Survival Analysis for the pbc Dataset

Zilan Cheng

April 6, 2022

1 Introduction

In this report, we work on the data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. The dataset contains the baseline measurements and survival data of 418 pbc patients, among which 312 subjects provide largely complete data, 106 out-of-trial cases contains basic measurements records and survival data. The dataset contains 20 variables. To specify, **age**, **albumin**, **alk.phos**, **ascites**, **ast**, **bili**,**chol**, **copper**, **edema**, **hepato**, **id**, **platelet**, **protime**, **sex**, **spiders**, **stage**, **status**, **time**, **trt**(D-penicillmain) and **trig**.
The main purpose of our study is to investigate the impact of variables, for instance, the drug D-penicillamine, to the lifetime of patients with Primary Biliary Cirrhosis (PBC). We are interested in both the qualitative and the quantitative analysis of the variables' impact.
The analysis on this report is based on **R 4.0.4**.

2 Exploratory data analysis

Before conducting survival analysis, we first take a glance at the data.
For univariate analysis, we visualize the numerical variables by the boxplots in **Figure 1**. Some basic statistics for these numerical variables are given in **Table 1**. In addition, the barplots of categorical variables are given in **Figure 2**. According to the table and figures, we find that there are several missing values (NA), which result from the presence of censoring.

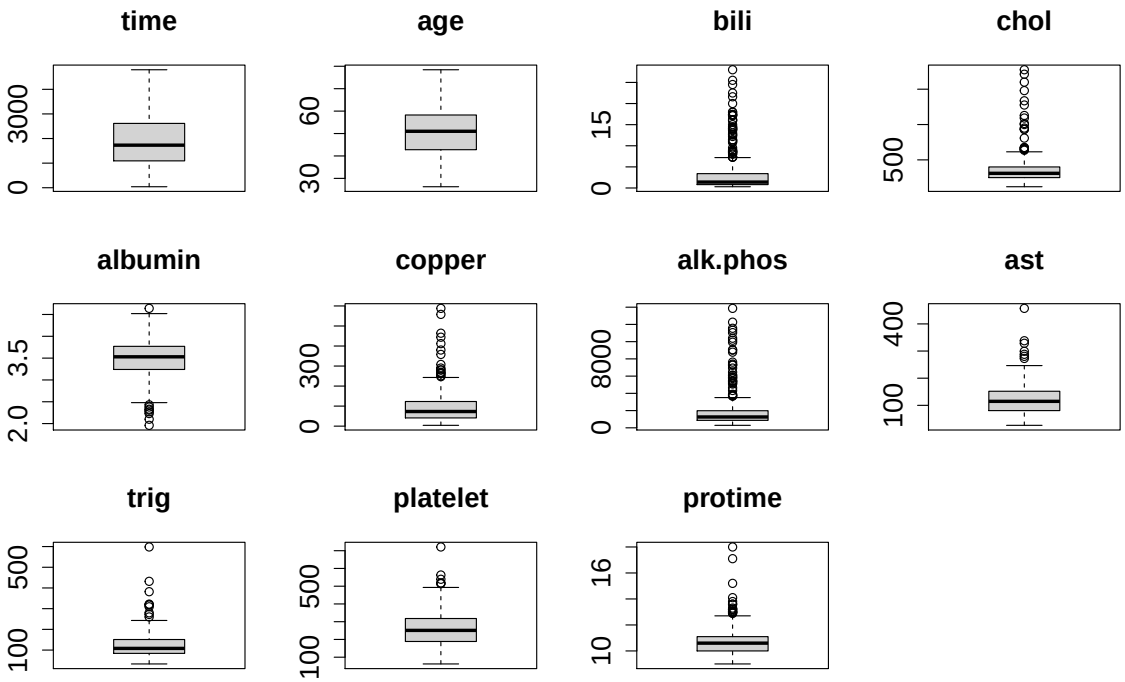

Figure 1: Boxplots for the numerical variables

Table 1: Summary statistics for the numerical variables

	time	age	bili	chol	albumin	copper	alk.phos	ast	trig	platelet	protime
Min.	41	26.28	0.300	120.0	1.960	4.00	289.0	26.35	33.00	62.0	9.00
1st Qu.	1093	42.83	0.800	249.5	3.243	41.25	871.5	80.60	84.25	188.5	10.00
Median	1730	51.00	1.400	309.5	3.530	73.00	1259.0	114.70	108.00	251.0	10.60
Mean	1918	50.74	3.221	369.5	3.497	97.65	1982.7	122.56	124.70	257.0	10.73
3rd Qu.	2614	58.24	3.400	400.0	3.770	123.00	1980.0	151.90	151.00	318.0	11.10
Max.	4795	78.44	28.000	1775.0	4.640	588.00	13862.4	457.25	598.00	721.0	18.00
NA's	-	-	-	134	-	108	106	106	136	11	2

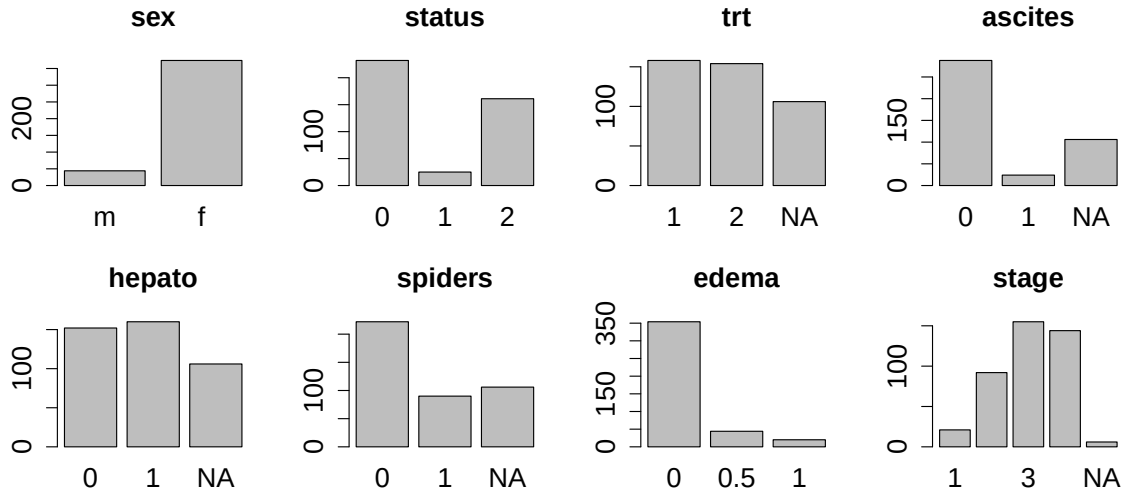

Figure 2: Barplot

We now implement the bivariate analysis. After prepossessing, we give the correlation matrix for numerical variables and the corresponding p-values in **Figure 3.** In the correlation matrix, positive correlations are displayed in blue and negative correlations in red color.The correlation coefficient values are leaved blank if p-values>0.01, since in this case the correlations lose statistical significance.

According to the correlation matrix, the absolute value of correlation coefficients between time and bili, albumin and copper are high, we are smaller than 0.5. We guess that they may have a relatively strong correlation, which will be checked in the following survival model.

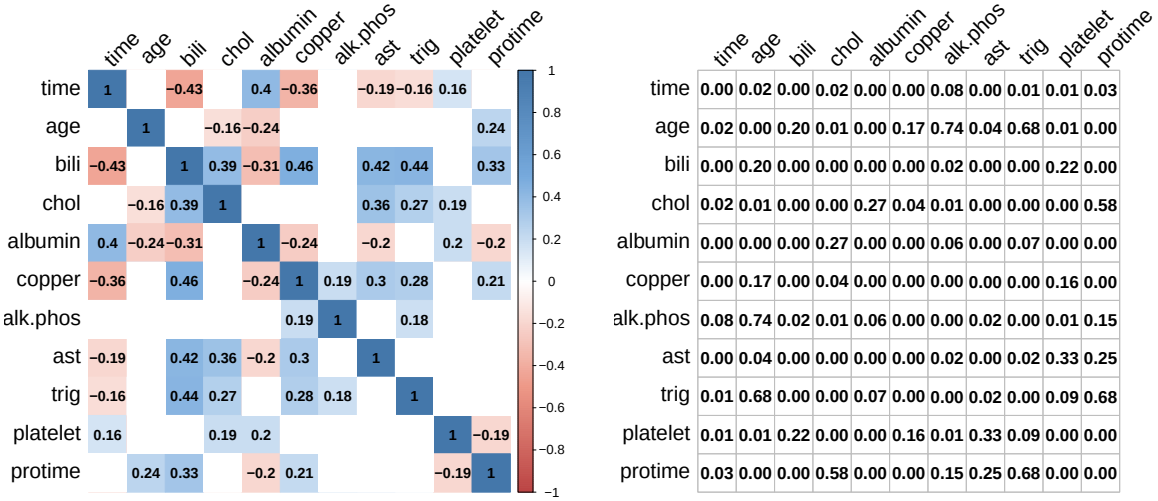

Figure 3: Correlation matrix and corresponding p-values

3 Survival curves

3.1 Simple survival curves

The survival probability, also known as the survival function $S(t)$, is the probability that an individual survives from the time origin. We employ the Kaplan-Meier(KM) estimator to estimate the survival function. The definition of the estimation is given as follows:

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} (1 - \frac{d_j}{r_j})$$

where $t_{(j)}$ is the jth largest unique survival time, r_j is the number of individuals at risk just before $t_{(j)}$ (including censored individuals at $t_{(j)}$), d_j is the number of individuals experiencing the event at time $t_{(j)}$.
We use the function **survfit()** to compute Kaplan-Meier survival estimate and the function **ggsurvplot()**to obtain the survival curves.

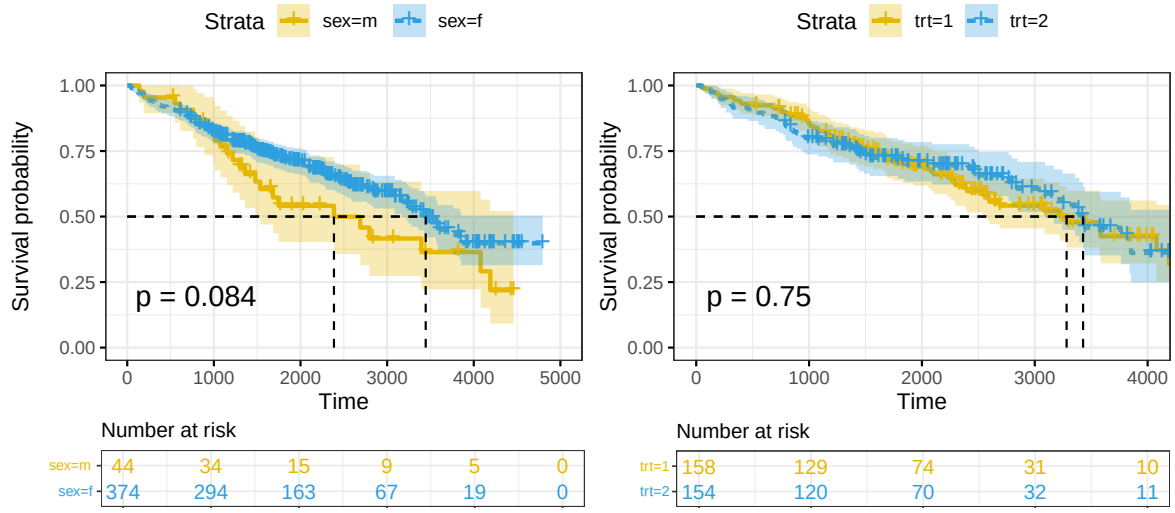

Figure 4: Survival probability by **trt**(left) and **sex**(right)

For the categorical data, we could produce the survival curves for multiple groups of subjects. For instance, the survival probability by trt and sex (respectively) is visualized in **Figure 4**. There appears to be a slight survival advantage for female with primary biliary cirrhosis compared to male. By checking the output the median survival time for male is 2386 days, as opposed to 3445 days for female. On the contrary, the survival functions of the two **trt** groups look like similar. However, to evaluate whether this difference is statistically significant requires a formal statistical test.
We conduct the log-rank test to compare multiple survival curves. The null hypothesis is that there is no difference in survival between different groups. If the p value is smaller than 0.05, we could make a conclusion that the null hypothesis is rejected.
For the trt and the sex groups, the log-rank test using inherent function **survdiff()** gives us the p-value of 0.7 and 0.08, respectively. Both are larger than 0.05. Hence we could not conclude that the two sex groups or the two trt groups differ significantly in survival.

3.2 Complex survival curves

Although **sex** or **trt** might not have statistically significant impact on the discrepancies of survival curves, we still would like to know if the result will be different on the case that computing the survival curves using the combination of multiple factors(**sex** and **trt** here). **Figure 5** gives the visualization of our output.

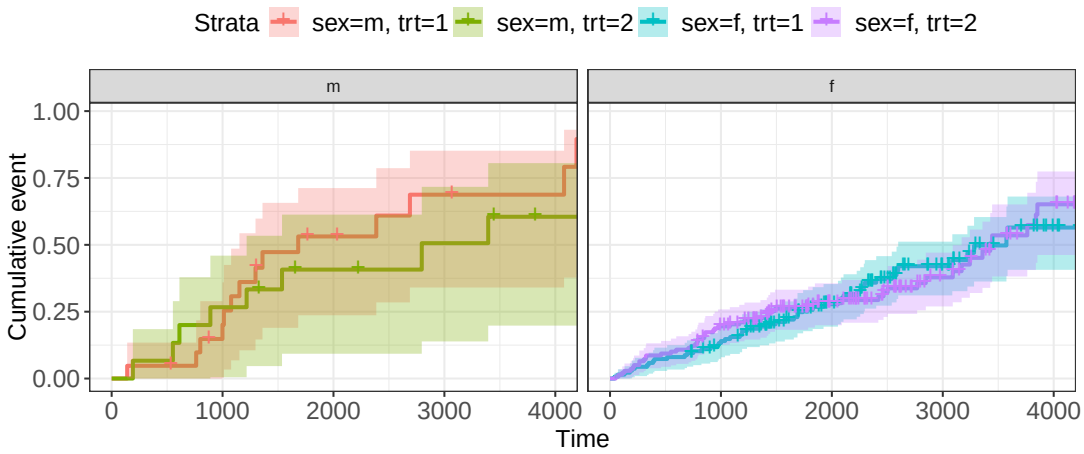

Figure 5: Survival curves using the combination of **sex** and **trt**

From **Figure 5**, the survival curves seem not to present obvious discrepancies. In order to check, we carry out the log-rank test as follows:

Table 2: Log-rank test on the complex survival curves

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
sex=m,trt=1	21	14	7.73	5.080	5.447
sex=m,trt=2	15	8	6.89	0.180	0.192
sex=f,trt=1	137	51	55.49	0.363	0.654
sex=f,trt=2	139	52	54.90	0.153	0.273
$\chi^2 = 5.8$ on 3 degrees of freedom, $p = 0.1$					

According to **Table 2** above, the p-value $p = 0.1 > 0.05$. Hence we could not conclude that the four groups have statistically significant impact on the lifetime of the patients.

4 Model fitting

The analysis of just survival curves above has limitations. For one thing, we could not solve the functional relation between the variables and the survival time but only obtain the qualitative results (whether the discrepancy between survival curves is statistically significant or not); for another thing, we could only work on the categorical data rather than the numerical data (unless we stratify the variables).
We try to fit a model. The first step is the qualitative step, that is, to choose the significant variables that will be included in the model. According to Collett(1), many advocate the approach of first doing a univariate analysis to 'screen' out potentially significant variables for consideration in the multivariate model. So we first employ univariate Cox model to obtain the potentially significant variables. Then we use Akaike information criterion(AIC) to select the final significant variables stepwisely based on the potential variables. Finally, we obtain the quantitative model by multivariate Cox model.

4.1 Univariate Cox model

The Cox proportional-hazards model can be written as:

$$h(t) = h(t_0)exp(\beta \mathbf{X})$$

where h_0 is the baseline function, i.e., the hazard function for individuals with all explanatory variables =0.
Below are the 3 assumptions of Cox proportional-hazards model :

- The hazards are proportional.
- The covariates have linear combinations (including possibly higher order items, interactions).
- There is no outliers. That is to say, the model never poorly predicted based on the dataset.

We apply the univariate coxph function to multiple covariates at once and obtain **Table 3** below:

Table 3: Univariate Cox Model

	beta	HR	(95% CI for HR)	wald.test	p.value
trt	-0.057	0.94	(0.66-1.3)	0.1	0.75
age	0.039	1	(1-1.1)	25	5.9e-07
sex	-0.38	0.68	(0.44-1.1)	2.9	0.086
ascites	2.1	7.8	(4.9-12)	75	5.3e-18
hepato	1.2	3.3	(2.2-4.8)	36	2e-09
spiders	0.97	2.6	(1.8-3.8)	28	1.1e-07
edema==1	2.2	9.2	(5.8-15)	85	2.9e-20
bili	0.14	1.2	(1.1-1.2)	150	1.4e-34
chol	0.0011	1	(1-1)	14	0.00024
albumin	-1.5	0.21	(0.15-0.31)	72	2.7e-17
copper	0.0067	1	(1-1)	77	2.1e-18
alk.phos	5e-05	1	(1-1)	2.6	0.11
ast	0.0061	1	(1-1)	30	4.4e-08
trig	0.0042	1	(1-1)	15	9.4e-05
platelet	-0.0026	1	(1-1)	8.8	0.003
protime	0.26	1.3	(1.2-1.4)	36	2.2e-09
stage	0.8	2.2	(1.8-2.7)	54	2.2e-13

The results can be interpreted as follows:

- Betas represent the regression coefficients. A positive sign means that the hazard (risk of death) is higher and thus the prognosis worse. Take **edema==1** as an example, the beta coefficient 2.2 indicates edema despite diuretic therapy corresponds to higher risk of death based on the dataset.
- HR(exp(beta)) denotes the hazard ratios, which gives the effect size of covariates. For example, edema despite diuretic therapy increases the hazard by a factor of 920%.
- 95% CI for HR gives the 95% confidence intervals of hazard ratios.
- The p-value is obtained by Wald test, which shows the statistical significance of each univariate Cox regressions. The variates **age**, **ascites**,**hepato**, **spiders**, **edema**, **bili**, **chol**, **albumin**, **copper**, **ast**, **trig**, **platelet**, **protime** and **stage** have highly statistically significant coefficients, while the coefficients for **trt**, **sex** and **alk.phos** are not significant.

We take the 6 covariates with biggest absolute values of beta and acceptable p values, that is, **albumin**, **ascites**, **edema**, **hepato**, **stage** and **spiders**, as our potential significant variables.

4.2 Akaike Information Criterion

Akaike information criterion(AIC) is one of the most popular and readily available methods for model fit criteria. The AIC method estimates the expected Kullback-Leibler (KL) information (Kullback and Leibler, 1951), a measure of the information lost when using an approximating distribution for estimation and inference instead of the true (unknown) distribution. The degree of freedom of the model give a bias correction to the expected KL information in large samples and act as a penalty on the numbers of parameters in the model. The optimal model minimizes AIC with respect to degree of freedom providing a balance between model fit (via the log-likelihood) and parsimony (df).
The formula of AIC is defined as follows:

$$AIC = -2\log L + 2df$$

in which df denotes the model effective degrees of freedom, n indicates the total sample size, $\log L$ is the log partial likelihood.
We use the stepwise AIC (using the inherent function **stepAIC** in R) to select the variables. Based on the model with potential significant variables, we take a greedy approach:
(1) calculate the AIC after elimination of each variable that having been included in the model;
(2) calculate the AIC after adding of each candidate having not been included in the model.
If the minimum AIC at one step is achieved by (1), then we take a backward step and eliminate the specific variable from the model; If the minimum AIC at one step is achieved by (2), then we take a forward step and add the corresponding candidate to the model. The procedures terminates when the AIC cannot descent whether adding a new candidate or eliminating an existing variable.

Table 4: Stepwise AIC

Step	Variables in the model	Change	AIC
0	albumin, ascites, edema==1, hepato, stage, spiders	-	1008.01
1	albumin, ascites, edema==1, hepato, stage, spiders, copper	+copper	993.46
2	albumin, ascites, edema==1, hepato, stage, spiders, copper, chol	+chol	986.68
3	albumin, ascites, edema==1, hepato, stage, spiders, copper, chol, age	+age	976.42
4	albumin, ascites, edema==1, hepato, stage, spiders, copper, chol, age, ast	+ast	973.63
5	albumin, ascites, edema==1, stage, spiders, copper, chol, age, ast	-hepato	971.98
6	albumin, ascites, edema==1, stage, copper, chol, age, ast	-spiders	970.65
7	albumin, edema==1, stage, copper, chol, age, ast	-chol	969.47
8	albumin, edema==1, stage, copper, chol, age, ast, trig	+trig	968.78

The automatic variable selection procedure is shown in **Table 4**. As shown in this table, we decide the variables in the final model, that is: **albumin**, **edema**, **stage**, **copper**, **chol**, **age**, **ast**, and **trig**.

4.3 Multivariate Cox analysis

With the 8 significant covariates, we consider the model as follows:

$$h(t) = h(t_0)exp(\beta \mathbf{X})$$

where

$$exp(\beta \boldsymbol{X}) = exp(\beta_1 \cdot \text{albumin} + \beta_2 \cdot (\text{edema} == 1) + \beta_3 \cdot \text{stage} + \beta_4 \cdot \text{copper} + \beta_5 \cdot \text{chol} + \beta_6 \cdot \text{age} + \beta_7 \cdot \text{ast} + \beta_8 \cdot \text{trig})$$

Table 5: Multivariate cox model

	coef	exp(coef)	se(coef)	z	Pr(> |z|)	lower .95	upper .95
albumin	-0.7721	0.4621	0.2650	-2.91	0.00358	0.2749	0.7767
edema==1	1.7128	5.5444	0.3355	5.11	3.30e-07	2.8726	10.7012
stage	0.5494	1.7322	0.1536	3.58	0.00035	1.2819	2.3407
copper	0.0034	1.0034	0.0010	3.52	0.00043	1.0015	1.0053
chol	0.0011	1.0011	0.0004	2.7	0.00693	1.0003	1.0018
age	0.0390	1.0398	0.0099	3.94	8.10e-05	1.0198	1.0602
ast	0.0037	1.0037	0.0016	2.29	0.02216	1.0005	1.0069
trig	0.0019	1.0019	0.0011	1.62	0.10472	0.9996	1.0041

Likelihood ratio test= 151 on 8 df, p= <2e-16
Wald test = 160 on 8 df, p= <2e-16
Score (logrank) test = 227 on 8 df, p= <2e-16

In **Table 5**, the p-value for all three overall tests (likelihood, Wald, and score) are significant, indicating that the model is significant. These tests evaluate the omnibus null hypothesis that all of the betas (β) are 0 and the test statistics shows that the omnibus null hypothesis is soundly rejected.
The covariate **trig** fails to be significant (p = 0.10, which is grater than 0.05). The value of HR is 1.0019, which means the 1mg/dl increase of triglycerides (mg/dl) will increase the hazard of death by 0.19%.
However, all other covariates are significant (p<0.05), which indicates a strong relationship between these covariates and risk of death. The HR of **albumin** is 0.4621, which means that high albumin level is related to good prognosis. The HR of **edema==1** is 5.5444, the HR of **satge** is 1.7322, the HR of **copper** is 1.0034, the HR of **chol** is 1.0011, the HR of **age** is 1.0398, the HR of **ast** is 1.0037: all of them are larger than 1. These 6 covariates also show a positive impact on the deaths.

5 Assessment of the Cox model

5.1 Testing proportional hazards (PH) assumption

The proportional hazards (PH) assumption can be checked using statistical tests and graphical diagnostics based on the scaled Schoenfeld residuals. For each covariate, the function **cox.zph()** correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole. The function **ggcoxzph()** gives the intuitive plots of Schoenfeld residuals.
According to the p-values shows in **Figure 6** (all>0.05), the test is not statistically significant for each of the covariates, and the global test is also not statistically significant. From the graphical inspection, the residuals center around 0. Besides, there is no pattern with time. The assumption of proportional hazards appears to be supported.

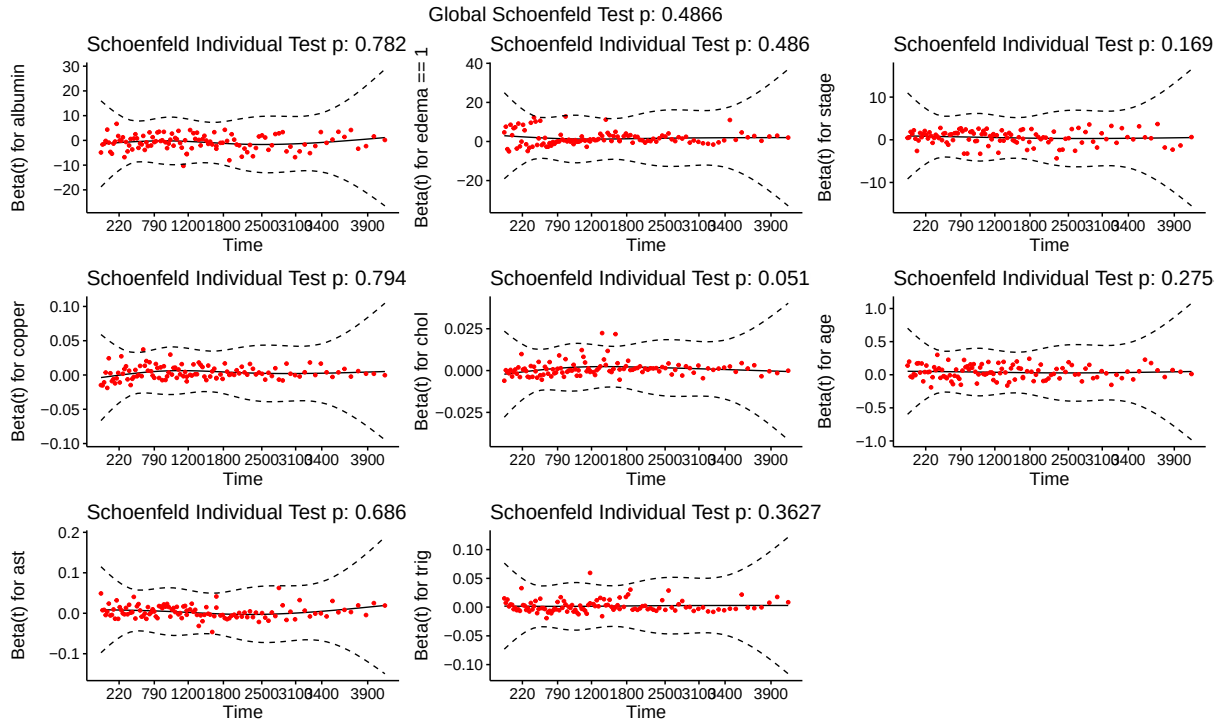

Figure 6: Schoenfeld residuals

5.2 Testing outliers

We employ the function **ggcoxdiagnostics()** with type="dfbeta" and type="deviance" to check the influential observations.

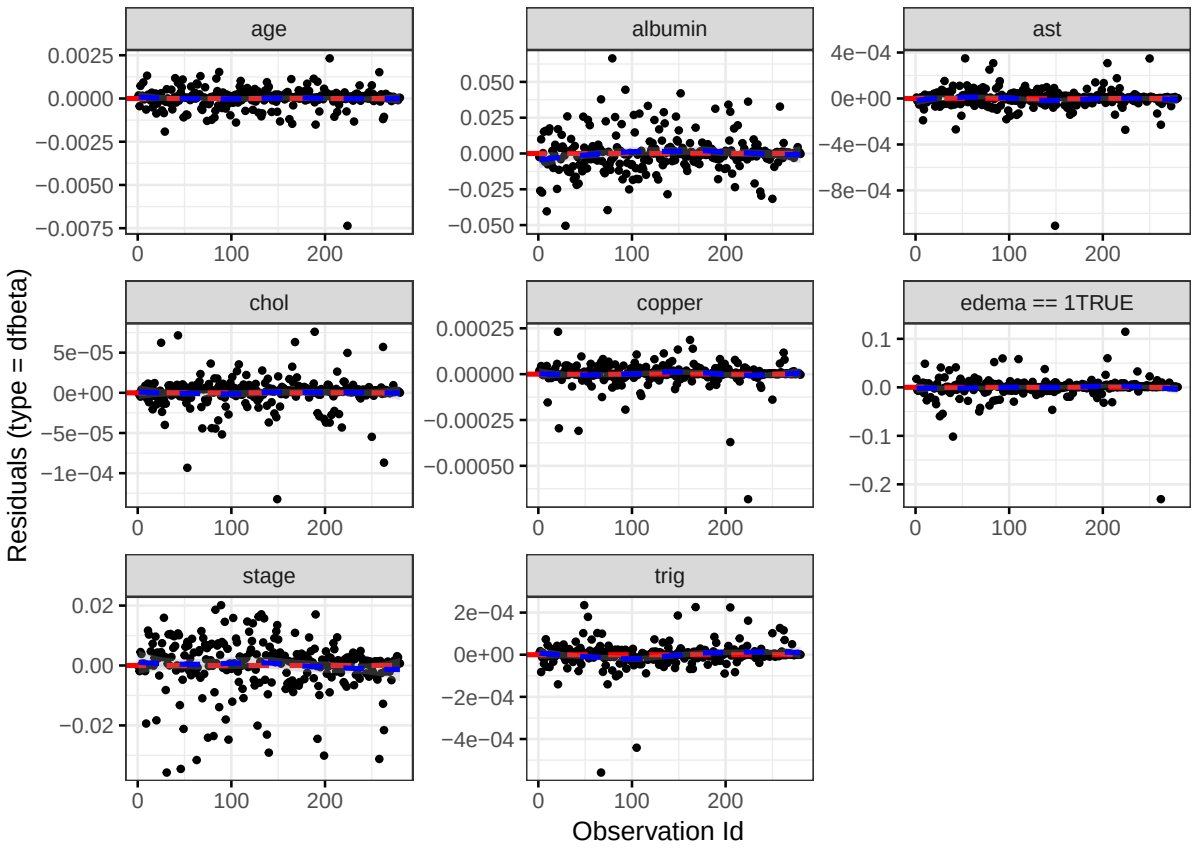

Figure 7: Dfbeta Residuals

Figure 7 gives the dfbeta residuals. It suggests that none of the observations is terribly influential individually, even though some of the absolute dfbeta values for age, ast, etc. are large compared with the others.

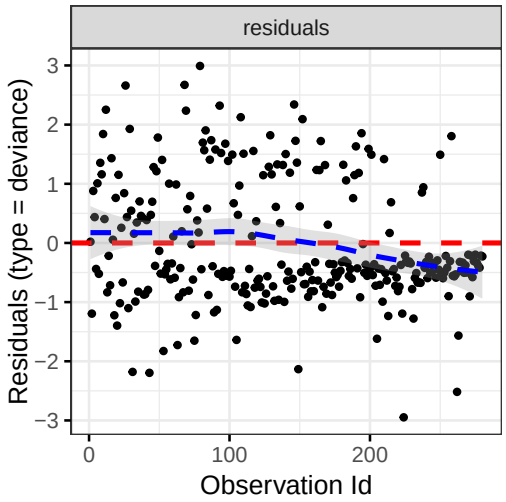

Figure 8: Deviance Residuals

Figure 8 gives the deviance residuals, which is a normalized transform of the martingale residuals. In the figure, the residuals roughly symmetrically distributed about 0. Positive values correspond to individuals that "died too soon" compared to expected survival times. Negative values correspond to individual that "lived too long". There still seems to be no extreme outliers (poorly predicted by the model).

5.3 Testing nonlinearity

Often, we assume that continuous covariates have a linear form. However, this assumption should be checked.

Plotting the Martingale residuals against continuous covariates is a common approach used to detect nonlinearity. In fact, there is only one continuous covariate, **albumin**, included in our final model.

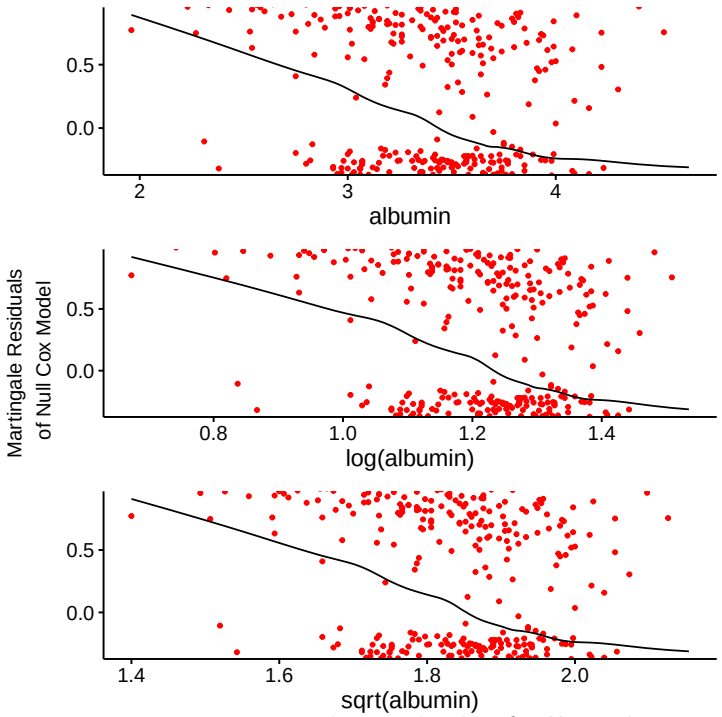

Figure 9: Martigale residuals of **albumin**

It appears in **Figure 9** that, the linearity holds.

6 Shortcomings

We preprocess the variable **edema** and obtain a logical value. 1 for edema despite diuretic therapy, 0 for no edema or untreated or successfully treated. It might be better if we refine the classification.

7 Conclusion

In this report, we analyze the Primary Biliary Cirrhosis (PBC) datasets. First of all, we implement the exploratory data analysis to check if making further analysis is reasonable. Then we calculate the survival functions using the Kaplan-Meier estimator. We compare both simple survival curves grouping based on only one variable and complex survival curves grouping based on multi variables by the low-rank test. Moreover, we establish a reasonable multivariate model. By the AIC criterion, we select several significant variables. The potential(original) variables are chosen by univariate Cox model and the selection method is stepwise selection (including both forward steps and backward steps). Finally, we develop the estimated multivariate Cox model and make the assessments according to the 3 assumptions of Cox PH model.
The final estimated model between basic measurements records and survival data is:

$$\hat{h(t)} = h(t_0)exp(\hat{\beta}\mathbf{X})$$

where

$$\begin{aligned} exp(\hat{\beta}\mathbf{X}) = exp(&-7.72 \times 10^{-1}\text{albumin} + 1.71(\text{edema} == 1) + 5.50 \times 10^{-1}\text{stage} \\ &+ 3.37 \times 10^{-3}\text{copper} + 1.06 \times 10^{-3}\text{chol} + 3.90 \times 10^{-2}\text{age} \\ &+ 3.71 \times 10^{-3}\text{ast} + 1.85 \times 10^{-3}\text{trig}) \end{aligned}$$

The negative coefficients indicate that the large amount of albumin is associated with good prognostic. However, the presence of edema despite diuretic therapy, the high level of copper, chol,ast, trig, the high histologic stage of disease and being old means higher hazard of deaths.
It is instructive for us to invent new drugs and conduct clinical medication to save the lives of the patients with Primary Biliary Cirrhosis. For instance, we could use ALB(Human serum albumin) in the treatment process.(2) Or we may consider other diagnosis for edema except diuretic therapy.

References

[1] Collett, David. Modelling Survival Data in Medical Research. CRC Press, 2015.

[2] Faloon, William W., Richard D. Eckhardt, T. Lynch Murphy, Arnold M. Cooper, and Charles S. Davidson. "AN EVALUATION OF HUMAN SERUM ALBUMIN IN THE TREATMENT OF CIRRHOSIS OF THE LIVER." The Journal of Clinical Investigation 28, no. 4 (July 1, 1949): 583–94. https://doi.org/10.1172/JCI102108.