# Introduction to Survival Analysis using R: Theory and Practice

Prof. Dr. Roel Braekers

Interuniversity Institute for Biostatistics and
statistical Bioinformatics (I-BioStat)
Hasselt University, Belgium.

15th January - 19th January 2018

These notes are based on the books:

- Collett (2015), *Modelling Survival data in Medical Research, Third Edition*

- Klein and Moeschberger (1997), *Survival analysis*, Springer-Verlag, New York.

**Some additional reading:**

- Liu (2012), *Survival Analysis, models and applications*
- Kalbfleisch and Prentice (2002), *The Statistical Analysis of Failure Time Data*, Wiley, New York.
- Andersen, Borgan, Gill, Keiding (1993), *Statistical Models Based on counting Processes*, Springer-Verlag, New York.
- Hougaard (2000), *Analysis of Multivariate Survival Data*, Springer-Verlag, New York.

# Example: Turbine disk

(Nelson (1982))

In an industrial trial, the lifetime is followed for 206 turbine disks. The data are observed in 100-hour intervals and after each inspection the number of failed disks is recorded.

Extra difficulty: the experiment was stopped after 2100 hours and some disks were still running at that time (censored).

# Example: Computer

A computer is build up out of many parts.

If we know the life expectancy of the hard disk, does this tell us sometime about the life expectancy of the computer ?

# Example: Leukemia

(Klein and Moeschberger (2003))

Bone marrow transplants are a standard treatment for acute leukemia. In this multi center trial, patients were followed from the moment they had a bone marrow transplant until the leukemia returns. The researchers investigated the time until relapse and looked for factors that influenced this time.

Some patients did not have a relapse during the study period or died without relapsing.

# Example: Vaginal cancer

(Kalbfleisch and Prentice (2002))

In a study on carcinogenesis, two groups of rats were exposed to a carcinogen DMBA. One group was keeped in a germ-free environment while the other group was not. In both group the number of days was recorded until these rats died of vaginal cancer. The researchers investigated whether the time until death was different in both groups.

However four rats did not die from cancer or were still alive at the end of the study.

# Example: Recidivism

(Schmidt and Witte (1988))

The North Caroline Department of Correction conducted two studies on recidivism. They looked at the time until a prisoner, who was released from a North Carolina prison, returned to a NC prison.

The goal of these studies was to find possible predicting factors like race, age at release, drug or alcohol problems, marital status, ...

We note that some released prisoners will never return.

# Example: Liability claims

(Klugman and Rioux (2006))

At an insurance company, liability claims are made after some damage has occurred. There are different policies which have deductibles of 100, 250 or 500 euro and maximum payments of 1000, 3000 or 5000 euro. This insurance company wants to know how much it should pay each year and is interested in the distribution of the claim sizes.

Because they only pay a maximum payment of 5000, it is impossible to express the true amount of the damage.

# Example: Wages

To gain insight into the economical status of a region, you may want to look at the income of every person in this region.

How would you deal with unemployment?

In reliability and survival analysis, we are interested in a non-negative random variable $T$ ($T \geq 0$). This variable $T$ can be discrete with values $\{0, 1, 2, \ldots\}$ or continuous on $(0, \infty)$.

This variable is known under a lot of different names:

- failure time
- lifetime
- time until an event
- loss
- ...

Sometimes we cannot fully observe this random variable $T$ but only observe some boundaries for this time. This is called censoring.

# Example: Leukemia

(Klein and Moeschberger (2003))

Let's go back to the starting examples...

Bone marrow transplants are a standard treatment for acute leukemia. In this multi center trial, patients were followed from the moment they had a bone marrow transplant until the leukemia returns. The researchers investigated the time until relapse and looked for factors that influenced this time.

Some patients did not have a relapse during the study period or died without relapsing.

# Example: Vaginal cancer

(Kalbfleisch and Prentice (2002))

In a study on carcinogenesis, two groups of rats were exposed to a carcinogen DMBA. One group was keeped in a germ-free environment while the other group was not. In both group the number of days was recorded until these rats died of vaginal cancer. The researchers investigated whether the time until death was different in both groups.

However four rats did not die from cancer or were still alive at the end of the study.

# Example: Recidivism

(Schmidt and Witte (1988))

The North Caroline Department of Correction conducted two studies on recidivism. They looked at the time until a prisoner, who was released from a North Carolina prison, returned to a NC prison.

The goal of these studies was to find possible predicting factors like race, age at release, drug or alcohol problems, marital status, ...

We note that some released prisoners will never return.

# Conclusion:

In each of the examples, we cannot fully observe the time until a certain event. Due to different practical reasons, we only observe in the examples a lower bound of the true time.

This is called right censoring.

In general, any situation in which you cannot fully observe a time until an event but only observe some boundaries for this time is called censoring.

In the examples, there is another non-negative random variable $C$ ($C \geq 0$) which we call the censoring variable and which obscures the observation of $T$.

For the moment, we only consider right censoring. There are three types of censoring.

- Type I or fixed censoring.
- Type II censoring
- Type III or random censoring.

# Type I or fixed censoring

Let $t_c \in \mathbb{R}$ be a fixed time point and take a sample lifetimes $T_1, \ldots, T_n$.

We only observe a lifetime $T_i$ if it is smaller than $t_c$, otherwise we get this fixed time point.

Hence, we get a sample $Y_1, \ldots, Y_n$ where

$$Y_i = \begin{cases} T_i & \text{, if } T_i \leq t_c \\ t_c & \text{, if } T_i > t_c \end{cases} \quad , i = 1, \ldots, n.$$

Example: Study stopped at a fixed time.

# Type II censoring

Let $r < n$ with $r \in \mathbb{N}$ and denote by $T_{(1)} < \ldots < T_{(n)}$ the ordered lifetimes.

We observe until the $r$-th system has failed.

Hence we get

$$Y_{(i)} = \begin{cases} T_{(i)} & \text{, if } T_{(i)} \leq T_{(r)} \\ T_{(r)} & \text{, if } T_{(i)} > T_{(r)} \end{cases} \quad \text{, } i = 1, \ldots, n.$$

Example: Industrial test trial.

# Type III or random censoring

Let $C_1, \ldots, C_n$ be a sample of censoring times.

We observe a sample of couples, $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ where, for $i = 1, \ldots, n$,

$$
\begin{aligned}
Y_i &= \min(T_i, C_i) = \begin{cases} T_i & , \text{if } T_i \leq C_i \\ C_i & , \text{if } T_i > C_i \end{cases} \\
\delta_i &= I(T_i \leq C_i) = \begin{cases} 1 & , \text{if } T_i \leq C_i \\ 0 & , \text{if } T_i > C_i \end{cases}
\end{aligned}
$$

In general we assume that, for $i = 1, \ldots, n$, $T_i$ and $C_i$ are independent.

Based on observations of $Y_1, \ldots, Y_n$, we want to estimate the distribution $F$.

In a similar way, we can look at other censoring schemes.

# Left censoring

We observe a sample $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ where, for $i = 1, \ldots, n$,

$$Y_i = \max(T_i, C_i) = \begin{cases} T_i & \text{, if } T_i \geq C_i \\ C_i & \text{, if } T_i < C_i \end{cases}$$

$$\delta_i = I(T_i \geq C_i) = \begin{cases} 1 & \text{, if } T_i \geq C_i \\ 0 & \text{, if } T_i < C_i \end{cases}$$

**Some examples:**

- Development of children behavior: At which age can a child perform a certain task. Some children can perform the task when they enter the study.
- Detection limits: A measuring device cannot give a correct value below a fixed limit.

# Interval-censoring

Instead of a sample of lifetimes $T_1, \ldots, T_n$, we get for each individual an interval in which the event occurred. Hence we get $(L_1, R_1], \ldots, (L_n, R_n]$.

**Example:**
In a clinical trial on breast cancer patients, the researchers were interested whether there was a difference in cosmetic effects for early breast cancer patients when they were treated with radiotherapy only or with radiotherapy and chemotherapy. At each visit, every 4 till 6 month, a clinician recorded a measure for breast retraction. Of interest was the time until moderate or severe breast retraction appeared.

# Doubly censoring

We observe a sample $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ where, for $i = 1, \ldots, n$,

$$\begin{aligned} Y_i &= \min(\max(T_i, L_i), R_i) \\ \delta_i &= = \begin{cases} 1 & \text{, if } Y_i = T_i \\ 0 & \text{, if } Y_i = R_i \\ -1 & \text{, if } Y_i = L_i \end{cases} \end{aligned}$$

**Example:**
In a study conducted at the Stanford-Palo Alto Peer counseling program, 191 California high school boys were asked: "When did you first use marijuana?" The answers were the exact ages (uncensored observations), "I never used it" (right censored observations), "I have used it but can not recall when the first time was" (left censored observations).

In a clinical study, we encounter three main mechanisms of censoring:

- Administrative censoring: due to the end of study.
- Withdrawal: a patient does not want to continue the treatment.
- Loss to follow-up: a patient do not show up anymore at the follow-up visits during trial.

Sometimes we note that the censoring variable is not independent of the lifetime, or the distribution of the censoring variable is linked to the distribution of the lifetime. We then call the censoring informative.

In the example, we see that

- Administrative censoring usually fulfills the independence condition.
- Withdrawals/losses to follow-up are potentially problematic: can be due to, e.g., disease progression or death.

In this course we will mainly deal with non-informative censoring and should try to avoid that we get informative censoring by checking the reasons why individuals leave a study.

To define a lifetime correctly, we need

- A time-origin.
- A time-scale.
- A definition of when the endpoint occurs.

However, sometimes there are several choices of time-origins and associated times-scales. Choosing the right one depends on the purpose of the study. We have in most cases:

- Study time: calendar time between the beginning and end of the study.
- Subject time: time spent in a study, measured from the subject's origin.

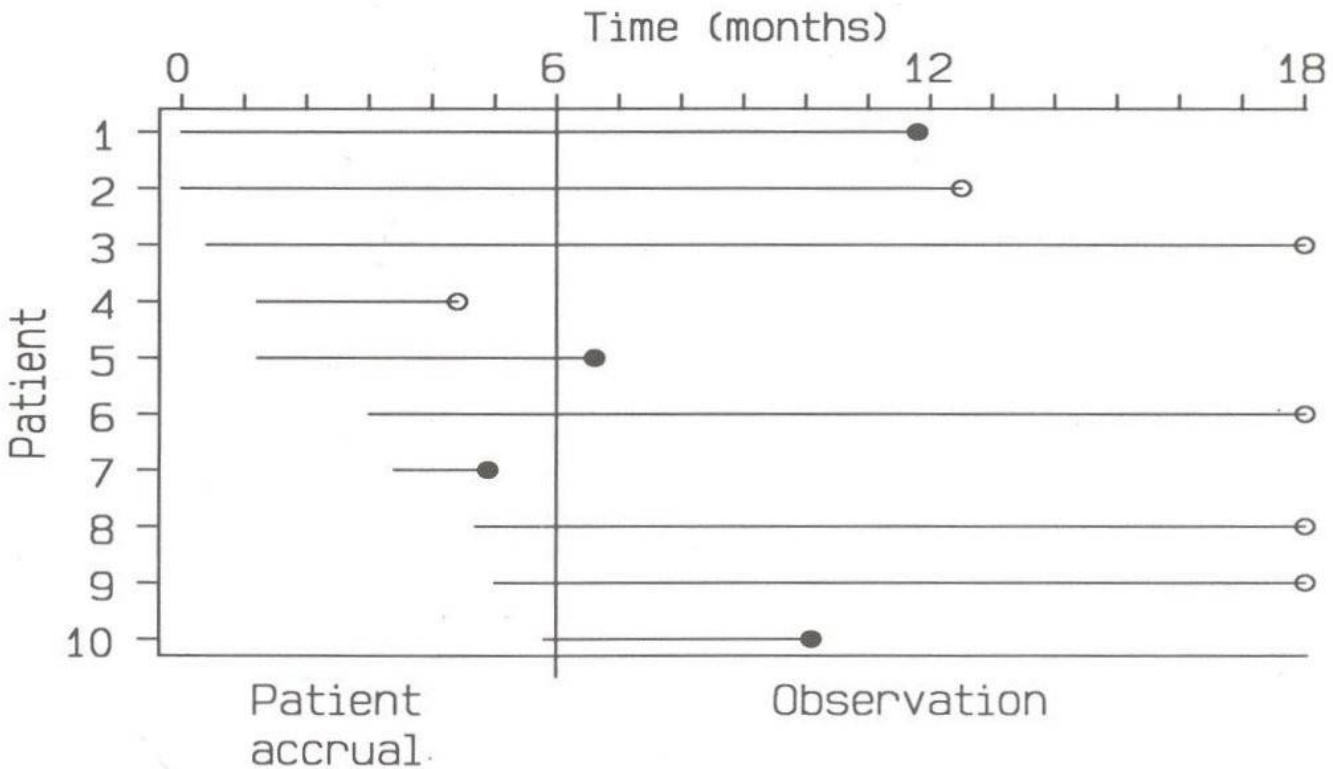



**Figure 13.1** Diagram showing patients entering a study at different times and the observation of known (●) and censored (○) survival times.

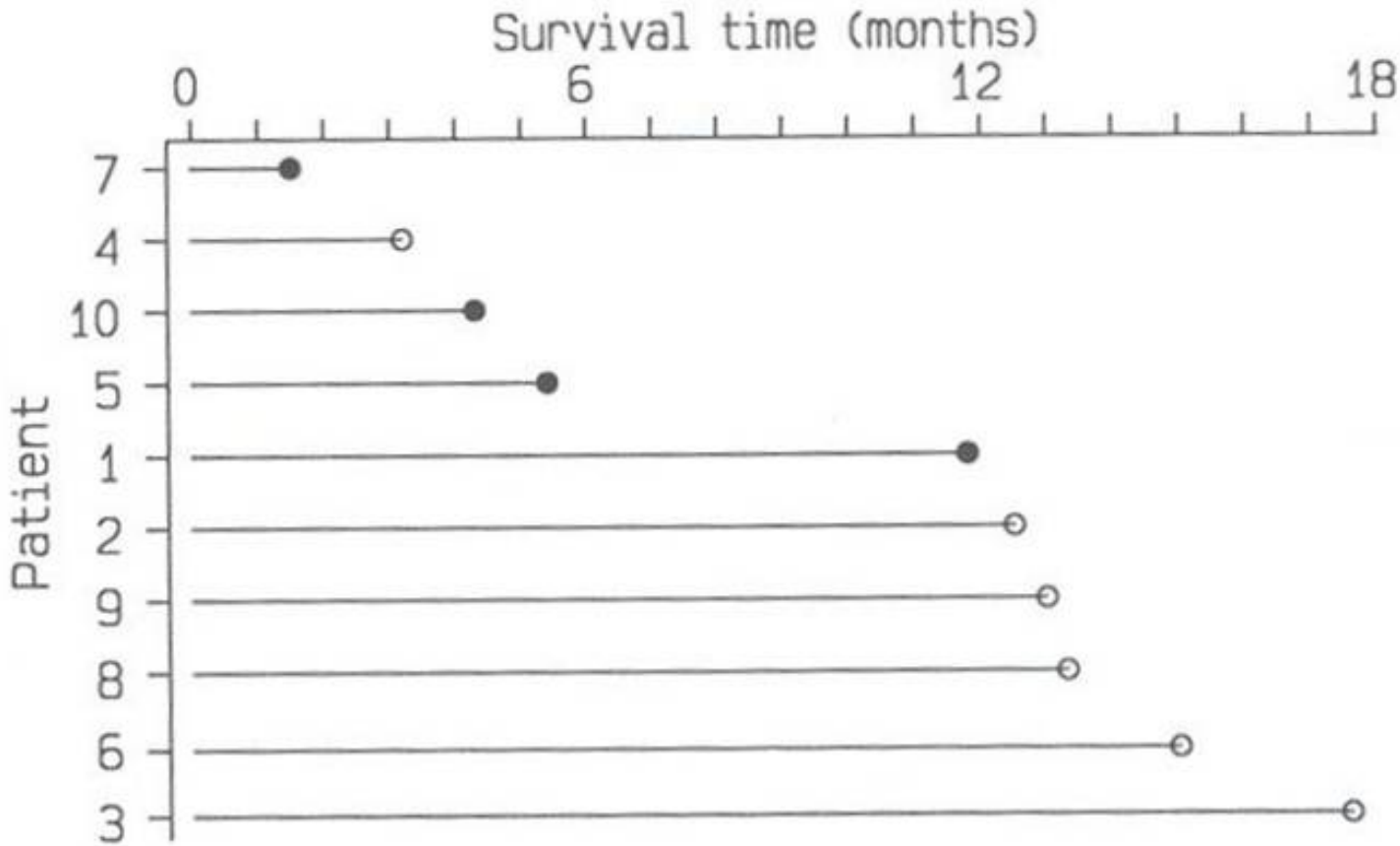



**Figure 13.2** Figure 13.1 reorganized to correspond to method of analysis.

# Survival function

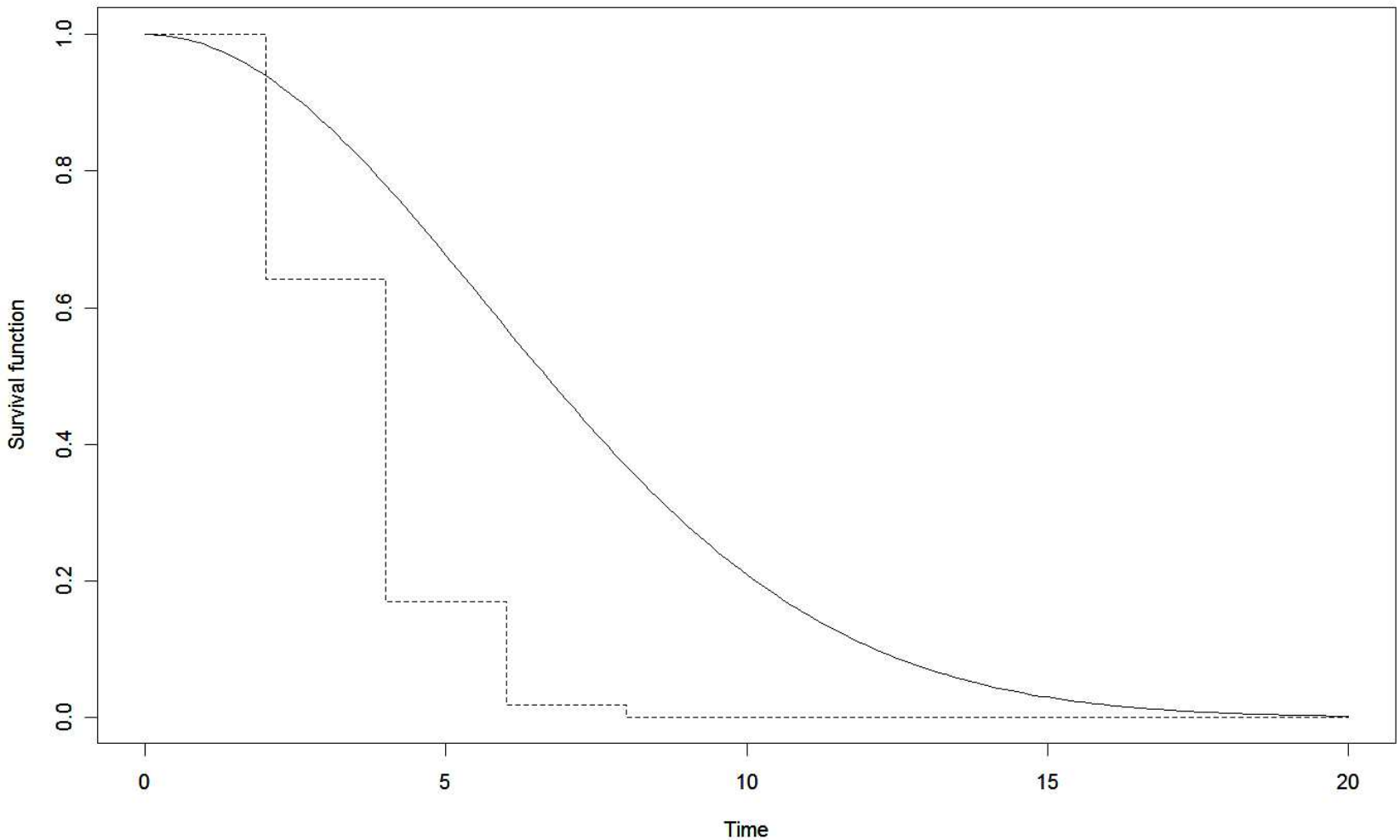
One of the important quantities in reliability and survival analysis is the survival function or reliability function

$$S(t) = P(T > t) = 1 - F(t).$$

It is the probability that an event has not occurred by time $t$.

Some properties:

- $S(0) = 1$.
- $S(+\infty) = \lim_{t \to +\infty} S(t) = 0$.
- $S(t)$ is a non-increasing function of $t$.


Survival function
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
Time

# $T$ is continuous

In this case, we have a density function

$$f(t) = -\frac{dS(t)}{dt} \quad \Rightarrow S(t) = \int_t^\infty f(x)dx.$$

A second important quantity in reliability and survival is the hazard function which is often also called (instantaneous) failure rate or intensity function,

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

It describes the probability for an event to take place in an small interval after time $t$, given that it has not occurred before $t$,

$$\lambda(t)h \approx P(t \leq T < t + h | T > t).$$

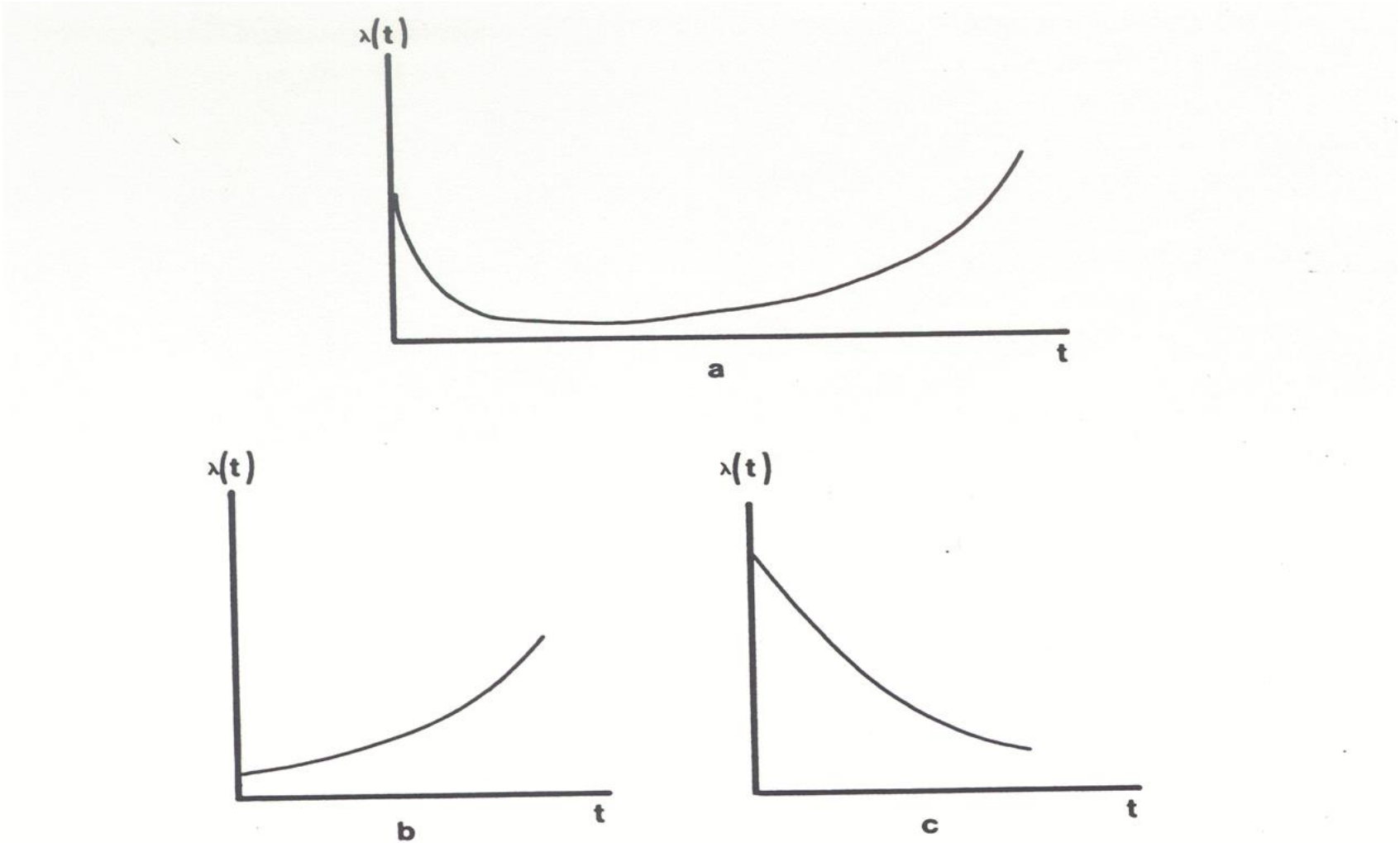



**Figure 1.1** Some types of hazard functions: (a) hazard for human mortality; (b) positive aging; (c) negative aging.

Relationship between $\lambda(t)$ and $S(t)$.

$$\begin{aligned}
\lambda(t) &= \lim_{h\to 0} \frac{P(t \le T < t+h | T > t)}{h} \\
&= \lim_{h\to 0} \frac{1}{h} \frac{P(t \le T < t+h)}{P(T > t)} \\
&= \lim_{h\to 0} \frac{1}{h} \frac{S(t) - S(t+h)}{S(t)} = \frac{f(t)}{S(t)}.
\end{aligned}$$

Conversely,

$$\begin{aligned}
\lambda(t) &= \frac{f(t)}{S(t)} = \frac{-1}{S(t)} \frac{d}{dt} S(t) = -\frac{d}{dt} \log S(t) \\
&\Rightarrow \log S(t) = -\int_0^t \lambda(x)dx \Rightarrow S(t) = \exp\left(-\int_0^t \lambda(x)dx\right)
\end{aligned}$$

We define the cumulative hazard function as

$$\Lambda(t) = \int_0^t \lambda(x)dx.$$

We note that $\Lambda(0) = 0$ and $\Lambda(+\infty) = \int_0^{+\infty} \lambda(x)dx = +\infty$.

In reliability and survival analysis, we are also interested in the residual lifetime $T_t$. This is the remaining lifetime of a system that has survived until time $t$. The distribution of the residual lifetime is given by

$$F_t(x) = P(T - t \leq x | T \geq t) = \frac{F(t+x) - F(t)}{S(t)}.$$

Looking at the average value, we get the mean residual life time,

$$r(t) = E[T - t | T \geq t] = \frac{\int_t^\infty (x-t) f(x) dx}{S(t)} = \frac{\int_t^\infty S(x) dx}{S(t)}.$$

Furthermore we can derive the following relationships

$$\begin{aligned} S(t) &= \frac{r(0)}{r(t)} \exp\left(-\int_0^t \frac{du}{r(u)}\right) \\ f(t) &= \left(\frac{d}{dt} r(t) + 1\right) \frac{r(0)}{r(t)^2} \exp\left(-\int_0^t \frac{du}{r(u)}\right) \\ \lambda(t) &= \frac{1}{r(t)} \left(\frac{d}{dt} r(t) + 1\right) \end{aligned}$$

We note for the mean life time that $r(0) = E[T]$.

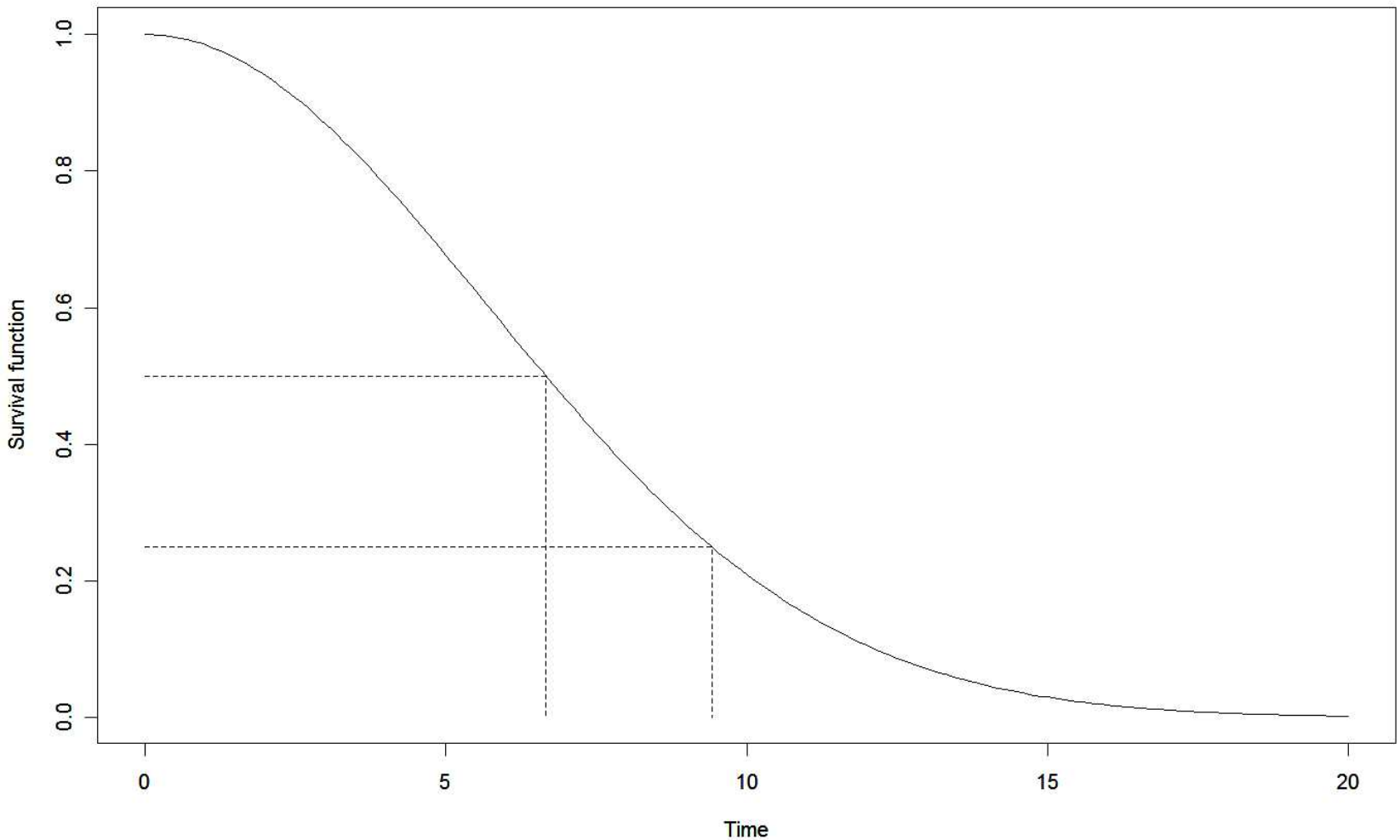
Hereby we can show that

$$r(0) = E[T] = \int_0^{+\infty} tf(t)dt = \int_0^{+\infty} S(t)dt.$$

Sometimes we are confronted in reliability and survival analysis that $E[T] = +\infty$. Therefore we introduce percentiles.

The $p$ -th percentile $t_p$ is the solution of the equation

$$S(t_p) = 1 - p.$$


Survival function
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
Time

# $T$ is discrete

The probability function is given by

$$f(a_i) = P(T = a_i), \quad i = 1, 2, \ldots \quad \Rightarrow S(t) = \sum_{a_j > t} f(a_j).$$

The hazard function is given by

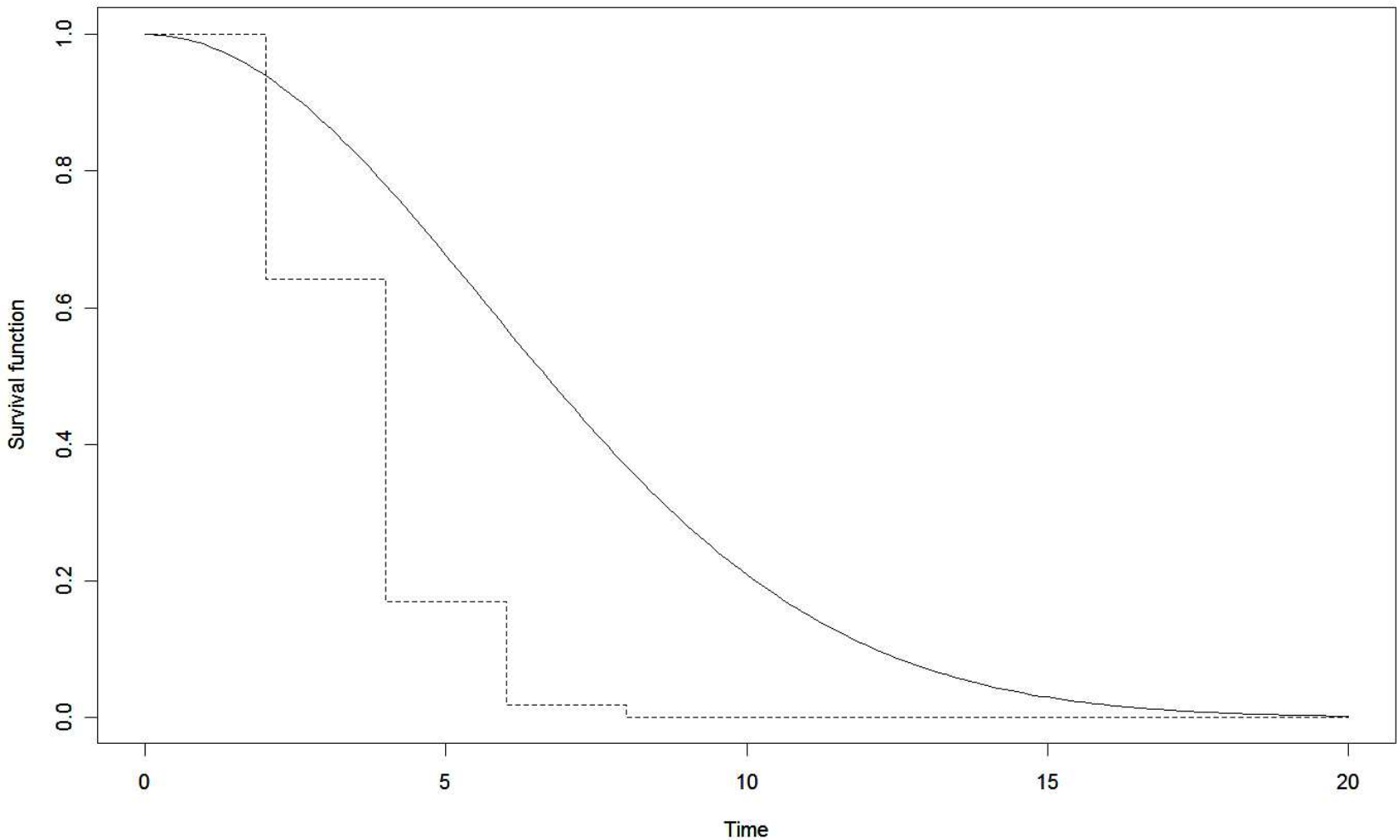
$$\lambda_i = \lambda(a_i) = P(T = a_i | T \geq a_i) = \frac{f(a_i)}{S(a_j^-)}, \quad i = 1, 2, \ldots$$

and

$$S(t) = \prod_{a_i \leq t} (1 - \lambda_i).$$

The $p$ -th percentile $t_p$ is the smallest $t_p$ such that $S(t_p) \leq 1 - p$,

$$t_p = \inf\{t : S(t) \leq 1 - p\}.$$


Survival function
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
Time

In an analysis, we first explore the data. Hereby we are also interested in estimating the survival function $S$ from right censored data.

We consider again the previous example.

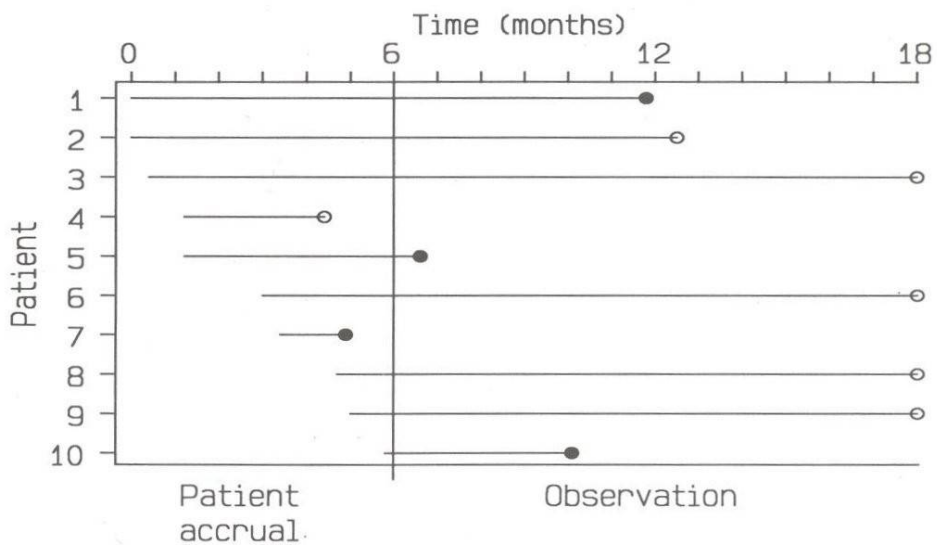



**Figure 13.1** Diagram showing patients entering a study at different times and the observation of known (•) and censored (○) survival times.

We extract the following data from the graph.

| Patient | Random. time (months) | Last obs. (months) | Died? | lifetime (months) | status (1 = yes) |
|---|---|---|---|---|---|
| 1 | 0.0 | 11.8 | yes | 11.8 | 1 |
| 2 | 0.0 | 12.5 | no | 12.5 | 0 |
| 3 | 0.4 | 18.0 | no | 17.6 | 0 |
| 4 | 1.2 | 4.4 | no | 3.2 | 0 |
| 5 | 1.2 | 6.6 | yes | 5.4 | 1 |
| 6 | 3.0 | 18.0 | no | 15.0 | 0 |
| 7 | 3.4 | 4.9 | yes | 1.5 | 1 |
| 8 | 4.7 | 18.0 | no | 13.3 | 0 |
| 9 | 5.0 | 18.0 | no | 13.0 | 0 |
| 10 | 5.8 | 10.1 | yes | 4.3 | 1 |

**Example:** We want to know the probability that a patient has died at 6 months?

If we ignore censoring, we use the empirical distribution function and estimate

$$\hat{p}(6) = \frac{1}{10}\sum_{i=1}^{10} I(\text{lifetime}_i \leq 6) = \frac{4}{10}.$$

However, one patient (no. 4) has survived 3.2 months (censored observation).

How to deal with this patient?

- Assume no. 4 died before 6 months → overestimation ($\frac{4}{10}$)
- Assume no. 4 survived 6 months → underestimation ($\frac{3}{10}$)
- Ignore no. 4 → loss of information ($\frac{3}{9}$)

Hence, censoring is causing problems!

To derive an estimator for the survival distribution of the lifetime, we introduce the following notation:

- The $j$-th time interval is $[a_{j-1}, a_j[$, $j = 1, \ldots, K+1$. $a_0 = 0$ and $a_{K+1} = +\infty$.
- $d_j$ is # failures in the $j$-th interval.
- $c_j$ is # censored observations in the $j$-th interval.
- $n_j$ is # individuals entering the $j$-th interval.

We decompose the survival function $S(a_j)$ as

$$\begin{aligned} S(a_j) &= P(T > a_j | T > a_{j-1}) S(a_{j-1}) \\ &= P(T > a_j | T > a_{j-1}) \dots P(T > a_1 | T > a_0) S(a_0) \end{aligned}$$

For each interval $[a_{j-1}, a_j[$, we estimate the conditional survival probability by

$$\begin{aligned} P(T > a_j | T > a_{j-1}) &= P(\text{Survive interval } [a_{j-1}, a_j[ | T > a_{j-1}) \\ &= 1 - \frac{\text{Number failures}_j}{\text{Number at risk}_j}. \end{aligned}$$

Some questions:

1. How do we define the number at risk $n'_j$ in the $j$-th interval?
2. What should we do with censored people?

Answer:

We assume that the censoring times are uniformly distributed over the interval, and hence occur on average half-way through the interval. For each interval, we define the number at risk $n'_j$ as

$$n'_j = n_j - \frac{c_j}{2}.$$

We call the obtained estimator, the Actuarial estimator.

Until now, we assumed that the individual lifetimes were observed. Sometimes they are grouped in fixed time intervals.

Some examples:

- Cohort life table: presents the actual mortality experience from birth to death for a group of people born at about the same time.
- Current life table: From a cross-sectional study, mostly census information, the number of individuals alive at each age is recorded, together with statistics on number of deaths in each age group.
- Clinical life table: applies to grouped survival data from studies in patients with a specific disease.

# An example: Time to weaning of breast-fed newborns

- In the National Labor Survey of Youth (NLSY) data set, youths between age 14 and 21 were yearly interviewed.
- Females were asked about pregnancies and breast-feeding.
- Data on 927 breast-fed first-born children.
- Duration of breast-feeding was recorded in weeks. An indicator whether breast-feeding was completed.

| Weeks | $n_j$ | $c_j$ | $d_j$ |
|---|---|---|---|
| $[0, 2[$ | 927 | 2 | 77 |
| $[2, 3[$ | 848 | 3 | 71 |
| $[3, 5[$ | 774 | 6 | 119 |
| $[5, 7[$ | 649 | 9 | 75 |
| $[7, 11[$ | 565 | 7 | 109 |
| $[11, 17[$ | 449 | 5 | 148 |
| $[17, 25[$ | 296 | 3 | 107 |
| $[25, 37[$ | 186 | 0 | 74 |
| $[37, 53[$ | 112 | 0 | 85 |
| $[53, \infty[$ | 27 | 0 | 27 |

# Using SAS software

```
data wean;
input Weeks Status Freq;
cards;
1 1 77
2.5 1 71
4 1 119
6 1 75
8 1 109
15 1 148
20 1 107
30 1 74
40 1 85
60 1 27
1 0 2
2.5 0 3
4 0 6
6 0 9
8 0 7
15 0 5
20 0 3
30 0 0
40 0 0
60 0 0 ;
```

```
proc lifetest data=wean method=lt intervals=(0 2 3 5 7 11 17 25 37 53);
time Weeks*Status(0);
freq Freq;
run;
```

The LIFETEST Procedure

Life Table Survival Estimates

| Interval [Lower, Upper) | Number Failed | Number Censored | Effective Sample Size | Conditional Probability of Failure | Conditional Probability Standard Error | Survival | Failure |
|---|---|---|---|---|---|---|---|
| 0 | 2 | 77 | 2 | 926.0 | 0.0832 | 0.00907 | 1.0000 | 0 |
| 2 | 3 | 71 | 3 | 846.5 | 0.0839 | 0.00953 | 0.9168 | 0.0832 |
| 3 | 5 | 119 | 6 | 771.0 | 0.1543 | 0.0130 | 0.8399 | 0.1601 |
| 5 | 7 | 75 | 9 | 644.5 | 0.1164 | 0.0126 | 0.7103 | 0.2897 |
| 7 | 11 | 109 | 7 | 561.5 | 0.1941 | 0.0167 | 0.6276 | 0.3724 |
| 11 | 17 | 148 | 5 | 446.5 | 0.3315 | 0.0223 | 0.5058 | 0.4942 |
| 17 | 25 | 107 | 3 | 294.5 | 0.3633 | 0.0280 | 0.3381 | 0.6619 |
| 25 | 37 | 74 | 0 | 186.0 | 0.3978 | 0.0359 | 0.2153 | 0.7847 |
| 37 | 53 | 85 | 0 | 112.0 | 0.7589 | 0.0404 | 0.1296 | 0.8704 |
| 53 | . | 27 | 0 | 27.0 | 1.0000 | 0 | 0.0313 | 0.9687 |

Evaluated at the Midpoint of the Interval

| Interval [Lower, | Upper) | Survival Standard Error | Median Residual Lifetime | Median Standard Error | PDF | PDF Standard Error | Hazard | Hazard Standard Error |
|---|---|---|---|---|---|---|---|---|
| 0 | 2 | 0 | 11.2078 | 0.5880 | 0.0416 | 0.00454 | 0.04338 | 0.004939 |
| 2 | 3 | 0.00907 | 10.6957 | 0.5639 | 0.0769 | 0.00877 | 0.087546 | 0.01038 |
| 3 | 5 | 0.0121 | 11.0717 | 0.5413 | 0.0648 | 0.00554 | 0.083626 | 0.007639 |
| 5 | 7 | 0.0149 | 11.3915 | 0.5006 | 0.0413 | 0.00457 | 0.061779 | 0.00712 |
| 7 | 11 | 0.0160 | 11.5839 | 0.8624 | 0.0305 | 0.00273 | 0.053748 | 0.005118 |
| 11 | 17 | 0.0166 | 11.5508 | 0.7793 | 0.0279 | 0.00209 | 0.066219 | 0.005335 |
| 17 | 25 | 0.0158 | 14.4748 | 1.3803 | 0.0154 | 0.00139 | 0.055498 | 0.005231 |
| 25 | 37 | 0.0138 | 15.5765 | 1.2836 | 0.00714 | 0.000790 | 0.041387 | 0.00466 |
| 37 | 53 | 0.0114 | 10.5412 | 0.9960 | 0.00615 | 0.000630 | 0.076439 | 0.00656 |
| 53 | . | 0.00591 | . | . | . | . | . | . |

Summary of the Number of Censored and Uncensored Values

| Total | Failed | Censored | Percent Censored |
|---|---|---|---|
| 927 | 892 | 35 | 3.78 |

NOTE: There were 3 observations with missing values, negative time values or frequency values less than 1.

The main quantity of interest is the probability that an event will not occur by time $t$:

$$S(t) = P(T > t).$$

Kaplan and Meier (1958) develop an estimator for the survival function

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)^{\delta_i} = \prod_{t_i \leq t} \left(\frac{n_i - d_i}{n_i}\right)^{\delta_i}.$$

where

- $d_i$ = number of patients died at $t_i$
- $n_i$ = number of patients at risk before $t_i$

This estimator is called Kaplan-Meier or Product-limit estimator.

Let $t_1 < t_2 < \ldots < t_k$ be the ordered lifetimes.

**The main idea:** conditional probability

To survive until time $t_{j+1}$, you need to first survive until time $t_j$ units, and then until $t_{j+1}$.

Symbolically:

$$\begin{aligned} S(t_{j+1}) &= P(T > t_{j+1} | T > t_j) S(t_j) \\ &= P(\text{Survive interval } ]t_j, t_{j+1}] | T > t_j) S(t_j). \end{aligned}$$

There are three possibilities for each interval:

- There is a censoring → We assume that they survive until the end of the interval. The conditional probability is 1.
- There is a death, but no censoring → conditional probability of surviving the interval is $1 - \frac{d}{r}$ where $d$ is the number of deaths within the interval and $r$ the number at risk at the beginning of the interval.
- There are tied deaths and censoring → We assume that censoring occurs at the end of the interval such that the conditional probability is $1 - \frac{d}{r}$.

| Lifetime (Months) | Status | Number Events ($d_i$) | Number at risk ($n_i$) | $\hat{S}(t)$ |
|---|---|---|---|---|
| 0 | | | | 1 |
| 1.5 | 1 | 1 | 10 | $S(0)\,[1-1/10] = 0.9000$ |
| 3.2 | 0 | 0 | 9 | $S(1.5) \times 1 = 0.9000$ |
| 4.3 | 1 | 1 | 8 | $S(3.2)\,[1-1/8] = 0.7875$ |
| 5.4 | 1 | 1 | 7 | $S(4.3)\,[1-1/7] = 0.6750$ |
| 11.8 | 1 | 1 | 6 | $S(5.4)\,[1-1/6] = 0.5625$ |
| 12.5 | 0 | 0 | 5 | $S(11.8) \times 1 = 0.5625$ |
| 13.0 | 0 | 0 | 4 | $S(12.5) \times 1 = 0.5625$ |
| 13.3 | 0 | 0 | 3 | $S(13.0) \times 1 = 0.5625$ |
| 15.0 | 0 | 0 | 2 | $S(13.3) \times 1 = 0.5625$ |
| 17.6 | 0 | 0 | 1 | $S(15.0) \times 1 = 0.5625$ |

Returning to the introduction, we note that the probability of surviving more than 6 months is

$$\hat{S}(6) = 67.5\%.$$

Comparing with some "naive" estimators.

- Assume patient with $t = 3.2$ died before 6 months $\rightarrow 60\%$.
- Assume patient with $t = 3.2$ survived 6 months $\rightarrow 70\%$.
- Ignore patient with $t = 3.2 \rightarrow 66.6\%$

# Some properties for the KM-estimator

- The KM-estimator is a step-function which only jumps at uncensored observations. The different jumps are random, dependent on censored observations.

$$W_j = \hat{S}(t_{j-1}) - \hat{S}(t_j) = \frac{d_j}{n_j} \prod_{t_i \leq t_{j-1}} \left(1 - \frac{d_i}{n_i}\right)^{\delta_i}.$$

- When the largest observation $t_k$ is censored, the KM-estimator does not converge to zero at infinity and is often taken as undefined.
- The KM-estimator is the nonparametric MLE of

$$L = \prod_{j=0}^{k} \left\{ \left[S(t_j^-) - S(t_j)\right]^{d_j} \prod_{l=1}^{m_j} S(t_{jl}) \right\}.$$

If there is no censoring, the KM-estimator reduces to the empirical survival function.

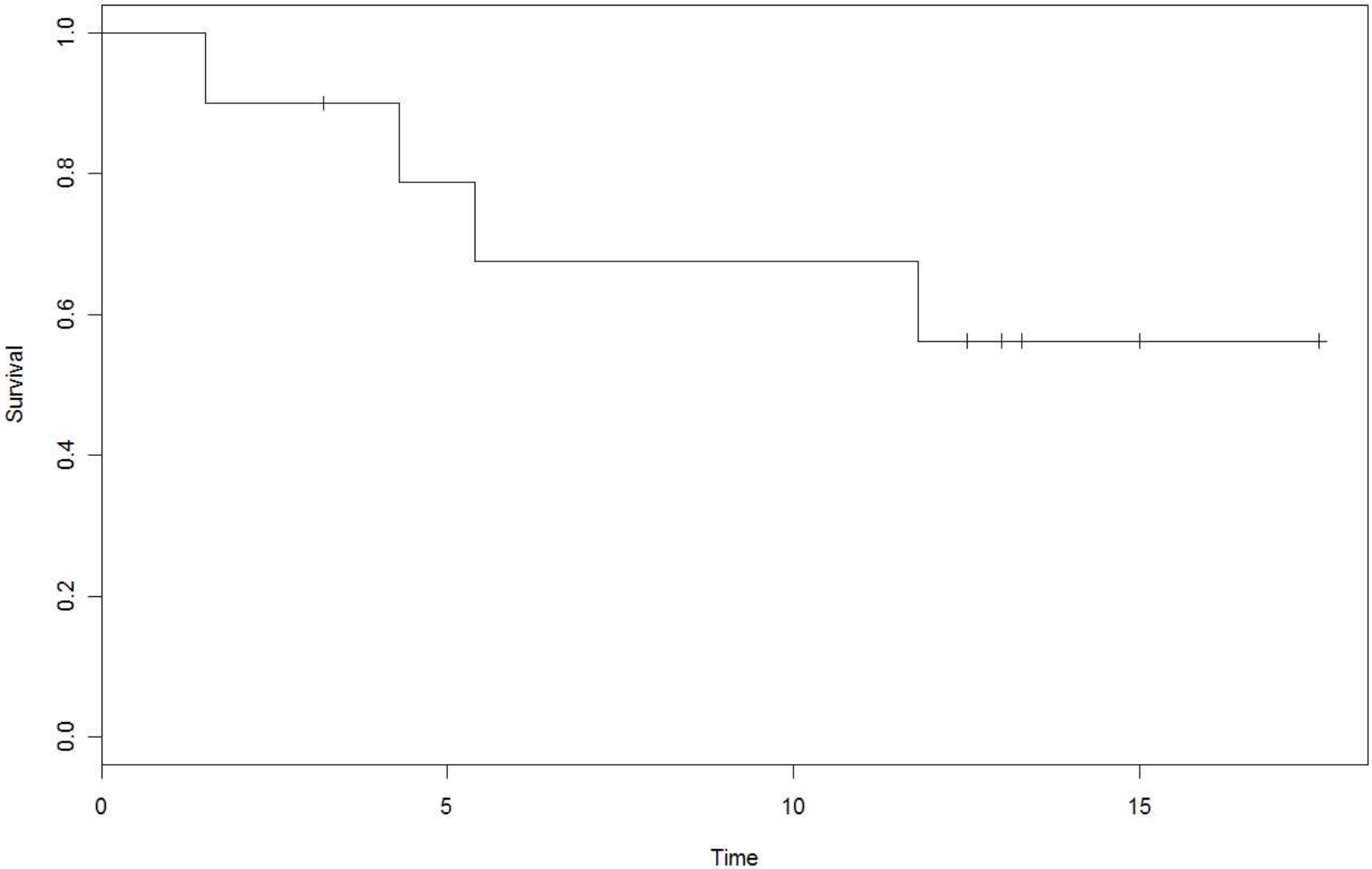
If $t_1 < t_2 < \ldots < t_k$ then $n_i = n - \sum_{j=1}^{i-1} d_j$.

Hence

$$\begin{aligned}
\hat{S}(t) &= \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \\
&= \left(\frac{n - d_1}{n}\right)\left(\frac{n - d_1 - d_2}{n - d_1}\right) \ldots \left(\frac{n - d_1 - \ldots - d_i}{n - d_1 - \ldots - d_{i-1}}\right) \\
&= \frac{n - \sum_{j=1}^{i} d_j}{n} = \frac{\#\text{ observations } > t}{n}.
\end{aligned}$$

# Using R software

```
> library(survival)

> Time<-c(1.5,3.2,4.3,5.4,11.8,12.5,13.0,13.3,15.0,17.6)
> Status<-c(1,0,1,1,1,0,0,0,0,0)

> Surv(Time,Status)
 [1]  1.5   3.2+  4.3   5.4   11.8  12.5+ 13.0+ 13.3+ 15.0+ 17.6+

> survfit(Surv(Time,Status)~1)
Call: survfit(formula = Surv(Time, Status) ~ 1)

records   n.max n.start  events  median 0.95LCL 0.95UCL
   10.0    10.0    10.0     4.0      NA     5.4      NA

> summary(survfit(Surv(Time,Status)~1))
Call: survfit(formula = Surv(Time, Status) ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  1.5     10       1    0.900  0.0949        0.732            1
  4.3      8       1    0.787  0.1340        0.564            1
  5.4      7       1    0.675  0.1551        0.430            1
 11.8      6       1    0.562  0.1651        0.316            1

> plot(survfit(Surv(Time,Status)~1,conf.type="none"),xlab="Time",ylab="Survival")
```


Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
Time

# Using SAS software

```
data clin;
input Time Status;
cards;
1.5 1
3.2 0
...
17.6 0
;
run;

proc lifetest data=clin;
time Time*Status(0);
run;
```

The LIFETEST Procedure

Product-Limit Survival Estimates

| Time | Survival | Failure | Survival Standard Error | Number Failed | Number Left |
|---|---|---|---|---|---|
| 0.0000 | 1.0000 | 0 | 0 | 0 | 10 |
| 1.5000 | 0.9000 | 0.1000 | 0.0949 | 1 | 9 |
| 3.2000* | . | . | . | 1 | 8 |
| 4.3000 | 0.7875 | 0.2125 | 0.1340 | 2 | 7 |
| 5.4000 | 0.6750 | 0.3250 | 0.1551 | 3 | 6 |
| 11.8000 | 0.5625 | 0.4375 | 0.1651 | 4 | 5 |
| 12.5000* | . | . | . | 4 | 4 |
| 13.0000* | . | . | . | 4 | 3 |
| 13.3000* | . | . | . | 4 | 2 |
| 15.0000* | . | . | . | 4 | 1 |
| 17.6000* | . | . | . | 4 | 0 |

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable Time

Quartile Estimates

| Percent | Point Estimate | 95% Confidence Interval [Lower | Upper) |
|---|---|---|---|
| 75 | . | 11.8000 | . |
| 50 | . | 5.4000 | . |
| 25 | 5.4000 | 1.5000 | . |

| Mean | Standard Error |
|---|---|
| 9.2063 | 1.4535 |

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values

| Total | Failed | Censored | Percent Censored |
|---|---|---|---|
| 10 | 4 | 6 | 60.00 |

An alternative estimator for the Kaplan-Meier estimator is the Nelson-Aalen estimator.

Hereby we used the link between the survival function and the cumulative hazard function,

$$S(t) = \exp(-\Lambda(t)).$$

Estimating first the cumulative hazard function, we then get another estimator for the survival function,

$$\hat{\Lambda}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \Rightarrow \hat{S}(t) = \exp\left(-\sum_{t_i \leq t} \frac{d_i}{n_i}\right).$$

```
proc lifetest data=clin nelson method=breslow;
time Time*Status(0);
run;
```

The LIFETEST Procedure

Survival Function and Cumulative Hazard Rate

| | Breslow | | | Nelson-Aalen | | | |
|---|---|---|---|---|---|---|---|
| Time | Survival | Failure | Survival Standard Error | Cumulative Hazard | Cum Haz Standard Error | Number Failed | Number Left |
| 0.0000 | 1.0000 | 0 | 0 | 0 | . | 0 | 10 |
| 1.5000 | 0.9048 | 0.0952 | 0.0954 | 0.1000 | 0.1000 | 1 | 9 |
| 3.2000* | . | . | . | . | . | 1 | 8 |
| 4.3000 | 0.7985 | 0.2015 | 0.1359 | 0.2250 | 0.1601 | 2 | 7 |
| 5.4000 | 0.6922 | 0.3078 | 0.1590 | 0.3679 | 0.2146 | 3 | 6 |
| 11.8000 | 0.5859 | 0.4141 | 0.1719 | 0.5345 | 0.2717 | 4 | 5 |
| 12.5000* | . | . | . | . | . | 4 | 4 |
| 13.0000* | . | . | . | . | . | 4 | 3 |
| 13.3000* | . | . | . | . | . | 4 | 2 |
| 15.0000* | . | . | . | . | . | 4 | 1 |
| 17.6000* | 0.5859 | 0.4141 | . | . | . | 4 | 0 |

We note that

$$\hat{S}(5.4) = \exp(-0.3679) = 0.6922.$$

# Using R software

```
> library(survival)

> Time<-c(1.5,3.2,4.3,5.4,11.8,12.5,13.0,13.3,15.0,17.6)
> Status<-c(1,0,1,1,1,0,0,0,0,0)

> Surv(Time,Status)
[1]  1.5   3.2+  4.3   5.4  11.8  12.5+ 13.0+ 13.3+ 15.0+ 17.6+

> survfit(Surv(Time,Status)~1)
Call: survfit(formula = Surv(Time, Status) ~ 1)

records   n.max n.start  events  median 0.95LCL 0.95UCL
10.0    10.0    10.0     4.0      NA     5.4      NA

> summary(survfit(Surv(Time,Status)~1,type="fleming-harrington"))
Call: survfit(formula = Surv(Time, Status) ~ 1, type = "fleming-harrington")

time n.risk n.event survival std.err lower 95% CI upper 95% CI
 1.5     10       1    0.905  0.0954        0.736            1
 4.3      8       1    0.799  0.1359        0.572            1
 5.4      7       1    0.692  0.1590        0.441            1
11.8      6       1    0.586  0.1719        0.330            1
```

Next to an estimate $\hat{S}(t)$ for the survival function $S(t)$ at $t$, we want to have an idea about the variability of this estimate.

This is given by Greenwood's formula

$$\widehat{\text{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

where $t_1, \ldots, t_k$ are the uncensored observed lifetimes.

Looking at the survival function, we rewrite as

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \hat{\lambda}_j\right)$$

where $\hat{\lambda}_j = \frac{d_j}{n_j}$.

The number of events $d_j$ has a binomial distribution, which is approximated by a normal distribution. Hence,

$$\widehat{\text{Var}}(\hat{\lambda}_j) = \frac{\hat{\lambda}_j \left(1 - \hat{\lambda}_j\right)}{n_j}.$$

In large samples, the $\hat{\lambda}_j$ are independent.

Instead of $\hat{S}(t)$, we look at

$$\log\left[\hat{S}(t)\right] = \sum_{t_j \leq t} \log\left(1 - \hat{\lambda}_j\right).$$

Using the delta-method, we get that

$$\begin{aligned}
\widehat{\text{Var}}\left[\log\left(\hat{S}(t)\right)\right] &= \sum_{t_j \leq t} \widehat{\text{Var}}\left[\log\left(1 - \hat{\lambda}_j\right)\right] \\
&= \sum_{t_j \leq t} \left(\frac{1}{1 - \hat{\lambda}_j}\right)^2 \frac{\hat{\lambda}_j(1 - \hat{\lambda}_j)}{n_j} \\
&= \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.
\end{aligned}$$

Since $\hat{S}(t) = \exp\left[\log(\hat{S}(t))\right]$, we get the result from the delta-method,

$$\widehat{\text{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

**Delta-method:**

If $Y$ is normally distributed with mean $\mu$ and variance $\sigma^2$, then $g(Y)$ is approximately normal with mean $g(\mu)$ and variance $g'(\mu)^2\sigma^2$.

With no censoring, $\hat{S}(t)$ is the empirical survival function and

$$\hat{S}(t) \approx N\left(S(t), \frac{S(t)(1-S(t))}{n}\right).$$

A pointwise $(1-\alpha)\%$ confidence interval for $S(t)$ is given by

$$\hat{S}(t) \pm z_{1-\frac{\alpha}{2}} \text{s.e.}(\hat{S}(t)).$$

With censoring,

- $\hat{S}(t)$ still approximately normal.
- the mean of $\hat{S}(t)$ is still the true $S(t)$.
- variance $\rightarrow$ Greenwood's formula.

However, the bounds of the C.I can be $< 0$ or $> 1$ !

Hence, better C.I's by using transformations of $S(t)$.

- Log-function: $\log[\hat{S}(t)] \pm z_{1-\frac{\alpha}{2}} \text{s.e.}\log[\hat{S}(t)]$

$$\Rightarrow \left[\hat{S}(t)\exp\left(z_{1-\frac{\alpha}{2}}\hat{\tau}(t)\right), \hat{S}(t)\exp\left(-z_{1-\frac{\alpha}{2}}\hat{\tau}(t)\right)\right]$$

with $\hat{\tau}^2(t) = \frac{\widehat{\text{Var}}(\hat{S}(t))}{\hat{S}(t)^2}$ (default: R).

- Log-log-function: $\log(-\log[\hat{S}(t)]) \pm z_{1-\frac{\alpha}{2}} \text{s.e.}\log(-\log[\hat{S}(t)])$

$$\Rightarrow \left[\hat{S}(t)^{\exp\left(z_{1-\frac{\alpha}{2}}\hat{\tau}(t)\right)}, \hat{S}(t)^{\exp\left(-z_{1-\frac{\alpha}{2}}\hat{\tau}(t)\right)}\right]$$

with $\hat{\tau}^2(t) = \frac{\widehat{\text{Var}}(\hat{S}(t))}{(\hat{S}(t)\log(\hat{S}(t)))^2}$ (default: SAS).

# A second example: Sea sickness

## Prediction of sea sickness, Burns, Aviat Space Environ Med (1984)

- 21 persons are subjected to 2-hr "rocking" with 0.167 Hz frequency and 0.111 G acceleration.
- Time until event: time when vomited for the first time.
- Two persons requested the stop of the experiment.

| Time (minutes) | Vomit (1=yes) |
|---|---|
| 30 | 1 |
| 50 | 1 |
| 50 | 0 |
| 51 | 1 |
| 66 | 0 |
| 82 | 1 |
| 92 | 1 |
| 120 | 0 |
| ... | ... |
| 120 | 0 |

```
proc lifetest data=vomit;
time time*vomit(0);
survival out=out1 conftype=log;
run;
proc print data=out1;
run;
```

The LIFETEST Procedure

Product-Limit Survival Estimates

| time | Survival | Failure | Survival Standard Error | Number Failed | Number Left |
|---|---|---|---|---|---|
| 0.000 | 1.0000 | 0 | 0 | 0 | 21 |
| 30.000 | 0.9524 | 0.0476 | 0.0465 | 1 | 20 |
| 50.000 | 0.9048 | 0.0952 | 0.0641 | 2 | 19 |
| 50.000* | . | . | . | 2 | 18 |
| 51.000 | 0.8545 | 0.1455 | 0.0778 | 3 | 17 |
| 66.000* | . | . | . | 3 | 16 |
| 82.000 | 0.8011 | 0.1989 | 0.0894 | 4 | 15 |
| 98.000 | 0.7477 | 0.2523 | 0.0981 | 5 | 14 |
| 120.000* | . | . | . | 5 | 13 |
| 120.000* | . | . | . | 5 | 12 |
| 120.000* | . | . | . | 5 | 11 |
| 120.000* | . | . | . | 5 | 10 |

| | | | | | |
|---|---|---|---|---|---|
| 120.000* | . | . | . | 5 | 9 |
| 120.000* | . | . | . | 5 | 8 |
| 120.000* | . | . | . | 5 | 7 |
| 120.000* | . | . | . | 5 | 6 |
| 120.000* | . | . | . | 5 | 5 |
| 120.000* | . | . | . | 5 | 4 |
| 120.000* | . | . | . | 5 | 3 |
| 120.000* | . | . | . | 5 | 2 |
| 120.000* | . | . | . | 5 | 1 |
| 120.000* | . | . | . | 5 | 0 |

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable time

Quartile Estimates

| Percent | Point Estimate | 95% Confidence Interval [Lower | Upper) |
|---|---|---|---|
| 75 | . | . | . |
| 50 | . | . | . |
| 25 | 98.000 | 51.000 | . |

| Mean | Standard Error |
|---|---|
| 89.259 | 4.789 |

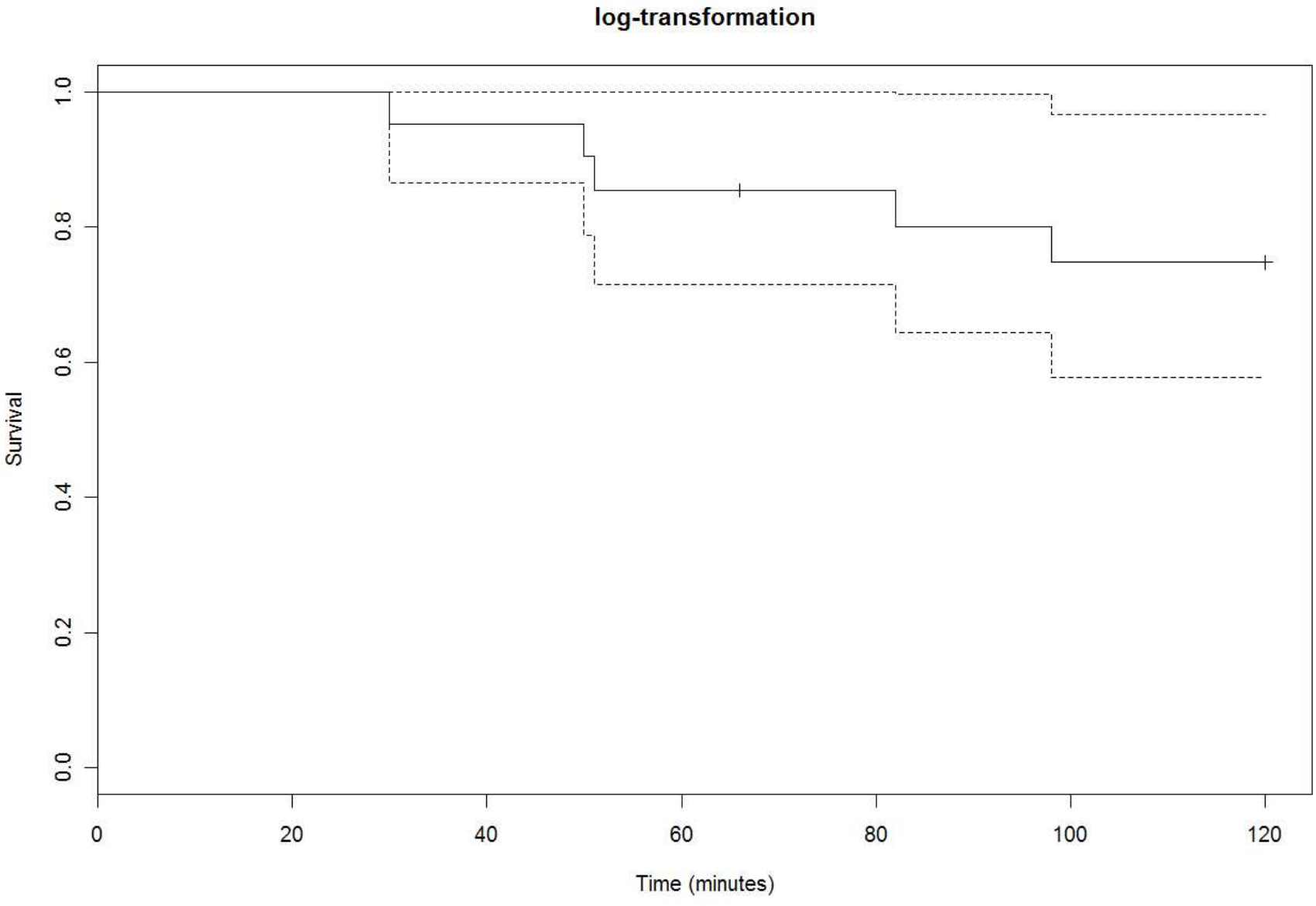
NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

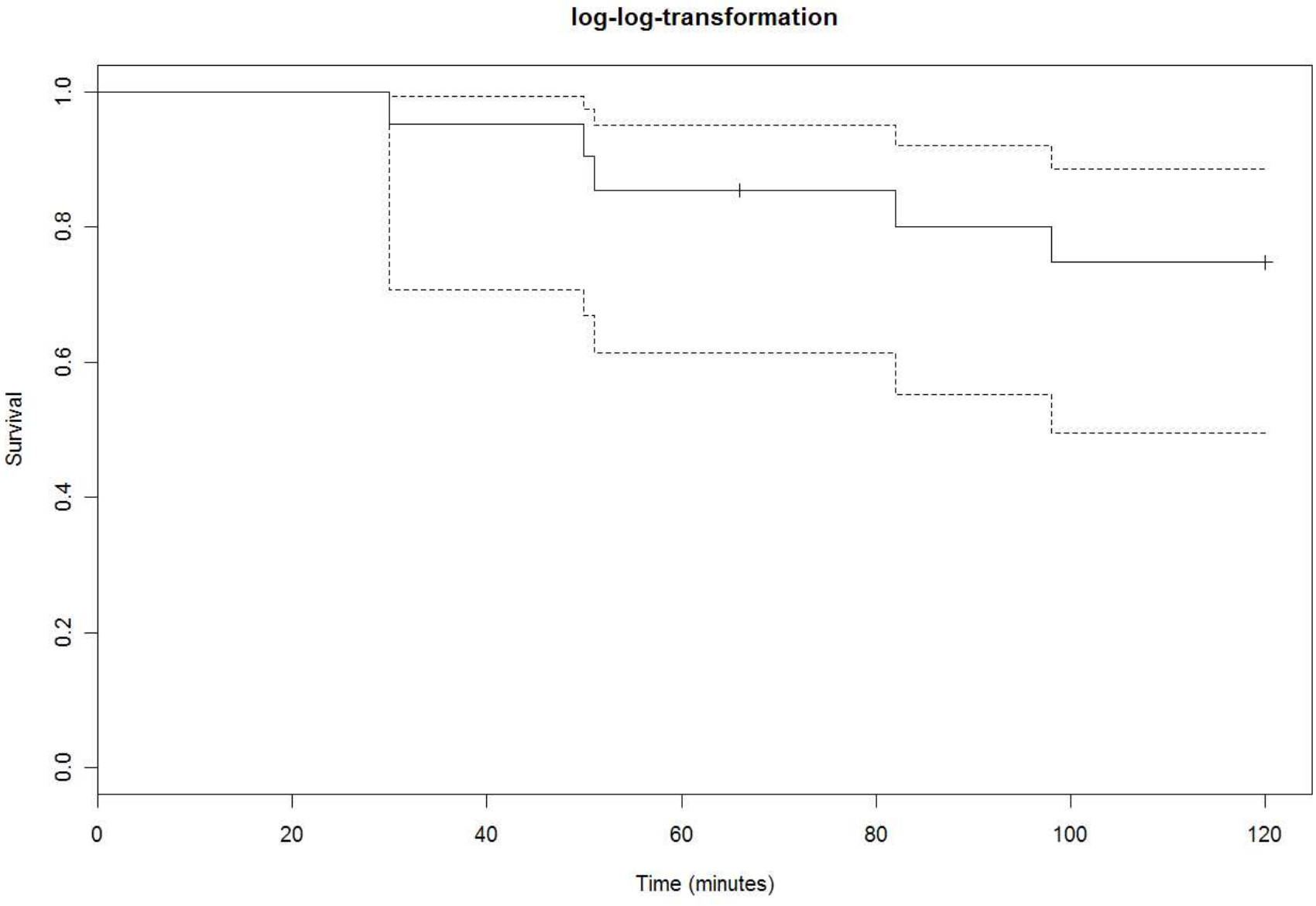
Summary of the Number of Censored and Uncensored Values

| Total | Failed | Censored | Percent Censored |
|---|---|---|---|
| 21 | 5 | 16 | 76.19 |

| Obs | time | _CENSOR_ | SURVIVAL | CONFTYPE | SDF_LCL | SDF_UCL |
|---|---|---|---|---|---|---|
| 1 | 0 | . | 1.00000 | | 1.00000 | 1.00000 |
| 2 | 30 | 0 | 0.95238 | LOG | 0.86552 | 1.00000 |
| 3 | 50 | 0 | 0.90476 | LOG | 0.78754 | 1.00000 |
| 4 | 50 | 1 | 0.90476 | | . | . |
| 5 | 51 | 0 | 0.85450 | LOG | 0.71491 | 1.00000 |
| 6 | 66 | 1 | 0.85450 | | . | . |
| 7 | 82 | 0 | 0.80109 | LOG | 0.64375 | 0.99689 |
| 8 | 98 | 0 | 0.74769 | LOG | 0.57817 | 0.96690 |
| 9 | 120 | 1 | . | | . | . |
| 10 | 120 | 1 | . | | . | . |
| 11 | 120 | 1 | . | | . | . |
| 12 | 120 | 1 | . | | . | . |
| 13 | 120 | 1 | . | | . | . |
| 14 | 120 | 1 | . | | . | . |
| 15 | 120 | 1 | . | | . | . |
| 16 | 120 | 1 | . | | . | . |
| 17 | 120 | 1 | . | | . | . |
| 18 | 120 | 1 | . | | . | . |
| 19 | 120 | 1 | . | | . | . |
| 20 | 120 | 1 | . | | . | . |
| 21 | 120 | 1 | . | | . | . |
| 22 | 120 | 1 | . | | . | . |

```
> survfit(Surv(time,vomit)~1)
Call: survfit(formula = Surv(time, vomit) ~ 1)

records   n.max n.start  events  median 0.95LCL 0.95UCL
     21      21      21       5      NA      NA      NA

> summary(survfit(Surv(time,vomit)~1))
Call: survfit(formula = Surv(time, vomit) ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   30     21       1    0.952  0.0465        0.866        1.000
   50     20       1    0.905  0.0641        0.788        1.000
   51     18       1    0.854  0.0778        0.715        1.000
   82     16       1    0.801  0.0894        0.644        0.997
   98     15       1    0.748  0.0981        0.578        0.967

> summary(survfit(Surv(time,vomit)~1,conf.type="log-log"))
Call: survfit(formula = Surv(time, vomit) ~ 1, conf.type = "log-log")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   30     21       1    0.952  0.0465        0.707        0.993
   50     20       1    0.905  0.0641        0.670        0.975
   51     18       1    0.854  0.0778        0.613        0.951
   82     16       1    0.801  0.0894        0.552        0.921
   98     15       1    0.748  0.0981        0.495        0.887

> plot(survfit(Surv(time,vomit)~1),xlab="Time (minutes)",ylab="Survival",main="log-transformation")
> plot(survfit(Surv(time,vomit)~1,conf.type="log-log"),xlab="Time (minutes)",ylab="Survival",
main="log-log-transformation")
```


log-transformation
Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
Time (minutes)


log-log-transformation
Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
Time (minutes)

# "Redistribute to the right"- algorithm

Efron developed an alternative method to compute the KM-estimator.

He set up an algorithm which starts by assuming no censoring and, as in the empirical distribution, each observation has the same mass $\frac{1}{n}$.

Afterwards in different steps, we distribute the mass of the censored observations over the observations which are still at risk.

**Example:** $3, 4, 5+, 6, 6+, 8+, 11, 14, 15, 16+$.

| Data | Step 0 | Step 1 | Step 2 | Step 3 | $\hat{S}(t)$ |
|---|---|---|---|---|---|
| 3 | $\frac{1}{10}$ | 0.100 | 0.100 | 0.100 | 0.900 |
| 4 | $\frac{1}{10}$ | 0.100 | 0.100 | 0.100 | 0.800 |
| 5+ | $\frac{1}{10}$ | 0.000 | 0.000 | 0.000 | 0.800 |
| 6 | $\frac{1}{10}$ | $\frac{1}{10} + \frac{1}{7}\frac{1}{10}$ $= 0.114$ | 0.114 | 0.114 | 0.686 |
| 6+ | $\frac{1}{10}$ | 0.114 | 0.000 | 0.000 | 0.686 |
| 8+ | $\frac{1}{10}$ | 0.114 | $0.114 + \frac{1}{5}0.114$ $= 0.137$ | 0.000 | 0.686 |
| 11 | $\frac{1}{10}$ | 0.114 | 0.137 | $0.137 + \frac{1}{4}0.137$ $= 0.171$ | 0.515 |
| 14 | $\frac{1}{10}$ | 0.114 | 0.137 | 0.171 | 0.343 |
| 15 | $\frac{1}{10}$ | 0.114 | 0.137 | 0.171 | 0.171 |
| 16+ | $\frac{1}{10}$ | 0.114 | 0.137 | 0.171 | 0.000* |

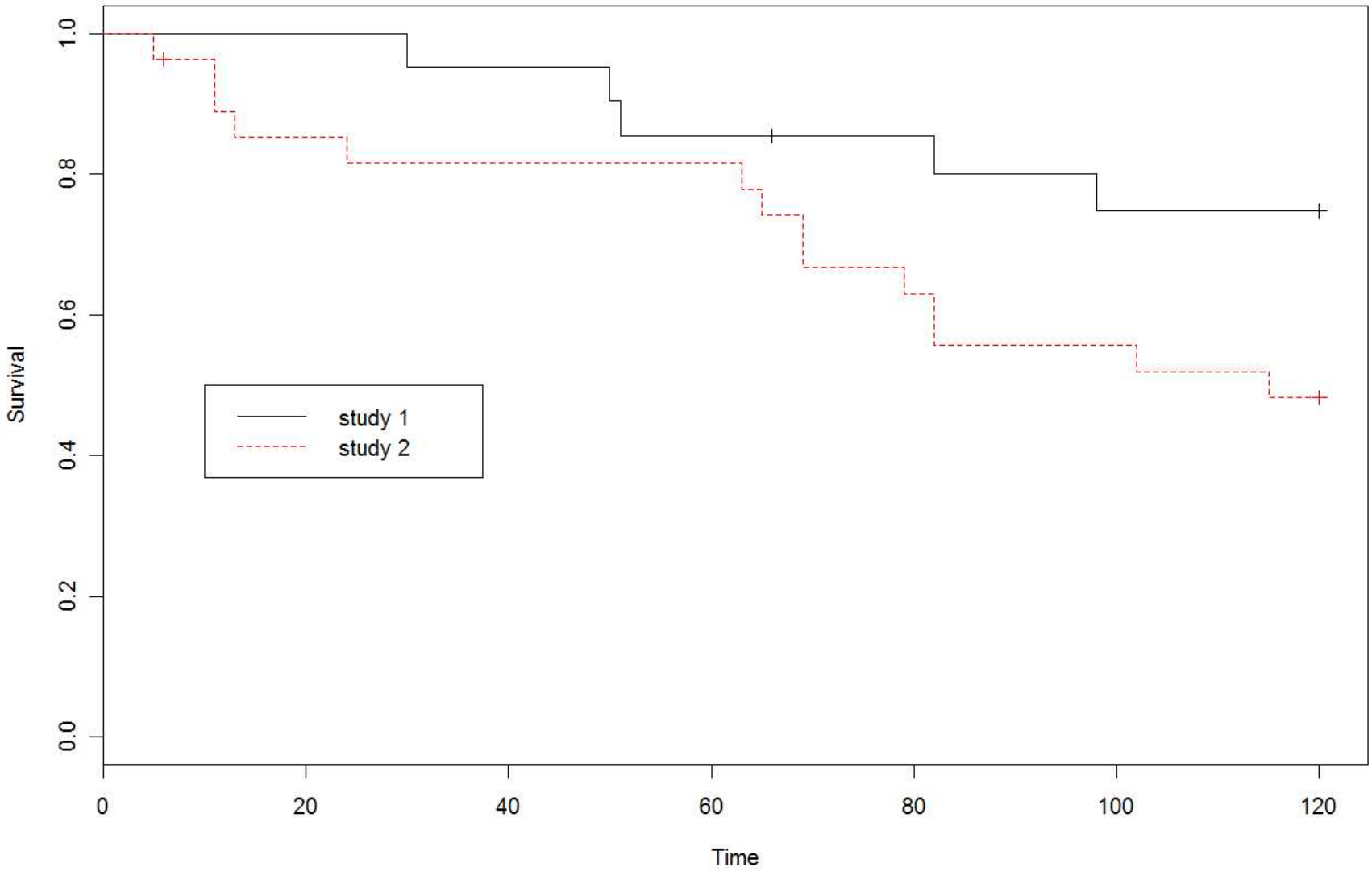
# Example: Sea sickness 2

- The "rocking" study on sea sickness was repeated with 28 subjects.
- 2 times higher frequency (0.333 Hz) and acceleration (0.222 G).
- From the table, it seems that the times are shorter in the second study!

| Time (min.) | Vomit (yes = 1) | Study |
|---|---|---|
| 30 | 1 | 1 |
| 50 | 1 | 1 |
| 50 | 0 | 1 |
| ... | ... | ... |
| 120 | 0 | 1 |
| 5 | 1 | 2 |
| 6 | 0 | 2 |
| 11 | 1 | 2 |
| ... | ... | ... |
| 120 | 0 | 2 |

```
> study<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,
+ 2,2,2,2,2,2,2,2)
> Time<-c(30,50,50,51,66,82,98,120,120,120,120,120,120,120,120,120,120,120,120,120,120,5,6,11,11,
+ 13,24,63,65,69,69,79,82,82,102,115,120,120,120,120,120,120,120,120,120,120,120,120)
> Status<-c(1,1,0,1,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,
+ 0,0,0,0,0,0,0,0)

> fit<-survfit(Surv(Time,Status)~study)
> summary(fit)
Call: survfit(formula = Surv(Time, Status) ~ study)

                study=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   30     21       1    0.952  0.0465        0.866        1.000
   50     20       1    0.905  0.0641        0.788        1.000
   51     18       1    0.854  0.0778        0.715        1.000
   82     16       1    0.801  0.0894        0.644        0.997
   98     15       1    0.748  0.0981        0.578        0.967

                study=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     28       1    0.964  0.0351        0.898        1.000
   11     26       2    0.890  0.0599        0.780        1.000
   13     24       1    0.853  0.0679        0.730        0.997
   24     23       1    0.816  0.0744        0.682        0.976
   63     22       1    0.779  0.0797        0.637        0.952
   65     21       1    0.742  0.0841        0.594        0.926
   69     20       2    0.668  0.0906        0.512        0.871
   79     18       1    0.630  0.0928        0.472        0.841
   82     17       2    0.556  0.0956        0.397        0.779
  102     15       1    0.519  0.0961        0.361        0.746
  115     14       1    0.482  0.0962        0.326        0.713
```

```
> plot(fit[1],xlab="Time",ylab="Survival")
> lines(fit[2],col="red",lty=2)
> legend(10,0.5,legend=c("study 1","study 2"),lty=c(1,2),col=c("black","red"))
```



```
data vomit2;
input time vomit study;
cards;
30  1 1
50  1 1
50  0 1
51  1 1
66  0 1
82  1 1
98  1 1
...
120 0 2
120 0 2
120 0 2
120 0 2
120 0 2
120 0 2
;
run;

proc lifetest data=vomit2;
by study;
time time*vomit(0);
run;
```

study=1

The LIFETEST Procedure

Product-Limit Survival Estimates

| time | Survival | Failure | Survival Standard Error | Number Failed | Number Left |
|---|---|---|---|---|---|
| 0.000 | 1.0000 | 0 | 0 | 0 | 21 |
| 30.000 | 0.9524 | 0.0476 | 0.0465 | 1 | 20 |
| 50.000 | 0.9048 | 0.0952 | 0.0641 | 2 | 19 |
| 50.000* | . | . | . | 2 | 18 |
| 51.000 | 0.8545 | 0.1455 | 0.0778 | 3 | 17 |
| 66.000* | . | . | . | 3 | 16 |
| 82.000 | 0.8011 | 0.1989 | 0.0894 | 4 | 15 |
| 98.000 | 0.7477 | 0.2523 | 0.0981 | 5 | 14 |
| 120.000* | . | . | . | 5 | 13 |
| 120.000* | . | . | . | 5 | 12 |
| 120.000* | . | . | . | 5 | 11 |
| 120.000* | . | . | . | 5 | 10 |
| 120.000* | . | . | . | 5 | 9 |
| 120.000* | . | . | . | 5 | 8 |
| 120.000* | . | . | . | 5 | 7 |
| 120.000* | . | . | . | 5 | 6 |
| 120.000* | . | . | . | 5 | 5 |
| 120.000* | . | . | . | 5 | 4 |
| 120.000* | . | . | . | 5 | 3 |
| 120.000* | . | . | . | 5 | 2 |
| 120.000* | . | . | . | 5 | 1 |
| 120.000* | . | . | . | 5 | 0 |

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable time

Quartile Estimates

| Percent | Point Estimate | 95% Confidence Interval [Lower | Upper) |
|---|---|---|---|
| 75 | . | . | . |
| 50 | . | . | . |
| 25 | 98.000 | 51.000 | . |

| Mean | Standard Error |
|---|---|
| 89.259 | 4.789 |

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values

| Total | Failed | Censored | Percent Censored |
|---|---|---|---|
| 21 | 5 | 16 | 76.19 |

study=2

Product-Limit Survival Estimates

| time | Survival | Failure | Survival Standard Error | Number Failed | Number Left |
|---|---|---|---|---|---|
| 0.000 | 1.0000 | 0 | 0 | 0 | 28 |
| 5.000 | 0.9643 | 0.0357 | 0.0351 | 1 | 27 |
| 6.000* | . | . | . | 1 | 26 |
| 11.000 | . | . | . | 2 | 25 |
| 11.000 | 0.8901 | 0.1099 | 0.0599 | 3 | 24 |
| 13.000 | 0.8530 | 0.1470 | 0.0679 | 4 | 23 |
| 24.000 | 0.8159 | 0.1841 | 0.0744 | 5 | 22 |
| 63.000 | 0.7788 | 0.2212 | 0.0797 | 6 | 21 |
| 65.000 | 0.7418 | 0.2582 | 0.0841 | 7 | 20 |
| 69.000 | . | . | . | 8 | 19 |
| 69.000 | 0.6676 | 0.3324 | 0.0906 | 9 | 18 |
| 79.000 | 0.6305 | 0.3695 | 0.0928 | 10 | 17 |
| 82.000 | . | . | . | 11 | 16 |
| 82.000 | 0.5563 | 0.4437 | 0.0956 | 12 | 15 |
| 102.000 | 0.5192 | 0.4808 | 0.0961 | 13 | 14 |
| 115.000 | 0.4821 | 0.5179 | 0.0962 | 14 | 13 |
| 120.000* | . | . | . | 14 | 12 |
| 120.000* | . | . | . | 14 | 11 |
| 120.000* | . | . | . | 14 | 10 |
| 120.000* | . | . | . | 14 | 9 |
| 120.000* | . | . | . | 14 | 8 |
| 120.000* | . | . | . | 14 | 7 |
| 120.000* | . | . | . | 14 | 6 |
| 120.000* | . | . | . | 14 | 5 |
| 120.000* | . | . | . | 14 | 4 |
| 120.000* | . | . | . | 14 | 3 |
| 120.000* | . | . | . | 14 | 2 |
| 120.000* | . | . | . | 14 | 1 |
| 120.000* | . | . | . | 14 | 0 |

Summary Statistics for Time Variable time

Quartile Estimates

| Percent | Point Estimate | 95% Confidence Interval [Lower | Upper) |
|---|---|---|---|
| 75 | . | . | . |
| 50 | 115.000 | 69.000 | . |
| 25 | 65.000 | 13.000 | 82.000 |

| Mean | Standard Error |
|---|---|
| 84.739 | 7.709 |

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values

| Total | Failed | Censored | Percent Censored |
|---|---|---|---|
| 28 | 14 | 14 | 50.00 |

# Difference in survival curves at fixed points

At a fixed point $t_0$, we look at the different survival probabilities,

$$H_0: \ S_1(t_0) = S_2(t_0) = \ldots = S_K(t_0) \quad \text{vs} \quad H_a : \text{at least one} \neq .$$

Like in regression, we rewrite this hypothesis as

$$H_0 : CS(t_0) = 0$$

with $K-1 \times K$ contrast matrix $C$ and vector $S(t_0)$ given by

$$C = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & -1 \\ 0 & 1 & 0 & \ldots & 0 & -1 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{bmatrix} \quad \text{and} \quad S(t_0) = \begin{bmatrix} S_1(t_0) \\ S_2(t_0) \\ \vdots \\ S_K(t_0) \end{bmatrix}.$$

Since we have that

$$\hat{S}_j(t_0) \approx N(S_j(t_0), \text{Var}(S_j(t_0))), \quad j = 1, \ldots, K$$

Hence, under $H_0$,

$$T = (C\hat{S}(t_0))^t [C\hat{V}C^t]^{-1} (C\hat{S}(t_0)) \approx \chi^2_{K-1}$$

with $\hat{V}$ diagonal matrix containing $\widehat{\text{Var}}(\hat{S}_j(t_0))), j = 1, \ldots, K$.

**Remarks:**

- Simple method to compare survival curves, but fixed point $t_0$ has to be chosen in advance!
- More time points $t_{01}, \ldots, t_{0m} \Rightarrow$ correction multiple testing (ex: Bonferroni).

Returning to the example, we compare the survival probability in both groups at time-points $t_0 = 30, 60, 90$.

We have the hypotheses,

$$H_0 : S_1(t_0) = S_2(t_0) \quad \text{vs} \quad H_a : S_1(t_0) \neq S_2(t_0).$$

and the test is, under $H_0$,

$$T = \frac{(\hat{S}_1(t_0) - \hat{S}_2(t_0))^2}{\widehat{\text{Var}}(\hat{S}_1(t_0)) + \widehat{\text{Var}}(\hat{S}_2(t_0))} \approx \chi^2_1.$$

| $t_0$ | $\hat{S}_1(t_0)$ | $\hat{S}_2(t_0)$ | $T$ | $p$-value |
|---|---|---|---|---|
| 30 | 0.9524 (0.0465) | 0.8159 (0.0744) | 2.4205 | 0.1198 |
| 60 | 0.8545 (0.0778) | 0.8159 (0.0744) | 0.1286 | 0.7199 |
| 90 | 0.7477 (0.0981) | 0.5563 (0.0956) | 1.9525 | 0.1623 |

# Log-rank test

Instead of looking at fixed time points, we want to compare the whole survival function of different groups,

$$H_0 : S_1(t) = S_2(t) = \ldots = S_K(t),\ 0 < t < \tau.$$

Since the true survival functions are unknown in each group, we go for a nonparametric test.

Before we derive the test statistic for this nulhypothesis. We look at a simpler setting with $K = 2$,

$$H_0 : S_1(t) = S_2(t),\ 0 < t < \tau.$$

In this setting it is more clear which idea's we will use here.

Suppose we observe, from populations $j = 0, 1$,

$$(T_{j1}, \delta_{j1}), (T_{j2}, \delta_{j2}), \ldots, (T_{jn_j}, \delta_{jn_j}).$$

Under $H_0$, both populations are equal. Hence we can order the uncensored lifetimes. Denote by $\tau_1, \ldots, \tau_k$ be the $k$ ordered, distinct death times.

At the $l$-th death time, we construct a $2 \times 2$ contingency table,

| Popul / Died | yes | no | Total |
|---|---|---|---|
| 0 | $d_{0l}$ | $n_{0l} - d_{0l}$ | $n_{0l}$ |
| 1 | $d_{1l}$ | $n_{1l} - d_{1l}$ | $n_{1l}$ |
| Total | $d_l$ | $n_l - d_l$ | $n_l$ |

where $d_{jl}$ is the number of deaths and $n_{jl}$ is the number at risk in population $j$ at this time.

Now, under $H_0$ and conditional on the marginals, $d_{jl}$ has a hypergeometric distribution.

$$d_{jl} \sim \text{Hypergeometric}(n_l, d_l, n_{jl}).$$

Therefore, we find that the mean and variance of $d_{jl}$ are given by

$$\begin{aligned} \text{E}[d_{jl}] &= \frac{n_{1l} d_l}{n_l} \\ \text{Var}(d_{jl}) &= \frac{n_l - n_{1l}}{n_l - 1} n_{1l} \frac{d_l}{n_l} \left(1 - \frac{d_l}{n_l}\right) \\ &= \frac{n_{1l} n_{0l} d_l (n_l - d_l)}{n_l^2 (n_l - 1)}. \end{aligned}$$

If the sample size at each death time is sufficiently large, we approximate the hypergeometric distribution by a normal distribution.

$$d_{1l} - \mathrm{E}[d_{1l}] = d_{1l} - \frac{n_{1l}d_l}{n_l} \approx N\left(0, \frac{n_{1l}n_{0l}d_l(n_l - d_l)}{n_l^2(n_l - 1)}\right).$$

Assuming that the contingency tables at different death times are independent, we find the log-rank test,

$$T = \frac{\left[\sum_{l=1}^{k}\left(d_{1l} - \frac{n_{1l}d_l}{n_l}\right)\right]^2}{\sum_{l=1}^{k}\frac{n_{1l}n_{0l}d_l(n_l - d_l)}{n_l^2(n_l - 1)}}$$

which is, under $H_0$, approximately $\chi^2$ distributed with df 1.

# Some remarks:

- By the hypergeometric distribution, we get in the denominator a finite sample correction factor.
- The numerator can be interpreted as $\sum_{l=1}^{k}(O_l - E_l)$ where $O_l$ is the observed number of deaths in group 1, and $E_l$ is the expected number, given the risk set. furthermore $E_l$ is the proportion of deaths in group 1 among those at risk.
- It does not matter which group we choose to sum over because $\sum_{\text{groups}}(O_l - E_l) = 0$.

```
proc lifetest data=vomit2;
time time*vomit(0);
strata study;
run;

The LIFETEST Procedure

Testing Homogeneity of Survival Curves for time over Strata

        Rank Statistics

study       Log-Rank    Wilcoxon

1            -3.8607     -149.00
2             3.8607      149.00

Covariance Matrix for the Log-Rank Statistics

study             1             2

1           4.64782      -4.64782
2          -4.64782       4.64782

...

       Test of Equality over Strata

                                    Pr >
Test      Chi-Square      DF    Chi-Square

Log-Rank      3.2069       1      0.0733
Wilcoxon      3.1816       1      0.0745
-2Log(LR)     3.4928       1      0.0616
```

The log-rank test is a special case of the Tarone-Ware class of tests.

$$T = \frac{\left[\sum_{l=1}^{k} w_l \left(d_{1l} - \frac{n_{1l} d_l}{n_l}\right)\right]^2}{\sum_{l=1}^{k} w_l^2 \frac{n_{1l} n_{0l} d_l (n_l - d_l)}{n_l^2 (n_l - 1)}}$$

where $w_l \geq 0$ are weights.

| Test | $w_l$ |
|---|---|
| Log-rank | 1 |
| Wilcoxon or Gehan | $n_l$ |
| Peto-peto | $\tilde{S}(t_i)$ |
| Harrington-Fleming (p,q) | $\hat{S}(t_i)^p (1 - \hat{S}(t_i))^q, \ p, q \geq 0$ |

with $\tilde{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i + 1}\right)$.

In a practical data analysis, the choice of the weights is important.

For example,

- Log-rank test has optimal power to detect alternatives in which the hazard are proportional.
- Wilcoxon-Gehan test is more "sensitive" to "early" differences in survival curves (at later times)
- Harrington-Fleming test with $p = 0, q > 0$ is "sensitive" to "late" differences.

**Note:** the decision about the choice of the test should be made before seeing the data!

```
proc lifetest data=vomit2;
time time*vomit(0);
strata study/test=(logrank wilcoxon peto);
run;

The LIFETEST Procedure

Testing Homogeneity of Survival Curves for time over Strata

              Rank Statistics

study       Log-Rank    Wilcoxon        Peto

1            -3.8607     -149.00     -3.0632
2             3.8607      149.00      3.0632

...

Covariance Matrix for the Peto Statistics

study             1             2

1           2.94876      -2.94876
2          -2.94876       2.94876

       Test of Equality over Strata

                                    Pr >
Test      Chi-Square      DF    Chi-Square

Log-Rank      3.2069       1      0.0733
Wilcoxon      3.1816       1      0.0745
Peto          3.1822       1      0.0744
```

```
> survdiff(Surv(Time,Status)~study)
Call: survdiff(formula = Surv(Time, Status) ~ study)

         N Observed Expected (O-E)^2/E (O-E)^2/V
study=1 21        5     8.86      1.68      3.21
study=2 28       14    10.14      1.47      3.21

Chisq= 3.2  on 1 degrees of freedom, p= 0.0733

> survdiff(Surv(Time,Status)~study,rho=1)
Call: survdiff(formula = Surv(Time, Status) ~ study, rho = 1)

         N Observed Expected (O-E)^2/E (O-E)^2/V
study=1 21     4.01      7.2      1.41      3.22
study=2 28    11.49      8.3      1.22      3.22

Chisq= 3.2  on 1 degrees of freedom, p= 0.0728
```

When we have $K$ populations, we generalize the previous results.

At the $l$-th ordered death time, we now construct a $K \times 2$ contingency table,

| Popul / Died | yes | no | Total |
|---|---|---|---|
| 1 | $d_{1l}$ | $n_{1l} - d_{1l}$ | $n_{1l}$ |
| 2 | $d_{2l}$ | $n_{2l} - d_{2l}$ | $n_{2l}$ |
| ... | | | ... |
| K | $d_{Kl}$ | $n_{Kl} - d_{Kl}$ | $n_{Kl}$ |
| Total | $d_l$ | $n_l - d_l$ | $n_l$ |

where $d_{jl}$ is the number of deaths and $n_{jl}$ is the number at risk in population $j$ at this time.

Under $H_0$, the vector $\mathbf{O}_l$ of the observed deaths in groups 1 to $K-1$ at time $l$,

$$\mathbf{O}_l = (d_{1l}, \ldots, d_{K-1})$$

has a multivariate hypergeometric distribution with mean $\mathbf{E}_l$ and covariance matrix $\mathbf{V}_l$, given by

$$\begin{aligned}
\mathbf{E}_l &= \left( \frac{d_l n_{1l}}{n_l}, \ldots, \frac{d_l n_{(K-1)l}}{n_l} \right) \\
\mathbf{V}_{jjl} &= \frac{n_{jl}(n_l - n_{jl}) d_l (n_l - d_l)}{n_l^2 (n_l - 1)} \\
\mathbf{V}_{jml} &= \frac{-n_{jl} n_{ml} d_l (n_l - d_l)}{n_l^2 (n_l - 1)}.
\end{aligned}$$

Approximating this distribution by a multivariate normal distribution and using an analogous construction as before, we get a test statistic $T$,

$$T = (\mathbf{O} - \mathbf{E})\mathbf{V}^{-1}(\mathbf{O} - \mathbf{E})^t$$

which is approximately $\chi^2$ distributed with df $K - 1$.

Hereby $\mathbf{O} = \sum_{l=1}^{k} \mathbf{O}_l$, $\mathbf{E} = \sum_{l=1}^{k} \mathbf{E}_l$, $\mathbf{V} = \sum_{l=1}^{k} \mathbf{V}_l$ are the sums over the $k$ distinct death times.

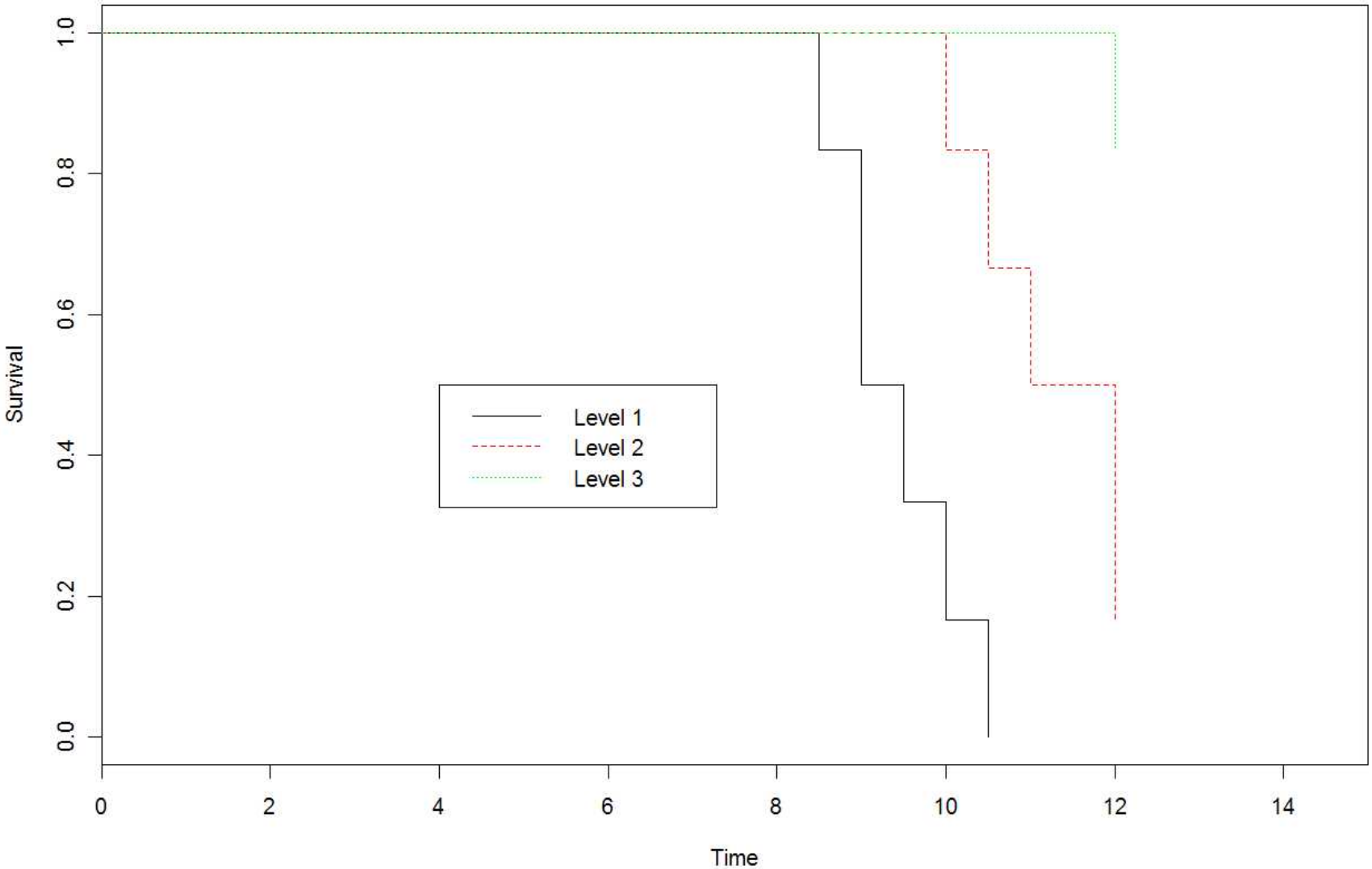
# Example: Performance testing

- Test-subject were asked to perform a certain test and the time needed was recorded.
- 3 different noise distractions are applied.
- Did the different noise distractions influence the time to finish the test?

| Noise level | | |
|---|---|---|
| 9.0 | 10.0 | 12.0 |
| 9.5 | 12.0 | $12.0^+$ |
| 9.0 | $12.0^+$ | $12.0^+$ |
| 8.5 | 11.0 | $12.0^+$ |
| 10.0 | 12.0 | $12.0^+$ |
| 10.5 | 10.5 | $12.0^+$ |

```
> Level<-c(1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3)
> Time<-c(9.0,9.5,9.0,8.5,10.0,10.5,10.0,12.0,12.0,11.0,12.0,10.5,12.0,12.0,12.0,12.0,12.0,12.0)
> Censor<-c(1,1,1,1,1,1,1,1,0,1,1,1,1,0,0,0,0,0)

> survfit(Surv(Time,Censor)~Level)
Call: survfit(formula = Surv(Time, Censor) ~ Level)

        n events median 0.95LCL 0.95UCL
Level=1 6      6   9.25     9.0     Inf
Level=2 6      5  11.50    10.5     Inf
Level=3 6      1    Inf     Inf     Inf

> fit<-survfit(Surv(Time,Censor)~Level)
> summary(fit)
Call: survfit(formula = Surv(Time, Censor) ~ Level)

                Level=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  8.5      6       1    0.833   0.152       0.5827        1.000
  9.0      5       2    0.500   0.204       0.2246        1.000
  9.5      3       1    0.333   0.192       0.1075        1.000
 10.0      2       1    0.167   0.152       0.0278        0.997
 10.5      1       1    0.000      NA           NA           NA

                Level=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
 10.0      6       1    0.833   0.152       0.5827        1.000
 10.5      5       1    0.667   0.192       0.3786        1.000
 11.0      4       1    0.500   0.204       0.2246        1.000
 12.0      3       2    0.167   0.152       0.0278        0.997

                Level=3
 time       n.risk      n.event     survival      std.err lower 95% CI upper 95% CI
 12.000      6.000        1.000        0.833        0.152        0.583        1.000
```

```
> plot(fit[1],xlab="Time",ylab="Survival",xlim=c(0.0,15.0))
> lines(fit[2],col="red",lty=2)
> lines(fit[3],col="green",lty=3)
> legend(4,0.5,legend=c("Level 1","Level 2","Level 3"),lty=c(1,2,3),col=c("black","red","green"))

> survdiff(Surv(Time,Censor)~Level)
Call: survdiff(formula = Surv(Time, Censor) ~ Level)

        N Observed Expected (O-E)^2/E (O-E)^2/V
Level=1 6        6     1.57   12.4463   17.2379
Level=2 6        5     4.53    0.0488    0.0876
Level=3 6        1     5.90    4.0660    9.4495

 Chisq= 20.4  on 2 degrees of freedom, p= 3.75e-05

> survdiff(Surv(Time,Censor)~Level,rho=1)
Call: survdiff(formula = Surv(Time, Censor) ~ Level, rho = 1)

        N Observed Expected (O-E)^2/E (O-E)^2/V
Level=1 6     5.17     1.39   10.2756   16.2355
Level=2 6     3.00     3.28    0.0235    0.0536
Level=3 6     0.50     4.00    3.0625    8.4750

Chisq= 18.3  on 2 degrees of freedom, p= 0.000105
```


Survival
1.0
0.8
0.6
0.4
0.2
0.0
Level 1
Level 2
Level 3
0
2
4
6
8
10
12
14
Time

```
data noise;
input Time censor level;
cards;
9.0 1 1
9.5 1 1
9.0 1 1
8.5 1 1
10.0 1 1
10.5 1 1
10.0 1 2
12.0 1 2
12.0 0 2
11.0 1 2
12.0 1 2
10.5 1 2
12.0 1 3
12.0 0 3
12.0 0 3
12.0 0 3
12.0 0 3
12.0 0 3
;
run;

proc lifetest data=noise;
time Time*censor(0);
strata level/test=(logrank wilcoxon peto);
run;
```

The LIFETEST Procedure

Testing Homogeneity of Survival Curves for Time over Strata

Rank Statistics

| level | Log-Rank | Wilcoxon | Peto |
|---|---|---|---|
| 1 | 4.4261 | 68.000 | 3.4232 |
| 2 | 0.4703 | -5.000 | -0.3476 |
| 3 | -4.8964 | -63.000 | -3.0756 |

...

Test of Equality over Strata

| Test | Chi-Square | DF | Pr > Chi-Square |
|---|---|---|---|
| Log-Rank | 20.3844 | 2 | <.0001 |
| Wilcoxon | 18.3265 | 2 | 0.0001 |
| Peto | 18.0014 | 2 | 0.0001 |

Sometimes we want to compare survival curves for three or more ordered groups. For example: different tumor stages (T0,T1,T2), age-groups,...

$$H_0: \quad S_1(t) = S_2(t) = \ldots = S_K(t), t \leq \tau$$

$$H_a: \quad \begin{cases} S_1(t) \geq S_2(t) \geq \ldots \geq S_K(t), t \leq \tau, \text{ at least one } > \\ \text{or} \\ S_1(t) \leq S_2(t) \leq \ldots \leq S_K(t), t \leq \tau, \text{ at least one } < \end{cases}$$

In such case one may use a logrank test for trend:

$$T = \frac{\sum_{j=1}^{K} a_j (O_j - E_j)}{\sqrt{\sum_{j=1}^{K} \sum_{k=1}^{K} a_j a_k V_{jk}}} \approx N(0,1), \text{ under } H_0,$$

where $a_1 < a_2 < \ldots < a_K$ are ordered of scores (mostly $a_j = j$) and

$$\begin{aligned} O_j &= \sum_{l=1}^{k} d_{jl} \\ E_j &= \sum_{l=1}^{k} \frac{n_{jl} d_l}{n_l}. \end{aligned}$$

For the ordered alternative hypothesis, the test for trend has got a higher statistical power than the "usual" logrank.

```
proc lifetest data=noise;
time Time*censor(0);
strata level/trend;
run;
```

The LIFETEST Procedure

Scores for Trend Test

| level | Score |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |

Trend Tests

| Test | Test Statistic | Standard Error | z-Score | Pr > \|z\| |
|---|---|---|---|---|
| Log-Rank | -9.3224 | 2.1960 | -4.2451 | <.0001 |
| Wilcoxon | -131.0000 | 32.2452 | -4.0626 | <.0001 |

```
> survdiff(Surv(Time,Censor)~Level)
Call: survdiff(formula = Surv(Time, Censor) ~ Level)

        N Observed Expected (O-E)^2/E (O-E)^2/V
Level=1 6        6     1.57   12.4463   17.2379
Level=2 6        5     4.53    0.0488    0.0876
Level=3 6        1     5.90    4.0660    9.4495

Chisq= 20.4  on 2 degrees of freedom, p= 3.75e-05

> OB<-survdiff(Surv(Time,Censor)~Level)$obs
> OB
[1] 6 5 1

> EX<-survdiff(Surv(Time,Censor)~Level)$exp
> EX
[1] 1.573950 4.529692 5.896359

> V<-survdiff(Surv(Time,Censor)~Level)$var
> V
[,1]       [,2]       [,3]
[1,]  1.1364441 -0.5619089 -0.5745352
[2,] -0.5619089  2.5244614 -1.9625525
[3,] -0.5745352 -1.9625525  2.5370877

> a<-c(1,2,3)
> test<-a%*%(OB-EX)
> stderror<-sqrt(t(a)%*%V%*%a)
> zscore<-test/stderror
> Pvalue<-2*pnorm(abs(zscore),lower.tail=FALSE)
> data.frame(test,stderror,zscore,Pvalue)
      test stderror    zscore       Pvalue
1 -9.322409 2.196042 -4.245095 2.185007e-05
```

With the log-rank test, we compare two or more survival curves.

However, sometimes there are confounding variables which also affect the outcome and for which we need to adjust for.

If the confounding variable has $M$ levels, we get

$$H_0 : S_{1m}(t) = S_{2m}(t) = \ldots = S_{Km}(t),\ t \leq \tau,\ m = 1, \ldots, M.$$

Hence, we let the shape of the survival function differ for each level of the confounding variable.

To set up a test statistic, we divide the data according the $M$ levels of the confounding variable and construct a $2 \times 2$ contingency table for each ordered death time at each level,

| Popul / Died | yes | no | Total |
|---|---|---|---|
| 1 | $d_{m0l}$ | $n_{m0l} - d_{m0l}$ | $n_{m0l}$ |
| 2 | $d_{m1l}$ | $n_{m1l} - d_{m1l}$ | $n_{m1l}$ |
| Total | $d_{ml}$ | $n_{ml} - d_{ml}$ | $n_{ml}$ |

Let $\mathbf{O}_m$ be the sum of the observed $O$'s by applying the log-rank calculations in level $m$. Similar for $\mathbf{E}_m$, the sum of the expected $E$'s and $\mathbf{V}_m$, the sum of the $v$'s.

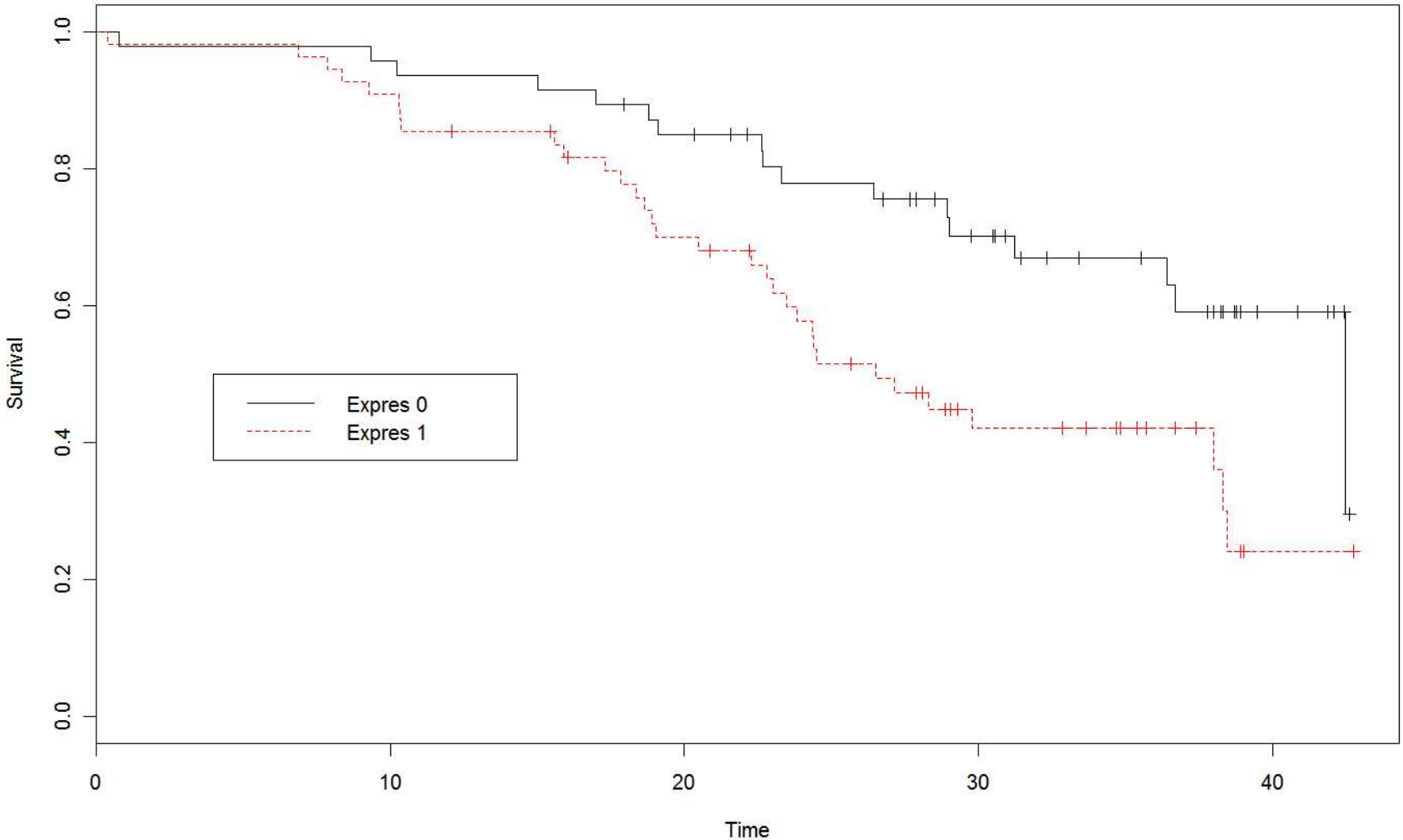
The stratified log-rank is

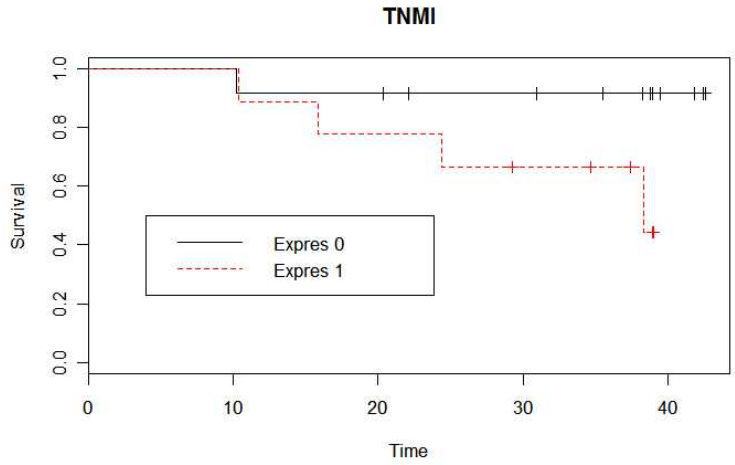
$$Z = \frac{\sum_{m=1}^{M} (\mathbf{O}_m - \mathbf{E}_m)}{\sqrt{\sum_{m=1}^{M} \mathbf{V}_m}}.$$

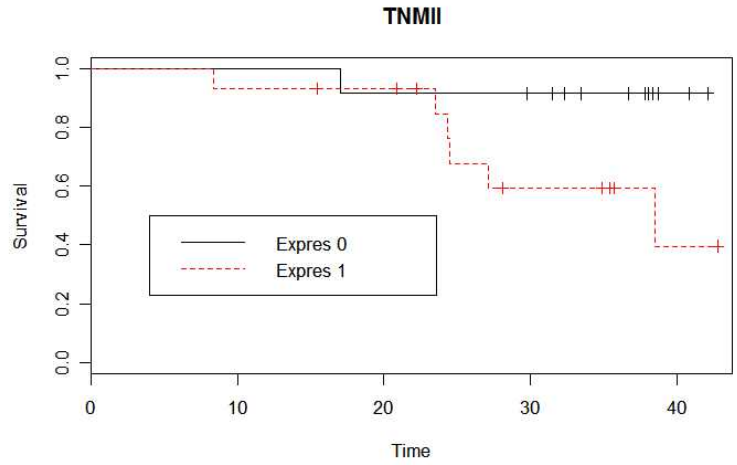
# NSCLC

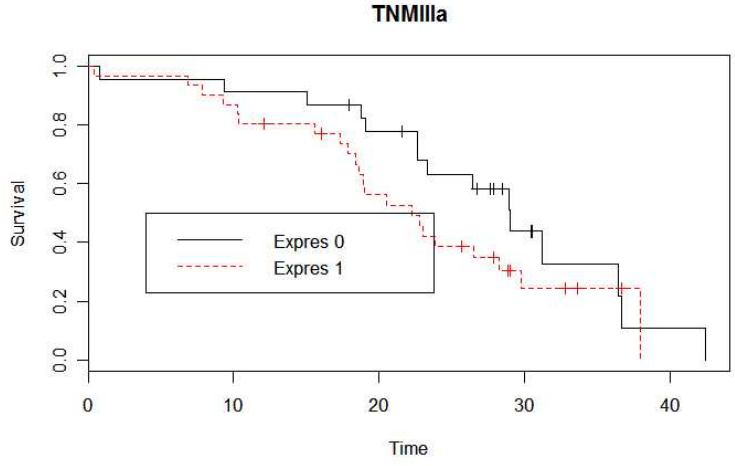
- Laudanski et al., Eur Respir J (2001).
- In this study, we had 102 patients who were operated from long cancer.
- The severity of the cancer was expressed in three TNM (Tumor, Nodes, Metastasis) categories: I, II, IIIa.
- The expression of the P53 protein was found from tumor biopsies.
- We are interested on the effect of this protein on the survival time of a patient.

```
> NSCLC<-read.table("C:/Werk/Roel/Onderwijs/Theorie/SurvivalLeuven/NSCLC.txt",header=T,sep="\t")
>
> fit<-survfit(Surv(survtime,survind)~expres,data=NSCLC)
> summary(fit)

> plot(fit[1],xlab="Time",ylab="Survival")
> lines(fit[2],col="red",lty=2)
> legend(4,0.5,legend=c("Expres 0","Expres 1"),lty=c(1,2),col=c("black","red"))
>
> survdiff(Surv(survtime,survind)~expres,data=NSCLC)
Call: survdiff(formula = Surv(survtime, survind) ~ expres, data = NSCLC)

          N Observed Expected (O-E)^2/E (O-E)^2/V
expres=0 47       17     26.2      3.25       7.1
expres=1 55       32     22.8      3.75       7.1

Chisq= 7.1  on 1 degrees of freedom, p= 0.00769

> survdiff(Surv(survtime,survind)~expres+strata(tnm),data=NSCLC)
Call: survdiff(formula = Surv(survtime, survind) ~ expres + strata(tnm), data = NSCLC)

          N Observed Expected (O-E)^2/E (O-E)^2/V
expres=0 47       17     25.5      2.82      6.05
expres=1 55       32     23.5      3.06      6.05

Chisq= 6.1  on 1 degrees of freedom, p= 0.0139
```


Survival
1.0
0.8
0.6
0.4
0.2
0.0
Expres 0
Expres 1
0
10
20
30
40
Time


TNMI
Survival
Time
Expres 0
Expres 1
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40


TNMII
Survival
Time
Expres 0
Expres 1
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40


TNMIIIa
Survival
Time
Expres 0
Expres 1
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40

```
> survdiff(Surv(survtime,survind)~tnm,data=NSCLC)
Call: survdiff(formula = Surv(survtime, survind) ~ tnm, data = NSCLC)

       N Observed Expected (O-E)^2/E (O-E)^2/V
tnm=1 21        5     13.2      5.13      7.30
tnm=2 27        7     15.8      4.91      7.32
tnm=3 54       37     19.9     14.60     26.23

Chisq= 26.3  on 2 degrees of freedom, p= 1.94e-06

> OB<-survdiff(Surv(survtime,survind)~tnm,data=NSCLC)$obs
> OB
[1]  5  7 37

> EX<-survdiff(Surv(survtime,survind)~tnm,data=NSCLC)$exp
> EX
[1] 13.24541 15.81566 19.93893

> V<-survdiff(Surv(survtime,survind)~tnm,data=NSCLC)$var
> V
[,1]      [,2]      [,3]
[1,]  9.309469 -4.417150 -4.892319
[2,] -4.417150 10.623178 -6.206028
[3,] -4.892319 -6.206028 11.098346

> a<-c(1,2,3)
> test<-a%*%(OB-EX)
> stderror<-sqrt(t(a)%*%V%*%a)
> zscore<-test/stderror
> Pvalue<-2*pnorm(abs(zscore),lower.tail=FALSE)
> data.frame(test,stderror,zscore,Pvalue)
     test   stderror  zscore       Pvalue
1 25.30649 5.494766 4.605562 4.113525e-06
```

```
proc lifetest data=nsclc;
time Survtime*Survind(0);
strata expres;
run;
```

Testing Homogeneity of Survival Curves for Survtime over Strata

Rank Statistics

| expres | Log-Rank | Wilcoxon |
|---|---|---|
| 0 | -9.2411 | -648.00 |
| 1 | 9.2411 | 648.00 |

...

Test of Equality over Strata

| Test | Chi-Square | DF | Pr > Chi-Square |
|---|---|---|---|
| Log-Rank | 7.1051 | 1 | 0.0077 |
| Wilcoxon | 6.5750 | 1 | 0.0103 |
| -2Log(LR) | 5.2939 | 1 | 0.0214 |

```
proc lifetest data=nsclc;
time Survtime*Survind(0);
test expres;
strata tnm;
run;
```

Univariate Chi-Squares for the Log-Rank Test

| Variable | Test Statistic | Standard Deviation | Chi-Square | Pr > Chi-Square |
|---|---|---|---|---|
| expres | -8.4782 | 3.4457 | 6.0541 | 0.0139 |

Covariance Matrix for the Log-Rank Statistics

| Variable | expres |
|---|---|
| expres | 11.8730 |

Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test

| Variable | DF | Chi-Square | Pr > Chi-Square | Chi-Square Increment | Pr > Increment |
|---|---|---|---|---|---|
| expres | 1 | 6.0541 | 0.0139 | 6.0541 | 0.0139 |

```
proc lifetest data=nsclc;
time Survtime*Survind(0);
strata tnm/trend;
run;
```

Summary of the Number of Censored and Uncensored Values

| Stratum | tnm | Total | Failed | Censored | Percent Censored |
|---|---|---|---|---|---|
| 1 | 1 | 21 | 5 | 16 | 76.19 |
| 2 | 2 | 27 | 7 | 20 | 74.07 |
| 3 | 3 | 54 | 37 | 17 | 31.48 |
| Total | | 102 | 49 | 53 | 51.96 |

Trend Tests

| Test | Test Statistic | Standard Error | z-Score | Pr > \|z\| |
|---|---|---|---|---|
| Log-Rank | 25.3065 | 5.4948 | 4.6056 | <.0001 |
| Wilcoxon | 1502.0000 | 403.6583 | 3.7210 | 0.0002 |

In the previous sections, we looked at non-parametric methods.

However, in most studies there are explanatory variables for which we want to look at their influence on the time till event.

The most commonly used classes of regression models in survival data are

- Cox's regression model or proportional hazard model

$$\lambda(t|\mathbf{X}) = \lambda_0(t)g(\beta_1 X_1 + \ldots + \beta_p X_p)$$

with $g \geq 0$ a known function, mostly $g(u) = e^u$.

- Accelerated failure time model (AFT)

$$\log(T) = \mu + \beta_1 X_1 + \ldots + \beta_p X_p + \sigma W$$

with $W$ a (parametric) error-distribution.

Here, we study the Cox's regression model or proportional hazard model, introduced by Cox (1972).

In this model, the conditional hazard of an individual, given the covariate values $X_1, \ldots, X_p$, is defined as

$$\lambda(t|\mathbf{X}) = \lambda_0(t)e^{\beta_1 X_1 + \ldots + \beta_p X_p} = \lambda_0(t)e^{\beta^t \mathbf{X}}.$$

where $\lambda_0(t)$ is called the baseline hazard.

The strength of this model is that $\lambda_0(t)$ is left unspecified (unknown function). It represents the hazard of an individual with covariates equal to zero.

We call this a semi-parametric regression model.

This model is often also called proportional hazard model.

Consider two individuals with covariates $\mathbf{X}$ and $\mathbf{X}^*$, the ratio of their hazards is

$$HR(t) = \frac{\lambda(t|\mathbf{X})}{\lambda(t|\mathbf{X}^*)} = \frac{\lambda_0(t)e^{\beta^t\mathbf{X}}}{\lambda_0(t)e^{\beta^t\mathbf{X}^*}} = e^{\sum_{i=1}^{p}\beta_i(X_i - X_i^*)}$$

which is a constant over time.

This quantity is called the hazard ratio and compares the hazard of having an event with covariate value $\mathbf{X}$ to the hazard of having an event with covariate value $\mathbf{X}^*$.

Often, the ratio is (inappropriately) called a *relative risk*.

If $X$ is an indicator for gender: Male ($X = 0$) vs. Female ($X = 1$), we get that

$$HR(t) = \frac{\lambda_F(t)}{\lambda_M(t)} = e^{\beta_1}.$$

In this model, we get for Males: $\lambda_M(t) = \lambda_0(t)$ and for Females: $\lambda_F(t) = \lambda_0(t)e^{\beta_1}$.

Using a different coding for gender: Male ($X = 1$) vs. Female ($X = 0$), we get for Males: $\lambda_M(t) = \lambda_0(t)e^{\beta_2}$ and for Females: $\lambda_F(t) = \lambda_0(t)$.

Hence, we get the same results for $HR(t) = e^{\beta_1} = e^{-\beta_2}$ but selecting the baseline is important.

For a continuous variable $X$, we note that the hazard ratio of level $x+1$ versus level $x$ is given by

$$\frac{\lambda(t|X=x+1)}{\lambda(t|X=x)} = \frac{\lambda_0(t)e^{\beta(x+1)}}{\lambda_0(t)e^{\beta x}} = e^{\beta}.$$

Hence, in Cox's regression model

$$\lambda(t|X) = \lambda_0(t)e^{\beta X}$$

So $e^{\beta}$ describes the proportional change of the hazard due to the increase of $X$ by one unit.

We note that

$$\beta > 0 \Rightarrow \text{hazard increases.}$$

$$\beta < 0 \Rightarrow \text{hazard decreases.}$$

# Example: NSCLC

- Laudanski et al., Eur Respir J (2001).
- In this study, we had 102 patients who were operated from long cancer.
- The severity of the cancer was expressed in three TNM (Tumor, Nodes, Metastasis) categories: I, II, IIIa.
- The expression of the P53 protein was found from tumor biopsies.

Let's take $X = 0, 1, 2$ for, respectively, TNM = I, II, IIIa.

Hence, we get

- hazard function for TNM I: $\lambda(t|X = 0) = \lambda_0(t)$.
- hazard function for TNM II: $\lambda(t|X = 1) = \lambda_0(t)e^{\beta}$.
- hazard function for TNM IIIa: $\lambda(t|X = 2) = \lambda_0(t)e^{2\beta}$.

If $\beta = 1$, we get that $e^{\beta} = e^1 = 2.73$. This means that an increase of the TNM stage by one level increases the hazard 2.73 times.

We note that the hazard functions for TNM II and IIIa are specified relative to the hazard function of TNM I.

For another parametrization, $X = 2, 1, 0$ for TNM= I, II, IIIa.

We get

- hazard function for TNM IIIa: $\lambda(t|X = 0) = \lambda_0(t)$.
- hazard function for TNM II: $\lambda(t|X = 1) = \lambda_0(t)e^{\beta}$.
- hazard function for TNM I: $\lambda(t|X = 2) = \lambda_0(t)e^{2\beta}$.

The hazard functions for TNM I and II are taken relative to the hazard of TNM IIIa.

The reference level $(X = 0)$ has changed!

# Estimation of $\beta$

In the Cox's model,

$$\lambda(t|\mathbf{X}) = \lambda_0(t)e^{\beta^t \mathbf{X}},$$

we treat the unknown baseline hazard $\lambda_0(t)$ as a nuisance parameter.

Therefore we need an estimation method for the parameters $\beta$ without estimating $\lambda_0(t)$.

Hence, we will construct a partial likelihood.

Our data consists a sample of triplets $(T_i, \delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ where $\mathbf{X}_i$ is a vector which contains the values of $\mathbf{X}$ for individual $i$.

We assume

- Given $\mathbf{X}_i$, the lifetime and the censoring time are independent (non-informative censoring).
- Let $\tau_1 < \tau_2 < \ldots < \tau_D$ be the $D$ ordered distinct death times.
- We assume that there are **no tied** death times.

Let us define by

- $I_j$ the identity of the individual who failed at time $\tau_j$.
- $V_j$ the time of the $j$th failure $(\tau_j)$ and all information about censoring in $[\tau_{j-1}, \tau_j[$.

The observable data $(T_i, \delta_i, \mathbf{X}_i)$ is represented by $\{I_j\}$ and $\{V_j\}$. Hence

$$\begin{aligned} P(\text{Data}) &= P(\{I_1, V_1, \ldots, I_D, V_D\}) \\ &= P(\{I_1, V_1\}) \times P(\{I_2, V_2\}|\{I_1, V_1\}) \times \ldots \\ &\quad \times P(\{I_D, V_D\}|\{I_1, V_1, \ldots, I_{D-1}, V_{D-1}\}) \\ &= \prod_{j=1}^{D} P(I_j|\{I_1, V_1, \ldots, I_{j-1}, V_{j-1}, V_j\}) \\ &\quad \times P(V_j|\{I_1, V_1, \ldots, I_{j-1}, V_{j-1}\}) \end{aligned}$$

Due to the non-informative censoring, the second term does not add much information about the parameters $\beta$.

Hence, we define the partial likelihood as

$$\begin{aligned} L^{\text{partial}}(\beta) &= \prod_{j=1}^{D} P(I_j | \{I_1, V_1, \ldots, I_{j-1}, V_{j-1}, V_j\}) \\ &= \prod_{j=1}^{D} P(I_j | H_j) \end{aligned}$$

where $H_j$ is the "history" of the data, up to $j$th failure and including the failure time, but not the identity of the failing.

At each failure, we note that the quantity $P(I_j|H_j)$ is the conditional probability that a specific individual fails at time $\tau_j$, given all the individuals that had not fail before $\tau_j$.

We denote by $\mathcal{R}(t)$ the set of all the individuals under study just prior to time $t$.

$$
\begin{aligned}
P(I_j|H_j) &= P(\text{individual } I_j \text{ fails}|\text{one individual fails in } \mathcal{R}(\tau_j)) \\
&= \frac{P(\text{individual } j \text{ fails}|\text{at risk at } \tau_j)}{\sum_{l\in\mathcal{R}(\tau_j)} P(\text{individual } l \text{ fails}|\text{at risk at } \tau_j)} \\
&= \frac{\lambda(\tau_j|\mathbf{X}_j)d\tau_j}{\sum_{l\in\mathcal{R}(\tau_j)} \lambda(\tau_j|\mathbf{X}_l)d\tau_j} = \frac{\lambda_0(\tau_j)e^{\beta^t\mathbf{X}_j}}{\sum_{l\in\mathcal{R}(\tau_j)} \lambda_0(\tau_j)e^{\beta^t\mathbf{X}_l}} \\
&= \frac{e^{\beta^t\mathbf{X}_j}}{\sum_{l\in\mathcal{R}(\tau_j)} e^{\beta^t\mathbf{X}_l}}
\end{aligned}
$$

We get as partial likelihood,

$$L^{\text{partial}}(\beta) = \prod_{j=1}^{D} \frac{e^{\beta^t \mathbf{X}_j}}{\sum_{l \in \mathcal{R}(\tau_j)} e^{\beta^t \mathbf{X}_l}}.$$

**Some remarks:**

- Contributions only at uncensored times.

$$L^{\text{partial}}(\beta) = \prod_{j=1}^{n} \left[ \frac{e^{\beta^t \mathbf{X}_j}}{\sum_{l \in \mathcal{R}(\tau_j)} e^{\beta^t \mathbf{X}_l}} \right]^{\delta_j}.$$

- The partial likelihood is **not** a product of independent terms, but of conditional probabilities. Hence, Partial likelihood $\neq$ Usual likelihood.

Although not a usual likelihood, the regular likelihood properties are still valid!

So we have that,

- Log-partial likelihood

$$l(\beta) = \sum_{j=1}^{D} \left[ \beta^t \mathbf{X}_j - \log \left[ \sum_{l \in \mathcal{R}(\tau_j)} e^{\beta^t \mathbf{X}_l} \right] \right].$$

- Partial likelihood score equation

$$U_i(\beta) = \frac{\partial}{\partial \beta_i} l(\beta) = \sum_{j=1}^{D} \left[ \mathbf{X}_{ji} - \frac{\sum_{l \in \mathcal{R}(\tau_j)} \mathbf{X}_{li} e^{\beta^t \mathbf{X}_l}}{\sum_{l \in \mathcal{R}(\tau_j)} e^{\beta^t \mathbf{X}_l}} \right].$$

- Information matrix

$$I_{ik} = -\frac{\partial^2}{\partial\beta_i\partial\beta_k} l(\beta) =$$

$$\sum_{j=1}^{D} \left[ \frac{\sum\limits_{l\in\mathcal{R}(\tau_j)} \mathbf{X}_{li}\mathbf{X}_{lk}e^{\beta^t\mathbf{X}_l}}{\sum\limits_{l\in\mathcal{R}(\tau_j)} e^{\beta^t\mathbf{X}_l}} - \frac{\sum\limits_{l\in\mathcal{R}(\tau_j)} \mathbf{X}_{li}e^{\beta^t\mathbf{X}_l} \sum\limits_{l\in\mathcal{R}(\tau_j)} \mathbf{X}_{lk}e^{\beta^t\mathbf{X}_l}}{\left(\sum\limits_{l\in\mathcal{R}(\tau_j)} e^{\beta^t\mathbf{X}_l}\right)^2} \right]$$

Furthermore, we note that

$$\hat{\beta} - \beta \sim N(0, I^{-1}).$$

Based on these quantities, we derive three main test for the global hypothesis $H_0 : \beta = \beta_0$. Under $H_0$, we get

- Wald-test

$$X^2_W = (b - \beta_0)^t I(b)(b - \beta_0) \sim \chi^2_p$$

- Score test

$$X^2_S = (U_1(\beta_0), \ldots, U_p(\beta_0))^t I^{-1}(\beta_0)(U_1(\beta_0), \ldots, U_p(\beta_0)) \sim \chi^2_p$$

- (Partial) Likelihood ratio

$$X^2_{LR} = 2l(b) - 2l(\beta_0) \sim \chi^2_p$$

# Using SAS

```
data nsclc;
input Survtime Survind tnm expres;
cards;
24.51000023      1      2      1
27.12999916      1      2      1
...
;
run;

proc phreg data=nsclc;
model Survtime*Survind(0)=expres;
run;
```

The PHREG Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.NSCLC |
| Dependent Variable | Survtime |
| Censoring Variable | Survind |
| Censoring Value(s) | 0 |
| Ties Handling | BRESLOW |

| | |
|---|---|
| Number of Observations Read | 102 |
| Number of Observations Used | 102 |

Summary of the Number of Event and Censored Values

| Total | Event | Censored | Percent Censored |
|---|---|---|---|
| 102 | 49 | 53 | 51.96 |

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

| Criterion | Without Covariates | With Covariates |
|---|---|---|
| -2 LOG L | 400.456 | 393.305 |
| AIC | 400.456 | 395.305 |
| SBC | 400.456 | 397.197 |

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 7.1509 | 1 | 0.0075 |
| Score | 7.1051 | 1 | 0.0077 |
| Wald | 6.7640 | 1 | 0.0093 |

Analysis of Maximum Likelihood Estimates

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| expres | 1 | 0.78590 | 0.30218 | 6.7640 | 0.0093 | 2.194 |

**Conclusion:** expression of P53 increases mortality hazard approximately 2 times as compared to no expression.

# Using R

```
> fit<-coxph(Surv(survtime,survind)~expres,data=NSCLC)
> summary(fit)
Call: coxph(formula = Surv(survtime, survind) ~ expres, data = NSCLC)

n= 102
        coef exp(coef) se(coef)   z      p
expres 0.786      2.19    0.302 2.6 0.0093

       exp(coef) exp(-coef) lower .95 upper .95
expres      2.19      0.456      1.21      3.97

Rsquare= 0.068   (max possible= 0.98 )
Likelihood ratio test= 7.15  on 1 df,   p=0.0075
Wald test            = 6.76  on 1 df,   p=0.0093
Score (logrank) test = 7.11  on 1 df,   p=0.00769
```

Using the normality of $\beta$, we have a 95% C.I for $\beta$:

$$[b - 1.96s.e., b + 1.96s.e.].$$

Hence, the 95% C.I. for the hazard ratio is:

$$[e^{b-1.96s.e.}, e^{b+1.96s.e.}].$$

# Adjustments for ties

Sofar, we assumed no tied death times.

In case of ties, we have to adjust the partial likelihood.
Therefore we define

- $\tau_1 < \tau_2 < \ldots < \tau_D$ as the $D$ ordered distinct death times.
- $d_j$ the number of failures at $\tau_j$.
- $I_{j_1}, \ldots, I_{j_d}$ the identities of the individuals who failed at time $\tau_j$.
- $H_j$ is the "history" of the data, up to $j$th failure and including the failure time, but not the identities of the failing.

The partial likelihood is now defined as

$$L^{\text{partial}}(\beta) = \prod_{j=1}^{D} P(I_{j_1}, \ldots, I_{j_d} | H_j).$$

If the lifetimes are continuous, we find the exact likelihood,

$$L_1 = \prod_{j=1}^{D} \left\{ \int_0^{\infty} \prod_{k \in D_j} \left[ 1 - \exp\left( \frac{e^{\beta^t X_k}}{\sum_{l \in \mathcal{R}^*(\tau_j)} e^{\beta^t X_l}} t \right) \right] \exp(-t) dt \right\}$$

where $\mathcal{R}^*(\tau_j)$ denote the set of individuals whose event or censored times exceed $\tau_j$ or whose censored times are equal to $\tau_j$ and $D_j$ is the set of individuals that fail at $\tau_j$.

If the number of ties increases, the denominator is impossible to calculate. Therefore several authors considered approximations

- Breslow

$$L_2(\beta) = \prod_{j=1}^{D} \frac{e^{\beta^t \sum_{k \in D_j} \mathbf{X}_k}}{\left( \sum_{l \in \mathcal{R}(\tau_j)} e^{\beta^t \mathbf{X}_l} \right)^{d_j}}.$$

- Efron

$$L_3(\beta) = \prod_{j=1}^{D} \frac{e^{\beta^t \sum_{k \in D_j} \mathbf{X}_k}}{\prod_{k=1}^{d_j} \left( \sum_{l \in \mathcal{R}(\tau_j)} e^{\beta^t \mathbf{X}_l} - \frac{k-1}{d_j} \sum_{l \in D_j} e^{\beta^t \mathbf{X}_l} \right)}.$$

If the lifetimes are discrete, we get the discrete partial likelihood,

$$L_4(\beta) = \prod_{j=1}^{D} \frac{e^{\beta^t \sum_{k \in D_j} \mathbf{X}_k}}{\sum_{q \in \mathcal{Q}(\tau_j)} e^{\beta^t \sum_{l=1}^{d_j} \mathbf{X}_{ql}}}$$

where $\mathcal{Q}(\tau_j)$ denote the set of all subsets of $d_j$ individuals selected from the risk set $\mathcal{R}(\tau_j)$.

We note that each of these partial likelihoods reduces to the original partial likelihood when there are no ties.

# Example: Performance testing

- Test-subject were asked to perform a certain test and the time needed was recorded.
- 3 different noise distractions are applied.
- Did the decreasing noise distractions influence the hazard of the time to finish the test?

| Noise level | | |
|---|---|---|
| 9.0 | 10.0 | 12.0 |
| 9.5 | 12.0 | $12.0^+$ |
| 9.0 | $12.0^+$ | $12.0^+$ |
| 8.5 | 11.0 | $12.0^+$ |
| 10.0 | 12.0 | $12.0^+$ |
| 10.5 | 10.5 | $12.0^+$ |

# In SAS, the Breslow method is the default.

```
data noise;
input Time censor level;
cards;
9.0 1 1
9.5 1 1
9.0 1 1
...
12.0 0 3
;
run;

proc phreg data=noise;
model Time*censor(0)=level;
run;

proc phreg data=noise;
model Time*censor(0)=level/ties=discrete;
run;

proc phreg data=noise;
model Time*censor(0)=level/ties=efron;
run;

proc phreg data=noise;
model Time*censor(0)=level/ties=exact;
run;
```

Ties Handling BRESLOW

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| level | 1 | -2.10429 | 0.62384 | 11.3781 | 0.0007 | 0.122 |

Ties Handling DISCRETE

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| level | 1 | -2.66767 | 0.80298 | 11.0370 | 0.0009 | 0.069 |

Ties Handling EFRON

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| level | 1 | -2.24966 | 0.63592 | 12.5149 | 0.0004 | 0.105 |

Ties Handling EXACT

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| level | 1 | -2.45815 | 0.73994 | 11.0364 | 0.0009 | 0.086 |

# In R, the Efron method is the default.

```
> summary(coxph(Surv(Time,Censor)~Level))
Call: coxph(formula = Surv(Time, Censor) ~ Level)

n= 18
       coef exp(coef) se(coef)     z     p
Level -2.25     0.105    0.636 -3.54 4e-04

      exp(coef) exp(-coef) lower .95 upper .95
Level     0.105       9.48    0.0303     0.367


> summary(coxph(Surv(Time,Censor)~Level,method="breslow"))
Call: coxph(formula = Surv(Time, Censor) ~ Level, method = "breslow")

n= 18
       coef exp(coef) se(coef)     z       p
Level -2.10     0.122    0.624 -3.37 0.00074

      exp(coef) exp(-coef) lower .95 upper .95
Level     0.122        8.2    0.0359     0.414


> summary(coxph(Surv(Time,Censor)~Level,method="exact"))
Call: coxph(formula = Surv(Time, Censor) ~ Level, method = "exact")

n= 18
       coef exp(coef) se(coef)     z       p
Level -2.67    0.0694    0.803 -3.32 0.00089

      exp(coef) exp(-coef) lower .95 upper .95
Level    0.0694       14.4    0.0144     0.335
```

# Example: NSCLC

- Laudanski et al., Eur Respir J (2001).
- In this study, we had 102 patients who were operated from long cancer.
- The severity of the cancer was expressed in three TNM (Tumor, Nodes, Metastasis) categories: I, II, IIIa.
- The expression of the P53 protein was found from tumor biopsies.

Sofar, expression of P53 has a positive influence on the hazard, however we did not take the confounder TNM into account.

# Model 1: $\lambda(t|\mathbf{X}) = \lambda_0(t)e^{\beta_1 \times \mathbf{expres} + \beta_2 \times \mathbf{TNM}}$

```
> fit<-coxph(Surv(survtime,survind)~expres+tnm,data=NSCLC)
> summary(fit)
Call: coxph(formula = Surv(survtime, survind) ~ expres + tnm, data = NSCLC)

n= 102, number of events= 49

         coef exp(coef) se(coef)     z Pr(>|z|)
expres 0.7345    2.0845   0.3032 2.423   0.0154 *
tnm    1.0725    2.9227   0.2546 4.213 2.52e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

       exp(coef) exp(-coef) lower .95 upper .95
expres     2.085     0.4797     1.151     3.776
tnm        2.923     0.3421     1.775     4.814

Concordance= 0.7  (se = 0.044 )
Rsquare= 0.259   (max possible= 0.98 )
Likelihood ratio test= 30.58  on 2 df,   p=2.286e-07
Wald test            = 23.75  on 2 df,   p=6.947e-06
Score (logrank) test = 26.95  on 2 df,   p=1.405e-06
```

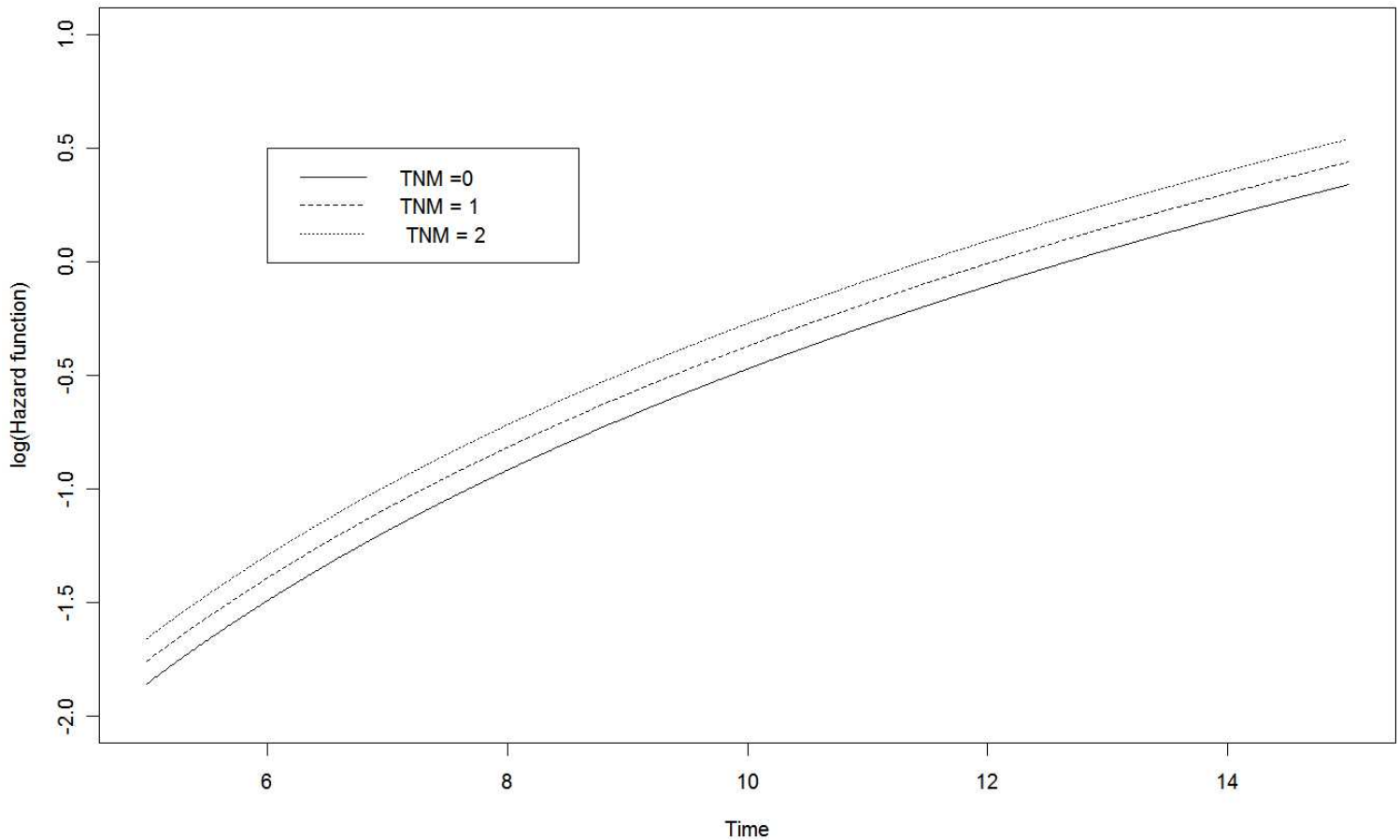
```
proc phreg data=nsclc;
model Survtime*Survind(0)=expres TNM;
run;
```

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| expres | 1 | 0.73452 | 0.30316 | 5.8702 | 0.0154 | 2.084 |
| tnm | 1 | 1.07226 | 0.25455 | 17.7438 | <.0001 | 2.922 |

**Conclusion:** Influence of expression P53 changes slightly, TNM has a large positive influence.

However, for different tumor-categories, the change in expression of P53 remains the same!

$$\frac{\lambda(t|\text{expres}=1,\text{TNM}=1)}{\lambda(t|\text{expres}=0,\text{TNM}=1)} = \frac{\lambda(t|\text{expres}=1,\text{TNM}=2)}{\lambda(t|\text{expres}=0,\text{TNM}=2)} = e^{\beta_1}.$$


TNM =0
TNM = 1
TNM = 2
log(Hazard function)
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
6
8
10
12
14
Time

# Model 2: $\lambda(t|\mathbf{X}) = \lambda_0(t)e^{\beta_1\mathbf{expres}+\beta_2\mathbf{TNM}+\beta_3\mathbf{expres}*\mathbf{TNM}}$

```
> fit<-coxph(Surv(survtime,survind)~expres*tnm,data=NSCLC)
> summary(fit)
Call: coxph(formula = Surv(survtime, survind) ~ expres * tnm, data = NSCLC)

n= 102, number of events= 49

              coef exp(coef) se(coef)      z Pr(>|z|)
expres      3.4814   32.5052   1.7444  1.996  0.04595 *
tnm         1.7948    6.0180   0.5612  3.198  0.00138 **
expres:tnm -1.0186    0.3611   0.6198 -1.643  0.10029
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

           exp(coef) exp(-coef) lower .95 upper .95
expres       32.5052    0.03076    1.0645   992.541
tnm           6.0180    0.16617    2.0034    18.078
expres:tnm    0.3611    2.76925    0.1072     1.217

Concordance= 0.709  (se = 0.044 )
Rsquare= 0.283   (max possible= 0.98 )
Likelihood ratio test= 33.87  on 3 df,   p=2.11e-07
Wald test            = 18.71  on 3 df,   p=0.0003131
Score (logrank) test = 27.02  on 3 df,   p=5.827e-06
```

```
data nsclc1;
set nsclc;
exTNM=expres*TNM;
run;

proc phreg data=nsclc1;
model Survtime*Survind(0)=expres TNM exTNM;
run;
```

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| expres | 1 | 3.48139 | 1.74435 | 3.9833 | 0.0460 | 32.505 |
| tnm | 1 | 1.79476 | 0.56120 | 10.2276 | 0.0014 | 6.018 |
| exTNM | 1 | -1.01857 | 0.61978 | 2.7009 | 0.1003 | 0.361 |

**Conclusion:** Influence of expression P53 changes drastically by interaction with TNM. However, interaction is not significant.

# Model 3: $\lambda(t|\mathbf{X}) = \lambda_0(t)e^{\beta_1\mathbf{expres}+\beta_2(\mathbf{TNM}=1)+\beta_3(\mathbf{TNM}=2)}$

```
> fit<-coxph(Surv(survtime,survind)~expres+(tnm==1)+(tnm==2),data=NSCLC)
> summary(fit)
Call: coxph(formula = Surv(survtime, survind) ~ expres + (tnm == 1) + (tnm == 2), data = NSCLC)

n= 102, number of events= 49

                coef exp(coef) se(coef)      z Pr(>|z|)
expres        0.7594    2.1369   0.3050  2.490 0.012790 *
tnm == 1TRUE -1.7225    0.1786   0.4995 -3.449 0.000563 ***
tnm == 2TRUE -1.5788    0.2062   0.4251 -3.714 0.000204 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

             exp(coef) exp(-coef) lower .95 upper .95
expres          2.1369      0.468   1.17532    3.8852
tnm == 1TRUE    0.1786      5.598   0.06711    0.4754
tnm == 2TRUE    0.2062      4.849   0.08963    0.4744

Concordance= 0.709  (se = 0.044 )
Rsquare= 0.279   (max possible= 0.98 )
Likelihood ratio test= 33.31  on 3 df,   p=2.768e-07
Wald test            = 27.71  on 3 df,   p=4.173e-06
Score (logrank) test = 32.63  on 3 df,   p=3.849e-07
```

```
> fit<-coxph(Surv(survtime,survind)~expres,data=NSCLC)
> summary(fit)
Call: coxph(formula = Surv(survtime, survind) ~ expres, data = NSCLC)

n= 102, number of events= 49

          coef exp(coef) se(coef)     z Pr(>|z|)
expres 0.7859    2.1945   0.3022 2.601   0.0093 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

       exp(coef) exp(-coef) lower .95 upper .95
expres     2.194     0.4557     1.214     3.968

Concordance= 0.599  (se = 0.039 )
Rsquare= 0.068   (max possible= 0.98 )
Likelihood ratio test= 7.15  on 1 df,   p=0.007493
Wald test            = 6.76  on 1 df,   p=0.009298
Score (logrank) test = 7.11  on 1 df,   p=0.007687

> fit1<-coxph(Surv(survtime,survind)~expres+(tnm==1)+(tnm==2),data=NSCLC)

> anova(fit1,fit)
Analysis of Deviance Table
Cox model: response is  Surv(survtime, survind)
Model 1: ~ expres + (tnm == 1) + (tnm == 2)
Model 2: ~ expres
   loglik  Chisq Df P(>|Chi|)
1 -183.57
2 -196.65 26.161  2 2.086e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
data nsclc1;
set nsclc;
TNM1=0;
TNM2=0;
If TNM=1 then TNM1=1;
If TNM=2 then TNM2=1;
run;

proc phreg data=nsclc1;
model Survtime*Survind(0)=expres TNM1 TNM2;
TNM: test TNM1=0,TNM2=0;
TNMC: test TNM2=2*TNM1;
run;
```

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| expres | 1 | 0.75936 | 0.30502 | 6.1980 | 0.0128 | 2.137 |
| TNM1 | 1 | -1.72246 | 0.49947 | 11.8927 | 0.0006 | 0.179 |
| TNM2 | 1 | -1.57884 | 0.42512 | 13.7931 | 0.0002 | 0.206 |

```
Linear Hypotheses Testing Results

Wald
Label    Chi-Square      DF    Pr > ChiSq

TNM         21.4783       2        <.0001
TNMC         3.4425       1        0.0635
```

**Conclusion:** Considering TNM as a class variable, we still see its significant influence. A local test shows that we can take it as a continuous variable.

Still, we assume that TNM has a proportional influence on the hazard function.

- For some covariates, the proportional hazard assumption can be fulfilled, but not for others.
- In such a situation, one can use the Cox model, using modified baseline hazards.

Assume that the hazard is proportional for a covariate $Y$, but for $X$.

We take a stratified Cox's model,

$$\lambda(t|X,Y) = \lambda_{X=x,0}(t)e^{\beta Y}.$$

Note that the baseline hazard functions are different for strata defined by the levels of $X$, but the effect of $Y$ is still expressed as the proportional change of the hazard function (for a fixed level of $X$).

# Model 4: $\lambda(t|\mathbf{X}) = \lambda_{\mathbf{TNM},0}(t)e^{\beta \times \mathbf{expres}}$

```
> fit<-coxph(Surv(survtime,survind)~expres+strata(tnm),data=NSCLC)
> summary(fit)
Call: coxph(formula = Surv(survtime, survind) ~ expres + strata(tnm), data = NSCLC)

n= 102, number of events= 49

         coef exp(coef) se(coef)     z Pr(>|z|)
expres 0.7404    2.0968   0.3076 2.407   0.0161 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

       exp(coef) exp(-coef) lower .95 upper .95
expres     2.097     0.4769     1.147     3.832

Concordance= 0.597  (se = 0.058 )
Rsquare= 0.059   (max possible= 0.949 )
Likelihood ratio test= 6.16  on 1 df,   p=0.0131
Wald test            = 5.79  on 1 df,   p=0.01609
Score (logrank) test = 6.05  on 1 df,   p=0.01387
```

```
proc phreg data=nsclc1;
model Survtime*Survind(0)=expres;
strata TNM;
run;
```

Summary of the Number of Event and Censored Values

| Stratum | tnm | Total | Event | Censored | Percent Censored |
|---|---|---|---|---|---|
| 1 | 1 | 21 | 5 | 16 | 76.19 |
| 2 | 2 | 27 | 7 | 20 | 74.07 |
| 3 | 3 | 54 | 37 | 17 | 31.48 |
| Total | | 102 | 49 | 53 | 51.96 |

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| expres | 1 | 0.74029 | 0.30763 | 5.7910 | 0.0161 | 2.097 |

Next to the estimation of the parameters $\beta$, we want to estimate the survival function at specific covariate values.

Hereto we consider two methods:

- Kalbfleisch-Prentice method (no ties):

$$\hat{S}(t|\mathbf{X}) = S_0(t)^{\exp(\hat{\beta}^t\mathbf{X})},\ S_0(t) = \prod_{t_i \leq t}\left[1 - \frac{\exp(\hat{\beta}^t\mathbf{X}_i)}{\sum_{l\in\mathcal{R}(t_i)}\exp(\hat{\beta}^t\mathbf{X}_l)}\right]$$

- Breslow method:

$$\hat{S}(t|\mathbf{X}) = S_0(t)^{\exp(\hat{\beta}^t\mathbf{X})},\ S_0(t) = \prod_{t_i \leq t}\exp\left[-\frac{d_i}{\sum_{l\in\mathcal{R}(t_i)}\exp(\hat{\beta}^t\mathbf{X}_l)}\right]$$

```
data teken;
input x;
datalines;
0.00
1.00
;

proc phreg data=test;
model time*status(0)=x/ties=efron;
baseline covariates=teken out=predd survival=_all_/method=pl;
run;

proc print data=predd;
run;

proc phreg data=test;
model time*status(0)=x/ties=efron;
baseline covariates=teken out=predd1 survival=_all_;
run;

proc print data=predd1;
run;
```

| Obs | x | time | Survival | StdErr Survival | Lower Survival | Upper Survival |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1.00000 | . | . | . |
| 2 | 0 | 1 | 0.89962 | 0.09941 | 0.72444 | 1.00000 |
| 3 | 0 | 2 | 0.63352 | 0.18915 | 0.35287 | 1.00000 |
| 4 | 0 | 3 | 0.45826 | 0.21153 | 0.18544 | 1.00000 |
| 5 | 0 | 4 | 0.00000 | 0.00000 | . | . |
| 6 | 1 | 0 | 1.00000 | . | . | . |
| 7 | 1 | 1 | 0.84562 | 0.13059 | 0.62478 | 1.00000 |
| 8 | 1 | 2 | 0.48502 | 0.17082 | 0.24320 | 0.96726 |
| 9 | 1 | 3 | 0.29027 | 0.14425 | 0.10960 | 0.76881 |
| 10 | 1 | 4 | 0.00000 | 0.00000 | . | . |

| Obs | x | time | Survival | StdErr Survival | Lower Survival | Upper Survival |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1.00000 | . | . | . |
| 2 | 0 | 1 | 0.90720 | 0.10024 | 0.73055 | 1 |
| 3 | 0 | 2 | 0.68444 | 0.20435 | 0.38123 | 1 |
| 4 | 0 | 3 | 0.54840 | 0.25313 | 0.22192 | 1 |
| 5 | 0 | 4 | 0.20175 | 0.22220 | 0.02330 | 1 |
| 6 | 1 | 0 | 1.00000 | . | . | . |
| 7 | 1 | 1 | 0.85695 | 0.13234 | 0.63315 | 1 |
| 8 | 1 | 2 | 0.54825 | 0.19309 | 0.27491 | 1 |
| 9 | 1 | 3 | 0.38585 | 0.19175 | 0.14568 | 1 |
| 10 | 1 | 4 | 0.07907 | 0.14905 | 0.00196 | 1 |

```
> test1 <- list(time=c(4,3,1,1,2,2,3),
+               status=c(1,1,1,0,1,1,0),
+               x=c(0,2,1,1,1,0,0))
```

```
> summary(survfit(coxph(Surv(time, status) ~ x, test1),type="kalbfleisch-prentice",
newdata=data.frame(x=c(0,1))))
Call: survfit(formula = coxph(Surv(time, status) ~ x, test1), newdata = data.frame(x = c(0,1)),
 type = "kalbfleisch-prentice")

time n.risk n.event survival1 survival2
1      7       1     0.900     0.846
2      5       2     0.634     0.485
3      3       1     0.458     0.290
4      1       1     0.000     0.000
```

```
> summary(survfit(coxph(Surv(time, status) ~ x, test1),type="breslow",newdata=data.frame(x=c(0,1))))
Call: survfit(formula = coxph(Surv(time, status) ~ x, test1), newdata = data.frame(x = c(0,1)),
 type = "breslow")

time n.risk n.event survival1 survival2
1      7       1     0.907    0.8569
2      5       2     0.684    0.5482
3      3       1     0.548    0.3859
4      1       1     0.202    0.0791
```

# Time-dependent covariates

Suppose:

- After a kidney transplantation, we look at the time until the host body rejects the organ.
- Every month, a patient has a check-up and blood pressure, white cell count,... are recorded.
- we are interested how these time-dependent covariates influence the hazard of rejection time.

Hence, we have $(T_i, \delta_i, \{\mathbf{X}_i(t), 0 \leq t \leq T_i\})$, $i = 1, \ldots, n$.

Note that fixed-time covariates are a special case, namely,

$$\mathbf{X}(t) = \mathbf{X}(0), \quad \forall t > 0.$$

Extending the Cox's model to accommodate for time-dependent covariates, we assume that

$$\lambda(t|\mathbf{X}(t)) = \lambda_0(t)e^{\beta^t \mathbf{X}(t)}.$$

To estimate the parameters $\beta$, we extend the partial likelihood.

We assume

- The value of $\mathbf{X}_i(t)$ is known for any time at which the subject is at risk.
- Given $\mathbf{X}_i(t)$, the lifetime and the censoring time are independent (non-informative censoring).
- Let $\tau_1 < \tau_2 < \ldots < \tau_D$ be the $D$ ordered distinct death times.
- We assume that there are **no tied** death times.

We get as partial likelihood,

$$L^{\text{partial}}(\beta) = \prod_{j=1}^{D} \frac{e^{\beta^t \mathbf{X}_j(\tau_j)}}{\sum_{l \in \mathcal{R}(\tau_j)} e^{\beta^t \mathbf{X}_l(\tau_j)}}.$$

**Some remarks:**

- For each individual, we only need the values of $\mathbf{X}(t)$ at the uncensored death times.
- When there are ties, we extend as before one of the previous partial likelihoods.

# Example: Cancer in rodents

- Forty-five rodents were randomly assigned to three dose groups of a tumor-promoting agent.
- The rodents were examined every week for the number of papillomas (weeks 27, 34, 37, 41, 43, 45, 46, 47, 49, 50, 51, 53, 65, 67 and 71).
- Researchers are interested in the time until death of cancer and how this was influenced by dose after adjusting for the number of papillomas.

```
> rodent<-read.table("C:/werk/Roel/Onderwijs/Theorie/G0B67AStatAnalReliaSurvData/
Cursus/Rodent.txt",header=T,sep=";")
> rodent[1:5,]
   Id Time Dead Dose P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15
 1  1   47    1    1  0  5  6  8 10 10 10 10 NA  NA  NA  NA  NA  NA  NA
 2  2   71    1    1  0  0  0  0  0  0  0  0  1   1   1   1   1   1   1
 3  3   81    0    1  0  1  1  1  1  1  1  1  1   1   1   1   1   1   1
 4  4   81    0    1  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0
 5  5   81    0    1  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0

> n<-dim(rodent)[1]
> Date0<-rep(0,n)
> P0<-rep(0,n)
> Date1<-rep(27,n)
> Date2<-rep(34,n)
> Date3<-rep(37,n)
> Date4<-rep(41,n)
> Date5<-rep(43,n)
> Date6<-rep(45,n)
> Date7<-rep(46,n)
> Date8<-rep(47,n)
> Date9<-rep(49,n)
> Date10<-rep(50,n)
> Date11<-rep(51,n)
> Date12<-rep(53,n)
> Date13<-rep(65,n)
> Date14<-rep(67,n)
> Date15<-rep(71,n)
> rodent1<-data.frame(rodent,P0,Date0,Date1,Date2,Date3,Date4,Date5,Date6,Date7,Date8,Date9,Date10,
Date11,Date12,Date13,Date14,Date15)
> rodentlong<-tmerge(rodent1[,1:4],rodent1,id=Id,status=event(Time,Dead),pap=tdc(Date0,P0),
+ pap=tdc(Date1,P1),pap=tdc(Date2,P2),pap=tdc(Date3,P3),pap=tdc(Date4,P4),pap=tdc(Date5,P5),
+ pap=tdc(Date6,P6),pap=tdc(Date7,P7),pap=tdc(Date8,P8),pap=tdc(Date9,P9),pap=tdc(Date10,P10),
+ pap=tdc(Date11,P11),pap=tdc(Date12,P12),pap=tdc(Date13,P13),pap=tdc(Date14,P14),pap=tdc(Date15,P15))
```

```
> rodentlong[1:20,4:9]
   Dose id tstart tstop status pap
1     1  1      0    27      0   0
2     1  1     27    34      0   0
3     1  1     34    37      0   5
4     1  1     37    41      0   6
5     1  1     41    43      0   8
6     1  1     43    45      0  10
7     1  1     45    46      0  10
8     1  1     46    47      1  10
9     1  2      0    27      0   0
10    1  2     27    34      0   0
11    1  2     34    37      0   0
12    1  2     37    41      0   0
13    1  2     41    43      0   0
14    1  2     43    45      0   0
15    1  2     45    46      0   0
16    1  2     46    47      0   0
17    1  2     47    49      0   0
18    1  2     49    50      0   1
19    1  2     50    51      0   1
20    1  2     51    53      0   1
```

```
> fit<-coxph(Surv(tstart,tstop,status)~Dose+pap,data=rodentlong)
> summary(fit)
Call: coxph(formula = Surv(tstart, tstop, status) ~ Dose + pap, data = rodentlong)

n= 428, number of events= 25

        coef exp(coef) se(coef)     z Pr(>|z|)
Dose 0.09132   1.09562  0.05582 1.636  0.10183
pap  0.10303   1.10852  0.03163 3.257  0.00113 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

     exp(coef) exp(-coef) lower .95 upper .95
Dose     1.096     0.9127    0.9821     1.222
pap      1.109     0.9021    1.0419     1.179

Concordance= 0.809  (se = 0.064 )
Rsquare= 0.045   (max possible= 0.321 )
Likelihood ratio test= 19.75  on 2 df,   p=5.139e-05
Wald test            = 18.27  on 2 df,   p=0.0001078
Score (logrank) test = 23.57  on 2 df,   p=7.616e-06
```

```
option nocenter;
data rodent;
infile datalines missover;
input ID Time Dead Dose P1-P15;
label ID='Subject ID'; datalines;
1 47 1  1.0  0  5  6  8 10 10 10 10
2 71 1  1.0  0  0  0  0  0  0  0  0  1  1  1  1 1 1 1
3 81 0  1.0  0  1  1  1  1  1  1  1  1  1  1  1 1 1 1
...
44 37 1 10.0  0  1  1
45 43 1 10.0  9 19 19 19 19
;

proc phreg data=rodent;
model Time*Dead(0)=Dose NPap;
array pp{*} P1-P14;
array tt{*} t1-t15;
t1 = 27;
t2 = 34;
t3 = 37;
t4 = 41;
t5 = 43;
t6 = 45;
t7 = 46;
t8 = 47;
t9 = 49;
t10= 50;
t11= 51;
t12= 53;
t13= 65;
t14= 67;
t15= 71;
if Time <  tt[1]  then NPap=0;
else if time >= tt[15] then NPap=P15;
else do i=1 to dim(pp);
if tt[i] <= Time < tt[i+1] then NPap= pp[i];
end;
run;
```

The PHREG Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.RODENT |
| Dependent Variable | Time |
| Censoring Variable | Dead |
| Censoring Value(s) | 0 |
| Ties Handling | BRESLOW |

| | |
|---|---|
| Number of Observations Read | 45 |
| Number of Observations Used | 45 |

Summary of the Number of Event and Censored Values

| Total | Event | Censored | Percent Censored |
|---|---|---|---|
| 45 | 25 | 20 | 44.44 |

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

| Criterion | Without Covariates | With Covariates |
|---|---|---|
| -2 LOG L | 166.793 | 143.269 |
| AIC | 166.793 | 147.269 |
| SBC | 166.793 | 149.707 |

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 23.5243 | 2 | <.0001 |
| Score | 28.0498 | 2 | <.0001 |
| Wald | 21.1646 | 2 | <.0001 |

Analysis of Maximum Likelihood Estimates

| Parameter Variable | DF | Standard Estimate | Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| Dose | 1 | 0.06885 | 0.05620 | 1.5010 | 0.2205 | 1.071 |
| NPap | 1 | 0.11714 | 0.02998 | 15.2705 | <.0001 | 1.124 |

# Model building

In the previous examples, we looked at how maximal two covariates influenced the survival time in the Cox's regression model.

However in studies with several covariates, we need a procedure to identify interesting covariates.

In nested models, we used the likelihood ratio test to check whether a covariate was significant or not. Here, we extend this idea and consider the $AIC$-criterium,

$$AIC = -2 \log \hat{L} + \alpha q,$$

in which $q$ is the number of unknown $\beta$ parameters.

In practice, model building is a combination of

- knowledge of the science
- trial and error, common sense
- automatic variable selection
  - stepwise: forward, backward, both
  - best subsets

A general strategy that is commonly used is,

1. Perform univariate analysis to "screen" potentially significant variables.
2. Fit a multiple model and discard variables that are non-significant.
3. Check whether variables that were dropped before, interactions or higher order terms should be added.

# Example: Leukemia

- Bone marrow transplants are a standard treatment for acute leukemia.
- The researchers investigated the time until relapse of leukemia.
- Several covariates: Disease Group, Patient Age (Years), Donor Age (Years), Patient Sex, Donor Sex, Patient CMV Status, Donor CMV Status, Waiting Time to Transplant In Days, FAB grade, Hospital, MTX Use.

```
proc phreg data=bmt1;
model T2*delta2(0)=g1 g2 Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 Z91 z92 z93 Z10/selection=forward details sle=0.1
include=2;
run;

proc phreg data=bmt1;
model T2*delta2(0)=g1 g2 Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 Z91 z92 z93 Z10/selection=backwards details sls=0.1
include=2;
run;

proc phreg data=bmt1;
model T2*delta2(0)=g1 g2 Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 Z91 z92 z93 Z10/selection=stepwise details sle=0.1
sls=0.07 include=2;
run;

proc phreg data=bmt1;
model T2*delta2(0)=g1 g2 Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 Z91 z92 z93 Z10/selection=score best=4 include=2;
run;
```

The PHREG Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.BMT1 |
| Dependent Variable | T2 |
| Censoring Variable | delta2 |
| Censoring Value(s) | 0 |
| Ties Handling | BRESLOW |

| | |
|---|---|
| Number of Observations Read | 137 |
| Number of Observations Used | 137 |

Summary of the Number of Event and Censored Values

| Total | Event | Censored | Percent Censored |
|---|---|---|---|
| 137 | 42 | 95 | 69.34 |

The following variable(s) will be included in each model:

g1 g2

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Analysis of Variables Not in the Model

| Variable | Score Chi-Square | Pr > ChiSq |
|---|---|---|
| Z1 | 0.0099 | 0.9209 |
| Z2 | 0.0580 | 0.8097 |
| Z3 | 2.0354 | 0.1537 |
| Z4 | 0.8051 | 0.3696 |
| Z5 | 0.8466 | 0.3575 |
| Z6 | 0.0231 | 0.8793 |
| Z7 | 2.6081 | 0.1063 |
| Z8 | 11.2959 | 0.0008 |
| z91 | 0.2108 | 0.6462 |
| z92 | 0.1297 | 0.7187 |
| z93 | 0.5913 | 0.4419 |
| Z10 | 0.8084 | 0.3686 |

Residual Chi-Square Test

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 19.1425 | 11 | 0.0586 |

Step 1. Variable Z8 is entered. The model contains the following explanatory variables:

g1 g2 Z8

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

| Criterion | Without Covariates | With Covariates |
|---|---|---|
| -2 LOG L | 380.487 | 353.210 |
| AIC | 380.487 | 359.210 |
| SBC | 380.487 | 364.423 |

The PHREG Procedure

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 27.2772 | 3 | <.0001 |
| Score | 27.8146 | 3 | <.0001 |
| Wald | 24.4659 | 3 | <.0001 |

Analysis of Maximum Likelihood Estimates

| Parameter Variable | DF | Standard Estimate | Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| g1 | 1 | -1.53927 | 0.51812 | 8.8260 | 0.0030 | 0.215 |
| g2 | 1 | -0.22154 | 0.48465 | 0.2090 | 0.6476 | 0.801 |
| Z8 | 1 | 1.31812 | 0.41855 | 9.9179 | 0.0016 | 3.736 |

Analysis of Variables Not in the Model

| Variable | Score Chi-Square | Pr > ChiSq |
|---|---|---|
| Z1 | 0.6505 | 0.4199 |
| Z2 | 0.0024 | 0.9613 |
| Z3 | 0.7547 | 0.3850 |
| Z4 | 0.9355 | 0.3334 |
| Z5 | 0.4415 | 0.5064 |
| Z6 | 0.0071 | 0.9329 |
| Z7 | 2.5716 | 0.1088 |
| z91 | 0.0330 | 0.8558 |
| z92 | 0.0003 | 0.9854 |
| z93 | 1.1483 | 0.2839 |
| Z10 | 0.7865 | 0.3751 |

Residual Chi-Square Test

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 7.7833 | 10 | 0.6500 |

NOTE: No (additional) variables met the 0.1 level for entry into the model.

Summary of Forward Selection

| Step | Variable Entered | Number In | Score Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| 1 | Z8 | 3 | 11.2959 | 0.0008 |

Model Information

| | |
|---|---|
| Data Set | WORK.BMT1 |
| Dependent Variable | T2 |
| Censoring Variable | delta2 |
| Censoring Value(s) | 0 |
| Ties Handling | BRESLOW |

| | |
|---|---|
| Number of Observations Read | 137 |
| Number of Observations Used | 137 |

Summary of the Number of Event and Censored Values

| Total | Event | Censored | Percent Censored |
|---|---|---|---|
| 137 | 42 | 95 | 69.34 |

The following variable(s) will be included in each model:

g1 g2

Step 0. The model contains the following variables:

g1 g2 Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 z91 z92 z93 Z10

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Analysis of Maximum Likelihood Estimates

| Parameter Variable | DF | Standard Estimate | Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| g1 | 1 | -1.97737 | 0.60510 | 10.6789 | 0.0011 | 0.138 |
| g2 | 1 | -0.49615 | 0.53394 | 0.8635 | 0.3528 | 0.609 |
| Z1 | 1 | 0.04161 | 0.03170 | 1.7223 | 0.1894 | 1.042 |
| Z2 | 1 | -0.03104 | 0.02816 | 1.2154 | 0.2703 | 0.969 |
| Z3 | 1 | -0.30270 | 0.35289 | 0.7358 | 0.3910 | 0.739 |
| Z4 | 1 | 0.39950 | 0.37151 | 1.1564 | 0.2822 | 1.491 |
| Z5 | 1 | 0.14540 | 0.38000 | 0.1464 | 0.7020 | 1.157 |
| Z6 | 1 | -0.07759 | 0.36061 | 0.0463 | 0.8296 | 0.925 |
| Z7 | 1 | -0.0009995 | 0.0007768 | 1.6555 | 0.1982 | 0.999 |
| Z8 | 1 | 1.44166 | 0.44917 | 10.3015 | 0.0013 | 4.228 |
| z91 | 1 | 0.63995 | 0.57090 | 1.2565 | 0.2623 | 1.896 |
| z92 | 1 | 0.51117 | 0.76634 | 0.4449 | 0.5048 | 1.667 |
| z93 | 1 | 0.86387 | 0.59507 | 2.1075 | 0.1466 | 2.372 |
| Z10 | 0 | 0 | . | . | . | . |

Step 1. Variable Z10 is removed because of its redundancy.

Step 2. Variable Z6 is removed. The model contains the following explanatory variables:

g1 g2 Z1 Z2 Z3 Z4 Z5 Z7 Z8 z91 z92 z93

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Analysis of Maximum Likelihood Estimates

| Parameter Variable | DF | Standard Estimate | Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| g1 | 1 | -1.96479 | 0.60164 | 10.6650 | 0.0011 | 0.140 |
| g2 | 1 | -0.48874 | 0.53240 | 0.8427 | 0.3586 | 0.613 |
| Z1 | 1 | 0.04197 | 0.03163 | 1.7606 | 0.1845 | 1.043 |
| Z2 | 1 | -0.03203 | 0.02772 | 1.3350 | 0.2479 | 0.968 |
| Z3 | 1 | -0.28731 | 0.34564 | 0.6909 | 0.4058 | 0.750 |
| Z4 | 1 | 0.40862 | 0.36893 | 1.2267 | 0.2680 | 1.505 |
| Z5 | 1 | 0.11851 | 0.35856 | 0.1092 | 0.7410 | 1.126 |
| Z7 | 1 | -0.0009920 | 0.0007736 | 1.6442 | 0.1998 | 0.999 |
| Z8 | 1 | 1.45514 | 0.44548 | 10.6698 | 0.0011 | 4.285 |
| z91 | 1 | 0.64716 | 0.56998 | 1.2892 | 0.2562 | 1.910 |
| z92 | 1 | 0.51320 | 0.76571 | 0.4492 | 0.5027 | 1.671 |
| z93 | 1 | 0.86823 | 0.59393 | 2.1370 | 0.1438 | 2.383 |

...

Step 11. Variable Z7 is removed. The model contains the following explanatory variables:

g1 g2 Z8

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Analysis of Maximum Likelihood Estimates

| Parameter Variable | DF | Standard Estimate | Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| g1 | 1 | -1.53927 | 0.51812 | 8.8260 | 0.0030 | 0.215 |
| g2 | 1 | -0.22154 | 0.48465 | 0.2090 | 0.6476 | 0.801 |
| Z8 | 1 | 1.31812 | 0.41855 | 9.9179 | 0.0016 | 3.736 |

NOTE: No (additional) variables met the 0.1 level for removal from the model.

Summary of Backward Elimination

| Step | Variable Removed | Number In | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| 1 | Z10 | 13 | . | . |
| 2 | Z6 | 12 | 0.0463 | 0.8296 |
| 3 | Z5 | 11 | 0.1092 | 0.7410 |
| 4 | z92 | 10 | 0.6167 | 0.4323 |
| 5 | Z3 | 9 | 0.6233 | 0.4298 |
| 6 | z91 | 8 | 0.5974 | 0.4396 |
| 7 | Z4 | 7 | 0.7501 | 0.3864 |
| 8 | z93 | 6 | 1.0015 | 0.3169 |
| 9 | Z2 | 5 | 1.2315 | 0.2671 |
| 10 | Z1 | 4 | 0.3947 | 0.5299 |
| 11 | Z7 | 3 | 2.4189 | 0.1199 |

Analysis of Variables Not in the Model

| Variable | Score Chi-Square | Pr > ChiSq |
|---|---|---|
| Z1 | 0.0099 | 0.9209 |
| Z2 | 0.0580 | 0.8097 |
| Z3 | 2.0354 | 0.1537 |
| Z4 | 0.8051 | 0.3696 |
| Z5 | 0.8466 | 0.3575 |
| Z6 | 0.0231 | 0.8793 |
| Z7 | 2.6081 | 0.1063 |
| Z8 | 11.2959 | 0.0008 |
| z91 | 0.2108 | 0.6462 |
| z92 | 0.1297 | 0.7187 |
| z93 | 0.5913 | 0.4419 |
| Z10 | 0.8084 | 0.3686 |

Residual Chi-Square Test

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 19.1425 | 11 | 0.0586 |

Step 1. Variable Z8 is entered. The model contains the following explanatory variables:

g1 g2 Z8

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Analysis of Variables Not in the Model

| Variable | Score Chi-Square | Pr > ChiSq |
|---|---|---|
| Z1 | 0.6505 | 0.4199 |
| Z2 | 0.0024 | 0.9613 |
| Z3 | 0.7547 | 0.3850 |
| Z4 | 0.9355 | 0.3334 |
| Z5 | 0.4415 | 0.5064 |
| Z6 | 0.0071 | 0.9329 |
| Z7 | 2.5716 | 0.1088 |
| z91 | 0.0330 | 0.8558 |
| z92 | 0.0003 | 0.9854 |
| z93 | 1.1483 | 0.2839 |
| Z10 | 0.7865 | 0.3751 |

Residual Chi-Square Test

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 7.7833 | 10 | 0.6500 |

NOTE: No (additional) variables met the 0.1 level for entry into the model.

Summary of Stepwise Selection

| Step | Variable Entered | Number Removed | Score In | Wald Chi-Square | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|
| 1 | Z8 | | 3 | 11.2959 | . | 0.0008 |

NOTE: The following variables are not used in the SCORE selection since they are a linear combination of other variables as shown.

Z10 = 1 * z92 + 1 * z93

Regression Models Selected by Score Criterion

| Number of Variables | Score Chi-Square | Variables Included in Model |
|---|---|---|
| 2 | 16.4771 | g1 g2 |
| 3 | 27.8146 | g1 g2 Z8 |
| 3 | 19.5130 | g1 g2 Z7 |
| 3 | 18.4199 | g1 g2 Z3 |
| 3 | 17.3021 | g1 g2 Z5 |
| 4 | 30.9681 | g1 g2 Z7 Z8 |
| 4 | 29.2600 | g1 g2 Z3 Z8 |
| 4 | 28.6084 | g1 g2 Z8 z93 |
| 4 | 28.4205 | g1 g2 Z4 Z8 |
| 5 | 31.8900 | g1 g2 Z3 Z7 Z8 |
| 5 | 31.6531 | g1 g2 Z7 Z8 z93 |
| 5 | 31.3819 | g1 g2 Z5 Z7 Z8 |
| 5 | 31.3151 | g1 g2 Z4 Z7 Z8 |
| 6 | 32.5001 | g1 g2 Z3 Z4 Z7 Z8 |
| 6 | 32.4470 | g1 g2 Z3 Z7 Z8 z93 |
| 6 | 32.4034 | g1 g2 Z3 Z5 Z7 Z8 |
| 6 | 32.0224 | g1 g2 Z4 Z7 Z8 z93 |

```
7     33.0644  g1 g2 Z3 Z4 Z7 Z8 z93
7     32.9565  g1 g2 Z3 Z4 Z5 Z7 Z8
7     32.8969  g1 g2 Z3 Z5 Z7 Z8 z93
7     32.6351  g1 g2 Z3 Z7 Z8 z91 z93

8     33.4602  g1 g2 Z3 Z4 Z5 Z7 Z8 z93
8     33.2658  g1 g2 Z3 Z4 Z7 Z8 z91 z93
8     33.1314  g1 g2 Z2 Z3 Z4 Z7 Z8 z93
8     33.1069  g1 g2 Z3 Z4 Z7 Z8 z92 z93

9     33.6562  g1 g2 Z3 Z4 Z5 Z7 Z8 z91 z93
9     33.5729  g1 g2 Z2 Z3 Z4 Z5 Z7 Z8 z93
9     33.4994  g1 g2 Z3 Z4 Z5 Z6 Z7 Z8 z93
9     33.4965  g1 g2 Z3 Z4 Z7 Z8 z91 z92 z93

10     33.7837  g1 g2 Z3 Z4 Z5 Z7 Z8 z91 z92 z93
10     33.7755  g1 g2 Z1 Z2 Z3 Z4 Z7 Z8 z91 z93
10     33.7660  g1 g2 Z2 Z3 Z4 Z5 Z7 Z8 z91 z93
10     33.7142  g1 g2 Z1 Z2 Z3 Z4 Z5 Z7 Z8 z93

11     34.0330  g1 g2 Z1 Z2 Z3 Z4 Z5 Z7 Z8 z91 z93
11     33.9245  g1 g2 Z1 Z2 Z3 Z4 Z7 Z8 z91 z92 z93
11     33.8610  g1 g2 Z2 Z3 Z4 Z5 Z7 Z8 z91 z92 z93
11     33.8210  g1 g2 Z3 Z4 Z5 Z6 Z7 Z8 z91 z92 z93

12     34.1143  g1 g2 Z1 Z2 Z3 Z4 Z5 Z7 Z8 z91 z92 z93
12     34.0635  g1 g2 Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 z91 z93
12     33.9279  g1 g2 Z1 Z2 Z3 Z4 Z6 Z7 Z8 z91 z92 z93
12     33.8841  g1 g2 Z2 Z3 Z4 Z5 Z6 Z7 Z8 z91 z92 z93

13     34.1454  g1 g2 Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 z91 z92 z93
```

After fitting a Cox's regression model to a practical data set, it is important to check whether the Cox's regression model is an correct model for this data set.

Like in ordinary regression, we use residuals to assess goodness of fit.

In survival analysis, several types of residuals can be determined:

- Cox-Snel residuals
- Schoenfeld residuals
- Martingale residuals

As a first type of residuals, we consider Cox-Snel residuals.

These residuals are defined as

$$r_i = \hat{H}_0(t_i) \exp(x_i \hat{\beta})$$

where $\hat{H}_0(t_i)$ is the baseline cumulative hazard function at $t_i$.

If the Cox's regression model is satisfied, we get that $r_i$ is a censored sample of an exponential distribution with lambda 1.

Testing for this, gives a test of model adequacy.

Graphically, we can do this by plotting a cumulative hazard estimate of these residuals. Since $H_0(t) = t$ for an exponential distribution, we should see a straight line.

We concentrate on martingale residuals.

For every individual, we get

$$r_i^m = \delta_i - \hat{H}(T_i), \quad i = 1, \ldots, n$$

where $\hat{H}$ is the fitted cumulative hazard function under the Cox's regression model.

We note that

- The martingale residuals sum to zero.
- In large sample, the martingale residuals are uncorrelated and have an expected value of zero.
- For each individual, the martingale residual looks like the difference of the observed number of deaths in interval $[0, t_i[$ minus expected number under the fitted value.

By plotting these values against index, fitted values or covariates we asses whether there are outliers, the models fits or the functional form of the covariates is satisfied.

A third type of residuals are the Schoenfeld residuals.

From the partial likelihood, we know that the parameter $\beta$ are estimated from

$$\sum_{i=1}^{d}(x_i - E[x_i|\mathcal{R}(t_i)]) = 0$$

with

$$E[x_i|\mathcal{R}(t_i)] = \frac{\sum_{l\in\mathcal{R}(t_i)} x_l \exp(x_l^t\beta)}{\sum_{l\in\mathcal{R}(t_i)} \exp(x_l^t\beta)}.$$

The Schoenfeld residual are defined as

$$r_i^s = x_i - E[x_i|\mathcal{R}(t_i)]$$

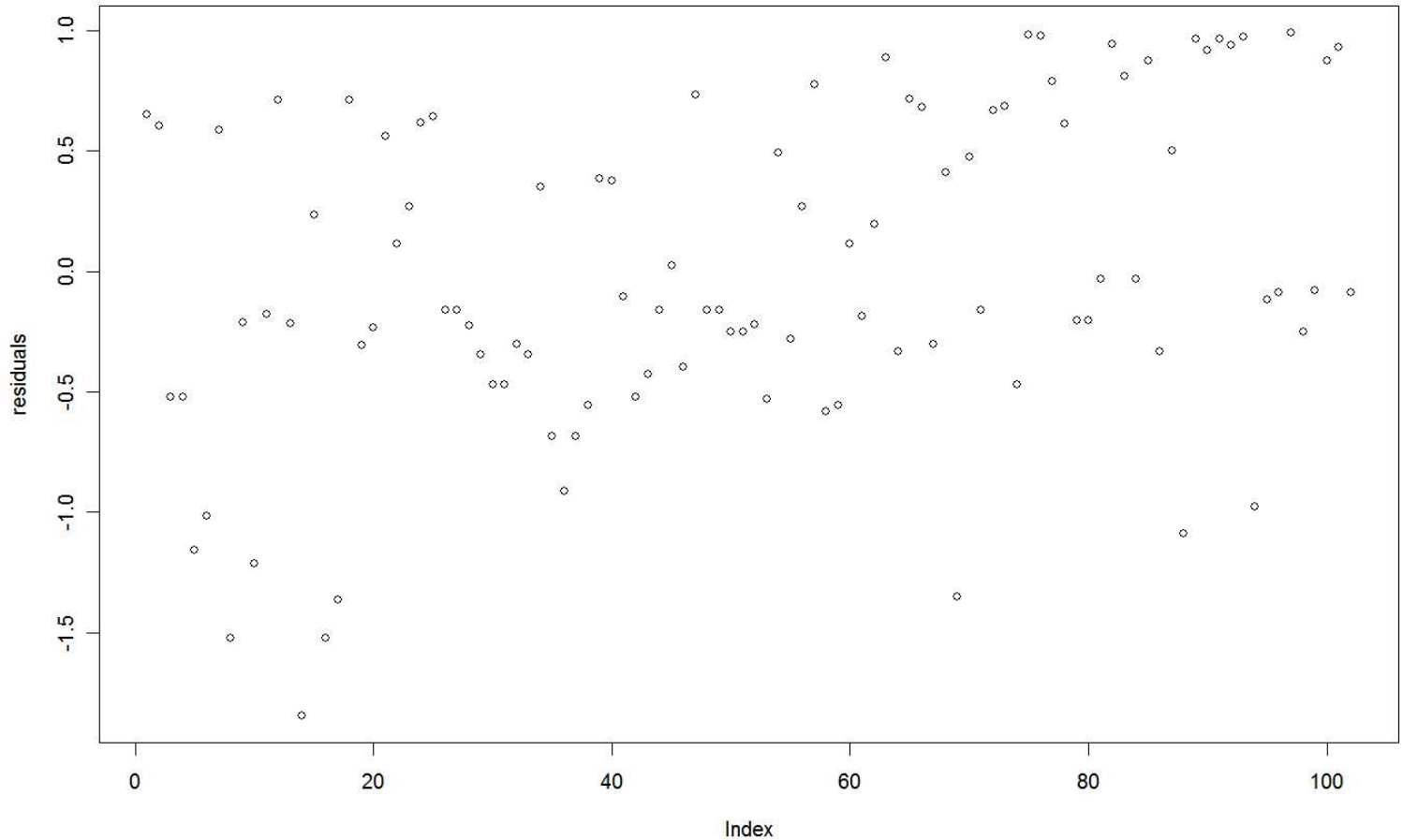
We note that this leads to a multivariate residual.

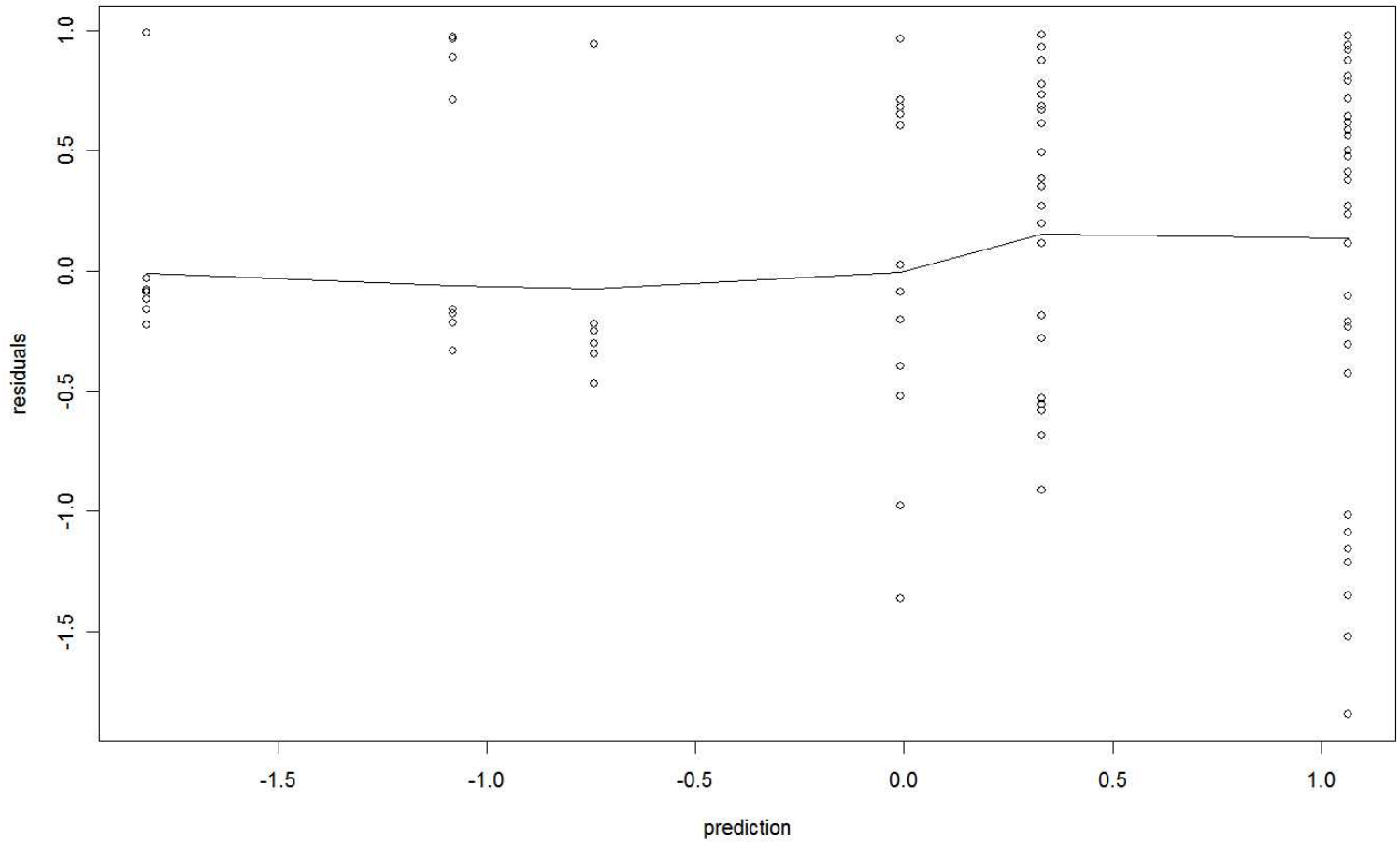
We use these residuals to asses whether the covariates satisfy the proportional hazards assumption.

Hereto we plot them versus ranks of the survival times.

# Example: NSCLC

```
> fit<-coxph(Surv(survtime,survind)~expres+tnm,data=NSCLC)
> summary(fit)
>
> fitresi<-residuals(fit,type="martingale")
> fitpred<-predict(fit)
>
> plot(fitresi,ylab="residuals")
>
> plot(fitpred,fitresi,xlab="prediction",ylab="residuals")
> lines(lowess(fitpred,fitresi))
>
> plot(NSCLC$expres,fitresi,xlab="expression",ylab="residuals")
> lines(lowess(NSCLC$expres,fitresi))
>
> plot(NSCLC$tnm,fitresi,xlab="TNM",ylab="residuals")
> lines(lowess(NSCLC$tnm,fitresi))
```


residuals
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
20
40
60
80
100
Index


residuals
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
prediction

Until now, we assumed that a covariate $X$ satisfied the proportional hazards assumption. However in practice we need to check for this.

Under this assumption, we have that

$$\Lambda(t|X) = \int_0^t \lambda(s|X)ds = e^{\beta X} \int_0^t \lambda_0(s)ds = e^{\beta X}\Lambda_0(t)$$

and

$$S(t|X) = e^{-\Lambda(t|X)} = e^{-\Lambda_0(t)e^{\beta X}} = S_0(t)^{e^{\beta X}}.$$

Hence,

$$\log(-\log(S(t|X))) = \log(-\log(S_0(t))) + \beta X.$$

We use this relationship for a graphical check of proportional hazards.

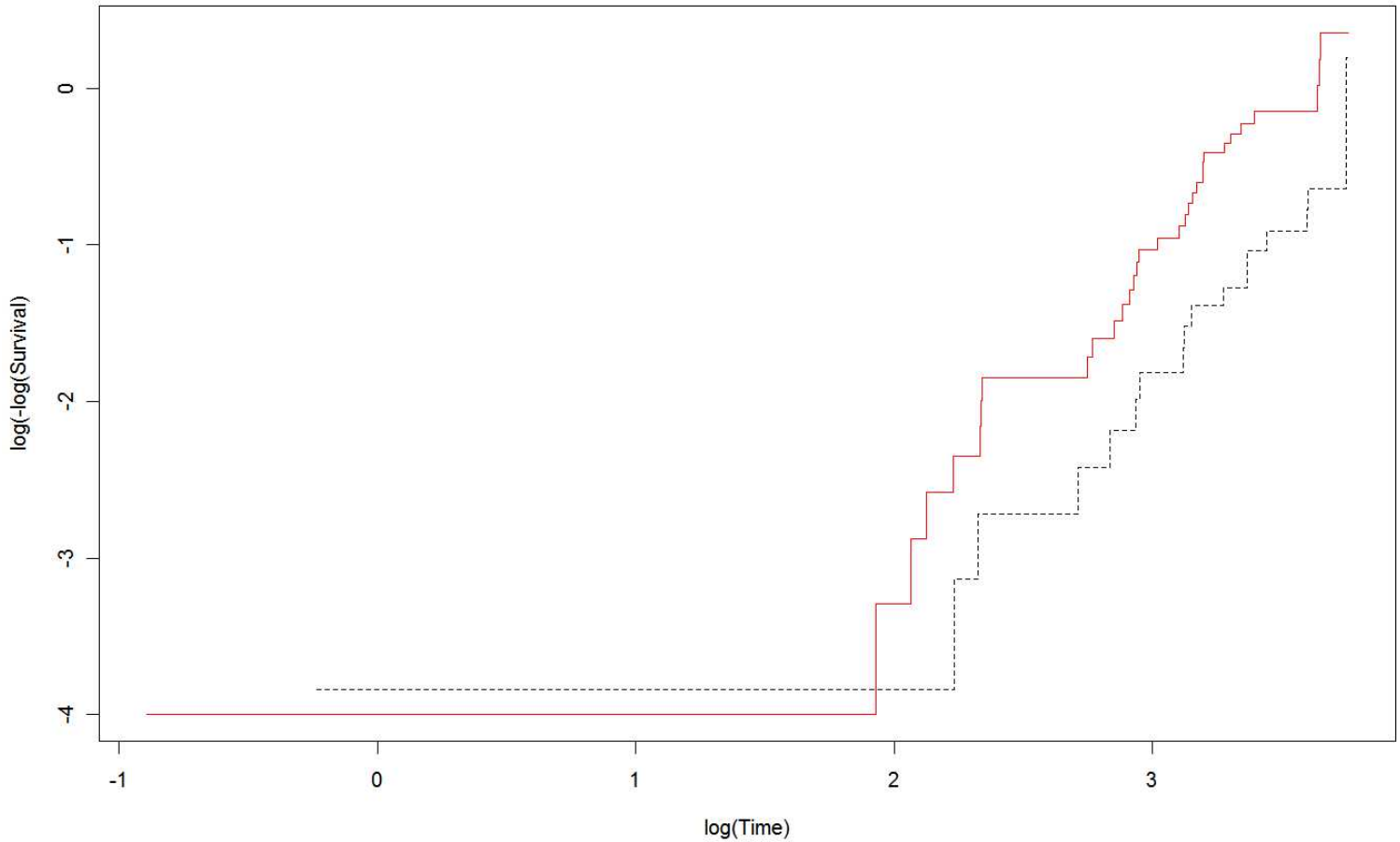
Consider a discrete covariate $X$ with levels $x_1, \ldots, x_K$.

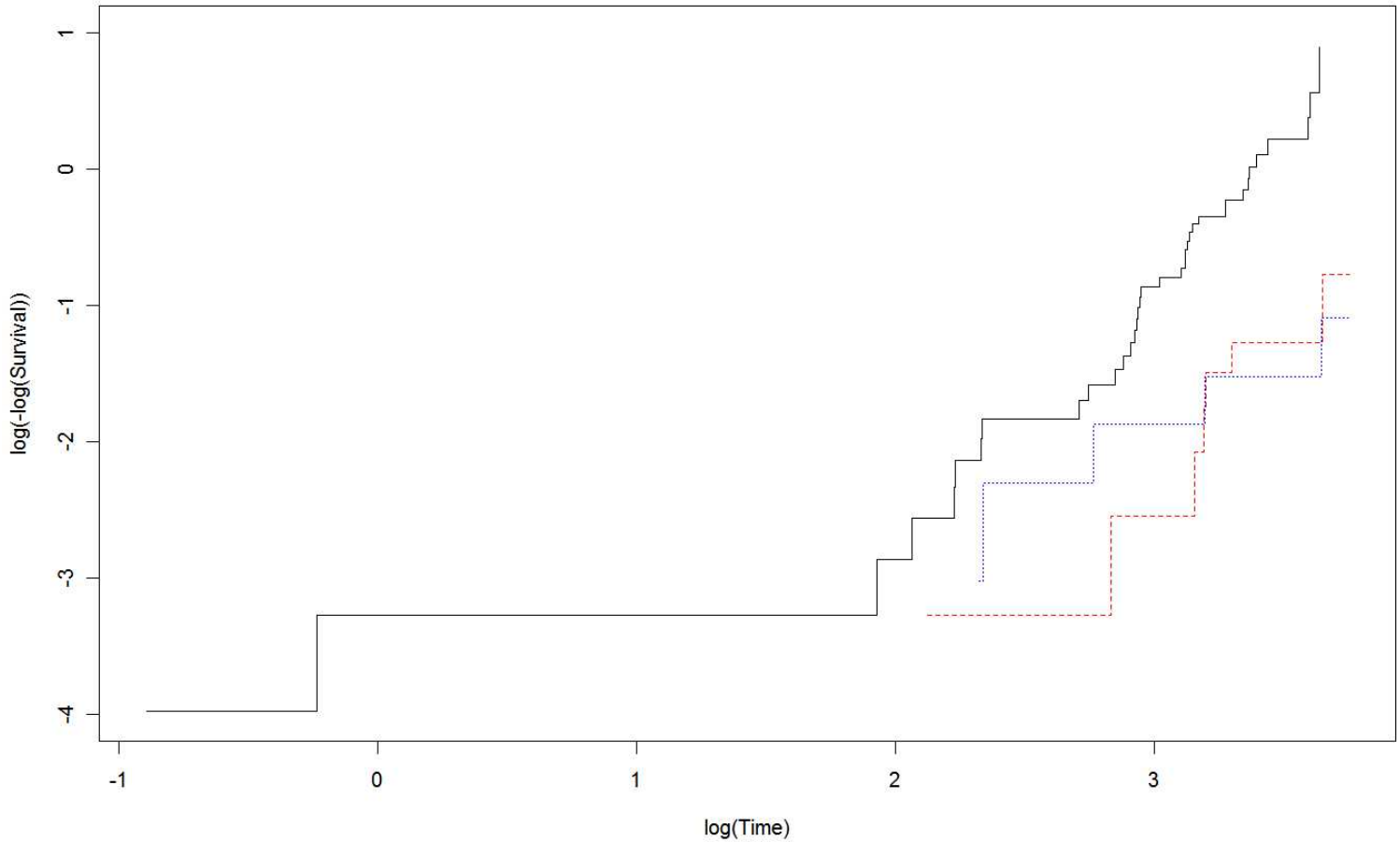
- Calculate KM curves for the various levels of $X$.
- Plot $\log(-\log(S(t|X = x))$ versus $\log(t)$.
- If they are parallel, the proportional hazard assumption is satisfied.

**Note:**

- For a continuous variable, we discretizes into categories.
- For more than one variable, we calculate a KM curve for each combination of values.

```
fit<-survfit(Surv(survtime,survind)~expres,data=NSCLC)
summary(fit)
plot(log(fit[2]$time),log(-log(fit[2]$surv)),xlab="log(Time)",ylab="log(-log(Survival)",
type="s",col="red")
lines(log(fit[1]$time),log(-log(fit[1]$surv)),lty=2,type="s")

fit<-survfit(Surv(survtime,survind)~tnm,data=NSCLC)
summary(fit)
plot(log(fit[3]$time),log(-log(fit[3]$surv)),xlab="log(Time)",ylab="log(-log(Survival))",
type="s",ylim=c(-4,1))
lines(log(fit[2]$time),log(-log(fit[2]$surv)),lty=2,type="s",col="red")
lines(log(fit[1]$time),log(-log(fit[1]$surv)),lty=3,type="s",col="blue")
```


log(-log(Survival))
0
-1
-2
-3
-4
-1
0
1
2
3
log(Time)


log(-log(Survival))
1
0
-1
-2
-3
-4
-1
0
1
2
3
log(Time)

Until now we assumed only one covariate but the idea also holds for multiple covariates.

Consider a discrete covariate $X$ with levels $x_1, \ldots, x_K$.

- Calculate the fitted curves containing other covariates for the various levels of $X$.
- Plot $\log(-\log(S(t|X = x))$ versus $\log(t)$.
- If they are parallel, the proportional hazard assumption is satisfied.

A major use of time-dependent covariate methodology is to test the proportionality assumption for a fixed-time covariate $X_1$.

- Define an **artificial** time-dependent covariate $X_2(t)$,

$$X_2(t) = X_1 \times g(t)$$

  where $g(t)$ is a known function of time $t$ (ex: $g(t) = \log(t)$).
- Test $H_0 : \beta_2 = 0$ in the Cox's model,

$$\lambda(t|X_1, X_2(t)) = \lambda_0(t)e^{\beta_1 X_1 + \beta_2 X_2(t)}.$$

Now,

$$HR(t) = e^{\beta_1(X_1 - X_1^*) + \beta_2 g(t)(X_1 - X_1^*)}$$

which is independent of $t$ when $\beta_2 = 0$.

# Example: NSCLC

```
> fit<-coxph(Surv(survtime,survind)~expres+tnm+tt(expres),data=NSCLC,tt=function(x,t,...)x*log(t))
> summary(fit)
Call: coxph(formula = Surv(survtime, survind) ~ expres + tnm + tt(expres),data = NSCLC,
 tt = function(x, t, ...) x * log(t))

n= 102, number of events= 49

              coef exp(coef) se(coef)     z Pr(>|z|)
expres     0.49387   1.63864  1.08088 0.457    0.648
tnm        1.07384   2.92660  0.25485 4.214 2.51e-05 ***
tt(expres) 0.08354   1.08712  0.36139 0.231    0.817
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

           exp(coef) exp(-coef) lower .95 upper .95
expres         1.639     0.6103    0.1970    13.631
tnm            2.927     0.3417    1.7760     4.823
tt(expres)     1.087     0.9199    0.5354     2.207

Concordance= 0.7  (se = 0.287 )
Rsquare= 0.259   (max possible= 0.98 )
Likelihood ratio test= 30.63  on 3 df,   p=1.015e-06
Wald test            = 23.79  on 3 df,   p=2.761e-05
Score (logrank) test = 27.03  on 3 df,   p=5.792e-06
```

```
> fit<-coxph(Surv(survtime,survind)~expres+tnm+tt(tnm),data=NSCLC,tt=function(x,t,...)x*log(t))
> summary(fit)
Call: coxph(formula = Surv(survtime, survind) ~ expres + tnm + tt(tnm),data = NSCLC,
 tt = function(x, t, ...) x * log(t))

n= 102, number of events= 49

          coef exp(coef) se(coef)     z Pr(>|z|)
expres  0.7399    2.0957   0.3035 2.438   0.0148 *
tnm     0.5151    1.6737   0.7837 0.657   0.5110
tt(tnm) 0.1917    1.2113   0.2612 0.734   0.4632
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

        exp(coef) exp(-coef) lower .95 upper .95
expres      2.096     0.4772    1.1560     3.799
tnm         1.674     0.5975    0.3603     7.776
tt(tnm)     1.211     0.8256    0.7259     2.021

Concordance= 0.695  (se = 0.287 )
Rsquare= 0.262   (max possible= 0.98 )
Likelihood ratio test= 31.05  on 3 df,   p=8.303e-07
Wald test            = 24.22  on 3 df,   p=2.251e-05
Score (logrank) test = 27.67  on 3 df,   p=4.254e-06
```

# Example: NSCLC

```
proc phreg data=nsclc;
model Survtime*Survind(0)=expres TNM exptime;
exptime=expres*log(Survtime);
run;
```

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| expres | 1 | 0.49389 | 1.08088 | 0.2088 | 0.6477 | 1.639 |
| tnm | 1 | 1.07358 | 0.25482 | 17.7503 | <.0001 | 2.926 |
| exptime | 1 | 0.08352 | 0.36139 | 0.0534 | 0.8172 | 1.087 |

```
proc phreg data=nsclc;
model Survtime*Survind(0)=expres TNM TNMtime;
TNMtime=TNM*log(Survtime);
run;
```

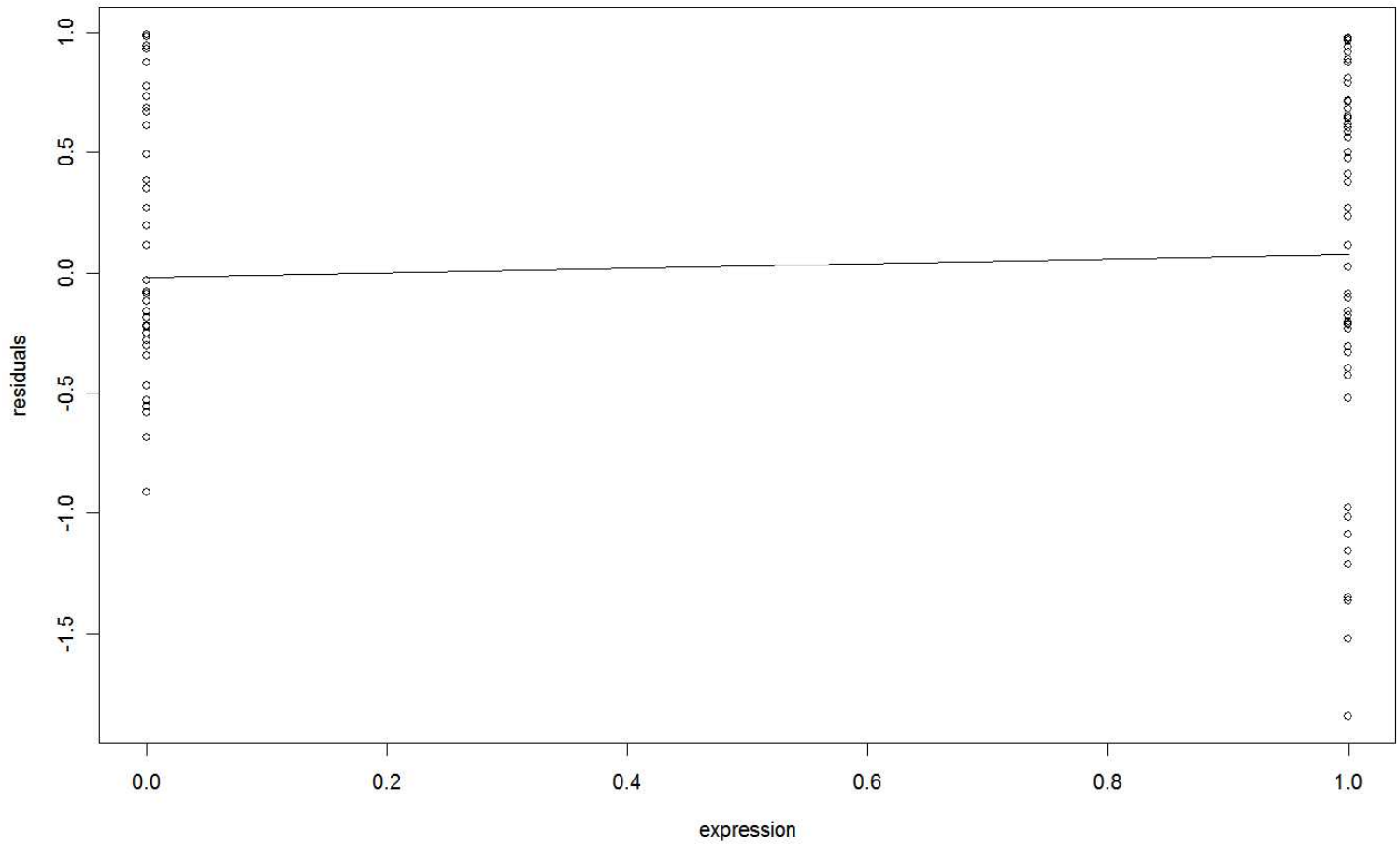
| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|
| expres | 1 | 0.73986 | 0.30352 | 5.9417 | 0.0148 | 2.096 |
| tnm | 1 | 0.51472 | 0.78357 | 0.4315 | 0.5113 | 1.673 |
| TNMtime | 1 | 0.19170 | 0.26120 | 0.5387 | 0.4630 | 1.211 |

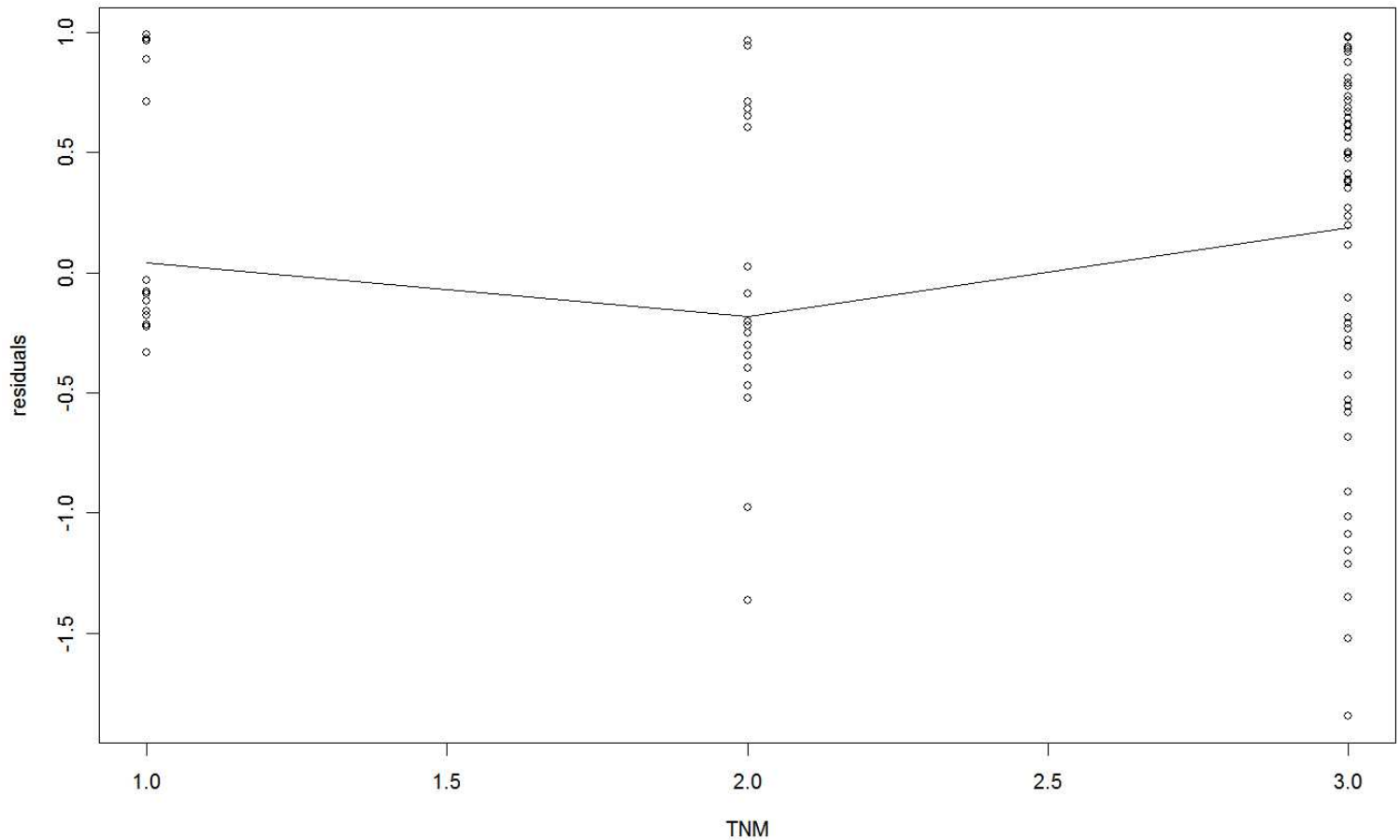
Sometimes, we are interested in checking which the best functional form of a covariate.

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp(f(x_i)\beta_1 + x_m^t\beta)$$

To check which functional form is best, we can estimate several functional forms and compare them by the partial likelihood ratio test.

By plotting the martingale residuals versus a covariate, we can verify whether the functional form is correct.


residuals
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0
0.2
0.4
0.6
0.8
1.0
expression


residuals
1.0
0.5
0.0
-0.5
-1.0
-1.5
1.0
1.5
2.0
2.5
3.0
TNM

In the previous lectures, we did not make any assumptions on the distribution of the lifetime variable $T$

Since this random variable is only defined for positive value and often is highly skewed, we intuitively feel that a normal distribution is not suited to describe this distribution.

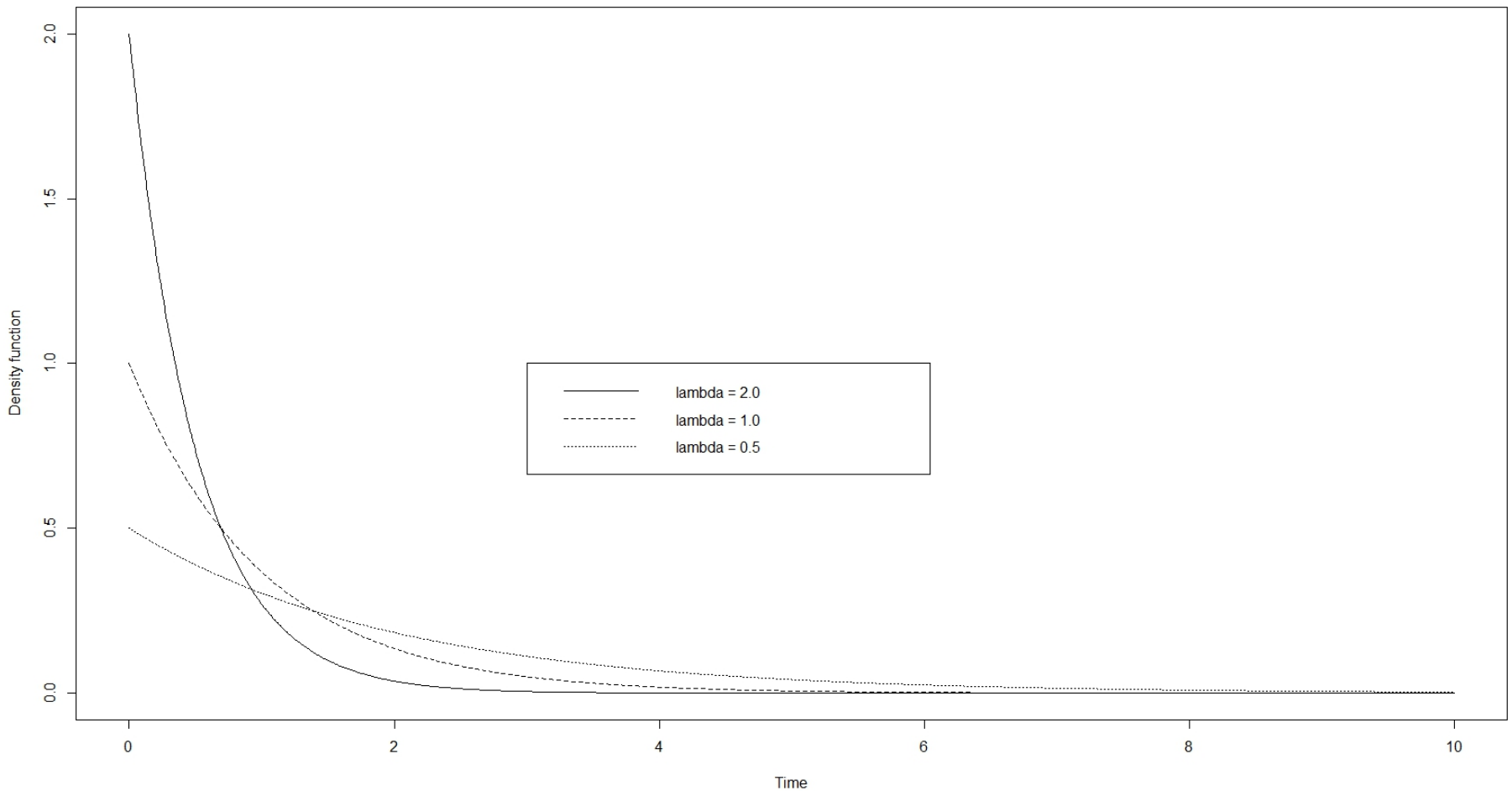
Therefore we propose some popular survival distributions.

As a first popular parametric distribution, we consider the exponential distribution.

A lifetime $T$ has an exponential distribution if for $\lambda > 0$,

$$\begin{aligned} f(t) &= \lambda \exp(-\lambda t), \quad t > 0 \\ S(t) &= \exp(-\lambda t), \quad t > 0 \\ \lambda(t) &= \lambda, \quad t > 0 \end{aligned}$$

This is the only distribution with a constant hazard.


Density function
0.0
0.5
1.0
1.5
2.0
0
2
4
6
8
10
Time
lambda = 2.0
lambda = 1.0
lambda = 0.5

Furthermore we can show that

$$E[T] = \frac{1}{\lambda} < +\infty.$$

The median $t_{50\%}$ is the time point at which 50% has failed,

$$S(t_{50\%}) = 0.5 \Leftrightarrow t_{50\%} = \frac{1}{\lambda} \log 2.$$

More generally, the $p$-th percentile of the survival distribution is given by

$$S(t_p) = 1 - p \Leftrightarrow t_p = -\frac{\log(1-p)}{\lambda}.$$

The exponential distribution has a lack of memory property, given by

$$P(T > t + x|T > x) = P(T > t).$$

In reality, most processes rarely follow an exponential distribution.

This is the result of having a hazard function which depends on time and age of an individual.

If we have events which tend to occur constantly over time, we still can use the exponential distribution.

For the mean residual lifetime, we get

$$r(t) = E[T - t|T > t] = E[T] = \frac{1}{\lambda}.$$

This means that the expected future lifetime for a system or individual is not influenced by the time $t$ that it has lived so far.

It is assumed that there is no ageing effect.

Also if we look a small time intervals, the exponential distribution gives good approximations.

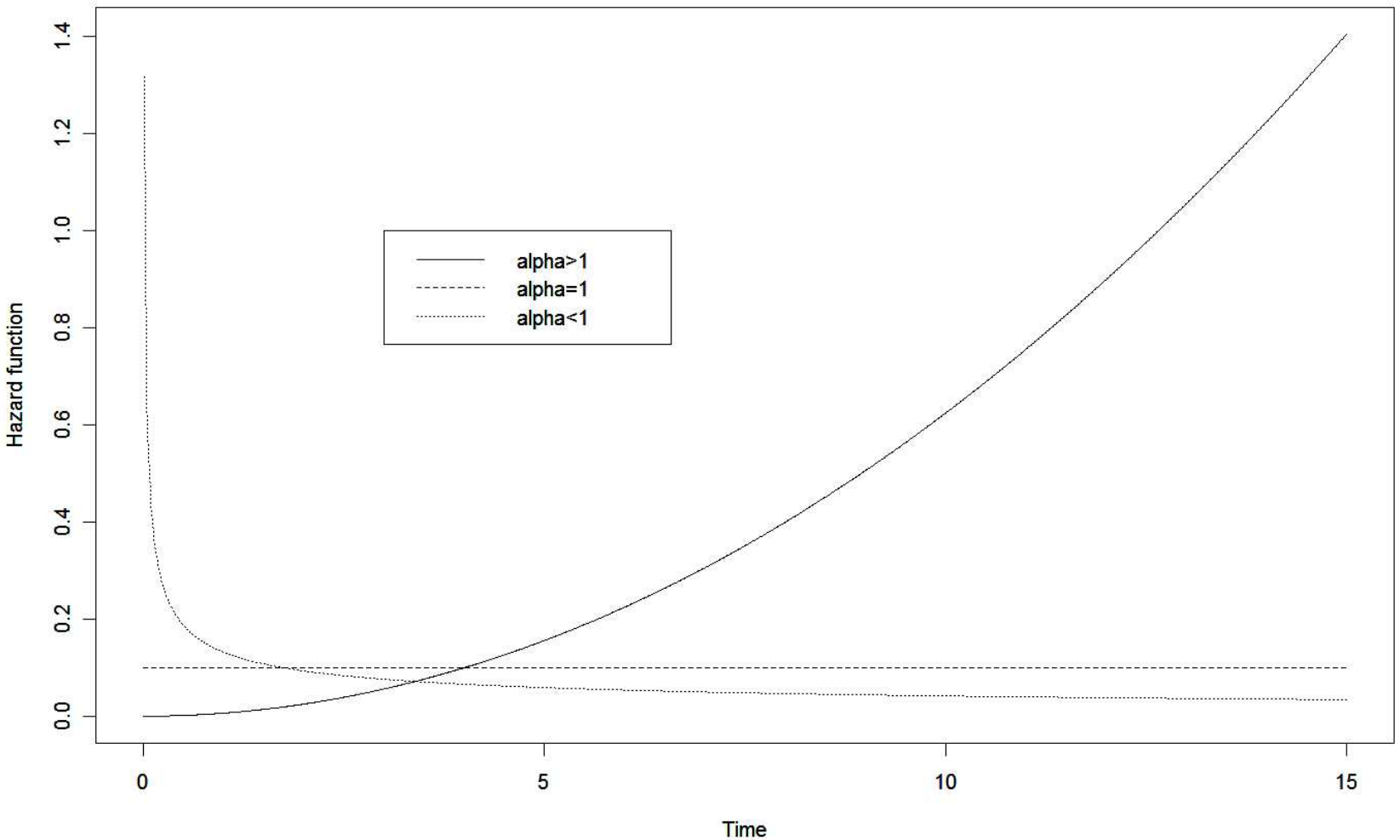
# Weibull distribution

A lifetime $T$ has a Weibull distribution if for $\lambda > 0$ and $\alpha > 0$,

$$
\begin{aligned}
f(t) &= \alpha\lambda \exp(-\lambda t^{\alpha-1}), \quad t > 0 \\
S(t) &= \exp(-\lambda t^{\alpha}), \quad t > 0 \\
\lambda(t) &= \lambda\alpha t^{\alpha-1}, \quad t > 0
\end{aligned}
$$

where $\lambda$ is a scale parameter while $\alpha$ is a shape parameter.

We note that the exponential distribution is a special case of the Weibull ($\alpha = 1$).


Hazard function
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
alpha>1
alpha=1
alpha<1
0
5
10
15
Time

As for the exponential distribution, we can show that

$$E[T] = \lambda^{-1/\alpha}\Gamma(\alpha^{-1} - 1)$$

with

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du.$$

The median $t_{50\%}$ is the time point at which 50% has failed,

$$S(t_{50\%}) = 0.5 \Leftrightarrow t_{50\%} = \left(\frac{1}{\lambda}\log 2\right)^{1/\alpha}.$$

More generally, the $p$-th percentile of the survival distribution is given by

$$S(t_p) = 1 - p \Leftrightarrow t_p = \left(-\frac{\log(1-p)}{\lambda}\right)^{1/\alpha}.$$

# Extreme value distribution

Sometimes it is more useful to work with the logarithm of the lifetimes.

If $T \sim \text{Weibull}(\lambda, \alpha)$, we get that

$$Y = \log(T) = \mu + \sigma E$$

with $\mu = (-\log \lambda)/\alpha$, $\sigma = 1/\alpha$ and $E$ the standard extreme value distribution

$$f_E(w) = \exp(w - e^w), \quad w \in I\!R.$$

We note that a reparametrization of the parameters gives $\mu \in I\!R$ and $\sigma > 0$.

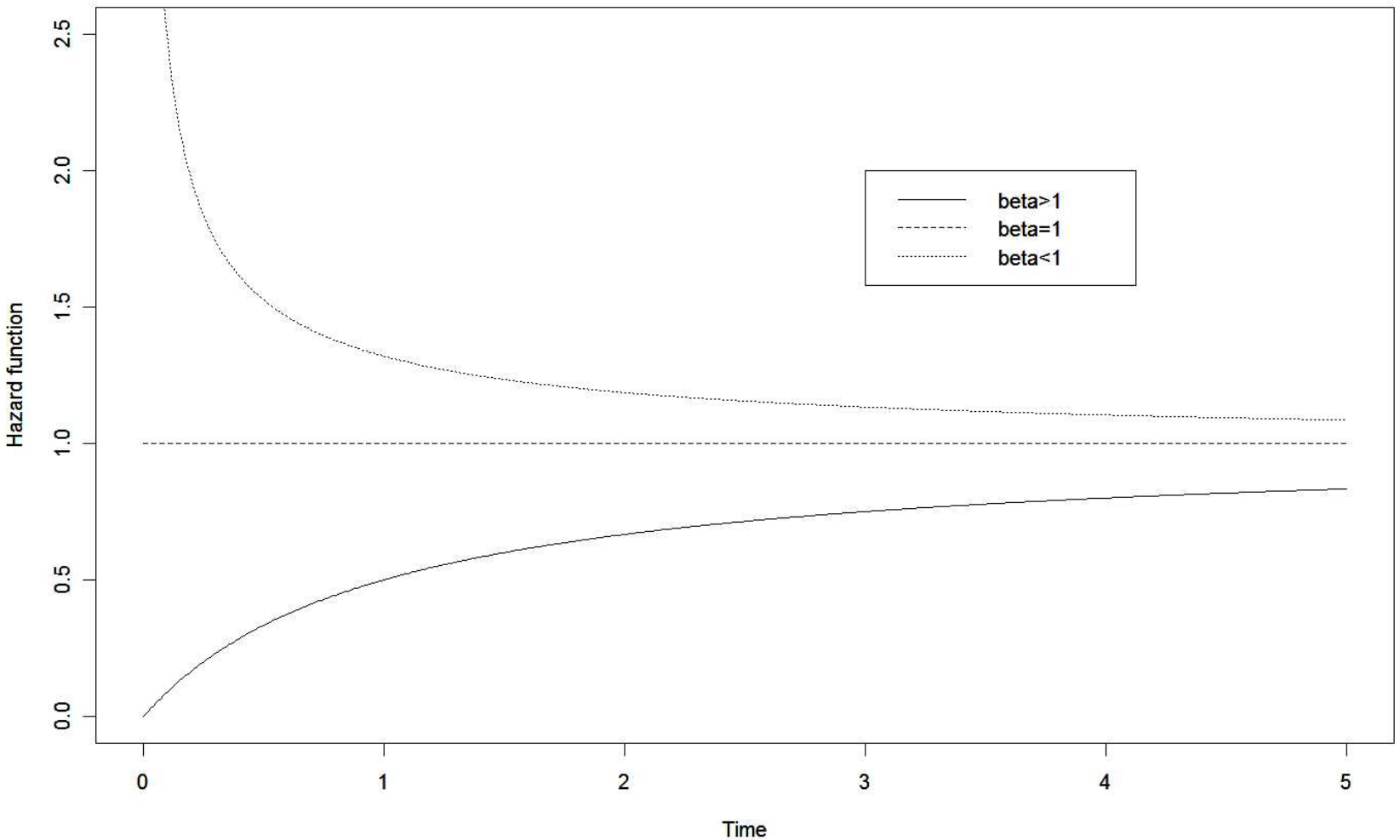
A lifetime $T$ has a gamma distribution if for $\lambda > 0$ and $\beta > 0$,

$$\begin{aligned} f(t) &= \frac{\lambda^{\beta} t^{\beta-1}}{\Gamma(\beta)} \exp(-\lambda t), \quad t > 0 \\ S(t) &= 1 - \frac{1}{\Gamma(\beta)} \int_0^{\lambda t} u^{\beta-1} \exp(-u) du, \quad t > 0 \end{aligned}$$

where $\lambda$ is a scale parameter while $\beta$ is a shape parameter.

We note that

$$E[T] = \frac{\beta}{\lambda} \qquad \text{and} \qquad \text{Var}(T) = \frac{\beta}{\lambda^2}.$$

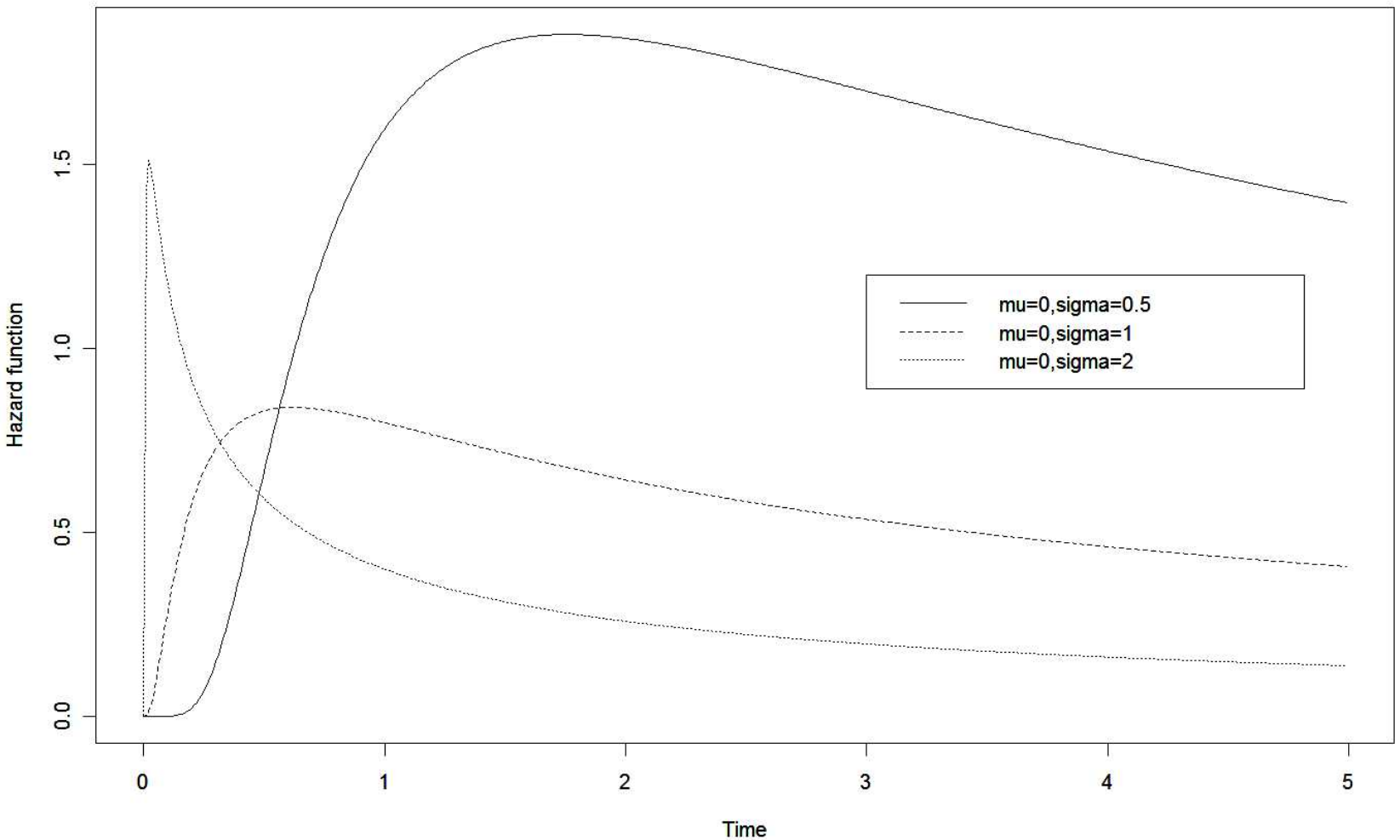

Hazard function
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4
5
Time
beta>1
beta=1
beta<1

A lifetime $T$ has a lognormal distribution if $\log(T)$ has a normal distribution.

This gives

$$\begin{aligned} f(t) &= \frac{1}{t\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right), \quad t > 0 \\ &= \frac{1}{t}\phi\left(\frac{\log t - \mu}{\sigma}\right), \quad t > 0 \\ S(t) &= 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right), \quad t > 0 \end{aligned}$$

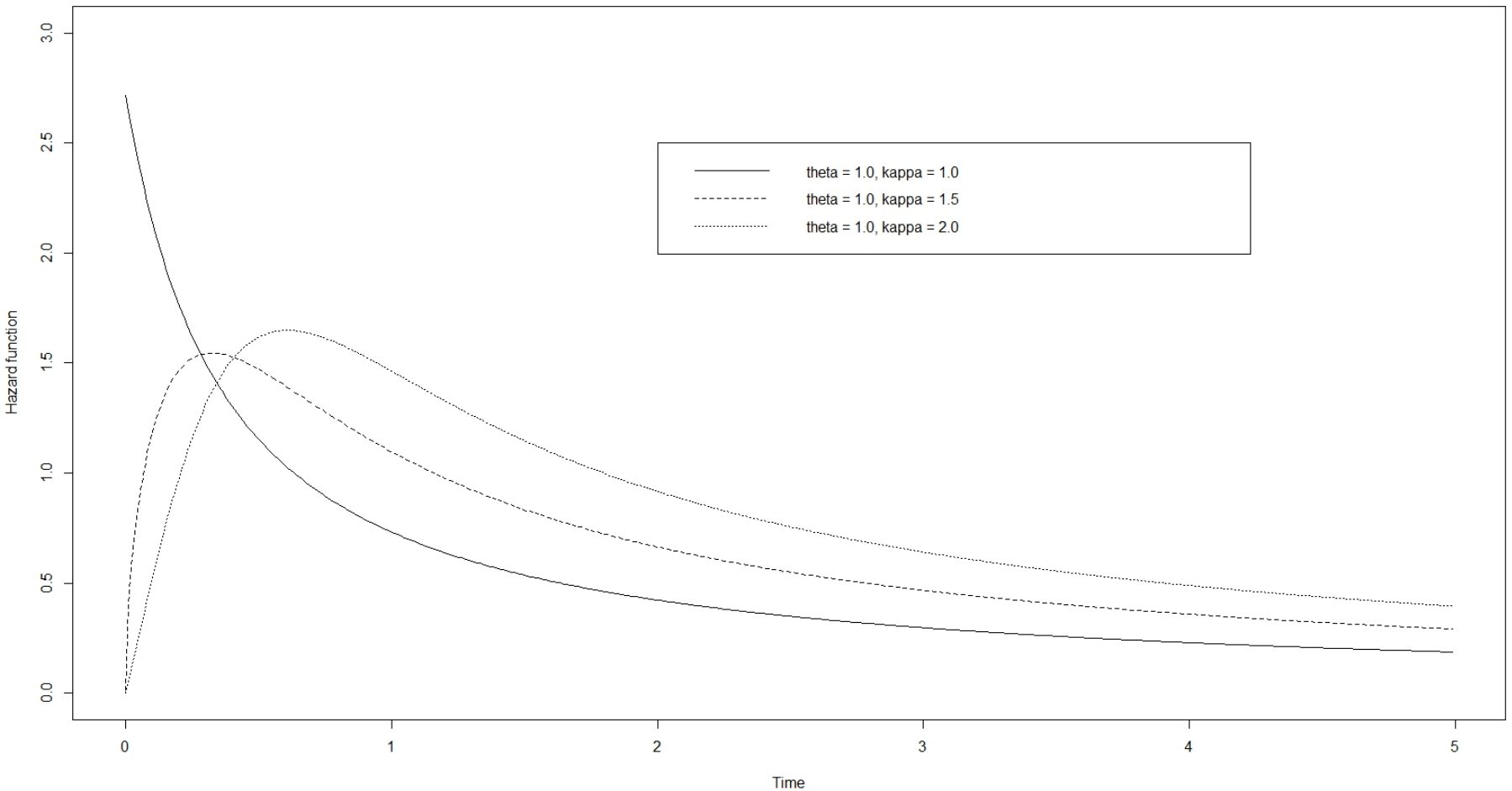
where $\mu \in \mathbb{R}$ and $\sigma > 0$.


Hazard function
0.0
0.5
1.0
1.5
0
1
2
3
4
5
Time
mu=0,sigma=0.5
mu=0,sigma=1
mu=0,sigma=2

A lifetime $T$ has a log-logistic distribution when for $\kappa > 0$, $\theta \in \mathbb{R}$,

$$\begin{aligned} f(t) &= \frac{e^{\theta}\kappa t^{\kappa-1}}{(1+e^{\theta}t^{\kappa})^2} \\ S(t) &= [1+e^{\theta}t^{\kappa}]^{-1} \\ \lambda(t) &= \frac{e^{\theta}\kappa t^{\kappa-1}}{1+e^{\theta}t^{\kappa}} \end{aligned}$$

As for the lognormal distribution, we note that $T$ has a log-logistic distribution if $\log(T)$ has a logistic distribution.


theta = 1.0, kappa = 1.0
theta = 1.0, kappa = 1.5
theta = 1.0, kappa = 2.0
Hazard function
Time
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5

A lifetime $T$ has a Gompertz distribution if

$$\begin{aligned} f(t) &= \theta e^{\alpha t} \exp\left(\frac{\theta}{\alpha}(1-e^{\alpha t})\right), \quad t>0 \\ S(t) &= \exp\left(\frac{\theta}{\alpha}(1-e^{\alpha t})\right), \quad t>0 \\ \lambda(t) &= \theta e^{\alpha t}, \quad t>0 \end{aligned}$$

where $\alpha > 0$ and $\theta > 0$.

We note that the hazard function is increasing in $t$ and that $\lim_{t\to 0} \lambda(t) = \theta \Rightarrow \theta$ is background hazard.

# Generalized Gamma distribution

We extend the Gamma distribution and define

$$f(t) = \frac{\theta \lambda^{\rho\theta} t^{\rho\theta - 1} \exp[-(\lambda t)^{\theta}]}{\Gamma(\rho)}.$$

- If $\theta = \rho = 1 \Rightarrow$ Exponential
- If $\theta = 1 \Rightarrow$ Gamma
- If $\rho = 1 \Rightarrow$ Weibull
- If $\rho \to +\infty \Rightarrow$ Lognormal

# Pareto distribution

A lifetime $T$ has a Pareto distribution if

$$\begin{aligned} f(t) &= \frac{\theta\lambda^\theta}{t^{\theta+1}}, \quad t > \lambda \\ S(t) &= \frac{\lambda^\theta}{t^\theta}, \quad t > \lambda \\ \lambda(t) &= \frac{\theta}{t}, \quad t > \lambda \end{aligned}$$

where $\lambda > 0$ and $\theta > 0$.

# Inverse Gaussian distribution

A lifetime $T$ has an Inverse Gaussian distribution if

$$f(t) = \sqrt{\frac{\lambda}{2\pi t^3}} \exp\left(\frac{\lambda(t-\mu^2)}{2\mu^2 t}\right), \quad t > 0$$

$$S(t) = \Phi\left(\sqrt{\frac{\lambda}{t}}\left(1-\frac{t}{\mu}\right)\right) - e^{\frac{2\lambda}{\mu}}\Phi\left(-\sqrt{\frac{\lambda}{t}}\left(1+\frac{t}{\mu}\right)\right), \quad t > 0$$

where $\lambda > 0$.

The hazard of this distribution has a complicated form!

For some parametric distributions, we can easily verify whether they are a good candidate.

- If $T \sim \text{Exp}(\lambda)$, we have that

$$\begin{aligned} S(t) &= \exp(-\lambda t) \\ \log(S(t)) &= -\lambda t. \end{aligned}$$

$\Rightarrow$ Plot $\log(\hat{S}(t))$ vs $t$.

- If $T \sim \text{Weibull}(\lambda, \alpha)$, we have that

$$\begin{aligned} S(t) &= \exp(-\lambda t^{\alpha}) \\ \log(S(t)) &= -\lambda t^{\alpha} \\ \log(-\log(S(t))) &= \log(\lambda) + \alpha \log(t). \end{aligned}$$

$\Rightarrow$ Plot $\log(-\log(\hat{S}(t)))$ vs $\log(t)$.

- If $T$ has a log-logistic distribution, we have that

$$\begin{aligned} \frac{S(t)}{1-S(t)} &= e^{-\theta} t^{-\kappa} \\ \log\left(\frac{S(t)}{1-S(t)}\right) &= -\theta - \kappa \log(t). \end{aligned}$$

$\Rightarrow$ Plot $\log\left(\frac{\hat{S}(t)}{1-\hat{S}(t)}\right)$ vs $\log(t)$.

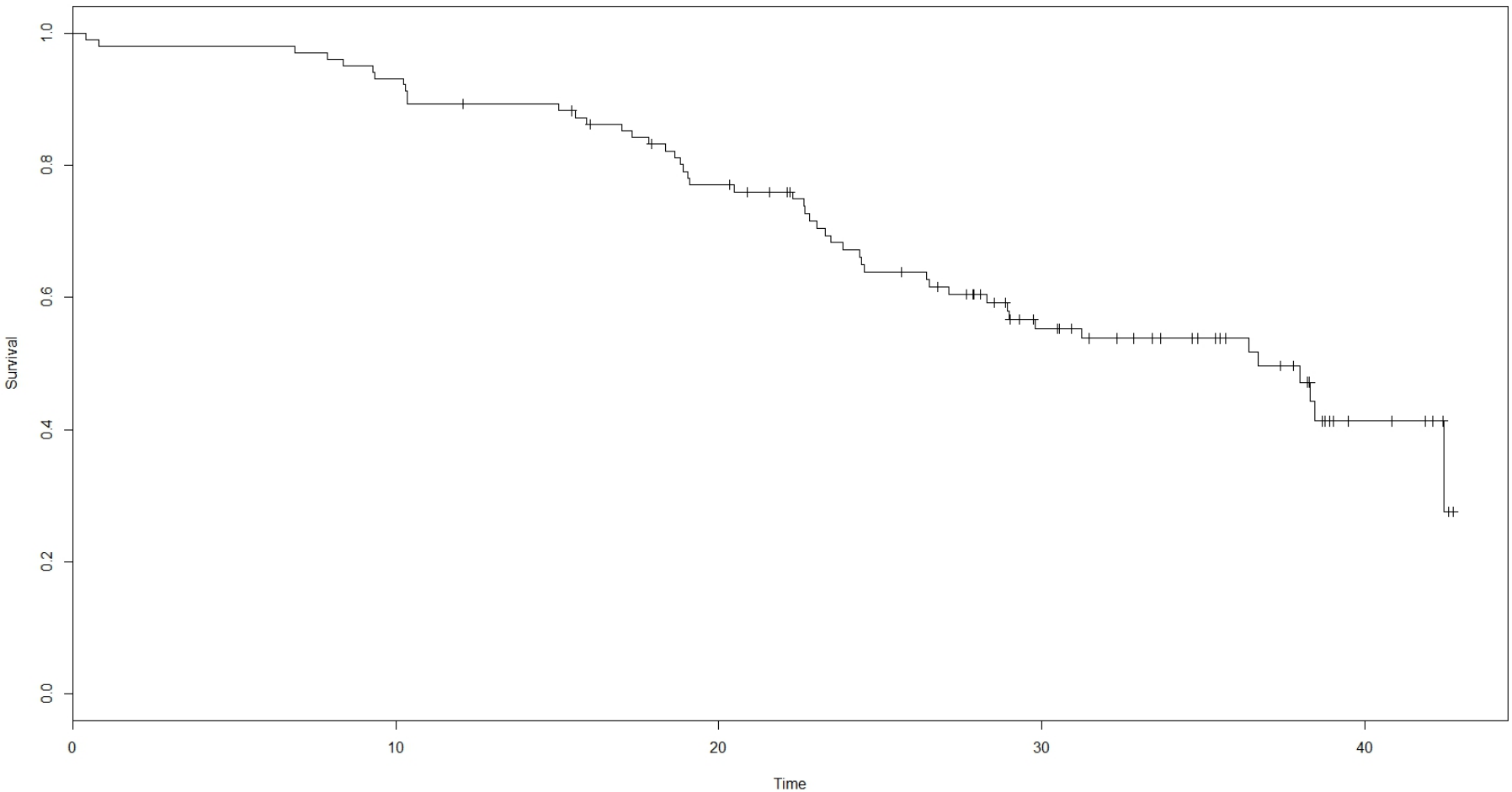
- If $T$ has a lognormal distribution, we have that

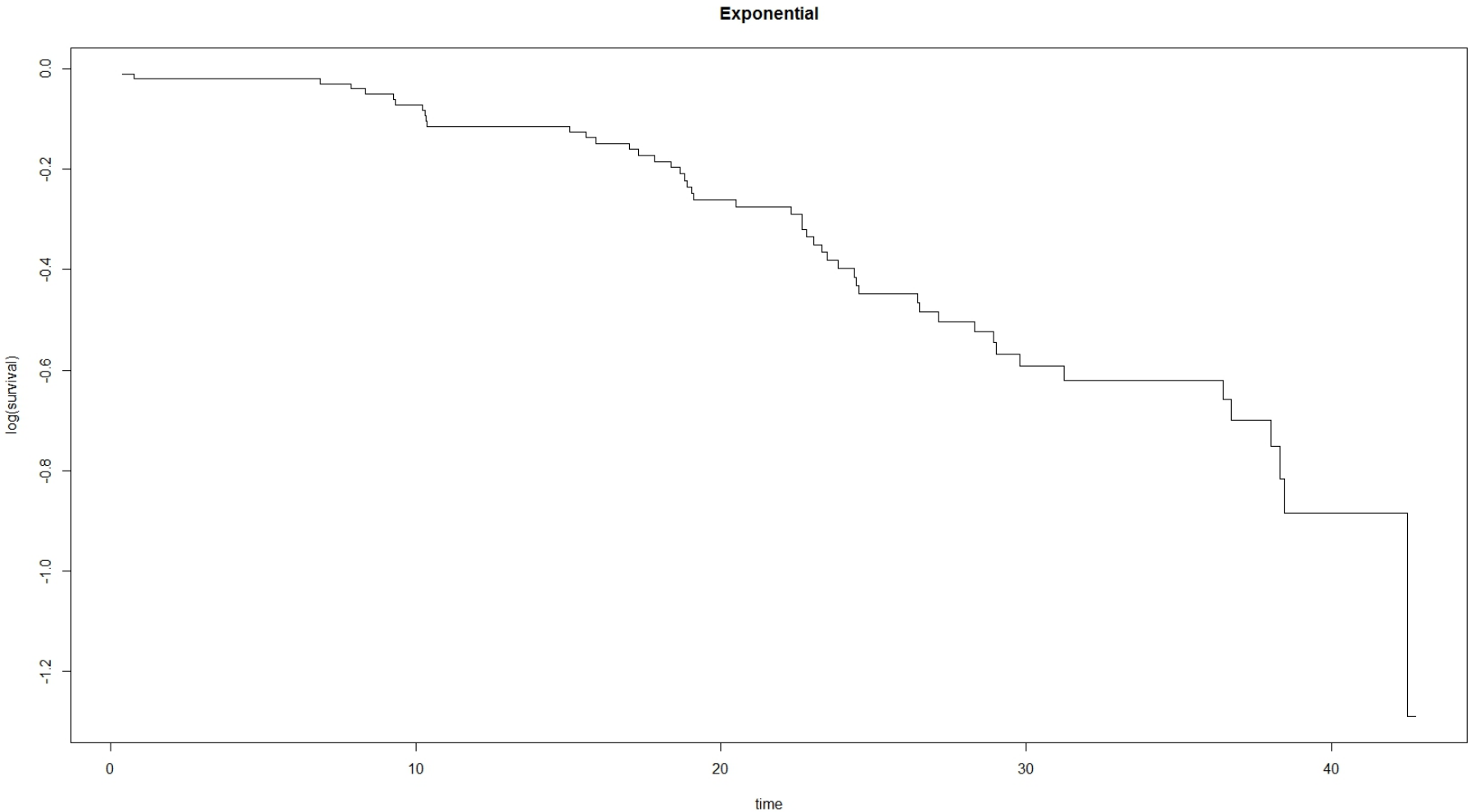
$$\begin{aligned} S(t) &= 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right) \\ \Phi^{-1}(1-S(t)) &= \frac{\log(t) - \mu}{\sigma}. \end{aligned}$$

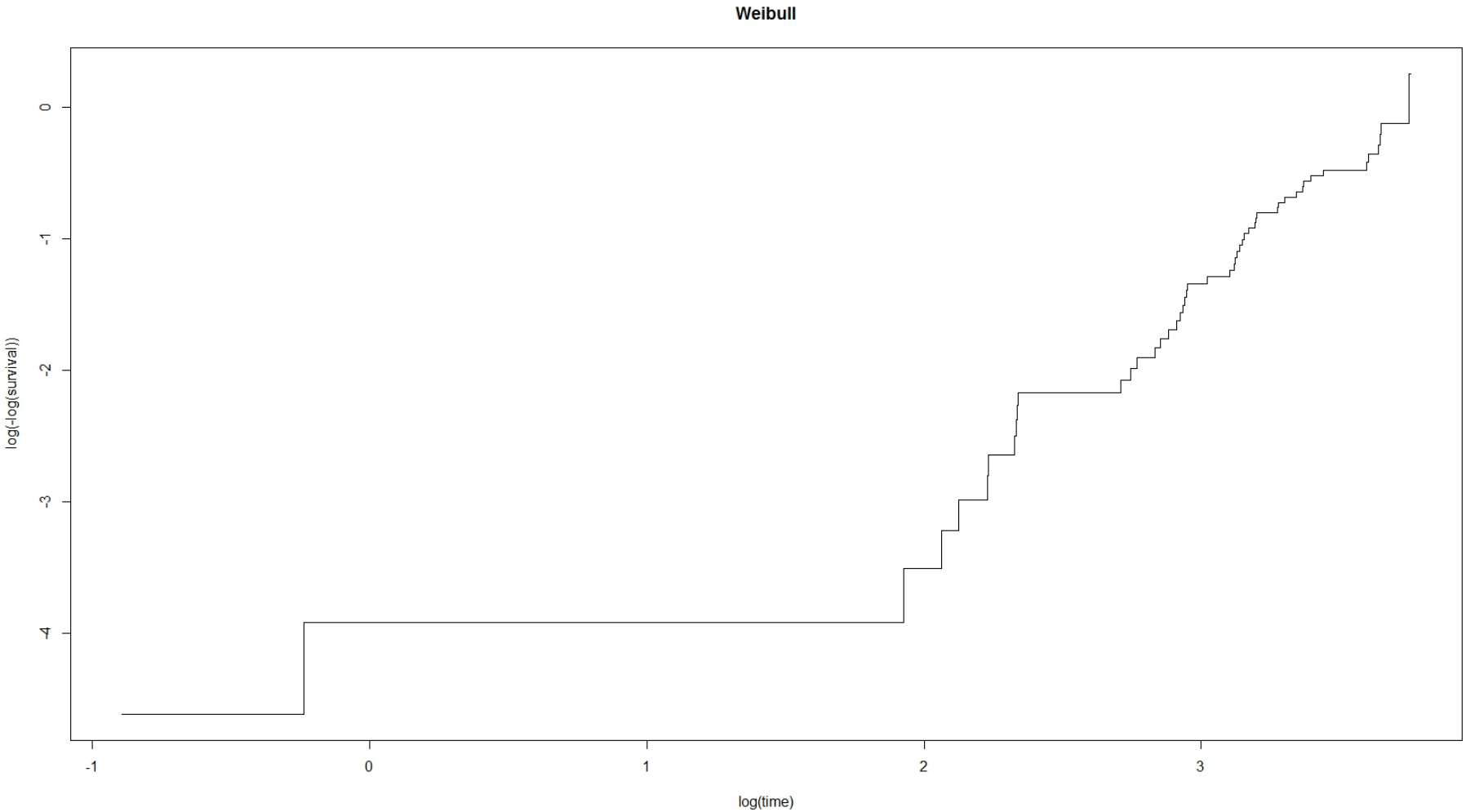
$\Rightarrow$ Plot $\Phi^{-1}(1-\hat{S}(t))$ vs $\log(t)$.

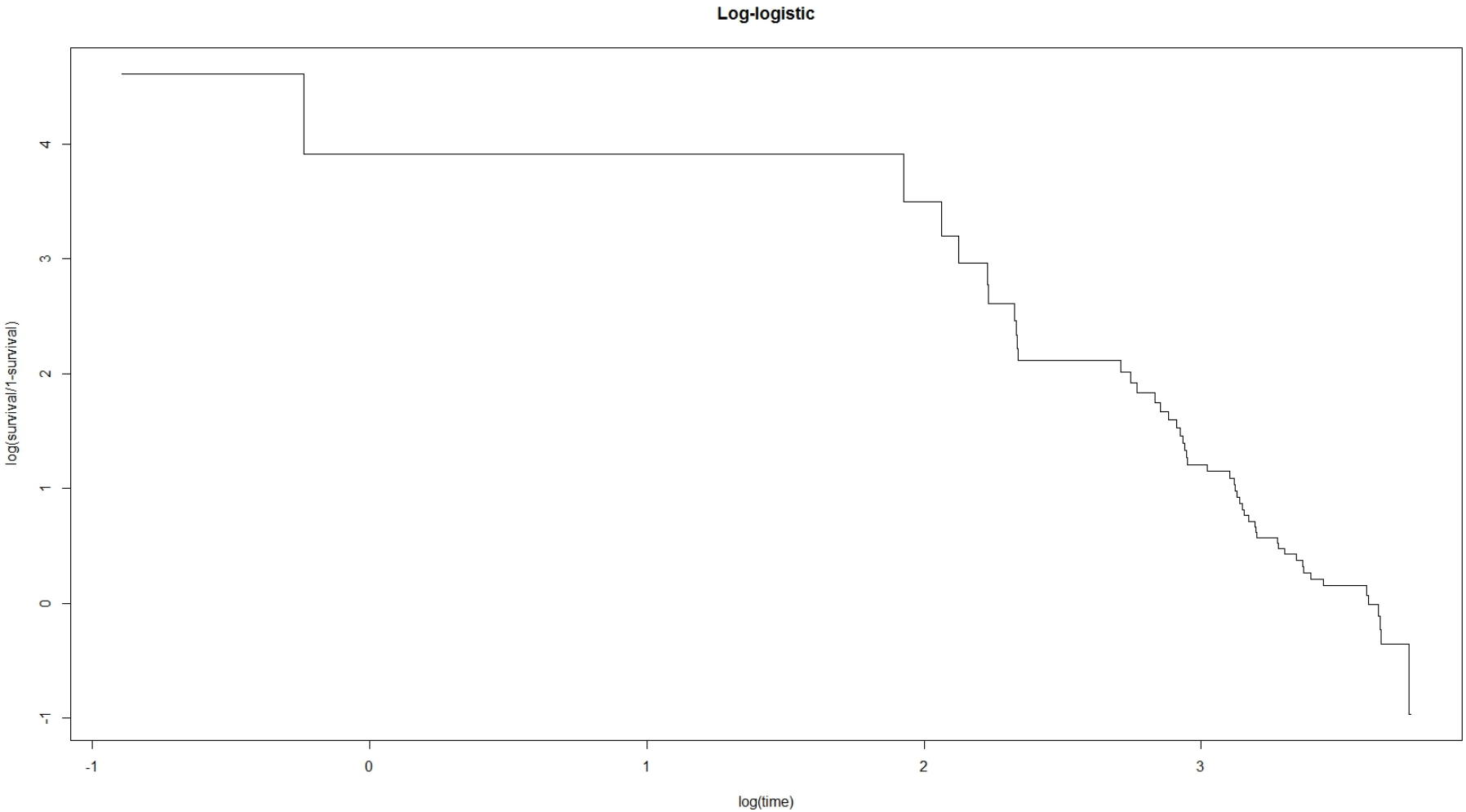
# Example: NSCLC

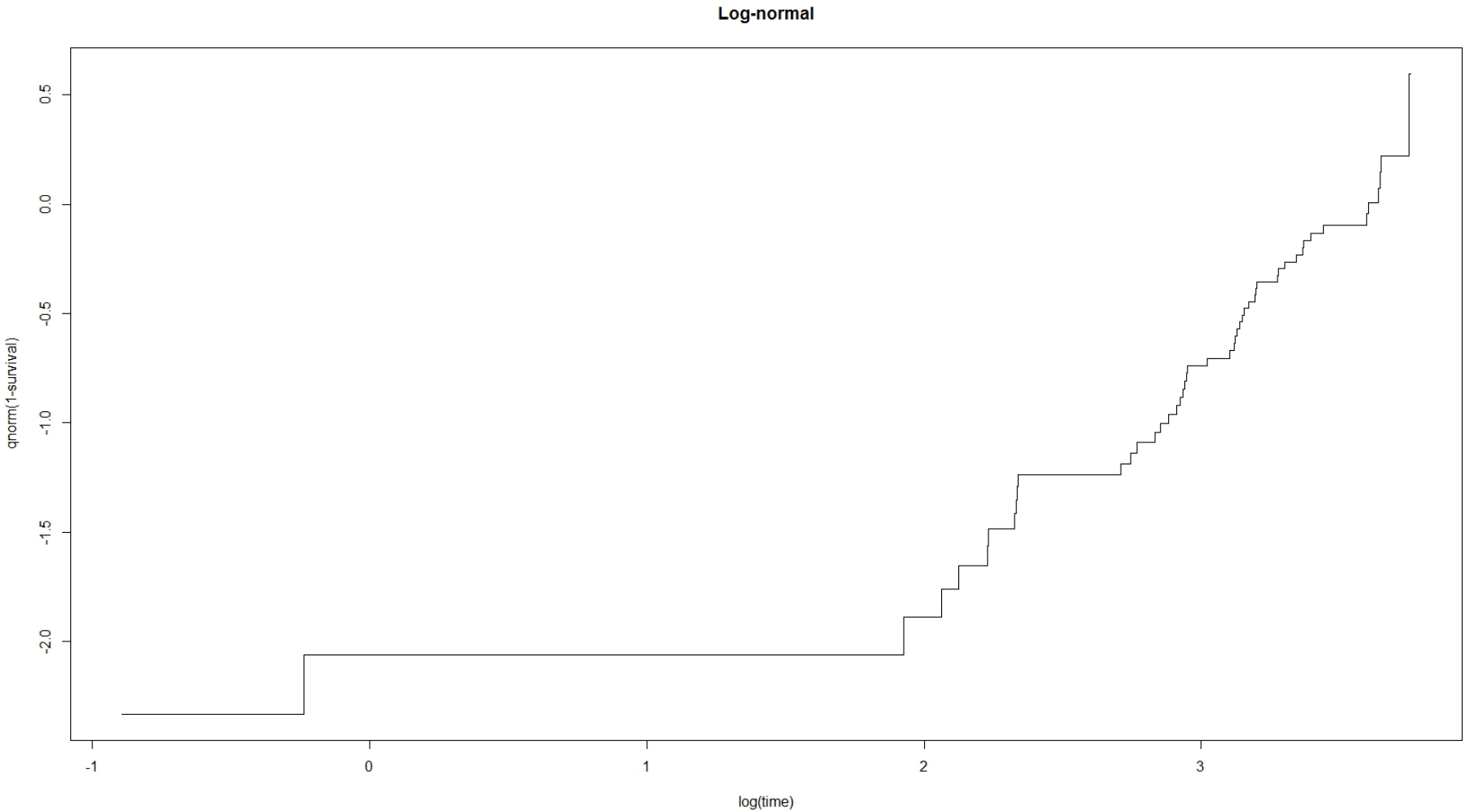
```
NSCLC<-read.table("C:/werk/Roel/Onderwijs/Theorie/G0B67AStatAnalReliaSurvData/Cursus/NSCLC.txt",
header=T,sep="\t")
fit<-survfit(Surv(survtime,survind)~1,data=NSCLC)
summary(fit)
plot(survfit(Surv(survtime,survind)~1,conf.type="none",data=NSCLC),xlab="Time",ylab="Survival")

#Exponential
#--------------
plot(fit$time,log(fit$surv),type="s",xlab="time",ylab="log(survival)",main="Exponential")

#Weibull
#--------------
plot(log(fit$time),log(-log(fit$surv)),type="s",xlab="log(time)",ylab="log(-log(survival))",
main="Weibull")

#log-logistic
#--------------
plot(log(fit$time),log(fit$surv/(1-fit$surv)),type="s",xlab="log(time)",
ylab="log(survival/1-survival)",main="Log-logistic")

#log-normal
#--------------
plot(log(fit$time),qnorm(1-fit$surv),type="s",xlab="log(time)",ylab="qnorm(1-survival)",
main="Log-normal")
```


Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
Time


Exponential
log(survival)
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
0
10
20
30
40
time


Weibull
log(-log(survival))
0
-1
-2
-3
-4
-1
0
1
2
3
log(time)


Log-logistic
log(survival/1-survival)
-1
0
1
2
3
4
log(time)
-1
0
1
2
3


Log-normal
qnorm(1-survival)
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-1
0
1
2
3
log(time)

In the previous section, we looked at commonly used parametric distributions for the response lifetime $T$.

However, in most situations, we are also interested in investigating the influence of other random variables, covariates, on the response lifetime.

In survival analysis, we can discriminate between two different approaches:

- the proportional hazards models
- the accelerated failure time models

In the proportional hazards models, we study the influence of covariates through the conditional hazard function.

We assume that the conditional hazard function of the lifetime $T$ is given by

$$\lambda(t|\mathbf{X}) = \lambda_0(t)e^{\beta^t \mathbf{X}}.$$

Hereby,

- we call $\lambda_0(t)$ the baseline hazard function.
  It represent the hazard function when the covariates are zero.

- the influence of the covariates is multiplicative through the $e^{\beta^t \mathbf{X}}$-term.

These models are called proportional hazards model because the ratio for two different individuals is constant.

$$\frac{\lambda(t|\mathbf{X})}{\lambda(t|\mathbf{X}')} = \frac{\lambda_0(t)e^{\beta^t\mathbf{X}}}{\lambda_0(t)e^{\beta^t\mathbf{X}'}} = e^{\beta^t(\mathbf{X}-\mathbf{X}')}.$$

In an analysis, we specify for the baseline hazard function $\lambda_0(t)$ always a parametric hazard function.

Using the relationship a hazard and a survival function, we get that

$$\begin{aligned} S(t|\mathbf{X}) &= \exp\left(-\int_0^t \lambda_0(u)e^{\beta^t\mathbf{X}}du\right) = \exp(-\exp(\beta^t\mathbf{X})\Lambda_0(t)) \\ &= S_0(t)^{\exp(\beta^t\mathbf{X})} \end{aligned}$$

where $S_0(t)$ is the baseline survival function.

An alternative regression model in survival data is the Accelerated failure time model.

Hereby we can consider on a log-scale that

$$\begin{aligned} Y = \log(T) &= \mu + \beta_1 X_1 + \ldots + \beta_p X_p + \sigma W \\ &= \beta^t \mathbf{X} + \sigma W \end{aligned}$$

with $W \sim F$ a parametric error distribution.

We note that this looks like the situation which we have in a standard generalized linear regression model.

When we denote by $S_0$ the survival function when $\mathbf{X} = 0$ then we find that

$$\begin{aligned} P(T > t|\mathbf{X}) &= P(Y > \log(t)|\mathbf{X}) \\ &= P(\mu + \sigma W > \log(t) - \beta^t \mathbf{X}|\mathbf{X}) \\ &= P(\exp(\mu + \sigma W) > t \exp(-\beta^t \mathbf{X})|\mathbf{X}) \\ &= S_0(t \exp(-\beta^t \mathbf{X})) \end{aligned}$$

The effect of the covariates on the survival function is that the time scale is changed by a factor $\exp(-\beta^t \mathbf{X})$.

We call this an acceleration factor.

We note that when

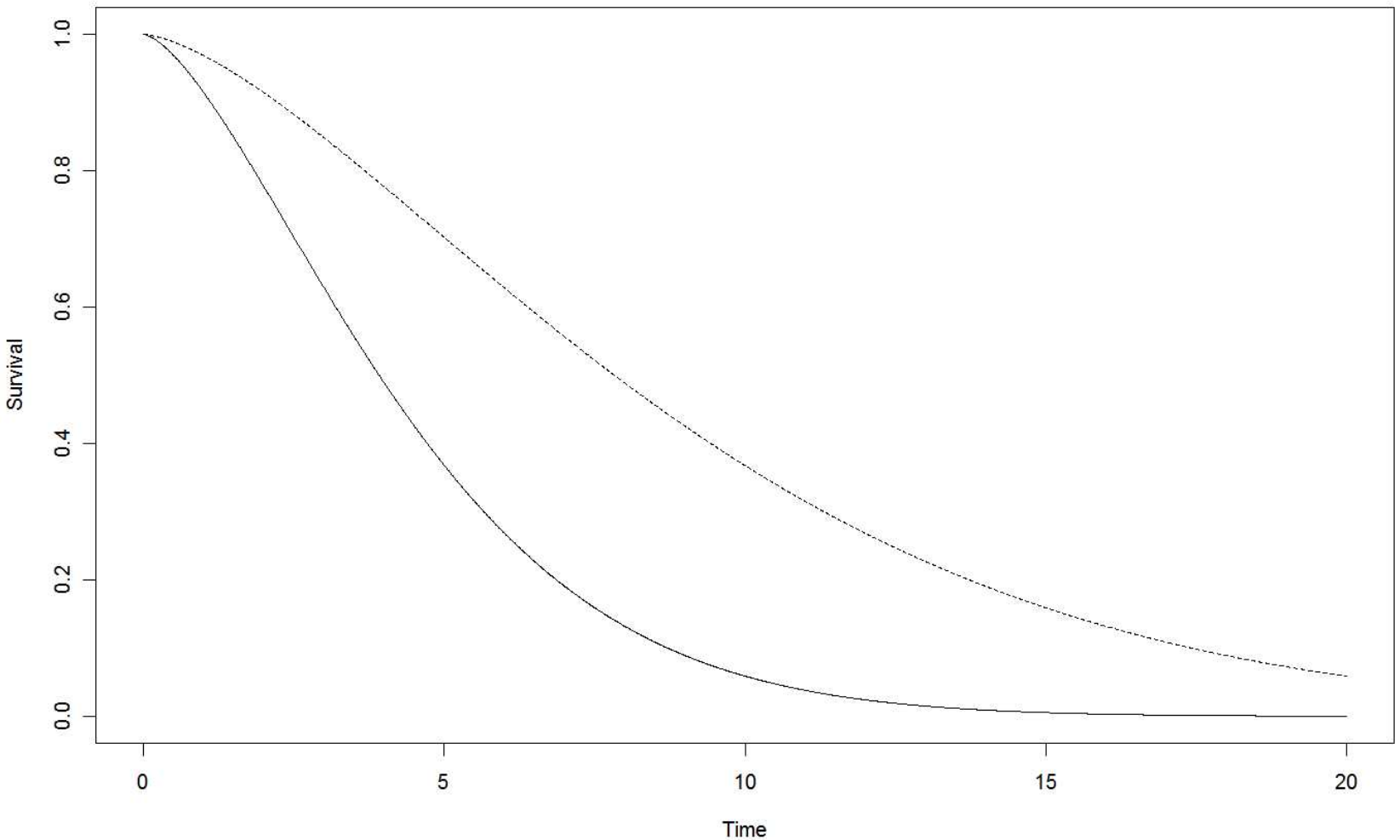
$$\exp(-\beta^t \mathbf{X}) > 1 \Rightarrow \text{the survival process accelerates.}$$
$$\exp(-\beta^t \mathbf{X}) < 1 \Rightarrow \text{the survival process decelerates.}$$

If $\mathbf{X}$ is an indicator variable, this is equivalent to

$$\beta > 0 \Rightarrow \text{Time shrinks.}$$
$$\beta < 0 \Rightarrow \text{Time accelerates.}$$

The hazard function is in this case is given by

$$\lambda(t|\mathbf{X}) \quad = \quad \lambda_0(te^{-\beta^t \mathbf{X}})e^{-\beta^t \mathbf{X}}$$


Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Time

In each distribution we have (several) parameter(s) which determined the shape of this distribution.

How will we estimate the parameters of the models in a practical data set?

In survival analysis, some observations are **censored**!

Hence, estimation method has to be adapted to censoring.

In parametric modeling, Maximum Likelihood Estimation is commonly used.

# No censoring

Let $T_1, \ldots, T_n$ be a sample from a population $T \sim F(t, \theta)$.
$F(t, \theta)$ a continuous distribution with density $f(t, \theta)$.

The maximum likelihood function is defined as

$$L(\theta) = \prod_{i=1}^{n} f(t_i, \theta).$$

We estimate $\theta$ by maximizing this expression ($\hat{\theta}$).

**Example:** $T_1, \ldots, T_n$ with $T_i \sim \text{Exp}(\lambda)$.

$$L(\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda t_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} t_i\right).$$

To find an estimate for $\lambda$, we calculate

$$l(\lambda) = n \log(\lambda) - \lambda \sum_{i=1}^{n} t_i.$$

and solve the equation

$$U(\lambda) = \frac{d}{d\lambda} l(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0 \qquad \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\bar{t}}.$$

To get a maximum, we check that

$$\frac{d^2}{d\lambda^2} l(\hat{\lambda}) = \frac{-n}{\hat{\lambda}^2} < 0.$$

# With censoring

Suppose we have a censored sample $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ where

$$Y_i = \min(T_i, C_i) \quad \text{and} \quad \delta_i = I(T_i \leq C_i), \quad i = 1, \ldots, n$$

with

- a sample $T_1, \ldots, T_n \sim f(t, \theta)$ lifetimes. We denote the survival function by $S(t, \theta)$.
- a sample $C_1, \ldots, C_n \sim g(c)$ censoring times with survival function $G(t)$.
- $T_i$ and $C_i$ are independent.

**Note:** We assume non-informative censoring.

For an uncensored observation ($\delta = 1$), we calculate the contribution to the likelihood by

$$\begin{aligned} P(Y \leq y, \delta = 1) &= P(\min(T, C) \leq y, T \leq C) \\ &= P(T \leq y, C \geq T) \\ &= \int_0^y C(t) f(t, \theta) dt \\ \Rightarrow f_{Y,\delta=1}(y, \theta) &= f(y, \theta) G(y). \end{aligned}$$

For a censored observation ($\delta = 0$), we get similarly

$$f_{Y,\delta=0}(y, \theta) = g(y) S(y, \theta).$$

Hence, we get the likelihood function

$$L(\theta) = \prod_{i,\delta_i=1} f(y_i,\theta)G(y_i) \times \prod_{i,\delta_i=0} g(y_i)S(y_i,\theta).$$

Since we assumed that the censoring is non-informative, we remove the terms $g$ and $G$. So, we get

$$\begin{aligned} L(\theta) &= \prod_{i,\delta_i=1} f(y_i,\theta) \times \prod_{i,\delta_i=0} S(y_i,\theta) \\ &= \prod_{i=1}^{n} f(y_i,\theta)^{\delta_i} S(y_i,\theta)^{1-\delta_i}. \\ &= \prod_{i=1}^{n} \lambda(y_i,\theta)^{\delta_i} S(y_i,\theta). \end{aligned}$$

Returning to the example $T_1, \ldots, T_n$ with $T_i \sim \text{Exp}(\lambda)$, we get

$$L(\lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} \exp(-\lambda y_i)$$

and corresponding log-likelihood function

$$l(\lambda) = \sum_{i=1}^{n} \delta_i \log(\lambda) - \lambda \sum_{i=1}^{n} y_i.$$

Furthermore we get

$$U(\lambda) = \frac{\sum_{i=1}^{n} \delta_i}{\lambda} - \sum_{i=1}^{n} y_i = 0.$$

Solving this equation, we get

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} y_i}$$

This is a maximum since

$$\frac{d^2}{d\lambda^2} l(\hat{\lambda}) = -\frac{\sum_{i=1}^{n} \delta_i}{\hat{\lambda}^2} < 0.$$

Finding a analytic solution is the exception rather than the rule!

Assume that the lifetimes are Weibull with $\lambda, \gamma > 0$. We get

$$
\begin{aligned}
L(\lambda, \gamma) &= \prod_{i=1}^{n} \left(\lambda\gamma y_i^{\gamma-1}\right)^{\delta_i} \exp(-\lambda y_i^{\gamma}) \\
l(\lambda, \gamma) &= \sum_{i=1}^{n} \delta_i \left[\log(\lambda) + \log(\gamma) + (\gamma - 1)\log(y_i)\right] - \lambda \sum_{i=1}^{n} y_i^{\gamma} \\
\frac{\partial}{\partial\lambda} l(\lambda, \gamma) &= \sum_{i=1}^{n} \frac{\delta_i}{\lambda} - \sum_{i=1}^{n} y_i^{\gamma} = 0 \Rightarrow \hat{\lambda} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} y_i^{\gamma} \\
\frac{\partial}{\partial\gamma} l(\lambda, \gamma) &= \sum_{i=1}^{n} \frac{\delta_i}{\gamma} + \sum_{i=1}^{n} \delta_i \log(y_i) - \lambda \sum_{i=1}^{n} y_i^{\gamma} \log(y_i) = 0
\end{aligned}
$$

# Regression models

Until now, we only looked at a (censored) failure time $T > 0$.

Often there are covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ in a study.

**Examples:**

- Different treatments in cancer or AIDS -trialS.
- Gender-effect in sociology studies.
- Temperature, pressure in industrial trials.
- ...

**How do these covariates influence $T$?**

When conditional on the covariates, the censoring $C$ is non-informative for the lifetime $T$,

we get a similar expression for the likelihood

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(t_i|\mathbf{x}_i,\theta)^{\delta_i} S(t_i|\mathbf{x}_i,\theta)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \lambda(t_i|\mathbf{x}_i,\theta)^{\delta_i} S(t_i|\mathbf{x}_i,\theta). \end{aligned}$$

We get estimate for the different parameter when we maximize this function.

# Asymptotic results

The asymptotic results of the maximum likelihood function remain in general valid.

- MLE $\hat{\theta}$ asymptotic normal
- Likelihood ratio test $\chi^2$ - distributed
- Score statistics
- ...

Next we give some examples.

The simplest regression model is the Exponential proportional hazard model.

$$\lambda(t|\mathbf{X}) = \lambda(\mathbf{X})$$

Given $\mathbf{X}$, the hazard function is constant.

Mostly, the influence is described multiplicative as

$$\lambda(t|\mathbf{X}) = \lambda c(\beta^t \mathbf{X})$$

where

$\lambda$ a constant
$\beta^t = (\beta_1, \ldots, \beta_p)$ regression parameters
$c$ is a known positive function
(Commonly: $c(s) = \exp(s)$)

We note that when in this model

$$\lambda(t|\mathbf{X}) = \lambda \exp(\beta^t \mathbf{X}),$$

the conditional distribution of $T$, given $\mathbf{X}$, is

$$\begin{aligned} f(t|\mathbf{X}) &= \lambda \exp(\beta^t \mathbf{X}) \exp[-\lambda t \exp(\beta^t \mathbf{X})] \\ S(t|\mathbf{X}) &= \exp[-\lambda t \exp(\beta^t \mathbf{X})]. \end{aligned}$$

This is a proportional hazards models since

$$\lambda(t|\mathbf{X}) = \lambda_0(t) e^{\beta^t \mathbf{X}} = \lambda \exp(\beta^t \mathbf{X}).$$

We can also write this models as an accelerated failure model.

$$\begin{aligned}\lambda(t|\mathbf{X}) &= \lambda_0(te^{-\beta_{aft}^t\mathbf{X}})e^{-\beta_{aft}^t\mathbf{X}} \\ &= \lambda\exp(\beta^t\mathbf{X})\end{aligned}$$

where the parameter estimates are related through

$$\beta = -\beta_{aft}.$$

Using the log-scale notation,

$$Y = \log(T) = \mu + \beta_{aft}^t\mathbf{X} + E$$

with $E$ an extreme value distribution, we get that

$$\lambda = \exp(-\mu) \quad \text{and} \quad \beta = -\beta_{aft}.$$

A generalization of the exponential model is the Weibull proportional hazard model.

Hereto we assume that

$$\lambda(t|\mathbf{X}) = \alpha\lambda t^{\alpha-1}\exp(\beta^t\mathbf{X}).$$

The conditional distribution of $T$, given $\mathbf{X}$, is now

$$\begin{aligned} f(t|\mathbf{X}) &= \lambda\alpha\lambda t^{\alpha-1}\exp(\beta^t\mathbf{X})\exp[-\lambda t^{\alpha}\exp(\beta^t\mathbf{X})] \\ S(t|\mathbf{X}) &= \exp[-\lambda t^{\alpha}\exp(\beta^t\mathbf{X})]. \end{aligned}$$

Also this model can be rewritten as accelerated failure model.

$$\begin{aligned} S(t|\mathbf{X}) &= S_0(t \exp(-\beta_{aft}^t \mathbf{X})) \\ &= \exp(-\lambda t^\alpha \exp(-\alpha \beta_{aft}^t \mathbf{X})) \end{aligned}$$

where the parameters $\lambda$ and $\alpha$ are the same and the other estimates are related through

$$\beta = -\alpha \beta_{aft}.$$

Using the log-scale notation, we could estimate the parameters of this model

$$Y = \log(T) = \mu + \beta_{aft}^t \mathbf{X} + \sigma E$$

with $E$ an extreme value distribution, we get that

$$\lambda = \exp\left(-\frac{\mu}{\sigma}\right), \quad \alpha = \frac{1}{\sigma} \quad \text{and} \quad \beta = -\frac{\beta_{aft}}{\sigma}.$$

The standard errors of these parameters can be estimated by the delta-method.

# Example: NSCLC

- Laudanski et al., Eur Respir J (2001).
- In this study, we had 102 patients who were operated from long cancer.
- The severity of the cancer was expressed in three TNM (Tumor, Nodes, Metastasis) categories: I, II, IIIa.
- The expression of the P53 protein was found from tumor biopsies.
- We are interested on the effect of this protein on the survival time of a patient.

```
> fit<-survreg(Surv(survtime,survind)~expres,data=NSCLC)
> summary(fit)

Call: survreg(formula = Surv(survtime, survind) ~ expres, data = NSCLC)
             Value Std. Error     z         p
(Intercept)  4.014      0.154 26.13 1.80e-150
expres      -0.450      0.177 -2.54  1.11e-02
Log(scale)  -0.553      0.127 -4.34  1.40e-05

Scale= 0.575

Weibull distribution
Loglik(model)= -235.8   Loglik(intercept only)= -239.3
Chisq= 7.12 on 1 degrees of freedom, p= 0.0076
Number of Newton-Raphson Iterations: 5
n= 102


> fit<-survreg(Surv(survtime,survind)~expres,dist="exponential",data=NSCLC)
> summary(fit)

Call: survreg(formula = Surv(survtime, survind) ~ expres, data = NSCLC,
dist = "exponential")
             Value Std. Error     z        p
(Intercept)  4.405      0.243 18.16 9.99e-74
expres      -0.673      0.300 -2.24 2.49e-02

Scale fixed at 1

Exponential distribution
Loglik(model)= -243.3   Loglik(intercept only)= -246
Chisq= 5.29 on 1 degrees of freedom, p= 0.021
Number of Newton-Raphson Iterations: 4
n= 102
```

```
> fit<-survreg(Surv(survtime,survind)~expres,dist="lognormal",data=NSCLC)
> summary(fit)

Call: survreg(formula = Surv(survtime, survind) ~ expres, data = NSCLC,
dist = "lognormal")
              Value Std. Error      z        p
(Intercept)  3.9573      0.205 19.263 1.10e-82
expres      -0.5417      0.249 -2.175 2.96e-02
Log(scale)   0.0771      0.108  0.712 4.76e-01

Scale= 1.08

Log Normal distribution
Loglik(model)= -244   Loglik(intercept only)= -246.4
Chisq= 4.83 on 1 degrees of freedom, p= 0.028
Number of Newton-Raphson Iterations: 3
n= 102


> fit<-survreg(Surv(survtime,survind)~expres,dist="loglogistic",data=NSCLC)
> summary(fit)

Call: survreg(formula = Surv(survtime, survind) ~ expres, data = NSCLC,
dist = "loglogistic")
             Value Std. Error     z         p
(Intercept)  3.837      0.154 24.84 3.25e-136
expres      -0.477      0.190 -2.52  1.19e-02
Log(scale)  -0.713      0.126 -5.66  1.54e-08

Scale= 0.49

Log logistic distribution
Loglik(model)= -237.1   Loglik(intercept only)= -240.4
Chisq= 6.52 on 1 degrees of freedom, p= 0.011
Number of Newton-Raphson Iterations: 4
n= 102
```

```
> fit<-survreg(Surv(survtime,survind)~expres+tnm,data=NSCLC)
> summary(fit)

Call: survreg(formula = Surv(survtime, survind) ~ expres + tnm, data = NSCLC)
             Value Std. Error     z        p
(Intercept)  5.174      0.383 13.51 1.45e-41
expres      -0.356      0.158 -2.25 2.42e-02
tnm         -0.517      0.132 -3.93 8.54e-05
Log(scale)  -0.664      0.126 -5.27 1.38e-07

Scale= 0.515

Weibull distribution
Loglik(model)= -224.8   Loglik(intercept only)= -239.3
Chisq= 29.09 on 2 degrees of freedom, p= 4.8e-07
Number of Newton-Raphson Iterations: 6
n= 102

> fit<-survreg(Surv(survtime,survind)~expres*tnm,data=NSCLC)
> summary(fit)

Call:survreg(formula = Surv(survtime, survind) ~ expres * tnm, data = NSCLC)
             Value Std. Error     z        p
(Intercept)  6.241      0.870  7.18 7.09e-13
expres      -1.844      0.923 -2.00 4.57e-02
tnm         -0.908      0.301 -3.02 2.53e-03
expres:tnm   0.552      0.326  1.69 9.05e-02
Log(scale)  -0.665      0.126 -5.27 1.35e-07

Scale= 0.514

Weibull distribution
Loglik(model)= -222.9   Loglik(intercept only)= -239.3
Chisq= 32.77 on 3 degrees of freedom, p= 3.6e-07
Number of Newton-Raphson Iterations: 6
n= 102
```

```
> fit<-survreg(Surv(survtime,survind)~expres+factor(tnm),data=NSCLC)
> summary(fit)

Call: survreg(formula = Surv(survtime, survind) ~ expres + factor(tnm),data = NSCLC)
                Value Std. Error     z        p
(Intercept)   4.4085      0.262 16.81 1.95e-63
expres       -0.3586      0.157 -2.29 2.22e-02
factor(tnm)2 -0.0509      0.299 -0.17 8.65e-01
factor(tnm)3 -0.8113      0.253 -3.20 1.35e-03
Log(scale)   -0.6734      0.126 -5.35 8.56e-08

Scale= 0.51

Weibull distribution
Loglik(model)= -223.5   Loglik(intercept only)= -239.3
Chisq= 31.67 on 3 degrees of freedom, p= 6.1e-07
Number of Newton-Raphson Iterations: 6
n= 102
```

We consider the accelerated failure time models with distributions: Weibull, exponential, generalized gamma and lognormal.

```
proc lifereg data=nsclc;
model Survtime*Survind(0)=expres;
run;

proc lifereg data=nsclc;
model Survtime*Survind(0)=expres/distribution=expontial;
run;

proc lifereg data=nsclc;
model Survtime*Survind(0)=expres/distribution=gamma;
run;
proc lifereg data=nsclc;
model Survtime*Survind(0)=expres/distribution=lognormal;
run;

proc lifereg data=nsclc;
model Survtime*Survind(0)=expres TNM expres*TNM;
run;

proc lifereg data=nsclc;
class TNM;
model Survtime*Survind(0)=expres TNM;
run;
```

The LIFEREG Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.NSCLC |
| Dependent Variable | Log(Survtime) |
| Censoring Variable | Survind |
| Censoring Value(s) | 0 |
| Number of Observations | 102 |
| Noncensored Values | 49 |
| Right Censored Values | 53 |
| Left Censored Values | 0 |
| Interval Censored Values | 0 |
| Name of Distribution | Weibull |
| Log Likelihood | -96.25063495 |

| | |
|---|---|
| Number of Observations Read | 102 |
| Number of Observations Used | 102 |

Algorithm converged.

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| expres | 1 | 6.4463 | 0.0111 |

Analysis of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 4.0138 | 0.1536 | 3.7127 | 4.3149 | 682.62 | <.0001 |
| expres | 1 | -0.4502 | 0.1773 | -0.7978 | -0.1027 | 6.45 | 0.0111 |
| Scale | 1 | 0.5753 | 0.0732 | 0.4482 | 0.7383 | | |
| Weibull Shape | 1 | 1.7383 | 0.2213 | 1.3545 | 2.2310 | | |

# The parameter Weibull Shape is the inverse of Scale.

| | |
|---|---|
| Name of Distribution | Exponential |
| Log Likelihood | -103.8227604 |

Analysis of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 4.4054 | 0.2425 | 3.9300 | 4.8807 | 329.92 | <.0001 |
| expres | 1 | -0.6730 | 0.3001 | -1.2612 | -0.0848 | 5.03 | 0.0249 |
| Scale | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |
| Weibull Shape | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

Lagrange Multiplier Statistics

| Parameter | Chi-Square | Pr > ChiSq |
|---|---|---|
| Scale | 99.7168 | <.0001 |

Name of Distribution Gamma
Log Likelihood -95.98435044

WARNING: Iteration limit exceeded.

Analysis of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 4.1238 | 0.1509 | 3.8280 | 4.4195 | 746.65 | <.0001 |
| expres | 1 | -0.3732 | 0.1602 | -0.6872 | -0.0593 | 5.43 | 0.0198 |
| Scale | 1 | 0.2492 | 0.0336 | 0.1914 | 0.3245 | | |
| Shape | 0 | 2.8349 | 0.0000 | 2.8349 | 2.8349 | | |

The parameter Shape is an extra parameter.

Name of Distribution Lognormal
Log Likelihood -104.473411

Analysis of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 3.9573 | 0.2054 | 3.5547 | 4.3600 | 371.06 | <.0001 |
| expres | 1 | -0.5417 | 0.2490 | -1.0298 | -0.0536 | 4.73 | 0.0296 |
| Scale | 1 | 1.0801 | 0.1168 | 0.8737 | 1.3352 | | |

| | |
|---|---|
| Name of Distribution | Weibull |
| Log Likelihood | -83.42497702 |

Algorithm converged.

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| expres | 1 | 3.9914 | 0.0457 |
| tnm | 1 | 9.1166 | 0.0025 |
| expres*tnm | 1 | 2.8662 | 0.0905 |

Analysis of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 6.2413 | 0.8695 | 4.5370 | 7.9455 | 51.52 | <.0001 |
| expres | 1 | -1.8437 | 0.9228 | -3.6524 | -0.0350 | 3.99 | 0.0457 |
| tnm | 1 | -0.9083 | 0.3008 | -1.4979 | -0.3187 | 9.12 | 0.0025 |
| expres*tnm | 1 | 0.5515 | 0.3258 | -0.0870 | 1.1900 | 2.87 | 0.0905 |
| Scale | 1 | 0.5145 | 0.0648 | 0.4019 | 0.6587 | | |
| Weibull Shape | 1 | 1.9437 | 0.2450 | 1.5182 | 2.4884 | | |

| | |
|---|---|
| Name of Distribution | Weibull |
| Log Likelihood | -83.97589782 |

Algorithm converged.

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| expres | 1 | 5.2336 | 0.0222 |
| tnm | 2 | 18.2350 | 0.0001 |

Analysis of Parameter Estimates

| Parameter | | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|
| Intercept | | 1 | 3.5972 | 0.1357 | 3.3313 | 3.8631 | 702.92 | <.0001 |
| expres | | 1 | -0.3586 | 0.1567 | -0.6658 | -0.0514 | 5.23 | 0.0222 |
| tnm | 1 | 1 | 0.8113 | 0.2532 | 0.3150 | 1.3075 | 10.27 | 0.0014 |
| tnm | 2 | 1 | 0.7603 | 0.2212 | 0.3268 | 1.1939 | 11.81 | 0.0006 |
| tnm | 3 | 0 | 0.0000 | . | . | . | . | . |
| Scale | | 1 | 0.5100 | 0.0641 | 0.3986 | 0.6525 | | |
| Weibull Shape | | 1 | 1.9610 | 0.2466 | 1.5326 | 2.5091 | | |

We saw earlier that a Weibull (or Exponential) proportional hazard model and accelerated failure time model are actually the same model, but with a transformation of the parameters.

$$\begin{aligned} \lambda(t|\mathbf{X}) &= \alpha\lambda t^{\alpha-1}\exp(\beta^t\mathbf{X}) \\ &\Rightarrow \quad S(t|\mathbf{X}) = \exp\left(-\lambda t^{\alpha} e^{\beta^t\mathbf{X}}\right) \end{aligned}$$

$$\begin{aligned} Y = \log(T) &= \mu + \beta_{aft}^t\mathbf{X} + \sigma E \\ &\Rightarrow \quad S(t|\mathbf{X}) = \exp\left(-t^{\frac{1}{\sigma}} e^{-\frac{\mu}{\sigma}} e^{-\frac{\beta_{aft}^t}{\sigma}\mathbf{X}}\right) \end{aligned}$$

with $E$ an extreme value distribution, we get that

$$\lambda = \exp\left(-\frac{\mu}{\sigma}\right), \quad \alpha = \frac{1}{\sigma} \quad \text{and} \quad \beta = -\frac{\beta_{aft}}{\sigma}.$$

Using the output of the accelerated failure model, we derive estimates for the parametric proportional hazards model by the delta-method,

$$\lambda = \exp\left(-\frac{\mu}{\sigma}\right) = \exp\left(-\frac{\mu}{\exp(l\sigma)}\right) = \exp\left(-\mu\exp(l\sigma)\right)$$

$$\alpha = \frac{1}{\sigma} = \frac{1}{\exp(l\sigma)} = \exp(-l\sigma)$$

$$\beta_i = -\frac{\beta_{i,aft}}{\sigma} = -\frac{\beta_{i,aft}}{\exp(l\sigma)} = -\beta_{i,aft}\exp(-l\sigma), i = 1, \ldots, m$$

```
> fit<-survreg(Surv(survtime,survind)~expres+factor(tnm),data=NSCLC)
> summary(fit)

Call: survreg(formula = Surv(survtime, survind) ~ expres + factor(tnm),data = NSCLC)
               Value Std. Error     z        p
(Intercept)   4.4085      0.262 16.81 1.95e-63
expres       -0.3586      0.157 -2.29 2.22e-02
factor(tnm)2 -0.0509      0.299 -0.17 8.65e-01
factor(tnm)3 -0.8113      0.253 -3.20 1.35e-03
Log(scale)   -0.6734      0.126 -5.35 8.56e-08

Scale= 0.51

Weibull distribution
Loglik(model)= -223.5   Loglik(intercept only)= -239.3
Chisq= 31.67 on 3 degrees of freedom, p= 6.1e-07
Number of Newton-Raphson Iterations: 6
n= 102


> para<-fit$coef
> lscale<-log(fit$scale)
> V<-fit$var
> para
(Intercept)       expres factor(tnm)2 factor(tnm)3
 4.40847577  -0.35859629  -0.05093227  -0.81126028
> lscale
[1] -0.6734365
> V
             (Intercept)        expres  factor(tnm)2  factor(tnm)3    Log(scale)
(Intercept)   0.06874934 -0.0161316354 -5.089700e-02 -0.0572241351  1.228617e-02
expres       -0.01613164  0.0245703252 -2.028463e-03 -0.0005509755 -4.008380e-03
factor(tnm)2 -0.05089700 -0.0020284634  8.933596e-02  0.0522539493 -1.198954e-05
factor(tnm)3 -0.05722414 -0.0005509755  5.225395e-02  0.0641076951 -8.665926e-03
Log(scale)    0.01228617 -0.0040083796 -1.198954e-05 -0.0086659265  1.581553e-02
```

```
> lambda<-exp(-para[1]*exp(-lscale))
> alpha<-exp(-lscale)
> beta<--para[-1]*exp(-lscale)
> x<-c(lambda,alpha,beta)
> names(x)[1]<-"lambda"
> names(x)[2]<-"alpha"

> m<-length(para[-1])
> G<-matrix(0,nrow=m+2,ncol=m+2)
> G[1,1]<--exp(-para[1]*exp(-lscale))*exp(-lscale)
> G[2:(m+1),3:(m+2)]<-diag(m)*(-exp(-lscale))
> G[m+2,1]<-exp(-para[1]*exp(-lscale))*para[1]*exp(-lscale)
> G[m+2,2]<--exp(-lscale)
> G[m+2,3:(m+2)]<-para[-1]*exp(-lscale)
> G
[,1]       [,2]       [,3]        [,4]       [,5]
[1,] -0.0003451855  0.000000  0.0000000  0.00000000  0.000000
[2,]  0.0000000000  0.000000 -1.9609647  0.00000000  0.000000
[3,]  0.0000000000  0.000000  0.0000000 -1.96096466  0.000000
[4,]  0.0000000000  0.000000  0.0000000  0.00000000 -1.960965
[5,]  0.0015217418 -1.960965 -0.7031946 -0.09987639 -1.590853

> PrVar<-t(G)%*%V%*%G
> PrVar
[,1]           [,2]           [,3]           [,4]           [,5]
[1,]  3.190825e-08 -3.887838e-05 -1.289976e-05 -3.639639e-05 -4.441547e-05
[2,] -3.887838e-05  6.081677e-02  6.394915e-03  3.051432e-03  1.601443e-02
[3,] -1.289976e-05  6.394915e-03  9.124817e-02 -7.491043e-03 -8.880563e-03
[4,] -3.639639e-05  3.051432e-03 -7.491043e-03  3.436840e-01  2.017147e-01
[5,] -4.441547e-05  1.601443e-02 -8.880563e-03  2.017147e-01  2.324762e-01
```

```
> PrStd<-sqrt(diag(PrVar))
> PrStd
[1] 0.0001786288 0.2466105616 0.3020731279 0.5862456807 0.4821578539

> PrChisq<-c(" "," ",(x[3:(m+2)]/PrStd[3:(m+2)])^2)
> PrPvalue<-c(" "," ",pchisq((x[3:(m+2)]/PrStd[3:(m+2)])^2,1,lower.tail=F))

> out<-data.frame(x,PrStd,PrChisq,PrPvalue)
> names(out)<-c("Estimate","StdError","Chisq","P-value")
> out
                Estimate     StdError              Chisq              P-value
lambda       0.0001760284 0.0001786288
alpha        1.9609646641 0.2466105616
expres       0.7031946467 0.3020731279   5.41909702045414   0.0199176582666829
factor(tnm)2 0.0998763852 0.5862456807 0.0290246050649941    0.864722237494367
factor(tnm)3 1.5908527380 0.4821578539   10.8863293368624 0.000968766257431309
```

For these models, we want to verify whether a specific distribution is satisfied. Hereto we generalize some previous ideas.

- If $T|\mathbf{X} \sim \text{Exp}(\lambda \exp(\beta^t \mathbf{X}))$, we have that

$$\begin{aligned} S(t|\mathbf{X}) &= \exp(-\lambda t \exp(\beta^t \mathbf{X})) \\ \log(S(t|\mathbf{X})) &= -\lambda \exp(\beta^t \mathbf{X}) t. \end{aligned}$$

$\Rightarrow$ Plot $\log(\hat{S}(t|\mathbf{X}))$ vs $t$.

- If $T|\mathbf{X} \sim \text{Weibull}(\lambda \exp(\beta^t \mathbf{X}), \alpha)$, we have that

$$\begin{aligned} S(t|\mathbf{X}) &= \exp(-\lambda t^\alpha \exp(\beta^t \mathbf{X})) \\ \log(S(t|\mathbf{X})) &= -\lambda t^\alpha \exp(\beta^t \mathbf{X}) \\ \log(-\log(S(t|\mathbf{X}))) &= \log(\lambda) + \beta^t \mathbf{X} + \alpha \log(t). \end{aligned}$$

$\Rightarrow$ Plot $\log(-\log(\hat{S}(t|\mathbf{X})))$ vs $\log(t)$.

Until now, we considered in the parametric models only right-censored data.

If we have a different censoring scheme, it is also possible to get results. The only change is the construction of the parametric likelihood function.

For example, if we have left-censored data (under non-informative censoring), the likelihood function has the following shape:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(t_i|\mathbf{x}_i,\theta)^{\delta_i} F(t_i|\mathbf{x}_i,\theta)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \{\lambda(t_i|\mathbf{x}_i,\theta) S(t_i|\mathbf{x}_i,\theta)\}^{\delta_i} \left(1 - S(t_i|\mathbf{x}_i,\theta)\right)^{1-\delta_i}. \end{aligned}$$

# Example: Tobit regression

Let us consider a sub-set of the Mroz(1987) data set.

| Hours | YrsEd | YrsExp |
|---|---|---|
| 0 | 8 | 9 |
| 0 | 8 | 12 |
| 0 | 9 | 10 |
| 0 | 10 | 15 |
| 0 | 11 | 4 |
| 0 | 11 | 6 |
| 1000 | 12 | 1 |
| 1960 | 12 | 29 |
| 0 | 13 | 3 |
| 2100 | 13 | 36 |
| 3686 | 14 | 11 |
| 1920 | 14 | 38 |
| 0 | 15 | 14 |
| 1728 | 16 | 3 |
| 1568 | 16 | 19 |
| 1316 | 17 | 7 |
| 0 | 17 | 15 |

Hereby we have that

- Hours = numbers of hours the wife worked outside the household in a given year
- YrsEd = numbers of years of education
- YrsExp = numbers of years of experience

We want to know how Hours depends on YrsEd and YrsExp:

$$\text{Hours} = \alpha + \beta_1 \text{YrsEd} + \beta_2 \text{YrsExp} + \varepsilon$$

with $\varepsilon \sim N(0, \sigma^2)$.

Since some women do not work outside the household, we observe only left-censored observations:

$$\text{Hours} = \begin{cases} \text{Hours}^* & , \text{Hours}^* > 0 \\ 0 & , \text{Hours}^* \leq 0 \end{cases} .$$

If we construct the likelihood function correctly, we can take this problem into account for a parametric regression.

```
data Mroz;
input Hours YrsEd YrsExp @@;
if Hours eq 0
 then Lower=.;
 else Lower=Hours;
datalines;
0 8 9 0 8 12 0 9 10 0 10 15 0 11 4 0 11 6
1000 12 1 1960 12 29 0 13 3 2100 13 36
3686 14 11 1920 14 38 0 15 14 1728 16 3
1568 16 19 1316 17 7 0 17 15
;
run;

proc lifereg data=Mroz;
model (lower, hours) = yrsed yrsexp / d=normal;
run;
```

The LIFEREG Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.MROZ |
| Dependent Variable | Lower |
| Dependent Variable | Hours |
| Number of Observations | 17 |
| Noncensored Values | 8 |
| Right Censored Values | 0 |
| Left Censored Values | 9 |
| Interval Censored Values | 0 |
| Name of Distribution | Normal |
| Log Likelihood | -74.9369977 |

| | |
|---|---|
| Number of Observations Read | 17 |
| Number of Observations Used | 17 |

Fit Statistics

| | |
|---|---|
| -2 Log Likelihood | 149.874 |
| AIC (smaller is better) | 157.874 |
| AICC (smaller is better) | 161.207 |
| BIC (smaller is better) | 161.207 |

Algorithm converged.

Analysis of Maximum Likelihood Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | -5598.64 | 2850.248 | -11185.0 | -12.2553 | 3.86 | 0.0495 |
| YrsEd | 1 | 373.1477 | 191.8872 | -2.9442 | 749.2397 | 3.78 | 0.0518 |
| YrsExp | 1 | 63.3371 | 38.3632 | -11.8533 | 138.5276 | 2.73 | 0.0987 |
| Scale | 1 | 1582.870 | 442.6732 | 914.9433 | 2738.397 | | |

```
> MROZ<-read.table("C:/werk/Roel/Onderwijs/Theorie/G0B67AStatAnalReliaSurvData/Cursus/MROZ.txt",
header=T,sep=";")

> start<-MROZ$Hours
> start[start==0]<--Inf
> stop<-MROZ$Hours

> Surv(start,stop,type="interval2")
[1]    0-    0-    0-    0-    0-    0- 1000  1960     0- 2100  3686  1920
[13]    0- 1728  1568  1316     0-

> MROZ1<-data.frame(MROZ,start,stop)
> fit<-survreg(Surv(start,stop,type="interval2")~YrsEd+YrsExp,dist="gaussian",data=MROZ1)
> summary(fit)

Call: survreg(formula = Surv(start, stop, type = "interval2") ~ YrsEd + YrsExp,
 data = MROZ1, dist = "gaussian")
               Value Std. Error     z         p
(Intercept) -5598.64    2850.25 -1.96  4.95e-02
YrsEd         373.15     191.89  1.94  5.18e-02
YrsExp         63.34      38.36  1.65  9.87e-02
Log(scale)      7.37       0.28 26.34 6.30e-153

Scale= 1583

Gaussian distribution
Loglik(model)= -74.9   Loglik(intercept only)= -78.6
Chisq= 7.33 on 2 degrees of freedom, p= 0.026
Number of Newton-Raphson Iterations: 4
n= 17
```