Drug D-penicillamine versus placebo and influence of some visible variables (age, hepato, ascites) on the patient survival probability for the pancreatic liver dataset pbc

Benjamin Lepers, DSTI

April 17, 2019

Abstract

We study a subset of the primary biliary cirrhosis (pbc) dataset by removing missing values and liver transplanted patients. A logrank test reveals that there is no statistical evidence of the efficiency of the D penicillamine on these patients (p = 0.56). Men have 67 % higher risk hazard than women ($p = 0.03$) and the group of > 50 years has higher risk than the < 50 years group ($p = 0.000138$). Visible symptomatic variables such as enlarged liver (hepato=1) and ascites (fluid in the abdomen) are also strong parameters that decreases the patient survival probability.

Contents

1 Introduction

The pbc dataset from the survival package of R contains 424 patients and 20 variables from the Mayo clinical trial in primary liver cirrhosis. This study was conducted bewtween 1974 and 1984 and met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. After removing all the missing values, we conduct this analysis on a subset of 258 patients and 20 features. We investigate the effect of the age, gender, ascites and hepato variables on the probability of survival.

2 Description of the data

The variables of the dataset are:

- age: in years
- albumin: serum albumin (g/dl)
- alk.phos: alkaline phosphotase (U/liter)
- ascites: presence of ascites, abnormal buildup of fluid in the abdomen (usually due to cirrhosis)
- ast: aspartate aminotransferase, once called SGOT (U/ml)
- bili: serum bilirunbin (mg/dl) (usually, with abscence of liver disease, bili is high level)
- chol: serum cholesterol (mg/dl)
- copper: urine copper (ug/day)
- edema: 0 no edema, 0.5 untreated or successfully treated
- 1 edema despite diuretic therapy
- hepato: presence of hepatomegaly (enlarged liver)
- id: case number
- platelet: platelet count
- protime: standardised blood clotting time
- sex: m/f
- spiders: blood vessel malformations in the skin
- stage: histologic stage of disease (needs biopsy)
- status: status at endpoint, 0/1/2 for censored, transplant, dead
- time: number of days between registration and the earlier of death,
- transplantion, or study analysis in July, 1986
- trt: 1/2/NA for D-penicillmain, placebo, not randomised
- trig: triglycerides (mg/dl)

2.1 Data set preparation

For simplicity, we remove all the rows of the dataset containing missing value, we also redefine the status variable as 0 for censored and 1 for death. We remove the patients that were transplanted (original status variable set to 1). The time variable is also converted in years. These tasks are made with following R commands below:

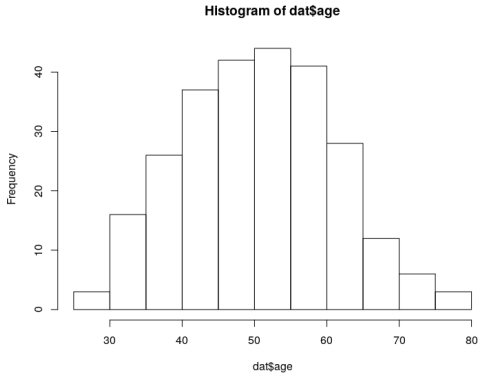

Figure 1: Distribution by age of the patients

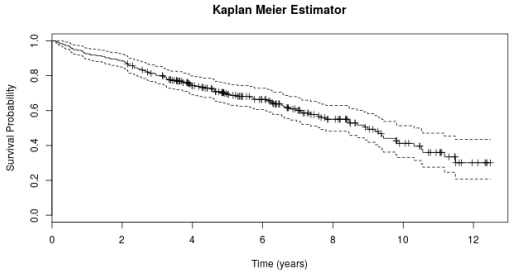

Figure 2: Kaplan meier estimator, survival probability versus time

```
# data set is pbc
dat <-pbc
a <- subset(dat, (dat$status == 0)|(dat$status == 2))
df <- mutate(a, status = as.integer(a$status == 2), id = as.character(id))
# remove all the 1's. all the rows with status = 1.
str(df)
dat <- df[complete.cases(df), ]
datyear <- mutate(dat, time = time/365)
```

2.2 Descriptive statistics

The dataset contains 31 men and 227 women. There are 147 censored patients and 111 events (deaths) occured during the 10 years period study. The age of the patients is ranging from 26 to 78 years. The age distribution follows approximatly a normal distribution and is shown in the figure 1.

2.3 Kaplan meier estimator

To take into account the censoring, we use the kaplan meier estimator. The following R command is used:

```
kmsurvival <- survfit(Surv(datyear$time,datyear$status) ~ 1,
 data = datyear)
 plot(kmsurvival, mark.time = TRUE, xlab =
 "Time (years)", ylab = "Survival Probability", main =
 'Kaplan Meier Estimator')
kmsurvival
# n  events  median 0.95LCL 0.95UCL
# 258.00  111.00    8.99    7.66   10.31
```

From the 258 patients recorded data, 111 events occured (111 deaths during the 10 years study). The last command gives a median survival time of 9 (7.6 - 10.3) years, in others words, 50 % of the patients died within 9 years. The survival curve with the confidence interval is shown in fig 2. 75 % of the patient are still alive within 3.9 years, 30 % of the patient are still alive within 11.4 years. $(S(0.75), S(0.5), S(0.3) = \{3.9, 9, 11.4\}$ years).

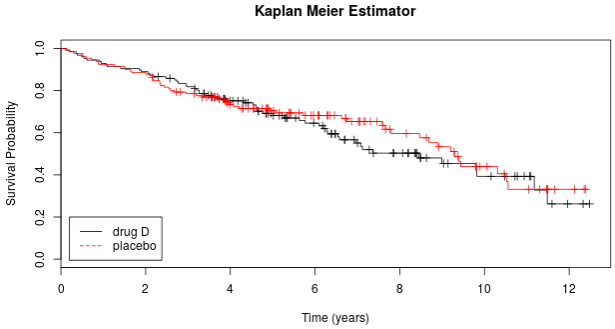

Figure 3: D-penicillamince versus placebo on the survival probability $p = 0.56$

3 Question

We would like to answer the question whether the drug D improves or not the survival probability of the patient. In a second step, we investigate the effect of the variables which are visible for the physician, ie the following variables: sex, age, ascites, hepato.

4 Method

The Kaplan Meier estimator is used to compare the survival probability between groups and to take into account the censored data. Then, the logrank tests is used to test the difference between the two groups (if the difference is significative or not). Finally, when possible, we will use a cox proportional hazard model to quantify the difference between the groups in terms of probability of survival.

5 Results

5.1 Comparaison between the drug D-penicillamine and the placebo

With the R survfit fonction, the kaplan meier estimator gives a median of $\{8.45, 9.3\}$ years for the drug D and for the placebo treatments respectively. Only the lower bound for the confidence interval is given, because the upper bound exceed the duration of the clinical study.

```
km.trt <- survfit(Surv(datyear$time,datyear$status) ~ datyear$trt,
 data=datyear)
km.trt
# datyear$trt=1 127      57   8.45    6.54      NA
# datyear$trt=2 131      54   9.30    8.47      NA
survdiff(Surv(datyear$time,datyear$status) ~ datyear$trt)
```

A log rank test gives a p value of $0.56 > 0.05$. Hence, there is no statistical evidence of the efficiency of the D penicillamine on these patients. The kaplan meier curves in figure 3 are quite close to each other (overlaping from 0 to 6 years and crossing around 10 years).

5.2 Comparaison by gender

Using the R survfit function, the kaplan meier estimator gives a median time of survival of 6.5 years (CI 3.16 - 11.5) for the men and 9.2 years (CI 8.45 - 10.6) for the women. The log rank test gives a p value of 0.03, which means that null hypothesis is rejected. So there is a significant difference between the male and women group in terms of survival probability

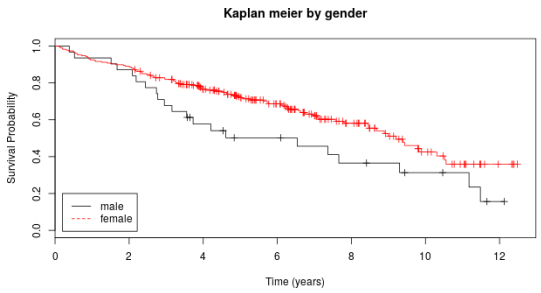

Figure 4: Survival probability, effect of Gender, $p = 0.03$

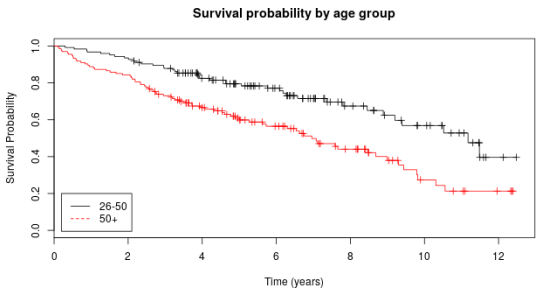

Figure 5: Survival probability for two age groups, $p = 0.000138$

(see figure 4). Because the 2 curves are almost parallel (the hazard risk function ratio between women and men is constant), we can use a cox proportional hazard model with the function coxph from R.

```
fit.cphsex <- coxph(formula = Surv(datyear$time,datyear$status)
                    ~ datyear$sex, data = datyear)
summary(fit.cphsex)
# exp(coef) exp(-coef) lower .95 upper .95
# datyear$sexf    0.5985      1.671    0.3711    0.9653
```

The risk hazard for men is 67 % higher than the risk for women.

5.3 Comparaison by age

A new categorical variable is added to the dataset by splitting the patients into groups with age higher or lower than 50 years old. As shown below, 124 patients are below 50 and 134 above 50.

```
datyear$ageCat <- cut(datyear$age, breaks = c(26, 50, Inf))
table(datyear$ageCat)
#  (26,50] (50,Inf]
#    124      134
```

Using Kaplan meier estimator, the median survival for the < 50 years group is about 11 years and 7 years for the group with age > 50. A log rank test give a p value of $p = 0.000138$, so the age is a significative variable for the survival time. The kaplan meier curves shown in fig 5 are almost parallel, so a cox proportional hazard model is used to quantify the difference of risk hazard for the two groups with the following R command:

```
fit.cpgroupage <- coxph(formula = Surv(datyear$time,datyear$status)
 ~ datyear$ageCat, data = datyear)
summary(fit.cpgroupage)
                          coef exp(coef) se(coef)     z Pr(>|z|)
datyear$ageCat(50,Inf] 0.7430    2.1023   0.1994 3.726 0.000194 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                       exp(coef) exp(-coef) lower .95 upper .95
datyear$ageCat(50,Inf]     2.102     0.4757     1.422     3.107
```

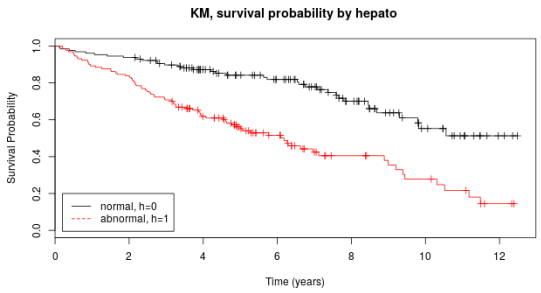

Figure 6: Survival probability depending of the variable hepato, $p = 8.39e - 08$

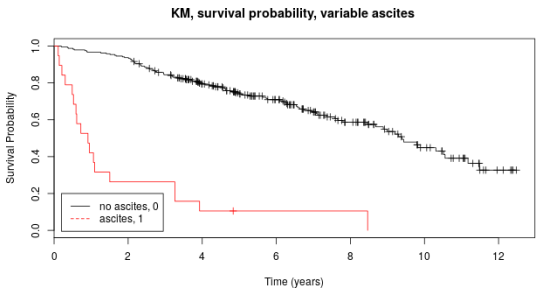

Figure 7: Survival probability depending of the variabele ascites, $p = 0$

At any given time, the risk hazard of death is 2 times higher for the > 50 years group than the < 50 years group. We can also performed a cox hazard regression with the variable age as below:

```
fit.cphage <- coxph(formula = Surv(datyear$time,datyear$status)
    ~ datyear$age, data = datyear)
summary(fit.cphage)
#              exp(coef) exp(-coef) lower .95 upper .95
# datyear$age     1.042     0.9595     1.023     1.062
```

The risk hazard increase by a factor of 1.042 / Year (or 1.5 /decade) for the patients. In other words, for any patient the risk of death increases by 4 % for each additional year, or equivalently, the hazard of death increases by 50 % per decade.

5.4 Comparaison by hepato

Presence or not of an enlarged liver is marked with the hepato variable (hepato = 0 means no enlarged liver and hepato = 1 means enlarged liver).

```
km.hepato <-survfit(Surv(datyear$time, datyear$status) ~
                    datyear$hepato, data = datyear)
km.hepato
# datyear$hepato=0 128     36     NA    9.79      NA
# datyear$hepato=1 130     75   6.18    4.77    8.99
```

The median survival time with and without enlarged liver are respectively 4.8 and 9.8 years. Using a log rank test, the p value is $p = 8.39e - 08$, so the null hypothesis is rejected, and the presence or not of an enlarged liver is a strong parameter on the overall survival. Using a cox proportional hazard model (the survival curves from fig 6 are almost parallel), the risk hazard for the patient with enlarged liver is 2.8 times higher than the patient with no enlarged liver.

```
n= 258, number of events= 111
                  coef exp(coef) se(coef)     z Pr(>|z|)
datyear$hepato 1.0504    2.8589   0.2046 5.133 2.85e-07 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
datyear$hepato     2.859     0.3498     1.914      4.27
```

5.5 Comparaison by ascites

Ascites is an abnormal build up of fluid in the abdomen and is often associated with liver cirrhosis.

```
km.ascites <- survfit(Surv(datyear$time,datyear$status) ~
    datyear$ascites, data=datyear)
km.ascites
#                      n events median 0.95LCL 0.95UCL
# datyear$ascites=0 239     93  9.392   8.685   11.18
# datyear$ascites=1  19     18  0.915   0.592    3.26
```

The median survival time for the patients with and without the presence of ascites are respectively 0.9 (CI 0.6 - 3.2) and 9.4 (CI 8.7 - 11.2) years respectively. This parameter is a very strong indicator for the prediction of the survival probability of the patient. A log rank test give a p value of 0. The figure 7 shows that for patient with presence of ascites, the survival probability are quite lower than for patient without. Here the curves are not parallel, the ratio hazard ratio between the 2 groups cannot be assumed to be constant.

6 Summary

For the subset of the pbc dataset (from the R survival library) obtained by removing the missing value and ignoring the transplant patient, the effect of the D penicillamine drug is not statistically significant (p value = 0.56). The mean overall survival time for these patient is 9 years (CI 7.6 - 10.3), ie 50 % of the patient died within 9 years. Men have a risk hazard 67 % higher than women and the risk hazard of death for the > 50 years group is 2 times higher than the younger group (< 50 years) for this dataset. The risk hazard of death for any patient increases by 4 % per additional year, (or increase by 50 % per decade). Enlarge liver (hepato = 1) is a significative variable for the survival probability, the presence of enlarge liver gives a risk hazard 2.7 times higher than normal liver. (hepato = 0). Ascites, which is an abnormal build up fluid in the abdomen is a very significative variable for the survival of the patient. (survival median 9.4 when ascites = 0, and 0.9 when ascites = 1).

7 Conclusion

For this subset dataset of the primary biliary cirrhosis pbc dataset, the effect of the drug D-penicillamine is not statistically significant (p = 0.56). Some visible variables (or symptomatic variables) are very strong indicators for the survival probability of the patients. The other variables (concentration of albumine, phosphotase, bili, ect) not studied here are most likely correlated to the symptomatic status of the patient. Hence for further study and to built a predictive model of the patient survival probability, using either symptomatic variables or the variables with the concentrations of molecules might be sufficient.

Appendix A R script

```
#  Projet:  R script, survival analysis, Dataset Pbc from
#  survival library. B.Lepers 15/04/2019
#  study of the visible variables (gender, age, hepato, ascites) on
#  the survival probability for
#  the pancreatic liver data set.

setwd('/home/ben/Documents/DSTI/Survival_analysis_R/project')
# import libraries
library(survival)
library(asaur)
library(tidyverse)

# dataset from the Mayo clinic trial in primary biliary cirrhosis of the liver
# trial conducted between 1974 and 1984.
# we have removed all rows containning any missing values
# the variables are:
# age: in years
# albumin:  serum albumin (g/dl)
# alk.phos:  alkaline phosphotase (U/liter)
# ascites:  presence of ascites, abnormal buildup of fluid in
# the abdomen (usually due to cirrhosis)
# ast:  aspartate aminotransferase, once called SGOT (U/ml)
# bili:  serum bilirunbin (mg/dl)  (usually, with abscence of
# liver disease, bili is high level)
# chol:  serum cholesterol (mg/dl)
# copper:  urine copper (ug/day)
# edema:  0 no edema, 0.5 untreated or successfully treated
# 1 edema despite diuretic therapy
# hepato:  presence of hepatomegaly (enlarged liver)
# id:  case number
# platelet:  platelet count
# protime:  standardised blood clotting time
# sex:  m/f
# spiders:  blood vessel malformations in the skin
# stage:  histologic stage of disease (needs biopsy)
# status:  status at endpoint, 0/1/2 for censored, transplant, dead
# time:  number of days between registration and the earlier of death,
# transplantion, or study analysis in July, 1986
# trt:  1/2/NA for D-penicillmain, placebo, not randomised
# trig:  triglycerides (mg/dl)
# redefine the status variable.  0 for censored and 1 for death.

# the status variable has 3 states 0/1/2 for censored, transplant and dead
# we removed all the rows with the status = 1, transplant  to get a binary
# value for the status variable
# we want to perform an survival analysis so we study 0 censored or 1 death.

dat <-pbc
a <- subset(dat, (dat$status == 0)|(dat$status == 2))
df <- mutate(a, status = as.integer(a$status == 2), id = as.character(id))
str(df)
#library(tidyr)



#dat<-df %>% drop_na()
dat <- df[complete.cases(df), ]
# make a data set with no missing values
# dim(dat)
# 258 rows, 20 columns
#
table(dat$status)
hist(dat$age)
table(dat$sex)
# m   f
# 31 227
# 0 censored, 1 death
#  0   1
# 147 111
summary(dat)
#
# 1.********* Use kaplan Meier survival estimate to take into
# account the status variable (censoring)
datyear <- mutate(dat, time = time/365)
png("R_fig_KM.png", width=640, height=360, res=72)
kmsurvival <- survfit(Surv(datyear$time,datyear$status) ~ 1, data = datyear)
plot(kmsurvival, mark.time = TRUE, xlab="Time (years)",
     ylab="Survival Probability", main =  'Kaplan Meier Estimator')
dev.off()
kmsurvival
# n  events  median 0.95LCL 0.95UCL
# 258.00  111.00    8.99    7.66   10.3

# 2.**** Drug or placebo variable trt: 1: D penicillmain,
# 2: placebo, NA: not randomised
km.trt <- survfit(Surv(datyear$time,datyear$status) ~ datyear$trt, data=datyear)
km.trt
```

```
# n events median 0.95LCL 0.95UCL
# datyear$trt=1 127     57   8.45    6.54      NA
# datyear$trt=2 131     54   9.30    8.47      NA
# logrank test
png("R_fig_KM_drug.png", width=640, height=360, res=72)
plot(km.trt, xlab="Time (years)", ylab="Survival Probability",
     mark.time = TRUE, col = c(1,2),main =  'Kaplan Meier Estimator')
legend(0.2, 0.2, c("drug D", "placebo"), lty=c(1:2),col = c('black', 'red'))
dev.off()
survdiff(Surv(datyear$time,datyear$status) ~ datyear$trt)

# N Observed Expected (O-E)^2/E (O-E)^2/V
# datyear$trt=1 127       57       54     0.171     0.333
# datyear$trt=2 131       54       57     0.161     0.333
# Chisq= 0.3  on 1 degrees of freedom, p= 0.564
# do not reject H0, the effect of D penicillmain is not significant on this dataset
plot(km.trt, xlab="Time (years)", ylab="Survival Probability", mark.time = TRUE, col = c(1,2))
legend(0.2, 0.2, c("1","2"), lty=c(1:2),col = c('black', 'red'))
# not significant

# 3.***** by gender
km.sex <- survfit(Surv(datyear$time,datyear$status) ~ datyear$sex, data = datyear)
km.sex
survdiff(Surv(datyear$time,datyear$status) ~ datyear$sex)
png("R_fig_KM_sex.png", width=640, height=360, res=72)
plot(km.sex, xlab="Time (years)", ylab="Survival Probability", mark.time = TRUE, col = c(1,2),
     main = 'Kaplan meier by gender')
legend(0.2, 0.2, c("male", "female"), lty=c(1:2),col = c('black', 'red'))
dev.off()
# p = 0.0033 < 0.05, so gender is significant. The 2 curves are
# almost parallel, so we can use a cox proportional hazard model
fit.cphsex <- coxph(formula = Surv(datyear$time,datyear$status)
                    ~ datyear$sex, data = datyear)
summary(fit.cphsex)
# exp(coef) exp(-coef) lower .95 upper .95
# datyear$sexf    0.5985      1.671    0.3711    0.9653

# 4.******* by age category, Kaplan Meier
# effect of age below and higher than 50
# *************** study by age group
# add a categorical variable age group column to the dataset
# first lets split the data in a young group < 50 and old group > 50
datyear$ageCat <- cut(datyear$age, breaks = c(26, 50, Inf))
table(datyear$ageCat)

km.age <- survfit(Surv(datyear$time,datyear$status) ~ datyear$ageCat, data=datyear)
km.age
# n events median 0.95LCL 0.95UCL
# datyear$ageCat=(26,50]  124     39  11.18    9.20      NA
# datyear$ageCat=(50,Inf] 134     72   6.96    5.63     9.3
# age seems significant variable
# survival probability between 2 groups of different age
survdiff(Surv(datyear$time,datyear$status) ~ datyear$ageCat)
# N Observed Expected (O-E)^2/E (O-E)^2/V
# datyear$ageCat=(26,50]  124       39       59      6.77      14.5
# datyear$ageCat=(50,Inf] 134       72       52      7.68      14.5

# Chisq= 14.5  on 1 degrees of freedom, p= 0.000138
# pvalue < 0.05, reject H0, age is very significant. The younger have higher chances
# of survival.
png("R_fig_KM_age.png", width=640, height=360, res=72)
plot(km.age, xlab="Time (years)", ylab="Survival Probability", mark.time = TRUE, col = c(1,2),
     main = 'Survival probability by age group')
legend(0.2, 0.2, c("26-50", "50+"), lty=c(1:2),col = c('black', 'red'))
dev.off()
# the 2 curves are almost parallel, we can use the coxregression model
fit.cphage <- coxph(formula = Surv(datyear$time,datyear$status) ~ datyear$age, data = datyear)
summary(fit.cphage)

# n= 258, number of events= 111
# coef exp(coef) se(coef)     z Pr(>|z|)
# datyear$age 0.041349  1.042216 0.009673 4.275 1.91e-05 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# exp(coef) exp(-coef) lower .95 upper .95
# datyear$age     1.042     0.9595     1.023     1.062

# Concordance= 0.631  (se = 0.03 )
# Rsquare= 0.068   (max possible= 0.985 )
# Likelihood ratio test= 18.23  on 1 df,   p=1.962e-05
# Wald test            = 18.27  on 1 df,   p=1.913e-05
# Score (logrank) test = 18.41  on 1 df,   p=1.777e-05
# So the risk hazar function increase by a factor of 1.042/ Year (or 1.5 /decade)
# for the older group compared to the young group.
# the pvalue shows also that the difference are significative.
# the hazard of death of the older group increases by 4% for each additional year.
# or equivalently, the hazard of death increases by 50 % per decade.
fit.cpgroupage <- coxph(formula = Surv(datyear$time,datyear$status) ~ datyear$ageCat, data = datyear)
summary(fit.cpgroupage)
```

```
# coef exp(coef) se(coef)     z Pr(>|z|)
# datyear$ageCat(50,Inf] 0.7430    2.1023   0.1994 3.726 0.000194 ***
#  ---
#  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
# exp(coef) exp(-coef) lower .95 upper .95
#datyear$ageCat(50,Inf]     2.102     0.4757     1.422     3.107

# 5.************edema
km.edema <- survfit(Surv(datyear$time,datyear$status) ~ datyear$edema, data=datyear)
km.edema
plot(km.edema, xlab="Time (years)", ylab="Survival Probability", mark.time = TRUE, col = c(1,2,3))
legend(0.2, 0.2, c("0","0.5","1"), lty=c(1:3),col = c('black', 'red', 'green'))
survdiff(Surv(datyear$time,datyear$status) ~ datyear$edema, data = datyear)
fit.cxedema <- coxph(Surv(datyear$time,datyear$status) ~ datyear$edema, data=datyear)
summary(fit.cxedema)
# hepato: presence of hepatomegaly
km.hepato <-survfit(Surv(datyear$time, datyear$status) ~
                    datyear$hepato, data = datyear)
km.hepato
png("R_fig_KM_hepato.png", width=640, height=360, res=72)
plot(km.hepato, xlab="Time (years)", ylab="Survival Probability", mark.time = TRUE, col = c(1,2),
     main = "KM, survival probability by hepato ")
legend(0.2, 0.2, c("normal, h=0","abnormal, h=1"), lty=c(1:2),col = c('black', 'red'))
dev.off()
# datyear$hepato=0 128     36     NA    9.79      NA
# datyear$hepato=1 130     75   6.18    4.77    8.99
# survival probability between 2 groups of different age
survdiff(Surv(datyear$time,datyear$status) ~ datyear$hepato)
#survdiff(formula = Surv(datyear$time, datyear$status) ~ datyear$hepato)

#N Observed Expected (O-E)^2/E (O-E)^2/V
#datyear$hepato=0 128       36     63.7      12.0      28.7
#datyear$hepato=1 130       75     47.3      16.2      28.7

#Chisq= 28.7  on 1 degrees of freedom, p= 8.39e-08
# reject H0, the difference is significative

# the curve are almost paralell, we assume that the ration between the hazard function
# is constant. We can use a cox ph model
fit.hepato <- coxph(formula = Surv(datyear$time,datyear$status) ~ datyear$hepato, data = datyear)
summary(fit.hepato)
# coxph(formula = Surv(datyear$time, datyear$status) ~ datyear$hepato,
# data = datyear)

# n= 258, number of events= 111

# coef exp(coef) se(coef)     z Pr(>|z|)
# datyear$hepato 1.0504    2.8589   0.2046 5.133 2.85e-07 ***
#  ---
#  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# exp(coef) exp(-coef) lower .95 upper .95
# datyear$hepato     2.859     0.3498     1.914      4.27

# Concordance= 0.633  (se = 0.026 )
# Rsquare= 0.105   (max possible= 0.985 )
# Likelihood ratio test= 28.53  on 1 df,   p=9.217e-08
# Wald test            = 26.35  on 1 df,   p=2.853e-07
# Score (logrank) test = 28.71  on 1 df,   p=8.391e-08
# The difference is quite significative.
# with hepato, the risk hazard is 2.8 times higher than without.

# 6. ******************** effect of ascites
km.ascites <- survfit(Surv(datyear$time,datyear$status) ~ datyear$ascites, data=datyear)
km.ascites
# n events median 0.95LCL 0.95UCL
# datyear$ascites=0 239     93  9.392   8.685   11.18
# datyear$ascites=1  19     18  0.915   0.592    3.26
# without ascites the mean survival time is 9.3 years (8.7 - 11.2)
# with ascites the mean survival time is 0.9 years ( 0.5 - 3.2)
#log rank test
survdiff(Surv(datyear$time,datyear$status) ~ datyear$ascites)
# survdiff(formula = Surv(datyear$time, datyear$status) ~ datyear$ascites)

# N Observed Expected (O-E)^2/E (O-E)^2/V
# datyear$ascites=0 239       93   108.55      2.23       102
# datyear$ascites=1  19       18     2.45     98.61       102

# Chisq= 102  on 1 degrees of freedom, p= 0
# here we reject H0
png("R_fig_KM_ascites.png", width=640, height=360, res=72)
plot(km.ascites, xlab="Time (years)", ylab="Survival Probability", mark.time = TRUE, col = c(1,2),
     main = 'KM, survival probability, variable ascites ')
legend(0.2, 0.2, c("no ascites, 0", "ascites, 1"), lty=c(1:2),col = c('black', 'red'))
dev.off()
fit.ascites <- coxph(formula = Surv(datyear$time,datyear$status) ~ datyear$ascites, data = datyear)
summary(fit.ascites)
# this is a very strong variable in therm of survival.

# this is not include in the report
```

```
# 7. **** variables selections
# can perform an AIC test of the full model to determine which variables are the most significant
Mfull <- coxph(Surv(time, status) ~   trt + age + sex + ascites + hepato +
                 spiders + edema + bili + chol + albumin + copper + alk.phos +
                 ast + trig + platelet + protime + stage,
               data = dat)

MAIC <- step(Mfull)

# we get the final model
# Step:  AIC=946.84
# Surv(time, status) ~ id + age + edema + bili + albumin + copper +
#   ast + protime + stage
summary(MAIC)
```