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

- For survival data, we consider rather
  \[ S(t) \]  
  \[ H(t) \]  
  \[ h(t) \]  
  \[ mrl(t) \]  
  survival function
  cumulative hazard function
  hazard function
  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):

\[ h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \]

\[ = \frac{1}{P(T \geq t)} \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \]

\[ = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t) = \frac{d}{dt} H(t) \]

- \( h(t) \) 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
Basic concepts

Nonparametric estimation

Hypothesis testing in a nonparametric setting

Proportional hazards models

Parametric survival models

Hazard functions of different shapes

- Exponential
- Weibull, $\rho=0.5$
- Weibull, $\rho=1.5$
- Bathtub
Mean residual life function:

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

This result is obtained from

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

Mean life time:

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

\[
T = \text{survival time} \\
C = \text{censoring time}
\]

⇒ Data : \((Y, \delta)\) with

\[
Y = \min(T, C) \\
\delta = I(T \leq C)
\]
Type I right censoring

- All subjects are followed for a fixed amount of time → 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).

→ 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
  - censoring is a random variable
  - 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

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

[Diagram showing patient data over time]

January 1, 2005 to December 31, 2008
**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

⇒ Data: \((Y_\ell, \delta_\ell)\) with

\[
Y_\ell = \max(T, C_\ell) \\
\delta_\ell = \mathbb{I}(T > C_\ell) \\
C_\ell = \text{censoring time}
\]

- 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
  → 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
  - 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

⇒ 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
  \[ T = \text{survival time} \]
  \[ R = \text{right truncation time} \]
- 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
    \[ \Rightarrow \text{Patients with long incubation time will not be part of the sample, nor patients that die from another cause before they develop AIDS} \]
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 $n$:
  
  $T_1, \ldots, T_n$ \hspace{1cm} survival time
  
  $C_1, \ldots, C_n$ \hspace{1cm} censoring time

  $\Rightarrow$ Observed data: $(Y_i, \delta_i) \ (i = 1, \ldots, n)$ 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 \((y_i = t_i, \delta_i = 1)\):

\[
\lim_{\epsilon \to 0} \frac{1}{2\epsilon} P(y_i - \epsilon < Y < y_i + \epsilon, \delta = 1) >
\]

\[
= \lim_{\epsilon \to 0} \frac{1}{2\epsilon} P(y_i - \epsilon < T < y_i + \epsilon, T \leq C)
\]

\[
= \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int \int dG(c) dF(t) \quad \text{(due to independence)}
\]

\[
= \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_{y_i - \epsilon}^{y_i + \epsilon} (1 - G(t)) dF(t)
\]

\[
= (1 - G(y_i)) f(y_i)
\]
Contribution to the likelihood of a right censored observation \((y_i = c_i, \delta_i = 0)\):

\[
\lim_{\epsilon \to 0} \frac{1}{2\epsilon} P (y_i - \epsilon < Y < y_i + \epsilon, \delta = 0) > 2 \\
= \lim_{\epsilon \to 0} \frac{1}{2\epsilon} P (y_i - \epsilon < C < y_i + \epsilon, T > C) > \\
= (1 - F(y_i))g(y_i)
\]

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.

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

⇒ 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 $T_i \geq a_i$:

$$f(t_i | T \geq a_i) = \lim_{\epsilon \to 0} \frac{1}{2\epsilon} P(t_i - \epsilon < T < t_i + \epsilon | T \geq a_i)$$

$$= \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \frac{P(t_i - \epsilon < T < t_i + \epsilon, T \geq a_i)}{P(T \geq a_i)}$$

$$= \frac{1}{P(T \geq a_i)} \lim_{\epsilon \to 0} \frac{P(t_i < T < t_i + \epsilon)}{\epsilon}$$

$$= \frac{f(t_i)}{S(a_i)}$$
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:

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

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

$$\hat{S}(t) = \frac{R(2)}{R(1)} \frac{R(3)}{R(2)} \cdots \frac{R(k+1)}{R(k)}$$

where $y(k) \leq t < y(k+1)$

$$= \frac{R(k+1)}{R(1)}$$

$$= \frac{\# \text{ subjects with survival time} \geq y(k+1)}{\# \text{ at risk before first death time}}$$

$$= \frac{1}{n} \sum_{i=1}^{n} I(y_i > t)$$
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

  \[
  \hat{H}(t) = \sum_{j: y(j) \leq t} \frac{d(j)}{R(j)} \quad \text{for } t \leq y(r)
  \]

- Its asymptotic variance can be estimated by
  \[
  \hat{V}_{As}(\hat{H}(t)) = \sum_{j: y(j) \leq t} \frac{d(j)}{R^2(j)}
  \]

- 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:

<table>
<thead>
<tr>
<th>Event time</th>
<th>6</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of events</td>
<td>$d_{(j)}$</td>
<td>1</td>
</tr>
<tr>
<td>Number at risk</td>
<td>$R_{(j)}$</td>
<td>4</td>
</tr>
<tr>
<td>KM contribution</td>
<td>$1 - d_{(j)}/R_{(j)}$</td>
<td>3/4</td>
</tr>
<tr>
<td>KM estimator</td>
<td>$\hat{S}(y_{(j)})$</td>
<td>3/4=0.75</td>
</tr>
<tr>
<td>NA contribution</td>
<td>$\exp\left(-d_{(j)}/R_{(j)}\right)$</td>
<td>0.7788</td>
</tr>
<tr>
<td>NA estimator</td>
<td>$\prod_{j:y_{(j)} \leq t} \exp\left(-d_{(j)}/R_{(j)}\right)$</td>
<td>0.7788</td>
</tr>
</tbody>
</table>
Basic concepts
Nonparametric estimation
Hypothesis testing in a nonparametric setting
Proportional hazards models
Parametric survival models

Estimated survival

Kaplan–Meier
Nelson–Aalen

Time

0 5 10 15 20 25 30 35
0.0 0.2 0.4 0.6 0.8 1.0

0 5 10 15 20 25 30 35
0.0 0.2 0.4 0.6 0.8 1.0

Time

Estimated survival

Kaplan–Meier
Nelson–Aalen
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 \(\infty\).

One can show that
\[
\frac{\log(-\log \hat{S}(t)) - \log(-\log S(t))}{\sqrt{\hat{V}_{As}(\log(-\log \hat{S}(t)))}} \overset{d}{\rightarrow} N(0, 1),
\]
where
\[
\hat{V}_{As}(\log(-\log \hat{S}(t))) = \frac{1}{(\log \hat{S}(t))^2} \sum_{j:y(j)\leq t} \frac{d(j)}{R(j)(R(j) - d(j))}
\]
Hence, CI for \( \log(-\log S(t)) \) is given by
\[
\log(-\log \hat{S}(t)) \pm z_{\alpha/2} \sqrt{\hat{V}_{As}(\log(-\log \hat{S}(t)))}
\]
By transforming back, we get the following CI for \( S(t) \):
\[
\hat{S}(t) \exp \left[ \pm z_{\alpha/2} \sqrt{\hat{V}_{As}(\log(-\log \hat{S}(t)))} \right]
\]
Point estimate of the mean survival time

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

\[ \mu = E(T) = \int_{0}^{\infty} xf(x) \, dx = \int_{0}^{\infty} S(x) \, dx \]

\[ \Rightarrow \text{We can estimate } \mu \text{ by replacing } S(x) \text{ by the KM estimator } \hat{S}(x) \]

- 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 \([0, t_{\text{max}}]\) and consider \( \hat{S}(t) = \hat{S}(y_n) \) for \( y_n \leq t \leq t_{\text{max}} \):
    \[ \hat{\mu}_{t_{\text{max}}} = \int_{0}^{t_{\text{max}}} \hat{S}(t) \, dt \]
\( \hat{\mu}_{y_n} \) and \( \hat{\mu}_{t_{\text{max}}} \) are inconsistent estimators of \( \mu \), but given the lack of data in the right tail, we cannot do better (at least not nonparametrically).

**Variance of \( \hat{\mu}_\tau \) (with \( \tau \) either \( y_n \) or \( t_{\text{max}} \)) :**

\[
\hat{\mathcal{V}}_{\text{As}}(\hat{\mu}_\tau) = \sum_{j=1}^{r} \left( \int_{y(j)}^{\tau} \hat{S}(t)dt \right)^2 \frac{d(j)}{R(j)(R(j) - d(j))}
\]

**A 100(1 - \( \alpha \))% CI for \( \mu \) is given by :**

\[
\hat{\mu}_\tau \pm z_{\alpha/2} \sqrt{\hat{\mathcal{V}}_{\text{As}}(\hat{\mu}_\tau)}
\]
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:
  $$\hat{x}_p = \inf \left\{ t \mid \hat{S}(t) \leq 1 - p \right\}$$
  ⇒ 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
⇒ We prefer not to use this variance estimator in the construction of a CI

Thanks to the asymptotic normality of $\hat{S}(x_\rho)$:

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

with obviously $S(x_\rho) = 1 - \rho$.

⇒ A $100(1 - \alpha)$% CI for $x_\rho$ is given by

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

◊ Schizophrenia is one of the major mental illnesses encountered in Ethiopia
  → disorganized and abnormal thinking, behavior and language + emotionally unresponsive
  → higher mortality rates due to natural and unnatural causes

◊ Project on schizophrenia in Butajira, Ethiopia
  → survey of the entire population (68491 individuals) in the age group 15-49 years

⇒ 280 cases of schizophrenia identified and followed for 5 years (1997-2001)
Table: Data on schizophrenia patients

<table>
<thead>
<tr>
<th>Patid</th>
<th>Time</th>
<th>Censor</th>
<th>Education</th>
<th>Onset</th>
<th>Marital</th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>37</td>
<td>3</td>
<td>1</td>
<td>44</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>26</td>
<td>1</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1</td>
<td>12</td>
<td>25</td>
<td>1</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>29</td>
<td>3</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>278</td>
<td>1787</td>
<td>0</td>
<td>2</td>
<td>16</td>
<td>2</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>279</td>
<td>1792</td>
<td>0</td>
<td>2</td>
<td>23</td>
<td>1</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>280</td>
<td>1794</td>
<td>1</td>
<td>2</td>
<td>28</td>
<td>1</td>
<td>1</td>
<td>35</td>
</tr>
</tbody>
</table>
In R: `survfit`

```r
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`

```sas
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;
```
Kaplan–Meier estimate of the survival function for Schizophrenic patients (confidence interval based on log–log transformation)
Nonparametric estimation

Proportional hazards models

Parametric survival models

Basic concepts

Hypothesis testing in a nonparametric setting

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

<table>
<thead>
<tr>
<th>n</th>
<th>events</th>
<th>median</th>
<th>0.95LCL</th>
<th>0.95UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>280</td>
<td>163</td>
<td>933</td>
<td>757</td>
<td>1099</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>time</th>
<th>n.risk</th>
<th>n.event</th>
<th>survival</th>
<th>std.err</th>
<th>lower 95% CI</th>
<th>upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>280</td>
<td>1</td>
<td>0.996</td>
<td>0.00357</td>
<td>0.9749</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td>279</td>
<td>1</td>
<td>0.993</td>
<td>0.00503</td>
<td>0.9717</td>
<td>0.998</td>
</tr>
<tr>
<td>4</td>
<td>277</td>
<td>1</td>
<td>0.989</td>
<td>0.00616</td>
<td>0.9671</td>
<td>0.997</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1770</td>
<td>13</td>
<td>1</td>
<td>0.219</td>
<td>0.03998</td>
<td>0.1465</td>
<td>0.301</td>
</tr>
<tr>
<td>1773</td>
<td>12</td>
<td>1</td>
<td>0.201</td>
<td>0.04061</td>
<td>0.1283</td>
<td>0.285</td>
</tr>
<tr>
<td>1784</td>
<td>8</td>
<td>2</td>
<td>0.151</td>
<td>0.04329</td>
<td>0.0782</td>
<td>0.245</td>
</tr>
<tr>
<td>1785</td>
<td>6</td>
<td>2</td>
<td>0.100</td>
<td>0.04092</td>
<td>0.0387</td>
<td>0.197</td>
</tr>
<tr>
<td>1794</td>
<td>1</td>
<td>1</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Kaplan–Meier estimate of the survival function for Schizophrenic patients (confidence interval based on Greenwood formula)
> 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 $R(j)$:

$$R(j) = \text{number of individuals at risk at time } y(j)$$
and under observation prior to time $y(j)$

$$= \# \{ i : l_i \leq y(j) \leq y_i \},$$

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

$$
\hat{S}_l(t) = \begin{cases} 
1 & \text{if } t < l \\
\prod_{j: l \leq y(j) \leq t} \left(1 - \frac{d(j)}{R(j)}\right) & \text{if } t \geq l
\end{cases}
$$

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 \([t - b, t + b]\):

\[
\hat{h}(t) = (2b)^{-1} \sum_{j=1}^{r} I(-b \leq t - y(j) \leq b) \Delta \hat{H}(y(j)),
\]

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:

<table>
<thead>
<tr>
<th>Name</th>
<th>Density function</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform</td>
<td>( K(x) = \frac{1}{2} )</td>
<td>(-1 \leq x \leq 1)</td>
</tr>
<tr>
<td>Epanechnikov</td>
<td>( K(x) = \frac{3}{4}(1 - x^2) )</td>
<td>(-1 \leq x \leq 1)</td>
</tr>
<tr>
<td>biweight</td>
<td>( K(x) = \frac{15}{16}(1 - x^2)^2 )</td>
<td>(-1 \leq x \leq 1)</td>
</tr>
</tbody>
</table>

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.

⇒ 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

![Graph showing survival differences between transplant + chemotherapy and only chemotherapy.](image)
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 \( h_0(t) \):

\[
H_0 : h(t) = h_0(t) \quad \text{for all } t \leq y_{(r)} \\
H_1 : h(t) \neq h_0(t) \quad \text{for some } t \leq y_{(r)}
\]

◊ 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**:
  \[ Z = \sum_{j=1}^{r} d_{(j)} - \int_{0}^{Y^{(r)}} R(s) h_0(s) ds \]
  \[ = \sum_{j=1}^{r} d_{(j)} - \sum_{i=1}^{n} \int_{0}^{Y_i} h_0(s) ds \]
  \[ = \sum_{j=1}^{r} d_{(j)} - \sum_{i=1}^{n} H_0(y_i) = O - E \]

- **Under $H_0$**:
  \[ V(Z) = \int_{0}^{Y^{(r)}} R(s) h_0(s) ds = E \]
  \[
  \frac{O - E}{\sqrt{E}} \approx N(0, 1)
  \]
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)

<table>
<thead>
<tr>
<th>Age</th>
<th>Follow-up</th>
<th>Status</th>
<th>Gender</th>
<th>Race</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>22</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>53</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>57</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>57</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>85</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>86</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Standardized actuarial table:

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

<table>
<thead>
<tr>
<th>Age</th>
<th>Caucasian male</th>
<th>Caucasian female</th>
<th>black male</th>
<th>black female</th>
</tr>
</thead>
<tbody>
<tr>
<td>50-54</td>
<td>0.6070</td>
<td>0.3608</td>
<td>1.3310</td>
<td>0.7156</td>
</tr>
<tr>
<td>55-59</td>
<td>0.9704</td>
<td>0.5942</td>
<td>1.9048</td>
<td>1.0558</td>
</tr>
<tr>
<td>60-64</td>
<td>1.5855</td>
<td>0.9632</td>
<td>2.8310</td>
<td>1.6048</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>80-84</td>
<td>9.3128</td>
<td>6.2880</td>
<td>10.4625</td>
<td>7.2523</td>
</tr>
<tr>
<td>85-84</td>
<td>17.7671</td>
<td>14.6814</td>
<td>16.0835</td>
<td>13.7017</td>
</tr>
</tbody>
</table>
◊ E.g. first patient: Caucasian female followed from 52 years on for 22 years:

<table>
<thead>
<tr>
<th></th>
<th>hazard for the $52^{th}$ year</th>
<th>=</th>
<th>0.3608</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>hazard for the $53^{th}$ year</td>
<td>=</td>
<td>0.3608</td>
</tr>
<tr>
<td>(2)</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>(22)</td>
<td>hazard for the $73^{th}$ year</td>
<td>=</td>
<td>2.3454</td>
</tr>
</tbody>
</table>

Total (cumulative hazard) = 25.637

⇒ 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):

\[ w(t) = R(t) S_0^p(t)(1 - S_0(t))^q \quad p, q \geq 0 \]

- \( p = q = 0 \) : 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 :

\[ H_0 : h_1(t) = h_2(t) \quad \text{for all } t \leq y(r) \]
\[ H_1 : h_1(t) \neq h_2(t) \quad \text{for some } t \leq y(r) \]

⋄ 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 \( k \) \((j = 1, \ldots, r \text{ and } k = 1, 2)\)
- \( R_{(j)k} \) : 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 $y_{(j)}$:

<table>
<thead>
<tr>
<th>Group</th>
<th>Event</th>
<th>No Event</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$d_{(j)1}$</td>
<td>$R_{(j)1} - d_{(j)1}$</td>
<td>$R_{(j)1}$</td>
</tr>
<tr>
<td>2</td>
<td>$d_{(j)2}$</td>
<td>$R_{(j)2} - d_{(j)2}$</td>
<td>$R_{(j)2}$</td>
</tr>
<tr>
<td>Total</td>
<td>$d_{(j)}$</td>
<td>$R_{(j)} - d_{(j)}$</td>
<td>$R_{(j)}$</td>
</tr>
</tbody>
</table>

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:

\[
U = \sum_{j=1}^{r} w(y_{(j)})(O_j - E_j)
\]

\[
= \sum_{j=1}^{r} w(y_{(j)}) \left( d_{(j)} - \frac{d_{(j)} R_{(j)}}{R_{(j)}} \right)
\]

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,

$$
V(U) = \sum_{j=1}^{r} w^2(y_{(j)}) V(d_{(j)1})
$$

$$
= \sum_{j=1}^{r} w^2(y_{(j)}) \frac{d_{(j)} \left( \frac{R_{(j)1}}{R_{(j)}} \right) \left( 1 - \frac{R_{(j)1}}{R_{(j)}} \right) \left( R_{(j)} - d_{(j)} \right)}{R_{(j)} - 1}
$$
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
\[ w(y_{(j)}) = f(R_{(j)}) \]

\[ \rightarrow \text{Tarone and Ware (1977)} \]

\[ \rightarrow \text{a suggested choice is } f(R_{(j)}) = \sqrt{R_{(j)}} \]

\[ \rightarrow \text{puts more weight on early departures from } H_0 \]

\[ w(y_{(j)}) = \hat{S}(y_{(j)}) = \prod_{y_k \leq y_{(j)}} \left(1 - \frac{d_k}{R_k + 1}\right) \]

\[ \rightarrow \text{Peto and Peto (1972) and Kalbfleisch and Prentice (1980)} \]

\[ \rightarrow \text{based on an estimate of the common survival function close to the pooled product limit estimate} \]

\[ w(y_{(j)}) = \left(\hat{S}(y_{(j-1)})\right)^p \left(1 - \hat{S}(y_{(j-1)})\right)^q \quad p \geq 0, q \geq 0 \]

\[ \rightarrow \text{Fleming and Harrington (1981)} \]

\[ \rightarrow \text{include weights of the log-rank as special case} \]

\[ q = 0, p > 0 : \text{more weight is put on early differences} \]

\[ p = 0, q > 0 : \text{more weight is put on late differences} \]
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:**
  \[ H_0 : h_1(t) = h_2(t) = \ldots = h_l(t) \text{ for all } t \leq y(r) \]
  \[ H_1 : h_i(t) \neq h_j(t) \text{ for at least one pair } (i, j) \text{ for some } t \leq y(r) \]

- **Notations:** same as earlier but now \( k = 1, \ldots, l \)

- **Test statistic based on the \( l \times 2 \) contingency tables for the different event times \( y(j) \)**

<table>
<thead>
<tr>
<th>Group</th>
<th>Event</th>
<th>No Event</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( d_{(j)1} )</td>
<td>( R_{(j)1} - d_{(j)1} )</td>
<td>( R_{(j)1} )</td>
</tr>
<tr>
<td>2</td>
<td>( d_{(j)2} )</td>
<td>( R_{(j)2} - d_{(j)2} )</td>
<td>( R_{(j)2} )</td>
</tr>
<tr>
<td>\ldots</td>
<td>( d_{(j)l} )</td>
<td>( R_{(j)l} - d_{(j)l} )</td>
<td>( R_{(j)l} )</td>
</tr>
<tr>
<td>l</td>
<td>( d_{(j)} )</td>
<td>( R_{(j)} - d_{(j)} )</td>
<td>( R_{(j)} )</td>
</tr>
<tr>
<td>Total</td>
<td>( d_{(j)} )</td>
<td>( R_{(j)} - d_{(j)} )</td>
<td>( R_{(j)} )</td>
</tr>
</tbody>
</table>
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:

$$U_k = \sum_{j=1}^{r} w(y(j)) \left( d(j)_k - \frac{d(j)_k R(j)_k}{R(j)} \right),$$

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

![Graph showing survival comparison]

- **Single**
- **Married**
- **Again alone**
- 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 = $1.5 \times 10^{-7}$ (based on a $\chi^2_2$)
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:

\[ U = \sum_{k=1}^{l} w_k U_k, \]

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 \( U \):

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

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.
  ⇒ Stratified test: obtain an overall assessment of the difference, by combining information over the different strata to gain power.
- Hypothesis test:
  \[ H_0 : h_{1b}(t) = h_{2b}(t) = \ldots = h_{lb}(t) \]
  for all \( t \leq y_{(r)} \) 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 \( k \) \((k = 1, \ldots, l)\) in stratum \( b \) \((b = 1, \ldots, m)\)
- Stratified summary statistic for population \( k \):
  \[ U_k. = \sum_{b=1}^{m} U_{kb} \]
- Define \( U. = (U_1., \ldots, U_{(l-1).})^t \)

Entries of the variance-covariance matrix \( V(U) \) of \( U. \):

\[
\text{Cov}(U_k., U_{k'.}) = \sum_{b=1}^{m} \text{Cov}(U_{kb}, U_{k'b})
\]

For large sample size and under \( H_0 \):

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

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):

<table>
<thead>
<tr>
<th></th>
<th>single</th>
<th>married</th>
<th>alone again</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_b$</td>
<td>5.81</td>
<td>5.98</td>
<td>6.06</td>
</tr>
<tr>
<td>$V(U_b)$</td>
<td>9.77</td>
<td>4.12</td>
<td>15.71</td>
</tr>
</tbody>
</table>

$\sum_{b=1}^3 U_b = 17.85$ 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:
  \[ H_0 : h_{1b}(t) = h_{2b}(t) \text{ for all } t \leq y_{(r)} \text{ and } b = 1, \ldots, m \]
**Basic concepts**

**Nonparametric estimation**

**Hypothesis testing in a nonparametric setting**

**Proportional hazards models**

**Parametric survival models**

◊ 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 ($j = 1, 2$)

⇒ 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
  - $\beta$ : 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 \( \beta \) 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 \( \beta \)

◊ 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 $\beta$ fixed, the likelihood can be rewritten as

$$L(h_{0(1)}, \ldots, h_{0(r)} \mid \beta)$$

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

$$\times \prod_{j=1}^{r} \exp \left( - h_{0(j)} \sum_{k \in R(y(j))} \exp \left( x_{k}^t \beta \right) \right),$$

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:

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

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

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

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

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

$$h_{0(j)} = \frac{1}{b_j} = \frac{1}{\sum_{k \in R(y_{(j)})} \exp \left( x_{(j)}^t \beta \right)}$$
Plug this solution into the likelihood, and ignore factors not containing any of the parameters:

\[
L(\beta) = \prod_{j=1}^{r} \frac{\exp(x_{(j)}^t \beta)}{\sum_{k \in R(y_{(j)})} \exp(x_{(j)}^t \beta)}
\]

= partial likelihood

This expression is used to estimate \( \beta \) through maximization

Logarithm of the partial likelihood:

\[
\ell(\beta) = \sum_{j=1}^{r} x_{(j)}^t \beta - \sum_{j=1}^{r} \log \left( \sum_{k \in R(y_{(j)})} \exp(x_{(j)}^t \beta) \right)
\]
Maximization is often done via the Newton-Raphson procedure, which is based on the following iterative procedure:

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

with

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

⇒ convergence is reached when \( \hat{\beta}_{\text{old}} \) and \( \hat{\beta}_{\text{new}} \) are sufficiently close together.
Score function $U(\beta)$:

$$U_h(\beta) = \frac{\partial \ell(\beta)}{\partial \beta_h}$$

$$= \sum_{j=1}^{r} x(j)_h - \sum_{j=1}^{r} \frac{\sum_{k \in R(y(j))} x_{kh} \exp(x_k^t \beta)}{\sum_{k \in R(y(j))} \exp(x_k^t \beta)}$$

Observed information matrix $I(\beta)$:

$$I_{hl}(\beta) = -\frac{\partial^2 \ell(\beta)}{\partial \beta_h \partial \beta_l}$$

$$= \sum_{j=1}^{r} \frac{\sum_{k \in R(y(j))} x_{kh} x_{kl} \exp(x_k^t \beta)}{\sum_{k \in R(y(j))} \exp(x_k^t \beta)}$$

$$- \sum_{j=1}^{r} \left[ \frac{\sum_{k \in R(y(j))} x_{kh} \exp(x_k^t \beta)}{\sum_{k \in R(y(j))} \exp(x_k^t \beta)} \right]$$

$$\times \sum_{j=1}^{r} \left[ \frac{\sum_{k \in R(y(j))} x_{kl} \exp(x_k^t \beta)}{\sum_{k \in R(y(j))} \exp(x_k^t \beta)} \right]$$
Remarks:

- Variance-covariance matrix of $\hat{\beta}$ can be approximated by the inverse of the information matrix evaluated at $\hat{\beta}$
  \[ V(\hat{\beta}_h) \text{ can be approximated by } [I(\hat{\beta})]^{-1} \]

- Properties (consistency, asymptotic normality) of $\hat{\beta}$ are well established (Gill, 1984)

- A 100(1-$\alpha$)% confidence interval for $\beta_h$ 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
  → 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
→ 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

<table>
<thead>
<tr>
<th>Pat ID</th>
<th>Time</th>
<th>Censoring</th>
<th>Gender</th>
<th>Age</th>
<th>Weight</th>
<th>Func. Status</th>
<th>Clin. Status</th>
<th>CD4</th>
<th>ART</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>699</td>
<td>0</td>
<td>1</td>
<td>42</td>
<td>37</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>455</td>
<td>1</td>
<td>2</td>
<td>30</td>
<td>50</td>
<td>1</td>
<td>3</td>
<td>111</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>705</td>
<td>0</td>
<td>1</td>
<td>32</td>
<td>57</td>
<td>0</td>
<td>3</td>
<td>165</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>694</td>
<td>0</td>
<td>2</td>
<td>50</td>
<td>40</td>
<td>1</td>
<td>3</td>
<td>95</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>86</td>
<td>0</td>
<td>2</td>
<td>35</td>
<td>37</td>
<td>0</td>
<td>4</td>
<td>34</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>98</td>
<td>101</td>
<td>0</td>
<td>1</td>
<td>39</td>
<td>37</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>98</td>
<td>709</td>
<td>0</td>
<td>2</td>
<td>35</td>
<td>66</td>
<td>2</td>
<td>3</td>
<td>103</td>
<td>1</td>
</tr>
<tr>
<td>99</td>
<td>464</td>
<td>0</td>
<td>1</td>
<td>27</td>
<td>37</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>2</td>
</tr>
<tr>
<td>100</td>
<td>537</td>
<td>1</td>
<td>2</td>
<td>30</td>
<td>76</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
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:

\[ \hat{\beta}_{\text{agecat}} = 0.226 \text{ (HR=1.25)} \]
\[ \hat{\beta}_{\text{gender}} = 1.120 \text{ (HR=3.06)} \]

- 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 ⇒ Censoring and event times can be tied
- If ties between censoring time(s) and an event time ⇒ we assume that
  - the censoring time(s) fall just after the event time ⇒ 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):

\[
L(\beta) = \prod_{j=1}^{r} \frac{\prod_{l: y_l = y(j), \delta_l = 1} \exp \left( x^t_l \beta \right)}{\left[ \sum_{k: y_k \geq y(j)} \exp \left( x^t_k \beta \right) \right]^{d(j)}}
\]

Approximation proposed by Efron (1977):

\[
L(\beta) = \prod_{j=1}^{r} \frac{\prod_{l: y_l = y(j), \delta_l = 1} \exp \left( x^t_l \beta \right)}{V_j(\beta)}
\]

where

\[
V_j(\beta) = \prod_{h=1}^{d(j)} \left( \sum_{k: y_k \geq y(j)} \exp \left( x^t_k \beta \right) \right) - \frac{h - 1}{d(j)} \sum_{l: y_l = y(j), \delta_l = 1} \exp \left( x^t_l \beta \right)
\]
Approximation proposed by Cox (1972):

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

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:

<table>
<thead>
<tr>
<th>Method</th>
<th>Approx. Max(partial likel.)</th>
<th>$\hat{\beta}$</th>
<th>s.e.($\hat{\beta}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breslow</td>
<td>-776.11</td>
<td>0.661</td>
<td>0.164</td>
</tr>
<tr>
<td>Efron</td>
<td>-775.67</td>
<td>0.661</td>
<td>0.164</td>
</tr>
<tr>
<td>Cox</td>
<td>-761.36</td>
<td>0.665</td>
<td>0.165</td>
</tr>
</tbody>
</table>

- 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 $\beta$

- **More specific tests:**
  - hypothesis tests regarding a subvector of $\beta$
  - hypothesis tests for contrasts and sets of contrasts
Global hypothesis tests

- Hypotheses regarding the $p$-dimensional vector $\beta$:
  
  $H_0 : \beta = \beta_0$
  
  $H_1 : \beta \neq \beta_0$

- Wald test statistic:
  
  $$U^2_W = (\hat{\beta} - \beta_0)^t I(\hat{\beta}) (\hat{\beta} - \beta_0)$$

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

  $\Rightarrow$ Under $H_0$, and for large sample size: $U^2_W \approx \chi^2_p$
Likelihood ratio test statistic:

\[ U_{LR}^2 = 2 \left( \ell(\hat{\beta}) - \ell(\beta_0) \right) \]

with

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

\( \Rightarrow \) 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 \( \beta_0 \)

\( \Rightarrow \) 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:

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

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

- \(U^2_W = 31.6\); p-value: \(P(\chi^2_3 > 31.6) = 6 \times 10^{-7}\)
- \(U^2_{LR} = 30.6\)
- \(U^2_{SC} = 33.5\)
Local hypothesis tests

- Let $\beta = (\beta_1^t, \beta_2^t)^t$, where $\beta_2$ contains the ‘nuisance’ parameters
- Hypotheses regarding the $q$-dimensional vector $\beta_1$:
  - $H_0 : \beta_1 = \beta_{10}$
  - $H_1 : \beta_1 \neq \beta_{10}$
- 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 $\beta_1$
  \[
  \Rightarrow I^{-1} = \begin{bmatrix}
  I_{11}^{11} & I_{11}^{12} \\
  I_{21}^{11} & I_{22}^{11}
  \end{bmatrix}
  \]
- Note that the complete information matrix is required to obtain $I_{11}^{11}$, except when $\hat{\beta}_1$ is independent of $\hat{\beta}_2$
Define

\[ \hat{\beta}_1 = \text{maximum likelihood estimator of } \beta_1 \]

\[ \hat{\beta}_2(\beta_{10}) = \text{maximum likelihood estimator of } \beta_2 \text{ with } \beta_1 \text{ put equal to } \beta_{10} \]

\[ U_1(\beta_{10}, \hat{\beta}_2(\beta_{10})) = \text{score subvector evaluated at } \beta_{10} \text{ and } \hat{\beta}_2(\beta_{10}) \]

\[ l^{11}(\beta_{10}, \hat{\beta}_2(\beta_{10})) = \text{matrix } l^{11} \text{ for } \beta_1 \text{ evaluated at } \beta_{10} \text{ and } \hat{\beta}_2(\beta_{10}) \]
Basic concepts
Nonparametric estimation
Hypothesis testing in a nonparametric setting
Proportional hazards models
Parametric survival models

- **Wald test** :
  \[
  U^2_W = (\hat{\beta}_1 - \beta_{10})^t (I^{11}(\hat{\beta}))^{-1} (\hat{\beta}_1 - \beta_{10}) \approx \chi^2_q
  \]

- **Likelihood ratio test** :
  \[
  U^2_{LR} = 2 \left( \ell(\hat{\beta}) - \ell(\beta_{10}, \hat{\beta}_2(\beta_{10})) \right) \approx \chi^2_q
  \]

- **Score test** :
  \[
  U^2_{SC} = U_1(\beta_{10}, \hat{\beta}_2(\beta_{10}))^t I^{11}(\beta_{10}, \hat{\beta}_2(\beta_{10})) \times U_1(\beta_{10}, \hat{\beta}_2(\beta_{10})) \approx \chi^2_q
  \]
Testing more specific hypotheses

- Consider a $p \times 1$ vector of coefficients $c$
- Hypothesis test:
  \[ H_0 : c^t \beta = 0 \]
- Wald test statistic:
  \[ U_{W}^2 = (c^t \hat{\beta})^t (c^t I^{-1}(\hat{\beta})c)^{-1} (c^t \hat{\beta}) \]
  Under $H_0$ and for large sample size:
  \[ U_{W}^2 \approx \chi^2_1 \]
- 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:

\[ U_{W}^2 = (C\hat{\beta})^t (CI^{-1}(\hat{\beta})C^t)^{-1} (C\hat{\beta}) \]

Under \( H_0 \) and for large sample size: \( U_{W}^2 \approx \chi^2_q \)

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: $P(\chi_1^2 > 1.18) = 0.179$

◇ $H_0 : \beta_{\text{married}} = \beta_{\text{alone again}} = 0$
  $\rightarrow C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
  $\rightarrow$ 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 $x_i = 0$)
- $\exp(\beta) =$ hazard of a subject $i$ with covar. $x_i$ relative to hazard of a subject $j$ 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:

\[ h_i(t) = h_0(t) \exp(\beta_{age} \cdot age_i) \]

- \( \beta_{age} = 0.00119 \) (s.e. = 0.00952).

- \( HR = \frac{\text{hazard for a subject of age } i \text{ (in years)}}{\text{hazard for a subject of age } i - 1} = 1.001 \)

- 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 $l$ levels:
  $$h_i(t) = h_0(t) \exp(\beta^t x_i),$$
  where
  - $\beta = (\beta_1, \ldots, \beta_l)$
  - $x_i$ 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)
  - \( \text{HR} = 0.814 \) (95% CI: [0.534, 1.240]), \( p = 0.34 \)

- Alone again vs single:
  - \( \hat{\beta}_{\text{alone again}} = 0.794 \) (s.e. = 0.185)
  - \( \text{HR} = 2.213 \) (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

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

<table>
<thead>
<tr>
<th>Covariate</th>
<th>( \hat{\beta} )</th>
<th>s.e.(( \hat{\beta} ))</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>-0.0154</td>
<td>0.0104</td>
<td>0.99</td>
<td>[0.97, 1.01]</td>
</tr>
<tr>
<td>married</td>
<td>-0.3009</td>
<td>0.2238</td>
<td>0.74</td>
<td>[0.48, 1.15]</td>
</tr>
<tr>
<td>alone again</td>
<td>0.8195</td>
<td>0.1857</td>
<td>2.269</td>
<td>[1.58, 3.27]</td>
</tr>
</tbody>
</table>
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
    \[ \Rightarrow \text{add } j - 1 \text{ parameters} \]
  - **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
    \[ \Rightarrow \text{add } (j - 1) \times (k - 1) \text{ parameters} \]
Example: Impact of marital status and age on survival of schizophrenic patients

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

<table>
<thead>
<tr>
<th>Covariate</th>
<th>( \hat{\beta} )</th>
<th>s.e.( (\hat{\beta}) )</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>-0.0238</td>
<td>0.0172</td>
<td>0.977</td>
<td>[0.94, 1.01]</td>
</tr>
<tr>
<td>married</td>
<td>-0.6811</td>
<td>0.8579</td>
<td>0.506</td>
<td>[0.09, 2.72]</td>
</tr>
<tr>
<td>alone again</td>
<td>0.3979</td>
<td>0.7475</td>
<td>1.489</td>
<td>[0.34, 6.44]</td>
</tr>
<tr>
<td>age</td>
<td>married</td>
<td>0.0129</td>
<td>0.0299</td>
<td>1.013</td>
</tr>
<tr>
<td>age</td>
<td>alone again</td>
<td>0.0133</td>
<td>0.0228</td>
<td>1.013</td>
</tr>
</tbody>
</table>
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:
\[ \exp(\hat{\beta}_{\text{age}} + \hat{\beta}_{\text{age}|\text{married}}) = \exp(-0.0238 + 0.0129) = 0.989 \]

Effect of age in the alone again group:
\[ \exp(\hat{\beta}_{\text{age}} + \hat{\beta}_{\text{age}|\text{alone again}}) = \exp(-0.0238 + 0.0133) = 0.990 \]

Likelihood ratio test for the interaction:
\[ U_{LR}^2 = 0.76 \]
\[ P(\chi^2_2 > 0.76) = 0.684 \]
\[ HR_{\text{married}} = \exp(\hat{\beta}_{\text{married}}) = 0.506 \]
\[ = HR \text{ of a married subject relative to a single subject at the age of 0 year} \]
⇒ more relevant to express the age as the difference between a particular age of interest (e.g. 30 years)
⇒ has impact on parameter estimates of differences between groups, but not on parameter estimates related to age

<table>
<thead>
<tr>
<th>Covariate</th>
<th>( \hat{\beta} )</th>
<th>s.e.(( \hat{\beta} ))</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>-0.0238</td>
<td>0.0172</td>
<td>0.977</td>
<td>[0.94,1.01]</td>
</tr>
<tr>
<td>married</td>
<td>-0.2928</td>
<td>0.2378</td>
<td>0.746</td>
<td>[0.47,1.19]</td>
</tr>
<tr>
<td>alone again</td>
<td>0.7971</td>
<td>0.1911</td>
<td>2.219</td>
<td>[1.53,3.23]</td>
</tr>
<tr>
<td>age</td>
<td>married</td>
<td>0.0129</td>
<td>0.0299</td>
<td>1.013</td>
</tr>
<tr>
<td>age</td>
<td>alone again</td>
<td>0.0133</td>
<td>0.0228</td>
<td>1.013</td>
</tr>
</tbody>
</table>
Example: Impact of marital status and gender on survival of schizophrenic patients

\[ h_i(t) = h_0(t) \exp(\beta_{\text{married}} \times x_{i2} + \beta_{\text{alone again}} \times x_{i3} \]
\[ + \beta_{\text{female}} \times \text{gender}_i \]
\[ + \beta_{\text{female}|\text{married}} \times x_{i2} \times \text{gender}_i \]
\[ + \beta_{\text{female}|\text{alone again}} \times x_{i3} \times \text{gender}_i \]

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\hat{\beta}$</th>
<th>s.e.$(\hat{\beta})$</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>0.520</td>
<td>0.286</td>
<td>1.681</td>
<td>[0.96, 2.95]</td>
</tr>
<tr>
<td>married</td>
<td>-0.253</td>
<td>0.26</td>
<td>0.776</td>
<td>[0.47, 1.29]</td>
</tr>
<tr>
<td>alone again</td>
<td>0.807</td>
<td>0.236</td>
<td>2.242</td>
<td>[1.41, 3.56]</td>
</tr>
<tr>
<td>female</td>
<td>married</td>
<td>0.389</td>
<td>0.46</td>
<td>1.476</td>
</tr>
<tr>
<td>female</td>
<td>alone again</td>
<td>-0.146</td>
<td>0.372</td>
<td>0.865</td>
</tr>
</tbody>
</table>

Likelihood ratio test for the interaction:

\[ U_{LR}^2 = 1.94; P(\chi_2^2 > 1.94) = 0.23 \]
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 \( \beta \):
  - Let \( x_k(y) \) be the covariate vector for subject \( k \) at time \( y \)
  - Define the partial likelihood:
    \[ L(\beta) = \prod_{i=1}^{n} \left[ \frac{\exp(x_i(y_i)^t \beta)}{\sum_{k \in R(y_i)} \exp(x_k(y_i)^t \beta)} \right]^{\delta_i} \]
  - Let
    \[ \hat{\beta} = \arg\max_{\beta} L(\beta) \]
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** $i$ ($i = 1, \ldots, n$):

  \[ (y_i, \delta_i, x_i(t_{i1}), \ldots, x_i(t_{ik_i})) \]

  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

\[ \hat{\beta} = -0.0273 \quad (\text{s.e.} = 0.0162) \]

\[ HR = \exp(-0.0273) = 0.973 \]

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 $x_i$:

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

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

$$\hat{H}_0(t) = \sum_{j:y(j)\leq t} \hat{h}_0(j),$$

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
\[ \frac{\hat{S}_i(t) - S_i(t)}{V^{1/2}(\hat{S}_i(t))} \xrightarrow{d} \mathcal{N}(0, 1) \]
Example: Survival function estimates for marital status groups in the schizophrenic patients data

![Survival function estimates for marital status groups](image-url)
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
  → 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
  $$h_{ij}(t) = h_{i0}(t) \exp(x_{ij}^t \beta)$$

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

  → 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
  - 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:
  \[ \hat{\beta} = -0.0588 \quad \text{(s.e. = 0.0198)} \]
  \[ HR = 0.943 \quad 95\% \ CI = [0.907, 0.980] \]

⇒ 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 $\beta_i x_i \times t$ to the PH model:
    \[
    h_i(t) = h_0(t) \exp(\beta x_i + \beta_i x_i \times t)
    \]
  - If $\beta_i \neq 0$, the PH assumption does not hold
  - Instead of adding $\beta_i x_i \times t$, one can also add $\beta_i x_i \times g(t)$ 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

\[ \hat{H}_1(t), \hat{H}_2(t), \ldots, \hat{H}_r(t) \] should be constant multiples of each other:

<table>
<thead>
<tr>
<th>Plot</th>
<th>PH assumption holds if</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log(\hat{H}_1(t)), \ldots, \log(\hat{H}_r(t)) ) vs ( t )</td>
<td>parallel curves</td>
</tr>
<tr>
<td>( \log(\hat{H}_j(t)) - \log(\hat{H}_1(t)) ) vs ( t )</td>
<td>constant lines</td>
</tr>
<tr>
<td>( \hat{H}_j(t) ) vs ( \hat{H}_1(t) )</td>
<td>straight lines through origin</td>
</tr>
</tbody>
</table>
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 \( \lambda > 0 \):
  \[
  \begin{align*}
  S_0(t) & = \exp(-\lambda t) \\
  f_0(t) & = \lambda \exp(-\lambda t) \\
  h_0(t) & = \lambda 
  \end{align*}
  \]
  → 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

- Basic concepts
- Nonparametric estimation
- Hypothesis testing in a nonparametric setting
- Proportional hazards models
- Parametric survival models

The left graph shows the hazard function for the exponential distribution with a lambda of 0.14. The right graph shows the survival function for the same distribution.
Weibull distribution :

- Characterized by a scale parameter $\lambda > 0$ and a shape parameter $\rho > 0$:
  
  $S_0(t) = \exp(-\lambda t^\rho)$
  
  $f_0(t) = \rho \lambda t^{\rho-1} \exp(-\lambda t^\rho)$
  
  $h_0(t) = \rho \lambda t^{\rho-1}$

  → hazard decreases if $\rho < 1$

  → hazard increases if $\rho > 1$

  → 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

Hazard and survival functions for Weibull distribution

Hazard and survival functions for Weibull distribution with parameters Lambda=0.31, Rho=0.5 and Lambda=0.06, Rho=1.5.
Log-logistic distribution:

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

$$S_0(t) = \frac{1}{1 + (t\lambda)^\kappa}$$

$$f_0(t) = \frac{\kappa t^{\kappa-1} \lambda^\kappa}{[1 + (t\lambda)^\kappa]^2}$$

$$h_0(t) = \frac{\kappa t^{\kappa-1} \lambda^\kappa}{1 + (t\lambda)^\kappa}$$

- The median event time is only a function of the parameter $\lambda$:

$$M(T) = \exp(1/\lambda)$$
Hazard and survival function for the log-logistic distribution

```
Hazard
0 2 4 6 8 10
0.0 0.2 0.4

Survival
0 2 4 6 8 10
0.0 0.2 0.4 0.6 0.8 1.0
```

Lambda=0.2, Kappa=1.5
Lambda=0.2, Kappa=0.5
Log-normal distribution:

- Resembles the log-logistic distribution but is mathematically less tractable
- A random variable $T$ has a log-normal distribution if $\log T$ has a normal distribution
- Characterized by two parameters $\mu$ 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 $\mu$:
  
  $$M(T) = \exp(\mu)$$
Hazard and survival function for the log-normal distribution

Hazard and survival function for the log-normal distribution with different parameters: 
- Hazard function with \( \mu = 1.609 \) and \( \gamma = 0.5 \)
- Hazard function with \( \mu = 1.609 \) and \( \gamma = 1.5 \)
- Survival function with \( \mu = 1.609 \) and \( \gamma = 0.5 \)
- Survival function with \( \mu = 1.609 \) and \( \gamma = 1.5 \)
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,

  \[ h_i(t) = \exp(\theta^t x_i) h_0(\exp(\theta^t x_i)t) \]
and

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

where \( M_i = \text{median of } S_i \), since

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

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

![Survival function graph](image-url)
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,

\[ S_i(t) = P(t_i > t) = P(\log t_i > \log t) = P(\mu + \sigma w_i > \log t - \gamma^t x_i) = S_0(\exp(\log t - \gamma^t x_i)) = S_0(t \exp(\theta^t x_i)) \]

⇒ The two models are equivalent
Weibull distribution

Consider the accelerated failure time model

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

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

\[ \Rightarrow S_i(t) = \exp(-\lambda \exp(\beta^t x_i)t^\alpha) \text{ with } \beta = \alpha \theta \]
\[ \Rightarrow f_i(t) = \lambda \alpha t^{\alpha-1} \exp(\beta^t x_i) \exp(-\lambda \exp(\beta^t x_i)t^\alpha) \]
\[ \Rightarrow h_i(t) = \alpha \lambda t^{\alpha-1} \exp(\beta^t x_i) = h_0(t) \exp(\beta^t x_i), \]

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

\[ \Rightarrow \text{We also have a Cox PH model} \]
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,

$$P(W > w) = P(\exp(\mu + \sigma W) > \exp(\mu + \sigma w)) = S_0(\exp(\mu + \sigma w)) = \exp(-\lambda \exp(\alpha \mu + \alpha \sigma w))$$

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

\[ S_i(t) = S_0 \left( \exp(\theta^t x_i) t \right), \]

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

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

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

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

\[ \Rightarrow \text{We also have a so-called proportional odds model} \]
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,

$$P(W > w) = P(\exp(\mu + \sigma W) > \exp(\mu + \sigma w))$$

$$= S_0(\exp(\mu + \sigma w))$$

$$= 1/[1 + \lambda \exp(\alpha \mu + \alpha \sigma w)]$$

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

\[
L(\mu, \gamma, \sigma) = \prod_{i=1}^{n} f_i(y_i)^{\delta_i} S_i(y_i)^{1-\delta_i}
\]

\[
= \prod_{i=1}^{n} \left[ \frac{1}{\sigma y_i} f_W \left( \frac{\log y_i - \mu - \gamma^t x_i}{\sigma} \right) \right]^{\delta_i}
\times \left[ S_W \left( \frac{\log y_i - \mu - \gamma^t x_i}{\sigma} \right) \right]^{1-\delta_i}
\]

Since \( W \) has a known distribution, this likelihood can be maximized w.r.t. its parameters \( \mu, \gamma, \sigma \)
Let

$$(\hat{\mu}, \hat{\gamma}, \hat{\sigma}) = \arg\max_{\mu, \gamma, \sigma} L(\mu, \gamma, \sigma)$$

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 \((\mu, \gamma)\)
- \( k = 0 \) 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