4/20/2020

INTRODUCTION TO SURVIVAL ANALYSIS

Study of Primary Biliary Cirrhosis

Fawaz Qutami
DATA SCIENCETECH INSTITUTE

Table of Contents

Introduction

The goal of the analysis is to use some methods namely: Kaplan–Meier estimator, Log-Rank Test, Cox Proportional Hazard, Machine Learning, and Cross-Validation, to assess the benefit of some selected covariates on survival.

The basic analysis, that I will do, can be done using PBC dataset and the functions in R package "survival", alongside some other packages.

I will incorporate my conclusions - in italic blue, in the document, wherever is that possible and as I go through the analysis.

Primary Biliary Cirrhosis (PBC) Description

"The data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval met eligibility criteria for the randomized placebo-controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants.

Format:

age:	in years
albumin	serum albumin (g/dl)
alk.phos	alkaline phosphotase (U/liter)
ascites	presence of ascites
ast	aspartate aminotransferase, once called SGOT (U/ml)
bili	serum bilirunbin (mg/dl)

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 or 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)

" *Source:* *https://stat.ethz.ch/R-manual/R-patched/library/survival/html/pbc.html*

Dataset Processing and Basic Statistics

Let us first load the needed libraries and read in our pbc dataset and make the necessary modifications to it.

```
library(survival)
library(ranger)
library(dplyr)
library(survminer)
library(psych)
library(gmodels)
library(survivalROC)
library(tidyverse)
library(glmnet)
data(pbc)
```

We are particularly interested in 'time' and 'status' features in the dataset. "time" represents the number of days after registration, which is the observed, potentially censored survival time, and "status" which indicates if the observation has been censored. In the complete dataset there are 418 patients, and the status shows that 161 have died, 232 have been censored and then 25 have got a transplant. We exclude those 25 observations (transplant) from the further analysis. Furthermore, I will change the status to a logical and change some categorial covariates to factors, as well as change the time to reflect years instead of days:

```
PbcData <- subset(pbc, status != 1)
PbcData <- transform(PbcData, status = as.logical(status))
PbcData$sex <- factor(PbcData$sex, levels = c("m","f"), labels = c("Male",
"Female"))
PbcData$trt <- factor(PbcData$trt, levels = c(1, 2), labels = c("D-
penicillmain", "placebo"))
PbcData$edema <- factor(PbcData$edema, levels = c(0,0.5, 1), labels = c("No
Edema", "Treated Successfully/Untreated", "Edema"))
PbcData <- mutate(PbcData, time = (PbcData$time / 365.25))

describe(pbc_data$years)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
X1	1	393	5.32	3.07	4.84	5.16	3.13	0.11	13.13	13.02	0.44	-0.56	0.15

Result of describe() shows different basic statistics such as: sample size- n, median, mean, min, max, etc.

What is the death rate in males and females?

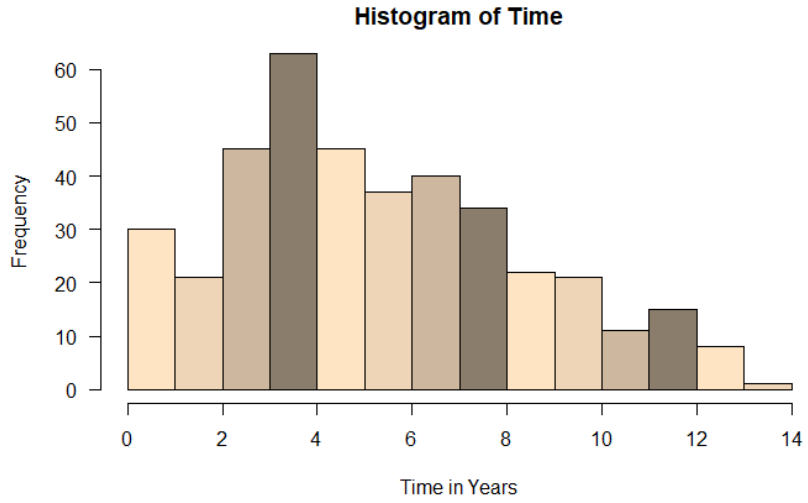
```
relative_frequency = round(100 * prop.table(table(PbcData$status,
PbcData$sex), 2), 1)
relative_frequency
```

```
        Male Female
  FALSE 41.5   61.1
  TRUE  58.5   38.9
```

Death rates in Males is 58.5% which is higher than in Females (38.9%).

What is the shape of time distribution?

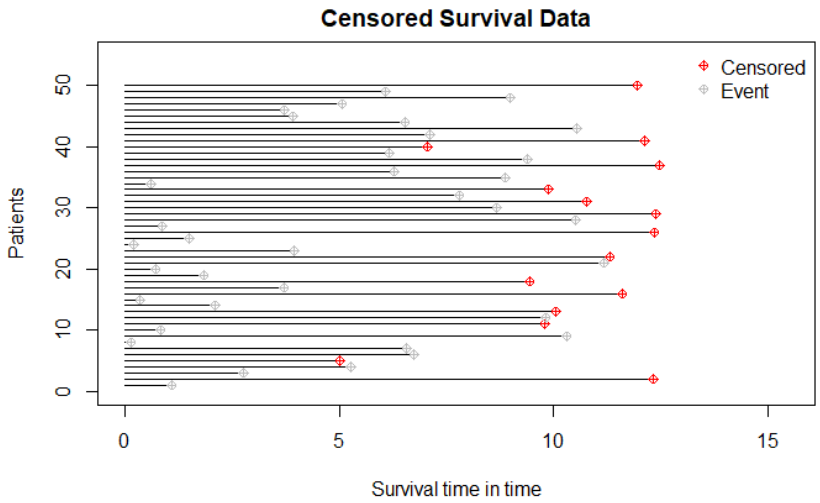
```
hist(PbcData$time ,las = 1, col = c("bisque1", " bisque2", "bisque3",
"bisque4"), main = "Histogram of Time", xlab = "Time in Years")
```



Distribution of follow-up times is skewed and may differ between censored patients and those with events.

Kaplan–Meier Estimator (KM)

Kaplan–Meier curve is important because the method can take into account some types of censored data, particularly right-censoring, which occurs if a patient withdraws from a study, is lost to follow-up, or is alive without event occurrence at last follow-up. In our case, the dataset is right censored.

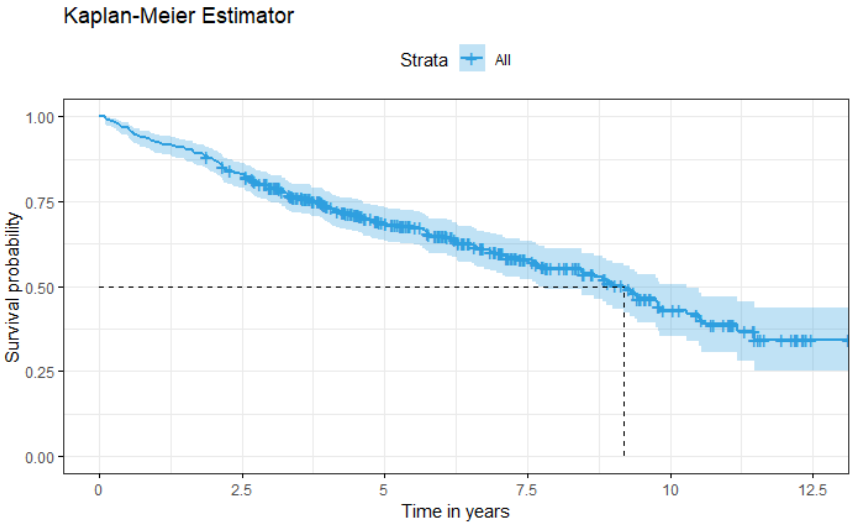


What is Kaplan-Meier median survival time overall the data?

```
Surv_OBJ <- Surv(PbcData$time, PbcData$status) # by default right censored
KM_Model <- survfit(Surv_OBJ ~ 1, data = PbcData, conf.type="log-log", type =
"kaplan-meier")
KM_Model
ggsurvplot(KM_Model,data = PbcData,xlab = "Time in years"
           ,surv.median.line = "hv",title="Kaplan-Meier Estimator"
           ,palette = "#2E9FDF",ggtheme = theme_bw())
```

```
Call: survfit(formula = Surv_OBJ ~ 1, data = PbcData, conf.type = "log-log",
    type = "kaplan-meier")

      n  events  median 0.95LCL 0.95UCL
 393.00  161.00    9.19    7.70   10.30
```



The result shows the number of patients – sample size (393), the number of events (161), the median survival time(9.19 year) and its 95% confidence interval (7.70 - 10.3 year).

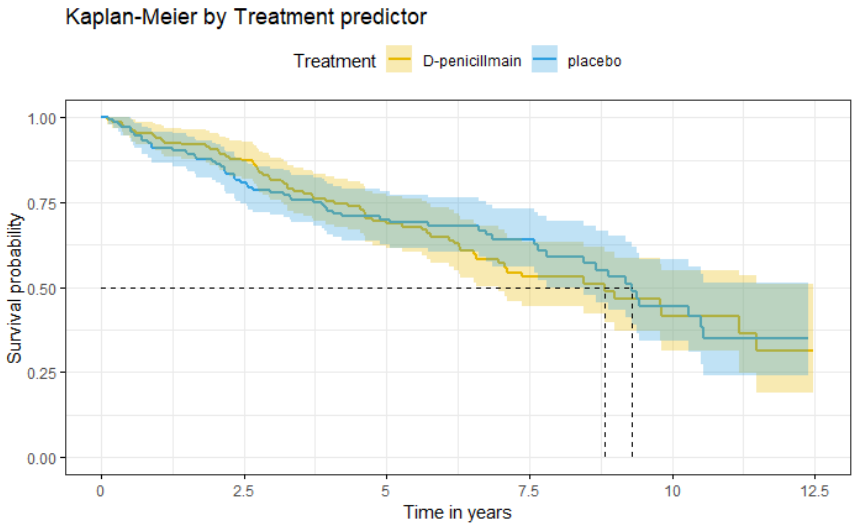
What is KM median survival time by treatment predictor?

```
KM_trt <- survfit(Surv_OBJ ~ trt, data = PbcData)
```

```
KM_trt
ggsurvplot(KM_trt,data = PbcData,conf.int = TRUE ,xlab = "Time in years",
surv.median.line = "hv",title="Kaplan-Meier by Treatment predictor"
           ,legend.title="Treatment",palette = c("#E7B800", "#2E9FDF")
           ,ggtheme = theme_bw(),censor = FALSE
           ,legend.labs = c("D-penicillmain", "placebo"))
```

```
Call: survfit(formula = Surv_OBJ ~ trt, data = PbcData)

   100 observations deleted due to missingness
                      n events median 0.95LCL 0.95UCL
trt=D-penicillmain 148     65   8.82    6.95      NA
trt=placebo        145     60   9.30    8.46      NA
```



The result shows that median survival time for "D-penicillmain" is 8.82 years for a sample size n =148, and in "placebo" is 9.30 years for a sample size n = 145.

Log-Rank Test

We can use the log-rank test to compare survival curves of two groups. Chi-squared distribution can be used to derive a p-value. P-values are used in statistical hypothesis testing to quantify statistical significance (**A result with p < 0.05 is usually considered significant)**.

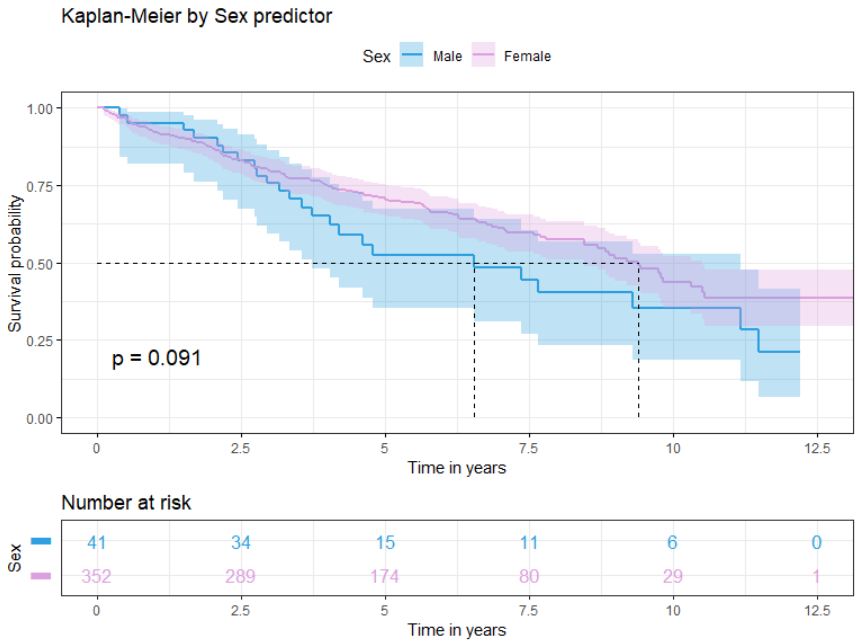
What is the probability of surviving over sex predictor?

```
pbc_sex = PbcData[complete.cases(PbcData$sex), ]
Surv_sex <- Surv(pbc_sex$time, pbc_sex$status)
logrank_Model_sex <- survdiff(Surv_OBJ ~ sex, data = pbc_sex)
logrank_Model_sex
KMLR_sex <- survfit(Surv_OBJ ~ sex, data = pbc_sex, conf.type="log-log")
ggsurvplot(KMLR_sex,data = pbc_sex,pval = TRUE,conf.int = TRUE
           ,xlab = "Time in years",surv.median.line = "hv"
           ,risk.table = TRUE,title="Kaplan-Meier by Sex predictor"
           ,risk.table.height=.25,tables.col = "strata"
           ,tables.y.text = FALSE,legend.title="Sex"
           ,palette = c("#2E9FDF", "#DE9FDF")
           ,ggtheme = theme_bw(),censor = FALSE
           ,legend.labs = c("Male", "Female"))
```

```
Call:
survdiff(formula = Surv_OBJ ~ sex, data = pbc_sex)

             N Observed Expected (O-E)^2/E (O-E)^2/V
sex=Male    41       24     17.4     2.529      2.85
sex=Female 352      137    143.6     0.306      2.85

 Chisq= 2.9  on 1 degrees of freedom, p= 0.09
```



Assumption of the null hypothesis indicates that Males and Females are non-significantly different in terms of survival . We have no statistically significant evidence that the survival distributions are not the same (p-value = 0.09 > 0.05).

What is the probability of surviving over Treatment predictor?

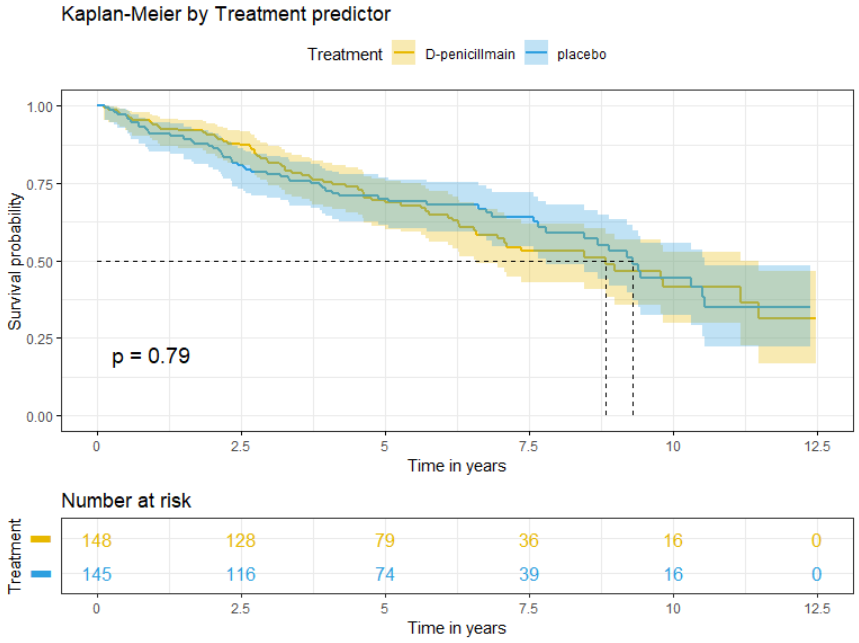
```
pbc_trt = PbcData[complete.cases(PbcData$trt), ]
Surv_trt <- Surv(pbc_trt$time, pbc_trt$status)
logrank_Model_trt <- survdiff(Surv_trt ~ trt, data = pbc_trt)
logrank_Model_trt
# Plot using KM and ggsurvplot
KMLR_trt <- survfit(Surv_trt ~ trt, data = pbc_trt, conf.type="log-log")
ggsurvplot(KMLR_sex,data = pbc_sex,pval = TRUE,conf.int = TRUE
           ,xlab = "Time in years",surv.median.line = "hv"
           ,risk.table = TRUE,title="Kaplan-Meier by Treatment predictor"
           ,risk.table.height=.25,tables.col = "strata"
           ,tables.y.text = FALSE,legend.title="Treatment"
           ,palette = c("#E7B800", "#2E9FDF")
           ,ggtheme = theme_bw(),censor = FALSE
           ,legend.labs = c("D-penicillmain", "placebo"))
```

```
Call:
survdiff(formula = Surv_trt ~ trt, data = pbc_trt)

                N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
trt=D-penicillmain 148       65     63.5    0.0354    0.0722
trt=placebo        145       60     61.5    0.0366    0.0722

 Chisq= 0.1  on 1 degrees of freedom, p= 0.8
```



Assumption of the null hypothesis indicates that "D-penicillmain", "placebo", are non-significantly different in terms of survival; We have no statistically significant evidence that the survival distributions are not the same (p-value = 0.8 > 0.05).

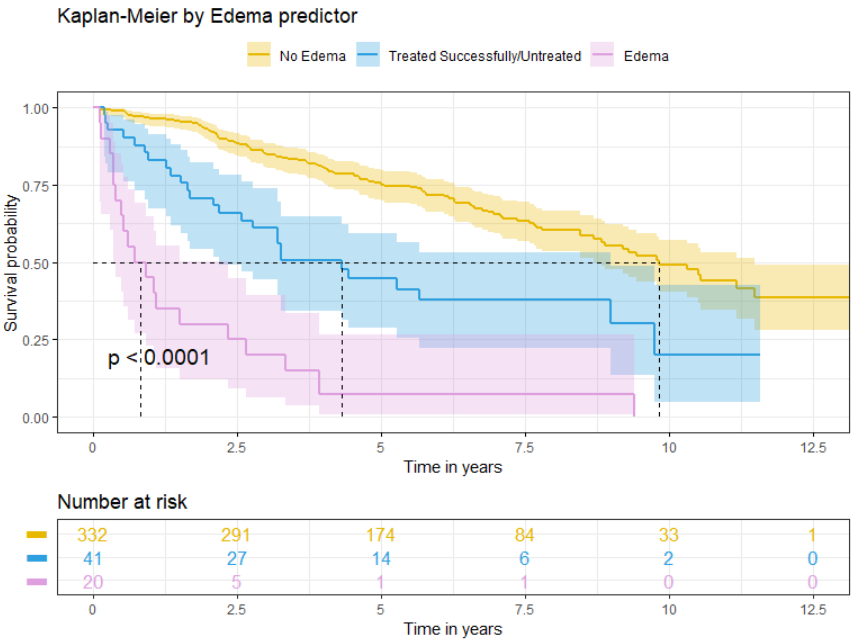
What is the probability of surviving over edema predictor?

```
pbc_edema = PbcData[complete.cases(PbcData$edema), ]
Surv_edema <- Surv(pbc_edema$time, pbc_edema$status)
logrank_Model_edema <- survdiff(Surv_edema ~ edema, data = pbc_edema)
logrank_Model_edema
KMLR_edema <- survfit(Surv_edema ~ edema, data = pbc_edema, conf.type="log-
log")
ggsurvplot(KMLR_edema,data = pbc_edema,pval = TRUE,conf.int = TRUE
           ,xlab = "Time in years" ,surv.median.line = "hv"
           ,risk.table = TRUE,title="Kaplan-Meier by Edema predictor"
           ,risk.table.height=.25,tables.col = "strata",tables.y.text = FALSE
           ,legend.title="",palette = c("#E7B800", "#2E9FDF", "#DE9FDF")
           ,ggtheme = theme_bw(),censor = FALSE
           ,legend.labs = c("No Edema", "Treated Successfully/Untreated",
"Edema"))
```

```
Call:
survdiff(formula = Surv_edema ~ edema, data = pbc_edema)

                                 N Observed Expected (O-E)^2/E (O-E)^2/V
edema=No Edema                 332      116   145.38      5.94      61.6
edema=Treated Succ./Untreated   41       26    13.00     12.99      14.2
edema=Edema                     20       19     2.61    102.79     105.4

 Chisq= 123  on 2 degrees of freedom, p= <2e-16
```



Assumption of the null hypothesis indicates that "No Edema", "Treated Successfully/Untreated", and "Edema" are significantly different in terms of survival; We have high statistically significant evidence that the survival distributions are not the same (p-value << 0.0001). From the plot it seems clear that individuals without edema have a larger survival time than those with edema, and those treated successfully or left untreated have a larger survival time than those treated unsuccessfully.

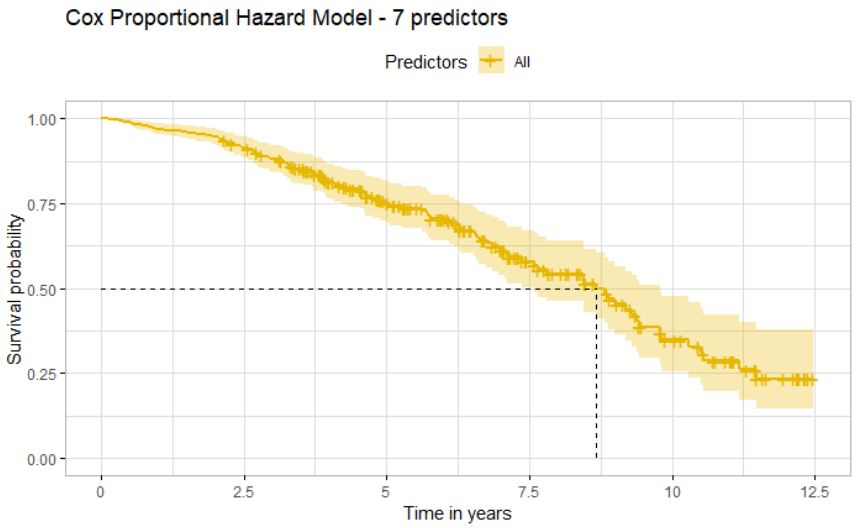
Cox Proportional Hazard Model Regression

We want to study the effect of a unit increase in a covariant with respect to the hazard rate. Let us create an initial cox model and show the result summary in a rearranged table:

```
Cox_Data = PbcData[complete.cases(PbcData), ]
Surv_Cox <- Surv(Cox_Data$time, Cox_Data$status)
COX_Model <- coxph(Surv_Cox ~ age + sex + edema + alk.phos
                   + albumin + bili + trt + protime + stage
                   , data = Cox_Data)
ggsurvplot(survfit(COX_Model),data=Cox_Data,xlab = "Time in time"
           ,title="Cox Proportional Hazard Model - 7 predictors"
           ,ggtheme = theme_light(),legend.title="Predictors"
           ,surv.median.line = "hv",palette = "#E7B800")
```

#	term	estimate	HR	std.error	p.value	conf.low	conf.high
1	age	0.019	1.019	0.011	0.0835	0.00	0.04
2	sex: Female vs Male	-0.660	0.517	0.270	0.0145	-1.19	-0.13
3	edema:Treated /Untreated vs No Edema	0.039	1.039	0.313	0.9021	-0.58	0.65
4	edema: Edema vs No Edema	1.100	3.004	0.352	0.0018	0.41	1.79
5	alk.phos	0.000	1.000	0.000	0.3781	0.00	0.00
6	albumin	-0.889	0.411	0.278	0.0014	-1.43	-0.35
7	bili	0.116	1.123	0.018	0.0000	0.08	0.15
8	trt: placebo vs D-penicillmain	-0.166	0.847	0.208	0.4248	-0.57	0.24

9	protime	0.199	1.220	0.102	0.0506	0.00	0.40
10	stage	0.462	1.588	0.142	0.0011	0.18	0.74



What is the interpretation of the above table?

The above table represents the information extracted from the model summary; however, we can conclude the following:

- *We can tell from the summary that the betas corresponding to the covariates are significantly different from 0 (Estimates).*
- *Associated with each 1-year increase in "age" is an increased risk of death (HR = 1.02). This is not statistically significant (p-value = 0.08 > 0 .05).*
- *"Female" vs "Males" shows that the risk in "Females" is less than in males (HR = 0.51). This is statistically significant (p-value = 0.0145 < 0.05).*
- *Associated with each 1 unit (1 g/dl) decrease in "albumin" is an increased risk of death (HR = 0.41). This is highly statistically significant (p-value = 0.0014 < 0.01).*
- *Associated with each 1 unit (1 mg/dl) increase in "bilirubin" is an increased risk of death (HR = 1.123). This is high statistically significant (p-value << 0.0001).*
- *Associated with each 1 unit (U/liter) increase in "alk.phos" is not an increased risk of death(HR = 1). This is not statistically significant (p-value = .38 > 0.05).*
- *Associated with each 1 unit increase in "protime" is an increased risk of death (HR = 1.22). This is not statistically significant (p-value = 0.00506 > 0.05) but may considered as so close.*
- *"Edema" vs "No Edema" there is an increased risk of death (HR = 3). This is statistically significant (p-value = 0.0018 < 0.01).*
- *Between the stages there is an increased risk of death (HR = 1.59). This is statistically significant (p-value = 0.0011 < 0.01).*
- *"placebo" vs "D-penicillmain", p value > 0.05 indicates non-significantly different in terms of survival.*
- *"Untreated" or "Treated successfully" vs "No Edema", p values > 0.05 indicates non-significantly different in terms of survival.*

What are the top 5 risky cases in this model?

```
new_COX_data <- PbcData[c("id", "age", "sex", "alk.phos", "edema", "albumin",
"bili", "trt", "stage", "protime")]
segmented <- new_COX_data %>% mutate(risk_score = predict(COX_Model, newdata
= new_COX_data, type = "lp"))
segmented %>% arrange(desc(risk_score)) %>% head(5)
```

	id <int>	age <dbl>	sex <fctr>	alk.phos <dbl>	edema <fctr>	albumin <dbl>	bili <dbl>	trt <fctr>	stage <int>	protime <dbl>	risk_score <dbl>
1	281	65.88364	Female	705.0	Edema	2.10	17.9	D-penicillmain	4	12.9	5.095756
2	1	58.76523	Female	1718.0	Edema	2.60	14.5	D-penicillmain	4	12.2	4.014771
3	223	61.24298	Female	1833.0	Edema	2.43	14.1	D-penicillmain	4	11.0	3.931687
4	23	55.96715	Female	6064.8	Edema	2.94	17.4	placebo	4	11.7	3.864631
5	191	52.69268	Female	3740.0	No Edema	3.35	24.5	placebo	4	15.2	3.782303

Model Comparing

Models can be compared using methods such as Likelihood Ratio Test (LRT), The Akaike information criterion (AIC) and Stepwise model selection based on AIC.

In our case, we will implement "**Stepwise model selection based on AIC**" to extract the best model:

```
auto_AIC <- step(COX_Model)
```

```
      …
Step:  AIC=956.41
Surv_Cox ~ age + sex + edema + albumin + bili + protime + stage
           Df    AIC
<none>        956.41
- protime   1 957.70
- age       1 957.74
- sex       1 960.03
- edema     2 961.45
- stage     1 965.62
- albumin   1 965.91
- bili      1 989.34
```

*The final result ended up with (**Surv_Cox ~ age + sex + edema + albumin + bili + protime + stage**) which is the best model in our case.*

Model Diagnostics

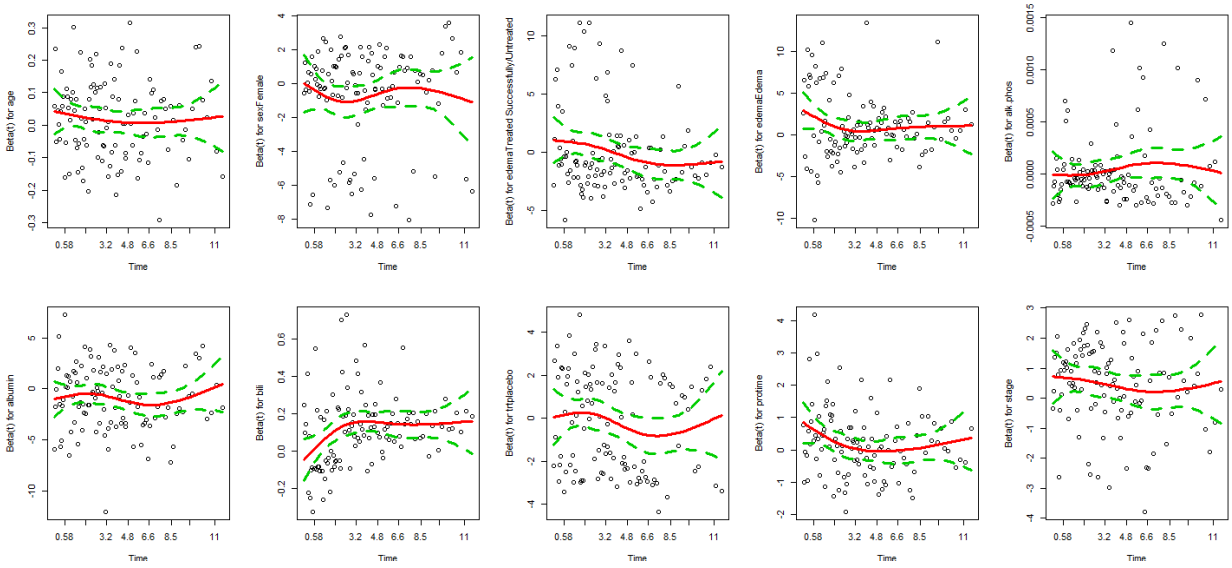
Diagnose a model includes different methods such as Martingale Residuals, Case Deletion Residuals, and Proportional Hazards Assumption(namely Complementary log-log and Schoenfeld Residuals). Schoenfeld residuals can be used for both continuous and categorical covariates, to test the proportional hazards assumption, I will implement Schoenfeld residuals to diagnose cox model:

```
SRPH  <- cox.zph(COX_Model, transform = "km")
SRPH
par(mfrow=c(2,5))
plot(SRPH, coef(COX_Model)[1], col = 2:3, lwd = 3)
```

```
                                      rho   chisq      p
age                               -0.0556  0.3658 0.5453
sexFemale                          0.0141  0.0226 0.8806
edemaTreated Successfully/Untreated -0.1991  5.3769 0.0204
```

```
edemaEdema          -0.0839  0.8316 0.3618
alk.phos             0.0985  0.9463 0.3307
albumin             -0.0296  0.1121 0.7378
bili                 0.2124  5.0585 0.0245
trtplacebo          -0.1226  1.8242 0.1768
protime             -0.1428  2.0434 0.1529
stage               -0.0977  0.9254 0.3361
GLOBAL                   NA 16.1751 0.0947
```

The function reports a test for non-proportionality (**the null hypothesis is that the corresponding variable satisfies PH**). It is more useful to inspect the PH assumption by plotting Schoenfeld residuals:

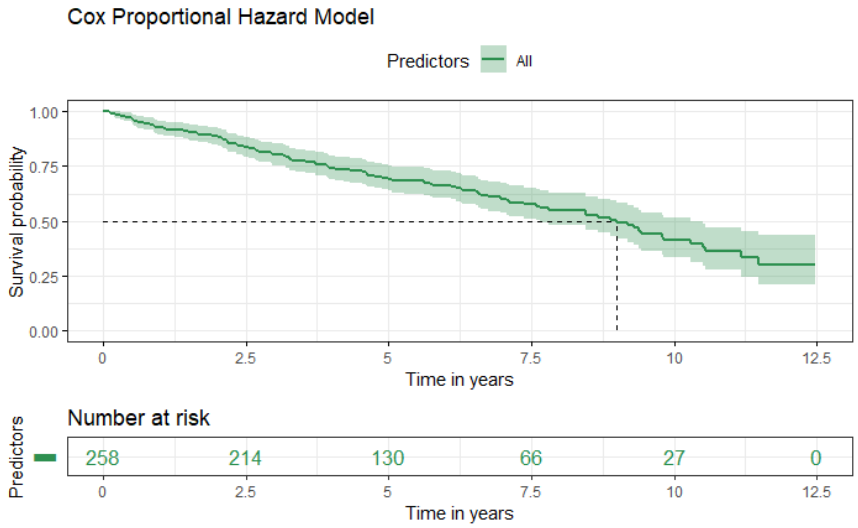


We observe that for all covariates the PH seems to be hold (except for "edema Treated Successfully/Untreated" and " bili").

Elastic Net Penalized Cox Proportional Hazards Regression

Let us first prepare the data for a machine learning task, by creating Y which is the survival time object, and X which is a numeric matrix of all other covariates:

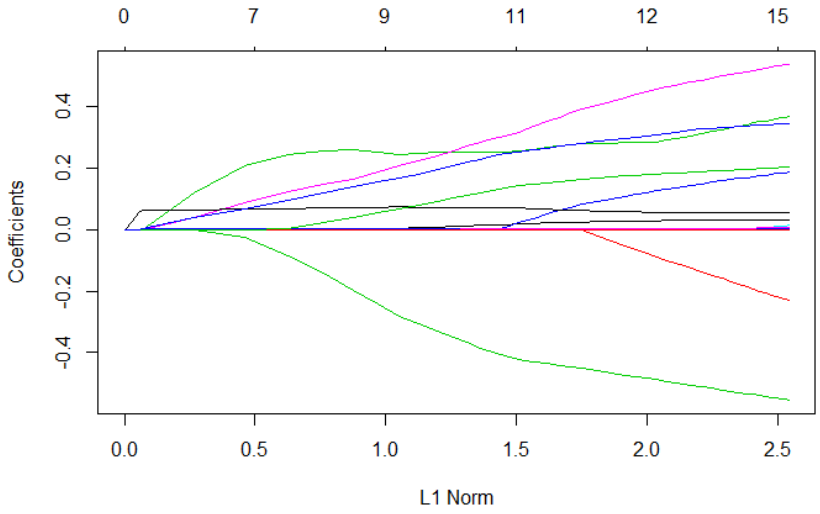
```
Y <- Surv(PbcData$time, PbcData$status)
X <- subset(PbcData, select = -c(time, status, id))
X <- as.matrix(sapply(X, as.numeric))
ggsurvplot(survfit(Y~1),data = PbcData,xlab = "Time in years"
,surv.median.line = "hv",risk.table = TRUE,title="Cox Proportional Hazard
Model" ,risk.table.height=.25,legend.title="Predictors" ,palette = "#2E9050"
,ggtheme = theme_bw(),censor = FALSE,tables.col = "strata",tables.y.text =
FALSE)
```



Next, we have to create train set and test set, then fit a generalized linear model (Cox regression model) via penalized maximum likelihood:

```
set.seed(1234)
train.idx <- sample(nrow(X), size = 200, replace = FALSE)
X.train <- X[train.idx,, drop = FALSE]
Y.train <- Y[train.idx,, drop = FALSE]
X.test <- X[-train.idx,, drop = FALSE]
Y.test <- Y[-train.idx,, drop = FALSE]
fit_sets <- glmnet(X.train, Y.train, family = "cox")
fit_sets
plot(fit_sets)
```

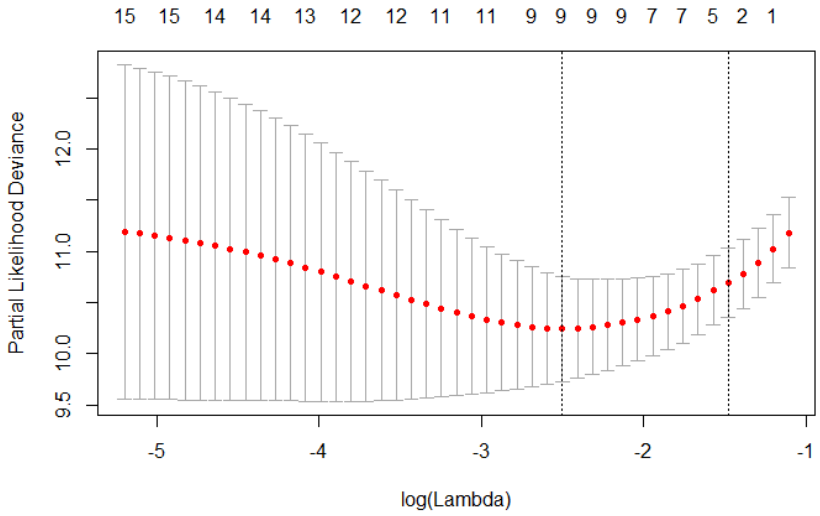
```
Call:  glmnet(x = X.train, y = Y.train, family = "cox")
      Df    %Dev   Lambda
 [1,]  0 0.00000 0.330700
 [2,]  1 0.02132 0.301300…
 …
 [48,] 16 0.15870 0.004172
 [49,] 16 0.15870 0.003802
```



Now, we need to compute the k-fold cross-validation for cox model; once the model is fit, we can view the optimal λ value and a cross validated error plot to help evaluate our model:

```
set.seed(1234)
cv.fit <- cv.glmnet(X.train, Y.train, family = "cox")
plot(cv.fit)
cv.fit$lambda.min
cv.fit$lambda.1se
```

```
[1] 0.0819061
[1] 0.2279086
```



The left vertical line in our plot shows us where the CV-error curve hits its minimum. The right vertical line shows us the most regularized model with CV-error within 1 standard deviation of the minimum. We can also extract such optimal λ's, and check the active covariates in our model and see their coefficients:

```
coef.min <- coef(cv.fit, s = cv.fit$lambda.min)
active.min <- which(coef.min != 0)
index.min <- coef.min[active.min]
index.min
coef.min
```

```
[1]  0.012642336  0.252354013  0.271287995  0.072844663 -0.368420972
0.002653440
[7]  0.001743331  0.112284291  0.218717515
17 x 1 sparse Matrix of class "dgCMatrix"
                      1
trt        .
age        0.012642336
sex        .
ascites    0.252354013
hepato     .
spiders    .
edema      0.271287995
bili       0.072844663
chol       .
albumin   -0.368420972
copper     0.002653440
alk.phos   .
```

```
ast       0.001743331
trig      .
platelet  .
protime   0.112284291
stage     0.218717515
```

How to predict and test the quality of the model?

Let us do a prediction and test the quality of this prediction via cox regression – here we compare a continuous predictor vs a right-censored time-to-failure outcome:

```
prediction.scores <- predict(cv.fit, newx = X.test, s = "lambda.min")
# Compute the interquartile range (IQR) for prediction.scores:
prediction.scores.scaled <- prediction.scores / IQR(prediction.scores)
# Test the prediction quality via Cox regression:
summary(coxph(Y.test ~ prediction.scores.scaled))
```

```
Call:
coxph(formula = Y.test ~ prediction.scores.scaled)
  n= 58, number of events= 24

                           coef exp(coef) se(coef)     z Pr(>|z|)
prediction.scores.scaled 2.0051    7.4269   0.3294 6.087 1.15e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                         exp(coef) exp(-coef) lower .95 upper .95
prediction.scores.scaled     7.427     0.1346     3.894      14.16

Concordance= 0.834  (se = 0.047 )
Rsquare= 0.474   (max possible= 0.944 )
Likelihood ratio test= 37.23  on 1 df,   p=1e-09
Wald test            = 37.06  on 1 df,   p=1e-09
Score (logrank) test = 51.91  on 1 df,   p=6e-13
```

The above result shows that our prediction is exceptionally good (HR = 7.4) with a positive coefficient (beta). This is high statistically significant (p-value << .0001)

How could one distinguish between low-risk and high-risk patients for this prediction?

Let us categorize the patients of our prediction – in terms of risk, to "Low Risk" and "High Risk":

```
Risk_Factor <-
  ifelse(prediction.scores.scaled <= median(prediction.scores.scaled)
         , "Low Risk", "High Risk")
table(Risk_Factor)
```

```
Risk_Factor
High Risk  Low Risk
       29        29
```

The result shows that "Low Risk" and "High Risk" have the same percentage (50%).

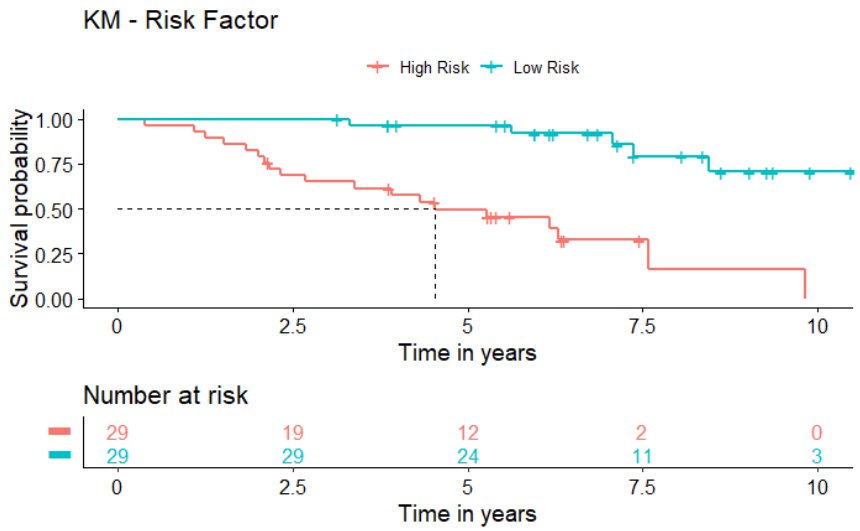
What is the median survival time by this Risk Factor?

```
KM_Risk <- survfit(Y.test ~ Risk_Factor, conf.type = "log-log")
KM_Risk
```

```
ggsurvplot(KM_Risk, data = Y.test, xlab = "Time in years"
           ,risk.table = TRUE,title="KM - Risk Factor", risk.table.height=.3
           ,surv.median.line = "hv", legend.title="", tables.col = "strata"
           ,tables.y.text = FALSE, legend.labs = c("High Risk", "Low Risk"))
```

```
Call: survfit(formula = Y.test ~ Risk_Factor, conf.type = "log-log")

                         n events median 0.95LCL 0.95UCL
Risk_Factor=High Risk 29     19   4.54    2.34    7.58
Risk_Factor=Low Risk  29      5     NA    8.45      NA
```



The result shows that median survival time for "High Risk" patients is 4.54 years and as unknown for the "Low Risk " patients.

What is Kaplan-Meier probability of surviving for 2, 4, and 6 years between "Low Risk" and "High Risk" patients?

```
summary(KM_Risk, time = c(2, 4, 6))
```

```
Call: survfit(formula = Y.test ~ X_risk, conf.type = "log-log")

                X_risk=High Risk
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     24       5    0.828  0.0701        0.634        0.924
    4     15       7    0.576  0.0934        0.374        0.733
    6      7       3    0.455  0.0966        0.263        0.628

                X_risk=Low Risk
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     29       0    1.000  0.0000        1.000        1.000
    4     24       1    0.964  0.0351        0.772        0.995
    6     19       1    0.920  0.0544        0.715        0.980
```

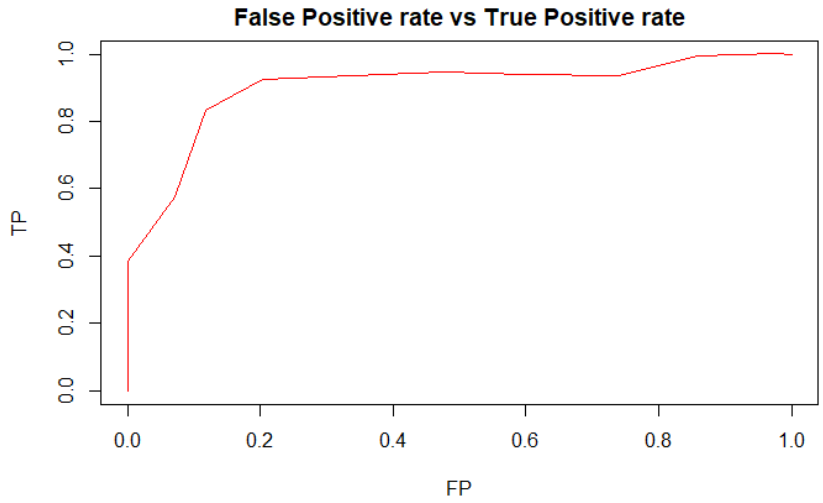
From the curve and summary, we observe that the possibility of surviving about 2 years for "High Risk" patient is roughly 82% vs 100% for "Low Risk" patient. The possibility of surviving about 4 years for "High Risk" patient is roughly 58% vs 97% for "Low Risk" patient. The possibility of surviving about 6 years for "High Risk" patient is roughly 46% vs 92% for "Low Risk" patient. At the same time, we also have the confidence interval ranges which show the margin of expected error.

What is the score of our prediction, is it good?

Finally, we want to test and plot the score of our prediction, and we can do that by using ROC (receiver operating characteristic curve) curve and AUC ("Area under the ROC Curve"):

```
cutOff = quantile(prediction.scores.scaled, prob = 0:10/10)
ROC <- survivalROC(Stime = Y.test[, 1],
                   status = Y.test[, 2],
                   marker = prediction.scores.scaled,
                   cut.values = cutOff,
                   predict.time = 5,
                   method = "KM")
ROC$AUC
with(ROC, plot(TP ~ FP, type = "l", xlim = c(0, 1), ylim = c(0, 1)))
```

```
1] 0.9023212
```



The result shows that our score is 90%, which is not bad at all!

Survival Forest

We will use the Ranger library to extract the variables importance:

```
Ranger_Data = PbcData[complete.cases(PbcData), ]
Ranger_Model <- ranger(Surv(Ranger_Data$time
                            , Ranger_Data$status) ~.
                       ,data = Ranger_Data
                       , num.trees = 500
                       , importance = "permutation"
                       ,seed = 1)
 data.frame(sort(Ranger_Model$variable.importance
                 ,decreasing = TRUE))
```

bili	0.080365	ast	0.0059025
copper	0.0223767	id	0.0053388
albumin	0.0135705	hepato	0.0027206
edema	0.0112782	alk.phos	0.0020533
age	0.0107358	spiders	0.0011826
ascites	0.0106054	trig	0.0007937

chol	0.0079491	sex	0.0005739
stage	0.0073557	trt	-0.000375
protime	0.0068208	platelet	-0.001779

Comparison between KM, Cox HP, Random Forest models

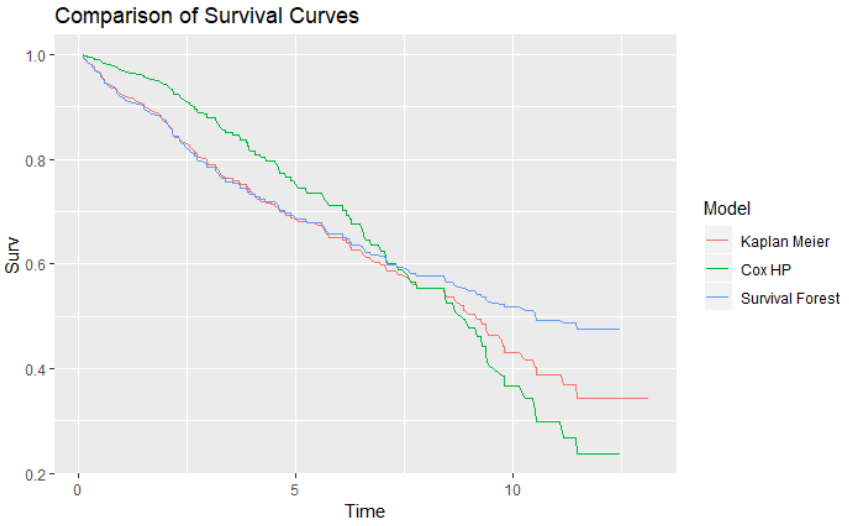
```
km <- rep("Kaplan Meier", length(KM_Model$time))
cox_surv <- survfit(COX_Model)
cox <- rep("Cox HP", length(cox_surv$time))
rf <- rep("Survival Forest",length(Ranger_Model$unique.death.times))

# Create a dataframe
km_df <- data.frame(KM_Model$time, KM_Model$surv, km)
cox_df <- data.frame(cox_surv$time, cox_surv$surv, cox)
rf_df <-
data.frame(Ranger_Model$unique.death.times,sapply(data.frame(Ranger_Model$sur
vival),mean),rf)

# Rename the columns so they are same for all dataframes
names(km_df) <- c("Time","Surv","Model")
names(cox_df) <- c("Time","Surv","Model")
names(rf_df) <- c("Time","Surv","Model")

# Combine the results
plot_combo <- rbind(km_df,cox_df,rf_df)

# Make a ggplot
plot_gg <- ggplot(plot_combo, aes(x = Time, y = Surv, color = Model))
plot_gg + geom_line() + ggtitle("Comparison of Survival Curves")
```



We see here that Cox model is the most unsteady one with the most data and features. It is higher for lower values and drops down sharply when the time increases. The survival forest is of the lowest range and resembles Kaplan-Meier curve. The difference might be because of Survival forest having less records because we removed all the NA values from the dataset (this might fetch us a better R2 and more stable curves).