Introduction to Survival Analysis

Smart Analytics for Big Data

Iragaël Joly
Grenoble INP - UMR GAEL

2020-12-09

Overview

Motivations

Basic concepts

Non-parametric estimation

Proportional Hazards - semi-parametric estimation

Fully-parametric Models

Motivations

Motivations

Survival analysis

- Statistical approaches investigating time it takes for an event of interest to occur.
- Operational in many fields (Medecine, sociology, economy, informatics, engineering,...)
- Duration until death, failure, healing ; duration of use, process duration, etc.

Survival analysis

- In basic analysis, we compare proportions (risks, rates, etc) between different groups.
  - assuming *constant rates over the period of the study*
- In longitudinal studies
  - Aim at tracking / observing samples or subjects from one time point (e.g., entry into a study, diagnosis, start of a treatment)
  - until occurence of some outcome event (e.g., death, onset of disease, relapse)

It doesn't make sense to assume the rates are constant over time.

For example :

- the “risk” to find a job is increasing in the first months of search, reaches a top and then decreases
- the risk of death after heart surgery is highest immediately post-op, decreases as the patient recovers, then rises slowly again as the patient ages.}

Survival analysis

Survival analysis is used to model time-to-event (time until an event occurs) or compare the time-to-event between different groups, or how time-to-event correlates with covariables

Overview

Most of survival analyses use the following methods :

- Kaplan-Meier plots to visualize survival curves
- Log-rank test to compare the survival curves of two or more groups
- Cox proportional hazards regression to describe the effect of variables on survival. The Cox model is discussed in the next chapter : Cox proportional hazards model.
- parametric hazard regression to model duration dependance + multivariate analysis

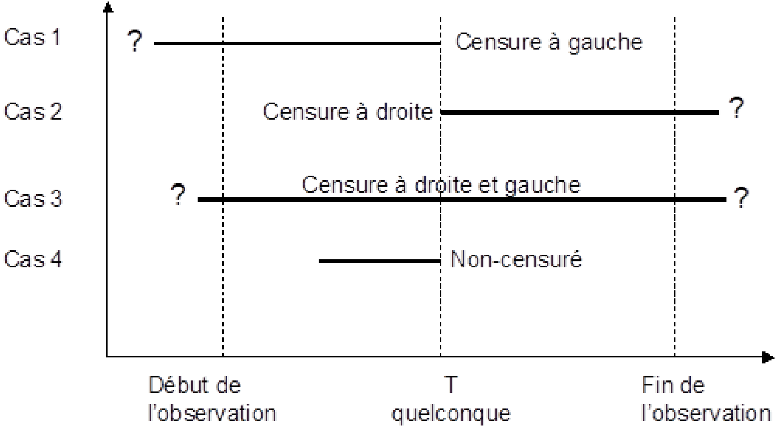
Basic concepts

Basic concepts

- Survival analysis focuses on the **expected duration** until occurrence
- During the period of observation, the event may not be observed for some individuals, producing censored observations
- Censoring is a type of missing data, unique to the survival analysis.
- Right censoring
  - Happens when you track and the event never occurs.
  - This could also happen due to the subject dropping out of the study (for other reasons than the event under study)
- Left censoring
  - occurs when the "start" is unknown
  - The data is at least t, we do not know anything about survival time after that.
- Interval censoring

We limit our introduction to *left censoring*

Censoring


Cas 1
?
Censure à gauche
Cas 2
Censure à droite
?
Cas 3
?
Censure à droite et gauche
?
Cas 4
Non-censuré
Début de l'observation
T quelconque
Fin de l'observation

Survival data

For example :

The `lung` data

- time : Survival time in days
- status : censoring status 1=censored, 2=dead

inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.c
3	306	2	74	1	1	90	100	11
3	455	2	68	1	0	90	90	12
3	1010	1	56	1	0	90	90	N
5	210	2	57	1	1	90	60	11
1	883	2	60	1	0	100	90	N
12	1022	1	74	1	1	50	80	5

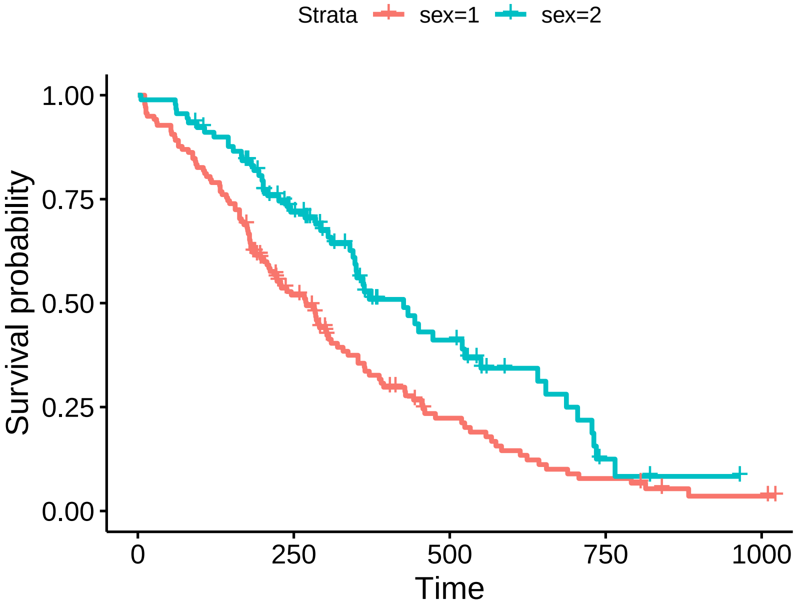
Survival Function

Survival function
$S(t)$ is the probability an event does not occur (an individual survives) up to and including time t.

$$S(t) = Pr(T > t)$$

where T is the time-to-event.

- S is a probability
- so $0 \leq S(t) \leq 1$
- since survival times are always positive ($T \geq 0$)


Strata
sex=1
sex=2
1.00
0.75
0.50
0.25
0.00
Survival probability
0
250
500
750
1000
Time

Hazard rate

Hazard function

Hazard is the instantaneous event rate at a particular time point t.

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

Survival analysis doesn't assume the hazard is constant over time.

Hazard Rate and Survival

Survival function can be writen :

$$S(t) = P(T \geq t) = 1 - F(t)$$

, with $F(t)$ a cumulative distribution function, associatd to the density function $f(t)$:

$$f(t) = lim_{\Delta \to 0} \frac{P(t \leq T < t + \Delta)}{\Delta}$$

Hazard Rate and Survival

Hence, hasard rate is :

$$h(t) = lim_{\Delta \to 0} \frac{F(t+\Delta) - F(t)}{\Delta(1 - F(t))}$$

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{\partial F(t)/\partial t}{S(t)} = \frac{-\partial S(t)/\partial t}{S(t)} = \frac{-\partial lnS(t)}{\partial t}$$

Cumulative hazard

The cumulative hazard is the total hazard experienced up to time t.

$$H(t) = \int_0^t h(t)dt = -lnS(t)$$

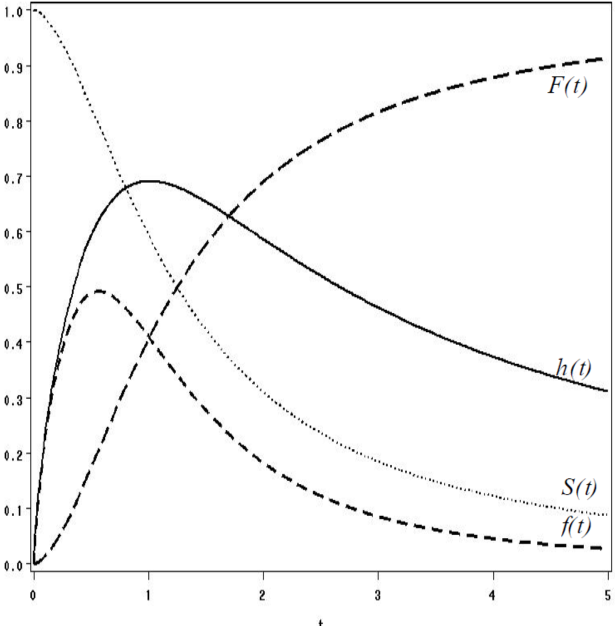
- Hasard function gives the risk of interruption of the duration process (the risk of occurence), knowing that the the process has lasts until t
- Only 'survivors' are observed until t
- Hasard may be different at each time
- Hasard gives the temporal dynamics of the process

Survival interpretation

Survival function :

- $S(t)$ is the probability an individual atteign the date t
- Median survival, at each data t, gives an estimate of the expected survival time, at each time
- Shape of $S(t)$ illustrates the temporal dynamic of the process

Survival, hazard, density functions



Estimation

Estimation techniques can be viewed as *non-parametric*, *semi-parametric* or *parametric*.

- Non-parametric :
  - mainly used to describe $h(t)$ ans (t)
  - usefull for bivariate analysis (test survival difference between groups)
- Parametric methods
  - consist in the fit of a multivariate functionnal form
  - taking into account effects of covariates (as in linear regression)
- semi-parametric methods :
  - are non parametric form for the time distribution
  - but introduce parametric form for the covariates effects : *Proportional hazards*

Parametric and semi-parametric models

Parametric and semi-parametric models are linked.

- Decompose the hazard and survival functions to
- Distinguish & identify
  - the effect of covariates on the hazard
  - the temporal dynamics (effect of elapsed time on the probability of occurence)

Parametric and semi-parametric models

Hence

- Baseline function ($h_0(t)$ and $S_0(t)$)
  - is linked to the assumed distribution function describing the time effect
  - the hazard / survival at time t for an individual where all covariables are 0
- Covariates functional form : $g(\beta' X)$, which affects either
  - the baseline function ($h_0(t)$) : **Proportional hazards models**
  - or directly the time t : **Accelerated failure time models**

Parametric estimation

Parametric estimation techniques permit

- several possible distributions to describe the **temporal dynamics** (constant, monotonic and non-monotonic hazards are allowed)
- estimation of the covariates effects
- estimation of the parameter of the temporal distribution
- gives precise estimations of both temporal dynamics and covariates effects (with all inference properties : CI, PV, etc)

Semi-parametric model

Semi-parametric approaches constrain the model to proportional hazard (separating temporal dynamics and covariates effects).

- covariates effects are precisely estimated, under proportionality assumption (which should be tested)
- temporal dynamics is unconstrained (no parametric distribution is assumed)

Preference between semi-parametric and parametric models is debate subject in litterature as each method has its pro and cons - strenghts and weakness.

Non-parametric estimation

Kaplan-Meier estimator

Survival function is estimated using the KM limit product (Kaplan and Meier (1958)). Estimated survival at time t is calculated as the product of the following proportions :

$$S_{KM}(t_j) = \prod_{k=1}^{j} \frac{n(t_k) - d(t_k)}{n(t_k)}$$

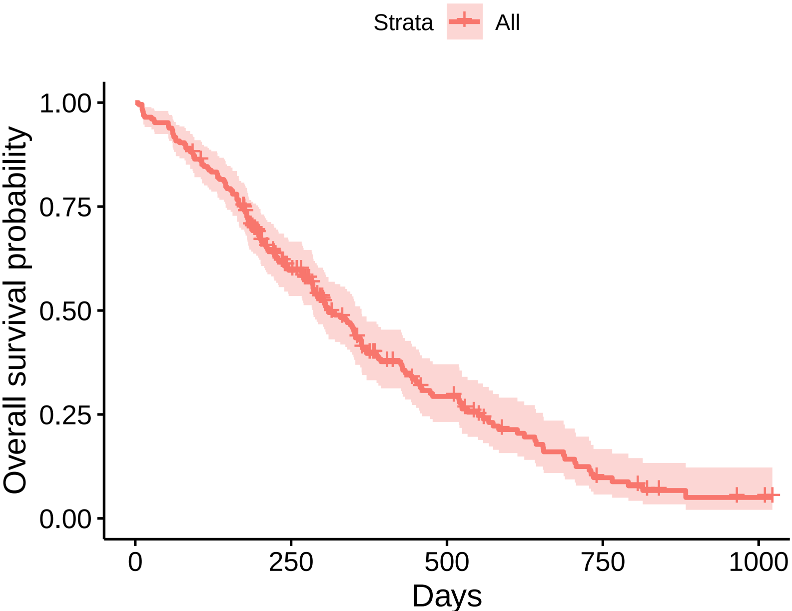
where

- $n(t_k)$ is the population at risk at time t_k.
- $d(t_k)$ is the number of events at time t_k.

or

$$S_{KM}(t_k) = S_{KM}(t_{k-1}) \cdot \left(1 - \frac{d(t_k)}{n(t_k)}\right)$$

Survival curve



Survival KM-Estimate

- $S_{KM}(t_k)$ is a step function illustrating the cumulative survival probability over time
- S_{KM} multiply each step probability to estimate the survival function
- Step are horizontal over periods where no event occurs
- $S_{KM}(t_k)$ drops vertically - change in the survival function at each time an event occurs
- Censored observation are taken into account until they are out of the sample, but they do not count as event
- S_{KM} is asymptotically normally distributed (Andersen et al. (1993), Fleming and Harrington (1991))

Confidence Interval of $S(t)$

Hence, asymptotic confidence interval is given by

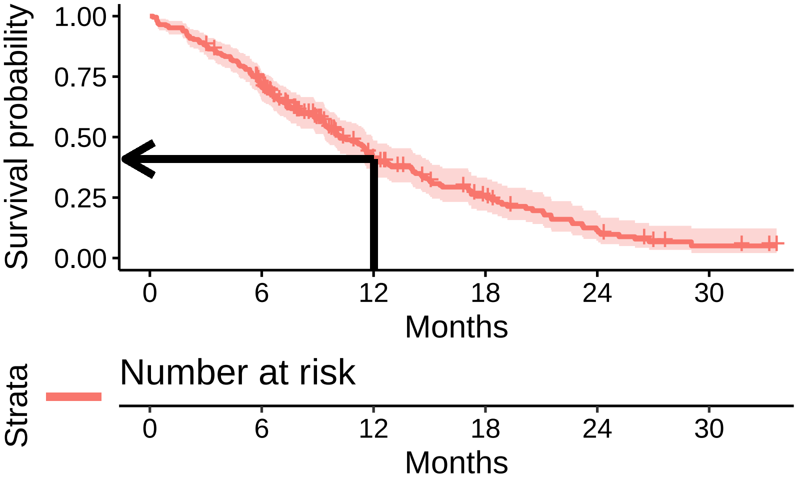
$$\hat{S}(t) \pm z_{1-\alpha/2}\hat{\sigma}_{\hat{S}(t)}$$

where $z_{1-\alpha/2}$ is the normal standard quantile and $\hat{\sigma}_{\hat{S}(t)}$ is the standard error obtained from the variance of the survival estimator (M. (1926)) :

$$\hat{V}(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

x-year survival and the survival curve

The 1-year survival probability is the point on the y-axis that corresponds to 1 year on the x-axis for the survival curve.



Survival probability
1.00
0.75
0.50
0.25
0.00
0
6
12
18
24
30
Months
Strata
Number at risk
0
6
12
18
24
30
Months

Testing survival equivalence between classes

- Survival equivalence are based on contingency table at each date t_i
- Differences are tested between events occurence in a class j : $d_j(t_i)$ and the number of predicted events : $\hat{e}_j(t_i)$ based on the estimation of a common survival function to each class

Contingency table is of the form :

Event	Class 1	Class 0	Total
Interruption	$d_1(t_i)$	$d_0(t_i)$	$d(t_i)$
Non interruption	$n_1(t_i) - d_1(t_i)$	$n_0(t_i) - d_0(t_i)$	$n(t_i) - d(t_i)$
Population at risk	$n_1(t_i)$	$n_0(t_i)$	$n(t_i)$

- $d_1(t_j)$ is the number of event at time j in group 1
- $d(t_j)$ is the total number of event in both groups at time j
- $n_1(t_j)$ is the number at risk just prior to time j
- $n(t_j)$ is the total number of cases that are at risk just prior to j

Estimation of the predicted number of events in class 1 at each date t_j is :

$$\hat{e}_{1j} = \hat{e}_1(t_j) = \frac{n_1(t_j)d(t_j)}{n(t_j)} = \frac{n_{1j}d_{1j}}{n_j}$$

The log-rank test

- For group 1 the log-rank statistic can be written as : $LRT = \sum_{j=1}^{r}(d_{1j} - e_{1j})/(\sqrt{Var(d_{1j})})$, where the summation is over all unique event time (from 1 to r).
- d_{1j} is the number of event occuring at time j in group 1.
- e_{1j} is the expected number of events in group 1 at time j.
- under H_0, $LRT \sim \chi^2(df)$, with $df = j - 1$

The Wilcoxon test

- derived from weighted LRT :

$$LRT = \sum_{j=1}^{r} w_j(d_{1j} - e_{1j})/(\sqrt{w_j^2 Var(d_{1j})})$$

- weights are n_j, the total number at risk at each time.
- Wilcoxon gives more weight to **early times** than late times (as n_j decreases)
- less sensitive than the LRT to differences occuring later
- Log-rank test is more powerful for detecting differences of the form : $S_1(t) = [S_2(t)]^\gamma$, where γ is a positive number other than 1.
- This equation gives proportional hazard model.
- Wilcoxon is more powerfull in situation where event times have log-normal distribution with commun variance

Proportional Hazards - semi-parametric estimation

Proportional Hazards

Proportional hazards assumption :

- PH doesn't assume hazard is constant
- PH assume that the **ratio of hazards** between groups is **constant** over time.
- cumulative hazard ratio between two groups remains constant over time.
- Non-parametric methods are
  - **Visualizing**
  - **Testing differences** in survival between two categorical groups
  - **Multivariate analysis** : link between covariates (both categorical and continuous variables) and hazard.

PH model & Cox model

Under proportional hazards assumption :

$$h(t|X) = h_0(t)g(X, \beta)$$

Where $g(X, \beta) = exp\{(X\beta\}$ (Cox (1972))

Hence : $h(t|X) = h_0(t)exp\{X\beta\}$

Positive coefficient associated with X implies a positive impact of the covariate on the hazard, and as consequence a decrease in survival time.

Cox model

Finally, the Cox model estimates :

$$lnh(t) = lnh_0(t) + X\beta$$

- $h_0(t)$: baseline hazard function depending on t
- covariates X impact in a multiplicative way the hazard
- baseline hazard $h_0(t)$ is 'shared' by all individuals
- $h_0(t)$ and $g(X, \beta)$ are such that $h(t)$ is positive
- $h(t) = h_0(t)$ when $g(X, \beta) = 1$ and $g(X = 0, \beta) = 1$
- $h_0(t)$ depends only on the survival time and represents the varying conditional probability of event with time independtly from the covariates

Note : Exponential and Weibull parametric models are compatible with HP assumption

Interpretation

Ratio of the hazards of individuals i and j (differing in terms of covariates X : X_i and X_j) is :

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t) \times g(X_i, \beta)}{h_0(t) \times g(X_j, \beta)} = \frac{h_0(t) exp\{X_i \beta\}}{h_0(t) exp\{X_j \beta\}} = exp\{(X_i - X_j)\beta\}$$

Coefficient are interpretable as effects on the hazards ratio or in terms of change of the log hazard with respect to the covariate (as relative variation of the hazard) :

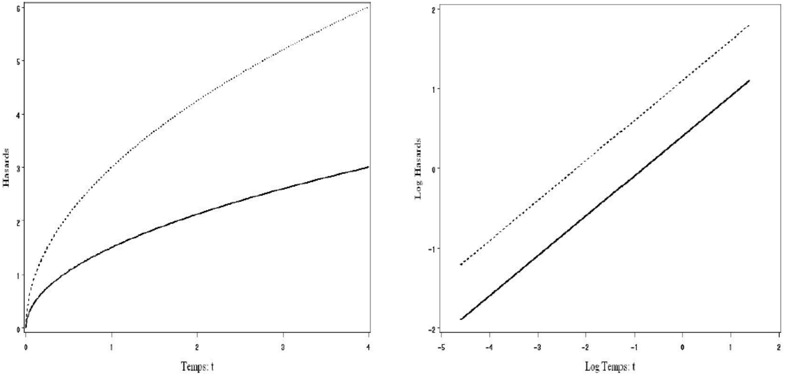
$$\frac{\partial h(t)}{\partial X_k} = \frac{\partial \ln [h_0(t) \cdot g(X, \beta)]}{\partial X_k} = \frac{\partial \ln g(X, \beta)}{\partial X_k} = \frac{\ln(exp\{X\beta\})}{\partial X_k} = \beta_k$$

Note that a positive $\beta > 0$ will leads to an decrease in time-to-event, as it increases the hazard

Interpretation

- for binaries variables : e^{β} gives the ratio of hazards
- for quantitative variables : $+1$ unit of X leads to a change in the hazard of $100 \times (e^{\beta} - 1)$ %.
- Elasticity of the hazard rate with respect to the variable X_k is :

$$\epsilon_k = \frac{X_k}{h} \times \frac{\partial h}{\partial X_k} = \frac{\partial lnh}{\partial lnX_k} = \beta_k X_k$$


Hasards
Temps: t
Log Hasards
Log Temps: t

Cox PH regression

Cox PH regression models the natural log of the hazard at time t, denoted $h(t)$, as a function of the baseline hazard $(h_0(t))$ and multiple covariates $(x_1, \ldots x_k)$.

The form of the Cox PH model is :

$$\ln(h(t)) = \ln(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Hazard Ratio

Assume a restricted model to a unique binary covariate (exposure : $x_1 = 1$ and non exposure $x_1 = 0$), we have (after exponentiation) :

$$h_1(t) = h_0(t) \cdot e^{\beta_1 x_1}$$

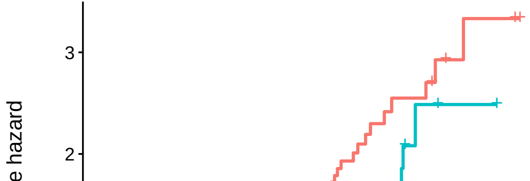
The hazard ratio comes :

$$HR(t) = \frac{h_1(t)}{h_0(t)} = e^{\beta_1}$$

Which shows the constant hazard over all t

Hazard rate and Hazard Ratio

- The quantity of interest from a Cox regression model is a **hazard rate ($h(t)$)**
- Hazard Ratio (HR) : ratio of $h(t)$ between two groups at any particular point in time.
- $h(t)$ is interpreted as the instantaneous rate of occurrence of the event of interest in those who are still at risk for the event
- Regression parameter β leads to HR $= \exp(\beta)$.
- A $HR < 1$ indicates reduced hazard of death whereas a $HR > 1$ indicates an increased hazard of death.



Estimation of the Cox model

- Estimation of the Cox model is performed through maximisation of *the partial Likelihood*[1]
- Use of the PL proposed by Cox (1972) avoids risk of mispecification of the distribution of T
- Estimates of β are considered as more reliable than in the fully parametric model with uncorrect assumed distribution (OAKES (1977)).
- Drawback of this method is theoritically, an increase in the estimates variances, compared to the one obtained in the fully parametric with correct distribution.
- Nevertheless, several studies have shown that this loss in precision is low (Hensher and Mannering (1994)).
- Efron (1977) and OAKES (1977) obtained variance-covariances

Fully-parametric Models

Accelerated Failure Time model

Parametric models assume a log-linear form : $lnt = g(X, \beta) + \sigma\epsilon$

Where

- X is the matrix of covariates (k columns X_k)
- β the associated vector of coefficients
- ϵ is the error term and σ a scale coefficient
- They assume the distribution of ϵ as known (normal, logistic or extrem value)
- The distribution of time-to-event T will depend on the chosen ϵ distribution.

Parametric models

- $g(X, \beta)$ may be different
- Commun specification, we will use here, is :

$$g(X, \mu, \beta) = \mu + \beta X$$

- This form eases interpretation
  - when $X = 0$ then μ : location parameter of the random variable lnT $(E(g(X, \mu, \beta) = \mu + E(X\beta))$
  - β : the variation of $E(lnT|X)$.
- Additive linear form $(X\beta)$ permits a ML estimation
- Log ensures positive predicted values

Interaction : t and X

Time-to-event T is deduced from ϵ.

$$t = exp\{g(X, \mu, \beta) \times (exp\{\epsilon\})^{\sigma}\}$$

- Flexible to model interaction between time and covariates
- Cox and Oakes (1988) : $\lambda = exp\{-g(X, \mu, \beta)\} = exp\{-X\beta\}$, λ has a scale factor role
- If $\lambda > 1$ the temporal scale is accelerated, and decreased when $\lambda < 1$

Covariates are assumed to interact with time :

$$S(t) = S_0(t \times exp\{-\beta' X\})$$

with $S_0(t)$ the baseline survival function.

Corresponding hazard function :

$$h(t) = \frac{-\partial S(t/X)/\partial t}{S(t/X)} = h_0(t \times exp\{-\beta' X\}) \times exp\{-\beta' X\})$$

AFT model

- AFT model : log-linear model

$$lnt = \beta' X + \epsilon$$

- ϵ follows a density function $f(\epsilon)$
- Choosing different $f(\epsilon)$ leads to different models and baseline survival functions

Interpretation

β interpreted in terms of effect on T : $\beta = \frac{\partial ln T}{\partial X_k}$

- binary variables X_k : e^{β_k} is the ratio of *survival times*
- quantitative variables $100 \times (e^{\beta_k} - 1)\%$ is the variation of the survival time associated with a change in covariate X_k

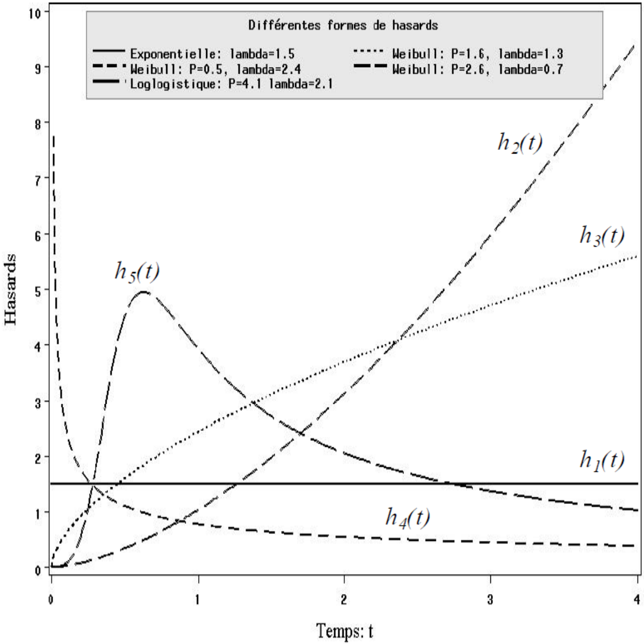
Usual distributions

Distributions des résidus et distributions des durées

Residuals Distribution (ϵ)	Duration Distribution (T)	Model Type
1 parameter Extrem values	Exponential	PH and AFT
2 parameters Extrem values	Weibull	PH and AFT
Logistic	Log-logistic	AFT
Normal	Log-normal	AFT
3 parameters Log Gamma	Generalised Gamma	AFT

Usual distributions

- Exponential hazard ($h_1(t)$) is constant over time and caracterises process that are independant with time
- Weibull hazard ($h_2(t), h_3(t), h_4(t)$), is monotonic
  - If it is positive, then the longer the time-to-event, the higher the probability of event.
- Log-logistic hazard ($h_5(t)$) admits monotonic and non-monotonic forms, given the variance parameter of the distribution.


Différentes formes de hasards
Exponentielle: lambda=1.5
Weibull: P=0.5, lambda=2.4
Loglogistique: P=4.1 lambda=2.1
Weibull: P=1.6, lambda=1.3
Weibull: P=2.6, lambda=0.7
$h_2(t)$
$h_3(t)$
$h_5(t)$
$h_1(t)$
$h_4(t)$
Hasards
Temps: t
0
1
2
3
4
5
6
7
8
9
10

Graphical Diagnostic of the hazard form

- Integrated hazard is usefull to evaluate graphically adequation of a model type to the data.
- KM estimates will help distribution choice
- Note $g(X, \beta) = X\beta$ and $\lambda = exp\{-g(X, \beta)\} = exp\{-X\beta\}$, and $\rho = 1/\sigma$
- Integrated hazard are :
  - **Exponential** case : $H(t) = \lambda \cdot t$
  - **Weibull** case : $H(t) = (\lambda t)^{\rho}$. Hence, its log is : $\ln H(t) = \rho \ln t - \rho X\beta$
  - **Log-logistic** case : $H(t) = \ln(1 + (\lambda t)^{\rho})$. Hence : $\ln(exp\{H(t)\} - 1) = \rho \ln(\lambda t) = \rho \ln t - \rho X\beta$

We can deduce that each couple $(t, H(t))$ or their preceding transformations should follow a linear form with a specific slope.

Likelihood ratio test of the models

- LR test permits to test restriction of a general model versus its constrained version
- Only the log-logistic model is excluded, all other models are nested (exponential, Weibull, log-normal and gamma)
- The LR test statistics is :

$$LR = 2 \cdot \left[\ln L(\hat{\theta_{H1}}) - \ln L(\hat{\theta_{H0}}) \right]$$

- where $\hat{\theta_{H1}}$ and $\hat{\theta_{H0}}$ are the parameters values that maximise the likelihood function associated to the tested assumption H_0 and H_1.
- Under H_0, $LR \ \chi^2(df)$ with $df =$ number of independant restrictions in H_0

LR test

Following restrictions are applicable to pass from the generalised gamma model to another model.

Restriction of the generalised gamma parameter and corresponding model

Constraint	Model	χ^2 distribution df
$\sigma = 1$	Gamma standard	1
$\delta = 1$ and $\sigma \neq 1$	Weibull	1
$\delta = 1$ and $\sigma = 1$	Exponential	2
$\delta \to 0$	Log-normal	1

LR test

Generalised gamma density is characterised by two parameters, σ and δ :

$$f(t) = \frac{\rho\lambda^{\frac{1}{\delta^2}} t^{\rho\frac{1}{\delta^2}-1} exp\{-(\lambda t)^{\rho}\}}{\Gamma(\frac{1}{\delta^2})}$$

Bibliography

Andersen, Per Kragh, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. 1993. *Statistical Models Based on Counting Processes*. Springer.

Bivand, Roger S., Edzer J. Pebesma, and Virgilio. Gómez-Rubio. 2008. *Applied Spatial Data Analysis with r*. New York ; London : Springer.

Cox, D. R. 1972. "Regression Models and Life Tables." *Journal of the Royal Statistic Society* B (34) : 187–202.

Cox, D. R., and D. Oakes. 1988. *Analysis of Survival Data*. Monographs on Statistics and Applied Probability. Chapman ; Hall. https://books.google.fr/books?id=p31BtAEACAAJ.

Bibliography

Efron, Bradley. 1977. "The Efficiency of Cox's Likelihood Function for Censored Data." *Journal of the American Statistical Association* 72 (359) : 557–65.

Fleming, Thomas R., and David P. Harrington. 1991. *Counting Processes and Survival Analysis*. John Wiley & Sons.

Hensher, D., and F. Mannering. 1994. "Hazard-Based Duration Models and Their Application to Transport Analysis." *Transportation Reviews* 14 (1) : 63–82.

———. 1994. "Hazard-Based Duration Models and Their Application to Transport Analysis." *Transportation Reviews* 14 (1) : 63–82.

Bibliography

Kaplan, E. L., and Paul Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53 (282) : 457–81.

Lovelace, R., J. Nowosad, and J. Muenchow. 2019. *Geocomputation with r*. Chapman & Hall/CRC the r Series. CRC Press. https://books.google.fr/books?id=0y6ODwAAQBAJ.

M., Greenwood. 1926. "The Natural Duration of Cancer." *Reports on Public Health and Medical Subjects, Her Majesty's Stationery Office, Londres*, no. 33 : 1–26.

OAKES, DAVID. 1977. "The asymptotic information in censored survival data." *Biometrika* 64 (3) : 441–48. https://doi.org/10.1093/biomet/64.3.441.