STA9797 Project - Survival Analysis

Chi Nguyen - Huong Nguyen - Phuong Dao

2022-11-29

Abstract

Heart failure is a chronic condition when the heart cannot pump blood and oxygen to support the patient's body. This is the leading cause of death in America and the rest of the world. In this study, we focus on survival analysis on heart failure patient mortality, learning the impact of multiple attributes such as age, anemia, diabetes, creatinine phosphokinase, ejection fraction (%), high blood pressure, serum creatinine, serum sodium, sex, and smoking on the death events. The two main methods that were used in this study includes Kaplan-Meier estimator and the Cox Proportional Hazard Model.

Keywords: heart failure, survival analysis, Kaplan-Meier, Cox Proportional Hazard Model.

Introduction and Data Description

The dataset was collected at the Institute of Cardiology and Allied hospital Faisalabad-Pakistan, and is available on the University of California Irvine Machine Learning Repository for academic starting from 2020. The dataset are the cardiovascular medical records from 299 patients (105 female and 194 male) from the age of 40 or above. The follow-up time was between 4 to 285 days. All the patients have left ventricular systolic dysfunction and belonged to NYHA class II and IV.

Variables	Types	Description	Range
Age	Years	Age of the patient	[40,..,95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
High blood pressure	If a patient has hypertension	Boolean	0, 1
Creatinine phosphokinase (CPK)	Level of the CPK enzyme in the blood	mcg/L	[23,…, 7861]
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14,…, 80]

Variables	Types	Description	Range
Sex	Woman or man	Binary	0, 1
Platelets	Platelets in the blood	kiloplatelets/mL	[25.01,…, 850.00]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50,…, 9.40]
Serum sodium	Level of sodium in the blood	mEq/L	[114,…, 148]
Smoking	If the patient smokes	Boolean	0, 1
Time	Follow-up period	Days	[4,…,285]
DEATH EVENT	If the patient died during the follow-up period	Boolean	0, 1

Methods

Survival analysis is a set of statistical methods to analyze the time until an event occurs in a population. In this study, our event of interest is death. The survival function is the probability of surviving (not experiencing death) as a function of time. There are several approaches, including parametric, nonparametric and semi-parametric approaches for survival analysis, that study the relationship of a factor of interest to the time to event, in the presence of multiple covariates such as age, gender, blood pressure, etc. In the scope of this study, we focus on two methods: Kaplan Meier and Cox Proportional Hazard Model.

a. Kaplan-Meier Method

Kaplan-Meier method is a nonparametric estimator of the survival function. It is one of the most popular approach and is widely used in the univariate setting, in obtaining univariate descriptive statistics for survival data, such as median survival time, or comparing the survival experience for two or more groups of subjects. The Kaplan-Meier estimator of the survivorship function (survival probability) $S(t) = Pr(T \geq t)$ is:

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

- $\tau_1, \ldots \tau_k$ is the set of K distinct uncensored failure times observed in the sample.
- d_j is the number of failures at τ_j
- r_j is the number of individuals at risk right before the j-th failure time (everone who died or cencored at or after that time).

Censoring is a type of missing data problem unique to survival analysis. When there is no censoring, KM estimator is the same as the empircal estomator. The KM curve illustrates the cumilative survival probability over time. The curve is horizontal over periods where no event occurs, then drops vertically coressponding to a change in the survival function at each time an event occurs.

b. Cox Proportional Hazards Model

The most common regression model for the analysis of survival data is the Cox proportional hazards regression model. It allows testing for differences in survival times of two or more groups of interest, while allowing to adjust for covariates of interest. The Cox regression model is a semiparametric model, making fewer assumptions than typical parametric methods but more assumptions than those nonparametric methods described above.

$$h(t) = h_0(t)exp(\beta_1 x_1 \ + \beta_2 x_2 + \ldots + \beta_p x_p)$$

- $h(t)$ is the hazard function includes a set of p covariates of interest. The covariates can be continuous factors (eg. age, blood pressure, etc), discrete factors (gender, marital status, etc) or possible interaction (eg. age by sex interaction).
- t represents the survival time
- $\beta_1, \beta_2, .., \beta_p$ are the coefficients which measure the effect or impact of the covariates.
- $h(0)$ is the baseline hazard model which relects the underlying hazard for subjects with all covariates equal to 0 (i.e., the “reference group”).

If we limit to just one single categorical exposure variable:

$$h_1(t) = h_0(t)exp(\beta_1 x_1 \implies HR(t) := \frac{h_1(t)}{h_0(t)} = e^{\beta_1}$$

- $HR(t)$ is the hazard ratio, comparing the exposed to unexposed individuals at time t. This ratio remains constant over time t. A value of hazard ratio above 1 indicates that the covariate is positively associated with event probability (aka increase in hazard or negatively associated with the length of survival). Hazard ratio, which is less than 1, indicates that the covariate is negatively associated with event probability or reduction in hazard. When $HR = 1$, it means there is no effect.

Detailed Output

```
#Create a “survival object” Surv().
#For right-censored data, only two arguments are needed in the Surv() function: a vector
of times and a vector indicating which times are observed and censored.
mySobject <- Surv(time = data$time,
                  event = data$DEATH_EVENT)
```

1. Kaplan-Meier Estimates

```
kmsurvival <- survfit(mySobject ~ 1,
                  conf.int = 0.95, conf.type = "log")
#conf.type specifies the transformation used for calculating the confidence interval
 (="log" by default)
kmsurvival
```

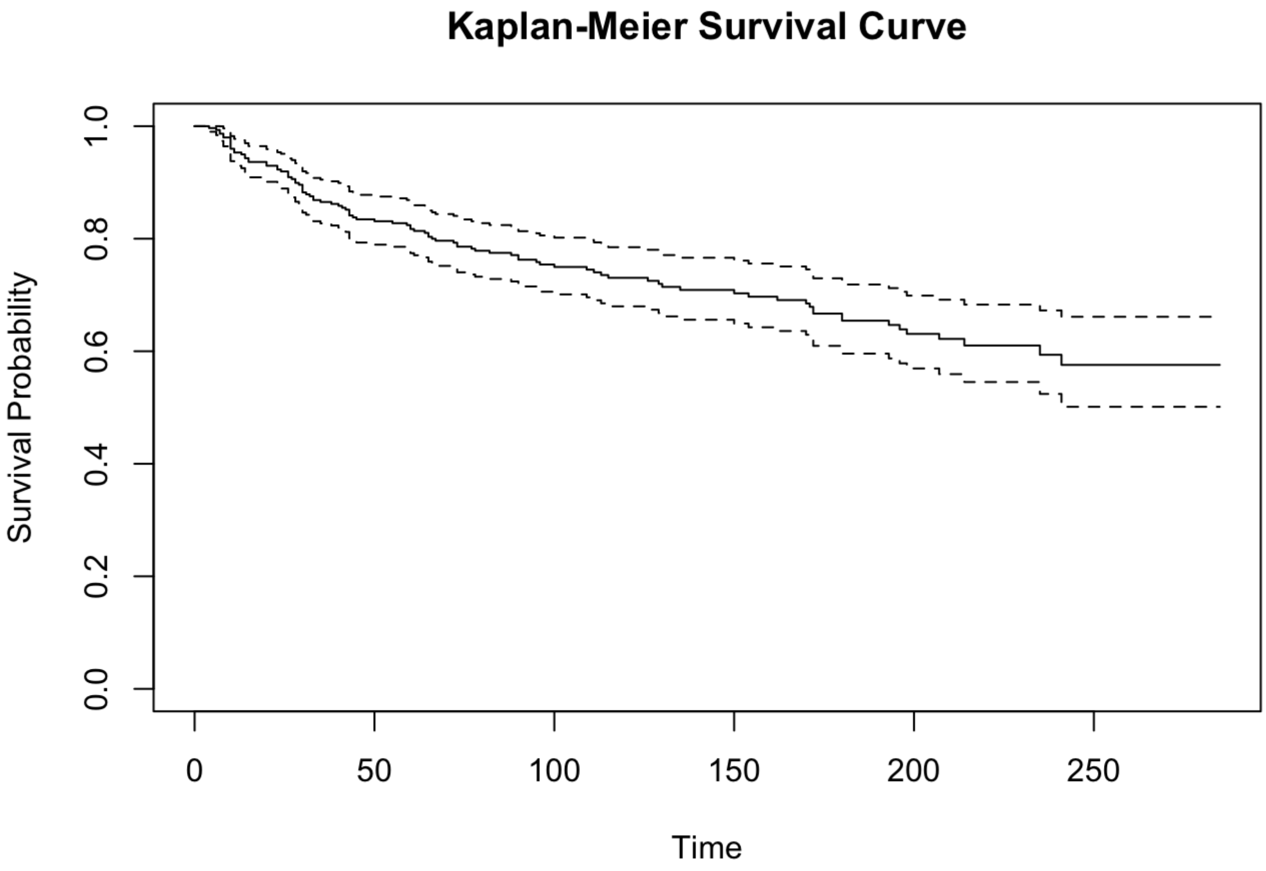
```
## Call: survfit(formula = mySobject ~ 1, conf.int = 0.95, conf.type = "log")
##
##        n events median 0.95LCL 0.95UCL
## [1,] 299     96     NA      NA      NA
```

```
#summary(kmsurvival)
```

```
#To get specific information out of the Kaplan-Meier model.
#summary(kmsurvival)$surv    # returns the Kaplan-Meier estimate at each t_i
#summary(kmsurvival)$time    # {t_i}
#summary(kmsurvival)$n.risk  # {Y_i}
#summary(kmsurvival)$n.event # {d_i}
#summary(kmsurvival)$std.err # standard error of the K-M estimate at {t_i}
#summary(kmsurvival)$lower   # lower pointwise estimates (alternatively, $upper)
```

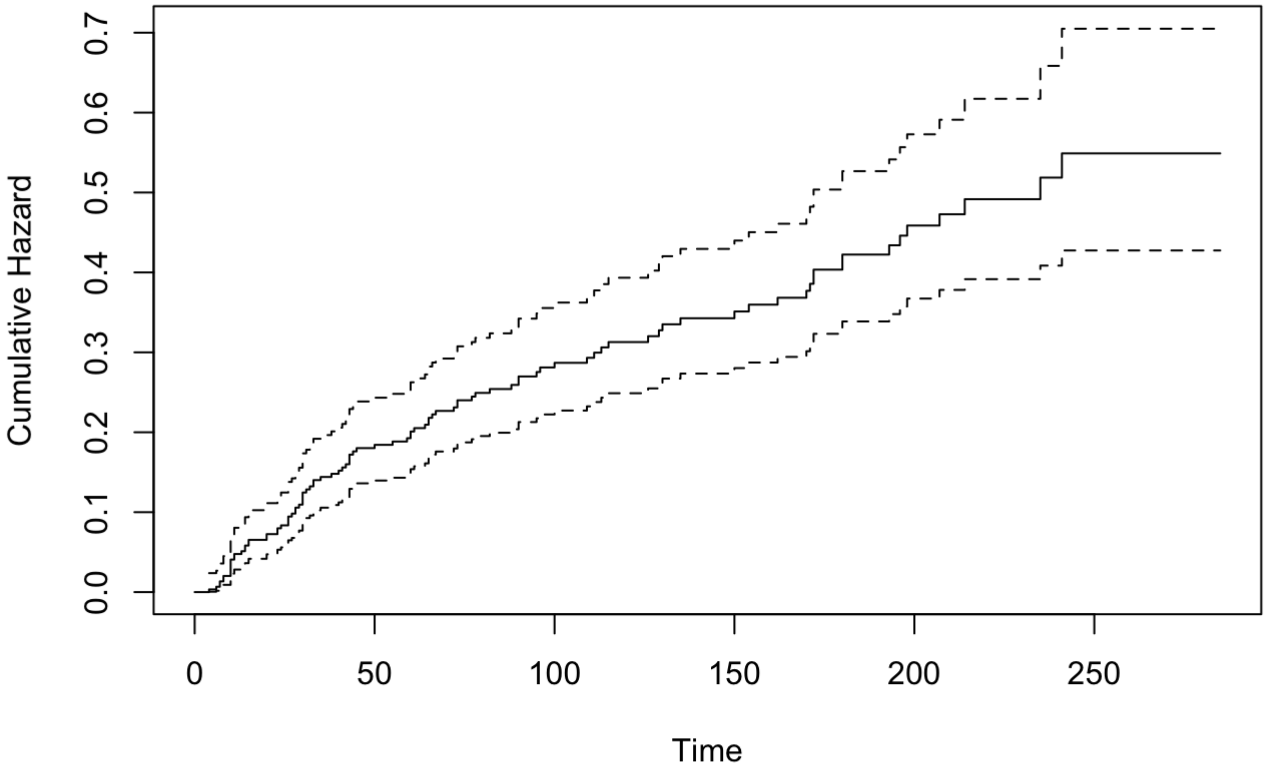
a. Plot the Kaplan-Meier Survival Curve

```
plot(kmsurvival,
     xlab="Time",
     ylab="Survival Probability",
     main="Kaplan-Meier Survival Curve")
```



b. Plot the cumulative hazard.

```
plot(kmsurvival, fun="cumhaz",
     xlab="Time",
     ylab="Cumulative Hazard")
```



###

c. Estimated survival of two groups

Categorical Variables

```
#categorical variables
#test the difference (using log-rank test)
survdiff(mySobject ~ data$anaemia, rho=0)
```

```
## Call:
## survdiff(formula = mySobject ~ data$anaemia, rho = 0)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## data$anaemia=0 170       50     57.9      1.07      2.73
## data$anaemia=1 129       46     38.1      1.63      2.73
##
##  Chisq= 2.7  on 1 degrees of freedom, p= 0.1
```

```
survdiff(mySobject ~ data$diabetes, rho=0)
```

```
## Call:
## survdiff(formula = mySobject ~ data$diabetes, rho = 0)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## data$diabetes=0 174       56       55    0.0172    0.0405
## data$diabetes=1 125       40       41    0.0231    0.0405
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.8
```

```
survdiff(mySobject ~ data$high_blood_pressure, rho=0)
```

```
## Call:
## survdiff(formula = mySobject ~ data$high_blood_pressure, rho = 0)
##
##                              N Observed Expected (O-E)^2/E (O-E)^2/V
## data$high_blood_pressure=0 194       57     66.4      1.34      4.41
## data$high_blood_pressure=1 105       39     29.6      3.00      4.41
##
##  Chisq= 4.4  on 1 degrees of freedom, p= 0.04
```

```
survdiff(mySobject ~ data$sex, rho=0)
```

```
## Call:
## survdiff(formula = mySobject ~ data$sex, rho = 0)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## data$sex=0 105       34     34.3   0.00254   0.00397
## data$sex=1 194       62     61.7   0.00141   0.00397
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

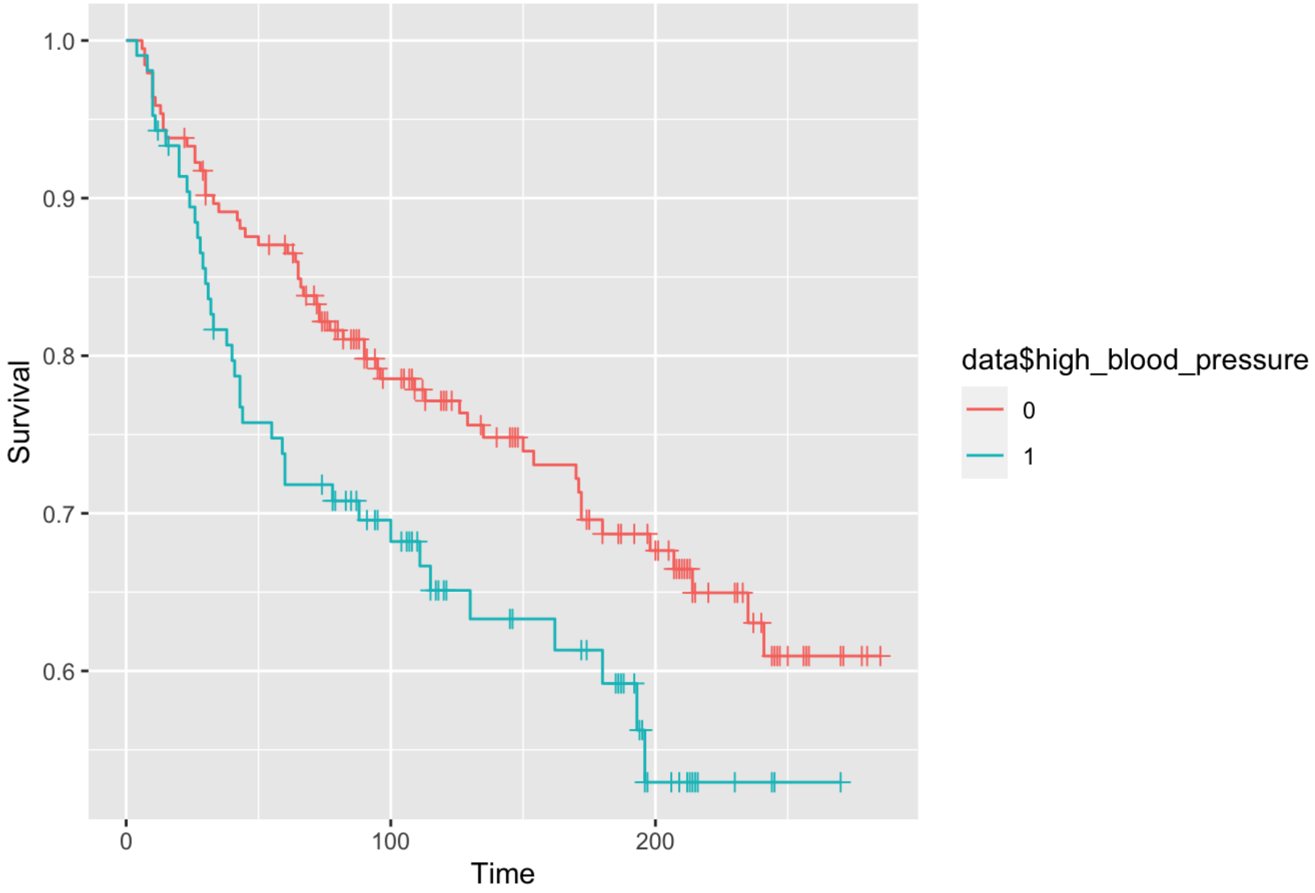
```
survdiff(mySobject ~ data$smoking, rho=0)
```

```
## Call:
## survdiff(formula = mySobject ~ data$smoking, rho = 0)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## data$smoking=0 203       66     65.8   0.00064   0.00204
## data$smoking=1  96       30     30.2   0.00139   0.00204
##
##  Chisq= 0  on 1 degrees of freedom, p= 1
```

```
#where rho=0 is the log-rank or Mantel-Haenszel test
#Note: The null hypothesis is that h1(t) = h2(t) for all t
```

‘anemia’, ‘smoking’, ‘sex’, and ‘diabetes’ are not statistically significant. For ‘high_blood_pressure’, we can reject null hypothesis and conclude that the survival curves are different for high blood pressure groups because the p-value = 0.04 (less than 0.05).

```
#plot the KM curve for high_blood_pressure
km_high_blood_pressure <- survfit(mySobject ~ data$high_blood_pressure)
ggsurv(km_high_blood_pressure)
```



Interpreting High Blood Pressure’s KM Plot

High BP adds to the heart’s workload. The KM estimate curve also shows a similar trend where patients with high blood pressure (blue line) are at an increased risk of survival due to heart failure with significant lower survival probability. The survival time for individuals with high blood pressure is significantly different from individuals with normal blood pressure.

Continuous Variables

```
#transform continuous variables to cargetorical
data$age_cat <- quantcut(data$age,4)
data$serum_sodium_cat <- quantcut(data$serum_sodium,4)
data$creatinine_phosphokinase_cat <- quantcut(data$creatinine_phosphokinase,4)
data$serum_creatinine_cat <- quantcut(data$serum_creatinine,4)
data$platelets_cat <- quantcut(data$platelets,3)
data$ejection_fraction_cat <- cut(data$ejection_fraction,
                       breaks=c(0, 30, 45, 100),
                       labels=c('EF < 30', '30 < EF < 45', 'EF > 45'))
```

```
#test the difference (using log-rank test)
survdiff(mySobject ~ data$age_cat, rho=0)
```

```
## Call:
## survdiff(formula = mySobject ~ data$age_cat, rho = 0)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## data$age_cat=[40,51] 78       20     26.8     1.705     2.380
## data$age_cat=(51,60] 84       24     28.0     0.573     0.813
## data$age_cat=(60,70] 85       21     27.9     1.702     2.408
## data$age_cat=(70,95] 52       31     13.3    23.343    27.399
##
##  Chisq= 27.6  on 3 degrees of freedom, p= 4e-06
```

```
survdiff(mySobject ~ data$serum_sodium_cat, rho=0)
```

```
## Call:
## survdiff(formula = mySobject ~ data$serum_sodium_cat, rho = 0)
##
##                                  N Observed Expected (O-E)^2/E (O-E)^2/V
## data$serum_sodium_cat=[113,134] 83       42     25.0    11.518     15.65
## data$serum_sodium_cat=(134,137] 94       24     30.3     1.311      1.93
## data$serum_sodium_cat=(137,140] 80       19     25.9     1.838      2.53
## data$serum_sodium_cat=(140,148] 42       11     14.8     0.964      1.14
##
##  Chisq= 15.7  on 3 degrees of freedom, p= 0.001
```

```
survdiff(mySobject ~ data$creatinine_phosphokinase_cat, rho=0)
```

```
## Call:
## survdiff(formula = mySobject ~ data$creatinine_phosphokinase_cat,
##     rho = 0)
##
##                                                   N Observed Expected (O-E)^2/E
## data$creatinine_phosphokinase_cat=[23,116]       75       19     22.7     0.593
## data$creatinine_phosphokinase_cat=(116,250]      76       28     23.2     1.000
## data$creatinine_phosphokinase_cat=(250,582]      84       32     27.8     0.631
## data$creatinine_phosphokinase_cat=(582,7.86e+03] 64       17     22.3     1.276
##                                                  (O-E)^2/V
## data$creatinine_phosphokinase_cat=[23,116]           0.780
## data$creatinine_phosphokinase_cat=(116,250]          1.325
## data$creatinine_phosphokinase_cat=(250,582]          0.892
## data$creatinine_phosphokinase_cat=(582,7.86e+03]     1.670
##
##  Chisq= 3.5  on 3 degrees of freedom, p= 0.3
```

```
survdiff(mySobject ~ data$serum_creatinine_cat, rho=0)
```

```
## Call:
## survdiff(formula = mySobject ~ data$serum_creatinine_cat, rho = 0)
##
##                                      N Observed Expected (O-E)^2/E (O-E)^2/V
## data$serum_creatinine_cat=[0.5,0.9] 81        9     28.1    12.972    18.428
## data$serum_creatinine_cat=(0.9,1.1] 82       24     27.1     0.365     0.512
## data$serum_creatinine_cat=(1.1,1.4] 64       18     20.7     0.341     0.436
## data$serum_creatinine_cat=(1.4,9.4] 72       45     20.1    30.808    39.185
##
##  Chisq= 44.7  on 3 degrees of freedom, p= 1e-09
```

```
survdiff(mySobject ~ data$platelets_cat, rho=0)
```

```
## Call:
## survdiff(formula = mySobject ~ data$platelets_cat, rho = 0)
##
##                                          N Observed Expected (O-E)^2/E
## data$platelets_cat=[2.51e+04,2.26e+05] 103       39     33.3     0.981
## data$platelets_cat=(2.26e+05,2.78e+05]  96       27     30.6     0.426
## data$platelets_cat=(2.78e+05,8.5e+05]  100       30     32.1     0.138
##                                        (O-E)^2/V
## data$platelets_cat=[2.51e+04,2.26e+05]     1.509
## data$platelets_cat=(2.26e+05,2.78e+05]     0.629
## data$platelets_cat=(2.78e+05,8.5e+05]      0.208
##
##  Chisq= 1.6  on 2 degrees of freedom, p= 0.5
```

```
survdiff(mySobject ~ data$ejection_fraction_cat, rho=0)
```

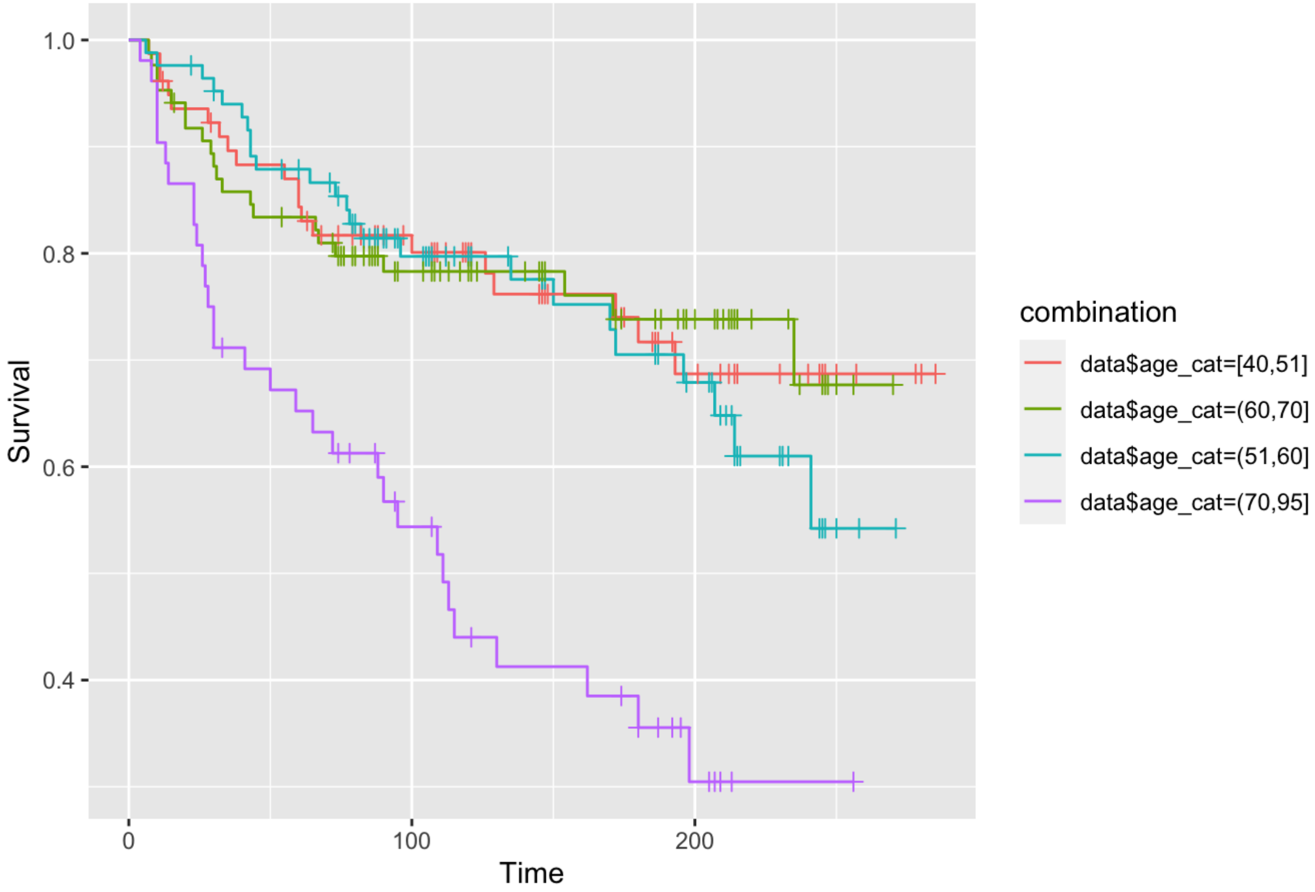
```
## Call:
## survdiff(formula = mySobject ~ data$ejection_fraction_cat, rho = 0)
##
##                                           N Observed Expected (O-E)^2/E
## data$ejection_fraction_cat=EF < 30       93       51     26.8     21.93
## data$ejection_fraction_cat=30 < EF < 45 146       31     51.2      7.97
## data$ejection_fraction_cat=EF > 45       60       14     18.0      0.90
##                                         (O-E)^2/V
## data$ejection_fraction_cat=EF < 30          30.63
## data$ejection_fraction_cat=30 < EF < 45     17.32
## data$ejection_fraction_cat=EF > 45           1.12
##
##  Chisq= 31.1  on 2 degrees of freedom, p= 2e-07
```

```
#where rho=0 is the log-rank or Mantel-Haenszel test
#Note: The null hypothesis is that h1(t) = h2(t) for all t
```

‘creatinine_phosphokinase_cat, and ’platelets’ are not statistically significant. For ‘ejection_fraction’, ‘serum_creatinine’, ‘serum_sodium’, and age’, we can reject null hypothesis and conclude that the survival curves are different for high blood pressure groups because the p-value = 0.04 (less than 0.05).

Interpreting Age’s KM Plot

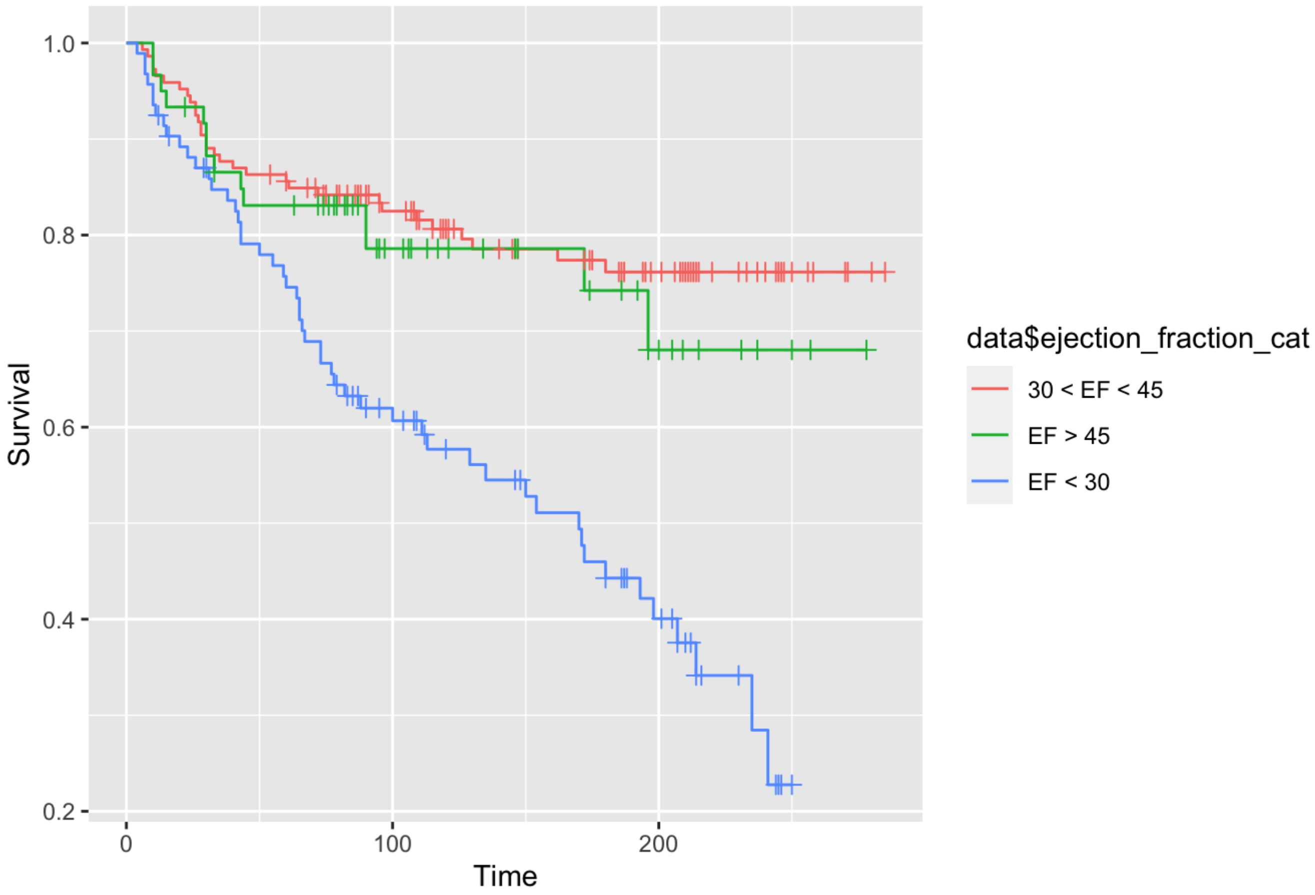
```
#Plot the KM curve for age variable
km_age <- survfit(mySobject ~ data$age_cat)
ggsurv(km_age)
```



The age of the population under consideration ranges from [40 - 95]. We divided this variables into 4 groups - [40 - 51], [51 - 60], [60 - 70] and [70 - 95]. The plot the survival probability for the above 70 group (purple line) is the least. The survival probability for the other two groups [40 - 51] and [51 - 60] (green and red line) are almost similar, and [60 - 70] (blue line) has a slightly lower chances of survival.

Interpreting Ejection Fraction's KM Plot

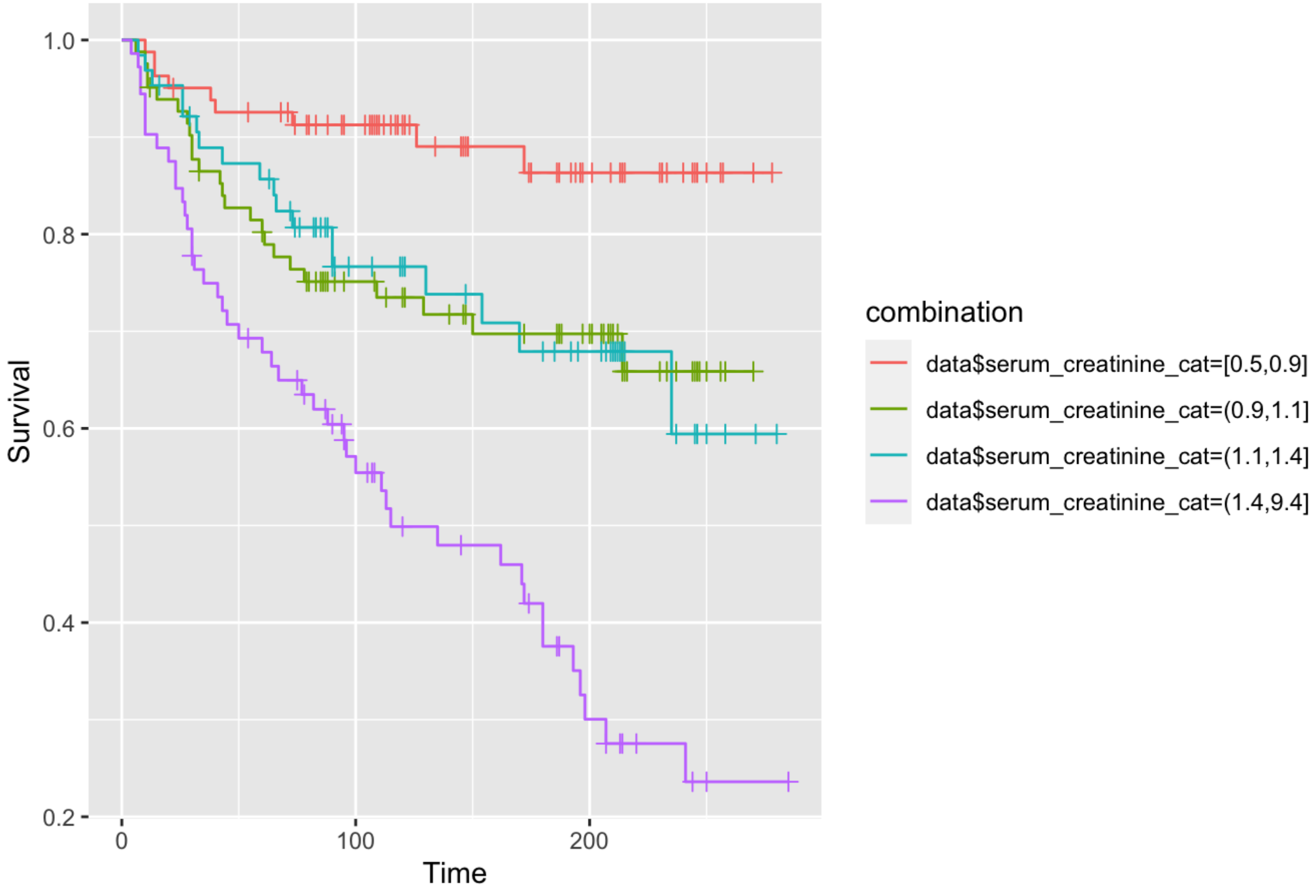
```
#Plot the KM curve for ejection_fraction
km_ejection_fraction <- survfit(mySobject ~ data$ejection_fraction_cat)
ggsurv(km_ejection_fraction)
```



Ejection Fraction (EF) is a measurement, expressed as a percentage, of how much blood the left ventricle pumps out with each contraction. In this study, ejection fraction ranges from 14 to 80. We converted into three buckets of [0 - 30], [30 - 45], and [45, 100]. KM estimate curves shows that patients with EF less than 30 have the least chance of surviving a heart failure condition. The blue line (EF <30) shows a significant different from other group. On the other hand, the patients with EF more than 30 show a comparable and decent chance of survival.

Interpreting Serum Creatinine's KM Plot

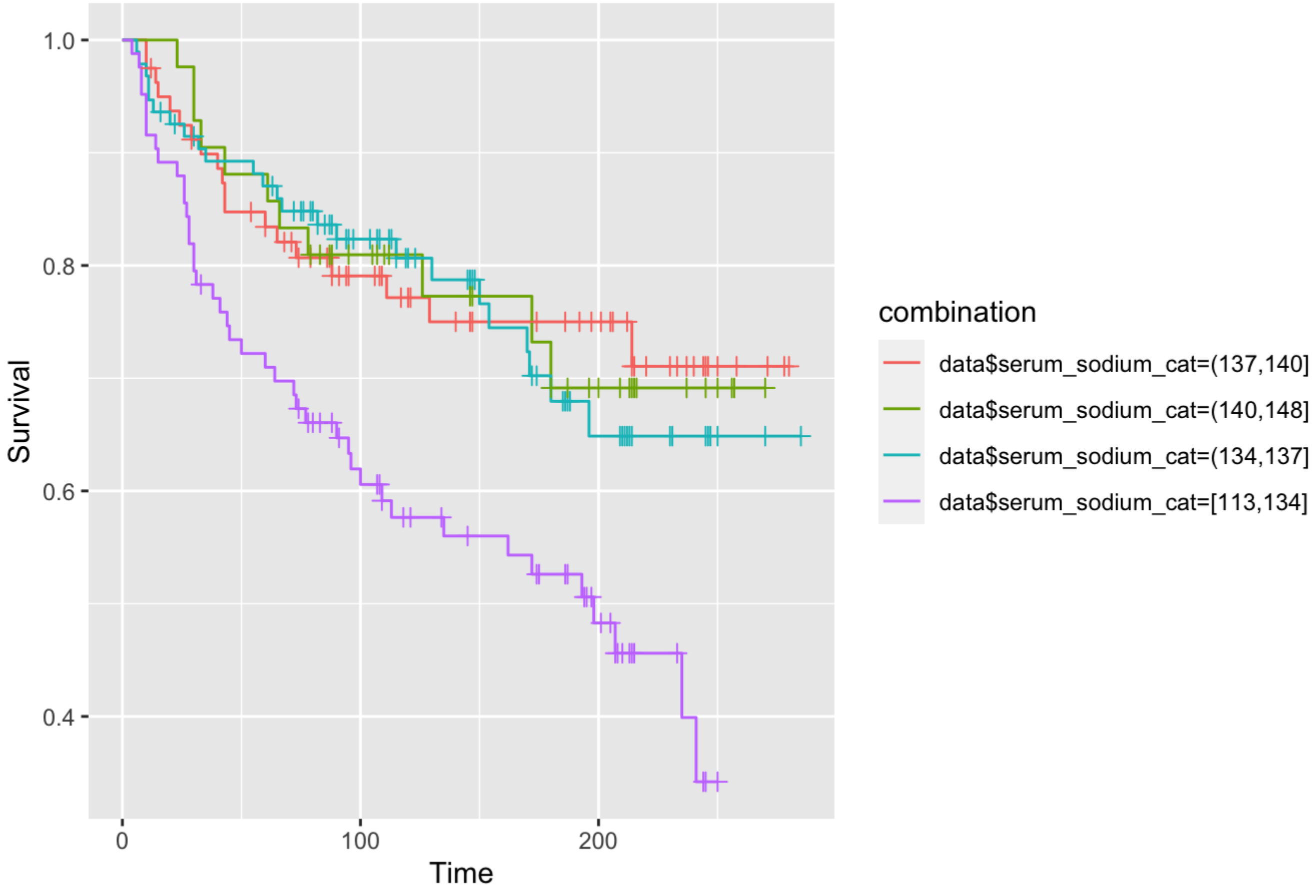
```
#Plot the KM curve for serum_creatinine variables

km_serum_creatinine <- survfit(mySobject ~ data$serum_creatinine_cat)
ggsurv(km_serum_creatinine)
```



Creatinine is a chemical waste product in the blood that passes through the kidneys to be filtered and eliminated in urine, measured in mg/dL. We divided serum_creatinine into four group by quantiles: [0.5 - 0.9], [0.9 - 1.1], [1.1 - 1.4] and [1.4 - 9.4]. The plot shows that the group with highesr level of serum creatinine [1.4 to 9.4] shows the lowest survival probability (purple line), while the group with lowest level of serum creatinine [0.5 - 0.9] shows the highest survival probability (red line).The middle groups, [0.9 - 1.1] and [1.1 - 1.4] shows comparable and similar survival probability.

Interpreting Serum Sodium's KM Plot

```
#Plot the KM curve for serum_sodium
km_serum_sodium <- survfit(mySobject ~ data$serum_sodium_cat)
ggsurv(km_serum_sodium)
```



The sodium blood test measures the concentration of sodium in the blood. In this study, the range of serum sodium levels for the population ranges from [113 - 148]. To plot the KM estimates curve, we divided the population into 4 buckets - [113 - 134], [134 - 137], [137 - 140] and [140 - 148]. From the KM plots, it is observed that survival probability is least for population falling under the 1st bucket that has the lowest sodium levels [113 - 134] - purple line. The survival time for individuals in the remaining three group with sodium levels more than 134 mEq/L appears significantly different, with higher probability of survival.

2. Cox's proportional hazards model

a. Univariate Cox Regression

```r
#apply the univariate coxph function to multiple covariates at once

covariates <- c("age", "anaemia", "creatinine_phosphokinase", "diabetes", "ejection_frac
tion", "high_blood_pressure", "platelets", "serum_creatinine", "serum_sodium", "sex", "s
moking")
univ_formulas <- sapply(covariates,
                        function(x) as.formula(paste('Surv(time, DEATH_EVENT)~', x)))

univ_models <- lapply( univ_formulas, function(x){coxph(x, data = data)})

# Extract data
univ_results <- lapply(univ_models,
                       function(x){
                          x <- summary(x)
                          p.value<-signif(x$wald["pvalue"], digits=2)
                          wald.test<-signif(x$wald["test"], digits=2)
                          beta<-signif(x$coef[1], digits=2);#coeficient beta
                          HR <-signif(x$coef[2], digits=2);#exp(beta)
                          HR.confint.lower <- signif(x$conf.int[,"lower .95"], 2)
                          HR.confint.upper <- signif(x$conf.int[,"upper .95"],2)
                          HR <- paste0(HR, " (",
                                       HR.confint.lower, "-", HR.confint.upper, ")")
                          res<-c(beta, HR, wald.test, p.value)
                          names(res)<-c("beta", "HR (95% CI for HR)", "wald.test",
                                        "p.value")
                          return(res)
                          #return(exp(cbind(coef(x),confint(x))))
                         })
res <- t(as.data.frame(univ_results, check.names = FALSE))
as.data.frame(res)
```

	beta <chr>	**HR (95% CI for HR)** <chr>	**wald.test** <chr>	**p.value** <chr>
age	0.042	1 (1-1.1)	24	8.4e-07
anaemia	0.34	1.4 (0.94-2.1)	2.7	0.1
creatinine_phosphokinase	0.00011	1 (1-1)	1.3	0.26
diabetes	-0.042	0.96 (0.64-1.4)	0.04	0.84
ejection_fraction	-0.046	0.95 (0.93-0.98)	18	1.7e-05
high_blood_pressure	0.44	1.5 (1-2.3)	4.3	0.037
platelets	-7.8e-07	1 (1-1)	0.53	0.47
serum_creatinine	0.29	1.3 (1.2-1.5)	28	1.2e-07
serum_sodium	-0.068	0.93 (0.9-0.97)	12	0.00052
sex	0.014	1 (0.67-1.5)	0	0.95

1-10 of 11 rows

Previous **1** 2 Next

The output above shows the regression beta coefficients, the effect sizes (given as hazard ratios), 95% confidence interval, and statistical significance for each of covariate in relation to overall survival. Each factor is assessed through separate univariate Cox regressions.

The attributes ‘age’, ‘ejection_fraction’,‘high_blood_pressure’,‘serum_creatinine’, ‘serum_sodium’ have highly statistically coefficients. This result similar with what we have from the previous Kaplan Meier method.

Among all statistically significant covariates, ‘age’, ‘high_blood_pressure’,‘serum_creatinine’ have positive beta coefficients. Thus, we can interpret that older age, higher blood pressure, or higher level of creatinine in the blood is associated with poorer survival.

On the other hand, ‘ejection_fraction’ and ‘serum_sodium’ have negative beta coefficients. Thus, having higher percentage of blood leaving the heart at each contraction or having higher level of sodium in the blood is associated with better survival chance.

b. Multivatirate Cox regression analysis

Model 1 with 5 factors (‘age’, ‘ejection_fraction’,‘high_blood_pressure’,‘serum_creatinine’, ‘serum_sodium’)

Now, we want to describe how the factors jointly impact on survival. To answer to this question, we’ll perform a multivariate Cox regression analysis. We’ll skip any varibles that are not significant in the univariate analysis. We’ll include the 5 factors (‘age’, ‘ejection_fraction’,‘high_blood_pressure’,‘serum_creatinine’, ‘serum_sodium’) into the multivariate model.

```
res.cox_multi <- coxph(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + high_blood_pr
essure + serum_creatinine + serum_sodium, data =  data)
summary(res.cox_multi)
```

```
## Call:
## coxph(formula = Surv(time, DEATH_EVENT) ~ age + ejection_fraction +
##     high_blood_pressure + serum_creatinine + serum_sodium, data = data)
##
##   n= 299, number of events= 96
##
##                          coef exp(coef)  se(coef)      z Pr(>|z|)
## age                  0.044564  1.045572  0.009004  4.950 7.43e-07 ***
## ejection_fraction   -0.045759  0.955272  0.010238 -4.470 7.83e-06 ***
## high_blood_pressure  0.494319  1.639382  0.211894  2.333   0.0197 *
## serum_creatinine     0.316546  1.372380  0.070615  4.483 7.37e-06 ***
## serum_sodium        -0.037959  0.962753  0.023609 -1.608   0.1079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## age                    1.0456     0.9564    1.0273    1.0642
## ejection_fraction      0.9553     1.0468    0.9363    0.9746
## high_blood_pressure    1.6394     0.6100    1.0822    2.4834
## serum_creatinine       1.3724     0.7287    1.1950    1.5761
## serum_sodium           0.9628     1.0387    0.9192    1.0083
##
## Concordance= 0.729  (se = 0.028 )
## Likelihood ratio test= 73.75  on 5 df,   p=2e-14
## Wald test            = 80.07  on 5 df,   p=8e-16
## Score (logrank) test = 81.1  on 5 df,   p=5e-16
```

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. In the above example, the test statistics are in close agreement, and the omnibus null hypothesis is soundly rejected.

In the multivariate Cox analysis, the covariates ‘age’, ‘ejection_fraction’,‘high_blood_pressure’ and ‘serum_creatinine’ remain significant (p-value < 0.05). However, the covariate ‘serum_sodium’ fails to be significant (p-value = 0.1079, which is grater than 0.05). Also the 95% confidence interval for HR is from 0.91 to 1.008, which includes 1, indicating ‘serum_sodium’ makes smaller contribution to the difference in HR.

Interpreting beta in terms of hazard ratio

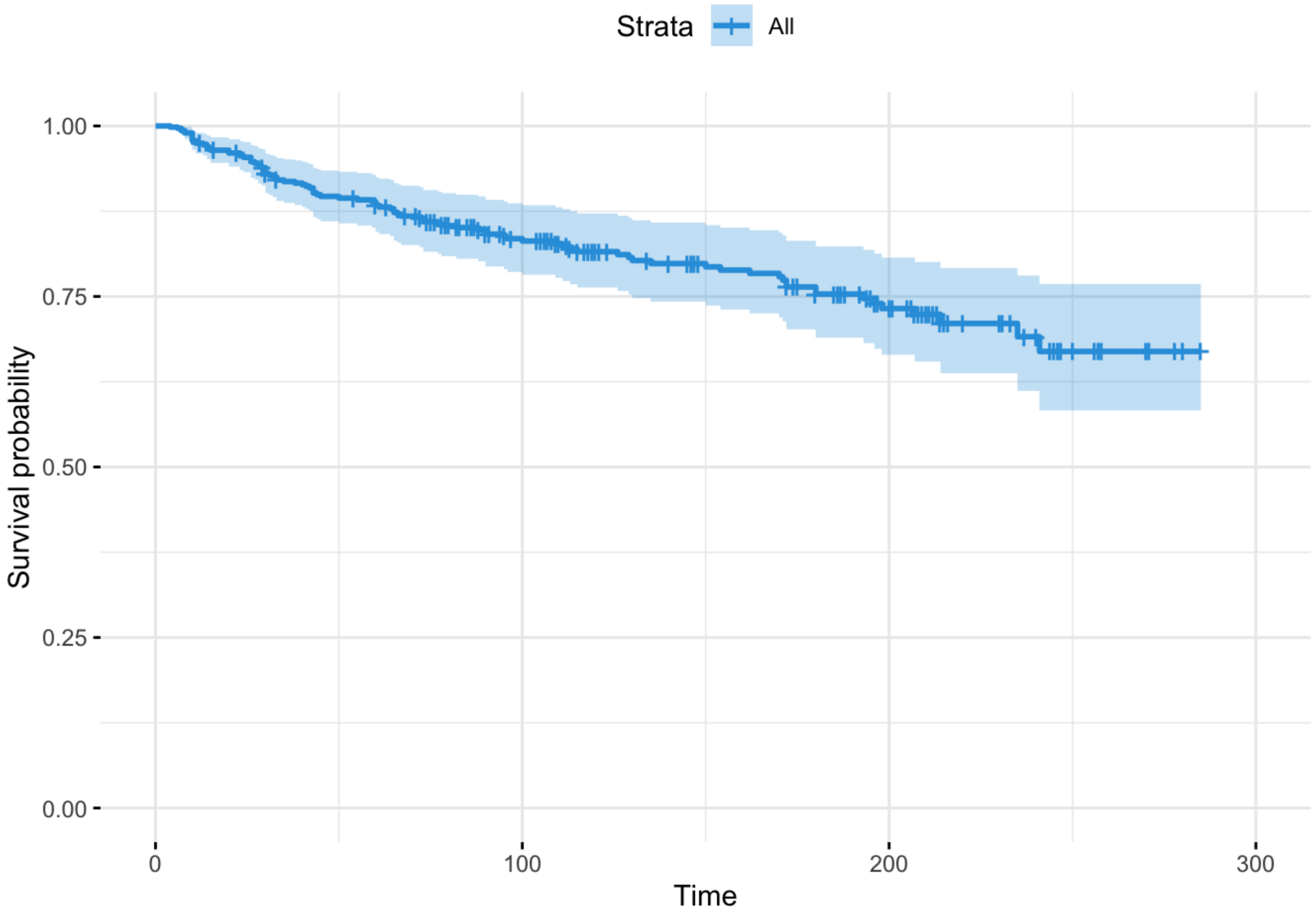
The p-value for ‘age’ is 7.43e-07, with a hazard ratio HR = exp(coef) = 1.0456, indicating a strong relationship between the patients’ age and increased risk of death. The hazard ratios of covariates are interpretable as multiplicative effects on the hazard. Thus, holding other variables constant, being older increases the hazard by a factor of 1.0456. We conclude that, having older age is associated with bad prognostic.

Similar to ‘age , hazard ratios HR = exp(coef) for ’high_blood_pressure’ and ‘serum_creatinine’ are also greater than 1, indicating the increase in hazard, meaning higher blood pressure or higher level of creatinine in the blood associate with higher risk of death. These two attributes are associated with bad prognostic.

On the other hand, the p-value for ‘ejection_fraction’ is 7.83e-06, with a hazard ratio HR = 0.9553, indicating a strong relationship between the ‘ejection_fraction’ value and decreased risk of death. Holding the other covariates constant, a higher value of ‘ejection_fraction’ is associated with a better survival. This is associated with good prognostic.

Plot the baseline harzard function

```
ggsurvplot(survfit(res.cox_multi), data = data, palette = "#2E9FDF",
           ggtheme = theme_minimal())
```



Model 2 with 4 factors (‘age’, ‘ejection_fraction’,‘high_blood_pressure’,‘serum_creatinine’)

```
#model 2 with only statistically significant attributes: age, ejection_fraction, high_bl
ood_pressure, serum_creatinine
res.cox_multi2 <- coxph(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + high_blood_p
ressure + serum_creatinine, data =  data)
summary(res.cox_multi2)
```

```
## Call:
## coxph(formula = Surv(time, DEATH_EVENT) ~ age + ejection_fraction + 
##     high_blood_pressure + serum_creatinine, data = data)
## 
##   n= 299, number of events= 96 
## 
##                          coef exp(coef)  se(coef)      z Pr(>|z|)    
## age                  0.044173  1.045163  0.009027  4.894 9.91e-07 ***
## ejection_fraction   -0.049587  0.951623  0.009968 -4.975 6.54e-07 ***
## high_blood_pressure  0.471236  1.601973  0.211410  2.229   0.0258 *  
## serum_creatinine     0.347001  1.414818  0.066705  5.202 1.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##                     exp(coef) exp(-coef) lower .95 upper .95
## age                    1.0452     0.9568    1.0268    1.0638
## ejection_fraction      0.9516     1.0508    0.9332    0.9704
## high_blood_pressure    1.6020     0.6242    1.0585    2.4244
## serum_creatinine       1.4148     0.7068    1.2414    1.6124
## 
## Concordance= 0.729  (se = 0.029 )
## Likelihood ratio test= 71.31  on 4 df,   p=1e-14
## Wald test            = 78.79  on 4 df,   p=3e-16
## Score (logrank) test = 78.04  on 4 df,   p=5e-16
```

From the first cox regression model, we removed the insignificant variable - ‘serum_sodium’, then fit remaining four variables into the second model. In the second multivariate Cox analysis, the covariates ‘age’, ‘ejection_fraction’,‘high_blood_pressure’ and ‘serum_creatinine’ remain significant as p-value < 0.05. The p-value for all three overall tests (likelihood, Wald, and score) are significant, indicating that the model is significant.

The final Cox Proportional Hazard model with these four variables can be written as:

$$h(t) = h_0(t)exp(\beta_1\ Age + \beta_2\ ejectrion\_fraction + \beta_3\ blood\_pressure + \beta_4\ serum\_creatinine$$

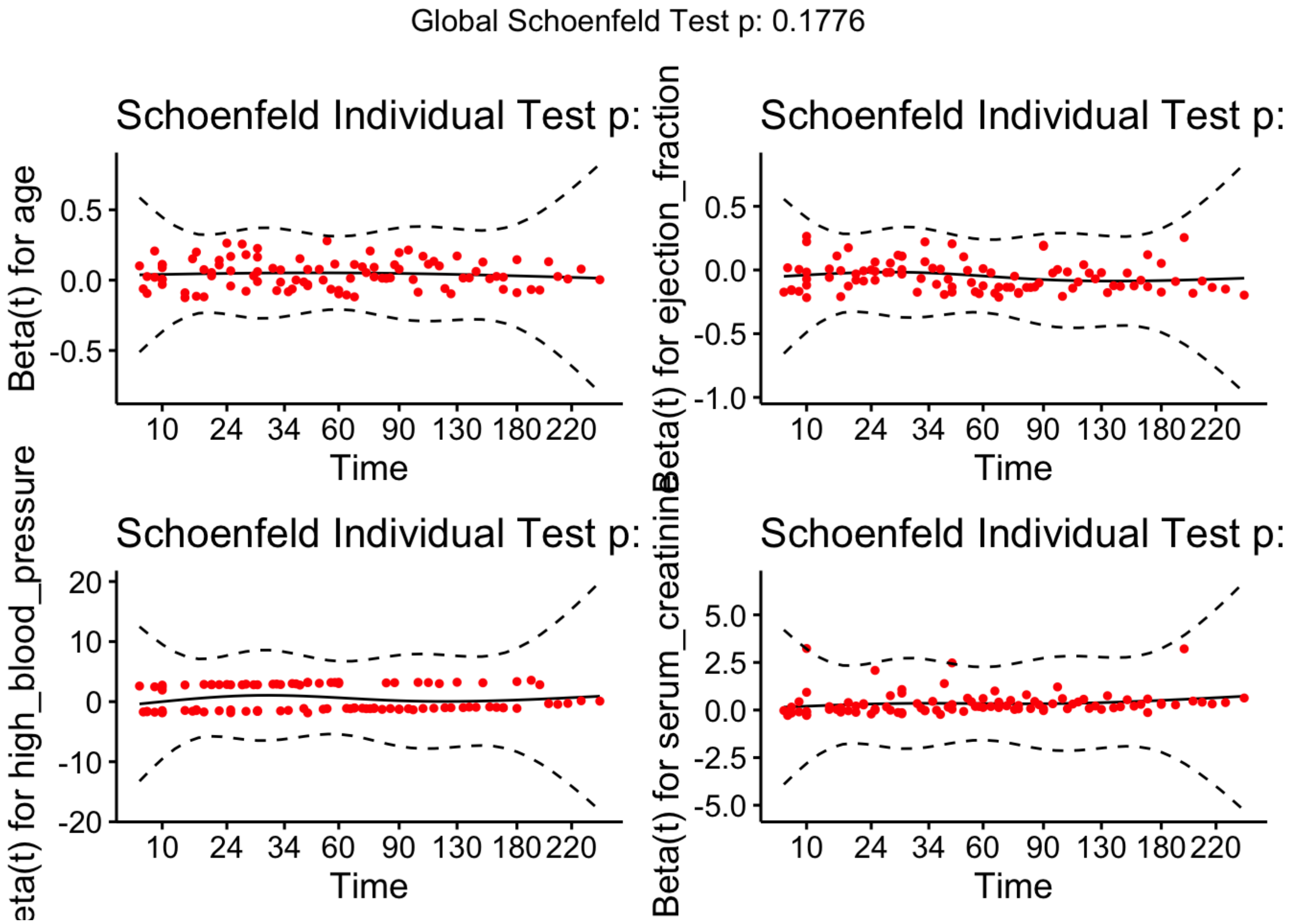
- The hazard ratio for age: HR = exp(coef) = 1.0452, indicating a strong relationship between the patients’ age and increased risk of death. Holding other variables constant, an additional year of ‘age’ increases the risk of death by a factor of $e^{\beta_1} = 1.0452$ - that is, by 4.52%.
- Similar to ‘age , hazard ratios HR = exp(coef) = 1.602 for ’high_blood_pressure’, which is greater than 1, indicating the increase in hazard. Each increase in higher blood pressure increases the risk of death by a factor of $e^{\beta_3} = 1.602$ - that is, by 60.2%.
- Hazard ratios HR = exp(coef) = 1.4148 for ‘serum_creatinine’ is greater than 1, indicating the increase in hazard. The increase in serum_creatinin by one unit increases the risk of death by a factor of $e^{\beta_4} = 1.4148$ - that is, by 41.48%.
- Lastly, the hazard ratios HR = exp(coef) = 0.9516 for ‘ejection_fraction’ is less than 1, indicating the decrease in hazard. Thus, the increase in ejection_fraction by one unit actually lower the risk of death by a factor of $e^{\beta_2} = 0.9516$ - that is, by 4.84%.
- ‘age’,‘high_blood_pressure’, and ‘serum_creatinine’ are associated with bad prognostic, while ‘ejection_fraction’ is associated with good prognostic.

Testing Proportional Hazards Assumption with Model 2

```
#model 2
test.ph = cox.zph(res.cox_multi2)
test.ph
```

```
##                     chisq df     p
## age                 0.271  1 0.603
## ejection_fraction   4.394  1 0.036
## high_blood_pressure 0.034  1 0.854
## serum_creatinine    1.289  1 0.256
## GLOBAL              6.304  4 0.178
```

```
ggcoxzph(test.ph)
```



From the output above, the test is not statistically significant for each of the covariates (except ejection_fraction), and the global test is also not statistically significant. Therefore, our proportional hazards assumption is reasonable. The graphs also shows no pattern with time. Thus, the assumption of proportional hazards appears to be supported for the covariates ‘age’,‘high_blood_pressure’,‘ejection_fraction’, and ‘serum_creatinine’.

Conclusion

With the Kaplan-Meier method, we conclude that five variables - ‘high_blodd_pressure’, ‘ejection_fraction’, ‘serum_creatinine’, ‘serum_sodium’, and age’ - each contributes significantly to the probability of death event in the study. On the other hand, under Cox Proportional Hazard Model, we fit a model with four significant varibles, including ‘high_blodd_pressure’, ‘ejection_fraction’, ‘serum_creatinine’, and ’age. From this study, we learned how different these two survival analysis methods work under the univariate and multivariate settings.

For the future, we would love to add and examine the significance of the interaction term or the power term in our Cox Proportional Hazard Model, and compare to find the best model for future prediction.