Project Survival Data Analysis onto
Documented Primary Biliary Cirrhosis Dataset :
SDA Models

Table of contents

PART.1 Exploratory Statistic Analysis

PART.2 Univariate Analysis

PART.3 : Multi-Variate Analysis

PART.4 : Random-Forest

Missing time for :

1. Project Profile

Project Package overview for documentation and codes:

- Title:	**Survival Data Analysis for PBC Mayo Clinical Trial data**
- Project platform:	**Rstudio**
- Submitted to:	**Dr. Antonio Di Narzo**
- Developed by:	**Michaël Faivre**

2. Introduction

2.1 Project Summary

The project encompasses both aspects of (I) testing univariate & multi-variate models (ii) testing respective covariate impacts and (iii) comparing the performances of all models. This is a stimulating project as it involves the use of both semi-parametric and non-parametric methods in a classical case clinical study. I decided to implement the algorithm of analysis in R due to quickly accessible references and more straightforward in coding compared with Python.

NB : All, many online references used are acknowledged in the codes.

Unfortunately, the task will demand me more to be achieved, and some sections remain incompleted currently in the present document.

2.2 Purpose of the assignment & Aim of the SDA analysis for clinical trial data

The assignment aims at : (1) choose and describe and right-censored dataset ; (2) proceed to first level exploratory analysis and data cleaning if needed ; (3) apply,test and compare Survival Analysis toolkit with semi-parametric and non-parametric approaches.

The main purpose of this study is to investigate the impact of D-penicillamine and serum bilirubin to lifetime of patients with Primary Biliary Cirrhosis (PBC). The scope of the study is also to make distinct assessement of survival time with respect to sex and treatment factorial variables.

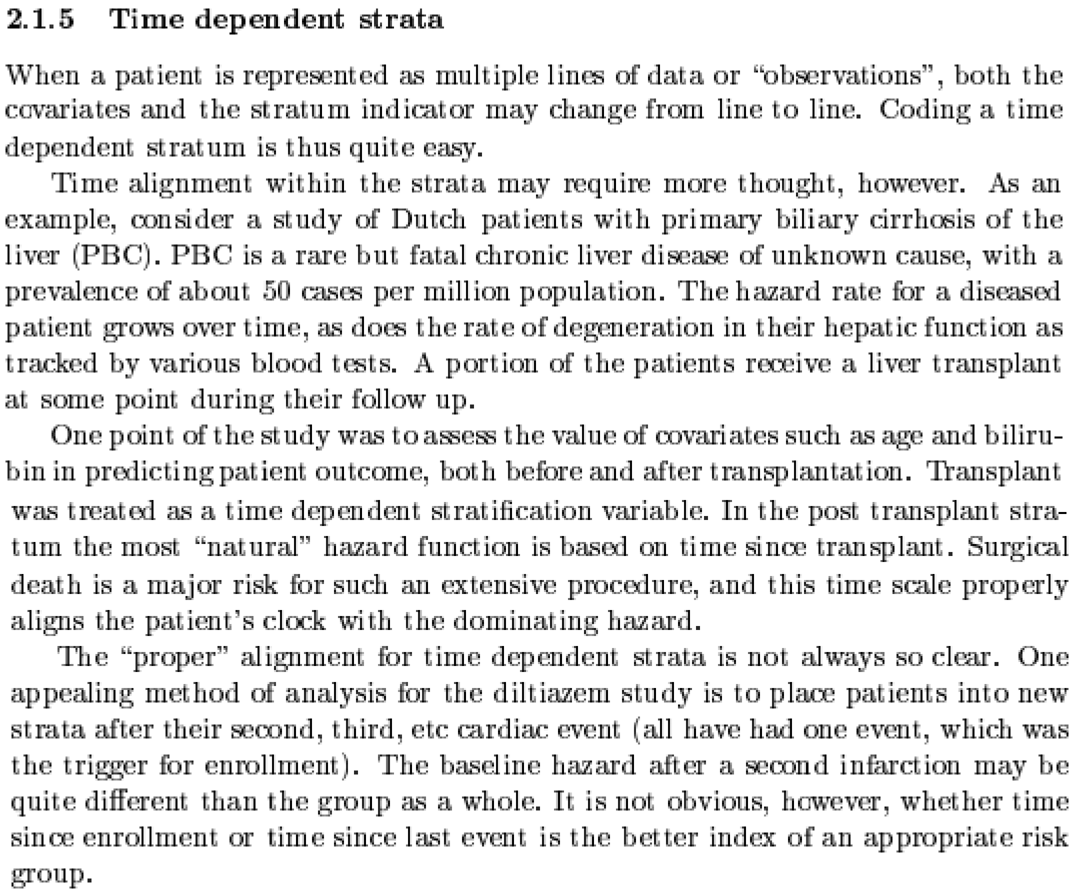
PART.1 : Exploratory Statistic Analysis

Dataset : Primary Biliary Cirrhosis from the Mayo Clinic trials

Dataset overview :

PBC MayoBiliary is a so called right censored reference dataset. It contains the records from 412 patients measured at the Mayo Clinic, Rochester, Minnesota, over the period 1974-1984.
I choose this dataset as I am not yet confident in my handling of the Survival Data Analysis protocol and the PBC turns out to be a benchmark data set with quite a number of associated publications.

From : Therneau Modeling Survival Data, Extending the Cox model



2.1.5 Time dependent strata

When a patient is represented as multiple lines of data or "observations", both the covariates and the stratum indicator may change from line to line. Coding a time dependent stratum is thus quite easy.

Time alignment within the strata may require more thought, however. As an example, consider a study of Dutch patients with primary biliary cirrhosis of the liver (PBC). PBC is a rare but fatal chronic liver disease of unknown cause, with a prevalence of about 50 cases per million population. The hazard rate for a diseased patient grows over time, as does the rate of degeneration in their hepatic function as tracked by various blood tests. A portion of the patients receive a liver transplant at some point during their follow up.

One point of the study was to assess the value of covariates such as age and bilirubin in predicting patient outcome, both before and after transplantation. Transplant was treated as a time dependent stratification variable. In the post transplant stratum the most "natural" hazard function is based on time since transplant. Surgical death is a major risk for such an extensive procedure, and this time scale properly aligns the patient's clock with the dominating hazard.

The "proper" alignment for time dependent strata is not always so clear. One appealing method of analysis for the diltiazem study is to place patients into new strata after their second, third, etc cardiac event (all have had one event, which was the trigger for enrollment). The baseline hazard after a second infarction may be quite different than the group as a whole. It is not obvious, however, whether time since enrollment or time since last event is the better index of an appropriate risk group.

DS1.1 Dataset description

```
    'data.frame':   418 obs. of  20 variables:
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ time    : int  400 4500 1012 1925 1504 2503 1832 2466 2400 51 ...
 $ status  : int  2 0 2 2 1 2 0 2 2 2 ...
 $ trt     : int  1 1 1 1 2 2 2 2 1 2 ...
 $ age     : num  58.8 56.4 70.1 54.7 38.1 ...
 $ sex     : Factor w/ 2 levels "m","f": 2 2 1 2 2 2 2 2 2 2 ...
 $ ascites : int  1 0 0 0 0 0 0 0 0 1 ...
 $ hepato  : int  1 1 0 1 1 1 1 0 0 0 ...
 $ spiders : int  1 1 0 1 1 0 0 0 1 1 ...
 $ edema   : num  1 0 0.5 0.5 0 0 0 0 0 1 ...
 $ bili    : num  14.5 1.1 1.4 1.8 3.4 0.8 1 0.3 3.2 12.6 ...
 $ chol    : int  261 302 176 244 279 248 322 280 562 200 ...
 $ albumin : num  2.6 4.14 3.48 2.54 3.53 3.98 4.09 4 3.08 2.74 ...
 $ copper  : int  156 54 210 64 143 50 52 52 79 140 ...
 $ alk.phos: num  1718 7395 516 6122 671 ...
 $ ast     : num  137.9 113.5 96.1 60.6 113.2 ...
 $ trig    : int  172 88 55 92 72 63 213 189 88 143 ...
 $ platelet: int  190 221 151 183 136 NA 204 373 251 302 ...
 $ protime : num  12.2 10.6 12 10.3 10.9 11 9.7 11 11 11.5 ...
 $ stage   : int  4 3 4 4 3 3 3 3 2 4 ...
```

20 variables splitted as follow :

- **1 patient index**
  2 response variables : time & protime

1 censor factor : status
16 explanatory covariates

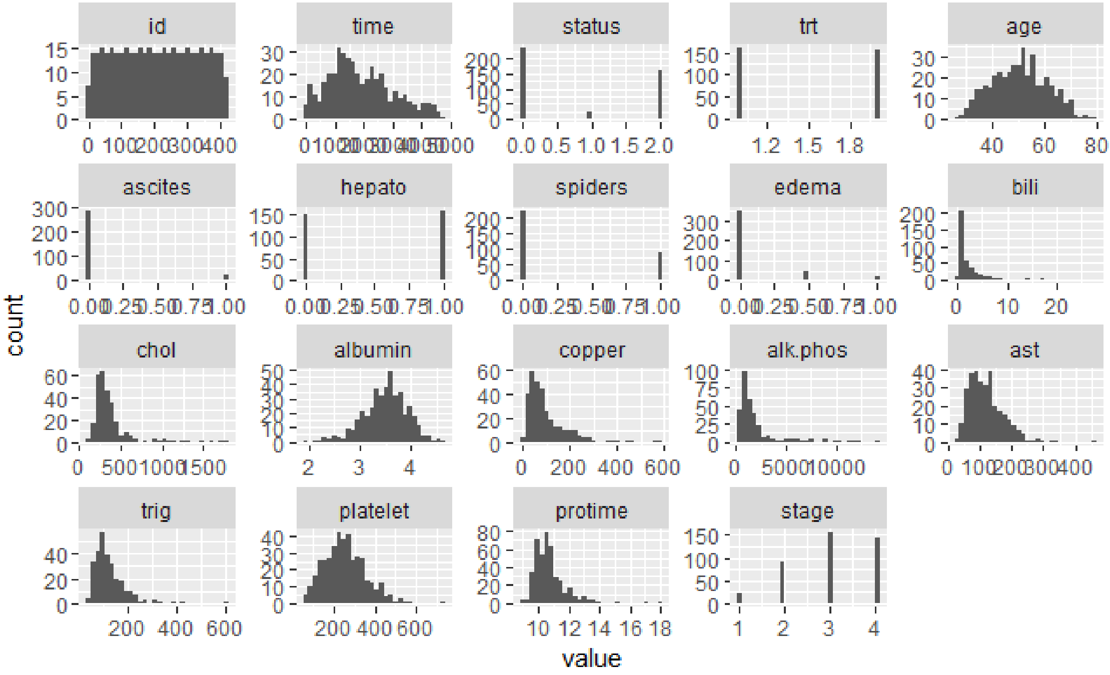


Figure.1 : covariates distributions. Note on variable distributions : exponential, uniform, normal-like distributions. On STATUS : there is a majority of Censored Data (flag=0) !!

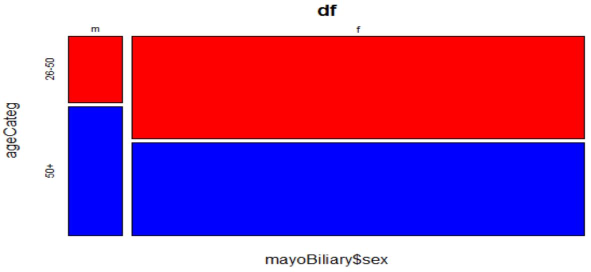


Figure.2 Age/sex proportions at a glance : female way more represented

DS1.2 Descriptives Statistics

```
id (Not to use)  time            status          trt             age             sex (factor)
 Min.   :  1.0   Min.   :  41    Min.   :0.0000  Min.   :1.000   Min.   :26.28   m: 44
 1st Qu.:105.2   1st Qu.:1093    1st Qu.:0.0000  1st Qu.:1.000   1st Qu.:42.83   f:374
 Median :209.5   Median :1730    Median :0.0000  Median :1.000   Median :51.00
 Mean   :209.5   Mean   :1918    Mean   :0.8301  Mean   :1.494   Mean   :50.74
 3rd Qu.:313.8   3rd Qu.:2614    3rd Qu.:2.0000  3rd Qu.:2.000   3rd Qu.:58.24
 Max.   :418.0   Max.   :4795    Max.   :2.0000  Max.   :2.000   Max.   :78.44
                                                 NA's   :106


    ascites            hepato          spiders           edema             bili
 Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 0.300
 1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.800
 Median :0.00000   Median :1.0000   Median :0.0000   Median :0.0000   Median : 1.400
 Mean   :0.07692   Mean   :0.5128   Mean   :0.2885   Mean   :0.1005   Mean   : 3.221
 3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 3.400
 Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :28.000
 NA's   :106       NA's   :106      NA's   :106
      chol            albumin          copper          alk.phos             ast
 Min.   : 120.0   Min.   :1.960   Min.   :  4.00   Min.   :  289.0   Min.   : 26.35
 1st Qu.: 249.5   1st Qu.:3.243   1st Qu.: 41.25   1st Qu.:  871.5   1st Qu.: 80.60
 Median : 309.5   Median :3.530   Median : 73.00   Median : 1259.0   Median :114.70
 Mean   : 369.5   Mean   :3.497   Mean   : 97.65   Mean   : 1982.7   Mean   :122.56
 3rd Qu.: 400.0   3rd Qu.:3.770   3rd Qu.:123.00   3rd Qu.: 1980.0   3rd Qu.:151.90
 Max.   :1775.0   Max.   :4.640   Max.   :588.00   Max.   :13862.4   Max.   :457.25
 NA's   :134                      NA's   :108      NA's   :106       NA's   :106
      trig           platelet        protime          stage
 Min.   : 33.00   Min.   : 62.0   Min.   : 9.00   Min.   :1.000
 1st Qu.: 84.25   1st Qu.:188.5   1st Qu.:10.00   1st Qu.:2.000
 Median :108.00   Median :251.0   Median :10.60   Median :3.000
 Mean   :124.70   Mean   :257.0   Mean   :10.73   Mean   :3.024
 3rd Qu.:151.00   3rd Qu.:318.0   3rd Qu.:11.10   3rd Qu.:4.000
 Max.   :598.00   Max.   :721.0   Max.   :18.00   Max.   :4.000
 NA's   :136      NA's   :11      NA's   :2       NA's   :6
```

- 2 things I can mention at this point : (1) to exclude non-relevant Id variable and (2) how to deal with the missing values in the Right-censored Survival analysis

- For the missing values processing, I refer to the work by Therneau
- Number of censored data (status=0) :
- **Categorial variable** 'Treatment' :NA(not randomised) → 0 flag
- **To address missing values, Imputation** is processed for all covarates with the R mice function. Please see the document imputed_PBC.pdf

Table.1 Variable explanation and categorical or continuous

Categorical	**Continuous**
ascites (as.factor)	albumin
age (cut in pseudo-uniform ranges)	Alkaline phosphotase
edema	ast
hepato	time
Sex	protime
spiders	Platelet count
stage	trig
	bilirunbin
	copper
	chol

As seen in the course, day3, **Categorical variables** are to treated as dummy variables. Ex : age 0,1 → age group1 <50 years & group2>50years.

DS1.3 Pre-processing operations Table

Table.2 List of Data cleaning & preprocessing and basic tests

Missing values processing	Yes missing values with distribution indicated in Table.1 2 techniques tested → 2 new datasets : 1 with imputation, 1 with spline approximation
Data transformation	-Time : days / 356,24 → years ;sex → factor(sex)
Oulier detection and processing	2 techniques tested : Cook's distance and influence measure
Variable selection and Dim reduction	PCA on var-cov matrix
Homoscedasticity	Not applied
Gaussianity of the errors	Not applied
Kernel Density Estimation (Parzen-Rosenblatt)	Tested for a couple of variables

DS1.4 Missing Data per covariate

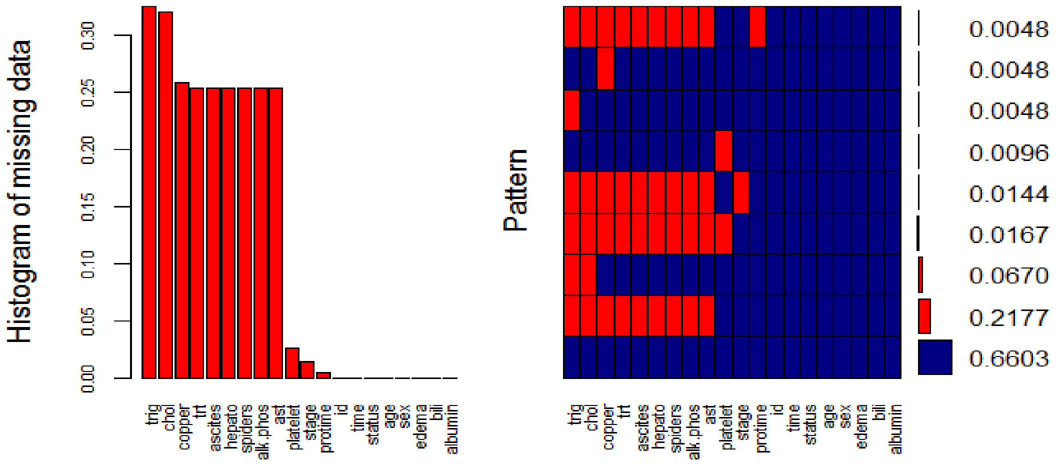


Figure.3 Missing values percentage per covariate. Not all covariates have missing values.

DS1.5 Dataviz Time dependence of some covariates

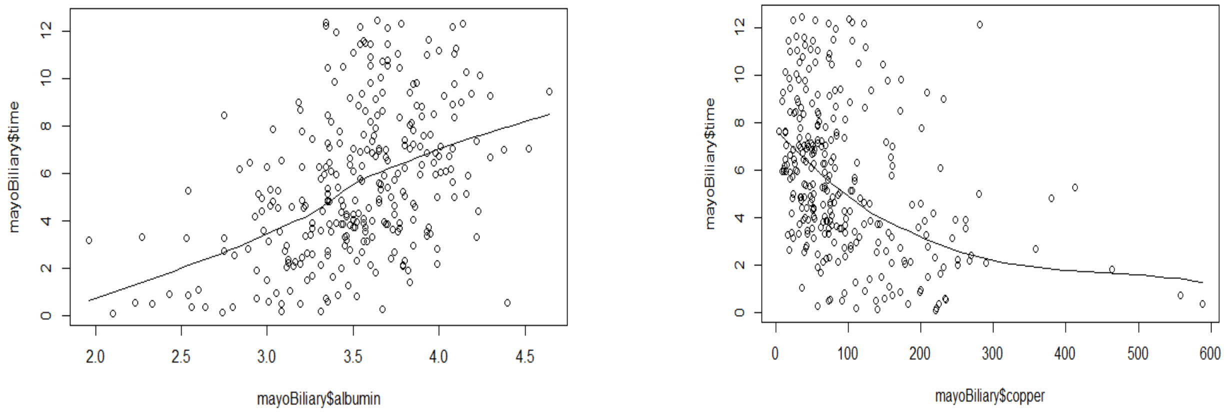


Figure.4 Survival time vs Albumin (left) and Copper (right)

A few covariates exhibit a significant variation vs survival time (Copper, Albumin). The Correlogram (Apendix.1) depicts the time depence for each covariate.

DS1.6 Log-Linearity of numerical variables

Log-linearity of the numeric covariates has been checked along the way using splines Beta(t) per covariate in Cox Models.

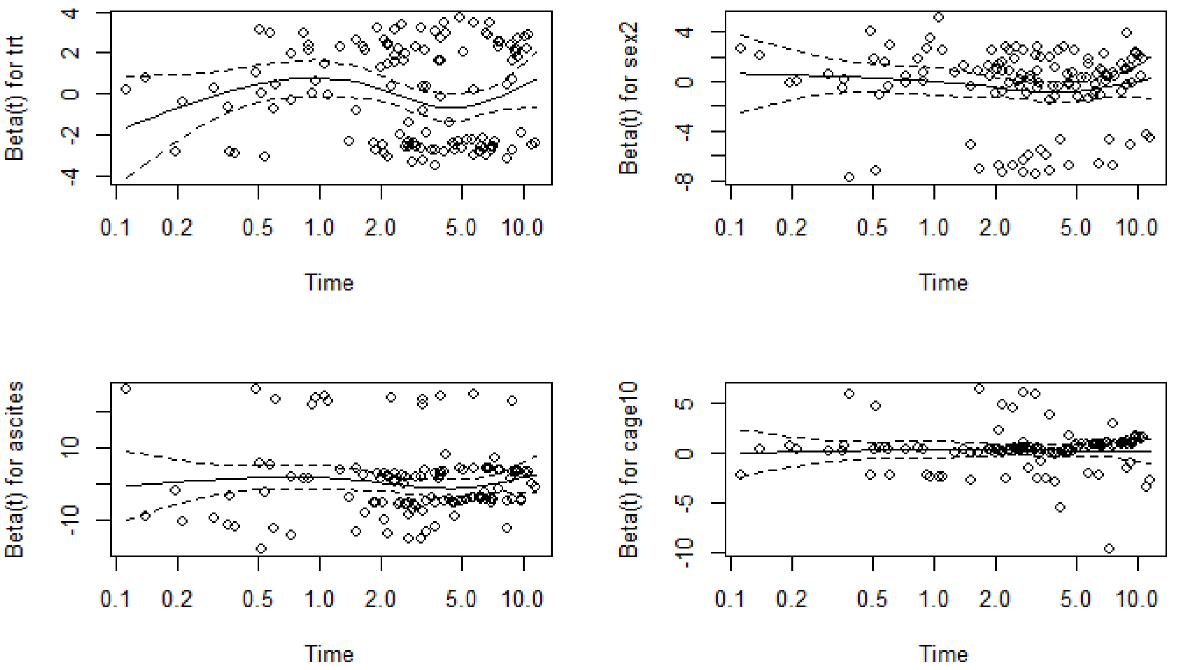


Figure. 5 scaltter plot of Beta(t) per covaraite over time to assess their log-linearity.

DS1.7 Box plot dispersion separated for each

Attention : Boxplot for censored data is not a good idea!!!

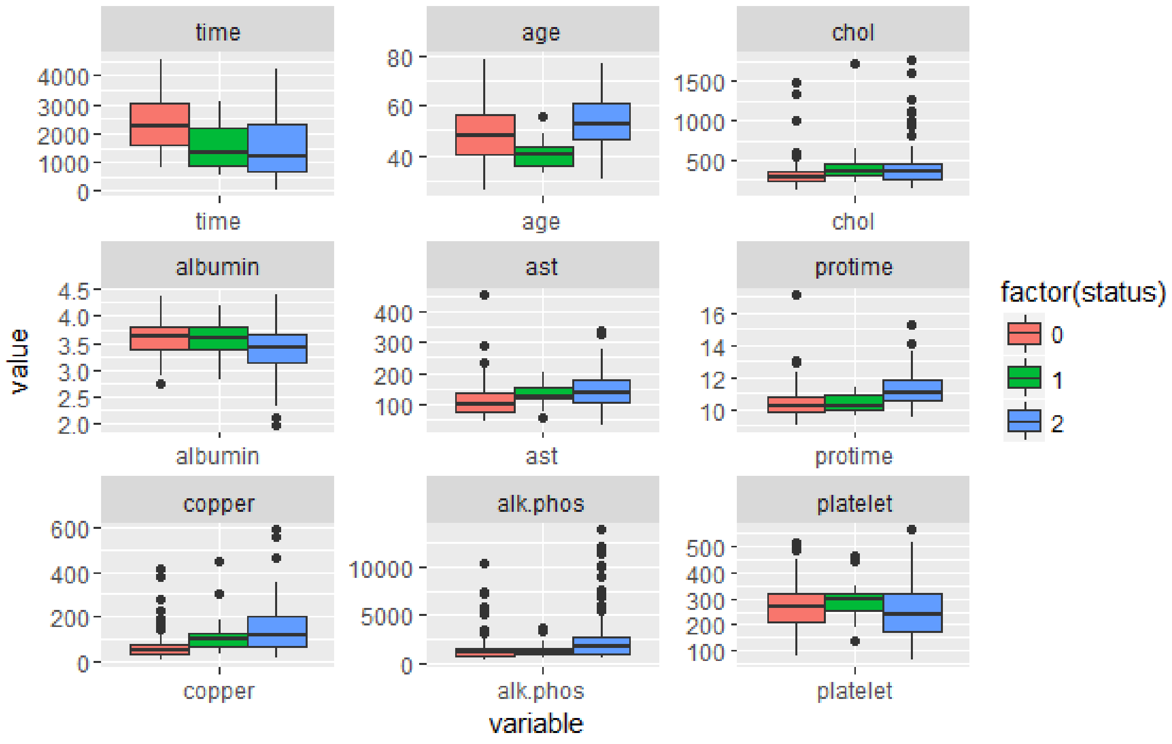


Figure.5 : Boxplot for a selection of covariates with respect to status. One can notice quite significant 'relationship' of some explanatory variables with respect to the 'status' : e.g. Copper, albumin, time, protime.

SDA Methods explored and their respective Aims :

Methods (Figure)	Purpose	Formula (from pdf)	Status
- **Survival time** curve **(Fig.4)**	At a given time *t* , what is the probability of survival	$S(t) = \exp\{-\int_0^t \lambda(x)dx\}.$	Yes tested
- Survival time function with respect treatment and sex separately **(Fig.5)**	Probability of failure time *probabilty of survival Time T being larger than a given time-lag t*	$S(t)=P(T>t)$	Yes
- **Kaplan-Meier estimator** (KME) non-parametric **(Fig.6)**	Estimate of survival rate at time *t* *considering the number of failure at times ti and the number of patients ar risk at the same time*	$S(t)=\prod_{ti<t}(1-\frac{\mathrm{di}}{\mathrm{ni}})$	Yes
CI 95% for Kaplan-Meier estimator			

- Nelson-Aalen (N-A) **non-parametric**	Consider estimating the cumulative hazard Λ(t). A simple approach is to start from an estimator of S(t) and take minus the log. An alternative approach is to estimate the cumulative hazard directly using the NelsonAalen estimator [ref.5]	$\hat{H}_{NA}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$ $\hat{S}_{NA}(t) = e^{\hat{H}_{NA}(t)}$	Yes
- Probability Description of Survival Data with **Maximum Likelihood Estimator**	Suppose then that we have n units with lifetimes governed by a survivor function S(t) with associated density f(t) and hazard λ(t). [Ref.9]	$L_i = f(t_i) = S(t_i)\lambda(t_i)$	No
- **Cumulative Hazard** figures **(Fig.6b)**	Given the hazard, we can always integrate to obtain the cumulative hazard and then exponentiate to obtain the survival function	$\Lambda(t) = \int_0^t \lambda(x)dx.$	Yes
- **Log-Rank Test (Fig.7)** the mostly used test in SDA	Test equality, *in terms of Hypothesis test, of 2 survival models (taking into account different covariates)*		Yes
- **Complementary log-log** plot (Fig.8) (differs from plot scale log-log)	How the proportional hazards assumption holds ?		
- **Cox PH semi-parametric approach** **it is like a Linear Regression** **But in the framework of SDA**	A continuous predictor vs a right-censored time-to-failure The Cox proportional hazards will show the increased rate of having an event in one curve versus the other. [Ref.9] The regression parameters are estimated by maximizing the partial log-likelihood defined by ℓ=∑flog(exp(β′xf)∑r(f)exp(β′xr))ℓ=∑flog(exp(β′xf)∑r(f)exp(β′xr))	$h_i(t) = h_0(t)\exp(x'_i\beta)$	Yes
- **Cox penalized**			No
- **Stratified Cox Regression**	This enables to explore the impact of one or several covariates on the Survival time with respect to a stratified covariate (age) for a refined vision.		Yes on categorical variable ageGoups
- Non-Parametric Maximum Likelihood		L = Ym i=1 [S(t(i−1)) − S(t(i))]diS(t(i)) ci ,	To be tested later

PART.2 : Univariate Analysis : Kaplan-Meier, Nelson-Aalen, Log-Rank Test, Cox PH.

3. Non-Parametric SDA methods

3.1 Survival Time curve

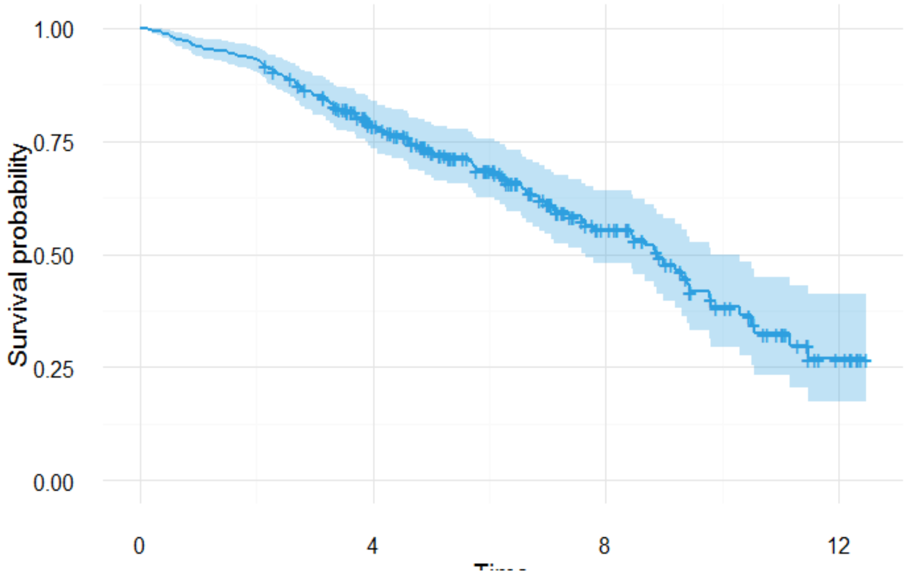


Figure.5 : Survival Time curve. The computed mean overall survival time is : 8.5 years.
The shaded area represent the 1 sigma ; the cross represent the censored data.

Interpretation : In *overall survival* curves, the event of interest is death from any cause. This provides a very broad, general sense of the mortality of the groups. [ref. 9]

At 12 years, the KME estimator gives a 25% survival rate.

3.2 Kaplan-Meier Estimator for univariate impact assessment

Interpreting the Kaplan-Meier curves : The non-continuous nature of the Kaplan-Meier curve emphasizes that they are not smooth functions, but rather step-wise estimates; thus, calculating a point survival can be difficult. The following is an example of a rough estimate of point survival;

(Figure.6) The cumulative probability of surviving a given time is seen on the Y-axis. For example, if you are in Group 1=D-Penicill, the probability of surviving 2 years is above 80%; conversely, if you are in Group 3=not randomized, the probability of surviving the same time is slightly more than 70%. It is obvious that the steepness of the curve is determined by the survival durations [Ref.9]

3.2.1 KME univariate Treatment **(0: not randomized; 1:D-penicillamin ; 2:placebo)**

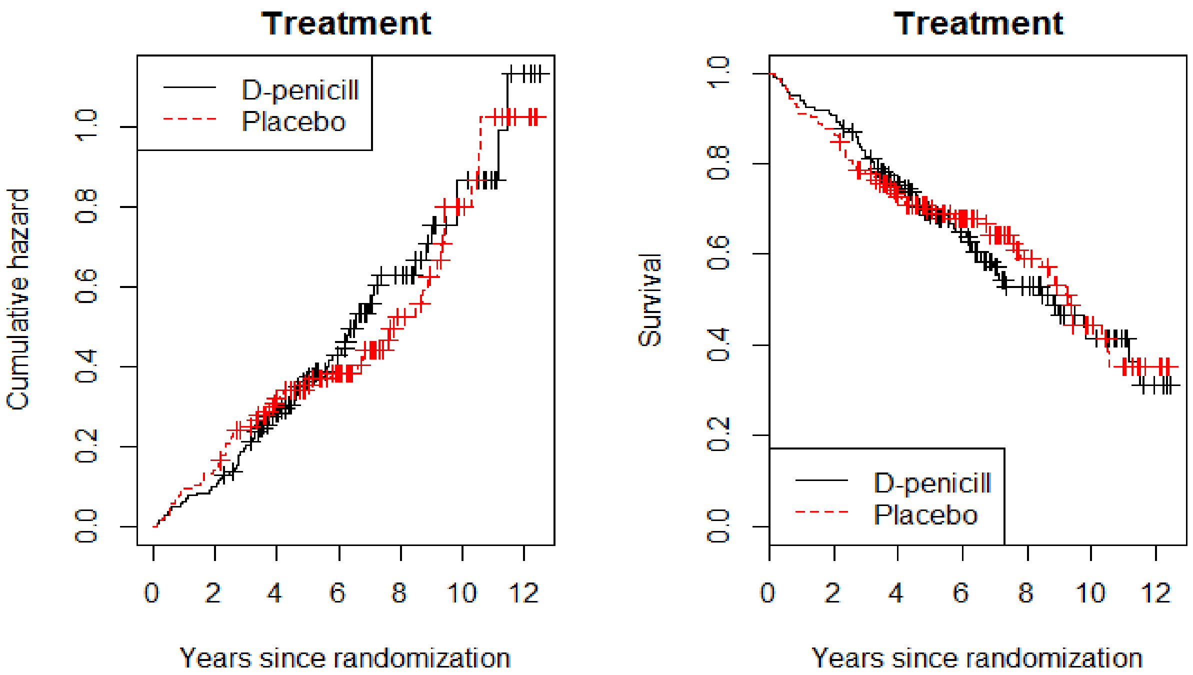


Figure.6 : Survival time curves w/r treatment (right) and Cumulativ Hazard (left).
Red stands for treatment D-penicillamine.

Result on KME univariate for Treatment : Strangely enough, survival curve estimated by Kaplan-Meier do **not exhibit clear evidence of improvement from D-Penicillamin vs Placebo.**

- **This preliminary result is to be confirmed by the CoxPH exp(coef) on treatement covariate.**

R code:

```
fit.KM.trt = survfit(responseSurv ~ factor(trt), data=mayoBiliary)
plot(fit.KM.trt, col=1:3)
title(main="KM Survival estimator for the covariate treatment",bty="n",cex.main=0.8,xlab="Time(year)",ylab="S(t)")
legend("bottomleft", col = 1:3, lty = 1, legend = c("not randomised","D-pen","Placebo"),bty="n")
fit.KM.sex = survfit(responseSurv ~ sex, data=mayoBiliary)
plot(fit.KM.sex, col=1:2)
title(main="KM Survival estimator for the covariate sex",bty="n",cex.main=0.8,xlab="Time(year)",ylab="S(t)")
legend("bottomleft", col = 1:2, lty = 1, legend = c("m","f"),bty="n")
```

3.2.2 KME univariate Sex

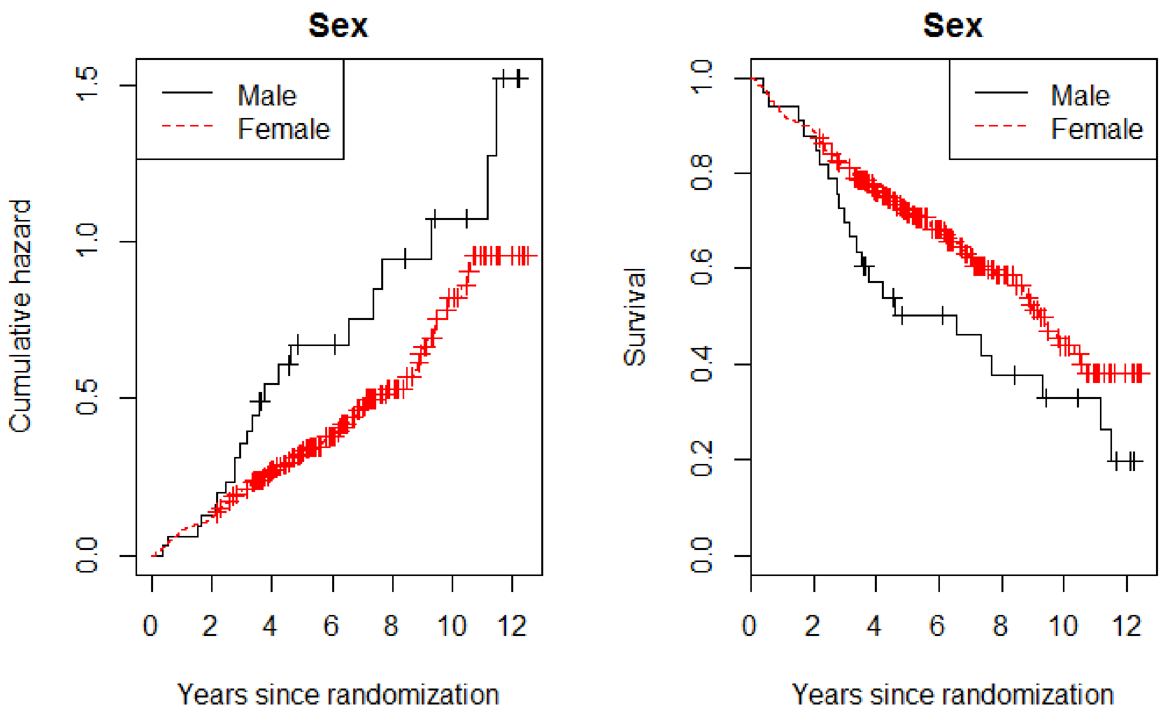


Figure.6b same as Figure6 for the right-hand side but from the reference code [Ref.7]

Result on KME univariate for Gender : survival curve estimated by Kaplan-Meier shows a string discrepancy associated with the 'sex' covariate.

- **This preliminary result is to be confirmed by the CoxPH exp(coef) on 'sex' covariate as shown in section 3.3.**

3.2.3 KME univariate Ascites

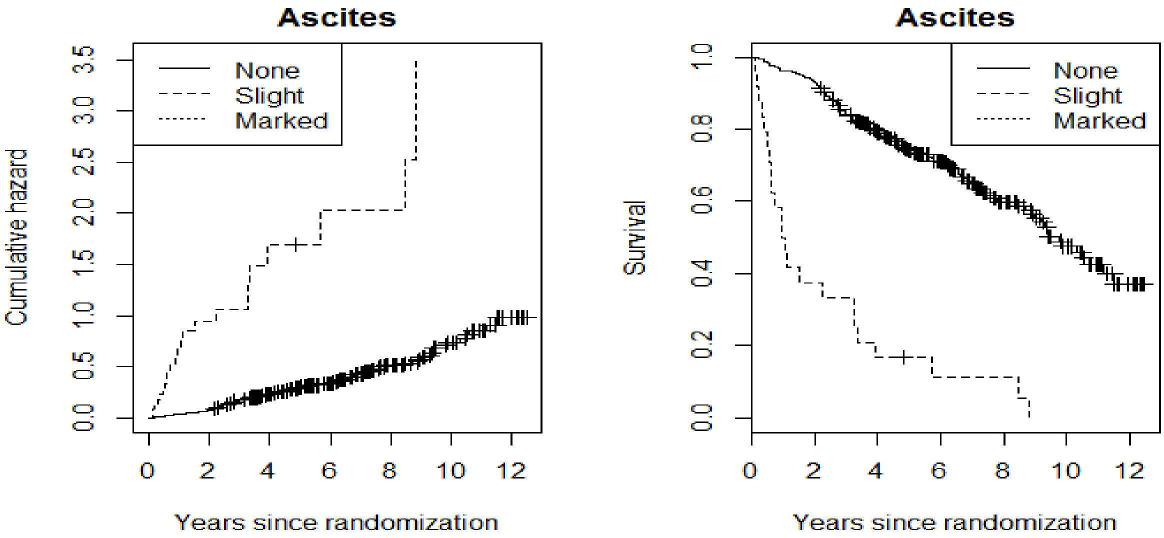


Figure.6c same as Figure6 for the right-hand side but from the reference code [Ref.7]

3.2.4 KME univariate Age intervals Strata (from Ref.7)

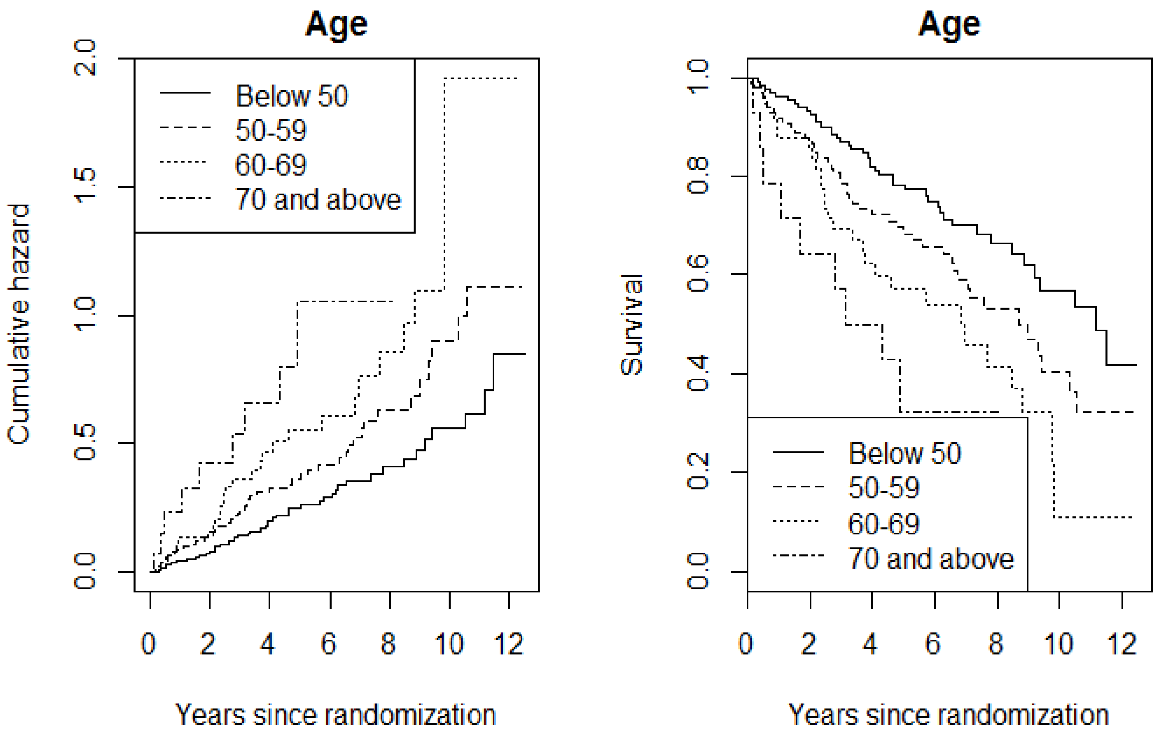


Figure.7 CHF (left) and Survival Time (right) estimated by KME with stratification on Age.

Interpretation : the KME method for **univariate stratitfied Age** enables to distinguish expected behavior with respect to the Age group. It clearly denotes a stronger cumulative hazard in order with respect to age groups. At 6-year survival time, for Agegroup < 50years, the survival rate is ~ 70%, whereas for the AgeGroup >70years, the survival rate drops to ~30%.

The **Cumulative Risk** can be interpreted as the cumulative force of mortality. Otherwise, it corresponds to the number of events that would be expected for each individual at time T if the event was a repetitive process.

3.3 Statistics associated with Kaplan-Meier survival curves univariate

3.3.1 K-M method : 95%CI Gender dependency

```
Call: survfit(formula = Surv(time, status) ~ sex, data = mayoBiliary,
    conf.type = "plain")
```

```
        n events median 0.95LCL 0.95UCL
sex=m  33     22   6.53    3.33    11.2
sex=f 260    103   9.19    8.46    10.5
Call:
survdiff(formula = Surv(time, status) ~ sex, data = mayoBiliary)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=m  33       22     14.6     3.781      4.33
sex=f 260      103    110.4     0.499      4.33

Chisq= 4.3  on 1 degrees of freedom, p= 0.0375
```

Results KME statistics 95%CI and log-rank test for covariate sex

The 95% Confidence interval of survival time for (a) males is [3.3, 11.2] years
when it reaches [8.4 , 10.5] for (b) female group of controlled patients (censored and not censored).

3.3.2 K-M method ; 95%CI Treatment dependency

```
Call: survfit(formula = Surv(time, status) ~ trt, data = mayoBiliary,
    conf.type = "plain")

n events median 0.95LCL 0.95UCL
trt=1 148     65   8.82    6.57    11.5
trt=2 145     60   9.30    7.79    10.5

Chisq= 0.1  on 1 degrees of freedom, p= 0.788
```

Not a significant difference on Treatment alone.

3.4 Kaplan-Meier estimator for Survival Time w/r histologic stage of disease

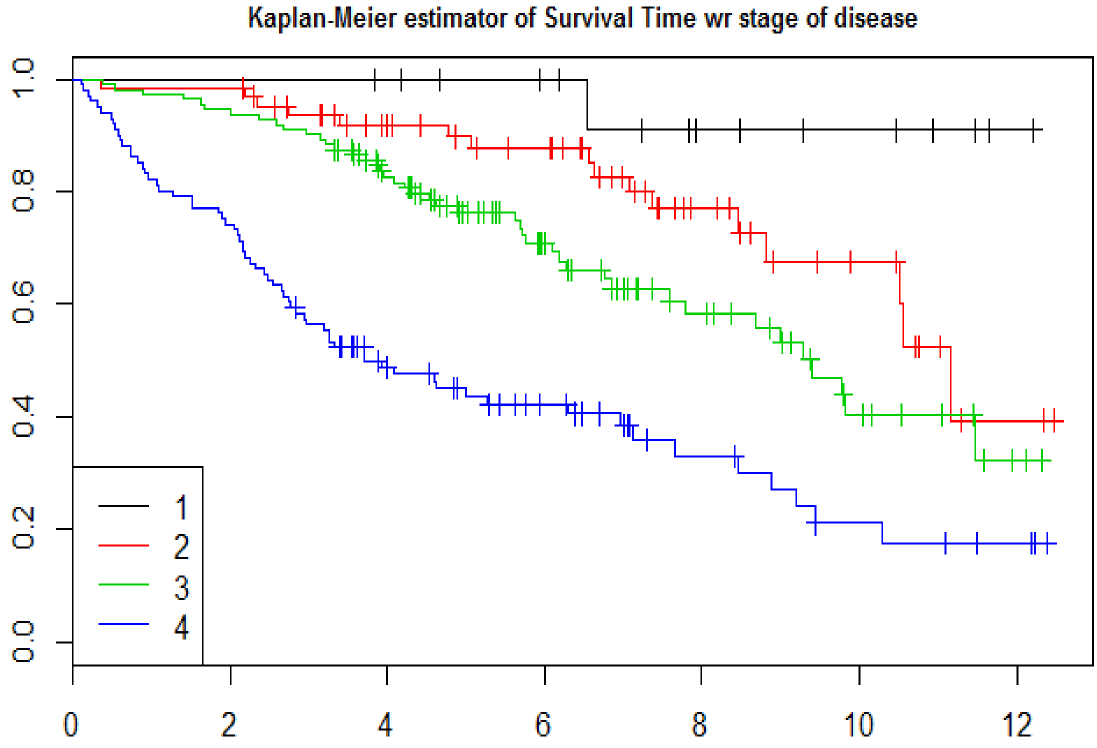


Figure.8 KM survival time estimated for each **stage of disease**
the histological scale is as follow : 1-> up to 4->

Result on KME :

- One can observe from Figure.8 a **strong discrepancy in KME estimation of Survival Time with respect to the histological stage of disease.** At 6-years survival time, it goes from >84% in class #2, down to 45% in class 4.
- The K-M model also infers statistically significant impacts of covariates : ageGroup, Ascites, Gender.

3.4 Nelson-Aalen for univariates : Treatment, Gender, Age, Edema

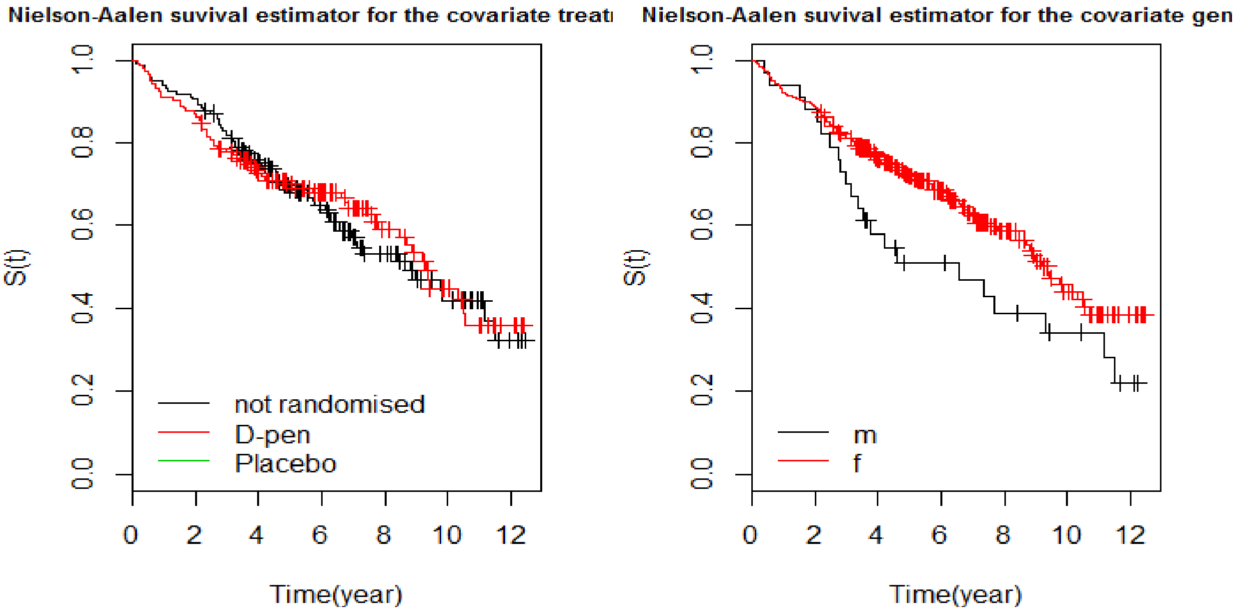


Figure.9 Nelson-Aalen Survival Time estimation curve for a univariate model (left: treatment) and (right : gender). The N-A Survival curves for Treatment and Gender are very simillar to those shown in Figure.6b. NB : Age and Edema Nelson-Aalen curves are shown in Appendix.2

Interpretation of Survival functions and curves :

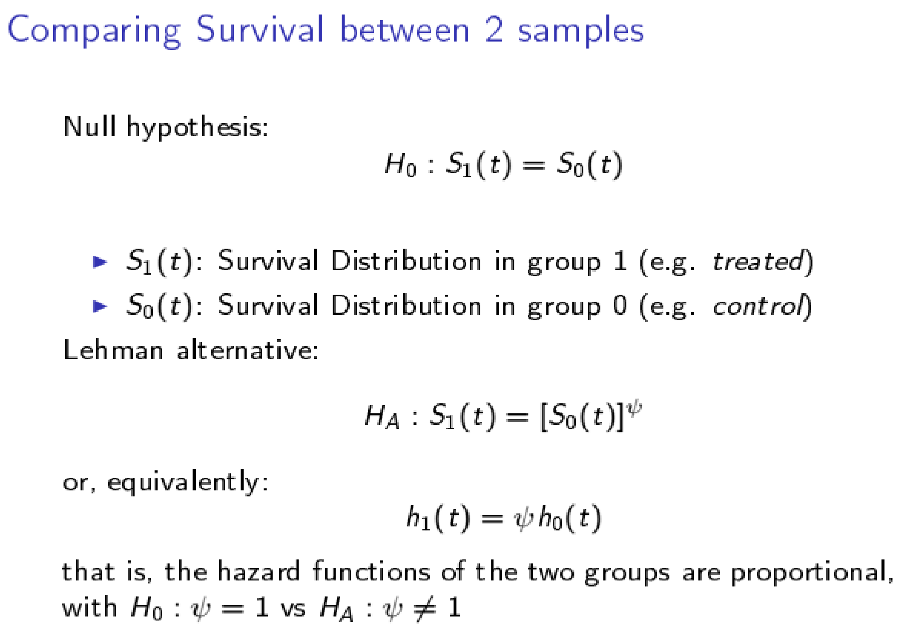
- **Survivor function** at the far right of a Kaplan-Meier survival curve should be interpreted cautiously, since there are *fewer* patients remaining in the study group and the survival estimates are not as accurate. [Ref.9]

- It should also be remembered that after the first patient is censored the **survival curve becomes an *estimate*,** since we do not know if censored patients would have experienced an event at some point later in their life. ***Thus, the more patients that are censored in a study (especially early in the study), the less reliable is the survival curve.*** Likewise, it is helpful to know *why* patients were censored. If many patients were censored in a given group(s), one must question how the study was carried out or how the type of treatment affected the patients. This stresses the importance of showing censored patients as tick marks in survival curves. [Ref.9]

4. Log-Rank test

Aim : to test if 2 groups from a given categorical covariate,differ significantly in terms of survival time.

Log-Rank is used to compare 2+ groups. χ^2 from the log-rank test will suggest whether two curves are statistically different. **[Ref.]**



Comparing Survival between 2 samples

Null hypothesis:

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

- $S_1(t)$: Survival Distribution in group 1 (e.g. *treated*)
- $S_0(t)$: Survival Distribution in group 0 (e.g. *control*)

Lehman alternative:

$$H_A : S_1(t) = [S_0(t)]^{\psi}$$

or, equivalently:

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

that is, the hazard functions of the two groups are proportional, with $H_0 : \psi = 1$ vs $H_A : \psi \neq 1$

4.1 Log-Rank test by Age groups :

```
Rcode :
# Log-rank test:
survdiff(Surv(time,status)~agegroup,data=mayoBiliary)
```

```
             n events median 0.95LCL 0.95UCL
agegroup=1 132     43  11.17    9.19      NA
agegroup=2  98     45   8.68    6.75   10.55
agegroup=3  49     28   6.85    3.72    8.82
agegroup=4  14      9   3.73    1.06      NA
Call:
survdiff(formula = Surv(time, status) ~ agegroup, data = mayoBiliary)

             N Observed Expected (O-E)^2/E (O-E)^2/V
agegroup=1 132       43    62.03      5.84    11.668
agegroup=2  98       45    41.89      0.23     0.347
agegroup=3  49       28    17.53      6.26     7.325
agegroup=4  14        9     3.54      8.39     8.702

 Chisq= 21  on 3 degrees of freedom, p= 0.000105
```

Interpretation : The **logrank tes**t is used to test the null hypothesis that there is no difference between the populations in the probability of an event (here a death) at any time point. The analysis is based on the times of events (here deaths). For each such time we calculate the observed number of deaths in each

group and the number expected if there were in reality no difference between the groups. **[Ref.11].** We can now use a χ2 test of the null hypothesis. The test statistic is the sum of (O - E)2/E for each group, where O and E are the totals of the observed and expected events.
Here (O-E)^2/E = 5,84 for Group1(age<50 years) vs 8,39 fr Group4 (age>70 years).
Stat(Group1)+Stat(Group1)=1.74.
From a table of the χ2 distribution we get P < 0.01, so that the difference between the groups is statistically significant. **??? to be verified.**

4.2 Log-Rank Treatment groups : D-Penicill vs Placebo

```
    [1] "survdiff(Surv(time,status)~trt,data=mayoBiliary)"
Call:
survdiff(formula = Surv(time, status) ~ trt, data = mayoBiliary)

        N Observed Expected (O-E)^2/E (O-E)^2/V
trt=1 148       65     63.5    0.0354    0.0722
trt=2 145       60     61.5    0.0366    0.0722
Chisq= 0.1  on 1 degrees of freedom, p= 0.788
```

Interpretation of log-rank test for Treatment :
as the p-value is very high, the impact of treatment is not statistically significant.

4.3 Log-Rank Gender groups :

Rcode :

survdiff(formula = Surv(time, status) ~ sex, data = mayoBiliary)

```
Call:
survdiff(formula = Surv(time, status) ~ sex, data = mayoBiliary)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=m  33       22     14.6     3.781      4.33
sex=f 260      103    110.4     0.499      4.33
Chisq= 4.3  on 1 degrees of freedom, p= 0.0375
```

Interpretation of log-rank test for Gender :
Here (O-E)^2/E = 3.8 for Group1(sex=male) vs 0.5 fr Group2 (sex=female).
Stat(Group1)+Stat(Group1)=1.74.
From a table of the χ2 distribution we get P <4%, so that the difference between the gender groups is statistically significant. **To be verified.**

5. Cox Proportional Hazard Univariate

The Cox proportional hazards will show the increased rate of having an event in one curve versus the other [Ref.9].

5.1 Cox PH of time-to-failure vs gender and treatement separately

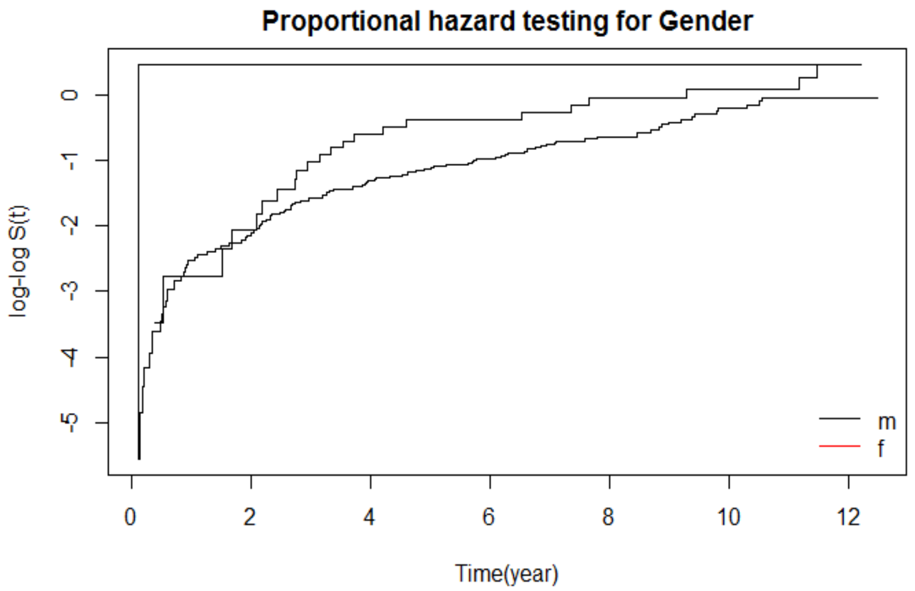


Figure.10 Proportational Hazard testing for gender (the smoother curve corresponds to 'female') in log-log(S(t))

Interpretation : to be documented.

R code :

```
plot(fit.KM.sex$time, log(-log(fit.KM.sex$surv)), col=c(1,2), type="s",xlab ="Time(year)", ylab = "log-log S(t)",
main = "Proportional hazard testing for Gender")
legend("bottomright", col = 1:2, lty = 1, legend = c("m","f"),bty="n")
print(dim(fit.KM.trt))
```

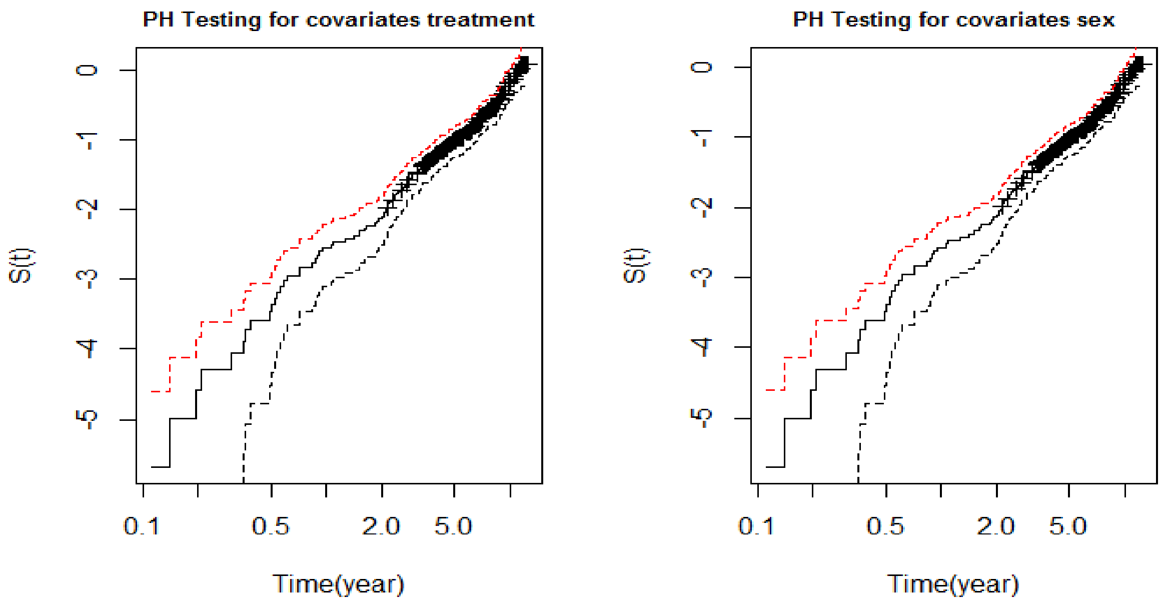


Figure11. Proportional Hazard Complementary log-log curves for treatement (left) and gender(right)

Interpretation on Proportional Hazard curves for univariate assessment :
The regression coefficients. **Beta coef** for Gender is equal to : -0. 49. The Beta coef for Treatment is .

5.2 Cox PH univariate of Gender

```
[1] "summary(cox.sex)"
Call:
coxph(formula = Surv(time, status) ~ sex, data = mayoBiliary)

  n= 293, number of events= 125

        coef exp(coef) se(coef)     z Pr(>|z|)
sex2 -0.4872    0.6143   0.2365 -2.06   0.0394 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
sex2    0.6143      1.628    0.3864    0.9766

Concordance= 0.532  (se = 0.015 )
Rsquare= 0.013   (max possible= 0.987 )
Likelihood ratio test= 3.82  on 1 df,   p=0.05064
Wald test            = 4.24  on 1 df,   p=0.0394
Score (logrank) test = 4.33  on 1 df,   p=0.03754
```

Interpretation of exp(coef) per variable from the statistics of Cox PH in terms of Hazard Ratios :

Hazard ratios. The exponentiated coefficients (exp(coef) = exp(-0,487) = 0,61), also known as *hazard ratios*, give the effect size of covariates. For example, being female (sex=2) reduces the hazard by a factor of 0.61, or 39%. Being female is associated with good prognostic.
The p-value associated with the Cox-PH statistical test is equal to 4% which yields in rejecting H0, e.g. the impact of covariate 'sex' is significant;

5.3 transformation on age variable and assessment

age → pspline(age)

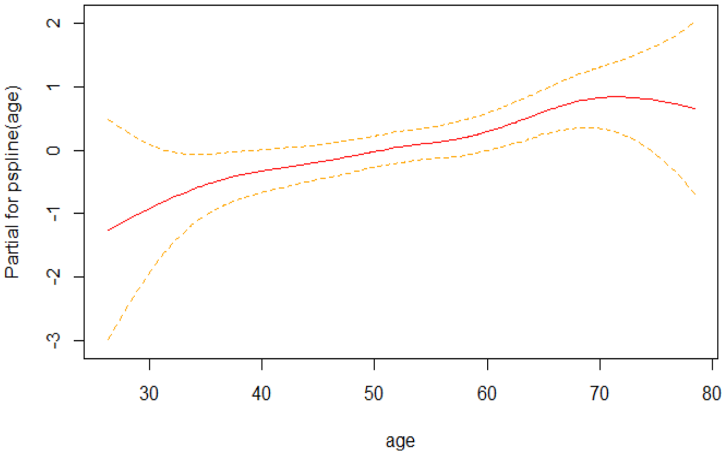


Figure.12 Assessment on transformation **pspline applied to 'age'** covariate

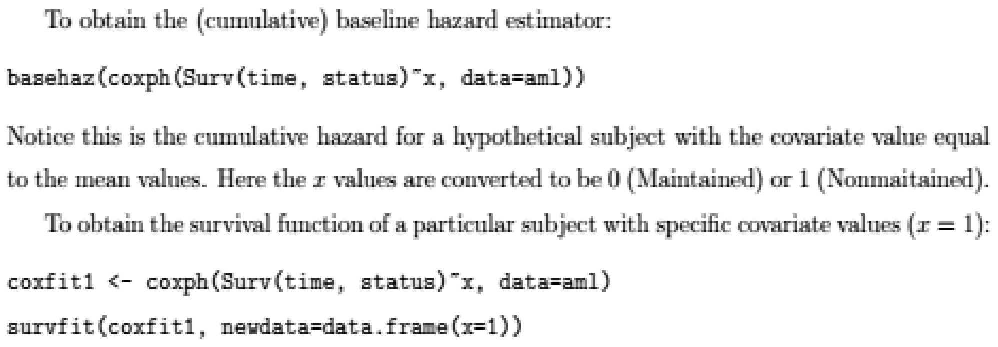
COX PH results on psline(Age) and Age 10 groups :

```
coef se(coef)      se2    Chisq   DF     p
pspline(age), linear 0.01294  0.00816 0.00816 2.51203 1.00 0.11
pspline(age), nonlin                          0.86135 3.04 0.84

Iterations: 4 outer, 10 Newton-Raphson
     Theta= 0.747
Degrees of freedom for terms= 4
Likelihood ratio test=3.85  on 4.04 df, p=0.433
  n=161 (232 observations deleted due to missingness)  #could not fixed this bug


coxph(formula = Surv(time, status) ~ age10, data = mayoBiliary)

  n= 161, number of events= 161
   (232 observations deleted due to missingness)

         coef exp(coef) se(coef)     z Pr(>|z|)
age10 0.12875   1.13740  0.08261 1.559    0.119

      exp(coef) exp(-coef) lower .95 upper .95
```

```
age10     1.137     0.8792    0.9674     1.337
```

5.4 Cumulative Baseline Hazard estimator

From rsurvTutorial :



To obtain the (cumulative) baseline hazard estimator:

```
basehaz(coxph(Surv(time, status)~x, data=aml))
```

Notice this is the cumulative hazard for a hypothetical subject with the covariate value equal to the mean values. Here the x values are converted to be 0 (Maintained) or 1 (Nonmaitained).

To obtain the survival function of a particular subject with specific covariate values ($x = 1$):

```
coxfit1 <- coxph(Surv(time, status)~x, data=aml)
survfit(coxfit1, newdata=data.frame(x=1))
```

5.5 Cox PH for Ascites coef

```
[1] "summary(cox.asc)"
Call:
coxph(formula = Surv(time, status) ~ ascites, data = mayoBiliary)

  n= 293, number of events= 125 

          coef exp(coef) se(coef)     z Pr(>|z|)    
ascites 1.9993    7.3842   0.2377 8.413   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
ascites     7.384     0.1354     4.635     11.77

Concordance= 0.598  (se = 0.01 )
Rsquare= 0.15   (max possible= 0.987 )
Likelihood ratio test= 47.52  on 1 df,   p=5.447e-12
Wald test            = 70.78  on 1 df,   p=0
Score (logrank) test = 97     on 1 df,   p=0
```

Interpretation : Model coxph(formula = Surv(time, status) ~ treatemnt, data = mayoBiliary) : the covariate treatment does have a significant impact on the censored response survival time, as **p-values are very small.**

PART.3 : Multi-Variate Analysis with Cox PH for covariate interactions

6. Cox PH Multivariate

We fit a Cox model with all the categorical covariates of interest.In a second step, we will assess pairwise interactions.

R code

```
cox.all=coxph(Surv(time,status)~trt+sex+ascites+age10+biligroup, data=mayoBiliary)
summary(cox.all)
```

6.1 Group of categorical covariates of interest

coxph(formula = Surv(time, status) ~ trt + sex + ascites + age10 + biligroup, data = mayoBiliary)

```
         coxph(formula = Surv(time, status) ~ trt + sex + ascites + age10 + 
    biligroup, data = mayoBiliary)

  n= 293, number of events= 125 

              coef exp(coef) se(coef)     z Pr(>|z|)    
trt        0.30106   1.35129  0.19492 1.545  0.12246    
sex2       0.17675   1.19333  0.25455 0.694  0.48747    
ascites    1.33755   3.80968  0.26676 5.014 5.33e-07 ***
age10      0.24839   1.28196  0.09112 2.726  0.00641 ** 
biligroup2 0.45097   1.56983  0.38792 1.163  0.24502    
biligroup3 1.59884   4.94731  0.32393 4.936 7.98e-07 ***
biligroup4 2.63308  13.91656  0.32249 8.165 3.33e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
trt            1.351    0.74003    0.9222     1.980
sex2           1.193    0.83799    0.7246     1.965
ascites        3.810    0.26249    2.2585     6.426
age10          1.282    0.78006    1.0723     1.533
biligroup2     1.570    0.63701    0.7339     3.358
biligroup3     4.947    0.20213    2.6220     9.335
biligroup4    13.917    0.07186    7.3966    26.184

Concordance= 0.817  (se = 0.028 )
Rsquare= 0.419   (max possible= 0.987 )
Likelihood ratio test= 159.1  on 7 df,   p=0
Wald test            = 158.3  on 7 df,   p=0
Score (logrank) test = 230.4  on 7 df,   p=0
```

Interpretation of exp(coef) per variable from the statistics of Cox PH:

oefficient reads as a regression coefficient (in the context of the Cox model, described hereafter) and its exponential **exp(coef)** gives the hazard in the treatment group.

R code

~~~
# trtment and ascitesites:
```{r}
cox.trt.ascites=coxph(Surv(time,status)~trt+sex+ascites+age10+ bili+trt:ascites, data=mayoBiliary)
anova(cox.all,cox.trt.ascites)
~~~

6.2 First order interactions between any pair of two covariates

a. treatment and sex:

```
Analysis of Deviance Table
 Cox model: response is  Surv(time, status)
 Model 1: ~ trt + sex + ascites + age10 + biligroup
 Model 2: ~ trt + sex + ascites + age10 + bili + trt:sex
   loglik  Chisq Df P(>|Chi|)
1 -555.01
2 -572.32 34.627  1 3.994e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Call:
coxph(formula = Surv(time, status) ~ trt + sex + ascites + age10 +
    bili + trt:sex, data = mayoBiliary)

  n= 293, number of events= 125

             coef exp(coef) se(coef)      z Pr(>|z|)
trt      -0.27807   0.75724  0.45190 -0.615  0.53833
sex2     -0.89345   0.40924  0.72514 -1.232  0.21791
ascites   1.29680   3.65759  0.26357  4.920 8.65e-07 ***
age10     0.25335   1.28834  0.09228  2.745  0.00604 **
bili      0.13869   1.14877  0.01421  9.761  < 2e-16 ***
trt:sex2  0.37789   1.45920  0.49077  0.770  0.44131
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
    exp(coef) exp(-coef) lower .95 upper .95
trt         0.7572     1.3206    0.3123     1.836
sex2        0.4092     2.4435    0.0988     1.695
ascites     3.6576     0.2734    2.1819     6.131
age10       1.2883     0.7762    1.0752     1.544
bili        1.1488     0.8705    1.1172     1.181
trt:sex2    1.4592     0.6853    0.5577     3.818

Concordance= 0.819  (se = 0.028 )
Rsquare= 0.346   (max possible= 0.987 )
```

```
Likelihood ratio test= 124.4  on 6 df,   p=0
Wald test            = 171    on 6 df,   p=0
Score (logrank) test = 257.2  on 6 df,   p=0
```

Interpretation : This COX PH model attempts to assess first order interaction on pair : treatment:gender. The test is significant. It has a 45% positive impact on suvival time.

Rcode:

```
cox.trt.sex=coxph(Surv(time,status)~trt+sex+ascites+age10+ bili+trt:sex, data=mayoBiliary)
anova(cox.all,cox.trt.sex)
summary(cox.trt.sex)

pred.trt.sex = survfit(cox.trt.sex, data=mayoBiliary, type="aalen")

plot(pred.trt.sex, col=1:2, fun = "cloglog")
title(main="PH Testing for covariates treat&sex",cex.main=0.8,xlab="Time(year)",ylab="S(t)")
```

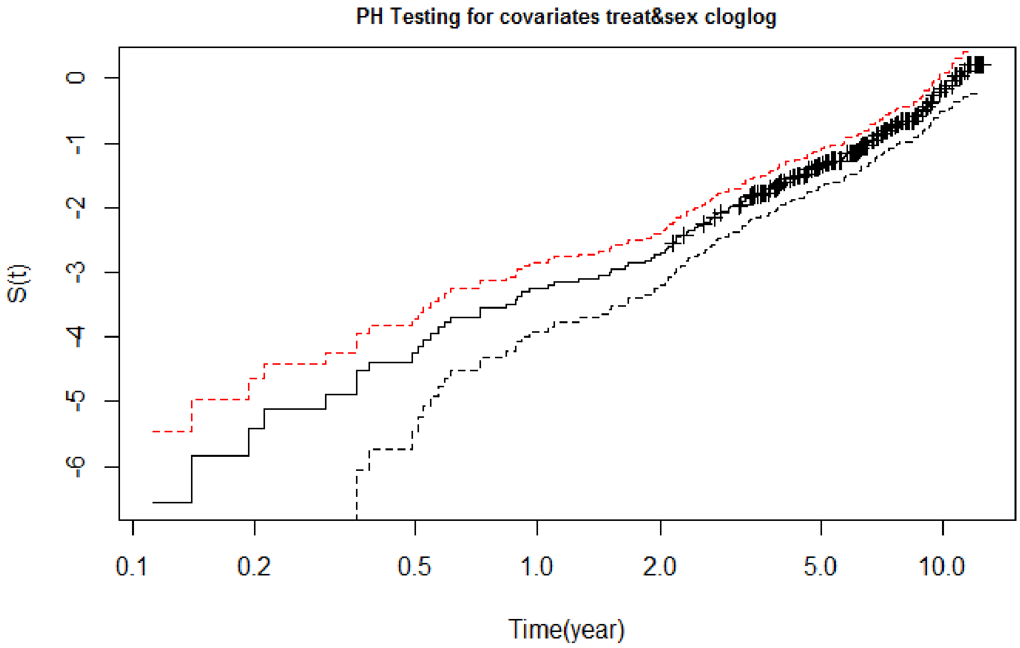


Figure.13 Proportional Hazard in complementary loglog scale for cox. Treatment and sex

b. Treatment and ascites

```
Analysis of Deviance Table
 Cox model: response is  Surv(time, status)
 Model 1: ~ trt + sex + ascites + age10 + biligroup
 Model 2: ~ trt + sex + ascites + age10 + bili + trt:ascites
   loglik  Chisq Df P(>|Chi|)
1 -555.01
2 -571.51 33.011  1 9.163e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Call:
coxph(formula = Surv(time, status) ~ trt + sex + ascites + age10 +
    bili + trt:ascites, data = mayoBiliary)

  n= 293, number of events= 125

                coef exp(coef) se(coef)      z Pr(>|z|)
trt         -0.07857   0.92444  0.20316 -0.387  0.69896
sex2        -0.37982   0.68398  0.24676 -1.539  0.12375
ascites      0.30769   1.36028  0.73511  0.419  0.67553
age10        0.26367   1.30170  0.09360  2.817  0.00485 **
bili         0.13972   1.14995  0.01421  9.833  < 2e-16 ***
trt:ascites  0.72636   2.06755  0.48078  1.511  0.13084
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

            exp(coef) exp(-coef) lower .95 upper .95
trt            0.9244     1.0817    0.6208     1.377
sex2           0.6840     1.4620    0.4217     1.109
ascites        1.3603     0.7351    0.3220     5.746
age10          1.3017     0.7682    1.0835     1.564
bili           1.1500     0.8696    1.1184     1.182
trt:ascites    2.0675     0.4837    0.8058     5.305
```

c. treatment & age10

```
coef exp(coef) se(coef)      z Pr(>|z|)
trt        -2.30718   0.09954  0.91096 -2.533  0.01132 *
sex2        0.24551   1.27827  0.25608  0.959  0.33770
ascites     1.48812   4.42877  0.26466  5.623 1.88e-08 ***
age10      -0.46365   0.62898  0.25715 -1.803  0.07138 .
biligroup2  0.39467   1.48390  0.38934  1.014  0.31073
biligroup3  1.64168   5.16384  0.32506  5.050 4.41e-07 ***
biligroup4  2.73485  15.40742  0.32633  8.381  < 2e-16 ***
trt:age10   0.50054   1.64961  0.17004  2.944  0.00324 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
trt          0.09954    10.0461    0.0167    0.5935
sex2         1.27827     0.7823    0.7738    2.1115
ascites      4.42877     0.2258    2.6364    7.4398
age10        0.62898     1.5899    0.3800    1.0412
```

```
biligroup2   1.48390     0.6739     0.6918     3.1828
biligroup3   5.16384     0.1937     2.7307     9.7649
biligroup4  15.40742     0.0649     8.1274    29.2083
trt:age10    1.64961     0.6062     1.1821     2.3021

Concordance= 0.817  (se = 0.028 )
Rsquare= 0.436   (max possible= 0.987 )
Likelihood ratio test= 167.7  on 8 df,   p=0
Wald test            = 158.9  on 8 df,   p=0
Score (logrank) test = 234.8  on 8 df,   p=0
```

Interpretation : treatment + ageGroup have a 65% positive impact on survival time. e.g. ? need to get more documented.

PART.4 : Machine Learning Train/Test with Random-Forest : Preliminary study

Introduction [Ref.12 : Github & Badr Tajini report]

Random Survival Forest (RSF) is a class of survival prediction models, those that use data on the subjects' life history (response) and their characteristics (predictive variables). In this case, it extends the RF algorithm for a target that is not a class, or a number, but a survival curve.

Rcode:

```
library("ggRandomForests")
#library("ggplot2")
library("dplyr")
data(pbc, package="randomForestSRC")
head(pbc[is.na(pbc$treatment),], n = 50)

#Please consider a more traditional train/test split, only with the 312 complete data:
pbc2 <- pbc[!is.na(pbc$treatment), ]

smp_size <- floor(0.70 * nrow(pbc2))

## set the seed to make your partition reproductible
set.seed(123)
train_ind <- sample(seq_len(nrow(pbc2)), size = smp_size)

pbc.train <- pbc2[train_ind, ]
pbc.test  <- pbc2[-train_ind, ]
```

```
nrow(pbc.train)
## [1] 218
nrow(pbc.test)
## [1] 94
##build model
rfsrc_pbc <- rfsrc(Surv(days, status) ~ .,
              data = pbc.train)
##plot the random survival forest
ggRFsrc <- plot(gg_rfsrc(rfsrc_pbc), alpha = 0.2) +
#scale_color_manual(values = strCol) +
theme(legend.position = "none") +
labs(y = "Survival Probability", x = "Time (Months)") +
coord_cartesian(ylim = c(-0.01, 1.01))
ggRFsrc
```

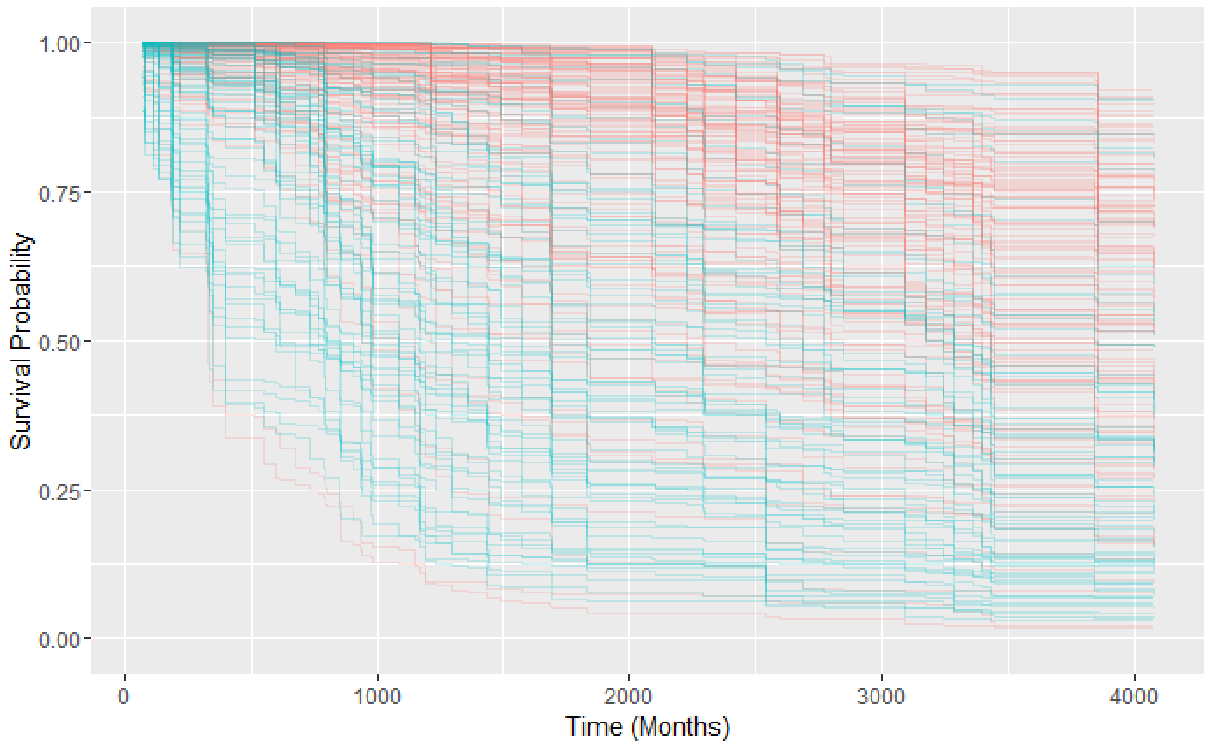


Figure.14 needs to get documented

```
Length Class      Mode
call           14   -none-     call
family          1   -none-     character
n               1   -none-     numeric
ntree           1   -none-     numeric
yvar            2   data.frame list
yvar.names      2   -none-     character
xvar           17   data.frame list
xvar.names     17   -none-     character
```

```
leaf.count     1000   -none-     numeric
proximity         0   -none-     NULL
forest           26   rfsrc      list
ptn.membership    0   -none-     NULL
membership        0   -none-     NULL
splitrule         1   -none-     character
inbag             0   -none-     NULL
var.used          0   -none-     NULL
imputed.indv     13   -none-     numeric
imputed.data     19   data.frame list
split.depth       0   -none-     NULL
node.stats        0   -none-     NULL
tree.err          1   -none-     logical
chf            6862   -none-     numeric
chf.oob           0   -none-     NULL
predicted        94   -none-     numeric
predicted.oob     0   -none-     NULL
```

Now, what to do with these promising results? I do not know yet. But I will study further the question on how RandForest perfoms in terms of survival time prediction for the Mayo Clinic right-censored data on PBC, compared with other methods previously seen.

SUMMARY

The present assignment for Survival Analysis on a classical right-censored clinical dataset. It allowed me to try several semi- and non-parametric methods we have seen during the course.
The main results I have obtained from non-parametric methods are :

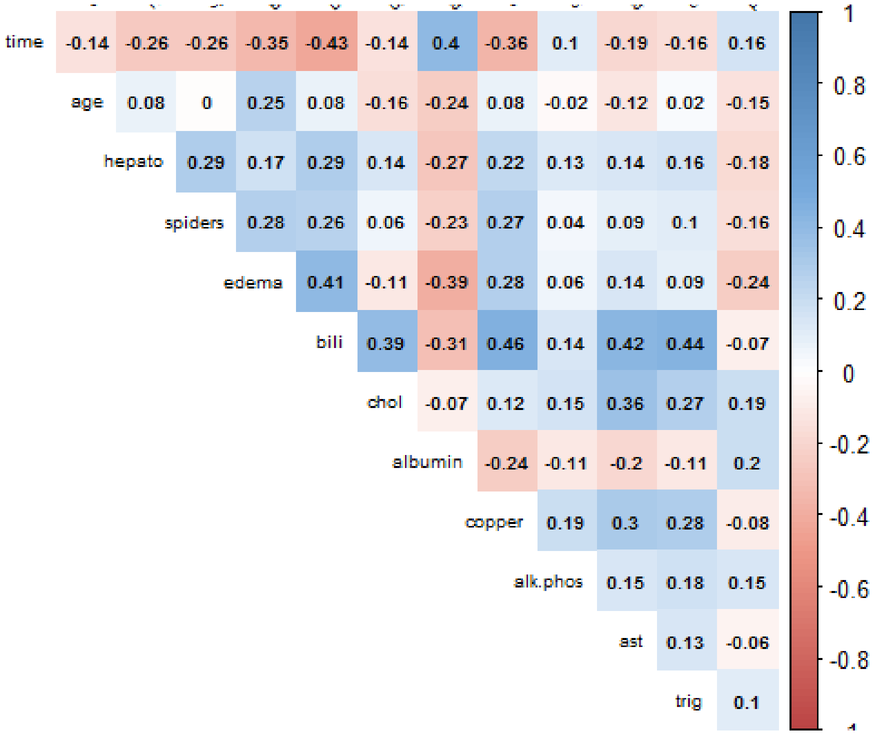
- Categorical covariates which have a significant impact of survival time are : Agegroups, Gender, Histologic stage, Ascites.
- Pairwise interactions have also been studied and significant interactions were found for : treatement+ageGroup, Treatement and ascites and reatment and gender.

Log-rank tests enabled to assess if 2 groups from a given categorical covariate are yielding to significantly different behaviors for survival times.

Survival data analysis is a quite complex domain which requires to understand several techniques, their scope of application, in addition to the data content itself.

Covariate interactions, for instance, need to be explored further, in order to have more clear ideas on the subject.

APPENDIX.1 : PBC correlogram



time	-0.14	-0.26	-0.26	-0.35	-0.43	-0.14	0.4	-0.36	0.1	-0.19	-0.16	0.16
age		0.08	0	0.25	0.08	-0.16	-0.24	0.08	-0.02	-0.12	0.02	-0.15
hepato			0.29	0.17	0.29	0.14	-0.27	0.22	0.13	0.14	0.16	-0.18
spiders				0.28	0.26	0.06	-0.23	0.27	0.04	0.09	0.1	-0.16
edema					0.41	-0.11	-0.39	0.28	0.06	0.14	0.09	-0.24
bili						0.39	-0.31	0.46	0.14	0.42	0.44	-0.07
chol							-0.07	0.12	0.15	0.36	0.27	0.19
albumin								-0.24	-0.11	-0.2	-0.11	0.2
copper									0.19	0.3	0.28	-0.08
alk.phos										0.15	0.18	0.15
ast											0.13	-0.06
trig												0.1

APPENDIX 2 More univariate results

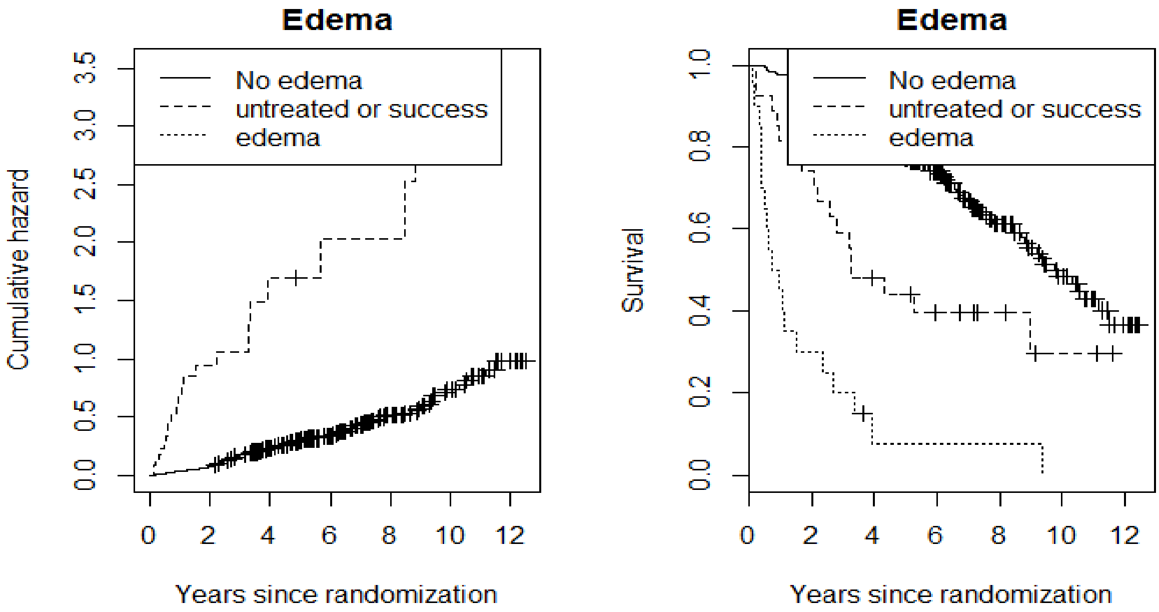


Figure.15 The result makes sense as the presence of edema results in shorter survival time

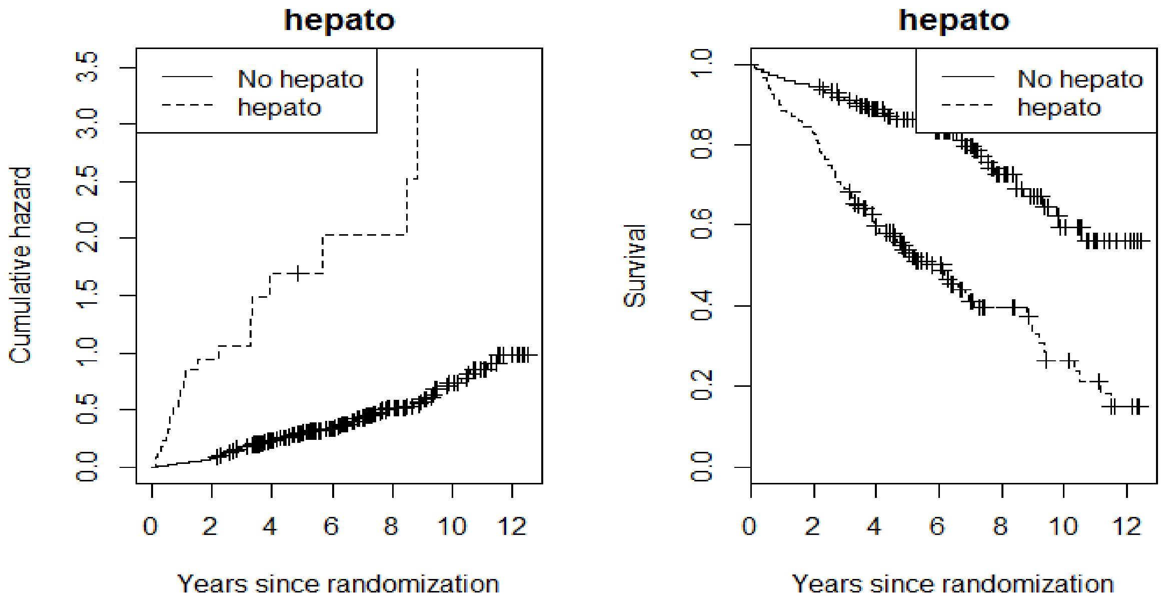


Figure.16 The result makes sense as the presence of hepatomegaly results in shorter survival time

APPENDIX.3 COXPH on treatment and age

coxph(formula = Surv(time, status) ~ trt, data = mayoBiliary)

```
  n= 161, number of events= 161 
   (232 observations deleted due to missingness)

      coef exp(coef) se(coef)     z Pr(>|z|)  
trt 0.2355    1.2656   0.1077 2.188   0.0287 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .95 upper .95
trt     1.266     0.7902     1.025     1.563
```

Concordance= 0.557 (se = 0.025)

```
Rsquare= 0.029   (max possible= 1 )
Likelihood ratio test= 4.71  on 1 df,   p=0.03007
Wald test            = 4.79  on 1 df,   p=0.02869
Score (logrank) test = 4.82  on 1 df,   p=0.02816
```

coxph(formula = Surv(time, status) ~ trt + age, data = mayoBiliary)

```
  n= 161, number of events= 161 
   (232 observations deleted due to missingness)

        coef exp(coef) se(coef)     z Pr(>|z|)  
trt 0.213050  1.237447 0.108085 1.971   0.0487 *
```

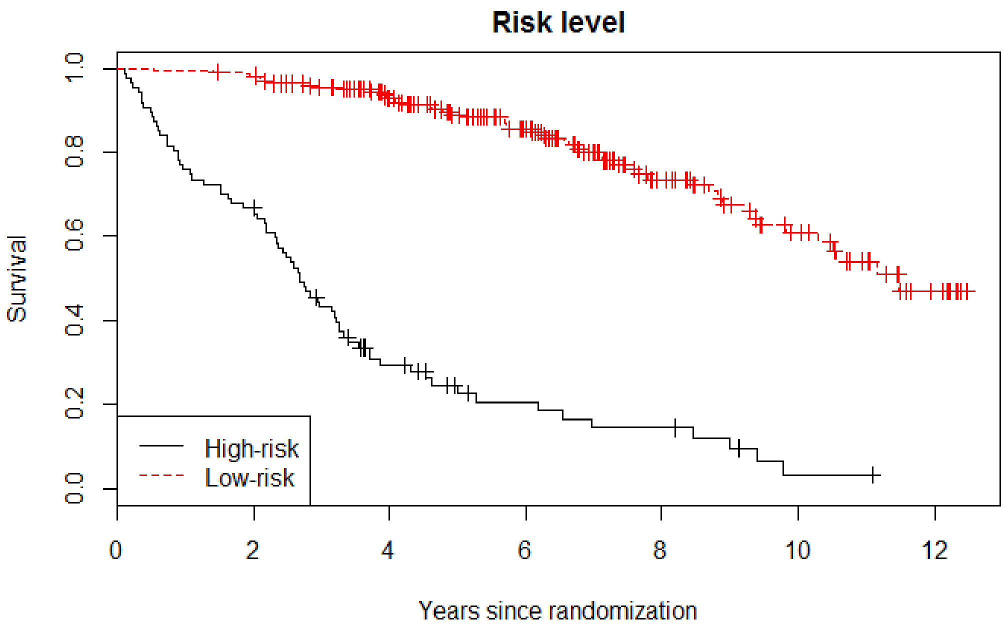
```
age 0.010513  1.010568 0.008384 1.254   0.2099
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .95 upper .95
trt     1.237     0.8081    1.0012     1.529
age     1.011     0.9895    0.9941     1.027

Concordance= 0.58  (se = 0.027 )
Rsquare= 0.038   (max possible= 1 )
Likelihood ratio test= 6.28 on 2 df,   p=0.04329
Wald test            = 6.42  on 2 df,   p=0.04036
Score (logrank) test= 6.46  on 2 df,   p=0.03961
```

Analysis of Deviance Table

```
 Cox model: response is  Surv(time, status)
•                 Model 1: ~ trt
•                 Model 2: ~ trt + age
   loglik     Chisq Df    P(>|Chi|)
1 -658.21
2 -657.43 1.5745  1    0.2096
```

Interpretation of Cox PH result for 2 models : as P(>|Chi|)>>5%, then we do not reject H0 : both models are not significantly different.

APPENDIX.4 mayo “low-risk”, “high risk” on 312 formal study participants in common with mayoBiliary dataset (+106 eligible nonenrolled subjects)



APPENDIX.5 Note on Log-Rank Test

The logrank test is based on the same assumptions as the Kaplan Meier survival curve[3]—namely, that censoring is unrelated to prognosis, the survival probabilities are the same for subjects recruited early and late in the study, and the events happened at the times specified. Deviations from these assumptions matter most if they are satisfied differently in the groups being compared, for example if censoring is more likely in one group than another.

The logrank test is most likely to detect a difference between groups when the risk of an event is consistently greater for one group than another. It is unlikely to detect a difference when survival curves cross, as can happen when comparing a medical with a surgical intervention. When analysing survival data, the survival curves should always be plotted.

Because the logrank test is purely a test of significance it cannot provide an estimate of the size of the difference between the groups or a confidence interval. For these we must make some assumptions about the data. Common methods use the hazard ratio, including the Cox proportional hazards model, which we shall describe in a future Statistics Note.

References

1. Extending the Cox Model – Thierry M. Therneau, Technical Report N° 58, 11/1996
2. All SDA exercise corrections by Pr. A. di Narzo
3. https://www.r-bloggers.com/outlier-detection-and-treatment-with-r/
4. http://publish.illinois.edu/spencer-guerrero/2014/12/11/2-dealing-with-missing-data-in-r-omit-approx-or-spline-part-1/
5. https://fr.slideshare.net/ASQwebinars/application-of-survival-data-analysis-introduction-and-discussion
6. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/
7. http://data.princeton.edu/pop509/NonParametricSurvival.pdf
8. http://www.uio.no/studier/emner/matnat/math/STK4080/h14/r-trial-project.txt
(this one really spared me a few hours of head scratching...)
9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3932959/
10. http://stats.stackexchange.com/questions/4528/what-is-the-difference-between-the-coef-and-expcoef-output-of-coxph-in-r
11. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403858/
12. http://www.sthda.com/english/wiki/cox-proportional-hazards-model
One of top references but seen to late...

APPENDIX.6 Full Rmd code

~~~
---
title: "03-PBC_MFaivre_correction"
output: html_document
---
#http://www.uio.no/studier/emner/matnat/math/STK4080/h14/r-trial-project.txt

# Detailed commands to the trial project
# ========================================

# This only gives the R commands for the trial project.
#It is not a draft solution of the project.
# Some comments on how to write a project report was given at the lectures.

# Read the data into R and attach the survival library

#load libraries
```{r}
library(survival)
library(survivalROC)
library(asaur)
library(ggplot2)
library(glmnet)
library(MASS)
```



#load ... dataset
```{r}
mayoBiliary <- pbc
head(mayoBiliary)
names(mayoBiliary)
summary(mayoBiliary)
print('str(mayoBiliary)')
str(mayoBiliary)

NA.hepato <- is.na(mayoBiliary$hepato)
# print(NA.hepato)
###dataHouse$bin.medv[is.na(dataHouse$bin.medv)] = nb_levels
missing.hepato = mayoBiliary$hepato[NA.hepato]
# print(mayoBiliary$status[NA.hepato]) ##values 0,1 or 2 met
```
##Data cleaning: filter out <NA> records
```{r}

```
#//////////////////// EXPLORATORY (READ DATA) ////////////////////
#==========================================================================================
#=== 1) Exploratory analysis : overall survival curve with right-censored status  ====
#==========================================================================================
#=== 1.a) summary
~~~

~~~~
```{r}
summary(mayoBiliary)

colnames <-names(mayoBiliary)
nb_obs   <-nrow(mayoBiliary)
nb_vars  <-ncol(mayoBiliary)
print(nb_vars)
print(nb_obs)
```
#=== 1.b) factor
```{r}
mayoBiliary$sex <- factor(mayoBiliary$sex)

```


#/////////////////// MISSING VALUES AND OUTLIERS ///////////////////

#====================================================================================
#=== 2. Data cleaning : replace missing data by imputation ; detection and filter-out outliers===
#====================================================================================
'''
=== 2.a) address missing values!
cannot just filter out missing which would result in data-loss
How can we trt missing data during survival data analysis?
The cook's distance for each observation i measures the change in Y Y^ (fitted Y) for all observations with and without the presence of observation i, so we know how much the observation i impacted the fitted values.
http://publish.illinois.edu/spencer-guerrero/2014/12/11/2-dealing-with-missing-data-in-r-omit-approx-or-spline-part-1/
https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/
'''

#///////////// MORE DATA PROCESSING FOR ANALYSIS /////////////////
#== 2.b STATUS : exclude transplant and as.logical
```{r}
# STATUS : exclude transplant from Dataset
mayoBiliary <- subset(mayoBiliary, status != 1) #exlcude transplant cases
#####
mayoBiliary <- transform(mayoBiliary, status = as.logical(status)) ##DO NOT FORGET THIS OR CENSORED DATA
#ARE FILTERED OUT !!!!
```
#=== 2.c TREATMENT
# trtMENT : NA-> 0 & filter out 'not randomized'
```{r}
mayoBiliary$trt[is.na(mayoBiliary$trt)] = 0   #'not randomized' flag
mayoBiliary <- subset(mayoBiliary, trt != 0)  #exlcude 'not randomized'
```


#=== . biligroup
```{r}
thresholds = quantile(mayoBiliary$bili)
print('thresholds')
print(thresholds)
~~~~

~~~
# mayoBiliary$biligroup=cut(cirrhosis$prot,breaks=c(0,49,69,89,150), labels=1:4)
mayoBiliary$biligroup=cut(mayoBiliary$bili,breaks=thresholds, labels=1:4)

str(mayoBiliary$biligroup)
```

#=== 2.d stats on missing data
```{r}
###install.packages("VIM")
library(VIM)

# need to address NA values !
n <- names(mayoBiliary)
formula1 <- as.formula(paste("time ~", paste(n[!n %in% c("protime","id")], collapse = " + ")))
mod <- lm(formula1, data=mayoBiliary)
cooksd <- cooks.distance(mod)
# print(cooksd)

# Dataset 1 : filter all records with NA rows
PBC_noNA = na.omit(mayoBiliary)
# print(dim(PBC_noNA))

# Dataset 2 : Proceed Imputation with mean value (as a first approx.) per each row
#Imputing the missing data
#The mice() function takes care of the imputing process
###install.packages("mice")
library(mice)
temp_NAimput <- mice(mayoBiliary, m=5, maxit=50, meth='pmm', seed=500)
PBC_NAimput  <- complete(temp_NAimput,1)
# summary(PBC_NAimput)
####densityPlot(temp_NAimput)

# Dataset 3 : apply a row-based spline to replace NA missing values
#PBC_NAspline =

#mayoBiliary = PBC_noNA
mayoBiliary <- PBC_NAimput
mean.survtime <- mean(mayoBiliary$time)
# print(mean.survtime/365.24)

# TIME IN YEARS
mayoBiliary$time <- mayoBiliary$time/365.24

# summary(mayoBiliary)
```
#/////////////////// MORE EXPLORATORY PLOTS //////////////////

#================================================
#=== 2.e scatter plots wr survival time       ===
#================================================
'''
```{r}
~~~

~~~
# 1. STime vs bilirubin
scatter.smooth(mayoBiliary$bili, mayoBiliary$time)

# 2. STime vs choloresterol
scatter.smooth(mayoBiliary$chol, mayoBiliary$time)

# 3. STime vs albumin
scatter.smooth(mayoBiliary$albumin, mayoBiliary$time)

# 4. STime vs copper
scatter.smooth(mayoBiliary$copper, mayoBiliary$time)

# 5. STime vs platelet
scatter.smooth(mayoBiliary$platelet, mayoBiliary$time)

# 6. STime vs alk.phos
scatter.smooth(mayoBiliary$alk.phos, mayoBiliary$time)
```

#=== 2.f Log-linearity of the numeric covariates Beta(t) per covariate in he model Cox has been checked along the way using splines
# [both for the unvariate and mulivariate Cox models (the latter commands not given here)]

# We also need to check for proportional hazards:
# We find an interaction between trtment and ascites, and between sex and age.
# To ease the interpretation of the interaction between sex and the numeric covariate age, it is useful
# to center age by subtracting 60 years (which is close to the mean age)
```{r}
mayoBiliary$cage10=(mayoBiliary$age-60)/10
cox.final=coxph(Surv(time,status)~trt+sex+ascites+cage10+bili+trt:ascites+sex:cage10, data=mayoBiliary)
print('summary(cox.final)')
summary(cox.final)

cox.test=cox.zph(cox.final,transform='log')
print(cox.test)
par(mfrow=c(2,2))
plot(cox.test)
```


#3. plot overall KME survival curve with censored data indicated
#================================================================
```{r}
plot(survfit(Surv(time, status) ~ 1, data=mayoBiliary), main='Overall KME survival time')

```

#/////////////////// UNVARIATE KAPLAN-MEIER & NELSON-AALEN ////////////////

#4. NON-PARAMETRIC METHOD UNIVARIATE
#===========================================================

# 4.a Simple univariate analyses for one covariate at a time
~~~

~~~
#============================================================

# Treatment
# ---------
# Nelson-Aalen and Kaplan-Meier plots stratified:
```{r}
par(mfrow=c(1,2))
fit.trt.naa=coxph(Surv(time,status)~strata(trt),data=mayoBiliary)
surv.trt.naa=survfit(fit.trt.naa)

plot(surv.trt.naa,fun="cumhaz", mark.time=T , col=1:3,lty=1:3,
    xlab="Years since randomization",ylab="Cumulative hazard", main="Treatment")
legend("topleft",legend=c("D-penicill","Placebo"),col=1:3,lty=1:3)

fit.trt.km=survfit(Surv(time,status)~trt,data=mayoBiliary, conf.type="plain")

plot(fit.trt.km, mark.time=T, col=1:3, lty=1:3,
    xlab="Years since randomization", ylab="Survival",main="Treatment")
legend("bottomleft",legend=c("D-penicill","Placebo"),col=1:3,lty=1:3)

# Estimates of five years survival probabilities with "plain" confidence intervals
# (alternatively we could have used the option "log-log")
print('summary(fit.trt.km,time=5)')
summary(fit.trt.km,time=5)

# Estimates of median survival time
print('fit.trt.km')
print(fit.trt.km)
#
# Alternatively, we may obtain the quartiles by the command:
print('quantile(fit.trt.km)')
quantile(fit.trt.km)

# Log-rank test:
print('survdiff(Surv(time,status)~trt,data=mayoBiliary)')
survdiff(Surv(time,status)~trt,data=mayoBiliary)
```

# Sex
# ---

# Nelson-Aalen and Kaplan-Meier plots stratified:
```{r}
par(mfrow=c(1,2))
fit.sex.naa=coxph(Surv(time,status)~strata(sex),data=mayoBiliary)
surv.sex.naa=survfit(fit.sex.naa)
plot(surv.sex.naa,fun="cumhaz", mark.time=T ,col=1:2, lty=1:2,
    xlab="Years since randomization",ylab="Cumulative hazard", main="Sex")
legend("topleft",legend=c("Male","Female"), col=1:2, lty=1:2)
fit.sex.km=survfit(Surv(time,status)~sex,data=mayoBiliary, conf.type="plain")
plot(fit.sex.km, mark.time=T, col=1:2, lty=1:2,
   xlab="Years since randomization", ylab="Survival",main="Sex")
legend("topright",legend=c("Male","Female"), col=1:2,lty=1:2)
~~~

~~~
# plot with error bars


# Estimates of five years survival probabilities with "plain" confidence intervals
print('summary(fit.sex.km,time=5)')
summary(fit.sex.km,time=5)

# Estimates of median survival time
print('fit.sex.km')
print(fit.sex.km)

# Log-rank test:
print('survdiff(Surv(time,status)~sex,data=mayoBiliary)')
survdiff(Surv(time,status)~sex,data=mayoBiliary)
```

# Ascites
# -------
```{r}
# Nelson-Aalen and Kaplan-Meier plots:
par(mfrow=c(1,2))
fit.asc.naa=coxph(Surv(time,status)~strata(ascites),data=mayoBiliary)
surv.asc.naa=survfit(fit.asc.naa)
plot(surv.asc.naa,fun="cumhaz", mark.time=T ,lty=1:3,
    xlab="Years since randomization",ylab="Cumulative hazard", main="Ascites")
legend("topleft",legend=c("None","Slight","Marked"),lty=1:3)
fit.asc.km=survfit(Surv(time,status)~ascites,data=mayoBiliary, conf.type="plain")
plot(fit.asc.km, mark.time=T, lty=1:3,
    xlab="Years since randomization", ylab="Survival",main="Ascites")
legend("topright",legend=c("None","Slight","Marked"),lty=1:3)

# Estimates of five years survival probabilities with "plain" confidence intervals
print('summary(fit.asc.km,time=5)')
summary(fit.asc.km,time=5)

# Estimates of median survival time
print('fit.asc.km')
print(fit.asc.km)

# Log-rank test:
print('survdiff(Surv(time,status)~ascites,data=mayoBiliary)')
survdiff(Surv(time,status)~ascites,data=mayoBiliary)
```


# Age
# ---
```{r}
# First we create a categorical variable for age group:
mayoBiliary$agegroup=cut(mayoBiliary$age,breaks=c(0,49,59,69,100), labels=1:4)

# Nelson-Aalen and Kaplan-Meier plots:
~~~

~~~
par(mfrow=c(1,2))
fit.age.naa=coxph(Surv(time,status)~strata(agegroup),data=mayoBiliary)
surv.age.naa=survfit(fit.age.naa)
plot(surv.age.naa,fun="cumhaz", mark.time=F ,lty=1:4,
    xlab="Years since randomization",ylab="Cumulative hazard", main="Age")
legend("topleft",legend=c("Below 50","50-59","60-69","70 and above"),lty=1:4)
#
fit.age.km=survfit(Surv(time,status)~agegroup,data=mayoBiliary, conf.type="plain")
plot(fit.age.km, mark.time=F, lty=1:4,
    xlab="Years since randomization", ylab="Survival",main="Age")
legend("bottomleft",legend=c("Below 50","50-59","60-69","70 and above"),lty=1:4)

# Estimates of five years survival probabilities with "plain" confidence intervals
summary(fit.age.km,time=5)

# Estimates of median survival time
print(fit.age.km)

# Log-rank test:
survdiff(Surv(time,status)~agegroup,data=mayoBiliary)
```
# Edema categorical univariate
# -------
```{r}
# Nelson-Aalen and Kaplan-Meier plots:
par(mfrow=c(1,2))
fit.edema.naa=coxph(Surv(time,status)~strata(edema),data=mayoBiliary)
surv.edema.naa=survfit(fit.asc.naa)

plot(surv.edema.naa,fun="cumhaz", mark.time=T ,lty=1:3,
    xlab="Years since randomization",ylab="Cumulative hazard", main="Edema")
legend("topleft",legend=c("No edema","untreated or success","edema"),lty=1:3)
fit.edema.km=survfit(Surv(time,status) ~ edema,data=mayoBiliary, conf.type="plain")

plot(fit.edema.km, mark.time=T, lty=1:3,
    xlab="Years since randomization", ylab="Survival",main="Edema")
legend("topright",legend=c("No edema","untreated or success","edema"),lty=1:3)

# Estimates of five years survival probabilities with "plain" confidence intervals
summary(fit.edema.km,time=5)

# Estimates of median survival time
print(fit.edema.km)

# Log-rank test:
survdiff(Surv(time,status)~edema,data=mayoBiliary)
```
#hepato
# -------
```{r}
# Nelson-Aalen and Kaplan-Meier plots:
par(mfrow=c(1,2))
fit.hepato.naa=coxph(Surv(time,status)~strata(hepato),data=mayoBiliary)
~~~

~~~~
surv.hepato.naa=survfit(fit.asc.naa)
plot(surv.hepato.naa,fun="cumhaz", mark.time=T ,lty=1:2,
    xlab="Years since randomization",ylab="Cumulative hazard", main="hepato")
legend("topleft",legend=c("No hepato","hepato"),lty=1:2)
fit.hepato.km=survfit(Surv(time,status) ~ hepato,data=mayoBiliary, conf.type="plain")
plot(fit.hepato.km, mark.time=T, lty=1:2,
    xlab="Years since randomization", ylab="Survival",main="hepato")
legend("topright",legend=c("No hepato","hepato"),lty=1:2)

# Estimates of five years survival probabilities with "plain" confidence intervals
summary(fit.hepato.km,time=5)

# Estimates of median survival time
print(fit.hepato.km)

# Log-rank test:
survdiff(Surv(time,status)~hepato,data=mayoBiliary)
```



#////////////////////// UNIVARIATE COX-REGRESSION ///////////////

# 5. Univariate Cox regressions
# ============================

# treatment
# ---------
```{r}
cox.trt=coxph(Surv(time,status)~trt,data=mayoBiliary)
print("summary(cox.trt)")
summary(cox.trt)
pred.trt = survfit(cox.trt, data=mayoBiliary, type="aalen")
```



# Sex
# ---
```{r}
cox.sex=coxph(Surv(time,status)~sex,data=mayoBiliary)
print('summary(cox.sex)')
summary(cox.sex)

pred.sex = survfit(cox.sex, data=mayoBiliary, type="aalen")

par(mfrow=c(1,2))
plot(pred.trt, col=1:2, fun = "cloglog")
title(main="PH Testing for covariates treatment",cex.main=0.8,xlab="Time(year)",ylab="S(t)")

plot(pred.sex, col=1:2, fun = "cloglog")
title(main="PH Testing for covariates sex",cex.main=0.8,xlab="Time(year)",ylab="S(t)")
```
~~~~

~~~
#Ascites:
# -------
```{r}
cox.asc=coxph(Surv(time,status)~ascites,data=mayoBiliary)
print('summary(cox.asc)')
summary(cox.asc)
```

# Age
# ---
```{r}
# For the numeric covariates age and prothrombin index, we need to decide how thay should be coded
# (as given on the data file, or suitably transformed, or grouped).

# To see how age should be coded, we fit a model using a spline for age:
cox.psage=coxph(Surv(time,status)~pspline(age),data=mayoBiliary)
print('cox.psage')
print(cox.psage)
par(mfrow=c(1,1))
termplot(cox.psage,se=T)

# Both the plot (from the templot-command) and the test (from the print-command) show that
# it is reasonable to assume a log-linear effect of age

# We will therefore fit a Cox model using age as a numeric covariate.
# It may be sensible to report the effect of age per 10 years, so we define a new covariate
# where age is given per 10 years and fit a Cox model with this covariate
mayoBiliary$age10= mayoBiliary$age/10
cox.age10=coxph(Surv(time,status)~age10,data=mayoBiliary)
print('cox.age10')
summary(cox.age10)
```

#6. KME PH univariate
#===================

#=== 6.a Plotting the CUMULATIVE HAZARD figures for Gender
```{r}
fit.KM.sex <- survfit(Surv(time, status) ~ factor(sex), data=mayoBiliary)
plot(fit.KM.sex$time, log(-log(fit.KM.sex$surv)), col=1:2, type="s",xlab ="Time(year)", ylab = "log-log S(t)", main = "Proportional hazard testing for Gender", lwd=1.4)
legend("bottomright", col = 1:2, lty = 1:2, legend = c("m","f"),bty="n")
```

#7. ROC "low-risk, high-risk" for 312 formal study participants from the matching 'mayo' dataset
#=====================================================================================
312 formal study participants in common with mayoBiliary dataset (+106 eligible nonenrolled subjects)
```{r}
library(survival)
library(survivalROC)

data(mayo)
##plot(survfit(Surv(time/365.25, censor) ~ 1, data=mayo))
~~~

~~~
print(str(mayo))
str(mayo)

ROC.4 = survivalROC(Stime = mayo$time,
            status= mayo$censor,
            marker= mayo$mayoscore4,
            predict.time = 365.25*5,
            method="KM")

ROC.5 = survivalROC(Stime = mayo$time,
            status= mayo$censor,
            marker= mayo$mayoscore5,
            predict.time = 365.25*5,
            method="KM")

'list from 2 models'
ROC = list(mayo4 = ROC.4, mayo5 = ROC.5)

cutoff = with(ROC$mayo5, min(cut.values[FP<=0.1]))

mayo$prediction = ifelse(mayo$mayoscore5 <= cutoff, "low_risk","high_risk")

'compare 2 groups : low_risk vs high_risk'
'Predict Survival Time taking into account censoring wr covariate=binary prediction{"high_risk","low_risk"}'
fit.KM = survfit(Surv(time/365.24,censor)~prediction, data=mayo)

plot(fit.KM,col=1:2,mark.time=T,lty=1:2,
     xlab="Years since randomization", ylab="Survival",main="Risk level")
legend("bottomleft",legend=c("High-risk","Low-risk"),col=1:2,lty=1:2)

'+ show censored data!'
'mayo$prediction = sapply(ROC, AUC)'
```


#////////////// MUTLI-VARIATE COX REGRESSION /////////////

#6. Multivariate Cox-PH
#=====================
# c. Multivariate Cox regression
# ============================
```{r}
# We then fit a Cox model with all the covariates
# cox.all=coxph(Surv(time,status)~trt+sex+ascites+age10+bili, data=mayoBiliary)
cox.all=coxph(Surv(time,status)~trt+sex+ascites+age10+biligroup, data=mayoBiliary)
summary(cox.all)
```
# In principle it may be the case that the coding of a numeric covariate that is appropriate
# for a univariate analysis, is not appropriate for a multivariate analysis and vice versa.
# But this does not seem to be the case here (commands not shown)

# We then check for first order interactions between any pair of two covariates
~~~

~~~
#        ----------------------------------------------------------------------

# treatment and sex:
```{r}
cox.trt.sex=coxph(Surv(time,status)~trt+sex+ascites+age10+ biligroup+trt:sex, data=mayoBiliary)
anova(cox.all,cox.trt.sex)
summary(cox.trt.sex)

pred.trt.sex = survfit(cox.trt.sex, data=mayoBiliary, type="aalen")

plot(pred.trt.sex, col=1:2)
title(main="PH Testing for covariates treat&sex",cex.main=0.8,xlab="Time(year)",ylab="S(t)")

plot(pred.trt.sex, col=1:2, fun = "cloglog")
title(main="PH Testing for covariates treat&sex cloglog",cex.main=0.8,xlab="Time(year)",ylab="S(t)")
```

# treatment and ascitesites:
```{r}
cox.trt.ascites=coxph(Surv(time,status)~trt+sex+ascites+age10+ biligroup+trt:ascites, data=mayoBiliary)
anova(cox.all,cox.trt.ascites)
summary(cox.trt.ascites)

pred.trt.asc = survfit(cox.trt.ascites, data=mayoBiliary, type="aalen")
```
# treatment and age:
```{r}
cox.trt.age=coxph(Surv(time,status)~trt+sex+ascites+age10+ biligroup+trt:age10, data=mayoBiliary)
anova(cox.all,cox.trt.age)
pred.trt.age = survfit(cox.trt.age, data=mayoBiliary, type="aalen")
summary(cox.trt.age)

par(mfrow=c(1,2))
plot(pred.trt.age, col=1:2)
title(main="PH Testing for covariates treat&age",cex.main=0.8,xlab="Time(year)",ylab="S(t)")

plot(pred.trt.asc, col=1:2)
title(main="PH Testing for covariates treat&ascites",cex.main=0.8,xlab="Time(year)",ylab="S(t)")

par(mfrow=c(1,2))
plot(pred.trt.age, col=1:2, fun = "cloglog")
title(main="PH Testing for covariates treat&age cloglog",cex.main=0.8,xlab="Time(year)",ylab="S(t)")

plot(pred.trt.asc, col=1:2, fun = "cloglog")
title(main="PH Testing for covariates treat&ascites cloglog",cex.main=0.8,xlab="Time(year)",ylab="S(t)")
```

# Sex and ascites:
```{r}
cox.sex.ascites=coxph(Surv(time,status)~trt+sex+ascites+age10+ biligroup+sex:ascites, data=mayoBiliary)
anova(cox.all,cox.sex.ascites)
pred.sex.asc = survfit(cox.sex.ascites, data=mayoBiliary, type="aalen")
summary(cox.sex.ascites)
~~~

~~~
```

# Sex and age:
```{r}
cox.sex.age=coxph(Surv(time,status)~trt+sex+ascites+age10+biligroup+sex:age10, data=mayoBiliary)
anova(cox.all,cox.sex.age)
pred.sex.age = survfit(cox.sex.age, data=mayoBiliary, type="aalen")
summary(cox.sex.age)

par(mfrow=c(1,2))
plot(pred.sex.age, col=1:2, fun = "cloglog")
title(main="PH Testing for covariates gender&age",cex.main=0.8,xlab="Time(year)",ylab="S(t)")

plot(pred.sex.asc, col=1:2, fun = "cloglog")
title(main="PH Testing for covariates gender&ascites",cex.main=0.8,xlab="Time(year)",ylab="S(t)")

# Plot the baseline survival function of this Cox PH model
install.packages("survminer")
library("survminer")
ggsurvplot(survfit(cox.sex.age), color = "#2E9FDF",
      ggtheme = theme_minimal(), title="Baseline survival function of Cox PH
model:Surv(time,status)~trt+sex+ascites+age10+bili+sex:age10")
```

# Sex and bilirubin:
```{r}
cox.sex.bili=coxph(Surv(time,status)~trt+sex+ascites+age10+biligroup+sex:biligroup, data=mayoBiliary)
anova(cox.all,cox.sex.bili)
summary(cox.sex.bili)
```
# ascitesites and age:
```{r}
cox.ascites.age=coxph(Surv(time,status)~trt+sex+ascites+age10+bili+ascites:age10, data=mayoBiliary)
anova(cox.all,cox.ascites.age)
summary(cox.ascites.age)
```
# Ascites and prothrombin:
# For checking interaction between ascites and prothrombin group, we need to merge the two highest
# prothrombin groups (since there is only one person with severe ascites in the highest prothrombin group)
# We find an interaction between treatment and ascites, and between sex and age.
# To ease the interpretation of the interaction between sex and the numeric covariate age, it is useful
# center age by subtracting 60 years (which is close to the mean age)
```{r}
mayoBiliary$cage10=(mayoBiliary$age-60)/10
cox.final=coxph(Surv(time,status)~trt+sex+ascites+cage10+biligroup+trt:ascites+sex:cage10, data=mayoBiliary)
summary(cox.final)
```

# Log-linearity of the numeric covariates has been checked along the way using splines
# [both for the unvariate and mulivariate Cox models (the latter commands not given here)]

# We also need to check for proportional hazards:
```{r}
~~~

~~~
cox.test=cox.zph(cox.final,transform='log')
print(cox.test)
par(mfrow=c(2,2))
plot(cox.test)
```



#/////////// MACHINE LEARNING : RANDOM FOREST ///////////

#7. Random Forest with radomforestSRC package
#==========================================

#for this part, we add the rm function to clean our memory because we don't
#need the latest objects to continue our study
```{r}
rm(list = ls())
##install.packages("ggRandomForests")
library("ggRandomForests")
#library("ggplot2")
library("dplyr")
data(pbc, package="randomForestSRC")
head(pbc[is.na(pbc$treatment),], n = 50)

#Please consider a more traditional train/test split, only with the 312 complete data:
pbc2 <- pbc[!is.na(pbc$treatment), ]

smp_size <- floor(0.70 * nrow(pbc2))

## set the seed to make your partition reproductible
set.seed(123)
train_ind <- sample(seq_len(nrow(pbc2)), size = smp_size)

pbc.train <- pbc2[train_ind, ]
pbc.test  <- pbc2[-train_ind, ]
nrow(pbc.train)
## [1] 218
nrow(pbc.test)
## [1] 94
##build model
rfsrc_pbc <- rfsrc(Surv(days, status) ~ .,
              data = pbc.train)
##plot the random survival forest
ggRFsrc <- plot(gg_rfsrc(rfsrc_pbc), alpha = 0.2) +
#scale_color_manual(values = strCol) +
theme(legend.position = "none") +
labs(y = "Survival Probability", x = "Time (Months)") +
coord_cartesian(ylim = c(-0.01, 1.01))
ggRFsrc
```

##test model - test data contains un-censored data
```{r}
~~~

```
test.pred.rfsrc <- predict(rfsrc_pbc, pbc.test,
na.action="na.impute") #added this so I get results for all test rows
#summary of our test model
summary(test.pred.rfsrc)
```