21} \cdots x_{2p}$
.	.	.	...
.	.	.	...
.	.	.	...
n	y_n	yes/no	$x_{n1} \cdots x_{np}$

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

Examples

- Time to disease onset
- Time from onset of a disease to death
- Duration of unemployment

Objective as before: Explain variation in time until "event of interest" by variation in $x_1, \cdots, x_p$.

We will often call the time until the event a *survival time*, also when the event in question is something else than death

New aspect: 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 denoted as **censoring**. Also these survival times contain important information and must be included in the analysis.

Example, clinical trials

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

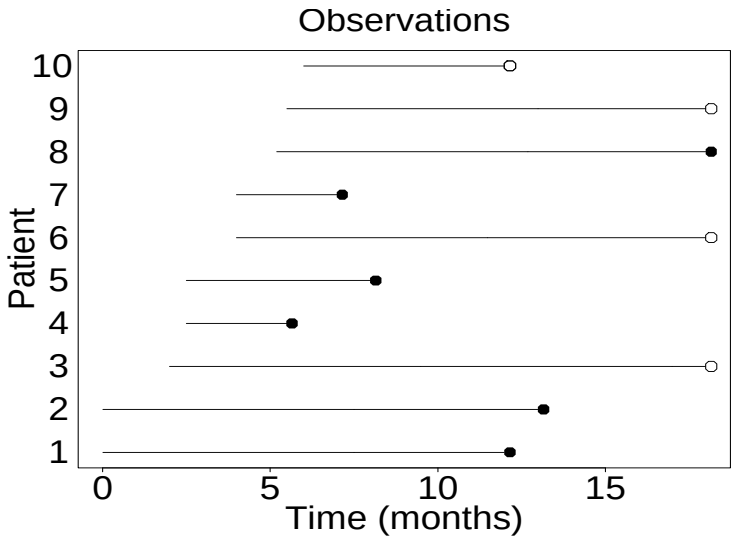
- New patients are diagnosed and included in the study

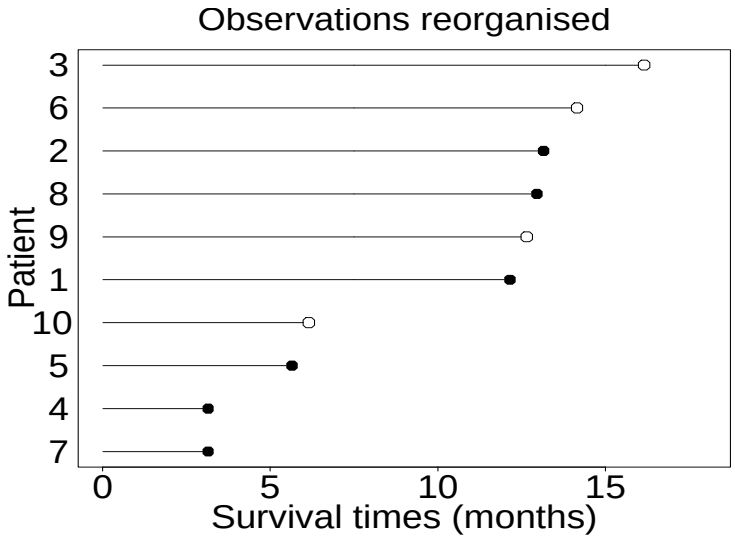
- The patients are followed until
  - death
  - no longer want to participate
  - study concluded

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

Upper figure : calender time

Lower figure: time on study.





Death: • and censoring: ○.

Censored survival times , formally

We introduce:

$$T_i \;=\; \text{survival time individual no } i$$

$$C_i \;=\; \text{time to censoring of individual } i$$

Do not observe T_i (or C_i), but only

$$X_i = \;\min\left(T_i, C_i\right) = \;\text{censored survival time}$$

$$D_i = \begin{cases} 1 & \text{if survival time is observed} \\ 0 & \text{if censoring time is observed} \end{cases}$$

The response for subject i is (X_i, D_i), i.e. a combination of the continuous variable X_i and the binary variable D_i.

Using X_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 for describing the distribution of survival times

- Density $f(t)$: $\quad \mathsf{P}(T \in [t, t+\Delta\rangle\,) \approx f(t)\Delta$
- Survival function $S(t) = P(T > t)$
- Hazard rate $\lambda(t)$:

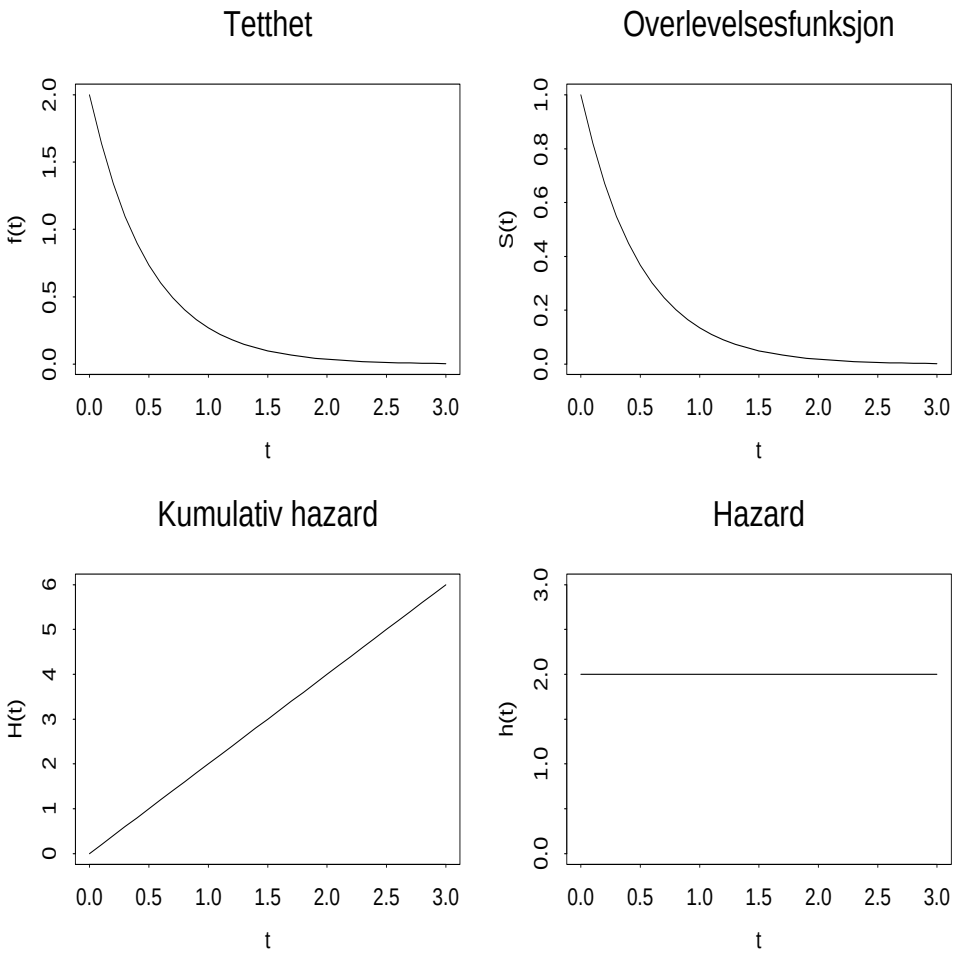
  $\mathsf{P}(T \in [t, t+\Delta\rangle \,|\, T \geq t) \approx \lambda(t)\Delta$
- Cumulative hazard $\Lambda(t) = \int_0^t \lambda(s)ds$

Note the following relations:

- $\lambda(t) = f(t)/S(t)$
- $S(t) = \exp(-\Lambda(t))$
- $\Lambda(t) = -\log(S(t))$
- $F(t) = \mathsf{P}(T \leq t) = 1 - S(t)$

The exponential distribution

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



Estimation of survival function

Example: Survival for patients with chronic active hepatitis (D=dead, A=alive)

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 survival function, contd.

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

For illustration we look at the prednisolone group. 19 of the 22 patients live more than 50 months.

Therefore:

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

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

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.

The 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	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}\cdot 0.456=0.304$

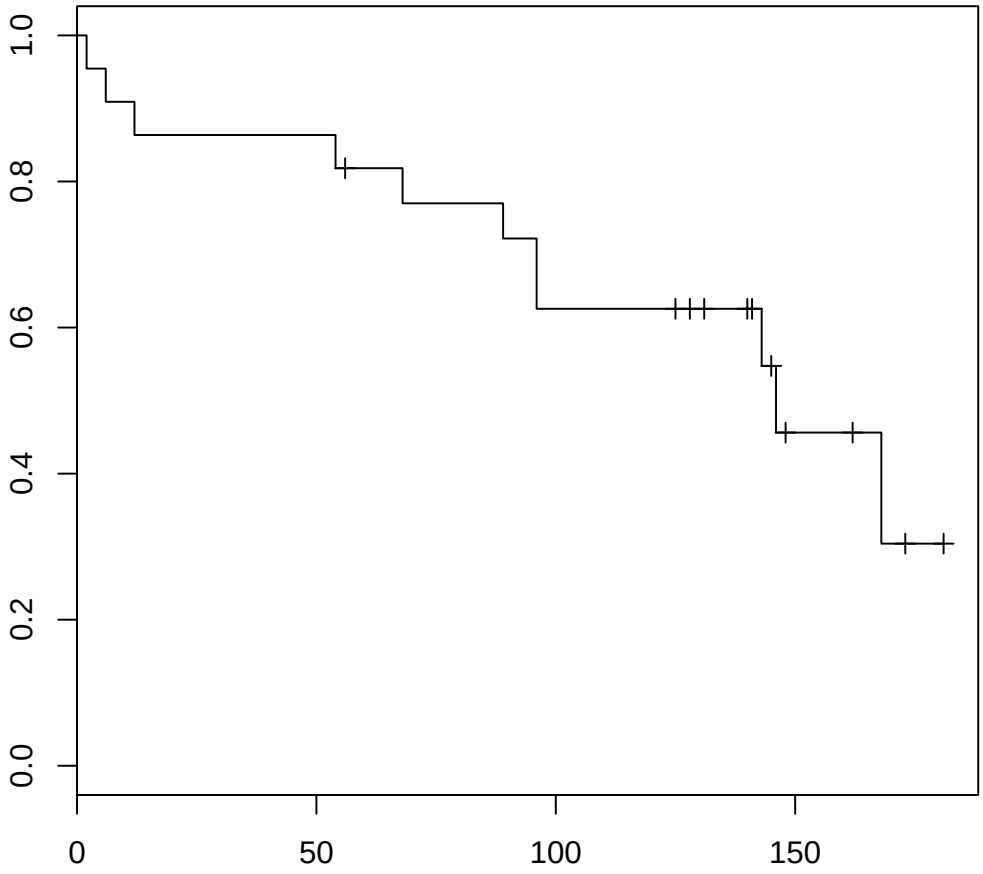
Example, contd. R-code

```
x<-c(2,6,12,54,56,68,89,96,96,125,128,131,140,141,
                  143,145,146,148,162,168,173,181)
d<-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(x,d)~1, conf.type="none")
summary(survpred)
```

```
 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
```

R-plot of Kaplan-Meier estimator

```
plot(survpred)
```



Kaplan-Meier: standard error and confidence intervals

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

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

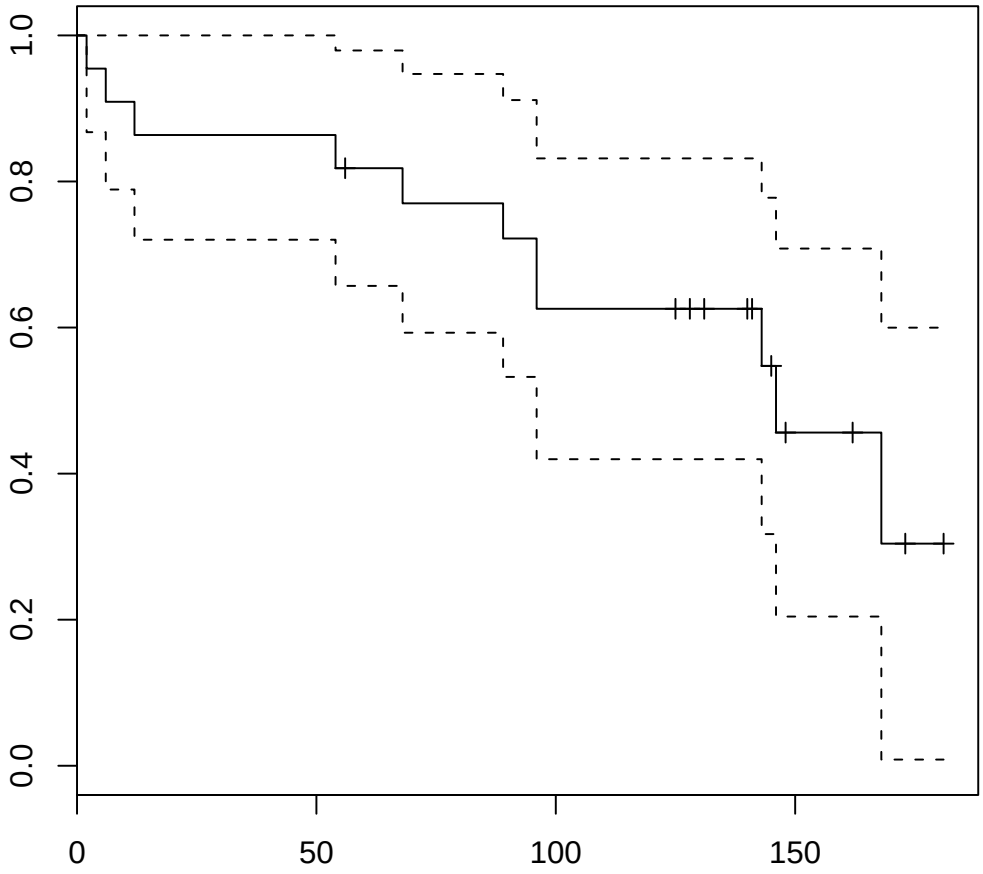
A 95% confidence interval for S(t) is given by

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

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

Example, contd. R-code

```
survpred2<-survfit(Surv(x,d)~1,conf.type="plain")
plot(survpred2)
```



Comparison of two groups

Want to compare survival in two groups (e.g. control and treatment):

Group 1 : (X_{i1}, D_{i1}); $i = 1, ..., n_1$

Group 2 : (X_{i2}, D_{i2}); $i = 1, ..., n_2$

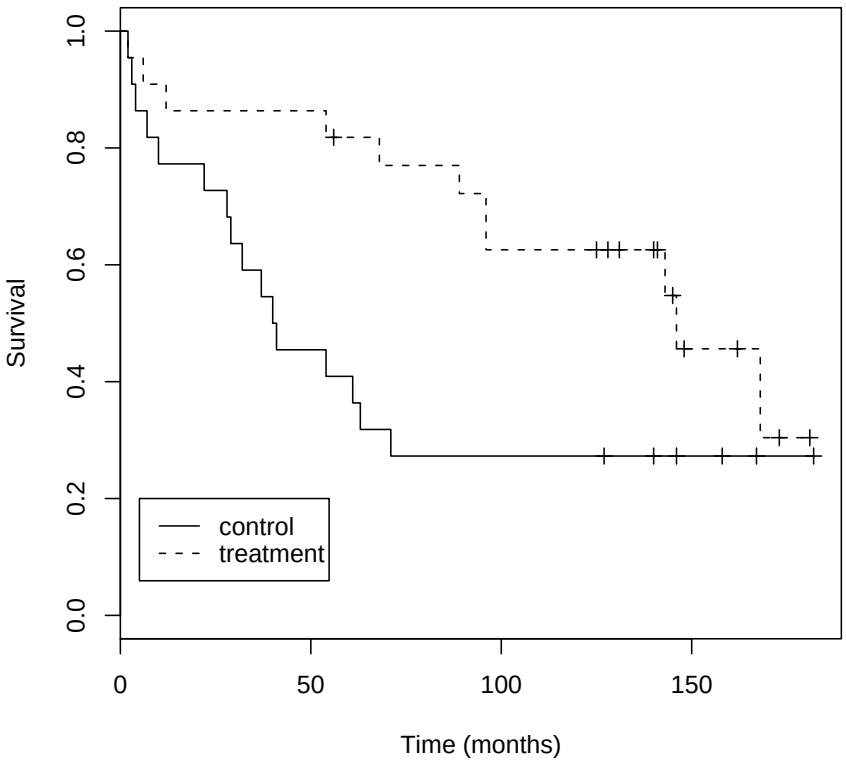
$\widehat{S}_k(t)$: Kaplan-Meier in group k $(k = 1, 2)$

Comparison of the groups:

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

Graphical comparison

```
x<-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)
d<-c(rep(1,16),rep(0,6),c(1,1,1,1,0,1,1,1,1,0,0,0,0,0,
1,0,1,0,0,1,0,0))
gr<-c(rep(1,22),rep(2,22))

survboth<-survfit(Surv(x,d)~gr)
plot(survboth,lty=1:2,xlab="Time (months)",ylab="Survival")
legend(5,0.2,c("control","treatment"),lty=1:2)
```



Log-rank test

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

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

Log-rank test, contd.

The test statistic

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

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

Equivalently:

$$Z^2 = \frac{(O_2 - E_2)^2}{\widehat{\mathrm{se}}(O_2 - E_2)^2}$$

is approximately χ^2_1-distributed under H_0

Log-rank test, contd.

```
survdiff(Surv(x,d)~gr)

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

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

We get a "conservative" version of the log rank test if we compare

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

with the χ^2_1-distribution

(A test is "conservative" if its P-value is too large.)

Log-rank test:

Comparison of $K > 2$ groups

H_0 : survival functions in all groups are equal

$O_k =$ number of events in group k

$E_k =$ expected number of events in group k

Log rank test statistic: $Z^2 \sim \chi^2_{K-1}$ under H_0.

The test statistic is based on a comparison of the O_ks and E_ks. Its expression is a bit complicated, but it is computed by statistical software

We get a "conservative" version of the log rank test if we compare

$$X^2 = \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j}$$

with the χ^2_{K-1}-distribution

Proportional hazards: one covariate

Hazard rate for subject with covariate x:

$$\lambda_x(t) = \lambda_0(t)\exp(\beta x)$$

The *baseline hazard* $\lambda_0(t)$ is the hazard for a subject with $x = 0$.

Interpretation: Hazard rate ratio (or loosely, relative risk, RR):

$$\text{RR} = \frac{\lambda_{x_1}(t)}{\lambda_{x_0}(t)} = \exp(\beta(x_1 - x_0))$$

In particular with x binary (i.e. values 0 and 1):

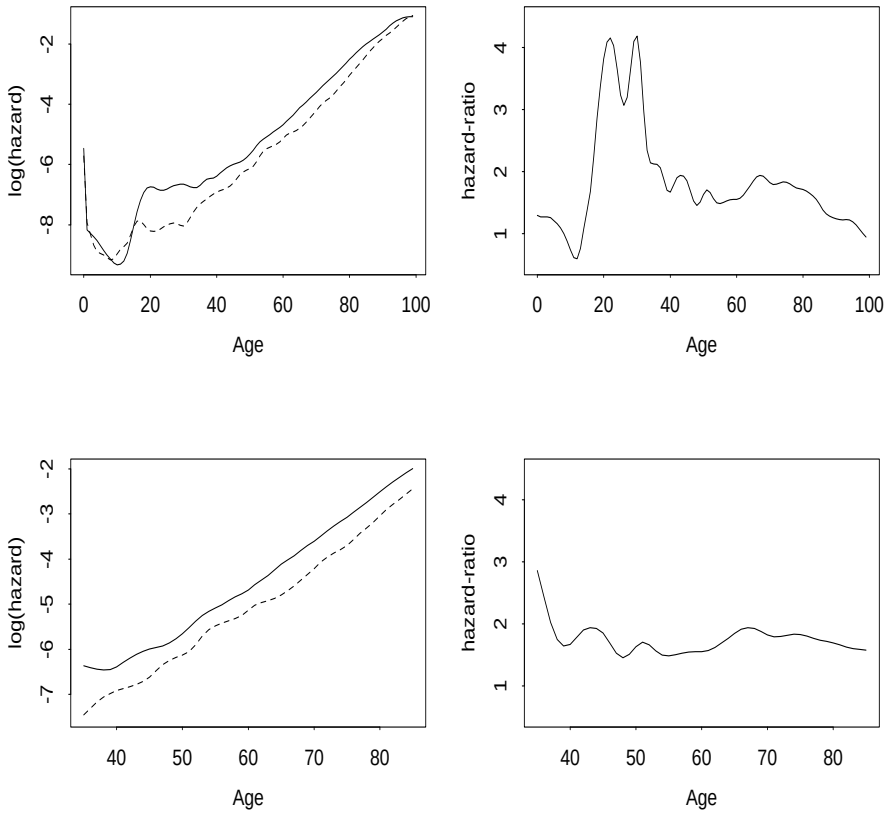
$$\text{RR} = \frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\beta)$$

Example: Mortality rates among men and women (Statistics Norway, 2000, smoothed)

Binary covariate: x indicator of men.

Propositional hazards model is **not** valid in age interval 0-100 years

Propositional hazards model roughly valid in interval 40-85 years with RR ≈ 1.8.



Example: Melanoma data

205 patients with malignant melanoma operated 1962-77. Followed until death or censoring (cf. Exercise)

Proportional hazards model:

$$\lambda_x(t) = \lambda_0(t)\exp(\beta x)$$

(i) Qualitative covariate:

x = indicator of ulceration

$RR = \exp(\beta)$ is hazard ratio between those with and without ulceration

(ii) Quantitative covariate:

x_1 = tumor thickness (in mm) subject 1,
x_2 = thickness subject 2 = $x_1 + 1$ mm:

$RR = \exp(\beta)$ = hazard ratio for 1 mm difference in thickness

Proportional hazards: several covariates

Hazard rate for individual with covariate vector $x = (x_1, x_2, ..., x_p)$:

$$\lambda_x(t) = \lambda_0(t) \exp\{\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p\}$$

The baseline hazard $\lambda_0(t)$ is the hazard for an individual with $x_1 = \cdots = x_p = 0$.

Interpretation: Hazard ratio (RR)

Another subject with $x' = (x'_1, x'_2, ..., x'_p)$ where $x'_1 = x_1 + 1$ and $x'_j = x_j$ otherwise:

$$\text{RR}_1 = \frac{\lambda_{x'}(t)}{\lambda_x(t)} = \exp\{\beta_1\}$$

Example: Melanoma data

Model:

$$\lambda_x(t) = \lambda_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$

x_1 = sex (M=1, F=0)

x_2 = indicator of ulceration

x_3 = thickness (in mm)

Consider $x = (x_1, 0, x_3)$ and $x' = (x_1, 1, x_3)$

Then

$$RR = \frac{\lambda_{x'}(t)}{\lambda_x(t)} = \exp(\beta_2)$$

is the hazard ratio between those with and without ulceration **adjusted** for sex and thickness.

Cox's regression model

For Cox's regression model the baseline hazard $\lambda_0(t)$ is an *arbitrary* non-negative function

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

The partial likelihood is of the form

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

where $t_1 < t_2 < \cdots < t_d$ 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)

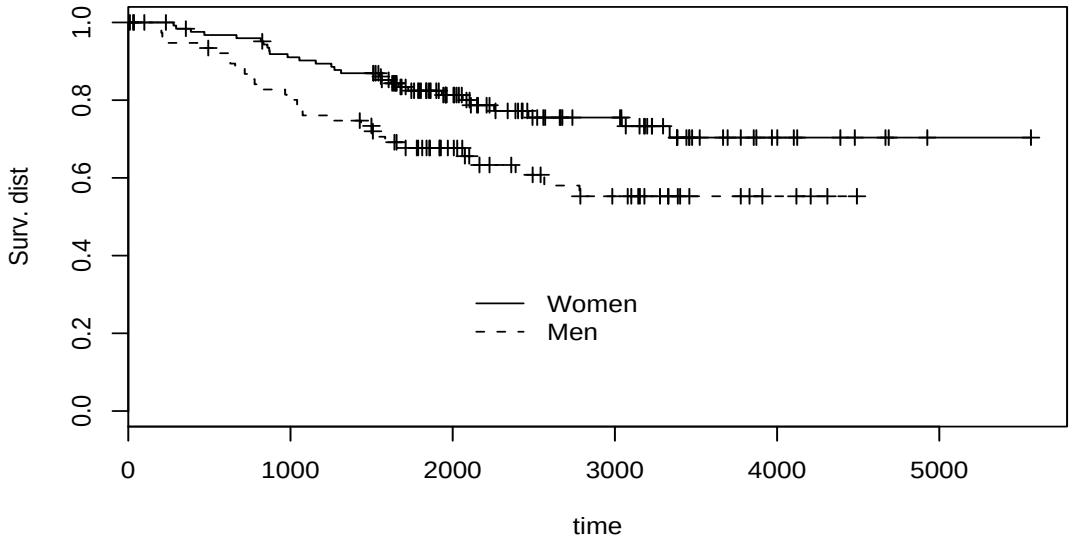
The partial likelihood has similar properties as an ordinary likelihood, and similar methods as for logistic regression and Poisson regression may be used. E.g. confidence intervals, Wald tests and tests based on the difference in deviance (i.e. twice the difference in log likelihoods)

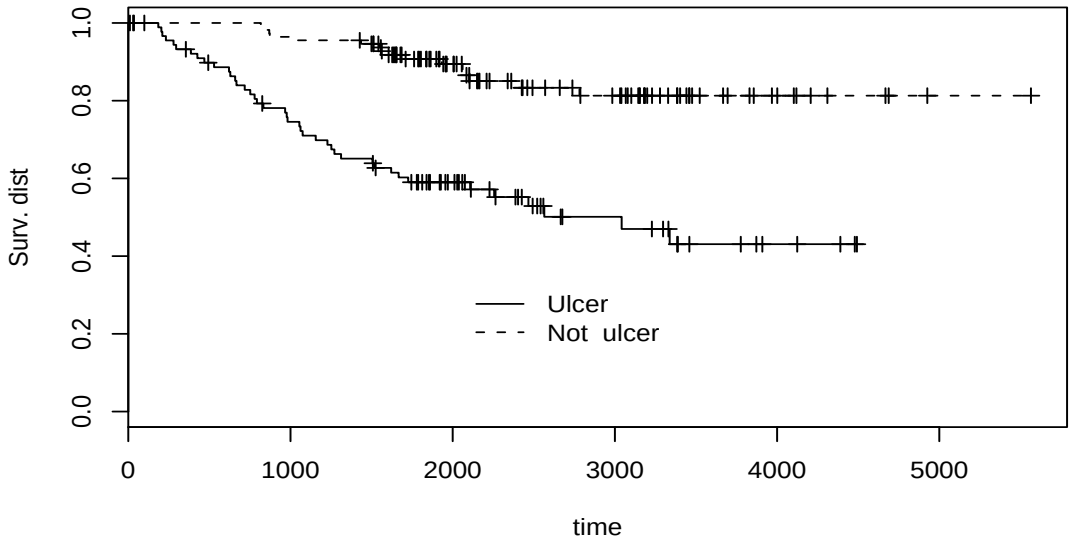
Example, melanoma data

Data:

- status: 1=died from melanoma, 2=censored, 4=died of other reasons
- lifetime: observed time to death/censoring (in years)
- ulcer: 1=ulcer, 2= no ulcer
- thickn: thickness in mm/100
- sex: 1=woman, 2=man
- age: age when operated

Example, contd.





Example, contd.

```
mod1<-coxph(Surv(lifetime,status==1)~thickn+ulcer+sex+age)

mod1

                  coef exp(coef) se(coef)     z       p
melanom$tumor   0.1089     1.115   0.0377  2.89 0.00390
melanom$ulcer -1.1645     0.312   0.3098 -3.76 0.00017
melanom$sex     0.4328     1.542   0.2674  1.62 0.11000
melanom$age     0.0122     1.012   0.0083  1.47 0.14000

Likelihood ratio test=41.6  on 4 df, p=2e-08  n= 205
```

Here the "likelihood ratio test" is the same as the "null deviance" in the output for a generalized linear model (e.g. logistic regression or Poisson regression)

Example, contd.

```
summary(mod1)
```

```
          coef exp(coef) se(coef)     z       p
thickn  0.1089     1.115   0.0377  2.89 0.00390
ulcer  -1.1645     0.312   0.3098 -3.76 0.00017
sex     0.4328     1.542   0.2674  1.62 0.11000
age     0.0122     1.012   0.0083  1.47 0.14000

       exp(coef) exp(-coef) lower .95 upper .95
thickn     1.115      0.897     1.036     1.201
ulcer      0.312      3.204     0.170     0.573
sex        1.542      0.649     0.913     2.604
age        1.012      0.988     0.996     1.029

Rsquare= 0.184   (max possible= 0.937 )
Likelihood ratio test= 41.6  on 4 df,   p=2e-08
Wald test            = 39.4  on 4 df,   p=5.72e-08
Score (logrank) test = 46.7  on 4 df,   p=1.79e-09
```

Example, contd.

```
mod0<-coxph(Surv(lifetime,status==1)~thickn+ulcer+sex)
mod0


         coef exp(coef) se(coef)     z       p
thickn  0.113     1.120   0.0379  2.99 0.00280
ulcer  -1.167     0.311   0.3115 -3.75 0.00018
sex     0.459     1.583   0.2668  1.72 0.08500

Likelihood ratio test=39.4  on 3 df, p=1.44e-08  n= 205


anova(mod0,mod1,test="Chisq")

Analysis of Deviance Table

Model 1: Surv(lifetime,status==1)~thickn+ulcer+sex
Model 2: Surv(lifetime,status==1)~thickn+ulcer+sex+age
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       202     527.01
2       201     524.78   1     2.23      0.14
```