POLITECNICO
MILANO 1863

Nonparametric estimation in Survival Analysis

Nonparametric Statistics
AA 2020-2021

Francesca Ieva
MOX – Department of Mathematics, Politecnico di Milano, Italy

Outline

1. Introduction
2. Kaplan-Meier estimator
3. Log-rank test
4. Hazard Ratio

Dataset: **Randomized clinical trial** of chemotherapy in 478 patients with osteosarcoma.

Outline

1. Introduction
2. Kaplan-Meier estimator
3. Log-rank test
4. Hazard Ratio

Introduction

Introduction

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is the **survival time**, a variable which measures the time from a particular starting time (origin event) to a particular endpoint (event of interest): **time-to-event**.

any event of interest: the time to this event to happen is the focus of the analysis. We would like some regression model that models such times according to some inputs.

- **Time** → years, months, weeks or days from the beginning of follow up of an individual until an event occur
- **Event** → death, disease incidence, relapse from remission, recovery or any designated experience of interest that may happen to an individual

Research fields: medicine, biology, public health, social sciences, economics, finance, engineering

Censoring

Censoring occurs when we have some information about individual survival time, but we do not know the survival time exactly.

↓

Partially observed data

Reasons why censoring may occur:

1. a person **does not experience the event** before the study ends;
2. a person is **lost** to follow up during the study period;
3. a person withdraws from the study because a **reason different** to the event of interest.

Right censoring: subjects may enter the study at different times and the real event time is greater than the observed time.

Time-to-event outcome

For each subject i, let:

- T_i^* be the non-negative r.v. denoting the **true event time**
- C_i be the non-negative r.v. denoting the time at which a **censoring** mechanism kicks in

What we actually observe is the **survival time**: $T_i = \min(T_i^*, C_i)$

We also define an indicator random variable δ_i for non-censoring:

$$\delta_i = I(T_i^* \leq C_i)$$

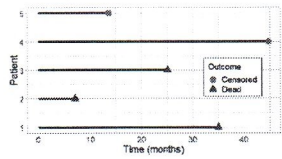
↓

Time-to-event data: $\mathcal{D}_N = \{(T_i, \delta_i), i = 1, ..., N\}$

Non informative censoring: $C_i \perp T_i^*$

Dataset

- Time = months
- Origin event = end date of last chemotherapy cycle
- Event of interest = death of the patient
- Outcome = {0 = *Censored*, 1 = *Dead*}



i	T_i^*	C_i	T_i	δ_i
5	?	13.53	13.53	0
4	?	44.80	44.80	0
3	25.10	-	25.10	1
2	7.03	-	7.03	1
1	35.03	-	35.03	1

Time-to-event data: $\mathcal{D}_5 = \{(T_i, \delta_i), i = 1, ..., 5\}$

Survival Functions

Let T denote the non-negative r.v. of survival time with probability density function $f(t)$ and distribution function $F(t) = Pr(T \leq t)$.

Survival function

The survival function at time t is defined as the the complement of the distribution function:

$$S(t) = Pr(T > t) = 1 - Pr(T \leq t)$$

Hazard function

The hazard function is the **instantaneous risk of failure** at time t, conditional on survival to that time:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

(*) $S(t_i) = \mathbb{P}(T \geqslant t_i) = \sum_{k: t_k > t_i} f(t_k)$

$h(t_i) = \mathbb{P}(T = t_i | T \geqslant t_i) = \frac{f(t_i)}{S(t_i)}$

$= \frac{\mathbb{P}(T \geqslant t_i) - \mathbb{P}(T \geqslant t_{i+1})}{S(t_i)}$

$= \frac{S(t_i) - S(t_{i+1})}{S(t_i)} = 1 - \frac{S(t_{i+1})}{S(t_i)}$

$S(t_{i+1}) = \frac{S(t_{i+1})}{S(t_i)} \cdot \frac{S(t_i)}{S(t_{i-1})} \cdots \frac{S(t_2)}{S(t_1)} \cdot \frac{S(t_1)}{S(t_0 = 0)}$

$S(0) = 1$

$= (1 - h(t_i)) \cdots (1 - h(t_1))(1 - h(0))$

$= \prod_{k: t_k < t_{i+1}} (1 - h(t_k))$

$h(0) = 0$

$\rightarrow S(t) = \prod_{i: t_i < t} (1 - h(t_i))$

Survival & Hazard Functions – T discrete

If the survival time T (discrete) has a probability mass function $P(T = t_i) = f(t_i)$, $i = 1, ..., n$, the survival function is

$$S(t) = Pr(T \geq t) = \sum_{i: t_i \geq t} f(t_i)$$

The hazard function $h(t)$ is defined as the conditional probability of failure at time t_i given that the individual has survived up to time t_i:

$$h_i = h(t_i) = Pr(T = t_i | T \geq t_i) \overset{(*)}{=} \frac{f(t_i)}{S(t_i)} = 1 - \frac{S(t_{i+1})}{S(t_i)}$$

Therefore:

$$S(t) = \prod_{i: t_i < t} (1 - h(t_i)) \Rightarrow f(t_i) = h(t_i) \cdot S(t_i)$$

Survival & Hazard Functions – T continuous

Let T denote the non-negative r.v. of survival time with probability density function $f(t)$, distribution function $F(t) = Pr(T \leq t)$ and survival function $S(t) = 1 - F(t)$.

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{Pr(T \in [t, t + \Delta t) \cap T \geq t) / Pr(T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{Pr(T \in [t, t + \Delta t))}{\Delta t} \cdot \frac{1}{Pr(T \geq t)} \\ &= \lim_{\Delta t \to 0} \frac{\int_t^{t+\Delta t} f(u) du}{\Delta t} \cdot \frac{1}{Pr(T > t)} \\ &= \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln S(t) \end{aligned}$$

Survival & Hazard Functions – T continuous

Hence the hazard function is

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\ln S(t)$$

Integrating from 0 to t using the boundary condition $S(0) = 1$, we obtain a formula for the survival probability as a function of the hazard:

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\}, \quad t \geq 0$$

$= e^{-H(t)}$ general form describing the Survival as a function of the time

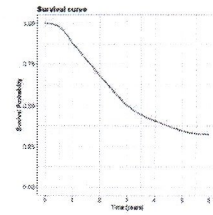
- The event rate is proportional to the rate at which the survival function $S(t)$ changes.
- If the survival function is decreasing sharply with time then the mortality rate is high (and vice versa).

The point of survival analysis will be to provide models, depending on covariates, of $H(t)$ in order to be able to explain the survival that we observed.
This is the framework for which also the survival time can be intended as any traditional regression model.

The Survival Curve

The survival function $S(t)$ is an estimate of the percentage of individuals in a cohort who are still *event free* at time t.

- $S(0) = 1$: All subjects are alive at beginning of the study
- $S(t)$ can only remain at same value or decrease as time progresses
- If all the subjects do not experience the event by the end of the study window, the curve may never reach zero

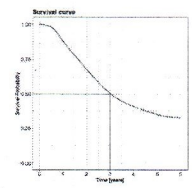


The survival function gives an estimate of these percentage for a given time between the beginning of the time interval and the end of the study.

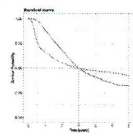
Median Survival Curve

The **median survival time** is estimated by the time at which 50% of the cohort being studied are still event free*

Median survival time: $t = 3$ years ▷



* If the Kaplan-Meier curve does not hit 50% exactly, the convention is to use the first event time where the curve drops below 50%



Caveat
Medians do not describe whole curve

different curves may have the same median survival time

Cumulative Incidence Function and Cumulative Hazard

- The **cumulative incidence**, or cumulative failure probability (CFP), is an estimate of the percentage of individuals in a cohort who have already *experienced the event* at time t and it is computed as:

$$CFP(t) = P(T \leq t) = 1 - S(t)$$

so can be estimated as $1 - \hat{S}(t)$.

- The **cumulative hazard function** at time t is:

$$H(t) = \int_0^t h(u)du = -\ln[S(t)]$$

It can be interpreted as the cumulative force of mortality.
Note that we can approach the estimation of the survival in a twofold way:
a) estimating H(t)--- N-A estimator and **b) directly estimating S(t)--- K-M estimator**

We will provide an estimation of the survival function in 2 ways:
- directly estimating the survival function through the Kaplan-Meier estimator
- estimating the cumulative hazard function through the Nelson-Aalen estimator and then retrieve the survival function

Nelson-Aalen estimator of $H(t)$

The cumulative hazard function $H(t)$ can be estimated using the **non-parametric Nelson-Aalen estimator** $\hat{H}(t)$ that is given by:

$$\hat{H}(t) = \sum_{j:t_j^* \leq t} \frac{d_j}{n_j} \quad \text{with} \quad \widehat{Var}(\hat{H}(t)) = \sum_{j:t_j^* \leq t} \frac{d_j}{n_j^2}$$

$$\hat{H}(t) = \sum_{j:t_j^* \leq t} \frac{d_j}{n_j}$$

$$\hat{Var}(H(t)) = \sum_{j:t_j^* \leq t} \frac{d_j}{n_j^2}$$

- j = failure (event) index $\in \{1, \ldots, J\}$
- J = total number of individuals who experienced the event
- $0 < t_1^* < \ldots < t_J^* < \infty$ = observed ordered event times
- n_j = number of event-free patient just before t_j^*, i.e. number of patients at risk at time t_j^*
- d_j = number of observed events at t_j^*

It can be interpreted as the ratio of the number of deaths to the number of exposed

Goals of Survival Analysis

Estimate the survival function for a **group** of individuals
↓
Kaplan-Meier estimator

for example "all the people that have a given age"

Compare survival functions between two or more groups
↓
Log-rank test & Hazard Ratio

Assess how the **covariates** affect the hazard function
↓
Cox, frailty, parametric models

16

Outline

Kaplan-Meier estimator

1. Introduction
2. **Kaplan-Meier estimator**
3. Log-rank test
4. Hazard Ratio

17

Kaplan-Meier estimator

The **Kaplan-Meier estimator**, also known as the **product limit estimator**, is a **non-parametric** statistic used to estimate the survival function $S(t)$ from lifetime data.

The Kaplan-Meier survival curve is defined as the probability of surviving in a given length of time while considering time in many small intervals.

There are three assumptions used in this analysis:

1. Censoring is unrelated to the outcome.
   Any time patients who are censored have the same survival prospects as those who continue to be followed.
2. The survival probabilities are the same for subjects recruited early and late in the study.
3. The events occurred at the **specified times.** (vs interval censoring estimation)

KM estimator

The Kaplan-Meier estimator

The Kaplan-Meier (K-M) estimator of the survival function $S(t)$ is:

$$\hat{S}(t) = \prod_{j:t_j^* \le t} p_j = \prod_{j:t_j^* \le t} \left(1 - \frac{d_j}{n_j}\right)$$

step function where we have jumps observed at the event times

- j = failure (event) index $\in \{1, \ldots, J\}$
- J = total number of individuals with events
- $0 < t_1^* < \ldots < t_J^* < \infty$ = observed ordered times of deaths
- p_j = conditional probability of surviving time t_j^*
- n_j = number of patient alive just before t_j^*, i.e. number of patients at risk at time t_j^*
- d_j = number of observed events at t_j^*

The KM estimator is a **step function** with *jumps* at the *observed death times*

19

KM computation: example

Ex: Survival [years] of patients with parathyroid cancer (N=20)

Alive	<1	<1 (*)	1	1	4	5	6	8	10	10	17
Dead	<1	2	6	6	7	9	9	11	14		

In order to build the ***life table***, for each year (or time unit) you need to compute:

- **# of person alive at start** (20)
- **# of withdrawn during the year** (2) (*)
- **# of person at risk for the year** (19 see below)
- **# of person dying** (1)

For the example in the table:

- **Number at risk** in the first year: 17 alive + 0.5+0.5 partially observed + 1 died = 19
- **Probability of dying** in the first year = 1/19
- **Probability of surviving** after the first year = 1-1/19 = 18/19

This procedure should be repeated for each year of the observational study.

Continue.....

20

"x"

~~Year~~ within the observation period we have to compute:

N_x = number of persons alive at start of the year x

W_x = number of withdrawn during the year x

$M_x = n_x$ = number at risk for the year x

d_x = number of events (people dying)

Ex. of the table

$N_1 = 20$

$W_1 = 2$

$M_1 = r_1 = N_1 - \frac{1}{2} W_1 = 19$

$d_1 = 1 \Rightarrow N_2 = 17$

$N_{x+1} = r_x - W_x - d_x$

Quantities of interest for building the KM ("product limit") estimator of the $S(t)$:

$q_x = \mathbb{P}(\text{dying during year } x) \overset{!}{=} \frac{d_x}{r_x}$

$p_x = \mathbb{P}(\text{surviving at } x) = 1 - q_x = 1 - \frac{d_x}{r_x}$

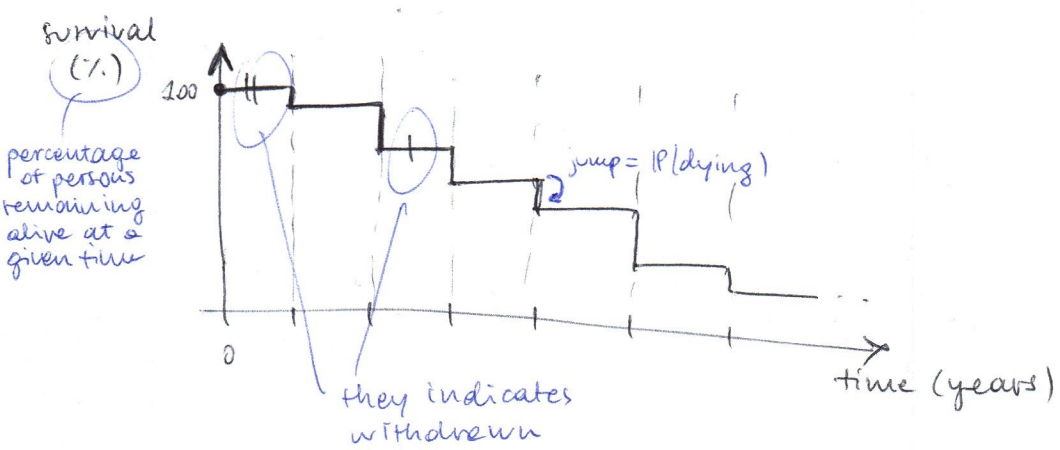
P_x = cumulative survival probability

1st year: $P_1 = p_1$

2nd year: $P_2 = p_2 p_1$

$= \mathbb{P}(\text{arriving to the start of the second year}) \cdot \mathbb{P}(\text{surviving during the second year})$

3rd year: $P_3 = p_3 P_2 \overset{!}{=} p_3 p_2 p_1$



Another method: estimate of the hazard function

KM: Non-Parametric Maximum Likelihood (NPML)

K-M estimator can be derived from maximum likelihood estimation of the hazard function h_j. The likelihood function $\mathcal{L}$ based on observed true times takes the form:

$$\mathcal{L}(h) = \prod_{j=1}^{J} P(T_j = t_j^*)^{d_j} P(T_j > t_j^*)^{w_j} \overset{\bigstar}{=} \prod_{j=1}^{J} h_j^{d_j} (1-h_j)^{n_j - d_j}$$

- n_j = number of patients at risk at time t_j^*
- d_j = number of observed events at t_j^*
- w_j = number of censored patients during $[t_j^*, t_{j+1}^*)$

The MLE of h_j is then:

$$\ell(h) = \ln(L) = \sum_{j=1}^{J} [d_j \ln(h_j) + (n_j - d_j) \ln(1-h_j)]$$

$$\frac{\partial \ell}{\partial \pi_j} = \frac{d_j}{h_j} - \frac{n_j - d_j}{1-h_j} = 0 \iff \boxed{\hat{h}_j = \frac{d_j}{n_j}}$$

— Estimate of the hazard function at time j = number of the observed death over the number of individuals at risk at time j

KM: Non-Parametric Maximum Likelihood (NPML)

The likelihood $\mathcal{L}$ of the hazard function based on observed true times is given by:

$$\mathcal{L}(h) = \prod_{j=1}^{J} P(T_j = t_j^*)^{d_j} P(T_j > t_j^*)^{w_j} \overset{\bigstar}{=} \prod_{j=1}^{J} f(t_j^*)^{d_j} S(t_j^*)^{w_j}$$
$$\overset{\bigstar}{=} \prod_{j=1}^{J} \left\{ h_j^{d_j} \prod_{k=1}^{j-1} (1-h_k)^{d_j} \prod_{k=1}^{j} (1-h_k)^{w_j} \right\}$$
$$= \prod_{j=1}^{J} h_j^{d_j} (1-h_j)^{n_j - d_j}$$

- j = event index $\in \{1, \ldots, J\}$, with J = total number of individuals with failures
- $0 < t_1^* < \ldots < t_J^* < \infty$ = observed ordered event times
- n_j = number of patients at risk at time t_j^*
- d_j = number of observed events at t_j^*
- w_j = number of censored patients during $[t_j^*, t_{j+1}^*)$

KM: NPMLE and Greenwood's formula

The K-M estimator $\hat{S}(t)$ follows from multiplying the conditional survival probabilities $(1-\hat{h}_j)$:

$$\hat{S}(t) = \prod_{j:t_j^* \le t} \left(1 - \frac{d_j}{n_j}\right)$$

It can also be proven that the variance of $\hat{S}(t)$ can be estimated using the *Greenwood's formula*:

$$\widehat{Var}\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 \widehat{Var}\left(\ln[\hat{S}(t)]\right) = \left[\hat{S}(t)\right]^2 \sum_{j:t_j^* \le t} \frac{d_j}{n_j (n_j - d_j)}$$

Remark: if no censoring, $\hat{S}(t)$ coincides with the empirical survival function

KM: 95% Confidence Intervals

The 95% **confidence interval** for the Kaplan-Meier survival estimator is given by:

$$CI_{0.95}(S(t)) = \left[\hat{S}(t) \pm z_{0.975} \cdot \widehat{se}(t)\right]$$

where the standard error is

$$\widehat{se}(t) = SE\left[\hat{S}(t)\right] = \hat{S}(t) \sqrt{\sum_{j:t_j^* \le t} \frac{d_j}{n_j (n_j - d_j)}}$$

PB: One nasty problem with the Greenwood formula for CI is that it may produce limits beyond the range of zero or one (it can produce negative point estimates or point estimates exceeding 100%). In those cases, we just clip the confince interval at zero.

KM: 95% Confidence Intervals – with R

To obtain the **plain confidence interval** given by:

$$CI_{0.95}(S(t)) = \left[\hat{S}(t) \pm z_{0.975} \cdot \widehat{se}(t)\right]$$

in R you have to specify:

```
survfit(Surv(T_i, δ_i) ~ 1, data, conf.type='plain')
```

Otherwise, the function returns the so-called **log confidence interval**, that is the exponential of $CI_{0.95}(\ln[S(t)])$ and it is given by:

$$CI_{0.95}(S(t)) = \left[S(t) \cdot e^{-z_{0.975}\sqrt{\widehat{Var}(\ln[\hat{S}(t)])}}; \hat{S}(t) \cdot e^{z_{0.975}\sqrt{\widehat{Var}(\ln[\hat{S}(t)])}}\right]$$

What confidence interval type should you use? There is no general consensus. The plain setting is great for its simplicity, the log setting produces variances that are stable.

— These produces more stable CI (even if it's more computationally demanding it's worth it)

Dataset: KM estimator using $\mathcal{D}_5$

	i	1	2	3	4	5
$\mathcal{D}_5$	T_i	35.03	7.03	25.10	44.80	13.53
	δ_i	1	1	1	0	0

In $\mathcal{D}_5$ there are $J = 3$ deaths. The ordered observed death times are:

$$0 < t_1^* = 7.03 < t_2^* = 25.10 < t_3^* = 35.03 < \infty$$

Lifetime table :

t_j^*	n_j	d_j	$p_j = (n_j - d_j)/n_j$	$\widehat{S}(t) = \prod_{j:t_j^* \le t} p_j$
7.03	5	1	0.800	0.800
25.10	3	1	0.667	0.533
35.03	2	1	0.500	0.267

(r_j) being at risk

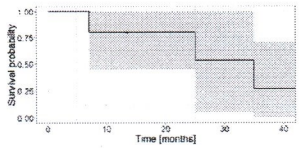
Remark: when computing the KM estimator, censored patients enter the computation of n_j only

26

Dataset: KM estimator using $\mathcal{D}_5$

▷ Survival probability $\hat{S}(t)$ plot computed using K-M estimator.

$t \in [t_j^*, t_{j+1}^*)$	$\hat{S}(t)$
$[0, 7.03)$	1.000
$[7.03, 25.10)$	0.800
$[25.10, 35.03)$	0.533
$[35.03, \infty)$	0.267

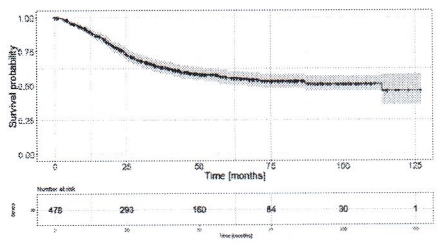


- One year survival rate: $\hat{S}(t = 12) = 80\%$
- Two years survival rate: $\hat{S}(t = 24) = 80\%$
- Three years survival rate: $\hat{S}(t = 36) = 26.7\%$
- Median survival time: $t = 35.03$ months

27

Dataset: KM estimator using $\mathcal{D}_N$

Considering the entire dataset $\mathcal{D}_N$, we have $J = 185$ deaths and, using K-M estimator, we obtain the following survival probability $\hat{S}(t)$ plot:



$\widehat{S}(t = 36) = 64\%$ $se(t = 36) = 2.37\%$

Median survival time: $t = 113.6$ months

Outline

Log-rank test

1. Introduction
2. Kaplan-Meier estimator
3. **Log-rank test**
4. Hazard Ratio

34

Log-rank test for 2 groups

The log-rank test (test di Mantel-Cox) is the most commonly-used **non-parametric statistical test** for comparing the **survival distributions** of two (or more) groups.

Log-rank test

$$H_0 : S_1(\cdot) = S_2(\cdot) \quad \text{vs} \quad H_1 : S_1(\cdot) \neq S_2(\cdot)$$

Let:

- $k = 1, 2$ groups
- $J =$ total number of dead patients (not censored)
- $0 < t_1^* < t_2^* < \cdots < t_J^* < \infty$ observed ordered times of deaths
- Intervals $[t_j^*, t_{j+1}^*)$, $j = 0, \ldots, J$ with $t_0^* = 0$ and $t_{J+1}^* = \infty$

One way to assess H_0 is to look at the difference between observed and expected numbers of events in each group on each time interval.

35

Log-rank test for 2 groups

Let us consider the following quantities:

- n_{kj} is the number patients in group k who are at risk at t_j^*
- n_j is the total number patients at risk at t_j^*: $n_j = n_{1j} + n_{2j}$
- d_{kj} is the number of observed events in group k at t_j^*
- d_j is the total number of observed events at t_j^*: $d_j = d_{1j} + d_{2j}$
- p_{d_j} is the probability of recurrence at t_j^* and it is given by:

$$p_{d_j} = \frac{d_{1j} + d_{2j}}{n_{1j} + n_{2j}} = \frac{d_j}{n_j}$$

- e_{kj} is the number of expected events in group k at t_j^* and it is given by:

$$e_{kj} = p_{d_j} n_{kj} = \frac{d_j n_{kj}}{n_j}$$

Observe that, defining w_{kj} as the number of withdrawals in group k during interval $[t_j^*, t_{j+1}^*)$, the number of patients in group k who are at risk at t_{j+1}^* is given by

$$n_{k,j+1} = n_{kj} - d_{kj} - w_{kj}$$

Log-rank test for 2 groups

Summing up over intervals j the number of observed and expected events in each group $k = 1, 2$ we obtain:

$$O_k = \sum_{j=1}^{J} d_{kj} \qquad E_k = \sum_{j=1}^{J} e_{kj}$$

The approximated **log-rank test statistic** is defined as:

$$\chi^2 = \sum_{k=1,2} \frac{(O_k - E_k)^2}{E_k} \sim \chi_1^2$$

Log-rank test

$H_0 : S_1(\cdot) = S_2(\cdot)$ vs $H_1 : S_1(\cdot) \neq S_2(\cdot)$

Decision rule: "We reject H_0 at statistical level α if $\chi^2 > \chi^2_{1,\alpha}$"

essentially observes the proportion of the rate of events over the time for each group, compare these observation with what should be expected if we consider the 2 groups as of the same group and then it makes an assessment using the χ^2 distribution

Log-rank test for 2 groups

The test statistic

$$\sum_{k=1,2} \frac{(O_k - E_k)^2}{E_k} \quad \text{with} \quad O_k = \sum_{j=1}^{J} d_{kj} \quad \text{and} \quad E_k = \sum_{j=1}^{J} e_{kj}$$

is an approximation of the real **log-rank test statistic** that is given by

$$\frac{O_k - E_k}{\sqrt{V}} \sim \mathcal{N}(0,1) \quad \text{or} \quad \chi^2 = \frac{(O_k - E_k)^2}{V} \sim \chi_1^2$$

with $\mathbb{E}[O_k] = E_k$ and $V = Var(O_k) = \sum_{j=1}^{J} Var(d_{kj}) = \sum_{j=1}^{J} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$.

This follows from the fact that in each interval $[t_j^*, t_{j+1}^*)$, the probability of having d deaths in n_{kj} at risk patients from a finite population of size n_j that contains exactly d_j dead patients is given by an hypergeometric distribution:

$$d_{kj} \sim Hypergeometric(n_j, d_j, n_{kj})$$

with $\mathbb{E}[d_{kj}] = \frac{d_j n_{kj}}{n_j} = e_{kj}$ and $Var(d_{kj}) = \frac{n_{kj}(n_j - n_{kj}) d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$

Log-rank test for K groups

For each interval $j = 1, \ldots, J$ compute:

- K is the number of total groups with $k = 1, 2, \ldots, K$
- n_{kj} is the number patients in group k who are at risk at t_j^*
- $n_j = \sum_{k=1}^{K} n_{kj}$ is the total number patients at risk at t_j^*
- d_{kj} is the number of observed events in group k at t_j^*
- $d_j = \sum_{k=1}^{K} d_{kj}$ is the total number of observed events at t_j^*
- $p_{d_j} = \frac{d_j}{n_j}$ is the probability of recurrence at t_j^*
- $e_{kj} = p_{d_j} n_{kj}$ is the number of expected events in group k at t_j^*

The approximated **log-rank test statistic** is defined as:

$$\chi^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} \sim \chi^2_{K-1} \quad \text{with} \quad O_k = \sum_{j=1}^{J} d_{kj} \quad \text{and} \quad E_k = \sum_{j=1}^{J} e_{kj}$$

Log-rank test for K groups

$H_0 : S_1(\cdot) = \cdots = S_K(\cdot)$ vs H_1 : survival curves are not identical

Decision rule: "We reject H_0 at statistical level α if $\chi^2 > \chi^2_{K-1,\alpha}$"

Dataset: covariates

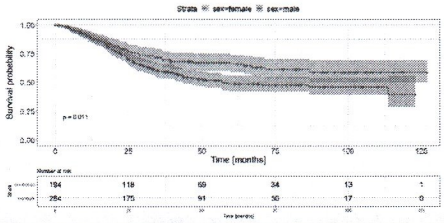
We consider $N = 478$ patients with $J = 185$ dead patients. For each patient i, we have the following information:

- $\texttt{gender}_i$ = **gender** at randomization
  $Female = 194\ (40.6\%)$ $Male = 284\ (59.4\%)$
- $\texttt{trt}_i$ = **chemotherapy** treatment
  $Conventional = 237\ (49.6\%)$ $Dose\ Intense = 241\ (50.4\%)$
- $\texttt{terminated}_i$ = therapy **terminated** or not
  $No = 102\ (21.3\%)$ $Yes = 376\ (78.7\%)$
- $\texttt{age}_i$ = **age** at randomization
  $Q_1 = 12.2$ $Q_2 = 15.2$ $Q_3 = 18.1$
  for which we consider four groups:
  $0 - 12 = 114\ (23.9\%)$ $13 - 15 = 116\ (24.3\%)$
  $16 - 18 = 113\ (23.6\%)$ $18+ = 135\ (28.2\%)$

each of them can be categorized and then we can see if the curves are different based on those categories

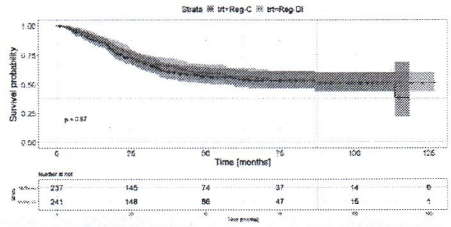
Dataset: log-rank test (I)

- Group $k = 1$: Female
- Group $k = 2$: **Male**
- $\chi^2 = 6.4 \rightarrow p - value = 0.011$
- Reject $H_0 \rightarrow S_{\text{Female}}(\cdot) \neq S_{\text{Male}}(\cdot)$



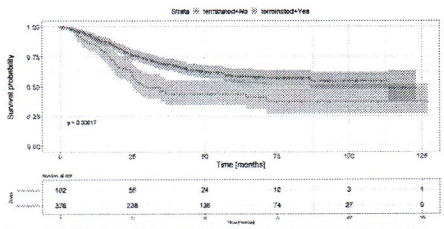
Dataset: log-rank test (II)

- Group $k = 1$: **Regimen C** (Conventional)
- Group $k = 2$: Regimen DI (Dose-Intense)
- $\chi^2 = 0.3 \rightarrow p - value = 0.57$
- Do not reject $H_0 \rightarrow S_{\text{C}}(\cdot) = S_{\text{DI}}(\cdot)$



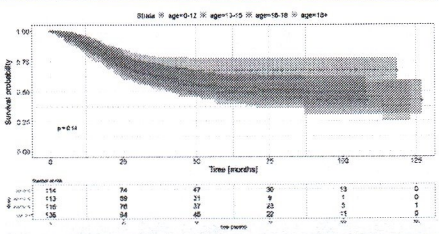
Dataset: log-rank test (III)

- Group $k = 1$: **Not terminated**
- Group $k = 2$: Terminated
- $\chi^2 = 14.1 \rightarrow p - value = 0.00017$
- Reject $H_0 \rightarrow S_{\text{No}}(\cdot) \neq S_{\text{Yes}}(\cdot)$



Dataset: log-rank test (IV)

- Group $k = 1$: **0-12** years old
- Group $k = 2$: 13-15 years old
- Group $k = 3$: 15-18 years old
- Group $k = 4$: 18+ years old
- $\chi^2 = 5.5 \rightarrow p - value = 0.14$ $(\chi^2_{3,0.05} = 7.81)$
- Do not reject $H_0 \rightarrow S_{0\text{-}12}(\cdot) = S_{13\text{-}15}(\cdot) = S_{15\text{-}18}(\cdot) = S_{18+}(\cdot)$



Outline

Hazard Ratio

1. Introduction
2. Kaplan-Meier estimator
3. Log-rank test
4. **Hazard Ratio**

Hazard Ratio

The **hazard ratio** is the ratio of the hazard rates corresponding to the conditions described by two levels of an explanatory variable.

To compute the HR starting from the log-rank test we have to make a **proportional hazards assumption**: we assume that the ratio is the same over time.

The HR is the ratio of the risk of death in group 1 to the risk of death in group 2 and can be calculated as:

$$HR = \frac{O_1/E_1}{O_2/E_2}$$

Remark: The ***risk of death*** is the number of observed deaths divided by the population at risk, but this keeps changing due to the censoring. This is then approximated with the number of expected deaths.

Hazard Ratio

The hazard ratio can be interpreted as the chance of an event occurring in group 1 divided by the chance of the event occurring in the group 2 (or vice versa) of a study:

$$HR = \frac{O_1/E_1}{O_2/E_2}$$

- $HR = 1$: No effect
- $HR < 1$: Reduction (increase) in the hazard (survival)
  Group 1 is a protective factor
- $HR > 1$: Increase (reduction) in hazard (survival)
  Group 1 is a risk factor

↓

You need to consider not only the exact value of the HR, but also its **confidence interval**. The direct calculation of the confidence interval for HR based on log-rank test is tedious (we omit it).

Dataset: Hazard Ratios

Group k	O_k	E_k
Female	59	75.9
Male	126	109.1

$$HR = \frac{59/75.9}{126/109.1} = 0.673$$

- The risk of deaths in females is 0.673 times the risk of death in males.
- $HR < 1$: females have higher survival probability than males.
- Being a female is a protective factor.

Group k	O_k	E_k
Not terminated	53	33.4
Terminated	132	151

$$HR = \frac{53/33.4}{132/151} = 1.822$$

- The risk of deaths in subjects who have not terminated the therapy is 1.822 times the risk of who have terminated the therapy.
- $HR > 1$: subjects who have not terminated the therapy have lower survival probability.
- Having not terminated the therapy is a risk factor.

References

- Aalen O. *Nonparametric estimation of partial transition probabilities in multiple decrement models.* The Annals of Statistics 1978; 6(3):534-545.
- Aalen O, Borgan O and Gjessing HK. *Survival and Event history Analysis: A Process Point of View.* Springer, New York, 2008.
- Bland M. *An Introduction to Medical Statistics - 4th Edition.* Oxford University Press, 2015.
- Greenwood M. *The natural duration of cancer.* Reports on Public Health and Medical. Her Majesty's Stationery Office, London. 1926; 33:1-26.
- Hosmer DW, Lemeshow S and May S. *Applied Survival Analysis: Regression Modeling of Time to Event Data* 2nd ed. Wiley-Interscience, USA, 2008.
- Kalbfleisch JD and Prentice RL. *The statistical Analysis of Failure Time Data.* 2nd ed. Wiley, New York, 2002.
- Kaplan E and Meier P. *Nonparametric estimation from incomplete observations.* Journal of American Statistical Association. 1958; 53:457-481.
- Kleinbaum DG and Klein M. *Survival Analysis: A Self-Learning Text.* Springer, New York, 1996.
- Mantel N. *Evaluation of survival data and two new rank order statistics arising in its consideration.* Cancer Chemotherapy Reports. 1966; 50(3):163-70.

Cox Proportional Hazard Model
with Clinical Application

Biostatistics Course
MSc in Bioinformatics and Computational Genomics – A.Y. 2019/2020

Practical session 4
Teaching Assistant: Marta Spreafico

POLITECNICO
MILANO 1863

Index

Dataset: **Randomized clinical trial** of chemotherapy in 478 patients with osteosarcoma.

Goals of survival analysis

Estimate the survival function for **a group** of individuals
↓
Kaplan-Meier estimator

Compare survival functions between two or more groups
↓
Log-rank test & Hazard Ratio

Assess how the **covariates** affect the hazard function
↓
Cox, frailty, parametric models

Cox PH model

= how to use covariates to model $h(\cdot)$ in:
$S(t) = \exp\{-\int_0^t h(u)du\}$ $t \geq 0$

(we want to model $h(\cdot)$ as we model any other thing (linear regression, etc.))

The Proportional-Hazard Cox Model (1972)

The Cox model is a statistical technique for exploring the relationship between the survival of a patient and several explanatory variables.

Hazard function for i-th patient

$$h_i(t|\boldsymbol{X}_i) = h_0(t)\exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}\}$$

hazard function (= function that describes the rate of an event to happen w.r.t. time)

We put into different parts the dependency of time and the dependency of the covariates

- $\boldsymbol{X}_i \in \mathbb{R}^p$ is the **covariates** vector of i-th patient
- $h_0(t)$ is an unspecified non-negative function of time called **baseline hazard**
- $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of **coefficients** that we want to estimate

The Cox PH model is:

- a **semiparametric** model since it has the property that the baseline hazard $h_0(t)$ is an unspecified function
- a **"robust"** model since it will closely approximate the correct parametric model

The Proportional-Hazard Cox Model

The quantities $\exp(\beta_l)$ are called Hazard Ratios (HR_l):

- $HR_l = 1$ ($\beta_l = 0$): No effect
- $HR_l < 1$ ($\beta_l < 0$): Reduction (increase) in the hazard (survival)
  $\rightarrow$ l-th covariate is a good prognostic factor
- $HR > 1$ ($\beta_l > 0$): Increase (reduction) in hazard (survival)
  $\rightarrow$ l-th covariate is a bad prognostic factor

The Cox model is also known as **Proportional-Hazard** model because the hazard ratio HR for two patients with fixed covariate vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_k$

$$HR = \frac{h_i(t|\boldsymbol{X}_i)}{h_k(t|\boldsymbol{X}_k)} = \frac{h_0(t)\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{h_0(t)\exp(\boldsymbol{X}_k^T\boldsymbol{\beta})} = \frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{X}_k^T\boldsymbol{\beta})} = \exp\{(\boldsymbol{X}_i - \boldsymbol{X}_k)^T\boldsymbol{\beta}\}$$

is constant over time.

Cox Partial Likelihood

Inference in Cox model is done on the so-called **partial likelihood**, that considers probabilities only for those subjects who fail, and does not explicitly consider probabilities for those subjects who are censored.

Let:

- $t_1, t_2, \ldots, t_N$ be the observed survival time for N individuals
- J be the total number of deaths in $\mathcal{D}_N$
- $0 < t_1^* < t_2^* < \cdots < t_J^* < \infty$ be the ordered observed deaths times
- $R(t_j^*)$ be the risk set just before t_j^*

The conditional probability that the j-th individual dies at t_j^* given that one individual from the risk set on $R(t_j^*)$ dies at t_j^* is (see Appendix A)

$$L_j = \frac{\exp(\boldsymbol{X}_j^T\boldsymbol{\beta})}{\sum_{k\in R(t_j^*)}\exp(\boldsymbol{X}_k^T\boldsymbol{\beta})} \quad \rightarrow \quad \text{Does not depend on } h_0(t)!$$

MLE of partial likelihood

Then the Cox partial likelihood $\mathcal{L}(\boldsymbol{\beta})$ is formulated as the product of each of the J conditional probabilities L_j:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{j=1}^{J} L_j = \prod_{j=1}^{J} \frac{\exp(\boldsymbol{X}_j^T\boldsymbol{\beta})}{\sum_{k\in R(t_j^*)}\exp(\boldsymbol{X}_k^T\boldsymbol{\beta})}$$

The corresponding log-likelihood is:

$$\ell(\boldsymbol{\beta}) = \ln(\mathcal{L}(\boldsymbol{\beta})) = \sum_{j=1}^{J}\left[\boldsymbol{X}_j^T\boldsymbol{\beta} - \ln\left(\sum_{k\in R(t_j^*)}\exp(\boldsymbol{X}_k^T\boldsymbol{\beta})\right)\right]$$

Therefore:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\, \ell(\boldsymbol{\beta})$$

Our Dataset: Covariates

We consider $N = 478$ patients with $J = 185$ dead patients. For each patient i, we have the following information:

$$\boldsymbol{X}_i = (\texttt{age}_i,\ \texttt{gender}_i,\ \texttt{trt}_i,\ \texttt{terminated}_i)$$

where

- $\texttt{age}_i$ = **age** at randomization
  $Q_1 = 12.2 \quad Q_2 = 15.2 \quad Q_3 = 18.4$
- $\texttt{gender}_i$ = **gender** at randomization
  $Female = 194\ (40.6\%) \quad Male = 284\ (59.4\%)$
- $\texttt{trt}_i$ = **chemotherapy** treatment
  $Conventional = 237\ (49.6\%) \quad Dose\ Intense = 241\ (50.4\%)$
- $\texttt{terminated}_i$ = therapy **terminated** or not
  $No = 102\ (21.3\%) \quad Yes = 376\ (78.7\%)$

Our Dataset: Cox model

We consider the following Cox PH model:

$$h_i(t|\boldsymbol{X}_i) = h_0(t)\exp\{\beta_1\texttt{age}_i + \beta_2\texttt{gender}_i + \beta_3\texttt{trt}_i + \beta_4\texttt{terminated}_i\}$$

	$\widehat{\beta}_l$	$se(\widehat{\beta}_l)$	$z = \widehat{\beta}_l/se(\widehat{\beta}_l)$	p	
age	0.0185	0.0114	1.6162	0.10604	
gender ($Male$)	**0.4189**	0.1591	2.6338	0.00844	**
trt (DI)	-0.0404	0.1479	-0.2728	0.78498	
terminated (Yes)	**-0.5932**	0.1669	-3.5535	0.00038	***

	$HR_l = \exp(\widehat{\beta}_l)$	$IC_{HR_l}(0.95)$
age	1.0186	[0.9961–1.0417]
gender ($Male$)	**1.5203**	[1.1131–2.0764]
trt (DI)	0.9604	[0.7187–1.2835]
terminated (Yes)	**0.5526**	[0.3984–0.7664]

Adjusted Survival Curves

Cox adjusted survival curves

The survival curves obtained through a Cox model that adjust for the explanatory variables are called **adjusted survival curves**.

For each patient i, the corresponding survival function is:

Adjusted survival function for i-th patient

$$S_i(t|\boldsymbol{X}_i) = [S_0(t)]^{\exp(\boldsymbol{X}_i^T\beta)} \quad \text{with} \quad S_0(t) = \exp\left\{-\int_0^t h_0(u)du\right\}$$

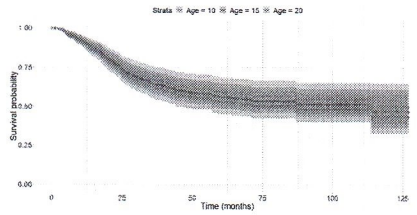
$$\downarrow$$

$$\widehat{S}_i(t|\boldsymbol{X}_i) = [\widehat{S}_0(t)]^{\exp(\boldsymbol{X}_i^T\widehat{\beta})}$$

where $\widehat{S}_0(t) = \prod_{j:t_j^* < t}\left(1 - \frac{1}{\sum_{k\in R(t_j^*)}\exp(\boldsymbol{X}_k^T\widehat{\beta})}\right)$

Our Dataset: Survival curves (I)

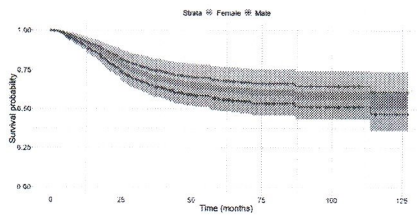
- Survival probability plot for three male patients with terminated chemotherapy in regimen DI aged 10, 15 and 20:



$S_{\mathbf{10}}(t = 36) = 67.2\%$ $S_{15}(t = 36) = 64.7\%$ $S_{\mathbf{20}}(t = 36) = 62\%$

Our Dataset: Survival curves (II)

- Survival probability plot for two patients aged 15 with terminated chemotherapy in regimen DI, stratified by gender:



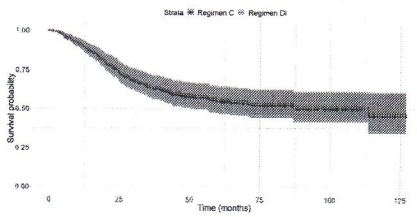
$$\downarrow$$

Being a female increases the survival probability

$S_{\text{Female}}(t = 36) = 75.1\%$ $S_{\mathbf{Male}}(t = 36) = 64.7\%$

Our Dataset: Survival curves (III)

- Survival probability plot for two male patients aged 15 with terminated therapy, stratified by chemotherapy treatment:



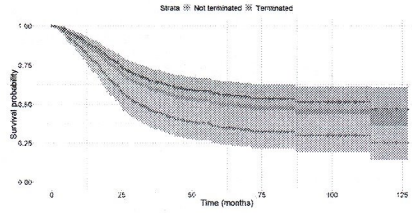
$$\downarrow$$

No treatment effect

$S_{\mathbf{C}}(t = 36) = 63.6\%$ $S_{\text{DI}}(t = 36) = 64.7\%$

Our Dataset: Survival curves (IV)

- Survival probability plot for two male patients aged 15 with chemotherapy in regimen DI, stratified by terminated therapy:



↓

Having terminated therapy increases the survival probability

$S_{\text{No}}(t = 36) = 45.5\%$ $S_{\text{Yes}}(t = 36) = 64.7\%$

Assessment of Cox model assumptions

Assessment of Cox model assumptions

When used inappropriately, statistical models may give rise to misleading conclusions. Therefore, it is important to check that a given model is an appropriate representation of the data.

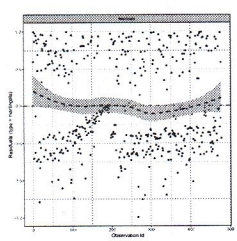
- **Goodness of Fit**
  - Martingale residuals (or deviance residuals)
  - Schoenfeld residuals
- **PH Assumptions**
  - Log-negative-log-KM
  - Test if β_l is constant over time ($l = 1, \ldots, p$)
  - Fit a regression model where variable interact with time

GOF – Martingale residuals

A first graphical option to check for goodness of fit is to check the Martingale Residuals:

$$M(t) = N(t) - \Lambda(t, \mathbf{X}, \boldsymbol{\beta})$$

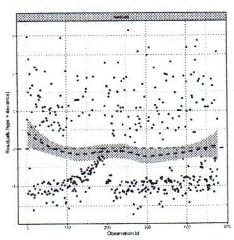
where $N(t)$ is the counting process for the event of interest, have **0 mean** along time.



Sometimes, martingale residuals are difficult to be interpreted.

GOF – Deviance residuals

The deviance residual is a **normalized transform of the martingale** residual. These residuals should be roughly **symmetrically distributed about zero** with a standard deviation of 1.

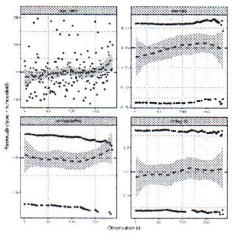


- Positive values correspond to individuals that "died too soon" compared to expected survival times.
- Negative values correspond to individual that "lived too long".
- Very large or small values are **outliers**, which are poorly predicted by the model.

GOF – Schoenfeld residuals

Shoenfeld residuals represent the difference between the **observed** covariate and the **expected given the risk** set at that time:

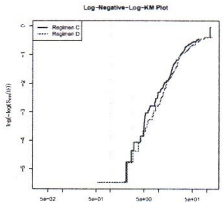
$$r_{il} \leftrightarrow \left(x_{il} - \widehat{\bar{x}}_{il}\right).$$



- They should be **flat**, centred about **zero**.
- In principle, they are independent of time.
- A plot that shows a non-random pattern against time is evidence of violation of the PH assumption.

PH Assumptions – Log-negative-log-KM

- If the **curves** seem **parallel**, PH assumption is satisfied.
- Only for **categorical** covariates. (If we categorize a numerical variables, different categorizations may give different graphical pictures).
- Subjective decision → Use a **conservative strategy**: assume PH is ok unless strong evidence of non-parallelism.



PH Assumptions – Quantitative strategies

As an alternative, some quantitative methods may be employed:

- Test if β_l is **constant over time** $(l = 1, \ldots, p)$
- Fit a Cox model where variable **interact with time** t (or $\log(t)$): not significant interactions provide evidence for PH assumption.
- `cox.zph()` function in `R`
  - For each covariate, the function correlates the corresponding set of **scaled Schoenfeld** residuals with time, to test for independence between residuals and time.
  - It also performs a **global test** for the model as a whole.
  - The PH assumption is supported by a **non-significant relationship between residuals and time**, and refuted by a significant relationship:
    H_0 : Hazards are proportional
    H_1 : Hazards are NOT proportional

Non-Proportional Hazards... and now what?

Possible approaches are possible in the context of the Cox model itself:

1. **Stratification**: covariates with nonproportional effects may be incorporated into the model as stratification factors rather than predictors.
   → *Warning:* Be careful! Stratification works naturally for categorical variables, however for quantitative variables you would have to discretize.
2. **Partition of the time axis**: if the proportional hazards assumption holds for short time periods but not for the entire study.
3. **Nonlinear effect**: continuous covariates with nonlinear effect may lead to nonproportional effects.

Stratified Cox PH Model

Stratified Cox PH Model

To stratify a Cox model we fit another Cox model in which we include predictors that satisfy the PH assumption and remove from it the predictor that is stratified, i.e. the one that violates PH.

Hazard function in the Stratified Cox model

$$h_k(t|\mathbf{X}) = h_{0k}(t)\exp\{\mathbf{X}^T\beta\}$$

- $k = 1, ..., K$ levels of the variable that is stratified
- $h_{0k}(t)$ is an unspecified non-negative function of time called **baseline hazard** for the k-th stratum
- $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are the vectors of covariates and coefficients, respectively.

In the stratified Cox model the regression coefficients are assumed to be the same for each stratum, although the baseline hazard functions may be different and completely unrelated.

References

- Aalen O. *Nonparametric estimation of partial transition probabilities in multiple decrement models.* The Annals of Statistics 1978; 6(3):534-545.
- Aalen O, Borgan O, Gjessing HK, *Survival and Event history Analysis: A Process Point of View.* Springer, New York, 2008.
- Bland M. *An Introduction to Medical Statistics - 4th Edition.* Oxford University Press, 2015.
- Cox DR. *Regression models and life-tables.* Journal of the Royal Statistical Society. 1972; 34:187-220.
- Cox DR. *Partial likelihood.* Biometrika. 1975; 62:269-276.
- Greenwood M. *The natural duration of cancer.* Reports on Public Health and Medical. Her Majesty's Stationery Office, London. 1926; 33:1-26.
- Hosmer DW, Lemeshow S and May S. *Applied Survival Analysis: Regression Modeling of Time to Event Data* 2nd ed. Wiley-Interscience, USA, 2008.
- Kalbfleisch JD, Prentice, RL. *The statistical Analysis of Failure Time Data.* 2nd ed. Wiley, New York, 2002.
- Kaplan E, Meier P. *Nonparametric estimation from incomplete observations.* Journal of American Statistical Association. 1958; 53:457-481.
- Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text.* Springer, New York, 1996.
- Mantel N. *Evaluation of survival data and two new rank order statistics arising in its consideration.* Cancer Chemotherapy Reports. 1966; 50(3):163-70.

Appendix A – Conditional Probability in Cox Model

The conditional probability that the j-th individual dies at t_j^* given that one individual from the risk set on $R(t_j^*)$ dies at t_j^* is given by

$$\begin{aligned}
L_j &= P\left(\text{individual } j \text{ dies at } t_j^* | \text{ one death from the risk set } R(t_j^*) \text{ at } t_j^*\right) \\
&= \frac{P\left(\text{individual } j \text{ dies at } t_j^*\right)}{P\left(\text{ one death from the risk set } R(t_j^*) \text{ at } t_j^*\right)} \\
&= \frac{P\left(\text{individual } j \text{ dies at } t_j^*\right)}{\sum_{k\in R(t_j^*)} P\left(\text{individual } k \text{ dies at } t_j^*\right)} \\
&= \frac{\lim_{\Delta t\to 0}\left[P\left\{\text{individual } j \text{ dies at } [t_j^*, t_j^* + \Delta t)\right\}/\Delta t\right]}{\lim_{\Delta t\to 0}\left[\sum_{k\in R(t_j^*)} P\left\{\text{individual } k \text{ dies at } [t_j^*, t_j^* + \Delta t)\right\}/\Delta t\right]} \\
&= \frac{h_i(t_j^*)}{\sum_{k\in R(t_j^*)} h_k\left(t_j^*\right)} = \frac{h_0(t_j^*)\exp(\boldsymbol{X}_j^T\boldsymbol{\beta})}{\sum_{k\in R(t_j^*)} h_0(t_j^*)\exp(\boldsymbol{X}_k^T\boldsymbol{\beta})} \\
&= \frac{\exp(\boldsymbol{X}_j^T\boldsymbol{\beta})}{\sum_{k\in R(t_j^*)}\exp(\boldsymbol{X}_k^T\boldsymbol{\beta})} \quad \rightarrow \text{Does not depend on } h_0(t)!
\end{aligned}$$

Time-dependent Cox, Frailty & Parametric AFT models

Biostatistics Course
MSc in Bioinformatics and Computational Genomics – A.Y. 2019/2020

Practical session 5
Teaching Assistant: Marta Spreafico

POLITECNICO
MILANO 1863

Index

Goals of survival analysis

Estimate the survival function for **a group** of individuals
↓
Kaplan-Meier estimator

Compare survival functions between two or more groups
↓
Log-rank test & Hazard Ratio

Assess how the **covariates** affect the hazard function
↓
Cox, frailty, parametric models

Cox model with time-dependent covariates

The Proportional-Hazard Cox Model (1972)

Cox PH model – Hazard function

$$h(t|\boldsymbol{X}) = h_0(t)\exp\{\boldsymbol{\beta}^T\boldsymbol{X}\}$$

- $\boldsymbol{X} \in \mathbb{R}^p$ is the vector of **covariates**
- $h_0(t)$ is an unspecified non-negative **baseline hazard** function
- $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of **coefficients** that we want to estimate

Important features of Cox model:

1. the baseline hazard depends on t, but not on the covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$
2. the hazard ratio $\exp(\boldsymbol{\beta}^T\boldsymbol{X})$ depends on the covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$, but not on time t (PH assumption)

But there are cases where if we measure some of the X_l over time, they may vary (e.g. patient's performance status, biomarkers, etc)

Cox model with time-dependent covariates

Hazard function

$$h(t|\boldsymbol{X}(t)) = h_0(t)\exp\{\boldsymbol{\beta}^T\boldsymbol{X}(t)\}$$

- $\boldsymbol{X}(t)$ is the vector of possibly time-dependent **covariates**
- $h_0(t)$ is an unspecified non-negative **baseline hazard** function
- $\boldsymbol{\beta}$ is the vector of **coefficients** that we want to estimate

Observe that:

1. The hazard at time t depends (only) on the value of the covariates at that time, i.e $\boldsymbol{X}(t)$.
2. The regression effect of $\boldsymbol{X}(\cdot)$ is constant over time.

Some people do not call this model proportional hazards any more, because the hazard ratio $\exp\{\boldsymbol{\beta}^T\boldsymbol{X}(t)\}$ varies over time.

Inference: partial likelihood

- We still use the partial likelihood $\mathcal{L}(\beta)$ to estimate $\boldsymbol{\beta}$:

$$\mathcal{L}(\beta) = \prod_{j=1}^{J} L_j = \prod_{j=1}^{J} \frac{\exp\{\boldsymbol{\beta}^T\boldsymbol{X}_j(t_j^*)\}}{\sum_{k\in R(t_j)}\exp\{\boldsymbol{\beta}^T\boldsymbol{X}_k(t_j^*)\}}$$

where L_j is still the conditional probability that the j-th individual dies at t_j^* given that one individual from the risk set on $R(t_j^*)$ dies at t_j^*.

- Only difference: the values of $\boldsymbol{X}$ now **changes at each risk set**.

- Do we need to worry about correlated data, since a given subject has multiple observations? No, we do not: the likelihood equations at any time point use only one copy of any subject, the program picks out the correct row of data at each time (programming trick).

- Estimated coefficients: $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p}\ell(\beta)$.

Example – Stanford Heart transplant

$N = 103$ subjects with $J = 78$ death. Variables:

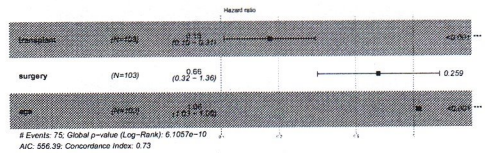
- `futime`: time from program enrolment until death or censoring
- `fustat`: indicator of death (1) or censoring (0)
- `transplant`: whether patient ever had transplant (1 if yes, 0 if no)
- `surgery`: previous heart surgery prior to program
- `age`: age at time of acceptance into program
- `wait.time`: time from acceptance into program until transplant

id	futime	fustat	transplant	surgery	age	wait.time
1	49	1	0	0	30.845	NA
2	5	1	0	0	51.836	NA
3	15	1	1	0	54.297	0
4	38	1	1	0	40.263	35
5	17	1	0	0	20.786	NA
6	2	1	0	0	54.595	NA

Example – Stanford Heart transplant

Initially, a Cox PH model was fitted for predicting survival time:

$$h(t) = h_0(t)\exp(\beta_1\texttt{transplant} + \beta_2\texttt{surgery} + \beta_3\texttt{age})$$



However, this model does not take into consideration that some patients had shorter waiting time to get transplants than others.

A model with a time dependent indicator of whether a patient had a transplant at each point in time might be more appropriate.

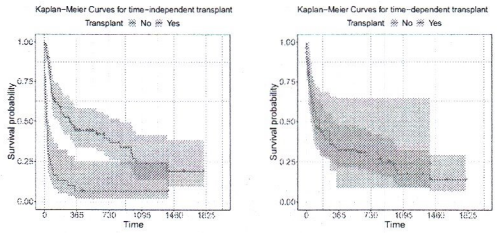
Example – Stanford Heart transplant

Long-format dataset:

- `start`: entry time for this interval of time
- `stop`: exit time for this interval of time
- `event`: indicator of death (1) or censoring (0)
- `transplant`: whether patient ever had transplant (1 if yes, 0 if no)
- `surgery`: previous heart surgery prior to program
- `age`: age at time of acceptance into program

id	start	stop	event	transplant	surgery	age
1	0	49	1	0	0	30.845
2	0	5	1	0	0	51.836
3	0	15	1	1	0	54.297
4	0	35	0	0	0	40.263
4	35	38	1	1	0	40.263
5	0	17	1	0	0	20.786

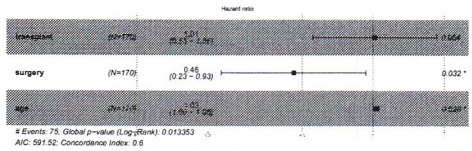
Example – Stanford Heart transplant



Example – Stanford Heart transplant

We considered the following Cox model for survival time with transplant time-dependent indicator:

$$h(t) = h_0(t)\exp(\beta_1 \texttt{transplant}(t) + \beta_2 \texttt{surgery} + \beta_3 \texttt{age})$$



Example – Stanford Heart transplant

Comparison with a single binary predictor:

- A Cox PH model with time-independent covariate would compare the **survival distributions between those without a transplant (ever) to those with a transplant**. A subject's transplant status at the end of the study would determine which category they were put into for the entire study follow-up.
- A Cox model with time-dependent covariate would compare the **risk of an event between transplant and non-transplant at each event time**, but would **re-evaluate** which risk group each person belonged in based on whether they'd had a transplant by that time.

Cox model with time-dependent coefficients

Cox model with time-dependent coefficients

Time dependent coefficients is another extension of a Cox model:

Hazard function

$$h(t|\boldsymbol{X}(t)) = h_0(t)\exp\{\boldsymbol{\beta}(t)^T\boldsymbol{X}\}$$

- $\boldsymbol{X}$ is the vector of **covariates**
- $h_0(t)$ is an unspecified non-negative **baseline hazard** function
- $\boldsymbol{\beta}(t)$ is the vector of possibly time-dependent **coefficients** that we want to estimate

Cox models with time-dependent coefficients are much less common, but represent one way to deal with **non-proportional hazards**.

The proportional hazard assumption is precisely that the coefficient does not change over time: $\beta(t) = c$.

Non-Proportional Hazards... and now what?

Possible approaches are possible in the context of the Cox model itself:

1. **Stratification**: covariates with nonproportional effects may be incorporated into the model as stratification factors rather than predictors.
   → *Warning:* Be careful! Stratification works naturally for categorical variables, however for quantitative variables you would have to discretize.
2. **Partition of the time axis**: if the proportional hazards assumption holds for short time periods but not for the entire study.
3. **Nonlinear effect**: continuous covariates with nonlinear effect may lead to nonproportional effects.

Partition of the time axis (I)

A partition of the time axis corresponds to the use of a **step function** for $\beta_l(t)$, i.e., different coefficients over different time intervals as follows:

$$\beta_l(t) = \sum_{k=1}^{K} \beta_{lk} I_{[t_{k-1};t_k)}(t)$$

where $0 = t_0 < t_1 < \cdots < t_K = \infty$ is the time partitioning, β_{lk} is the value assumed by $\beta_l(t)$ on interval $[t_{k-1}; t_k)$ and $I_{[t_{k-1};t_k)}(t)$ is the indicator function for interval $[t_{k-1}; t_k)$.

- Depending on how we dividing the intervals, the piecewise constant model **can approximate any shape** of $\beta_l(t)$.
- It has **simple interpretations**: the hazard ratio is β_{lk} from time t_{k-1} to time t_k.
- Without any other indications, we often take equal number of events per interval.

Partition of the time axis (II)

When $\beta_l(t)$ is **piecewise constant**, the non-PH model can be written as a *Cox model with time-dependent covariates*, as follows:

$$\begin{aligned}\beta_l(t) \cdot Z_l &= \sum_{k=1}^{K} \left(\beta_{lk} I_{[t_{k-1};t_k)}(t)\right) \cdot Z_l \\ &= \sum_{k=1}^{K} \beta_{lk} \cdot \left(I_{[t_{k-1};t_k)}(t) Z_l\right) \\ &= \sum_{k=1}^{K} \beta_{lk} \cdot Z_{lk}(t)\end{aligned}$$

where $Z_{lk}(t) = Z_l I_{[t_{k-1};t_k)}$.

This type of model is relatively easy to fit: we can use the `survSplit()` function to break the data set into time dependent parts (long-format dataset).

Nonlinear effect (I)

We can also assume that $\beta_l(t)$ has a **nonlinear functional form** as follows:

$$\beta_l(t) = \beta_{l1} + \beta_{l2} \cdot f(t)$$

and then fit an ordinary Cox model.

- A test of the parameter $\beta_{l2} = 0$ is a **test of the PH assumption**.
- The hazard ratio at time t for X_l is $\exp\{\beta_{l1} + \beta_{l2} f(t)\}$.
- Two simple particular forms that are often taken are

$$\beta_l(t) = \beta_{l1} + \beta_{l2} \cdot t \quad \text{or} \quad \beta_l(t) = \beta_{l1} + \beta_{l2} \cdot \log(t)$$

Nonlinear effect (II)

Also in that case we can look at $\beta_l(t)$ with functional form, as a *Cox model with time-dependent covariates*, as follows

$$\begin{aligned}\beta_l(t) \cdot Z_l &= (\beta_{l1} + \beta_{l2} \cdot f(t)) \cdot Z_l \\ &= \beta_{l1} \cdot Z_l + \beta_{l2} \cdot f(t) \cdot Z_l\end{aligned}$$

where $Z_l \cdot f(t)$ is the interaction term of l-th covariate Z_l with time or a function of time (time-dependent covariate).

The specification of the form $f(\cdot)$ for the interaction term $Z_l \cdot f(t)$ can be done using the **time-transform** functionality of `coxph`:

```
coxph(..., tt = function(x, t, ...)  x * f(t))
```

Frailty Cox models

Biological variation

- It is a basic observation of medical statistics that individuals are dissimilar.
- The natural course of a disease varies a lot from person to person. So does the effect of treatment, or the influence of various risk factors.
- This **heterogeneity** is often termed **biological variation** and it is generally recognized as one of the most important sources of variation in medicine and biology.
- In many clinical applications, the study population needs to be considered as a heterogeneous sample or as a cluster of homogeneous groups of individuals (families, geographical areas, etc).

Frailty survival models

The **frailty approach** is a statistical modelling method which aims to account for the heterogeneity caused by unmeasured covariates. It does so by adding **random effects** which act multiplicatively on the hazard function. [`frailtypack` package]

- **Shared frailty models**
  Used when observations are **clustered into groups** such as hospitals or cities, or when observations are **recurrent events times** (cancer relapses).
- **Nested frailty models**
  Used when observations are clustered at **several hierarchical levels** such as in geographical areas.
- **Joint frailty model**
  Used to assess the effect of recurrent events on the time-to-event of interest while taking into account the association.
- **Additive frailty model**
  Used to both looking at the heterogeneity between trials of underlying risk and treatment effects.

Shared frailty model

The hazard rate for the j-th recurrence of the i-th individual is

Hazard function

$$h_{ij}(t) = w_i h_0(t) \exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ij}\}$$

- $h_0(t)$ is an unspecified non-negative baseline hazard function
- $\boldsymbol{X}_{ij}$ is the the vector of explanatory variables, with coefficients $\boldsymbol{\beta}$
- w_i is the **shared frailty** for the i-th individual

We assume that the w_i are i.i.d from a distribution with $\mathbb{E}(w_i) = 1$ and $Var(w_i) = \theta$. The distribution has to be **positive** (Gamma, Log-normal). The parameter θ will allow us to assess the **variability between individuals**.

Remark. We used recurrent event terminology; nevertheless, grouped data can also be treated: j-th individual ($j = 1, \ldots, n_i$) of the i-th group ($i = 1, \ldots, G$).

Nested frailty model

The hazard rate for the k-th subject ($k = 1, \ldots, K_{ij}$) from the j-th subgroup ($j = 1, \ldots, n_i$) of the i-th group ($i = 1, \ldots, G$) is:

Hazard function

$$h_{ijk}(t) = w_i z_{ij} h_0(t) \exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ijk}\}$$

- $h_0(t)$ is an unspecified non-negative baseline hazard function
- $\boldsymbol{X}_{ijk}$ is the the vector of explanatory variables, with coefficients $\boldsymbol{\beta}$
- w_i is the **cluster random effect**
- z_{ij} is the **subcluster random effect**

The cluster random effect w_i and the subcluster random effect z_{ij} are both i.i.d and can be assumed gamma-distributed:

- $w_i \sim \Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$ i.i.d, with $\mathbb{E}(w_i) = 1$ and $Var(w_i) = \theta$
- $z_{ij} \sim \Gamma\left(\frac{1}{\eta}, \frac{1}{\eta}\right)$ i.i.d, with $\mathbb{E}(z_{ij}) = 1$ and $Var(z_{ij}) = \eta$

Parametric (AFT) models

Parametric Models

Basic Idea
The survival time follows a distribution.

Goal
Use data to estimate parameters of this distribution

⇒ completely specified model
⇒ prediction of time-quantiles

Parametric Survival Models vs Cox PH Models

Parametric Survival Model
+ Completely specified $h(t)$ and $S(t)$
+ More consistent with theoretical $S(t)$
+ Time-quantile prediction possible
− Assumption on underlying distribution

Cox PH Model
− Distribution of survival time unkonwn
− Less consistent with theoretical $S(t)$ (typically step function)
+ Does not rely on distributional assumptions
+ Baseline hazard not necessary for estimation of hazard ratio

Weibull Survival Model

The parametric **Weibull model** assumes that $T \sim Weibull(\lambda, k)$ with weibull probability density function

$$f(t) = \lambda k t^{k-1} \exp(-\lambda t^k) \qquad k > 0,\ \lambda > 0$$

where the survival and the hazard functions are given by

$$h(t) = \lambda k t^{k-1} \quad \text{and} \quad S(t) = \exp(-\lambda t^k).$$

- λ is the **rate** parameter (the **scale** parameter is $1/\lambda$)
- k is the **shape** parameter
  - If $k > 1$ the hazard increases
  - If $k = 1$ the hazard is constant (exponential model)
  - If $k < 1$ the hazard decreases

Exponential Survival Model

- The parametric **Exponential model** assumes that $T \sim Exp(\lambda)$ with exponential probability density function

$$f(t) = \lambda \exp(-\lambda t) \qquad \lambda > 0$$

where the survival and the hazard functions are given by

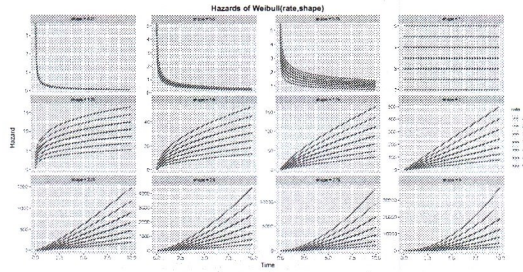
$$h(t) = \lambda \quad \text{and } S(t) = \exp(-\lambda t).$$

- λ is the **rate** parameter

- The exponential survival model is a **Proportional Hazard** model since its hazard $h(t)$ is constant over time.

- $T \sim Exp(\lambda) = Weibull(\lambda, 1)$.

Hazards function of $Weibull(\lambda, k)$



- λ is the rate parameter (the scale parameter is $1/\lambda$)
- k is the shape parameter

Accelerated Failure Time (AFT) models

- In the statistical area of survival analysis, an **accelerated failure time model** (AFT model) is a **parametric** model that provides an alternative to the commonly used proportional hazards models.
- Whereas a proportional hazards model assumes that the effect of a covariate is to multiply the hazard by some constant, an AFT model assumes that the effect of a covariate is to **accelerate** or **decelerate** the life course of a disease by some constant.
- This is especially appealing in a technical context where the 'disease' is a result of some mechanical process with a known sequence of intermediary stages.
- T Models describe **stretching out or contraction of survival** time as a function of predictor variables.

AFT model – Idea

Example: Smokers vs Nonsmokers

Let $S_S(t)$ and $S_{NS}(t)$ denote the survival functions for smokers and nonsmokers respectively.

- in terms of survival function:

$$S_{NS}(t) = S_S(\gamma t) \quad \text{for } t \geq 0$$

- in terms of random variables for survival time:

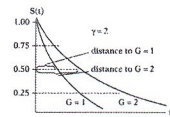
$$\gamma T_{NS} = T_S$$

where T_{NS} is a random variable following some distribution representing the survival time for nonsmokers and T_S the analogous one for smokers.

- $\gamma > 0$ is the constant **acceleration factor**

Acceleration factor

- The acceleration factor γ allows to evaluate the effect of predictor variables on the survival time.
- The acceleration factor is a **ratio of time-quantiles** corresponding to any fixed value of $S(t)$



- Assuming the event to occur is negative for an individual, comparing two groups leads to the following general interpretation:
  - $\gamma > 1 \rightarrow$ exposure benefits survival
  - $\gamma < 1 \rightarrow$ exposure harmful to survival
  - $\gamma = 1 \rightarrow$ no effect from exposure

General Form of AFT models

• Let T denote a continuous non-negative random variable representing survival time, with probability density function $f(t)$.
• A clinically plausible way to describe survival time is the equation given by the **AFT model** (case of single covariate):

$$T = \exp\{\beta_0 + \beta_1 X\} \cdot \varepsilon$$

that can be linearized as follows

$$\ln(T) = \beta_0 + \beta_1 X + \ln(\varepsilon) = \beta_0 + \beta_1 X + \varepsilon^*$$

where ε^* is a random error following some distribution.

- AFT model is:
  - **multiplicative** in terms of T
  - **additive** in terms of $\ln(T)$

General Form of AFT models

We assume that the error component ε^* follows the **extreme minimum value distribution**:

$$\varepsilon^* \sim G(0, \sigma)$$

- If $\sigma = 1 \quad \rightarrow \quad T \sim Exp(\lambda)$
- If $\sigma \neq 1 \quad \rightarrow \quad T \sim Weibull(\lambda, k)$

As alternatives, we could assume:

- $\varepsilon^* \sim Logistic \quad \rightarrow \quad T \sim Log - logistic$
- $\varepsilon^* \sim Normal \quad \rightarrow \quad T \sim Log - Normal$

Weibull AFT model

In the **Weibull AFT model**: $\lambda = \exp[-k(\beta_0 + \beta_1 X)]$ and $k = \frac{1}{\sigma}$.

- Survival function:

$$S(t, X, \boldsymbol{\beta}) = \exp\left\{-\lambda t^k\right\} = \exp\left\{-t^k \exp[-k(\beta_0 + \beta_1 X)]\right\}$$

- Hazard function:

$$h(t, X, \boldsymbol{\beta}) = \lambda k t^{k-1} = k t^{k-1} \exp[-k(\beta_0 + \beta_1 X)] = \frac{k t^{k-1}}{\exp[\beta_0 + \beta_1 X]^k}$$

- The acceleration factor for X dichotomous covariate is:

$$\gamma = \exp(\beta_1)$$

- The hazard ratio for X dichotomous covariate is:

$$HR = \exp(-\beta_1 \cdot k) = \exp\left(-\frac{\beta_1}{\sigma}\right)$$

Exponential AFT model

In the **Exponential AFT model**: $\lambda = \exp[-(\beta_0 + \beta_1 X)]$.

- Survival function:

$$S(t, X, \boldsymbol{\beta}) = \exp\{-\lambda t\} = \exp\{-t \exp[-(\beta_0 + \beta_1 X)]\}$$

- Hazard function:

$$h(t, X, \boldsymbol{\beta}) = \lambda = \exp[-(\beta_0 + \beta_1 X)]$$

- The acceleration factor for X dichotomous covariate is:

$$\gamma = \exp(\beta_1)$$

- The hazard ratio for X dichotomous covariate is:

$$HR = \exp(-\beta_1)$$

PH assumption

Both Weibull and Exponential AFT models preserve the **PH assumption**.

Indeed, the following proposition can be proved

Proposition – Weibull model

AFT assumption holds $\Leftrightarrow$ PH assumption holds (given that k is fixed)

Proof for the dichotomous example ($X = 1$ and $X = 0$):

- $[\Rightarrow]$ $\gamma = \exp(\beta_1)$
  Assume γ is constant $\Rightarrow \beta_1$ is constant
  $HR = \exp(-k\beta_1) \Rightarrow HR$ is constant

- $[\Leftarrow]$ $HR = \exp(-k\beta_1)$
  Assume HR is constant $\Rightarrow \beta_1$ is constant
  $\gamma = \exp(\beta_1) \Rightarrow \gamma$ is constant

References

- Aalen O, Borgan O, Gjessing HK, *Survival and Event history Analysis: A Process Point of View*. Springer, New York, 2008.
- Bland M. *An Introduction to Medical Statistics - 4th Edition*. Oxford University Press, 2015.
- Cox DR. *Regression models and life-tables*. Journal of the Royal Statistical Society. 1972; 34:187-220.
- Cox DR. *Partial likelihood*. Biometrika. 1975; 62:269-276.
- Hosmer DW, Lemeshow S and May S. *Applied Survival Analysis: Regression Modeling of Time to Event Data* 2nd ed. Wiley-Interscience, USA, 2008.
- Kalbfleisch JD, Prentice, RL. *The statistical Analysis of Failure Time Data*. 2nd ed. Wiley, New York, 2002.
- Kaplan E, Meier P. *Nonparametric estimation from incomplete observations*. Journal of American Statistical Association. 1958; 53:457-481.
- Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text*. Springer, New York, 1996.
- Rondeau V, Mazroui Y, Gonzalez JR. (2012). *frailtypack: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation*. Journal of Statistical Software. 2012; 47(4), 1-28.
- Therneau TM. *Package 'survival'. A Package for Survival Analysis in R*. R package version 3.1-12, 2020.