Survival Analysis

APTS 2016/17

Ingrid Van Keilegom
ORSTAT
KU Leuven

Glasgow, August 21-25, 2017

apts.ac.uk

Basic concepts

What is ‘Survival analysis’ ?

- Survival analysis (or duration analysis) is an area of statistics that models and studies the time until an event of interest takes place.
- In practice, for some subjects the event of interest cannot be observed for various reasons, e.g.
  - the event is not yet observed at the end of the study
  - another event takes place before the event of interest
  - ...
- In survival analysis the aim is
  - to model ‘time-to-event data’ in an appropriate way
  - to do correct inference taking these special features of the data into account.

Examples

- Medicine :
  - time to death for patients having a certain disease
  - time to getting cured from a certain disease
  - time to relapse of a certain disease
- Agriculture :
  - time until a farm experiences its first case of a certain disease
- Sociology (‘duration analysis’) :
  - time to find a new job after a period of unemployment
  - time until re-arrest after release from prison
- Engineering (‘reliability analysis’) :
  - time to the failure of a machine

Common functions in survival analysis

- Let T be a non-negative continuous random variable, representing the time until the event of interest.
- Denote

$F(t) = P(T \leq t)$	distribution function
$f(t)$	probability density function

- For survival data, we consider rather

$S(t)$	survival function
$H(t)$	cumulative hazard function
$h(t)$	hazard function
$mrl(t)$	mean residual life function

- Knowing one of these functions suffices to determine the other functions.

Survival function :

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

- ◇ Probability that a randomly selected individual will survive beyond time t
- ◇ Decreasing function, taking values in $[0, 1]$
- ◇ Equals 1 at $t = 0$ and 0 at $t = \infty$

Cumulative hazard function :

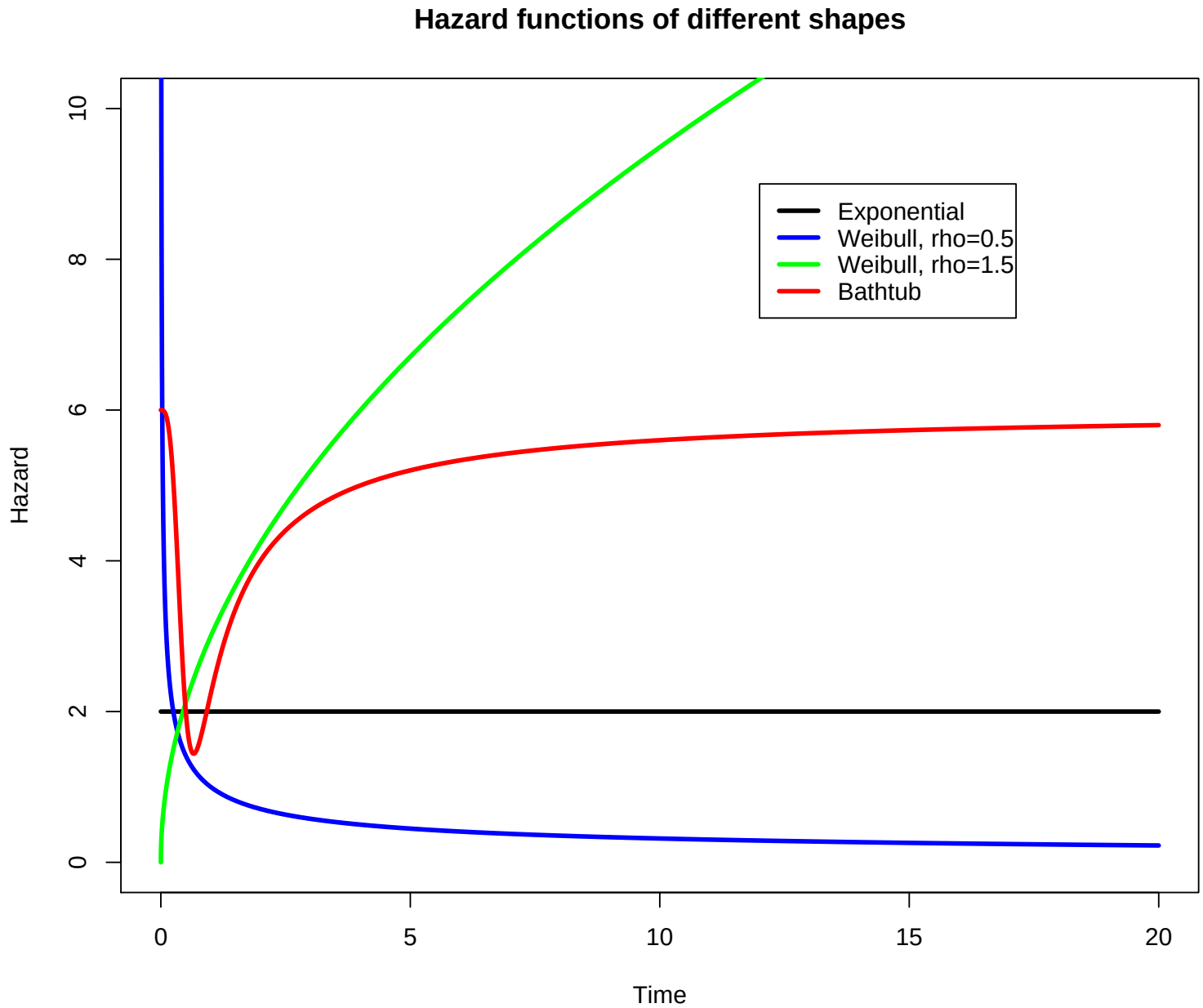
$$H(t) = -\log S(t)$$

- ◇ Increasing function, taking values in $[0, +\infty]$
- ◇ $S(t) = \exp(-H(t))$

Hazard function (or hazard rate) :

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \frac{1}{P(T \geq t)} \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \\ &= \frac{f(t)}{S(t)} = \frac{-d}{dt} \log S(t) = \frac{d}{dt} H(t) \end{aligned}$$

- $h(t)$ measures the instantaneous risk of dying right after time t given the individual is alive at time t
- Positive function (not necessarily increasing or decreasing)
- The hazard function $h(t)$ can have many different shapes and is therefore a useful tool to summarize survival data



Mean residual life function :

- The mrl function measures the expected remaining lifetime for an individual of age t. As a function of t, we have

$$\text{mrl}(t) = \frac{\int_t^\infty S(s)ds}{S(t)}$$

- This result is obtained from

$$\text{mrl}(t) = E(T - t \mid T > t) = \frac{\int_t^\infty (s-t)f(s)ds}{S(t)}$$

- Mean life time :

$$E(T) = \text{mrl}(0) = \int_0^\infty sf(s)ds = \int_0^\infty S(s)ds$$

Incomplete data

- Censoring :
  - For certain individuals under study, the time to the event of interest is only known to be within a certain interval
  - Ex : In a clinical trial, some patients have not yet died at the time of the analysis of the data
    $\Rightarrow$ Only a lower bound of the true survival time is known (right censoring)

- Truncation :
  - Part of the relevant subjects will not be present at all in the data
  - Ex : In a mortality study based on HIV/AIDS death records, only subjects who died of HIV/AIDS and recorded as such are included (right truncation)

Censoring and truncation do not only take place in 'time-to-event' data.

Examples

- Insurance : Car accidents involving costs below a certain threshold are often not declared to the insurance company
  $\Rightarrow$ Left truncation
- Ecology : Chemicals in river water cannot be detected below the detection limit of the laboratory instrument
  $\Rightarrow$ Left censoring
- Astronomy : A star is only observable with a telescope if it is bright enough to be seen by the telescope
  $\Rightarrow$ Left truncation

Right censoring

Only a lower bound for the time of interest is known

$$T = \text{survival time}$$

$$C = \text{censoring time}$$

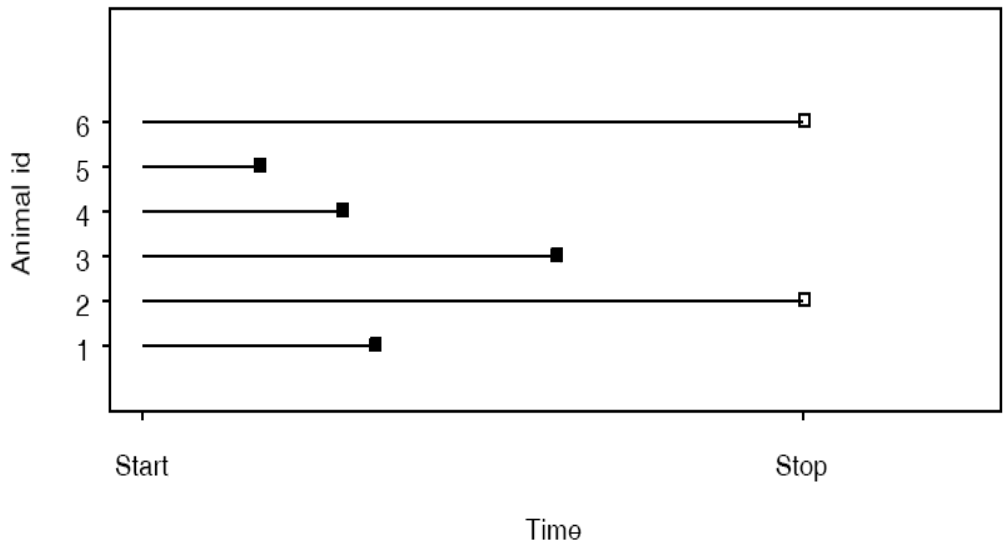
$\Rightarrow$ Data : (Y, δ) with

$$Y = \min(T, C)$$

$$\delta = I(T \leq C)$$

Type I right censoring

- All subjects are followed for a fixed amount of time
  $\rightarrow$ all censored subjects have the same censoring time
- Ex : Type I censoring in animal study



Type II right censoring

- ◇ All subjects start to be followed up at the same time and follow up continues until r individuals have experienced the event of interest (r is some predetermined integer) $\rightarrow$ The $n - r$ censored items all have a censoring time equal to the failure time of the r^{th} item.
- ◇ Ex : Type II censoring in industrial study : all lamps are put on test at the same time and the test is terminated when r of the n lamps have failed.

Random right censoring

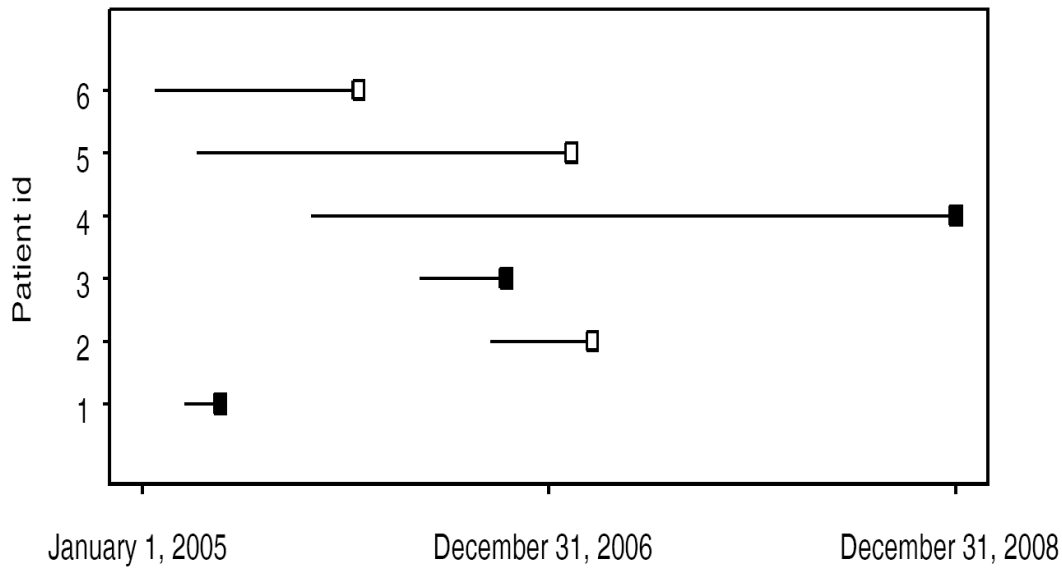
- The study itself continues until a fixed time point but subjects enter and leave the study at different times
  - $\rightarrow$ censoring is a random variable
  - $\rightarrow$ censoring can occur for various reasons:
    - end of study
    - lost to follow up
    - competing event (e.g. death due to some cause other than the cause of interest)
    - patient withdrawing from the study, change of treatment, ...
- Ex : Random right censoring in a cancer clinical trial

Example : Random right censoring in HIV study

- Study enrolment: January 2005 - December 2006
- Study end: December 2008
- Objective: HIV patients followed up to death due to AIDS or AIDS related complication (time in month from confirmed diagnosis)
- Possible causes of censoring :
  - death due to other cause
  - lost to follow up / dropped out
  - still alive at the end of study

Table: Data of 6 patients in HIV study

Patient id	Entry Date	Date last seen	Status	Time	Censoring
1	18 March 2005	20 June 2005	Dropped out	3	0
2	19 Sept 2006	20 March 2007	Dead due to AIDS	6	1
3	15 May 2006	16 Oct 2006	Dead due to accident	5	0
4	01 Dec 2005	31 Dec 2008	Alive	37	0
5	9 Apr 2005	10 Feb 2007	Dead due to AIDS	22	1
6	25 Jan 2005	24 Jan 2006	Dead due to AIDS	12	1



Left censoring

- Some subjects have already experienced the event of interest at the time they enter in the trial
- Only an upper bound for the time of interest is known
  $\Rightarrow$ Data : (Y_ℓ, δ_ℓ) with

$$\begin{aligned} Y_\ell &= \max(T, C_\ell) \\ \delta_\ell &= I(T > C_\ell) \\ C_\ell &= \text{censoring time} \end{aligned}$$

- Ex : Left censoring in malaria trial
  - Children between 2 and 10 years are followed up for malaria
  - Once children have experienced malaria, they will have antibodies in their blood against the Plasmodium parasite
  - Children entered at the age of 2 might have already been in touch with the parasite

Interval censoring

- The event of interest is only known to occur within a certain interval (L, U)
- Contrary to right and left censoring, we never observe the exact survival time
- Typically occurs if diagnostic tests are used to assess the event of interest
- Ex : Interval censoring in malaria trial
  $\rightarrow$ The exact time to malaria is between the last negative and the first positive test

Truncation : Individuals of a subset of the population of interest do not appear in the sample

Left truncation

- Occurs often in studies where a subject must first meet a particular condition before he/she can enter in the study and followed up for the event of interest
  $\Rightarrow$ Subjects that experience the event of interest before the condition is met, will not appear in the study
- Data : (T, L) observed if $T \geq L$, with

  T = survival time

  L = left truncation time

- Ex : Left truncation in HIV study
  - Incubation period between HIV infection and seroconversion
  - An individual is considered to have been infected with HIV only after seroconversion
    $\Rightarrow$ If we study HIV infected individuals and follow them for survival, all subjects that died between HIV infection and seroconversion will not be considered for inclusion in the study

Right truncation

- Occurs when only subjects who have experienced the event of interest are included in the sample
- Data : (T, R) observed if $T \leq R$, with

  T = survival time

  R = right truncation time
- Ex : Right truncation in AIDS study
  - Consider time between HIV seroconversion and development of AIDS
  - Often use a sample of AIDS patients, and ascertain retrospectively time of HIV infection
    $\Rightarrow$ Patients with long incubation time will not be part of the sample, nor patients that die from another cause before they develop AIDS

Remark

- Censoring :
  At least some information is available for a ‘complete’ random sample of the population
- Truncation :
  No information at all is available for a subset of the population

Nonparametric estimation

We will develop nonparametric estimators of the

- survival function
- cumulative hazard function
- hazard rate

for censored and truncated data

All these estimators will be based on the nonparametric likelihood function :

- Different from the likelihood for completely observed data due to the presence of censoring and truncation
- We will derive the likelihood function for :
  - right censored data
  - any type of censored data (right, left and interval censoring)
  - truncated data

Likelihood for randomly right censored data

- Random sample of individuals of size n :

  $T_1, \ldots, T_n$ survival time
  $C_1, \ldots, C_n$ censoring time

  $\Rightarrow$ Observed data : (Y_i, δ_i) $(i = 1, \ldots, n)$ with

  $$Y_i = \min(T_i, C_i)$$
  $$\delta_i = I(T_i \leq C_i)$$

- Denote

  $f(\cdot)$ and $F(\cdot)$ for the density and distribution of T
  $g(\cdot)$ and $G(\cdot)$ for the density and distribution of C

  and we assume that T and C are independent (called independent censoring)

Contribution to the likelihood of an event $(y_i = t_i, \delta_i = 1)$:

$$\begin{aligned}
&\lim_{\substack{\epsilon \to 0 \\ >}} \frac{1}{2\epsilon} P\left(y_i - \epsilon < Y < y_i + \epsilon, \delta = 1\right) \\
&= \lim_{\substack{\epsilon \to 0 \\ >}} \frac{1}{2\epsilon} P\left(y_i - \epsilon < T < y_i + \epsilon, T \leq C\right) \\
&= \lim_{\substack{\epsilon \to 0 \\ >}} \frac{1}{2\epsilon} \int_{y_i - \epsilon}^{y_i + \epsilon} \int_{t}^{\infty} dG(c) dF(t) \quad \text{(due to independence)} \\
&= \lim_{\substack{\epsilon \to 0 \\ >}} \frac{1}{2\epsilon} \int_{y_i - \epsilon}^{y_i + \epsilon} (1 - G(t)) dF(t) \\
&= (1 - G(y_i)) f(y_i)
\end{aligned}$$

Contribution to the likelihood of a right censored observation $(y_i = c_i, \delta_i = 0)$:

$$\begin{aligned}
&\lim_{\substack{\epsilon \to 0 \\ >}} \frac{1}{2\epsilon} P\left(y_i - \epsilon < Y < y_i + \epsilon, \delta = 0\right) \\
&\quad = \lim_{\substack{\epsilon \to 0 \\ >}} \frac{1}{2\epsilon} P\left(y_i - \epsilon < C < y_i + \epsilon, T > C\right) \\
&\quad = (1 - F(y_i))g(y_i)
\end{aligned}$$

This leads to the following formula of the likelihood :

$$\prod_{i=1}^{n} \left[(1 - G(y_i))f(y_i)\right]^{\delta_i} \left[(1 - F(y_i))g(y_i)\right]^{1-\delta_i}$$

We assume that the censoring is uninformative, i.e. the distribution of the censoring times does not depend on the parameters of interest related to the survival function.

$\Rightarrow$ The factors $(1 - G(y_i))^{\delta_i}$ and $g(y_i)^{1-\delta_i}$ are non-informative for inference on the survival function

$\Rightarrow$ They can be removed from the likelihood, leading to

$$\prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} = \prod_{i=1}^{n} h(y_i)^{\delta_i} S(y_i)$$

- This likelihood can also be written as

$$L = \prod_{i \in D} f(y_i) \prod_{i \in R} S(y_i)$$

with D the index set of survival times and R the index set of right censored times

- It is straightforward to see that the same survival likelihood is also valid in the case of fixed censoring times (type I and type II)

Likelihood for right, left and/or interval censored data

Generalization of the previous likelihood to include right, left and interval censoring :

$$L = \prod_{i \in D} f(y_i) \prod_{i \in R} S(y_i) \prod_{i \in L} (1 - S(y_i)) \prod_{i \in I} (S(l_i) - S(r_i)),$$

with

- D index set of survival times
- R index set of right censored times
- L index set of left censored times
- I index set of interval censored times (with l_i the lower limit and r_i the upper limit)

Likelihood for left truncated data

Suppose that the survival time T_i is left truncated at a_i

$\Rightarrow$ We have to consider the conditional distribution of T_i given $T_i \geq a_i$:

$$
\begin{aligned}
f(t_i | T \geq a_i) &= \lim_{\substack{\epsilon \to 0 \\ >}} \frac{1}{2\epsilon} P(t_i - \epsilon < T < t_i + \epsilon \mid T \geq a_i) \\
&= \lim_{\substack{\epsilon \to 0 \\ >}} \frac{1}{2\epsilon} \frac{P(t_i - \epsilon < T < t_i + \epsilon, T \geq a_i)}{P(T \geq a_i)} \\
&= \frac{1}{P(T \geq a_i)} \lim_{\substack{\epsilon \to 0 \\ >}} \frac{P(t_i < T < t_i + \epsilon)}{\epsilon} \\
&= \frac{f(t_i)}{S(a_i)}
\end{aligned}
$$

This leads to the following likelihood, accommodating left truncation and any type of censoring :

$$L = \prod_{i \in D} \frac{f(t_i)}{S(a_i)} \prod_{i \in R} \frac{S(t_i)}{S(a_i)} \prod_{i \in L} \frac{S(a_i) - S(t_i)}{S(a_i)} \prod_{i \in I} \frac{S(l_i) - S(r_i)}{S(a_i)}$$

For right truncated data :

- Consider the conditional density obtained by replacing $S(a_i)$ by $1 - S(b_i)$, where b_i is the right truncation time for subject i
- The likelihood function can then be constructed in a similar way

Nonparametric estimation of the survival function

- The survival (or distribution) function is at the basis of many other quantities (mean, quantiles, ...)
- The survival function is also useful to identify an appropriate parametric distribution
- For estimating the survival function in a nonparametric way, we need to take censoring and truncation into account

Kaplan-Meier estimator of the survival function

- Kaplan and Meier (*JASA*, 1958)
- Nonparametric estimation of the survival function for right censored data
- Based on the order in which events and censored observations occur

Notations :

- n observations $y_1, \ldots, y_n$ with censoring indicators $\delta_1, \ldots, \delta_n$
- r distinct event times ($r \leq n$)
- ordered event times : $y_{(1)}, \ldots, y_{(r)}$ and corresponding number of events: $d_{(1)}, \ldots, d_{(r)}$
- $R_{(j)}$ is the size of the risk set at event time $y_{(j)}$

- Log-likelihood for right censored data :

$$\sum_{i=1}^{n} \left[\delta_i \log f(y_i) + (1 - \delta_i) \log S(y_i) \right]$$

- Replacing the density function $f(y_i)$ by $S(y_{i-}) - S(y_i)$, yields the nonparametric log-likelihood :

$$\log L = \sum_{i=1}^{n} \left[\delta_i \log(S(y_{i-}) - S(y_i)) + (1 - \delta_i) \log S(y_i) \right]$$

- Aim : finding an estimator $\hat{S}(\cdot)$ which maximizes $\log L$
- It can be shown that the maximizer of $\log L$ takes the following form :

$$\hat{S}(t) = \prod_{j: y_{(j)} \leq t} (1 - h_{(j)}),$$

for some $h_{(1)}, \ldots, h_{(r)}$

- Plugging-in $\hat{S}(\cdot)$ into the log-likelihood, gives after some algebra :

$$\log L = \sum_{j=1}^{r} \left[d_{(j)} \log h_{(j)} + \left(R_{(j)} - d_{(j)} \right) \log(1 - h_{(j)}) \right]$$

- Using this expression to solve

$$\frac{d}{dh_{(j)}} \log L = 0$$

leads to

$$\hat{h}_{(j)} = \frac{d_{(j)}}{R_{(j)}}$$

- Plugging in this estimate $\hat{h}_{(j)}$ in $\hat{S}(t) = \prod_{j:y_{(j)} \leq t}(1 - h_{(j)})$ we obtain :

$$\hat{S}(t) = \prod_{j:y_{(j)} \leq t} \frac{R_{(j)} - d_{(j)}}{R_{(j)}} = \text{ Kaplan-Meier estimator}$$

- Step function with jumps at the event times
- If the largest observation, say y_n, is censored :
  - $\hat{S}(t)$ does not attain 0
  - Impossible to estimate $S(t)$ consistently beyond y_n
  - Various solutions :
    - Set $\hat{S}(t) = 0$ for $t \geq y_n$
    - Set $\hat{S}(t) = \hat{S}(y_n)$ for $t \geq y_n$
    - Let $\hat{S}(t)$ be undefined for $t \geq y_n$

Uncensored case

When all data are uncensored, the Kaplan-Meier estimator reduces to the empirical distribution function

Consider case without ties for simplicity :

- If no censoring, $R_{(j)} - d_{(j)} = R_{(j+1)}$ for $j = 1, \ldots, r$
- We can rewrite the KM estimator as

$$\begin{aligned}
\hat{S}(t) &= \frac{R_{(2)}}{R_{(1)}} \frac{R_{(3)}}{R_{(2)}} \cdots \frac{R_{(k+1)}}{R_{(k)}} \quad \text{where} \quad y_{(k)} \leq t < y_{(k+1)} \\
&= \frac{R_{(k+1)}}{R_{(1)}} \\
&= \frac{\#\text{ subjects with survival time } \geq y_{(k+1)}}{\#\text{ at risk before first death time}} \\
&= \frac{1}{n} \sum_{i=1}^{n} I(y_i > t)
\end{aligned}$$

Asymptotic normality of the KM estimator

- Asymptotic variance of the KM estimator :

$$V_{As}(\hat{S}(t)) = n^{-1} S^2(t) \int_0^t \frac{dH^u(s)}{(1 - H(s))(1 - H(s-))},$$

where

- $H(t) = P(Y \leq t) = 1 - S(t)(1 - G(t))$
- $H^u(t) = P(Y \leq t, \delta = 1)$

- This variance can be consistently estimated as (Greenwood formula)

$$\hat{V}_{As}(\hat{S}(t)) = \hat{S}^2(t) \sum_{j: y_{(j)} \leq t} \frac{d_{(j)}}{R_{(j)}(R_{(j)} - d_{(j)})}$$

- Asymptotic normality of $\hat{S}(t)$:

$$\frac{\hat{S}(t) - S(t)}{\sqrt{\hat{V}_{As}(\hat{S}(t))}} \xrightarrow{d} N(0, 1)$$

Nelson-Aalen estimator of the cumulative hazard function

- Proposed independently by Nelson (*Technometrics*, 1972) and Aalen (*Annals of Statistics*, 1978) :

$$\hat{H}(t) = \sum_{j:y_{(j)} \leq t} \frac{d_{(j)}}{R_{(j)}} \qquad \text{for } t \leq y_{(r)}$$

- Its asymptotic variance can be estimated by

$$\hat{V}_{As}(\hat{H}(t)) = \sum_{j:y_{(j)} \leq t} \frac{d_{(j)}}{R_{(j)}^2}$$

- Asymptotic normality :

$$\frac{\hat{H}(t) - H(t)}{\sqrt{\hat{V}_{As}(\hat{H}(t))}} \xrightarrow{d} N(0,1)$$

Alternative for KM estimator

- An alternative estimator for $S(t)$ can be obtained based on the Nelson-Aalen estimator using the relation

$$S(t) = \exp(-H(t)),$$

leading to

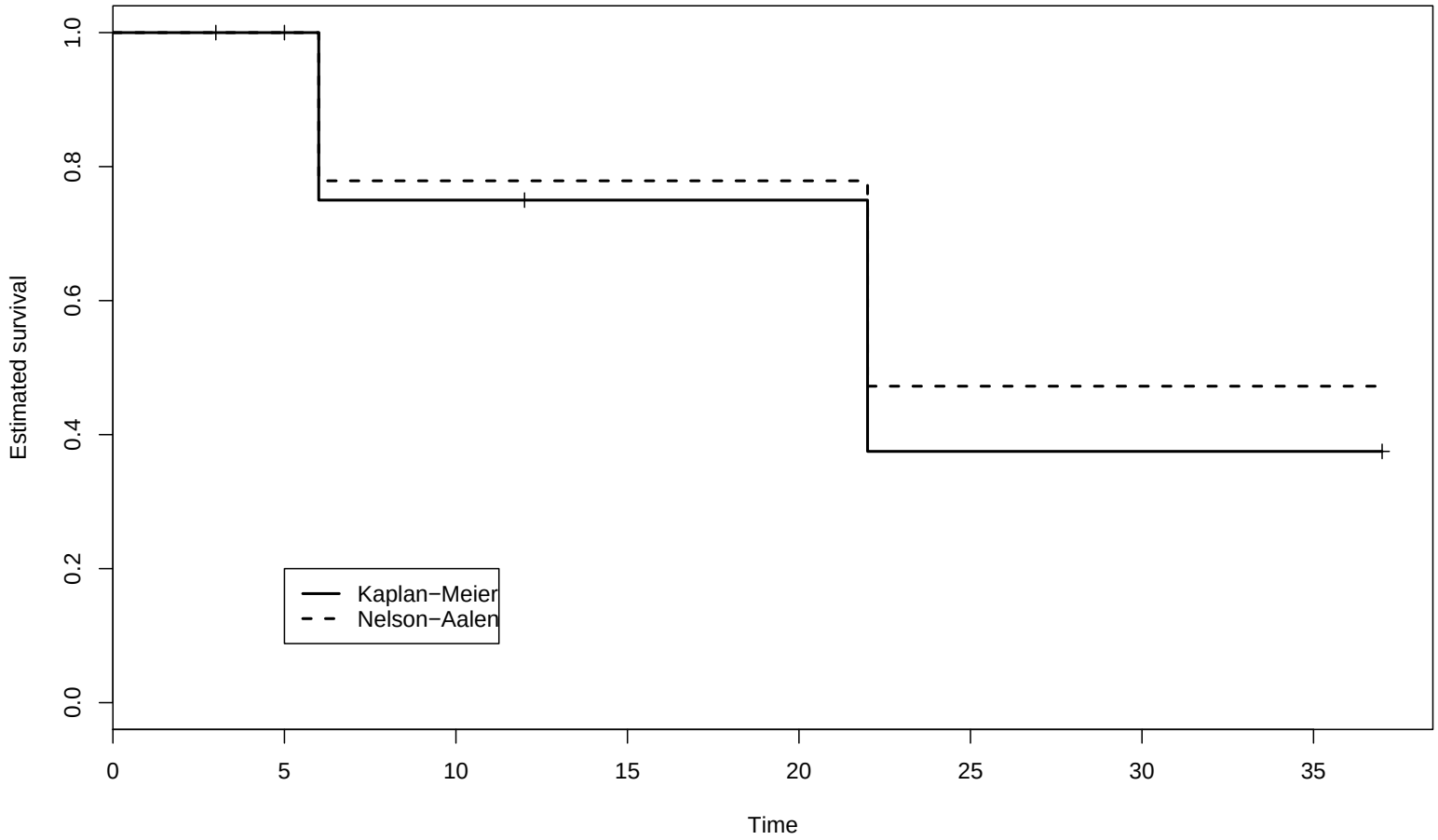
$$\hat{S}_{alt}(t) = \prod_{j:y_{(j)} \leq t} \exp\left(-\frac{d_{(j)}}{R_{(j)}}\right)$$

- $\hat{S}(t)$ and $\hat{S}_{alt}(t)$ are asymptotically equivalent
- $\hat{S}_{alt}(t)$ performs often better than $\hat{S}(t)$ for small samples

Example : Survival function for 6 HIV diagnosed patients

- Ordered observed times: 3*, 5*, 6, 12*, 22, 37*
- Only two contributions to KM and NA estimator :

		Event time	
		6	22
Number of events	$d_{(j)}$	1	1
Number at risk	$R_{(j)}$	4	2
KM contribution	$1 - d_{(j)}/R_{(j)}$	3/4	1/2
KM estimator	$\hat{S}(y_{(j)})$	3/4=0.75	3/8=0.375
NA contribution	$\exp(-d_{(j)}/R_{(j)})$	0.7788	0.6065
NA estimator	$\prod_{j:y_{(j)}\leq t} \exp(-d_{(j)}/R_{(j)})$	0.7788	0.4723



Confidence intervals for the survival function

- From the asymptotic normality of $\hat{S}(t)$, a $100(1-\alpha)\%$ confidence interval (CI) for $S(t)$ (t fixed) is given by :

$$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\hat{V}_{As}(\hat{S}(t))}$$

- However, this CI may contain points outside the $[0, 1]$ interval

  $\Rightarrow$ Use an appropriate transformation to determine the CI on the transformed scale and then transform back

- A popular transformation is $\log(-\log S(t))$, which takes values between $-\infty$ and ∞.
- One can show that

$$\frac{\log(-\log \hat{S}(t)) - \log(-\log S(t))}{\sqrt{\hat{V}_{As}\left(\log(-\log \hat{S}(t))\right)}} \xrightarrow{d} N(0,1),$$

where

$$\hat{V}_{As}\left(\log(-\log \hat{S}(t))\right) = \frac{1}{\left(\log \hat{S}(t)\right)^2} \sum_{j:y_{(j)} \le t} \frac{d_{(j)}}{R_{(j)}(R_{(j)} - d_{(j)})}$$

- Hence, CI for $\log(-\log S(t))$ is given by

$$\log(-\log \hat{S}(t)) \pm z_{\alpha/2} \sqrt{\hat{V}_{As}\left(\log(-\log \hat{S}(t))\right)}$$

- By transforming back, we get the following CI for $S(t)$:

$$\hat{S}(t)^{\exp\left[\pm z_{\alpha/2} \sqrt{\hat{V}_{As}\left(\log(-\log \hat{S}(t))\right)}\right]}$$

Point estimate of the mean survival time

- Nonparametric estimator can be obtained using the Kaplan-Meier estimator, since

$$\mu = E(T) = \int_0^\infty xf(x)dx = \int_0^\infty S(x)dx$$

$\Rightarrow$ We can estimate μ by replacing $S(x)$ by the KM estimator $\hat{S}(x)$

- But, $\hat{S}(t)$ is inconsistent in the right tail if the largest observation (say y_n) is censored
  - Proposal 1 : assume y_n experiences the event immediately after the censoring time :

$$\hat{\mu}_{y_n} = \int_0^{y_n} \hat{S}(t)dt$$

  - Proposal 2 : restrict integration to a predetermined interval $[0, t_{max}]$ and consider $\hat{S}(t) = \hat{S}(y_n)$ for $y_n \leq t \leq t_{max}$:

$$\hat{\mu}_{t_{max}} = \int_0^{t_{max}} \hat{S}(t)dt$$

- $\hat{\mu}_{y_n}$ and $\hat{\mu}_{t_{max}}$ are inconsistent estimators of μ, but given the lack of data in the right tail, we cannot do better (at least not nonparametrically)
- Variance of $\hat{\mu}_\tau$ (with τ either y_n or t_{max}) :

$$\hat{V}_{As}(\hat{\mu}_\tau) = \sum_{j=1}^{r} \left(\int_{y_{(j)}}^{\tau} \hat{S}(t)dt \right)^2 \frac{d_{(j)}}{R_{(j)}(R_{(j)} - d_{(j)})}$$

- A $100(1-\alpha)\%$ CI for μ is given by :

$$\hat{\mu}_\tau \pm z_{\alpha/2}\sqrt{\hat{V}_{As}(\hat{\mu}_\tau)}$$

Point estimate of the median survival time

- Advantages of the median over the mean :
  - As survival function is often skewed to the right, the mean is often influenced by outliers, whereas the median is not
  - Median can be estimated in a consistent way (if censoring is not too heavy)
- An estimator of the p^{th} quantile x_p is given by :

$$\hat{x}_p = \inf \left\{ t \mid \hat{S}(t) \leq 1 - p \right\}$$

$\Rightarrow$ An estimate of the median is given by $\hat{x}_{p=0.5}$

- Asymptotic variance of $\hat{x}_p$:

$$\hat{V}_{As}(\hat{x}_p) = \frac{\hat{V}_{As}(\hat{S}(x_p))}{\hat{f}^2(x_p)},$$

where $\hat{f}$ is an estimator of the density f

- Estimation of f involves smoothing techniques and the choice of a bandwidth sequence
  $\Rightarrow$ We prefer not to use this variance estimator in the construction of a CI
- Thanks to the asymptotic normality of $\hat{S}(x_p)$:

$$P\left(-z_{\alpha/2} \leq \frac{\hat{S}(x_p) - S(x_p)}{\sqrt{\hat{V}_{As}(\hat{S}(x_p))}} \leq z_{\alpha/2} \right) \approx 1 - \alpha,$$

with obviously $S(x_p) = 1 - p$.

$\Rightarrow$ A $100(1-\alpha)\%$ CI for x_p is given by

$$\left\{ t : -z_{\alpha/2} \leq \frac{\hat{S}(t) - (1-p)}{\sqrt{\hat{V}_{As}(\hat{S}(t))}} \leq z_{\alpha/2} \right\}$$

Example : Schizophrenia patients

- Schizophrenia is one of the major mental illnesses encountered in Ethiopia

  $\rightarrow$ disorganized and abnormal thinking, behavior and language + emotionally unresponsive

  $\rightarrow$ higher mortality rates due to natural and unnatural causes

- Project on schizophrenia in Butajira, Ethiopia

  $\rightarrow$ survey of the entire population (68491 individuals) in the age group 15-49 years

$\Rightarrow$ 280 cases of schizophrenia identified and followed for 5 years (1997-2001)

Table: Data on schizophrenia patients

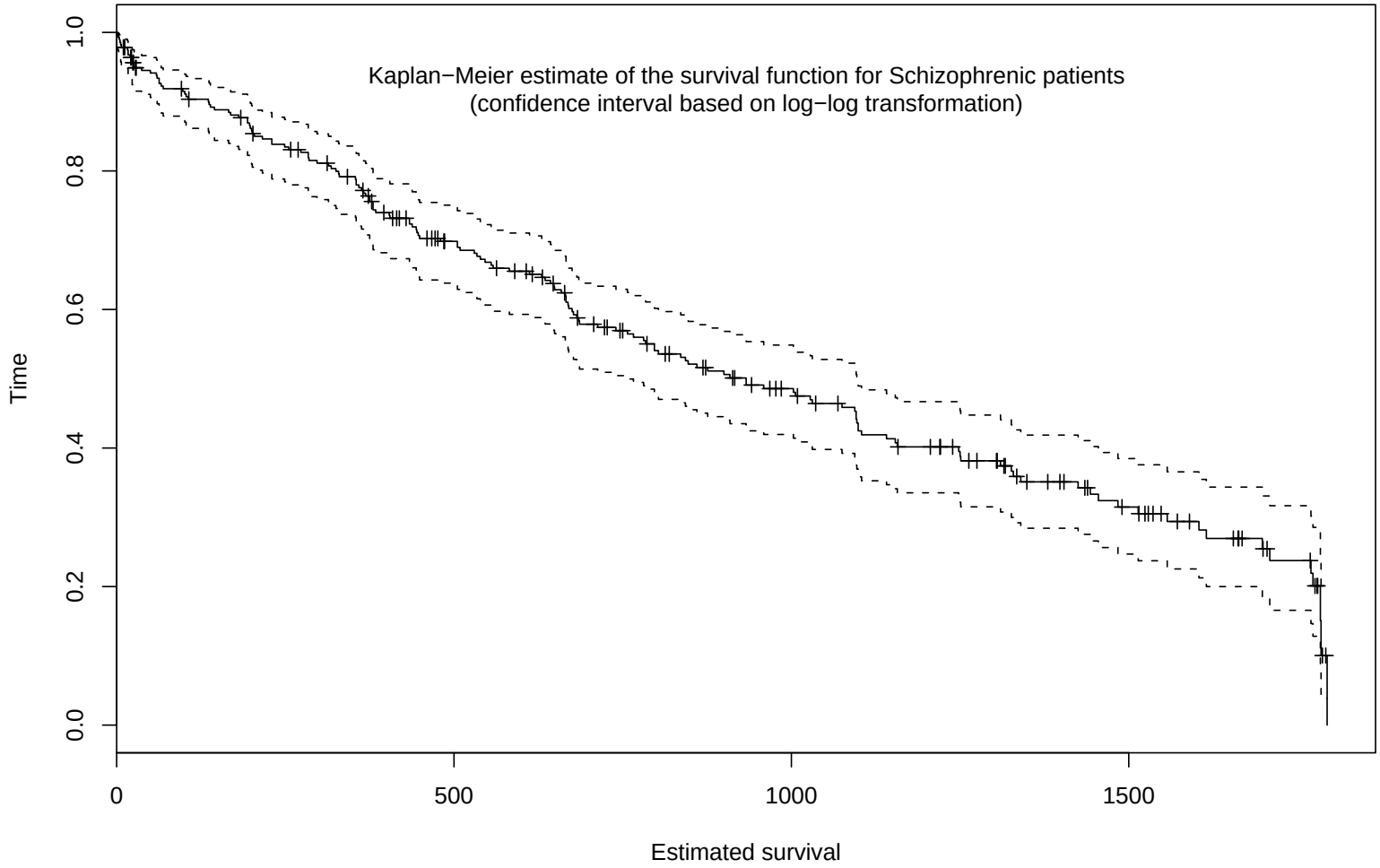
Patid	Time	Censor	Education	Onset	Marital	Gender	Age
1	1	1	1	37	3	1	44
2	3	1	3	15	2	2	23
3	4	1	6	26	1	1	33
4	5	1	12	25	1	1	31
5	5	0	5	29	3	1	33
...							
278	1787	0	2	16	2	1	18
279	1792	0	2	23	1	1	25
280	1794	1	2	28	1	1	35

- In R : `survfit`

```
schizo<-read.table("c://...//Schizophrenia.csv",
header=T,sep=";")
KM_schizo_l<-survfit(Surv(Time,Censor)~1,data=schizo,
type="kaplan-meier", conf.type="log-log")
plot(KM_schizo_l, conf.int=T, xlab="Estimated survival",
ylab="Time", yscale=1)
mtext("Kaplan-Meier estimate of the survival function
for Schizophrenic patients", 3,-3)
mtext("(confidence interval based on log-log
transformation)", 3,-4)
```

- In SAS : `proc lifetest`

```
title1 'Kaplan-Meier estimate of the survival function
for Schizophrenic patients';
proc lifetest method=km width=0.5 data=schizo;
time Time*Censor(0);
run;
```

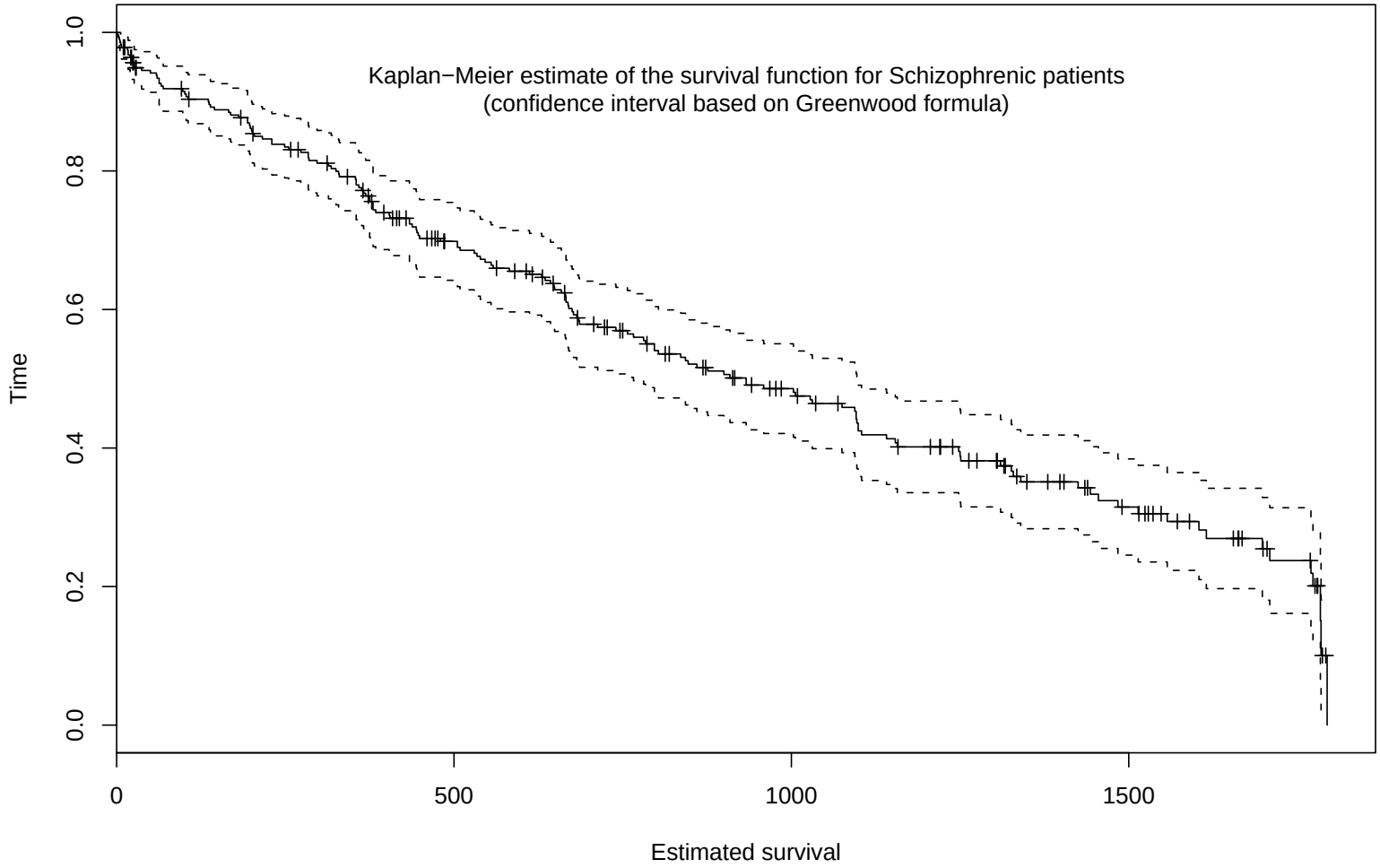


```
> KM_schizo_l
Call: survfit(formula = Surv(Time, Censor) ~ 1, data = schizo, type =
"kaplan-meier", conf.type = "log-log")

      n  events  median 0.95LCL 0.95UCL
    280     163     933     757    1099


> summary(KM_schizo_l)
Call: survfit(formula = Surv(Time, Censor) ~ 1, data = schizo, type =
"kaplan-meier", conf.type = "log-log")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    280       1    0.996 0.00357       0.9749        0.999
    3    279       1    0.993 0.00503       0.9717        0.998
    4    277       1    0.989 0.00616       0.9671        0.997
…
 1770     13       1    0.219 0.03998       0.1465        0.301
 1773     12       1    0.201 0.04061       0.1283        0.285
 1784      8       2    0.151 0.04329       0.0782        0.245
 1785      6       2    0.100 0.04092       0.0387        0.197
 1794      1       1    0.000      NA           NA           NA
```



```
> KM_schizo_g
Call: survfit(formula = Surv(Time, Censor) ~ 1, data = schizo, type =
"kaplan-meier", conf.type = "plain")

      n  events  median 0.95LCL 0.95UCL
    280     163     933     766    1099

> summary(KM_schizo_g)
Call: survfit(formula = Surv(Time, Censor) ~ 1, data = schizo, type =
"kaplan-meier", conf.type = "plain")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    280       1    0.996 0.00357       0.9894        1.000
    3    279       1    0.993 0.00503       0.9830        1.000
    4    277       1    0.989 0.00616       0.9772        1.000
 …
 1770     13       1    0.219 0.03998       0.1409        0.298
 1773     12       1    0.201 0.04061       0.1214        0.281
 1784      8       2    0.151 0.04329       0.0659        0.236
 1785      6       2    0.100 0.04092       0.0203        0.181
 1794      1       1    0.000      NA           NA           NA
```

- Median survival time is estimated to be 933 days
- 95% CI for the median : [757, 1099]
- Survival at, e.g., 505 days is estimated to be 0.6897 with std error 0.0290
- 95% CI for $S(505)$: [0.6329, 0.7465] (without transformation)
- 95% CI for $S(505)$: [0.6290, 0.7426] (using log-log transformation)

Estimation of the survival function for left truncated and right censored data

- We need to redefine $R_{(j)}$:

$$\begin{aligned} R_{(j)} &= \text{number of individuals at risk at time } y_{(j)} \\ &\quad \text{and under observation prior to time } y_{(j)} \\ &= \#\{i : l_i \leq y_{(j)} \leq y_i\}, \end{aligned}$$

  where l_i is the truncation time.

- We cannot estimate $S(t)$, but only a conditional survival function

$$S_l(t) = P(T \geq t \mid T \geq l)$$

  for some fixed value $l \geq \min(l_1, \ldots, l_n)$

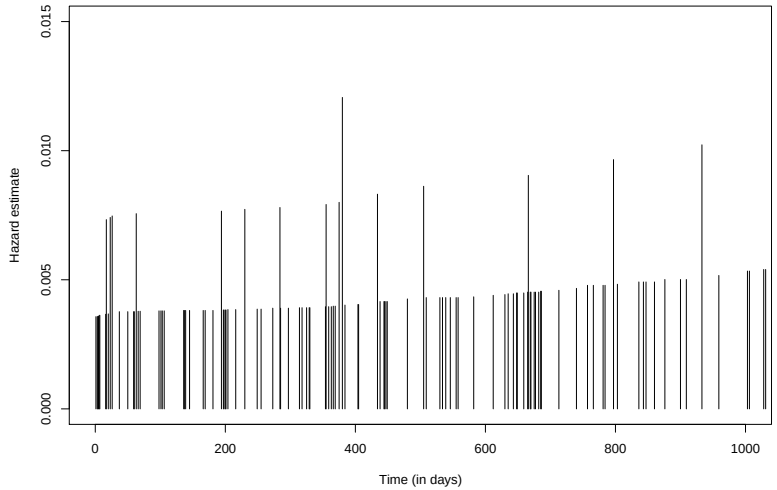
- The conditional survival function $S_l(t)$ is estimated by

$$\hat{S}_l(t) = \begin{cases} 1 & \text{if } t < l \\ \prod_{j:l \le y_{(j)} \le t} \left(1 - \frac{d_{(j)}}{R_{(j)}}\right) & \text{if } t \ge l \end{cases}$$

- Proposed and named after Lynden-Bell (1971), an astronomer

Estimation of the hazard function for right censored data

- Usually more informative about the underlying population than the survival or the cumulative hazard function
- Crude estimator : take the size of the jumps of the cumulative hazard function
- Ex : Crude estimator of the hazard function for data on schizophrenic patients



- Smoothed estimator of $h(t)$: (weighted) average of the crude estimator over all time points in the interval $[t-b, t+b]$ for a certain value b, called the bandwidth
- Uniform weight over interval $[t-b, t+b]$:

$$\hat{h}(t) = (2b)^{-1} \sum_{j=1}^{r} I\left(-b \leq t - y_{(j)} \leq b\right) \Delta\hat{H}(y_{(j)}),$$

where
- $\hat{H}(t) =$ Nelson-Aalen estimator
- $\Delta\hat{H}(y_{(j)}) = \hat{H}(y_{(j)}) - \hat{H}(y_{(j-1)})$

- General weight function :

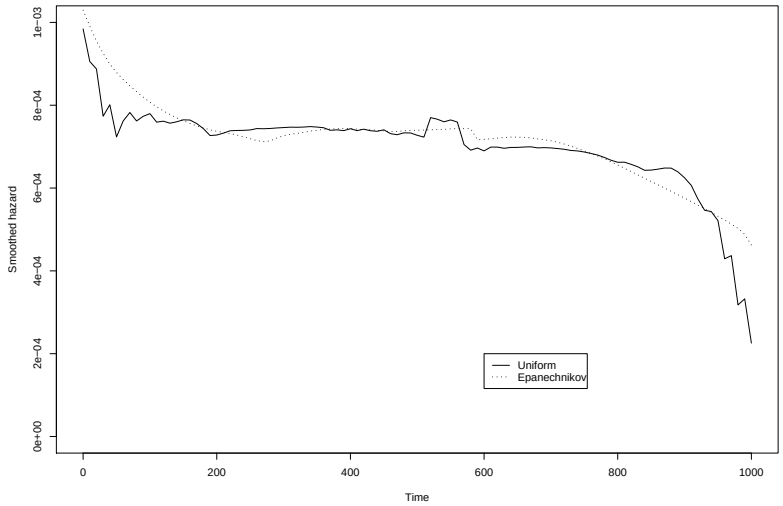
$$\hat{h}(t) = b^{-1} \sum_{j=1}^{r} K\left(\frac{t - y_{(j)}}{b}\right) \Delta\hat{H}(y_{(j)}),$$

where $K(\cdot)$ is a density function, called the kernel

- Example of kernels :

Name	Density function	Support
uniform	$K(x) = \frac{1}{2}$	$-1 \leq x \leq 1$
Epanechnikov	$K(x) = \frac{3}{4}(1 - x^2)$	$-1 \leq x \leq 1$
biweight	$K(x) = \frac{15}{16}(1 - x^2)^2$	$-1 \leq x \leq 1$

- Ex : Smoothed estimator of the hazard function for data on schizophrenic patients



- The choice of the kernel does not have a major impact on the estimated hazard rate, but the choice of the bandwidth does
  $\Rightarrow$ It is important to choose the bandwidth in an appropriate way, by e.g. plug-in, cross-validation, bootstrap, ... techniques
- Variance of $\hat{h}(t)$ can be estimated by

$$\hat{V}_{As}(\hat{h}(t)) = b^{-2} \sum_{j=1}^{r} K\left(\frac{t - y_{(j)}}{b}\right)^2 \Delta \hat{V}_{As}(\hat{H}(y_{(j)})),$$

where $\Delta \hat{V}_{As}(\hat{H}(y_{(j)})) = \hat{V}_{As}(\hat{H}(y_{(j)})) - \hat{V}_{As}(\hat{H}(y_{(j-1)}))$

Hypothesis testing in a nonparametric setting

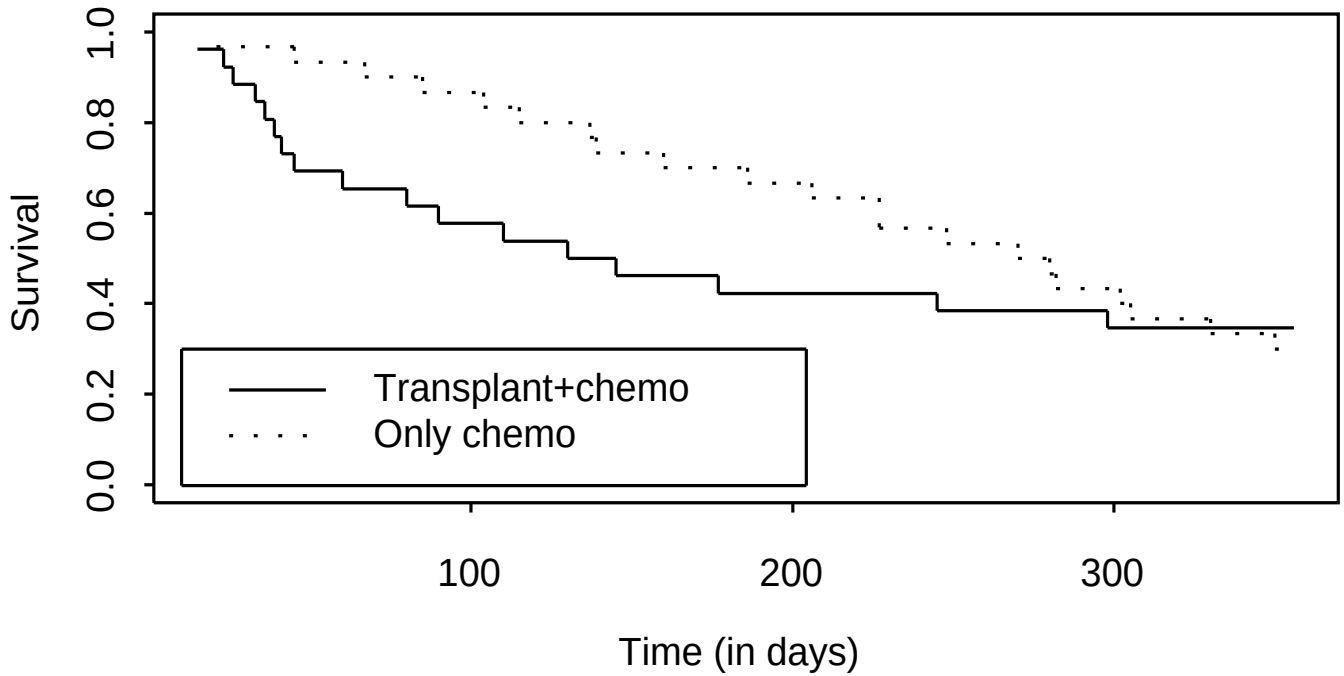
Hypothesis testing in a nonparametric setting

- Hypotheses concerning the hazard function of one population
- Hypotheses comparing the hazard function of two or more populations

Note that

- It is important to consider overall differences over time
- We will develop tests that look at weighted differences between observed and expected quantities (under H_0)
- Weights allow to put more emphasis on certain part of the data (e.g. early or late departure from H_0)
- Particular cases : log-rank test, Breslow's test, Cox Mantel test, Peto and Peto test, ...

Ex : Survival differences in leukemia patients : chemotherapy vs. chemotherapy + autologous transplantation



Hypotheses for the hazard function of one population

- Test whether a censored sample of size n comes from a population with a known hazard function $h_0(t)$:

$$H_0 : h(t) = h_0(t) \qquad \text{for all } t \leq y_{(r)}$$

$$H_1 : h(t) \neq h_0(t) \qquad \text{for some } t \leq y_{(r)}$$

- Based on the NA estimator of the cumulative hazard function, a crude estimator of the hazard function at time $y_{(j)}$ is

$$\frac{d_{(j)}}{R_{(j)}}$$

- Under H_0, the hazard function at time $y_{(j)}$ is $h_0(y_{(j)})$

- Let $w(t)$ be some weight function, with $w(t) = 0$ for $t > y_{(r)}$
- Test statistic :

$$Z = \sum_{j=1}^{r} w(y_{(j)}) \frac{d_{(j)}}{R_{(j)}} - \int_0^{y_{(r)}} w(s) h_0(s) ds$$

- Under H_0 :

$$V(Z) = \int_0^{y_{(r)}} w^2(s) \frac{h_0(s)}{R(s)} ds$$

with $R(s)$ corresponding to the number of subjects in the risk set at time s

- For large samples :

$$\frac{Z}{\sqrt{V(Z)}} \approx N(0, 1)$$

One sample log-rank test

- Weight function : $w(t) = R(t)$
- Test statistic :

$$\begin{aligned} Z &= \sum_{j=1}^{r} d_{(j)} - \int_0^{y_{(r)}} R(s) h_0(s) ds \\ &= \sum_{j=1}^{r} d_{(j)} - \sum_{i=1}^{n} \int_0^{y_i} h_0(s) ds \\ &= \sum_{j=1}^{r} d_{(j)} - \sum_{i=1}^{n} H_0(y_i) = O - E \end{aligned}$$

- Under H_0 :

$$V(Z) = \int_0^{y_{(r)}} R(s) h_0(s) ds = E$$

and

$$\frac{O - E}{\sqrt{E}} \approx N(0, 1)$$

Example : Survival in patients with Paget disease

- Benign form of breast cancer
- Compare survival in a sample of patients to the survival in the overall population
  - Data : Finkelstein et al. (2003)
  - Hazard function of the population : standardized actuarial table
- Compute the expected number of deaths under H_0 using
  - follow-up information of the group of patients with Paget disease
  - relevant hazard function from standardized actuarial table

Paget disease data:

- age (in years) at diagnosis
- time to death or censoring (in years)
- censoring indicator
- gender (1=male, 2=female)
- race (1=Caucasian, 2=black)

Age	Follow-up	Status	Gender	Race
52	22	0	2	1
53	4	0	2	1
57	8	0	2	1
57	7	0	2	1
...				
85	6	1	2	1
86	1	0	2	1

Standardized actuarial table :

- age (in years)
- hazard (per 100 subjects) for respectively Caucasian males, Caucasian females, black males, and black females

	Hazard function			
Age	Caucasian male	Caucasian female	black male	black female
50-54	0.6070	0.3608	1.3310	0.7156
55-59	0.9704	0.5942	1.9048	1.0558
60-64	1.5855	0.9632	2.8310	1.6048
...				
80-84	9.3128	6.2880	10.4625	7.2523
85-	17.7671	14.6814	16.0835	13.7017

- E.g. first patient : Caucasian female followed from 52 years on for 22 years :

(1)	hazard for the 52^{th} year	=	0.3608
(2)	hazard for the 53^{th} year	=	0.3608
...	...		...
(22)	hazard for the 73^{th} year	=	2.3454
	Total (cumulative hazard)	=	25.637
$\Rightarrow$	for one particular patient (/100)	=	0.25637

and do the same for all patients

- Expected number of deaths under H_0 : $E = 9.55$
- Observed number of deaths : $O = 13$
- Test statistic :

$$\frac{O - E}{\sqrt{E}} = \frac{13 - 9.55}{\sqrt{9.55}} = 1.116$$

- Two-sided hypothesis test :

$$2P(Z > 1.116) = 0.264$$

$\Rightarrow$ We do not reject H_0

Other weight functions

Weight function proposed by Harrington and Fleming (1982):

$$w(t) = R(t)S_0^p(t)(1 - S_0(t))^q \qquad p, q \geq 0$$

- $p = q = 0$: log-rank test
- $p > q$: more weight on early deviations from H_0
- $p < q$: more weight on late deviations from H_0
- $p = q > 0$: more weight on deviations in the middle
- $p = 1, q = 0$: generalization of the one-sample Wilcoxon test to censored data

Comparing the hazard functions of two populations

- Hypothesis test :

$$H_0 : h_1(t) = h_2(t) \qquad \text{for all } t \leq y_{(r)}$$

$$H_1 : h_1(t) \neq h_2(t) \qquad \text{for some } t \leq y_{(r)}$$

- Notations :
  - $y_{(1)}, y_{(2)}, \ldots, y_{(r)}$: ordered event times in the pooled sample
  - $d_{(j)k}$: number of events at time $y_{(j)}$ in sample k ($j = 1, \ldots, r$ and $k = 1, 2$)
  - $R_{(j)k}$: number of individuals at risk at time $y_{(j)}$ in sample k
  - $d_{(j)} = \sum_{k=1}^{2} d_{(j)k}$ and $R_{(j)} = \sum_{k=1}^{2} R_{(j)k}$

- Derive a 2×2 contingency table for each event time $y_{(j)}$:

Group	Event	No Event	Total
1	$d_{(j)1}$	$R_{(j)1} - d_{(j)1}$	$R_{(j)1}$
2	$d_{(j)2}$	$R_{(j)2} - d_{(j)2}$	$R_{(j)2}$
Total	$d_{(j)}$	$R_{(j)} - d_{(j)}$	$R_{(j)}$

- Test the independence between the rows and the columns, which corresponds to the assumption that the hazard in the two groups at time $y_{(j)}$ is the same
- Test statistic with group 1 as reference group :

$$O_j - E_j = d_{(j)1} - \frac{d_{(j)} R_{(j)1}}{R_{(j)}}$$

with $O_j =$ observed number of events in the first group

$E_j =$ expected number of events in the first group assuming that $h_1 \equiv h_2$

- Test statistic : weighted average over the different event times :

$$\begin{aligned} U &= \sum_{j=1}^{r} w(y_{(j)})(O_j - E_j) \\ &= \sum_{j=1}^{r} w(y_{(j)})\left(d_{(j)1} - \frac{d_{(j)} R_{(j)1}}{R_{(j)}}\right) \end{aligned}$$

  Different weights can be used, but choice must be made before looking at the data

- For large samples and under the null hypothesis :

$$\frac{U}{\sqrt{V(U)}} \approx N(0,1)$$

Variance of U :

- Can be obtained by observing that conditional on $d_{(j)}$, $R_{(j)1}$ and $R_{(j)}$, the statistic $d_{(j)1}$ has a hypergeometric distribution
- Hence,

$$\begin{aligned} V(U) &= \sum_{j=1}^{r} w^2(y_{(j)}) V(d_{(j)1}) \\ &= \sum_{j=1}^{r} w^2(y_{(j)}) \frac{d_{(j)} \left(\frac{R_{(j)1}}{R_{(j)}}\right) \left(1 - \frac{R_{(j)1}}{R_{(j)}}\right) (R_{(j)} - d_{(j)})}{R_{(j)} - 1} \end{aligned}$$

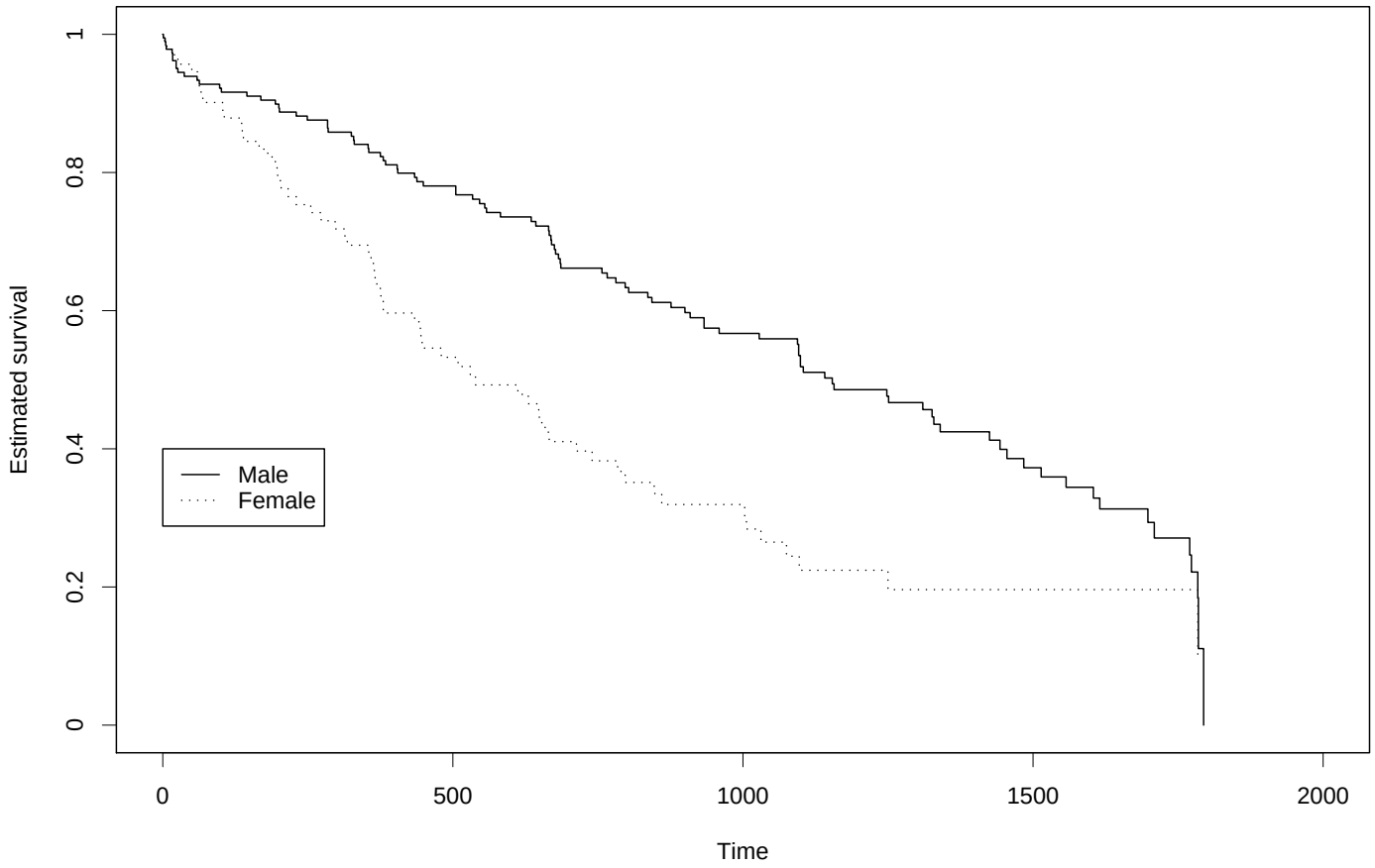
Weights :

- $w(y_{(j)}) = 1$
  - ↪ log-rank test
  - ↪ optimum power to detect alternatives when the hazard rates in the two populations are proportional to each other

- $w(y_{(j)}) = R_{(j)}$
  - ↪ generalization by Gehan (1965) of the two sample Wilcoxon test
  - ↪ puts more emphasis on early departures from H_0
  - ↪ weights depend heavily on the event times and the censoring distribution

- $w(y_{(j)}) = f(R_{(j)})$
  - ↪ Tarone and Ware (1977)
  - ↪ a suggested choice is $f(R_{(j)}) = \sqrt{R_{(j)}}$
  - ↪ puts more weight on early departures from H_0
- $w(y_{(j)}) = \hat{S}(y_{(j)}) = \prod_{y_{(k)} \leq y_{(j)}} \left(1 - \frac{d_{(k)}}{R_{(k)}+1}\right)$
  - ↪ Peto and Peto (1972) and Kalbfleisch and Prentice (1980)
  - ↪ based on an estimate of the common survival function close to the pooled product limit estimate
- $w(y_{(j)}) = \left(\hat{S}(y_{(j-1)})\right)^p \left(1 - \hat{S}(y_{(j-1)})\right)^q \quad p \geq 0, q \geq 0$
  - ↪ Fleming and Harrington (1981)
  - ↪ include weights of the log-rank as special case
  - ↪ $q = 0, p > 0$: more weight is put on early differences
  - ↪ $p = 0, q > 0$: more weight is put on late differences

Example : Comparing survival for male and female schizophrenic patients



- Observed number of events in female group : 93
- Expected number of events under H_0 : 62
- Log-rank weights :
  - $U/\sqrt{V(U)} = 4.099$
  - p-value (2-sided) = 0.000042
- Peto and Peto weights :
  - $U/\sqrt{V(U)} = 4.301$
  - p-value (2-sided) = 0.000017

Comparing the hazard functions of more than 2 populations

- Hypothesis test :

$$H_0 : h_1(t) = h_2(t) = \ldots = h_I(t) \text{ for all } t \leq y_{(r)}$$

$$H_1 : h_i(t) \neq h_j(t) \text{ for at least one pair } (i, j)$$

$$\text{for some } t \leq y_{(r)}$$

- Notations : same as earlier but now $k = 1, \ldots, I$
- Test statistic based on the $I \times 2$ contingency tables for the different event times $y_{(j)}$

Group	Event	No Event	Total
1	$d_{(j)1}$	$R_{(j)1} - d_{(j)1}$	$R_{(j)1}$
2	$d_{(j)2}$	$R_{(j)2} - d_{(j)2}$	$R_{(j)2}$
...			
I	$d_{(j)I}$	$R_{(j)I} - d_{(j)I}$	$R_{(j)I}$
Total	$d_{(j)}$	$R_{(j)} - d_{(j)}$	$R_{(j)}$

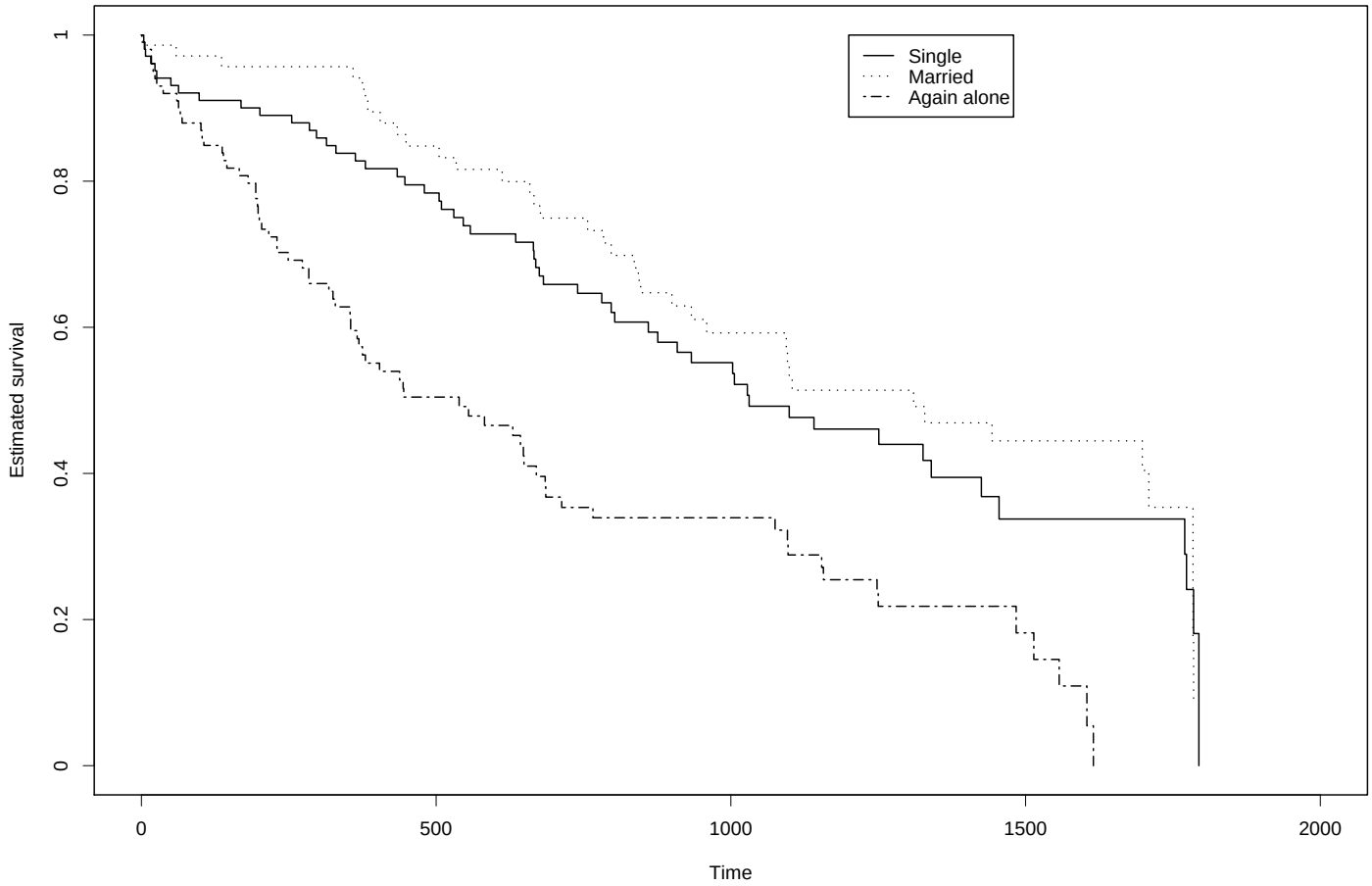
- The random vector $d_{(j)} = (d_{(j)1}, \ldots, d_{(j)l})^t$ has a multivariate hypergeometric distribution
- We can define analogues of the test statistic U defined previously :

$$U_k = \sum_{j=1}^{r} w(y_{(j)}) \left(d_{(j)k} - \frac{d_{(j)} R_{(j)k}}{R_{(j)}} \right),$$

  which is a weighted sum of the differences between the observed and expected number of events under H_0
- The components of the vector $(U_1, \ldots, U_l)$ are linearly dependent because $\sum_{k=1}^{l} U_k = 0$
  $\Rightarrow$ define $U = (U_1, \ldots, U_{l-1})^t$
  $\Rightarrow$ derive $V(U)$, the variance-covariance matrix of U
- For large sample size and under H_0 :

$$U^t V(U)^{-1} U \approx \chi^2_{l-1}$$

Example : Comparing survival for schizophrenic patients according to their marital status



- Observed number of events : 55 (single), 37 (married), 71 (alone again)
- Expected number of events under H_0 : 67, 55, 41
- Test statistic : $U^t V(U)^{-1} U = 31.44$
- p-value = 1.5×10^{-7} (based on a χ^2_2)

Test for trend

- Sometimes there exists a natural ordering in the hazard functions
- If such an ordering exists, tests that take it into consideration have more power to detect significant effects
- Test for trend :

$$H_0 : h_1(t) = h_2(t) = \ldots = h_l(t) \text{ for all } t \leq y_{(r)}$$

$$H_1 : h_1(t) \leq h_2(t) \leq \ldots \leq h_l(t) \text{ for some } t \leq y_{(r)} \text{ with at least one strict inequality}$$

(H_1 implies that $S_1(t) \geq S_2(t) \geq \ldots \geq S_l(t)$ for some $t \leq y_{(r)}$ with at least one strict inequality)

- Test statistic for trend :

$$U = \sum_{k=1}^{I} w_k U_k,$$

with

  - U_k the summary statistic of the k^{th} population
  - w_k the weight assigned to the k^{th} population, e.g. $w_k = k$ (corresponds to a linear trend in the groups)

- Variance of U :

$$V(U) = \sum_{k=1}^{I} \sum_{k'=1}^{I} w_k w_{k'} \text{Cov}(U_k, U_{k'})$$

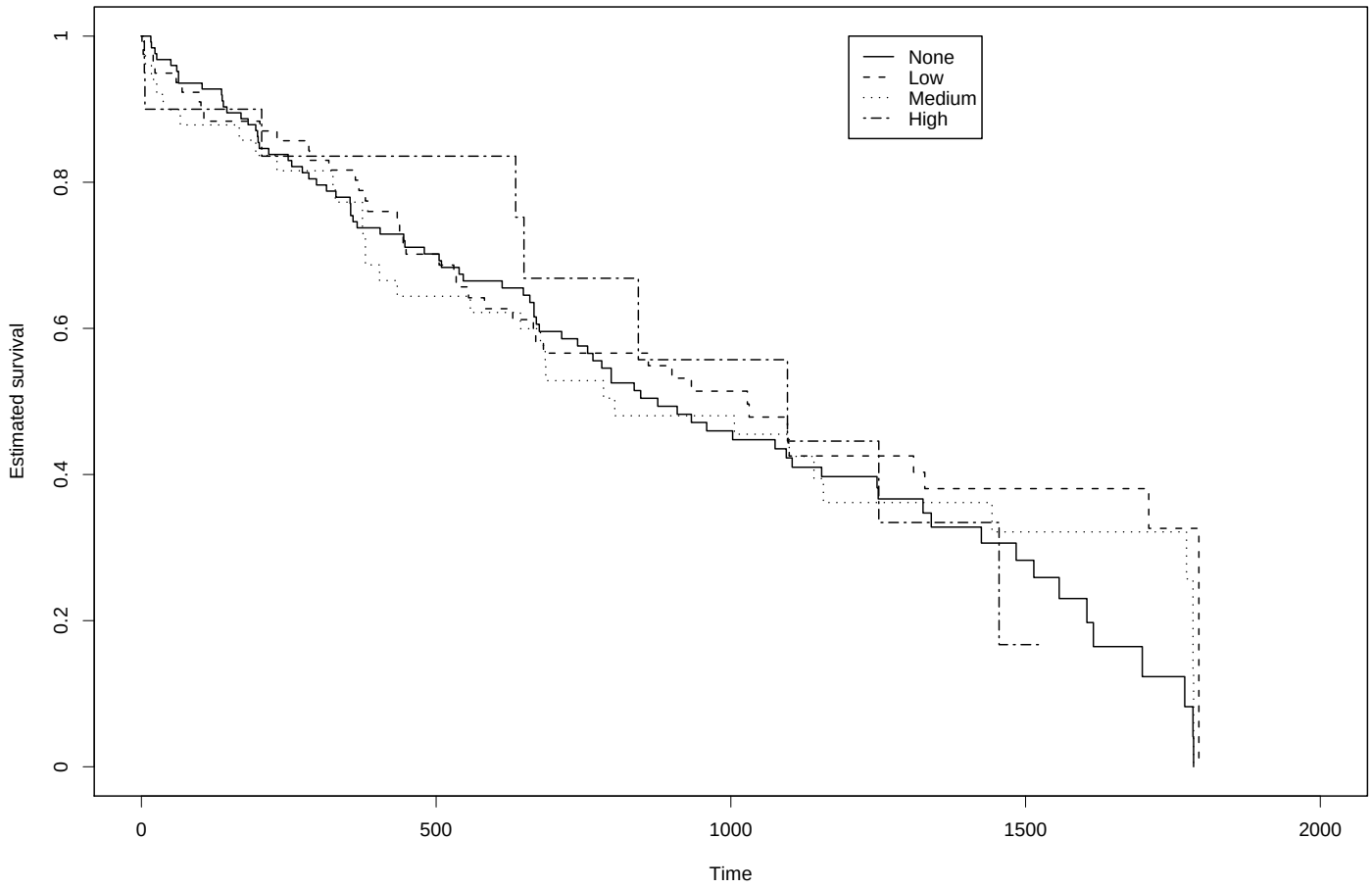
- For large sample size and under H_0 :

$$\frac{U}{\sqrt{V(U)}} \approx N(0,1)$$

- If $w_k = k$, we reject H_0 for large values of $U/\sqrt{V(U)}$ (one-sided test)

Example : Comparing survival for schizophrenic patients according to their educational level

4 educational groups : none, low, medium, high



- Observed number of events : 79 (none), 43 (low), 32 (medium), 9 (high)
- Expected number of events under H_0 : 71.3, 51.6, 31.1, 9.0
- Consider $H_1 : h_1(t) \geq \ldots \geq h_4(t)$
- Using weights 0, 1, 2, 3 we have :
  - $U = -6.77$ and $V(U) = 134$ so $U/\sqrt{V(U)} = -0.58$
  - One-sided p-value :

$$P(Z < -0.58) = 0.28$$

- p-value for 'global test' : $p = 0.49$

Stratified tests

- In some cases, subjects in a study can be grouped according to particular characteristics, called strata
  Ex : prognosis group (good, average, poor)
- It is often advisable to adjust for strata as it reduces variance
  $\Rightarrow$ Stratified test : obtain an overall assessment of the difference, by combining information over the different strata to gain power
- Hypothesis test :

$$H_0 : h_{1b}(t) = h_{2b}(t) = \ldots = h_{lb}(t)$$
$$\text{for all } t \leq y_{(r)} \text{ and } b = 1, \ldots, m,$$

  where $h_{kb}(\cdot)$ is the hazard of group k and stratum b ($k = 1, \ldots, l; b = 1, \ldots, m$)

- Test statistic :
  - U_{kb} = summary statistic for population k ($k = 1, \ldots, I$) in stratum b ($b = 1, \ldots, m$)
  - Stratified summary statistic for population k : $U_{k.} = \sum_{b=1}^{m} U_{kb}$
  - Define $U_{.} = (U_{1.}, \ldots, U_{(I-1).})^t$
- Entries of the variance-covariance matrix $V(U)$ of $U_{.}$:

$$\mathrm{Cov}(U_{k.}, U_{k'.}) = \sum_{b=1}^{m} \mathrm{Cov}(U_{kb}, U_{k'b})$$

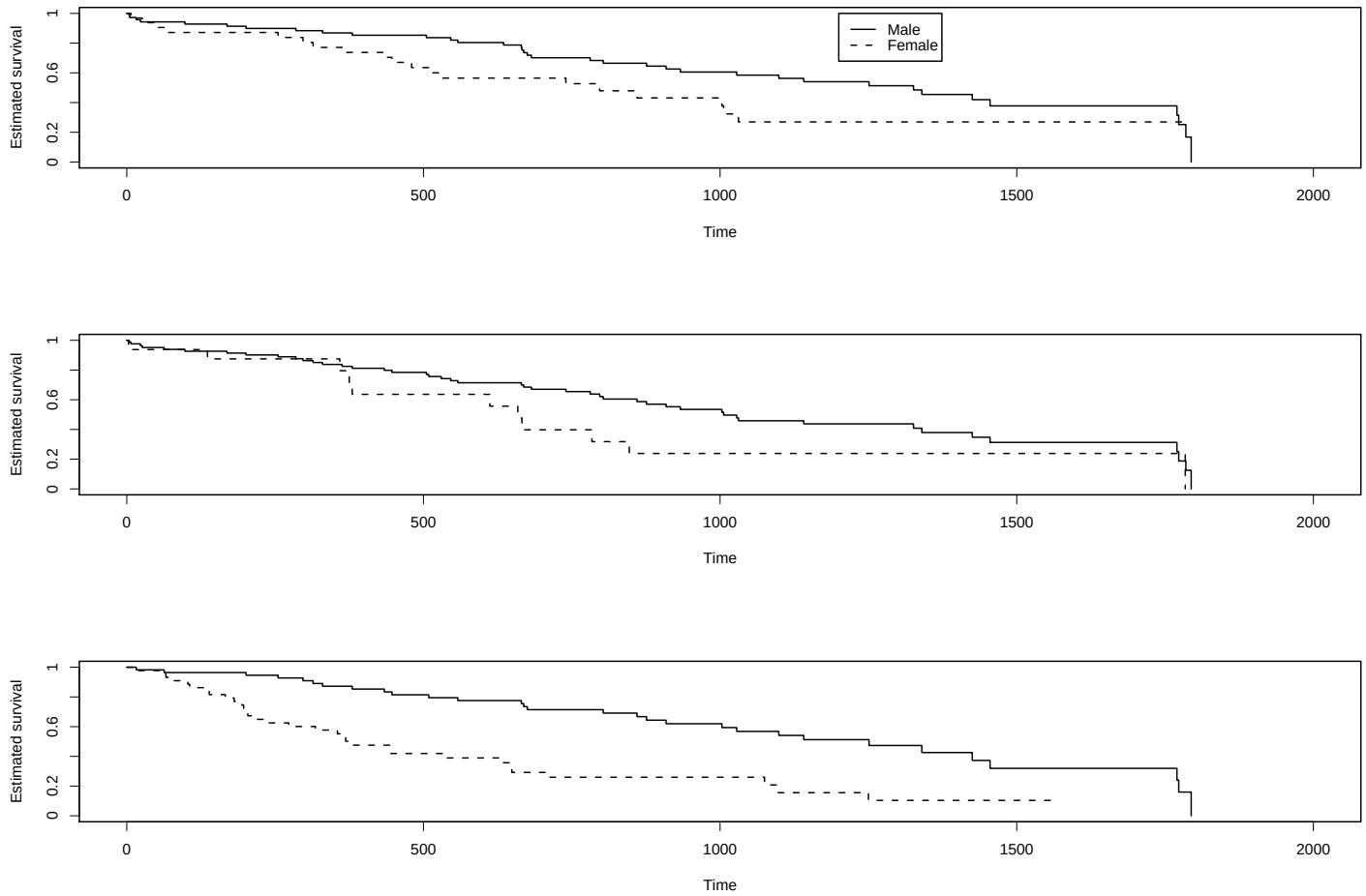
- For large sample size and under H_0 :

$$U_{.}^t V(U)^{-1} U_{.} \approx \chi^2_{I-1}$$

- If only two populations :

$$\frac{\sum_{b=1}^{m} U_b}{\sqrt{\sum_{b=1}^{m} V(U_b)}} \approx N(0,1)$$

Example : Comparing survival for schizophrenic patients according to gender stratified by marital status



- Log-rank test (weights=1) :

	single	married	alone again
U_b	5.81	5.98	6.06
$V(U_b)$	9.77	4.12	15.71

- $\sum_{b=1}^{3} U_b = 17.85$ and $\sum_{b=1}^{3} V(U_b) = 29.60$
- Test statistic :

$$\frac{\sum_{b=1}^{3} U_b}{\sqrt{\sum_{b=1}^{3} V(U_b)}} = \sqrt{10.76}$$

- p-value (2-sided) = 0.00103

Matched pairs test

- Particular case of the stratified test when each stratum consists of only 2 subjects
- m matched pairs of censored data : $(y_{1b}, y_{2b}, \delta_{1b}, \delta_{2b})$ for $b = 1, \ldots, m$, with
  - 1^{st} subject of the pair receiving treatment 1
  - 2^{nd} subject of the pair receiving treatment 2
- Hypothesis test :

$$H_0 : h_{1b}(t) = h_{2b}(t) \text{ for all } t \leq y_{(r)} \text{ and } b = 1, \ldots, m$$

- It can be shown that under H_0 and for large m :

$$\frac{U_.}{\sqrt{V(U_.)}} = \frac{D_1 - D_2}{\sqrt{D_1 + D_2}} \approx N(0,1),$$

where D_j = number of matched pairs in which the individual from sample j dies first ($j = 1, 2$)

$\Rightarrow$ Weight function has no effect on final test statistic in this case

Proportional hazards models

The semiparametric proportional hazards model

- Cox, 1972
- Stratified tests not always the optimal strategy to adjust for covariates :
  - Can be problematic if we need to adjust for several covariates
  - Do not provide information on the covariate(s) on which we stratify
  - Stratification on continuous covariates requires categorization
- We will work with semiparametric proportional hazards models, but there also exist parametric variations

Simplest expression of the model

- Case of two treatment groups (Treated vs. Control) :

  $$h_T(t) = \psi h_C(t),$$

  with $h_T(t)$ and $h_C(t)$ the hazard function of the treated and control group
- Proportional hazards model :
  - Ratio $\psi = h_T(t)/h_C(t)$ is constant over time
  - $\psi < 1$ ($\psi > 1$): hazard of the treated group is smaller (larger) than the hazard of the control group at any time
  - Survival curves of the 2 treatment groups can never cross each other

More generalizable expression of the model

- Consider a treatment covariate x_i (0 = control, 1 = treatment) and an exponential relationship between the hazard and the covariate x_i :

$$h_i(t) = \exp(\beta x_i) h_0(t),$$

with

  - $h_i(t)$: hazard function for subject i
  - $h_0(t)$: hazard function of the control group
  - $\exp(\beta) = \psi$: hazard ratio

- Other functional relationships can be used between the hazard and the covariate

More complex model

- Consider a set of covariates $x_i = (x_{i1}, \ldots, x_{ip})^t$ for subject i :

  $$h_i(t) = h_0(t) \exp(\beta^t x_i),$$

  with
  - β : the $p \times 1$ parameter vector
  - $h_0(t)$: the baseline hazard function (i.e. hazard for a subject with $x_{ij} = 0$, $j = 1, \ldots, p$)
- Proportional hazards (PH) assumption : ratio of the hazards of two subjects with covariates x_i and x_j is constant over time :

  $$\frac{h_i(t)}{h_j(t)} = \frac{\exp(\beta^t x_i)}{\exp(\beta^t x_j)}$$
- Semiparametric PH model : leave the form of $h_0(t)$ completely unspecified and estimate the model in a semiparametric way

Fitting the semiparametric PH model

- Based on likelihood maximization
- As $h_0(t)$ is left unspecified, we maximize a so-called partial likelihood instead of the full likelihood :
  - Derive the partial likelihood for data without ties
  - Extend to data with tied observations

Partial likelihood for data without ties

- Can be derived as a profile likelihood :

  First β is fixed, and the likelihood is maximized as a function of $h_0(t)$ only to find estimators for the baseline hazard in terms of β

- Notations :
  - r observed event times ($r = d$ as no ties)
  - $y_{(1)}, \ldots, y_{(r)}$ ordered event times
  - $x_{(1)}, \ldots, x_{(r)}$ corresponding covariate vectors
- Likelihood :

$$\prod_{j=1}^{r} h_{0(j)} \exp\left(x_{(j)}^t \beta\right) \prod_{i=1}^{n} \exp\left(-H_0(y_i) \exp(x_i^t \beta)\right),$$

with $h_{0(j)} = h_0(y_{(j)})$

- It can be seen that the likelihood is maximized when $H_0(y_i)$ takes the following form :

$$H_0(y_i) = \sum_{y_{(j)} \leq y_i} h_0(y_{(j)})$$

(i.e. $h_0(t) = 0$ for $t \neq y_{(1)}, \ldots, y_{(r)}$, which leads to the largest contribution to the likelihood)

- With β fixed, the likelihood can be rewritten as

$$\begin{aligned} & L(h_{0(1)}, \ldots, h_{0(r)} \mid \beta) \\ &= \prod_{j=1}^{r} h_{0(j)} \prod_{j=1}^{r} \exp\left(x_{(j)}^t \beta\right) \\ &\quad \times \prod_{j=1}^{r} \exp\left(- h_{0(j)} \sum_{k \in R(y_{(j)})} \exp\left(x_k^t \beta\right)\right), \end{aligned}$$

where $R(y_{(j)})$ is the risk set at time $y_{(j)}$

- Maximize the likelihood with respect to $h_{0(j)}$ by setting the partial derivatives wrt $h_{0(j)}$ equal to 0 :

$$\frac{\partial L\left(h_{0(1)}, \ldots, h_{0(r)} \mid \beta\right)}{\partial h_{0(1)}}$$

$$= \prod_{j=1}^{r} \exp\left(x_{(j)}^{t}\beta\right) \prod_{j=1}^{r} \exp\left(-h_{0(j)}b_j\right)$$

$$\times \left(h_{0(2)} \ldots h_{0(r)} - h_{0(1)}h_{0(2)} \ldots h_{0(r)}b_1\right) = 0$$

$$\Longleftrightarrow 1 - h_{0(1)}b_1 = 0,$$

with $b_j = \sum_{k \in R(y_{(j)})} \exp\left(x_k^t \beta\right)$, and in general

$$h_{0(j)} = \frac{1}{b_j} = \frac{1}{\sum_{k \in R(y_{(j)})} \exp\left(x_k^t \beta\right)}$$

- Plug this solution into the likelihood, and ignore factors not containing any of the parameters :

$$\begin{aligned} L(\beta) &= \prod_{j=1}^{r} \frac{\exp\left(x_{(j)}^t \beta\right)}{\sum_{k \in R(y_{(j)})} \exp\left(x_k^t \beta\right)} \\ &= \text{partial likelihood} \end{aligned}$$

- This expression is used to estimate β through maximization
- Logarithm of the partial likelihood :

$$\ell(\beta) = \sum_{j=1}^{r} x_{(j)}^t \beta - \sum_{j=1}^{r} \log\left(\sum_{k \in R(y_{(j)})} \exp\left(x_k^t \beta\right) \right)$$

- Maximization is often done via the Newton-Raphson procedure, which is based on the following iterative procedure :

$$\hat{\beta}_{new} = \hat{\beta}_{old} + I^{-1}(\hat{\beta}_{old})U(\hat{\beta}_{old}),$$

with

  - $U(\hat{\beta}_{old}) =$ vector of scores
  - $I^{-1}(\hat{\beta}_{old}) =$ inverse of the observed information matrix

$\Rightarrow$ convergence is reached when $\hat{\beta}_{old}$ and $\hat{\beta}_{new}$ are sufficiently close together

◇ Score function $U(\beta)$:

$$\begin{aligned} U_h(\beta) &= \frac{\partial \ell(\beta)}{\partial \beta_h} \\ &= \sum_{j=1}^{r} x_{(j)h} - \sum_{j=1}^{r} \frac{\sum_{k \in R(y_{(j)})} x_{kh} \exp\left(x_k^t \beta\right)}{\sum_{k \in R(y_{(j)})} \exp\left(x_k^t \beta\right)} \end{aligned}$$

◇ Observed information matrix $I(\beta)$:

$$\begin{aligned} I_{hl}(\beta) &= -\frac{\partial^2 \ell(\beta)}{\partial \beta_h \partial \beta_l} \\ &= \sum_{j=1}^{r} \frac{\sum_{k \in R(y_{(j)})} x_{kh} x_{kl} \exp\left(x_k^t \beta\right)}{\sum_{k \in R(y_{(j)})} \exp\left(x_k^t \beta\right)} \\ &\quad - \sum_{j=1}^{r} \left[\frac{\sum_{k \in R(y_{(j)})} x_{kh} \exp\left(x_k^t \beta\right)}{\sum_{k \in R(y_{(j)})} \exp\left(x_k^t \beta\right)} \right] \\ &\quad \times \sum_{j=1}^{r} \left[\frac{\sum_{k \in R(y_{(j)})} x_{kl} \exp\left(x_k^t \beta\right)}{\sum_{k \in R(y_{(j)})} \exp\left(x_k^t \beta\right)} \right] \end{aligned}$$

Remarks :

- Variance-covariance matrix of $\hat{\beta}$ can be approximated by the inverse of the information matrix evaluated at $\hat{\beta}$ $\rightarrow$ $V(\hat{\beta}_h)$ can be approximated by $[I(\hat{\beta})]^{-1}_{hh}$
- Properties (consistency, asymptotic normality) of $\hat{\beta}$ are well established (Gill, 1984)
- A 100(1-α)% confidence interval for β_h is given by

  $$\hat{\beta}_h \pm z_{\alpha/2}\sqrt{V(\hat{\beta}_h)}$$

  and for the hazard ratio $\psi_h = \exp(\beta_h)$:

  $$\exp\left(\hat{\beta}_h \pm z_{\alpha/2}\sqrt{V(\hat{\beta}_h)}\right),$$

  or alternatively via the Delta method

Example : Active antiretroviral treatment cohort study

- CD4 cells protect the body from infections and other types of disease
  $\rightarrow$ if count decreases beyond a certain threshold the patients will die
- As HIV infection progresses, most people experience a gradual decrease in CD4 count
- Highly Active AntiRetroviral Therapy (HAART)
  - AntiRetroviral Therapy (ART) + 3 or more drugs
  - Not a cure for AIDS but greatly improves the health of HIV/AIDS patients

- After introduction of ART, death of HIV patients decreased tremendously
  $\rightarrow$ investigate now how HIV patients evolve after HAART
- Data from a study conducted in Ethiopia :
  - 100 individuals older than 18 years and placed under HAART for the last 4 years
  - only use data collected for the first 2 years

Table: Data of HAART Study

Pat ID	Time	Censo-ring	Gen-der	Age	Weight	Func. Status	Clin. Status	CD4	ART
1	699	0	1	42	37	2	4	3	1
2	455	1	2	30	50	1	3	111	1
3	705	0	1	32	57	0	3	165	1
4	694	0	2	50	40	1	3	95	1
5	86	0	2	35	37	0	4	34	1
...									
97	101	0	1	39	37	2	.	.	1
98	709	0	2	35	66	2	3	103	1
99	464	0	1	27	37	.	.	.	2
100	537	1	2	30	76	1	4	1	1

How is survival influenced by gender and age ?

- Define agecat $= 1$ if age < 40 years
  $= 2$ if age ≥ 40 years
- Define gender $= 1$ if male
  $= 2$ if female
- Fit a semiparametric PH model including gender and agecat as covariates :
  - $\hat{\beta}_{\text{agecat}} = 0.226$ (HR=1.25)
  - $\hat{\beta}_{\text{gender}} = 1.120$ (HR=3.06)
  - Inverse of the observed information matrix :
    $$I^{-1}(\hat{\beta}) = \begin{bmatrix} 0.4645 & 0.1476 \\ 0.1476 & 0.4638 \end{bmatrix}$$
  - 95% CI for $\hat{\beta}_{\text{agecat}}$: [-1.11, 1.56]
    95% CI for HR of old vs. young : [0.33, 4.77]
  - 95% CI for $\hat{\beta}_{\text{gender}}$: [-0.21, 2.45]
    95% CI for HR of female vs. male : [0.81, 11.64]

Partial likelihood for data with tied observations

- Events are typically observed on a discrete time scale
  $\Rightarrow$ Censoring and event times can be tied
- If ties between censoring time(s) and an event time
  $\Rightarrow$ we assume that
  - the censoring time(s) fall just after the event time
    $\Rightarrow$ they are still in the risk set of the event time
- If ties between event times of two or more subjects : Kalbfleish and Prentice (1980) proposed an appropriate likelihood function, but
  - rarely used due to its complexity
  - different approximations have been proposed

Approximation proposed by Breslow (1974) :

$$L(\beta) = \prod_{j=1}^{r} \frac{\prod_{l:y_l=y_{(j)},\delta_l=1} \exp\left(x_l^t\beta\right)}{\left[\sum_{k:y_k\geq y_{(j)}} \exp\left(x_k^t\beta\right)\right]^{d_{(j)}}}$$

Approximation proposed by Efron (1977) :

$$L(\beta) = \prod_{j=1}^{r} \frac{\prod_{l:y_l=y_{(j)},\delta_l=1} \exp\left(x_l^t\beta\right)}{V_j(\beta)}$$

where

$$V_j(\beta) = \prod_{h=1}^{d_{(j)}} \left(\sum_{k:y_k\geq y_{(j)}} \exp\left(x_k^t\beta\right) - \frac{h-1}{d_{(j)}} \sum_{l:y_l=y_{(j)},\delta_l=1} \exp\left(x_l^t\beta\right) \right)$$

Approximation proposed by Cox (1972) :

$$L(\beta) = \prod_{j=1}^{r} \frac{\prod_{l:y_l=y_{(j)},\delta_l=1} \exp\left(x_l^t \beta\right)}{\sum_{q \in Q_j} \sum_{h \in q} \exp\left(x_h^t \beta\right)},$$

with Q_j the set of all possible combinations of $d_{(j)}$ subjects from the risk set $R(y_{(j)})$

Example : Effect of gender on survival of schizophrenic patients

- Fit a semiparametric PH model including gender as covariate :

Approx.	Max(partial likel.)	$\hat{\beta}$	s.e.($\hat{\beta}$)
Breslow	-776.11	0.661	0.164
Efron	-775.67	0.661	0.164
Cox	-761.36	0.665	0.165

- HR for female vs. male: 1.94
- 95% CI : [1.41; 2.69]

- Contribution to the partial likelihood at time 1096 days
  - males : 68 at risk, 2 events
  - females : 12 at risk, no event
  - Breslow :

$$\frac{\exp(2 \times 0)}{(68 + 12 \exp \beta)^2} = 0.000120$$

  - Efron :

$$\frac{\exp(2 \times 0)}{(68 + 12 \exp \beta)\,(67 + 12 \exp \beta)} = 0.000121$$

  - Cox :

$$\frac{\exp(2 \times 0)}{\left[\exp(2\beta)\binom{12}{2} + \exp(\beta)\binom{12}{1}\binom{68}{1} + \binom{68}{2}\right]} = 0.000243$$

Testing hypotheses in the framework of the semiparametric PH model

- Global tests :
  - hypothesis tests regarding the whole vector β
- More specific tests :
  - hypothesis tests regarding a subvector of β
  - hypothesis tests for contrasts and sets of contrasts

Global hypothesis tests

- Hypotheses regarding the p-dimensional vector β :

$$H_0 : \beta = \beta_0$$

$$H_1 : \beta \neq \beta_0$$

- Wald test statistic :

$$U_W^2 = \left(\hat{\beta} - \beta_0\right)^t I\left(\hat{\beta}\right)\left(\hat{\beta} - \beta_0\right)$$

with

  - $\hat{\beta} =$ maximum likelihood estimator
  - $I\left(\hat{\beta}\right) =$ observed information matrix

$\Rightarrow$ Under H_0, and for large sample size : $U_W^2 \approx \chi_p^2$

◇ Likelihood ratio test statistic :

$$U_{LR}^2 = 2\Big(\ell(\hat{\beta}) - \ell(\beta_0)\Big)$$

with

- $\ell(\hat{\beta}) =$ log likelihood evaluated at $\hat{\beta}$
- $\ell(\beta_0) =$ log likelihood evaluated at β_0

$\Rightarrow$ Under H_0, and for large sample size : $U_{LR}^2 \approx \chi_p^2$

◇ Score test statistic :

$$U_{SC}^2 = U(\beta_0)^t I^{-1}(\beta_0) U(\beta_0)$$

with

- $U(\beta_0) =$ score vector evaluated at β_0

$\Rightarrow$ Under H_0, and for large sample size : $U_{SC}^2 \approx \chi_p^2$

Example : Effect of age and marital status on survival of schizophrenic patients

- Model the survival as a function of age and marital status :

$$H_0 : \beta = \begin{pmatrix} \beta_{\text{age}} \\ \beta_{\text{married}} \\ \beta_{\text{alone again}} \end{pmatrix} = 0$$

($\beta_{\text{single}} = 0$ to avoid overparametrization)

- $U_W^2 = 31.6$; p-value : $P(\chi_3^2 > 31.6) = 6 \times 10^{-7}$

  $U_{LR}^2 = 30.6$

  $U_{SC}^2 = 33.5$

Local hypothesis tests

- Let $\beta = (\beta_1^t, \beta_2^t)^t$, where β_2 contains the 'nuisance' parameters
- Hypotheses regarding the q-dimensional vector β_1 :

$$H_0 : \beta_1 = \beta_{10}$$

$$H_1 : \beta_1 \neq \beta_{10}$$

- Partition the information matrix as

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$$

with $I_{11} =$ matrix of partial derivatives of order 2 with respect to the components of β_1

$$\Rightarrow I^{-1} = \begin{bmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{bmatrix}$$

- Note that the complete information matrix is required to obtain I^{11}, except when $\hat{\beta}_1$ is independent of $\hat{\beta}_2$

- Define

$$\begin{aligned}
\hat{\beta}_1 &= \text{maximum likelihood estimator of } \beta_1 \\
\hat{\beta}_2(\beta_{10}) &= \text{maximum likelihood estimator of } \beta_2 \text{ with } \beta_1 \text{ put equal to } \beta_{10} \\
U_1\left(\beta_{10}, \hat{\beta}_2(\beta_{10})\right) &= \text{score subvector evaluated at } \beta_{10} \text{ and } \hat{\beta}_2(\beta_{10}) \\
I^{11}\left(\beta_{10}, \hat{\beta}_2(\beta_{10})\right) &= \text{matrix } I^{11} \text{ for } \beta_1 \text{ evaluated at } \beta_{10} \text{ and } \hat{\beta}_2(\beta_{10})
\end{aligned}$$

- Wald test :

$$U_W^2 = (\hat{\beta}_1 - \beta_{10})^t (I^{11}(\hat{\beta}))^{-1} (\hat{\beta}_1 - \beta_{10}) \approx \chi_q^2$$

- Likelihood ratio test :

$$U_{LR}^2 = 2\Big(\ell(\hat{\beta}) - \ell\big(\beta_{10}, \hat{\beta}_2(\beta_{10})\big)\Big) \approx \chi_q^2$$

- Score test :

$$U_{SC}^2 = U_1\Big(\beta_{10}, \hat{\beta}_2(\beta_{10})\Big)^t I^{11}\Big(\beta_{10}, \hat{\beta}_2(\beta_{10})\Big) \times U_1\Big(\beta_{10}, \hat{\beta}_2(\beta_{10})\Big) \approx \chi_q^2$$

Testing more specific hypotheses

- Consider a $p \times 1$ vector of coefficients c
- Hypothesis test :

$$H_0 : c^t \beta = 0$$

- Wald test statistic :

$$U_W^2 = \left(c^t \hat{\beta}\right)^t \left(c^t I^{-1}(\hat{\beta}) c\right)^{-1} \left(c^t \hat{\beta}\right)$$

Under H_0 and for large sample size :

$$U_W^2 \approx \chi_1^2$$

- Likelihood ratio test and score test can be obtained in a similar way

- If different linear combinations of the parameters are of interest, define

$$C = \begin{pmatrix} c_1^t \\ \vdots \\ c_q^t \end{pmatrix}$$

with $q \leq p$ and assume that the matrix C has full rank

- Hypothesis test :

$$H_0 : C\beta = 0$$

- Wald test statistic :

$$U_W^2 = \left(C\hat{\beta}\right)^t \left(CI^{-1}(\hat{\beta})C^t\right)^{-1} \left(C\hat{\beta}\right)$$

Under H_0 and for large sample size : $U_W^2 \approx \chi_q^2$

- Likelihood ratio test and score test can be obtained in a similar way

Example : Effect of age and marital status on survival of schizophrenic patients

- $H_0 : \beta_{\text{married}} = 0$
  - $\rightarrow$ $c^t = (0, 1, 0)$
  - $\rightarrow$ Wald test statistic : 1.18; p-value: $P(\chi^2_1 > 1.18) = 0.179$
- $H_0 : \beta_{\text{married}} = \beta_{\text{alone again}} = 0$
  - $\rightarrow$ $C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
  - $\rightarrow$ Test statistics : $U^2_W = 31.6$; $U^2_{LR} = 30.6$; $U^2_{SC} = 33.5$
  - $\rightarrow$ p-value (Wald) : $P(\chi^2_2 > 31.6) = 1 \times 10^{-7}$

Building multivariable semiparametric models

- including a continuous covariate
- including a categorical covariate
- including different types of covariates
- interactions between covariates
- time-varying covariates

Including a continuous covariate in the semiparametric PH model

- For a single continuous covariate x_i :

$$h_i(t) = h_0(t)\exp(\beta x_i)$$

where

- $h_0(t) =$ baseline hazard (refers to a subject with $x_i = 0$)
- $\exp(\beta) = \dfrac{\text{hazard of a subject } i \text{ with covar. } x_i}{\text{hazard of a subject } j \text{ with covar. } x_j = x_i - 1}$ and is independent of the covariate x_i and of t
- $\exp(r\beta) =$ hazard ratio of two subjects with a difference of r covariate units

$\Rightarrow$ $\hat{\beta} =$ increase in log-hazard corresponding to a one unit increase of the continuous covariate

Example : Impact of age on survival of schizophrenic patients

- Introduce age as a continuous covariate in the semiparametric PH model :

$$h_i(t) = h_0(t) \exp(\beta_{\text{age}} age_i)$$

- $\beta_{\text{age}} = 0.00119$ (s.e. = 0.00952).
- $HR = \dfrac{\text{hazard for a subject of age } i \text{ (in years)}}{\text{hazard for a subject of age } i-1} = 1.001$

  95% CI : [0.983, 1.020]
- Other quantities can be calculated, e.g.

$$\frac{\text{hazard for a subject of age 40}}{\text{hazard for a subject of age 30}} = \exp(10 \times 0.00119) = 1.012$$

Including a categorical covariate in the semiparametric PH model

- For a single categorical covariate x_i with I levels :

$$h_i(t) = h_0(t)\exp(\beta^t x_i),$$

where

  - $\beta = (\beta_1, \ldots, \beta_I)$
  - x_i is the covariate for subject i

- This model is overparametrized $\Rightarrow$ restrictions :
  - Set $\beta_1 = 0$ so that $h_0(t)$ corresponds to the hazard of a subject with the first level of the covariate
    - $\exp(\beta_j) =$ HR of a subject at level j relative to a subject at level 1
    - $\exp(\beta_j - \beta_{j'}) =$ HR between level j and j' (note that $V(\hat{\beta}_j - \hat{\beta}_{j'}) = V(\hat{\beta}_j) + V(\hat{\beta}_{j'}) - 2Cov(\hat{\beta}_j, \hat{\beta}_{j'})$)
  - Other choices of restrictions are possible

Example : Impact of marital status on survival of schizophrenic patients

- Introduce marital status as a categorical covariate in the semiparametric PH model

$$h_i(t) = h_0(t)\exp(\beta_{\text{married}} x_{i2} + \beta_{\text{alone again}} x_{i3}),$$

where

  - $x_{i2} = 1$ if patient is married, 0 otherwise
  - $x_{i3} = 1$ if patient is alone again, 0 otherwise

- Married vs single :
  - $\hat{\beta}_{\text{married}} = -0.206$ (s.e. $= 0.214$)
  - $HR = 0.814$ ($95\% CI : [0.534, 1.240]$), $p = 0.34$

- Alone again vs single :
  - $\hat{\beta}_{\text{alone again}} = 0.794$ (s.e. $= 0.185$)
  - $HR = 2.213$ ($95\% CI : [1.540, 3.180]$), $p = 1.7 \times 10^{-5}$

- Married vs alone again :
  - $\exp(\hat{\beta}_{\text{married}} - \hat{\beta}_{\text{alone again}}) = 0.368$
  - Variance-covariance matrix :

$$V\begin{pmatrix} \hat{\beta}_{\text{married}} \\ \hat{\beta}_{\text{alone again}} \end{pmatrix} = \begin{pmatrix} 0.0460 & 0.0183 \\ 0.0183 & 0.0342 \end{pmatrix}$$

  - $V(\hat{\beta}_{\text{married}} - \hat{\beta}_{\text{alone again}}) = 0.0436$
  - 95% CI : [0.244, 0.553]

Including different covariates in the semiparametric PH model

- Estimates for a particular parameter will then be adjusted for the other parameters in the model
- Estimates for this particular parameter will be different from the estimate obtained in a univariate model (except when the covariates are orthogonal)

Example : Impact of marital status and age on survival of schizophrenic patients

$$h_i(t) = h_0(t) \exp(\beta_{\text{age}} age_i + \beta_{\text{married}} x_{i2} + \beta_{\text{alone again}} x_{i3})$$

Covariate	$\hat{\beta}$	s.e.($\hat{\beta}$)	HR	95% CI
age	-0.0154	0.0104	0.99	[0.97,1.01]
married	-0.3009	0.2238	0.74	[0.48,1.15]
alone again	0.8195	0.1857	2.269	[1.58,3.27]

Interaction between covariates

- Interaction : the effect of one covariate depends on the level of another covariate
- Continuous / categorical (j levels) : different hazard ratios are required for the continuous covariate at each level of the categorical covariate

  $\Rightarrow$ add $j - 1$ parameters
- Categorical (j levels) / categorical (k levels) : for each level of one covariate, different HR between the levels of the other covariate with the reference are required

  $\Rightarrow$ add $(j - 1) \times (k - 1)$ parameters

Example : Impact of marital status and age on survival of schizophrenic patients

$$h_i(t) = h_0(t)\exp(\beta_{\text{married}} \times x_{i2} + \beta_{\text{alone again}} \times x_{i3} + \beta_{\text{age}} \times age_i + \beta_{\text{age | married}} \times x_{i2} \times age_i + \beta_{\text{age | alone again}} \times x_{i3} \times age_i)$$

Covariate	$\hat{\beta}$	s.e.($\hat{\beta}$)	HR	95% CI
age	-0.0238	0.0172	0.977	[0.94,1.01]
married	-0.6811	0.8579	0.506	[0.09,2.72]
alone again	0.3979	0.7475	1.489	[0.34,6.44]
age|married	0.0129	0.0299	1.013	[0.96,1.07]
age|alone again	0.0133	0.0228	1.013	[0.97,1.06]

- Effect of age in the reference group (single) :

$$\exp(\hat{\beta}_{\text{age}}) = \exp(-0.0238) = 0.977$$

- Effect of age in the married group :

$$\begin{aligned} \exp(\hat{\beta}_{\text{age}} + \hat{\beta}_{\text{age|married}}) &= \exp(-0.0238 + 0.0129) \\ &= 0.989 \end{aligned}$$

- Effect of age in the alone again group :

$$\begin{aligned} \exp(\hat{\beta}_{\text{age}} + \hat{\beta}_{\text{age|alone again}}) &= \exp(-0.0238 + 0.0133) \\ &= 0.990 \end{aligned}$$

- Likelihood ratio test for the interaction :

$$U_{LR}^2 = 0.76$$

$$P(\chi_2^2 > 0.76) = 0.684$$

- $HR_{\text{married}} = \exp(\hat{\beta}_{\text{married}}) = 0.506$
  $= HR$ of a married subject relative to a single subject at the age of 0 year

$\Rightarrow$ more relevant to express the age as the difference between a particular age of interest (e.g. 30 years)

$\Rightarrow$ has impact on parameter estimates of differences between groups, but not on parameter estimates related to age

Covariate	$\hat{\beta}$	s.e.($\hat{\beta}$)	HR	95% CI
age	-0.0238	0.0172	0.977	[0.94,1.01]
married	-0.2928	0.2378	0.746	[0.47,1.19]
alone again	0.7971	0.1911	2.219	[1.53,3.23]
age|married	0.0129	0.0299	1.013	[0.96,1.07]
age|alone again	0.0133	0.0228	1.013	[0.97,1.06]

Example : Impact of marital status and gender on survival of schizophrenic patients

$$h_i(t) = h_0(t)\exp(\beta_{\text{married}} \times x_{i2} + \beta_{\text{alone again}} \times x_{i3} + \beta_{\text{female}} \times gender_i + \beta_{\text{female}|\text{married}} \times x_{i2} \times gender_i + \beta_{\text{female}|\text{alone again}} \times x_{i3} \times gender_i)$$

Covariate	$\hat{\beta}$	s.e.($\hat{\beta}$)	HR	95% CI
female	0.520	0.286	1.681	[0.96, 2.95]
married	-0.253	0.26	0.776	[0.47, 1.29]
alone again	0.807	0.236	2.242	[1.41, 3.56]
female|married	0.389	0.46	1.476	[0.60, 3.64]
female|alone again	-0.146	0.372	0.865	[0.42, 1.79]

↪ Likelihood ratio test for the interaction :

$$U^2_{LR} = 1.94;\ P(\chi^2_2 > 1.94) = 0.23$$

Time varying covariates

- In some applications, covariates of interest change with time
- Extension of the Cox model :

$$h_i(t) = h_0(t) \exp(\beta^t x_i(t))$$

$\Rightarrow$ Hazards are no longer proportional

- Estimation of β :
  - Let $x_k(y)$ be the covariate vector for subject k at time y
  - Define the partial likelihood :

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\left(x_i(y_i)^t \beta\right)}{\sum_{k \in R(y_i)} \exp\left(x_k(y_i)^t \beta\right)} \right]^{\delta_i}$$

  - Let

$$\hat{\beta} = \text{argmax}_\beta L(\beta)$$

Example : Time varying covariates in data on the first time to insemination for cows

- Aim : find constituent in milk that is predictive for the hazard of first insemination
  - one possible predictor is the ureum concentration
  - milk ureum concentration changes over time
- Information for an individual cow i $(i = 1, \ldots, n)$:

  $$\left(y_i, \delta_i, x_i\left(t_{i1}\right), \ldots, x_i\left(t_{ik_i}\right)\right)$$

  Covariate is determined only once a month

  $\Rightarrow$ Value at time t is determined by linear interpolation

- Ureum concentration is introduced as a time-varying covariate in the semiparametric PH model :

$$h_i(t) = h_0(t)\exp(\beta x_i(t)),$$

where

  - $h_i(t) =$ hazard of first insemination at time t for cow i having at time t ureum concentration equal to $x_i(t)$
  - $\beta =$ linear effect of the ureum concentration on the log-hazard of first insemination

- $\hat{\beta} = -0.0273 \qquad (\text{s.e.} = 0.0162)$

$HR = \exp(-0.0273) = 0.973$

95% CI $= [0.943, 1.005]$

p-value $= 0.094$

Model building strategies for the semiparametric PH model

- Often not clear what criteria should be used to decide which covariates should be included
- Should be based first on meaningful interpretation and biological knowledge
- Different strategies exist :
  - Forward selection
  - Backward selection
  - Forward stepwise selection
  - Backward stepwise selection
  - AIC selection

- Forward procedure :
  - First, include the covariate with the smallest p-value
  - Next, consider all possible models containing the selected covariate and one additional covariate, and include the covariate with the smallest p-value
  - Continue doing this until all remaining non-selected covariates are non-significant
- Backward procedure :
  - First, start from the full model that includes all covariates
  - Next, consider all possible models containing all covariates except one, and remove the covariate with the largest p-value
  - Continue doing this until all remaining covariates in the model are significant

- Forward / backward stepwise procedure :

  Start as in the forward / backward procedure, but an included / removed covariate can be excluded / included at a later stage, if it is no longer significant / non-significant with other covariates in the model

- Note that the above p-values can be based on either the Wald, likelihood ratio or score test

- Akaike's information criterion (AIC) : instead of including / removing covariates based on their p-value, we look at the AIC :

  $$AIC = -2\log(L) + kp$$

  where

  - $p =$ number of parameters in the model
  - $L =$ likelihood
  - $k =$ constant (often 2)

Example : Model building in the schizophrenic patients dataset

- Univariate models :

Marital status	$p = 6.7 \times 10^{-7}$
Gender	$p = 9.7 \times 10^{-5}$
Educational status	$p = 0.663$
Age	$p = 0.9$

- Forward procedure :
  - Start with a model containing marital status
  - Fit model containing marital status and one of the three remaining covariates
    $\Rightarrow$ Gender has smallest p-value
  - Fit model containing marital status, gender and one of the two remaining covariates
    $\Rightarrow$ None of the remaining covariates (educational status and age) is significant
    $\Rightarrow$ Final model contains marital status and gender

Survival function estimation in the semiparametric model

- Survival function for subject with covariate x_i :

$$\begin{aligned} S_i(t) &= \exp(-H_i(t)) \\ &= \exp(-H_0(t)\exp(\beta^t x_i)) \\ &= (S_0(t))^{\exp(\beta^t x_i)} \end{aligned}$$

  with $S_0(t) = \exp(-H_0(t))$ and $H_0(t) = \int_0^t h_0(s)ds$

- Estimate the baseline cumulative hazard $H_0(t)$ by

$$\hat{H}_0(t) = \sum_{j:y_{(j)} \le t} \hat{h}_{0(j)},$$

  where

$$\hat{h}_{0(j)} = \frac{d_{(j)}}{\sum_{k \in R(y_{(j)})} \exp\left(x_k^t \hat{\beta}\right)}$$

  extends the Breslow estimator to the case of tied observations

- Define

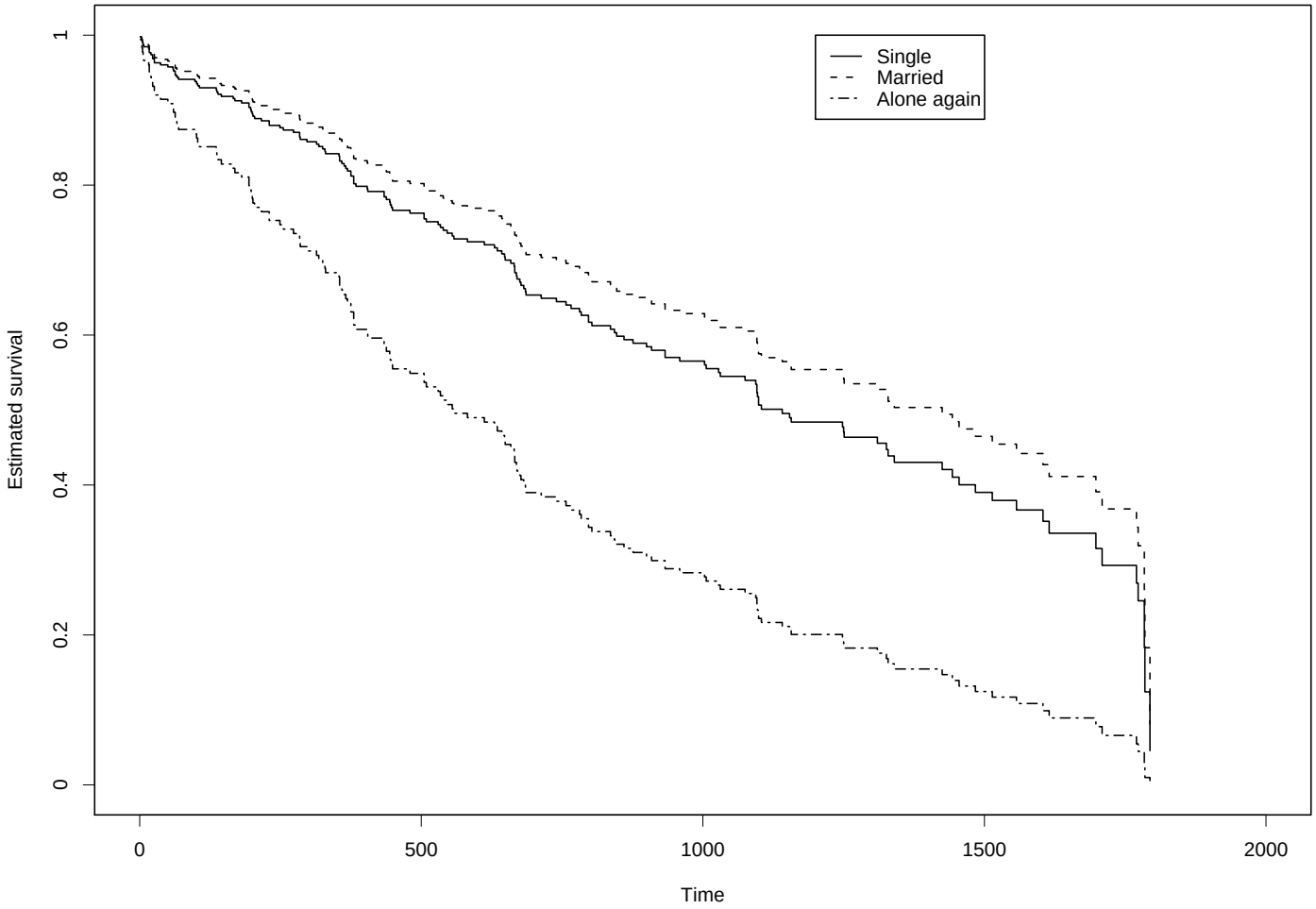
$$\hat{S}_i(t) = \left(\hat{S}_0(t)\right)^{\exp(\hat{\beta}^t x_i)},$$

with $\hat{S}_0(t) = \exp(-\hat{H}_0(t))$

- It can be shown that

$$\frac{\hat{S}_i(t) - S_i(t)}{V^{1/2}(\hat{S}_i(t))} \xrightarrow{d} N(0,1)$$

Example : Survival function estimates for marital status groups in the schizophrenic patients data



Consider e.g. survival at 505 days :

Single group :	0.755	95% CI : [0.690, 0.827]
Married group :	0.796	95% CI : [0.730, 0.867]
Alone again group :	0.537	95% CI : [0.453, 0.636]

Stratified semiparametric PH model

- The assumption that $h_0(t)$ is the same for all subjects might be too strong in practice
  $\Rightarrow$ Possible solution : consider groups (strata) of subjects with the same baseline hazard
- Stratified PH model : the hazard of subject j ($j = 1, \ldots, n_i$) in stratum i ($i = 1, \ldots, s$) is given by

$$h_{ij}(t) = h_{i0}(t) \exp\left(x_{ij}^t \beta\right)$$

- Extension of the partial likelihood :

$$L(\beta) = \prod_{i=1}^{s} \prod_{j=1}^{n_i} \left[\frac{\exp(x_{ij}^t \beta)}{\sum_{l \in R_i(y_{ij})} \exp(x_{il}^t \beta)} \right]^{\delta_{ij}}$$

  $\Rightarrow$ Risk set for a subject contains only the subjects still at risk within the same stratum

Example : Stratified PH model for the time to first insemination dataset

- Cows are coming from different farms
  $\Rightarrow$ baseline hazard might differ considerably between farms (even if the effect of the ureum concentration is similar)
- Consider the effect of the ureum concentration in milk on the time to first insemination, stratifying on the farms :

$$\hat{\beta} = -0.0588 \quad (\text{s.e.} = 0.0198)$$

$$HR = 0.943 \quad 95\%\ \text{CI} = [0.907, 0.980]$$

  $\Rightarrow$ By stratifying on the farms, ureum concentration becomes significant

Checking the proportional hazards assumption

- PH assumption : HR between two subjects with different covariates is constant over time
- Formal tests and diagnostic plots have been developed to check this assumption
- Formal test :
  - Add $\beta_l x_i \times t$ to the PH model :

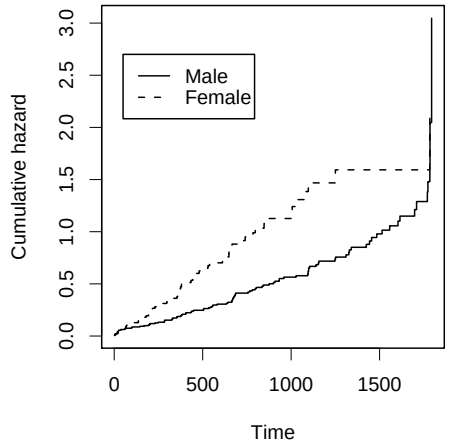
$$h_i(t) = h_0(t) \exp(\beta x_i + \beta_l x_i \times t)$$

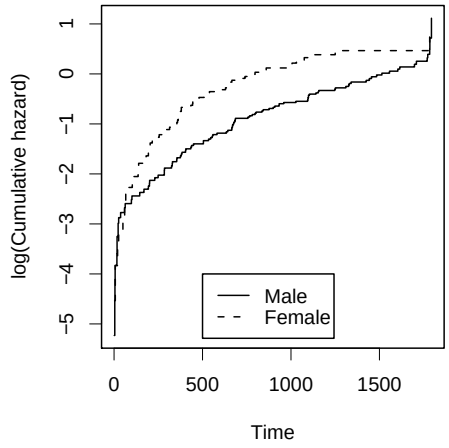
  - If $\beta_l \neq 0$, the PH assumption does not hold
  - Instead of adding $\beta_l x_i \times t$, one can also add $\beta_l x_i \times g(t)$ for some function g

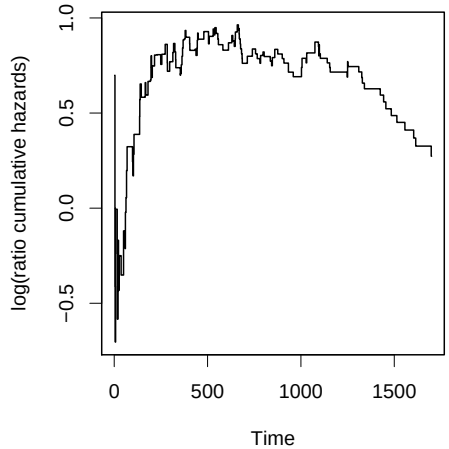
- Diagnostic plots :
  - Consider for simplicity the case of a covariate with r levels
  - Estimate the cumulative hazard function for each level of the covariate by means of the Nelson-Aalen estimator $\Rightarrow \hat{H}_1(t), \hat{H}_2(t), \ldots, \hat{H}_r(t)$ should be constant multiples of each other :

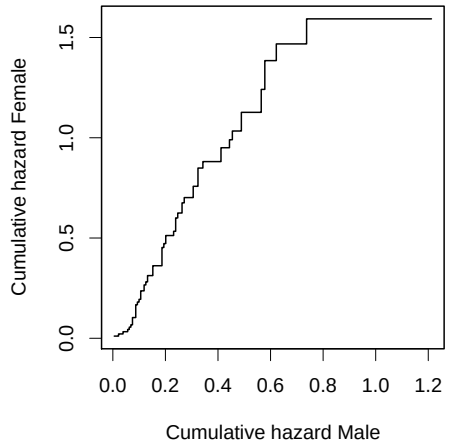
Plot	PH assumption holds if
$\log(\hat{H}_1(t)), ..., \log(\hat{H}_r(t))$ vs t	parallel curves
$\log(\hat{H}_j(t)) - \log(\hat{H}_1(t))$ vs t	constant lines
$\hat{H}_j(t)$ vs $\hat{H}_1(t)$	straight lines through origin

Example : PH assumption for the gender effect in the schizophrenic patients dataset









Parametric survival models

Some common parametric distributions

Exponential distribution :

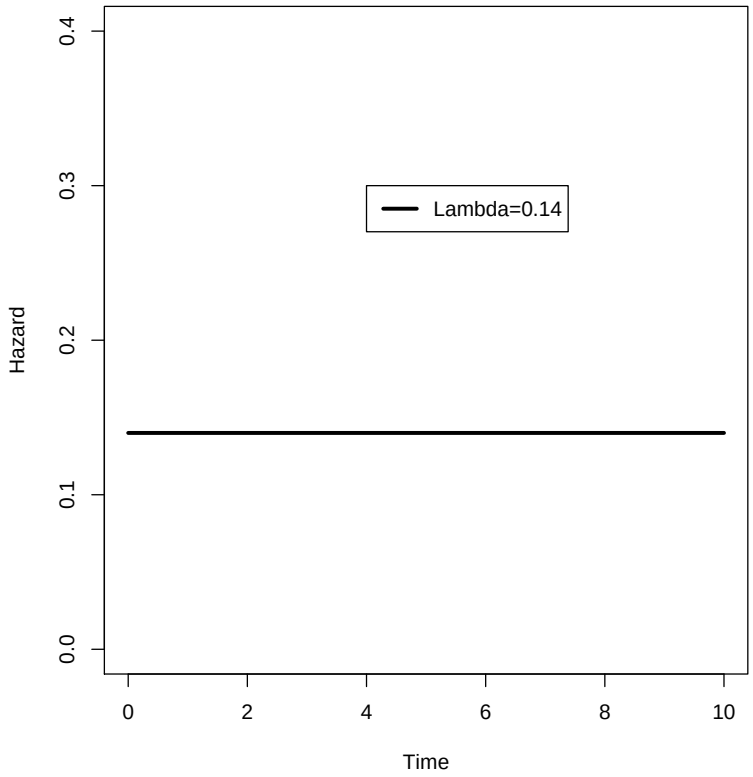
- Characterized by one parameter $\lambda > 0$:

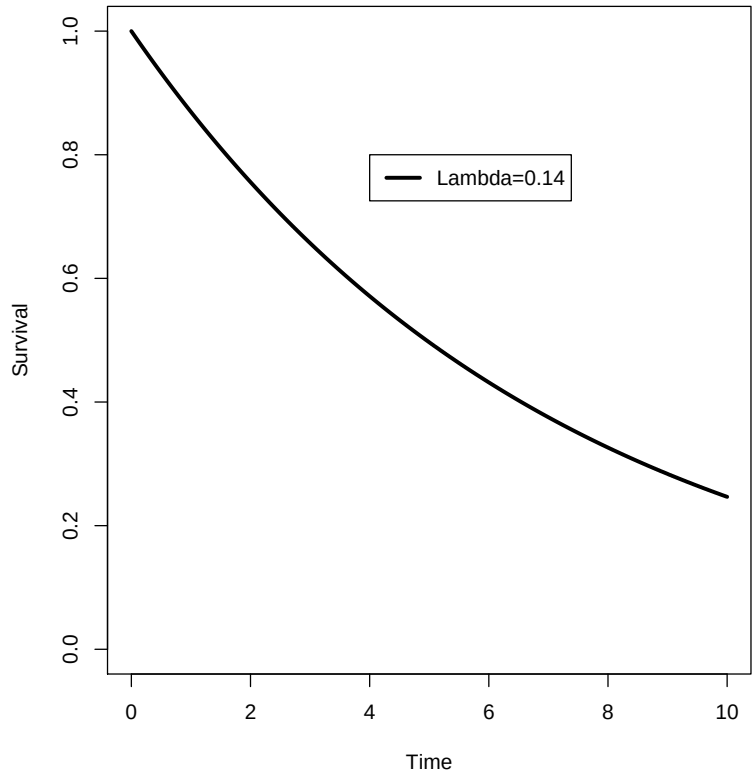
$$\begin{aligned} S_0(t) &= \exp(-\lambda t) \\ f_0(t) &= \lambda \exp(-\lambda t) \\ h_0(t) &= \lambda \end{aligned}$$

  $\rightarrow$ leads to a constant hazard function

- Empirical check : plot of the log of the survival estimate versus time

Hazard and survival function for the exponential distribution





Weibull distribution :

- Characterized by a scale parameter $\lambda > 0$ and a shape parameter $\rho > 0$:

$$\begin{aligned} S_0(t) &= \exp(-\lambda t^\rho) \\ f_0(t) &= \rho\lambda t^{\rho-1} \exp(-\lambda t^\rho) \\ h_0(t) &= \rho\lambda t^{\rho-1} \end{aligned}$$

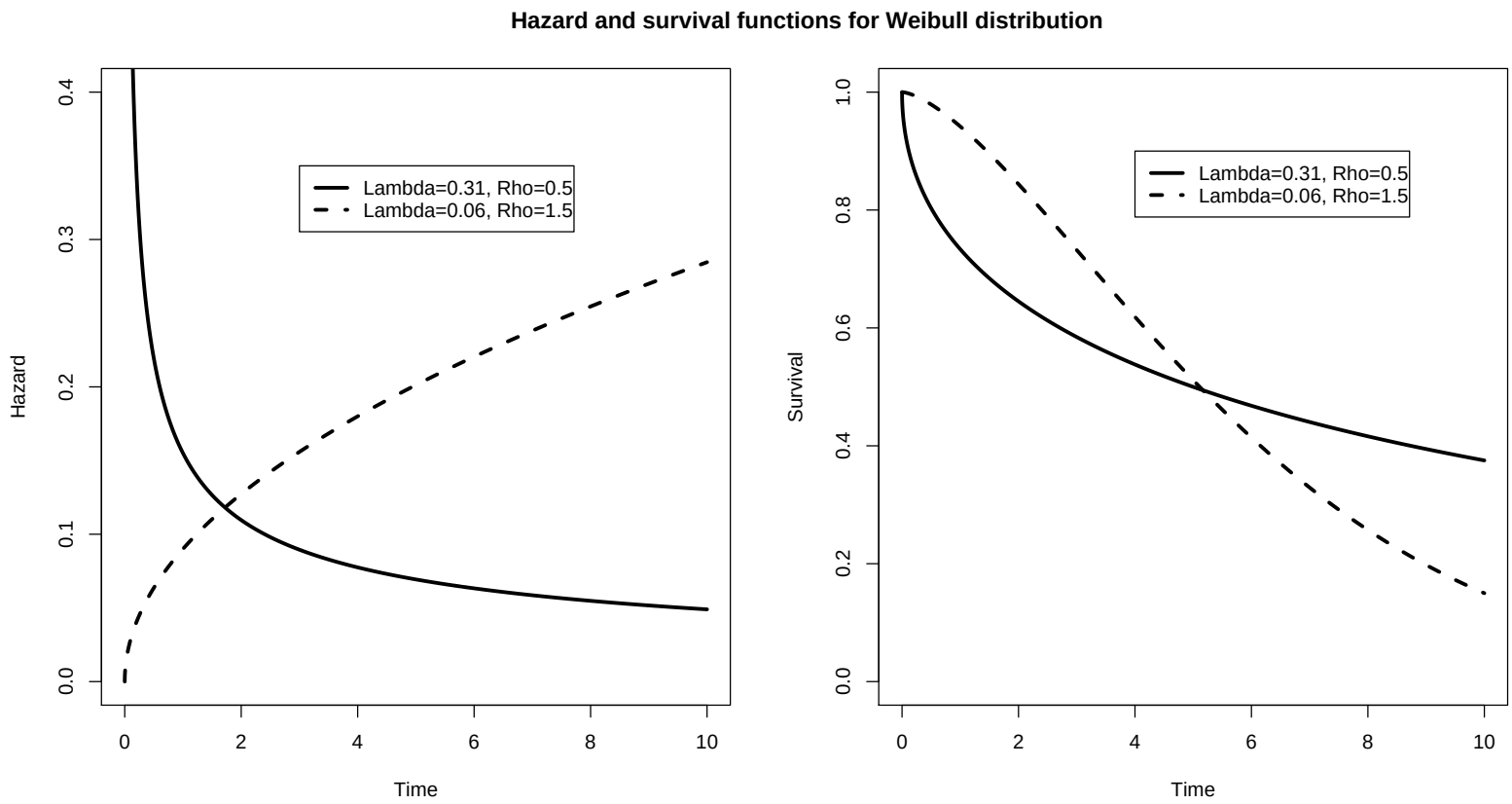
  $\rightarrow$ hazard decreases if $\rho < 1$

  $\rightarrow$ hazard increases if $\rho > 1$

  $\rightarrow$ hazard is constant if $\rho = 1$ (exponential case)

- Empirical check : plot log cumulative hazard versus log time

Hazard and survival function for the Weibull distribution



Log-logistic distribution :

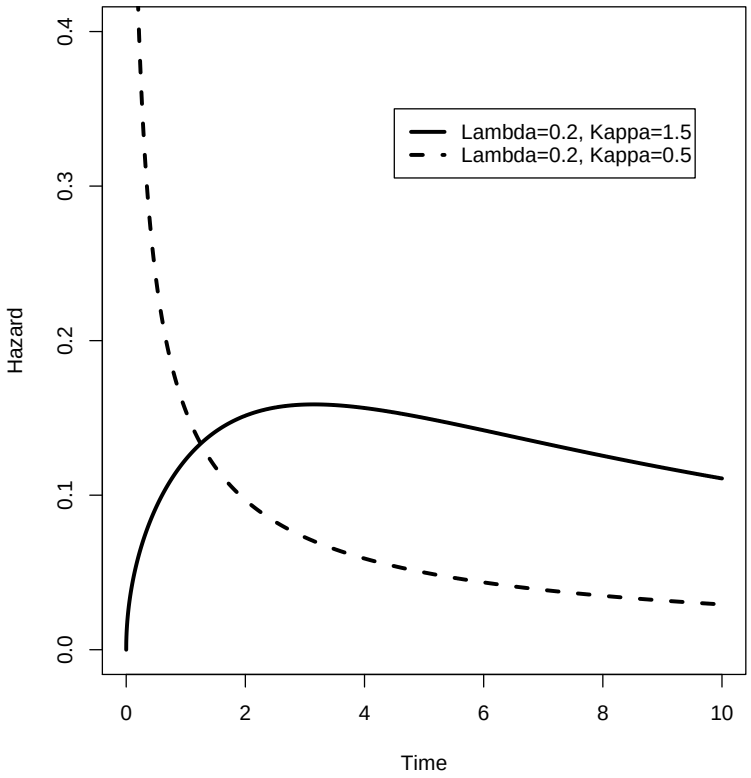
- A random variable T has a log-logistic distribution if $logT$ has a logistic distribution
- Characterized by two parameters λ and $\kappa > 0$:

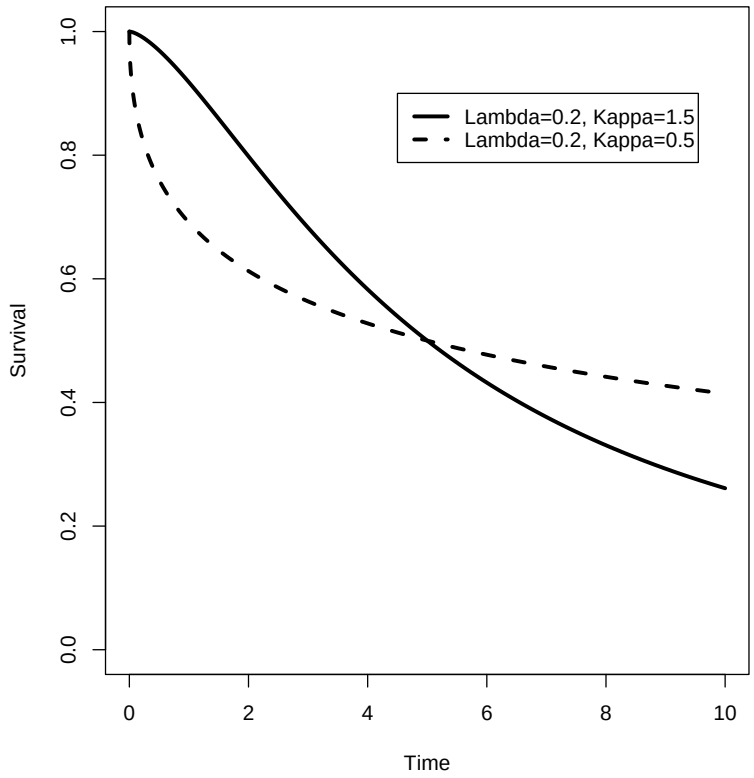
$$
\begin{aligned}
S_0(t) &= \frac{1}{1+(t\lambda)^\kappa} \\
f_0(t) &= \frac{\kappa t^{\kappa-1}\lambda^\kappa}{[1+(t\lambda)^\kappa]^2} \\
h_0(t) &= \frac{\kappa t^{\kappa-1}\lambda^\kappa}{1+(t\lambda)^\kappa}
\end{aligned}
$$

- The median event time is only a function of the parameter λ :

$$M(T) = \exp(1/\lambda)$$

Hazard and survival function for the log-logistic distribution





Log-normal distribution :

- Resembles the log-logistic distribution but is mathematically less tractable
- A random variable T has a log-normal distribution if $logT$ has a normal distribution
- Characterized by two parameters μ and $\gamma > 0$:

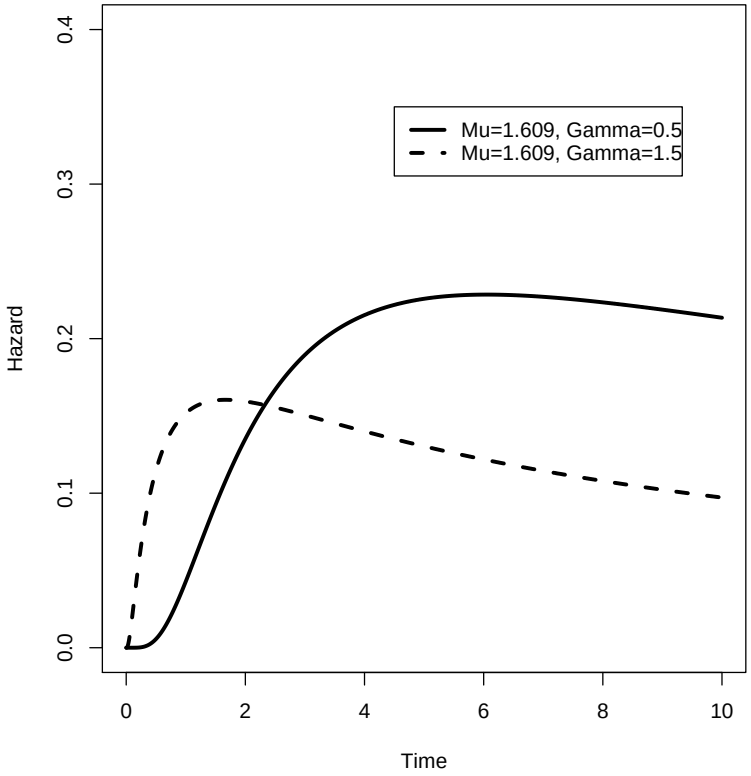
$$S_0(t) = 1 - F_N\left(\frac{\log(t) - \mu}{\sqrt{\gamma}}\right)$$

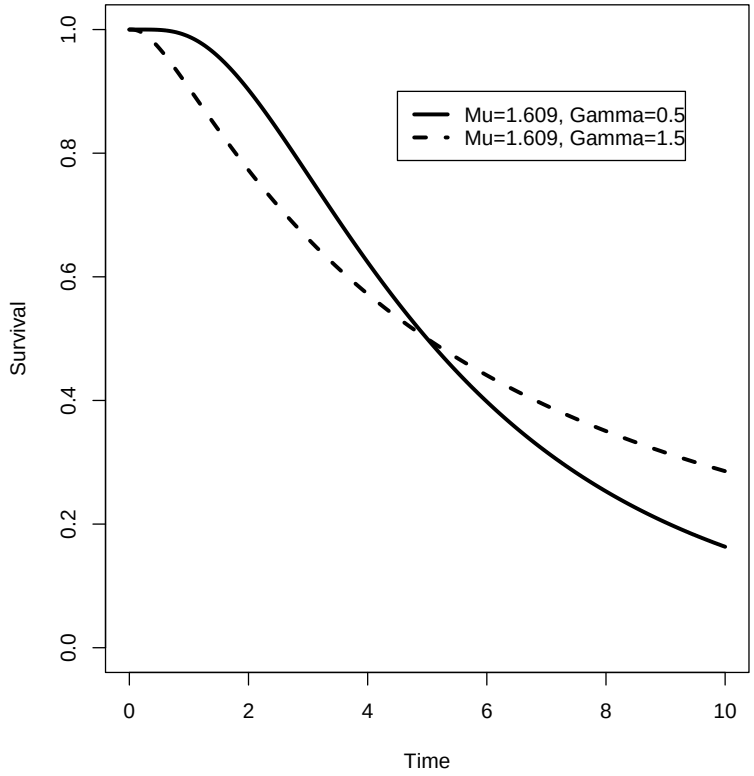
$$f_0(t) = \frac{1}{t\sqrt{2\pi\gamma}} \exp\left[-\frac{1}{2\gamma}(\log(t) - \mu)^2\right]$$

- The median event time is only a function of the parameter μ :

$$M(T) = \exp(\mu)$$

Hazard and survival function for the log-normal distribution





Parametric survival models

The parametric models considered here have two representations :

◇ Accelerated failure time model (AFT) :

$$S_i(t) = S_0(\exp(\theta^t x_i)t),$$

where

- $\theta = (\theta_1, \ldots, \theta_p)^t =$ vector of regression coefficients
- $\exp(\theta^t x_i) =$ acceleration factor
- S_0 belongs to a parametric family of distributions

Hence,

$$h_i(t) = \exp\left(\theta^t x_i\right) h_0\left(\exp(\theta^t x_i)t\right)$$

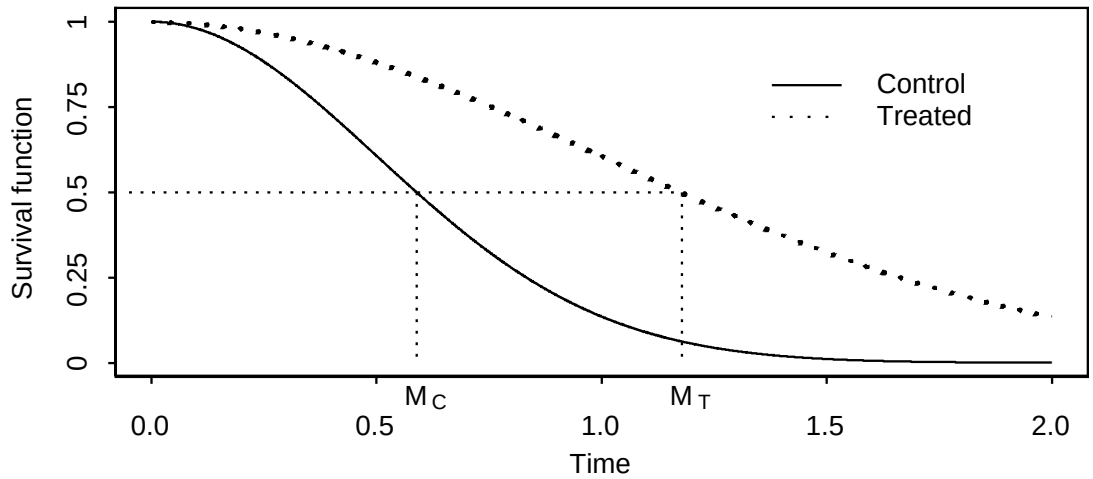
and

$$M_i = \exp(-\theta^t x_i) M_0$$

where M_i = median of S_i, since

$$S_0(M_0) = \frac{1}{2} = S_i(M_i) = S_0\big(\exp(\theta^t x_i) M_i\big)$$

Ex : For one binary variable (say treatment (T) and control (C)), we have $M_T = \exp(-\theta) M_C$:



- Linear model :

$$\log t_i = \mu + \gamma^t x_i + \sigma w_i,$$

where

  - $\mu =$ intercept
  - $\gamma = (\gamma_1, \ldots, \gamma_p)^t =$ vector of regression coefficients
  - $\sigma =$ scale parameter
  - W has known distribution

- These two models are equivalent, if we choose
  - $S_0 =$ survival function of $\exp(\mu + \sigma W)$
  - $\theta = -\gamma$

Indeed,

$$\begin{aligned} S_i(t) &= P(t_i > t) \\ &= P(\log t_i > \log t) \\ &= P(\mu + \sigma w_i > \log t - \gamma^t x_i) \\ &= S_0\big(\exp(\log t - \gamma^t x_i)\big) \\ &= S_0\big(t \exp(\theta^t x_i)\big) \end{aligned}$$

$\Rightarrow$ The two models are equivalent

Weibull distribution

- Consider the accelerated failure time model

$$S_i(t) = S_0\big(\exp(\theta^t x_i)t\big),$$

where $S_0(t) = \exp(-\lambda t^\alpha)$ is Weibull

$$\Rightarrow S_i(t) = \exp\big(-\lambda \exp(\beta^t x_i) t^\alpha\big) \text{ with } \beta = \alpha\theta$$

$$\Rightarrow f_i(t) = \lambda\alpha t^{\alpha-1} \exp(\beta^t x_i) \exp\big(-\lambda \exp(\beta^t x_i) t^\alpha\big)$$

$$\Rightarrow h_i(t) = \alpha\lambda t^{\alpha-1} \exp(\beta^t x_i) = h_0(t) \exp(\beta^t x_i),$$

with $h_0(t) = \alpha\lambda t^{\alpha-1}$ the hazard of a Weibull

$\Rightarrow$ We also have a Cox PH model

- The above model is also equivalent to the following linear model :

$$\log t_i = \mu + \gamma^t x_i + \sigma w_i,$$

where W has a standard extreme value distribution, i.e. $S_W(w) = \exp(-e^w)$. Indeed,

$$\begin{aligned} P(W > w) &= P\big(\exp(\mu + \sigma W) > \exp(\mu + \sigma w)\big) \\ &= S_0\big(\exp(\mu + \sigma w)\big) \\ &= \exp\big(-\lambda \exp(\alpha\mu + \alpha\sigma w)\big) \end{aligned}$$

Since W has a known distribution, it follows that $\lambda \exp(\alpha\mu) = 1$ and $\alpha\sigma = 1$, and hence

$$P(W > w) = \exp(-e^w)$$

- It follows that

  Weibull accelerated failure time model

  = Cox PH model with Weibull baseline hazard

  = Linear model with standard extreme value error distribution

  and

  - $\theta = -\gamma = \beta/\alpha$
  - $\alpha = 1/\sigma$
  - $\lambda = \exp(-\mu/\sigma)$

- Note that the Weibull distribution is the only continuous distribution that can be written as an AFT model and as a PH model

Log-logistic distribution

- Consider the accelerated failure time model

$$S_i(t) = S_0\big(\exp(\theta^t x_i)t\big),$$

where $S_0(t) = 1/[1 + \lambda t^\alpha]$ is log-logistic

$$\Rightarrow S_i(t) = \frac{1}{1 + \lambda \exp(\beta^t x_i) t^\alpha} \text{ with } \beta = \alpha\theta$$

$$\Rightarrow \frac{S_i(t)}{1 - S_i(t)} = \frac{1}{\lambda \exp(\beta^t x_i) t^\alpha}$$

$$= \exp(-\beta^t x_i) \frac{S_0(t)}{1 - S_0(t)}$$

$\Rightarrow$ We also have a so-called proportional odds model

- The above model is also equivalent to the following linear model :

$$\log t_i = \mu + \gamma^t x_i + \sigma w_i,$$

where W has a standard logistic distribution, i.e. $S_W(w) = 1/[1 + \exp(w)]$. Indeed,

$$\begin{aligned} P(W > w) &= P\big(\exp(\mu + \sigma W) > \exp(\mu + \sigma w)\big) \\ &= S_0\big(\exp(\mu + \sigma w)\big) \\ &= 1/\big[1 + \lambda \exp(\alpha\mu + \alpha\sigma w)\big] \end{aligned}$$

Since W has a known distribution, it follows that $\lambda \exp(\alpha\mu) = 1$ and $\alpha\sigma = 1$, and hence

$$P(W > w) = \frac{1}{1 + \exp(w)}$$

- It follows that

  Log-logistic accelerated failure time model

  = Proportional odds model with log-logistic baseline survival

  = Linear model with standard logistic error distribution

  and

  - $\theta = -\gamma = \beta/\alpha$
  - $\alpha = 1/\sigma$
  - $\lambda = \exp(-\mu/\sigma)$

- Note that the log-logistic distribution is the only continuous distribution that can be written as an AFT model and as a proportional odds model

Other distributions

- Log-normal :

  Log-normal accelerated failure time model
  = Linear model with standard normal error distribution

- Generalized gamma :

  t_i follows a generalized gamma distribution if

  $$\log t_i = \mu + \gamma^t x_i + \sigma w_i,$$

  where w_i has the following density :

  $$f_W(w) = \frac{|\theta| \left(\theta^{-2} \exp(\theta w)\right)^{1/\theta^2} \exp\left(-\theta^{-2} \exp(\theta w)\right)}{\Gamma(1/\theta^2)}$$

  If $\theta = 1 \Rightarrow$ Weibull model

  If $\theta = 1$ and $\sigma = 1 \Rightarrow$ exponential model

  If $\theta \to 0 \Rightarrow$ log-normal model

Estimation

- It suffices to estimate the model parameters in one of the equivalent model representations. Consider e.g. the linear model :

$$\log t_i = \mu + \gamma^t x_i + \sigma w_i$$

- The likelihood function for right censored data equals

$$\begin{aligned} L(\mu, \gamma, \sigma) &= \prod_{i=1}^{n} f_i(y_i)^{\delta_i} S_i(y_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sigma y_i} f_W\left(\frac{\log y_i - \mu - \gamma^t x_i}{\sigma} \right) \right]^{\delta_i} \\ &\quad \times \left[S_W\left(\frac{\log y_i - \mu - \gamma^t x_i}{\sigma} \right) \right]^{1-\delta_i} \end{aligned}$$

Since W has a known distribution, this likelihood can be maximized w.r.t. its parameters μ, γ, σ

- Let

$$(\hat{\mu}, \hat{\gamma}, \hat{\sigma}) = \text{argmax}_{\mu,\gamma,\sigma} L(\mu, \gamma, \sigma)$$

- It can be shown that
  - $(\hat{\mu}, \hat{\gamma}, \hat{\sigma})$ is asymptotically unbiased and normal
  - The estimators of the accelerated failure time model (or any other equivalent model) and their asymptotic distribution can be obtained from the Delta-method

Model selection

To select the best parametric model, we present two methods

- Selection of nested models :
  Consider the generalized gamma model as the 'full' model, and test whether
  - $\theta = 1 \Rightarrow$ Weibull model
  - $\theta = 1$ and $\sigma = 1 \Rightarrow$ exponential model
  - $\theta = 0 \Rightarrow$ log-normal model

  The test can be done using the Wald, likelihood ratio or score test statistic derived from the likelihood for censored data

- AIC selection :

$$AIC = -2 \log L + 2(p + 1 + k),$$

where

- $p + 1 =$ dimension of (μ, γ)
- $k = 0$ for the exponential model
- $k = 1$ for the Weibull, log-logistic, log-normal model
- $k = 2$ for the generalized gamma model

and minimize the AIC among all candidate parametric models

The End