ESA MASTER
Econométrie et Statistique Appliquée
Université d'Orléans

Modèles de Durée

Gilles de Truchis

Master 2 ESA

September 10, 2025

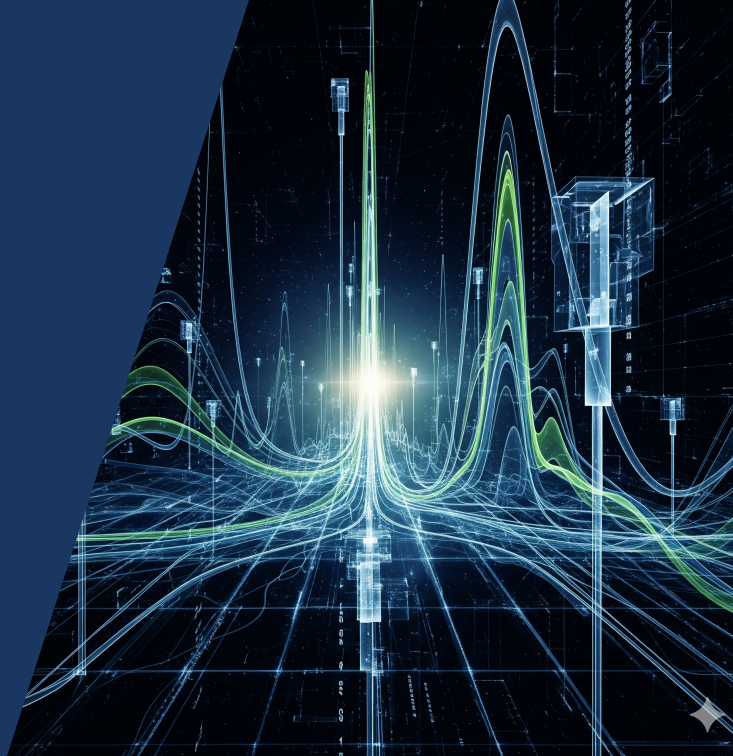


Table of contents

Chapter 1

Chapitre 1 : Survival Data

Survival Analysis

- Study of survival times of a particular phenomenon...

... and the factor that influence them

- Data with survival outcomes are numerous
  - ⇒ Clinical trials
  - ⇒ Biomedical studies
  - ⇒ Industrial settings (failure of a device)
  - ⇒ Labor market
  - ⇒ **Credit default**
- Statistical analysis of survival data requires
  - ⇒ Estimation of survival distribution
  - ⇒ Comparisons of various survival distributions
  - ⇒ Elucidations of the factors that influence survival times (regressions)

Survival Data

- The variable of interest has key characteristics
  - $\Rightarrow$ Non-negative discrete (or continuous) random variable
  - $\Rightarrow$ Represents the time from a well-defined origin to a well-defined event
  - $\Rightarrow$ Often subject to censoring : the starting or ending event is not observed
- Example of right censoring
  - — Let T^* be a random variable representing the time to failure
  - — Let U be a random variable representing the time to censoring event
  - — The recorded event will be $T = \min(T^*, U)$ and we can define
    $$\delta = I(T^* < U)$$
    a censoring indicator taking value 1 or 0
  - $\Rightarrow$ $\delta = 1$ if T is an observed failure time and $\delta = 0$ if T is a censored time

Note 1 Left censoring are possible albeit less frequent

Note 2 Interval censoring are also possible : the failure time has occurred within an unobserved time interval

Censoring classification

- There are 3 types of censoring times :

Type I Pre-specified censored times

e.g. In a study with a pre-specified ending time, if an individual has not experienced the event of interest before the end, it is censored at that time

Type II Pre-specified fraction of failure

e.g. If the study runs until a pre-specified fraction of failure is reached (e.g. 25 %), individuals or objects that have not failed (75%) are censored

Random Censoring that occurs randomly and independently of the study

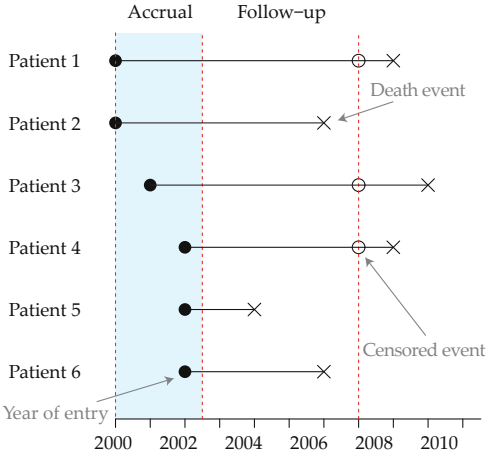
e.g. In a biomedical study, patient dropout that are unrelated to the disease process (e.g. death unrelated to the disease under investigation)

Note The random nature of this type of censoring is crucial to avoid bias

Type I censored data

- In biomedical studies, administrative censoring is of type I

$\Rightarrow$ It occurs when patients are still alive at the end of the follow-up period



Patient time structure

- Survival database are generally structured as follows

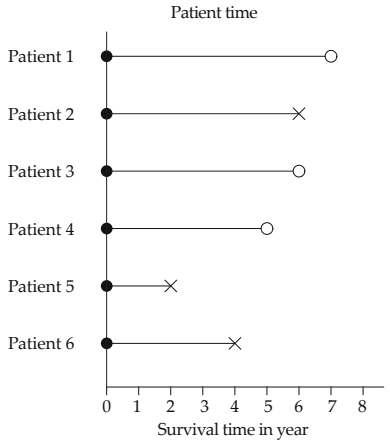
$\Rightarrow$ For each individual, the survival time and δ ("Status") are reported

Table: Survival data example

Patient	Survtime	Status
1	7	0
2	6	1
3	6	0
4	5	0
5	2	1
6	4	1

Patient time representation

- The patient time graphical representation is as follows



Database example (1)

- Additional informations can include additional outcomes
  - — individual characteristics
  - — competing risks factors
  - $\Rightarrow$ Below, $\delta \in \{0, 1, 2\}$ where 2 to indicate death from other causes

Table: Survival prospects of prostate cancer patients with high-risk disease

Patient	grade	stage	ageGroup	survTime	status
88	poor	T2	75-79	33	0
89	mode	T2	75-79	6	0
90	mode	T1c	75-79	15	2
91	mode	T2	70-74	6	2
92	mode	T1ab	80+	93	1
93	poor	T2	80+	60	2
94	mode	T2	80+	1	0
95	mode	T1ab	75-79	34	0

Database example (2)

- Comparisons survival data is also of crucial interest

  e.g. triple-medication v.s. nicotine patch therapy alone

Note 1 δ is set to 0 for individuals who remained non-smokers for 6 months

Note 2 Below, the variable ttr is time until return to smoking

$\Rightarrow$ The objective is to compare the two treatment therapies by identifying the factors related to this outcome

Table: Comparison of medical therapies to aid smokers to quit

	ttr	relapse	grp	age	gender	morphotype	employment
1	182	0	patchOnly	36	Male	white	ft
2	14	1	patchOnly	41	Male	white	other
3	5	1	combination	25	Female	white	other
4	16	1	combination	54	Male	white	ft
5	0	1	combination	45	Male	white	other
6	182	0	combination	43	Male	hispanic	ft

Hazard and Survival Functions

- Survival Analysis relies on the survival distribution that is specified by

either the Survival Function (SF)

or the Hazard Function (HF)

- The SF is defined as the probability of surviving up to a point t

$$S(t) = \mathbb{P}(T > t), \quad 0 < t < \infty$$

$\Rightarrow$ $S(t)$ is right continuous, equals 1 at time 0 and decreases over time

Note In some cases, $S(t)$ can also remain constant and never reach 0

- The HF is defined as the instantaneous failure rate

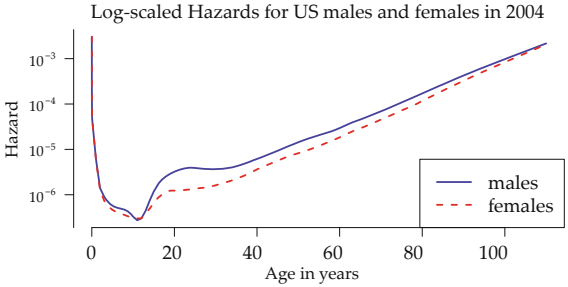
$$h(t) = \lim_{\Delta \to 0} \frac{\mathbb{P}(t < T < t + \Delta | T > t)}{\Delta}$$

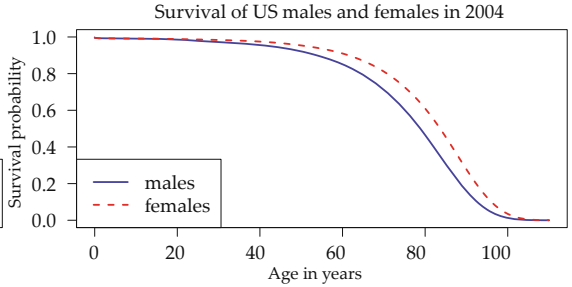
$\Rightarrow$ $h(t)$ is the probability of failing in the next interval of time Δ, given that the subject has survived up to time t, divided by that interval

Hazard and Survival Functions representation

Data: the daily hazard rates of men and women by age from 1940 to 2004

- The initial days and weeks of life are particularly dangerous
- The hazard increases during the teen years, then levels off
- It starts a steady increase in midlife





Other representations of the Survival Distribution

- The complement of the SF is just the so-called CDF

$$F(t) = \mathbb{P}(T \leq t), \quad 0 < t < \infty$$

$\Rightarrow$ known as cumulative risk function in the survival analysis

- The PDF is also an obvious alternative representation

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

$\Rightarrow$ it is the rate of change of $F(t)$ or minus the rate of change of $S(t)$

- $f(t)$ is also related to $h(t)$ by

$$h(t) = \frac{f(t)}{S(t)}$$

$\Rightarrow$ the hazard at time t is the probability that an event occurs in the neighborhood of t divided by the probability that the subject is alive at t

The Survival Function as function of the Hazard Function

- The area under the HF up to time t is the cumulative HF

$$H(t) = \int_0^t h(u)\,du$$

- Then, one can define the survival function in terms of the CHF

$$S(t) = \exp\left(- \int_0^t h(u)\,du \right) = \exp(-HF)$$

Mean and Median Survival time

- The expected value of the survival time is simply

$$\mathbb{E}(T) = \int_0^\infty t f(t) dt = \mu$$

- An alternative equivalent measurement is

$$\mu = \int_0^\infty S(t) dt$$

Note 1 it is defined ($\mu < \infty$) only if $S(\infty) = 0$: all subjects eventually fail

$\Rightarrow$ this might not be the case if, e.g., the survival outcome is time to cancer recurrence and a fraction c of subjects are completely cured

- The Median survival time is the time τ such that $S(\tau) = 1/2$

Note 2 If $S(t)$ is a step function, it is not continuous at $1/2$ and the Median is the smallest t such that $S(t) \leq 1/2$

Note 3 If $S(t)$ never drop below $c = 1/2$ during the observation period, the Median is undefined

Introduction to parametric Survival Distributions

- In view of modeling the survival process, we need to specify a distribution
- The simplest survival distribution is the exponential one

$$f(t) = \lambda e^{-\lambda t},$$

- The definitions of S12 allows to compute the SF

$$S(t) = e^{-\lambda t}$$

and alternative representations of S14 give

$$h(t) = \lambda$$

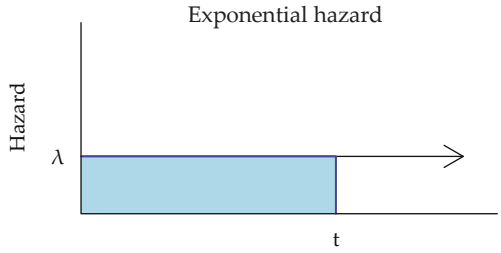
$\Rightarrow$ This SD has constant hazard function $h(t) = \lambda$

The Exponential Survival Distribution

- The cumulative hazard function is hence

$$H(t) = \int_0^t h(u)du = \int_0^t \lambda du = \lambda t$$

and is represented by the shaded area below



- The mean survival time is simply

$$\mathbb{E}(T) = \int_0^\infty S(t)dt = \int_0^\infty e^{-\lambda t}dt = 1/\lambda$$

The Weibull Survival Distribution

- The constant hazard is a strong assumption in many practical cases

$\Rightarrow$ a first generalization is obtained by considering

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

the hazard function derived from the Weibull distribution

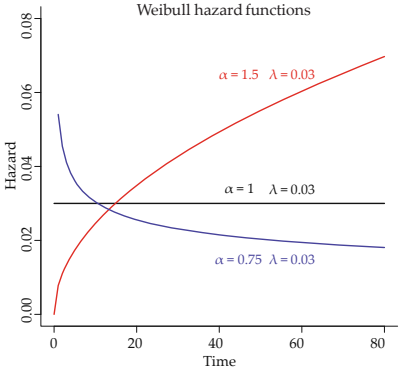
Note For $\alpha = 1$ it comes down to the exponential distribution

- From $h(t)$ one can easily derive $H(t) = (\lambda t)^{\alpha}$ and hence

$$S(t) = e^{-(\lambda t)^{\alpha}}$$

The Weibull Hazard Function

- For several parameter choices the behavior of $h(t)$ is represented below



- The mean survival time formula is not obvious

$$\mathbb{E}(T) = \int_0^\infty S(t)\,dt = \frac{\Gamma(1+1/\alpha)}{\lambda}$$

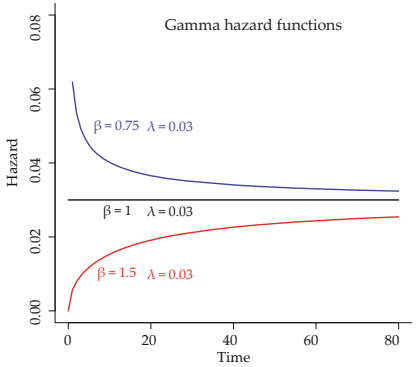
and the median survival time is given by $\tau = \log(2)^{1/\alpha}/\lambda$

The Gamma Hazard Function

- Another choice for survival modeling is the Gamma distribution

$$f(t) = \frac{\lambda^\beta t^{\beta-1} \exp(-\lambda t)}{\Gamma(\beta)}$$

which comes down to the exponential one for $\beta = 1$ as $\Gamma(1) = 1$



Note No closed form exist for the HF and SF $\Rightarrow$ numerical computations

Numerical approximation to the Hazard and Survival Functions

- In some cases (see e.g. S13), the distribution is much more complicated
- An alternative way is numerical computation :
  1. Take people dead at birth, after 1 day, week, month, year, 2 years, ...
  2. Take the data in difference to obtained rectangles
  3. Compute the cumulated sum of data in each rectangle to get $\widehat{H}(t)$
  4. The SF is simply given by $\widehat{S}(t) = \exp(-\widehat{H}(t))$
- One can use $\widehat{S}(t)$ to compute the mean that is

$$73.80$$

for the male and

$$78.90$$

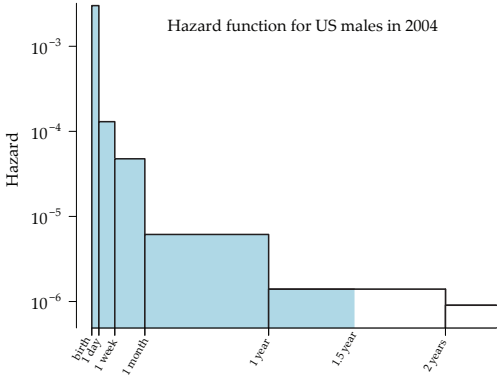
for the women when considering the US lifetime data of S13

Example of CHF approximation

- Step 1 to 3 allow to approximate the integral of $H(t)$

e.g. The male lifetime CHF up to 1.5 years is given by the blue area

$\Rightarrow$ Applying this method beyond 2 years leads to the blue CHF curve in S13



Unknown distribution parameters

- In general, we have poor knowledge upon
$$S(t)$$
the underlying Survival Distribution
- We only have realizations
$$t_1, t_2, \ldots, t_n$$
of random variables for which a distributional assumption is done

e.g. Under exponential distribution hypothesis, the parameter
$$\lambda$$
is unobserved and we would like to estimate it

$\Rightarrow$ A natural candidate is the Maximum Likelihood estimator

MLE principle for Survival data

- As in time series analysis, the likelihood function take the general form

$$L(\theta; t_1, t_2, \ldots, t_n) = f(t_1, \theta) \cdot f(t_2, \theta) \ldots f(t_n, \theta) = \prod_{i=1}^{n} f(t_i, \theta)$$

  with $\theta = \lambda$ in the exponential distribution case

Note However, particular attention has to be paid to censored data

e.g. For right-censored data we use δ and the Survival Function

$$S(t_i, \theta)^{1-\delta_i}$$

to indicate that observation i is known only to exceed t_i as

$$S(t_i, \theta) = \mathbb{P}(T_i > t_i)$$

$\Rightarrow$ The likelihood is hence transformed to

$$L(\theta; t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} f(t_i, \theta)^{\delta_i} S(t_i, \theta)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i, \theta)^{\delta_i} S(t_i, \theta)$$

Note For left-censored data we use δ and $1 - S(t_i, \theta) = \mathbb{P}(T_i \geq t_i) = F(t_i, \theta)$

MLE principle for exponential distribution

- In the particular case of the exponential distribution,

$$L(\theta; t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} \left(\lambda e^{-t_i/\mu}\right)^{\delta_i} \left(e^{-\lambda t_i}\right)^{1-\delta_i} = \lambda^d e^{-\lambda V}$$

  where $d = \delta_1 + \ldots + \delta_n$ is the total number of failure and

$$V = t_1 + \ldots + t_n$$

  is the total amount of time of patients

- The MLE is given by the value of λ that maximizes $L(\lambda; t_1, t_2, \ldots, t_n)$
- As log-transformation simplifies the likelihood function we prefer

$$\ell(\lambda) = \log L(\theta; t_1, t_2, \ldots, t_n) = d \log \lambda - \lambda V$$

- Under regularity conditions, the MLE is asymptotically Gaussian

Solution of exponential-based MLE

- The first derivative (score function) give

$$\ell'(\lambda) = \frac{d}{\lambda} - V$$

  and hence the maximum likelihood estimate is $\widehat{\lambda} = d/V$

- The second derivative (Hessian function) is

$$\ell''(\lambda) = -\frac{d}{\lambda^2} = -I(\lambda)$$

  where $I(\lambda) > 0$ is the Fisher information

- As $\ell''(\lambda) < 0$ the solution is a maximum and inversing $I(\lambda)$ we obtain

$$\mathbb{V}(\hat{\lambda}) = \sigma^2_\lambda \approx I^{-1}(\lambda) = \lambda^2/d$$

- In practice we will use

$$\widehat{\sigma}^2_\lambda \approx I^{-1}(\lambda) = \widehat{\lambda}^2/d = d/V^2$$

Note For most of distributions, no explicit solutions exist $\Rightarrow$ numerical resolution

Exercise

- Consider the data of Table 1
- Plot the log-likelihood and compute the MLE of λ and $\mathbb{V}(\hat{\lambda})$

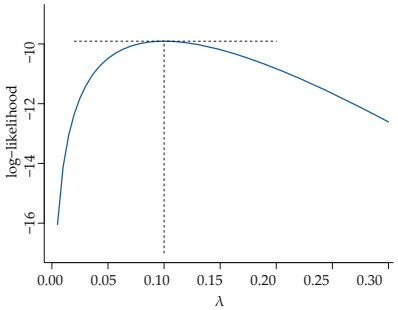
Exercise

- Consider the data of Table 1
- Plot the log-likelihood and compute the MLE of λ and $\mathbb{V}(\hat{\lambda})$
- Simple observation of the data gives $d = 3$ and $V = 7 + 6 + 6 + 5 + 2 + 4 = 30$

$\Rightarrow$ The log-likelihood function is

$$\ell(\lambda) = 3 \log \lambda - 30\lambda$$

and hence we obtain $\widehat{\lambda} = 3/30 = 0.1$ with $\widehat{\sigma}^2_{\lambda} \approx 3/(30^2) = 0.0033$



The Kaplan-Meier estimator (KPE)

- In practice, the distribution/survival/hazard function is hard to choose

$\Rightarrow$ The parametric approach is likely to be misspecified

- Nonparametric estimation procedures offer more flexibility

$\Rightarrow$ The most widely used of these procedures is the Kaplan-Meier estimator

$$\widehat{S}(t) = \prod_{t_i \leq t} (1 - \widehat{q}_i) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where d_i is the number of failure at time t_i and n_i the number of individuals at risk at that time

$\Rightarrow$ $\widehat{S}(t)$ is the product over failure times of the conditional probabilities of surviving to the next failure time

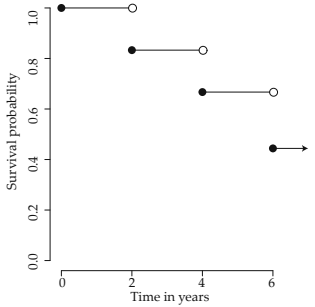
Application of the KPE

- By using the data of the Table 1, one can easily obtain

Table: Kaplan-Meier estimator

t_i	n_i	d_i	q_i	$1-q_i$	$\widehat{S}_i$
2	6	1	0.167	0.833	0.833
4	5	1	0.200	0.800	0.667
6	3	1	0.333	0.667	0.444

- One can use $\widehat{S}_i$ to reconstruct graphically the Survival Function



Interpretation of $\widehat{S}_t$

- $\widehat{S}_t$ is a non-increasing right continuous step function
  - t_i is the failure time
  - n_i is the number of individuals at risk at time t_i
  - d_i is the number of individuals who fail at time t_i
  - $q_i = d_i/n_i$ is the failure probability
  - $1 - q_i$ is the conditional survival probability
  - S_i is the Survival Function at time t_i
- The right-continuity is illustrated by open and closed circles

e.g. $S(4) = 0.667$ while $S(3.99) = 0.833$

Note The median is obtained for

$$t_i = \widehat{\tau} = 6,$$

that is the smallest time such that $S(t) \leq 1/2$ $(\widehat{S}(\tau) = 0.444)$

KPE and inference

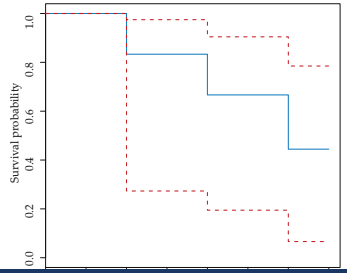
- The variance of the KPE can be approximated by

$$\mathbb{V}(\widehat{S}_t) \approx \widehat{S}_t^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

- Unfortunately, CI derived from $\mathbb{V}(\widehat{S}_t)$ may extend above 1 or below 0 but $S(t) \in [0, 1]$

$\Rightarrow$ One often overcome this issue by using a $\log$-$\log$ transformation of $\widehat{S}(t)$

$$\mathbb{V}(\log(-\log \widehat{S}_t)) \approx \frac{1}{(\log \widehat{S}_t)^2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$



Nelson-Altschuler estimate of the SF

- An alternative estimator is the one of Nelson-Altschuler based on $H(t)$

$$\widehat{S}_t = e^{-\widehat{H}(t)}, \quad \widehat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

Table: Nelson-Altschuler estimator

t_i	n_i	d_i	q_i	$\widehat{H}_i$	$\widehat{S}_i$
2	6	1	0.167	0.167	0.846
4	5	1	0.200	0.367	0.693
6	3	1	0.333	0.700	0.497

- Confidence intervals can be obtained in a similar way to KPE

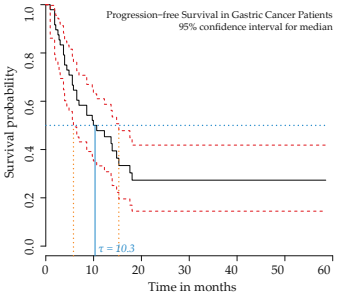
Median and inference

- As stated previously, the median is $\widehat{\tau} = \inf\{t : \widehat{S}(t) \leq 1/2\}$
- For a risk level α confidence intervals are given by

$$-z_{\alpha/2} \leq \frac{g\big(\widehat{S}(t)\big) - g\big(1/2\big)}{\mathbb{V}\Big(L\big(\widehat{S}(t)\big)\Big)^{1/2}} \leq z_{\alpha/2}$$

with $g(x) = \log(-\log(x))$ and $z_{\alpha/2}$ a Standard Normal quantile

e.g. Consider the data of Table 2 and the KPE : $\widehat{\tau} = 10.3$



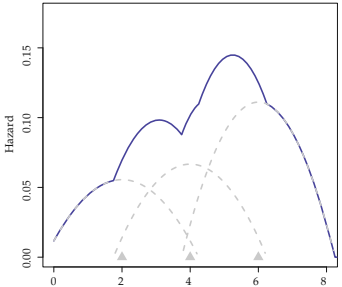
Kernel smoothing and Hazard Function estimation

- The Nelson-Altschuler estimate of $h(t)$ can be rough and quite instable
- A kernel function can be used to smooth $\widehat{h}(t)$

$$\widehat{h}(t) = \frac{1}{b} \sum_{i=1}^{D} \mathcal{K}\Big(\frac{t - t_i}{b}\Big) \frac{d_i}{n_i}$$

where $t_1 < \ldots < t_D$ are ordered failure times and b a tuning parameter

Note Many kernel function exist but the Epanechnikov kernel is very common
$\mathcal{K}(x) = 3/4(1 - x^2), \quad -1 \leq x \leq 1$



Corrected Kernel smoothing and Hazard Function estimation

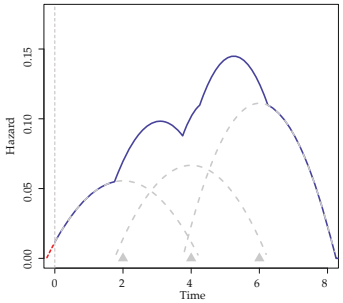
- Without corrections $\mathcal{K}(x)$ is likely to be $\neq 0$ at time $t < 0$

$\Rightarrow$ The first kernel below is centered at $t = 2$ and $b = 2.5$ meaning that

$$t - b = -0.5 \quad t + b = 4.5$$

and hence, the actual area under the first kernel is too small

$\Rightarrow$ The modified Epanechnikov kernel is recommended



- Another approach: setting a time-varying b

Example Kernel smoothing and Hazard Function estimation

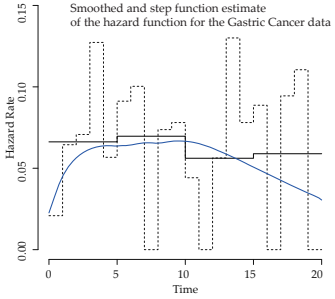
- Consider again the data of Table 2

$\Rightarrow$ Choose the modified Epanechnikov kernel with $b = 20$

Note Selection of b can be critical :

— if b is too small, the estimate may gyrate widely

— if b is too wide, the hazard function may be too smooth to observe real variations in the hazard function

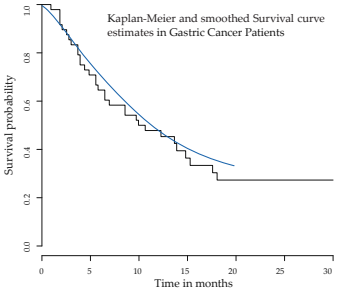


Example Kernel smoothing and Survival Function estimation

- One can use $\widehat{h}$ to obtain a smooth estimate of $S(t)$

$$\widehat{S}(t) = \exp\Big(- \int_{u=0}^{t} \widehat{h}(u)\, du \Big)$$

- In practice the integral is approximated by the rectangles method



Comparing Two Groups of Survival Times

- Comparison of distributional features is of crucial interest

e.g. In medical trials you need to compare treatment and control groups

$$H_0 : S_1(t) = S_0(t)$$

- Let $S_1(t)$ be the SF of the treatment group

$\Rightarrow$ Two alternative hypotheses can be specified (one-sided or two-sided)

$$H_1 : S_1(t) > S_0(t) \text{ or } H_1 : S_1(t) \neq S_0(t)$$

$\Rightarrow$ Unfortunately, Survival data imply several serious issues

— Survival distributions can be similar for some t and differ for others

— Survival distributions can cross

Lehman alternatives

- One solution is to consider Lehman-type alternatives defined as

$$H_1 : S_1(t) = \big(S_0(t)\big)^{\psi}$$

where $\psi \neq 1$ unless under

$$H_0 : S_1(t) = \big(S_0(t)\big)^{1}$$

$\Rightarrow$ The one-sided alternative is now

$$H_1 : \psi < 1$$

and imposes that $S_1(t)$ is uniformly higher than $S_0(t)$

- Theses hypotheses can be formulated in terms of proportional hazards

$$h_1(t) = \psi h_0(t)$$

The 2-by-2 Table representation

- In the spirit of the rank tests à la Mann-Whitney H_0 can be tested against Lehman alternatives

Note Complications arise from the presence of censoring

$\Rightarrow$ To solve this issue consider a two-by-two table representation of the data

Table: 2-by-2 Table representation

	Control	Treatment	Sums
Failure	d_{0i}	d_{1i}	d_i
Non-failure	$n_{0i} - d_{0i}$	$n_{1i} - d_{1i}$	$n_i - d_i$
At risk	n_{0i}	n_{1i}	n_i

- Numbers at risk for the control and treatment groups are n_{0i} and n_{1i}
- Numbers of failure for the control and treatment groups are d_{0i} and d_{1i}
- This representation is adopted for any distinct ordered failure time t_i

Hypergeometric distribution

- If one holds d_i, n_{0i} and n_{1i} fixed (and hence n_i too) we can derive

$$\mathbb{P}(d_{0i}|n_{0i}, n_{1i}, d_i) = \binom{n_{0i}}{d_{0i}}\binom{n_{1i}}{d_{1i}}\binom{n_i}{d_i}^{-1}$$

  the hypergeometric distribution of d_{0i} where

$$\binom{n_i}{d_i} = \frac{n_i!}{d_i!(n_i - d_i)!}$$

  represents the number of combinations of n items taken d at time t_i

- The 2 first moments of that distribution are

$$\mathbb{E}(d_{0i}) = \frac{n_{0i}d_i}{n_i} = \mu_{0i}$$

  and

$$\mathbb{V}(d_{0i}) = \frac{n_{0i}n_{1i}d_i(n_i - d_i)}{n_i^2(n_i - 1)} = \sigma_{0i}^2$$

The log-rank test statistics

- Based on the 2-by-2 representation and $\mathbb{E}(d_{0i})$ one can define

$$U_0 = \sum_{i=1}(d_{0i} - \mathbb{E}(d_{0i}))$$

  a simple linear test statistic and its variance

$$\mathbb{V}(U_0) = \sum_{i=1} \mathbb{V}(d_{0i})$$

- One can show that $U_0/\sqrt{\mathbb{V}(U_0)} \sim \mathcal{N}(0,1)$ or equivalently

$$\frac{U_0^2}{\mathbb{V}(U_0)} \sim \chi_1^2$$

- This test statistic is known as the log-rank test of group comparison

Note 1 This test is also known as the Mantel-Haenzel test

Note 2 A comparison of k groups is possible and modify the distribution to

$$\chi_{k-1}^2$$

but is slightly different from the stratified tests discussed in S51

Exercise: application of the $\log$-rank test

- Consider the following survival data
- C and T stand for Control and Treatment groups respectively

Table: Survival data

Patient	Survtime	Censor	Group
1	6	1	C
2	7	0	C
3	10	1	T
4	15	1	C
5	19	0	T
6	25	1	T

- When required, construct the 2-by-2 tables
- Compute the $\log$-rank test and interpret the result

Exercise: computation

- Failures appear at $t = 6, 10, 15, 25$ and result in four 2-by-2 tables

Table: 2-by-2 tables for $t = 6, 10, 15, 25$

	t = 6			t = 10			t = 15			t = 25		
	C	T	$\sum$	C	T	$\sum$	C	T	$\sum$	C	T	$\sum$
Failure	1	0	1	0	1	1	1	0	1	0	1	1
Non-failure	2	3	5	1	2	3	0	2	2	0	0	0
At risk	3	3	6	1	3	4	1	2	3	0	1	1

Table: Intermediate calculus to compute the log-rank test statistic

t_i	n_i	d_i	n_{0i}	d_{0i}	n_{1i}	d_{1i}	μ_{0i}	σ^2_{0i}
6	6	1	3	1	3	0	0.500	0.2500
10	4	1	1	0	3	1	0.250	0.1875
15	3	1	1	1	2	0	0.333	0.2222
25	1	1	0	0	1	1	0.000	0.0000
$\sum$				2		2	1.083	0.6597

Exercise: interpretation

- From Tables in previous slide we easily obtain

$$U_0 = \sum_i d_{0i} - \sum_i \mu_{0i} = O_0 - E_0 = 2 - 1.083 = 0.917$$

and $\mathbb{V}(U_0) = \sum_i \sigma^2_{0i} = V_0 = 0.6597$

$\Rightarrow$ The $\log$-rank test statistic is

$$\frac{U_0^2}{\mathbb{V}(U_0)} \approx 1.26$$

which we compare to a χ^2_1 distribution

$\Rightarrow$ The corresponding p-value is

$$p = 0.259$$

meaning that we cannot reject H_0 and hence the group difference is not statistically significant

Note When applying the test to d_{1i}, the result is identical as it also sums to 2

The generalized log-rank test statistics

- An important generalization of the log-rank test is

$$U_0(w) = \sum_{i=1} w_i(d_{0i} - \mathbb{E}(d_{0i}))$$

with the corresponding variance $\mathbb{V}(U_0) = \sum_{i=1} w_i^2 \mathbb{V}(d_{0i})$

- This leads to the so called Fleming-Harrington $G(\rho)$ test

$$G(\rho) = \frac{U_0(w)^2}{\mathbb{V}(U_0(w))}$$

- The most common way of setting weights is à la Gehan-Wilcoxon

$$w_i = \mathcal{F}\big(\widehat{S}(t_i)\big)^\rho, \quad \mathcal{F}(.) \text{ being a certain function}$$

Note 1 When $\rho = 1$ we get the Prentice modification : places higher weight on earlier survival times

Note 2 When $w_i = \sqrt{n_i}$ we get the Tarone-Ware modification : intermediate weight compared to $\rho = 0$ and $\rho > 0$

Note 3 When $w_i = \widehat{S}(t_i)^p(1 - \widehat{S}(t_i))^q$ we get the Harrington-Fleming(p, q) test : more flexible

Example : Prentice modification of Gehan-Wilcoxon test

- Let consider pancreatic cancer data from a clinical trial (41 patients)
- We are interested in the progression-free survival (PFS)

$\Rightarrow$ the time from assignment in the trial to disease progression or death

Table: Locally Advanced Pancreatic Cancer or Metastatic Pancreatic Cancer

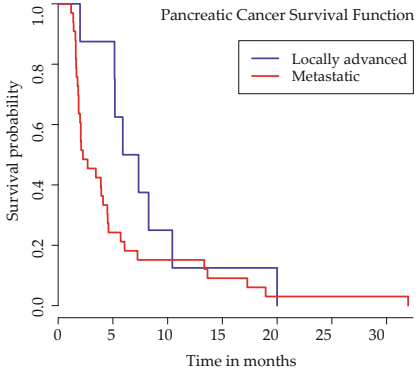
	stage	onstudy	progression	death
1	MPC	16/12/2005	02/02/2006	19/10/2006
2	MPC	06/01/2006	26/02/2006	19/04/2006
3	LAPC	03/02/2006	02/08/2006	19/01/2006
4	MPC	30/03/2006	"NA"	11/05/2006
5	LAPC	27/04/2006	11/03/2007	29/05/2007
6	MPC	07/05/2006	25/06/2006	11/10/2006
⋮	⋮	⋮	⋮	⋮

Note 1 "NA" means that the patient died with no recorded progression and the PFS is time to death

Note 2 For all other patients, the PFS is time to the date of progression

Example : Prentice modification of Gehan-Wilcoxon test

- The graphical analysis of SF reveals :
  - the LAPC group shows an early survival advantage over the MPC
  - but the survival curves converge after about 10 months



Example : Prentice modification of Gehan-Wilcoxon test

- When computing the Gehan-Wilcoxon test for $\rho = 0$ (i.e. the log-rank test) and

$$\rho = 1$$

i.e. the Prentice modification, we obtain

Table: Fleming-Harrington $G(\rho)$ for $\rho = 0$ and $\rho = 1$, with $k = \{0, 1\}$

$\rho = 0$	N	O_k	E_k	$(O_k - E_k)^2 / V_k$
LAPC	8	8	1.49	2.25
MPC	33	33	0.64	2.25
We cannot reject H_0 (no difference) as				p-value $= 0.134$

$\rho = 1$	N	O_k	E_k	$(O_k - E_k)^2 / V_k$
LAPC	8	2.34	2.13	4.71
MPC	33	18.76	0.82	4.71
We reject H_0 as				p-value $= 0.0299$

- The two tests produce conflicting results as they are optimized for different alternatives

$\Rightarrow$ For $\rho = 1$, the test places higher weight on earlier survival times

Stratified tests

- To compare two groups while adjusting for another covariate, one can
  1. include the other covariate as regression terms for the hazard function (see next Chapter)
  2. construct a stratified log-rank test if the covariate we are adjusting for is categorical

$\Rightarrow$ denote h_{0j} the population hazard of level $j = 1, 2, \ldots, G$, with G small

- For the G categories of the covariate we can test

$$H_0 : h_{0j}(t) = h_{1j}(t), \quad j = 1, 2, \ldots, G$$

- Accordingly, the stratified version of the log-rank test statistic is

$$X^2 = \frac{\left(\sum_{g=1}^{G} U_{0g}\right)^2}{\sum_{g=1}^{G} V_{0g}} \sim \chi_1^2$$

Example 1 of stratified test

- Consider the dataset of Table 3 (time to return smoking)
- We first compare the 2 treatment groups by means of the log-rank test

$\rho = 0$	N	O_k	E_k	$(O_k - E_k)^2 / V_k$
Combination	61	37	49.9	8.03
Patch only	64	52	39.1	8.03
We reject H_0 (no difference) as				p-value $= 0.00461$

- If now we are interested by the influence of the age we may define

$$g = 1 : 21 - 49 \;||\; g = 2 : 50 \text{ or more}$$

a categorical variable that divides the subjects in 2 groups

- The resulting stratified log-rank test is close to the unadjusted test

$\Rightarrow$ the stratification based on the age seems unnecessary

$\rho = 0$	N	O_k	E_k	$(O_k - E_k)^2 / V_k$
Combination	61	37	49.1	7.03
Patch only	64	52	39.9	7.03
We reject H_0 (no difference) as				p-value $= 0.008$

Example 2 of stratified test

- Consider simulated data representing an artificial clinical trial
- This trial compares a standard therapy (control) and an experimental one (treatment)
- The survival times are simulated as exponentially distributed and produces no censoring
- A confounding genotype factor is also simulated with only 2 levels

$$g = 1: \text{ wild type genotype } || \ g = 2: \text{ mutant genotype}$$

with $g = 2$ leading to poorer prognosis as the hazard rate is

$$\lambda = 0.03 \text{ per day}$$

for a mutant patient in the control group whilst the effect of treatment leads to

$$\lambda = 0.0165$$

- For wild type patients $\lambda = 0.006$ whilst the effect of treatment leads to

$$\lambda = 0.0033$$

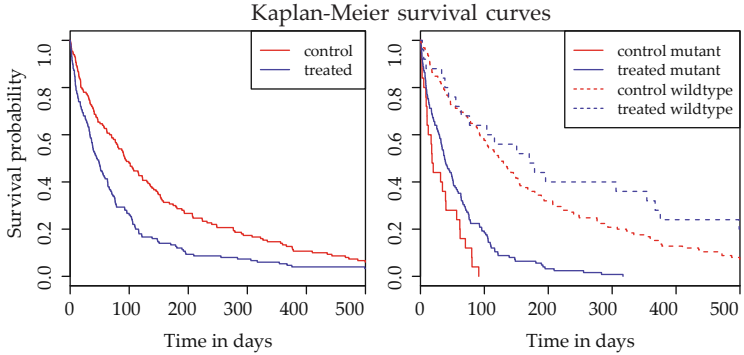
Example 2 of stratified test

- The Kaplan-Meier survival curves are computed both naively and accounting for the gene confounder

Note 1 The naive estimate concludes against the experimental therapy

Note 2 When accounting for the gene confounder the results are at the opposite

$\Rightarrow$ within each genotype, the treatment is actually superior to the control



Example 2 of stratified test

- The stratified log-rank test is now used to confirm the graphical analysis

Unadjusted	N	O_k	E_k	$(O_k - E_k)^2 / V_k$
Control	150	150	183	15.9
Treatment	150	150	117	15.9
We reject H_0 (no difference) as				p-value $= 0.00006$

Note 1 The unadjusted test shows that the treatment reduces survival

Stratified	N	O_k	E_k	$(O_k - E_k)^2 / V_k$
Control	150	150	133	7.57
Treatment	150	150	167	7.57
We reject H_0 (no difference) as				p-value $= 0.00595$

Note 2 The stratified test confirms that the treatment improves survival compared to the control

Note 3 Patients carrying the wild type form of the gene have better survival than do patients carrying the mutation

Note 4 There are more mutation-carrying patients in the treatment group than in the control group, whereas the reverse is true for wild type patients

Chapter 2

Chapitre 2 : Hazards Model

Non parametric models

- As discussed in Chapter 1, Lehman-type alternatives are defined as

$$H_1 : S_1(t) = \big(S_0(t)\big)^{\psi}$$

where $\psi \neq 1$ unless under

$$H_0 : S_1(t) = \big(S_0(t)\big)^{1}$$

$\Rightarrow$ theses hypotheses can be formulated in terms of proportional hazards

$$h_1(t) = \psi h_0(t)$$

- The latter Eq. is the key to quantify the difference between two hazard functions by means of the so-called proportional hazards model
- We can extend the model to include covariate information x as follows

$$\psi = e^{x\beta}$$

- Other functional are possible albeit this is the most common in practice

Note The estimation is complicated in absence of parametric form for

$$h_0(t),$$

and require the concept of partial likelihood developed by Cox

Introduction to the partial likelihood

- Let j denotes the j'th failure time (sorted from lowest to highest)
- Let $h_i(t_j)$ be the hazard function for subject i at failure time t_j

$\Rightarrow$ The Cox proportional hazards (semi-parametric) model is

$$h_i(t_j) = \psi_i h_0(t_j), \quad \psi_i = e^{x_i'\beta}$$

Note ψ_i characterize the hazard ratio $h_i(t_j)/h_0(t_j)$

- In the simplest case where we compare two groups (dummy variable)

$$x_i = \{0, 1\}$$

- In the particular case of control vs treatment group we expect

$$\beta < 0$$

as the experimental group is less likely than control patients to fail

$\Rightarrow$ Hence, $\psi_i < 1$ ($\psi_i = 1$) is expected in the treatment (control) group

The partial likelihood

- Consider the first failure time t_1 and let

$$R_1$$

be the set of all subjects at risk for failure at this time (the risk set)

- The probability that the subject i fails is its hazard divided $\sum h_k(t_1)$

$$\mathbb{P}_1 = \frac{h_i(t_1)}{\sum_{k\in R_1} h_k(t_1)} = \frac{\psi_i h_0(t_1)}{\sum_{k\in R_1} \psi_k h_0(t_1)} = \frac{\psi_i}{\sum_{k\in R_1} \psi_k}$$

where $h_0(t_1)$ is the hazard for a subject from the control group

- At failure time t_2 a new (smaller) risk set R_2 is considered

$\Rightarrow$ We repeat this calculation to obtain p_2 and so on up to t_n

- The partial likelihood is the product

$$\mathcal{L}(\psi) = \mathbb{P}_1\mathbb{P}_2\ldots\mathbb{P}_n$$

Example of partial likelihood computation

- Consider the following (artificial) data (see also Chapter 1)

Table: Survival data

Patient	Survtime	Censor	Group
1	6	1	$C(x_1 = 0)$
2	7	0	$C(x_2 = 0)$
3	10	1	$T(x_3 = 1)$
4	15	1	$C(x_4 = 0)$
5	19	0	$T(x_5 = 1)$
6	25	1	$T(x_6 = 1)$

- Consider the following (artificial) data (see also Chapter 1)

$\Rightarrow$ the first failure time is at $t = 6$ and for each patient we have either

$$\psi_1 = \psi_2 = \psi_4 = 1 \text{ or } \psi_3 = \psi_5 = \psi_6 = \psi$$

i.e. we have 6 patients at risk (3 in the "C" group for which $\psi = 1$) and

$$\mathbb{P}_1 = \frac{\psi_1 h_0(t_1)}{3\psi h_0(t_1) + 3h_0(t_1)} = \frac{1}{3 \times \psi + 3}$$

Example of partial likelihood computation

- The second failure time is at $t = 10$ because at $t = 7$ there is no failure

Note At $t = 7$ we have a "C" patient that dropped out due to censoring

$\Rightarrow$ Of the 6 patients at risk at the first time, only 4 remains in R_2 and

$$\mathbb{P}_2 = \frac{\psi}{3\psi + 1}$$

where ψ appears in the numerator as the patient 3 was in the "T" group

- The third failure time (t_3) is at $t = 15$ with 3 patients in R_3 and

$$\mathbb{P}_3 = \frac{1}{2\psi + 1}$$

- The last failure time (t_4) is at $t = 25$ with 1 patient in R_4 and

$$\mathbb{P}_4 = \frac{\psi}{\psi} = 1$$

as she is in the "T" group

Example of partial likelihood computation

- Now we are ready to compute the partial likelihood

$$\mathcal{L}(\psi) = \mathbb{P}_1\mathbb{P}_2\mathbb{P}_3\mathbb{P}_4 = \frac{\psi}{(3\psi+3)(3\psi+1)(2\psi+1)}$$

- In the case of a Cox model the log partial likelihood is

$$\ell(\beta) = \beta - \log(3\exp(\beta)+3) - \log(3\exp(\beta)+1) - \log(2\exp(\beta)+1)$$

as ψ is assumed to be of exponential form : $\psi = e^{\beta}$

$\Rightarrow$ The maximum partial likelihood estimate is

$$\widehat{\beta}$$

the value of β that maximizes this function

Note 1 As discussed above, it is nonparametric because the hazard function

$$h_0(t)$$

does not enter the partial likelihood and hence requires no specification

Note 2 Unlike traditional likelihood, $\mathcal{L}(\psi)$ is not a probability but allows to estimate β

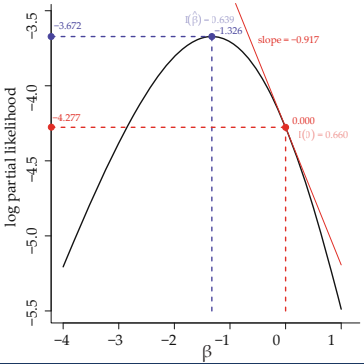
Example of partial likelihood computation

- $\widehat{\beta} = -1.3261$ is obtain by numerical optimization
- We anticipate on the next slide and report some test statistics

Note 1 The null hypothesis ($\beta = 0$) is reported for comparison

Note 2 The slope of the tangent is given by the LM statistic $S(\beta) = \ell'(\beta)$

Note 3 $I(\beta) = -S'(\beta) = -\ell''(\beta)$ denotes the fisher information



Partial likelihood hypothesis tests

- As in standard likelihood one can derive 3 types of test for $H_0 : \beta = 0$
  - The Wald test
  - The LM test
  - The LR test
- The limit theory of theses tests can differ and is often more difficult to derive
- In view of presenting them, define
  - $S(\beta) = \ell'(\beta)$, the score function
  - $I(\beta) = -S'(\beta) = -\ell''(\beta)$, the fisher information
  - $I(\widehat{\beta})$, the observed information

The Wald test

- The Wald test is of form

$$Z_W = \frac{\widehat{\beta}}{\sigma_{\widehat{\beta}}}$$

where $\sigma^2_{\widehat{\beta}}$ is obtained numerically from the negative inverse of the Hessian

$$I(\widehat{\beta})^{-1} = -\ell''(\widehat{\beta})^{-1}$$

Note As the second derivative reflects the curvature of the likelihood, a sharper curve (i.e. more information) leads to lower variance

- Under the null hypothesis $H_0 : \beta = 0$, this normalized statistic if Gaussian

$\Rightarrow$ We reject H_0 if $|Z_W| > z_{\alpha/2}$ or $Z_W^2 > \chi^2_{\alpha,1}$

- The asymptotic normality can be used to construct confidence intervals

$$\widehat{\beta} \pm z_{\alpha/2} \times \sigma_{\widehat{\beta}}$$

The Lagrange Multiplier (score) test

- The LM test is based on the score of the partial log-likelihood

$\Rightarrow$ The variance of this test is hence directly $I(\beta)$

- The test is computed under the null hypothesis as follows

$$Z_{LM} = \frac{S(\beta = 0)}{\sqrt{I(\beta = 0)}}$$

$\Rightarrow$ We reject $H_0 : \beta = 0$ if $|Z_{LM}| > z_{\alpha/2}$ or $Z_W^2 > \chi^2_{\alpha,1}$

Note 1 This test can be computed without finding the MPLE

Note 2 This test is equivalent to the log-rank test statistic U_0 discussed in Chapter 1

$\Rightarrow$ With the same artificial data of Table 12, U_0 was equal to $0.917 \equiv -S(0)$

The Likelihood Ratio test

- The LR test is based on the asymptotic behavior of

$$Z_{LR} = 2\big(\ell(\beta = \widehat{\beta}) - \ell(\beta = 0)\big) \sim \chi^2_1$$

- Z_{LR} is invariant to monotonic transformations of β (unlike the LM and Wald tests)

$\Rightarrow$ Whether the test is computed in terms of β or $\psi = \exp(\beta)$ has no effect on the p-value

$\Rightarrow$ We reject H_0 if $Z^2_{LR} > \chi^2_{\alpha,1}$

Exercise : computation of partial likelihood hypothesis tests

- Consider the MPLE results plotted on S63

$\Rightarrow$ All elements needed to compute Z_W, Z_{LM}, Z_{LR} are there

Exercise : computation of partial likelihood hypothesis tests

- Consider the MPLE results plotted on S63

$\Rightarrow$ All elements needed to compute Z_W, Z_{LM}, Z_{LR} are there

- For Z_{LM} we have

$$Z_{LM}^2 = \Big(\frac{S(\beta=0)}{\sqrt{I(\beta=0)}}\Big)^2 = \frac{(-0.917)^2}{0.660} = 1.274$$

  Any software can compute the p-value which is $p = 0.2591$

- For Z_W we have

$$Z_W^2 = \Big(\frac{\widehat{\beta}}{\sigma_{\widehat{\beta}}}\Big)^2 = \Big(\frac{-1.326129}{\sqrt{1/0.639}}\Big)^2 = 1.124$$

  Any software can compute the p-value which is $p = 0.2891$

- Finally, for Z_{LR} we have

$$Z_{LR} = 2\big(\ell(\beta=\widehat{\beta}) - \ell(\beta=0)\big) = 2(-3.672 + 4.277) = 1.209$$

  Any software can compute the p-value which is $p = 0.2715$

Pseudo-R^2 statistic

- At this stage one can also use
$$\ell(\beta = \widehat{\beta}) \text{ and } \ell(\beta = 0)$$
to compute an adaptation of the R^2 statistic to survival analysis
- The R^2_{CS} statistic (Cox and Snell) is defined as follows
$$R^2_{CS} = 1 - \left(\frac{\ell(0)}{\ell(\beta)} \right)^{2/n}$$

$\Rightarrow$ R^2_{CS} reflects the improvement in the fit of the model with the covariate compared to $\beta = 0$

Note R^2_{CS} has a major drawback as it is capped to 0.75 but alternatives are not consensual

The partial likelihood with multiple covariates

- To achieve greater generality we now consider the case where

$$x_i = (x_{i,1}, \cdots, x_{i,p})'$$

  is a vector of p dummy covariates for each individual i
- To save place we use ψ_i in place of $\psi_i(x_i, \beta)$, where β is now a vector of p coefficients
- In the particular case of the Cox model, the hazard ratio is $\exp(x_i'\beta)$
- As in S59, before the first failure time, all of the subjects are said to be at risk

$\Rightarrow$ Among them one will fail at time t_1 in the risk set R_1

- More generally, at time t_j, the risk set is R_j leading to

$$\mathcal{L}(\beta) = \prod_{j=1}^{D} \frac{h_i(t_j)}{\sum_{k \in R_j} h_k(t_j)} = \prod_{j=1}^{D} \frac{\psi_j h_0(t_j)}{\sum_{k \in R_j} \psi_k h_0(t_j)} = \prod_{j=1}^{D} \frac{\psi_j}{\sum_{k \in R_j} \psi_k}$$

  for the Cox proportional hazard model, with D the number of failures

The log partial likelihood with multiple covariates

- The log partial likelihood is simply given by

$$\ell(\beta) = \sum_{j=1}^{D} \left(\log(\psi_j - \log \left(\sum_{k \in R_j} \psi_k \right) \right) = \sum_{j=1}^{D} x_j' \beta - \sum_{j=1}^{D} \log \left(\sum_{k \in R_j} \exp(x_k' \beta) \right)$$

- The score function has p components, one for each of the p covariates

$\Rightarrow$ For the l'th component the score is given by

$$S_l(\beta) = \frac{\partial \ell(\beta)}{\partial \beta_l} = \sum_{j=1}^{D} \left(x_{jl} - \frac{\sum_{k \in R_j} x_{jk} \exp(x_j' \beta)}{\sum_{k \in R_j} \exp(x_j' \beta)} \right)$$

Note We may view the score function as the sum of "residuals"

$\Rightarrow$ The observed value x_{jl} of the covariate l minus an "expected" value

Recall When x_j is a single binary covariate, $S(\beta = 0)$ is the log-rank statistic

Note The Fisher information matrix is now a matrix

$$I(\beta; x) = -\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta'} = -\frac{S(\beta)}{\partial \beta}$$

Wald, LR and LM tests with multiple covariates

- In presence of multiple covariates the usual tests are as follows
- The Wald test under $H_0 : \beta = 0$ is
$$Z_W^2 = \widehat{\beta}' I(\widehat{\beta}; x)\widehat{\beta}$$
- The LM test :
$$Z_{LR}^2 = S'(\beta = 0; x)I(\beta = 0; x)^{-1}S(\beta = 0; x)$$
- The LR test :
$$Z_{LM}^2 = 2\big(\ell(\beta = \widehat{\beta}) - \ell(\beta = 0)\big)$$
- Under H_0, all 3 statistics are asymptotically χ^2_{k-1}

Exercise with multiple covariates

- Consider the exponential survival data simulated in Chapter 1

$\Rightarrow$ A confounding binary genotype factor was introduced :

$$g = 1 \text{ (wild type) or } g = 2 \text{ (mutant type)}$$

- When estimating the Cox model to compare trivially the "T" and "C" group we obtain

$$\widehat{\beta} = 0.464 (\sigma_{\widehat{\beta}} = 0.117) \text{ with } LR = 15.5 (p = 0.00000)$$

$\Rightarrow$ How to interpret those results ?

Exercise with multiple covariates

- Consider the exponential survival data simulated in Chapter 1

$\Rightarrow$ A confounding binary genotype factor was introduced :

$$g = 1 \text{ (wild type) or } g = 2 \text{ (mutant type)}$$

- When estimating the Cox model to compare trivially the "T" and "C" group we obtain

$$\widehat{\beta} = 0.464 (\sigma_{\widehat{\beta}} = 0.117) \text{ with } LR = 15.5 (p = 0.00000)$$

$\Rightarrow$ How to interpret those results ?

Note 1 It suggests higher hazards for the "T" group ($\widehat{\beta} > 0$) with a significant difference with the "C" group

Note 2 Also, $\exp(\widehat{\beta}) = 1.59$ indicates that the "T" group is associated with a 59% additional risk of death over the "C" group

Exercise with multiple covariates

- As for the log-rank test, it is possible to stratified the data
- When estimating the stratified Cox model to compare the "T" and "C" group we obtain

$$\widehat{\beta} = -0.453 (\sigma_{\widehat{\beta}} = 0.164) \text{ with } LR = 7.66 (p = 0.00566)$$

$\Rightarrow$ How to interpret those results ?

Exercise with multiple covariates

- As for the log-rank test, it is possible to stratified the data
- When estimating the stratified Cox model to compare the "T" and "C" group we obtain

$$\widehat{\beta} = -0.453 (\sigma_{\widehat{\beta}} = 0.164) \text{ with } LR = 7.66 (p = 0.00566)$$

$\Rightarrow$ How to interpret those results ?

Note 1 It suggests higher hazards for the "C" ($\widehat{\beta} < 0$) group with a significant difference with the "T" group

Note 2 Also, $\exp(\widehat{\beta}) = 0.636$ indicates that the "T" group is associated with

$$1 - 0.636 = 36\%$$

less risk of death over the "C" group

Exercise with multiple covariates

- Finally, we introduce the genotype as a covariate
- When estimating the Cox model with the two covariates we obtain

$$\widehat{\beta}_{grp} = -0.453 (\sigma_{\widehat{\beta}_{grp}} = 0.163)$$

and

$$\widehat{\beta}_{gen} = -1.568 (\sigma_{\widehat{\beta}_{gen}} = 0.183)$$

with

$$LR = 93.4 (p = 0.00000)$$

$\Rightarrow$ How to interpret those results ?

Exercise with multiple covariates

- Finally, we introduce the genotype as a covariate
- When estimating the Cox model with the two covariates we obtain

$$\widehat{\beta}_{grp} = -0.453 (\sigma_{\widehat{\beta}_{grp}} = 0.163)$$

and

$$\widehat{\beta}_{gen} = -1.568 (\sigma_{\widehat{\beta}_{gen}} = 0.183)$$

with

$$LR = 93.4 (p = 0.00000)$$

$\Rightarrow$ How to interpret those results ?

Note 1 As for the stratified Cox model, the correct treatment effect is identified

Note 2 Indeed, we see higher hazards for the "C" ($\widehat{\beta} < 0$) group with a significant difference with the "T" group

Tied survival times

- Tied survival time are failure that occurs simultaneously

Note 1 In continuous time data this is likely to arise due to rounding

Note 2 In discrete time data this can genuinely appear

Note 3 If censoring times are tied with failure times, the convention is to consider the failures to precede the censoring

Example Consider a continuous time process and the following reports

Table: Survival data with tied survival times

Patient	Survtime	Censor	Group
1	1	1	T
2	1	1	T
3	2	1	C
4	3	0	T
5	4	1	T
6	4	1	C
7	5	0	C
8	6	1	C
9	6	0	C
10	7	0	C

Tied survival times and partial likelihood

- As the underlying times are actually continuous we use the Cox model

$$h(t;x) = e^{x\beta} h_0(t)$$

where $x = 1$ or 0 for the treatment or control group, respectively

- As in the regular case, the likelihood is the product of probabilities

$\mathbb{P}_1$ At $t = 1$, all 10 patients are at risk and two of them fail, both from the "T" group, and either of those two patients may have failed first

$\Rightarrow$ We account for those two possibilities when constructing $\mathbb{P}_1$

$$\mathbb{P}_1 = \frac{\exp(\beta)}{4\exp(\beta)+6}\frac{\exp(\beta)}{3\exp(\beta)+6} + \frac{\exp(\beta)}{4\exp(\beta)+6}\frac{\exp(\beta)}{3\exp(\beta)+6} = A \times B + C \times D$$

- The first (second) product assumes that patient 1 (2) fails first

Note 1 In B, 4 becomes 3 as patient 1 has failed

Note 2 In D, 4 becomes 3 as patient 2 has failed

Note 2 As both patients are in the "T" group the $A \times B$ and $C \times D$ are symmetric

Exercise : tied survival times and partial likelihood

- We want to derived the remaining terms of the partial likelihood

Exercise : tied survival times and partial likelihood

- We want to derived the remaining terms of the partial likelihood

$\mathbb{P}_2$ At $t = 2$, 8 patients are at risk (2 and 6 in the "T" and "C" group resp.)

$\Rightarrow$ As there is only 1 failure in the "C" group we have

$$\mathbb{P}_2 = \frac{1}{2\exp(\beta) + 6}$$

Exercise : tied survival times and partial likelihood

- We want to derived the remaining terms of the partial likelihood

$\mathbb{P}_2$ At $t = 2$, 8 patients are at risk (2 and 6 in the "T" and "C" group resp.)

$\Rightarrow$ As there is only 1 failure in the "C" group we have

$$\mathbb{P}_2 = \frac{1}{2\exp(\beta) + 6}$$

$\mathbb{P}_3$ At $t = 4$, 6 patients are at risk (as at $t = 3$ patient 4 is censored)

$\Rightarrow$ We have two failures, one in each group, and

$$\mathbb{P}_3 = \frac{1}{\exp(\beta) + 5} \times \frac{\exp(\beta)}{\exp(\beta) + 4} + \frac{\exp(\beta)}{\exp(\beta) + 5} \times \frac{1}{5}$$

to account for all scenarios of failure (patient 5 first or patient 6 first)

Exercise : tied survival times and partial likelihood

- We want to derived the remaining terms of the partial likelihood

$\mathbb{P}_2$ At $t = 2$, 8 patients are at risk (2 and 6 in the "T" and "C" group resp.)

$\Rightarrow$ As there is only 1 failure in the "C" group we have

$$\mathbb{P}_2 = \frac{1}{2\exp(\beta) + 6}$$

$\mathbb{P}_3$ At $t = 4$, 6 patients are at risk (as at $t = 3$ patient 4 is censored)

$\Rightarrow$ We have two failures, one in each group, and

$$\mathbb{P}_3 = \frac{1}{\exp(\beta) + 5} \times \frac{\exp(\beta)}{\exp(\beta) + 4} + \frac{\exp(\beta)}{\exp(\beta) + 5} \times \frac{1}{5}$$

to account for all scenarios of failure (patient 5 first or patient 6 first)

- Only 1 constant factor remains as patients 7 and 10 are censored and

$$\mathbb{P}_4 = \frac{1}{3}$$

as at $t = 6$, by convention, the censored patient 9 failed after patient 8

$\Rightarrow$ One may express the partial likelihood as $\mathcal{L}(\beta) = \mathbb{P}_1\mathbb{P}_2\mathbb{P}_3$ or $\mathbb{P}_1\mathbb{P}_2\mathbb{P}_3\mathbb{P}_4$

Discrete tied survival times

- Consider now that times are in fact discrete in the table below

$\Rightarrow$ In such a case, the Cox model is transformed to a discrete logistic model

$$\frac{h(t;x)}{1-h(t;x)} = e^{x\beta}\frac{h_0(t)}{1-h_0(t)}$$

Table: Survival data with tied survival times

Patient	Survtime	Censor	Group
1	1	1	T
2	1	1	T
3	2	1	C
4	3	0	T
5	4	1	T
6	4	1	C
7	5	0	C
8	6	1	C
9	6	0	C
10	7	0	C

Discrete tied survival times and partial likelihood

- At $t = 1$, as 2 patients fail among the 10 patients at risk we now have

$$\binom{10}{2} = \frac{10!}{2!(n-k)!} = 45$$

  pairs that could represent the two failures
- All factors are summarized in the matrix below and lead to

$$\mathbb{P}_1 = \frac{e^{2\beta}}{6e^{2\beta} + 24e^{\beta} + 15}$$

Table: Pairs that could represent two failures among 10 patients

	e^β	e^β	e^β	e^β	1	1	1	1	1	1
e^β	•									
e^β	$e^{2\beta}$	•								
e^β	$e^{2\beta}$	$e^{2\beta}$	•							
e^β	$e^{2\beta}$	$e^{2\beta}$	$e^{2\beta}$	•						
1	e^β	e^β	e^β	e^β	•					
1	e^β	e^β	e^β	e^β	1	•				
1	e^β	e^β	e^β	e^β	1	1	•			
1	e^β	e^β	e^β	e^β	1	1	1	•		
1	e^β	e^β	e^β	e^β	1	1	1	1	•	
1	e^β	e^β	e^β	e^β	1	1	1	1	1	•

Exercise : discrete tied survival times and partial likelihood

- We want to compute the remaining factors

Exercise : discrete tied survival times and partial likelihood

- We want to compute the remaining factors
- At $t = 2$, there is only 1 failure in the "C" group $\Rightarrow \mathbb{P}_2 = 1/(2e^{\beta} + 6)$

Exercise : discrete tied survival times and partial likelihood

- We want to compute the remaining factors
- At $t = 2$, there is only 1 failure in the "C" group $\Rightarrow \mathbb{P}_2 = 1/(2e^{\beta} + 6)$
- At $t = 4$, there are 2 failures and 6 patients are at risk such that we have

$$\binom{6}{2} = 15$$

possible pairs, of which 1 is from the "T" group and 1 from the "C" group

$$\mathbb{P}_3 = \frac{\exp(\beta) \times 1}{5\exp(\beta) + 10}$$

$\Rightarrow$ Again, one may simply express the partial likelihood as $\mathcal{L}(\beta) = \mathbb{P}_1\mathbb{P}_2\mathbb{P}_3$

Table: Pairs that could represent two failures among 6 patients

	e^{β}	1	1	1	1	1
e^{β}	•					
1	e^{β}	•				
1	e^{β}	1	•			
1	e^{β}	1	1	•		
1	e^{β}	1	1	1	•	
1	e^{β}	1	1	1	1	•

Approximation in presence of tied survival times

- With many ties, the discrete and continuous methods are cumbersome

$\Rightarrow$ Two approximation methods can be implemented

Breslow It adjusts the denominator to simply reflect all patients at risk

$\Rightarrow$ In the previous example, $\mathbb{P}_1$ and $\mathbb{P}_3$ becomes

$$\mathbb{P}_1 = \frac{2e^{2\beta}}{(6e^{\beta}+4)^2} \text{ and } \mathbb{P}_3 = \frac{2(e^{\beta}\times 1)}{(e^{\beta}+5)^2}$$

Efron It is better as it reflects all patients at risk before and after the failure

$\Rightarrow$ In the previous example, $\mathbb{P}_1$ and $\mathbb{P}_3$ becomes

$$\mathbb{P}_1 = \frac{e^{\beta}}{(6e^{\beta}+4)}\frac{e^{\beta}}{(0.5e^{\beta}+0.5e^{\beta}+4e^{\beta}+4)}$$

and

$$\mathbb{P}_3 = \frac{e^{\beta}}{(e^{\beta}+5)}\frac{1}{(0.5+0.5e^{\beta}+3)}$$

with the weight 0.5 reflecting that each of the 2 patients has a chance of $1/2$ of being in the second denominator since 1 of them would have been the first failure

Left truncated data

- Consider the data of Table 12 with left truncation information

e.g. A patient can be diagnosed before entering a trial (i.e. backwards recurrence times is $\neq 0$)

Note 1 The standard way to compare the 2 groups is to ignore "back times"

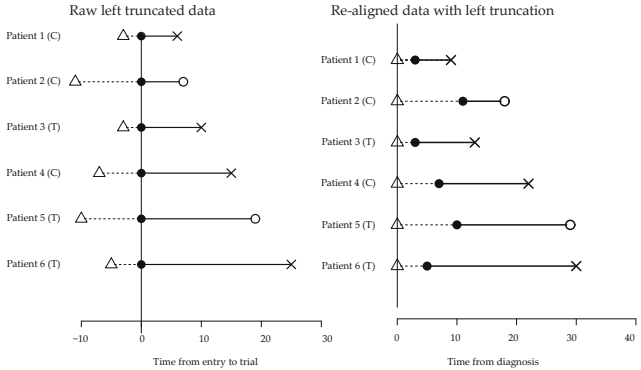
$\Rightarrow$ Nothing wrong (i.e. no bias) in that way to proceed but starting from diagnosis could be of interest

Note 2 To account for backwards recurrence times, one can re-configure the data so that they start at 0

Table: Survival left truncated data

Patient	Survtime	Censor	Group	Back time
1	6	1	C	-3
2	7	0	C	-11
3	10	1	T	-3
4	15	1	C	-7
5	19	0	T	-10
6	25	1	T	-5

Left truncation and re-configured data



- In that case, estimation results are similar for the two data sets

$\Rightarrow$ No statistical difference between "C" and "T" (but n is too small)

— Raw data : $\widehat{\beta} = -1.33 (\sigma_{\widehat{\beta}} = 1.25)$ with $LR = 1.21 (p = 0.271)$

— Re-configured data : $\widehat{\beta} = -1.07 (\sigma_{\widehat{\beta}} = 1.24)$ with $LR = 0.81 (p = 0.368)$

Categorical and Continuous Covariates

- All covariates considered until now are dummy variables

Note An exception is the confounder "genotype" that is categorical

$$g \in \{1, 2\}$$

but can easily be transformed to $\{0, 1\}$ as it is dichotomous

- More generally one can encode categorical variables with dummies

e.g. If we have a 3-level variable we need : "Ba (x_1), Ma (x_2), no-diploma (x_3)"

$\Rightarrow$ If "Ba" is the reference, then $x_1 = 1$, $x_2 = x_3 = 0$

$\Rightarrow$ An individual without any diploma implies $x_1 = x_2 = 0$ and $x_3 = 1$

- Continuous variables are also frequent and have to be considered

e.g. income, age, etc.

The Cox model with categorical and continuous covariates

- For a set of k covariates (categorical or/and continuous) the model is

$$\log(\psi_i) = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{ki}\beta_k = x_i'\beta$$

- For the covariate x_j, β_j is the log hazard ratio for the effect of that parameter on survival, adjusting for the other covariates
- For continuous covariates, it represents the effect of a unit change in the covariate
- For dummy covariates, it represents the effect of the corresponding level as compared to the reference

Note 1 As for logistic regression, a variable can enter non-linearly the model

Note 2 Interaction terms can be introduced

Note 3 At this stage, all covariate are assumed to be fixed in time

Note 4 This model differs from the logistic model as there is no intercept term : if there were one, it would cancel out just as $h_0(t)$ canceled out

Example of Cox model estimation with categorical and continuous covariates

- Consider artificial survival data with two covariates : age and diploma

$\Rightarrow$ individual at risk can loose their job

- Ages are between 40 and 80 at random
- We set the diploma variable so that there are 20 of each 3 categories
- We assume an exponential distribution with parameter as follows
  - We set the $\log$-rate parameter to have baseline -4.5
  - The diploma variable take the values 1 and 2 for "Ba" and "No diploma" when compared to "Ma"
  - We let "age" decrease the $\log$ rate by 0.05 per year
- We do not introduce censoring in the data set and $n = 60$

Example of Cox model estimation with categorical and continuous covariates

- When applying the Cox model we obtain the following estimates

$$\widehat{\beta}_{Ba} = 1.151,\ (\sigma_{\widehat{\beta}_{Ba}} = 0.368), \quad z = 3.113\ (p = 0.00173)$$

and

$$\widehat{\beta}_{No} = 2.499,\ (\sigma_{\widehat{\beta}_{No}} = 0.429), \quad z = 5.820\ (p = 0.00000)$$

and

$$\widehat{\beta}_{age} = -0.078,\ (\sigma_{\widehat{\beta}_{age}} = 0.014), \quad z = 5.385\ (p = 0.00000)$$

$\Rightarrow$ Estimates of log hazard ratios are close to the true values (1, 2 and 0.05)

- When looking at exponential coefficient, $\exp(\beta)$, we conclude that
  - Individuals with Bachelor degree have $\exp(\beta_{Ba}) = 3.16$ times the risk of being fired as do subject with Ma degree
  - Individuals without diploma have $\exp(\beta_{No}) = 12.17$ times the risk of being fired as do subject with Ma degree

Note The z statistics is a generalizations of the 2-group comparison Wald tests

Nested models

- When comparing models we have to determine whether that are nested
- Here is an illustration of nested models in terms of covariates
  - Model A : “Age”
  - Model B : “Employment”
  - Model C : “Age” + “Employment”

$\Rightarrow$ Model A is nested in Model C as well as model B

- To test for the presence of nested models we can compute LR tests

Note Models A and B are not nested and requires specific testing procedures

Example of nested models

- Consider the data on therapies to aid smokers to quit (Chapter 1)
- In this study, "Age" and "Employment" have 4 and 3 levels
  - Age : "21-34", "35-49", "50-64" and "65+"
  - Employment : "ft" (full-time), "other" and "pt" (part-time)

$\Rightarrow$ By default we choose the first level as the reference level

- Estimation of the Cox model on model A, B and C

	coef	exp(coef)	se(coef)	z	p
LR : 12.2 ($p = 0.006$)			Model A		
age35-49	0.0293	1.030	0.309	0.0947	0.920
age50-64	-0.7914	0.453	0.336	-2.3551	0.019
age65+	-0.3173	0.728	0.444	-0.7153	0.470
LR : 2.06 ($p = 0.357$)			Model B		
other	0.198	1.22	0.237	0.836	0.40
pt	0.450	1.57	0.323	1.394	0.16
LR : 16.8 ($p = 0.005$)			Model C		
age35-49	-0.130	0.878	0.321	-0.404	0.6900
age50-64	-1.024	0.359	0.359	-2.856	0.0043
age65+	-0.782	0.457	0.505	-1.551	0.1200
other	0.526	1.692	0.275	1.913	0.0560
pt	0.500	1.649	0.332	1.508	0.1300

Example of nested models

- From the Wald test (z) for Model C we see that some levels are significant

  e.g. The "50-64" age group has a lower hazard when compared to the reference "21-34" with $\widehat{\beta} = -1.024$

  e.g. The "other" employment group has higher hazard when compared to the reference "ft" with $\widehat{\beta} = 0.526$

- However, we cannot easily see whether "Age" or "Employment" should be part of the model

$\Rightarrow$ We assess this issue using (partial) likelihood ratio tests based on
$\ell(\widehat{\beta})$ Model A : -380.043, Model B : -385.123, Model C : -377.759

LR : A|C $2(\ell(\widehat{\beta}_C) - \ell(\widehat{\beta}_A) = 4.567$ compare to $\chi^2_{\nu=5-3}$ which leads to $p = 0.1019$

$\Rightarrow$ "Age" is not significant when "Employment" is included in the model

LR : B|C $2(\ell(\widehat{\beta}_C) - \ell(\widehat{\beta}_B) = 14.727$ compare to $\chi^2_{\nu=5-2}$ which leads to $p = 0.0020$

$\Rightarrow$ "Employment" is significant when "Age" is included in the model

Example of nested models

- These results raise the question of including "Age" in model A

$\Rightarrow$ To test this hypothesis we consider the null model N

$$\ell(\widehat{\beta}_N) = -386.153$$

free of any covariate

LR : N|A $2(\ell(\widehat{\beta}_A) - \ell(\widehat{\beta}_N) = 12.220$ compare to $\chi^2_{\nu=3-0}$ which leads to $p = 0.0066$

$\Rightarrow$ "Age" is significant when included in the model N

When a large number of potential factors can enter the model

$\Rightarrow$ The forward stepwise model selection

Step 1 fit univariate models (1 for each covariate) and retain the one with the smallest p-value

Step 2 apply Step 1 again but with the selected covariate included

Step 3 continue until no additional covariate has a p-value less than a pre-defined threshold (e.g. 5%)

$\Rightarrow$ The backward stepwise model selection

Step 1 fit a model with all covariates

Step 2 remove one by one the covariates, each time removing the one with the largest p-value

Step 3 continue the procedure until the p-values are all below a pre-defined threshold (e.g. 5%)

- The stepwise approach can be automatized but has 2 main drawbacks

  — Due to multiple comparisons, the p-values produced from one stage to the next are misleading

Note Corrections like the one of Bonferroni exist

  — Also, p-values are only valid for nested models and hence this approach is not recommended for non-nested models

Non-nested models and criterion based selection

- Information criteria apply to partial log likelihood
- We discuss some examples based on the so-called AIC

$$AIC = -2\ell(\widehat{\beta}) + 2k$$

where k is the number of parameters in the model

- One can view the AIC as balancing two quantities
  - — The goodness of fit $-2\ell(\widehat{\beta})$ (smaller for models that fit the data well)
  - — The complexity measure that enter the criterion as a penalty term $2k$
- Applying the AIC to the previous model selection issue we obtain

$\ell(\widehat{\beta})$ Model A : 766.086, Model B : 774.246, Model C : 765.519

$\Rightarrow$ The model C is the one that minimizes the AIC and offers the best fit

Note The BIC (or SIC) also applies to survival analysis

$$BIC = -2\ell(\widehat{\beta}) + k\log(n)$$

and as it penalizes by a factor of $\log(n)$, it will tend to select models with fewer parameters as compared to AIC

Information criterion and the stepwise approach

- We can implement the backward stepwise procedure with the AIC
- Let consider additional covariates for the smokers therapies
  - — “yearsSmoking”+“levelSmoking”+“priorAttempts”+“longestNoSmoke”
  - \+ “gender”+ “morphotype”+ “age”+ “employment”

Note 1 (+) & (-) show the effect on AIC of adding or removing the covariate

Note 2 Covariates are listed in order from the one which, when removed, yields the greatest AIC reduction to the smallest reduction

Information criterion and the stepwise approach

- When starting the procedure, all covariates are there (AIC = 770.2)
  - ⇒ "(-) morpho" is at the top of the list and will be removed first
- Intermediate results are unreported but proceed in the same way
- At final step (AIC = 758.42) and all per-covariate are above 758.42
  - ⇒ The sign (-) remains for employment & age and reveal that removing them would be detrimental
  - ⇒ At the opposite, variables for which a "(+)" appears indicate that adding would deteriorate the fit of the model

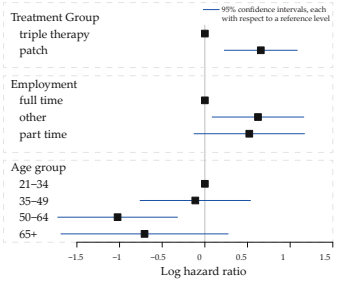
Sign	Covariate	Level	AIC	Sign	Covariate	Level	AIC
Step 1			770.2	Final Step			758.42
-	morpho	3	766.98		<none>		758.42
-	years	1	768.20	+	longest	1	759.10
-	gender	1	768.20	-	employment	2	760.31
-	prior	1	768.24	+	years	1	760.34
-	level	1	768.47	+	gender	1	760.39
-	longest	1	769.04	+	prior	1	760.40
	none		770.20	+	level	1	760.41
-	employment	2	772.45	+	morpho	3	761.53
-	age	3	774.11	-	age	3	767.24

Forest plot

Final model	coef	exp(coef)	se(coef)	z	p
grppatchOnly	0.656	1.928	0.220	2.986	0.0028
employmentother	0.623	1.865	0.276	2.254	0.0240
employmentpt	0.521	1.684	0.332	1.570	0.1200
ageGroup435-49	-0.112	0.894	0.322	-0.348	0.7300
ageGroup450-64	-1.023	0.359	0.360	-2.845	0.0044
ageGroup465+	-0.707	0.493	0.502	-1.410	0.1600

- The Forest plot offers an alternative representation :

e.g. 1 triple therapy is better than the patch alone
e.g. 2 subjects with full-time work have a better success rate than others
e.g. 3 the upper age groups have better results than younger patients



Smooth estimates of continuous covariates

- For continuous covariates, the relationship with the log-hazard can be

... linear, quadratic, or of any other nonlinear nature

e.g. in the previous study, the age has been split into 4 groups and

... the forest plot reveals different effects and hence nonlinearities

$\Rightarrow$ An alternative way to capture this nonlinearity is via pieces of

... polynomial functions (Splines) that are stitched to form a smooth curve

- The points where these pieces are joined are called "knots"

... and a crucial issue is to determined their locations

$\Rightarrow$ The Splines enter the penalized partial likelihood via a penalty term

$$\mathcal{P}(\beta, \omega) = \ell(\beta, \omega) - g(\omega, \theta)$$

with ω the set of constrained parameters and θ some tuning parameters

Penalized Cox model and Spline fit

- Splines with many knots increase the complexity of the likelihood

... but also improve the goodness of fit

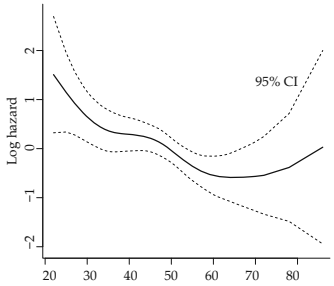
$\Rightarrow$ $\mathcal{P}(\beta, \omega)$, when maximized, balances goodness of fit against complexity

e.g. When plotting the penalized spline fit from the Cox model we observe

— a decreasing relationship with age with a slight upward turn after age 65

— but for most of the part, the effect seems not significant

Figure: Splines



Penalized Cox model and Spline fit

- The penalized Cox model estimation results are reported below

	coef	exp(coef)	se(coef)	χ^2	ν	p
grppatchOnly	0.651	0.221	0.219	8.67	1.00	0.0032
employmentother	0.633	0.277	0.275	5.21	1.00	0.0220
employmentpt	0.570	0.340	0.333	2.81	1.00	0.0940
pspline(age,linear)	-0.034	0.010	0.010	11.07	1.00	0.0009
pspline(age,nonlinear)				4.20	3.08	0.2500

- For the 3 first factors the coefficient are stable as compared to S97
- The Splines are decomposed in two parts : linear and nonlinear
  - the linear one is highly significant
  - the nonlinear one is not significant (probably because the data set is sparse)

Martingale residuals

- Assessing goodness of fit using residuals also applies to survival analysis
- Residual analysis essentially relies on graphical analysis

$\Rightarrow$ Typically, residuals are plotted versus some quantity

- To construct the residuals sequence, we compare the censoring indicator

$$\delta_i$$

to the expected value of the indicator under the Cox model

$\Rightarrow$ In absence of time dependent covariates and for right-censored data

$$\widehat{m}_i = \delta_i - \widehat{H}_0(t_i)\exp(x_i'\widehat{\beta})$$

- These Martingale residuals range in value from $-\infty$ to 1 and $\mathbb{E}(\widehat{m}_i) = 0$
- However these residuals can be asymmetric and hence cannot be used as a measure of goodness of fit

Deviance residuals

- An alternative is the so-called deviance residual defined as

$$\widehat{d}_i = \text{sign}(\widehat{m}_i)\Big(-2\big(\widehat{m}_i + \delta_i \log(\delta_i - \widehat{m}_i)\big)\Big)^{1/2}$$

- d_i residuals are symmetrically distributed with $\mathbb{E}(\widehat{d}_i)$

Note 1 The sum of squares of $\widehat{d}_i$ is the value of the partial likelihood ratio test

- While their properties might seem preferable to those of $\widehat{m}_i$, only $\widehat{m}_i$ have the property of showing us the functional form of a covariate

$\Rightarrow$ In practice, the martingale residuals are more useful

Note 2 Other types of residuals will be discussed later

Example : Martingale versus deviance residuals

- Consider again the Cox model for smoking therapies data
- As discussed earlier, the null model (N) is the one without covariates

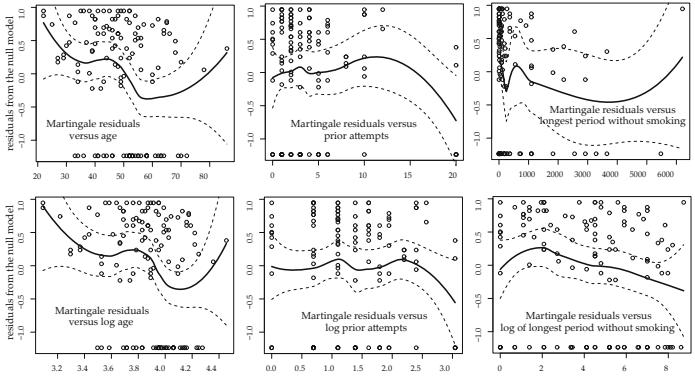
$\Rightarrow$ We may plot $\widehat{m}_i$ against continuous covariates to get a preliminary assessment of which of them should be in the model

Note 1 We also include the log of covariates and use a LOESS curve to identify patterns

Note 2 LOESS (LOcally Estimated Scatterplot Smoothing) is a nonparametric regression based on the nearest neighbor method

Note 3 The 95% confidence intervals for the LOESS curve are also reported

Example : Martingale versus deviance residuals



- For the raw covariates we observe strong non-linearities

  e.g. For "age", we find something similar to Figure 1 (Spline fit)

  $\Rightarrow$ This null model residual based approach is an alternative way to identify nonlinearity
- For the log-transformed covariates we observe less non-linearities

  e.g. For "LongestNoSmoke", the log seems sufficient to remove the non-linearity

Example : Martingale versus deviance residuals

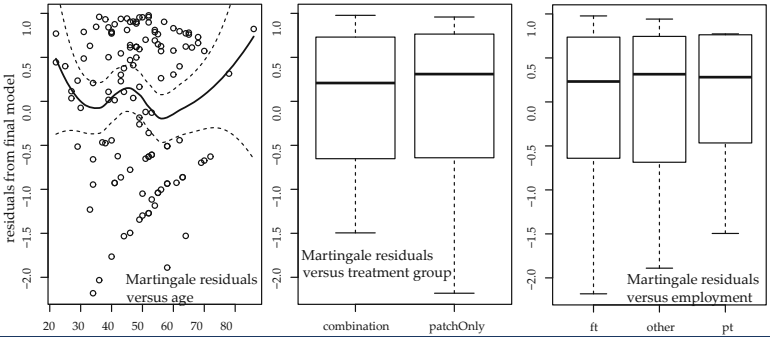
- We apply the stepwise approach with the log of "LongestNoSmoke"

⇒ The results are unchanged (only "age" and "employment" are retained)

- We compute the final model residuals and obtain the following plots

⇒ Some non-linearity remains for "age" albeit less than for the null model

- The residual distributions of both "group" and "employ" are reasonably comparable, indicating that these variables are modeled successfully



Jackknife residuals

- Some subject may have a huge influence on the parameter estimates

$\Rightarrow$ As this may indicate a problem with the data

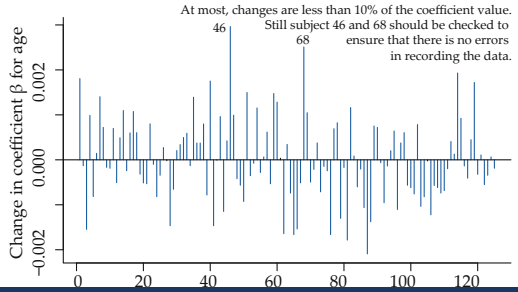
... we need tools that can identify those individuals

- The Jackknife residuals are computed as the difference in the value of

$$\widehat{\beta}$$

when all data are used and when an individual is deleted from the data

$\Rightarrow$ Then, we can plot the change in coefficients for each subject



Log cumulative hazard plots

- When comparing survival times between two groups

... the proportional hazards assumption is of importance

$$S_1(t) = \big(S_0(t)\big)^{\exp(\beta)}$$

with $\exp(\beta)$ the proportional hazards constant

$\Rightarrow$ This is the foundation of Lehman alternatives and the Cox model

- The log-transformation gives

$$\log(S_1(t)) = \exp(\beta)\log(S_0(t))$$

with all logs being negative as survival functions are less than 1

- $g(u) = \log(-\log(u))$ changes the range of u from $(0, 1)$ to $(-\infty, \infty)$

$\Rightarrow$ The so-called log cumulative hazard plot, that is a plot of

$$g(S_1(t)) \text{ and } g(S_0(t)) \text{ versus } \log(t)$$

should lead to parallel curves separated by β if the assumption is correct

Example of log cumulative hazard plots

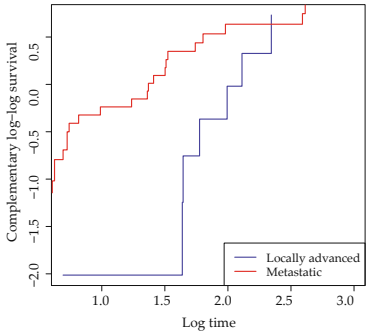
- Consider the pancreatic cancer data (see in Chapter 1)

Recall We performed the Prentice-modification test and found a stronger group difference than did the log-rank test

$\Rightarrow$ As this test places higher weight on earlier survival times it suggests non-proportional hazards

- This is confirmed by the log cumulative hazard plot

Note However, statistical inference is unavailable and this approach is limited



Schoenfeld residuals

- Schoenfeld residuals can assess the proportional hazards assumption more rigorously
- To compute them, let start from the partial log-likelihood

$$\ell(\beta) = \sum_{i \in D} \left(\log(\psi_i) - \log\Big(\sum_{k \in R_i} \psi_k \Big) \right) = \sum_{i \in D} \left(x_i \beta - \log\Big(\sum_{k \in R_i} \exp(x_k \beta) \Big) \right)$$

and its derivative (the score function)

$$\ell(\beta)' = \sum_{i \in D} \Big(x_i - \sum_{k \in R_i} x_k p(\beta, x_k) \Big), \quad p(\beta, x_k) = \exp(x_k \beta) \Big(\sum_{j \in R_k} \exp(x_j \beta) \Big)^{-1}$$

where the second term can be viewed as the weighted expected value $\mathbb{E}(X_i) = \bar{x}(t_i)$

- The Schoenfeld residuals are the individual terms of the score

$$\hat{r}_i = x_i - \sum_{k \in R_i} x_k p(\beta, x_k) = x_i - \bar{x}(t_i)$$

- A plot of $\hat{r}_i$ versus x_i will yield a pattern of points

$\Rightarrow$ They are centered on 0 if the proportional hazards assumption is correct

Example of Schoenfeld residuals

- Consider the artificial data of S12 ($\hat{\beta} = -1.32$) and compute the weights

t_i	n_{0i}	n_{1i}	$p(\beta, x_k = 0)$	$p(\beta, x_k = 1)$	Grp
6	3	3	$1/(3 + 3e^{\hat{\beta}})$	$e^{\hat{\beta}}/(3 + 3e^{\hat{\beta}})$	C
10	1	3	$1/(1 + 3e^{\hat{\beta}})$	$e^{\hat{\beta}}/(1 + 3e^{\hat{\beta}})$	T
15	1	2	$1/(1 + 2e^{\hat{\beta}})$	$e^{\hat{\beta}}/(1 + 2e^{\hat{\beta}})$	C
25	0	1	$1/e^{\hat{\beta}}$	$e^{\hat{\beta}}/e^{\hat{\beta}} = 1$	T

- It remains to compute $\mathbb{E}(X_i)$ and $\hat{r}_i$ which for $t_i = 6$ gives

Example of Schoenfeld residuals

- Consider the artificial data of S12 ($\hat{\beta} = -1.32$) and compute the weights

t_i	n_{0i}	n_{1i}	$p(\beta, x_k = 0)$	$p(\beta, x_k = 1)$	Grp
6	3	3	$1/(3 + 3e^{\hat{\beta}})$	$e^{\hat{\beta}}/(3 + 3e^{\hat{\beta}})$	C
10	1	3	$1/(1 + 3e^{\hat{\beta}})$	$e^{\hat{\beta}}/(1 + 3e^{\hat{\beta}})$	T
15	1	2	$1/(1 + 2e^{\hat{\beta}})$	$e^{\hat{\beta}}/(1 + 2e^{\hat{\beta}})$	C
25	0	1	$1/e^{\hat{\beta}}$	$e^{\hat{\beta}}/e^{\hat{\beta}} = 1$	T

- It remains to compute $\mathbb{E}(X_i)$ and $\hat{r}_i$ which for $t_i = 6$ gives

$$\mathbb{E}(X_i) = 3 \times 0 \times 1/(3 + 3e^{\hat{\beta}}) + 3 \times 1 \times e^{\hat{\beta}}/(3 + 3e^{\hat{\beta}}) = 0.2098 \Rightarrow \hat{r}_i = 0 - 0.2098$$

Example of Schoenfeld residuals

- Consider the artificial data of S12 ($\hat{\beta} = -1.32$) and compute the weights

t_i	n_{0i}	n_{1i}	$p(\beta, x_k = 0)$	$p(\beta, x_k = 1)$	Grp
6	3	3	$1/(3+3e^{\hat{\beta}})$	$e^{\hat{\beta}}/(3+3e^{\hat{\beta}})$	C
10	1	3	$1/(1+3e^{\hat{\beta}})$	$e^{\hat{\beta}}/(1+3e^{\hat{\beta}})$	T
15	1	2	$1/(1+2e^{\hat{\beta}})$	$e^{\hat{\beta}}/(1+2e^{\hat{\beta}})$	C
25	0	1	$1/e^{\hat{\beta}}$	$e^{\hat{\beta}}/e^{\hat{\beta}} = 1$	T

- It remains to compute $\mathbb{E}(X_i)$ and $\hat{r}_i$ which for $t_i = 6$ gives

$$\mathbb{E}(X_i) = 3 \times 0 \times 1/(3+3e^{\hat{\beta}}) + 3 \times 1 \times e^{\hat{\beta}}/(3+3e^{\hat{\beta}}) = 0.2098 \Rightarrow \hat{r}_i = 0 - 0.2098$$

- For $t_i = 10$: $\mathbb{E}(X_i) = 1 \times 0 \times 1/(1+3e^{\hat{\beta}}) + 3 \times 1 \times e^{\hat{\beta}}/(1+3e^{\hat{\beta}}) = 0.4434$

$$\Rightarrow \hat{r}_i = 1 - 0.4434 = 0.5566$$

- For $t_i = 15$: $\mathbb{E}(X_i) = 1 \times 0 \times 1/(1+2e^{\hat{\beta}}) + 2 \times 1 \times e^{\hat{\beta}}/(1+2e^{\hat{\beta}}) = 0.3468$

$$\Rightarrow \hat{r}_i = 0 - 0.3468 = -0.3468$$

- For $t_i = 25$ we have $\mathbb{E}(X_i) = 1$

$$\Rightarrow \hat{r}_i = 1 - 1 = 0$$

Grambsch and Therneau residuals

- They propose to scale each residual by an estimate of its variance

$$\widehat{r}_i^* = \widehat{r}_i \times d \times \mathbb{V}(\widehat{\beta})$$

where d is the total number of death

- Then, Grambsch and Therneau show that if hazards are non proportional

$$\mathbb{E}(r_i^*) \approx \beta + \beta(t)$$

i.e. a survival time dependent β (unknown) enter the $\mathbb{E}(\hat{r}_i^*)$ whereas

$$\mathbb{E}(r_i^*) = \beta$$

in presence of proportional hazards

$\Rightarrow$ $\beta(t)$ can be approximated by

$$\widehat{\beta}(t) \approx \widehat{r}_i^* - \widehat{\beta}$$

where $\widehat{\beta}$ is estimated from the Cox model

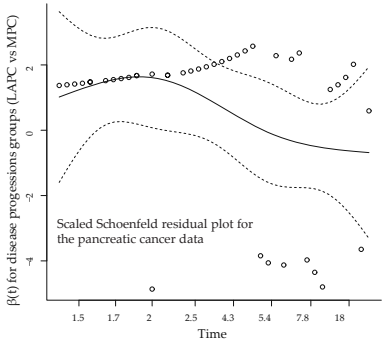
Note Statistical inference is now possible to test $H_0 : \beta(t) = 0$

Example for Grambsch and Therneau residuals

- We compute $\widehat{\beta}(t)$ for the pancreatic cancer data and plot it versus time

Note 1 we also compute the LOESS curve and its 95% confidence intervals

Note 2 the time axis is scaled to match the Kaplan-Meier-transformed time



- The curve reveals a slight increase, followed by a steady decline

Note 3 Zero seems to be almost always in the confidence intervals

Example for Grambsch and Therneau residuals

- A more formal test can be obtained by fitting as straight line to $\widehat{r}_i^*$
- This score-type test statistic, denoted $\widehat{\rho} \sim \chi^2_1$, gives

$$\widehat{\rho} = -0.328, \quad p = 0.0496$$

$\Rightarrow$ we reject the null of a constant β (i.e. we reject the proportional hazards)

- The way we defined the time axis matters (Kaplan-Meier-transformed time here)

e.g. If we consider time ordered by the ranks survival times we obtain

$$\widehat{\rho} = -0.330, \quad p = 0.0486$$

$\Rightarrow$ very similar results

e.g. If we consider the untransformed time line we obtain

$$\widehat{\rho} = -0.197, \quad p = 0.2390$$

$\Rightarrow$ here we cannot reject the null of proportional hazards

Note This latter approach is not to be preferred when the failure times are sparse and not uniformly spaced over time

What are time dependent covariates ?

- The partial likelihood theory assumes that covariates are time invariant

$\Rightarrow$ The value of z at $t = 0$ is the same at any $t_i > 0$

- In some cases this assumption is unrealistic

e.g. In credit scoring analysis, the employment status is likely to change

e.g. In job market analysis, the skills are likely to evolve

$\Rightarrow$ Time dependent covariates require special measures to obtain valid parameter estimates

Impact of time dependent covariates

- Unfortunately, we cannot predict survival using future covariate values
- This deceptively principle can ensnare even experienced research

$\Rightarrow$ We illustrate this with the following example :

e.g. consider data on patients enrolled in a transplant program

- Here are the results of the survival study :

	coef	exp(coef)	se(coef)	z-test	p
transplant	-1.71711	0.17958	0.27853	-6.165	7.05e-10
age	0.05889	1.06065	0.01505	3.913	9.12e-05
surgery	-0.41902	0.65769	0.37118	-1.129	0.259

$\Rightarrow$ It seems that heart transplanted patients live longer than others

- The covariate "transplant" equals 1 for transplanted patients

$\Rightarrow$ The issue is that "transplant" is time dependent as patients in a transplant program have to live long enough to be transplanted

$\Rightarrow$ It only shows that patients who live long enough to receive a transplant have longer lives than patients who do not live as long (tautology)

Landmark time

- In that particular case, a simple fix is to define a landmark time τ
- It divide patients into two groups: intervention and comparison groups

Intervention those who received the intervention prior to τ

Comparison those who did not received the intervention prior to τ

- If only patients who survive up to τ are included

and all patients remain in their assigned group, this method is valid

Note Hence, patients transplanted after τ remain in the comparison group

$\Rightarrow$ the comparison group could be renamed "no transplant within τ days"

Example of landmark time

- If we set $\tau = 30$ days, 79 of the 103 patients lived this long
- Of these 79 patients, 33 had a transplant before τ and 46 did not
- Of these 46 patients, 30 subsequently had a transplant

Note we still count them in the comparison group

$\Rightarrow$ we have hence created a new variable "transplant30" which has a fixed value for all patients in the set of 30-day survivors

- Here are the valid results of survival study :

	coef	exp(coef)	se(coef)	z-test	p
transplant30	-0.04214	0.95874	0.28377	-0.148	0.8820
age	0.03720	1.03790	0.01714	2.170	0.0300
surgery	-0.81966	0.44058	0.41297	-1.985	0.0472

- The "transplant" covariate is no longer significant

Note However, one could discuss the choice of the landmark τ

Beyond the landmark approach

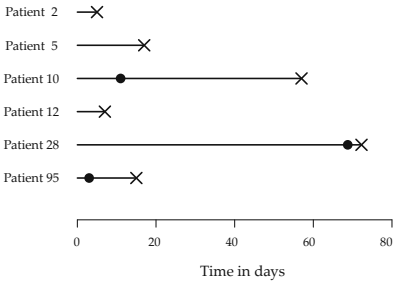
- Unfortunately there is no clear way to select the landmark τ

$\Rightarrow$ we prefer another approach based on adjustments of the Cox model

- Let consider a subset of 6 patients to illustrate this approach
- 3 of them received a transplant and 3 of them did not

id	wait.time	futime	fustat	transplant
2	–	5	1	0
5	–	17	1	0
10	11	57	1	1
12	–	7	1	0
28	70	71	1	1
95	1	15	1	1

futime : following-up (failure) time
fustat : 0 if censored, 1 otherwise
● and waiting time : time of transplant



Modified partial likelihood

- We first model incorrectly the data

	coef	exp(coef)	se(coef)	z-test	p
transplant	-1.6878	0.1849	1.1718	-1.44	0.150

- To correct the model we allow the contributions of each subject to change from one failure time to the next

$\Rightarrow$ The hazard function is now given by

$$h(t) = h_0(t)e^{x_k(t_i)\beta}$$

with $x_k(t_i)$ the time-varying covariate for the kth subject at time t_i

- This leads to the modified partial likelihood

$$\mathcal{L}(\beta) = \prod_{i=1}^{D} \psi_{ii} \Big(\sum_{k \in R_i} \psi_{ki} \Big)^{-1}$$

with $\psi_{ki} = e^{x_k(t_i)\beta}$

- In the fixed-time case we were able, as time passes, to successively delete ψ_i for subject that failed at that time
- We here have to recalculate the entire denominator at each failure time

Example of modified partial likelihood computation

- Let compute $\mathcal{L}(\beta)$ for the six patients (labeled 2, 5, 10, 12, 28, 95)

$\mathcal{L}_1(\beta)$ P2 fails at $t = 5$, all 6 being at risk, the P95 being the only 1 transplanted

$$\mathcal{L}_1(\beta) = \frac{1}{5 + e^{\beta}}$$

Example of modified partial likelihood computation

- Let compute $\mathcal{L}(\beta)$ for the six patients (labeled 2, 5, 10, 12, 28, 95)

$\mathcal{L}_1(\beta)$ P2 fails at $t = 5$, all 6 being at risk, the P95 being the only 1 transplanted

$$\mathcal{L}_1(\beta) = \frac{1}{5 + e^{\beta}}$$

$\mathcal{L}_2(\beta)$ P12 fails at $t = 7$, 5 being at risk, still 1 patient being transplanted

$$\mathcal{L}_2(\beta) = \frac{1}{4 + e^{\beta}}$$

Example of modified partial likelihood computation

- Let compute $\mathcal{L}(\beta)$ for the six patients (labeled 2, 5, 10, 12, 28, 95)

$\mathcal{L}_1(\beta)$ P2 fails at $t = 5$, all 6 being at risk, the P95 being the only 1 transplanted

$$\mathcal{L}_1(\beta) = \frac{1}{5 + e^{\beta}}$$

$\mathcal{L}_2(\beta)$ P12 fails at $t = 7$, 5 being at risk, still 1 patient being transplanted

$$\mathcal{L}_2(\beta) = \frac{1}{4 + e^{\beta}}$$

$\mathcal{L}_3(\beta)$ P95 fails at $t = 15$, 4 being at risk, but the P10 "transplant" status has switched to 1

$$\mathcal{L}_3(\beta) = \frac{e^{\beta}}{2 + 2e^{\beta}}$$

Example of modified partial likelihood computation

- Let compute $\mathcal{L}(\beta)$ for the six patients (labeled 2, 5, 10, 12, 28, 95)

$\mathcal{L}_1(\beta)$ P2 fails at $t = 5$, all 6 being at risk, the P95 being the only 1 transplanted

$$\mathcal{L}_1(\beta) = \frac{1}{5 + e^{\beta}}$$

$\mathcal{L}_2(\beta)$ P12 fails at $t = 7$, 5 being at risk, still 1 patient being transplanted

$$\mathcal{L}_2(\beta) = \frac{1}{4 + e^{\beta}}$$

$\mathcal{L}_3(\beta)$ P95 fails at $t = 15$, 4 being at risk, but the P10 "transplant" status has switched to 1

$$\mathcal{L}_3(\beta) = \frac{e^{\beta}}{2 + 2e^{\beta}}$$

$\mathcal{L}_4(\beta)$ P5 fails at $t = 17$, 3 being at risk, still 2 patients being transplanted

$$\mathcal{L}_4(\beta) = \frac{1}{2 + e^{\beta}}$$

Example of modified partial likelihood computation

- Let compute $\mathcal{L}(\beta)$ for the six patients (labeled 2, 5, 10, 12, 28, 95)

$\mathcal{L}_1(\beta)$ P2 fails at $t = 5$, all 6 being at risk, the P95 being the only 1 transplanted

$$\mathcal{L}_1(\beta) = \frac{1}{5 + e^{\beta}}$$

$\mathcal{L}_2(\beta)$ P12 fails at $t = 7$, 5 being at risk, still 1 patient being transplanted

$$\mathcal{L}_2(\beta) = \frac{1}{4 + e^{\beta}}$$

$\mathcal{L}_3(\beta)$ P95 fails at $t = 15$, 4 being at risk, but the P10 "transplant" status has switched to 1

$$\mathcal{L}_3(\beta) = \frac{e^{\beta}}{2 + 2e^{\beta}}$$

$\mathcal{L}_4(\beta)$ P5 fails at $t = 17$, 3 being at risk, still 2 patients being transplanted

$$\mathcal{L}_4(\beta) = \frac{1}{2 + e^{\beta}}$$

$\mathcal{L}_5(\beta)$ P10 fails at $t = 57$, 2 being at risk, still 2 patients being transplanted

$$\mathcal{L}_5(\beta) = e^{\beta}(1 + e^{\beta})^{-1}$$

Example of modified partial likelihood computation

- Let compute $\mathcal{L}(\beta)$ for the six patients (labeled 2, 5, 10, 12, 28, 95)

$\mathcal{L}_1(\beta)$ P2 fails at $t = 5$, all 6 being at risk, the P95 being the only 1 transplanted

$$\mathcal{L}_1(\beta) = \frac{1}{5 + e^{\beta}}$$

$\mathcal{L}_2(\beta)$ P12 fails at $t = 7$, 5 being at risk, still 1 patient being transplanted

$$\mathcal{L}_2(\beta) = \frac{1}{4 + e^{\beta}}$$

$\mathcal{L}_3(\beta)$ P95 fails at $t = 15$, 4 being at risk, but the P10 "transplant" status has switched to 1

$$\mathcal{L}_3(\beta) = \frac{e^{\beta}}{2 + 2e^{\beta}}$$

$\mathcal{L}_4(\beta)$ P5 fails at $t = 17$, 3 being at risk, still 2 patients being transplanted

$$\mathcal{L}_4(\beta) = \frac{1}{2 + e^{\beta}}$$

$\mathcal{L}_5(\beta)$ P10 fails at $t = 57$, 2 being at risk, still 2 patients being transplanted

$$\mathcal{L}_5(\beta) = e^{\beta}(1 + e^{\beta})^{-1}$$

$\mathcal{L}_6(\beta)$ P28 is the last to fail ($t = 71$), just after having been transplanted $\mathcal{L}_6(\beta) = e^{\beta}/e^{\beta} = 1$

Example of modified partial likelihood computation

- Overall, the modified partial likelihood is

$$\mathcal{L}(\beta) = \frac{1}{5+e^{\beta}} \times \frac{1}{4+e^{\beta}} \times \frac{e^{\beta}}{2+2e^{\beta}} \times \frac{1}{2+e^{\beta}} \times \frac{e^{\beta}}{1+e^{\beta}} \times 1$$

- On the numerical side, $\mathcal{L}(\beta)$ is based on the start-stop format
  - It divides the time data for patients with a time-varying covariate

  e.g. As P10 was a non-transplant patient until day 11, its future as a non-transplant patient is unknown

  $\Rightarrow$ we censor that portion of the patient's life experience at $t = 11$:

  start: $t = 0$, stop: $t = 11$

  $\Rightarrow$ we start a new record of P10 (which is left-truncated at $t = 11$)

  start: $t = 11$, stop: $t = 57$

  - For our subset of 6 patient it results in new lines in the database

P#	start	stop	death	transpl
2	0	5	1	0
5	0	17	1	0
10	0	11	0	0
10	11	57	1	1
12	0	7	1	0
28	0	70	0	0
28	70	71	1	1
95	0	1	0	0

Example of modified partial likelihood computation

- Once the data are in this start-stop format the Cox model applies
- For our subset of 6 patient the conclusions remain unchanged

	coef	exp(coef)	se(coef)	z-test	p
transplant	0.2846	1.3292	0.9609	0.296	0.767

- When considering the whole data set and all covariates we obtain

	coef	exp(coef)	se(coef)	z-test	p
transplant	0.01405	1.01415	0.30822	0.046	0.9636
surgery	-0.77326	0.46150	0.35966	-2.150	0.0316
age	0.03055	1.03103	0.01389	2.199	0.0279

- As with the landmark analysis we confirm that there is no evidence that receiving a heart transplant increases survival

Predictable time dependent variables

- An alternative way to model non-proportional hazard is to allows for

$$\beta = \beta(t)$$

for a particular covariate

- If there is only one covariate we have

$$h(t) = h_0 e^{x_k \beta(t)}$$

- Characterizing the functional form of $\beta(t)$ is challenging

$\Rightarrow$ A way to proceed is to define a new time dependent variable with fixed coefficients

Note As this variable is defined by the econometrician, it is referred as predictable variable

- The pattern of the Schoenfeld residuals are helpful to identify an appropriate time dependent function

Time transfer function

- Consider again the pancreatic cancer data as in S112
- A simple estimation of the Cox model gives

	coef	exp(coef)	se(coef)	z-test	p
stage of progress	0.593	1.81	0.401	1.48	0.14

Recall the Schoenfeld plot revealed that the hazard ratio might vary

- An alternative way is to define a time dependent covariate as

$$g(t) = \theta_0 + \theta_1 \times \log(t)$$

where θ_0 denotes the usual time-invariant group indicator

$\Rightarrow$ Plugging $g(t)$ in the Cox model, the fitted time transfer function is

$$\beta(t) = 6.01 - 1.09 \log(t)$$

	coef	exp(coef)	se(coef)	z-test	p
l(stage)	6.01	407.339	3.060	1.96	0.050
nl(stage)	-1.09	0.338	0.589	-1.84	0.065

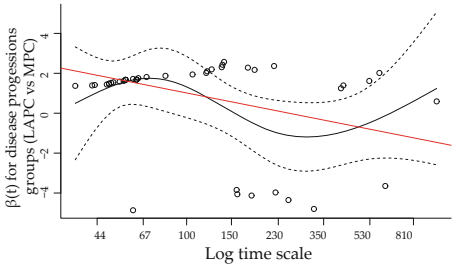
- The LR test that compares the two groups accounting $\beta(t)$ gives

$$LR = 6.33 \;\; (p = 0.0423)$$

$\Rightarrow$ As θ_0 and θ_1 are significant, this suggests that the group indicator combined with a time-varying

Visualization of the time transfer function

- We can use the Schoenfeld residuals plot of S112 to visualize $\theta_1 \times \log(t)$



- The red curve, $-1.09 \log(t)$, is linear as the time axis is in log

$\Rightarrow$ It indicates that overall, the log hazard ratio decreases over time

Note The results are dependent of the functional and e.g. no longer old for

$$g(t) = \theta_0 + \theta_1 \times t$$

stage.n	1.27810	3.590	0.66103	1.93	0.053
tt(stage.n)	-0.00366	0.996	0.00253	-1.44	0.150
LR test	4.56	p=0.102			

Variables that linearly increase with time

- A common source of confusion is whether the age variable is time dependent
- Indeed, the age increases with time itself

$\Rightarrow$ the age is definitely a time dependent variable

But it has no effect on the model if one includes it as time varying covariate

- To see why this happens defined the current age of a subject by

$$x(t) = x(0) + t$$

where $x(0)$ denotes the age at entry into the study

$\Rightarrow$ Then, the hazard function is given by

$$h(t) = h_0(t)e^{\beta x(t)} = \left(h_0(t)e^{\beta t}\right)e^{\beta x(0)}$$

such that once we insert $h(t)$ in the partial likelihood,

$$e^{\beta t}$$

appears in both the numerator and the denominator of each factor

$\Rightarrow$ Hence, it cancels out as does the baseline hazard

Chapter 3

Chapitre 3 : Beyond the Cox Model

Clustered survival times

- Until now, we have considered data with a single cause of failure
- Also, we have assumed that survival times were independent

⇒ How to deal with events that are dependent across individuals ?

— Covid-19 propagation is an example of what we call clustered data

⇒ contamination are more likely to occur for people in a same unit

e.g. children in the same school, employees in the same office, etc.

— In such a case, survival times within a cluster are more similar to each other than to those from other clusters

⇒ the independence assumption no longer holds

⇒ How to deal with an event that can occur repeatedly ?

— The seizure (crise d'épilespsie) is another example of clustered data

⇒ the event may repeat indefinitely per person

Washington Ashkenazi study : dependent data

- This study examined the mutations of a particular gene (the BRCA)

$\Rightarrow$ Is there an effect of mutations on risk of breast cancer ?

- The study was confined to volunteers from the Ashkenazi population
- Each volunteer was controlled for BRCA mutations
- A subset of 1960 families is available (at most two relatives per family)
- For each volunteer, information of two female relatives are collected
  - age of onset of breast cancer (current age for women without cancer)
- The BRCA mutation status of the volunteer is also collected

Washington Ashkenazi study

- Here is a subsample of 3 families
  - — for each volunteers there are 2 rows

e.g. F#1 consists of 2 first degree female relatives (ages 73 and 40)

... neither of them has ever had breast cancer

... nor the volunteer attached to F#1 have a BRCA mutation

Note 1 The survival variable is age of onset

Note 2 The censoring variable is "brcancer" and "mutant" is the covariate

Note 3 As family members share genetic characteristics, they are not independent

Table: Clustered survival data

	famID	brcancer	age	mutant
1	1	0	73	0
2	1	0	40	0
7	9	0	89	0
8	9	1	60	0
87	94	1	44	1
88	94	0	45	1

Marginal Survival Models (MSM)

- This approach ignores clustered data when estimating the model

$\Rightarrow$ Clusters are accounted for when computing standard errors of $\widehat{\beta}$

- MSM relies on standard Cox model estimation
  - — Assume there is one covariate with parameter estimate $\widehat{\beta}$ and $\sigma^2_{\widehat{\beta}} = \mathbb{V}(\widehat{\beta})$
  - — $\sigma^2_{\widehat{\beta}}$ can be misleading as it assumes that all subjects are independent
  - $\Rightarrow$ It has to be corrected for the clustering impact
- The correction requires to first define the following score residuals

$$s_{ij} = \delta_{ij}\big(x_{ij} - \bar{x}(t_{ij})\big) - \sum_{t_u \leq t_{ij}} \big(x_i - \bar{x}(t_{ij})\big) e^{x_i \beta} \Big(\widehat{H}_0(t_u) - \widehat{H}_0(t_{u-1})\Big)$$

  where we can notice that the first part is the Schoenfeld residuals

- The variance correction is then given by

$$C = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \sum_{m=1}^{n_i} s_{ij} s_{im}, \quad G \text{ and } n_i \text{ are defined in the next slide}$$

  where the cluster-adjusted standard error for $\widehat{\beta}$ is $\sigma^*_{\widehat{\beta}} = (\mathbb{V}(\widehat{\beta}) \times C)^{1/2}$

Cluster-adjusted standard errors

- When there are q covariates in the Cox model, β is a vector
- We hence have to apply the correction to the whole estimated covariance matrix of β
- The score residuals are now $1 \times q$ matrices and C is a $q \times q$ matrix as

$$C = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \sum_{m=1}^{n_i} s'_{ij} s_{im}$$

  where G is the number of clusters (assumed to be known here) and n_i is the number of failure in the ith cluster

- Then, the cluster-adjusted covariance matrix is given by

$$V^* = \mathbb{V}(\widehat{\beta}) C \mathbb{V}(\widehat{\beta})$$

  the traditional sandwich estimator

$\Rightarrow$ Adjusted standard errors are then derived as follows

$$se(\widehat{\beta}) = \text{diag}(V^*)^{1/2}$$

Frailty survival models : recall

- Another approach is to generalize to clustered data the likelihood

Recall Under the independence assumption, we may write (see Chapter 1)

$$\mathcal{L}(\beta; x_i) = \prod_{i=1}^{n} f(t_i, \beta)^{\delta_i} S(t_i, \beta)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i, \beta)^{\delta_i} S(t_i, \beta)$$

Recall Under proportional hazards assumption (Cox) it becomes

$$\mathcal{L}(\beta; x_i) = \prod_{i=1}^{n} \left(h_0(t_i) e^{x_i \beta} \right)^{\delta_i} e^{-H_0(t_i) \exp(x_i \beta)}$$

where

$$H_0(t_i) = -\int_0^{t_i} h_0(v) dv$$

is the baseline cumulative hazard

Frailty survival models : principle

- The idea is to assign each individual in a cluster a common factor

$\Rightarrow$ this common factor is known as frailty or random effect and denoted

$$\omega_i$$

for the ith cluster

- Then, for the jth subject in the ith cluster, the hazard function is

$$h_{ij}(t_{ij}) = h_0(t_{ij})\omega_i e^{x_{ij}\beta}$$

- We allow for ω_i to vary from one cluster to another

$\Rightarrow$ a common model that governs this variability is a gamma distribution

$$g(\omega, \theta) = \frac{\omega^{1/\theta - 1} e^{-\omega/\theta}}{\Gamma(1/\theta)\theta^{1/\theta}}$$

- An alternative is to use a standard normal distribution

$$h_{ij}(t_{ij}) = h_0(t_{ij})\omega_i e^{x_{ij}\beta} = h_0(t_{ij}) e^{x_{ij}\beta + u_i\sigma}, \text{ as } \omega_i = e^{u_i\sigma}$$

such that the random and fixed effects are put on the same level

Frailty survival models : unfeasible estimation

- Assuming that the frailties ω_i are observed, the joint likelihood is

$$\mathcal{L}_{ij}(\beta, \theta; \omega_i, t_{ij}, \delta_{ij}, x_{ij}) = g(\omega_i, \theta)\Big(h_0(t_{ij})\omega_i e^{x_{ij}\beta}\Big)^{\delta_{ij}} e^{-H_0(t_{ij})\omega_i \exp(x_{ij}\beta)}$$

  and the full likelihood is

$$\mathcal{L}_{ij}(\beta, \theta) = \sum_{i=1}^{G}\sum_{j=1}^{n_i} \mathcal{L}_{ij}(\beta, \theta; \omega_i, t_{ij}, \delta_{ij}, x_{ij})$$

- MLE of β and θ is feasible under assumption that

$$\omega_i, \quad t_{ij}, \quad \delta_{ij}, \quad x_{ij}$$

  are observed

- Although we can have an idea of the number of clusters, the frailties

$$\omega_i$$

  are in general not observed directly

Frailty survival models : EM algorithm

- In the more realistic case where ω_i are unknown

... one can use the Expectation-Maximization (EM) algorithm

$\Rightarrow$ It alternates between finding expected values for ω_i based on current estimates of

$$\beta \text{ and } \theta$$

and using these expected values to find updated estimates for

$$\beta \text{ and } \theta$$

until convergence

- If we use a parametric distribution for

$$f(t, \beta)$$

setting up the EM algorithm is fairly direct

- Generalizing this to the semi-parametric Cox model is more complex

Example: standard Cox model

- Consider the whole Ashkenazi data set

First Fit the standard Cox model to explain the age of onset of breast cancer

	coef	exp(coef)	se(coef)	z	p
mutant (BRCA)	1.1907	3.2895	0.1984	6.002	1.95e-09

- The likelihoods of the null versus mutant BRCA models are

Example: standard Cox model

- Consider the whole Ashkenazi data set

First Fit the standard Cox model to explain the age of onset of breast cancer

	coef	exp(coef)	se(coef)	z	p
mutant (BRCA)	1.1907	3.2895	0.1984	6.002	1.95e-09

- The likelihoods of the null versus mutant BRCA models are

$$-3579.707 \text{ and } -3566.745$$

respectively and leads to the following LR test statistics

$$LR = 2(-3566.745 + 3579.707) = 25.924$$

that we compare to a χ^2_1 and results in $p < 0.0001$

$\Rightarrow$ this confirms the need of including the BRCA status of the volunteer

Example: MSM

- We now implement the MSM to account for the clustering
- The clusters are defined through the family ID in the database
- We expect here the coefficient to be the same but the adjusted standard error to be different if the cluster are impacting

	coef	exp(coef)	se(coef)	robust se	z	p
mutant (BRCA)	1.1907	3.2895	0.1984	0.2023	6.002	1.95e-09

- The robust standard error is only slightly higher than the unadjusted one

$\Rightarrow$ the effect of clustering within first-degree relatives is small

$\Rightarrow$ the estimation of the MSM reveals that having a first-degree relative with a BRCA mutation increases the hazard of developing breast cancer by a factor of 3.30

Example: frailty

- Finally we implement the frailty model with a gamma distribution
- We expect here the standard error to be different if the clusters matter

$\Rightarrow$ the coefficient is also likely to vary as the likelihood is modified

	coef	se(coef)	se2	Chisq	df	p
mutant	1.272	0.2317	0.2004	30.13	1.0	4.0e-08
frailty(famID)				221.50	211.6	3.1e-01

- Softwares often returns 2 different standard errors
  - — the first is directly derived from the Hessian and is generally preferable
  - — the second is an alternative estimate based on a variation of the sandwich estimator
- The results are close to those obtained with the MSM and Cox models

$\Rightarrow$ having a first-degree relative with a BRCA mutation increases the hazard of developing breast cancer by a factor of $\exp(1.272) = 3.56$

Example: frailty

- The likelihoods of the fixed (no cluster) vs random effects models are

$$-3566.745 \text{ and } -3564.622$$

respectively and leads to the following LR test statistics

Example: frailty

- The likelihoods of the fixed (no cluster) vs random effects models are

$$-3566.745 \text{ and } -3564.622$$

respectively and leads to the following LR test statistics

$$LR = 2(-3564.622 + 3566.745) = 4.246$$

that we compare to a χ^2_1 and results in $p = 0.03934$

- When comparing the null model with the random effects model we have

Example: frailty

- The likelihoods of the fixed (no cluster) vs random effects models are

$$-3566.745 \text{ and } -3564.622$$

respectively and leads to the following LR test statistics

$$LR = 2(-3564.622 + 3566.745) = 4.246$$

that we compare to a χ^2_1 and results in $p = 0.03934$

- When comparing the null model with the random effects model we have

$$-3579.707 \text{ and } -3564.622$$

respectively which leads to the following LR test statistics

$$LR = 2(-3579.707 + 3566.745) = 30.17$$

and that we compare to a χ^2_1 and results in $p < 0.00001$

Cause-specific hazards

- Until now we have considered a single, well-defined outcome
- In some study we may face multiple causes of failure

e.g. an employee can quit the job for different reasons: fired, retirement, ...

- A naive solution is to focus on a particular type of failure

... and treat the others as a type of censoring

- This is questionable as censoring relies on an independence assumption

$\Rightarrow$ What we face here are competing risks, and we have to examine them

Note 1 Interpretation of survival analyses in the presence of competing risks will always be subject to at least some ambiguity due to uncertainty about the degree of dependence among the competing outcomes

Note 2 For a particular subject, we observe only one cause of failure

Kaplan-Meier estimation with competing risks

- Consider first the naive solution : for each type of failure

... while considering others as a type of censoring

- As presumably, the independence assumption is violated, we can question the consequences on Kaplan-Meier estimation

Note Conversely to Cox, KM estimator considers that censoring occurs first

- We illustrate this issue with the prostate cancer data (see Chapter 1)

$\Rightarrow$ focus on patients ages 80+, stage T2, poorly differentiated

Note old patients, with grade 3 advanced cancer

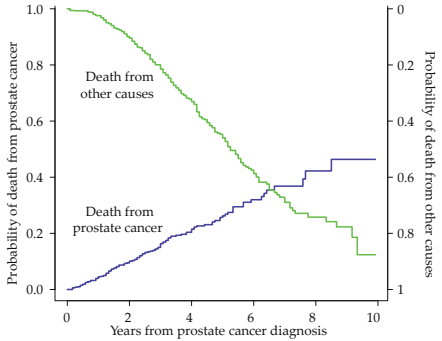
Table: Cancer prostate data for patients ages 80+

	grade	stage	ageGroup	survTime	δ (status)	Δ (other)	$1-\Delta$ (prost)
13	poor	T2	80+	21	0	0	0
38	poor	T2	80+	105	0	0	0
41	poor	T2	80+	2	1	0	1
47	poor	T2	80+	67	2	1	0
78	poor	T2	80+	2	0	0	0
93	poor	T2	80+	60	2	1	0

Example : Kaplan-Meier and competing risks

- In Table 19, when $\delta = 2$ we create a new censoring variable Δ

$\Rightarrow$ we apply twice 2 the KME : $\delta = 2$ as censored and $\delta = 1$ as censored



Note 1 At 10 years, e.g., the $\mathbb{P}$(of dying of prostate cancer) is 0.46 versus 0.88
Note 2 If one assume those 2 probabilities to be independent there is no issue
Note 3 If there are not, as they sums to $1.34 > 1$, this reveals a severe bias
Note 4 Unfortunately, this hypothesis cannot be tested from the data

The cumulative incidence functions

- How to formally address this issue in a non-parametric framework ?
- To develop a formal model to accommodate competing risks,

... assume that there are $K < \infty$ distinct causes of failure

- Also assume that the subject can experience at most one of the K causes
- Then, for each cause of interest, we defined as sub-distribution function

$$F_j(t) = \mathbb{P}(T \leq t, C = j) = \int_0^t h_j(u)S(u)\,du$$

also known as cumulative risk (or incidence) function for the jth cause

- It is increasing as any cumulative distribution function

... but goes, in the limit, to the probability of failure from the jth cause rather than to 1

$$F_j(\infty) = \mathbb{P}(C = j)$$

The cause-specific hazard

- The cause-specific hazard is hence defined conditionally to $C = j$

$$h_j = \lim_{\delta \to 0} \left(\frac{\mathbb{P}(t < T < t + \delta, C = j | T > t)}{\delta} \right)$$

- One can obtain the whole hazards function as follows

$$h(t) = \sum_{j=1}^{K} h_j(t)$$

$\Rightarrow$ The risk of failure at a particular time is simply the sum of the risks of all specific causes at that time

- Now assume that we have D distinct ordered failure times $t_1, t_2, \ldots, t_D$
- We may estimate the hazard at the ith time t_i using

$$\widehat{h}(t_i) = d_i / n_i$$

and the cause-specific hazard for the kth type cause as

$$\widehat{h}_k(t_i) = d_{ik} / n_i$$

i.e. the # of events of type k at t_i divided by the # of subjects at risk

Estimating cause-specific hazards

- The sum over all cause-specific hazards is estimated as

$$\widehat{h}(t_i) = n_i^{-1} \sum_{j=1}^{K} d_{ik}$$

- The probability of failure from any cause at t_i is

$$\widehat{S}(t_{i-1}) \times \widehat{h}(t_i)$$

and hence, for a particular cause k we have

$$\widehat{S}(t_{i-1}) \times \widehat{h}_k(t_i)$$

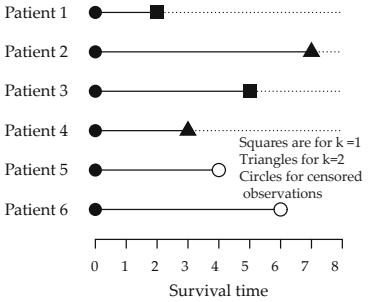
from which we obtain an estimate of the cumulative incidence function

$$\widehat{F}_k(t) = \sum_{t_i \leq t} \widehat{S}(t_{i-1}) \times \widehat{h}_k(t_i)$$

Example : estimation of the cumulative incidence function

- Consider the following artificial data and compute $\widehat{F}_k(t)$ given that

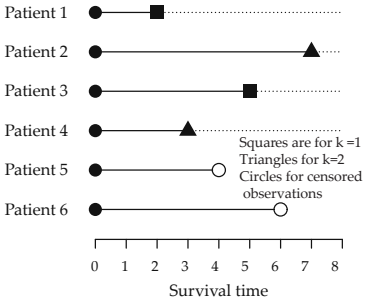
$$\widehat{S}(0, 2, 3, 5, 7) = (1, 0.833, 0.667, 0.444, 0.000)'$$



Example : estimation of the cumulative incidence function

- Consider the following artificial data and compute $\widehat{F}_k(t)$ given that

$$\widehat{S}(0, 2, 3, 5, 7) = (1, 0.833, 0.667, 0.444, 0.000)'$$



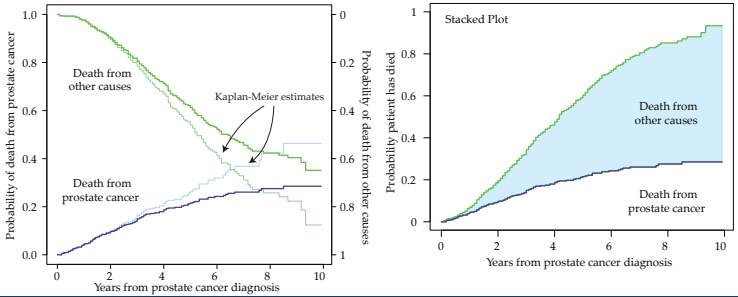
t_i	n_i	d_{i1}	d_{i2}	d_i	$\widehat{S}(t_{i-1})$	$\widehat{h}_1(t_i)$	$\widehat{h}_2(t_i)$	$\widehat{F}_1(t_i)$	$\widehat{F}_2(t_i)$
0	6	0	0	0	1	/	/	0.000	0.000
2	6	1	0	1	0.833	$1/6$	0	0.167	0.000
3	5	0	1	1	0.667	0	$1/5$	0.167	0.167
5	3	1	0	1	0.444	$1/3$	0	0.389	0.167
7	1	0	1	1	0.000	0	1	0.389	0.611

Example : cumulative incidence function for the prostate cancer

- An estimate of $\widehat{F}_k(t)$ for the prostate cancer data gives

t_i	$\widehat{S}(t_{i-1})$	$\widehat{F}_1(t_i)$	$\widehat{F}_2(t_i)$
1	0	1.000	0.00000
2	1	0.994	0.00000
3	2	0.988	0.00602
4	3	0.984	0.00848
5	4	0.983	0.00973
6	5	0.978	0.01477

- When comparing with KME, we see that $\widehat{F}_1(t_i)$ and $\widehat{F}_2(t_i)$ never cross



Regression methods for cause-specific hazards

- Capturing the influence of covariates is challenging in the semi-parametric model of Cox

$\Rightarrow$ How to define the $h_k(t_i)$ on which the covariates should operate ?

- In the spirit of the naive method for the KME, one can consider other causes as censoring and vice versa
- When fitting the Cox model for prostate cancer death we obtain

	coef	exp(coef)	se(coef)	z	p
gradepoor	1.2199	3.3867	0.1004	12.154	2e-16
ageGroup70-74	-0.2860	0.7513	0.2595	-1.102	0.2704
ageGroup75-79	0.4027	1.4958	0.2257	1.784	0.0744
ageGroup80+	0.9728	2.6454	0.2148	4.529	5.92e-06

Note 1 Patients having poorly differentiated disease have much worse prognosis than do patients with moderately differentiated disease

Note 2 The hazard of dying from prostate cancer increases with increasing age of diagnosis (the reference is the youngest age group, 65-69)

Regression methods for cause-specific hazards

- When fitting the Cox model for death from other causes we obtain

	coef	exp(coef)	se(coef)	z	p
gradepoor	0.28104	1.32451	0.05875	4.784	1.72e-06
ageGroup70-74	0.09462	1.09924	0.12492	0.757	0.44879
ageGroup75-79	0.31330	1.36793	0.11709	2.676	0.00746
ageGroup80+	0.79012	2.20367	0.11204	7.052	1.76e-12

Note 1 Patients with poorly differentiated cancer have a higher risk of death from non-prostate-cancer related disease than do those with moderately differentiated disease

Note 3 These results are highly suspect as they rely on the independence assumption

The Fine-Gray method for cause-specific hazards

- A solution that can overcome this issue is to set

$$h_k(t) = \lim_{\delta \to 0} \left(\frac{\mathbb{P}(t < T_k < t + \delta | E)}{\delta} \right)$$

  i.e. to define the effects of covariates on the cause specific hazards where

$$E = \Big((T_k > t \text{ or } (T_{k'} \leq t \text{ and } k' \neq k) \Big)$$

  denotes the conditional event

- The effects of the covariates enter the sub-distribution hazard as follows

$\Rightarrow$ the conditioning set specifies not only $T_k > t$ but also allows other events

... in which case we must have $T_{k'} \leq t$

$\Rightarrow$ the risk set includes not only those currently alive and at risk for the kth event type but also those who failed earlier of causes of type k'

The Fine-Gray method and the model of Cox

- The Fine-Gray framework meets the proportional hazard models by setting

$$h_k(t) = -\frac{\delta \log(1 - F_k(t))}{\delta t}$$

- A proportional Cox-type equation is then apply to sub-distribution hazard

$$h_k(t, x, \beta) = h_{0,k}(t) e^{x\beta}$$

$\Rightarrow$ the sub-distribution hazard for a subject with covariate x is proportional to a baseline sub-distribution function $h_{0,k}(t)$

Example : the Fine-Gray method

- We first fit the Fine-Gray model with the prostate cancer as death cause

	coef	exp(coef)	se(coef)	z	p
gradepoor	1.132	3.102	0.101	11.20	0.00000
ageGroup70-74	-0.272	0.762	0.253	-1.08	0.28000
ageGroup75-79	0.367	1.443	0.219	1.67	0.09400
ageGroup80+	0.799	2.224	0.208	3.85	0.00012

- Second, we estimate the model for death from other causes

	coef	exp(coef)	se(coef)	z	p
gradepoor	0.126	1.13	0.0584	2.154	3.1e-02
ageGroup70-74	0.103	1.11	0.1252	0.824	4.1e-01
ageGroup75-79	0.273	1.31	0.1176	2.323	2.0e-02
ageGroup80+	0.667	1.95	0.1128	5.917	3.3e-09

Note 1 Again we see that poorly differentiated patients have higher risk for death from other causes

Note 2 The risk ratio being 0.126 the effect size is however smaller than we obtained with the naive method (0.281)

Note 3 The estimated effect on death from prostate cancer of having poorly differentiated disease is similar for both methods

Example: comparing the effects of covariates on different causes of death

- One could be interested in comparing the effect of the grade and the age on both causes of death

e.g. the risk of death increases with age but can differ from one cause to another

- Here is a summary of the numbers of events of each type for the dataset

from/to	event-free	prostate	other	no event	total entering
event-free	0	410	1345	4165	5920

- Now we can stratify on cause of death and get estimates of

... the effect of "grade" on cause of death under the assumption that they affect

1 both causes equally

	coef	exp(coef)	se(coef)	z	p
gradepoor	0.515	1.673	0.050	10.372	2.0e-16
ageGroup70-74	0.027	1.027	0.112	0.238	0.81210
ageGroup75-79	0.332	1.394	0.104	3.198	0.00139
ageGroup80+	0.833	2.301	0.099	8.396	2.0e-16

Note This first model is not really useful as we expect that cancer grade affects prostate cancer death differently than it does death from other causes

Example: comparing the effects of covariates on different causes of death

2 or the "grade" status affects both causes differently

	coef	exp(coef)	se(coef)	z	p
gradepoor	1.239	3.451	0.100	12.391	2.0e-16
factor(trans)2	NA	NA	0.000	NA	NA
ageGroup70-74	0.026	1.027	0.112	0.235	0.81431
ageGroup75-79	0.333	1.395	0.104	3.201	0.00137
ageGroup80+	0.833	2.301	0.099	8.394	2.0e-16
gradepoor: factor(trans)2	-0.963	0.382	0.116	-8.327	2.0e-16

- The estimate for "grade" (1.239) is the effect of grade on prostate cancer death, and is similar to what we got earlier (see S149)
- However, the last row is an estimate for the difference between the effect on prostate cancer death and death from other causes

$\Rightarrow$ -0.963, represents the additional effect of poor grade on risk of death from other causes relative to its effect on prostate cancer death

Note 1 Specifically, the hazard of death from other causes is $\exp(1.239 - 0.963) = 1.318$, and hence increased by 32% (much less than the 3.451 factor of death from prostate cancer)

Example: comparing the effects of covariates on different causes of death

- Regarding the age, here are the results we obtain

	coef	exp(coef)	se(coef)	z	p
gradepoor	1.220	3.387	0.100	12.154	2.0e-16
ageGroup70-74	-0.286	0.751	0.260	-1.102	0.2704
ageGroup75-79	0.403	1.496	0.226	1.784	0.0744
ageGroup80+	0.973	2.645	0.215	4.529	5.92e-06
trans2	NA	NA	0.000	NA	NA
gradepoor:trans2	-0.939	0.391	0.116	-8.072	6.66e-16
ageGroup70-74:trans2	0.380	1.463	0.288	1.322	0.1863
ageGroup75-79:trans2	-0.089	0.914	0.254	-0.351	0.7252
ageGroup80+:trans2	-0.183	0.833	0.242	-0.754	0.4508

Note None of these differences are statistically significant

$\Rightarrow$ we conclude that there is no difference in the effect of age on the two death causes, after adjusting for grade

The parametric approach

- In what follows, we develop further the MLE section of Chapter 1
- Non-parametric (e.g. KME) and semi-parametric (e.g. Cox model) approaches are powerful

but they accommodate complex censoring and truncation less directly

$\Rightarrow$ In the parametric framework, the standard likelihood theory applies

but its validity depends on the appropriateness of the selected model

- Here we essentially review
  - — the exponential distribution
  - — the Weibull distribution
  - — the log-normal distribution
  - — the log-logistic distribution

The exponential distribution

- In Ch. 1, we saw that the simple distribution to work with is the exponential one
- It has constant hazard function $h(t) = \lambda$ ($\Rightarrow$ memory-less property)

$\Rightarrow$ The risk of facing the event of interest is the same at any point in time

i.e. neither declines nor increases in time

Recall The p.d.f and survival functions are

$$f(t;\lambda) = \lambda e^{\lambda t} \text{ and } S(t;\lambda) = e^{-\lambda t}$$

- In general, it is not flexible enough but it can help in some specific applications

$\Rightarrow$ power and size calculations

$\Rightarrow$ The Weibull distribution, of which the exponential distribution is a special case, offers more flexibility

The Weibull distribution

Recall The hazard and survival functions are

$$h(t) = \alpha\lambda t^{\alpha-1} \text{ and } S(t) = e^{-(\lambda t)^{\alpha}}$$

- In view of introducing covariates in the parametric model, let define

$$\mu = -log\lambda \text{ and } \sigma = 1/\alpha$$

a location and scale parameter for the distribution

$\Rightarrow$ One can hence rewrite the hazard and survival functions as

$$h(t) = \frac{1}{\sigma} e^{-\mu/\sigma} t^{1/\sigma-1} \text{ and } S(t) = e^{-e^{-\mu/\sigma} t^{1/\sigma}}$$

Note Obviously, when $\sigma = 1$, this reduces to the exponential distribution

Diagnostic tool for the Weibull distribution

- Consider now the $g(u) = \log(-\log(u))$ transformation function for $S(t)$

$$g(S(t_i)) = \alpha \log(\lambda) + \alpha \log(t_i) = -\frac{\mu}{\sigma} + \frac{1}{\sigma} \log(t_i)$$

- This will allow for assessing how well a set of survival data follow a Weibull distribution

1 First compute the KME $\widehat{S}(t_i)$ and define

$$y_i = g(\widehat{S}(t_i))$$

2 Then, plot y_i versus $\log(t_i)$ and fit the linear equation

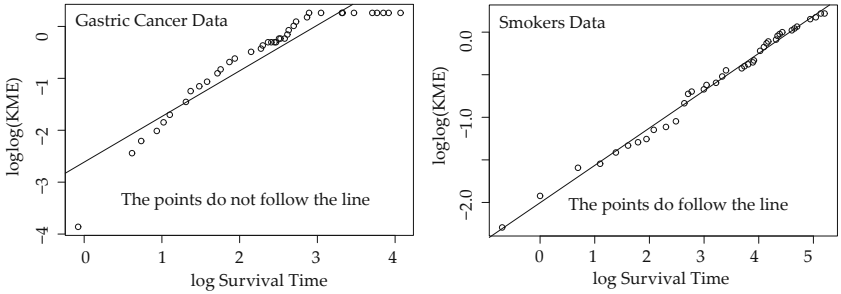
$$y = b + m \log t$$

where $m = 1/\sigma$ and $b = -\mu/\sigma$

$\Rightarrow$ If the plotted points fall along this fitted line, a Weibull distribution should approximate well the distribution of the data

Example: diagnostic tool for the Weibull distribution

- Consider the some databases introduced in Chapter 1



- For the second data set, the Weibull distribution seems plausible
- The fitted straight line parameters are : $b = -2.0032$ and $m = 0.4385$

$\Rightarrow$ Weibull scale and location parameter estimates are:

$$\widehat{\mu} = -b/m = 2.0032/0.4385 = 4.568 \text{ and } \widehat{\sigma} = 1/m = 1/0.4385 = 2.280$$

MLE of Weibull parameters for a single group of survival data

- The linear approach is limited but provides good entries for the MLE
- The log-likelihood function is (see Ch. 1)

$$\ell(\lambda, \alpha) = \sum_{i=1}^{n} \Big(\delta_i \log h(t_i) + \log S(t_i) \Big)$$

- Substituting the expressions for $h(t_i)$ and $S(t_i)$ we get

$$\ell(\lambda, \alpha) = d \log \alpha + d\alpha \log \lambda + (\alpha - 1) \sum_{i=1}^{n} \delta_i \log t_i - \lambda^{\alpha} \sum_{i=1}^{n} t_i^{\alpha}$$

  with $d = \sum_{i=1}^{n} \delta_i$

- The expression can of course be expressed in terms of μ and σ
- Once implemented and applied to the smokers data, we obtain

$$\widehat{\mu}_{MLE} = 4.656329 \text{ and } \widehat{\sigma}_{MLE} = 2.041061$$

$\Rightarrow$ The results are not so far from the linear approach

Note In general, the standard errors are computed for $\widehat{\mu}_{MLE}$ and $\log \widehat{\sigma}_{MLE}$

$$\widehat{\sigma}_{\mu} = 0.2170 \text{ and } \widehat{\sigma}_{\log \sigma} = 0.0919$$

Profile Weibull likelihood

- Suppose that a survival random variable T, follows a Weibull distribution

$$T \sim \text{Weib}(\alpha)$$

- For a given value of α one can define a new random variable

$$T^* = T^\alpha \sim \exp(\lambda^\alpha)$$

- In such a case (see Ch. 1), the analytic solution of the MLE is known

$$\widehat{\lambda}(\alpha) = (d/V)^{1/\alpha}$$

with $V = \sum t_i^\alpha$ and d the total number of deaths

- Since the MLE $\widehat{\lambda}(\alpha)$ can easily be obtained, we can define as

$$\ell^*(\alpha) = \ell(\widehat{\lambda}(\alpha), \alpha)$$

the Weibull profile likelihood

- Maximizing $\ell^*(\alpha)$ yield the MLE of α and the MLE for $\lambda(\alpha)$ is

$$\widehat{\lambda}(\widehat{\alpha}) = (d/V)^{1/\widehat{\alpha}}$$

Example : the profile Weibull likelihood

- When applied to the smokers data, we obtained

$$\widehat{\sigma} = 1/\widehat{\alpha} = 2.041063$$

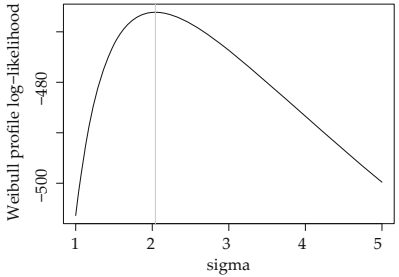
which is almost identical to $\widehat{\sigma}_{MLE}$

- Then, $\widehat{\alpha}$ is used to obtain $\widehat{\lambda}$ and finally

$$\widehat{\mu} = 4.656329$$

which is indistinguishable from $\widehat{\mu}_{MLE}$

$\Rightarrow$ As the MLE only relies on 1 parameter we can plot the profile likelihood



The Accelerated Failure Time model

- When comparing patients of two groups (e.g. treatment and control)

$$e^{\beta},$$

  i.e. the hazard ratio, was the quantity we used

- It was assumed to be time-invariant (proportional hazards hypothesis)

$\Rightarrow$ If the treatment group is effective in increasing survival

$$\beta < 0,$$

  i.e. the log-hazard ratio, such that the hazard ratio is less than 1

- An alternative way of comparing two groups is called AFT
- We assume here that the survival time of the first group is a multiple

$$\theta = e^{\gamma}$$

  of what the survival time would have been had if the patient was in the second group

More intuition on AFT models

- The AFT approach assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant

e.g. If $\theta = 2$ everything in the life history of patient happens twice as fast

$\Rightarrow$ If the model concerns the development of a tumor, this implies that

1 all of the stages progress twice as fast as for the unexposed individual

2 the expected time until the failure event is 0.5 of the baseline time

$\Rightarrow$ Formally, the survival distributions for the AFT models are given by

$$S_1 = S_0(e^{-\gamma}t)$$

and the hazards are given by

$$h_1(t) = e^{-\gamma}h_0(e^{-\gamma}t)$$

The AFT model with Weibull distribution

- In the case of the Weibull distribution we have (see S159)

$$h_1(t) = e^{-\gamma} h_0(e^{-\gamma} t) = e^{-\gamma} \frac{1}{\sigma} e^{-\mu/\sigma} (e^{-\gamma} t)^{1/\sigma - 1}$$

- Rearranging, we have

$$h_1(t) = e^{-\gamma/\sigma} \frac{1}{\sigma} e^{-\mu/\sigma} t^{1/\sigma - 1} = e^{-\gamma/\sigma} h_0(t) = e^{\beta} h_0(t)$$

that is, the AFT model is equivalent to the Cox model with $\beta = -\gamma/\sigma$

Note This equivalence only exists for the Weibull distribution

Comparison of two groups with the parametric Weibull model

- Consider again the smokers data and the comparison of the triple therapy treatment group to the patch therapy group
- When estimating the AFP model we obtain the following results

	coeffs	se	z	p
(Intercept)	5.286	0.3320	15.92	4.59e-57
grppatchOnly	-1.251	0.4348	-2.88	4.00e-03
Log(scale)	0.689	0.0911	7.56	3.97e-14

$\Rightarrow$ $\widehat{\gamma} = -1.251$ indicates that by a factor of

$$\widehat{\theta} = e^{\widehat{\gamma}} = 0.286$$

the patch therapy group has shorter times to relapse (life course to relapse decelerates for the triple therapy group)

$\Rightarrow$ The scale parameter estimate is $\widehat{\sigma} = \exp(0.689) = 1.992$, leading to

$$\widehat{\beta} = -\widehat{\gamma}/\widehat{\sigma} = 0.629$$

for the log proportional hazard in the Cox model

Note In comparison, a Cox-model-based estimation of β gives $\widehat{\beta} = 0.6050$

Interpreting the intercept in the AFT model

- The Cox-model fit provides only 1 estimate, $\widehat{\beta}$
- The AFT Weibull model provides 3 estimates, 2 of them being linked to the baseline Weibull distribution
- In particular, the intercept μ, cannot be estimated in the Cox approach

$\Rightarrow$ it would cancel out of the partial likelihood (as the baseline hazard does)

- The AFT model allows for direct estimation of the baseline hazard as

$$\widehat{\mu} = 5.286 \text{ and } \widehat{\sigma} = 1.992$$

lead to $\widehat{\alpha} = 1/1.992 = 0.502$ and $\widehat{\lambda} = \exp(-5.286) = 0.00506$ and finally

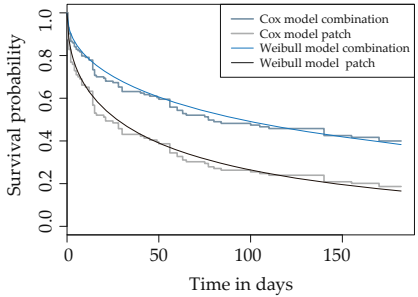
$$\widehat{S}_0(t) = e^{-(\widehat{\lambda}t)^{\widehat{\alpha}}}$$

Comparison of two groups based on survival functions

- The survival function for the combination group is

$$S_1(t) = \big(S_0(t)\big)^{e^{-\gamma/\sigma}}$$

  and can be estimated by replacing all quantities by their estimates
- We can compare $\widehat{S}_0(t)$ and $\widehat{S}_1(t)$ with those obtained from the Cox model
- Notice that the parametric nature of the AFT produces smooth curves



AFT-Weibull-based regression

- An alternative way of looking at Weibull AFT model is to define

$$\log(T) = \mu + \gamma x + \sigma \varepsilon^*$$

  i.e. to model the $\log$-survival time as a location-scale model where

$$\varepsilon^* = \log \varepsilon$$

  with ε, a unit exponential distribution and x a vector of covariates

- Then, the survival function is given by

$$\begin{aligned} S(t) &= \mathbb{P}(T > t) = \mathbb{P}(\varepsilon^* > \frac{\log(t) - \mu - \gamma x}{\sigma}) \\ &= S_0(te^{-\gamma x}) \end{aligned}$$

- This formulation is quite general as different choices for

$$\varepsilon \sim \mathcal{L}(\theta)$$

  can lead to other parametric survival models

Example: AFT-Weibull-based regression

- Consider again the smokers data but with the covariates (see Ch. 2)

Recall For the Cox model we obtained the results below

Recall In this model, $\widehat{\beta}_{patch} = 0.608$ means that the hazard is higher for this treatment group by a constant factor of $\exp(0.608) = 1.83654$

	coef	exp(coef)	se(coef)	z	p
grppatchOnly	0.60788	1.83654	0.21837	2.784	0.00537
age	-0.03529	0.96533	0.01075	-3.282	0.00103
employmentother	0.70348	2.02077	0.26929	2.612	0.00899
employmentpt	0.65369	1.92262	0.32732	1.997	0.04581

- For the AFT Weibull model we obtain

Note $\widehat{\gamma}_{patch} = -1.1902$ means that patients with the patch only have shorter times to relapse by a ~~deceleration~~ acceleration factor of $\exp(-1.1902) = 0.304$

	coeffs	se	z	p
(Intercept)	2.4024	0.9653	2.490	1.28e-02
grppatchOnly	-1.1902	0.4133	-2.880	3.98e-03
age	0.0697	0.0203	3.430	6.02e-04
employmentother	-1.3890	0.5029	-2.760	5.74e-03
employmentpt	-1.3143	0.6132	-2.140	3.21e-02
Log(scale)	0.6313	0.0900	7.020	2.26e-12

Exercise: AFT-Weibull-based regression

- Express the results of the AFT Weibull model in terms of proportional hazards coefficients
- Then, compare these coefficients we those obtained from the Cox model

Exercise: AFT-Weibull-based regression

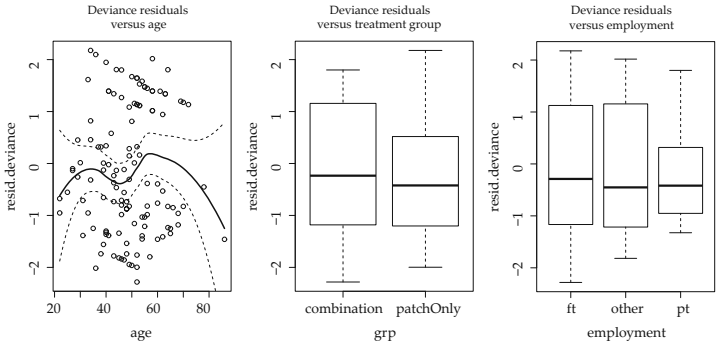
- Express the results of the AFT Weibull model in terms of proportional hazards coefficients
- Then, compare these coefficients we those obtained from the Cox model
- For each regression coefficient γ_j, we have $\beta_j = -\gamma_j/\sigma$

	weib.coef.ph	coxph.coef
grppatchOnly	0.63301278	0.60788405
age	-0.03708786	-0.03528934
employmentother	0.73878031	0.70347664
employmentpt	0.69903157	0.65369019

Model selection and residual analysis

- Many of the facilities for model selection and residual analysis of Ch. 2 remain valid

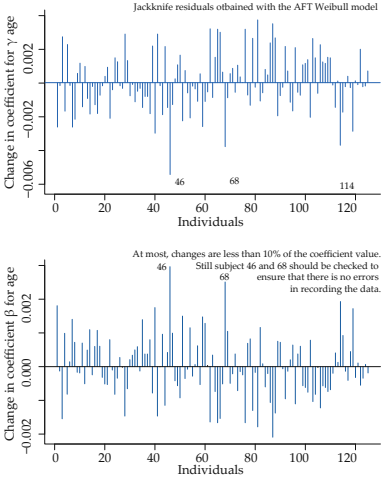
e.g. We plot below the deviance residuals from the previous Weibull model



- The residual distributions of both "grp" and "employ" are reasonably comparable, indicating that these variables are modeled successfully
- For "age", the distribution may be consistent with a linear model, when one considers the width of the 95% confidence intervals

Jackknife residuals

Recall These residuals are computed as the difference in the value of $\widehat{\gamma}$ when all data are used and when an individual is deleted from the data



AFT log-normal model

- Various distribution can be considered in the AFT framework
- For instance, when

$$\varepsilon \sim \mathcal{N}(0, 1)$$

ε^* follows a log-normal distribution and we obtain the following results

	coeffs	se	z	p
(Intercept)	1.6579	1.0084	1.64	1.00e-01
grppatchOnly	-1.2623	0.4523	-2.79	5.25e-03
age	0.0648	0.0203	3.20	1.39e-03
employmentother	-1.1711	0.5316	-2.20	2.76e-02
employmentpt	-0.9543	0.7198	-1.33	1.85e-01
Log(scale)	0.8754	0.0796	10.99	4.15e-28
Scale	2.4			

- All estimates are quite different from what we obtain with the Weibull model albeit with similar signs

AFT log-logistic model

- If ε has a logistic distribution, with survival distribution given by

$$S(u) = \frac{1}{1+e^u}$$

then, T has a $\log$-logistic distribution and the results are now

	coeffs	se	z	p
(Intercept)	1.9150	0.9708	1.97	4.85e-02
grppatchOnly	-1.3260	0.4588	-2.89	3.85e-03
age	0.0617	0.0196	3.15	1.66e-03
employmentother	-1.2605	0.5392	-2.34	1.94e-02
employmentpt	-1.0991	0.7050	-1.56	1.19e-01
Log(scale)	0.3565	0.0884	4.03	5.47e-05
Scale	1.43			

- Again, the estimates are different from what we obtain with the two other models

Large set of covariates

- One of the purpose analysis is to understand how covariates contributes to survival times
- Sometimes we focus on specific covariates such as age, employment, etc.
- By contrast, we can focus on the predictive ability of a set of covariates

$\Rightarrow$ In many cases, dozens or thousands or predictors may be available

- In such a study, many of them are unrelated with survival

and those that are relevant may be strongly correlated amongst themselves

$\Rightarrow$ this multicollinearity is likely to complicate estimation and inference

$\Rightarrow$ Penalized methods such as the Lasso method are useful in such situation

The Lasso method for survival models

- This approach maximizes the partial likelihood function but now with

... the additional stipulation that the L_1 norm of β_j satisfies

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

for a constant t and with p the number of parameters

$\Rightarrow$ This may be shown to be equivalent to maximizing the penalized likelihood

$$\ell_p(\beta) = \ell(\beta) - \lambda \sum_{j=1}^{p} |\beta_j|$$

for λ a pre-specified value of λ

Note 1 Adding this constraint on coefficients shrinks them toward zero (as compared to non-penalized MLE)

Note 2 A too large λ will result in no covariates at all in the model

Note 3 A too small λ will result in a large number of covariates in the model

Optimization issues with the Lasso method

- A complication is that $\ell_p(\beta)$ may not be strictly concave (weakly concave or flat)

$\Rightarrow$ this causes convergence problems

- A crucial issue is hence to select λ

$\Rightarrow$ As in other econometric fields, cross validation procedures are helpful

1. we randomly divide the data set into 5 subsets of equal size
2. we select 1 subset to be the so-called "validation" set
3. we combine the remaining subsets in the so-called "training" set
4. we use the training set ($\approx 80\%$ of the data) to build the Lasso model
5. we use this model to predict the survival times in the validation set
6. we use a partial-likelihood-based measure of goodness-of-fit to this set
7. we repeat steps 1-6 with each of the remaining 4 subsets in turn playing the role of the validation set
8. we derive an average partial-likelihood goodness-of-fit
9. we repeat the whole process for a wide range of values of λ
10. we select the value of λ that produces the optimum goodness-of-fit: λ^*

Example: biomarkers data

- Consider 227 patients with hepatocellular carcinoma (cancer du foie)
- For each patients, a wide range of clinical and biomarker covariates is collected
- The dataset is composed of 48 clinical and biomarker measurements
- Of the 227 patients, 117 have levels of a variety of chemokines markers

$\Rightarrow$ some represent the levels in the tumor itself

Note In medical study, building a predictive model is a complex process that involves interplay between the known medical science and the optimal predictive model

$\Rightarrow$ as we are economists we omit this dimension and consider 26 biomarkers

5 chemokines markers for 3 patients as an example

	OS	Death	CD4T	CD4N	CD8T	CD8N	CD20T
1	83	0	2.600000	0.000000	190.6000	126.80	20.950000
76	20	1	14.450000	2.758621	2.1500	38.95	26.100000
131	35	1	2.821133	8.294828	8.0064	62.64	2.821133

Example: Lasso method for selecting biomarkers

- First, we select the 117 patients for which all biomarkers are available
- Before implement the Lasso, we standardize the covariates

$\Rightarrow$ as the biomarker ranges vary widely

- Then, we set $\lambda = 10$ and fit the Lasso model using 26 biomarkers

$\Rightarrow$ we see that 7 are retained and here are there coefficient estimates

CD8N	CD68T	CD4TR	CD8TR	CD68TR	Ki67	CD34
0.104	0.258	-0.035	-0.096	0.111	0.285	-0.013

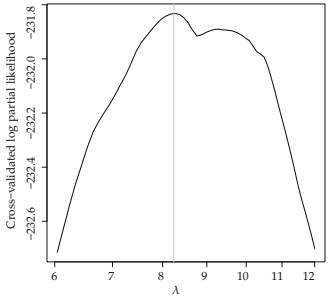
- As $\lambda = 10$ has been specified arbitrarily, we can question the results

$\Rightarrow$ To investigate this we implement the cross validation procedure

Example: Cross validation procedure for the Lasso method

- The cross-validated partial log-likelihood can be plotted to visualize λ^*

$\Rightarrow$ we see that the global maximum is obtained for $\lambda^* = 8.24$



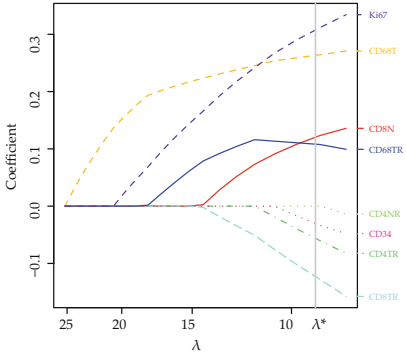
- The results where obtained for $\lambda \in [2, 12]$

CD8N	CD68T	CD4NR	CD4TR	CD8TR	CD68TR	Ki67	CD34
0.133	0.269	-0.009	-0.076	-0.149	0.102	0.328	-0.044

Example: Cross validation procedure for the Lasso method

- One can also be interested in the impact of λ on the estimates

$\Rightarrow$ we can plot the selected markers estimated coefficients for $\lambda \in [20, \lambda^*]$



- The 8 paths that intersect the λ^* vertical line are
  - positive coefficients Ki67, CD68T, CD8N, and CD68TR
  - negative coefficients CD4NR, CD34, CD4TR, and CD8TR

Interpretation of Lasso-based estimates

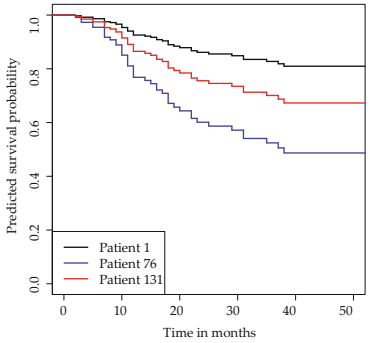
- All $\widehat{\beta}_j$ obtained with λ^* are not interpretable in terms of hazard ratio

1 the lasso procedure has shrunken them

2 they are standardized to have standard deviation one

$\Rightarrow$ However, they can be use to predict the survival profile of patients

Note These patients are those who were in the sample Table of S181



Survival Analysis with R

- We will review the implementation of most of the examples, in R
- All packages we use are listed below (databases are packed in "asaur")

Note This package is attached to the book of Dirk F. Moore (Springer, 2016) that I mainly use for this course

- "asaur" package
- "bshazard" package
- "cmprsk" package
- "coxme" package
- "forestplot" package
- "muhaz" package
- "numDeriv" package
- "Hmisc" package
- "kmconfband," package
- "stats" package
- "penalize" package
- "survival" package

Loading packages and visualizing data

- One of the first dataset we introduce is the aid to smokers to quit
- To visualize the data we need the "asaur" package
- Then we display for the 6 first subjects some columns (2 to 8)

```
> library (asaur)
> pharmacoSmoking[1:6, 2:8]
```

- In the same package we also have, e.g., the pancreatic cancer data
- To quickly visualize the first observations of the database we use

```
> head(pancreatic)
```

Manipulating and visualizing parametric survival distributions

- In the second section we introduce some parametric distributions
- For example, we can plot the Weibull survival function as follows

```
weibSurv <- function(t, shape, scale) pweibull(t, shape=shape,
scale=scale, lower.tail=F)

curve(weibSurv(x, shape=1.5, scale=1/0.03), from=0, to=80,
ylim=c(0,1), ylab='Survival probability', xlab='Time')
```

- We can also plot the Weibull hazard function as follows

```
weibHaz <- function(x, shape, scale) dweibull(x, shape=shape,
scale=scale)/pweibull(x, shape=shape, scale=scale,
lower.tail=F)

curve(weibHaz(x, shape=1.5, scale=1/0.03), from=0, to=80,
ylab='Hazard', xlab='Time', col=''red'')
```

Manipulating and visualizing parametric survival distributions

- If needed we can simulate data from Weibull distribution as

```
tt.weib <- rweibull(1000, shape=1.5, scale=1/0.03)
```

- We can then check whether some empirical quantities converge to their theoretical values

```
> mean(tt.weib)
[1] 31.35497
> median(tt.weib)
[1] 26.84281

> gamma(1 + 1/1.5)/0.03
[1] 30.09151
> (log(2)^(1/1.5))/0.03
[1] 26.10733
```

Computation of the Survival function from the Hazard function

- As discussed in Ch. 1, one can use the hazard function to approximate the survival function
- Then, we can compute the empirical mean and median estimates

```
> library (survival)
> tm <- c(0,1/365,7/365,28/365,1:107)
> tm.diff <- diff(tm)
> survMale <- exp(-cumsum(hazMale*tm.diff)*365.24)
> survFemale <- exp(-cumsum(hazFemale*tm.diff)*365.24)
> sum(survMale*tm.diff)
[1] 71.99964
> sum(survFemale*tm.diff)
[1] 76.98838
```

- At this stage, to get an estimate of the hazard function we rely on

```
> tm <- c(0, birth
1/365, first day of life
7/365, seventh day of life
28/365,fourth week of life
1:106) subsequent years
> hazMale <- survexp.us[,"male","2004"]
> hazFemale <- survexp.us[,"female","2004"]
```

The Kaplan-Meier estimator

- The Kaplan-Meier estimator is the most used non-parametric estimator of the survival function
- In the course we first apply it to artificial data

```
> library(survival)
> tt <- c(7,6,6,5,2,4)
> cens <- c(0,1,0,0,1,1)
> Surv(tt, cens)
[1] 7+ 6 6+ 5+ 2 4
```

- Then, the KME rely on the following function of the survival library

```
result.km <- survfit(Surv(tt, cens) ~ 1, conf.type="log-log")
> summary(result.km)
> plot(result.km)
```

Thank you!