SUMMER SCHOOL IN STATISTICS 2015
HANOï VIETNAM

Survival Analysis

Part I

Jean-François Dupuy
Jean-Francois.Dupuy@ina-rennes.fr
and
Agnès Lagnoux
lagnoux@univ-tlse2.fr

Contents

Chapter 1

Introduction

In this lecture, our goal is to modelize times elapsed before the occurrence of an event. Historically, the studied events corresponded to deaths (that is why the term "lifetime" is currently used) but any event can be studied in this framework.

Numerous and various domains are concerned:

- medicine and health: survival time of the patient, remission time of a disease...
- reliability: component lifetimes,
- economy: loss of employment,
- psychology: learning time of some specific skill,
- sociology: succession life events (wedding, birth, divorce,...).

The statistical community has been particularly active on this research domain during the 20th century:

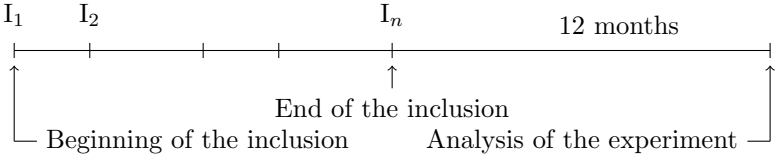
- Kaplan and Meier (1958): non parametric estimator of the survival function,
- Peto and Peto (1972): logrank test to compare the survival of two groups,
- Cox (1972): semi-parametric model that now is the most popular in survival data analysis,
- Gill (1980): introduction to the theory of processes ad martingales.

1.1 Why specific methods?

Example 1: We are interested in the survival of a group of mice after the injection of cancer cells; all the mice receive the injection at the same moment.
We can expect the death of all the mice and the exact lifetimes are observed and thus the standard methods apply.

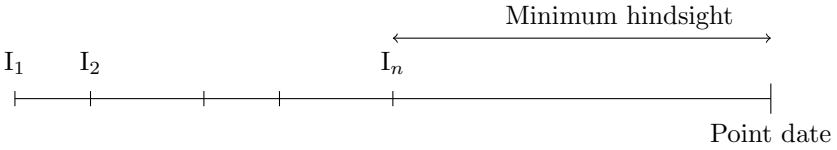
Example 2: Therapeutic trial (clinical research)
We want to test the efficiency of a vaccine against the hepatitis B.

- 2 randomized groups: treatment/placebo.
- Criterion: arisen a hepatitis B during the year that follows the injection.
- Various injection dates for the different patients.



We observed 6 hepatitis on the 184 patients that received the vaccine but 7 patients has not been controlled during an entire year.
How can we estimate the rate of hepatitis after 12 months?
$\frac{6}{184}$? $\frac{6}{184-7}$?

Example 3: Assume that we are interested in the survival after surgery of men operated for a bronchial cancer.
It is not realistic to wait for the death of all the subjects. Thus we fix a stopping time called **point date**.



Conclusion: Examples 2 and 3 deal with individuals for whom the event time is not known; the only thing that we know is that the event has occured in a certain interval: this notion is called **censoring**. Only dealing with the non censored observations led to a loss of information which can be not insignificant.

$\hookrightarrow$ One has to introduce specific methods to take into account the censored data.

1.2 Lifetime models

1.2.1 Lifetime and related functions

Let X be a non negative random variable having a known continuous distribution function. Assume that X corresponds to the delay between a fixed original time and the occurrence of an event of interest.

Example: X is a survival variable and the event under concern corresponds to the death.

The usual functions such that the cumulative distribution function (c.d.f) F and the density probability function (p.d.f.) f are not used in this research domain. One prefers working with more adapted functions that are easier to interpret:

- The **survival (or reliability) function** is defined as the probability to survive until time t:

$$S(t) = 1 - F(t) = \mathbb{P}(X > t)$$

The survival function is sometimes denoted $\bar{F}(\cdot)$.

The mathematical expectation of X can be written in terms of the survival function:

$$\mathbb{E}[X] = \int_0^{+\infty} x dF(x) = \int_0^{+\infty} S(x)dx.$$

Indeed, by an integration by parts,

$$\begin{aligned}\mathbb{E}[X] &= \int_0^{+\infty} x dF(x) \\ &= [-S(x)x]_0^{+\infty} + \int_0^{+\infty} S(x)dx \\ &= \int_0^{+\infty} S(x)dx.\end{aligned}$$

- The **hazard rate (or risk) function** is given by

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leqslant X < t + dt \mid X \geqslant t)}{dt}$$

for $t \geqslant 0$. $h(t)dt$ represents the probability to die during the interval t and $t + dt$ for an individual conditionally that he is still alive at time t; in other words, the hazard rate at point t is the instantaneous probability of failure (or death) at time t given that failure (or death) has not occurred before. The hazard rate function may have different shapes: the most famous is called the **bathup curve**.

The function $h(\cdot)$ satisfies the following relations:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \ln S(t).$$

As a consequence, one has

$$S(t) = \exp\left\{-\int_0^s h(s)ds\right\},$$
$$f(t) = h(t) \exp\left\{-\int_0^s h(s)ds\right\}.$$

Examples of hazard rate functions:

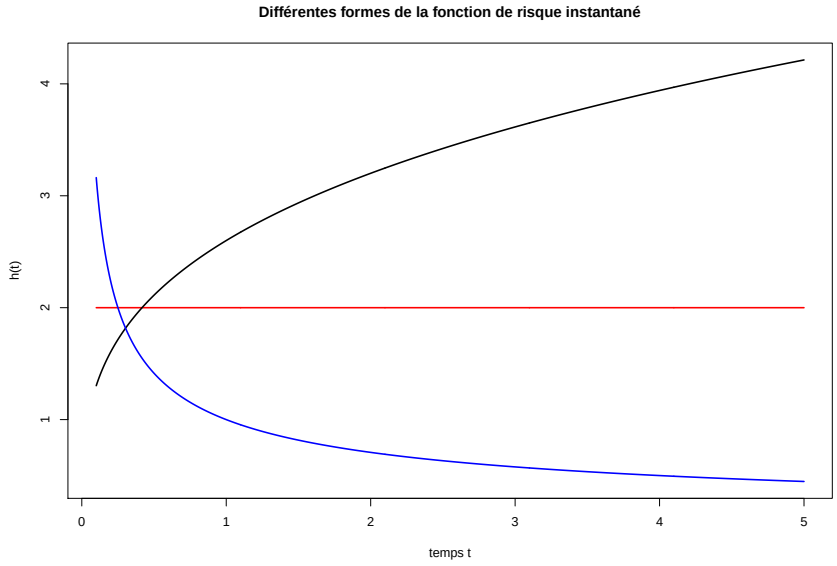
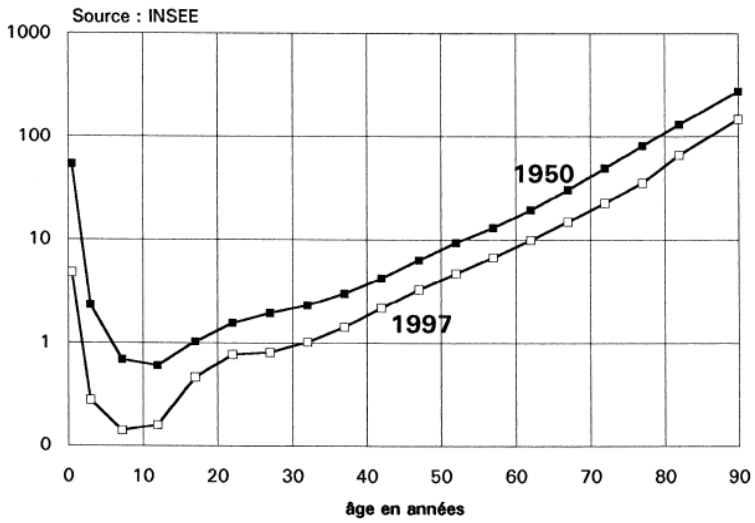
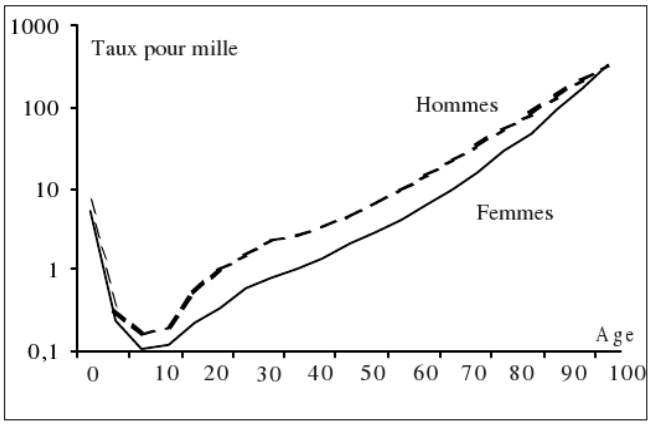


Figure 10 - TAUX DE MORTALITE PAR AGE EN FRANCE METROPOLITAINE (pour 1000 personnes) EN 1950 ET EN 1997



Taux de mortalité en Ile-de-France par sexe et par âge en 1993-95



Source : INSEE Exploitation ORS
Représentation logarithmique

- The **cumulative hazard rate function** is given by

$$H(t) = \int_0^t h(s)ds, \quad \text{for all} \quad t.$$

One has the following relationships:

$$S(t) = \exp\{-H(t)\},$$
$$f(t) = h(t)S(t) = h(t)\exp\left\{-\int_0^t h(s)ds\right\}.$$

Particular case: If the random variable X is discrete (i.e. valued in a countable set $\{x_1, x_2, \cdots, x_n, \cdots\}$ with $x_1 < x_2 < \cdots < x_n, \cdots$), we have

$$F(t) = \sum_{i:x_i \leqslant t} p_i$$

where $p_i = \mathbb{P}(X = x_i)$. The hazard rate function is given by

$$h(x_i) = \mathbb{P}(X = x_i | X \geqslant x_i) = \frac{p_i}{S(x_{i-1})}$$

and the cumulative rate function by

$$H(t) = \sum_{i:x_i \leqslant t} h(x_i).$$

Finally,

$$S(t) = \prod_{i:x_i \leqslant t} [1 - h(x_i)].$$

- The **residual life at time** t, denoted by τ_t, is the random variable with distribution:

$$\mathbb{P}(\tau_t > s) = \mathbb{P}(X - t > s | X \geqslant t) = \frac{S(t+s)}{S(t)}.$$

- The **mean residual function** is defined for $t \geqslant 0$ by

$$m(t) = \mathbb{E}[\tau_t] = \mathbb{E}[X - t | X \geqslant t] = \frac{\int_t^{+\infty} S(s)ds}{S(t)}.$$

1.2.2 Usual lifetime distributions

Let X be a non negative random variable. We are interested in random variables whose support is $\mathbb{R}^+$. As a consequence, the Gaussian distribution is no longer the reference distribution.

- The **exponential** $\mathcal{E}(\lambda)$ where $\lambda > 0$. One has, for any $t \geqslant 0$,

$$\begin{aligned} S(t) &= e^{-\lambda t} \\ f(t) &= \lambda e^{-\lambda t} \mathbb{1}_{t \geqslant 0} \\ h(t) &= \lambda. \end{aligned}$$

  The risk is then a constant function with respect to the time: this distribution is currently used in models without aging. One has $\mathbb{E}[X] = \frac{1}{\lambda}$ and $\mathrm{Var}(X) = \frac{1}{\lambda^2}$ and

$$\forall x, y > 0, \quad \mathbb{P}(X > x + y \mid X > y) = \mathbb{P}(X > x).$$

  The exponential is said to have the memoryless property: the distribution of the survival times for times greater than y is not affected by the knowledge that the individual has survived until y.

- The **Weibull distribution** $\mathcal{W}(\alpha, \lambda)$, where $\lambda > 0$ and $\alpha > 0$, is a generalization of the exponential distribution (that is a particular case when $\alpha = 1$). It allows to have increasing or decreasing risks. The parameter λ is the scale parameter and α is the shape parameter.

  Here one has, for any $t \geqslant 0$,

$$\begin{aligned} S(t) &= e^{-(\lambda t)^\alpha}, \\ f(t) &= \alpha \lambda^\alpha t^{\alpha-1} e^{-(\lambda t)^\alpha} \mathbb{1}_{t>0}, \\ h(t) &= \alpha \lambda^\alpha t^{\alpha-1}. \end{aligned}$$

  The risk is then a power of the time.

- when $0 < \alpha < 1$: the risk decreases from $+\infty$ to 0;
- when $\alpha = 1$: the risk is constant ($\mathcal{W}(1, \lambda) = \mathcal{E}(\lambda)$);
- when $\alpha \geqslant 1$: the risk increases from 0 to $+\infty$.

Property: If $X \sim \mathcal{W}(\alpha, \lambda)$ then $X^\alpha \sim \mathcal{E}(\lambda^\alpha)$.

Warning: in the software R, the scale parameter is $\frac{1}{\lambda}$.

- The **Gamma distribution** $\gamma(a, \lambda)$ where $\lambda > 0$ and $a > 0$:

$$f(t) = \frac{\lambda^a}{\Gamma(a)} t^{a-1} e^{-\lambda t} \mathbb{1}_{t>0}$$

where $\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx$.

The parameter λ is the scale parameter and α is the shape parameter.

Reminders on the Gamma function Γ:

$$\begin{aligned}
&\Gamma(1) = 1,\\
&\Gamma(\frac{1}{2}) = \sqrt{\pi},\\
&\forall a > 0,\ \Gamma(a+1) = a\Gamma(a),\\
&\forall n \in I\!N*, \Gamma(n) = (n-1)!.
\end{aligned}$$

In this context, there is no explicit expression for S nor h.
One has: $\mathbb{E}[X] = \frac{a}{\lambda}$ and $\mathrm{Var}(X) = \frac{a}{\lambda^2}$.

- when $0 < a < 1$: the risk decreases from $+\infty$ to $\frac{1}{\lambda}$;
- when$a = 1$: the risk is constant ($\gamma(\lambda, 1) = \mathcal{E}(\lambda)$);
- when $a \geqslant 1$: the risk increases from 0 to λ.

There exists other distributions that lead to monotone risks. To get non monotone risk functions, one can use the following distributions:

- the log-normal distribution,
- the log-logistic distribution,
- the inverse Gaussian distribution,
- the generalized Weibull distribution.

1.3 Censored data

1.3.1 Censoring and truncation

1. The data are often censored, that means that the event under concern may not be observed and one only knows that the delay of interest lies in a given time interval. There exists 3 different types of **censoring**:
   - right censoring (the most frequent): the event occurs after a given date C: $X > C$,

- left censoring: the event occurs before a given date C: $X < C$,
- censoring by interval: the event occurs between two given dates C_1 and C_2: $C_1 < X < C_2$.

2. The **truncation** is a condition that hide some individuals in such a way that the statistician is not aware of their existence: the individuals that not satisfy the condition are not included in the study.

   - Left truncation: the condition $X > Y$ guarantees that the individual takes part to the study where Y corresponds to the delay of occurrence of another event.

     Example: one is interested in the survival of individuals that have contracted a given disease and the sample is taken in an retirement home. The individuals prematurely dead (before being retired) are not taken into account. Here, the survival X (that is the age at death) is left truncated by Y that represents the age at the entry of the retirement home.

   - Right truncation: only the individuals for whom the event under concern has occurred are included in the study.

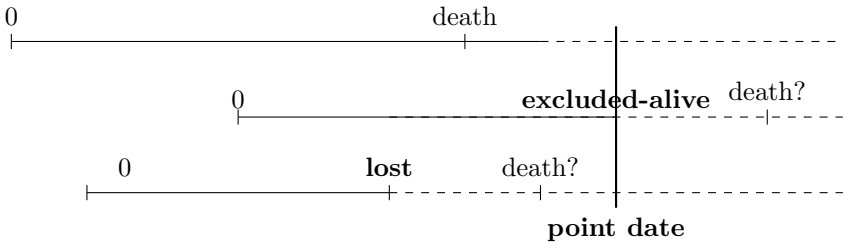
     Example: Considering AIDS, one is interested in the distribution of the delay to develop the disease after the inclusion of the HIV. In this study, only the patients that have developed the HIV take part to the study: the healthy individuals that carry HIV are not known by the clinician.

Warning: Beware of not making the confusion between truncation and censoring!

In medicine, a right censoring and a left truncation appear in the most famous models.

1.3.2 Some examples of censoring

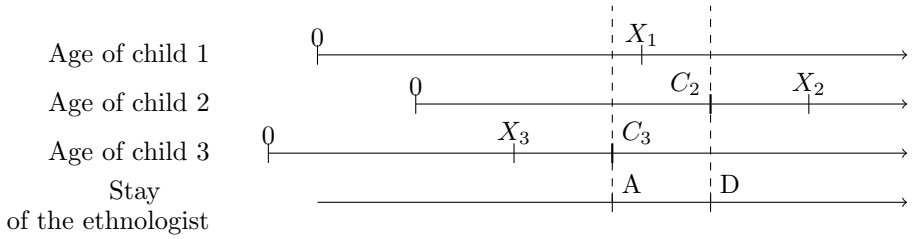
Example of right censoring:



We distinguish two different types of patients:

- the excluded-alive: patients still alive at the point date,
- the lost individuals: patients whose state is unknown at the point date.

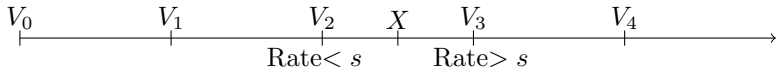
Example of left censoring: the ethnologist



- X_i is the age of child i when is able to do a given task,

- A and D are the arrival and departure times of the ethnologist,
- C_2: age of child 2 when the ethnologist leaved,
- C_3: age of child 3 at the arrival of the ethnologist.
- X_1 is not censored,
- X_2 is right censored by C_2: $X_2 > C_2$,
- X_3 is left censored by C_3: $X_3 < C_3$.

Examples of censoring by interval: we consider a group of patients that are controlled frequently during a clinical survey. Let X be the delay necessary for a biological variable to reach some level s. One only knows that X lives in an interval between two successive controls: here, $V_2 < X < V_3$.



The censoring by interval appears also in reliability if periodical inspections are lead to check the good functioning of the machines.

1.3.3 Right censoring

Since the right censoring is the most encountered in practice, we study it in details.

1. The **censoring of type I**: fixed

   The delay X_i of the i-th individual is observed if and only if $X_i \leqslant c$ where c is a **fixed duration**. Otherwise, one only knows that $X_i > c$.
   In other words, the censoring time c is known and fixed. One only observes the random variables $(T_i, \delta_i)_{i=1\cdots n}$ defined by

   $$\begin{cases} T_i = \min(X_i, c) \\ \delta_i = \mathbb{1}_{\{X_i \leqslant c\}} \end{cases}$$

   for $i = 1, \cdots, n$.

2. The **progressive censoring of type I**

   The censoring times c_i for $i = 1, \ldots, n$ are known and fixed. One only observes the random variables $(T_i, \delta_i)_{i=1\cdots n}$ defined by

   $$\begin{cases} T_i = \min(X_i, c_i) \\ \delta_i = \mathbb{1}_{\{X_i \leqslant c_i\}} \end{cases}$$

   for $i = 1, \cdots, n$.

   Example: all the patients enter the study the same day. There are censored (excluded-alive) at the point date.

3. The **censoring of type II** : waiting (commonly used in reliability)

   One considers the survival time X_i of n individuals until r events occur and then stops the study: the aim is to save time and money! The censoring occurs at $X_{(r)}$ the r-th order statistics. In other words, the censoring time is given by the time of the r-th failure observed in the sample. One only observes the random variables $(T_i, \delta_i)_{i=1\cdots n}$ defined by

   $$\begin{cases} T_i = \min(X_i, X_{(r)}) \\ \delta_i = \mathbb{1}_{\{X_i \leqslant X_{(r)}\}} \end{cases}$$

   for $i = 1, \cdots, n$ and $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(r)}$ where are the r first order statistics.

4. The **censoring of type III**: random

   To each patient, we associate a survival time X_i and a censoring delay C_i. There is a tradeoff between these two delays and one only observes the smaller denoted T_i:

   - the observed delay is $T_i = \min(X_i, C_i)$;
   - the event indicator is $\delta_i = \mathbb{1}_{X_i \leqslant C_i}$.

   As a consequence, the sample consists in pairs (T_i, δ_i), $i = 1, \cdots, n$ and one only observes the random variables $(T_i, \delta_i)_{i=1 \cdots n}$ defined by

   $$\begin{cases} T_i = \min(X_i, C_i) \\ \delta_i = \mathbb{1}_{\{X_i \leqslant C_i\}} \end{cases}$$

   for $i = 1, \cdots, n$ and where $C_1, \cdots, C_n$ are random variables with c.d.f. $G_1, \cdots, G_n$ respectively. The random variables $X_1, \cdots, X_n$ and $C_1, \cdots, C_n$ are generally assumed to be independent. In this case, the c.d.f. of the random variable T_i is given by $S_{T_i} = S_X S_{C_i}$ for $i = 1, \cdots, n$.

In the sequel, we consider a random censoring and we assume that X and C are two independent random variables. This assumption may be not satisfied in some practical examples: e.g. the treatment is stopped if some secondary effects appear or by tiredness of the failure in the studies on the infertility.

Vocabulary from medicine

We denote X the delay of occurrence of the event under concern. To define correctly this delay, one needs an origin and a date of end.

- The **origin date** (OD) corresponds to the original time of the individual. This is the principal source of the randomness of the censoring.

  Example: date of birth, date of first symptoms of a disease, date of the diagnosis by the doctor, date of surgery, date of inclusion in a clinical experiment.

- The **date of last news** (DLN) corresponds to the date of death if the patient died. Otherwise it corresponds to the latest date of inspection, control, monitoring...

- The **monitoring duration** is the delay between the origin and the date of latest news.

- The **point date** (PD) corresponds to the date (common to all individuals) when we decide to check the state of the patients.

  Remark: one does not take into account information posterior to the point date; otherwise a bias is introduced. Usually, one has recent news of non representative patients sub-groups (e.g. information on patients that relapse and no news of the ones that are in remission).

- The **participation time** corresponds to the "survival time" reported at the end of the analysis that is
  - if DDN $\leqslant$ PD: the time of participation is the monitoring duration.
    - if the patient is dead at the latest news, the time of participation is not censored.
    - if the patient is still alive at the latest news, he is lost and his time of participation is censored.
  - if DDN $>$ PD: the time of participation is censored and corresponds to the delay between the origin and the date point. Th patient is then considered as an excluded-alive.

- The **hindsight** of an individual is the delay between the origin and the date point.

1.4 Likelihood of a censored sample

1.4.1 Expression of the likelihood

Let X be a lifetime random variable whose density and survival functions are given by $f(t)$ and $S(t)$. Let C be a censoring whose density and cumulative distribution functions are given by $g(t)$ and $G(t)$. We assume that X and C are independent (we then deal with a model of censoring of type III).

As said before, the sample consists in n pairs $(T_i, \delta_i)_{1 \leqslant i \leqslant n}$ where $T_i = \min(X_i, C_i)$ corresponds to the observed delay and $\delta_i = \mathbb{1}_{X_i \leqslant C_i}$ is the event indicator for the patient i. The distribution of the pair (T_i, δ_i) is given by:

$$f(t, \delta) = [f(t)G(t)]^{\delta}[g(t)S(t)]^{1-\delta}.$$

The likelihood of the sample becomes

$$L_n((T_1, \delta_1), \cdots, (T_n, \delta_n)) = \prod_{i=1}^{n} f(T_i, \delta_i).$$

If the distribution of the censoring does not depend on the parameters of the survival, the censoring is non informative and the "useful" part of the likelihood is simply:

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

1.4.2 Estimation in parametric models

Let us consider a parametric model

$$\mathcal{F} = \{\mathbb{P}_\theta,\ \theta \in \Theta\} \quad \text{or} \quad \mathcal{F} = \{f_\theta,\ \theta \in \Theta\} \quad \text{or} \quad \mathcal{F} = \{F_\theta,\ \theta \in \Theta\}$$

for the lifetime X.

The question is how to estimate the (possibly multidimensional) parameter θ? Many different methods are available...

- Without censoring, you already know how to do with a complete sample.
- With censoring, it is more difficult. Under the assumption of independence between lifetime and censoring, the likelihood is given by

$$L_n((T_1, \delta_1), \cdots, (T_n, \delta_n); \theta) = \prod_{i=1}^{n} \left[f_\theta(T_i)(1 - G(T_i))\right]^{\delta_i} \left[S_\theta(T_i)g(T_i)\right]^{1-\delta_i},$$

  where g (respectively G) still represents the p.d.f. (resp. c.d.f.) of the censoring C.

  If the distribution of the censoring time does not depend on the parameter of interest θ, one can consider the "useful" likelihood given by

$$L_n((T_1, \delta_1), \cdots, (T_n, \delta_n); \theta) = \prod_{i=1}^{n} \left[f_\theta(T_i)\right]^{\delta_i} \left[S_\theta(T_i)\right]^{1-\delta_i}.$$

  The maximum likelihood estimator (MLE) is the value of θ that maximizes the likelihood.

1.5 The point process $N(t)$

Instead of studying the lifetime X, we may define the point process $N(t)$ that is equal to 0 while the event has yet occurred yet and is equal to 1 after:

$$N(t) = \mathbb{1}_{\{X \leqslant t\}}.$$

The process $N(t)$ jumps by 1 in $t = x$ when $X = x$. We denote $dN(t)$ the process variation on the interval $[t, t + dt[$:

$$\begin{aligned} &dN(t) = N((t+dt)^-) - N(t^-) \\ &\lim_{dt \to 0} \frac{\mathbb{P}(dN(t) = 1/N(t^-) = 0)}{dt} = h(t) \\ &\mathbb{P}(dN(t) = 1/N(t^-) = 1) = 0. \end{aligned}$$

Then one has

$$\lim_{dt \to 0} \frac{P(dN(t) = 1/N(t^-))}{dt} = h(t)\mathbb{1}_{X \geqslant t}.$$

We denote $\lambda(t) = h(t)\mathbb{1}_{X \geqslant t}$: λ is the intensity of the counting process N. The cumulative intensity is the function Λ defined by

$$\Lambda(t) = \int_0^t \lambda(u)du = \int_0^t \lambda(u)\mathbb{1}_{u \leqslant X} du = H(t \wedge X).$$

Example: For a lifetime X exponentially distributed with parameter $\theta > 0$,

$$\begin{aligned} &f_X(t) = \theta e^{-\theta t}\mathbb{1}_{t>0},\ S_X(t) = e^{-\theta t}, t > 0 \\ &h(t) = \theta,\ H(t) = \theta t, \\ &\lambda(t) = \theta \mathbb{1}_{X \geqslant t},\ \Lambda(t) = \theta(t \wedge X). \end{aligned}$$

The difference between $N(t)$ and $\Lambda(t)$ is a martingale: $M(t) = N(t) - \Lambda(t)$.

$$\mathbb{E}[dM(t)/\mathcal{F}_{t^-}] = \mathbb{E}[dN(t) - \lambda(t)dt/\mathcal{F}_{t^-}] = \mathbb{E}[dN(t)/\mathcal{F}_{t^-}] - \lambda(t)dt = 0$$

where $\mathcal{F}_{t^-} = \sigma(N(u), u < t)$.

Point processes for censored data:

In the previous example, we add a right censoring and one observes $T = \min(X, C)$ instead of X. Let $\delta = \mathbb{1}_{\{X \leqslant C\}}$. The indicator of "presence at risk of a subject" becomes:

$$Y(t) = \mathbb{1}_{\{t \leqslant X \wedge C\}}.$$

If there is n patients, we denote for $i = 1, \ldots, n$,

$$\begin{aligned} Y_i(t) &= \mathbb{1}_{T_i \geqslant t} \\ N_i(t) &= \mathbb{1}_{T_i \leqslant t, \delta_i = 1}. \end{aligned}$$

We denote $\bar{Y} = \sum_{i=1}^n Y_i(t)$ and $\bar{N}(t) = \sum_{i=1}^n N_i(t)$.
Moreover, if there is a left truncation (the subject is seen in the study only if $X > U$), we get

$$Y(t) = \mathbb{1}_{U \leqslant t \leqslant X \wedge C}.$$

1.6 Bibliography

Andersen, Per Kragh, Borgan, Ørnulf, Gill, Richard D. and Keiding, Niels. *Statistical models based on counting processes.* Springer Series in Statistics. Springer-Verlag, New York, 1993.

Cox, D. R.and Oakes, D. *Analysis of survival data.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1984.

Fleming, Thomas R. and Harrington, David P. *Counting processes and survival analysis.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1991.

Kalbfleisch, John D. and Prentice, Ross L. *The Statistical Analysis of Failure Time Data.* Second Edition Wiley Series in Probability and Statistics, 2011.

Klein, John P., Moeschberger, Melvin L. *Survival Analysis. Techniques for Censored and Truncated Data.* Springer Science & Business Media, 2003.

Therneau, Terry M. and Grambsch, Patricia M. *Modeling survival data: extending the Cox model.* Statistics for Biology and Health. Springer-Verlag, New York, 2000.

Chapter 2

Non parametric estimations

2.1 Without censoring

Assume that we observe a n sample $X_1,\ X_2,\ \cdots,\ X_n$ of the lifetime X. The empirical cumulative distribution function

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leqslant x}$$

is an estimator of the c.d.f. $F(\cdot)$. As a consequence, a natural estimator of $S(\cdot) = 1 - F(\cdot)$ is given through the empirical cumulative distribution function $\hat{F}_n$:

$$\hat{S}_n(x) = 1 - \hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i > x}.$$

It is equivalent to estimate the hazard rate $h(\cdot)$ by

$$\begin{aligned} &\hat{h}_n(X_{(i)}) = \frac{1}{n-i+1} \quad \text{for} \quad i = 1, \cdots, n \\ &\hat{h}_n(x) = 0 \qquad\qquad\qquad\quad \text{for all } x \text{ that is not an observation} \end{aligned}$$

where the sample has been ordered: $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$.

The cumulative hazard rate function can be estimated by

$$\hat{H}_n(x) = \frac{1}{n}\sum_{i:X_{(i)} \leqslant x} \frac{1}{n-i+1}.$$

Theorem 2.1.1 (Glivenko-Cantelli Theorem) *The empirical c.d.f. is consistent: as n goes to infinity*

$$\sup_{x \geqslant 0} \left| \hat{F}_n(x) - F(x) \right| \overset{a.s.}{\to} 0$$

Theorem 2.1.2 (Donsker Theorem) *The empirical cdf is weakly convergent:*

$$\sqrt{n}\left(\hat{F}_n(\cdot) - F(\cdot)\right) \overset{\mathcal{L}}{\to} B(\cdot)$$

where $B(\cdot)$ is a centered Gaussian process with covariance function

$$\mathrm{Cov}(B(s), B(t)) = F(s \wedge t) - F(t)F(s).$$

Remark 2.1.3 *$B(\cdot)$ is called the **Brownian bridge** and is also defined by $B(t) = W(t) - tW(1)$ for any $0 \leqslant t \leqslant 1$ and $W(\cdot)$ is a standard Brownian motion.*

Now one can deduce a confidence band or a pointwise confidence interval for $F(\cdot)$. For example,

$$\left[\hat{F}_n(x) - z_{1-\alpha/2}\sqrt{\frac{\hat{F}_n(x)\hat{S}_n(x)}{n}}, \hat{F}_n(x) + z_{1-\alpha/2}\sqrt{\frac{\hat{F}_n(x)\hat{S}_n(x)}{n}}\right]$$

is an asymptotic $(1-\alpha)$ confidence interval for $F(x)$ where z_α is the α-quantile of the standard Gaussian distribution.

2.2 Kaplan-Meier estimator of the survival function

With right censored observations, the estimation of S only based on the uncensored observations does not lead to an estimator that converges to S. More precisely, assume that we observe a sample of possibly right censored data (T_i, δ_i), $i = 1, \cdots, n$. Let $n_1 = \sum_{i=1}^n \delta_i$ be the number of uncensored data in the sample. One may want to use

$$\hat{S}_n^{(1)}(t) = \frac{1}{n_1}\sum_{i=1}^n \mathbb{1}_{T_i>t,\, \delta_i=1}$$

as an estimator of $S(\cdot)$. But this is not a good idea since one can show that

$$\hat{S}_n^{(1)}(t) \longrightarrow \int_t^{+\infty}(1-G(t))dF(t) \neq S(t)$$

except if $G(t) = 0$ for $x > t$ which means that there is no censoring.

2.2.1 Construction of the estimator

The estimator of Kaplan-Meier (1958) is based on the following statement: being alive after t amounts to being alive just before t and not dying at time t. Hence one has,

$$\begin{aligned} S(t) &= \mathbb{P}(X \geqslant t) \\ &= \mathbb{P}(X \geqslant t \mid X \geqslant t-1)\mathbb{P}(X \geqslant t-1) \\ &= \cdots \\ &= \mathbb{P}(X \geqslant t \mid X \geqslant t-1)\mathbb{P}(X \geqslant t-1 \mid X \geqslant t-2)\cdots\mathbb{P}(X \geqslant 1 \mid X \geqslant 0)\mathbb{P}(X \geqslant 0) \\ &= Q_t \times Q_{t-1}\cdots \times Q_1 \times 1 \end{aligned} \tag{2.1}$$

where $Q_j = \mathbb{P}(X \geqslant j \mid X \geqslant j-1)$ represents the survival probability at j conditioned on being alive just before j.

Then one estimates $S(t)$ by the product of the estimations $\hat{Q}_j$ of Q_j where $\hat{Q}_j$ is the observed proportion of patients that are still alive after the j-th day among those alive just before j.

- If m_j deaths have occurred at j, then

$$\hat{Q}_j = \frac{n_i - m_i}{n_i}.$$

- If at day j, no death occurred, then $\hat{Q}_j = 1$.

As a consequence, only the death times appear in the estimation of $S(\cdot)$ and the estimator of $S(\cdot)$ is constant between two death times.

Examples

1. We observe 5 patients dead at days 3, 4, 6, 6 and 7. We want to estimate the survival probability after the 6-th day. We get

$$\begin{aligned}
&\hat{Q}_1 = \hat{Q}_2 = 5/5 = 1 \\
&\hat{Q}_3 = 4/5 = 0.80 \qquad m_3 = 1 \quad n_3 = 5 \\
&\hat{Q}_4 = 3/4 = 0.75 \qquad m_4 = 1 \quad n_4 = 4 \\
&\hat{Q}_5 = 3/3 = 1 \\
&\hat{Q}_6 = 1/3 = 0.33 \qquad m_6 = 2 \quad n_6 = 3
\end{aligned}$$

   Then

$$\hat{S}_5(6) = \hat{Q}_6\hat{Q}_5\hat{Q}_4\hat{Q}_3\hat{Q}_2\hat{Q}_1 = \hat{Q}_6\hat{Q}_4\hat{Q}_3 = 4/5 \times 3/4 \times 1/3 = 1/5.$$

   In this example, the probability to survive after day 6 is 20%. Since the data are uncensored, this probability can be computed making the ratio of the persons that are still alive after the day 6 and the total number of people.

2. Assume now that a right censoring has occurred at time 4. The survival times are then 3, 4^+, 6, 6 and 7. One has

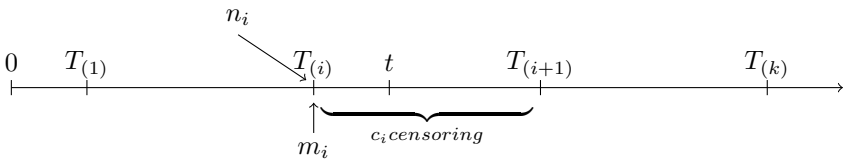
$$\begin{aligned}
&\hat{Q}_1 = \hat{Q}_2 = 5/5 = 1 \\
&\hat{Q}_3 = 4/5 = 0.80 \qquad m_3 = 1 \quad n_3 = 5 \\
&\hat{Q}_4 = 4/4 = 1 \\
&\hat{Q}_5 = 3/3 = 1 \\
&\hat{Q}_6 = 1/3 = 0.33 \qquad m_6 = 2 \quad n_6 = 3
\end{aligned}$$

   Then

$$\hat{S}_5(6) = \hat{Q}_6\hat{Q}_3 = 4/5 \times 1/3 = 0.27.$$

   Then the probability to survive after day 6 is 27%.

More generally, to estimate the survival function with a n-sample, one has to order the observations increasingly according to their participation time. Let $T_{(1)} \leqslant \cdots \leqslant T_{(n)}$ be the n ordered observed times and $\delta_{(1)}, \cdots, \delta_{(n)}$ the corresponding indicators. Denote n_i the number of subjects at risk at $T_{(i)}$ (not dead nor censored) and m_i the number of patients dead in $T_{(i)}$.



At the origin, $T_{(0)} = 0$, $m_0 = 0$ and c_0 is the number of censored observations between 0 and $T_{(1)}$. As a consequence, one has

$$n_i = n_{i-1} - m_{i-1} - c_{i-1}$$

and thus

$$n_i = n_0 - \sum_{j=1}^{i-1} m_j - \sum_{j=1}^{i-1} c_j$$

that means that the number of subjects at risk at $T_{(i)}$ is the number of subjects at the origin minus the number of subjects dead before $T_{(i)}$ minus the number of subjects censored before $T_{(i)}$. A natural estimator of the survival function is given by

$$\hat{S}_{n,KM}(t) = \prod_{i:T_{(i)} \leqslant t} \hat{Q}_{T_{(i)}} = \prod_{i:T_{(i)} \leqslant t} \frac{n_i - m_i}{n_i} = \prod_{i:T_{(i)} \leqslant t} \left(1 - \frac{m_i}{n_i}\right).$$

using the decomposition (2.1) in product of S with $i = T_{(i)}$.
For $t < T_{(1)}$, one has by convention $\hat{S}_{n,KM}(t) = 1$.

If there is no ex-æquo in the sample

One has $n_i = n - i + 1$ and $m_i = 1$ at each death time and one can estimate the hazard rate function $h(\cdot)$ by

$$\begin{aligned} \hat{h}_n(T_{(i)}) &= \frac{\delta_{(i)}}{n-i+1} \quad \text{for} \quad i = 1, \cdots, n \\ \hat{h}_n(t) &= 0 \qquad\qquad\qquad \text{otherwise.} \end{aligned}$$

The **Nelson-Aalen estimator** of the cumulative hazard rate function $H(\cdot)$ is then naturally given by

$$\hat{H}_{n,NA}(t) = \sum_{i:T_{(i)} \leqslant t} \frac{\delta_{(i)}}{n-i+1}.$$

We present in Section 2.4 the expression of the estimator in the general case.

The **Kaplan-Meier estimator** of the survival function $S(\cdot)$ is

$$\hat{S}_{n,KM}(t) = \prod_{i:T_{(i)} \leqslant t} \left(1 - \frac{\delta_{(i)}}{n-i+1}\right) = \prod_{i:T_{(i)} \leqslant t} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}}.$$

If there are some ex-æquo in the sample

Let $T'_1 \leqslant \cdots \leqslant T'_k$ be the k distinct and ordered observed times in the sample. The **Kaplan-Meier estimator** of the survival function $S(\cdot)$ is written in its general form as

$$\hat{S}_{n,KM}(t) = \prod_{i:T'_{(i)} \leqslant t} \left(1 - \frac{m_i}{n_i}\right).$$

If there is no censoring in the sample

If there is no censoring before $T_{(i)}$,

$$\begin{aligned} n_1 &= n_0 - m_0 = n_0, \\ n_j &= n_0 - \sum_{k=1}^{j-1} m_k \quad \text{for any} \quad 1 < j \leqslant i. \end{aligned}$$

Thus, for any $T_{(j)} < T_{(i)}$, one has

$$\begin{aligned} \hat{S}_{n,KM} &= \frac{n_0 - m_1}{n_0} \frac{n_0 - m_1 - m_2}{n_0 - m_1} \times \ldots \times \frac{n_0 - m_1 - m_2 - \ldots - m_j}{n_0 - m_1 - m_2 - \ldots - m_{j-1}} \\ &= \frac{n_0 - m_1 - m_2 - \ldots - m_j}{n_0}. \end{aligned}$$

As a consequence, without censoring, $\hat{S}_{n,KM}$ is the observed proportion of the subjects still alive at t that is simply the empirical survival function $\hat{S}_n = 1 - \hat{F}_n$.

Remark 2.2.1 *1. This estimator is also called limit-product estimator since it can be obtained as the limit of a product.*

2. *The estimator $\hat{S}_{n,KM}$ is a stepwise function with jumps at the observed death times. It reaches 0 only if the last observed delay $T_{(n)}$ of the sample corresponds to a real event (death or failure..) instead of a censoring i.e. if $\delta_{(n)} = 1$. When 0 is not reached, one observes a stage that traduces in general a few number of subjects that are controlled during a long period. That corresponds to a lack of information instead to a disappearance of the long-term risk.*

3. *Expression with the point processes:*

$$\hat{S}_{n,KM}(t) = \prod_{i:T_{(i)} \leqslant t} \left(1 - \frac{\Delta \bar{N}(T_{(i)})}{\bar{Y}(T_{(i)})}\right)$$

where

- $\bar{Y}(t) = \sum_{i=1}^{n} Y_i(t)$ *is the number of subjects at risk before* t,
- $\bar{N}(t) = \sum_{i=1}^{n} N_i(t)$ *is the number of observed deaths before* t,
- $\Delta \bar{N}(t)$ *corresponds to the number of observed deaths in* t.

Remark 2.2.2 *One can find another expression of the Kaplan-Meier estimator in the literature:*

$$\hat{S}_{n,KM}(t) = \prod_{i:T_{(i)} \leqslant t} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} \mathbb{1}_{\{t \leqslant T_{(n)}\}}.$$

Example of Freireich data: Freireich, in 1963, realized a therapeutic study in order to compare the remission durations (in weeks) of patients that suffer from leukemia. The patients are divided into two subgroups: some of them received a medicine (6 M-P) and the others a placebo. The results are the following:

6-MP	6, 6, 6, 6^+, 7, 9^+, 10, 10^+, 11^+, 13, 16, 17^+,
	19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+.
Placebo	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11,
	11, 12, 12, 15, 17, 22, 23.

The patients with a + sign correspond to lost subjects at the considered time of observation: they are censored, "excluded-alive" of the study and one only knows that their remission duration is greater than the observed delay.

2.2.2 Properties of the estimator

Proposition 2.2.3 (Optimality) *$\hat{S}_{n,KM}$ maximizes the non parametric likelihood.*

Theorem 2.2.4 (Consistency) *Under some assumptions (fulfilled in particular as soon as X is absolutely continuous) and if the censoring times are i.i.d. with common c.d.f. G, we have*

$$\sup_{t \geqslant 0} \left|\hat{F}_{n,KM}(t) - F(t)\right| \xrightarrow{a.s.} 0,$$
$$\sup_{t \geqslant 0} \left|\hat{S}_{n,KM}(t) - S(t)\right| \xrightarrow{a.s.} 0,$$

Theorem 2.2.5 (Asymptotic normality) *Under some assumptions (fulfilled in particular as soon as X is absolutely continuous) and if the censoring times are i.i.d. with common c.d.f. G, we have*

$$\sqrt{n}\left(\hat{F}_{n,KM}(\cdot) - F(\cdot)\right) \xrightarrow{\mathcal{L}} S(\cdot)B(W(\cdot)),$$
$$\sqrt{n}\left(\hat{S}_{n,KM}(\cdot) - S(\cdot)\right) \xrightarrow{\mathcal{L}} S(\cdot)B(W(\cdot)),$$

where $B(\cdot)$ is a Brownian motion on $\mathbb{R}^+$ and $W(\cdot)$ is the function defined by

$$W(t) = \int_0^t \frac{dF(u)}{S^2(u)(1 - G(u^-))}, \quad \textit{for all} \quad t > 0.$$

One can see that the process $Z_2(\cdot) := S(\cdot)B(W(\cdot))$ is a centered Gaussian process with covariance function

$$\mathrm{Cov}(Z_2(s), Z_2(t)) = S(s)S(t)\int_0^{s\wedge t} \frac{dF(u)}{S^2(u)(1-G(u^-))}.$$

2.2.3 Variance estimation and confidence intervals

Variance estimation

Let us denote the **Greenwood estimator** $\hat{\sigma}_n^2(t)$ by

$$\hat{\sigma}^2_{n,KM}(t) = \hat{S}_{n,KM}(t)^2 \sum_{i:T'_{(i)}\leqslant t} \frac{m_i}{n_i(n_i-m_i)}.$$

In particular, if there is no ex-æquo, this expression becomes

$$\hat{\sigma}^2_{n}(t) = \hat{S}_{n,KM}(t)^2 \sum_{i:T'_{(i)}\leqslant t} \frac{\delta_i}{(n-i)(n-i+1)}.$$

One can show that the Greenwood estimator $\hat{\sigma}_n^2(t)$ is an estimation of the variance of the Kaplan-Meier estimator and is a uniformly consistent estimator of the asymptotic variance function

$$S(t)^2\int_0^t \frac{dF(u)}{S^2(u)(1-G(u^-))}.$$

Without censoring, one recover the estimation of the variance of a proportion:

$$\frac{\hat{S}_{n,KM}(t)(1-\hat{S}_{n,KM}(t))}{n_0}.$$

In Section 2.3, we explain the origin of this estimator of the variance of $\hat{S}_{n,KM}(t)$.

Confidence interval for $S(t)$

Assuming the gaussianity of the asymptotic behavior of $\hat{S}_{n,KM}(t)$ an asymptotic $(1-\alpha)$-confidence interval for $S(t)$ is given by

$$\left[\hat{S}_{n,KM}(t) - z_{1-\alpha/2}\hat{\sigma}_{n,KM}(t), \hat{S}_{n,KM}(t) + z_{1-\alpha/2}\hat{\sigma}_{n,KM}(t)\right]$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Gaussian distribution.

This interval is the one given by the option `conf.type = plain` of the `survfit` function that is available in R.

As an estimator of the expectation $\mathbb{E}[X]$ one can use

$$\hat{\mu}_n = \int_0^{+\infty} t d\hat{F}_n(t) = \sum_{i=1}^k T'_{(i)}\Delta\hat{F}_n(T'_{(i)}) = \sum_{i=1}^k T'_{(i)}\frac{m_i}{n_i}\prod_{j=1}^{i-1}\left(1-\frac{m_i}{n_i}\right).$$

We have

$$\hat{\mu}_n \xrightarrow{a.s.} \int x dF(x) = \mathbb{E}[X], \quad \text{as} \quad n\to+\infty.$$

2.3 Breslow estimator of the cumulative hazard rate function

One can deduce from $\hat{S}_{n,KM}$ an estimator of the cumulative hazard rate function using the relation $H(t) = -\ln S(t)$:

$$\hat{H}_{n,BR}(t) = -\ln \hat{S}_{n,KM}(t) = -\sum_{i:T_{(i)} \leqslant t} \ln \hat{Q}_{T_{(i)}} = -\sum_{i:T_{(i)} \leqslant t} \ln \frac{n_i - m_i}{n_i}.$$

Theorem 2.3.1 (Consistency) *Under some assumptions (fulfilled in particular as soon as X is absolutely continuous) and if the censoring times are i.i.d. with common c.d.f. G, we have*

$$\sup_{t \geqslant 0} \left| \hat{H}_{n,BR}(t) - H(t) \right| \xrightarrow{a.s.} 0.$$

Theorem 2.3.2 (Asymptotic normality) *Under some assumptions (fulfilled in particular as soon as X is absolutely continuous) and if the censoring times are i.i.d. with common c.d.f. G, we have*

$$\sqrt{n} \left(\hat{H}_{n,BR}(\cdot) - H(\cdot) \right) \xrightarrow{\mathcal{L}} B(W(\cdot)),$$

where $B(\cdot)$ is a Brownian motion on $\mathbb{R}^+$ and $W(\cdot)$ is the function defined by

$$W(t) = \int_0^t \frac{dF(u)}{S^2(u)(1 - G(u^-))}, \quad \textit{for all} \quad t > 0.$$

Variance estimation

One has

$$\mathrm{Var}\left(\hat{H}_{n,BR}(t) \right) = \mathrm{Var}\left(\sum_{i:T_{(i)} \leqslant t} \ln \hat{Q}_{T_{(i)}} \right).$$

The random variables $\hat{Q}_{T_{(i)}}$ are not independent, but assuming it is the case, on gets the following variance approximation:

$$\mathrm{Var}\left(\hat{H}_{n,BR}(t) \right) \approx \sum_{i,T_{(i)} \leqslant t} \mathrm{Var}\left(\ln \hat{Q}_{T_{(i)}} \right).$$

Now we apply the delta method [1] to approximate $\mathrm{Var}\left(\ln \hat{Q}_{T_{(i)}} \right)$: for any regular function f,

$$\mathrm{Var}(f(X)) \approx f'(\mathbb{E}[X])^2 \mathrm{Var}(X)$$

and we use the fact that the random variables $n_i \hat{Q}_{T_{(i)}}$ are binomial with parameters n_i and $Q_{T_{(i)}}$. We get

$$\mathrm{Var}\left(\hat{H}_{n,BR}(t) \right) \approx \sum_{i,T_{(i)} \leqslant t} \frac{1 - \hat{Q}_{T_{(i)}}}{n_i \hat{Q}_{T_{(i)}}} \approx \sum_{i,T_{(i)} \leqslant t} \frac{m_i}{n_i(n_i - m_i)}.$$

Thus an approximation of the variance of $\hat{H}_{n,BR}(t)$ is given by

$$\hat{\sigma}^2_{n,BR}(t) = \sum_{i:T_{(i)} \leqslant t} \frac{m_i}{n_i(n_i - m_i)}.$$

Confidence intervals for $H(t)$ and $S(t)$

Assuming the gaussianity of the asymptotic behavior of $\hat{H}_{n,BR}(t)$, we deduce an asymptotic $(1 - \alpha)$-confidence interval for $H(t)$:

$$\left[\hat{H}_{n,BR}(t) - z_{1-\alpha/2} \hat{\sigma}_{n,BR}(t), \hat{H}_{n,BR}(t) + z_{1-\alpha/2} \hat{\sigma}_{n,BR}(t) \right],$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard Gaussian distribution.

Using the relation that links S to H: $S(t) = \exp\{-H(t)\}$, one deduces an asymptotic $(1-\alpha)$-confidence interval for $S(t)$:

$$\left[\exp\left\{-(\hat{H}_{n,BR}(t) + z_{1-\alpha/2}\hat{\sigma}_{n,BR}(t))\right\}, \exp\left\{-(\hat{H}_{n,BR}(t) - z_{1-\alpha/2}\hat{\sigma}_{n,BR}(t))\right\}\right]. \tag{2.2}$$

This interval is the one given by the option `conf.type = log` of the `survfit` function that is available in R.

Remark 2.3.3 (Estimation of the variance of $\hat{S}_{n,KM}$) *The Greenwood estimator $\hat{\sigma}^2_{n,KM}(t)$ of the variance of $\hat{S}_{n,KM}$ can be deduced from the previous variance. It has been obtained by the following considerations. We use once again the delta method [1] to $S(t) = \exp\{-\ln S(t)\}$ with $f(x) = e^x$ and $X = -\ln S(t)$ to get the following approximation*

$$\begin{aligned}
\mathrm{Var}\left(\hat{S}_{n,KM}(t)\right) &\approx S(t)^2\ \mathrm{Var}\left(\ln \hat{S}_{n,KM}(t)\right) \\
&\approx \hat{S}_{n,KM}(t)^2\ \mathrm{Var}\left(\ln \hat{S}_{n,KM}(t)\right) \\
&= \hat{S}_{n,KM}(t)^2\ \mathrm{Var}\left(\hat{H}_{n,BR}(t)\right) \\
&\approx \hat{S}_{n,KM}(t)^2 \sum_{i,T_{(i)}<t} \frac{m_i}{n_i(n_i-m_i)} \\
&\approx \hat{S}_{n,KM}(t)^2\ \hat{\sigma}^2_{n,BR}(t).
\end{aligned}$$

One deduces an asymptotic $(1-\alpha)$-confidence interval for $S(t)$:

$$\left[\hat{S}_{n,KM}(t)\left(1 - z_{1-\alpha/2}\hat{\sigma}_{n,BR}(t)\right), \hat{S}_{n,KM}(t)\left(1 + z_{1-\alpha/2}\hat{\sigma}_{n,BR}(t)\right)\right]$$

and since, $\hat{S}_{n,KM}(t) = \exp(-\hat{H}_{n,BR}(t))$, this interval corresponds also to:

$$\left[\exp\{-\hat{H}_{n,BR}(t)\}\left(1 - z_{1-\alpha/2}\hat{\sigma}_{n,BR}(t)\right), \exp\{-\hat{H}_{n,BR}(t)\}\left(1 + z_{1-\alpha/2}\hat{\sigma}_{n,BR}(t)\right)\right]. \tag{2.3}$$

If $\hat{\sigma}_{n,BR}(t)$ is close to 0, one gets an equivalent expression for both intervals in (2.2) and (2.3).

2.4 Nelson-Aalen estimator of the cumulative hazard function

For ds small, one as

$$H(s+ds) - H(s) \approx h(s)ds \approx \mathbb{P}(s < X < s+ds \mid X > s) \tag{2.4}$$

that is naturally estimated by $\frac{m_i}{n_i}$ that is the increasing of $H(t)$ in $T_{(i)}$ where m_i deaths are observed among n_i patients at risk.

Summing these quantities over the sub intervals of $(0,t]$ and making their length tends to 0 in such a way that they contains only at most one element, we derive the **Nelson-Aalen estimator** of the cumulative hazard function:

$$\hat{H}_{n,NA}(t) = \sum_{i:T_{(i)}\leqslant t} \frac{m_i}{n_i}.$$

By convention, $\hat{H}_{n,NA}(t) = 0$ for $t < T_{(1)}$.

Equivalent expression with the point processes

In terms of point processes, the difference (2.4) is naturally estimated by $\left(\bar{N}(s+ds) - \bar{N}(s)\right)/\bar{Y}(s)$. As a consequence

$$\hat{H}_{n,NA}(t) = \int_0^t \frac{d\bar{N}(s)}{\bar{Y}(s)}$$

that can be written in a discrete form:

$$\hat{H}_{n,NA}(t) = \sum_{i:T_{(i)} \leqslant t} \frac{d\bar{N}(T_{(i)})}{\bar{Y}(T_{(i)})}.$$

Property and interpretation:

1. $\hat{H}_{n,NA}$ is a stepwise function whose jumps are equal to $\frac{m_i}{n_i}$ at each observed death $T_{(i)}$.
2. $H(t)$ represents the mean number of deaths (or failures) on $(0, t]$ for an individual perpetually at risk: an electric component is working during t hours; at each failure, we replace the current component by another component that has worked the same duration than the one we are replacing (in reliability, this method is called **protocol of minimal reparation**).
3. The derivative of H represents the hazard rate function h. Since the estimator $\hat{H}_{n,NA}$ is a stepwise function, it is not possible to compute its derivative. As a consequence, to estimate h, as for the estimation of any density probability function, one needs to proceed to the smoothing of $\hat{H}_{n,NA}$.

Variance estimation

A consistent estimator of the variance of $\hat{H}_{n,NA}(t)$ is given by

$$\sigma^2_{n,NA}(t) = \sum_{i:T_{(i)} \leqslant t} \frac{d\bar{N}(t_i)}{\bar{Y}(t_i)^2} = \sum_{i:T_{(i)} \leqslant t} \frac{m_i}{n_i^2}.$$

Confidence interval for $H(t)$

Assuming the gaussianity of $\hat{H}_{n,NA}(t)$, one deduces an asymptotic $(1-\alpha)$-confidence interval for $H(t)$:

$$\left[\hat{H}_{n,NA}(t) - z_{1-\alpha/2}\sigma_{n,NA}(t), \hat{H}_{n,NA}(t) + z_{1-\alpha/2}\sigma_{n,NA}(t)\right].$$

2.5 Harrington and Fleming estimator of the survival function

Since $S(t) = \exp\{-H(t)\}$, we deduce from the Nelson-Aalen estimator of the cumulative hazard rate function the **Harrington and Fleming estimator** of the survival function:

$$\hat{S}_{n,HF}(t) = \exp\left\{-\hat{H}_{n,NA}(t)\right\}.$$

The relation $\hat{H}_{n,NA}(t) = \sum_{i:T_{(i)} \leqslant t} \frac{m_i}{n_i}$ yields

$$\hat{S}_{n,HF}(t) = \exp\left\{-\sum_{i:T_{(i)} \leqslant t} \frac{m_i}{n_i}\right\} = \prod_{i:T_{(i)} \leqslant t} \exp\left\{-\frac{m_i}{n_i}\right\}$$

that we should compare to the Kaplan-Meier estimation $\hat{S}_{n,KM}(t) = \prod_{i:T_{(i)} \leqslant t} \frac{n_i - m_i}{n_i}$ of the survival function.

Variance estimation

By the delta method [1], $\mathrm{Var}(\hat{S}_{n,HF}(t)) = \hat{S}_{n,HF}(t)^2 \mathrm{Var}(\hat{H}_{n,NA}(t))$ that leads to an estimator of the variance of $\hat{S}_{n,HF}$:

$$\sigma^2_{n,HF}(t) \exp\left\{-2\sum_{i:T_{(i)} \leqslant t} \frac{m_i}{n_i}\right\} \sum_{i:T_{(i)} \leqslant t} \frac{m_i}{n_i^2}.$$

Confidence interval for $S(t)$

Assuming the gaussianity of $\hat{S}_{n,HF}(t)$, one deduces an asymptotic $(1-\alpha)$-confidence interval for $S(t)$:

$$\left[\hat{S}_{n,HF}(t) - z_{1-\alpha/2}\sigma_{n,HF}(t), \hat{S}_{n,HF}(t) + z_{1-\alpha/2}\sigma_{n,HF}(t)\right].$$

2.6 Comments

We introduced two different estimators of the cumulative hazard rate function H:

1. the Breslow estimator deduced for the Kaplan-Meier estimation of the survival function,
2. the Nelson-Aalen estimator.

One may find in the literature a third estimation of H that comes from the actuarial method. These three different estimators of H are asymptotically equivalent and in practice one may prefer use the first one because of its relation with the Kaplan-Meier estimator.

Chapter 3

Comparison of the survival functions of two (or more) groups

3.1 The weighted logrank tests

These tests allows us to test the equality of K survival distributions through a sample of censored data. Here,

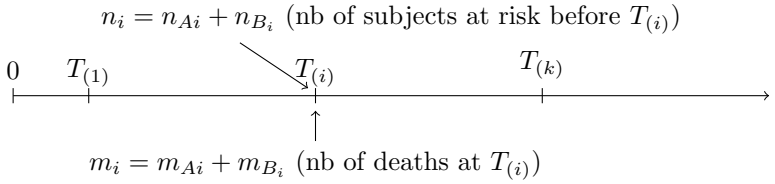
$$H_0 : S_1(t) = S_2(t) = \ldots = S_K(t) \quad \forall t \quad \text{versus}$$
$$H_1 : \text{at least two groups do not share the same survival function.}$$

3.1.1 Comparison of two groups

The goal is to test

$$H_0 : S_A(t) = S_B(t) \quad \forall t \quad \text{versus}$$
$$H_1 : \exists t \quad \text{such that} \quad S_A(t) \neq S_B(t).$$

Let $T_{(1)} < T_{(2)} < \ldots < T_{(k)}$ be the k ordered times of death of the sample that gathers the data of the two groups A and B.



where n_i represents the number of subjects at risk in $T_{(i)}$ and m_i the number of subjects that die in $T_{(i)}$.

For any time of death $T_{(i)}$, one can resume the information in the following 2×2 tabular:

	Dead	Alive after $T_{(i)}$	Total
Group A	m_{Ai}	$n_{Ai} - m_{Ai}$	n_{Ai}
Group B	m_{Bi}	$n_{Bi} - m_{Bi}$	n_{Bi}
Total	m_i	$n_i - m_i$	n_i

Under the null hypothesis H_0, in $T_{(i)}$, the proportion of dead among the subjects at risk is identical in both groups. Considering that the marginals of the tabular are fixed, under the null hypothesis H_0, m_{Ai} is distributed as an hypergeometric random variable whose parameters are (m_i, n_{Ai}, n_{Bi}). Then,

$$\mathbb{E}(m_{Ai}) = e_{Ai} = m_i \frac{n_{Ai}}{n_i},$$
$$\mathrm{Var}(m_{Ai}) = v_{Ai} = m_i \frac{n_i - m_i}{n_i - 1} \frac{n_{Ai} n_{Bi}}{n_i^2}.$$

The random variable $m_{Ai} - e_{Ai}$ is centered under H_0 and on can prove that m_{Ai} and m_{Aj} are uncorrelated as soon as $i \neq j$. Then we naturally define the statistics of the weighted logrank test by

$$U_A = \sum_{i=1}^{k} w_i \left(m_{Ai} - m_i \frac{n_{Ai}}{n_i} \right)$$

where the weight w_i depends on $T_{(i)}$. Obviously,

$$LR = \frac{U_A}{\sqrt{\mathrm{Var}(U_A)}} = \frac{\displaystyle\sum_{i=1}^{k} w_i \left(m_{Ai} - m_i \frac{n_{Ai}}{n_i} \right)}{\sqrt{\displaystyle\sum_{i=1}^{k} w_i^2 \left(m_i \frac{n_i - m_i}{n_i - 1} \frac{n_{Ai} n_{Bi}}{n_i^2} \right)}}$$

converges under H_0 to a standard Gaussian random variable and equivalently

$$\frac{U_A^2}{\mathrm{Var}(U_A)} \xrightarrow[H_0]{\mathcal{L}} \chi_1^2.$$

Remark 3.1.1 *1. The weighted logrank statistics are ordered statistics: they only depend on the ranks of the observations instead of their exact value.*

2. Breslow proposed a simplified form of the logrank statistics that comes from another way to take into account the ex-æquo:

$$LR_{BR} = \frac{U_A}{\sqrt{\mathrm{Var} U_A}} = \frac{\displaystyle\sum_{i=1}^{k} w_i \left(m_{Ai} - m_i \frac{n_{Ai}}{n_i} \right)}{\sqrt{\displaystyle\sum_{i=1}^{k} w_i^2 m_i \frac{n_{Ai} n_{Bi}}{n_i^2}}}.$$

Examples of weighting:

1. The **logrank test** (also called **Mantel-Haenszel test**): $w_i = 1 \ \forall i$.
   All the deaths have the same weights. It is the simplest weighting. The test amounts to compare the observed number O_A of deaths in the group A to yhe expected number E_A of deaths. Under H_0, we have

   $$O_A = \sum_{i=1}^{k} m_{Ai} \quad \text{and} \quad E_A = \sum_{i=1}^{k} e_{Ai}.$$

   The test statistics is then

   $$\frac{(O_A - E_A)^2}{\sum_{i=1}^{k} v_i}$$

2. The **Gehan test**: $w_i = n_i \ \forall i$ (available in the software `SAS` under the name Wilcoxon since it is a generalization of the Wilcoxon- Mann-Whitney test in the context of censoring).
   The first deaths have greater weights.

3. The **Peto and Prentice test** (close to the Kaplan-Meier estimator): $w_i = S_i \quad \forall i$ with $S_i^* = \prod_{j=1}^{i} \frac{n_j}{n_j + m_j}$ (available in `SAS`).

   The first deaths have greater weights.

4. The **Tarone and Ware test**: $w_i = \sqrt{n_i} \quad \forall i$.

5. The **Harrington and Fleming test**: $w_i = \hat{S}_{n,KM}(T_{(i)})^\rho \quad \forall i$ with $0 \leqslant \rho \leqslant 1$.
   The logrank test is a particular case of the Harrington and Fleming test when $\rho = 0$ and for $\rho = 1$, we get a test that is close to the one of Peto and Prentice.

These tests are available in `R` (function `survdiff`).

Approximated logrank test
The observed total number of deaths in both groups is equal, under H_0, to the expected total number of deaths:

$$O_A + O_B = E_A + E_B$$

and

$$\mathrm{Var}(O_A - E_A) = \mathrm{Var}(O_B - E_B).$$

The logrank statistics can be rewritten equivalently

$$\frac{(O_A - E_A)^2}{\mathrm{Var}(O_A - E_A)} \quad \text{or} \quad \frac{(O_B - E_B)^2}{\mathrm{Var}(O_B - E_B)}.$$

One can show that the statistics

$$(LRA)^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}$$

is always less than or equal to the logrank one.

1. This statistics is easier to compute than the previous one and is called the **approximated logrank statistics**.

2. It reminds the χ_2 statistics used to compare two observed proportions.

3. It is a conservative statistics: anytime it rejects H_0, the classical logrank statistics would have given the same conclusion. However the approximated logrank test is less powerful than the classical one.

Comparison criterion
In order to quantify the mortality difference between the groups and in analogy with the notion of relative risk, the quantity

$$RR_1 = \frac{O_B/E_B}{O_A/E_A}$$

has been proposed as an estimation of the ratio of the risk functions in each group. One might be cautious using this estimation since it is biased and the bias increases with the real ratio of the risk functions and the sample size.

One may compute this relative risk using the Gehan or Peto and Prentice weighting introducing

$$O_g = \sum_{i=1}^{k} w_i m_{gi} \quad \text{and} \quad E_g = \sum_{i=1}^{k} w_i e_{gi}.$$

the letter g stands for the group: A or B.

Other estimators have been introduced: for example,

$$RR_2 = \frac{\sum_{i=1}^{k} m_{Bi}(n_{Ai} - m_{Ai})/n_i}{\sum_{i=1}^{k} m_{Ai}(n_{Bi} - m_{Bi})/n_i} \quad \text{or} \quad RR_3 = \frac{\sum_{i=1}^{k} m_{Bi} n_{Ai}/n_i}{\sum_{i=1}^{k} m_{Ai} n_{Bi}/n_i}.$$

Weighting choice
As usual, we are looking for the most powerful test. As a consequence the choice of the weighting depends on the alternative hypothesis H_1. One can prove that, for a given alternative, the optimal asymptotic weights are proportional to $\ln\left(\frac{h_A(t)}{h_B(t)}\right)$.

- Logrank test ($w_i = 1$): optimal for the proportional alternatives

  $$H_0 : h_A(t) = h_B(t) \quad \forall t \quad \text{versus} \quad H_1 : h_B(t) = r h_A(t) \quad \forall t.$$

  These alternatives correspond to hypothesis of teh Cox model (we see in the chapter dedicated to the Cox model that the logrank test corresponds to the scoring test of the Cox model with two groups).

  An equivalent expression is given by: $S_B(t) = S_A(t)^r$.

- Peto and Prentice test: optimal for the alternatives such that $\ln\left(\frac{h_A(t)}{h_B(t)}\right) = S_A(t)$.

  The Gehan and Peto and Prentice tests are pretty well adapted when there are numerous premature deaths.

3.1.2 Generalization to the comparison of K groups

The goal is to test

$$H_0 : S_1(t) = S_2(t) = \ldots = S_K(t) \quad \forall t$$

versus

$$H_1 : \quad \text{at least two groups do not share the same survival distribution.}$$

We compute the statistics $U_1, U_2, \ldots, U_{K-1}$, the weighted sums at each observed death time of the difference between the observed number of deaths and the expected one under H_0.
The test statistic is given by

$$X = (U_1, U_2, \ldots, U_{K-1})\Sigma^{-1}(U_1, U_2, \ldots, U_{K-1})'$$

where Σ is the variance-covariance matrix of the vector $(U_1, U_2, \ldots, U_{K-1})$.
Under the null hypothesis H_0, X converges to a χ^2 distribution with $K-1$ degrees of freedom.

3.2 Comparison with adjustment: stratified logrank test

Assume that the survival is strongly linked to the age of the patient (which is quite natural), then it is more relevant to compare groups at a fixed given age.

In that view, we introduce a qualitative factor with M strata where each strata s corresponds to a sub sample.

One defines

$$U_s = \sum_{i=1}^{k_s} \left(m_{Asi} - m_{si}\frac{n_{Asi}}{n_{si}} \right);$$

then

$$\mathrm{Var}(U_s) = \sum_{i=1}^{k_s} \left(m_{si}\frac{n_{si} - m_i}{n_{si} - 1}\frac{n_{Asi} n_{Bsi}}{n_{si}^2} \right)$$

The stratified logrank test statistics is then defined by

$$U_{\text{stratified}} = \frac{\sum_{s=1}^{M} U_s}{\sqrt{\sum_{s=1}^{M} \text{Var}(U_s)}} \xrightarrow[H_0]{\mathcal{L}} \mathcal{N}(0,1).$$

Expression through the point processes
The corresponding notation are:

$$\begin{aligned}
&\bar{Y_A}(T_{(i)}) \equiv n_{Ai},\\
&\bar{Y_B}(T_{(i)}) \equiv n_{Bi},\\
&\bar{Y}(T_{(i)}) \equiv n_i,\\
&\Delta\bar{N_A}(T_{(i)}) \equiv m_{Ai},\\
&\Delta\bar{N_B}(T_{(i)}) \equiv m_{Bi},\\
&\Delta\bar{N}(T_{(i)}) \equiv m_i.
\end{aligned}$$

We rewrite the logrank statistics as the sum of the differences at each death time between the observed hazard rate in group A and the one expected under the null hypothesis H_0. More precisely, the observed hazard rate group A in $T_{(i)}$ is

$$\frac{\Delta\bar{N_A}(T_{(i)})}{\bar{Y_A}(T_{(i)})}$$

while the expected hazard rate under H_0 is

$$\frac{\Delta\bar{N}(T_{(i)})}{\bar{Y}(T_{(i)})}.$$

Then the weighted logrank test statistics is

$$U_A = \sum_{i=1}^{k} W_i(T_{(i)}) \left(\frac{\Delta\bar{N_A}(T_{(i)})}{\bar{Y_A}(T_{(i)})} - \frac{\Delta\bar{N}(T_{(i)})}{\bar{Y}(T_{(i)})} \right)$$

where $W_i(T_{(i)})$ is a non negative weighting function that has usually the following form $W_i(T_{(i)}) = \bar{Y_A}(T_{(i)})W(T_{(i)})$ with W a common weight to both groups.

One directly gets the following equivalent expression:

$$U_A = \sum_{i=1}^{k} W(T_{(i)}) \left(\Delta\bar{N_A}(T_{(i)}) - \bar{Y_A}(T_{(i)}) \frac{\Delta\bar{N}(T_{(i)})}{\bar{Y}(T_{(i)})} \right).$$

Chapter 4

Parametric regression models

The parametric regression models assume that the distribution of the lifetime random variable is a parametric function of time, possibly depending on one or more risk factors (covariates). Usually, it is the risk function $h(\cdot)$ that is assumed to parametric.

Suppose that we have a n sample of independent observations of the triplet (T, δ, Z) where

$$T = \min(X, C), \quad \delta = \mathbb{1}_{\{X \leqslant C\}} \quad \text{and} \quad Z = (Z_1, \ldots, Z_p)$$

is a vector of p covariates.

We consider the risk function conditioned on Z:

$$h(t \mid Z) = \lim_{dt \to 0} \frac{\mathbb{P}(t < X \leqslant t + dt \mid X > t, Z)}{dt}.$$

4.1 The exponential model

It is the simplest parametric model: the risk function is a constant with respect to the time. We want to compare two groups A and B with risk functions:

$$\left\{ \begin{array}{rcl} h_A(t) & \equiv & h_A \\ h_B(t) & \equiv & h_B. \end{array} \right\}$$

Let e^β the ratio between the risk functions:

$$h_B = h_A e^\beta,$$

which is called the **relative risk of group B with respect to group A**.

To compare the survival distributions, one has to estimate the parameter β and test the null hypothesis $\{\beta = 0\}$. One proceeds computing the likelihood of the observations.

The exponential model is generalized taking into account one or more risk factors. Let $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})'$ the vector of the p covariates of subject i. The generalization is done considering the risk function of the subject i

$$h(t \mid Z_i) = h_0 e^{\beta' Z_i}$$

where h_0 is constant and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the vector of parameters quantifying the effect of each covariate on the survival.

The particular case of the comparison of two groups corresponds to the introduction of one binary covariate Z taking the value 0 (respectively 1) when the individual belongs to group A (resp. B).

4.2 The Weibull model

It generalizes the exponential model. The risk function of the Weibull distribution depends on two parameters h_0 and α and writes as mentioned in Section 1.2.2:

$$h(t) = \alpha h_0^\alpha t^{\alpha-1}.$$

where h_0 is the scale parameter and α is the shape parameter.

The Weibull model assumes that the risk function of an individual whose covariate is Z_i writes

$$h(t \mid Z_i) = \alpha h_0^\alpha t^{\alpha-1} e^{\beta' Z_i}$$

where β is, as in the case of the exponential model, the vector of unknown parameters quantifying the effect of each covariate on the survival.

4.3 The semi parametric Cox model

The Cox model allows a parametric relation between the risk function and the risk factors (covariates) without expliciting nor precising the form of the survival distribution. As a consequence, it is a semi parametric model. This model is the most popular to model the relation between a lifetime random variable and covariates.

The Cox model modellizes the risk function of individual i whose covariate is $Z_i = (Z_{i1}, \ldots, Z_{ip})$ by

$$h(t \mid Z_i) = h_0(t) e^{\beta' Z_i}$$

where β is the vector of the p unknown parameters and h_0 is some given risk function.

Remark 4.3.1 *1. This model generalizes the previous parametric models:*

- *exponential where the reference risk function is constant: $h_0(t) \equiv h_0$;*
- *and Weibull where the reference risk function is polynomial: $h_0(t) \equiv \alpha h_0^\alpha t^{\alpha-1}$.*

2. The function exp may be changed into any positive function $g(\beta' Z_i)$.

Property In the Cox model, the ratio of the risk functions of two individuals i and i' is a constant with respect to the time:

$$\frac{h(t \mid Z_i)}{h(t \mid Z_{i'})} = e^{\beta'(Z_i - Z_{i'})}.$$

As a consequence, this model is said to be **of proportional risk functions** and is often named " model of proportional risks".

Interpretation of the model parameters:

$$\frac{h(t \mid Z = Z_i)}{h(t \mid Z = Z_{i'})} = e^{\beta'(Z_i - Z_{i'})}.$$

Example 4.3.1 *Comparison of two groups A and B*

- *If $Z = 0$ for group A and $Z = 1$ for group B, as assumed in the previous section, one gets*

$$\frac{h(t \mid Z = 1)}{h(t \mid Z = 0)} = e^{\beta}.$$

*The constant $r = e^\beta$ is named the **relative risk (RR) of group B with respect to group A.***

(i) if $\beta > 0$, $RR > 1$: the risk of death is higher in group B;

(ii) if $\beta = 0$, $RR = 1$: the risks of death are equal in both groups;

(iii) if $\beta < 0$, $RR < 1$: the risk of death is smaller in group B.

- *If Z is more general, one has*
$$e^{\beta} = \frac{h(t \mid Z+1)}{h(t \mid Z)}.$$
$RR = e^{\beta}$ quantifies the ratio between the risks of two individuals whose covariates differs from one unit.

For example, if Z represents the age, if $\beta > 0$, the risk of death increases with the age and $RR = e^{\beta}$ is the ratio between the risks of an individual and a one year younger individual (that is constant whatever their ages!).

More generally, in a model with several covariates, β_j measures the effect of the covariate Z_j on the increasing of the risk considering the other covariates fixed.

Remark 4.3.2 *Considering the survival function, it gives*
$$S(t \mid Z_i) = S_0(t)^{\theta}$$
where $\theta = e^{\beta' Z_i}$.

Now we want to estimate the parameters β_j that represent the effects on each covariate Z_j on the survival and realize significance tests on these parameters with no assumption on the distribution of the lifetime variable X. In this setting, h_0 will be considered as a nuisance parameter. The way to proceed in such a context has been introduced by Cox (1972) and leads to consider the **partial likelihood**.

4.3.1 The Cox partial likelihood

Let $T_{(1)} < T_{(2)} < \ldots < T_{(k)}$ be the k ordered observed times of death. We denote $\mathcal{R}_i$ the set of individuals at risk at time $T_{(i)}$.

The Cox partial likelihood writes as a product of conditional likelihoods computed at each time $T_{(i)}$, considered as fixed.

The contribution $V_i(\beta)$ depending on β to the likelihood of the individual i whose covariate is Z_i and observed death time $T_{(i)}$ is equal to the probability, conditioned on $\mathcal{R}_i$, that this individual dies precisely at $T_{(i)}$ among those at risk at $T_{(i)}$ i.e. among those in $\mathcal{R}_i$.

As a consequence, one has
$$V_i(\beta) = \frac{h(t_i \mid Z_i)}{\sum_{j \in \mathcal{R}_i} h(t_i \mid Z_j)}.$$
Under the Cox model, the nuisance term $h_0(t_i)$ (the risk function of reference) cancels and the expression of V_i reduces to
$$V_i(\beta) = \frac{e^{\beta' Z_i}}{\sum_{j \in \mathcal{R}_i} e^{\beta' Z_j}}.$$
The Cox partial likelihood is then the product of the contributions of the dead individuals:
$$V(\beta) = \prod_{i=1}^{k} V_i(\beta)$$
and the partial log-likelihood $L = \ln V$ is
$$L(\beta) = \sum_{i=1}^{k} \left(\beta' Z_i - \ln \left(\sum_{j \in \mathcal{R}_i} e^{\beta' Z_j} \right) \right).$$

Presence of ex æquo

The model considered is a continuous one and thus, theoretically, there is no ex æquo. Nevertheless, in practice, if the time discretization is coarse (e.g. if the survival is measured in months), it may lead to ex æquo death times. In this case, there exists several ways to proceed and to take into account the ex æquo in the likelihood.

Consider the following example. Four individuals die at $T_{(1)} = T_{(2)} < T_{(3)} < T_{(4)}$. We write r_i for $e^{\beta' Z_i}$ to lighten notation. If the data have been more precise (i.e. if the discretization have been fine enough to prevent the possibility of ex æquo), then the contributions of the individuals 1 and 2 on the likelihood would have been:

$$\text{either } \left(\frac{r_1}{r_1+r_2+r_3+r_4}\right)\left(\frac{r_2}{r_2+r_3+r_4}\right) \quad 1 \text{ dead before } 2$$
$$\text{or } \left(\frac{r_2}{r_1+r_2+r_3+r_4}\right)\left(\frac{r_1}{r_1+r_3+r_4}\right) \quad 2 \text{ dead before } 1.$$

Since the real order of the deaths of the two first individuals is unknown, we should consider the mean of these terms or at least an approximation of the mean.

1. **Breslow likelihood:** Breslow proposes to use the complete sum $r_1+r_2+r_3+r_4$ in both denominators
$$\left(\frac{r_1}{r_1+r_2+r_3+r_4}\right)\left(\frac{r_2}{r_1+r_2+r_3+r_4}\right)$$
which leads to the following partial log-likelihood:
$$L(\beta) = \sum_{i=1}^{k}\left(\beta' \sum_{j=1}^{m_i} Z_j - m_i \ln\left(\sum_{j\in\mathcal{R}_i} e^{\beta' Z_j}\right)\right),$$
where m_i is the number of deaths in $T_{(i)}$.

   Problem: the dead individuals are counted twice in the denominator and a bias appears that leads to the under estimation of the parameter β.

2. **Efron likelihood**: Efron suggests to use the mean of the r_i's of the ex æquo individuals in the second denominator
$$\left(\frac{r_1}{r_1+r_2+r_3+r_4}\right)\left(\frac{r_2}{\frac{1}{2}r_1+\frac{1}{2}r_2+r_3+r_4}\right).$$
   Naturally, if the individuals 1,2 and 3 are ex æquo, we get
$$\left(\frac{r_1}{r_1+r_2+r_3+r_4}\right)\left(\frac{r_2}{\frac{2}{3}r_1+\frac{2}{3}r_2+\frac{2}{3}r_3+r_4}\right)\left(\frac{r_3}{\frac{1}{3}r_1+\frac{1}{3}r_2+\frac{1}{3}r_3+r_4}\right)$$
   .
   The Efron method is the one used in the function `coxph` implemented in the software `R`.

4.3.2 Estimation of the parameters of the model

The estimation $\hat{\beta}$ of β comes from the maximization of the partial likelihood. One can prove that $\hat{\beta}$ shares the same asymptotic properties with the classical MLE: it is consistent and asymptotically normal.

Assume that $m_i = 1\ \forall i = 1,\dots,k$, differentiating L with respect to the components of β, the score vector of length p is

$$U(\beta) := \left(\frac{d}{d\beta_1}L(\beta),\dots,\frac{d}{d\beta_p}L(\beta)\right)' = \sum_{i=1}^{k}\left(Z_i - \frac{\sum_{j\in\mathcal{R}_i} Z_j e^{\beta' Z_j}}{\sum_{j\in\mathcal{R}_i} e^{\beta' Z_j}}\right).$$

The estimator $\hat{\beta}$ is then obtained solving the system $U(\beta) = 0$ of p equations.

4.3.3 Significance tests

Let $I(\beta)$ the information matrix: $I(\beta) = -\mathbb{E}\left[\frac{\partial^2 L(\beta)}{\partial \beta^2}\right]$. To test the null hypothesis H_0: $\beta = 0$, the usual statistics considered are

- for the scoring test: $U'(0)I^{-1}(0)U(0)$,
- for the Wald test: $\hat{\beta}' I(\hat{\beta})\hat{\beta}$,
- for the likelihood ratio test: $-2\left(L(0) - L(\hat{\beta})\right)$.

These statistics are asymptotically equivalent. Moreover, under H_0, they are distributed as a χ^2 random variable with p degrees of freedom.

4.3.4 Estimation of the cumulative risk H_0 associated to h_0

The partial likelihood does not depend on the risk function of reference h_0; thus it is not estimated by solving the system of the p equations (the derivatives of the likelihood). Breslow proposes an estimator of the cumulative risk H_0 of reference that generalizes the Nelson-Aalen one used for homogeneous samples

$$\hat{H}_{NA}(t) = \int_0^t \frac{d\bar{N}(s)}{\bar{Y}(s)} \quad \text{that is} \quad \hat{H}_{NA}(t) = \sum_{i:t_i \leqslant t} \frac{m_i}{n_i}.$$

Now let $\hat{\beta}$ be the maximum partial likelihood estimator (MPLE) of β. The Breslow estimator writes

$$\hat{H}_0(t, \hat{\beta}) = \int_0^t \frac{d\bar{N}(s)}{\sum_{i=1}^n Y_i(s) e^{\hat{\beta}' Z_i}}$$

with a similar estimation of the variance than the one used in the Nelson-Aalen estimator:

$$\widehat{Var(\hat{H}_0(t, \hat{\beta}))} = \int_0^t \frac{d\bar{N}(s)}{[\sum_{i=1}^n Y_i(s) e^{\hat{\beta}' Z_i}]^2}.$$

Equivalent expression without the point processes

$$\hat{H}_0(t, \hat{\beta}) = \sum_{i:t_i \leqslant t} \frac{m_i}{\sum_{j \in R_i} e^{\hat{\beta}' Z_j}}$$

Estimation of the cumulative risk

The estimator of the cumulative risk for an individual whose covariate is $\tilde{Z}$ is naturally given by:

$$\hat{H}(t, \hat{\beta}, \tilde{Z}) = \hat{H}_0(t, \hat{\beta}) e^{\hat{\beta}' \tilde{Z}}$$

from which we derive a semi parametric estimation of the survival function

$$\hat{S}(t, \hat{\beta}, \tilde{Z}) = \exp\left(-\hat{H}(t, \hat{\beta}, \tilde{Z})\right).$$

One may also construct a consistent estimation of the variance of $\hat{H}(t, \hat{\beta}, \tilde{Z})$ (see Andersen *et al.*, 1993).

4.4 Validation of the Cox model

To check the adequation of a set of survival data to the Cox model, four aspects may be examined.

1. For each covariate, one may try to find the best functional form that explains its influence on the survival function $(Z, Z^2, \ln(Z), \ldots)$.
2. For each covariate, one may check whether the assumption of proportional risks holds.

3. One may quantify the efficiency of the model to predict the time interval until the occurence of an event of interest for a given individual.

4. One may identify the **outliers** and their influence on the estimation of the model parameters.

The procedures of model validation are often based on the study of the **residuals**: the values computed for each individual whose behavior is known (at least approximately) when the fitness to the model is correct.

Four principal types of residuals are defined in the Cox model setting:

1. the martingale residuals: to choose the best functional form of the covariates;

2. the deviation residuals: to detect the outliers;

3. the scoring residuals: to measure the influence of the individuals on the estimation of the model parameters (one may use `dfbeta` in `R`);

4. the Schoenfeld residuals: useful to test the hypothesis of proportional risks (function `cox.zph` in `R`).

The Cox-Snell residuals (less used) are not specific to the Cox model. They only allow to appreciate the global adjustment to the model but do not give information on the deviation type.

Proposition 4.4.1 *If X is a random variable with cumulative risk function H, then the random variable $H(X)$ is exponentially distributed with parameter 1.*

Proof *If F (respectively S) represents the distribution function (resp. the survival function) of X, one has*

$$\begin{aligned} F_{H(X)}(t) &= \mathbb{P}(H(X) \leqslant t) = \mathbb{P}(X \leqslant H^{-1}(t)) = F(H^{-1}(t)) \\ &= 1 - S(H^{-1}(t)) = 1 - e^{H(H^{-1}(t))} = 1 - e^{-t} \end{aligned}$$

which is the required result.

If the model adjustment is correct, then $\hat{H}(X_i)$ is a realization of a standard exponential distributed random variable.

4.4.1 The Cox-Snell residuals

The **Cox-Snell residual** of the individual i is given by

$$r_i^{CS} = \hat{H}_0(T_{(i)}) e^{\hat{\beta}' Z_i}$$

where $\hat{H}_0(T_{(i)})$ is the Breslow estimation of the cumulative risk H_0 of reference at time $T_{(i)}$.

If the r_i^{CS} are distributed as a sample of an $\mathcal{E}(1)$ random variable, then the cumulative risk function should be close to $H(t) = t$. To check if it is the case:

- we compute the estimation $\hat{H}_{n,NA}$ of the cumulative risk of the r_i^{CS} and we plot each $\hat{H}_{NA}(r_i^{CS})$,
- if the model adjustment is correct, $\hat{H}_{n,NA}$ is close to the first bisector.

Remark 4.4.2 *The standard exponential distribution $\mathcal{E}(1)$ is the reference with the true values of β and H_0. When we replace them by their estimations to compute the residuals, errors with respect of the law $\mathcal{E}(1)$ can be observed due to the uncertainty in the estimation of β and H_0. This fact is particularly verified when the sample size is small.*

4.4.2 The martingale residuals

The **martingale residuals** help the statistician to choose an adapted functional form to use for each covariate to correctly explain the survival random variable. These residuals come from the individual martingales

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta' Z_i(s)} h_0(s) ds.$$

One may interpret $M_i(t)$ at any time t as the difference on $(0, t]$ between the number of observed events for the individual i and the number of expected events under the assumption of a Cox model.

Remark 4.4.3 *The classical decomposition*

$$\textbf{data} = \textbf{model} + \textbf{noise}$$

has an analog: the so-called Doob decomposition

$$\textbf{counting process} = \textbf{compensator} + \textbf{martingale}.$$

Once the model has been fitted to the data, we have the following decompositions

$$\textbf{data} = \textbf{adjusted model} + \textbf{residuals}$$

and the analog with the point processes is:

$$\textbf{counting process} = \textbf{estimated compensator} + \textbf{martingale residual}.$$

Replacing β (respectively H_0) by $\hat{\beta}$ (resp. $\hat{H}_0$), one obtains the residual process:

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\hat{\beta}' Z_i(s)} d\hat{H}_0(s).$$

Now we define the martingale residual by

$$\hat{M}_i = \hat{M}_i(\infty) = N_i(\infty) - \int_0^\infty Y_i(s) e^{\hat{\beta}' Z_i(s)} d\hat{H}_0(s).$$

When the covariates do not depend on time, we get

$$\hat{M}_i = \delta_i - e^{\hat{\beta}' Z_i} \hat{H}_0(T_i) = \delta_i - r_i^{CS}.$$

Interpretation: the martingale residual is the difference between the number of observed events and the number of expected events under the adjusted model.

Properties 4.4.4 *One can show the following properties that are analog to the ones obtained in the linear model:*

1. $\mathbb{E}[M_i] = 0$*: the expected value of the residual is zero with the true parameter* β*;*
2. $\mathbb{E}[\hat{M}_i] \to 0$ *when* $n \to \infty$*;*
3. $\sum_{i=1}^n \hat{M}_i = 0$*,*
4. $\mathrm{Cov}(M_i, M_j) = 0$*;*
5. $\mathrm{Cov}(\hat{M}_i, \hat{M}_j) \leqslant 0$ *but this covariance goes to 0 as* $n \to \infty$*.*

To determine the functional form of a covariate in the Cox model:

- we adjust a Cox model with no covariates and we compute the martingale residuals;
- we represent them with respect to each covariate separately and we fit a smoothing function;
- if the correct model for Z_j is $\exp\{\beta_j f(Z_j)\}$, then the obtained smoothing function of the residuals with respect to Z_j has the form of f.

It comes from $\mathbb{E}[\hat{M}_i \mid Z_j] = c f(Z_j)$ where c is independent of Z_j and depends on the censoring rate.

Bibliography

[1] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.