Event History/Survival Analysis

Janez Stare

Faculty of Medicine, Ljubljana, Slovenia

Ljubljana, 2018

Some literature

1. David Collett. Modelling Survival Data in Medical Research. Chapman and Hall 2003.
2. David W. Hosmer, Stanley Lemeshow , Susanne May. Applied Survival Analysis. Wiley-Interscience 2008.

Characterization of processes we are interested in

1. there is a collection of units, each moving among a finite number of states;
2. changes (events) may occur at any point in time;
3. measurements are often (almost always) censored.
4. there are factors, possibly time-dependent, influencing the events.
5. effects of covariates may change in time.

In short ...

In event history analysis we are interested in time to a certain event. Or, putting it differently, we are interested in time between two states.

Examples of events are:

- job changes
- regime changes
- promotions
- marriages, divorces
- time in office
- crimes, arrests
- equipment failures
- deaths, remissions ...

Examples of events are:

- job changes
- regime changes
- promotions
- marriages, divorces
- time in office
- crimes, arrests
- equipment failures
- deaths, remissions ...

For now we will assume there can be only ONE event per subject, all events being of the SAME TYPE.

Other names for Event History/Survival Analysis are

- Failure Time Data Analysis
- Reliability Analysis

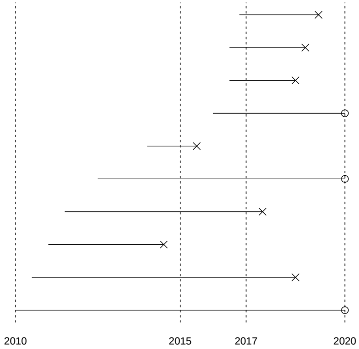
Censoring

Often times are not fully observed.

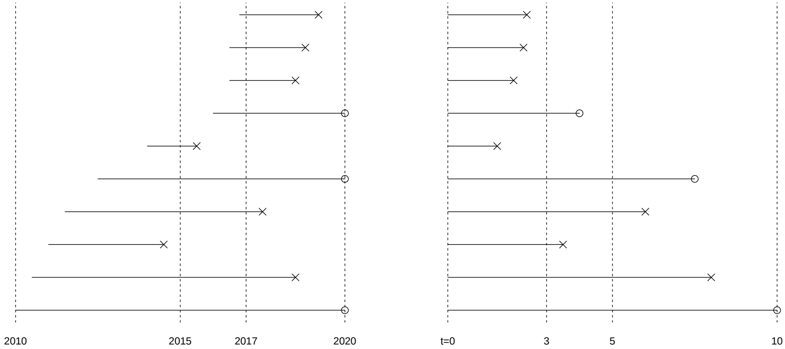
- the study may end before the event occurs
- a person may be lost during observational period
- another event may prevent the event of interest to occur (e.g. death in a car accident of a diseased person)

Such observations are called **censored**.

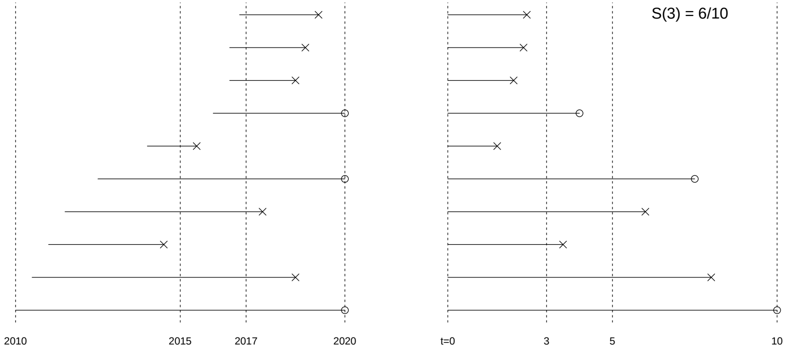
A typical situation



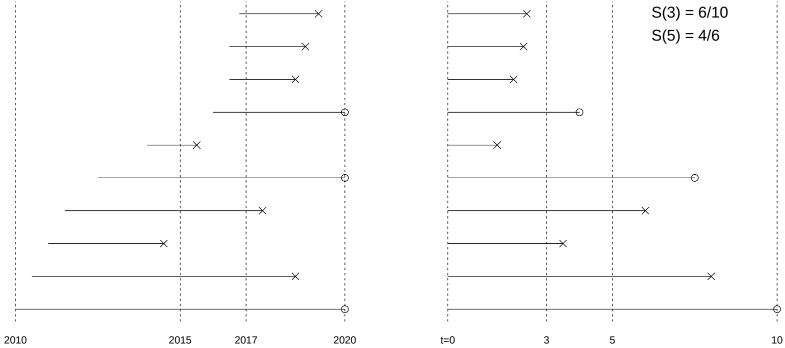
A typical situation



A typical situation



A typical situation



Censoring

The need for special methods comes (mostly) from **censoring.** There are different types of censoring.

T - time variable of interest (time to event)

C - censoring variable.

Right censoring: we only see $\min(T_i, C_i)$

Types of censoring

- Type I: censoring time fixed in advance (all C_i equal)
- Type II: data are censored after r events (when a given proportion fails)
- Type III: random censoring (most common)

Why is censoring a problem

With censored data we can't even calculate a simple arithmetic mean (in the usual way) or draw a histogram.

Why is censoring a problem

With censored data we can't even calculate a simple arithmetic mean (in the usual way) or draw a histogram.

So, the situation seems pretty much hopeless.

Why is censoring a problem

With censored data we can't even calculate a simple arithmetic mean (in the usual way) or draw a histogram.

So, the situation seems pretty much hopeless.

Luckily, it is not, although it took some time to come up with methods that deliver want we want.

Why is censoring a problem

With censored data we can't even calculate a simple arithmetic mean (in the usual way) or draw a histogram.

So, the situation seems pretty much hopeless.

Luckily, it is not, although it took some time to come up with methods that deliver want we want.

What do we want?

The Goals of Event History Analysis

The Goals of Event History Analysis

1. Estimation of the distribution (survival) function.

The Goals of Event History Analysis

1. Estimation of the distribution (survival) function.
2. Comparison of distribution (survival) functions.

The Goals of Event History Analysis

1. Estimation of the distribution (survival) function.
2. Comparison of distribution (survival) functions.
3. Finding association between the outcome (survival time) and prognostic variables.

Survival function - T continuous

- T a non-negative continuous random variable representing the survival times in a population
- $F(t)$ distribution function of T
- $f(t)$ density of T.

The **(cumulative) distribution function** is

$$F(t) = P(T \leq t) = \int_0^t f(x)dx.$$

The distribution function gives the **proportion of people having the event** until time t.

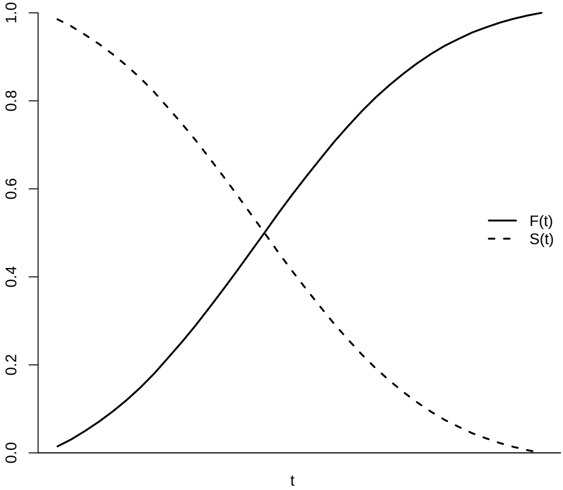
Survival function - T continuous

In EHA we are looking at **survival function**

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx.$$

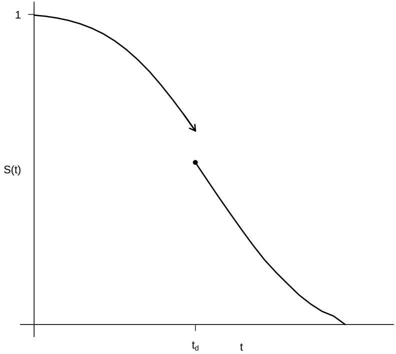
The survival function gives the **proportion of people NOT having the event** (e.g. surviving) until time t.

Survival function and distribution function



Survival function - T continuous

The function $S(t)$ is continuous from right.



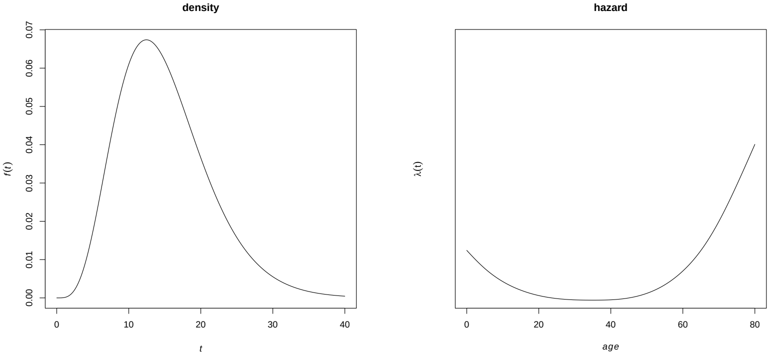
Hazard function (transition rate) - T continuous

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

Note that this is different from the definition of the density which is

$$f(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t}.$$

Do you distinguish between the probability in the definition of $f(t)$ and the conditional probability in $\lambda(t)$?


density
f(t)
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
10
20
30
40
t
hazard
λ(t)
0
20
40
60
80
age

Relations among $S(t)$ and $\lambda(t)$ - T continuous

Remembering that $P(A|B) = P(AB)/P(B)$ we can deduce

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d \ln S(t)}{dt} \tag{1}$$

and from this

$$S(t) = e^{-\int_0^t \lambda(x) dx}. \tag{2}$$

So, if we have the hazard, we have the survival function!!

Survival function - T discrete

T is now a discrete random variable taking values

$$a_1 < a_2 < \cdots$$

The corresponding probability function is

$$f(a_i) = P(T = a_i), \qquad i = 1{,}2, \ldots$$

and the survival function is

$$S(t) = \sum_{j|a_j>t} f(a_j)$$

(not a very useful expression!)

Hazard function (transition rate) - *T* discrete

Hazard is defined as the conditional probability of the event at a_i given that the event had not occurred before a_i. So

$$\lambda_i = P(T = a_i | T \geq a_i)$$

Cumulative hazard is

$$\Lambda(t) = \sum_{j|a_j \leq t} \lambda_j.$$

Note: cumulative hazard is a sum of conditional probabilities, but it is NOT a probability. It can be VERY large!

Relations among $S(t)$, $f(t)$ and $\lambda(t)$ - T discrete

From the definition of the hazard function we have

$$\lambda_i = P(T = a_i | T \geq a_i) = \frac{f(a_i)}{S(a_i^-)}$$

where we write $S(a^-)$ for $\lim_{t \to a^-} S(t)$.

The connection between the survival function and the hazard function is much more important. Let $a_j \leq t < a_{j+1}$. Then

$$\begin{aligned} S(t) &= P(T > a_1, T > a_2, \ldots, T > a_j) \\ &= P(T > a_1 | T \geq a_1) P(T > a_2 | T \geq a_2), \ldots, P(T > a_j | T \geq a_j) \\ &= \prod_{i | a_i \leq t} (1 - \lambda_i) \end{aligned} \tag{3}$$

Measures of central tendency - the mean

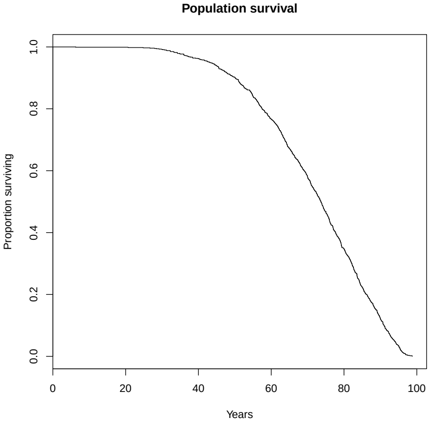
By definition **the mean** is

$$E(T) = \int_0^\infty tf(t)dt,$$

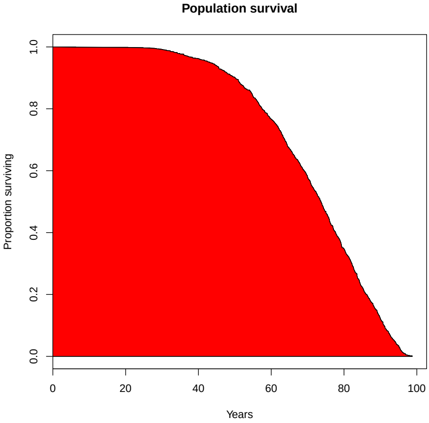
and with some effort we can show that

$$E(T) = \int_0^\infty S(t)dt$$

Measures of central tendency - the mean



Measures of central tendency - the mean



Measures of central tendency - the mean

The mean residual time is

$$mrt(u) = E(T - u | T > u),$$

for which we have

$$mrt(u) = \frac{\int_u^\infty S(t)dt}{S(u)}.$$

Measures of central tendency - the mean

The mean residual time is

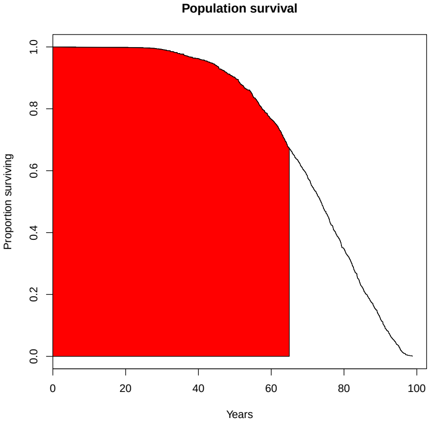
$$mrt(u) = E(T - u | T > u),$$

for which we have

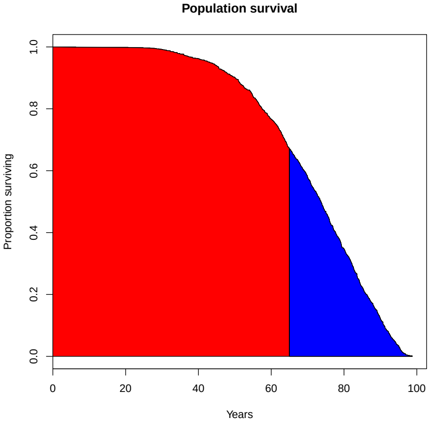
$$mrt(u) = \frac{\int_u^\infty S(t)dt}{S(u)}.$$

Nobel Prize winners, Academy award winners, famous conductors, Slovenian pension reform ...

Measures of central tendency - the mean



Measures of central tendency - the mean



mean = 71.58221

restricted mean = 61.09878

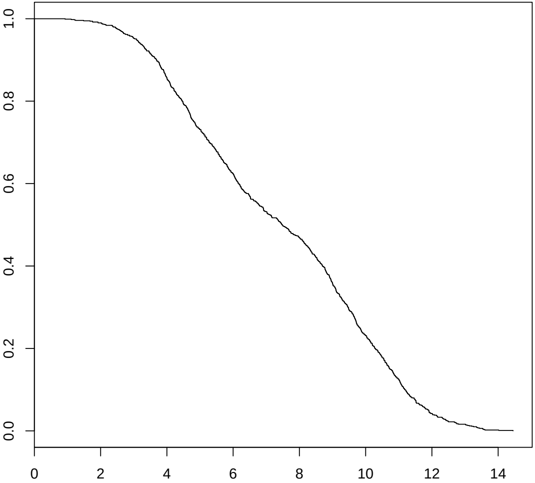
residual mean = 15.60035

Measures of central tendency - the median

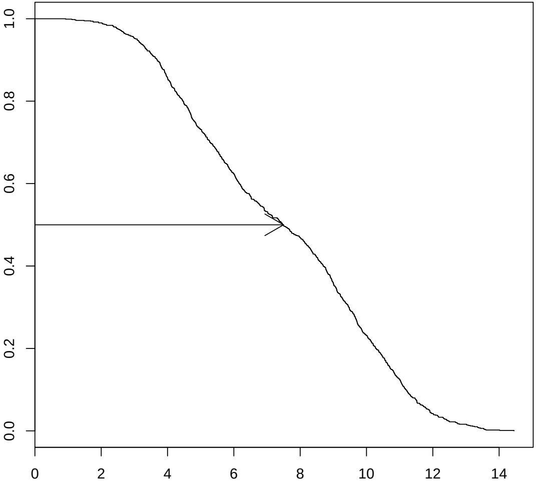
The median is the value τ, for which

$$S(\tau) = 0.5.$$

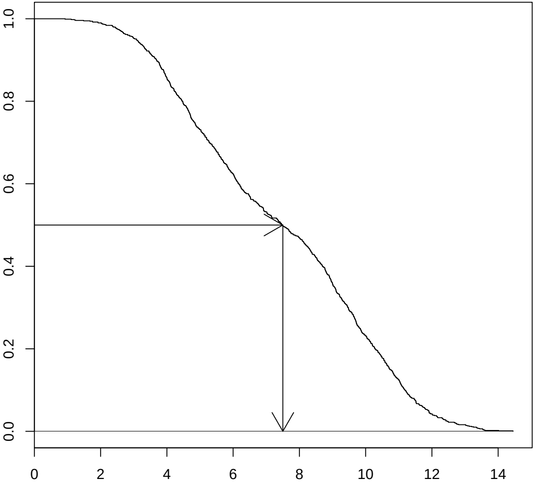
Measures of central tendency - the median



Measures of central tendency - the median



Measures of central tendency - the median



Likelihood function

Let C be a random variable representing censoring times. Denote the density of T by f and its survival function by S. Every individual thus has survival time T_i and censoring time C_i. We observe the pair (Y_i,δ_i), where

$$Y_i = \min(T_i,C_i) \quad \text{and} \quad \delta_i = \begin{cases} 1 & \text{if } T_i \le C_i \\ 0 & \text{if } C_i < T_i \end{cases}.$$

Likelihood function

If we observed n individuals, we have n realizations of the random variable Y, giving values y_i, and we can try to write the likelihood of this event.

If $\delta_i = 1$ (event at y_i), then at y_i we require high density $f(y_i)$.

If $\delta_i = 0$ (no event at y_i), then at y_i we require high probability of that person still not having the event (e.g. still alive), meaning that his survival, $S(y_i)$, function should be as high as possible.

Likelihood function

Both requirements can be united in the requirement of maximizing the expression

$$f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}.$$

The product of these values for all i gives us the likelihood of the observed event

$$L = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}. \tag{4}$$

and taking into account (1), we get

$$L = \prod_{i=1}^{n} \lambda(y_i)^{\delta_i} S(y_i). \tag{5}$$

Parametric models - Exponential distribution

The simplest function to assume for the hazard function is a constant,

$$\lambda(t) = \lambda > 0$$

on the domain of T.

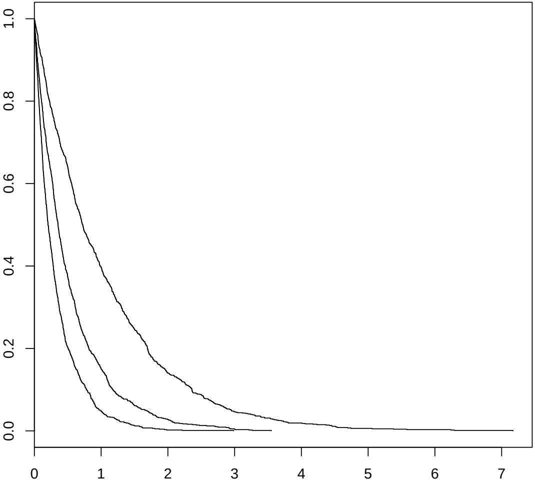
It follows that the conditional probability of an event in a given interval does not depend on the beginning of the interval. This property is sometimes called the *lack of memory* property.

The survival function, the density and the distribution function are

$$S(t) = e^{-\lambda t}, \quad f(t) = \lambda e^{-\lambda t} \quad \text{and} \quad F(t) = 1 - e^{-\lambda t}.$$

This means that T has an *exponential distribution*.

Graphs of exponential survival function for $\lambda = 1,2,3$



Parametric models - Weibull distribution

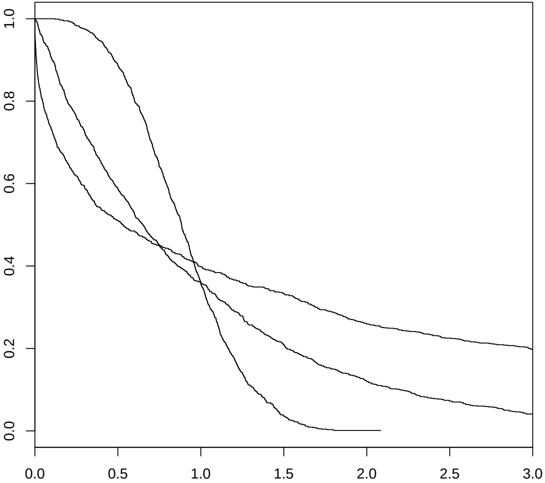
Exponential distribution is not very useful because of the constant hazard assumption. It is much more realistic to assume that the hazard is either decreasing or increasing. Such a hazard can be modelled as

$$\lambda(t) = \lambda\gamma(\lambda t)^{\gamma-1},$$

where λ and γ are positive constants. For $\gamma < 1$ the hazard is monotonically decreasing, and for $\gamma > 1$ it is increasing. The survival function is

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

Graphs of the Weibull survival function for $\gamma = 0.5, 1, 3$



Parametric models - Weibull distribution

From

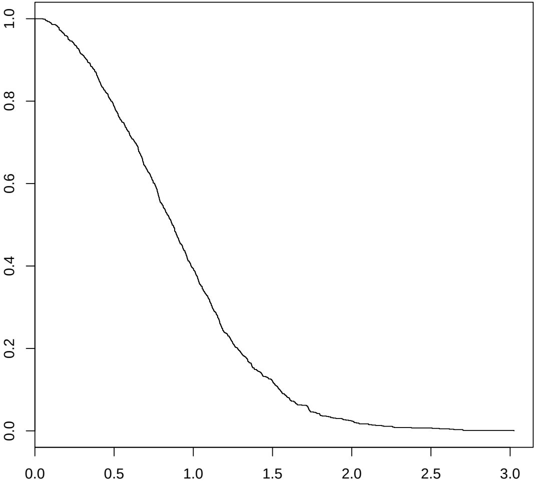
$$S(t) = e^{-(\lambda t)^{\gamma}}$$

we see that

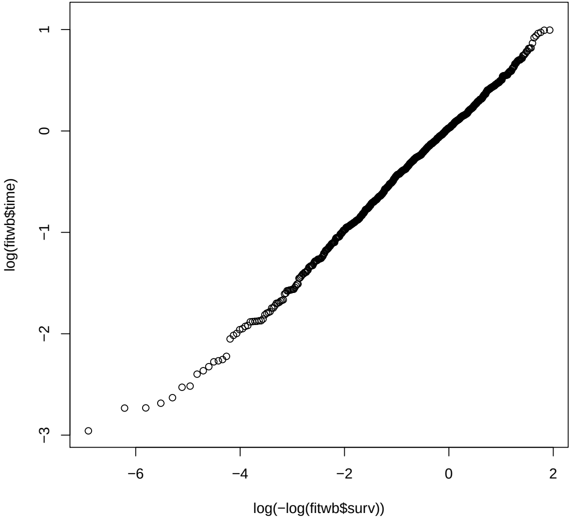
$$\log[-\log S(t)] = \gamma(\log t + log\lambda).$$

If we have an estimate of $S(t)$, then the graph of $\log[-\log \hat{S}(t)]$ versus logarithm of time should be approximately a straight line.

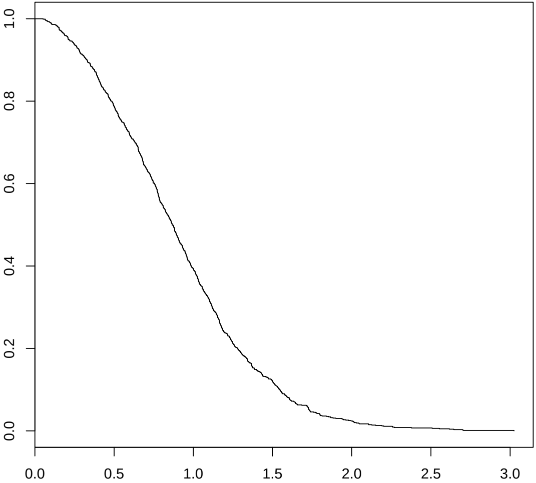
Graphical check of the fit for the Weibull distribution



Graphical check of the fit for the Weibull distribution



But where do curves like this come from?



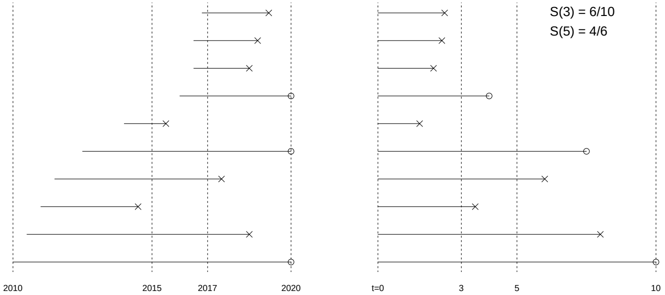
Estimating the survival function

If there was no censoring, we could easily estimate the survival function at time *t* by

$$\hat{S}(t) = \frac{\text{Number of cases for which T} > \text{t}}{\text{Number of all cases}}$$

But what if there is censoring? Do we just throw those observations away? It should be obvious that this would mean underestimating the survival function (or overestimating proportion of events), since we would use the data on those that suffered the event (died, say), but not on those censored (even if they stayed event-free (lived) for long).

Back to our example



We need something better!

Estimating the survival function parametric approach

One possibility is to assume a certain parametric distribution for the survival function and then estimate the parameters using the maximum likelihood method.

We will briefly look at the simplest possibility.

Estimating the survival function - the exponential model

Assume that T has exponential distribution and that we have n measured times t_i, of which some are censored. We will estimate the parameter λ using the maximum likelihood method.

For the exponential distribution we have (5)

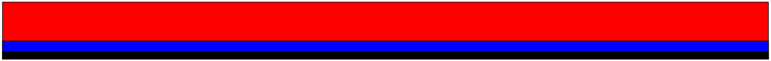
$$L = \prod_{i=1}^{n} (\lambda e^{-\lambda t_i})^{\delta_i} (e^{-\lambda t_i})^{1-\delta_i} = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i}.$$

Taking logarithms, differentiating with respect to λ and equating the result to 0 (extreme values only occur at points where the derivatives are 0!), we see that the maximum likelihood estimate of λ is

$$\hat{\lambda} = \frac{d}{\sum y_i},$$

where d is the number of all events (deaths).

Illustration of the $\hat{\lambda}$ estimator in the exponential model

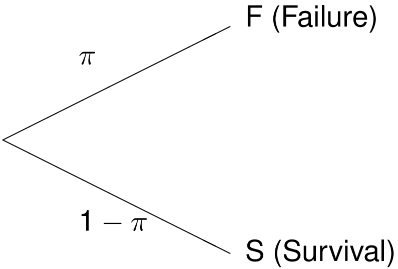


Estimating the survival function

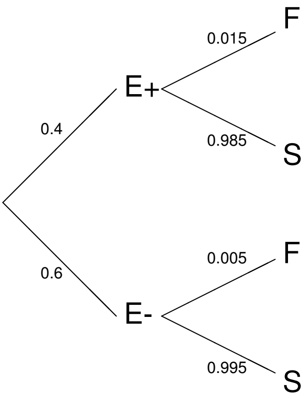
The exponential model is of course very simple, and most of the time unrealistic in describing actual distributions.

There are many other parametric possibilities, much more flexible than the exponential model, but guessing the right distribution is usually hard, or even impossible. It can be safely said that distributions, typically found in political and social sciences (and also in medicine), do not have nice parametric forms. It is much better to use a *nonparametric* alternative.

Estimating the survival function



Estimating the survival function



Probability

0.006

0.003

Estimating the survival function

More formally, we are using the formula for the probability of a product of events.

Estimating the survival function

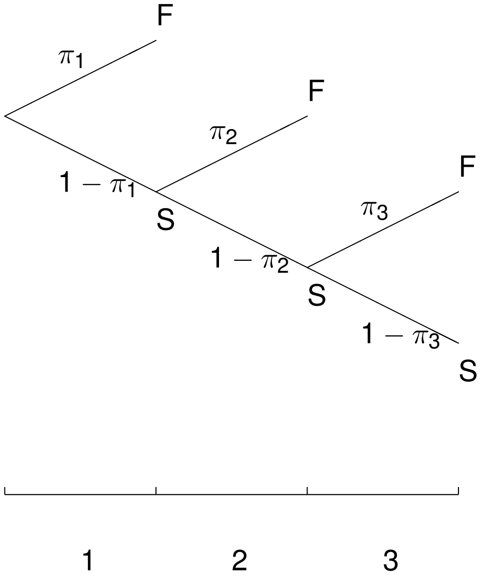
More formally, we are using the formula for the probability of a product of events.

If A and B are two events, then the probability of the product AB is

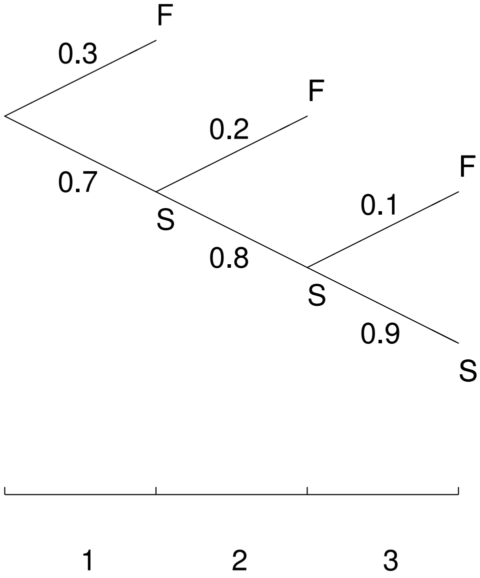
$$P(AB) = P(A)P(B|A)$$

where $P(B|A)$ is the conditional probability of B given A.

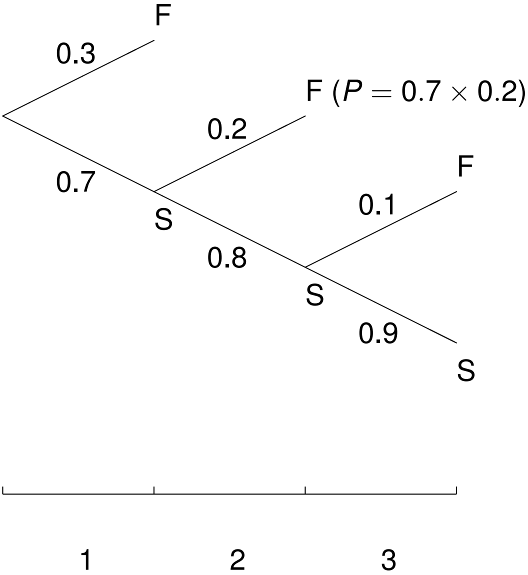
Estimating the survival function



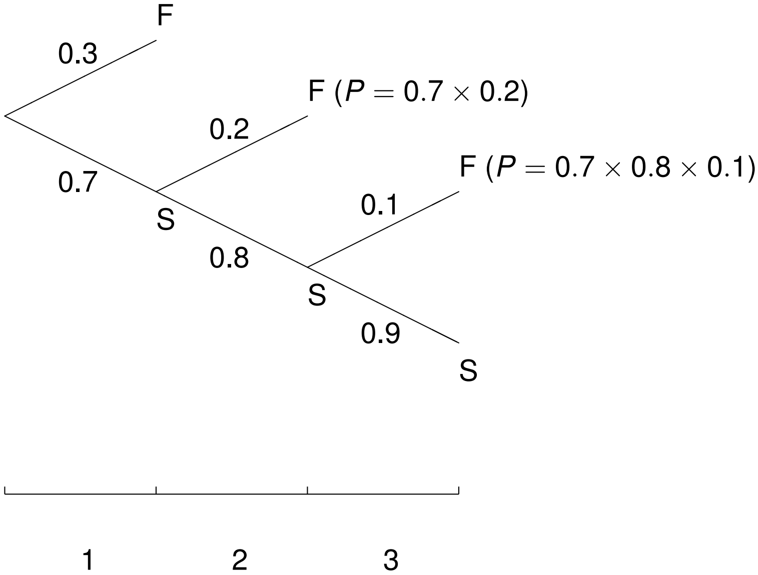
Estimating the survival function



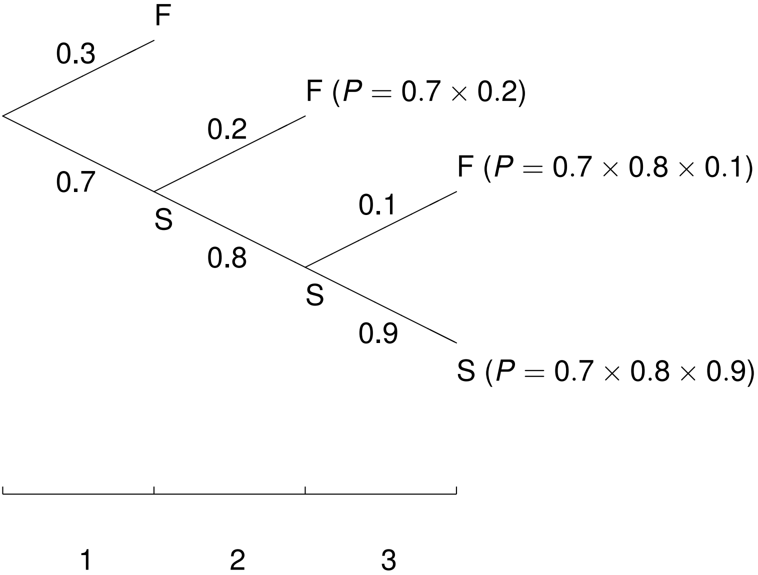
Estimating the survival function



Estimating the survival function



Estimating the survival function


F
0.3
F ($P = 0.7 \times 0.2$)
0.2
F ($P = 0.7 \times 0.8 \times 0.1$)
0.7
S
0.1
0.8
S
0.9
S ($P = 0.7 \times 0.8 \times 0.9$)
1
2
3

Estimating the survival function

We can use this principle in calculating survival even with **censored data**.

Estimating the survival function

We can use this principle in calculating survival even with **censored data**.

We first divide the time scale into intervals in such a way that events or censorings occur on the boarders of the intervals.

Estimating the survival function

We can use this principle in calculating survival even with **censored data**.

We first divide the time scale into intervals in such a way that events or censorings occur on the boarders of the intervals.

Then we calculate (conditional) probabilities of surviving each interval and obtain probability of surviving any time by simply multiplying the probabilities of survival up to the given point in time.

Estimating the survival function

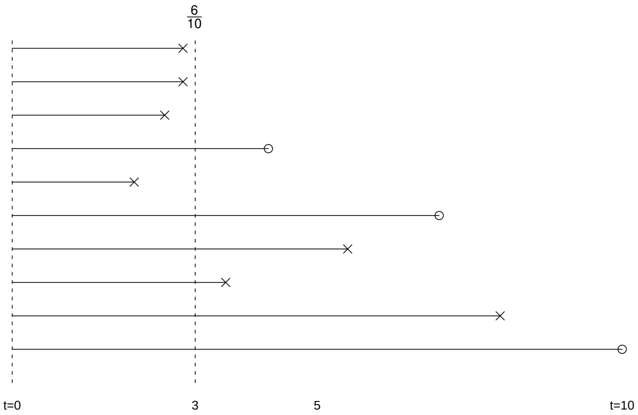
We can use this principle in calculating survival even with **censored data**.

We first divide the time scale into intervals in such a way that events or censorings occur on the boarders of the intervals.

Then we calculate (conditional) probabilities of surviving each interval and obtain probability of surviving any time by simply multiplying the probabilities of survival up to the given point in time.

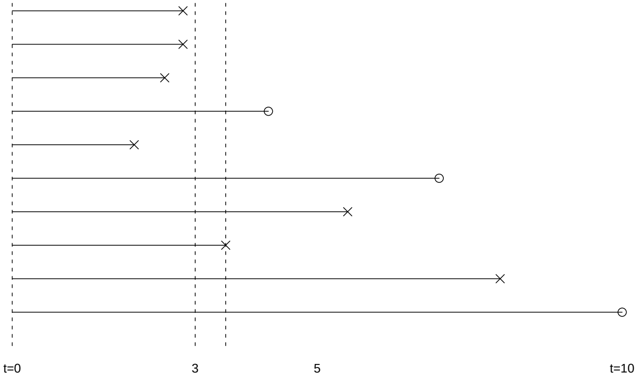
The method is named after **Kaplan and Meier**.

The Kaplan-Meier method



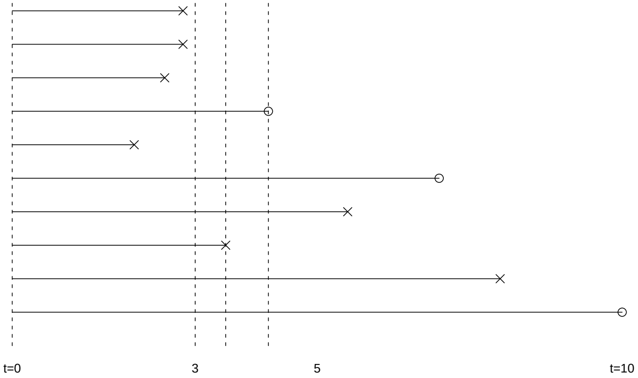
The Kaplan-Meier method

$$\frac{6}{10} \cdot \frac{5}{6} = \frac{5}{10}$$



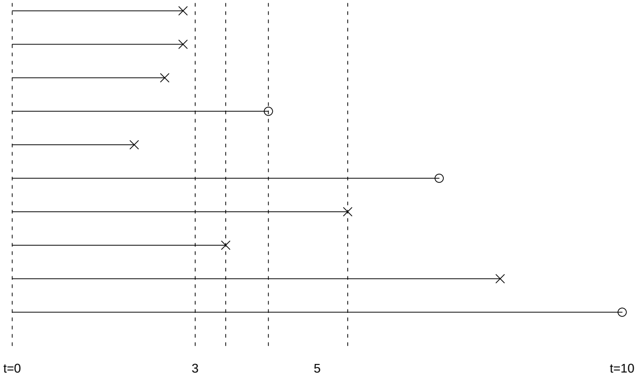
The Kaplan-Meier method

$$\frac{5}{10} \cdot \frac{5}{5} = \frac{5}{10}$$

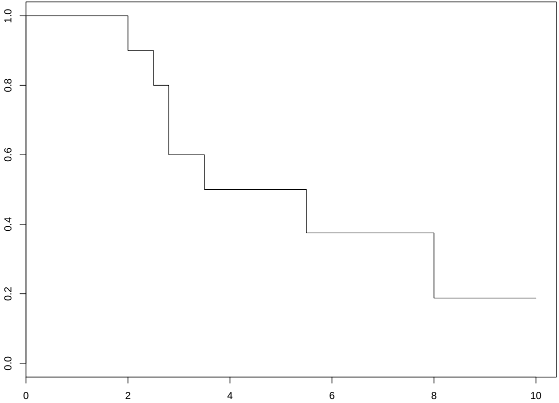


The Kaplan-Meier method

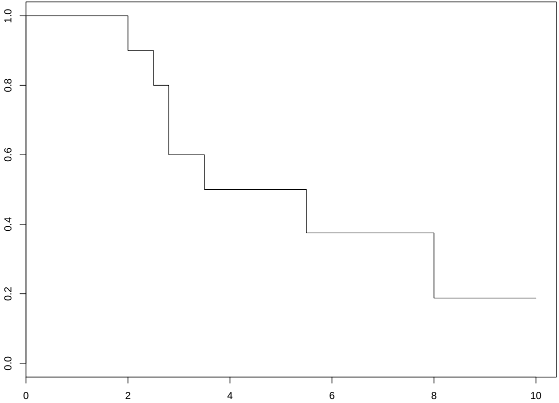
$$\frac{5}{10} \cdot \frac{3}{4} = \frac{3}{8}$$



The Kaplan - Meier curve for our example



The Kaplan - Meier curve for our example



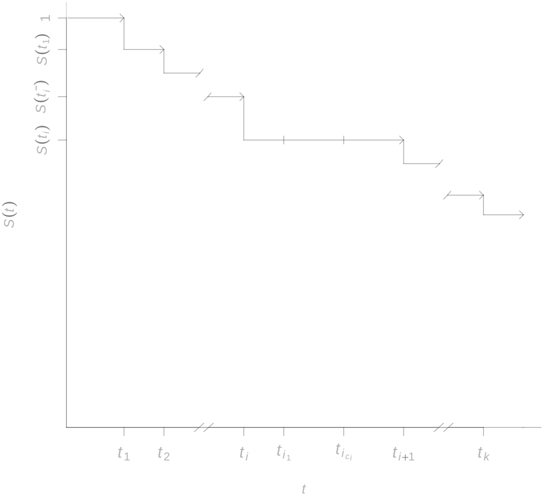
What do flat regions on the curve mean?

Kaplan-Meier method more formally

Let us now try to estimate $S(t)$ without assuming any particular functional form.

Let $0 < t_1 < t_2 < \cdots < t_k < \infty$ be measured times of events in a sample of size n. Obviously $k \leq n$. Let d_i be the number of events at t_i and let c_i represent the number of censored observations in the interval $[t_i, t_{i+1}), \quad i = 0, \ldots, k$, and the exact censoring times being $t_{i1}, \ldots, t_{ic_i}$. We have $t_0 = 0$ in $t_{k+1} = \infty$.

Kaplan-Meier method more formally



Kaplan-Meier method more formally

Since we only have information about event times at t_i, the estimated function will have to be a (right continuous) step function, with steps at measured times of events. Probability of an event at t_i is

$$P(T = t_i) = S(t_i^-) - S(t_i),$$

and the probability of not experiencing an event before or at t_i is $S(t_i)$. Since $S(t)$ is a step function, we have $S(t_{ij}) = S(t_i)$ for $j = 1, \ldots, c_i$, meaning that the function does not change at the censoring times. We can then write the probability (and therefore the likelihood) of the observed values as

$$L = \prod_{i=0}^{k} \left\{ [S(t_i^-) - S(t_i)]^{d_i} \prod_{j=1}^{c_i} S(t_{ij}) \right\} = \prod_{i=0}^{k} [S(t_i^-) - S(t_i)]^{d_i} S(t_i)^{c_i}$$

Kaplan-Meier method more formally

Remembering that $S(t_i^-) = \prod_{j=1}^{i-1}(1-\lambda_j)$ and $S(t_i) = \prod_{j=1}^{i}(1-\lambda_j)$ and bearing in mind that the first factor in L is equal to 1 (why?), we get

$$
\begin{aligned}
L &= \prod_{i=1}^{k}\left[\lambda_i^{d_i}\prod_{j=1}^{i-1}(1-\lambda_j)^{d_i}\prod_{j=1}^{i}(1-\lambda_j)^{c_i}\right] \\
&= \prod_{i=1}^{k}\lambda_i^{d_i}(1-\lambda_i)^{c_i}\prod_{j=1}^{i-1}(1-\lambda_j)^{d_i+c_i} \\
&= \prod_{i=1}^{k}\lambda_i^{d_i}(1-\lambda_i)^{n_i-d_i}. \qquad (6)
\end{aligned}
$$

In the last simplification we used the fact that $n_i = \sum_{j=i}^{k}(d_i + c_i)$ and that in this sum we are missing d_i.

Kaplan-Meier method more formally

Maximizing L (taking logarithms, taking derivatives with respect to λ_i, equaling those derivatives to 0 and solving the respective equations) gives us the following estimates of λ_i

$$\hat{\lambda}_i = \frac{d_i}{n_i}$$

so that

$$\hat{S}(t) = \prod_{i|t_i \le t} \frac{n_i - d_i}{n_i}. \tag{7}$$

Kaplan-Meier method - final result

Estimates of λ_i are

$$\hat{\lambda}_i = \frac{d_i}{n_i}$$

and estimator of the survival function is

$$\hat{S}(t) = \prod_{i|t_i \leq t} \frac{n_i - d_i}{n_i}. \tag{8}$$

Kaplan-Meier method more formally

We first used ‘common sense’ to get to (8). The formula (8) simply says that to calculate the probability of surviving past *t* we have to multiply the probabilities of surviving the intervals which we used to partition the time scale. This partition is done in such a way that each interval contains only one *t* (one day) at which an event was observed.

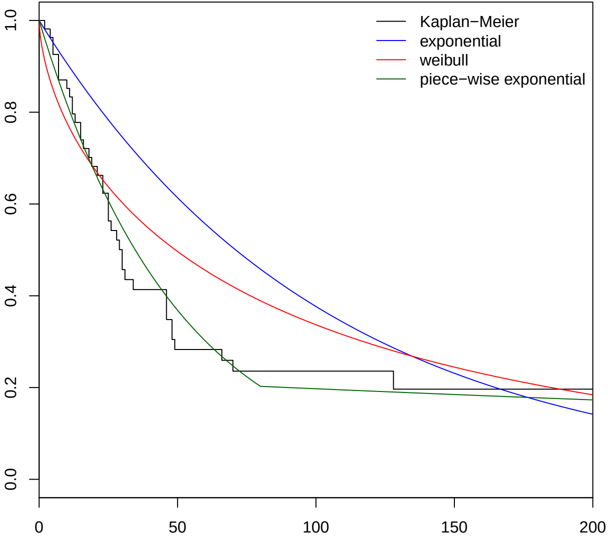
Example: estimation of survival curves

We analyze data on the duration of United Nation (UN) peacekeeping missions from 1948 to 2001.

There were 54 peacekeeping missions, 15 were still ongoing at the end of the study (censoring).

The figure shows the Kaplan-Meier survival curve along with the exponential model, Weibull model and piece-wise exponential model.

Example: estimation of survival curves



Exercise: estimation of survival curves

Do Labs Section 2 (Kaplan Meier)

Variance of $\hat{S}(t)$

We will use the **delta method** to calculate the variance of $\hat{S}(t)$. The method helps in calculating $\mathrm{var}(g(Y))$ when we know $\mathrm{var}(Y) = \sigma^2$ and $E(Y) = \mu$. Then we have (more or less precisely)

$$\mathrm{var}(g(Y)) \approx (g'(\mu))^2\sigma^2.$$

Variance of $\hat{S}(t)$

We start with the equation

$$\hat{S}(t) = \prod_{i|t_i \le t} (1 - \hat{\lambda}_i).$$

Taking logarithms

$$\ln \hat{S}(t) = \sum_{i|t_i \le t} \ln(1 - \hat{\lambda}_i)$$

we calculate

$$\begin{aligned} \mathrm{var}(\ln \hat{S}(t)) &= \sum_{i|t_i \le t} \left(\frac{1}{1 - \hat{\lambda}_i} \right)^2 \mathrm{var}(\hat{\lambda}_i) = \sum_{i|t_i \le t} \left(\frac{1}{1 - \hat{\lambda}_i} \right)^2 \frac{\hat{\lambda}_i(1 - \hat{\lambda}_i)}{n_i} \\ &= \sum_{i|t_i \le t} \frac{\hat{\lambda}_i}{(1 - \hat{\lambda}_i) n_i} = \sum_{i|t_i \le t} \frac{d_i}{(n_i - d_i) n_i}. \end{aligned}$$

Variance of $\hat{S}(t)$

Since $S(t) = e^{\ln(S(t))}$, we have (again using the delta method)

$$\mathrm{var}(\hat{S}(t)) = [\hat{S}(t)]^2 \,\mathrm{var}(\ln \hat{S}(t)) = [\hat{S}(t)]^2 \sum_{i|t_i \le t} \frac{d_i}{(n_i - d_i)n_i}. \tag{9}$$

Formula (9) is called the **Greenwood's formula.**

Variance of $\hat{S}(t)$

The confidence interval (at a given t) for $\hat{S}(t)$ is then:

$$[\hat{S}(t) - z_\alpha \mathrm{se}(\hat{S}(t)), \hat{S}(t) + z_\alpha \mathrm{se}(\hat{S}(t))],$$

where $\mathrm{se}(\hat{S}(t))$ is the standard error obtained with the Greenwood's formula.

Illustration - survival after myocardial infarction

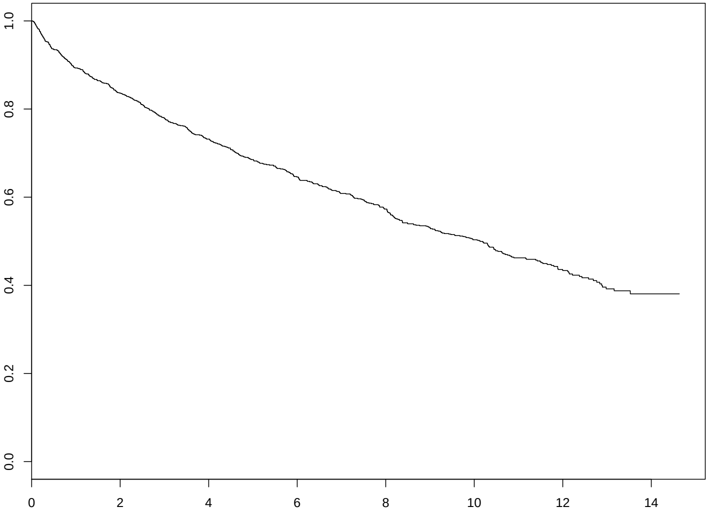
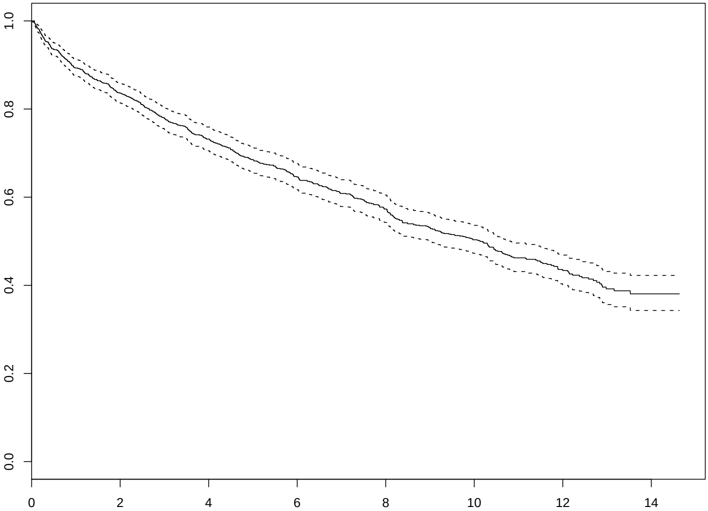


Illustration - survival after myocardial infarction



Other ways of calculating $\mathrm{var}(\hat{S}(t))$

The confidence interval obtained using the Greenwood formula is symmetric and can therefore be greater than 1 or smaller than 0. We can avoid this in the following way.

We introduce $L(t) = \ln(-\ln(S(t)))$ and calculate the confidence interval for $L(t)$, again using the delta method. Say this is $[\hat{L}(t) - A, \hat{L}(t) + A]$. Since $S(t) = e^{-e^{L(t)}}$, the confidence interval for $\hat{S}(t)$ is

$$[e^{-e^{\hat{L}(t)+A}}, e^{-e^{\hat{L}(t)-A}}],$$

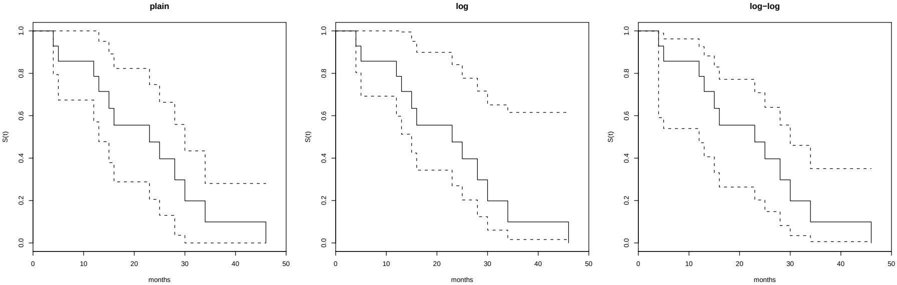
which can also be written as

$$[\hat{S}(t)^{e^{A}}, \hat{S}(t)^{e^{-A}}].$$

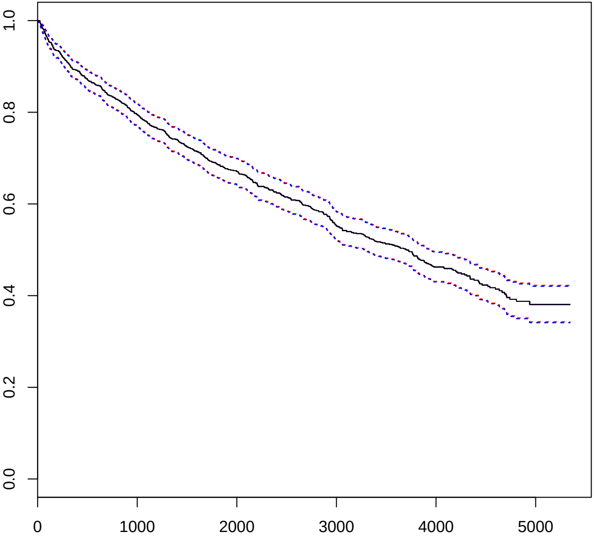
This interval is always between 0 and 1.

Note: what we did above was to calculate the confidence intervals for $\hat{S}(t)$ at any **given** t! This is **NOT** the same as a confidence interval for the whole $\hat{S}(t)$!

Other ways of calculating $\mathrm{var}(\hat{S}(t))$



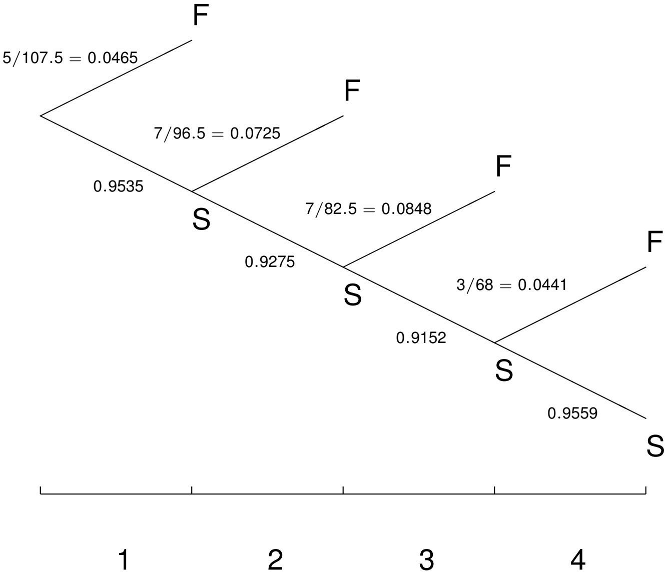
Different confidence intervals for a larger dataset



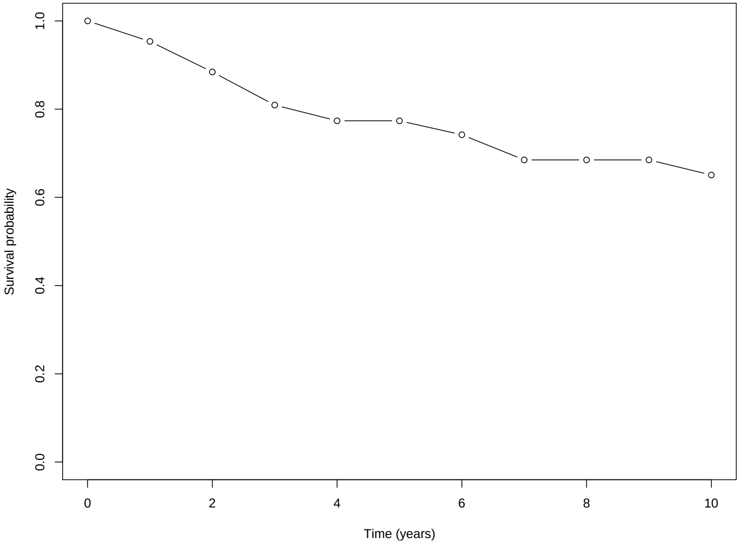
Life tables

Year	N	D	L
1	110	5	5
2	100	7	7
3	86	7	7
4	72	3	8
5	61	0	7
6	54	2	10
7	42	3	6
8	33	0	5
9	28	0	4
10	24	1	8

Life tables


F
5/107.5 = 0.0465
F
7/96.5 = 0.0725
0.9535
S
F
7/82.5 = 0.0848
0.9275
S
F
3/68 = 0.0441
0.9152
S
0.9559
S
1
2
3
4

Plotting survival curves from life tables



Nelson - Aalen estimate of the survival function

$S(t)$ can be estimated by first estimating $\Lambda(t)$, the cumulative hazard, and then calculating $\hat{S}(t)$. The cumulative hazard is obtained as a sum of hazards

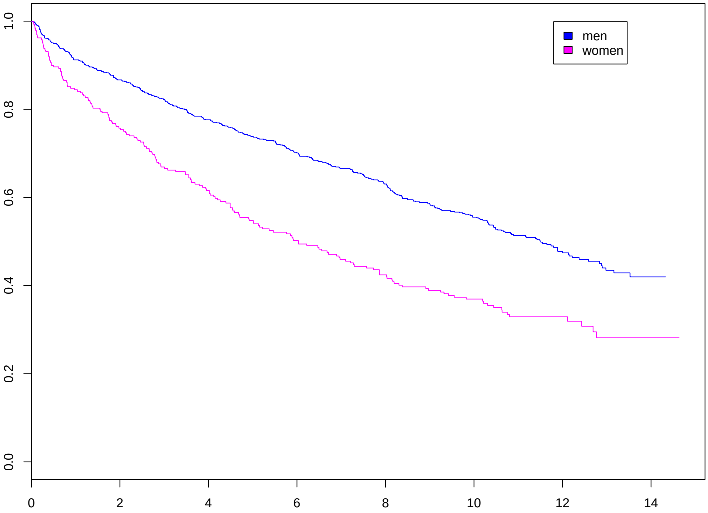
$$\hat{\Lambda}(t) = \sum_{i|t_i \leq t} \hat{\lambda}_i = \sum_{i|t_i \leq t} \frac{d_i}{n_i}$$

and the estimate of the survival function is then

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

We will skip the calculation of the variance of this estimate (for which we would again use the delta method). But let me point out that this estimate, contrary to the Kaplan Meier estimate, will never be 0.

Comparison of survival curves



The statistical test for the null hypothesis (that the two samples come from the same population) is based on the usual idea:

The statistical test for the null hypothesis (that the two samples come from the same population) is based on the usual idea:

Under the null hypothesis we expect that people will be dying proportionally to the group size.

The statistical test for the null hypothesis (that the two samples come from the same population) is based on the usual idea:

Under the null hypothesis we expect that people will be dying proportionally to the group size.

Based on this we calculate the **expected number of deaths** in each group and compare it to the **observed number of deaths.**

The statistical test for the null hypothesis (that the two samples come from the same population) is based on the usual idea:

Under the null hypothesis we expect that people will be dying proportionally to the group size.

Based on this we calculate the **expected number of deaths** in each group and compare it to the **observed number of deaths.**

The name of the test is **log rank test** for some strange reasons.

The statistical test for the null hypothesis (that the two samples come from the same population) is based on the usual idea:

Under the null hypothesis we expect that people will be dying proportionally to the group size.

Based on this we calculate the **expected number of deaths** in each group and compare it to the **observed number of deaths.**

The name of the test is **log rank test** for some strange reasons.

The *p*-value for the log rank test for the previous example is $3.1 \cdot 10^{-9}$.

Comparison od survival functions

Let us first remember this (why? - well, you'll see!): if an urn contains b black balls and c balls of some other colour, then the probability that in a random sample of n balls k of them will be black is

$$P_n(B = k) = \frac{\binom{b}{k}\binom{c}{n-k}}{\binom{N}{n}},$$

where $N = b + c$. This distribution is called the *hypergeometric distribution*. If a random variable B is distributed according to the hypergeometric distribution, then its expected value and variance are

$$E(B) = np, \qquad \mathrm{var}(B) = \frac{npq(N-n)}{N-1}.$$

Here we used $p = b/N$ and $q = c/N$.

Comparison od survival functions

Now let's look at a 2 × 2 contingency table.

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{..}$

Dots denote summation over relevant indices.

If we can assume the marginal frequencies fixed, then by choosing one of the values n_{11}, n_{12}, n_{21} and n_{22} we also fix the other three. In other words, the probability distributions of the random variables N_{11}, N_{12}, N_{21} and N_{22} are all the same.

Comparison od survival functions

And how is N_{11} distributed? We can look at the problem in the following way: when sampling $n_{1.}$ persons from $n_{..}$ persons (without replacement), the probability to choose n_{11} persons from $n_{.1}$ persons and the rest, that is $n_{12} = n_{1.} - n_{11}$ persons, from $n_{.2}$ persons, is equal to

$$P_{n_{1.}}(N_{11} = k) = \frac{\binom{n_{.1}}{n_{11}}\binom{n_{.2}}{n_{12}}}{\binom{n_{..}}{n_{1.}}},$$

With this notation, the expected value and the variance of N_{11} are

$$E(N_{11}) = n_{1.}\frac{n_{.1}}{n_{..}}, \qquad \mathrm{var}(N_{11}) = \frac{n_{1.}n_{2.}n_{.1}n_{.2}}{n_{..}^2(n_{..}-1)}.$$

The denominator in the expression of variance has the sums of rows and columns.

Comparison od survival functions

How can the above help in comparing survival curves? Assume we were observing two groups of people of sizes n_1 and n_2, and assume also that d_1 and d_2 people have experienced the event (d is usually used for *death*, but the event can of course be anything). Let's put this data in a table:

$$\begin{array}{cc|c} d_1 & n_1 - d_1 & n_1 \\ d_2 & n_2 - d_2 & n_2 \\ \hline d & n - d & n \end{array}$$

So we have

$$E(D_1) = \frac{n_1 d}{n}$$

and

$$\mathrm{var}(D_1) = \frac{n_1 n_2 d(n-d)}{n^2(n-1)}.$$

Comparison od survival functions

If the null hypothesis is true, we have

$$\chi^2_{MH} = \frac{[d_1 - n_1 d/n]^2}{\frac{n_1 n_2 d(n-d)}{n^2(n-1)}} \sim \chi^2_1$$

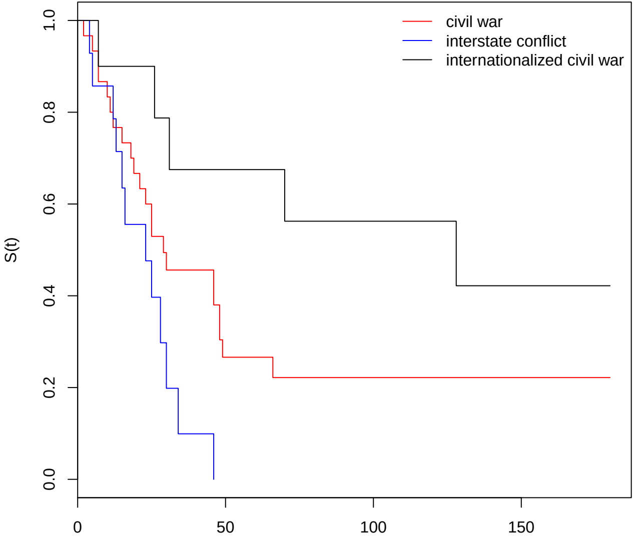
MH stands for Mantel in Haenszel, two statisticians who are credited with this test.

A table like the one above can be constructed at every event time. If we index times with *j* and there are *k* different times, the test of the null hypothesis is

$$\chi^2_{logrank} = \frac{\left[\sum_{j=1}^k (d_{1j} - n_{1j} d_j/n_j)\right]^2}{\sum_{j=1}^k [n_{2j} n_{1j} d_j (n_j - d_j)/[n_j^2(n_j - 1)]]}$$

For some obscure reasons the test is called the **log-rank test.** It can be naturally extended to the several groups case.

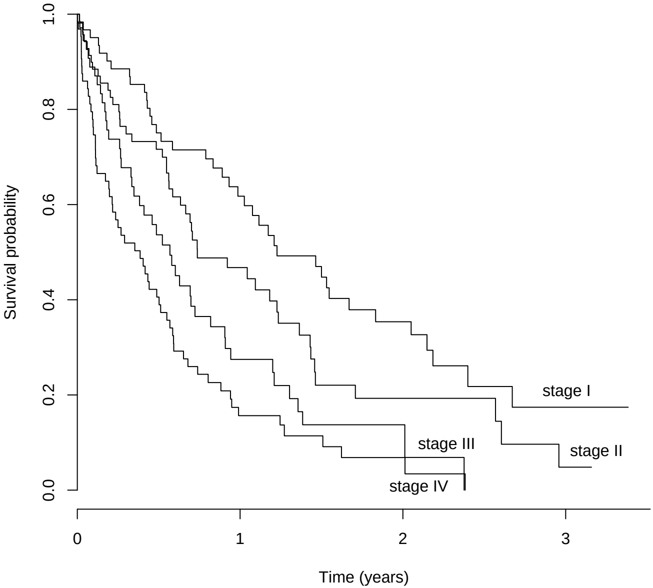
Example: log-rank test



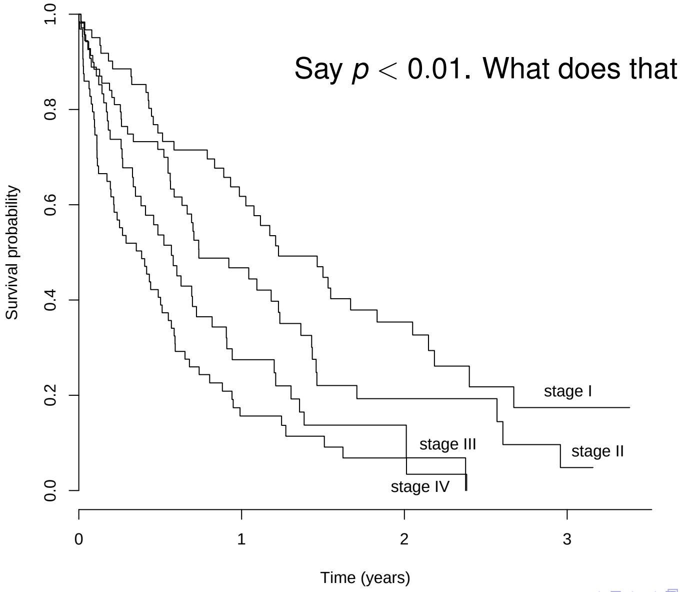
Example cont.: log-rank test

We used log-rank test to test the difference in duration of the mission for different types of the conflict precipitating a UN peacekeeping force. There were 30, 14 and 10 missions as a result of a civil war, interstate conflict and internationalized civil war respectively. P-value obtained from the log-rank test was 0.0095, so we can reject the null hypothesis that the duration of the missions for different types is equal.

Another example: log rank test



Another example: log rank test



Say $p < 0.01$. What does that mean?

Regression models

In the chapter on parametric estimation of survival curves we assumed that our measurements all come from the same distribution.

In real life this is seldom true.

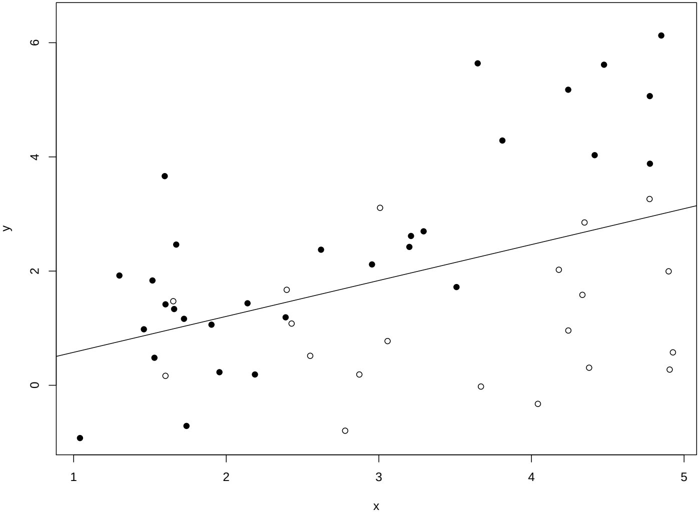
Distributions will often change with values of different variables, which is why we need to look at conditional distributions.

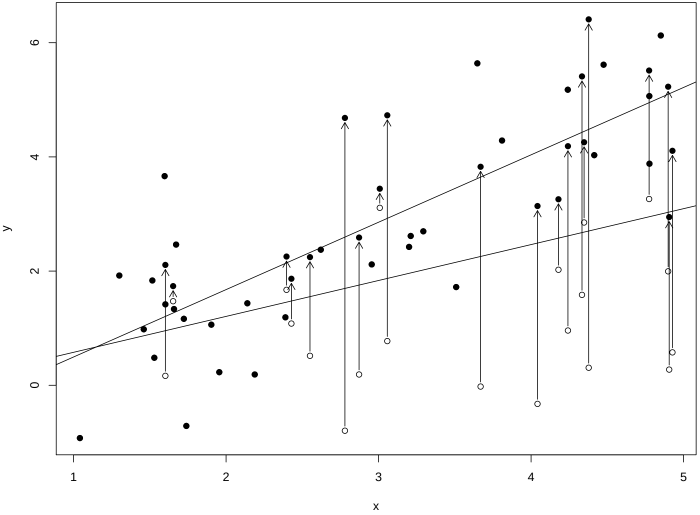
Regression models

When the outcome is a numerical variable, it is common to use the linear regression model

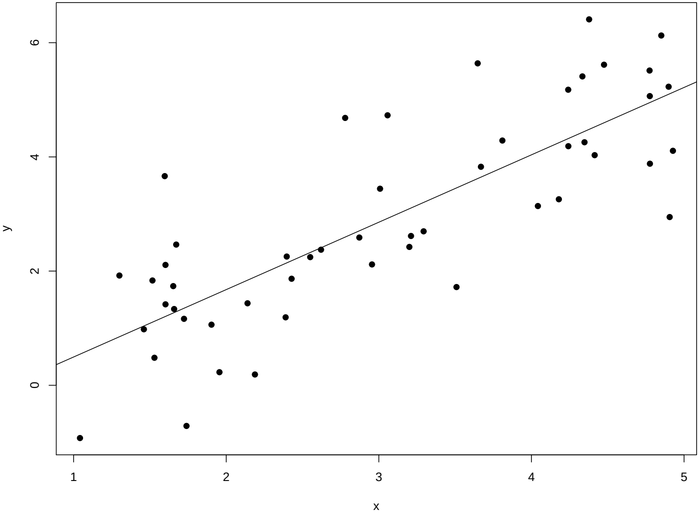
$$Y \sim \mathcal{N}(\alpha + \sum \beta_i X_i, \sigma^2)$$

This relates the values of Y to the values of X_i. We cannot do this in survival because of censoring.


y
6
4
2
0
x
1
2
3
4
5



y
0
2
4
6
x
1
2
3
4
5



y
6
4
2
0
1
2
3
4
5
x

Solution is the hazard function

Just to remind you

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

Solution is the hazard function

Just to remind you

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

$$S(t) = e^{-\int_0^t \lambda(u) du}$$

Exponential regression model

If we are confident that a certain distribution applies for our data, we assume a specific form for the hazard function. This is a parametric approach. In this course we will only look at the simplest parametric regression model, the exponential model.

Let $X_1, \ldots, X_p$ be variables, sometimes called *prognostic factors*, measured on each individual at $t = 0$.

Exponential regression model

In an exponential model the hazard is constant, but we can generalize our model by making this constant dependent on prognostic factors, for example like this

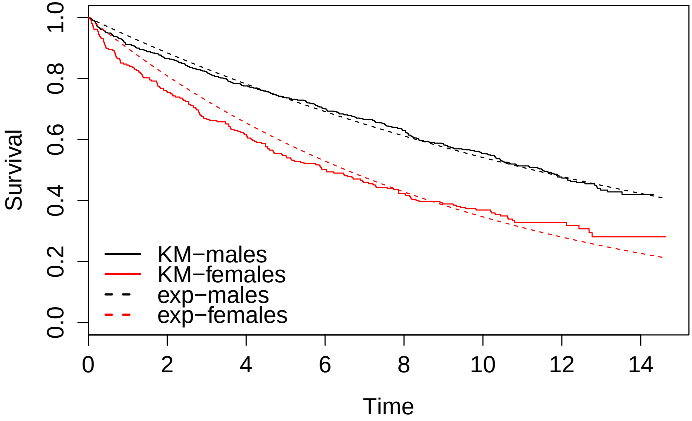
$$\lambda(t,x) = e^{\beta' x},$$

where $\beta' = \beta_0, \ldots, \beta_p$ is a vector of regression coefficients, and we add 1 as the first component of the vector X.

The density $f(t,x)$ is then given by

$$f(t,x) = e^{\beta' x} e^{-te^{\beta' x}}$$

Exponential regression model - MI example



Other possibilities

We will not look at this in any detail, I would just like to mention that in industrial settings the Weibull model is used a lot.

In socio/political area the piecewise exponential model was popular in the past.

Cox model (proportional hazards model)

The basic form of the model looks like this

$$\lambda(t,x) = \lambda_0(t)e^{\beta' x}, \tag{10}$$

where $\lambda_0(t)$ is the so called *baseline hazard*, and β and x have the usual meaning (no β_0!).

For two different x values we have

$$\frac{\lambda(t,x_1)}{\lambda(t,x_2)} = e^{\beta'(x_1 - x_2)}, \tag{11}$$

which is why the model is also called *the proportional hazards model*.

Cox model

From (11) we see that the hazard ratio for two subjects whose values differ by 1 in the *i*th covariate, with other values of covariates being equal, is simply exp(β_i).

From (11) it also follows that

$$\ln \lambda(t,x_1) - \ln \lambda(t,x_2) = \beta'(x_1 - x_2).$$

(This has some diagnostic value).

Cox model and monotone transformations of time

Assume now that T follows the proportional hazards model

$$\lambda(t,x) = \lambda_0(t)e^{\beta' x}.$$

Then

$$S(t,x) = e^{-e^{\beta' x}\int_0^t \lambda_0(u)du} = S_0(t)^{e^{\beta' x}}.$$

Cox model and monotone transformations of time

Let $T^* = g(T)$, where g is a monotonically increasing function. Let's calculate $S_{T^*}(t)$.

$$S_{T^*}(t) = P(T^* > t) = P(g(T) > t) = P(T > g^{-1}(t)) = S_T(g^{-1}(t)).$$

Then

$$S_{T^*}(t,x) = S_{T0}(g^{-1}(t))^{e^{\beta' x}}$$

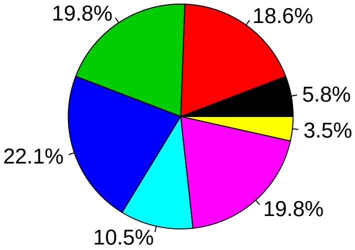
or

$$\lambda_{T^*}(t,x) = \lambda_{T0}(g^{-1}(t))e^{\beta' x}.$$

This means that T^* also follows the proportional hazards model. In other words, monotone transformations of time change the baseline hazard, but not $exp(\beta' x)$. If we're only interested in coefficients β, then the true values of the times of events are not important, only their ranks matter.

Estimation of coefficients in the Cox model - intuitive derivation

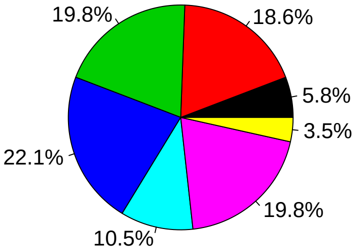
If you were randomly shooting at the target below, proportions of hits for different areas would be as shown.



Estimation of coefficients in the Cox model - intuitive derivation

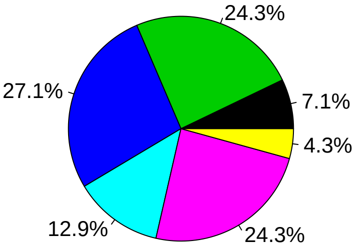
These are probabilities of hits, calculated simply as

$$P(\text{given colour}) = \frac{\text{Area(given colour)}}{\sum_i \text{Area}(\text{colour}_i)}$$



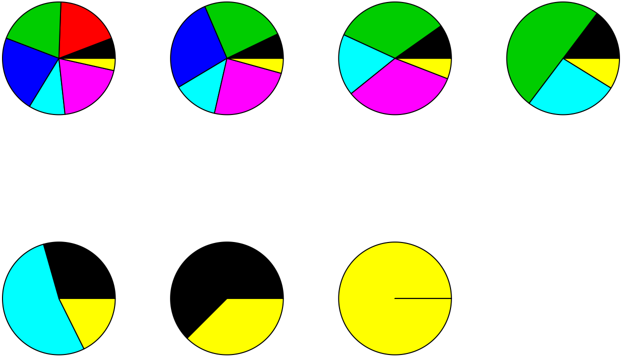
Estimation of coefficients in the Cox model - intuitive derivation

Imagine that the colour hit in the first try is removed from the target and we are shooting at the target with colours that are left



Estimation of coefficients in the Cox model - intuitive derivation

And so on . . .



Estimation of coefficients in the Cox model - intuitive derivation

If we had a model for the area, with some unknown parameters, we could estimate those parameters by maximizing the product of these probabilities!

Estimation of coefficients in the Cox model - intuitive derivation

We can also imagine that areas represent hazards of people we are following.

Estimation of coefficients in the Cox model - intuitive derivation

We can then calculate our probabilities at these different times

$$P_j(\text{given colour}, t_j) = \frac{\text{Area}(\text{given colour}, t_j)}{\sum_i \text{Area}(\text{colour}_i, t_j)}$$

Estimation of coefficients in the Cox model - intuitive derivation

If we had a model for the size of the areas, say

$$P_j(\text{given colour}, t_j) = f(t_j, \beta)$$

we could use the method of maximum likelihood to estimate β.

The product

$$\prod_j P_j(\text{given colour}, t_j) = \prod_j f(t_j, \beta)$$

Estimation of coefficients in the Cox model - formal derivation

Say that we measured (T_i,δ_i,X_i) on n subjects, where

T_i are measured times, censored or not,

δ_i are indicators of censoring (1 = event, 0 = censoring),

X_i is a vector of prognostic variables.

Let $t_1, \ldots, t_k$ be ordered, distinct times of events, so that at any t_i only one event occurs. Denote by $R(t) = \{i : t_i \geq t\}$ a set of subjects still at risk at t.

Estimation of coefficients in the Cox model

Let's write down the probability that at time t_i the subject i of those in $R(t_i)$ experiences the event.

$$\begin{aligned} P((i \text{ fails at } t_i | i \text{ in } R(t_i)) \quad & | \quad \text{one failure from } R(t_i)) \\ &= \frac{P(i \text{ fails} | i \text{ in } R(t_i))}{\sum_{j \in R(t_i)} P(j \text{ fails} | j \text{ in } R(t_i))} \\ &= \frac{\lambda(t_i, x_i)}{\sum_{j \in R(t_i)} \lambda(t_i, x_j)} \\ &= \frac{e^{\beta' x_i}}{\sum_{j \in R(t_i)} e^{\beta' x_j}} \end{aligned} \tag{12}$$

The last expression follows because the baseline hazards cancel out.

Estimation of coefficients in the Cox model

Cox suggested that the product of such probabilities is used as a criterion for estimating the parameters.

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{\beta' x_i}}{\sum_{j \in R(t_i)} e^{\beta' x_j}} = \prod_{i=1}^{n} \left[\frac{e^{\beta' x_i}}{\sum_{j \in R(t_i)} e^{\beta' x_j}} \right]^{\delta_i} . \qquad (13)$$

The criterion (13) is called *the partial likelihood*, but quite some water has passed under the bridges before it was proven that the partial likelihood can be treated as the full likelihood. The Cox model was well used in practice before we had a rigorous proof.

Estimation of coefficients in the Cox model

It is easy to see that (13) is indeed only a part of the full likelihood. The full likelihood, as we know from (5), is

$$L(\beta) = \prod_{i=1}^{n} \lambda(t_i,x_i)^{\delta_i} S(t_i,x_i).$$

If each factor in the above product is multiplied and divided by $\left[\sum_{j\in R(t_j)} \lambda(t_j,x_j)\right]^{\delta_i}$, we get

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\lambda(t_i,x_i)}{\sum_{j\in R(t_j)} \lambda(t_j,x_j)}\right]^{\delta_i} \left[\sum_{j\in R(t_j)} \lambda(t_j,x_j)\right]^{\delta_i} S(t_i,x_i)$$

from where we see that the expression (13) is rather far from the full likelihood. Cox has heuristically shown that (13) contains almost all the information about the coefficients β and, as already mentioned, he was later proven right.

Estimation of coefficients in the Cox model

So, if we can consider (13) as the usual likelihood function, we can use the well beaten path to the estimation of parameters. First we take logarithms

$$\ell(\beta) = \ln \prod_{i=1}^{n} \left[\frac{e^{\beta' x_i}}{\sum_{j \in R(t_i)} e^{\beta' x_j}} \right]^{\delta_i} = \sum_{i=1}^{n} \delta_i \left[\beta' x_i - \ln \left[\sum_{j \in R(t_i)} e^{\beta' x_j} \right] \right]$$

and then derivatives. If β has p components, we get for each $k = 1, \ldots, p$

$$U(\beta_k) = \frac{\partial}{\partial \beta_k} \ell(\beta) = \sum_{i=1}^{n} \delta_i \left[x_{ik} - \frac{\sum_{j \in R(t_i)} x_{jk} e^{\beta' x_j}}{\sum_{j \in R(t_i)} e^{\beta' x_j}} \right].$$

Estimation of coefficients in the Cox model

These derivatives are then equated to 0, and the corresponding equations solved (numerically).

(Mention ties here)

We now have our estimates of the coefficients. The next two results follow from the standard likelihood theory

$$\frac{\hat{\beta}_k - \beta_k}{se(\hat{\beta}_k)} \sim N(0,1)$$

$$\mathrm{var}(\hat{\beta}_k) \approx \left(-\frac{\partial^2}{\partial \beta_k^2} \ell(\beta) \right)^{-1}$$

We can use the above to calculate the confidence intervals for β.

Hypotheses testing - Likelihood ratio

Assume we measured $p + q$ variables

$$X_1, \ldots, X_p, X_{p+1}, \ldots, Xp + q$$

and that we want to compare models

$$\lambda(t, X) = \lambda_0 e^{\beta_1 X_1 + \cdots + \beta_p X_p}$$

and

$$\lambda(t, X) = \lambda_0 e^{\beta_1 X_1 + \cdots + \beta_{p+q} X_{p+q}}.$$

Hypotheses testing - Likelihood ratio

The hypothesis $H_0 : \beta_{p+1} = \cdots = \beta_{p+q} = 0$ can be tested using the **likelihood ratio test** in which we use the fact that

$$-2\left[\ln(\hat{L}(1)) - \ln(\hat{L}(2))\right] \sim \chi^2(q),$$

meaning that the left hand side expression above follows the χ^2 distribution with q degrees of freedom (we took this from the general theory of testing). Here $\hat{L}$ is the maximized likelihood, and the numbers in parentheses refer to the respective models without and with the last q variables.

Hypotheses testing - The Wald test

The null hypothesis for each variable separately

$$H_0 : \beta_j = 0$$

is usually tested using the **Wald test** which involves calculating

$$Z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad \text{or} \quad \chi^2 = \left(\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\right)^2 .$$

Hypotheses testing - The Wald test

Z has the standardized normal distribution, and χ^2 the χ^2 distribution with one degree of freedom. If we want to test the hypothesis about more than one coefficient being zero (we're interested in a group of variables or in a categorical variable which is represented with more dummy variables), we use the χ^2 test with corresponding *k* degrees of freedom (*k* being the number of the coefficients)

$$\chi^2_k = \hat{\beta}' \operatorname{var}(\hat{\beta})^{-1} \hat{\beta},$$

where $\hat{\beta}$ is now the corresponding vector of the estimated coefficients, and $\operatorname{var}(\hat{\beta})$ their covariance matrix.

Hypotheses testing - The score test

We'll skip this one, let me just mention that the derivative of the logarithm of the likelihood is called the **score** or score function.

Estimating the survival function for given values of covariates

Parametric models give us a complete specification of the hazard function, from which we can directly calculate the survival function using (2). In the Cox model the hazard has to be estimated separately. The method usually used is named after Breslow and is a generalizations of the Nelson-Aalen estimator of the survival curve. Another method, a generalization of the Kaplan-Meier estimator, was proposed by Kalbfleisch in Prentice R has both options implemented.

Once we have estimated the baseline hazard, we get the survival function estimate, given the values of covariates, in the following way

$$S(t,x) = e^{-e^{\beta x} \int_0^t \lambda_0(u)du} = (e^{-\int_0^t \lambda_0(u)du})^{e^{\beta x}} = S_0(t)^{e^{\beta x}}$$

Modeling techniques

Modeling techniques in the Cox model are really no different than such techniques are for any other regression model. We'll not spend much time on them, but we will mention some basics which are commonly used with categorical variables and in relaxing the linearity assumption for the effect of continuous covariates.

Categorical variables in the Cox model

Example: using the Cox model to compare survival curves

We take stage IV to be the reference category.

Stage	Stage I	Stage II	Stage III
I	1	0	0
II	0	1	0
III	0	0	1
IV	0	0	0

	coef	exp(coef)	se(coef)	z	p
Stage III	-0.316	0.729	0.202	-1.57	0.120
Stage II	-0.779	0.459	0.199	-3.92	< 0.001
Stage I	-1.203	0.300	0.213	-5.64	< 0.001

Continuous variables in the Cox model

Let the variable X be continuous. The Cox model assumes linear association between the logarithm of the hazard and X. This need not necessarily be true. What do we do? We might try to add the quadratic term. Then

$$\lambda(t,X) = \lambda_0(t)\exp(\beta_0 X + \beta_1 X^2).$$

Other variables are not important here, so let's forget about them.

With the form of the model above we can test the null hypothesis

H_0 : the model is linear in X

versus

the alternative hypothesis

H_a : the model is quadratic in X

by testing

$H_0 : \beta_1 = 0.$

Continuous variables in the Cox model

We can of course further complicate things by adding terms of a higher degree, but it usually turns out that polynomials, because of their hills and valleys, are not the best choice of a functional form. For example, if the true form is a logarithmic function, then polynomials will be far off.

Instead we may want to try a transformation of X ($\ln(X)$, say). But guessing the correct transformation can be a difficult task. It is much better to use *splines*.

Continuous variables in the Cox model - splines

Splines are polynomials, defined on the subintervals of the domain of X and connected at the borders of those intervals. The simplest splines are linear splines, piecewise linear functions. If the x axis was partitioned with points a, b and c (call the modes), a linear spline is defined as

$$f(X) = \beta_0 + \beta_1 X + \beta_2 (X - a)_+ + \beta_3 (X - b)_+ + \beta_4 (X - c)_+,$$

where

$$(u)_+ = \begin{cases} u, & u > 0 \\ 0, & u \leq 0. \end{cases}$$

In the Cox model β_0 is of course not needed.

Continuous variables in the Cox model - splines

Linear splines are simple, but they are not smooth at the joints and they will also not fit well if the underlying function has strong curvature. It turns out that one can do well by using polynomials of the third degree, which are glued in the nodes. For example, again for three nodes, we have

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X-a)_+^3 + \beta_5 (X-b)_+^3 + \beta_6 (X-c)_+^3 .$$

Of course we have to bear in mind that using such a procedure with three nodes we have six variables instead of one (for example $X_4 = (X-a)_+^3$), which has its consequences in sample size requirements.

Restricted cubic splines

Choice of nodes

The stratified Cox model

What do we do if the assumption of proportional hazards doesn't hold for a certain variable? Example: in a clinical trial we have two groups of patients, treated with two different treatments. Since we can not recruit enough patients in one hospital, we run the trial in several hospitals at the same time (a multi center trial). A weakness of such an approach is that treatment effects may be different in different centers. If we want to incorporate this in our model, we have to introduce the variable *center* into the model. But, its effect may not necessarily be proportional, and this causes problems.

The stratified Cox model

We can solve the problem by allowing different baseline hazards in different centers. So, if we have *M* centers, the model is

$$\lambda_m(t,x) = \lambda_{0m}(t)e^{\beta x}, \qquad m = 1,\ldots,M.$$

The partial likelihood is changed in such a way that each individual is only compared to the patients from the same center. In general we talk about strata and we call the above model *the stratified Cox model*.

$$PL = \prod_{i=1}^{M}\prod_{j=1}^{n_i}\left(\frac{e^{\beta x_{ij}}}{\sum_{k\in R_i(t_{ij})} e^{\beta x_{ik}}}\right)^{\delta_{ij}}$$

Time dependent variables in the Cox model

Variables that can influence the time until the event, can change in time.

- a person can stop smoking (and start again),
- a patient may have a transplant,
- marital status can change ...

Time dependent variables in the Cox model

If such changes have an effect on survival time, then a model which takes into account only the initial values, will not adequately reflect the influence of these variables. When we want to stress that we allow time dependent variables in the model, we write

$$\lambda(t,x(t)) = \lambda_0(t)e^{\beta x(t)}.$$

Of course, only some components of $x(t)$ may be time dependent.

Time dependent variables in the Cox model

The conditional probabilities at any time point are the same as before, except that the values of $x(t)$ may change for some individuals. This has to be accounted for in the calculation of the likelihood.

And for this we have to know $x(t)$ at all event times! (important to know when planning!)

Time dependent variables in the Cox model

Example: Assume now that we have three patients who were receiving treatment like described in the table below.

Patient	Time	Treatment
	(months)	
1	6	A always
2	18	A one year, then B
3	30	B two years, then A

Let $x(t) = 0/1$, if treatment A/B. Then the partial likelihood is

$$
\begin{aligned}
PL &= \frac{e^{\beta x_1(t_1)}}{e^{\beta x_1(t_1)} + e^{\beta x_2(t_1)} + e^{\beta x_3(t_1)}} \times \frac{e^{\beta x_2(t_2)}}{e^{\beta x_2(t_2)} + e^{\beta x_3(t_2)}} \times \frac{e^{\beta x_3(t_3)}}{e^{\beta x_3(t_3)}} \\
&= \frac{e^{\beta \times 0}}{e^{\beta \times 0} + e^{\beta \times 0} + e^{\beta \times 1}} \times \frac{e^{\beta \times 1}}{e^{\beta \times 1} + e^{\beta \times 1}} \times \frac{e^{\beta \times 0}}{e^{\beta \times 0}}
\end{aligned}
$$

In short, this is just the usual partial likelihood where we are careful to enter values of variables as they are at each time point.

Checking the proportional hazards assumption

The proportional hazards assumption is of course very important and has to be checked. There are different possibilities, and we will look at three here.

To simplify the notation, let X now be just a single variable. If the effect of X does not change in time, then the coefficient β_2 in the model

$$\lambda(t,x) = \lambda_0(t)e^{\beta_1 x + \beta_2 xt}$$

should be 0. In other words, the test for this coefficient should not be significant. If this is not true, the proportional hazards assumption for this variable does not hold.

Checking the proportional hazards assumption

Another possibility is graphical. Since we have

$$S(t,x) = S_0(t)^{e^{\beta x}} \tag{14}$$

we get for two different x_1 and x_2

$$\ln(S(t,x_1)) = e^{\beta x_1} \ln(S_0(t)) \quad \text{and} \quad \ln(S(t,x_2)) = e^{\beta x_2} \ln(S_0(t))$$

and from here

$$-\ln(S(t,x_1)) = -\frac{e^{\beta x_1}}{e^{\beta x_2}} \ln(S(t,x_2)).$$

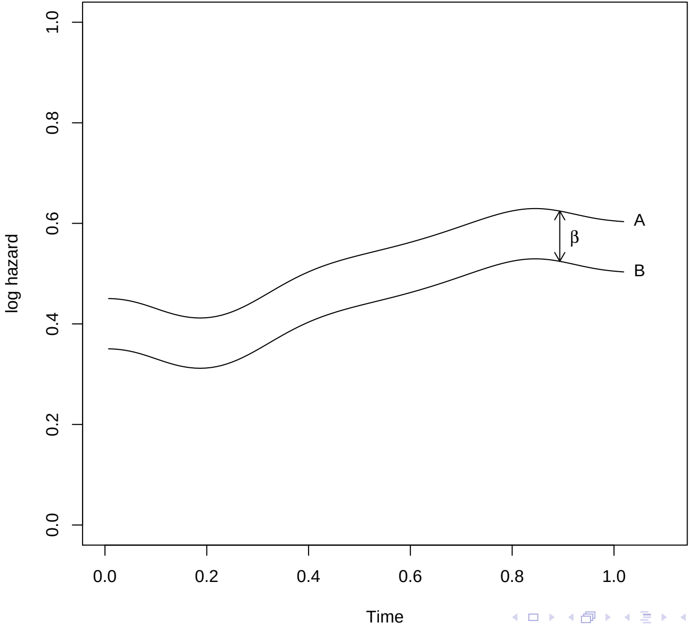
We put minuses so that we have positive quantities on both sides of the equation. In the logarithm of survival we recognize the cumulative hazard (or do we?).

Checking the proportional hazards assumption

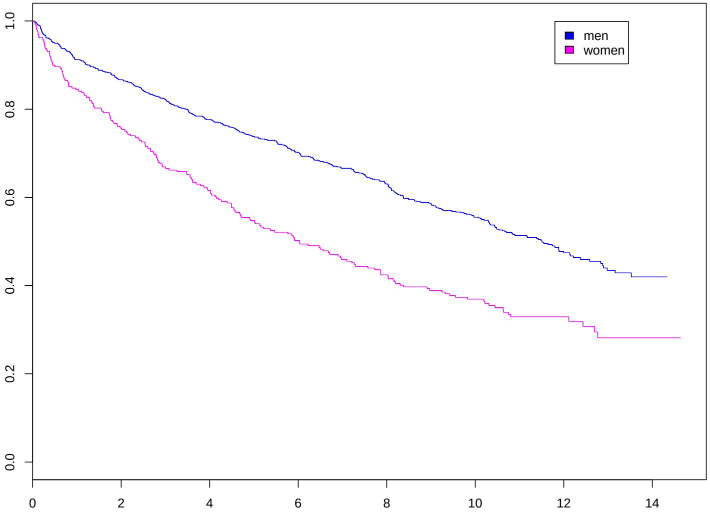
The above tells us that cumulative hazards are in linear relationship, where the coefficient is the hazard ratio between subjects with covariates x_1 and x_2.

We can of course take the logarithm of equation 14 twice, and then we have

$$\ln(-\ln(S(t,x))) = \beta x + \ln(-\ln(S_0(t)))$$


log hazard
0.0
0.2
0.4
0.6
0.8
1.0
A
B
β
0.0
0.2
0.4
0.6
0.8
1.0
Time

Example: checking the fit for MI data (try to guess!)



Checking the proportional hazards assumption - Schoenfeld residuals

Remember now the derivatives of the log(partial)likelihood for the Cox model.

If β has p components (we have p covariates), we get for each $k = 1, \ldots, p$

$$U(\beta_k) = \frac{\partial}{\partial \beta_k} \ell(\beta) = \sum_{i=1}^{n} \delta_i \left[x_{ik} - \frac{\sum_{j \in R(t_i)} x_{jk} e^{\beta' x_j}}{\sum_{j \in R(t_i)} e^{\beta' x_j}} \right].$$

In the expression in the brackets we recognize the difference between the value of the kth covariate of the subject who had the event at time t_i and the expected value (average) of those values, given the risk set. So

$$x_{ik} - \sum_{j \in R(t_i)} x_{jk} p_j = x_{ik} - \bar{x}_{ik}$$

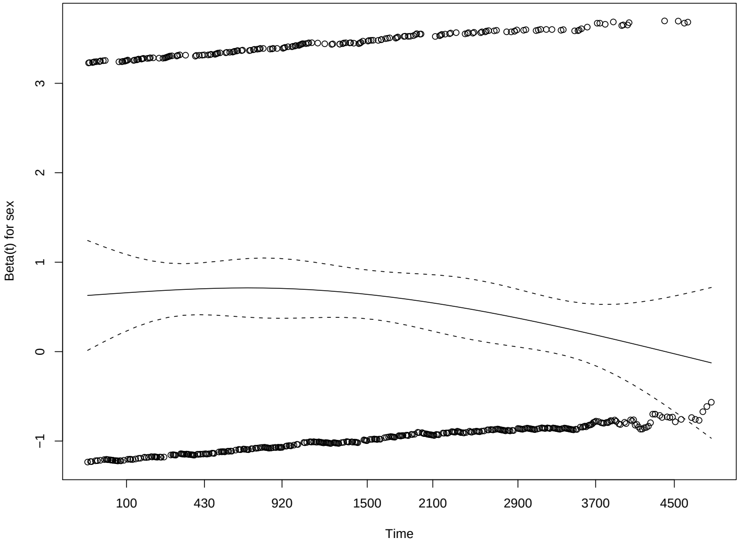
Checking the proportional hazards assumption - Schoenfeld residuals

These differences are called Schoenfeld residuals.

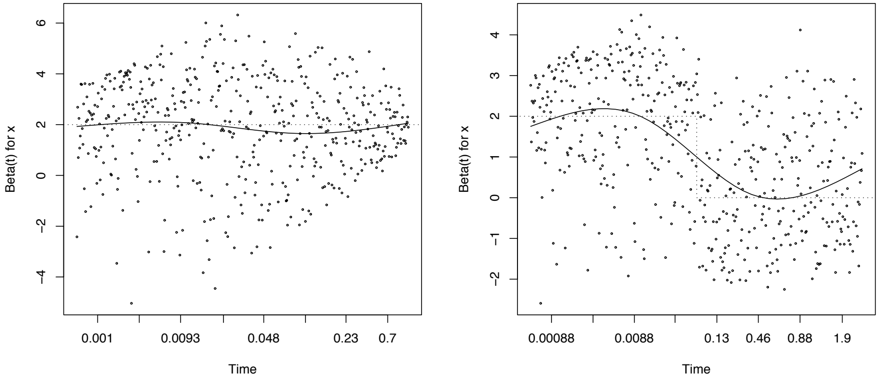
Remember, we have residuals for each individual who had the event, for each covariate).

If the model is correct, than these residuals should vary around 0 when plotted against time, or around estimated coefficient if such a coefficient is added to the residuals (which is what most packages give us). A non-random pattern is evidence of the proportional hazards assumption.

Example: checking the fit for the MI data



Example: good and bad fit



Schoenfeld residuals - an example

Say we fitted a Cox model with two covariates, gender and age (g and a).

And say the coefficients that we get are 2 for gender and 0.05 for age. Our model is then

$$\lambda(t,gender,age) = \lambda_0(t)e^{2*g+0.05*a}.$$

Schoenfeld residuals - an example

We know from (12) that the (conditional) probability of the subject *i* to have an event is

$$\frac{e^{2*g_i+0.05*a_i}}{\sum_j e^{2*g_j+0.05*a_j}}$$

where the sum in the denominator is over all the subjects still at risk at the time of the event.

Schoenfeld residuals - an example

At a certain time point *t* we have 5 subjects left with the following values of the covariates (male = 1, female = 0).

g	a	probability
1	58	0.27
0	55	0.03
1	45	0.14
0	67	0.06
1	70	0.50

The last column in the table below gives their probabilities of having the event, calculated using the formula on the previous slide.

Schoenfeld residuals - an example

What is the expected value of AGE for the person who has the event?

Based on the probabilities given by our model it is

$$58(.27) + 55(.03) + 45(.14) + 67(.06) + 70(.50) = 62.63$$

And if the one having the event was the 58 years old male, the corresponding Schoenfeld residual is $58 - 62.63 = -4.63$.

The expected value of GENDER is

$$1(.27) + 0(.03) + 1(.14) + 0(.06) + 1(.50) = 0.91$$

and the corresponding Schoenfeld residual is $1 - 0.91 = 0.09$

Time dependent effects

Let us now allow the coefficients in the Cox model to change with time. Then the model looks like this

$$\lambda(t,x) = \lambda_0(t)e^{\beta(t)x(t)}.$$

If we were now to estimate $\beta(t)$ at each time of an event, we would have too many parameters in the model, so it is necessary to limit the number of coefficients to a sensible number. Such estimation is a difficult problem, an area of active research at this time, here we will only look at a special case when $\beta(t)$ changes only once. The procedure can be generalized to more changes.

Time dependent effects

Assume that we know that $\beta(t)$ changes at time τ (a rather unrealistic assumption). Then we can do the following: we censor all the times that are greater than τ, and we then estimate $\beta(t)$. This will give us the coefficient up to the time τ. We then return to the original data and censor all observations that are less than or equal to τ. Estimating the coefficient on this data will give us $\beta(t)$ for the period after τ. We will of course achieve the same goal if the variable x, whose coefficient is changing in time, is introduced into the model like this

$$\lambda(t,x) = \lambda_0(t) e^{\beta_1 x_1(t) + \beta_2 x_2(t)},$$

where x_1 is equal to x until τ and after that it is 0, and x_2 is equal to 0 until τ and after that is equal to x. The procedure can be easily generalized to several changes.

Frailties

In a real world we can hardly expect all the subjects to be the same, meaning that their values of T would all come from the same distribution. We say that the population is heterogeneous.

Assume that each individual has some specific *frailty* z. Also assume that this frailty has a multiplicative effect on the hazard, so that

$$\lambda(t,z) = z\lambda(t).$$

The survival function is then

$$S(t,z) = S(t)^z,$$

and therefore different for each z. This of course is not surprising, since Z is simply a prognostic factor. But we have to remember that we do not really know Z and that we are looking at our subjects as a homogeneous group.

Frailties

Let us first calculate the average value of the hazard with respect to Z, at a given time t. This is

$$E(\lambda(t,Z)) = \lambda(t)E(Z).$$

Since subjects with larger values of z will experience the event earlier, then the average value of Z will decrease with time. Assuming that at $t = 0$ we have $E(Z) = 1$ (we can always do this), we then see that the ratio between $\lambda(t,z)$ and $\lambda(t)$ decreases when time increases.

The Gamma function

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx \qquad (a > 0)$$

The **gamma distribution** has the density

$$h(x) = \frac{\lambda^\eta x^{\eta-1} e^{-\lambda x}}{\Gamma(\eta)}$$

Let us try to calculate the *q*th moment of the gamma distribution.

$$E(X^q) = \int_0^\infty x^q h(x) dx = \frac{\lambda^\eta}{\Gamma(\eta)} \int_0^\infty x^q x^{\eta-1} e^{-\lambda x} dx = \frac{\Gamma(\eta+q)}{\lambda^q \Gamma(\eta)},$$

where we introduced a new variable $u = \lambda x$ in the last integral. From here we easily find the mean and the variance (since $E(X^2) = \Gamma(\eta+2)/(\lambda^2\Gamma(\eta)) = (\eta+1)\eta/\lambda^2$)

$$E(X) = \frac{\Gamma(\eta+1)}{\lambda\Gamma(\eta)} = \frac{\eta}{\lambda} \qquad \mathrm{var}(X) = \frac{\eta}{\lambda^2}$$

Frailties

Assume now that

$$\lambda(t,x,z) = z\lambda_0(t)e^{\beta x},$$

where the unknown values z come from the gamma distribution with the density $h(z)$. Of course we also have $\Lambda(t,x,z) = z\Lambda_0(t)e^{\beta x}$ and $S(t,x,z) = e^{-z\Lambda_0(t)e^{\beta x}}$. Since we do not know z, we shall in fact only see the marginal distribution, and therefore

$$S(t,x) = \int_0^\infty S(t,x,z)h(z)dz = \int_0^\infty e^{-z\Lambda_0(t)e^{\beta x}} \frac{\lambda^\eta z^{\eta-1}e^{-\lambda z}}{\Gamma(\eta)}dz.$$

Some rearranging gives

$$S(t,x) = \left(\frac{\lambda}{\lambda + e^{\beta x}\Lambda_0(t)}\right)^\eta.$$

Frailties

If $\eta = \lambda$, the above formula becomes

$$S(t,x) = \left(\frac{\eta}{\eta + e^{\beta x}\Lambda_0(t)}\right)^{\eta} = \frac{1}{(1+\xi e^{\beta x}\Lambda_0(t))^{\eta}},$$

where $\xi = 1/\eta = \mathrm{var}(Z)$.

From here

$$f(t,x) = -S'(t,x) = \frac{e^{\beta x}\lambda_0(t)}{(1+\xi e^{\beta x}\Lambda_0(t))^{\eta+1}}$$

and

$$\lambda(t,x) = \frac{f(t,x)}{S(t,x)} = \frac{e^{\beta x}\lambda_0(t)}{1+\xi e^{\beta x}\Lambda_0(t)} = \frac{e^{\beta x}\lambda_0(t)}{1+\mathrm{var}(Z)e^{\beta x}\Lambda_0(t)}.$$

Frailties

We then see that the ratio $\lambda(t,x)/\lambda_0(t)$ is smaller if $\mathrm{var}(Z)$ is bigger. And we see something else from the above formula: that the ratio must necessarily be decreasing with time, since $\Lambda_0(t)$ must be increasing. It is only constant when $\mathrm{var}(Z)=0$, meaning there is no frailty.

Let X be a binary prognostic variable with values 0 and 1. Let's look at the hazard ratio between these two groups. We first have

$$\lambda(t,1)=\frac{e^{\beta}\lambda_0(t)}{1+\mathrm{var}(Z)e^{\beta}\Lambda_0(t)} \quad \texttt{in} \quad \lambda(t,0)=\frac{\lambda_0(t)}{1+\mathrm{var}(Z)\Lambda_0(t)}$$

and the ratio is (where we denote e^{β} with r) is

$$\frac{\lambda(t,1)}{\lambda(t,0)}=\frac{r+r\,\mathrm{var}(Z)\Lambda_0(t)}{1+r\,\mathrm{var}(Z)\Lambda_0(t)}$$

This means that the ratio is approaching 1 as $t\to\infty$.

Frailties - uniqueness of Z

If we multiply z in $\lambda(t,x,z) = z\lambda_0(t)e^{\beta x}$ with some constant and divide $\lambda_0(t)$ with the same number, nothing changes. This means that the distribution of Z is not uniquely determined. It is common to work with $Z \sim \Gamma(\eta,\eta)$, which has the mean equal to 1.

Repeated events

Examples of repeated (or recurrent) events are: changes of marital status, changes of job status, arrests, reelections, heart attacks ...

There are different ways of dealing with repeated events:

1. assuming independence (not recommended, but done often)
2. fitting each transition separately
3. using the shared frailty model (used often)
4. using the stratified model where we stratify by event number (less efficient than frailty, but more general)
5. using the number of previous events as a covariate

Kidney data

Kidney patients have catheters inserted and time is measured until infection occurs, or catheter is removed for some other reason (censored).

variable	**codes and units**
time	in days
status	1 for infection, 0 for censoring
age	in years
disease	0 = GN (glomerulonephritis)
	1 = AN (acute nephritis)
	2 = PKD (polycystic kidney disease)
	3 = other

Kidney data

	id	time	status	age	sex	disease	frail
1	1	8	1	28	1	Other	2.30
2	1	16	1	28	1	Other	2.30
3	2	23	1	48	2	GN	1.90
4	2	13	0	48	2	GN	1.90
5	3	22	1	32	1	Other	1.20
6	3	28	1	32	1	Other	1.20
7	4	447	1	31	2	Other	0.50
8	4	318	1	32	2	Other	0.50
9	5	30	1	10	1	Other	1.50
10	5	12	1	10	1	Other	1.50

Kidney data - simple analysis

```
coxph(formula = Surv(time, status) ~ age + sex, data = kidney)

  n= 76, number of events= 58

         coef exp(coef)  se(coef)      z Pr(>|z|)
age  0.002032  1.002034  0.009246  0.220  0.82607
sex -0.829314  0.436349  0.298955 -2.774  0.00554 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

    exp(coef) exp(-coef) lower .95 upper .95
age    1.0020      0.998    0.9840     1.020
sex    0.4363      2.292    0.2429     0.784

Concordance= 0.662  (se = 0.046 )
Rsquare= 0.089   (max possible= 0.993 )
Likelihood ratio test= 7.12  on 2 df,   p=0.02849
Wald test            = 8.02  on 2 df,   p=0.01814
Score (logrank) test = 8.45  on 2 df,   p=0.01466
```

Kidney data - another simple analysis

```
coxph(formula = Surv(time, status) ~ age + sex + disease, data = kidney)

  n= 76, number of events= 58

                coef exp(coef)  se(coef)      z Pr(>|z|)
age         0.003181  1.003186  0.011146  0.285   0.7754
sex        -1.483137  0.226925  0.358230 -4.140 3.47e-05 ***
diseaseGN   0.087957  1.091941  0.406369  0.216   0.8286
diseaseAN   0.350794  1.420195  0.399717  0.878   0.3802
diseasePKD -1.431108  0.239044  0.631109 -2.268   0.0234 *


           exp(coef) exp(-coef) lower .95 upper .95
age           1.0032     0.9968   0.98151    1.0253
sex           0.2269     4.4067   0.11245    0.4579
diseaseGN     1.0919     0.9158   0.49238    2.4216
diseaseAN     1.4202     0.7041   0.64880    3.1088
diseasePKD    0.2390     4.1833   0.06939    0.8235

Concordance= 0.697  (se = 0.046 )
Rsquare= 0.207   (max possible= 0.993 )
Likelihood ratio test= 17.65  on 5 df,   p=0.003423
```

Kidney data - analysis with frailty

```
coxph(formula = Surv(time, status) ~ age + sex + disease + frailty(id),
         data = kidney)

  n= 76, number of events= 58

            coef      se(coef) se2     Chisq DF p
age          0.003181 0.01115  0.01115  0.08 1  7.8e-01
sex         -1.483138 0.35823  0.35823 17.14 1  3.5e-05
diseaseGN    0.087957 0.40637  0.40637  0.05 1  8.3e-01
diseaseAN    0.350794 0.39972  0.39972  0.77 1  3.8e-01
diseasePKD  -1.431107 0.63111  0.63111  5.14 1  2.3e-02
frailty(id)                             0.00 0  9.3e-01

           exp(coef) exp(-coef) lower .95 upper .95
age           1.0032     0.9968   0.98151    1.0253
sex           0.2269     4.4068   0.11245    0.4579
diseaseGN     1.0919     0.9158   0.49238    2.4216
diseaseAN     1.4202     0.7041   0.64880    3.1088
diseasePKD    0.2390     4.1833   0.06939    0.8235

     Variance of random effect= 5e-07   I-likelihood = -179.1
Degrees of freedom for terms= 1 1 3 0
Concordance= 0.699  (se = 0.046 )
Likelihood ratio test= 17.65  on 5 df,   p=0.003423
```

Kidney data - analysis with frailty but without disease

```
coxph(formula = Surv(time, status) ~ age + sex + frailty(id),
    data = kidney)

  n= 76, number of events= 58

            coef      se(coef) se2      Chisq DF    p
age          0.005253 0.01189  0.008795  0.20  1.00 0.66000
sex         -1.587489 0.46055  0.351996 11.88  1.00 0.00057
frailty(id)                             23.13 13.01 0.04000

    exp(coef) exp(-coef) lower .95 upper .95
age    1.0053     0.9948    0.9821    1.0290
sex    0.2044     4.8914    0.0829    0.5042

Iterations: 7 outer, 65 Newton-Raphson
     Variance of random effect= 0.4121647   I-likelihood = -181.6
Degrees of freedom for terms=  0.5  0.6 13.0
Concordance= 0.814  (se = 0.046 )
Likelihood ratio test= 46.76  on 14.14 df,   p=2.312e-05
```

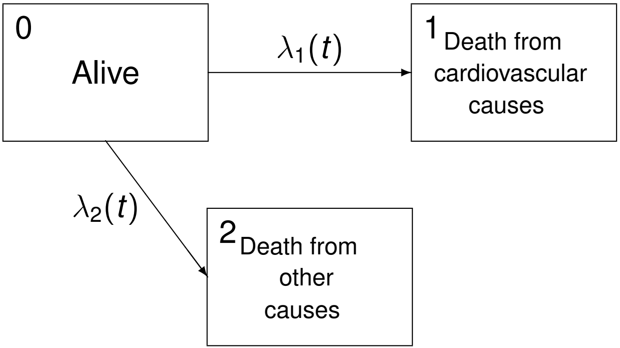
Why are above analyses different?

```
> fit<-(coxph(Surv(time,status)~age+sex+disease,data=kidney))
> cox.zph(fit)
                  rho   chisq     p
age           0.03945 0.09544 0.757
sex           0.18642 2.56162 0.109
diseaseGN    -0.02908 0.05037 0.822
diseaseAN     0.02794 0.04168 0.838
diseasePKD   -0.00472 0.00187 0.965
GLOBAL             NA 4.33109 0.503
```

```
> fit<-(coxph(Surv(time,status)~age+sex,data=kidney))
> cox.zph(fit)
          rho  chisq        p
age    0.0878  0.524 0.468996
sex    0.4363 11.470 0.000707
GLOBAL     NA 11.564 0.003083
```

Competing risks

Up to now we have assumed an individual can only experience one event. Suppose we are interested in several different kinds of events, a patient might die from different causes, the end of an unemployment spell might mean getting a job or exiting the labour market.



Competing risks

When studying such data, we are interested in the cause-specific hazard function

$$\lambda_j(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t, J = j | T \geq t)}{\Delta t},$$

where $J = j$ indicates a failure from cause j. Assuming that only one of the m failures types of interest can occur simultaneously, then

$$\lambda(t) = \sum_{j=1}^{m} \lambda_j(t).$$

Therefore, the overall survival function can be written as

$$S(t) = e^{-\int_0^t \sum\limits_{j=1}^{m} \lambda_j(u) du}.$$

Competing risks

The density function for time of failure *j* equals

$$f_j(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t, J = j)}{\Delta t},$$

and the corresponding cumulative distribution function (called the *cumulative incidence function*) is

$$F_j(t) = P(T \leq t, J = j)$$

Note that the function $S_j(t)$ would have no sensible interpretation and that the correspondence between the above functions is a bit different than in the case of only one possible event.

Competing risks

Following the same idea as in (1) we have

$$\lambda_j(t) = \frac{f_j(t)}{S(t-)},$$

and the cumulative distribution function F_j can be calculated as

$$F_j(t) = \int_0^t S(u-)\lambda_j(u)du, \tag{15}$$

but since the rightmost part of (1) no longer holds, the quantity $\exp(-\int_0^t \lambda_j(u)du)$ has no sensible interpretation.

Competing risks

To estimate the effect of covariates in the competing risks setting, one can again use the Cox model and specify the hazard functions as

$$\lambda_j(t,x) = \lambda_{0j}(t) e^{\beta_j x}. \tag{16}$$

Denoting by $t_{j1}, \ldots, t_{jn_j}$ the ordered, distinct times of failures of type j, the corresponding partial likelihood is

$$L(\beta) = \prod_{j=1}^{m} \prod_{i=1}^{n_j} \frac{e^{\beta_j' x_i}}{\sum_{k \in R(t_{ji})} e^{\beta_j' x_k}}$$

If we allow different β_j coefficients for each of the failure types, each part of the above product can be estimated separately. To estimate the coefficients in (16), one can therefore use the usual Cox model routine and censor all events but the event of interest.

Competing risks

Note that the estimated coefficients have to be interpreted in terms of the hazard function and can not be directly translated into probabilities (cumulative incidence functions) as (15) includes $S(t)$ that depends on hazards for the other failures as well. To model those directly, one has to turn to other methods, for example the Fine and Gray model that models the effect of the covariates through the function (this function does not have the hazard function interpretation!):

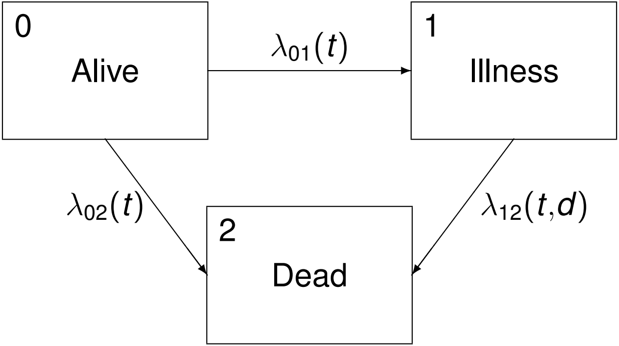
$$\lambda'_j(t,x) = \lambda'_{0j}(t) e^{\beta_j x}.$$

that corresponds to the equation

$$F_j(t) = 1 - e^{\int_0^t \lambda'_{0j}(u) e^{\beta_j x} du}$$

Multi-state models

In the previous section we dealt with several different types of events, with the common property that all of them brought an individual to a final state. One can also consider states that are transitional, i.e. states from which the individual can exit, an example is given in Figure 166.



Multi-state models

In such models, we follow the stochastic process X in time: in the example given in Figure 166, $X(t) = 0$ indicates that an individual is alive at time t, $X(t) = 1$ indicates he is ill and $X(t) = 2$ indicates he is dead at time t. The quantities of interest are for example the state occupation probability

$$P_j(t) = P(\text{individual is in state } j \text{ at time } t) = P(X(t) = j)$$

and the state transition probability

$$P_{hj}(s,t) = P(X(t) = j|P(X(s) = h),$$

where s and t denote two consequent time points and h and j two possible states of the model.

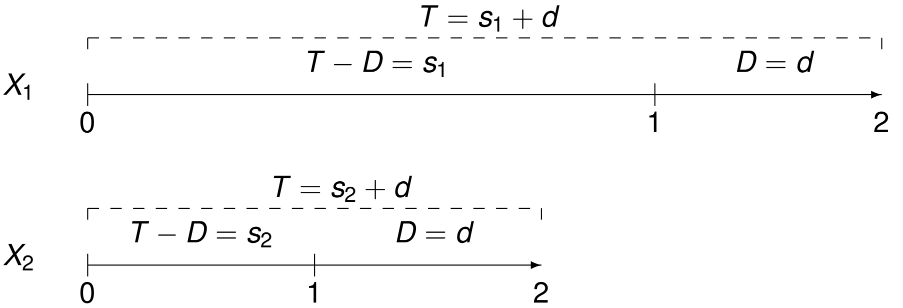
Multi-state models

The methods of estimating the above quantities are rather complicated and still represent a very active area of research, we will therefore comment only on the estimation of the hazard functions. These can be (as in the competing risks model) estimated using the standard methods (e.g. Cox model) by censoring all the events but the one we are interested in. The data set has to be split into the time-dependent form, with one line for each transition (start time, stop time, exiting state and entering state).

Multi-state models

For example, to estimate $\lambda_{01}(t)$ in Figure 166, one should focus on the data exiting state 0 and censor all the individuals that do not enter state 1.

As in the case of the competing risks, the coefficients estimated in this way must be interpreted in terms of the hazard function.



$T = s_1 + d$

X_1 $T - D = s_1$ $D = d$

0 1 2

$T = s_2 + d$

X_2 $T - D = s_2$ $D = d$

0 1 2

Multi-state models

When estimating the hazard function from a transient state, some care must be given to the time scale. In Figure 166 we can for example deal with time since origin (T) or the duration time in state 1 (D).
To estimate $\lambda_{12}(d)$ we focus on all individuals that were at some point in state 1, two examples of individual time-lines are given in Figure 169. Both individuals have the same duration time, but it is often sensible to include the time from origin to state 1 as a covariate in the model, for example:

$$\lambda_{12}(d,t,x) = \lambda_{0_{12}}(d)e^{\beta x+\gamma(t-d)}.$$

Multi-state models

On the other hand, if we are interested in the time since origin T, the two individuals in Figure 169 will never be directly compared in the partial likelihood function. As an additional covariate, one should in this case include time in state 1 (D) or, to make coding easier (no time-dependent covariates), time from origin to state 1 ($T - D$):

$$\lambda_{12}(t,d,x) = \lambda_{0_{12}}(t)e^{\beta x+\gamma(t-d)}.$$