Survival Analysis Final Project

The Analysis of Heart Failure Clinical Records Dataset

Yunjie Xu 250992343
Zhiyi Zhang 250902651

Contents

1. The Background of the Study

1.1 Objective of the Study

For this study, the first objective is analysis the survival rate of people who have heart diseases and make the features' statistics for each feature. The second objective is to find how the features will be effect on people's death rate. The third objective is to find a reasonable model to fit the dataset.

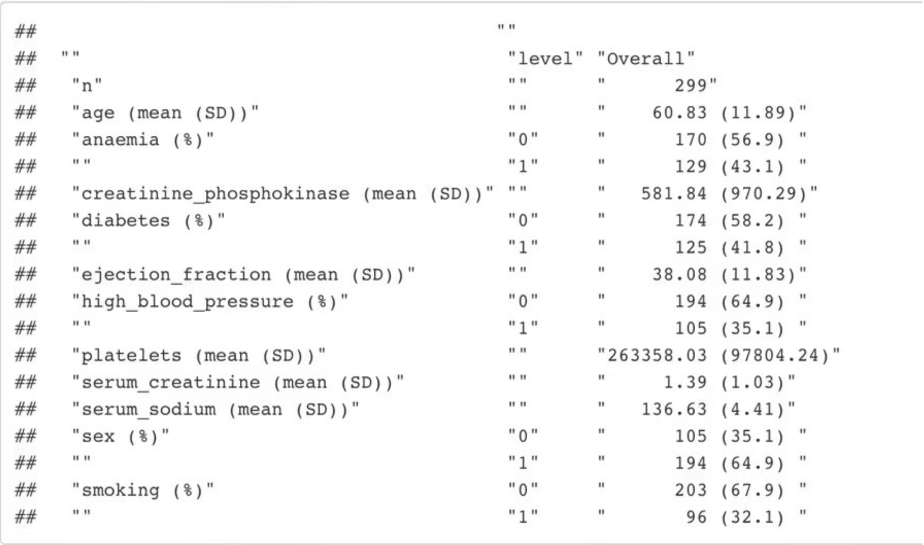
1.2 Background

Cardiovascular diseases (CVDs) are the most serious deceases which cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide.

Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure. Most cardiovascular diseases can be prevented by addressing risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, or already established disease) need early detection and management. From this study, we will figure out how these factors influence the survival rate. And find the model can explain this dataset.

2. The Data We Have

2.1 Baseline data analysis

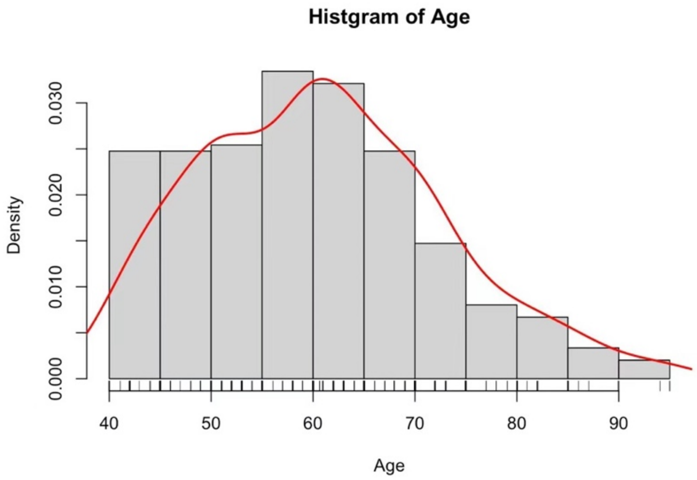


```
##                                       ""
##  ""                                    "level" "Overall"
##   "n"                                  ""      "      299"
##   "age (mean (SD))"                    ""      "    60.83 (11.89)"
##   "anaemia (%)"                        "0"     "      170 (56.9) "
##   ""                                   "1"     "      129 (43.1) "
##   "creatinine_phosphokinase (mean (SD))" ""      "   581.84 (970.29)"
##   "diabetes (%)"                       "0"     "      174 (58.2) "
##   ""                                   "1"     "      125 (41.8) "
##   "ejection_fraction (mean (SD))"      ""      "    38.08 (11.83)"
##   "high_blood_pressure (%)"            "0"     "      194 (64.9) "
##   ""                                   "1"     "      105 (35.1) "
##   "platelets (mean (SD))"              ""      "263358.03 (97804.24)"
##   "serum_creatinine (mean (SD))"       ""      "     1.39 (1.03)"
##   "serum_sodium (mean (SD))"           ""      "   136.63 (4.41)"
##   "sex (%)"                            "0"     "      105 (35.1) "
##   ""                                   "1"     "      194 (64.9) "
##   "smoking (%)"                        "0"     "      203 (67.9) "
##   ""                                   "1"     "       96 (32.1) "
```

From the table, we can see the number of observations are 299. Here are 11 features in the model. And there are 5 binary variables (anaemia, diabetes, high blood pressure, sex, smoking). The left 6 variables are continuous variables (age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium). From the table we can see the mean value and standard deviation of the continuous variables. For binary variable, we can see how many people have this habit and the proportion of patients has this bad habit. For example, there are 203 patients are smoker which make up 67.9% of people.

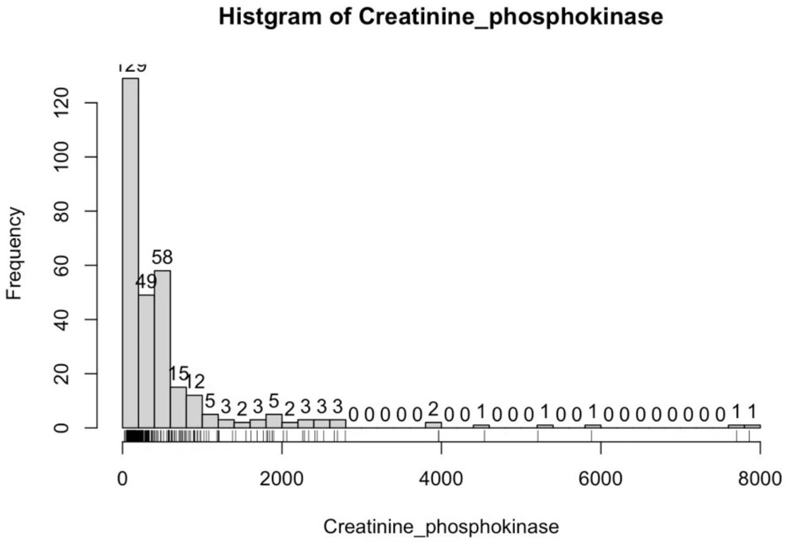
2.2 statistics for continuous variable

a) Age



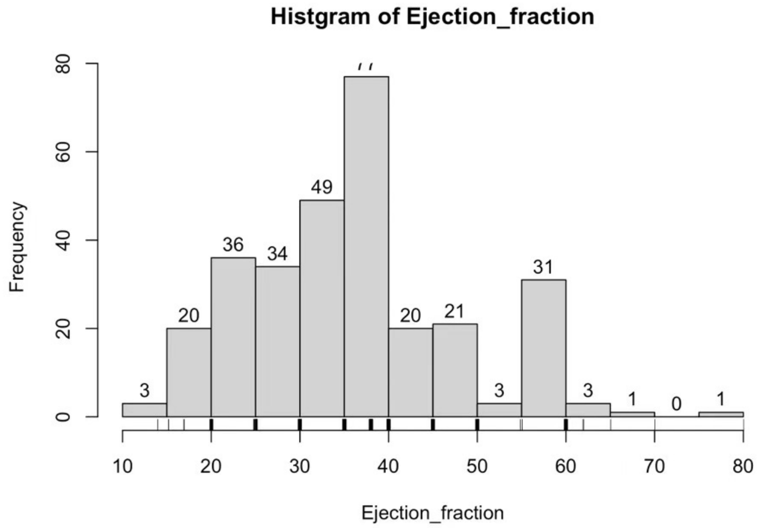
The age of patients in this dataset start from 40 and end to 95. More than 70% patients are from 40 to 70 years old. Patients in 55 to 65 years old is the most common age group in dataset. The second common age group are 40-55 and 60-65.

b) creatinine Phosphokinase (CPK)



Creatine phosphokinase (CPK) is an enzyme in the body. It is found mainly in the heart, brain, and skeletal muscle. The normal level of CPK is 10-120 mcg/L. From the plot, we can find that only 129(43.14%) people's CPK are in normal level. We guess the CPK may have effect on survival rate of patients.

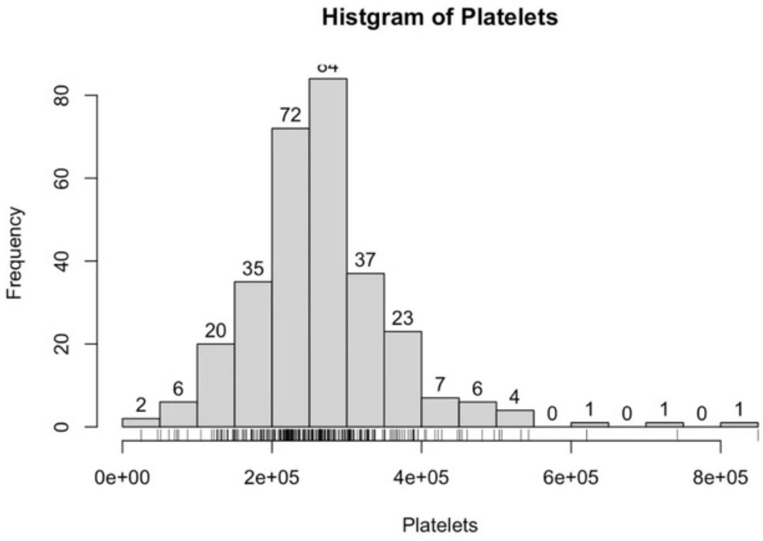
C) Ejection Fraction (EF)



Ejection fraction (EF) is a measurement, expressed as a percentage, of how much blood the left ventricle pumps out with each contraction. The normal level of EF is 50 to 70 percent. If people who have very low level of EF, that may cause serious diseases.

From the plot, we find only 39(12.71%) of patients' EF level are in normal level. And 250(83.61%) people's EF level below the normal level. We guess the EF factor may influence the patients survival rate.

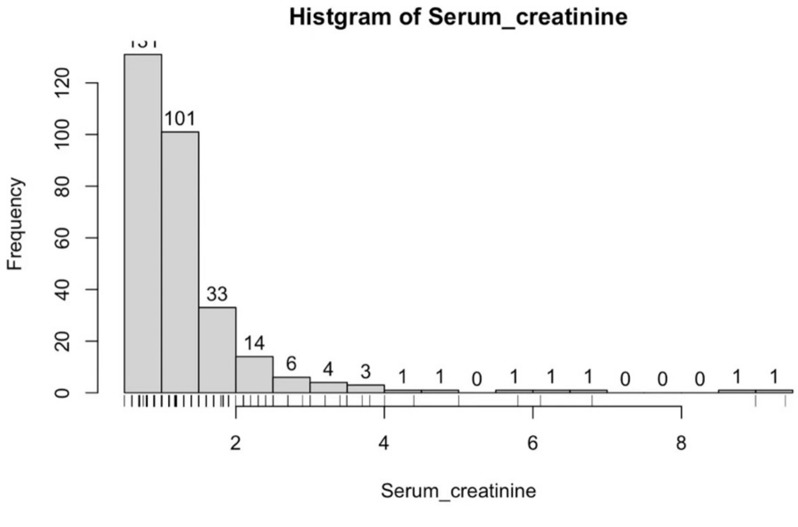
d) Platelets



Platelets are tiny blood cells that help your body form clots to stop bleeding. If one of your blood vessels gets damaged, it sends out signals to the platelets. The platelets then rush to the site of damage. they form a plug (clot) to fix the damage. And the normal platelets count is 150000 to 450000 plate per microliter of blood.

From the plot, we can see most of patients' platelets level are in normal level. So we guess platelets level is not a significant features in model.

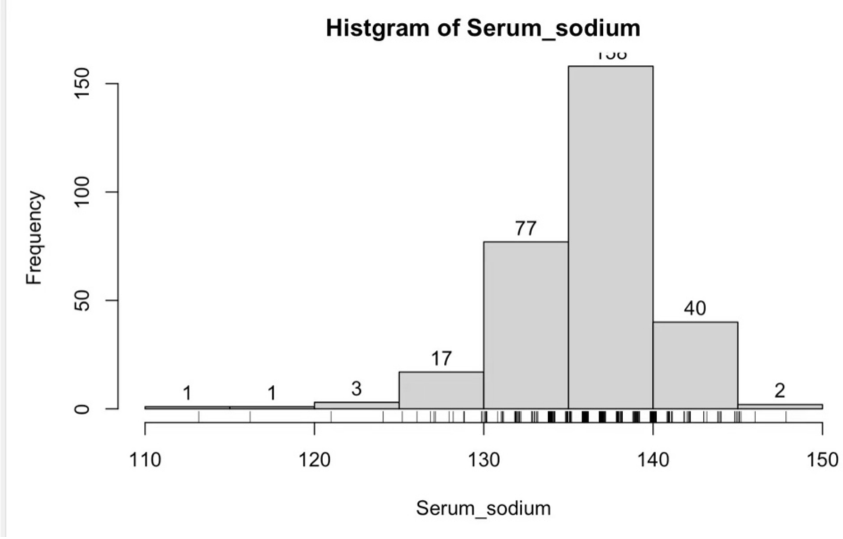
e) Serum Creatinine



In general, however, normal creatinine levels range from 0.6 to 1.3 mg/dL who are 18 to 60 years old. Normal levels are roughly the same for people over 60. High serum creatinine levels in the blood indicate that the kidneys aren't functioning properly.

In fact, more than half of patients (168,56.19%) in this dataset has higher serum creatinine. And we guess the serum creatinine may is the significant features.

f) serum Sodium

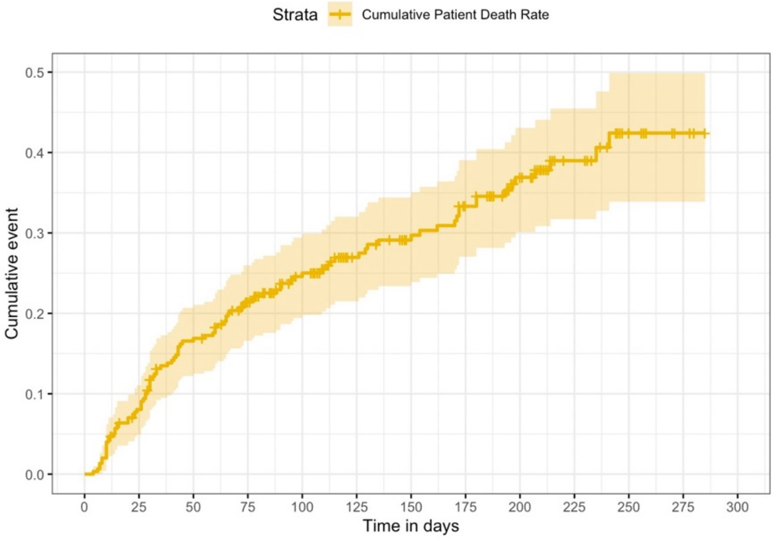


Measurement of serum sodium is routine in assessing electrolyte, acid-base, and water balance, as well as renal function. The reference range for serum sodium is 135-147 mmol/L. From the plot, we can see most of patients in this dataset has normal level of serum sodium. So we guess serum sodium will not effect on the survival rate of patients.

3. Non-Parametric Estimation

3.1 General survival analysis for the patients

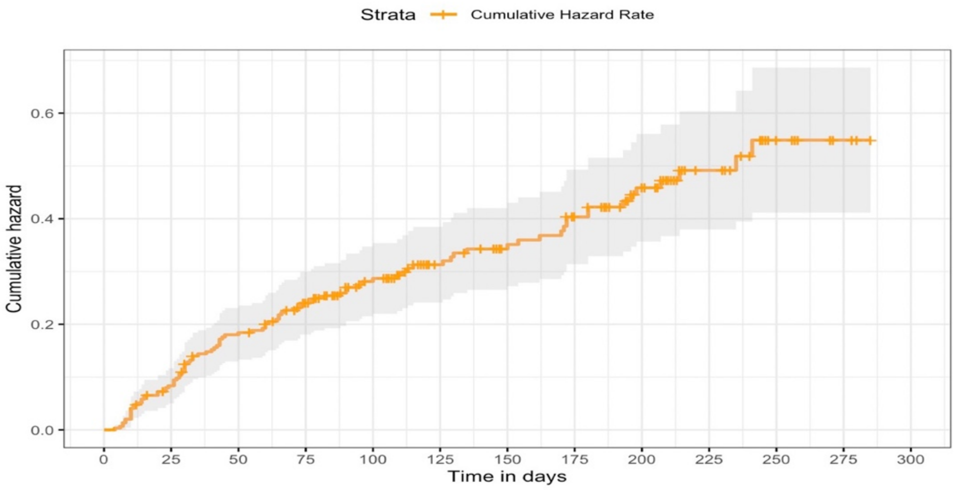
a) Cumulative Patient Death Rate

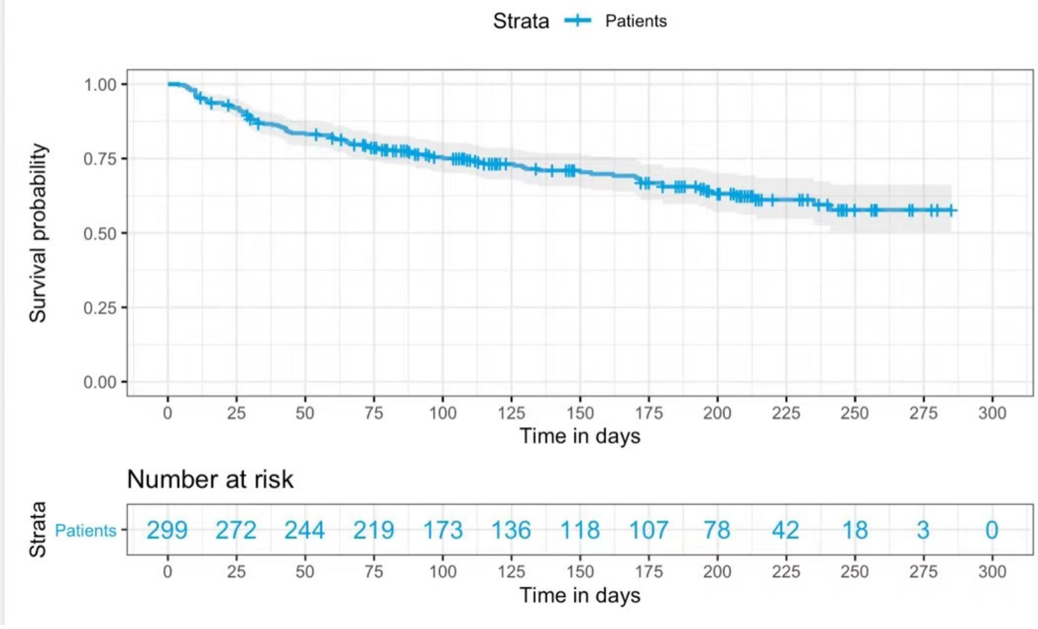


First, we plot the cumulative death rate plot by using the Kaplan-Meier estimation to know how many patients are died in this dataset. The result shows that the death rate more than 40% and only 57.5% people can survive after the experiments. That can show

CVDs is a very critical illness especially for people who have some bad habits.

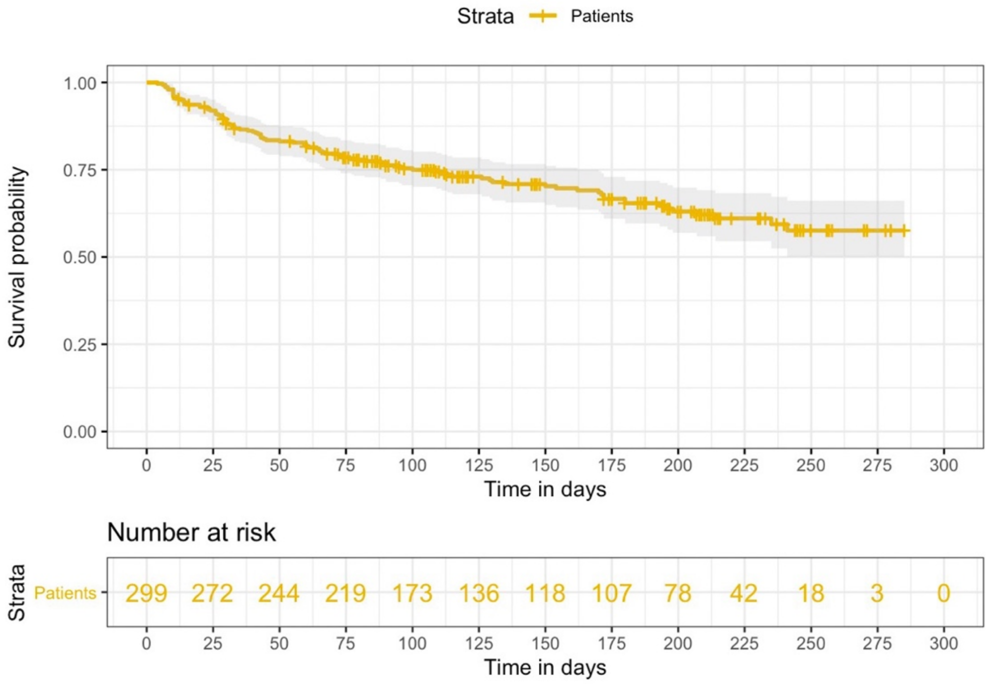
b) Nelson-Aalen Estimate



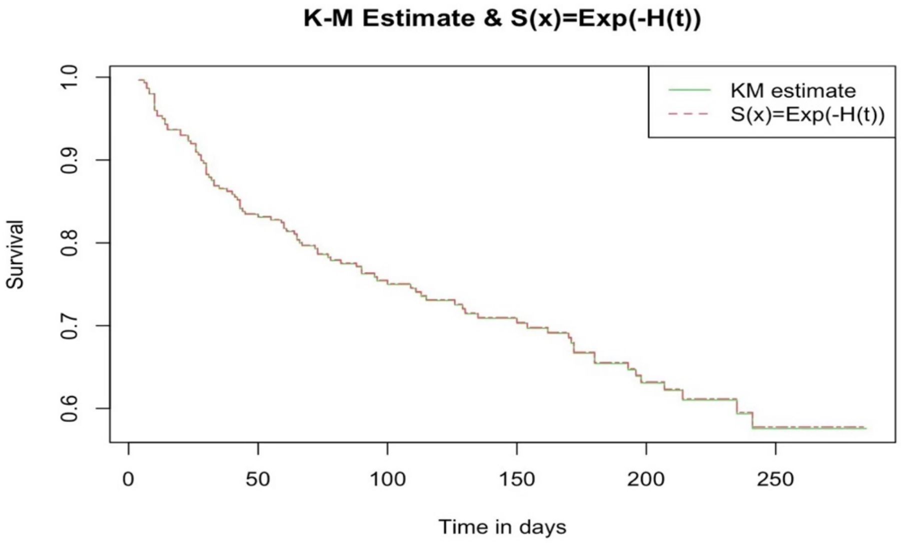


In this part, we use the Nelson-Aalen Estimate to construct the cumulative hazard function by using formula . And then we use the formula to plot the survival rate.

c) Kaplan-Meier Estimate & Survival Rate Comparison

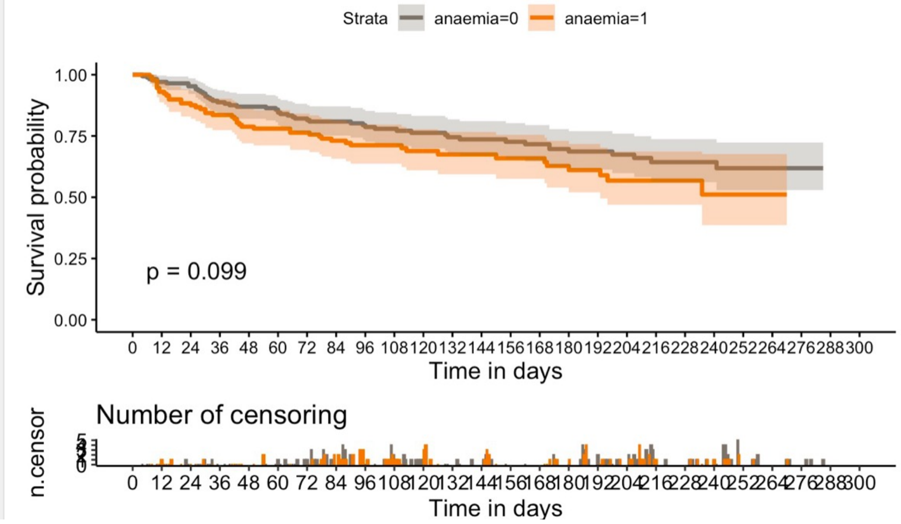


In this part, we use the Kaplan-Meier estimate to do the survival probability of patients. The formula we use =. The results is the same as using . The Survival rate after 285 days is 57.6%. Below the chart will how you the comparison between two survival plot by two these two different calculations.



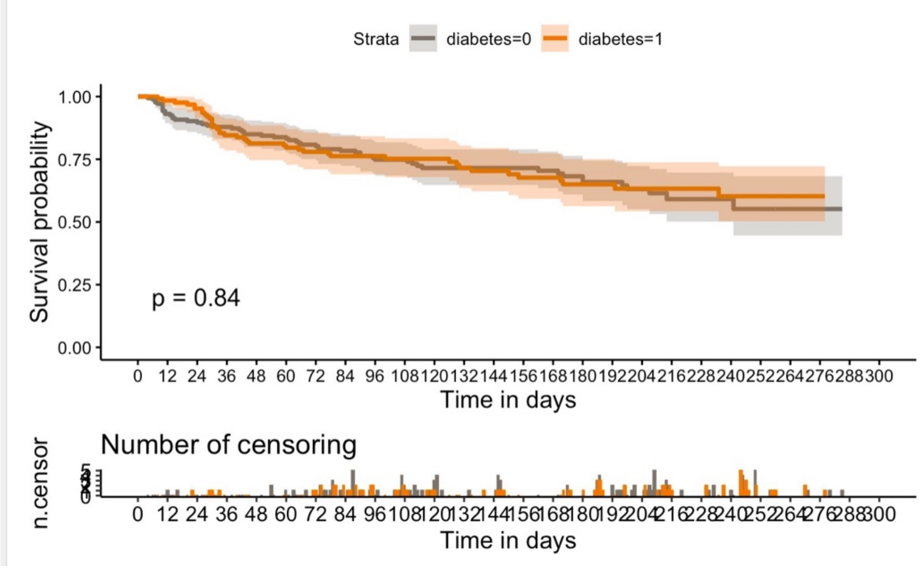
3.2 Kaplan-Meier estimate for each binary factor

a) Anaemia



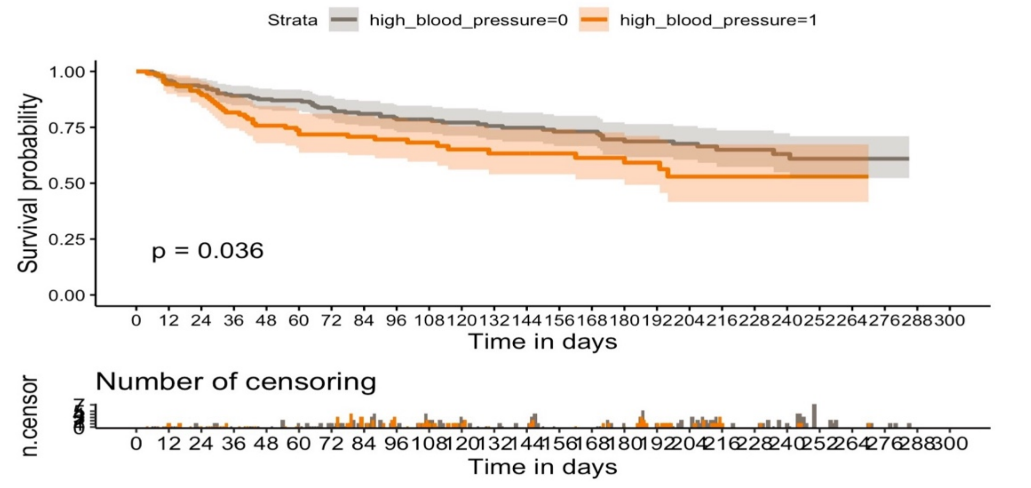
Anemia is a decrease in the total amount of red blood cells (RBCs) in the blood. This plot shows that people who suffers from anaemia will have higher survival rate than who do not have anemia conditions. Though anemia is a not good condition overall, however, in patients who have CVDs, anemia may help these patients live longer.

b) Diabetes



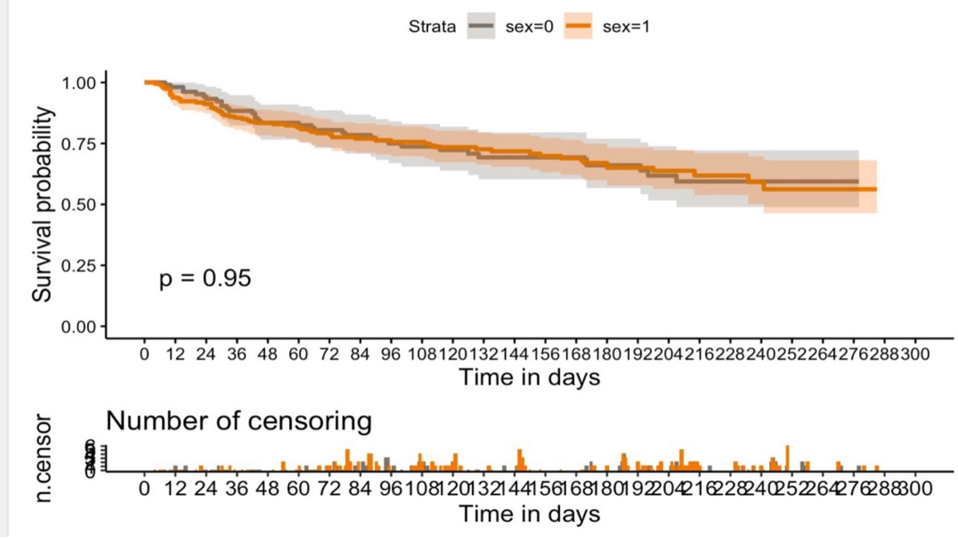
Diabetes is a metabolic disease that causes high blood sugar. From this plot, there is no significant difference of survival rate between patients who have diabetes or not.

c) High-Blood Pressure



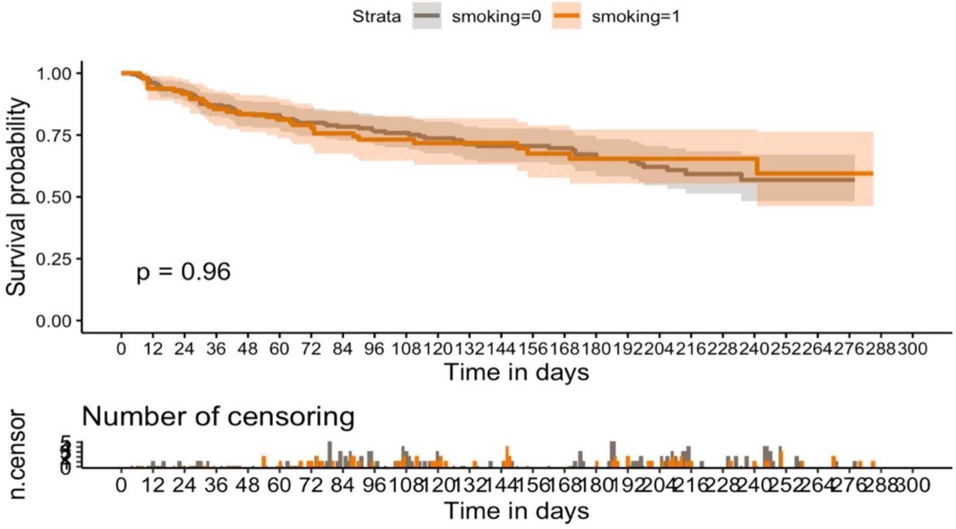
From the chart, patients who have high blood pressure have lower survival rate than people who do not have it. We guess the high blood pressure may pay a negative effect on survival rate.

d) sex



From the plot, the gender will not influence much on survival rate. The differences between female and male are very small.

e) smoking

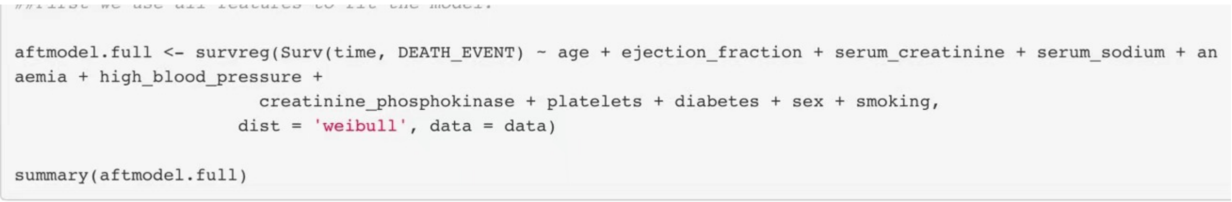


Although we all know smoking is a very bad habit, but for people who have CVDs, smoking seems will not be effect on patients' survival rate.

4. Parametric Analysis

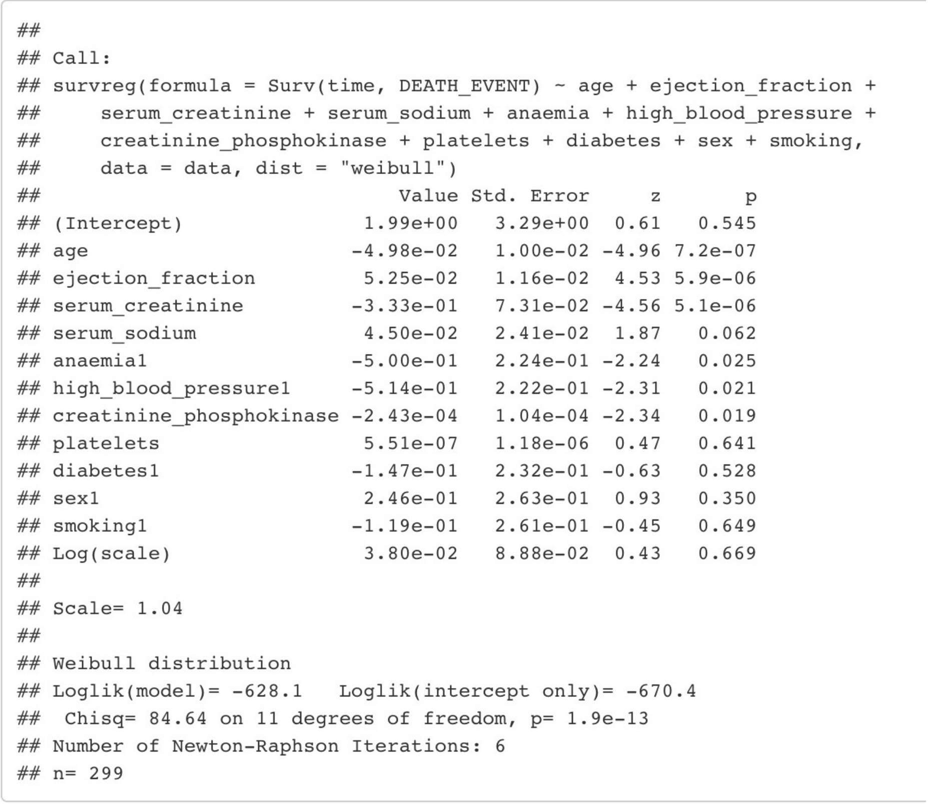
4.1 Accelerated failure time model (AFT model)

As we learned in class, the use of explanatory variables (covariates) in a regression model is an important way to represent heterogeneity in a population. By considering that age will be probably an effect on model, memoryless property is a very important property of exponential distribution. So we use Weibull distribution and all features to fit the full model.



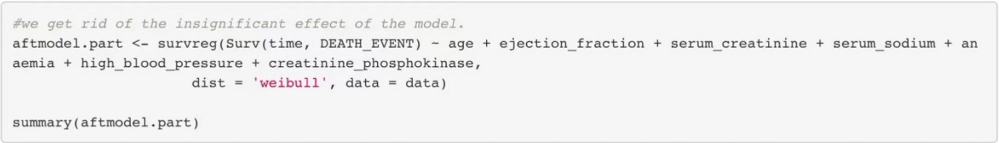
```
##First we use all features to fit the model.

aftmodel.full <- survreg(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine + serum_sodium + an
aemia + high_blood_pressure +
                     creatinine_phosphokinase + platelets + diabetes + sex + smoking,
                   dist = 'weibull', data = data)

summary(aftmodel.full)
```

Below will show the fit result:

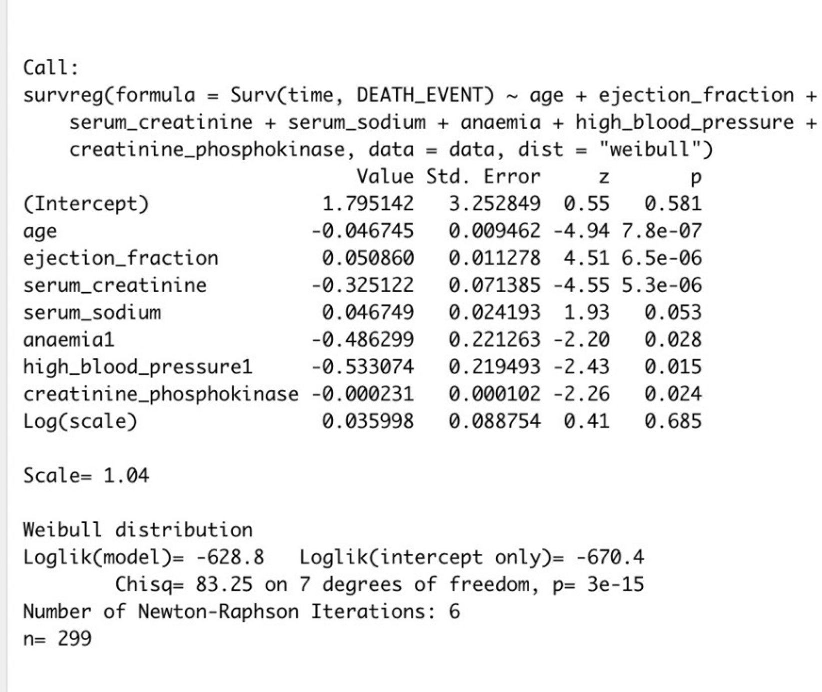


```
##
## Call:
## survreg(formula = Surv(time, DEATH_EVENT) ~ age + ejection_fraction +
##     serum_creatinine + serum_sodium + anaemia + high_blood_pressure +
##     creatinine_phosphokinase + platelets + diabetes + sex + smoking,
##     data = data, dist = "weibull")
##                              Value Std. Error     z       p
## (Intercept)               1.99e+00   3.29e+00  0.61   0.545
## age                      -4.98e-02   1.00e-02 -4.96 7.2e-07
## ejection_fraction         5.25e-02   1.16e-02  4.53 5.9e-06
## serum_creatinine         -3.33e-01   7.31e-02 -4.56 5.1e-06
## serum_sodium              4.50e-02   2.41e-02  1.87   0.062
## anaemia1                 -5.00e-01   2.24e-01 -2.24   0.025
## high_blood_pressure1     -5.14e-01   2.22e-01 -2.31   0.021
## creatinine_phosphokinase -2.43e-04   1.04e-04 -2.34   0.019
## platelets                 5.51e-07   1.18e-06  0.47   0.641
## diabetes1                -1.47e-01   2.32e-01 -0.63   0.528
## sex1                      2.46e-01   2.63e-01  0.93   0.350
## smoking1                 -1.19e-01   2.61e-01 -0.45   0.649
## Log(scale)                3.80e-02   8.88e-02  0.43   0.669
##
## Scale= 1.04
##
## Weibull distribution
## Loglik(model)= -628.1   Loglik(intercept only)= -670.4
##  Chisq= 84.64 on 11 degrees of freedom, p= 1.9e-13
## Number of Newton-Raphson Iterations: 6
## n= 299
```

we get rid of insignificant factor if their p-value > 0.1, then we use the left factors to fit the part model.



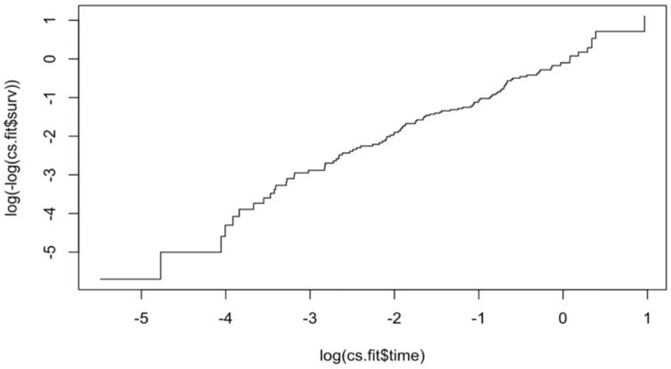
```
#we get rid of the insignificant effect of the model.
aftmodel.part <- survreg(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine + serum_sodium + an
aemia + high_blood_pressure + creatinine_phosphokinase,
                    dist = 'weibull', data = data)

summary(aftmodel.part)
```



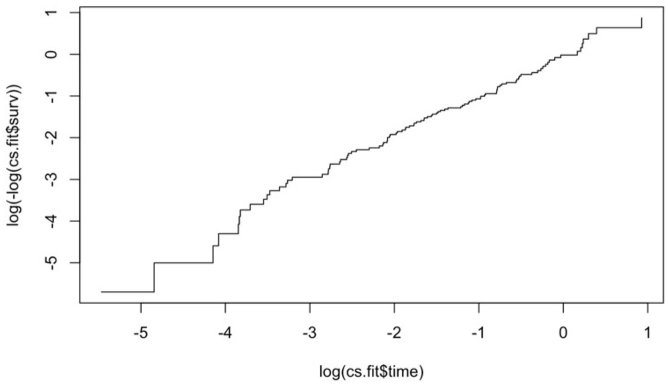
```
Call:
survreg(formula = Surv(time, DEATH_EVENT) ~ age + ejection_fraction +
    serum_creatinine + serum_sodium + anaemia + high_blood_pressure +
    creatinine_phosphokinase, data = data, dist = "weibull")
                            Value Std. Error     z       p
(Intercept)              1.795142   3.252849  0.55   0.581
age                     -0.046745   0.009462 -4.94 7.8e-07
ejection_fraction        0.050860   0.011278  4.51 6.5e-06
serum_creatinine        -0.325122   0.071385 -4.55 5.3e-06
serum_sodium             0.046749   0.024193  1.93   0.053
anaemia1                -0.486299   0.221263 -2.20   0.028
high_blood_pressure1    -0.533074   0.219493 -2.43   0.015
creatinine_phosphokinase -0.000231   0.000102 -2.26   0.024
Log(scale)               0.035998   0.088754  0.41   0.685

Scale= 1.04

Weibull distribution
Loglik(model)= -628.8   Loglik(intercept only)= -670.4
        Chisq= 83.25 on 7 degrees of freedom, p= 3e-15
Number of Newton-Raphson Iterations: 6
n= 299
```

We do the residual test to do the adequation check. If the model is reasonable model, then there will be a linear pattern between log(fit$time) and log(-log(fit$surv)).

Full model:



Part model:



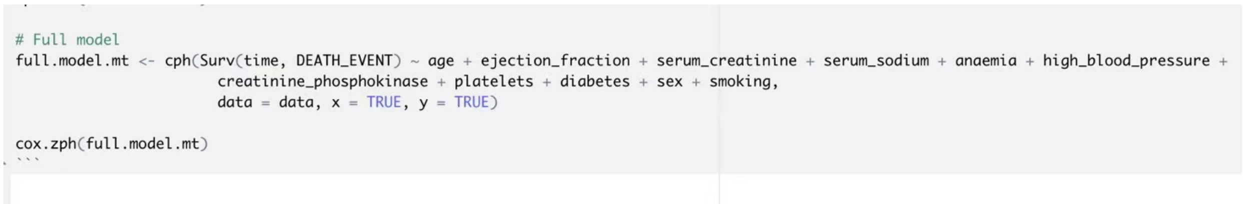
From the plot both show the linear pattern, thus we think both full model and part model are reasonable model to explain the dataset. We choose the part model as our final AFT model.

4.2 Cox PH model

The AFT model is based on time to model, and the PH model is based on hazard rate. And the Weibull regression is the only regression that satisfies both model assumptions. That means we can use the result from above AFT model to build Cox PH model. We use all features and part features selected by AFT model to fit the full cox PH model and the part cox PH model. And we do the PH assumption checking the model.

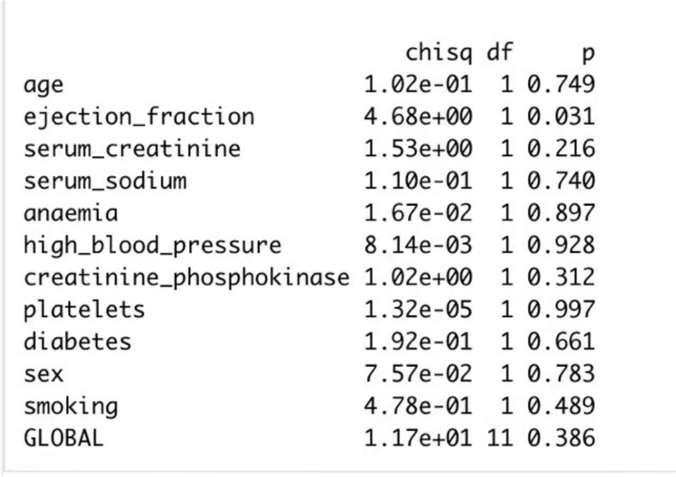
As we know, PH hypothesis can be checked by the residual plot. Under PH assumption, Schoenfeld residuals should not related with time. So if the residuals have pattern with time, that means this model violates the model assumption. However, we can use the “Cox.zph” to check the results easily by instead of residual plot.

a) Full model



```
# Full model
full.model.mt <- cph(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine + serum_sodium + anaemia + high_blood_pressure +
                         creatinine_phosphokinase + platelets + diabetes + sex + smoking,
                         data = data, x = TRUE, y = TRUE)

cox.zph(full.model.mt)
```

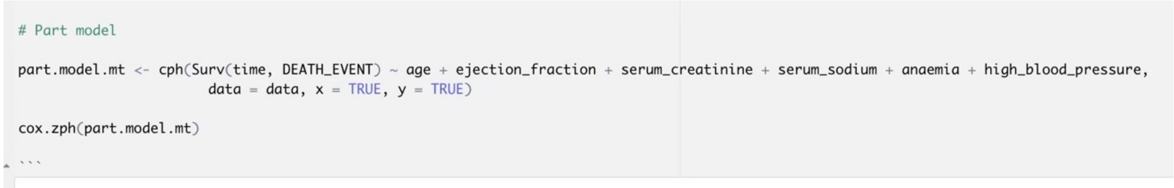
We do the proportional hazard assumption checking the model adequation. Below is the result:



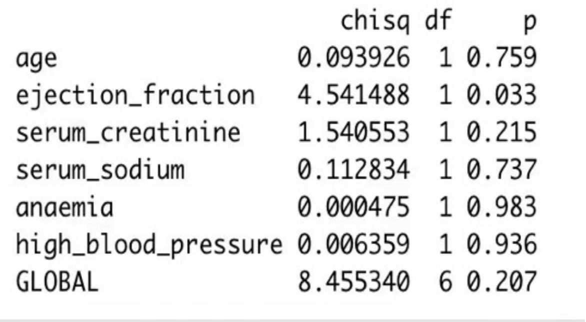
```
                            chisq df     p
age                      1.02e-01  1 0.749
ejection_fraction        4.68e+00  1 0.031
serum_creatinine         1.53e+00  1 0.216
serum_sodium             1.10e-01  1 0.740
anaemia                  1.67e-02  1 0.897
high_blood_pressure      8.14e-03  1 0.928
creatinine_phosphokinase 1.02e+00  1 0.312
platelets                1.32e-05  1 0.997
diabetes                 1.92e-01  1 0.661
sex                      7.57e-02  1 0.783
smoking                  4.78e-01  1 0.489
GLOBAL                   1.17e+01 11 0.386
```

From the output, all covariates’ p-value are larger than 0.05. And the whole model’s P-value is 0.386 which is also larger than 0.05. Thus we think the Cox PH model is reasonable.

b) Part model



~~~
# Part model

part.model.mt <- cph(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine + serum_sodium + anaemia + high_blood_pressure,
                     data = data, x = TRUE, y = TRUE)

cox.zph(part.model.mt)

```
~~~

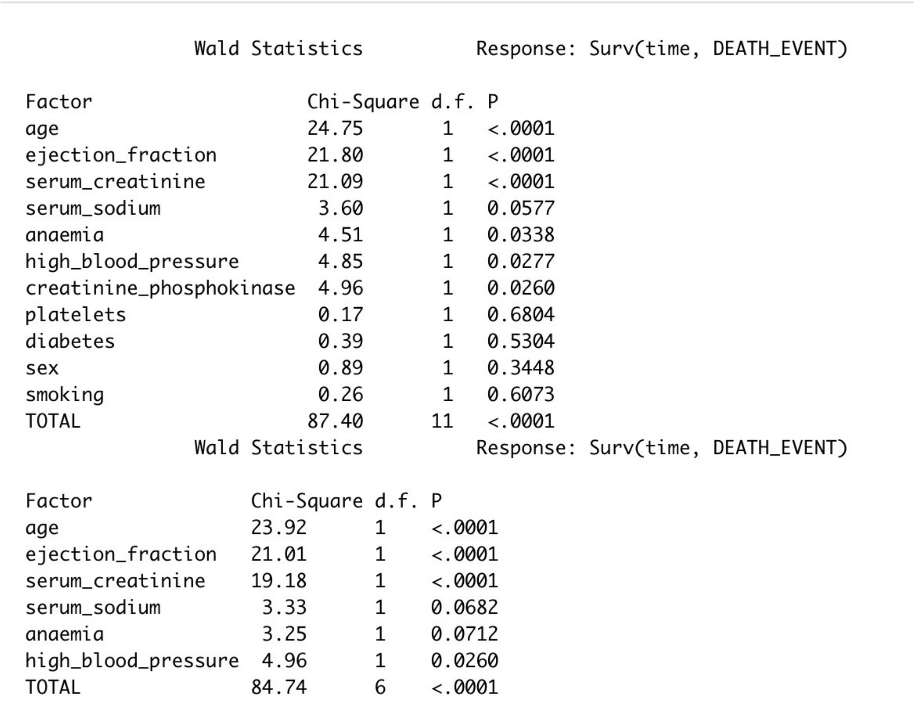


```
                       chisq df     p
age                 0.093926  1 0.759
ejection_fraction   4.541488  1 0.033
serum_creatinine    1.540553  1 0.215
serum_sodium        0.112834  1 0.737
anaemia             0.000475  1 0.983
high_blood_pressure 0.006359  1 0.936
GLOBAL              8.455340  6 0.207
```

similarly, we do the same test on part model. All covariates' P-value also greater than 0.05. And the whole model's P-value is 0.207. Thus part model is also acceptable.

c) Wald Test

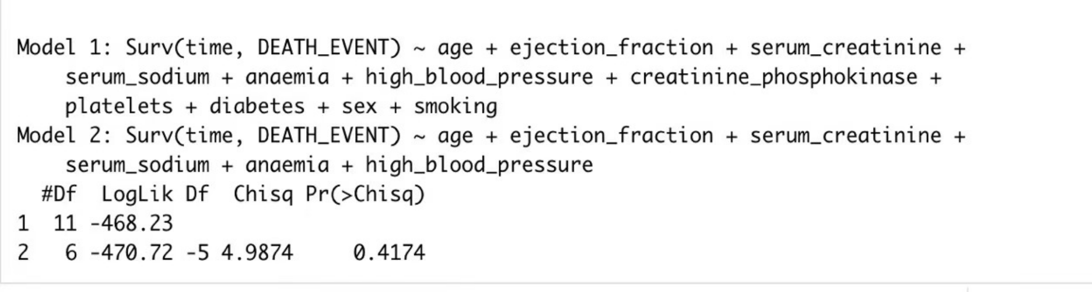
We also did the Wald Test to recheck the covariates' significance for both part model and full model. The results show all covariates in both models are significant.



```
               Wald Statistics          Response: Surv(time, DEATH_EVENT) 

 Factor                   Chi-Square d.f. P     
 age                      24.75       1   <.0001
 ejection_fraction        21.80       1   <.0001
 serum_creatinine         21.09       1   <.0001
 serum_sodium              3.60       1   0.0577
 anaemia                   4.51       1   0.0338
 high_blood_pressure       4.85       1   0.0277
 creatinine_phosphokinase  4.96       1   0.0260
 platelets                 0.17       1   0.6804
 diabetes                  0.39       1   0.5304
 sex                       0.89       1   0.3448
 smoking                   0.26       1   0.6073
 TOTAL                    87.40      11   <.0001
               Wald Statistics          Response: Surv(time, DEATH_EVENT) 

 Factor              Chi-Square d.f. P     
 age                 23.92      1    <.0001
 ejection_fraction   21.01      1    <.0001
 serum_creatinine    19.18      1    <.0001
 serum_sodium         3.33      1    0.0682
 anaemia              3.25      1    0.0712
 high_blood_pressure  4.96      1    0.0260
 TOTAL               84.74      6    <.0001
```

d) Likelihood Ratio Test

Because two model both are acceptable. Then we do the likelihood ratio test to check the model difference between those two models by using “Lrtest” function. Here is the result:

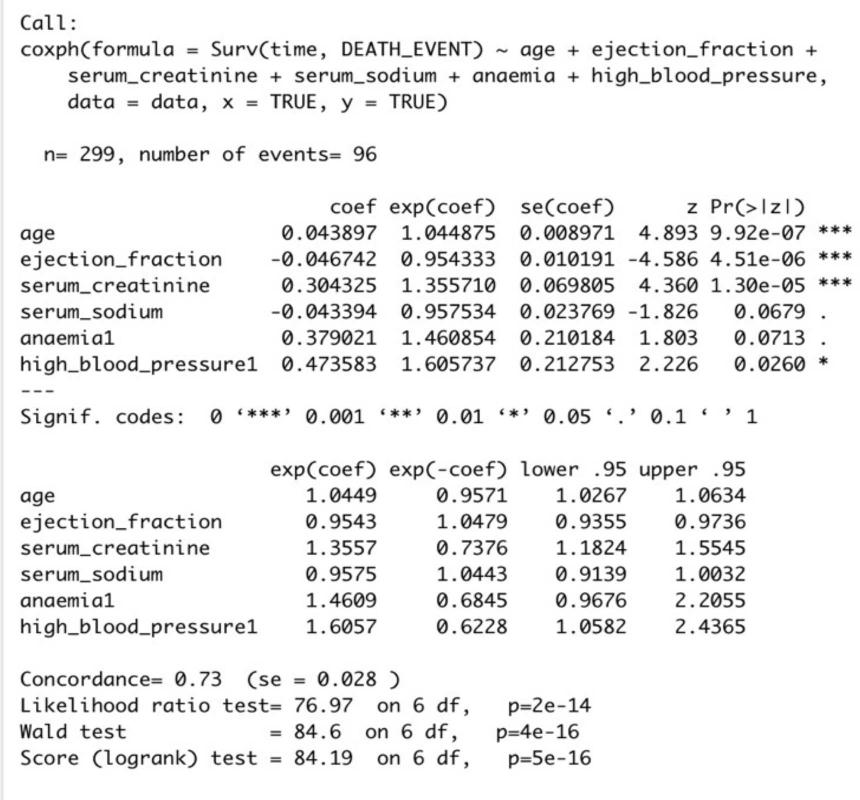


```
Model 1: Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine +
    serum_sodium + anaemia + high_blood_pressure + creatinine_phosphokinase +
    platelets + diabetes + sex + smoking
Model 2: Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine +
    serum_sodium + anaemia + high_blood_pressure
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  11 -468.23
2   6 -470.72 -5 4.9874     0.4174
```

From the output, the Pr(>Chisq) = 0.4174 that means there is no differences between two model. Thus we would like to select the part model as our final model.

e) Model Result

we find that the continuous variables age, ejection fraction and serum creatinine, the death risk of patients was 1.0449, 0.9543 and 1.3557 times higher for each unit increase. And for high blood pressure factors, patients with high blood pressure were 1.6057 times more likely to die than those without high blood pressure.



```
Call:
coxph(formula = Surv(time, DEATH_EVENT) ~ age + ejection_fraction +
    serum_creatinine + serum_sodium + anaemia + high_blood_pressure,
    data = data, x = TRUE, y = TRUE)

  n= 299, number of events= 96

                         coef exp(coef)  se(coef)      z Pr(>|z|)
age                  0.043897  1.044875  0.008971  4.893 9.92e-07 ***
ejection_fraction   -0.046742  0.954333  0.010191 -4.586 4.51e-06 ***
serum_creatinine     0.304325  1.355710  0.069805  4.360 1.30e-05 ***
serum_sodium        -0.043394  0.957534  0.023769 -1.826   0.0679 .
anaemia1             0.379021  1.460854  0.210184  1.803   0.0713 .
high_blood_pressure1  0.473583  1.605737  0.212753  2.226   0.0260 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                     exp(coef) exp(-coef) lower .95 upper .95
age                     1.0449     0.9571    1.0267    1.0634
ejection_fraction       0.9543     1.0479    0.9355    0.9736
serum_creatinine        1.3557     0.7376    1.1824    1.5545
serum_sodium            0.9575     1.0443    0.9139    1.0032
anaemia1                1.4609     0.6845    0.9676    2.2055
high_blood_pressure1    1.6057     0.6228    1.0582    2.4365

Concordance= 0.73  (se = 0.028 )
Likelihood ratio test= 76.97  on 6 df,   p=2e-14
Wald test            = 84.6  on 6 df,   p=4e-16
Score (logrank) test = 84.19  on 6 df,   p=5e-16
```

5. Others

5.1 Limitation of the study

a) Because the limitation of our ability, we did not consider the interaction effects in the model. Thus our model may not be the best model to fit the dataset.

b) some factors will naturally accelerate the death process and that may not be related to cardiovascular diseases (CVDs). Like the mortality will naturally increase when people grow old.

5.2 conclusion

a) The cardiovascular diseases (CVDs) will cuases very high death probability. About 40% of the patients are dead during.

b) Age, Ejection fraction, Serum creatine, Anaemia, High blood pressure, Creatinine phosphokinase are significant effect on the survival rate of patients.

c) We choose the part AFT model and part PH model as our final model of this study. The model will be included: age, ejection fraction, serum creatine and high blood pressure. From the output below, we find that the continuous variables age, ejection fraction and serum creatinine, the death risk of patients was 1.0449, 0.9543 and 1.3557 times higher for each unit increase and for high blood pressure factors, patients with high blood pressure were 1.6057 times more likely to die than those without high blood pressure.

5.3 Team member role

Yunjie Xu: 2.1/2.2/3.1/3.2/4.2
Zhiyizhang: 1.1/1.2/4.1/5.1/5.2

Final Project of AS4823

```r
rm(list = ls())
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.2
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```r
library("rms")
```

```
## Warning: package 'rms' was built under R version 3.6.3
```

```
## Loading required package: Hmisc
```

```
## Warning: package 'Hmisc' was built under R version 3.6.3
```

```
## Loading required package: survival
```

```
## Warning: package 'survival' was built under R version 3.6.3
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:caret':
##
##     cluster
```

```
## Loading required package: Formula
```

```
## Warning: package 'Formula' was built under R version 3.6.3
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## Loading required package: SparseM
```

```
## Warning: package 'SparseM' was built under R version 3.6.2
```

```
##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve
```

```r
library(tableone)
library(ggplot2)
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 3.6.3

## Loading required package: ggpubr

## Warning: package 'ggpubr' was built under R version 3.6.3
```

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'lmtest'

## The following object is masked from 'package:rms':
##
##     lrtest
```

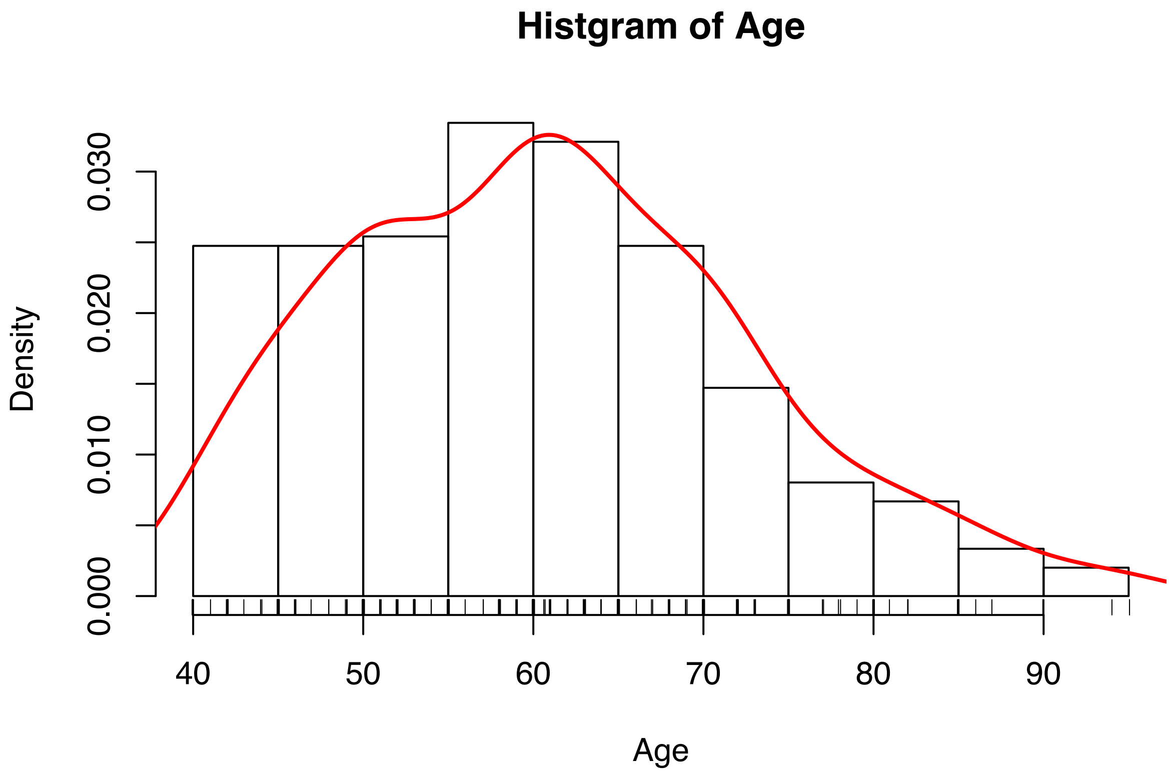
```r
#Read data set
#Data_set:Heart_Failure_Clinical_Records_Dataset
data <- read.csv(file="C:/Users/jaosn/Desktop/heart_failure_clinical_records_dataset.csv")
data$anaemia <- factor(data$anaemia)
data$diabetes <- factor(data$diabetes)
data$high_blood_pressure <- factor(data$high_blood_pressure)
data$sex <- factor(data$sex)
data$smoking <- factor(data$smoking)

names(data)
```

```
##  [1] "age"                      "anaemia"
##  [3] "creatinine_phosphokinase" "diabetes"
##  [5] "ejection_fraction"        "high_blood_pressure"
##  [7] "platelets"                "serum_creatinine"
##  [9] "serum_sodium"             "sex"
## [11] "smoking"                  "time"
## [13] "DEATH_EVENT"
```

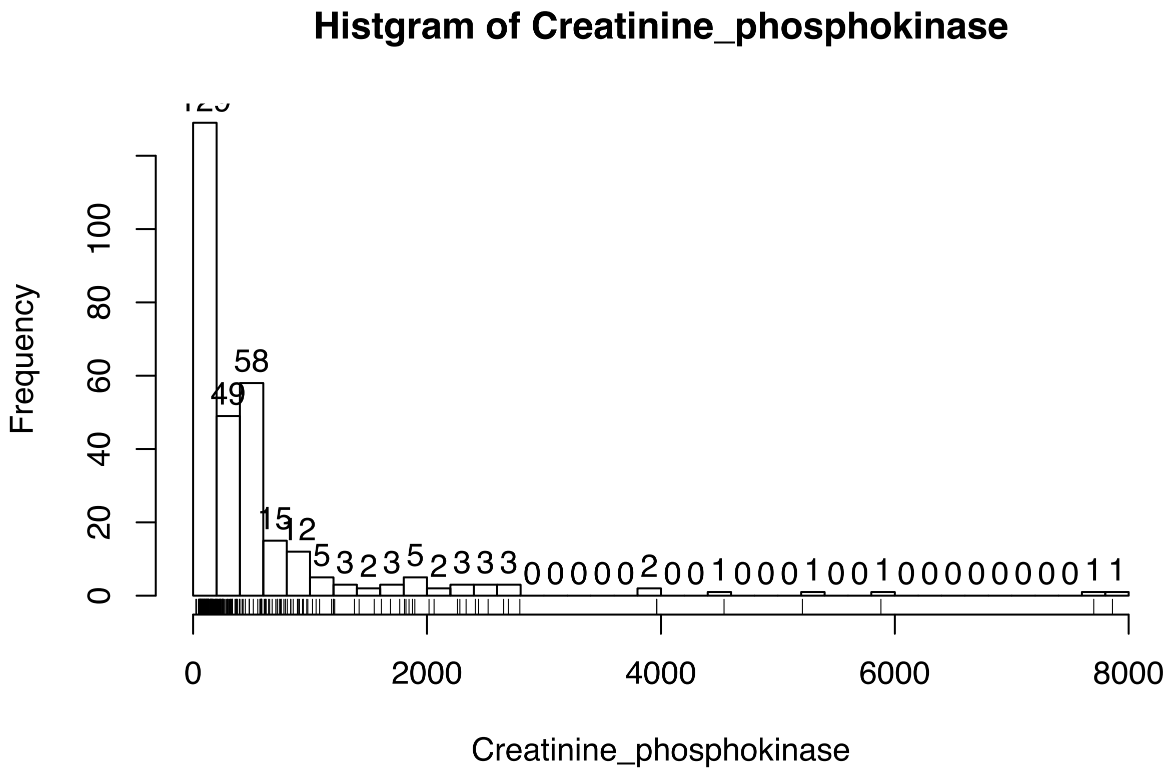
```r
#Patients feature statistics

cols <- c("age", "anaemia", "creatinine_phosphokinase", "diabetes",
          "ejection_fraction", "high_blood_pressure", "platelets", "serum_creatinine",
          "serum_sodium", "sex", "smoking")
print(CreateTableOne(var = cols,
                     factorVars = c("anaemia", "diabetes", "high_blood_pressure", "sex", "smoking"),
                     # strata = "residence",
                     data = data),
      showAllLevels = TRUE,
      quote = TRUE)
```

```
##                                          ""
##  ""                                      "level" "Overall"
##   "n"                                    ""      "      299"
##   "age (mean (SD))"                      ""      "    60.83 (11.89)"
##   "anaemia (%)"                          "0"     "      170 (56.9) "
##   ""                                     "1"     "      129 (43.1) "
##   "creatinine_phosphokinase (mean (SD))" ""      "   581.84 (970.29)"
##   "diabetes (%)"                         "0"     "      174 (58.2) "
##   ""                                     "1"     "      125 (41.8) "
##   "ejection_fraction (mean (SD))"        ""      "    38.08 (11.83)"
##   "high_blood_pressure (%)"              "0"     "      194 (64.9) "
##   ""                                     "1"     "      105 (35.1) "
##   "platelets (mean (SD))"                ""      "263358.03 (97804.24)"
##   "serum_creatinine (mean (SD))"         ""      "     1.39 (1.03)"
##   "serum_sodium (mean (SD))"             ""      "   136.63 (4.41)"
##   "sex (%)"                              "0"     "      105 (35.1) "
##   ""                                     "1"     "      194 (64.9) "
##   "smoking (%)"                          "0"     "      203 (67.9) "
##   ""                                     "1"     "       96 (32.1) "
```

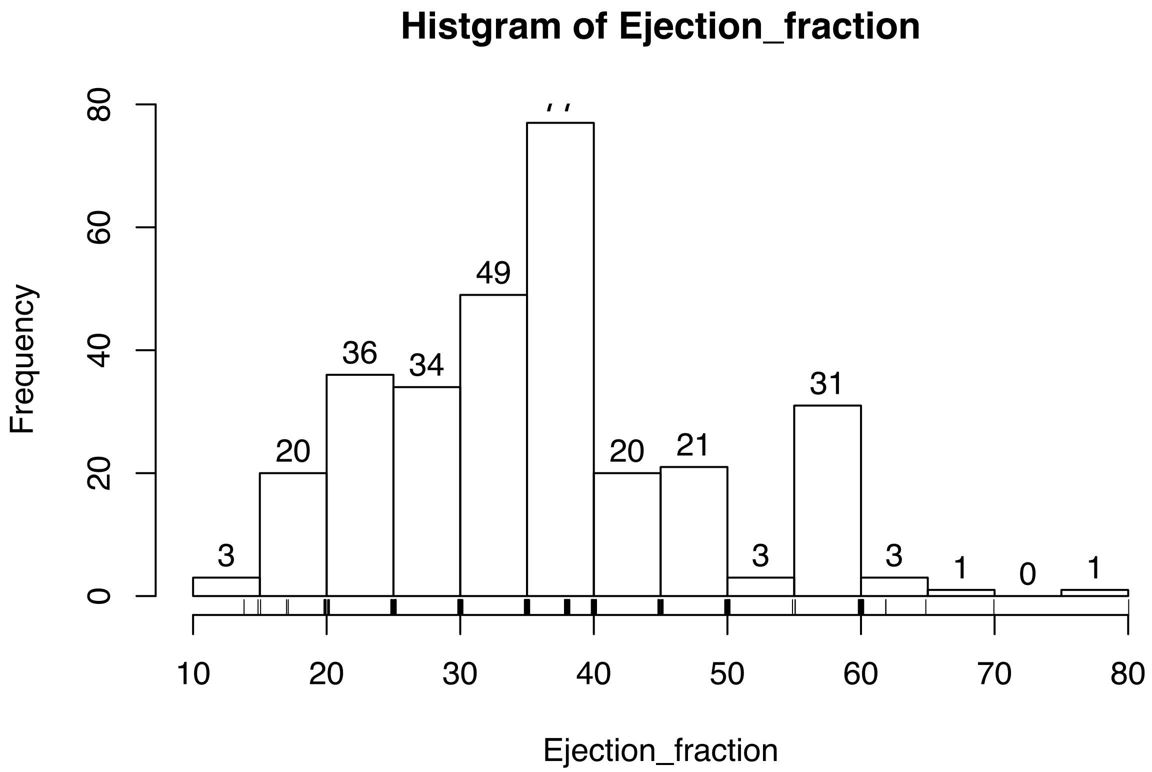
```r
##barplot for continous variable

#statistics for feature age
hist(data$age, freq = FALSE, xlab = "Age",main = "Histgram of Age",)
rug(jitter(data$age))
lines(density(data$age), col= "red",lwd=2)
```



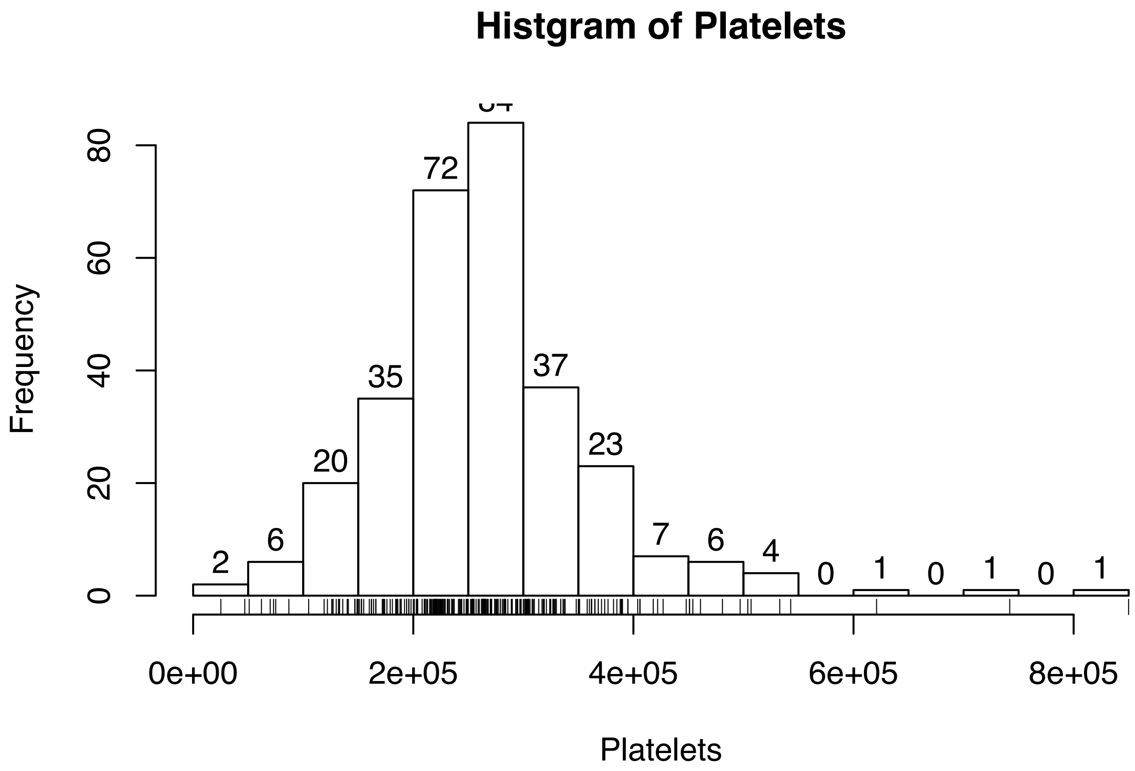
```
#statistics for feature creatinine_phosphokinase
hist(data$creatinine_phosphokinase,xlab = "Creatinine_phosphokinase", main ="Histgram of Creatinine_phos
rug(jitter(data$creatinine_phosphokinase))
```



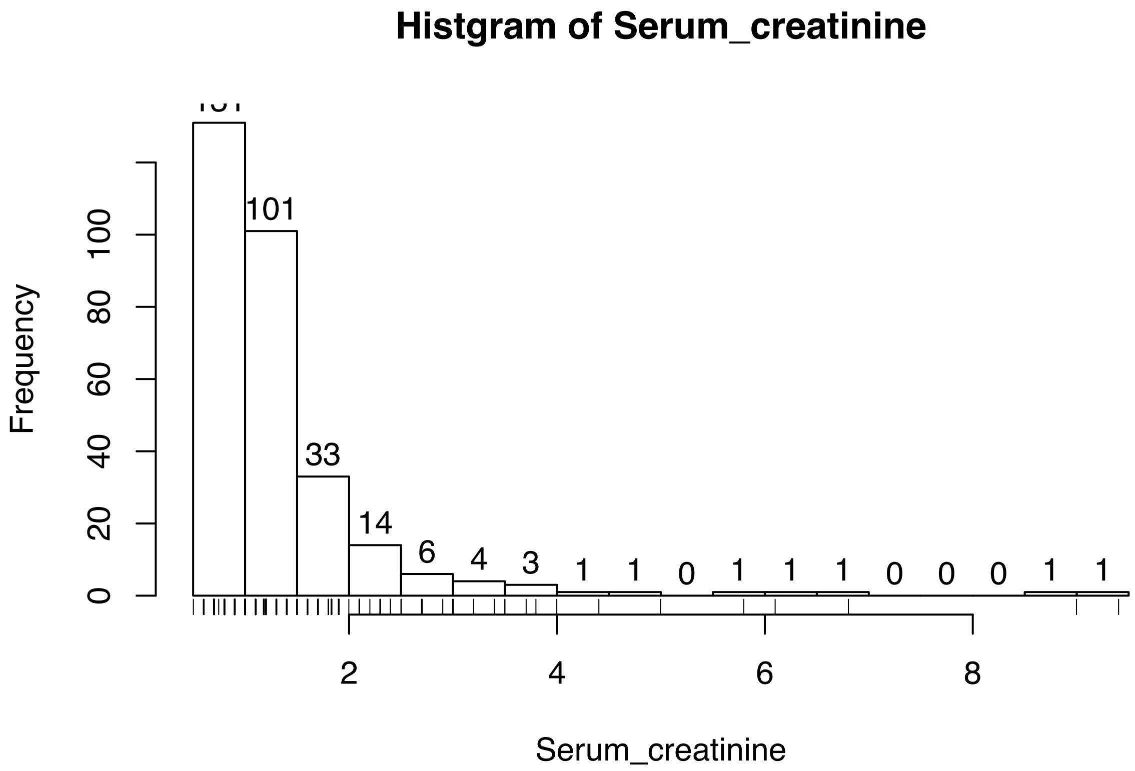
```
#statistics for feature ejection_fraction
hist(data$ejection_fraction,xlab = "Ejection_fraction", main ="Histgram of Ejection_fraction",labels = T
rug(jitter(data$ejection_fraction))
```



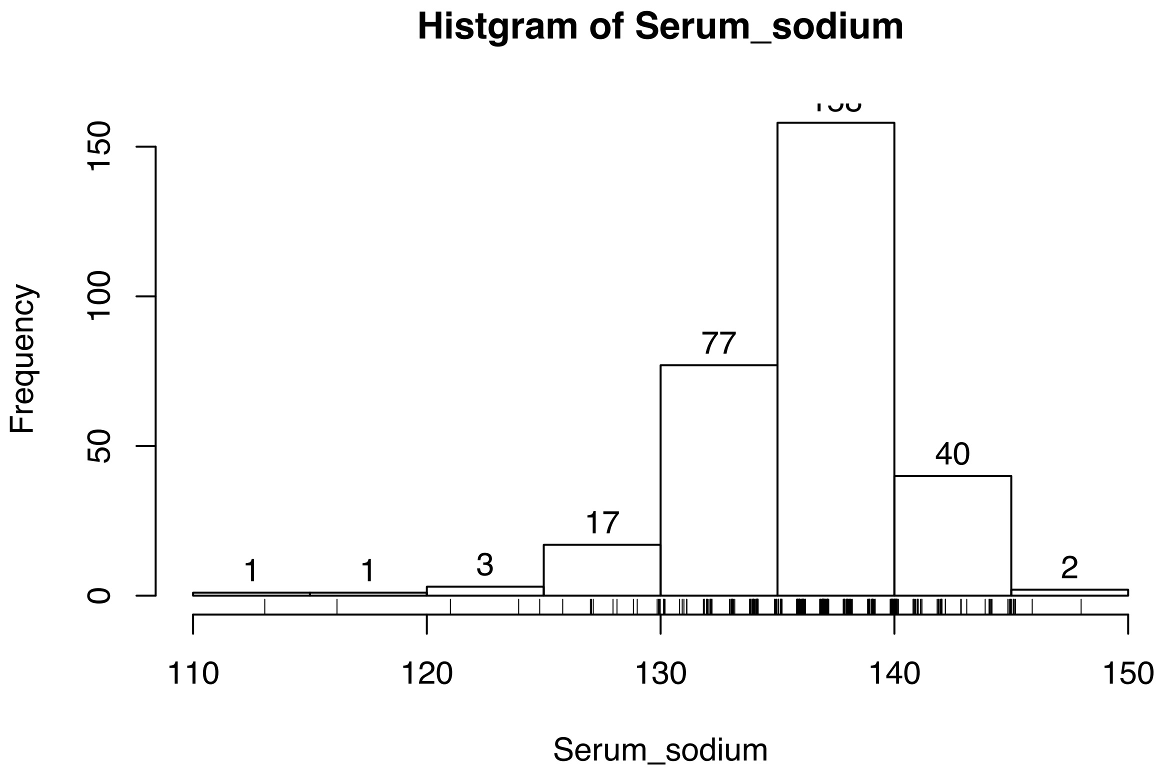
```
#statistics for feature platelets
hist(data$platelets,xlab = "Platelets", main ="Histgram of Platelets",labels = TRUE, breaks = 15)
rug(jitter(data$platelets))
```



```
#statistics for feature serum_creatinine
hist(data$serum_creatinine,xlab = "Serum_creatinine", main ="Histgram of Serum_creatinine",labels = TRU
rug(jitter(data$serum_creatinine))
```

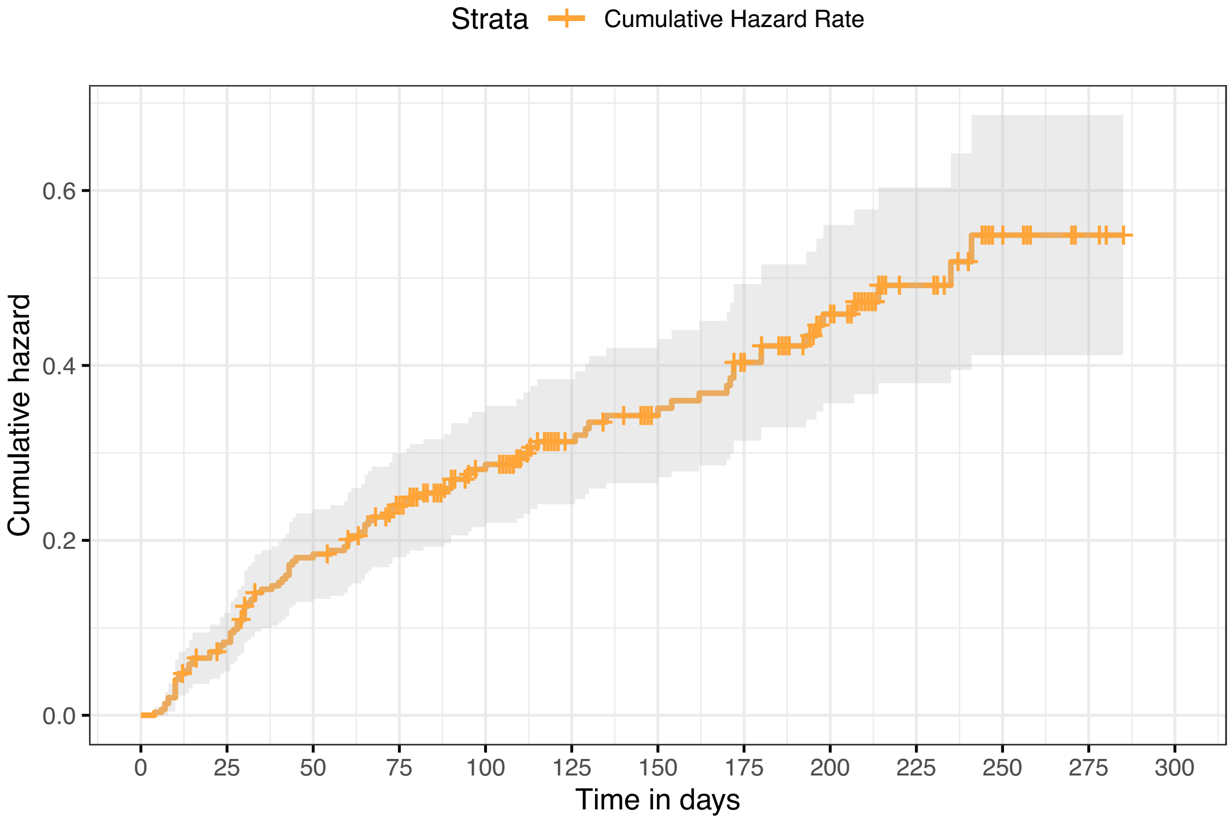


```
#statistics for feature serum_sodium
hist(data$serum_sodium,xlab = "Serum_sodium", main ="Histgram of Serum_sodium",labels = TRUE)
rug(jitter(data$serum_sodium))
```



```r
#Cumulative Hazard Function & Cumulative Death Probability

#Survival Probability by using cumulative hazard function

surv.na <- survfit(Surv(time, DEATH_EVENT) ~ 1, data = data, type = "fl")

ggsurvplot(surv.na,
           data = data,
           fun = 'cumhaz',
           pval = TRUE,
           conf.int = 0.95,
           conf.int.style ='ribbon',
           xlab = 'Time in days',
           break.time.by = 25,
           palette = c("#ffa436"),
            ggtheme = theme_bw(),
           legend.labs = c('Cumulative Hazard Rate')
           )
```

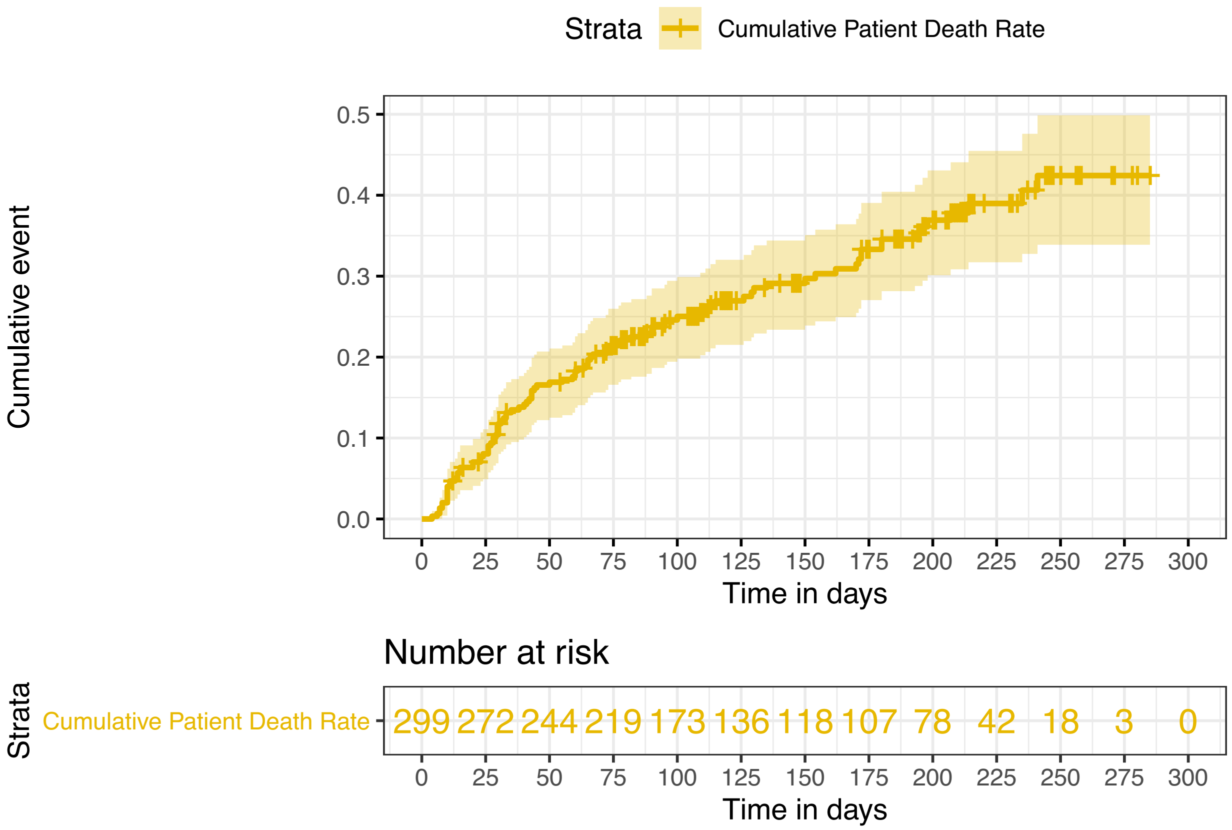
```
## Warning in .pvalue(fit, data = data, method = method, pval = pval, pval.coord = pval.coord, : There a
##  This is a null model.
```



```
h.sort.of <- surv.na$n.event / surv.na$n.risk

H.na <- cumsum(h.sort.of)

data.frame( time = surv.na$time, cumulative_hazard = H.na)
```

```
##     time cumulative_hazard
## 1      4       0.003344482
## 2      6       0.006700186
## 3      7       0.013434193
## 4      8       0.020213854
## 5     10       0.040691670
## 6     11       0.047660311
## 7     12       0.047660311
## 8     13       0.051181438
## 9     14       0.058248575
## 10    15       0.065366013
## 11    16       0.065366013
## 12    20       0.072560258
## 13    22       0.072560258
## 14    23       0.079832985
## 15    24       0.083495989
## 16    26       0.094525400
## 17    27       0.098242873
## 18    28       0.105705559
## 19    29        0.109464958
## 20    30        0.124616473
## 21    31        0.128477477
## 22    32        0.132353446
## 23    33        0.140135547
## 24    35        0.144072555
## 25    38        0.148025124
## 26    40        0.151993378
## 27    41        0.155977442
## 28    42        0.159977442
## 29    43        0.172025634
## 30    44        0.176090675
## 31    45        0.180172308
## 32    50        0.184270668
## 33    54        0.184270668
## 34    55        0.188420046
## 35    59        0.192586713
## 36    60        0.200954913
## 37    61        0.205192202
## 38    63        0.205192202
## 39    64        0.209465706
## 40    65        0.218049397
## 41    66        0.222378401
## 42    67        0.226726227
## 43    68        0.226726227
## 44    71        0.226726227
## 45    72        0.231131514
## 46    73        0.240020402
## 47    74        0.240020402
## 48    75        0.240020402
## 49    76        0.240020402
## 50    77        0.244628697
## 51    78        0.249258327
## 52    79        0.249258327
## 53    80        0.249258327
## 54    82        0.254089245
## 55    83        0.254089245
## 56    85        0.254089245
## 57    86        0.254089245
## 58    87        0.254089245
## 59    88        0.259243884
## 60    90        0.269825895
## 61    91        0.269825895
## 62    94        0.269825895
## 63    95        0.275381450
## 64    96        0.281095736
## 65    97        0.281095736
## 66   100        0.286876083
## 67   104        0.286876083
## 68   105        0.286876083
## 69   106        0.286876083
## 70   107        0.286876083
## 71   108        0.286876083
## 72   109        0.293165391
## 73   110       0.293165391
## 74   111       0.299617004
## 75   112       0.299617004
## 76   113       0.306195951
## 77   115       0.312862618
## 78   117       0.312862618
## 79   118       0.312862618
## 80   119       0.312862618
## 81   120       0.312862618
## 82   121       0.312862618
## 83   123       0.312862618
## 84   126       0.320215559
## 85   129       0.327622966
## 86   130       0.335085653
## 87   134       0.335085653
## 88   135       0.342661411
## 89   140       0.342661411
## 90   145       0.342661411
## 91   146       0.342661411
## 92   147       0.342661411
## 93   148       0.342661411
## 94   150       0.351135987
## 95   154       0.359682995
## 96   162       0.368303685
## 97   170       0.376999337
## 98   171       0.385771267
## 99   172       0.403470382
## 100  174       0.403470382
## 101  175       0.403470382
## 102  180       0.422338307
## 103  185       0.422338307
## 104  186       0.422338307
## 105  187       0.422338307
## 106  188       0.422338307
## 107  192       0.422338307
## 108  193       0.433966214
## 109  194       0.433966214
## 110  195       0.433966214
## 111  196       0.446014406
## 112  197       0.446014406
## 113  198       0.458672634
## 114  200       0.458672634
## 115  201       0.458672634
## 116  205       0.458672634
## 117  206       0.458672634
## 118  207       0.472757141
## 119  208       0.472757141
## 120  209       0.472757141
## 121  210       0.472757141
## 122  211       0.472757141
## 123  212       0.472757141
## 124  213       0.472757141
## 125  214       0.491625066
## 126  215       0.491625066
## 127  216       0.491625066
## 128  220       0.491625066
## 129  230       0.491625066
## 130  231       0.491625066
## 131  233       0.491625066
## 132  235       0.518652093
## 133  237       0.518652093
## 134  240       0.518652093
## 135  241       0.548955123
## 136  244       0.548955123
## 137  245       0.548955123
## 138  246       0.548955123
## 139  247       0.548955123
## 140  250       0.548955123
## 141  256       0.548955123
## 142  257       0.548955123
## 143  258       0.548955123
## 144  270       0.548955123
## 145  271       0.548955123
## 146  278       0.548955123
## 147  280       0.548955123
## 148  285       0.548955123
```

```
#Cumulative Death

surv.da <- survfit(Surv(time, DEATH_EVENT) ~ 1, data = data)
ggsurvplot(surv.da,
           data = data,
           risk.table.col = "strata",
           palette = c("#E7B800"),
           xlab = 'Time in days',
           break.time.by = 25,
            ggtheme = theme_bw(),
           legend.labs = c('Cumulative Patient Death Rate'),
           fun = 'event',
           risk.table = TRUE)
```

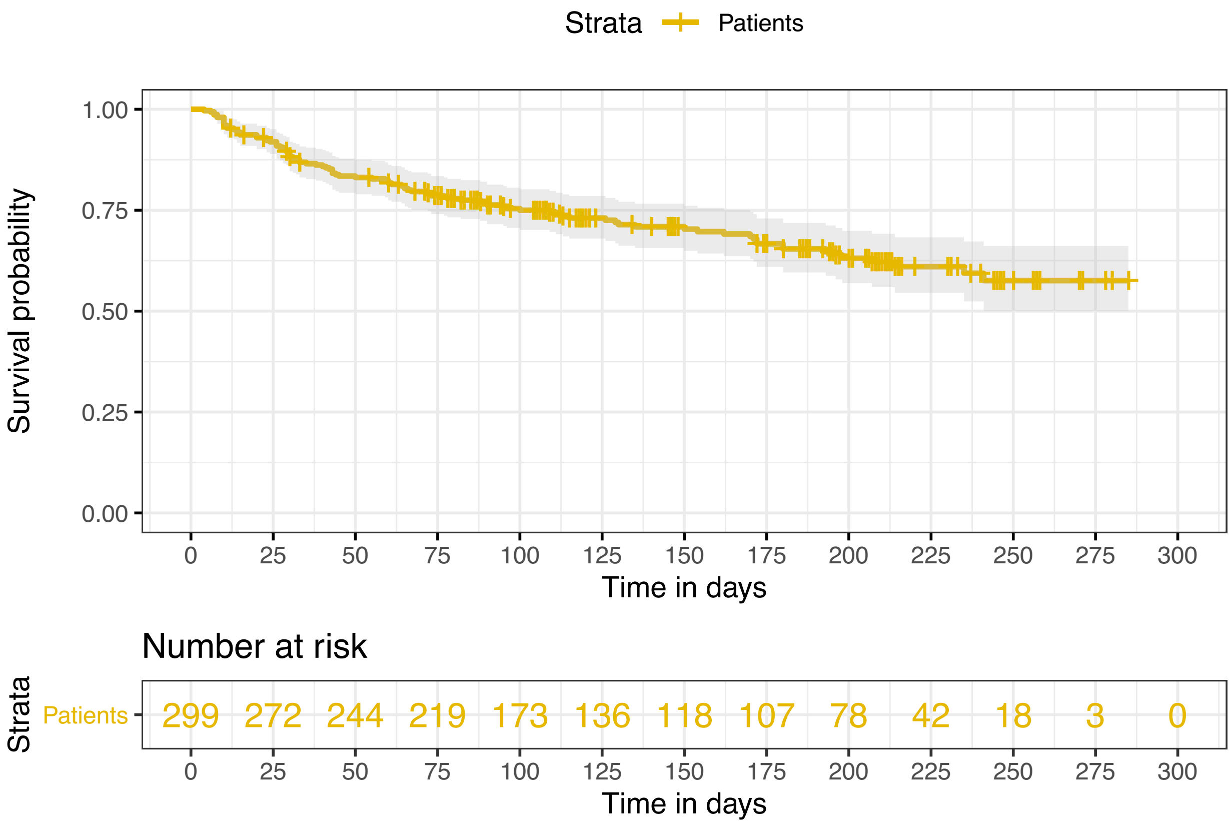


```
summary(surv.da)
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ 1, data = data)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     4    299       1    0.997 0.00334        0.990        1.000
##     6    298       1    0.993 0.00471        0.984        1.000
##     7    297       2    0.987 0.00664        0.974        1.000
##     8    295       2    0.980 0.00811        0.964        0.996
##    10    293       6    0.960 0.01135        0.938        0.982
##    11    287       2    0.953 0.01222        0.930        0.977
##    13    284       1    0.950 0.01263        0.925        0.975
##    14    283       2    0.943 0.01340        0.917        0.970
##    15    281       2    0.936 0.01412        0.909        0.964
##    20    278       2    0.930 0.01480        0.901        0.959
##    23    275       2    0.923 0.01545        0.893        0.954
##    24    273       1    0.920 0.01575        0.889        0.951
##    26    272       3    0.909 0.01663        0.877        0.943
##    27    269       1    0.906 0.01691        0.873        0.940
##    28    268       2    0.899 0.01745        0.866        0.934
##    29    266       1    0.896 0.01771        0.862        0.931
##    30    264       4    0.882 0.01869        0.846        0.920
##    31    259       1    0.879 0.01893        0.843        0.917
##    32    258       1    0.875 0.01916        0.839        0.914
##    33    257       2    0.869 0.01961        0.831        0.908
##    35    254       1    0.865 0.01983        0.827        0.905
##    38    253       1    0.862 0.02004        0.823        0.902
##    40    252       1    0.858 0.02025        0.820        0.899
##    41    251       1    0.855 0.02046        0.816        0.896
##    42    250       1    0.852 0.02066        0.812        0.893
##    43    249       3    0.841 0.02124        0.801        0.884
##    44    246       1    0.838 0.02143        0.797        0.881
##    45    245       1    0.834 0.02161        0.793        0.878
##    50    244       1    0.831 0.02179        0.789        0.875
##    55    241       1    0.828 0.02197        0.786        0.872
##    59    240       1    0.824 0.02215        0.782        0.869
##    60    239       2    0.817 0.02250        0.774        0.863
##    61    236       1    0.814 0.02267        0.771        0.859
##    64    234       1    0.810 0.02283        0.767        0.856
##    65    233       2    0.803 0.02316        0.759        0.850
##    66    231       1    0.800 0.02332        0.755        0.847
##    67    230       1    0.796 0.02348        0.752        0.844
##    72    227       1    0.793 0.02364        0.748        0.841
##    73    225       2    0.786 0.02394        0.740        0.834
##    77    217       1    0.782 0.02411        0.736        0.831
##    78    216       1    0.779 0.02427        0.732        0.828
##    82    207       1    0.775 0.02444        0.728        0.824
##    88    194       1    0.771 0.02464        0.724        0.821
##    90    189       2    0.763 0.02504        0.715        0.813
##    95    180       1    0.758 0.02526        0.711        0.810
##    96    175       1    0.754 0.02548        0.706        0.806
##   100    173       1    0.750 0.02571        0.701        0.802
##   109    159       1    0.745 0.02597        0.696        0.798
##   111    155       1    0.740 0.02625        0.691        0.794
##   113    152       1    0.735 0.02652        0.685        0.789
##   115    150       1    0.730 0.02679        0.680        0.785
##   126    136       1    0.725 0.02713        0.674        0.780
##   129    135       1    0.720 0.02746        0.668        0.776
##   130    134       1    0.714 0.02777        0.662        0.771
##   135    132       1    0.709 0.02808        0.656        0.766
##   150    118       1    0.703 0.02848        0.649        0.761
##   154    117       1    0.697 0.02886        0.643        0.756
##   162    116       1    0.691 0.02923        0.636        0.751
##   170    115       1    0.685 0.02959        0.629        0.745
##   171    114       1    0.679 0.02993        0.623        0.740
##   172    113       2    0.667 0.03059        0.610        0.730
##   180    106       2    0.654 0.03128        0.596        0.719
##   193     86       1    0.647 0.03183        0.587        0.712
##   196     83       1    0.639 0.03238        0.578        0.706
##   198     79       1    0.631 0.03297        0.569        0.699
##   207     71       1    0.622 0.03368        0.559        0.692
##   214     53       1    0.610 0.03503        0.545        0.683
##   235     37       1    0.594 0.03776        0.524        0.673
##   241     33       1    0.576 0.04068        0.501        0.661
```

```
#Kaplan-Meier Estimate
surv.km <- survfit(Surv(time, DEATH_EVENT) ~ 1, data = data) #KM estimate

ggsurvplot(surv.km,
           data = data,
           pval = TRUE,
           conf.int = 0.95,
           conf.int.style ='ribbon',
           xlab = 'Time in days',
           break.time.by = 25,
           risk.table = TRUE,
           risk.table.y.text.col = TRUE,
           risk.table.col = 'strata',
           linetype = 'strata',
            ggtheme = theme_bw(),
           legend.labs = c('Patients'),
           palette = c("#E7B800"),

           )
```

```
## Warning in .pvalue(fit, data = data, method = method, pval = pval, pval.coord = pval.coord, : There a
##  This is a null model.
```

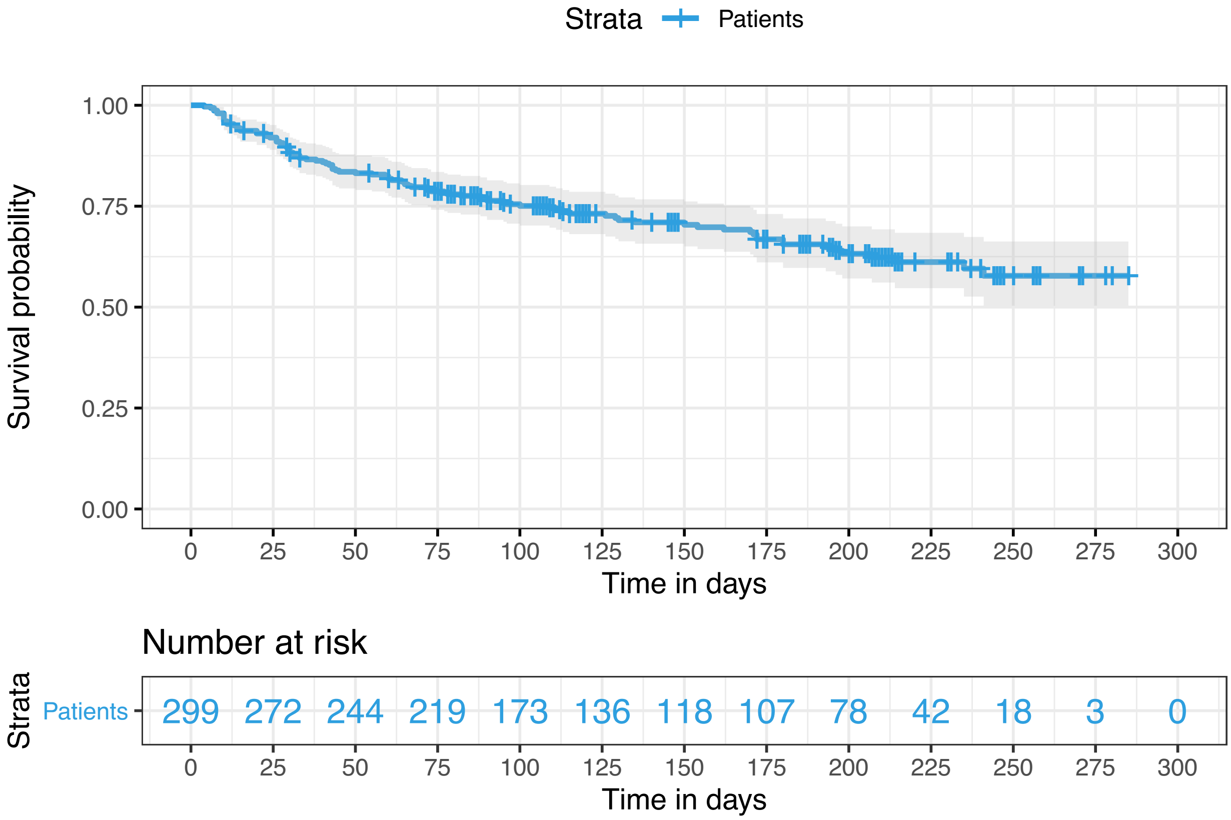


```
summary(surv.km)
```

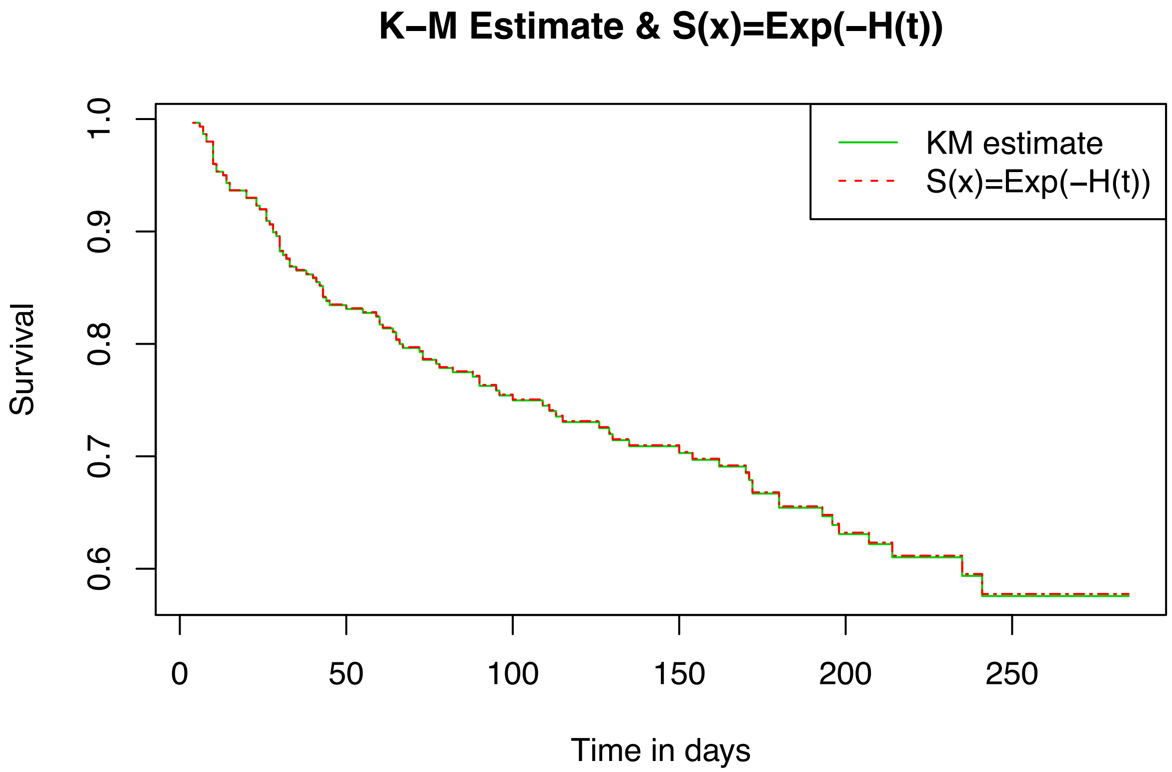
```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ 1, data = data)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     4    299       1    0.997 0.00334        0.990        1.000
##     6    298       1    0.993 0.00471        0.984        1.000
##     7    297       2    0.987 0.00664        0.974        1.000
##     8    295       2    0.980 0.00811        0.964        0.996
##    10    293       6    0.960 0.01135        0.938        0.982
##    11    287       2    0.953 0.01222        0.930        0.977
##    13    284       1    0.950 0.01263        0.925        0.975
##    14    283       2    0.943 0.01340        0.917        0.970
##    15    281       2    0.936 0.01412        0.909        0.964
##    20    278       2    0.930 0.01480        0.901        0.959
##    23    275       2    0.923 0.01545        0.893        0.954
##    24    273       1    0.920 0.01575        0.889        0.951
##    26    272       3    0.909 0.01663        0.877        0.943
##    27    269       1    0.906 0.01691        0.873        0.940
##    28    268       2    0.899 0.01745        0.866        0.934
##    29    266       1    0.896 0.01771        0.862        0.931
##    30    264       4    0.882 0.01869        0.846        0.920
##    31    259       1    0.879 0.01893        0.843        0.917
##    32    258       1    0.875 0.01916        0.839        0.914
##    33    257       2    0.869 0.01961        0.831        0.908
##    35    254       1    0.865 0.01983        0.827        0.905
##    38    253       1    0.862 0.02004        0.823        0.902
##    40    252       1    0.858 0.02025        0.820        0.899
##    41    251       1    0.855 0.02046        0.816        0.896
##    42    250       1    0.852 0.02066        0.812        0.893
##    43    249       3    0.841 0.02124        0.801        0.884
##    44    246       1    0.838 0.02143        0.797        0.881
##    45    245       1    0.834 0.02161        0.793        0.878
##    50    244       1    0.831 0.02179        0.789        0.875
##    55    241       1    0.828 0.02197        0.786        0.872
##    59    240       1    0.824 0.02215        0.782        0.869
##    60    239       2    0.817 0.02250        0.774        0.863
##    61    236       1    0.814 0.02267        0.771        0.859
##    64    234       1    0.810 0.02283        0.767        0.856
##    65    233       2    0.803 0.02316        0.759        0.850
##    66    231       1    0.800 0.02332        0.755        0.847
##    67    230       1    0.796 0.02348        0.752        0.844
##    72    227       1    0.793 0.02364        0.748        0.841
##    73    225       2    0.786 0.02394        0.740        0.834
##    77    217       1    0.782 0.02411        0.736        0.831
##    78    216       1    0.779 0.02427        0.732        0.828
##    82    207       1    0.775 0.02444        0.728        0.824
##    88    194       1    0.771 0.02464        0.724        0.821
##    90    189       2    0.763 0.02504        0.715        0.813
##    95    180       1    0.758 0.02526        0.711        0.810
##    96    175       1    0.754 0.02548        0.706        0.806
##   100    173       1    0.750 0.02571        0.701        0.802
##   109    159       1    0.745 0.02597        0.696        0.798
##   111    155       1    0.740 0.02625        0.691        0.794
##   113    152       1    0.735 0.02652        0.685        0.789
##   115    150       1    0.730 0.02679        0.680        0.785
##   126    136       1    0.725 0.02713        0.674        0.780
##   129    135       1    0.720 0.02746        0.668        0.776
##   130    134       1    0.714 0.02777        0.662        0.771
##   135    132       1    0.709 0.02808        0.656        0.766
##   150    118       1    0.703 0.02848        0.649        0.761
##   154    117       1    0.697 0.02886        0.643        0.756
##   162    116       1    0.691 0.02923        0.636        0.751
##   170    115       1    0.685 0.02959        0.629        0.745
##   171    114       1    0.679 0.02993        0.623        0.740
##   172    113       2    0.667 0.03059        0.610        0.730
##   180    106       2    0.654 0.03128        0.596        0.719
##   193     86       1    0.647 0.03183        0.587        0.712
##   196     83       1    0.639 0.03238        0.578        0.706
##   198     79       1    0.631 0.03297        0.569        0.699
##   207     71       1    0.622 0.03368        0.559        0.692
##   214     53       1    0.610 0.03503        0.545        0.683
##   235     37       1    0.594 0.03776        0.524        0.673
##   241     33       1    0.576 0.04068        0.501        0.661
```

```r
#we use the formula S(x)=Exp(-H(t)) to estimate the survival probaility


ggsurvplot(surv.na,
           data = data,
           pval = TRUE,
           conf.int = 0.95,
           conf.int.style ='ribbon',
           xlab = 'Time in days',
           break.time.by = 25,
           risk.table = TRUE,
           risk.table.y.text.col = TRUE,
           risk.table.col = 'strata',
           linetype = 'strata',
           ggtheme = theme_bw(),
           legend.labs = c('Patients'),
           palette = c("#2E9FDF")
           )
```

```
## Warning in .pvalue(fit, data = data, method = method, pval = pval, pval.coord = pval.coord, : There
##  This is a null model.
```



```
#Comparison between two way to calculate the survival probability.

plot(surv.km$time,
     surv.km$surv,
     type="s",
     xlab="Time in days",
     ylab="Survival",col="3",
     )

lines(surv.na$time, surv.na$surv, type="s", lty=6,col="2")

legend("topright", legend=c("KM estimate","S(x)=Exp(-H(t))"), lty=1:4,col=3:2)
title(main="K-M Estimate & S(x)=Exp(-H(t))")
```



```
#There is no difference between two ways
```

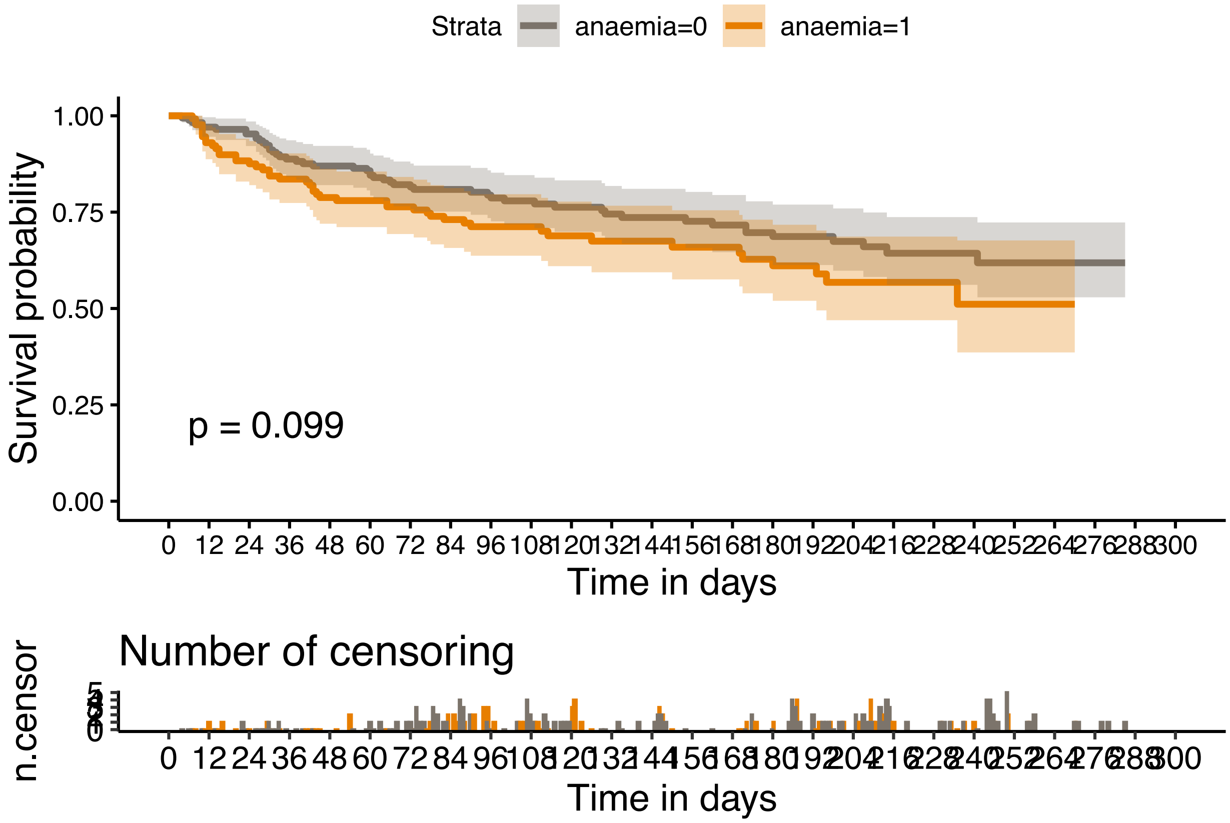
```
##All binary variable we have in this data_set

feat_con <- c("age", "creatinine_phosphokinase", "ejection_fraction", "platelets", "serum_creatinine",
feat_cat <- setdiff(cols, feat_con)
cols_exclude <- setdiff(names(data), cols)

feat_cat
```

```
## [1] "anaemia"              "diabetes"             "high_blood_pressure"
## [4] "sex"                  "smoking"
```

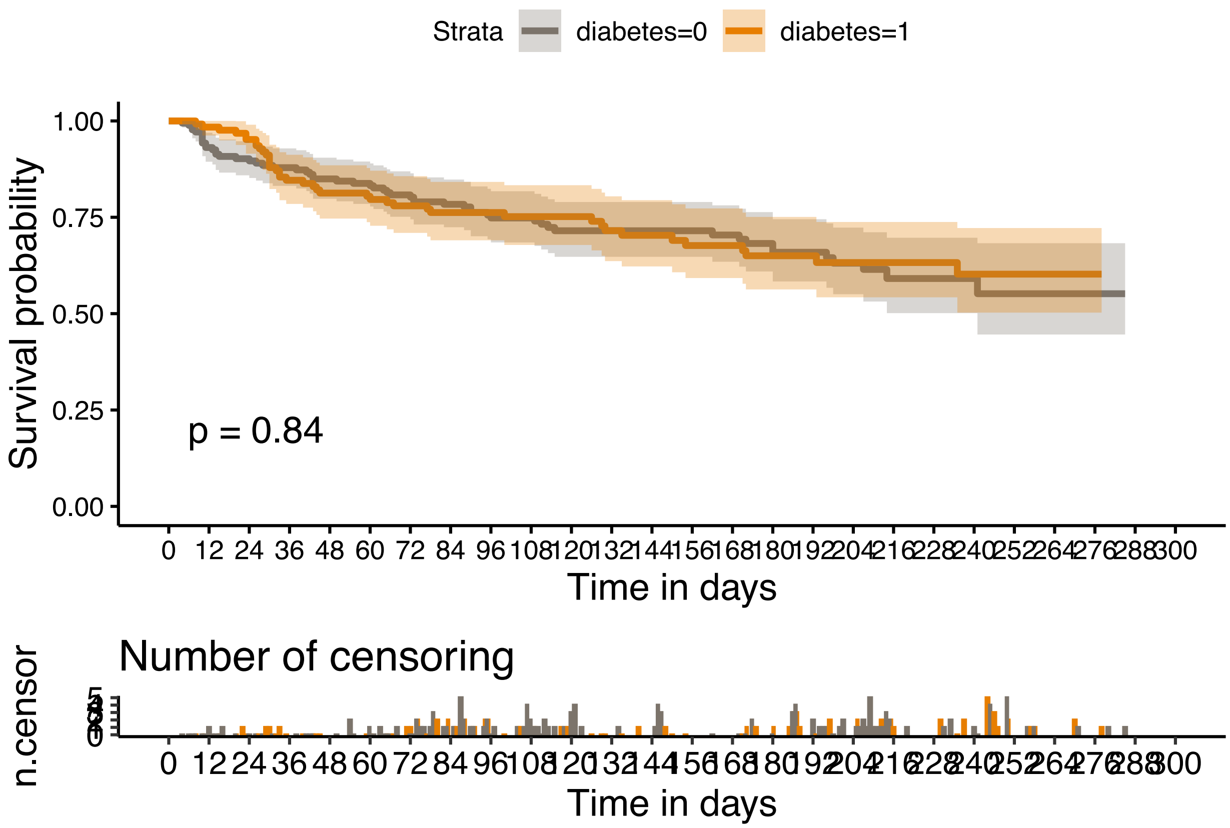
```
cols_exclude
```

```
## [1] "time"        "DEATH_EVENT"
```

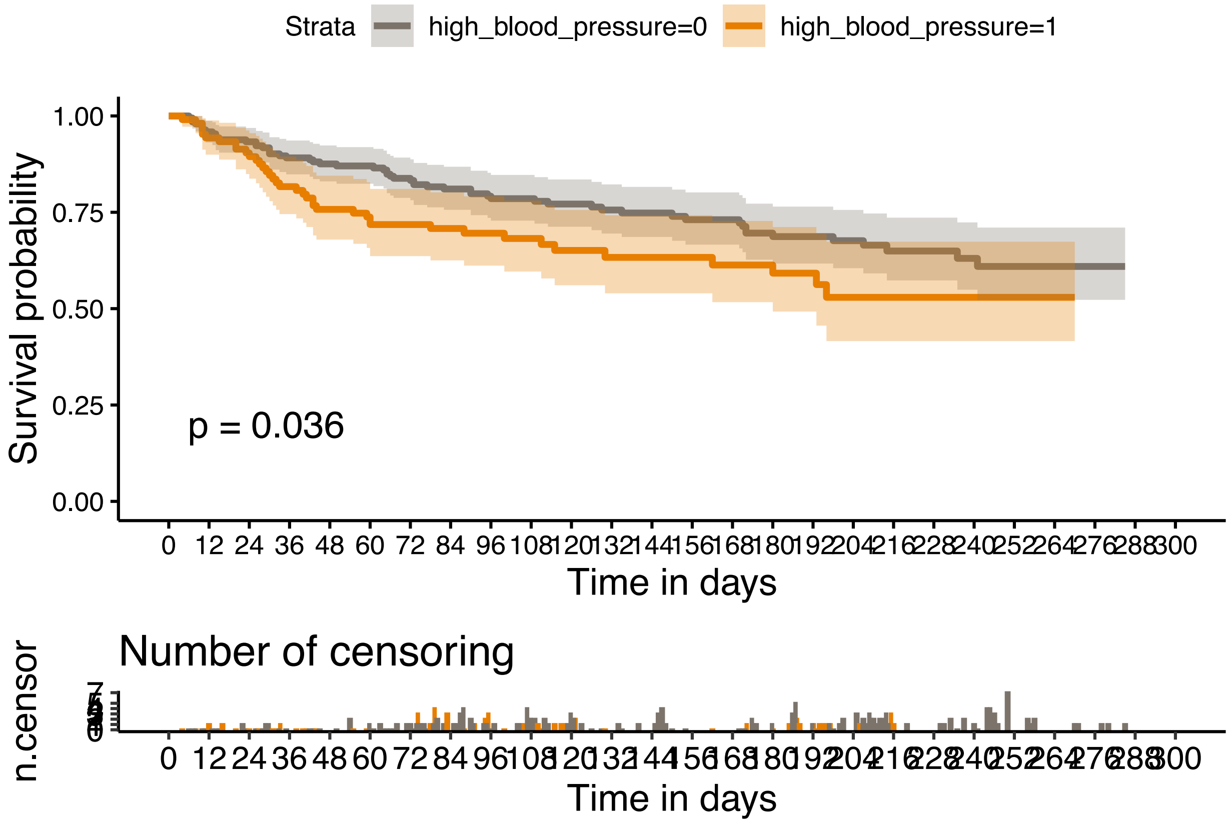
```
#Features of Anaemia

ggsurvplot(survfit(Surv(time, DEATH_EVENT) ~ anaemia, data = data, start.time = 0),
           data = data,
           pval = TRUE,
           conf.int = TRUE,
           xlab = 'Time in days',
           censor.shape = "",
           ncensor.plot = TRUE,
           legend = "top",
           break.x.by = 12,
           font.tickslab = 10,
           palette = c("#7c746b","#e77e00"),
           size = 1.2)
```



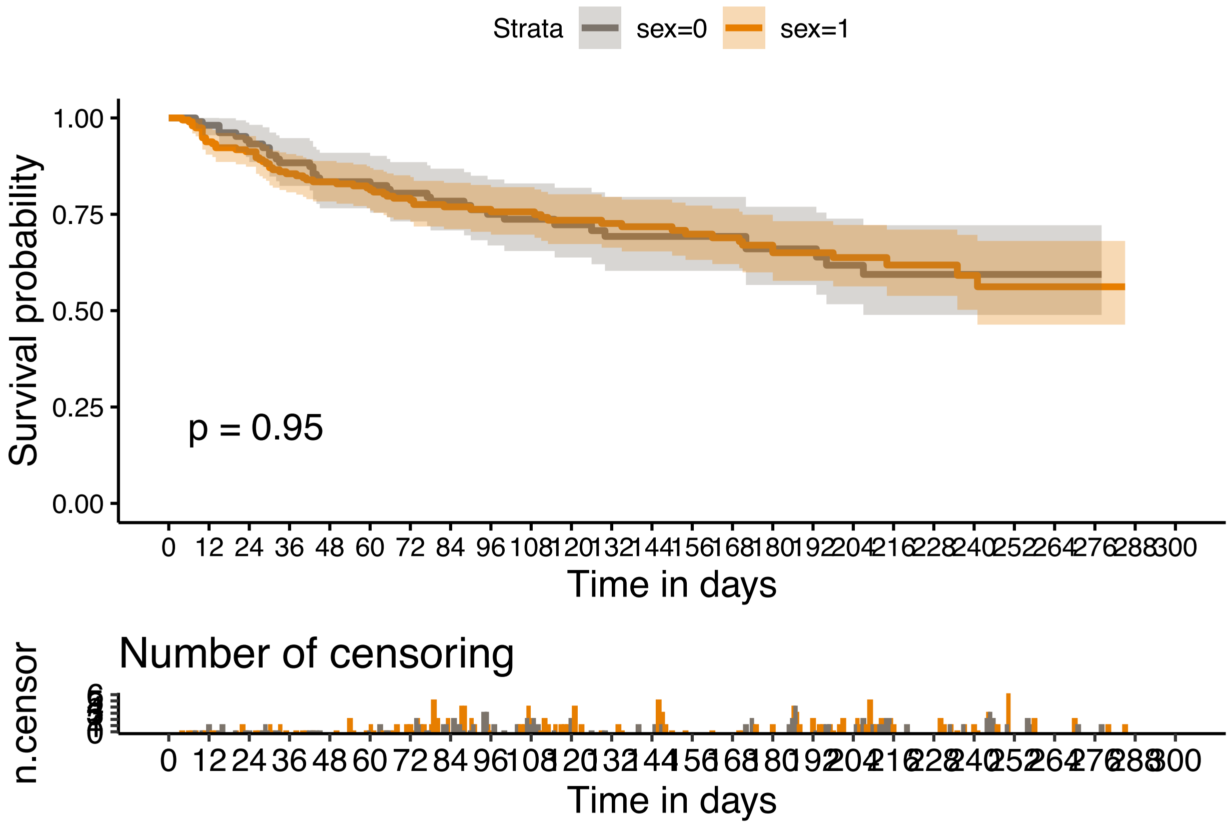
```
#Features of Diabetes

ggsurvplot(survfit(Surv(time, DEATH_EVENT) ~ diabetes, data = data, start.time = 0),
           data = data,
           pval = TRUE,
           conf.int = TRUE,
           xlab = 'Time in days',
           censor.shape = "",
           ncensor.plot = TRUE,
           legend = "top",
           break.x.by = 12,
           font.tickslab = 10,
           palette = c("#7c746b","#e77e00"),
           size = 1.2)
```



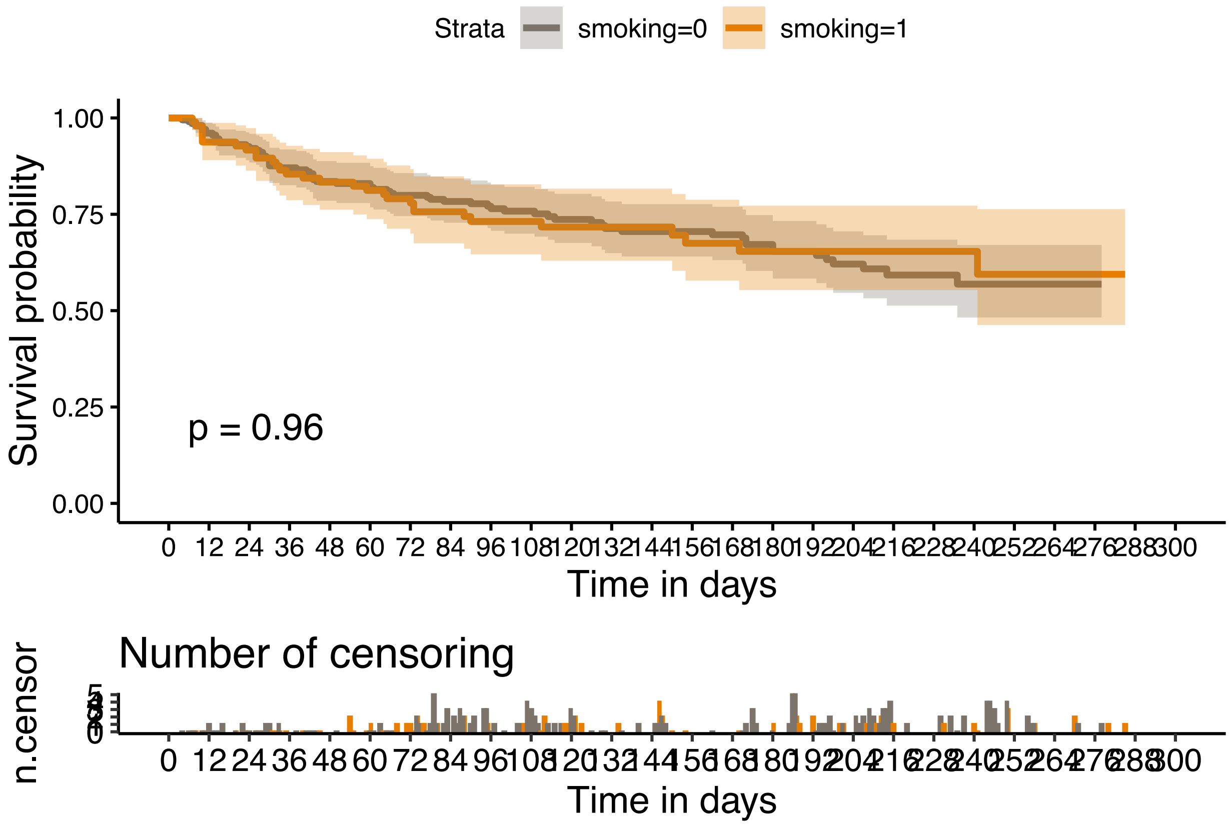
```r
#Features of High_blood_Pressure

ggsurvplot(survfit(Surv(time, DEATH_EVENT) ~ high_blood_pressure, data = data, start.time = 0),
           data = data,
           pval = TRUE,
           conf.int = TRUE,
           xlab = 'Time in days',
           censor.shape = "",
           ncensor.plot = TRUE,
           legend = "top",
           break.x.by = 12,
           font.tickslab = 10,
           palette = c("#7c746b","#e77e00"),
           size = 1.2)
```



```r
#Features of Sex

ggsurvplot(survfit(Surv(time, DEATH_EVENT) ~ sex, data = data, start.time = 0),
           data = data,
           pval = TRUE,
           conf.int = TRUE,
           xlab = 'Time in days',
           censor.shape = "",
           ncensor.plot = TRUE,
           legend = "top",
           break.x.by = 12,
           font.tickslab = 10,
           palette = c("#7c746b","#e77e00"),
           size = 1.2)
```



```
#Features of Smoking

ggsurvplot(survfit(Surv(time, DEATH_EVENT) ~ smoking, data = data, start.time = 0),
           data = data,
           pval = TRUE,
           conf.int = TRUE,
           xlab = 'Time in days',
           censor.shape = "",
           ncensor.plot = TRUE,
           legend = "top",
           break.x.by = 12,
           font.tickslab = 10,
           palette = c("#7c746b","#e77e00"),
           size = 1.2)
```



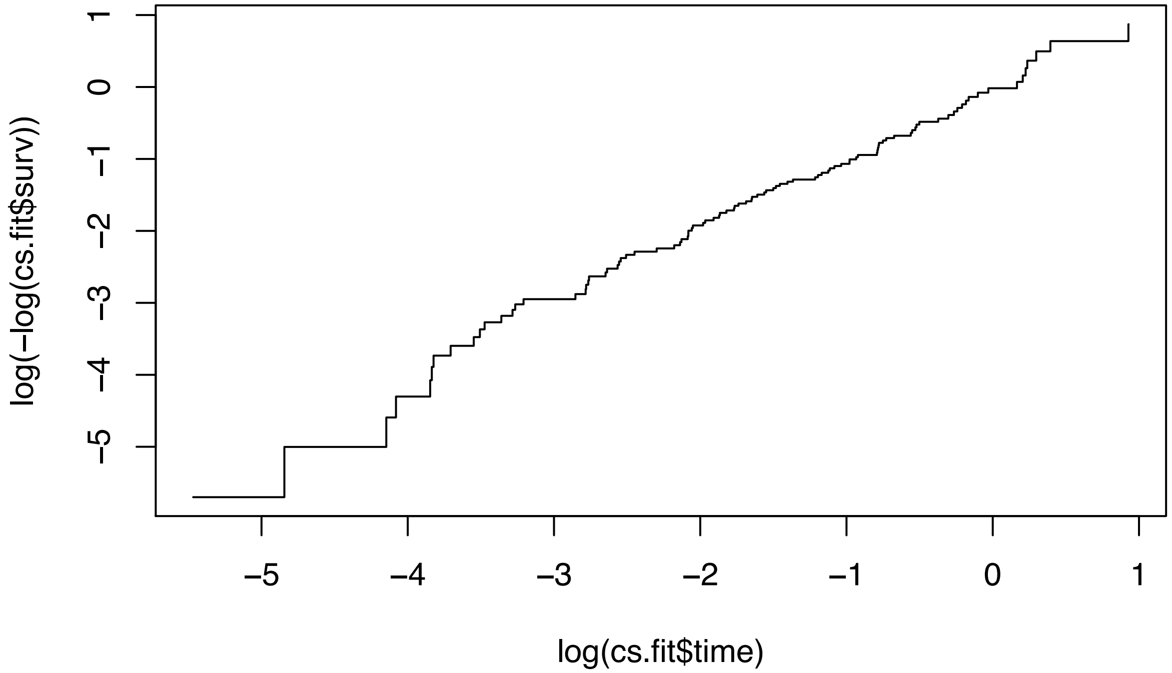
```
##AFT model
##becase by considering that age will be an effect,so we did not use expoentional distribution to fit th
##First we use all features to fit the model.

aftmodel.full <- survreg(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine + serum_so
                       creatinine_phosphokinase + platelets + diabetes + sex + smoking,
                     dist = 'weibull', data = data)

summary(aftmodel.full)
```

```
##
## Call:
## survreg(formula = Surv(time, DEATH_EVENT) ~ age + ejection_fraction +
##     serum_creatinine + serum_sodium + anaemia + high_blood_pressure +
##     creatinine_phosphokinase + platelets + diabetes + sex + smoking,
##     data = data, dist = "weibull")
##                              Value Std. Error     z       p
## (Intercept)               1.99e+00   3.29e+00  0.61   0.545
## age                      -4.98e-02   1.00e-02 -4.96 7.2e-07
## ejection_fraction         5.25e-02   1.16e-02  4.53 5.9e-06
## serum_creatinine         -3.33e-01   7.31e-02 -4.56 5.1e-06
## serum_sodium              4.50e-02   2.41e-02  1.87   0.062
## anaemia1                 -5.00e-01   2.24e-01 -2.24   0.025
## high_blood_pressure1     -5.14e-01   2.22e-01 -2.31   0.021
## creatinine_phosphokinase -2.43e-04   1.04e-04 -2.34   0.019
## platelets                 5.51e-07   1.18e-06  0.47   0.641
## diabetes1                 -1.47e-01   2.32e-01 -0.63   0.528
## sex1                       2.46e-01   2.63e-01  0.93   0.350
## smoking1                  -1.19e-01   2.61e-01 -0.45   0.649
## Log(scale)                 3.80e-02   8.88e-02  0.43   0.669
##
## Scale= 1.04
##
## Weibull distribution
## Loglik(model)= -628.1   Loglik(intercept only)= -670.4
##  Chisq= 84.64 on 11 degrees of freedom, p= 1.9e-13
## Number of Newton-Raphson Iterations: 6
## n= 299
```

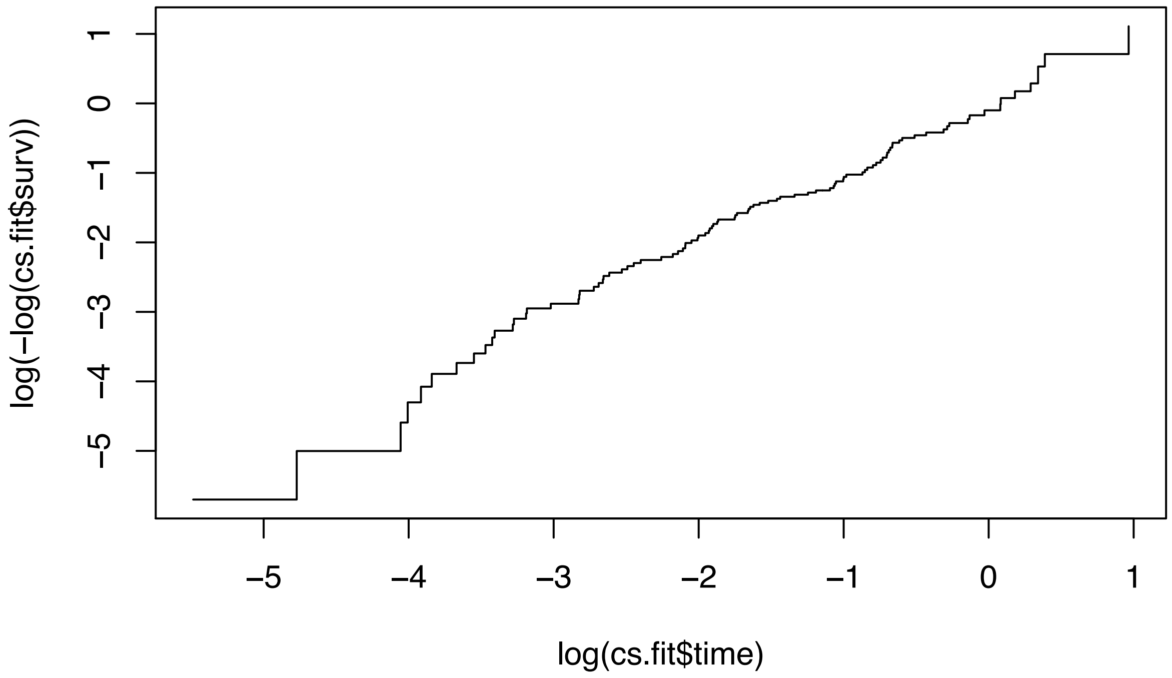
```
#we get rid of the insignificant effect of the model.
aftmodel.part <- survreg(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine + serum_s
                     dist = 'weibull', data = data)

summary(aftmodel.part)
```

```
##
## Call:
## survreg(formula = Surv(time, DEATH_EVENT) ~ age + ejection_fraction +
##     serum_creatinine + serum_sodium + anaemia + high_blood_pressure +
##     creatinine_phosphokinase, data = data, dist = "weibull")
##                              Value Std. Error     z       p
## (Intercept)               1.795142   3.252849  0.55   0.581
## age                      -0.046745   0.009462 -4.94 7.8e-07
## ejection_fraction         0.050860   0.011278  4.51 6.5e-06
## serum_creatinine         -0.325122   0.071385 -4.55 5.3e-06
## serum_sodium              0.046749   0.024193  1.93   0.053
## anaemia1                 -0.486299   0.221263 -2.20   0.028
## high_blood_pressure1     -0.533074   0.219493 -2.43   0.015
## creatinine_phosphokinase -0.000231   0.000102 -2.26   0.024
## Log(scale)                0.035998   0.088754  0.41   0.685
##
## Scale= 1.04
##
## Weibull distribution
## Loglik(model)= -628.8   Loglik(intercept only)= -670.4
##  Chisq= 83.25 on 7 degrees of freedom, p= 3e-15
## Number of Newton-Raphson Iterations: 6
## n= 299
```

```
## linear predictor \beta * x
linpred <- aftmodel.full$linear.predictor
# Residuals (transfer residual to be the survival time scale, not log(time) case)
cs.res <- exp(-aftmodel.full$linear.predictor/aftmodel.full$scale)* (Surv(data$time, data$DEATH_EVENT)[
cs.fit <- survfit(Surv(cs.res, data$DEATH_EVENT) ~ 1, type="fh2")
cs.fit <- survfit(Surv(cs.res, data$DEATH_EVENT) ~ 1, type="fleming-harrington")


plot(log(cs.fit$time), log(-log(cs.fit$surv)), type="s")  ## if Weibull is OK, this will have a linear
```



```
## linear predictor \beta * x
linpred <- aftmodel.part$linear.predictor
# Residuals (transfer residual to be the survival time scale, not log(time) case)
cs.res <- exp(-aftmodel.part$linear.predictor/aftmodel.part$scale)* (Surv(data$time, data$DEATH_EVENT)[
cs.fit <- survfit(Surv(cs.res, data$DEATH_EVENT) ~ 1, type="fh2")
cs.fit <- survfit(Surv(cs.res, data$DEATH_EVENT) ~ 1, type="fleming-harrington")


plot(log(cs.fit$time), log(-log(cs.fit$surv)), type="s")  ## if Weibull is OK, this will have a linear
```



```r
options(contrasts=c("contr.treatment", "contr.treatment"))
dd <- datadist(data)
options(datadist='dd')

# Full model
full.model.mt <- cph(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine + serum_sodiu
                        creatinine_phosphokinase + platelets + diabetes + sex + smoking,
                        data = data, x = TRUE, y = TRUE)

#full model checking

cox.zph(full.model.mt)
```

```
##                              chisq df     p
## age                       1.02e-01  1 0.749
## ejection_fraction         4.68e+00  1 0.031
## serum_creatinine          1.53e+00  1 0.216
## serum_sodium              1.10e-01  1 0.740
## anaemia                   1.67e-02  1 0.897
## high_blood_pressure       8.14e-03  1 0.928
## creatinine_phosphokinase  1.02e+00  1 0.312
## platelets                 1.32e-05  1 0.997
## diabetes                  1.92e-01  1 0.661
## sex                       7.57e-02  1 0.783
## smoking                   4.78e-01  1 0.489
## GLOBAL                    1.17e+01 11 0.386
```

```
# Part model

part.model.mt <- cph(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine + serum_sodiu
                       data = data, x = TRUE, y = TRUE)

# Part model checking

cox.zph(part.model.mt)
```

```
##                          chisq df     p
## age                   0.093926  1 0.759
## ejection_fraction     4.541488  1 0.033
## serum_creatinine      1.540553  1 0.215
## serum_sodium          0.112834  1 0.737
## anaemia               0.000475  1 0.983
## high_blood_pressure   0.006359  1 0.936
## GLOBAL                8.455340  6 0.207
```

```
anova(full.model.mt)
```

```
##                  Wald Statistics          Response: Surv(time, DEATH_EVENT)
##
##  Factor                   Chi-Square d.f. P
##  age                      24.75       1   <.0001
##  ejection_fraction        21.80       1   <.0001
##  serum_creatinine         21.09       1   <.0001
##  serum_sodium              3.60       1   0.0577
##  anaemia                   4.51       1   0.0338
##  high_blood_pressure       4.85       1   0.0277
##  creatinine_phosphokinase  4.96       1   0.0260
##  platelets                 0.17       1   0.6804
##  diabetes                  0.39       1   0.5304
##  sex                       0.89       1   0.3448
##  smoking                   0.26       1   0.6073
##  TOTAL                    87.40      11   <.0001
```

```
anova(part.model.mt)
```

```
##                  Wald Statistics          Response: Surv(time, DEATH_EVENT)
##
##  Factor              Chi-Square d.f. P
##  age                 23.92      1    <.0001
##  ejection_fraction   21.01      1    <.0001
##  serum_creatinine    19.18      1    <.0001
##  serum_sodium         3.33      1    0.0682
##  anaemia              3.25      1    0.0712
##  high_blood_pressure  4.96      1    0.0260
##  TOTAL               84.74      6    <.0001
```

```
#likelihood ratio test check

lrtest(full.model.mt, part.model.mt)
```

```
## Likelihood ratio test
##
## Model 1: Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine +
##     serum_sodium + anaemia + high_blood_pressure + creatinine_phosphokinase +
##     platelets + diabetes + sex + smoking
## Model 2: Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine +
##     serum_sodium + anaemia + high_blood_pressure
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  11 -468.23
## 2   6 -470.72 -5 4.9874     0.4174
```

```
#the final Model
model.cox <- coxph(Surv(time, DEATH_EVENT) ~ age + ejection_fraction + serum_creatinine + serum_sodium
                   data = data, x = TRUE, y = TRUE)
summary(model.cox)
```

```
## Call:
## coxph(formula = Surv(time, DEATH_EVENT) ~ age + ejection_fraction +
##     serum_creatinine + serum_sodium + anaemia + high_blood_pressure,
##     data = data, x = TRUE, y = TRUE)
##
##   n= 299, number of events= 96
##
##                          coef exp(coef)  se(coef)      z Pr(>|z|)
## age                  0.043897  1.044875  0.008971  4.893 9.92e-07 ***
## ejection_fraction   -0.046742  0.954333  0.010191 -4.586 4.51e-06 ***
## serum_creatinine     0.304325  1.355710  0.069805  4.360 1.30e-05 ***
## serum_sodium        -0.043394  0.957534  0.023769 -1.826   0.0679 .
## anaemia1             0.379021  1.460854  0.210184  1.803   0.0713 .
## high_blood_pressure1  0.473583  1.605737  0.212753  2.226   0.0260 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## age                    1.0449     0.9571    1.0267    1.0634
## ejection_fraction      0.9543     1.0479    0.9355    0.9736
## serum_creatinine       1.3557     0.7376    1.1824    1.5545
## serum_sodium           0.9575     1.0443    0.9139    1.0032
## anaemia1               1.4609     0.6845    0.9676    2.2055
## high_blood_pressure1    1.6057     0.6228    1.0582    2.4365
##
## Concordance= 0.73  (se = 0.028 )
## Likelihood ratio test= 76.97  on 6 df,   p=2e-14
## Wald test            = 84.6  on 6 df,   p=4e-16
## Score (logrank) test = 84.19  on 6 df,   p=5e-16
```