Titme to heart failiure analysis

Christophe Mpaga, Ahmed Oulad Amara, Adrien Parruitte

Contents

Context 1

Introduction 2

data acquisition preparation and investigation. 2

data dictionary . . . 2

univariate analysis : group comparison 3

overall Kaplan-Meyer estimator . . . 3

Male vs. Female . . . 8

Smoking . . . 9

blood pressure . . . 10

diabetes . . . 12

the impacet of age . . . 14

Ejection Fraction . . . 15

Cox Proportional Hazards Model 19

Context

Heart failure is a chronic condition where the heart is unable to effectively pump enough blood to meet the body's needs. It occurs when the heart muscle becomes weakened or damaged, leading to symptoms like shortness of breath, fatigue, and fluid retention. Heart failures happen from a variety of reason such as coronary disease, diabetes, obesity etc. In this study we try to determine the importance of various parameter on the survival of patient having heart failure. The event we analyse is then the death of the patient.

```r
library(survival)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
```

```
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(survminer)
```

```
## Loading required package: ggpubr
##
## Attaching package: 'survminer'
##
## The following object is masked from 'package:survival':
##
##     myeloma
```

Introduction

The individuals of the data were patients admitted to Institute of Cardiology and Allied hospital Faisalabad-Pakistan during April-December (2015). From the 299 patients of the dataset, 105 are women and are 194 men. They are between 40 and 95 years. All have left ventricular systolic dysfunction, belonging to New York Heart Association (NYHA) class III and IV. Class III means patients have marked limitations of physical activity. They are comfortable at rest but experience symptoms with less than ordinary physical activity. Class IV means patients are unable to carry out any physical activity without discomfort. They may have symptoms even at rest and are often bedridden. From the 299 patients of the dataset, 105 are women and are 194 men. They are between 40 and 95 years.

data acquisition preparation and investigation.

data dictionary

The dataset has 13 features: Age, Anemia, High Blood Pressure, Creatinine phosphokinase, Diabetes, Ejection Fraction, Sex, Platelets, Serum Creatinine, Serum Sodium, Smoking, Time and Death Event. We explain some of the non-evident features:
- anemia : lower than normal haemoglobin concentration in blood - Creatinine phosphokinase : an enzyme notably found in the heart. Can leak in the blood in case of heart damage. - serum creatinine : a waste formed by the functioning of muscle. It is present in the blood and eliminated by the kidney through urine. As it is a serum creatinine the amount is not error induced by the taking of supplement creatine
- ejection fraction : percentage of blood pump out of the left ventricle with each contraction. If the EF is less than 40%, it indicates an heart failure or cardiomyopathy. - platelets : the normal amount of platelets ranges between 150,000 to 450,000 per muL of blood.
- serum sodium : sodium amount in blood is a well known indicator of heart failure. Is normal value ranges between 135-145 milli equivalents per liter. The presence of time and death event make this dataset perfectly adapted for a survival analysis. he unit of time is day. As all the patient didn't die, the dataset has right censored data.

```r
data <- read_csv('data/heart_failure_clinical_records_dataset.csv')
```

```
## Rows: 299 Columns: 13
## -- Column specification --------------------------------------------------------
```

```
## Delimiter: ","
## dbl (13): age, anaemia, creatinine_phosphokinase, diabetes, ejection_fractio...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(data)
```

```
## # A tibble: 6 x 13
##     age anaemia creatini~1 diabe~2 eject~3 high_~4 plate~5 serum~6 serum~7   sex
##   <dbl>   <dbl>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>
## 1    75       0        582       0      20       1 265000      1.9     130     1
## 2    55       0       7861       0      38       0 263358.     1.1     136     1
## 3    65       0        146       0      20       0 162000      1.3     129     1
## 4    50       1        111       0      20       0 210000      1.9     137     1
## 5    65       1        160       1      20       0 327000      2.7     116     0
## 6    90       1         47       0      40       1 204000      2.1     132     1
## # ... with 3 more variables: smoking <dbl>, time <dbl>, DEATH_EVENT <dbl>, and
## #   abbreviated variable names 1: creatinine_phosphokinase, 2: diabetes,
## #   3: ejection_fraction, 4: high_blood_pressure, 5: platelets,
## #   6: serum_creatinine, 7: serum_sodium
```

```r
#summary(data)
```

```r
data$sex <- factor(data$sex, labels= c("female", "male"))
```

```r
data$anaemia <- factor(data$anaemia)
```

```r
data$diabetes <- factor(data$diabetes)
data$high_blood_pressure <- factor(data$high_blood_pressure)
data$smoking <- factor(data$smoking)
```

```r
# Define the breakpoints for the three levels
breakpoints <- c(-Inf, 30, 45, Inf)
# Divide EF into three levels
data$EF_levels <- cut(data$ejection_fraction, breaks = breakpoints, labels = c("EF <= 30", "30 < EF <= 4
# avoid special character difficult for Latex here , or which needs much attention plz.
```

univariate analysis : group comparison

overall Kaplan-Meyer estimator

We estimate the survival probability with the Kaplan-meier estimator.

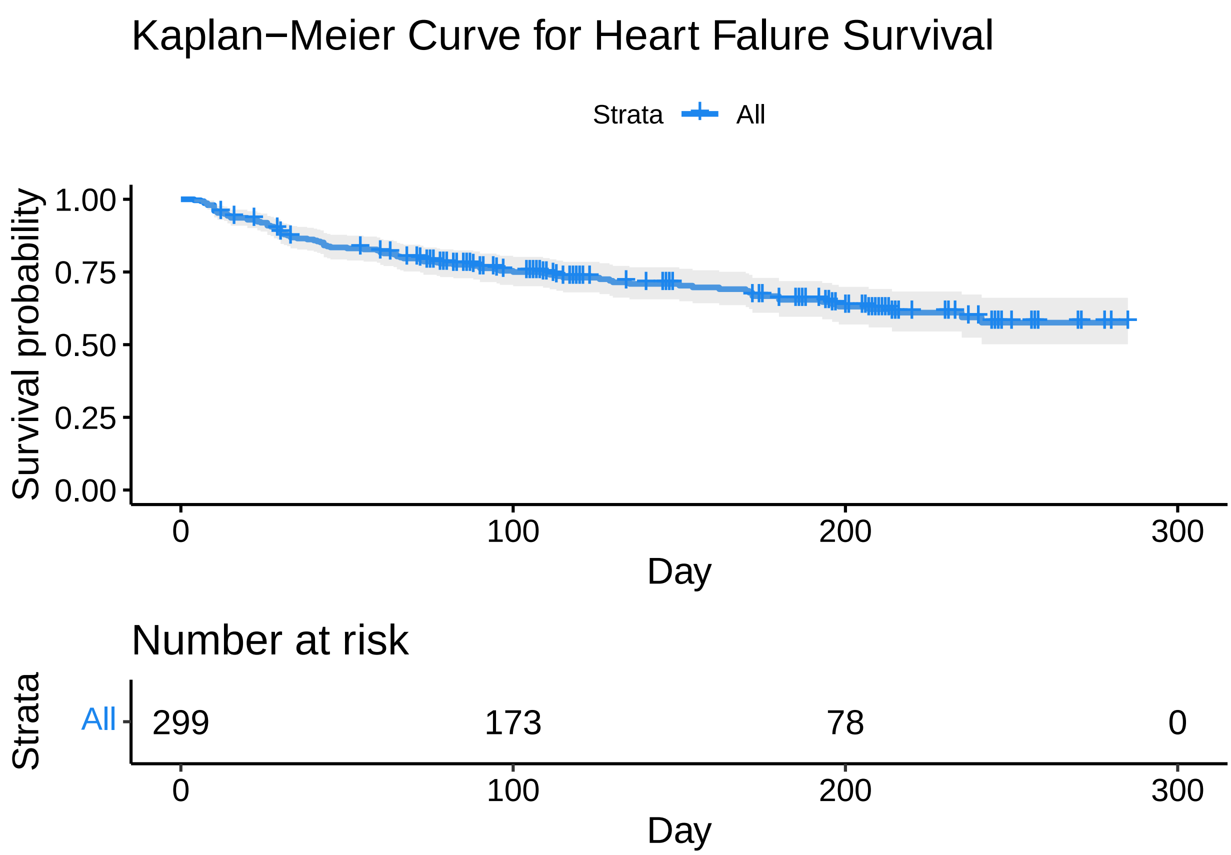
```r
fit.KM <- survfit(Surv(time,DEATH_EVENT) ~ 1, data = data)
fit.KM
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ 1, data = data)
##
##        n events median 0.95LCL 0.95UCL
## [1,] 299     96     NA      NA      NA
```

96 (32%) patients died due to the Cardiovascular Heart Disease (CHD). The median, 0.95LCL and 0.95UCl are NA because too many data are right censored. We need to go deeper in the analysis.

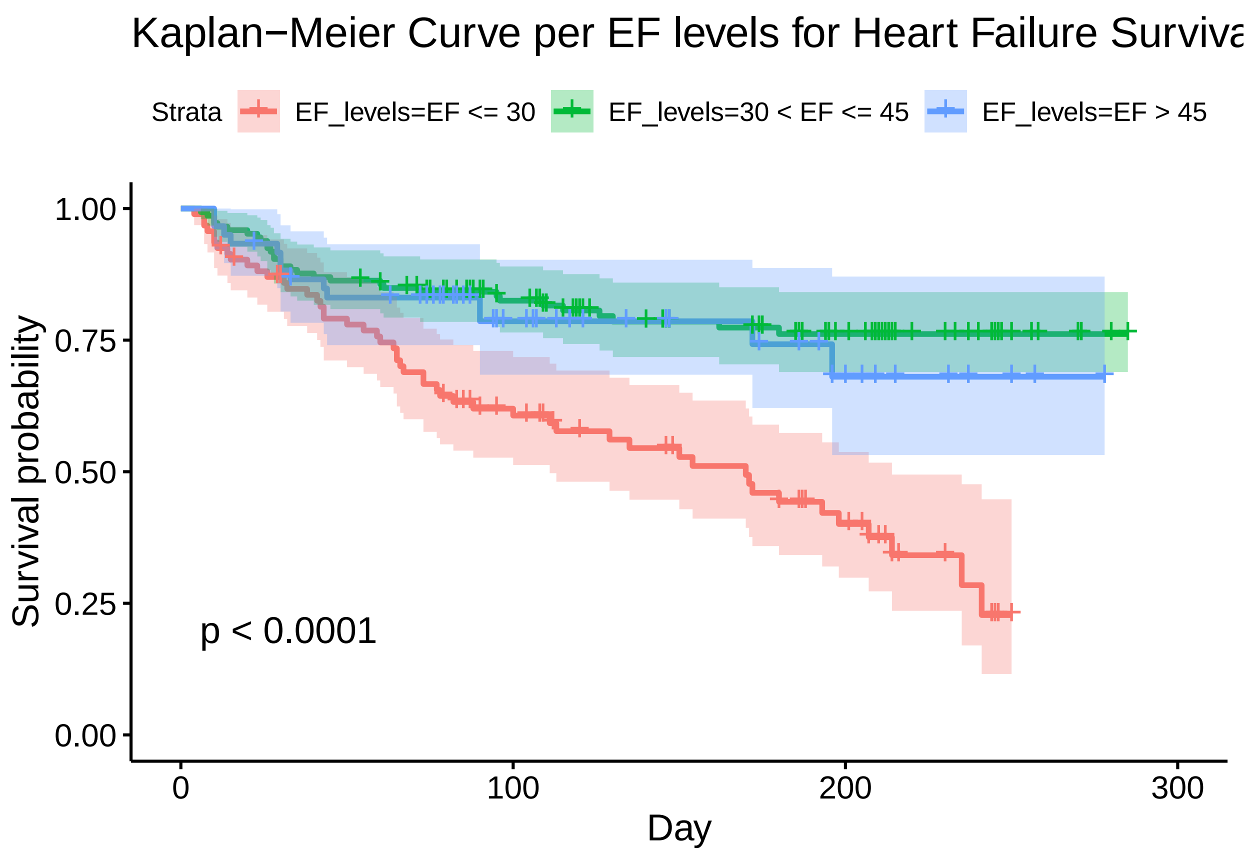
```
ggsurvplot(fit.KM,
           conf.int=TRUE,
           pval=TRUE,
           risk.table=TRUE,

           palette=c("dodgerblue2", "orchid2"),
           title="Kaplan-Meier Curve for Heart Falure Survival", xlab = "Day",
           risk.table.height=.30, data=data)
```

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

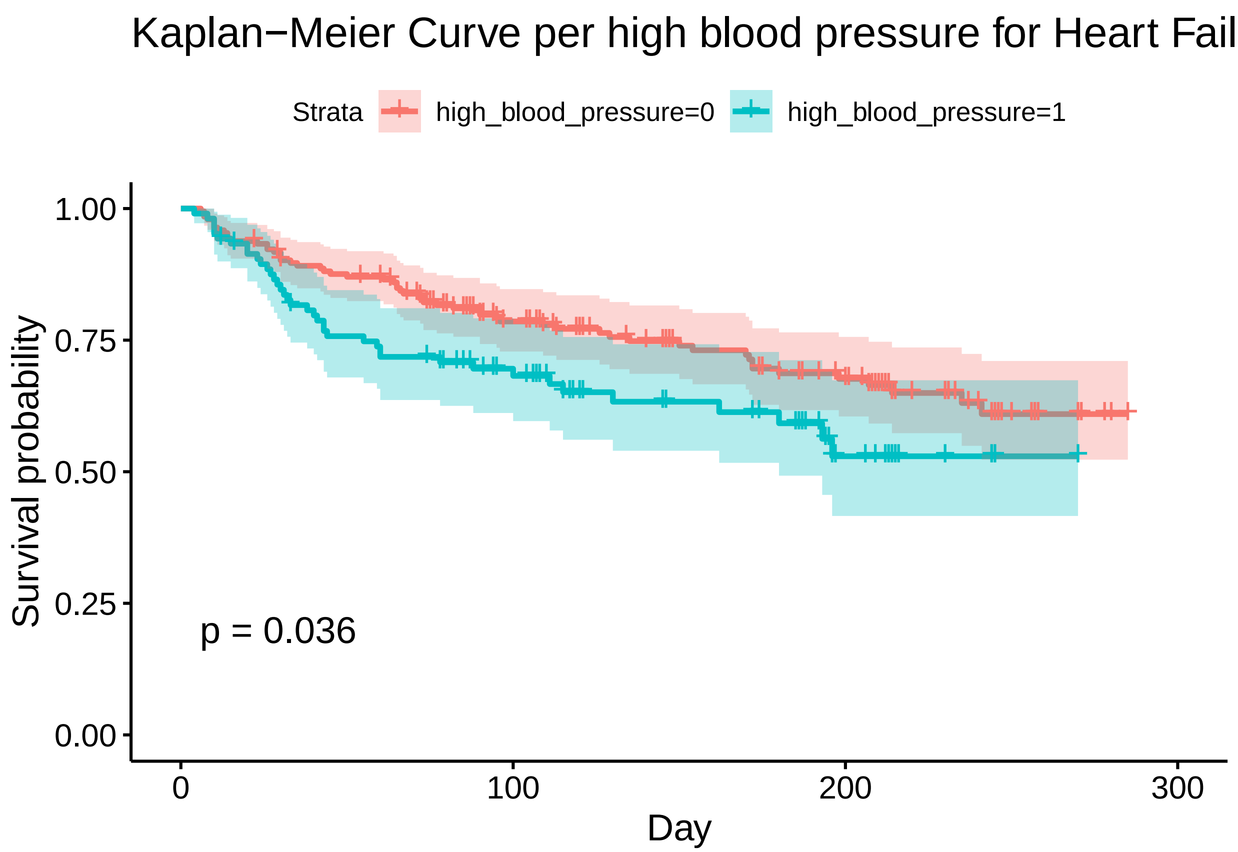


As the EF level and the high tension are the directly heart related covariates, we use them for the Kaplan Meier survival estimate

```
fit.KM1 <- survfit(Surv(time,DEATH_EVENT) ~ EF_levels, data = data)
ggsurvplot(fit.KM1, conf.int=TRUE, pval=TRUE,
           title="Kaplan-Meier Curve per EF levels for Heart Failure Survival", xlab = "Day",
           risk.table.height=.30, data=data)
```



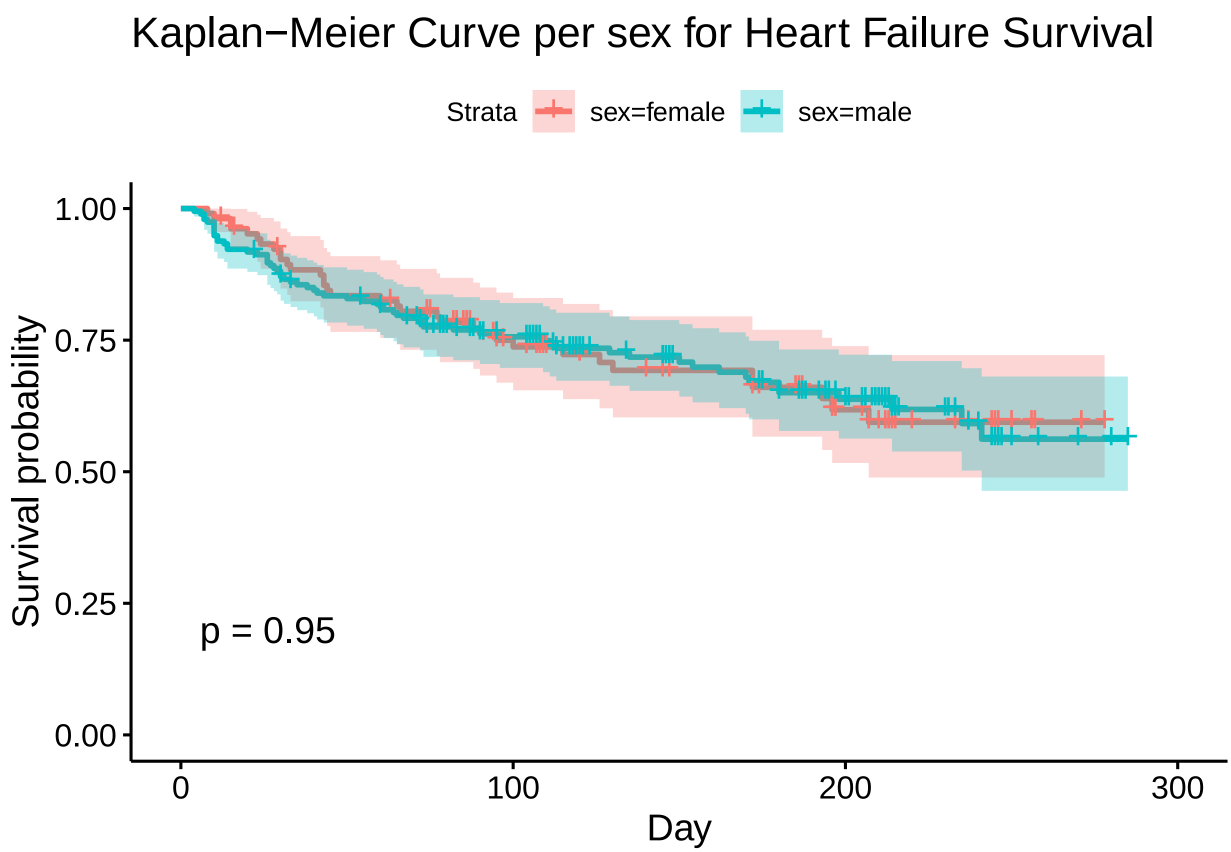

```
fit.KM2 <- survfit(Surv(time,DEATH_EVENT) ~ high_blood_pressure, data = data)
ggsurvplot(fit.KM2, conf.int=TRUE, pval=TRUE,
           title="Kaplan-Meier Curve per high blood pressure for Heart Failure Survival", xlab = "Day",
           risk.table.height=.30, data=data)
```



The EF levels are indeed heavily correlated to the death for patient with heart failure as shown bby the plot and the p-value. The high pressure is less correlated.

An other seemingly obvious covariates is the sex :

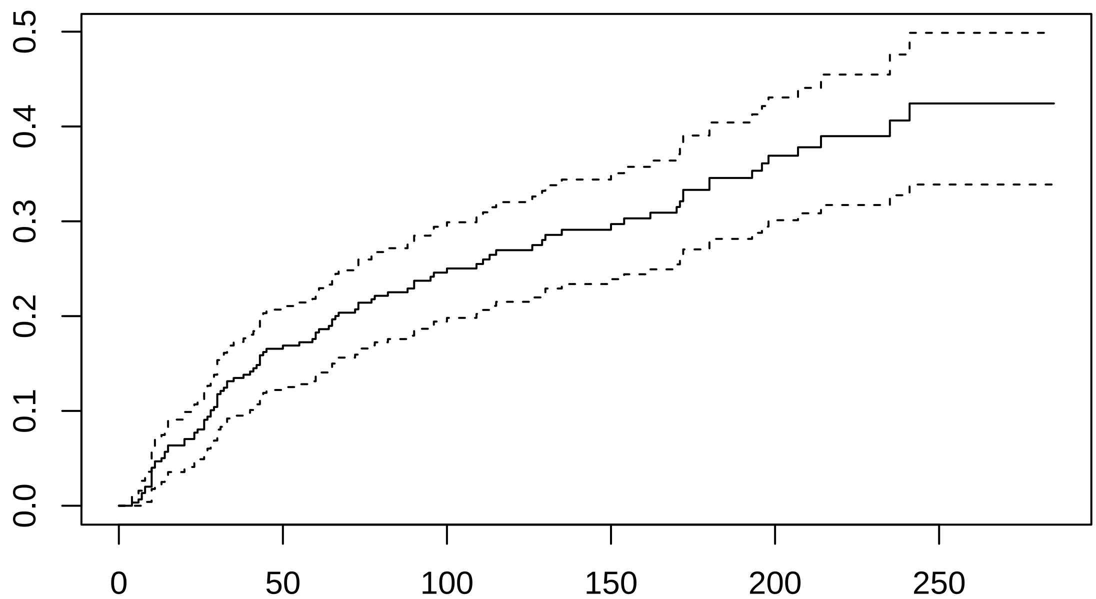
```
fit.KM3 <- survfit(Surv(time,DEATH_EVENT) ~ sex, data = data)
ggsurvplot(fit.KM3, conf.int=TRUE, pval=TRUE,
           title="Kaplan-Meier Curve per sex for Heart Failure Survival", xlab = "Day",
           risk.table.height=.30, data=data)
```



The sex is not correlated to death

CDF:

```
plot(fit.KM, fun = "F")
```

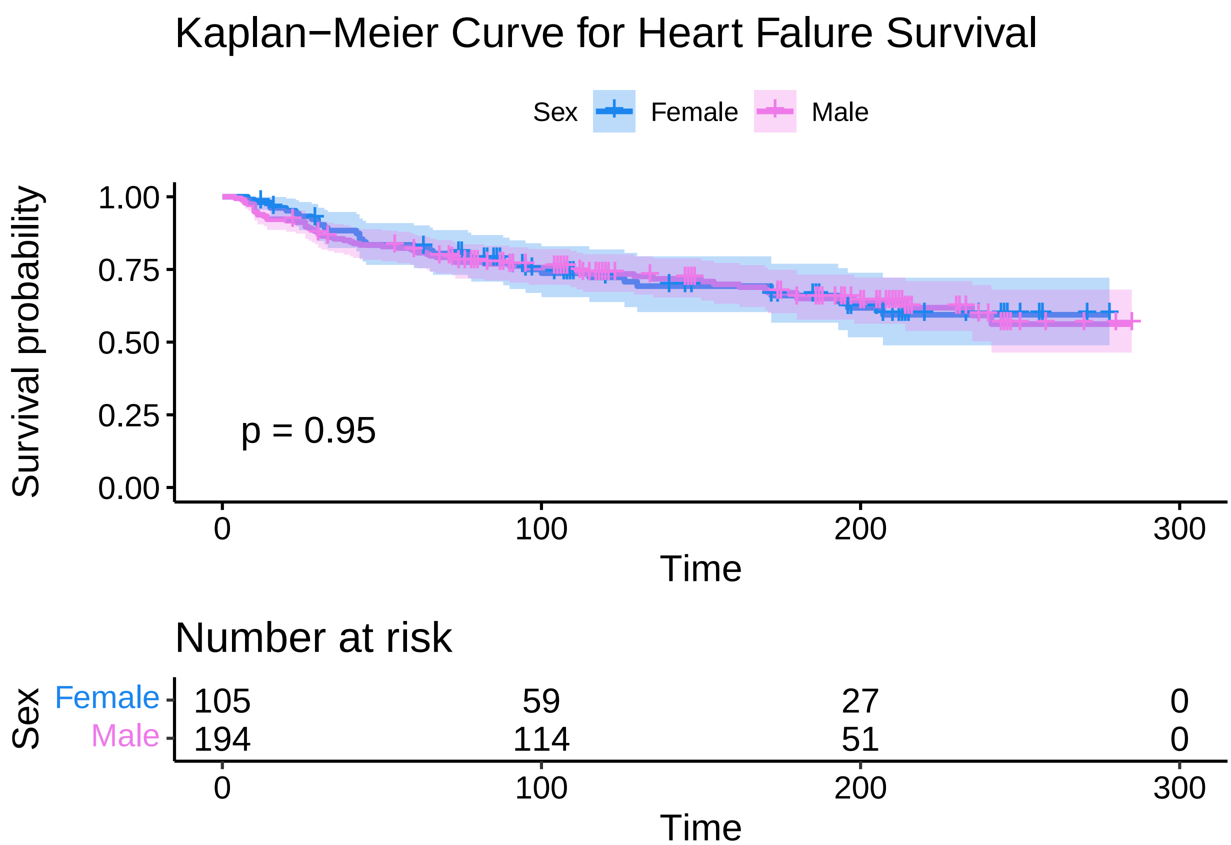


Male vs. Female

```
sfit <- survfit(Surv(time, DEATH_EVENT) ~ sex , data = data)
sfit
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ sex, data = data)
##
##               n events median 0.95LCL 0.95UCL
## sex=female 105     34     NA     207      NA
## sex=male   194     62     NA     241      NA
```

```
ggsurvplot(sfit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
           legend.labs=c("Female", "Male"), legend.title="Sex",
           palette=c("dodgerblue2", "orchid2"),
           title="Kaplan-Meier Curve for Heart Falure Survival",
           risk.table.height=.30, data=data)
```




```
fit.logrank <- survdiff(Surv(time, DEATH_EVENT) ~ sex, data = data)
fit.logrank
```

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

from survival curves of male and female and using the long rank test we can concluded that the sex has no significant impact

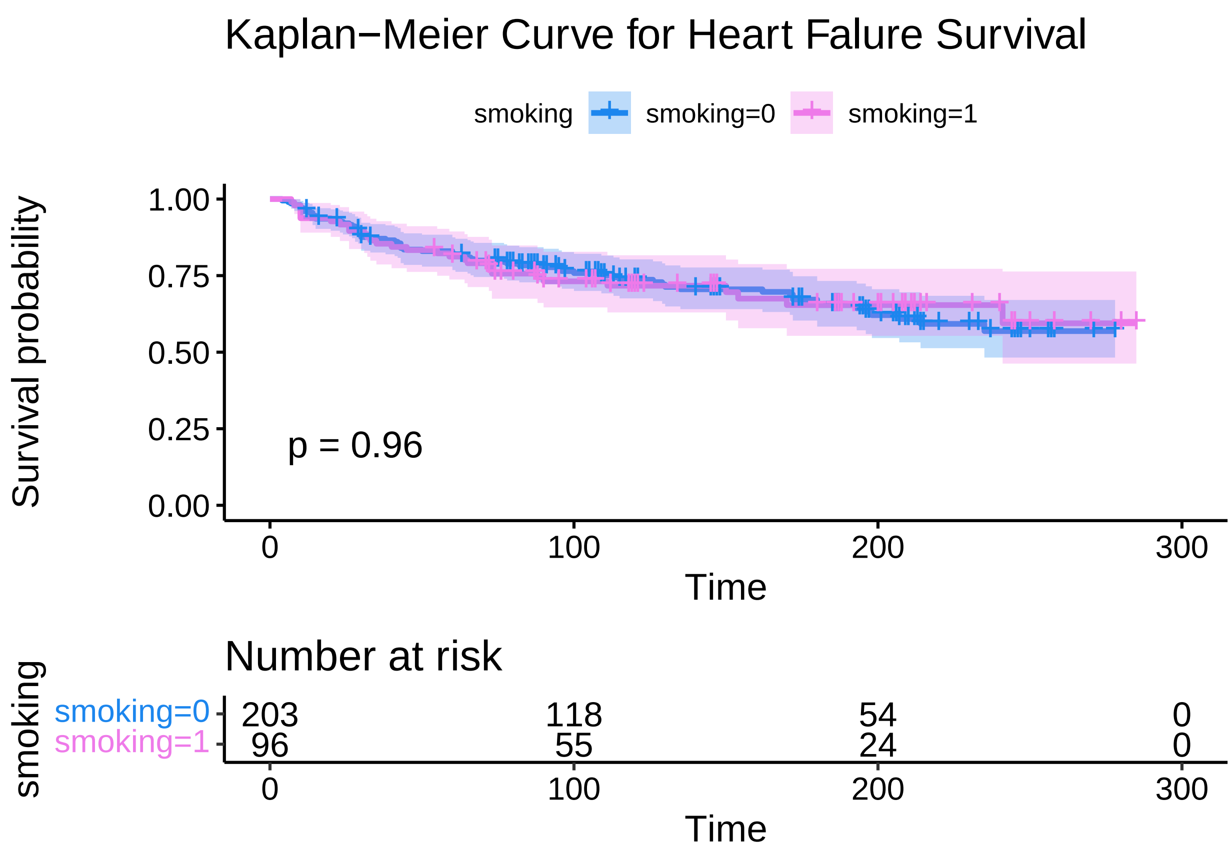
Smoking

```
sfit <- survfit(Surv(time, DEATH_EVENT) ~ smoking , data = data)
sfit
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ smoking, data = data)
##
##               n events median 0.95LCL 0.95UCL
```

```
## smoking=0 203     66     NA     235      NA
## smoking=1  96     30     NA     241      NA
```

```
ggsurvplot(sfit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
            legend.title="smoking",
           palette=c("dodgerblue2", "orchid2"),
           title="Kaplan-Meier Curve for Heart Falure Survival",
           risk.table.height=.28, data=data)
```

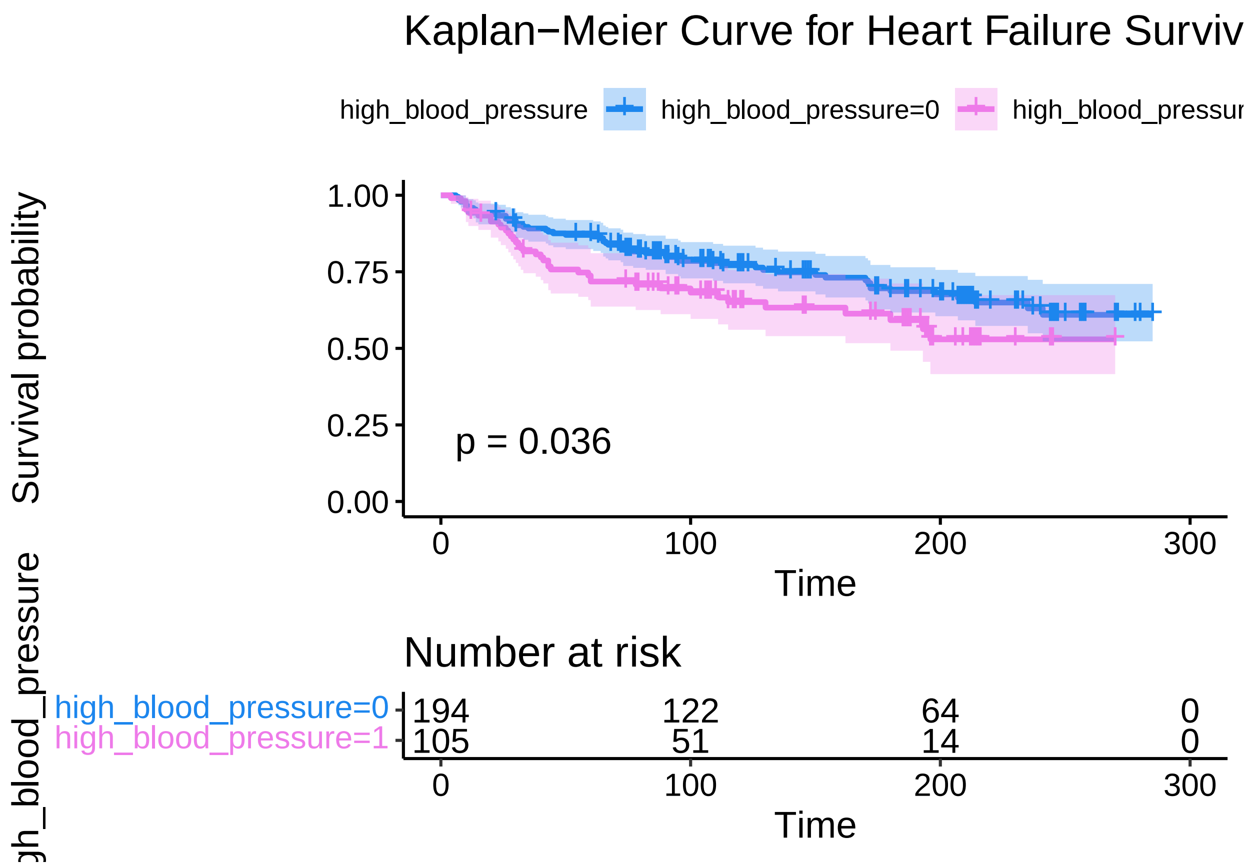


blood pressure

```
sfit <- survfit(Surv(time, DEATH_EVENT) ~ high_blood_pressure , data = data)
sfit
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ high_blood_pressure,
##     data = data)
##
##                         n events median 0.95LCL 0.95UCL
## high_blood_pressure=0 194     57     NA      NA      NA
## high_blood_pressure=1 105     39     NA     180      NA
```

```r
ggsurvplot(sfit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
            legend.title="high_blood_pressure",
           palette=c("dodgerblue2", "orchid2"),
           title="Kaplan-Meier Curve for Heart Failure Survival",
           risk.table.height=.28, data=data)
```



The logrank test

```r
fit.logrank <- survdiff(Surv(time, DEATH_EVENT) ~ high_blood_pressure, data = data)
fit.logrank
```

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

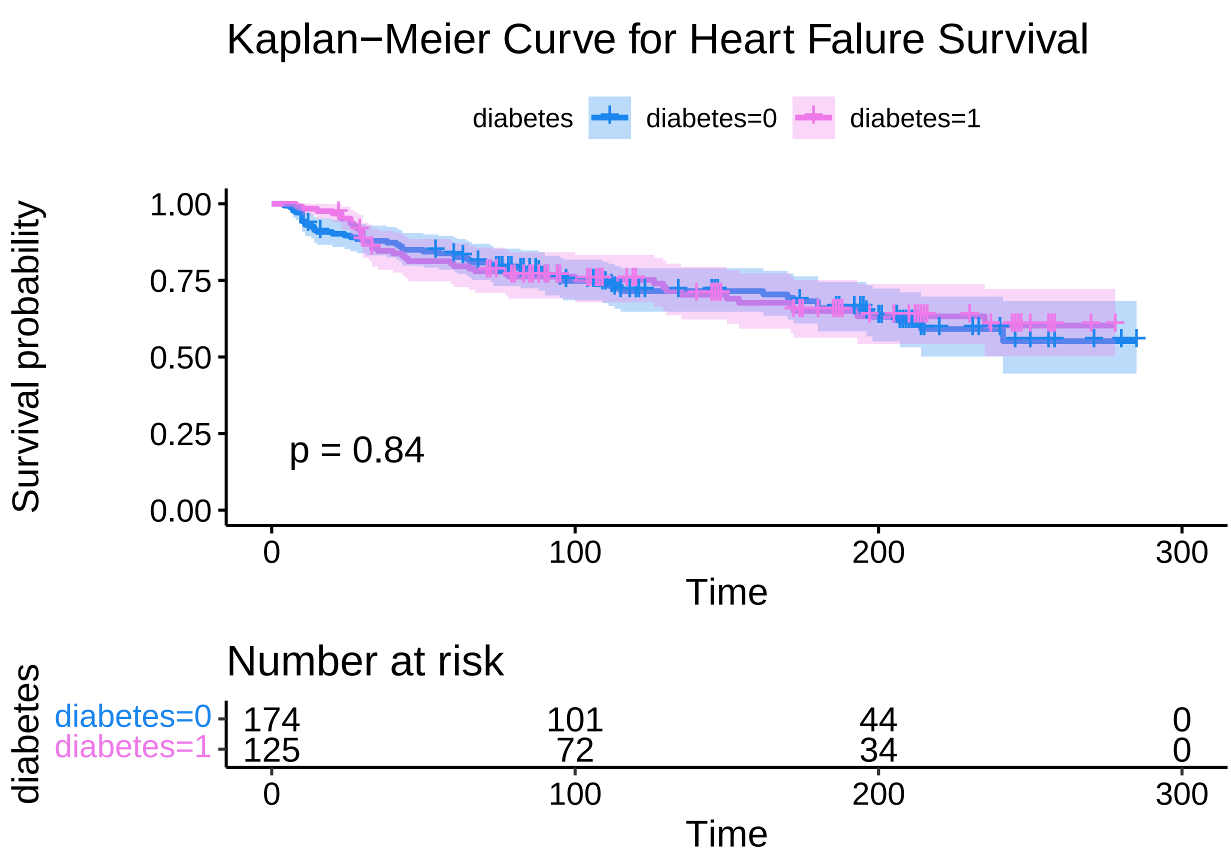
he p-value is 0.04, which is less than 0.05. Therefore, you can conclude that there is evidence of a statistically significant difference in survival between the two groups based on the presence or absence of high blood pressure

diabetes

```
sfit <- survfit(Surv(time, DEATH_EVENT) ~ diabetes , data = data)
sfit
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ diabetes, data = data)
##
##               n events median 0.95LCL 0.95UCL
## diabetes=0 174     56     NA     241      NA
## diabetes=1 125     40     NA      NA      NA
```

```
ggsurvplot(sfit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
            legend.title="diabetes",
           palette=c("dodgerblue2", "orchid2"),
           title="Kaplan-Meier Curve for Heart Falure Survival",
           risk.table.height=.28, data=data)
```



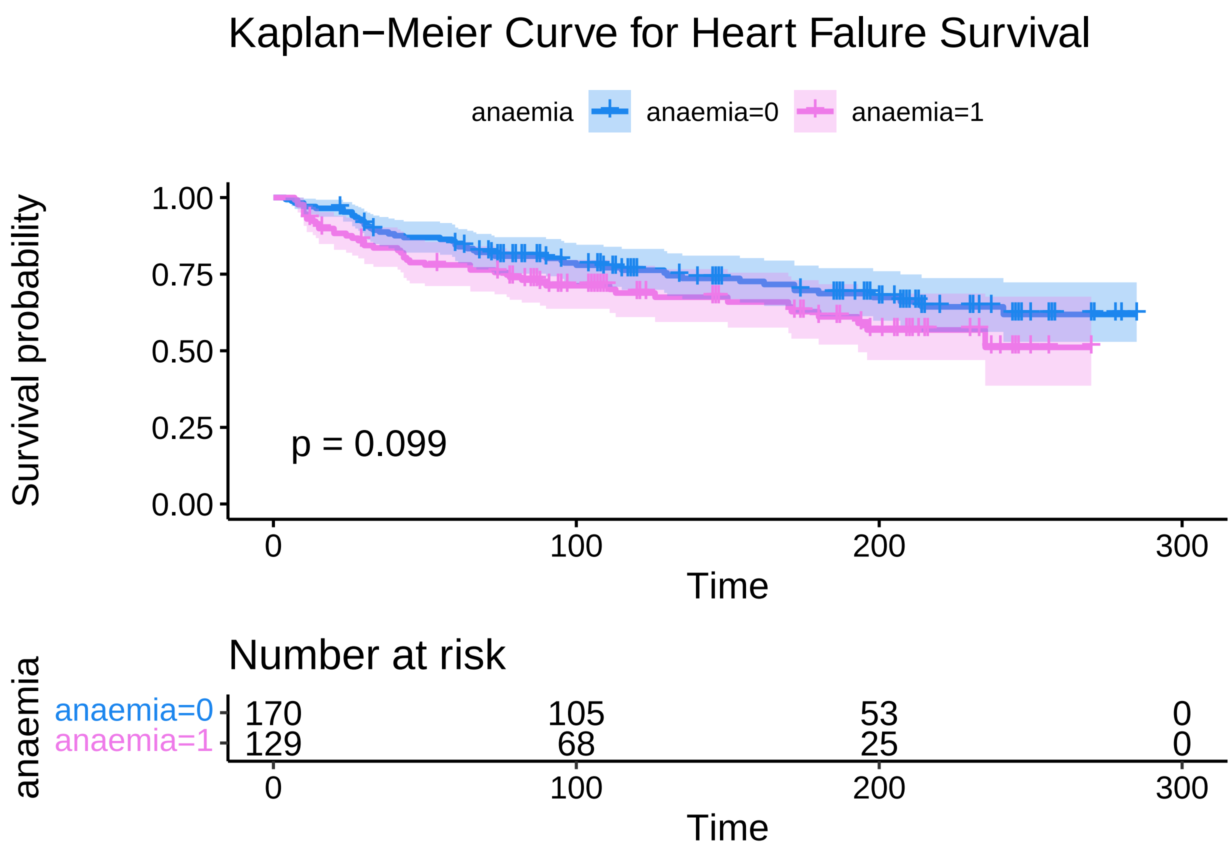
anaemia

```
sfit <- survfit(Surv(time, DEATH_EVENT) ~ anaemia , data = data)
sfit
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ anaemia, data = data)
##
```

```
##              n events median 0.95LCL 0.95UCL
## anaemia=0 170     50     NA      NA      NA
## anaemia=1 129     46     NA     193      NA
```

```r
ggsurvplot(sfit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
            legend.title="anaemia",
           palette=c("dodgerblue2", "orchid2"),
           title="Kaplan-Meier Curve for Heart Falure Survival",
           risk.table.height=.28, data=data)
```



The logrank test

```r
fit.logrank <- survdiff(Surv(time, DEATH_EVENT) ~ anaemia, data = data)
fit.logrank
```

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

the impacet of age

age is continuous variable we need first to discretise it, over 60 because 60 is median

```
d_age60 <-
  data |>
  mutate(age60 = factor(age <= 60,
                        labels = c("<=60", ">60")))

table(d_age60$age60)
```

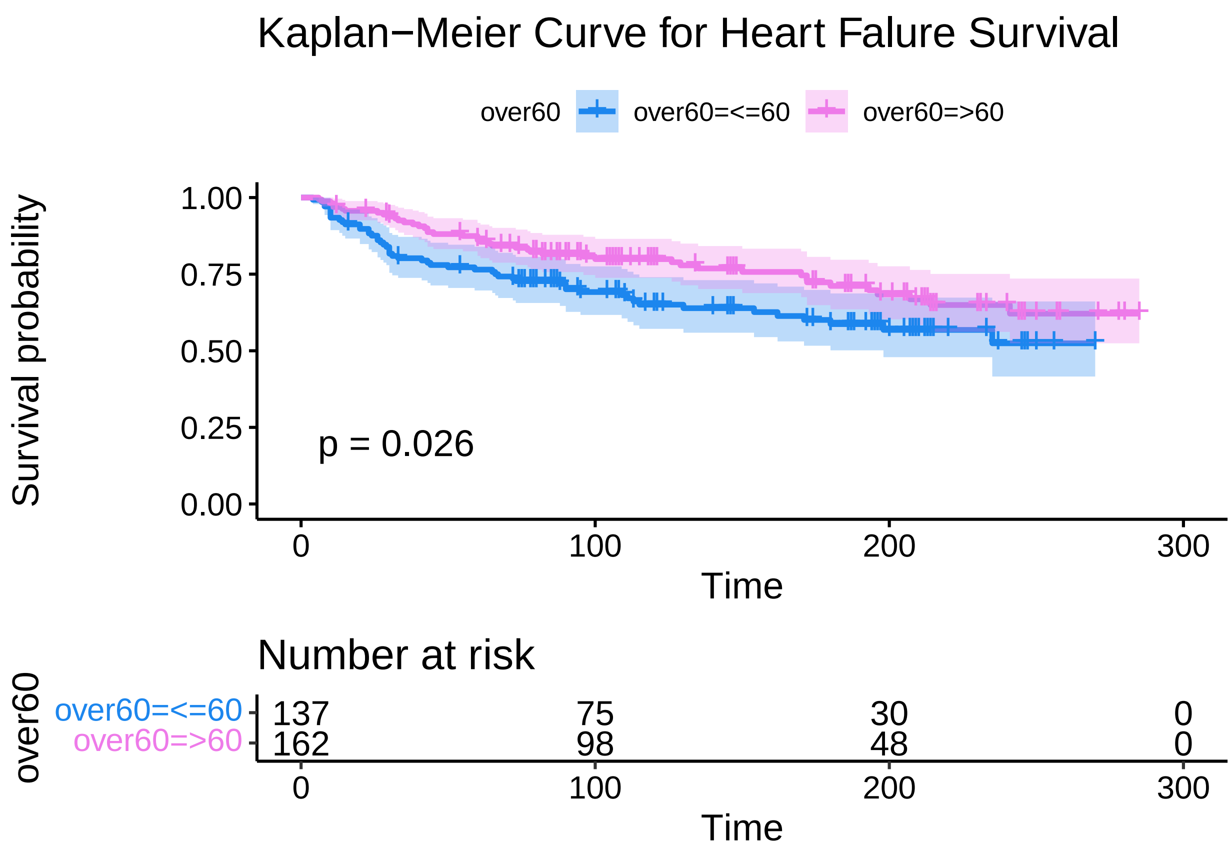
```
##
## <=60  >60
##  137  162
```

```
data$over60 = d_age60$age60
```

```
sfit <- survfit(Surv(time, DEATH_EVENT) ~ over60 , data = data)
sfit
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ over60, data = data)
##
##                n events median 0.95LCL 0.95UCL
## over60=<=60 137     52     NA     198      NA
## over60=>60  162     44     NA      NA      NA
```

```
ggsurvplot(sfit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
            legend.title="over60",
           palette=c("dodgerblue2", "orchid2"),
           title="Kaplan-Meier Curve for Heart Falure Survival",
           risk.table.height=.28, data=data)
```



The logrank test

```r
fit.logrank <- survdiff(Surv(time, DEATH_EVENT) ~ over60, data = data)
fit.logrank
```

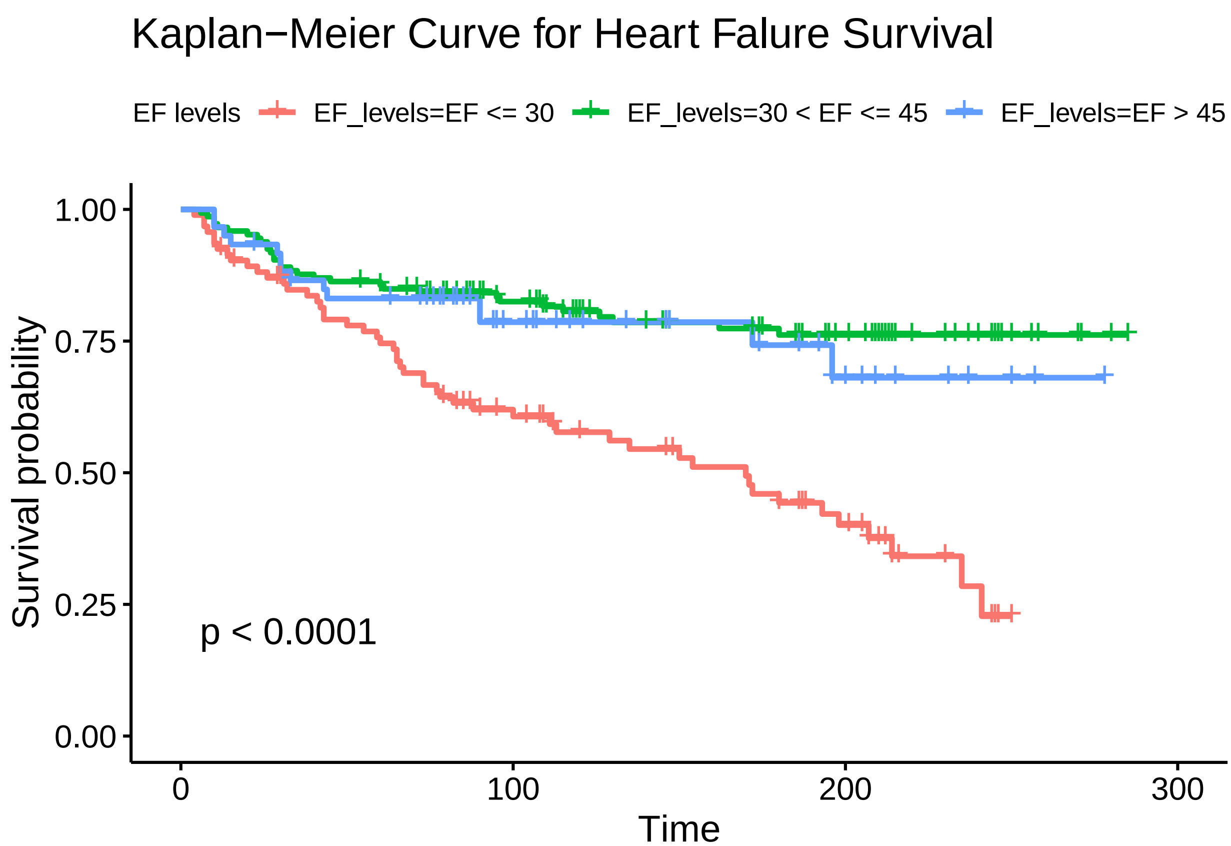
```
## Call:
## survdiff(formula = Surv(time, DEATH_EVENT) ~ over60, data = data)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## over60=<=60 137       52     41.2      2.81      4.95
## over60=>60  162       44     54.8      2.11      4.95
##
##  Chisq= 5  on 1 degrees of freedom, p= 0.03
```

Ejection Fraction

```r
sfit <- survfit(Surv(time, DEATH_EVENT) ~ EF_levels , data = data)
sfit
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ EF_levels, data = data)
##
##                            n events median 0.95LCL 0.95UCL
## EF_levels=EF <= 30        93     51    170     111     214
## EF_levels=30 < EF <= 45 146     31     NA      NA      NA
## EF_levels=EF > 45         60     14     NA      NA      NA
```

```r
ggsurvplot(sfit, pval = TRUE , data = data,
           legend.title="EF levels",
           title="Kaplan-Meier Curve for Heart Falure Survival",
           )
```



The logrank test

```r
fit.logrank <- survdiff(Surv(time, DEATH_EVENT) ~ EF_levels, data = data)
fit.logrank
```

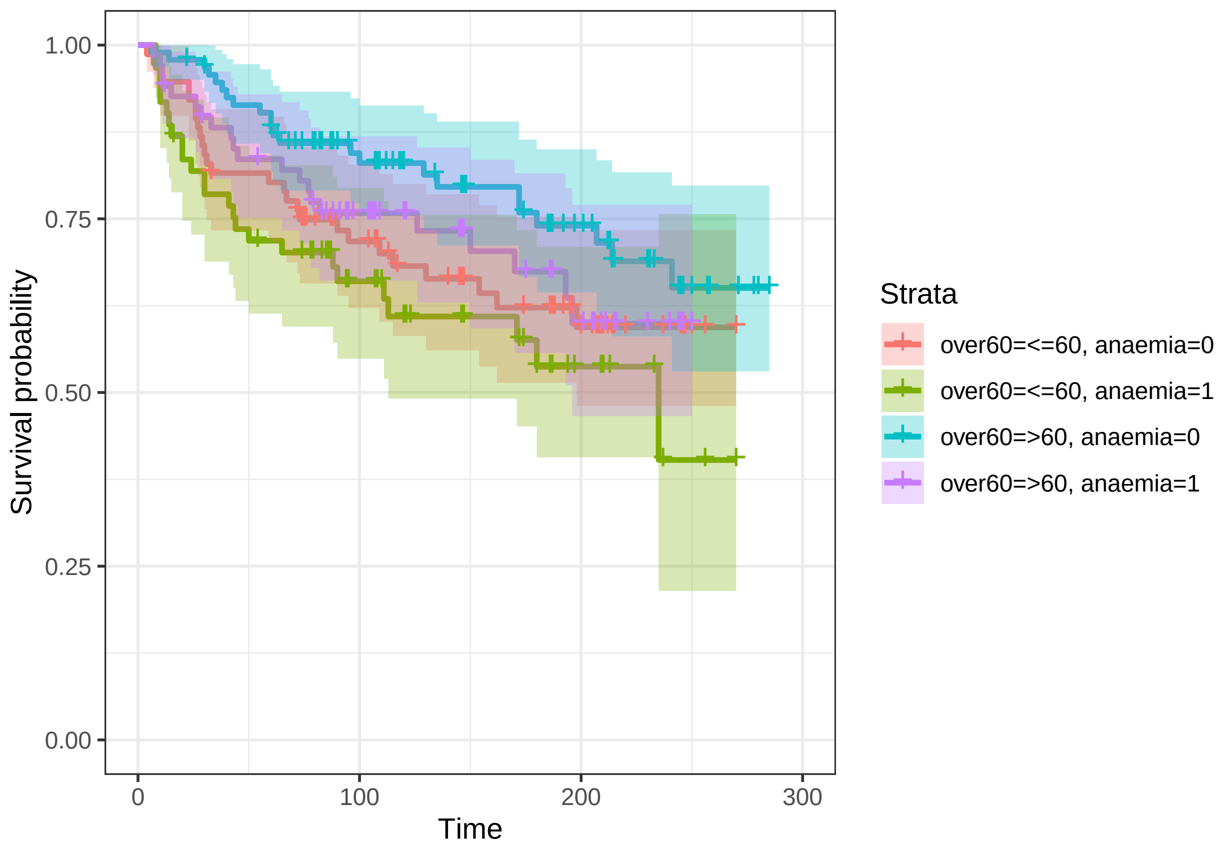
```
## Call:
## survdiff(formula = Surv(time, DEATH_EVENT) ~ EF_levels, data = data)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## EF_levels=EF <= 30        93       51     26.8     21.93     30.63
## EF_levels=30 < EF <= 45 146       31     51.2      7.97     17.32
## EF_levels=EF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

Since the p-value is less than 0.05 we reject the null hypothesis, this means there is a statistically significant difference in survival between the three groups.
age and anaemia

```r
sfit <- survfit(Surv(time, DEATH_EVENT) ~ over60 + anaemia , data = data)
sfit
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ over60 + anaemia,
##     data = data)
##
##                            n events median 0.95LCL 0.95UCL
## over60=<=60, anaemia=0 76     27     NA     198      NA
## over60=<=60, anaemia=1 61     25    235     113      NA
## over60=>60, anaemia=0  94     23     NA      NA      NA
## over60=>60, anaemia=1  68     21     NA     196      NA
```

```r
ggsurv <- ggsurvplot(sfit,  conf.int = TRUE,data=data,
                     ggtheme = theme_bw())

ggsurv$plot +theme_bw() +
  theme (legend.position = "right")
```



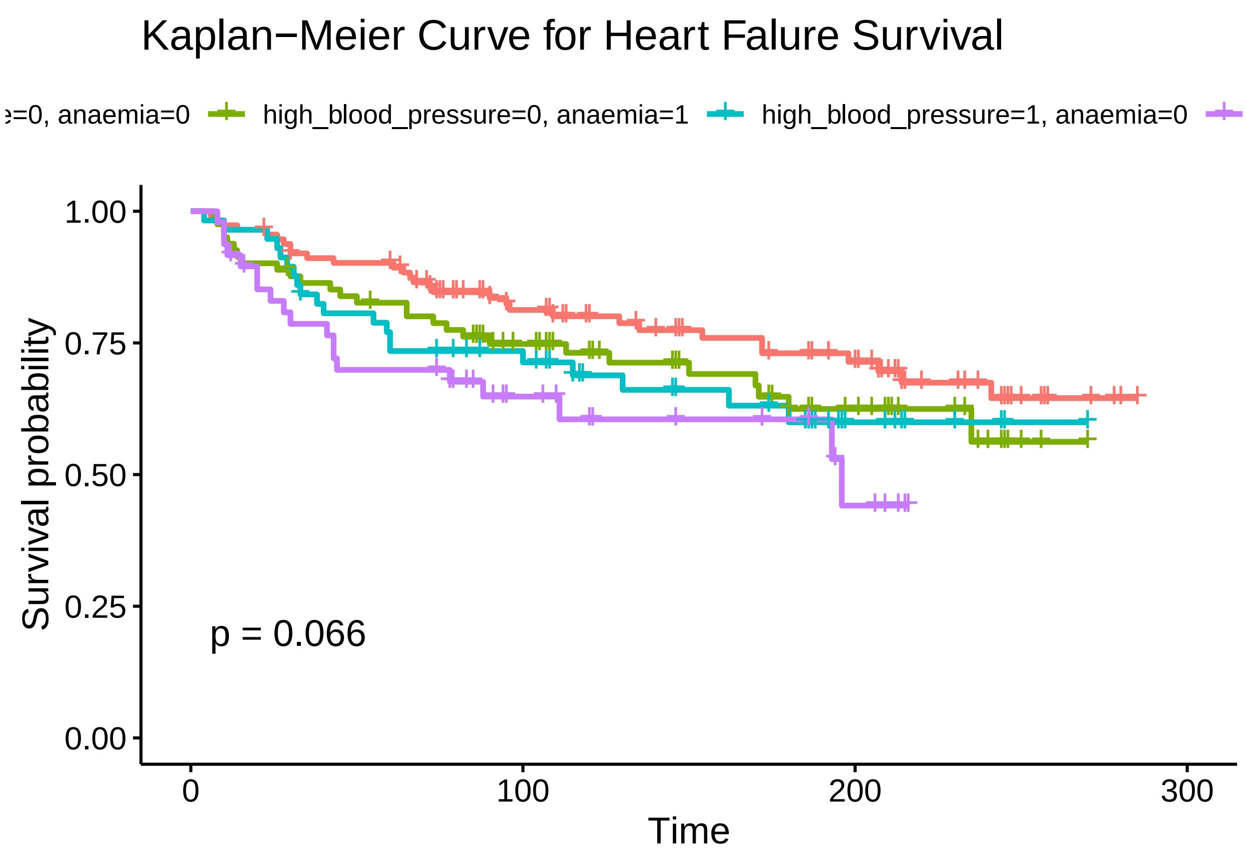
high_blood_pressure and anaemia

```r
sfit <- survfit(Surv(time, DEATH_EVENT) ~ high_blood_pressure + anaemia , data = data)
sfit
```

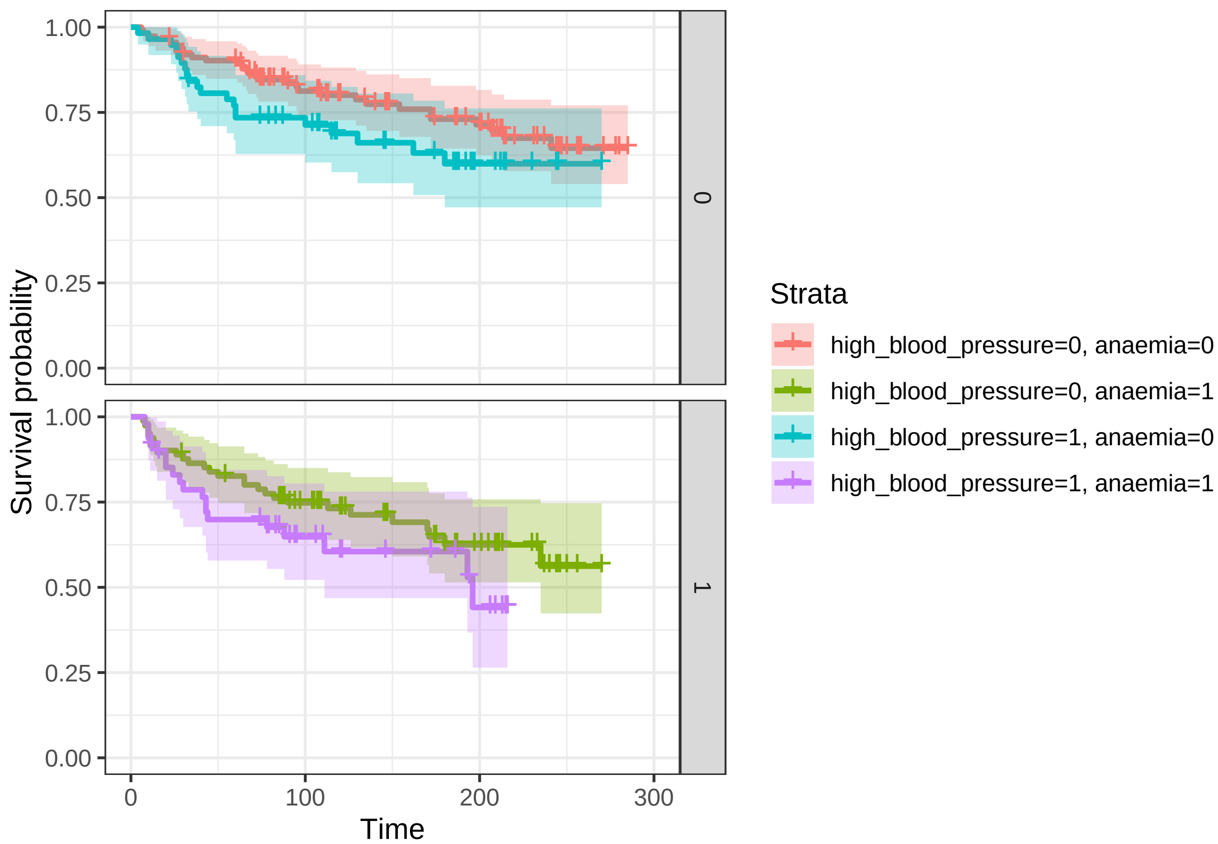
```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ high_blood_pressure +
##     anaemia, data = data)
```

```
##
##                                          n events median 0.95LCL 0.95UCL
## high_blood_pressure=0, anaemia=0 113     30     NA      NA      NA
## high_blood_pressure=0, anaemia=1  81     27     NA     235      NA
## high_blood_pressure=1, anaemia=0  57     20     NA     180      NA
## high_blood_pressure=1, anaemia=1  48     19    196     111      NA
```

```r
ggsurvplot(sfit, pval = TRUE , data = data,
           legend.title="",
           title="Kaplan-Meier Curve for Heart Falure Survival",
           )
```




```r
ggsurv <- ggsurvplot(sfit,  conf.int = TRUE,data=data,
                     ggtheme = theme_bw())

ggsurv$plot +theme_bw() +
  theme (legend.position = "right")+
  facet_grid('anaemia')
```



Cox Proportional Hazards Model

```r
head(data)
```

```
## # A tibble: 6 x 15
##     age anaemia creatini~1 diabe~2 eject~3 high_~4 plate~5 serum~6 serum~7 sex
##   <dbl> <fct>        <dbl> <fct>     <dbl> <fct>     <dbl>   <dbl>   <dbl> <fct>
## 1    75 0              582 0            20 1       265000      1.9     130 male
## 2    55 0             7861 0            38 0       263358.     1.1     136 male
## 3    65 0              146 0            20 0       162000      1.3     129 male
## 4    50 1              111 0            20 0       210000      1.9     137 male
## 5    65 1              160 1            20 0       327000      2.7     116 fema~
## 6    90 1               47 0            40 1       204000      2.1     132 male
## # ... with 5 more variables: smoking <fct>, time <dbl>, DEATH_EVENT <dbl>,
## #   EF_levels <fct>, over60 <fct>, and abbreviated variable names
## #   1: creatinine_phosphokinase, 2: diabetes, 3: ejection_fraction,
## #   4: high_blood_pressure, 5: platelets, 6: serum_creatinine, 7: serum_sodium
```

```r
fit_cph <- coxph(Surv(time, DEATH_EVENT) ~ high_blood_pressure + anaemia + EF_levels +over60, data = da
fit_cph
```

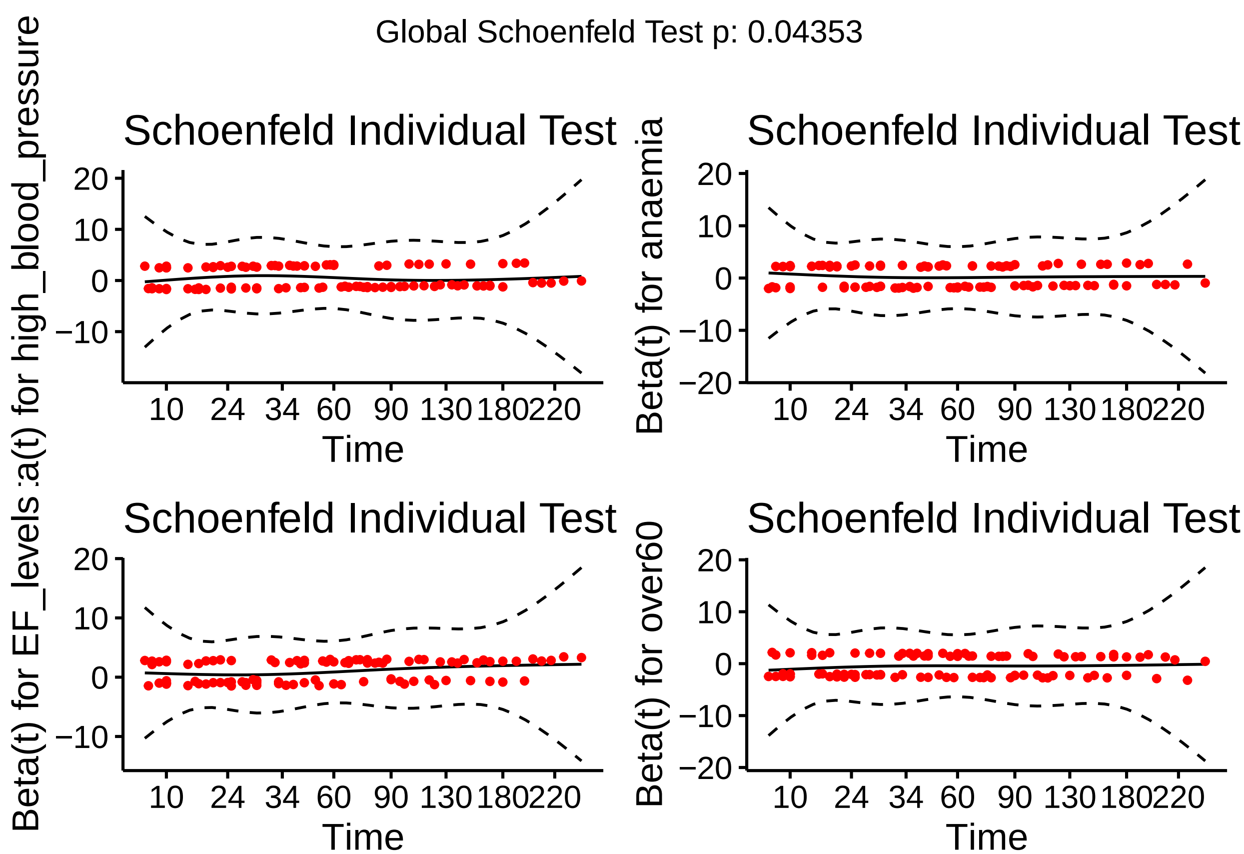
```
## Call:
```

```
## coxph(formula = Surv(time, DEATH_EVENT) ~ high_blood_pressure +
##     anaemia + EF_levels + over60, data = data)
##
##                             coef exp(coef) se(coef)      z        p
## high_blood_pressure1      0.4290    1.5358   0.2111  2.032 0.042152
## anaemia1                  0.2856    1.3305   0.2059  1.387 0.165553
## EF_levels30 < EF <= 45 -1.1775    0.3080   0.2304 -5.111 3.21e-07
## EF_levelsEF > 45         -1.0567    0.3476   0.3057 -3.456 0.000548
## over60>60                -0.5511    0.5763   0.2074 -2.658 0.007866
##
## Likelihood ratio test=40.89  on 5 df, p=9.886e-08
## n= 299, number of events= 96
```

```
test.ph <- cox.zph(fit_cph)
test.ph
```

```
##                      chisq df      p
## high_blood_pressure  0.177  1 0.6739
## anaemia              0.243  1 0.6222
## EF_levels            9.729  2 0.0077
## over60               2.067  1 0.1505
## GLOBAL              11.428  5 0.0435
```

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



[DATA Source] (Ahmad, Tanvir; Munir, Assia; Bhatti, Sajjad Haider; Aftab, Muhammad; Ali Raza, Muhammad (2017). DATA_MINIMAL.. PLOS ONE. Dataset. https://doi.org/10.1371/journal.pone.0181001.s001)