Survival analysis

— Primary Biliary Cirrhosis(pbc)

Content:

Description of the data

The dataset is from the R package “Mayo Clinic Primary Biliary Cirrhosis Data”. This 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: T Therneau and P Grambsch (2000),Â *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York. ISBN: 0-387-98784-3.

Status are reverted to censored/translate=0, dead=1; for variable sex: female=1, male=0, for simplicity, we turned days into years

```{r}
# recode status&sex
dat$status = ifelse(dat$status == 2, 1, 0)
dat$sex = ifelse(dat$sex == "f", 1, 0)
table(dat$status)/length(dat$status)

# convert the time in days to years for simplicity
dat$time = floor(dat$time / 365)
```

Descriptive statistics

- Sample Size :

There are 418 rows of data in total, while after remove all rows with missing values from the dataframe, 276 rows are left.

```{r}
# remove all rows with missing values from the dataframe
dat <- dat[complete.cases(dat), ]
```

- Variables distribution:

We studied the distribution of 2 variables, sex and age, results are shown below:

Sex: Female= 1; Male= 0

```
 0   1
34 242
```

Age:

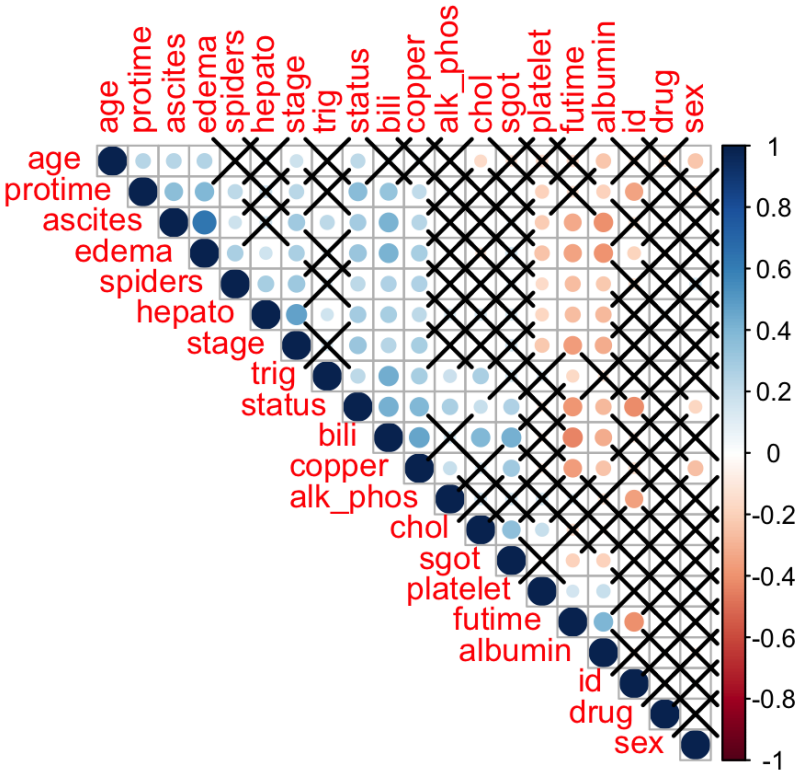

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
26.28   41.51   49.71   49.80   56.58   78.44
```

As shown by the distribution schema, we can conclude that the prevalence rate of woman are about 7 times of men's.
The ages of PCB patients are normally distributed. The median age is approximately 49 years. Patients range from ages 26 to 78.The age distribution of the most potential patients lies on the range [42,57].

```{r}
# Create a histogram of age
ggplot(dat, aes(x=age)) +
    geom_histogram(binwidth=2, colour="black", fill="pink") + ggtitle("Distribution of Age")
+ xlab("Age (years)") +  ylab("Frequency") +   ggtitle("Histogram of Age") +
  theme(plot.title = element_text(hjust = 0.5)) + scale_x_continuous(breaks = seq(0, 80,
by = 3))
summary(dat$age)
table(d$sex)
```

- **Correlation matrix:**

To study the association of the parameters, we calculated the correlation matrix below using the pearson correlation coefficient (PCC)which measures the linear dependence between two variables.

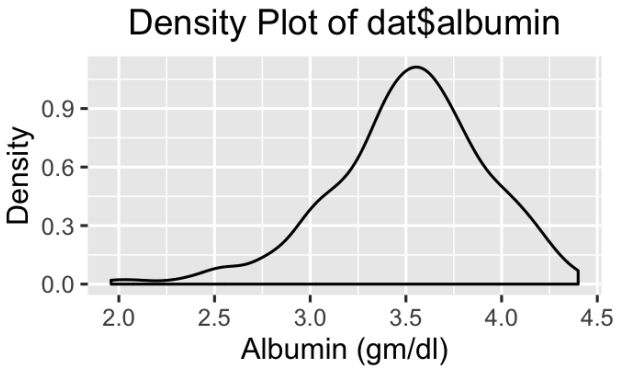


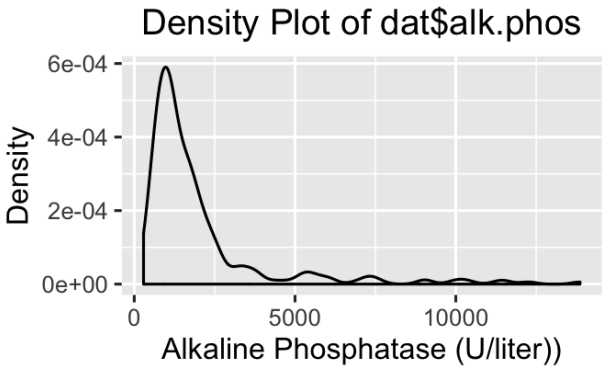
The most strongly associated covariates are `ascites` and `edema` which have a PCC about 0.7. It makes sense that these variables are positively correlated: they are both symptoms of liver disease. For other covariates, we can see that for some darker points, PCC is around 0.4, especially for some symptoms and health indices. That also makes sense because a disease usually have several symptoms and cause a serial of biological changes.

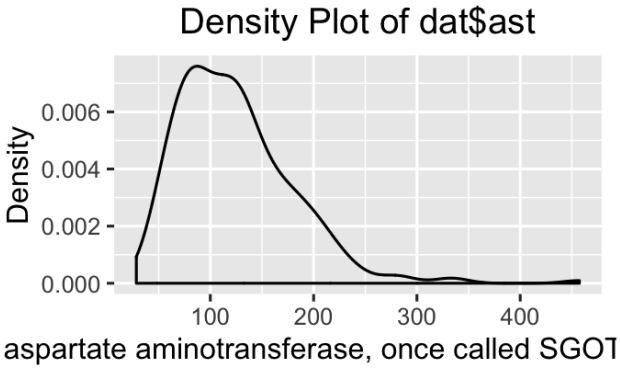
```{r}
# Correlation Matrix
cor.mtest <- function(mat) {
    mat <- as.matrix(mat)
    n <- ncol(mat)
    p.mat<- matrix(NA, n, n)
    diag(p.mat) <- 0
    for (i in 1:(n - 1)) {
        for (j in (i + 1):n) {
            tmp <- cor.test(mat[, i], mat[, j])
            p.mat[i, j] <- p.mat[j, i] <- tmp$p.value
        }
    }
  colnames(p.mat) <- rownames(p.mat) <- colnames(mat)
  p.mat
}
# matrix of correlation statistic
corr<-cor(dat)
# matrix of the p-value of the correlation
p.mat <- cor.mtest(dat)

# Specialized the insignificant value according to the significant level
corrplot(corr, type="upper", order="hclust",
         p.mat = p.mat, sig.level = 0.05)
```

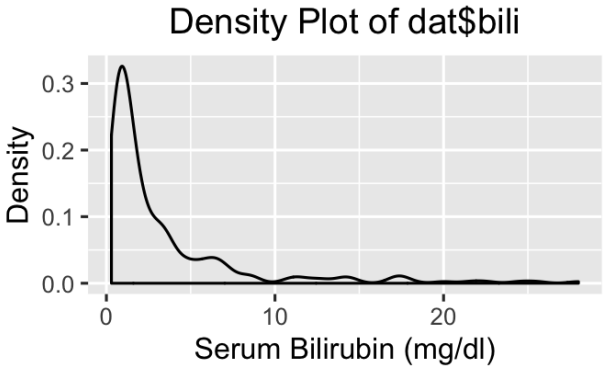
- **Density Distribution of Quantitative Variables**

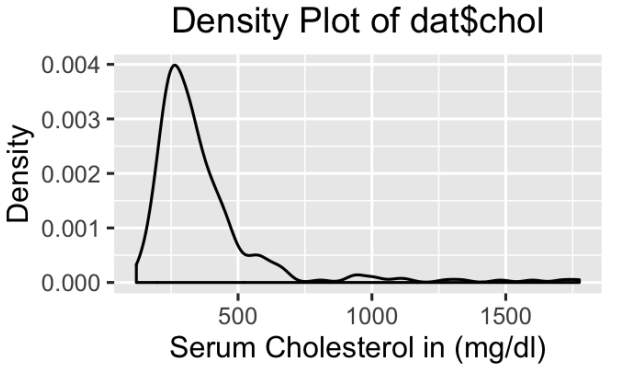
After observing the distribution of each variable, we found that expect the variable “albumin”, other variables are not normally distributed. On the contrary, the density in much intense at smaller values. Therefore we decided to apply a log transformation on certain variables.

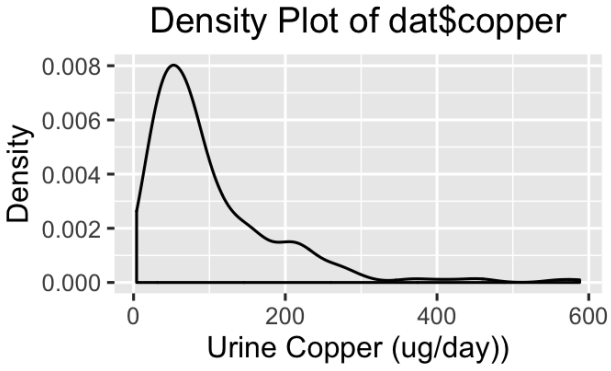

Density Plot of dat$albumin
Density
0.0
0.3
0.6
0.9
2.0
2.5
3.0
3.5
4.0
4.5
Albumin (gm/dl)

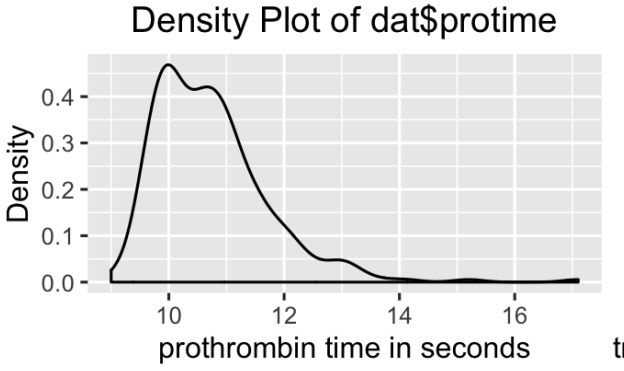

Density Plot of dat$alk.phos
Density
0e+00
2e-04
4e-04
6e-04
0
5000
10000
Alkaline Phosphatase (U/liter))

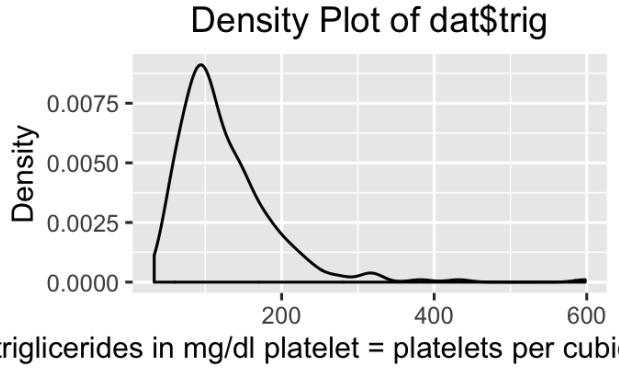

Density Plot of dat$ast
Density
0.000
0.002
0.004
0.006
100
200
300
400
aspartate aminotransferase, once called SGOT


Density Plot of dat$bili
Density
0.0
0.1
0.2
0.3
0
10
20
Serum Bilirubin (mg/dl)


Density Plot of dat$chol
Density
0.000
0.001
0.002
0.003
0.004
500
1000
1500
Serum Cholesterol in (mg/dl)


Density Plot of dat$copper
Density
0.000
0.002
0.004
0.006
0.008
0
200
400
600
Urine Copper (ug/day))


Density Plot of dat$protime
Density
0.0
0.1
0.2
0.3
0.4
10
12
14
16
prothrombin time in seconds


Density Plot of dat$trig
Density
0.0000
0.0025
0.0050
0.0075
200
400
600
triglicerides in mg/dl platelet = platelets per cubi

```{r}
# Arrange and display the plots
grid.arrange(density_plot(l[1]), density_plot(l[2]),
             density_plot(l[3]),density_plot(l[4]),
             density_plot(l[5]), density_plot(l[6]),
             density_plot(l[7]), density_plot(l[8]),
ncol=2)
```

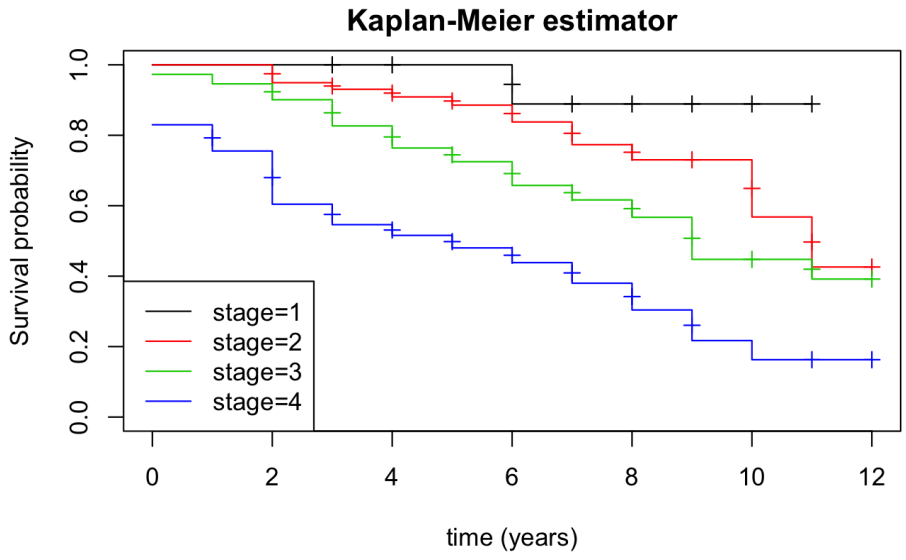
Methods&Results

Four methods are exploited: **nonparametric estimation, Logrank test, Cox regression, Automatic model selection based on AIC**

- **nonparametric estimation**
- Kaplan-Meyer estimator

Here we specifically estimate the impact of different 'histologic stage of disease' on the survival curve, below are the code and curves.

```{r}
#  Kaplan-Meyer estimator
fit.KM <- survfit(Surv(time, status) ~ stage, data = dat)
plot(fit.KM, mark.time = TRUE,
    main = "Kaplan-Meier estimator",
    ylab = "Survival probability",
    xlab = "time (years)" ,col = 1:4)
legend("bottomleft", lty = 1, col = 1:4, legend = names(fit.KM$strata))
```



Only one event for stage one, which causes the discontinuity of the curve. Generally, we can conclude that the early the stage, the less the risk of death. And the trend for each stage is almost linear. What's more, according to the curve, patients of all stages have the possibility to survival more than 12 years.

- Median survival

We explored the median value of this model to see the time when 50% patients still survive.

```{r}
fit.KM
```

Result:

```
          n events median 0.95LCL 0.95UCL
stage=1  12      1     NA      NA      NA
stage=2  59     14     11      10      NA
stage=3 111     41      9       8      NA
stage=4  94     55      5       3       8
```

As shown above, we don't know the median value of stage one because only one patient died at the 6th year, no other death case for this stage. For stage 2, 3 and 4, more than 50% patients can survive more than 11, 9, 5 years respectively.

- Test the possibility to survive more than 10 years

If we are interested in the possibility of survival more than 10 years, we can do as following.

```{r}
summary(fit.KM, time = 10)
```

Result：

```
            stage=1
      time       n.risk      n.event     survival      std.err lower 95% CI
    10.000        4.000        1.000        0.889        0.105        0.706
upper 95% CI
     1.000
             stage=2
      time       n.risk      n.event     survival      std.err lower 95% CI
    10.000        9.000       13.000        0.568        0.117        0.380
upper 95% CI
```

```
      0.850
             stage=3
       time       n.risk      n.event     survival      std.err lower 95% CI
     10.000       10.000       40.000        0.448        0.072        0.327
upper 95% CI
      0.614
             stage=4
       time       n.risk      n.event     survival      std.err lower 95% CI
    10.0000       4.0000      55.0000       0.1629       0.0716       0.0688
upper 95% CI
     0.3854
```

As shown above, for stage 1 - 4, the possibility to survive more than 10 years are respectively 88.9%, 56.8%, 44.8% and 16.3%.

- **Logrank test**

Logrank test aimes to compare two or more simples with a simple function survdiff(). For this report, we still test the quantity of test for 4 stages

```
```{r}
fit.logrank <- survdiff(Surv(time, status) ~ stage, data = dat)
fit.logrank
```
```

Result：

```
             N Observed Expected (O-E)^2/E (O-E)^2/V
stage=1  12        1     7.31     5.446      6.34
stage=2  59       14    28.39     7.297     10.60
stage=3 111       41    47.26     0.828      1.55
stage=4  94       55    28.04    25.924     37.86

 Chisq= 43.4  on 3 degrees of freedom, p= 2e-09
```

We can observe that the stage 3 is best fitted. And the p-value is equal to 2e-09, far more less than 0.05, we can take the stage as variable with significant impact.

- **Cox regression**

- Testing the Proportional-Hazards Assumption with **Schoenfeld residuals**

The proportional hazards assumption is supported by a non-significant relationship between the Schoenfeld residuals and time. The two must be independent.

```{r}
res.cox <- coxph(Surv(time, status) ~ trt + age + sex + ascites + hepato +
                 spiders + edema + bili + chol + albumin + copper + alk.phos +
                 ast + trig + protime + stage , data = dat)
test.ph <- cox.zph(res.cox)
test.ph
```

Result:

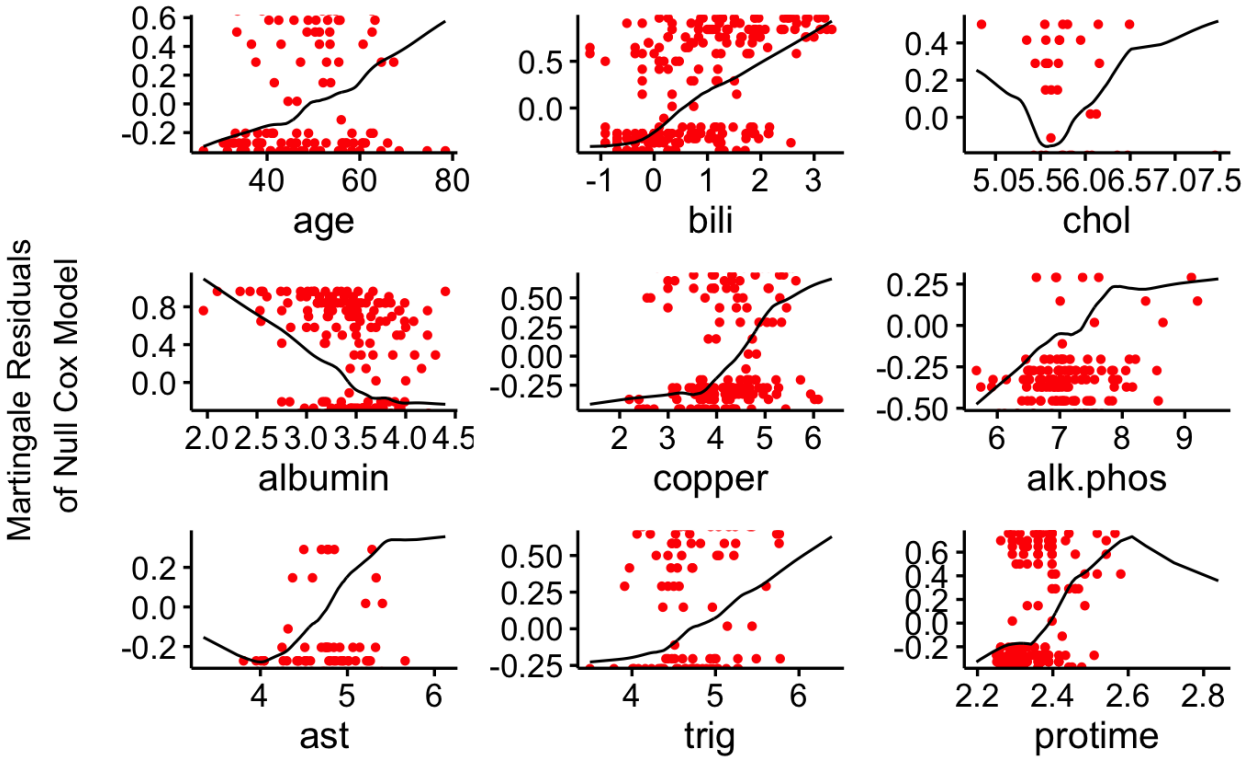
	rho	chisq	p
trt	-0.04190	0.22910	0.632
age	-0.04698	0.24400	0.621
sex	-0.05520	0.42768.	0.513
ascites	-0.03009	0.13280.	0.716
hepato	0.03591	0.16370	0.686
spiders	0.09601	1.34126	0.247
edema	-0.13052	2.49736	0.114
bili	0.07415	0.75332.	0.385
chol	0.01091	0.01546	0.901
albumin	-0.08089	0.96330	0.326
copper	-0.11371	1.73216	0.188
alk.phos	0.02596	0.07344	0.786
ast	-0.00648	0.00487	0.944
trig	0.05576	0.35282	0.553
protime	-0.06537	0.53208	0.466
stage	-0.11573	1.53406	0.216
GLOBAL	NA	14.00482	0.598

The output above shows that the test is not statistically significant for each of the covariates, and the global test is also not statistically significant.

- Testing the Proportional-Hazards Assumption with **Martingale residuals**

The Cox Model assumes that continuous covariates have a nonlinear form. We can test the nonlinearity assumption by plotting the Martingale residuals against the continuous covariates.

```{r}
ggcoxfunctional(Surv(time, status) ~  age + bili + chol + albumin + copper +
                alk.phos + ast + trig + protime, data = dat)
```



The assumption of non linearity is supported.

- Fit the Cox model

Since the Proportional-Hazards Assumption is satisfied by checking with Schoenfeld residuals and Martingale residuals, now we can fit the model with all the covariates by the code below.

```{r}
fit.cox <- coxph( Surv(time, status)  ~ sex + age + stage + trt + albumin + alk.phos +
ascites + ast + bili+ chol + copper + edema + hepato + platelet + protime + spiders + trig
, data = dat)
summary(fit.cox)
```

Result:

```
 n= 276, number of events= 111

              coef  exp(coef)   se(coef)      z Pr(>|z|)
sex      -0.1520740  0.8589247  0.3165963 -0.480  0.63099
age       0.0325391  1.0330743  0.0113855  2.858  0.00426 **
stage     0.3519564  1.4218465  0.1749122  2.012  0.04420 *
trt       0.0250927  1.0254101  0.2088758  0.120  0.90438
albumin  -0.4340506  0.6478795  0.2895648 -1.499  0.13388
alk.phos  0.1364856  1.1462384  0.1434828  0.951  0.34149
ascites   0.3350538  1.3980157  0.3771175  0.888  0.37429
ast       0.2949209  1.3430201  0.2992770  0.985  0.32441
bili      0.5813874  1.7885180  0.1778557  3.269  0.00108 **
chol      0.2747018  1.3161382  0.2834469  0.969  0.33247
copper    0.3179805  1.3743494  0.1774128  1.792  0.07308 .
edema     0.8647877  2.3745020  0.3844490  2.249  0.02449 *
hepato    0.0168308  1.0169732  0.2492129  0.068  0.94616
platelet  0.0003873  1.0003873  0.0011780  0.329  0.74235
protime   2.3315421 10.2938036  1.3289898  1.754  0.07937 .
spiders   0.0652191  1.0673928  0.2376536  0.274  0.78375
trig     -0.1977972  0.8205363  0.2513733 -0.787  0.43136
—
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Concordance= 0.867  (se = 0.018 )
```

The concordance of the model is 0.867 with 17 parameters, the four most significant variables are age, stage, bill as well as edema who have p-value less than 0.05.

- Automatic model selection based on AIC

As can be shown by cox fitting, only certain variables contribute significant impact on the result, so the next step we try to use automatic model selection based n AIC to filter significant variables.

```
```{r}
Mfull <- coxph(Surv(time, status) ~ sex + age + stage + trt + albumin + alk.phos +
ascites + ast + bili+ chol+copper+edema+hepato+platelet+protime+sex+spiders+trig,
data = train)
MAIC <- step(Mfull)
summary(MAIC)
```
```

Result:
— The last step:
Step: AIC=634.75
Surv(time, status) ~ sex + age + stage + albumin + alk.phos + ast + bili + chol + edema

```
          Df    AIC
<none>       634.75
- sex       1 634.96
- ast       1 635.29
- alk.phos  1 635.49
- albumin   1 636.84
- stage     1 637.06
- chol      1 637.88
- age       1 638.91
- bili      1 641.77
- edema     1 642.25
```

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
sexf	-5.095e-01	6.008e-01	3.324e-01	-1.533	0.12532	
age	3.426e-02	1.035e+00	1.381e-02	2.480	0.01313	*
stage	3.739e-01	1.453e+00	1.838e-01	2.035	0.04190	*
albumin	-6.939e-01	4.996e-01	3.383e-01	-2.052	0.04022	*
alk.phos	7.329e-05	1.000e+00	4.194e-05	1.747	0.08056	.
ast	4.199e-03	1.004e+00	2.595e-03	1.618	0.10557	
bili	9.886e-02	1.104e+00	3.054e-02	3.237	0.00121	**
chol	1.082e-03	1.001e+00	4.424e-04	2.446	0.01444	*
edema	1.479e+00	4.387e+00	4.587e-01	3.224	0.00127	**

—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Concordance= 0.852 (se = 0.025)

9 parameters are left with concordance=0.852, then we try the 6 most important variables and see the result.

```{r}
#try 6 most important variables
try<-coxph(Surv(time, status) ~  age + stage + albumin +  bili+ chol+edema, data = train)
summary(try)
```

Result:

```
             coef  exp(coef)   se(coef)      z Pr(>|z|)
age      0.0376801  1.0383990  0.0128872  2.924  0.00346 **
stage    0.3524443  1.4225404  0.1743310  2.022  0.04321 *
albumin -0.8026603  0.4481352  0.3237188 -2.479  0.01316 *
bili     0.1209979  1.1286225  0.0258149  4.687 2.77e-06 ***
chol     0.0009907  1.0009912  0.0004220  2.348  0.01890 *
edema    1.2059054  3.3397815  0.4398107  2.742  0.00611 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Concordance= 0.839  (se = 0.027 )
```

Conclusion

According to the comparison of the model built with 9 variables and the one with 6 variables, we can conclude that for this model, less parameters will save time and space of computation, while on the contrary, we found that the concordance is less than the model with 9 variables. This shows a trade-off between complexity of model and concordance. Therefore the model must be adapted by the certain exigence of concordance or complexity, we can only find a relative good model in certain condition. We can temporarily choose the last model of 6 variables as the best model so far since there is no big difference of concordance between this model and the one with 9 parameters. And also as shown in the complexity matrix, several parameters are related with each other because of the biological complexity of human body. Hence, several parameters are deleted may not change a lot since the related ones are remained in the modified model.

The disadvantages during the modeling lies on the insufficiency of data. As mentioned above, only one death case is recorded in the dataset, which causes the lack of persuasion for the result. And this dataset in all without the consideration of null values contains only 276 cases, which is relatively small as a dataset. Therefore, conclusively, one important way to improve the model quality is to enlarge the dataset.