Data Analysis and Modeling Techniques
Survival Analysis

Haibe-Kains B[1,2]   Bontempi G[2]

[1]Microarray Unit, Institut Jules Bordet

[2]Machine Learning Group, Université Libre de Bruxelles

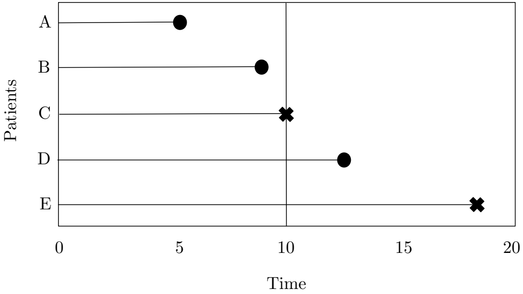
November 27, 2006

Introduction

- Study of the occurrence and timing of events
- Examples : death of patients, failure of a machine, . . .
- Two types of observation plans :
  - prospective : the events are recorded when they occur
  - retrospective : look back at some history recording events of interest
- Usually use of retrospective data with some potential limitations :
  - prone to errors (some events may be forgotten)
  - sampling may be a biased subsample of the initial population of interest

Censoring Data

- You may have only partial information for some cases
- Example : a patient leaves a study before an event occurs



- Cases are right-censored because observation is terminated before the event occurs
- Censoring is random when observations are terminated for reasons that are not under control

Censoring Data

Matrix

- If the study is not limited in time

Patient id	Survival time	Event
A	5	1
B	8	1
C	10	0
D	13	1
E	18	0

- If the study is limited at 10 years

Patient id	Survival time	Event
A	5	1
B	8	1
C	10	0
D	**10**	**0**
E	**10**	**0**

Survival Distribution

- Time of event are realizations of a random variable $\mathbf{t}$
- Two common ways to describe the probability distribution of $\mathbf{t}$
  - survivor function
  - hazard function

Survivor Function

- Probability of surviving beyond t
- $S(t) = \Pr\{\mathbf{t} > t\}$
- Because **t** cannot be negative, $S(0) = 1$
- $S(t)$ can be estimated by the Kaplan-Meier method [Kaplan and Meier, 1958]

$$\widehat{S}(t) = \prod_{j:t_j \leq t} \left[1 - \frac{d_j}{N_j} \right]$$

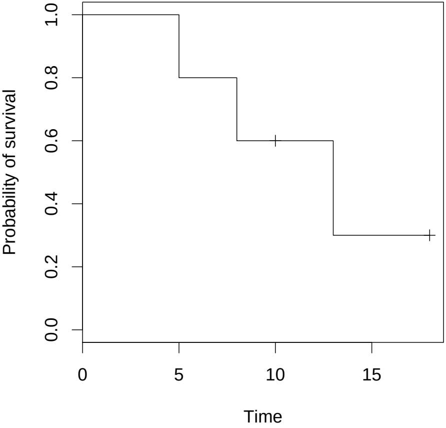
where N_j is the number of cases *at risk* of an event at time t_j and d_j is the number of events at times t_j

Survival Curve

- Censoring data

Patient id	Survival time	Event
A	5	1
B	8	1
C	10	0
D	13	1
E	18	0

- "+" sign represents the censoring on the survival curve



Hazard Function

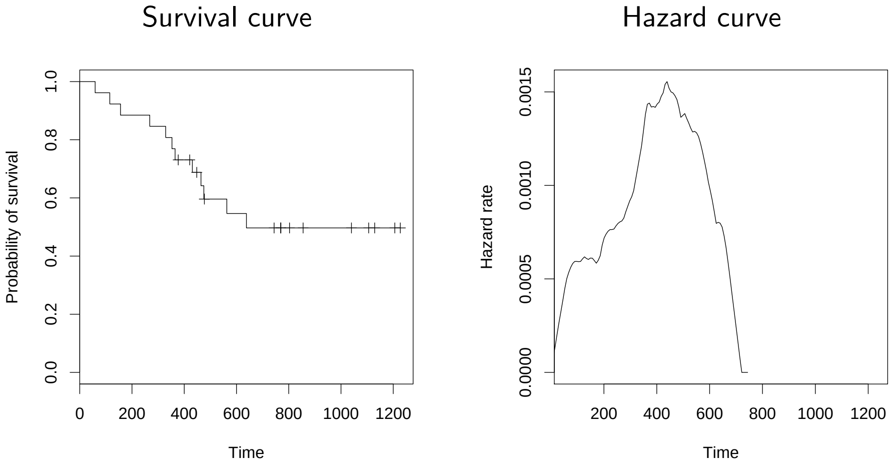
- Instantaneous risk that an event occurs in the small interval between t and $t + \Delta t$

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

- A hazard is a rate not a probability
- $h(t)$ can be estimated by kernel methods [Mueller and Wang, 1994] but you need a sufficient number of cases

Hazard Curve

- Example using a dataset of 26 cases



Regression Model for Survival Data

- There exists different models to fit survivor or hazard functions
  - Parametric model : assumption about the noise distribution that implies specific distribution of $\mathbf{t}$
  - Semiparametric model : no assumption about the distribution of $\mathbf{t}$
- The most widely used method is the Cox regression introduced in [Cox, 1972] that is a semiparametric model

NB : This paper is the most highly cited paper in the entire literature of statistics !

Cox Model

- Let be x_{ij} be the jth covariate for the ith individual with $j \in \{1, 2, \dots, n\}$ and $i \in \{1, 2, \dots, N\}$
- Basic model :

$$h_i(t) = \lambda_0(t) \exp\left(\beta_1 x_{i1} + \cdots + \beta_n x_{in}\right)$$

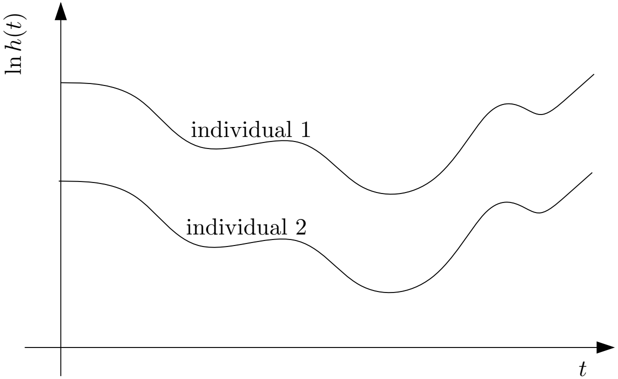
  - $\lambda_0(t)$ is the baseline hazard function
  - linear combination of n covariates which is exponentiated
- This model is called the *proportional hazards model* because the hazard of any individual is a fixed proportion of the hazard of any other individual :

$$\frac{h_i(t)}{h_k(t)} = \exp\left\{\beta_1(x_{i1} - x_{k1}) + \cdots + \beta_n(x_{in} - x_{kn})\right\}$$

  - As you can see, $\lambda_0(t)$ cancels out

Cox Model

Proportional Hazards



- There exist several tests to assess if this assumption is plausible [Therneau and Grambsch, 2000]

Cox Model

Maximum Partial Likelihood

- Fitting the proportional hazards model to an observed set of survival data :
  - estimation of $\beta_1, \beta_2, \dots, \beta_n$, of the covariates $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$
  - does not depend on the baseline hazard function
- Fitting can be performed in maximizing the *partial likelihood*

$$PL = \prod_{i=1}^{N} L_i$$

  - L_i is the likelihood for the ith event

Cox Model

Partial Likelihood

- Definition of L_i : "Given that an event occurred at time t, what is the probability that it happened to case i rather than any other cases ?"

$$L_i = \frac{h_i(t)}{h_i(t) + h_{i+1}(t) + \cdots + h_N(t)}$$

- General expression for the partial likelihood :

$$PL = \prod_{i=1}^{N} \left[\frac{e^{\beta X_i}}{\sum_{j=1}^{N} y_{ij} e^{\beta X_j}} \right]^{\delta_i}$$

  - δ_i is an indicator variable for censoring
  - y_{ij} such that $y_{ij} = 1$ if $t_j \geq t_i$ and $y_{ij} = 0$ if $t_j < t_i$
  - $X_i = [x_{1i}, x_{2i}, \ldots, x_{ni}]$ is a vector of n covariate values

Cox Model

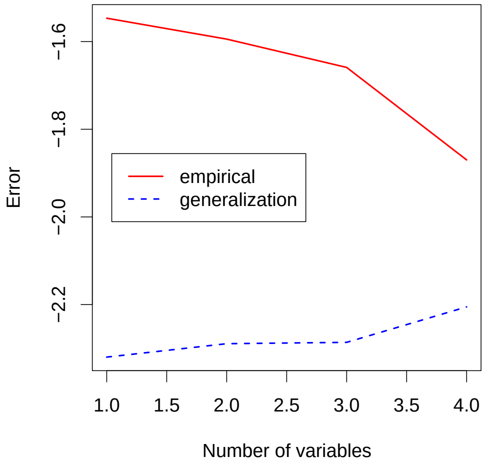
Cross-validated Partial Likelihood

- How to compute an error in cross-validation for the Cox model ?
- Use the CVPL introduced in [Verwij and Van Houwelingen, 1993]

$$CVPL = -\frac{1}{N}\sum_{i=1}^{N}\left[l\left(\hat{f}^{(-s)}\right) - l^{(-s)}\left(\hat{f}^{(-s)}\right)\right]$$

  - l is the log partial likelihood
  - $\hat{f}$ is a fitted Cox model
  - s is a set of cases
  - The index $(-s)$ means that we consider all the cases except those in set s

Example of forward feature selection with CVPL



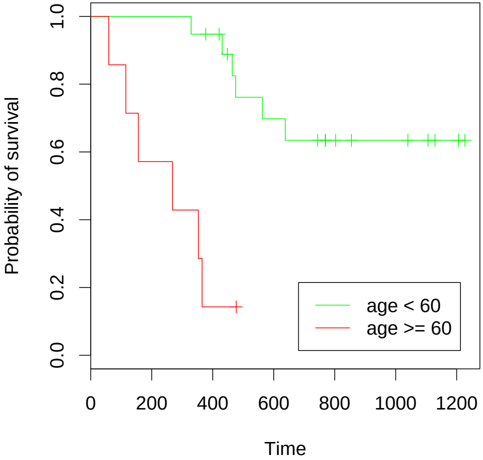
Difference in Survival

- Let say we have two groups of patients defined by their age
  - patients younger than 60 in group 0
  - otherwise in group 1
- You can :
  - test the difference between two survival curves (logrank test)
  - estimate the difference in risk between the two groups (Cox regression)
- These results can be extended for more groups

Testing for Difference in Survival

Logrank Test

We can estimate a survival curve for each group using the Kaplan-Meier estimator :



- Are these two curves statistically different ?
- use of the logrank test : p-value = 3.56e-05

Estimate the Difference in Survival

Hazard Ratio

- Hazard ratio is a relative risk between two conditions
- Summary of the difference between two survival curves
- This difference is constant over time assuming the proportional hazards
- How to compute it ?
  - Let **g** be an indicator variable to specify the group
  - Let g_i be the value of G for the ith individual

$$h_i(t) = \lambda_0(t) \exp(\beta g_i)$$

  - The hazard function for an individual in group 0 is $\lambda_0(t)$
  - The hazard function for an individual in group 1 is $\lambda_0(t) \exp(\beta)$
  - So the hazard ratio is $\exp(\beta)$

Estimate the Difference in Survival

Hazard Ratio : Example

- Most statistical programs report the following information for a fitted Cox model :

	coef	exp(coef)	se(coef)	z	p	N
age $\geq$ 60	2.33	10.2	0.673	3.46	5.5E$-$04	26

  - The indicator variable **g** is noted as "age $\geq$ 60"
  - "coef" is the coefficient
  - "exp(coef)" is the hazard ratio
  - "se(coef)" is the standard error of the coefficient
  - "z" is the common statistic that follows a χ^2 distribution with 1 degree of freedom
  - "p" is the p-value computed from the z statistic
  - "N" is the number of cases

Survival Analysis and Bioinformatics

- Commonly used in Medical fields
- Use of survival and microarray data to study what are the important genes for the appearance of a specific event
  - death
  - tumor
- Survival analysis cane be used with feature selection, regularization, cross-validation, . . .
- The performance assessment is not straightforward

Links

- Course web page : http://www.bioinfomaster.ulb.ac.be/cursus/index_html/en#DATANA
- Personal homepage : http://www.ulb.ac.be/di/map/bhaibeka/
- This presentation : http://www.ulb.ac.be/di/map/bhaibeka/bioinfo_courses/surv_analysis_pres_hkb.pdf

Thank you for your attention.

Part I

Bibliography

Personal Bibliography

- **Definition of clinically distinct molecular subtypes in estrogen receptor positive breast carcinomas through use of genomic grade**.Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, Harris A, Bergh J, Foekens JA, Klijn J, Larsimont D, Buyse M Bontempi G, Delorenszi M, Piccart MJ, Sotiriou C in Journal of clinical Oncology, 2007 (*accepted*)
- **Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis**, Sotiriou C, Wirapati P, Loi S, Harris A, Bergh J, Smeds J, Farmer P, Praz V, Haibe-Kains B, Lallemand F, Buyse M, Piccart M and Delorenzi M in Journal of National Cancer Institute, volume 98, pages 262-272, February 2006
- **Gene regulation by phorbol 12-myristate 13-acetate in MCF-7 and MDA-MB-231, two breast cancer cell lines exhibiting highly different phenotypes**, Lacroix M, Haibe-Kains B, Hennuy B, Laes JF, Lallemand F, Gonze I, Cardoso F, Piccart M, Leclercq G and Sotiriou C in Oncology Reports, volume 12, number 4, pages 701-708, October 2004

Cox, D. R. (1972).
Regression models and life tables.
Journal of the Royal Statistical Society Series B, 34:187–220.

Kaplan, E. L. and Meier, P. (1958).
Nonparametric estimation from incomplete observations.
Journal of American Statistical Asscoiation, 53:457–451.

Mueller, H. G. and Wang, J. L. (1994).
Hazard rates estimation under random censoring with varying kernels and bandwidths.
Biometrics, 50:61–76.

Therneau, T. M. and Grambsch, P. M. (2000).
Modeling Survival Data: Extending the Cox Model.
Springer.

Verwij, P. J. and Van Houwelingen, J. C. (1993).
Cross validation in survival analysis.
Statistics in Medicine, 12:2305–2314.

Part II

Appendix

Softwares

- **R** is a widely used open source language and environment for statistical computing and graphics
  - Software and documentation are available from http://www.r-project.org

R Code

R code to generate a survival curve using the ”ovarian” data :

```
library(survival)
fit <- survfit(Surv(futime, fustat), data=ovarian, conf.type="none")
par(cex=1.5)
plot(fit, xlab="Time", ylab="Probability of survival")
```

R code to generate a hazard curve using the ”ovarian” data :

```
library(survival)
library(muhaz)
fit <- muhaz(times=ovarian$futime, delta=ovarian$fustat, bw.pilot=10)
par(cex=1.5)
plot(fit, xlim=range(ovarian$futime), xlab="Time", ylab="Hazard rate")
```

R Code

R code to compute the forward feature selection in the ”ovarian” dataset and report the empirical and the generalization (CVPL) errors :

```
library(survival)
library(bensurvfoo)
library(gplots)
rr <- fw.cvpl(data=ovarian[ ,c("age","resid.ds","rx","ecog.ps")],
 surv.time=ovarian$futime, surv.event=ovarian$fustat,
 strata.cox=NULL, setseed=12345, na.rm=TRUE, verbose=TRUE)
gen.err <- - unlist(lapply(rr$perf, function(x) { return(x[[1]]) }))
emp.err <- NULL
for(i in 1:length(rr$sel)) {
emp.err <- c(emp.err, coxph(Surv(ovarian$futime, ovarian$fustat) ~ .,
 data=ovarian[ ,rr$sel[1:i],drop=FALSE])$loglik[2] / sum(ovarian$fustat))
}
plot(gen.err, ylim=range(c(gen.err, emp.err)), type="l",
 col="red", lwd=2, lty=1, xlab="Number of variables", ylab="Error")
lines(emp.err, col="blue", lwd=2, lty=2)
smartlegend(x="left", y="center", c("empirical", "generalization"),
 lty=c(1,2), lwd=c(2,2), col=c("red", "blue"))
```

R Code

R code to generate two survival curves using the "ovarian" data and testing their difference using the logrank test :

```
library(survival)
library(bensurvfoo)
par(cex=1.5)
mysurvivalplot(group=ovarian$age >= 60, surv.time=ovarian$futime,
 surv.event=ovarian$fustat, na.rm=TRUE,
 group.name=c("age < 60", "age >= 60"), global=FALSE,
 stat.info=c(FALSE, FALSE), strata.cox=NULL, main="",
 group.col=c("green", "red"))
survdiff(Surv(ovarian$futime, ovarian$fustat) ˜ ovarian$age >= 60)
```

R code to fit a Cox model using the "ovarian" data :

```
library(survival)
coxph(Surv(ovarian$futime, ovarian$fustat) ˜ ovarian$age >= 60)
```