POLITECNICO
MILANO 1863

Nonparametric estimation in Survival Analysis

Nonparametric Statistics

AA 2022-2023

Francesca Ieva

MOX – Department of Mathematics, Politecnico di Milano, Italy

Outline

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

Outline

Introduction

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

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**.

- **Time** $\rightarrow$ years, months, weeks or days from the beginning of follow up of an individual until an event occur
- **Event** $\rightarrow$ 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:

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

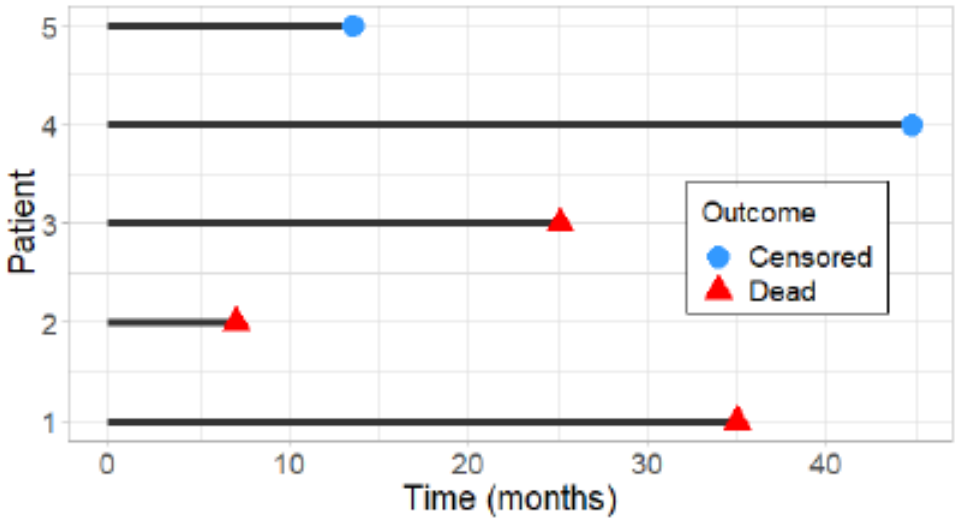
$$\downarrow$$

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 = $\{0 = \textit{Censored}, 1 = \textit{Dead}\}$



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:

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

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:

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

Therefore:

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

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:

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

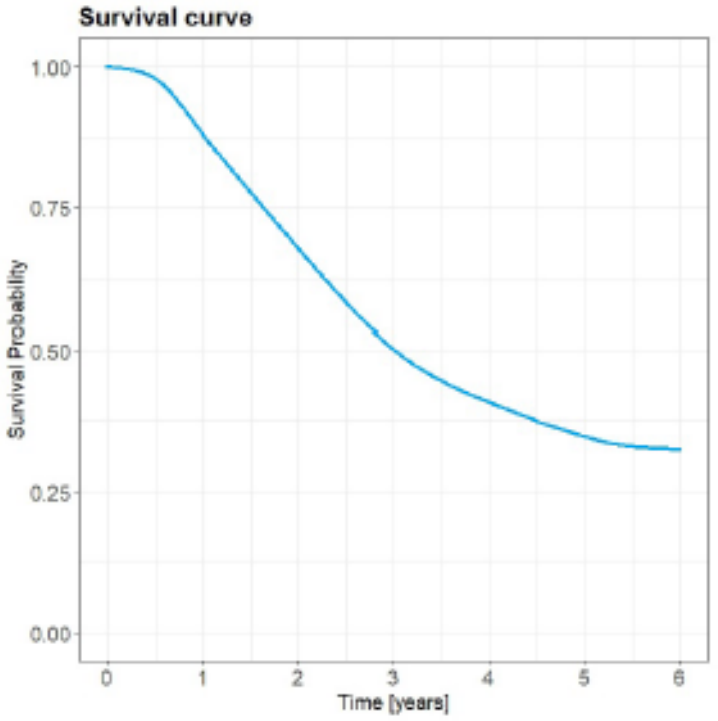
$= e^{-H(t)}$ *general form describing the Survival as a function of the time*

- 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 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



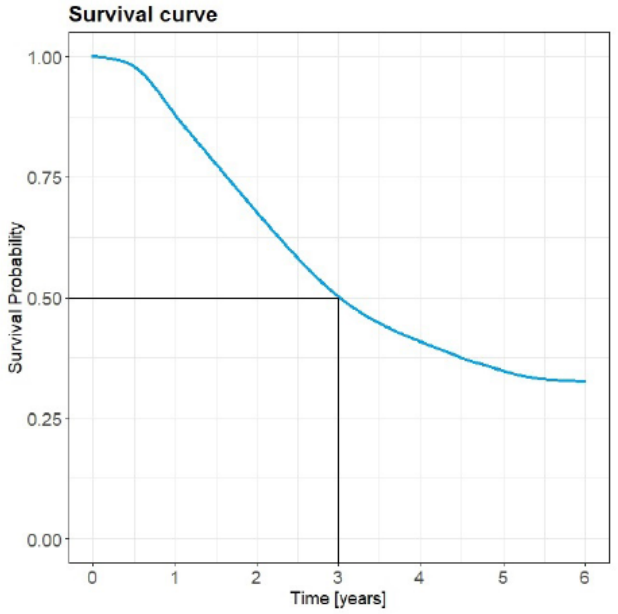
The survival function gives an estimate of these percentage for a given time between the beginning of the time interval and the end of the study.

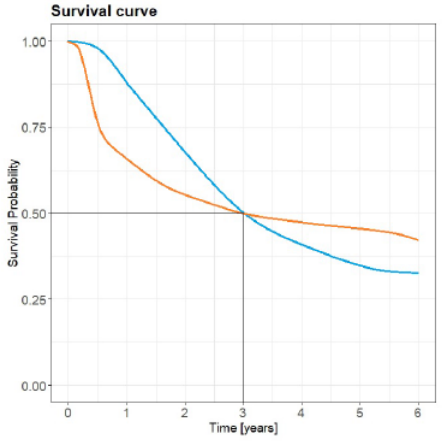
Median Survival Curve

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

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**

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:

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

- j = failure (event) index $\in \{1, ..., J\}$
- J = total number of individuals who experienced the event
- $0 < t_1^* < ... < t_J^* < \infty$ = 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

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

Kaplan-Meier estimator

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

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:

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

- j = failure (event) index $\in \{1, ..., J\}$
- J = total number of individuals with events
- $0 < t_1^* < ... < t_J^* < \infty$ = 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*

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

Continue……..

KM: NPMLE and Greenwood's formula

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

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

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

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

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

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

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:

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

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.

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

$\mathcal{D}_5$	i	1	2	3	4	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:

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

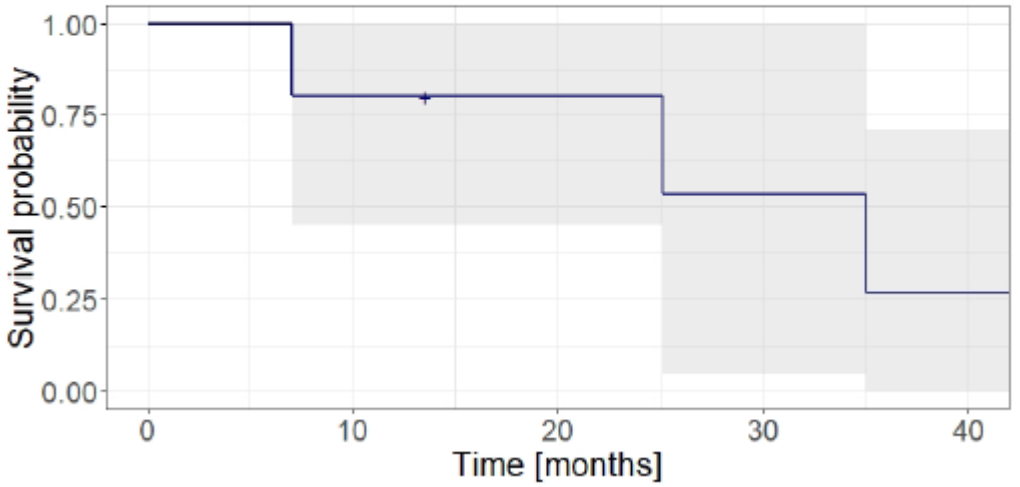
t_j^*	n_j	d_j	$p_j = (n_j - d_j)/n_j$	$\widehat{S}(t) = \prod_{j:t_j^* \leq t} p_j$
7.03	5	1	0.800	0.800
25.10	3	1	0.667	0.533
35.03	2	1	0.500	0.267

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

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

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

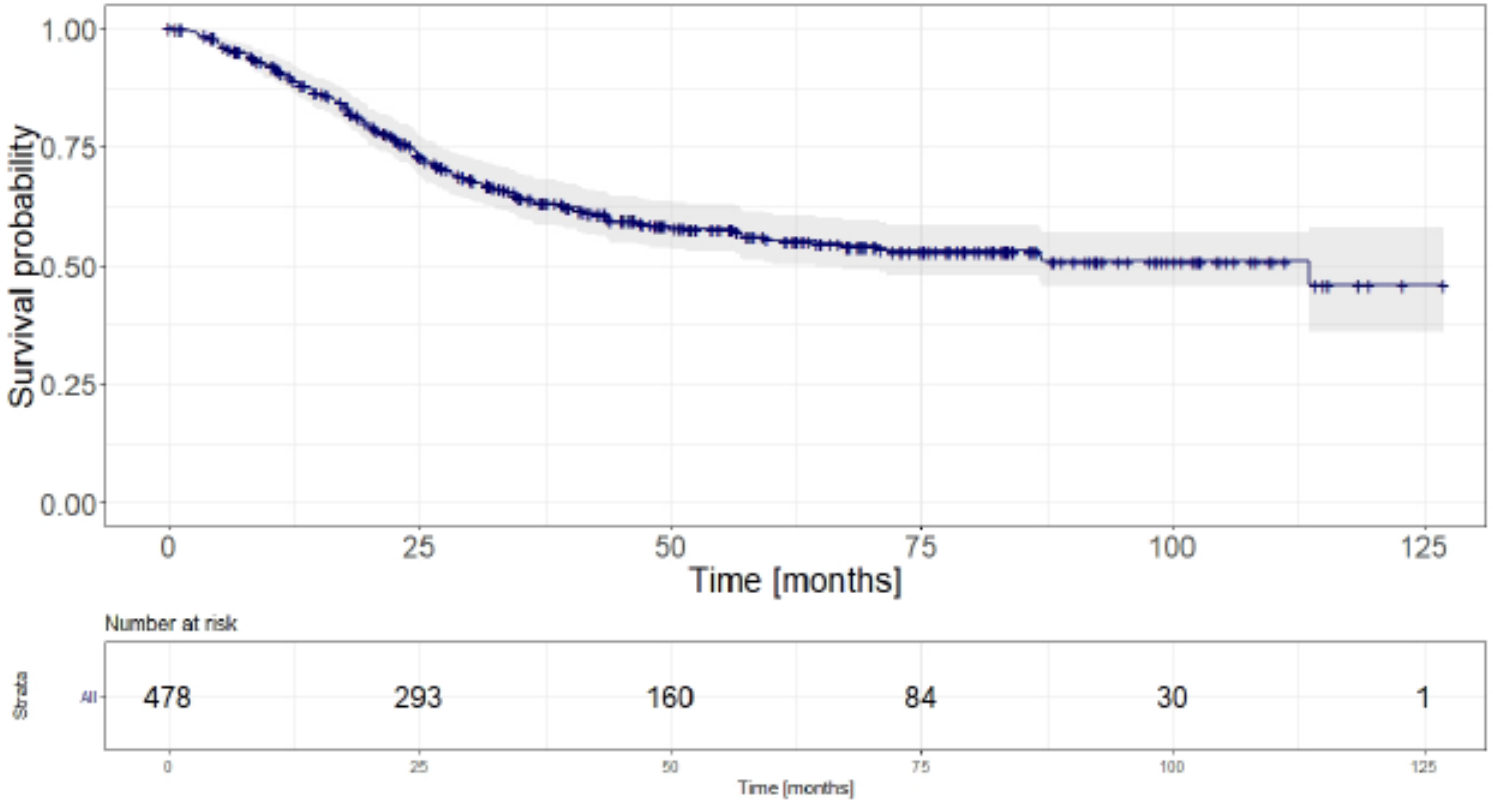
$t \in [t_j^*, t_{j+1}^*)$	$\hat{S}(t)$
$[0, 7.03)$	1.000
$[7.03, 25.10)$	0.800
$[25.10, 35.03)$	0.533
$[35.03, \infty)$	0.267



- 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

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:



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

Median survival time: $t = 113.6$ months

Log-rank test

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

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 $[t_j^*, t_{j+1}^*)$, $j = 0, ..., J$ 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.

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:

$$\chi^2 = \sum_{k=1,2} \frac{(O_k - E_k)^2}{E_k} \sim \chi^2_1$$

Log-rank test

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

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

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

$$\frac{O_k - E_k}{\sqrt{V}} \sim \mathcal{N}(0,1) \quad \text{or} \quad \chi^2 = \frac{(O_k - E_k)^2}{V} \sim \chi^2_1$$

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 $j = 1, ..., J$ compute:

- K is the number of total groups with $k = 1, 2, ..., K$
- 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

  $No = 102\ (21.3\%)$ $Yes = 376\ (78.7\%)$

- age_i = **age** at randomization

  $Q_1 = 12.2$ $Q_2 = 15.2$ $Q_3 = 18.4$

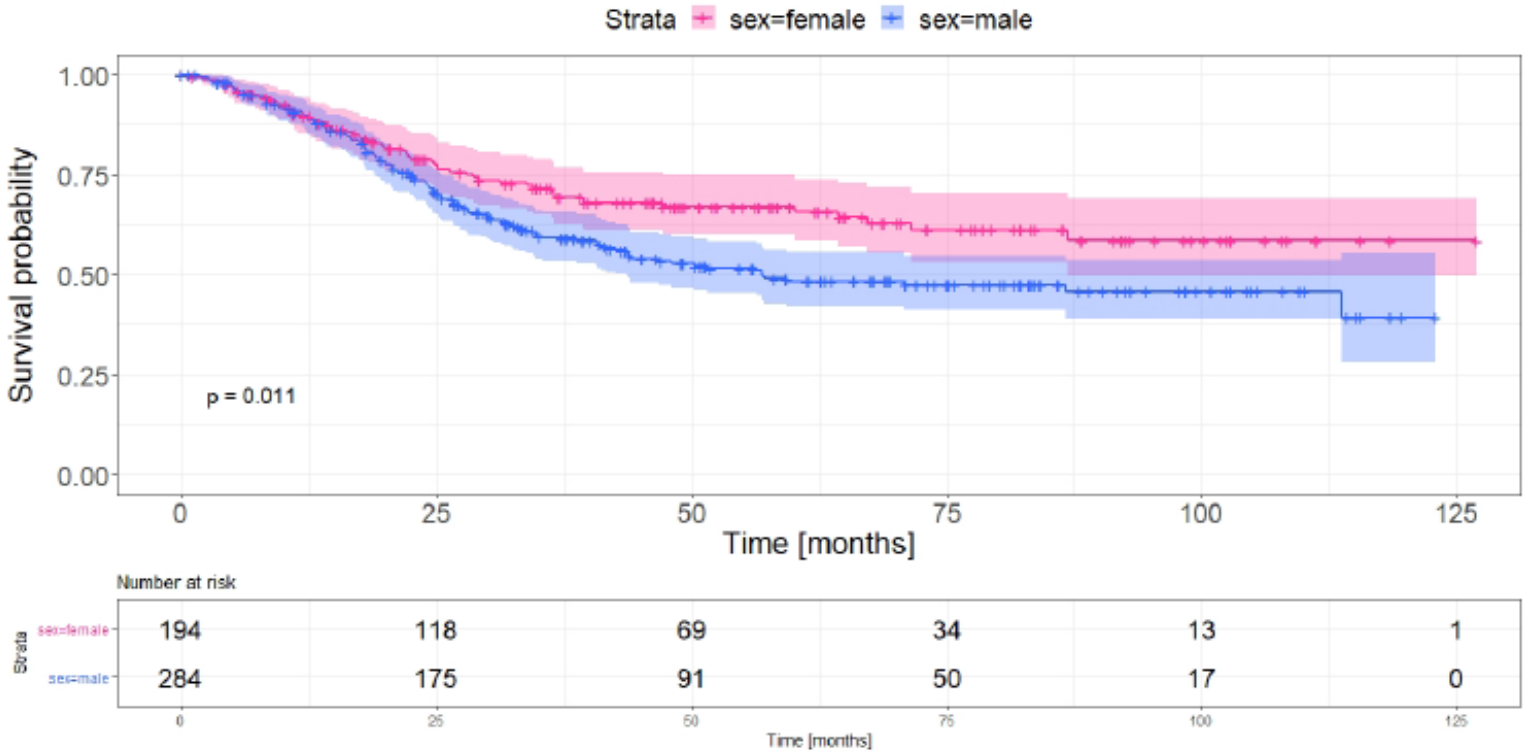
  for which we consider four groups:

  $0 - 12 = 114\,(23.9\%)$ $13 - 15 = 116\,(24.3\%)$

  $16 - 18 = 113\,(23.6\%)$ $18+ = 135\,(28.2\%)$

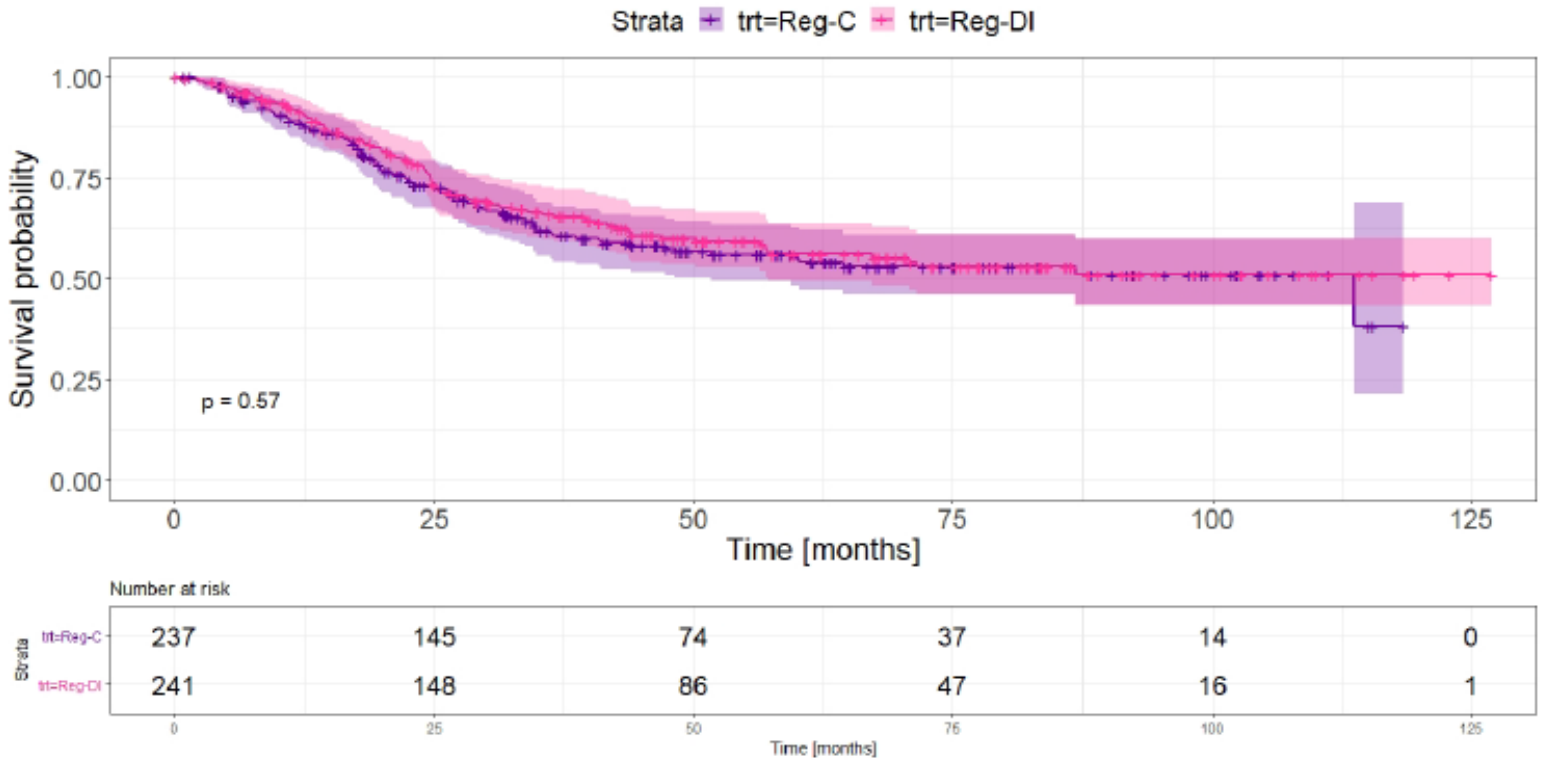
Dataset: log-rank test (I)

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



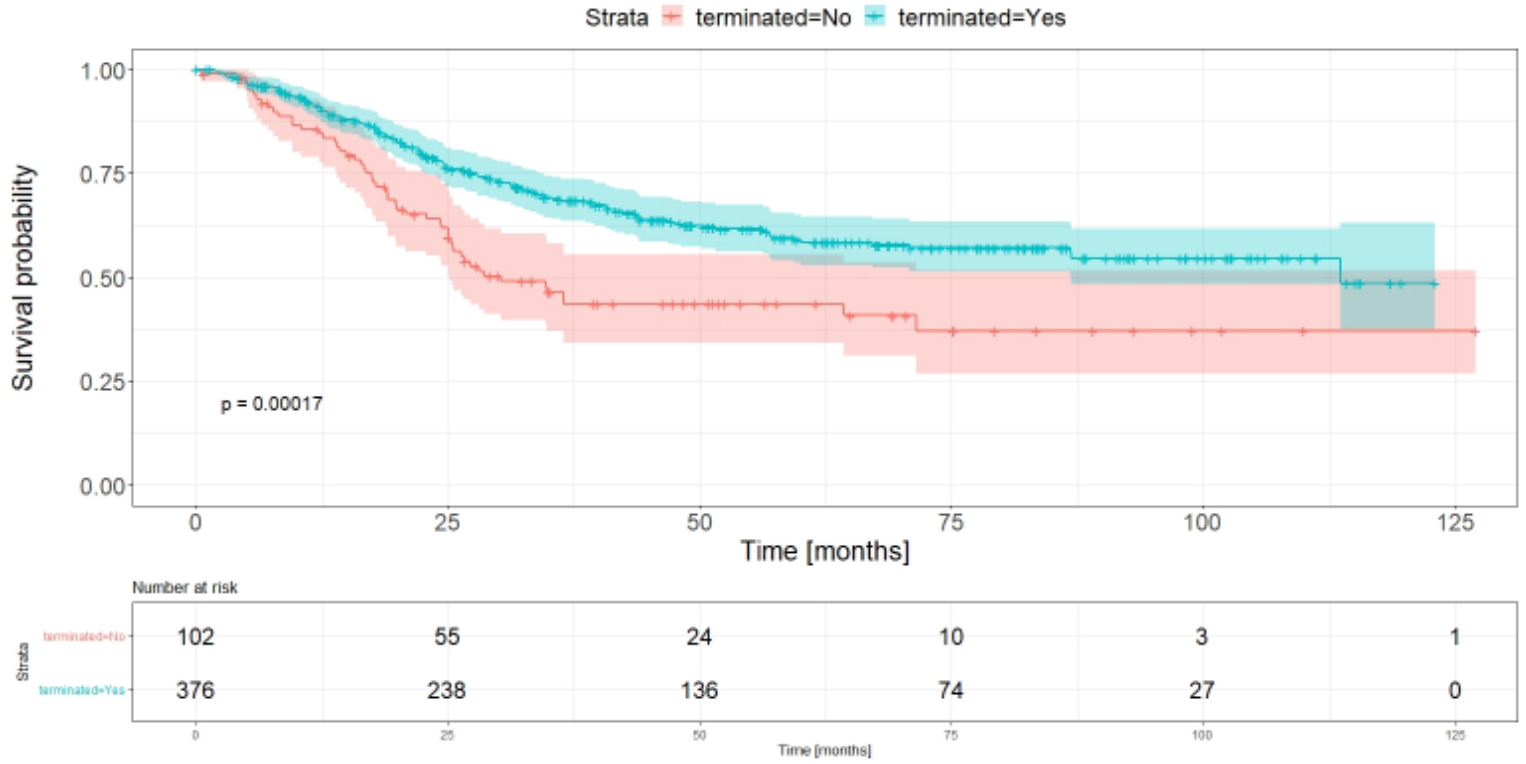
Dataset: log-rank test (II)

- 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 $H_0 \rightarrow S_{\mathbf{C}}(\cdot) = S_{\mathbf{DI}}(\cdot)$



Dataset: log-rank test (III)

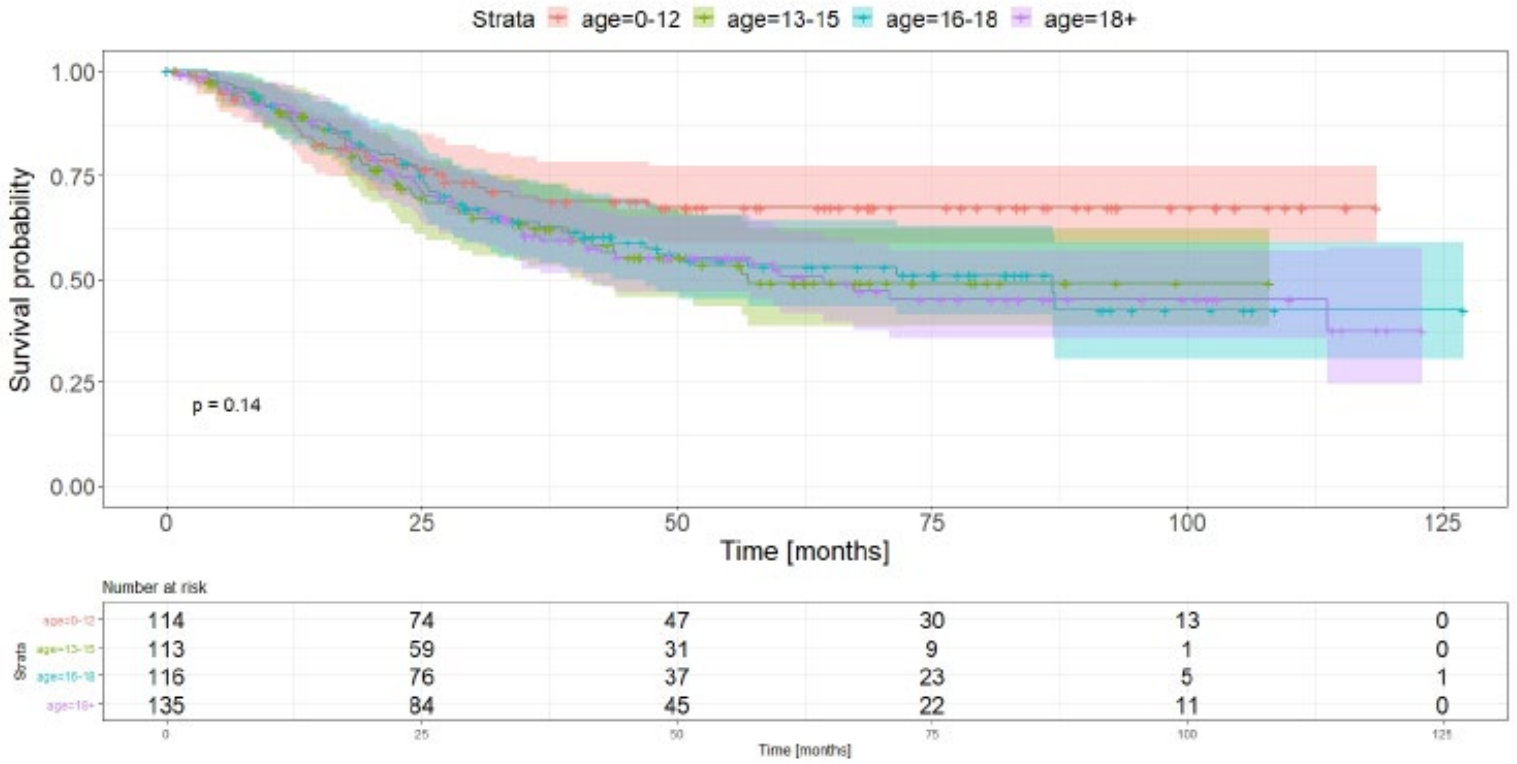
- Group $k = 1$: **Not terminated**
- Group $k = 2$: **Terminated**
- $\chi^2 = 14.1 \rightarrow p - value = 0.00017$
- Reject $H_0 \rightarrow S_{\text{No}}(\cdot) \neq S_{\text{Yes}}(\cdot)$



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

- $\chi^2 = 5.5 \rightarrow p - value = 0.14 \quad (\chi^2_{3,0.05} = 7.81)$
- Do not reject $H_0 \rightarrow S_{\text{0-12}}(\cdot) = S_{\text{13-15}}(\cdot) = S_{\text{15-18}}(\cdot) = S_{\text{18+}}(\cdot)$



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.

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

1. **Cox Proportional Hazard model**
2. Adjusted Survival Curves
3. Assessment of Cox model assumptions
4. Stratified Cox PH Model

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

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

- $\boldsymbol{X}_i \in \mathbb{R}^p$ 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 (1972)

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:

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

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]

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, York, 1996.
- Mantel N. *Evaluation of survival data and two new rank order statistics aris in its consideration.* Cancer Chemotherapy Reports. 1966; 50(3):163-70.

Appendix A

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

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